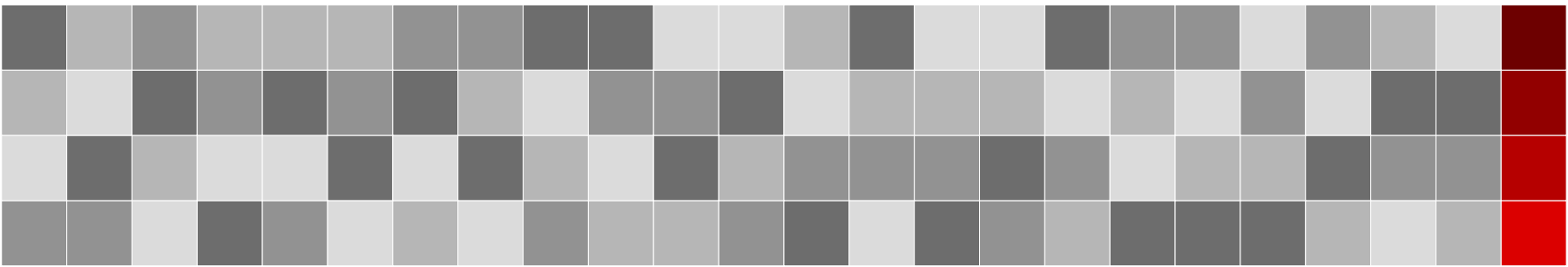
Principles of Statistical Analysis



Statistical Inference based on Randomization

Ery Arias-Castro

Publication

The present ebook format is optimized for reading on a screen and not for printing on paper. A somewhat different version will be available in print as a volume in the Institute of Mathematical Statistics Textbooks series, published by Cambridge University Press.

Version of August 18, 2019

Dedication

I would like to dedicate this book to some professors that have, along the way, inspired and supported me in my studies and academic career, and to whom I am eternally grateful:

David L. Donoho
my doctoral thesis advisor

Persi Diaconis
my first co-author on a research article

Yves Meyer
my master's thesis advisor

Robert Azencott
my undergraduate thesis advisor

In controlled experimentation it has been found not difficult to introduce explicit and objective randomization in such a way that the tests of significance are demonstrably correct. In other cases we must still act in the faith that Nature has done the randomization for us. [...] We now recognize randomization as a postulate necessary to the validity of our conclusions, and the modern experimenter is careful to make sure that this postulate is justified.

Ronald A. Fisher
International Statistical Conferences, 1947

CONTENTS

Preface

The book is intended for the mathematically literate reader who wants to understand how to analyze data in a principled fashion. The language of mathematics allows for a more concise, and arguably clearer exposition that can go quite deep, quite quickly, and naturally accommodates an axiomatic and inductive approach to data analysis, which is the raison d'être of the book.

The compact treatment is indeed grounded in mathematical theory and concepts, and is fairly rigorous, even though measure theoretic matters are kept in the background, and most proofs are left as problems. In fact, much of the learning is accomplished through embedded problems. Some problems call for mathematical derivations, and assume a certain comfort with calculus, or even real analysis. Other problems require some basic programming on a computer. We can recommend R, although some readers might prefer other languages such as Python.

Structure The book is divided into three parts, each with multiple chapters. The introduction to probability, in Part I, stands as the mathematical foundation for statistical inference. Indeed, without a solid foundation in probability, and in particular a good understanding of

how experiments are modeled, there is no clear distinction between a descriptive and an inferential analysis. The exposition is quite standard. It starts with Kolmogorov's axioms, moves on to define random variables, which in turn allows for the introduction of some major distributions, and some basic concentration inequalities and limit theorems are presented. A construction of the Lebesgue integral is not included. The part ends with a brief discussion of some stochastic processes.

Some utilitarian aspects of probability and statistics are discussed in Part II. These include probability sampling and pseudo-random number generation — the practical side of randomness; as well as survey sampling and experimental design — the practical side of data collection.

Part III is the core of the book. It attempts to build a theory of statistical inference from first principles. The foundation is randomization, either controlled by design or assumed to be natural. In either case, randomization provides the essential randomness needed to justify probabilistic modeling. It naturally leads to conditional inference, and allows for causal inference. In this framework, permutation tests play a special, almost canonical role. Monte Carlo sampling, performed on a computer, is presented as an alternative to complex mathematical derivations, and the bootstrap is then introduced as an accommodation when the sampling distribution is not directly available and has to be estimated.

What is not here I do not find normal models to be particularly compelling: Unless there is a central limit theorem at play, there is no real reason to believe some numerical data are normally distributed. Normal models are thus only mentioned in passing. More generally, parametric models are not emphasized — except for those that arise naturally in some experiments.

The usual emphasis on parametric inference is, I find, misplaced, and also misleading, as it can be (and often is) introduced independently of how the data were gathered, thus creating a chasm that separates the design of experiments and the analysis of the resulting data. Bayesian modeling is, consequently, not covered beyond some basic definitions in the context of average risk optimality. Linear models and time series are not discussed in any detail. As is typically the case for an introductory book, especially of this length and at this level, there is only a hint of abstract decision theory, and multivariate analysis is omitted entirely.

How to use this book The idea for this book arose with a dissatisfaction with how statistical analysis is typ-

ically taught at the undergraduate and master's levels, coupled with an inspiration for weaving a narrative which I find more compelling.

This narrative was formed over years of teaching statistics at the University of California, San Diego, and although the book has not been used as a textbook for any of the courses I have taught, it should be useful as such. It would appear to contain enough material for a whole year, in particular if starting with probability theory, and is meant to provide a solid foundation in statistical inference.

The student is invited to read the book in the order in which the material is presented, working on the problems as they come, and saving those parts that seem harder for later. The book is otherwise better suited for independent study under the guidance of an experienced instructor.

The instructor is encouraged to give the students additional problems to work on beyond those given here, and also reading assignments, such as research articles in the sciences that make use of concepts and tools introduced in the book.

Intention The book introduces some essential concepts that I would want a student graduating with a bachelor's or master's degree in statistics to have been exposed to, even if only in passing.

My main hope in writing this book is that it seduces mathematically minded people into learning more about statistical analysis, at least for their personal enrichment, particularly in this age of artificial intelligence, machine learning, and more generally, data science.

Ery Arias-Castro
San Diego, Summer of 2019

Acknowledgements

Ian Abramson, Emmanuel Candès, Persi Diaconis, David Donoho, Larry Goldstein, Gábor Lugosi, Dimitris Politis, Philippe Rigollet, Jason Schweinsberg, and Nicolas Verzelen, provided feedback or pointers on particular topics. Clément Berenfeld and Zexin Pan, as well as two anonymous referees retained by Cambridge University Press, helped proofread the manuscript (all remaining errors are, of course, mine). Diana Gillooly was my main contact at Cambridge University Press, offering expert guidance through the publishing process. I am grateful for all this generous support.

I looked at a number of textbooks in Probability and Statistics, and also lecture notes. The most important ones are cited in the text, in particular to entice the reader to learn more about some topics.

I used a number of freely available resources contributed by countless people all over the world. The manuscript was typeset in LaTeX (using TeXShop, BibDesk, TeXstudio, JabRef) and the figures were produced using R via RStudio. I made extensive use of Google Scholar, Wikipedia, and some online discussion lists such as StackExchange.

About the author

Ery Arias-Castro is a professor of Mathematics at the University of California, San Diego, where he specializes in Statistics and Machine Learning. His education includes a bachelor's degree in Mathematics and a master's degree in Artificial Intelligence, both from École Normale Supérieure de Cachan, France, as well as a PhD in Statistics from Stanford University.

Part A

Elements of Probability Theory

CHAPTER 1

AXIOMS OF PROBABILITY THEORY

Probability theory is the branch of mathematics that models and studies random phenomena. Although randomness has been the object of much interest over many centuries, the theory only reached maturity with *Kolmogorov's axioms*[1] in the 1930's [240].

As a mathematical theory founded on Kolmogorov's axioms, Probability Theory is essentially uncontroversial at this point. However, the notion of probability (i.e., chance) remains somewhat controversial. We will adopt here the frequentist notion of probability [238], which defines the chance that a particular experiment results in a given outcome as the limiting frequency of this event as the experiment is repeated an increasing number of times. The problem of giving probability a proper definition as it concerns real phenomena is discussed in [93] (with a good dose of humor).

[1] Named after Andrey Kolmogorov (1903 - 1987).

1.1 Elements of set theory

Kolmogorov's formalization of probability relies on some basic notions of *Set Theory*.

A *set* is simply an abstract collection of 'objects', sometimes called *elements* or *items*. Let Ω denote such a set. A *subset* of Ω is a set made of elements that belong to Ω. In what follow, a set will be a subset of Ω.

We write $\omega \in \mathcal{A}$ when the element ω belongs to the set $\mathcal{A}$. And we write $\mathcal{A} \subset \mathcal{B}$ when set $\mathcal{A}$ is a subset of set $\mathcal{B}$. This means that $\omega \in \mathcal{A} \Rightarrow \omega \in \mathcal{B}$. (Remember that $\Rightarrow$ means "implies".) A set with only one element ω is denoted $\{\omega\}$ and is called a *singleton*. Note that $\omega \in \mathcal{A} \Leftrightarrow \{\omega\} \subset \mathcal{A}$. (Remember that $\Leftrightarrow$ means "if and only if".) The *empty set* is defined as a set with no elements and is denoted $\varnothing$. By convention, it is included in any other set.

Problem 1.1. Prove that $\subset$ is transitive, meaning that if $\mathcal{A} \subset \mathcal{B}$ and $\mathcal{B} \subset \mathcal{C}$, then $\mathcal{A} \subset \mathcal{C}$.

The following are some basic set operations.

- *Intersection and disjointness:* The intersection of two sets $\mathcal{A}$ and $\mathcal{B}$ is the set with all the elements belonging to both $\mathcal{A}$ and $\mathcal{B}$, and is denoted $\mathcal{A} \cap \mathcal{B}$. $\mathcal{A}$ and $\mathcal{B}$ are said to be *disjoint* if $\mathcal{A} \cap \mathcal{B} = \varnothing$.
- *Union:* The union of two sets $\mathcal{A}$ and $\mathcal{B}$ is the set with elements belonging to $\mathcal{A}$ or $\mathcal{B}$, and is denoted $\mathcal{A} \cup \mathcal{B}$.
- *Set difference and complement:* The set difference of $\mathcal{B}$ minus $\mathcal{A}$ is the set with elements those in $\mathcal{B}$ that are not in $\mathcal{A}$, and is denoted $\mathcal{B} \setminus \mathcal{A}$. It is sometimes called the complement of $\mathcal{A}$ in $\mathcal{B}$. The complement of $\mathcal{A}$ in the whole set Ω is often denoted $\mathcal{A}^{\mathsf{c}}$.
- *Symmetric set difference:* The symmetric set difference of $\mathcal{A}$ and $\mathcal{B}$ is defined as the set with elements either in $\mathcal{A}$ or in $\mathcal{B}$, but not in both, and is denoted $\mathcal{A} \triangle \mathcal{B}$.

Sets and set operations can be visualized using a *Venn diagram*. See Figure 1.1 for an example.

Problem 1.2. Prove that $\mathcal{A} \cap \varnothing = \varnothing$, $\mathcal{A} \cup \varnothing = \mathcal{A}$, and $\mathcal{A} \setminus \varnothing = \mathcal{A}$. What is $\mathcal{A} \triangle \varnothing$?

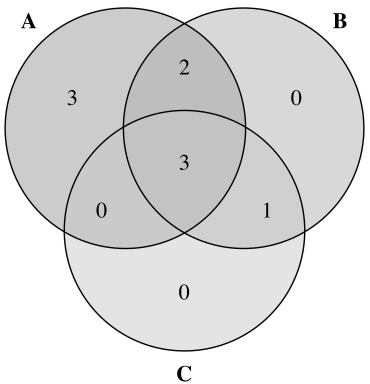
Problem 1.3. Prove that the complement is an involution, meaning $(\mathcal{A}^{\mathsf{c}})^{\mathsf{c}} = \mathcal{A}$.

Problem 1.4. Show that the set difference operation is not symmetric in the sense that $\mathcal{B} \setminus \mathcal{A} \neq \mathcal{A} \setminus \mathcal{B}$ in general. In fact, prove that $\mathcal{B} \setminus \mathcal{A} = \mathcal{A} \setminus \mathcal{B}$ if and only if $\mathcal{A} = \mathcal{B} = \varnothing$.

Proposition 1.5. *The following are true:*

(i) The intersection operation is commutative, meaning $\mathcal{A} \cap \mathcal{B} = \mathcal{B} \cap \mathcal{A}$*, and associative, meaning* $(\mathcal{A} \cap \mathcal{B}) \cap \mathcal{C} =$

Figure 1.1: A Venn diagram helping visualize the sets $\mathcal{A} = \{1,2,4,5,6,7,8,9\}$, $\mathcal{B} = \{2,3,4,5,7,9\}$, and $\mathcal{C} = \{3,4,5,9\}$. The numbers shown in the figure represent the size of each subset. For example, the intersection of these three sets contains 3 elements, since $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C} = \{4,5,9\}$.



$\mathcal{A} \cap (\mathcal{B} \cap \mathcal{C})$. *The same is true for the union operation.*

(ii) The intersection operation is distributive over the union operation, meaning $(\mathcal{A} \cup \mathcal{B}) \cap \mathcal{C} = (\mathcal{A} \cap \mathcal{C}) \cup (\mathcal{B} \cap \mathcal{C})$.

(iii) It holds that $(\mathcal{A} \cap \mathcal{B})^c = \mathcal{A}^c \cup \mathcal{B}^c$. *More generally,* $\mathcal{C} \setminus (\mathcal{A} \cap \mathcal{B}) = (\mathcal{C} \setminus \mathcal{A}) \cup (\mathcal{C} \setminus \mathcal{B})$.

We thus may write $\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}$ and $\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}$, that is, without parentheses, as there is no ambiguity. More generally, for a collection of sets $\{\mathcal{A}_i : i \in I\}$, where I is some index set, we can therefore refer to their intersection and union, denoted

$$\text{(intersection)} \quad \bigcap_{i \in I} \mathcal{A}_i, \qquad \text{(union)} \quad \bigcup_{i \in I} \mathcal{A}_i \ .$$

Remark 1.6. For the reader seeing these operations for the first time, it can be useful to think of $\cap$ and $\cup$ in analogy with the product $\times$ and sum $+$ operations on the integers. In that analogy, $\varnothing$ plays the role of 0.

Problem 1.7. Prove Proposition 1.5. In fact, prove the following identities:

$$\Big(\bigcup_{i \in I} \mathcal{A}_i\Big) \cap \mathcal{B} = \bigcup_{i \in I} (\mathcal{A}_i \cap \mathcal{B}),$$

and

$$\Big(\bigcup_{i \in I} \mathcal{A}_i\Big)^c = \bigcap_{i \in I} \mathcal{A}_i^c, \quad \text{as well as} \quad \Big(\bigcap_{i \in I} \mathcal{A}_i\Big)^c = \bigcup_{i \in I} \mathcal{A}_i^c,$$

for any collection of sets $\{\mathcal{A}_i : i \in I\}$ and any set $\mathcal{B}$.

1.2 Outcomes and events

Having introduced some elements of Set Theory, we use some of these concepts to define a probability experiment and its possible outcomes.

1.2.1 Outcomes and the sample space

In the context of an *experiment*, all the possible *outcomes* are gathered in a *sample space*, denoted Ω henceforth. In mathematical terms, the sample space is a set and the outcomes are elements of that set.

Example 1.8 (Flipping a coin)**.** Suppose that we flip a coin three times in sequence. Assuming the coin can only land heads (H) or tails (T), the sample space Ω consists of all possible ordered sequences of length 3, which in lexicographic order can be written as

$$\Omega = \Big\{\text{HHH}, \text{HHT}, \text{HTH}, \text{HTT}, \text{THH}, \text{THT}, \text{TTH}, \text{TTT}\Big\}.$$

Example 1.9 (Drawing from an urn)**.** Suppose that we draw two balls from an urn in sequence. Assume the urn contains red (R), green (G), and (B) blue balls. If the urn contains at least two balls of each color, or if at each trial the ball is placed back in the urn, the sample space Ω consists of all possible ordered sequences of length 2, which in the RGB order can be written as

$$\Omega = \Big\{\text{RR}, \text{RG}, \text{RB}, \text{GR}, \text{GG}, \text{GB}, \text{BR}, \text{BG}, \text{BB}\Big\}. \tag{1.1}$$

If the urn (only) contains one red ball, one green ball, and two or more blue balls, and a ball drawn from the urn is not returned to the urn, the number of possible outcomes is reduced and the resulting sample space is now

$$\Omega = \Big\{\text{RG}, \text{RB}, \text{GR}, \text{GB}, \text{BR}, \text{BG}, \text{BB}\Big\}.$$

Problem 1.10. What is the sample space when we flip a coin five times? More generally, can you describe the sample space, in words and/or mathematical language, corresponding to an experiment where the coin is flipped n times? What is the size of that sample space?

Problem 1.11. Consider an experiment that consists in drawing two balls from an urn that contains red, green, blue, and yellow balls. However, yellow balls are ignored, in the sense that if such a ball is drawn then it is discarded. How does that change the sample space compared to Example 1.9?

While in the previous examples the sample space is finite, the following is an example where it is (countably) infinite.

Example 1.12 (Flipping a coin until the first heads)**.** Consider an experiment where we flip a coin repeatedly until it lands heads. The sample space in this case is

$$\Omega = \Big\{\text{H}, \text{TH}, \text{TTH}, \text{TTTH}, \dots\Big\}.$$

Problem 1.13. Describe the sample space when the experiment consists in drawing repeatedly without replacement from an urn with red, green, and blue balls, three of each color, until a blue ball is drawn.

Remark 1.14. A sample space is in fact only required to contain all possible outcomes. For instance, in Example 1.9 we may always take the sample space to be (1.1) even though in the second situation that space contains outcomes that will never arise.

1.2.2 Events

Events are subsets of Ω that are of particular interest.

Example 1.15. In the context of Example 1.8, consider the event that the second toss results in heads. As a subset of the sample space, this event is defined as

$$\mathcal{E} = \Big\{\text{HHH}, \text{HHT}, \text{THH}, \text{THT}\Big\}.$$

Example 1.16. In the context of Example 1.9, consider the event that the two balls drawn from the urn are of the same color. This event corresponds to the set

$$\mathcal{E} = \Big\{\text{RR}, \text{GG}, \text{BB}\Big\}.$$

Example 1.17. In the context of Example 1.12, the event that the number of total tosses is even corresponds to the set

$$\mathcal{E} = \Big\{\text{TH}, \text{TTTH}, \text{TTTTTH}, \dots\Big\}.$$

Problem 1.18. In the context of Example 1.8, consider the event that at least two tosses result in heads. Describe this event as a set of outcomes.

1.2.3 Collection of events

Recall that we are interested in particular subsets of the sample space Ω and that we call these 'events'. Let Σ denote the collection of events. We assume throughout that Σ satisfies the following conditions:

- The entire sample space is an event, meaning

$$\Omega \in \Sigma. \tag{1.2}$$

- The complement of an event is an event, meaning

$$\mathcal{A} \in \Sigma \;\Rightarrow\; \mathcal{A}^{\mathsf{c}} \in \Sigma. \tag{1.3}$$

- A countable union of events is an event, meaning

$$\mathcal{A}_1, \mathcal{A}_2, \cdots \in \Sigma \;\Rightarrow\; \bigcup_{i \geq 1} \mathcal{A}_i \in \Sigma. \tag{1.4}$$

A collection of subsets that satisfies these conditions is called a *σ-algebra.* [2]

Problem 1.19. Suppose that Σ is a σ-algebra. Show that $\varnothing \in \Sigma$ and that a countable intersection of subsets of Σ is also in Σ.

From now on, Σ will be assumed to be a σ-algebra over Ω unless otherwise specified. When Σ is a σ-algebra of subsets of a set Ω, the pair (Ω, Σ) is sometimes called a *measurable space*.

Remark 1.20 (The power set)**.** The *power set* of Ω, often denoted 2^Ω, is the collection of all its subsets. (Problem 1.49 provides a motivation for this name and notation.) The power set is trivially a σ-algebra. In the context of an experiment with a discrete sample space, it is customary to work with the power set as σ-algebra, because this can always be done without loss of generality (Chapter 2). When the sample space is not discrete, the situation is more complex and the σ-algebra needs to be chosen with care (Section 4.2).

[2] This refers to the algebra of sets presented in Section 1.1.

1.3 Probability axioms

Before observing the result of an experiment, we speak of the probability that an event will happen. The Kolmogorov axioms formalize this assignment of probabilities to events. This has to be done carefully so that the resulting theory is both coherent and useful for modeling randomness.

A *probability distribution* (aka *probability measure*) on (Ω, Σ) is any real-valued function $\mathbb{P}$ defined on Σ satisfying the following properties (axioms): [3]

- *Non-negativity:*

$$\mathbb{P}(\mathcal{A}) \geq 0, \quad \forall \mathcal{A} \in \Sigma. \tag{1.5}$$

- *Unit measure:*

$$\mathbb{P}(\Omega) = 1. \tag{1.6}$$

- *Additivity on disjoint events:* For any discrete collection of disjoint events $\{\mathcal{A}_i : i \in I\}$,

$$\mathbb{P}\Big(\bigcup_{i \in I} \mathcal{A}_i\Big) = \sum_{i \in I} \mathbb{P}(\mathcal{A}_i). \tag{1.7}$$

[3] Throughout, we will often use 'distribution' or 'measure' as shorthand for 'probability distribution'.

A triplet $(\Omega, \Sigma, \mathbb{P})$, with Ω a sample space (a set), Σ a σ-algebra over Ω, and $\mathbb{P}$ a distribution on Σ, is called a *probability space*. We consider such a triplet in what follows.

Problem 1.21. Show that $\mathbb{P}(\varnothing) = 0$ and that

$$0 \le \mathbb{P}(\mathcal{A}) \le 1, \quad \mathcal{A} \in \Sigma.$$

Thus, although a probability distribution is said to have values in $\mathbb{R}_+$, in fact it has values in $[0, 1]$.

Proposition 1.22 (Law of Total Probability)**.** *For any two events $\mathcal{A}$ and $\mathcal{B}$,*

$$\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{A} \cap \mathcal{B}) + \mathbb{P}(\mathcal{A} \cap \mathcal{B}^{\mathsf{c}}). \tag{1.8}$$

Problem 1.23. Prove Proposition 1.22 using the 3rd axiom.

The 3rd axiom applies to events that are disjoint. The following is a corollary that applies more generally. (In turn, this result implies the 3rd axiom.)

Proposition 1.24 (Law of Addition)**.** *For any two events $\mathcal{A}$ and $\mathcal{B}$, not necessarily disjoint,*

$$\mathbb{P}(\mathcal{A} \cup \mathcal{B}) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}). \tag{1.9}$$

In particular,

$$\mathbb{P}(\mathcal{A}^{\mathsf{c}}) = 1 - \mathbb{P}(\mathcal{A}), \tag{1.10}$$

and,

$$\mathcal{A} \subset \mathcal{B} \;\Rightarrow\; \mathbb{P}(\mathcal{B} \setminus \mathcal{A}) = \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A}). \tag{1.11}$$

Proof. We first observe that we can get (1.11) from the fact that $\mathcal{B}$ is the disjoint union of $\mathcal{A}$ and $\mathcal{B} \setminus \mathcal{A}$ and an application of the 3rd axiom.

We now use this to prove (1.9). We start from the disjoint union

$$\mathcal{A} \cup \mathcal{B} = (\mathcal{A} \setminus \mathcal{B}) \cup (\mathcal{B} \setminus \mathcal{A}) \cup (\mathcal{A} \cap \mathcal{B}).$$

Applying the 3rd axiom yields

$$\mathbb{P}(\mathcal{A} \cup \mathcal{B}) = \mathbb{P}(\mathcal{A} \setminus \mathcal{B}) + \mathbb{P}(\mathcal{B} \setminus \mathcal{A}) + \mathbb{P}(\mathcal{A} \cap \mathcal{B}).$$

Then $\mathcal{A} \setminus \mathcal{B} = \mathcal{A} \setminus (\mathcal{A} \cap \mathcal{B})$, and applying (1.11), we get

$$\mathbb{P}(\mathcal{A} \setminus \mathcal{B}) = \mathbb{P}(\mathcal{A}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}),$$

and exchanging the roles of $\mathcal{A}$ and $\mathcal{B}$,

$$\mathbb{P}(\mathcal{B} \setminus \mathcal{A}) = \mathbb{P}(\mathcal{B}) - \mathbb{P}(\mathcal{A} \cap \mathcal{B}).$$

After some cancellations, we obtain (1.9), which then immediately implies (1.10). □

Problem 1.25 (Uniform distribution)**.** Suppose that Ω is finite. For $\mathcal{A} \subset \Omega$, define $\mathbb{U}(\mathcal{A}) = |\mathcal{A}|/|\Omega|$, where $|\mathcal{A}|$ denotes the number of elements in $\mathcal{A}$. Show that $\mathbb{U}$ is a probability distribution on Ω (equipped with its power set, as usual).

1.4 Inclusion-exclusion formula

The inclusion-exclusion formula is an expression for the probability of a discrete union of events. We start with some basic inequalities that are directly related to the inclusion-exclusion formula and useful on their own.

Boole's inequality Also know as the *union bound*, this inequality [4] is arguably one of the simplest, yet also one of the most useful, inequalities of Probability Theory.

Problem 1.26 (Boole's inequality)**.** Prove that for any countable collection of events $\{\mathcal{A}_i : i \in I\}$,

$$\mathbb{P}\Big(\bigcup_{i\in I} \mathcal{A}_i\Big) \le \sum_{i\in I} \mathbb{P}(\mathcal{A}_i). \tag{1.12}$$

(Note that the right-hand side can be larger than 1 or even infinite.) [One possibility is to use a recursion on the number of events, together with Proposition 1.24, to prove the result for any finite number of events. Then pass to the limit to obtain the result as stated.]

Bonferroni's inequalities [5] These inequalities comprise Boole's inequality. For two events, we saw the Law of Addition (Proposition 1.24), which is an exact expression for the probability of their union. Consider now three events $\mathcal{A}, \mathcal{B}, \mathcal{C}$. Boole's inequality (1.12) gives

$$\mathbb{P}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}) \le \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{C}).$$

The following provides an inequality in the other direction.

Problem 1.27. Show that

$$\begin{aligned}\mathbb{P}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}) \ge \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{C}) \\ - \mathbb{P}(\mathcal{A} \cap \mathcal{B}) - \mathbb{P}(\mathcal{B} \cap \mathcal{C}) - \mathbb{P}(\mathcal{C} \cap \mathcal{A}).\end{aligned}$$

[Drawing a Venn diagram will prove useful.]

In the proof, one typically proves first the identity

$$\begin{aligned}\mathbb{P}(\mathcal{A} \cup \mathcal{B} \cup \mathcal{C}) = \mathbb{P}(\mathcal{A}) + \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{C}) \\ - \mathbb{P}(\mathcal{A} \cap \mathcal{B}) - \mathbb{P}(\mathcal{B} \cap \mathcal{C}) - \mathbb{P}(\mathcal{C} \cap \mathcal{A}) \\ + \mathbb{P}(\mathcal{A} \cap \mathcal{B} \cap \mathcal{C}),\end{aligned}$$

which generalizes the Law of Addition to three events.

[4] Named after George Boole (1815 - 1864).

[5] Named after Carlo Emilio Bonferroni (1892 - 1960).

Proposition 1.28 (Bonferroni's inequalities)**.** *Consider any collection of events $\mathcal{A}_1, \dots, \mathcal{A}_n$, and define*

$$S_k := \sum_{1 \le i_1 < \cdots < i_k \le n} \mathbb{P}(\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_k}).$$

Then

$$\mathbb{P}(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n) \le \sum_{j=1}^{k} (-1)^{j-1} S_j, \quad k \text{ odd}; \tag{1.13}$$

$$\mathbb{P}(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n) \ge \sum_{j=1}^{k} (-1)^{j-1} S_j, \quad k \text{ even}. \tag{1.14}$$

Problem 1.29. Write down all of Bonferroni's inequalities for the case of four events $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3, \mathcal{A}_4$.

Inclusion-exclusion formula The last Bonferroni inequality (at $k = n$) is in fact an equality, the so-called *inclusion-exclusion formula*,

$$\mathbb{P}(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_n) = \sum_{j=1}^{n} (-1)^{j-1} S_j. \tag{1.15}$$

(In particular, the last inequality in Problem 1.29 is an equality.)

1.5 Conditional probability and independence

1.5.1 Conditional probability

Conditioning on an event $\mathcal{B}$ restricts the sample space to $\mathcal{B}$. In other words, although the experiment might yield other outcomes, conditioning on $\mathcal{B}$ focuses the attention on the outcomes that made $\mathcal{B}$ happen. In what follows we assume that $\mathbb{P}(\mathcal{B}) > 0$.

Problem 1.30. Show that $\mathbb{Q}$, defined for $\mathcal{A} \in \Sigma$ as $\mathbb{Q}(\mathcal{A}) = \mathbb{P}(\mathcal{A} \cap \mathcal{B})$, is a probability distribution if and only if $\mathbb{P}(\mathcal{B}) = 1$.

To define a bona fide probability distribution we renormalize $\mathbb{Q}$ to have total mass equal to 1 (required by the 2nd axiom) as follows

$$\mathbb{P}(\mathcal{A} \,|\, \mathcal{B}) = \frac{\mathbb{P}(\mathcal{A} \cap \mathcal{B})}{\mathbb{P}(\mathcal{B})}, \quad \text{for } \mathcal{A} \in \Sigma.$$

We call $\mathbb{P}(\mathcal{A} \,|\, \mathcal{B})$ the *conditional probability* of $\mathcal{A}$ given $\mathcal{B}$.

Problem 1.31. Show that $\mathbb{P}(\cdot \,|\, \mathcal{B})$ is indeed a probability distribution on Ω.

Problem 1.32. In the context of Example 1.8, assume that any outcome is equally likely. Then what is the

probability that the last toss lands heads if the previous tosses landed heads? Answer that same question when the coin is tossed $n \geq 2$ times, with n arbitrary and possibly large. [Regardless of n, the answer is 1/2.]

The conclusions of Problem 1.32 may surprise some readers. And indeed, conditional probabilities can be rather unintuitive. We will come back to Problem 1.32, which is an example of the *Gambler's Fallacy*. Here is another famous example.

Example 1.33 (Monty Hall Problem)**.** This problem is based on a television show in the US called *Let's Make a Deal* and named after its longtime presenter, Monty Hall. The following description is taken from a New York Times article [232]:

> *Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the other doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to take the switch?*

Not many problems in probability are discussed in the New York Times, to say the least. This problem is so simple to state and the answer so counter-intuitive that it generated quite a controversy (read the article). The problem can mislead anyone, including professional mathematicians, let alone the commoner appearing on television!

There is an entire book on the Monty Hall Problem [196]. The textbook [113] discusses this problem among other paradoxes arising when dealing with conditional probabilities.

1.5.2 Independence

Two events $\mathcal{A}$ and $\mathcal{B}$ are said to be *independent* if knowing that $\mathcal{B}$ happens does not change the chances (i.e., the probability) that $\mathcal{A}$ happens. This is formalized by saying that the probability of $\mathcal{A}$ conditional on $\mathcal{B}$ is equal to its (unconditional) probability, or in formula,

$$\mathbb{P}(\mathcal{A} \,|\, \mathcal{B}) = \mathbb{P}(\mathcal{A}). \tag{1.16}$$

The wording in English would imply a symmetric relationship. That's indeed the case because (1.16) is equivalent to $\mathbb{P}(\mathcal{B} \,|\, \mathcal{A}) = \mathbb{P}(\mathcal{B})$. The following equivalent definition of independence makes the symmetry transparent.

Proposition 1.34. *Two events $\mathcal{A}$ and $\mathcal{B}$ are independent if and only if*

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{A})\,\mathbb{P}(\mathcal{B}). \tag{1.17}$$

The identity (1.17) is often taken as a definition of independence.

Problem 1.35. Show that any event that never happens (i.e., having zero probability) is independent of any other event. In particular, $\varnothing$ is independent of any event.

Problem 1.36. Show that any event that always happens (i.e., having probability one) is independent of any other event. In particular, Ω is independent of any event.

The identity (1.17) only applies to independent events. However, it can be generalized as follows. (Notice the parallel with the Law of Addition (1.9).)

Problem 1.37 (Law of Multiplication)**.** Prove that, for any events $\mathcal{A}$ and $\mathcal{B}$,

$$\mathbb{P}(\mathcal{A} \cap \mathcal{B}) = \mathbb{P}(\mathcal{A} \,|\, \mathcal{B})\, \mathbb{P}(\mathcal{B}). \tag{1.18}$$

Problem 1.38 (Independence and disjointness)**.** The notions of independence and disjointness are often confused by the novice, even though they are very different. For example, show that two disjoint events are independent only when at least one of them either never happens or always happens.

Problem 1.39. Combine the Law of Total Probability (1.8) and the Law of Multiplication (1.18) to get

$$\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{A} \,|\, \mathcal{B})\, \mathbb{P}(\mathcal{B}) + \mathbb{P}(\mathcal{A} \,|\, \mathcal{B}^{\mathsf{c}})\, \mathbb{P}(\mathcal{B}^{\mathsf{c}}) \tag{1.19}$$

Problem 1.40. Suppose we draw without replacement from an urn with r red balls and b blue balls. At each stage, every ball remaining in the urn is equally likely to be picked. Use (1.19) to derive the probability of drawing a blue ball on the 3rd trial.

1.5.3 Mutual independence

One may be interested in several events at once. Some events, $\mathcal{A}_i, i \in I$, are said to be *mutually independent* (or *jointly independent*) if

$$\mathbb{P}(\mathcal{A}_{i_1} \cap \cdots \cap \mathcal{A}_{i_k}) = \mathbb{P}(\mathcal{A}_{i_1}) \times \cdots \times \mathbb{P}(\mathcal{A}_{i_k}), \quad \text{for any } k\text{-tuple } 1 \le i_1 < \cdots < i_k \le r. \tag{1.20}$$

They are said to be *pairwise independent* if

$$\mathbb{P}(\mathcal{A}_i \cap \mathcal{A}_j) = \mathbb{P}(\mathcal{A}_i)\, \mathbb{P}(\mathcal{A}_j), \quad \text{for all } i \ne j.$$

Obviously, mutual independence implies pairwise independence. The reverse implication is false, as the following counter-example shows.

Problem 1.41. Consider the uniform distribution on

$$\big\{(0,0,0),(0,1,1),(1,0,1),(1,1,0)\big\}.$$

Let $\mathcal{A}_i$ be the event that the ith coordinate is 1. Show that these events are pairwise independent but not mutually independent.

The following generalizes the Law of Multiplication (1.18). It is sometimes referred to as the *Chain Rule*.

Proposition 1.42 (General Law of Multiplication)**.** *For any mutually independent events, $\mathcal{A}_1,\dots,\mathcal{A}_r$,*

$$\mathbb{P}(\mathcal{A}_1\cap\cdots\cap\mathcal{A}_r)=\prod_{k=1}^{r}\mathbb{P}\big(\mathcal{A}_k\,|\,\mathcal{A}_1\cap\cdots\cap\mathcal{A}_{k-1}\big). \tag{1.21}$$

For example, for any events $\mathcal{A},\mathcal{B},\mathcal{C}$,

$$\mathbb{P}(\mathcal{A}\cap\mathcal{B}\cap\mathcal{C})=\mathbb{P}(\mathcal{C}\,|\,\mathcal{A}\cap\mathcal{B})\,\mathbb{P}(\mathcal{B}\,|\,\mathcal{A})\,\mathbb{P}(\mathcal{A}).$$

Problem 1.43. In the same setting as Problem 1.32, show that the result of the tosses are mutually independent. That is, define $\mathcal{A}_i$ as the event that the ith toss results in heads and show that $\mathcal{A}_1,\dots,\mathcal{A}_n$ are mutually independent. In fact, show that the distribution is the uniform distribution (Problem 1.25) if and only if the tosses are fair and mutually independent.

1.5.4 Bayes formula

The *Bayes formula*[6] can be used to "turn around" a conditional probability.

Proposition 1.44 (Bayes formula)**.** *For any two events $\mathcal{A}$ and $\mathcal{B}$,*

$$\mathbb{P}(\mathcal{A}\,|\,\mathcal{B})=\frac{\mathbb{P}(\mathcal{B}\,|\,\mathcal{A})\,\mathbb{P}(\mathcal{A})}{\mathbb{P}(\mathcal{B})}. \tag{1.22}$$

Proof. By (1.18),

$$\mathbb{P}(\mathcal{A}\cap\mathcal{B})=\mathbb{P}(\mathcal{A}\,|\,\mathcal{B})\,\mathbb{P}(\mathcal{B}),$$

and also

$$\mathbb{P}(\mathcal{A}\cap\mathcal{B})=\mathbb{P}(\mathcal{B}\,|\,\mathcal{A})\,\mathbb{P}(\mathcal{A}),$$

which yield the result when combined. □

The denominator in (1.22) is sometimes expanded using (1.19) to get

$$\mathbb{P}(\mathcal{A}\,|\,\mathcal{B})=\frac{\mathbb{P}(\mathcal{B}\,|\,\mathcal{A})\,\mathbb{P}(\mathcal{A})}{\mathbb{P}(\mathcal{B}\,|\,\mathcal{A})\,\mathbb{P}(\mathcal{A})+\mathbb{P}(\mathcal{B}\,|\,\mathcal{A}^{\mathsf{c}})\,\mathbb{P}(\mathcal{A}^{\mathsf{c}})}. \tag{1.23}$$

This form is particularly useful when $\mathbb{P}(\mathcal{B})$ is not directly available.

[6] Named after Thomas Bayes (1701 - 1761).

Problem 1.45. Suppose we draw without replacement from an urn with r red balls and b blue balls. What is the probability of drawing a blue ball on the 1st trial when drawing a blue ball on the 2nd trial?

Base Rate Fallacy Consider a medical test for the detection of a rare disease. There are two types of mistakes that the test can make:

- *False positive* when the test is positive even though the subject does not have the disease;
- *False negative* when the test is negative even though the subject has the disease.

Let α denote the probability of a false positive; $1-\alpha$ is sometimes called the *sensitivity*. Let β denote the probability of a false negative. $1-\beta$ is sometimes called the *specificity*. For example, the study reported in [183] evaluates the sensitivity and specificity of several HIV tests.

Suppose that the incidence of a certain disease is π, meaning that the disease affects a proportion π of the population of interest. A person is chosen at random from the population and given the test, which turns out to be positive. What are the chances that this person actually has the disease? Ignoring the *base rate* (i.e., the disease's prevalence) would lead one to believe these chances to be $1-\beta$. This is an example of the *Base Rate Fallacy*.

Indeed, define the events

$$\mathcal{A} = \text{'the person has the disease'}, \tag{1.24}$$

$$\mathcal{B} = \text{'the test is positive'}. \tag{1.25}$$

Thus our goal is to compute $\mathbb{P}(\mathcal{A}\,|\,\mathcal{B})$. Because the person was chosen at random from the population, we know that $\mathbb{P}(\mathcal{A}) = \pi$. We know the test's sensitivity, $\mathbb{P}(\mathcal{B}^{\mathsf{c}}\,|\,\mathcal{A}^{\mathsf{c}}) = 1-\alpha$, and its specificity, $\mathbb{P}(\mathcal{B}\,|\,\mathcal{A}) = 1-\beta$. Plugging this into (1.23), we get

$$\mathbb{P}(\mathcal{A}\,|\,\mathcal{B}) = \frac{(1-\beta)\pi}{(1-\beta)\pi + \alpha(1-\pi)}. \tag{1.26}$$

Mathematically, the Base Rate Fallacy arises from confusing $\mathbb{P}(\mathcal{A}\,|\,\mathcal{B})$ (which is what we want) with $\mathbb{P}(\mathcal{B}\,|\,\mathcal{A})$. We saw that the former depends on the latter and on the base rate $\mathbb{P}(\mathcal{A})$.

Problem 1.46. Show that $\mathbb{P}(\mathcal{A}\,|\,\mathcal{B}) = \mathbb{P}(\mathcal{B}\,|\,\mathcal{A})$ if and only if $\mathbb{P}(\mathcal{A}) = \mathbb{P}(\mathcal{B})$.

Example 1.47 (Finding terrorists)**.** In a totally different setting, Sageman [202] makes the point that a system for identifying terrorists, even if 99% accurate, cannot be ethically deployed on an entire population.

Fallacies in the courtroom Suppose that in a trial for murder in the US, some blood of type O- was found on the crime scene, matching the defendant's blood type. That blood type has a prevalence of about 1% in the US [7]. This leads the prosecutor to conclude that the suspect is guilty with 99% chance. This is an example of the *Prosecutor's Fallacy*.

In terms of mathematics, the error is the same as in the Base Rate Fallacy. In practice, the situation is even worse here because it is not even clear how to define the base rate. (Certainly, the base rate cannot be the unconditional probability that the defendant is guilty.)

In the same hypothetical setting, the defense could argue that, assuming the crime took place in a city with a population of about half a million, the defendant is only one among five thousand people in the region with the same blood type and that therefore the chances that he is guilty are 1/5000 = 0.02%. The argument is actually correct if there is no other evidence and it can be argued that the suspect was chosen more or less uniformly at random from the population. Otherwise, in particular if the latter is doubtful, this is is an example of the *Defendant's Fallacy*.

[7] redcrossblood.org/learn-about-blood/blood-types.html

Example 1.48. *People vs. Collins* is robbery case [8] that took place in Los Angeles, California in 1968. A witness had seen a Black male with a beard and mustache together with White female with a blonde ponytail fleeing in a yellow car. The Collins (a married couple) exhibited all these attributes. The prosecutor argued that the chances of another couple matching the description were 1 in 12,000,000. This lead to a conviction. However, the California Supreme Court overturned the decision. This was based on the questionable computations of the base rate as well as the fact that the chances of another couple in the Los Angeles area (with a population in the millions) matching the description were much higher.

For more on the use of statistics in the courtroom, see [230].

1.6 Additional problems

Problem 1.49. Show that if $|\Omega| = N$, then the collection of all subsets of Ω (including the empty set) has cardinality 2^N. This motivates the notation 2^Ω for this collection and also its name, as it is often called the power set of Ω.

[8] courtlistener.com/opinion/1207456/people-v-collins/

Problem 1.50. Let $\{\Sigma_i, i \in I\}$ denote a family of σ-algebras over a set Ω. Prove that $\bigcap_{i\in I} \Sigma_i$ is also a σ-algebra over Ω.

Problem 1.51. Let $\{\mathcal{A}_i, i \in I\}$ denote a family of subsets of a set Ω. Show that there is a unique smallest (in terms of inclusion) σ-algebra over Ω that contains each of these subsets. This σ-algebra is said to be *generated* by the family $\{\mathcal{A}_i, i \in I\}$.

Problem 1.52 (General Base Rate Fallacy)**.** Assume that the same diagnostic test is performed on m individuals to detect the presence of a certain pathogen. Due to variation in characteristics, the test performed on Individual i has sensitivity $1-\alpha_i$ and specificity $1-\beta_i$. Assume that a proportion π of these individuals have the pathogen. Show that (1.26) remains valid as the probability that an individual chosen uniformly at random has the pathogen given that the test is positive, with $1-\alpha$ defined as the average sensitivity and $1-\beta$ defined as the average specificity, meaning $\alpha = \frac{1}{m}\sum_{i=1}^{m} \alpha_i$ and $\beta = \frac{1}{m}\sum_{i=1}^{m} \beta_i$.

CHAPTER 2

DISCRETE PROBABILITY SPACES

We consider in this chapter the case of a probability space $(\Omega, \Sigma, \mathbb{P})$ with discrete sample space Ω. As we noted in Remark 1.20, it is customary to let Σ be the power set of Ω. We do so anytime we are dealing with a discrete probability space, as this can be done without loss of generality.

2.1 Probability mass functions

Given a probability distribution $\mathbb{P}$, define its *mass function* as

$$f(\omega) := \mathbb{P}(\{\omega\}), \quad \omega \in \Omega. \tag{2.1}$$

In general, we say that f is a mass function on Ω if it is a real-valued function on Ω satisfying the following conditions:

- *Non-negativity:*

$$f(\omega) \geq 0 \text{ for any } \omega \in \Omega.$$

- *Unit measure:*
$$\sum_{\omega\in\Omega} f(\omega) = 1. \tag{2.2}$$

Note that, necessarily, such an f takes values in $[0,1]$.

Problem 2.1. Show that (2.1) defines a probability mass function on Ω. Conversely, show that a probability mass function f on Ω defines a probability distribution on Ω as follows:
$$\mathbb{P}(\mathcal{A}) := \sum_{\omega\in\mathcal{A}} f(\omega), \quad \text{for } \mathcal{A} \subset \Omega. \tag{2.3}$$

Problem 2.2. Show that, if $\mathbb{P}$ has mass function f, and $\mathcal{B}$ is an event with $\mathbb{P}(\mathcal{B}) > 0$, then the conditional distribution given $\mathcal{B}$, meaning $\mathbb{P}(\cdot\,|\,\mathcal{B})$, has mass function
$$f(\omega\,|\,\mathcal{B}) := \frac{f(\omega)}{\mathbb{P}(\mathcal{B})}\,\{\omega \in \mathcal{B}\}, \quad \text{for } \omega \in \Omega,$$
where $\{\omega \in \mathcal{B}\} = 1$ if $\omega \in \mathcal{B}$ and $= 0$ otherwise.

2.2 Uniform distributions

Assume the sample space Ω is finite. For a set $\mathcal{A}$, denote its cardinality (meaning the number of elements it contains) by $|\mathcal{A}|$. The *uniform distribution!discrete* on Ω is defined as
$$\mathbb{P}(\mathcal{A}) = \frac{|\mathcal{A}|}{|\Omega|}, \quad \text{for } \mathcal{A} \subset \Omega. \tag{2.4}$$

We saw in Problem 1.25 that this is indeed a probability distribution on Ω (equipped with its power set).

Equivalently, the uniform distribution on Ω is the distribution with constant mass function. Because of the requirements a mass function satisfies by definition, this necessarily means that the mass function is equal to $1/|\Omega|$ everywhere, meaning
$$f(\omega) := \frac{1}{|\Omega|}, \quad \text{for } \omega \in \Omega. \tag{2.5}$$

Such a probability space thus models an experiment where all outcomes are equally likely.

Example 2.3 (Rolling a die)**.** Consider an experiment where a die, with six faces numbered 1 through 6, is rolled once and the result is recorded. The sample space is $\Omega = \{1,2,3,4,5,6\}$. The usual assumption that the die is fair is modeled by taking the distribution to be the uniform distribution, which puts mass $1/6$ on each outcome.

Remark 2.4 (Combinatorics)**.** The uniform distribution is intimately related to *Combinatorics*, which is the branch of Mathematics dedicated to counting. This is because of its definition in (2.4), which implies that computing the probability of an event $\mathcal{A}$ reduces to computing its cardinality $|\mathcal{A}|$, meaning counting the number of outcomes in $\mathcal{A}$.

2.3 Bernoulli trials

Consider an experiment where a coin is tossed repeatedly. We speak of *Bernoulli trials*[9] when the probability of heads remains the same at each toss regardless of the previous tosses. We consider a biased coin with probability of landing heads equal to $p \in [0,1]$. We will call this a p-coin.

Problem 2.5 (The roulette)**.** An American roulette has 38 slots: 18 are colored black (B), 18 are colored red (R), 2 slots colored green (G). A ball is rolled, and eventually lands in one of the slots. One way a player can gamble is to bet on a color, red or black. If the ball lands in a slot with that color, the player doubles his bet. Otherwise, the player loses his bet. Assuming the wheel is fair, show that the probability of winning in a given trial is $p = 18/38$ regardless of the color the player bets on. (Note that $p < 1/2$, and $1/2 - p = 1/38$ is the casino's margin.)

Remark 2.6 (Beating the roulette)**.** In the game of roulette, the odds are of course against the player. We will see later in Section 9.3.1 that this guaranties any gambler will lose his fortune if he keeps on playing. This is so if the mathematics underlying this statement are an accurate description of how the game is played in real life. But this is not necessarily the case. For one thing, the equipment can be less than perfect. This was famously exploited in the late 1940's by Hibbs and Walford, and lead the casinos to use much better made roulettes that had no obvious defects. Still, Thorp and Shannon took on the challenge of beating the roulette in the late 1950's and early 1960's. For that purpose, they built one of the first wearable computers to help them predict where the ball would end based on an appraisal of the ball's position and speed at a certain time. Their system afforded them advantageous odds against the casino. For more on this, see [116].

[9] Named after Jacob Bernoulli (1655 - 1705).

2.3.1 Probability of a sequence of given length

Assume we toss a p-coin n times (or simply focus on the first n tosses if the coin is tossed an infinite number of times). In this case, in contrast with the situation in Problem 1.43, the distribution over the space of n-tuples of heads and tails is not the uniform distribution, unless $p = 1/2$. We derive the distribution in closed form for an arbitrary p. We do this for the sequence HTHHT (so that $n = 5$) to illustrate the main arguments. Let $\mathcal{A}_i$ be the event that the ith trial results in heads. Then applying

the Chain Rule (1.21), we have

$$\begin{aligned}\mathbb{P}(\text{HTHHT}) = \mathbb{P}(\mathcal{A}_5^c \mid \mathcal{A}_1 \cap \mathcal{A}_2^c \cap \mathcal{A}_3 \cap \mathcal{A}_4) \\ \times \mathbb{P}(\mathcal{A}_4 \mid \mathcal{A}_1 \cap \mathcal{A}_2^c \cap \mathcal{A}_3) \\ \times \mathbb{P}(\mathcal{A}_3 \mid \mathcal{A}_1 \cap \mathcal{A}_2^c) \\ \times \mathbb{P}(\mathcal{A}_2^c \mid \mathcal{A}_1) \times \mathbb{P}(\mathcal{A}_1). \quad (2.6)\end{aligned}$$

By assumption, a toss results in heads with probability p regardless of the previous tosses, so that

$$\mathbb{P}(\mathcal{A}_5^c \mid \mathcal{A}_1 \cap \mathcal{A}_2^c \cap \mathcal{A}_3 \cap \mathcal{A}_4) = \mathbb{P}(\mathcal{A}_5^c) = 1 - p, \quad (2.7)$$

$$\mathbb{P}(\mathcal{A}_4 \mid \mathcal{A}_1 \cap \mathcal{A}_2^c \cap \mathcal{A}_3) = \mathbb{P}(\mathcal{A}_4) = p, \quad (2.8)$$

$$\mathbb{P}(\mathcal{A}_3 \mid \mathcal{A}_1 \cap \mathcal{A}_2^c) = \mathbb{P}(\mathcal{A}_3) = p, \quad (2.9)$$

$$\mathbb{P}(\mathcal{A}_2^c \mid \mathcal{A}_1) = \mathbb{P}(\mathcal{A}_2^c) = 1 - p, \quad (2.10)$$

$$\mathbb{P}(\mathcal{A}_1) = p. \quad (2.11)$$

Plugging this into (2.6), we obtain

$$\mathbb{P}(\text{HTHHT}) = p(1-p)pp(1-p) = p^3(1-p)^2, \quad (2.12)$$

after rearranging factors.

Problem 2.7. In the same example, show that the assumption that a toss results in heads with the same probability regardless of the previous tosses implies that the events $\mathcal{A}_i$ are mutually independent.

Beyond this special case, the following holds.

Proposition 2.8. *Consider n independent tosses of a p-coin. Regardless of the order, if k denotes the number of heads in a given sequence of n trials, the probability of that sequence is*

$$p^k(1-p)^{n-k}. \quad (2.13)$$

Problem 2.9. Prove Proposition 2.8 by induction on n

Remark 2.10. Although n Bernoulli trials do not necessarily result in a uniform distribution, Proposition 2.8 implies that, conditional on the number of heads being k, the distribution is uniform over the subset of sequences with exactly k heads.

Remark 2.11 (Gambler's Fallacy)**.** Consider a casino roulette (Problem 2.5). Assume that you have just observed five spins that all resulted in red (i.e., RRRRR). What color would you bet on? Many a gambler would bet on B in this situation, believing that "it is time for the ball to land black". In fact, unless you have reasons to suspect otherwise, the natural assumption is that each spin of the wheel is fair and independent of the previous ones, and with this assumption the probability of B remains the same as the probability of R (that is, 18/38).

Example 2.12. The paper [39] (which received a good

amount of media attention) studies how the sequence in which unrelated cases are handled affects the decisions that are made in "refugee asylum court decisions, loan application reviews, and Major League Baseball umpire pitch calls", and finds evidence of the Gambler's Fallacy at play.

2.3.2 Number of heads in a sequence of given length

Suppose again that we toss a p-coin n times independently. We saw in Proposition 2.8 that the number of heads in the sequence dictates the probability of observing that sequence. Thus it is of interest to study that quantity. In particular, we want to compute the probability of observing exactly k heads, where $k \in \{0, \dots, n\}$ is given.

Factorials For a positive integer n, define its *factorial* to be

$$n! := \prod_{i=1}^{n} i = n \times (n-1) \times \cdots \times 1.$$

For example, $5! = 5 \times 4 \times 3 \times 2 \times 1 = 120$. By convention, $0! = 1$.

Proposition 2.13. *$n!$ is the number of orderings of n distinct items.*

This can be generalized as follows. For two non-negative integers $k \le n$, define the *falling factorial*

$$(n)_k := n(n-1)\cdots(n-k+1). \tag{2.14}$$

For example, $(5)_3 = 5 \times 4 \times 3 = 60$. By convention, $(n)_0 = 1$. In particular, $(n)_n = n!$.

Proposition 2.14. *Given positive integers $k \le n$, $(n)_k$ is the number of ordered subsets of size k of a set with n distinct elements.*

Proof. There are n choices for the 1st position, then $n-1$ remaining choices for the 2nd position, etc, and $n-(k-1) = n-k+1$ remaining choices for the kth position. These numbers of choices multiply to give the answer. (Why they multiply and not add is crucial, and is explained at length, for example, in [113].) □

Binomial coefficients For two non-negative integers $k \le n$, define the *binomial coefficient*

$$\binom{n}{k} := \frac{n!}{k!(n-k)!} = \frac{(n)_k}{k!}. \tag{2.15}$$

The binomial coefficient (2.15) is often read "n choose k", as it corresponds to the number of ways of choosing k

distinct items out of a total of n, disregarding the order in which they are chosen.

Proposition 2.15. *Given positive integers $k \le n$, $\binom{n}{k}$ is the number of (unordered) subsets of size k of a set with n distinct elements.*

Proof. Fix k and n, and let N denote the number of (unordered) subsets of size k of a set with n distinct elements. Each such subset can be ordered in $k!$ ways by Proposition 2.13, so that there are $N \times k!$ ordered subsets of size k. Hence, $Nk! = (n)_k$ by Proposition 2.14, resulting in $N = (n)_k/k! = \binom{n}{k}$. □

Problem 2.16 (Pascal's triangle)**.** Adopt the convention that $\binom{n}{k} = 0$ when $k < 0$ or when $k > n$. Then prove that, for any integers k and n,

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

Do you see how this formula can be used to compute binomial coefficients recursively?

Exactly k heads out of n trials Fix an integer $k \in \{0, \dots, n\}$. By (2.3) and Proposition 2.8,

$$\mathbb{P}(\text{'exactly } k \text{ heads'}) = Np^k(1-p)^{n-k},$$

where N is the number of sequences of length n with exactly k heads. The trials are numbered 1 through n, and such a sequence is identified by the k trials among these where the coin landed heads. Thus N is equal to the number of (unordered) subsets of size k of $\{1, \dots, n\}$, which by Proposition 2.15 corresponds to the binomial coefficient $\binom{n}{k}$. Thus,

$$\mathbb{P}(\text{'exactly } k \text{ heads'}) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.16}$$

2.4 Urn models

We already discussed an experiment involving an urn in Example 1.9. We consider the more general case of an urn that contains balls of only two colors, say, red and blue. Let r denote the number of red balls and b the number of blue balls. We will call this an (r, b)-urn. The experiment consists in drawing n balls from such an urn. We saw in Example 1.9 that this is enough information to define the sample space. However, the sampling process is not specific enough to define the sampling distribution. We discuss here the two most basic variants: sampling with replacement and sampling without replacement. We make the crucial assumption that at every stage each ball in the urn is equally likely to be drawn.

2.4.1 Sampling with replacement

As the name indicates, this sampling scheme consists in repeatedly drawing a ball from the urn, every time returning the ball to the urn. Thus, the urn is the same before each draw. Because of our assumptions, this means that the probability of drawing a red ball remains constant, equal to $r/(r+b)$. Thus we conclude that sampling with replacement from an urn with r red balls and b blue balls is analogous to flipping a p-coin with $p = r/(r+b)$. In particular, based on (2.13), the probability of any sequence of n draws containing exactly k red balls is

$$\Big(\frac{r}{r+b}\Big)^k\Big(\frac{b}{r+b}\Big)^{n-k}.$$

Remark 2.17 (From urn to coin)**.** We note that, unlike a general p-coin as described in Section 2.3, where p is in principle arbitrary in $[0,1]$, the parameter p that results from sampling with replacement from a finite urn is necessarily a rational number. However, because the rationals are dense in $[0,1]$, it is possible to use an urn to approximate a p-coin. For that, it simply suffices to choose (integers) r and b be such that $r/(r+b)$ approaches the desired $p \in [0,1]$.

2.4.2 Sampling without replacement

This sampling scheme consists in repeatedly drawing a ball from the urn, without ever returning the ball to the urn. Thus the urn changes with each draw. For example, consider an urn with $r = 2$ red balls and $b = 3$ blue balls, and assume we draw a total of $n = 2$ balls from the urn without replacement. On the first draw, the probability of a red ball is $2/(2+3) = 2/5$, while the probability of drawing a blue ball is $3/5$.

- After first drawing a red ball, the urn contains 1 red ball and 3 blue balls, so the probability of a red ball on the 2nd draw is $1/4$.
- After first drawing a blue ball, the urn contains 2 red ball and 2 blue balls, so the probability of a red ball on the 2nd draw is $2/4$.

Although the resulting distribution is different than when the draws are executed with replacement, it is still true that all sequences with the same number of red balls have the same probability.

Proposition 2.18. *Assume that $n \le r+b$, for otherwise the sampling scheme is not feasible. Also, assume that $k \le r$. The probability of any sequence of n draws containing*

exactly k red balls is

$$\frac{r(r-1)\cdots(r-k+1)b(b-1)\cdots(b-n+k+1)}{(r+b)(r+b-1)\cdots(r+b-n+1)}. \tag{2.17}$$

Remark 2.19. The usual convention when writing a product like $r(r-1)\cdots(r-k+1)$ is that it is equal to 1 when $k=0$, equal to r when $k=1$, equal to $r(r-1)$ when $k=2$, etc. A more formal way to write such products is using factorials (Section 2.3.2).

We do not provide a formal proof of this result, but rather examine an example with $n=5$, $k=3$, and r and b arbitrary. We consider the outcome $\omega = \text{RBRRB}$. Let $\mathcal{A}_i$ be the event that the ith draw is red and note that ω corresponds to $\mathcal{A}_1 \cap \mathcal{A}_2^c \cap \mathcal{A}_3 \cap \mathcal{A}_4 \cap \mathcal{A}_5^c$. Then, applying the Chain Rule (Proposition 1.42), we have

$$\begin{aligned}\mathbb{P}(\text{RBRRB}) = \mathbb{P}(\mathcal{A}_5^c \mid \mathcal{A}_1 \cap \mathcal{A}_2^c \cap \mathcal{A}_3 \cap \mathcal{A}_4) \\ \times \mathbb{P}(\mathcal{A}_4 \mid \mathcal{A}_1 \cap \mathcal{A}_2^c \cap \mathcal{A}_3) \\ \times \mathbb{P}(\mathcal{A}_3 \mid \mathcal{A}_1 \cap \mathcal{A}_2^c) \\ \times \mathbb{P}(\mathcal{A}_2^c \mid \mathcal{A}_1) \times \mathbb{P}(\mathcal{A}_1).\end{aligned} \tag{2.18}$$

Then, for example, $\mathbb{P}(\mathcal{A}_3 \mid \mathcal{A}_1 \cap \mathcal{A}_2^c)$ is the probability of drawing a red after having drawn a red and then a blue. At that point there are $r-1$ reds and $b-1$ blues in the urn, so that probability is $(r-1)/(r-1+b-1) = (r-1)/(r+b-2)$.

Reasoning in the same way with the other factors, we obtain

$$\begin{aligned}\mathbb{P}(\text{RBRRB}) = \left(\frac{b-1}{r+b-4}\right) \times \left(\frac{r-2}{r+b-3}\right) \\ \times \left(\frac{r-1}{r+b-2}\right) \times \left(\frac{b}{r+b-1}\right) \times \left(\frac{r}{r+b}\right).\end{aligned} \tag{2.19}$$

Rearranging factors, we recover (2.17) specialized to the present case.

Problem 2.20. Repeat the above with any other sequence of 5 draws with exactly 3 reds. Verify that all these sequences have the same probability of occurring.

Problem 2.21 (Sampling without replacement from a large urn)**.** We already noted that sampling with replacement from an (r,b)-urn amounts to tossing a p-coin with $p = r/(r+b)$. Assume now that we are sampling without replacement, but that the urn is very large. This can be considered in an asymptotic setting where r and b diverge to infinity in such a way that $r/(r+b) \to p$. Show that, in the limit, sampling without replacement from the urn also amounts to tossing a p-coin. Do so by proving that, for any n and k fixed, (2.17) converges to (2.13).

2.4.3 Number of heads in a sequence of given length

As in Section 2.3.2, we derive the probability of observing exactly k red balls in n draws without replacement. We show that

$$\mathbb{P}(\text{`exactly } k \text{ heads'}) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{r+b}{n}}. \tag{2.20}$$

Indeed, since we assume that each ball is equally likely to be drawn at each stage, it follows that any subset of balls of size n is equally likely. We are thus in the uniform case (Section 2.2), and therefore the probability is given by the number of outcomes in the event divided by the total number of outcomes.

The denominator in (2.20) is the number of possible outcomes, namely, subsets of balls of size n taken from the urn. [10]

The numerator in (2.20) is the number of outcomes with exactly k red balls. Indeed, any such outcome can be uniquely obtained by first choosing k red balls out of r in total — there are $\binom{r}{k}$ ways to do that — and then choosing $n-k$ blue balls out of b in total — there are $\binom{b}{n-k}$ ways to do that.

[10] Although the balls could be indistinguishable except for their colors, we use a standard trick in Combinatorics which consists in making the balls identifiable. This is only needed as a thought experiment. One could imagine, for example, numbering the balls 1 through $r+b$.

2.4.4 Other urn models

There are many urn models as, despite their apparent simplicity, their theoretical study is surprisingly rich. We already presented the two most fundamental sampling schemes above. We present a few more here. In each case, we consider an urn with a finite number of balls of different colors.

Pólya urn model [11] In this sampling scheme, after each draw not only is the ball returned to the urn but together with an additional ball of the same color.

Problem 2.22. Consider an urn with r red balls and b blue balls. Show by example, as in Section 2.4.2, that the probability of any outcome sequence of length n with exactly k red balls is

$$\frac{r(r+1)\cdots(r+k-1)b(b+1)\cdots(b+n-k-1)}{(r+b)(r+b+1)\cdots(r+b+n-1)}.$$

[11] Named after George Pólya (1887 - 1985).

(Recall Remark 2.19.) Thus, once again, the central quantity is the number of red balls.

Moran urn model [12] In this sampling scheme, at each stage two balls are drawn: the first ball is returned to the urn together with an additional ball of the same color, while the second ball is not returned to the urn.

Note that if at some stage all the balls in the urn are of the same color, then the urn remains constant forever after. This can be shown to happen eventually and a question of interest is to compute the probability that the urn becomes all red.

Problem 2.23. Argue that if $r = b$, then that probability is $1/2$. In fact, argue that, if $\tau(r, b)$ denotes the probability that the process starting with r red and b blue balls ends up with only red balls, then $\tau(r, b) = 1 - \tau(b, r)$.

Problem 2.24. Derive the probabilities $\tau(1, 2)$, $\tau(2, 3)$, and $\tau(3, 4)$.

Wright–Fisher urn model [13] Assume that the urn contains N balls in total. In this sampling scheme, at each step the entire urn is reconstituted by sampling N balls uniformly at random with replacement from the urn.

Problem 2.25. Start with an urn with r red balls and b blue balls. Give the distribution of the number of red balls does the urn contain after one step.

Remark 2.26. The Wright–Fisher and the Moran urn models were proposed as models of genetic drift, which is the change in the frequency of gene variants (i.e., alleles) in a given population. In both models, the size of the population remains constant.

2.5 Further topics

2.5.1 Stirling's formula

The factorial, as a function on the integers, increases very rapidly.

Problem 2.27. Prove that the factorial sequence $(n!)$ increases faster to infinity than any power sequence, meaning that $a^n/n! \to 0$ for any real number $a > 0$. Moreover, show that $n! \leq n^n$ for $n \geq 1$.

In fact, the size of $n!$ is known very precisely. The following describes the first order asymptotics. More refined results exist.

[12] Named after Patrick Moran (1917 - 1988).

[13] Named after Sewall Wright (1889 - 1988) and Ronald Aylmer Fisher (1890 - 1962).

Theorem 2.28 (Stirling's formula[14])**.** *Letting e denote the Euler number,*

$$n! \sim \sqrt{2\pi n}\,(n/e)^n, \quad \text{as } n \to \infty. \tag{2.21}$$

In fact,

$$1 \le \frac{n!}{\sqrt{2\pi n}\,(n/e)^n} \le e^{1/(12n)}, \quad \text{for all } n \ge 1. \tag{2.22}$$

2.5.2 More on binomial coefficients

Binomial coefficients appear in many important combinatorial identities. Here are a few examples.

Problem 2.29. Show that there are $\binom{n+1}{k}$ binary sequences with exactly k ones and n zeros such that no two 1's are adjacent.

Problem 2.30 (The binomial identity)**.** Prove that

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}, \quad \text{for } a, b \in \mathbb{R}.$$

[One way to do so uses the fact that the binomial distribution needs to satisfy the 2nd axiom.]

[14] Named after James Stirling (1692 - 1770).

Problem 2.31. Show that

$$2^n = \sum_{k=0}^{n} \binom{n}{k}.$$

This can be done by interpreting this identity in terms of the number of subsets of $\{1, \dots, n\}$.

Partitions of an integer Consider the number of ways of decomposing a non-negative integer m into a sum of $s \ge 1$ non-negative integers. Importantly, we count different permutations of the same numbers as distinct possibilities. For example, here are the possible decompositions of $m = 4$ into $s = 3$ non-negative integers

$4+0+0$	$3+1+0$	$3+0+1$	$2+2+0$	$2+1+1$
$2+0+2$	$1+3+0$	$1+2+1$	$1+1+2$	$1+0+3$
$0+4+0$	$0+3+1$	$0+2+2$	$0+1+3$	$0+0+4$

Problem 2.32. Show that this number is equal to $\binom{m+s-1}{s-1}$. How does this change when the integers in the partition are required to be positive?

Catalan numbers Closely related to the binomial coefficients are the *Catalan numbers.*[15] The nth Catalan

[15] Named after Eugène Catalan (1814 - 1894).

number is defined as

$$C_n := \frac{1}{n+1}\binom{2n}{n}. \tag{2.23}$$

These numbers have many different interpretations of their own. One of them is that C_n is the number of balanced bracket expressions of length $2n$. Here are all such expressions of length 6 ($n = 3$):

$$()()() \quad (()()) \quad ()(()) \quad (())() \quad (()())$$

Problem 2.33. Prove that

$$C_n = \binom{2n}{n} - \binom{2n}{n+1}.$$

Problem 2.34. Prove the recursive formula

$$C_0 = 1, \quad C_{n+1} = \sum_{i=0}^{n} C_i\, C_{n-i}, \quad n \geq 0. \tag{2.24}$$

2.5.3 Two Envelopes Problem

Two envelopes containing money are placed in front of you. You are told that one envelope contains double the amount of the other. You are allowed to choose an envelope and look inside, and based on what you see you have to decide whether to keep the envelope that you just opened or switch for the other envelope. See [173] for an article-length discussion including different perspectives.

A flawed reasoning goes as follows.

> *If you see x in the envelope, then the amount in the other envelope is either $x/2$ or $2x$, each with probability $1/2$. The average gain if you switch is therefore $(1/2)(x/2) + (1/2)(2x) = (5/4)x$, so you should switch.*

The issue is that there are no grounds for the "with probability 1/2" claim since the distribution that generated x was not specified.

This illustrates the maxim (echoed in [113, Exa 4.28]).

Proverb. *Computing probabilities requires a well-defined probability model.*

See Problem 7.104 and Problem 7.105 for two different probability models for this situation that lead to different conclusions.

2.6 Additional problems

Problem 2.35. A rule of thumb in Epidemiology is that, in the context of examining the safety of a given drug, if one hopes to identify a (severe) side effect affecting 1 out

of every 1000 people taking the drug, then a trial needs to include at least 3000 individuals. In that case, what is the probability that in such a trial at least one person will experience the side effect?

Problem 2.36 (With or without replacement)**.** Suppose that we are sampling n balls with replacement from an urn containing N balls numbered $1, \ldots, N$. Compute the probability that all balls drawn are distinct. Now consider an asymptotic setting where $n = n(N)$ and $N \to \infty$, and let q_N denote that probability. Show that

$$\lim_{N\to\infty} q_N = \begin{cases} 0 & \text{if } n/\sqrt{N} \to \infty, \\ 1 & \text{if } n/\sqrt{N} \to 0. \end{cases}$$

Problem 2.37 (A stylized Birthday Problem)**.** Compute the minimum number of people, taken at random from those born in the year 2000, needed so that at least two share their birthday with probability at least 1/2. Model the situation using Bernoulli trials and assume that the a person is equally likely to be born any given day of a 365-day year.

Problem 2.38. Continuing with Problem 2.37, perform simulations in R to confirm your answer.

Problem 2.39 (More on the Birthday Problem)**.** The use of Bernoulli trials to model the situation in Problem 2.37 amounts to assuming that 1) each person is equally likely to be born any day of the year and 2) that the population is very large. Both are approximations. Keeping 2) in place, show that 1) only makes the number of required people larger.

Problem 2.40. Consider two independent draws with replacement from an urn containing N distinct items. Let p_i denote the probability of drawing item i. What is the probability that the same item is drawn twice? Show that this is minimized when the p_i are all equal, meaning, when drawing uniformly at random. [Use the method of Lagrange multipliers.]

Problem 2.41. Suppose that you have access to a computer routine that takes as input (n, N) and generates n independent draws with replacement from an urn with balls numbered $\{1, \ldots, N\}$.

(i) Explain how you would use that routine to generate n draws from an urn with r red balls and b blue balls with replacement.

(ii) Explain how you would use that routine to generate n draws from an urn with r red balls and b blue balls without replacement.

First answer these questions in writing. Then answer them by writing an program in R for each situation, using

the R function runid as the routine. (This is only meant for pedagogical purposes since the function sample can be directly used to fulfill the purpose in both cases.)

Problem 2.42 (Simpson's reversal)**.** Provide a simple example of a finite probability space $(\Omega, \mathbb{P})$ and events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ such that

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B}) < \mathbb{P}(\mathcal{A} \mid \mathcal{B}^{\mathsf{c}}),$$

while

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B} \cap \mathcal{C}) \geq \mathbb{P}(\mathcal{A} \mid \mathcal{B}^{\mathsf{c}} \cap \mathcal{C})$$

and

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B} \cap \mathcal{C}^{\mathsf{c}}) \geq \mathbb{P}(\mathcal{A} \mid \mathcal{B}^{\mathsf{c}} \cap \mathcal{C}^{\mathsf{c}}).$$

Show that this is not possible when $\mathcal{B}$ and $\mathcal{C}$ are independent of each other.

Problem 2.43 (The Two Children)**.** This is a classic problem that appeared in [101]. You are on the airplane and start a conversation with the person next to you. In the course of the conversation, you learn that (I) the person has two children; (II) one of them is a daughter; (III) and she is the oldest. After (I), what is the probability that the person has two daughters? How does that change after (II)? How does that change after (III)? [Make some necessary simplifying assumptions.]

Problem 2.44. Simulate 5 realizations of an experiment consisting of sampling with replacement $n = 10$ times from an urn containing $r = 7$ red balls and $b = 13$ blue balls. Repeat, now without replacement.

Problem 2.45. Write an R function polya that takes in a sequence length n, and the composition of the initial urn in terms of numbers of red and blue balls, r and b, and generates a sequence of that length from the Pólya process starting from that urn. Call the function on $(n, r, b) = (200, 5, 3)$ a large number of times, say $M = 1000$, each time compute the number of red balls in the resulting sequence, and tabulate the fraction of times that this number is equal to k, for all $k \in \{0, \ldots, 200\}$. Plot the corresponding bar chart.

Problem 2.46. Write an R function moran that takes in the composition of the urn in terms of numbers of red and blue balls, r and b, and runs the Moran urn process until the urn is of one color and returns that color and the number of stages that it took to get there. (You may want to bound the number of stages and then stop the process after that, returning a symbol indicating non-convergence.) Use that function to confirm your answers to Problem 2.24 following the guidelines of Problem 2.45.

CHAPTER 3

DISCRETE DISTRIBUTIONS

Throughout the chapter, we consider a discrete probability space $(\Omega, \Sigma, \mathbb{P})$, where Σ is taken to be the power set of Ω, as usual.

3.1 Random variables

It is often the case that measurements are taken from the experiment. Such a measurement is modeled as a function on the sample space. More formally, a *random variable* on Ω is a real-valued function

$$X : \Omega \to \mathbb{R}. \tag{3.1}$$

Remark 3.1. We will abbreviate $\{\omega \in \Omega : X(\omega) \in \mathcal{U}\}$ as $\{X \in \mathcal{U}\}$. In particular, $\{X = x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) = x\}$ and, similarly, $\{X \le x\}$ is shorthand for $\{\omega \in \Omega : X(\omega) \le x\}$, etc.

3.2 Discrete distributions

A random variable X on a discrete probability space $(\Omega, \Sigma, \mathbb{P})$ defines a distribution on $\mathbb{R}$ equipped with its power set,

$$\mathsf{P}_X(\mathcal{U}) := \mathbb{P}(X \in \mathcal{U}), \quad \text{for } \mathcal{U} \subset \mathbb{R}, \tag{3.2}$$

with mass function

$$f_X(x) := \mathbb{P}(X = x), \quad \text{for } x \in \mathbb{R}. \tag{3.3}$$

Problem 3.2. Show that P_X is a *discrete distribution* in the sense that there is a discrete subset $\mathcal{S} \subset \mathbb{R}$ such that $\mathsf{P}_X(\mathcal{U}) = 0$ whenever $\mathcal{U} \cap \mathcal{S} = \varnothing$. In this case, $\mathcal{S}$ corresponds to the *support* of P_X.

Remark 3.3. For a random variable X and a distribution P, we write $X \sim \mathsf{P}$ when X has distribution P, meaning that $\mathsf{P}_X = \mathsf{P}$.

This chapter presents some classical examples of discrete distributions on the real line. In fact, we already saw some of them in Chapter 2.

Remark 3.4. Unless specified otherwise, a discrete random variable will be assumed to take integer values. This can be done without loss of generality, and it also covers most cases of interest.

3.3 Binomial distributions

Consider the setting of Bernoulli trials as in Section 2.3 where a p-coin is tossed repeatedly n times. Unless otherwise stated, we assume that the tosses are independent. Letting $\omega = (\omega_1, \dots, \omega_n)$ denote an element of $\Omega := \{\mathrm{H}, \mathrm{T}\}^n$, for each $i \in \{1, \dots, n\}$, define

$$X_i(\omega) = \begin{cases} 1 & \text{if } \omega_i = \mathrm{H}, \\ 0 & \text{if } \omega_i = \mathrm{T}. \end{cases} \tag{3.4}$$

(Note that X_i is the indicator of the event $\mathcal{A}_i$ defined in Section 2.3.) In particular, the distribution of X_i is given by

$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = 0) = 1 - p. \tag{3.5}$$

X_i has the so-called *Bernoulli distribution* with parameter p. We will denote this distribution by $\mathrm{Ber}(p)$.

The X_i are *independent* (discrete) random variables in the sense that

$$\mathbb{P}(X_1 \in \mathcal{V}_1, \dots, X_r \in \mathcal{V}_r) = \mathbb{P}(X_1 \in \mathcal{V}_1) \cdots \mathbb{P}(X_r \in \mathcal{V}_r), \quad \forall \mathcal{V}_1, \dots, \mathcal{V}_r \subset \mathbb{R}, \forall r \geq 2,$$

or, equivalently,

$$\mathbb{P}(X_1 = x_1, \dots, X_r = x_r) = \mathbb{P}(X_1 = x_1) \cdots \mathbb{P}(X_r = x_r), \quad \forall x_1, \dots, x_r \in \mathbb{R}, \forall r \geq 2.$$

(In this particular case, the x_i can be taken to be in $\{0,1\}$.) We will discuss independent random variables in more detail in Section 6.2.

Let Y denote the number of heads in the sequence of n tosses, so that

$$Y = \sum_{i=1}^{n} X_i. \tag{3.6}$$

We note that Y is a random variable on the same sample space Ω. Y has the so-called *binomial distribution* with parameters (n,p). We will denote this distribution by $\mathrm{Bin}(n,p)$.

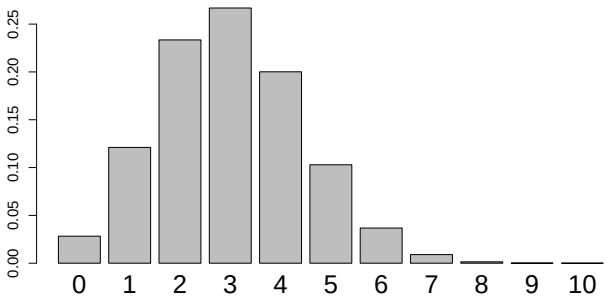
We already saw in Proposition 2.8 that Y plays a central role in this experiment. And in (2.16), we derived its distribution.

Proposition 3.5 (Binomial distribution)**.** *The binomial distribution with parameters (n,p) has mass function*

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k \in \{0,\dots,n\}. \tag{3.7}$$

Discrete mass functions are often drawn as bar plots. See Figure 3.1 for an illustration.

Figure 3.1: A bar plot of the mass function of the binomial distribution with parameters $n = 10$ and $p = 0.3$.



3.4 Hypergeometric distributions

Consider an urn model as in Section 2.4. Suppose, as before, that the urn has r red balls and b blue balls. We sample from the urn n times and, as before, let $X_i = 1$ if the ith draw is red, and $X_i = 0$ otherwise. (Note that X_i is the indicator of the event $\mathcal{A}_i$ defined in Section 2.4.) If we sample with replacement, we know that the experiment corresponds to Bernoulli trials with parameter $p := r/(r+b)$. We assume therefore that we are sampling without replacement. To be able to sample n times without replacement, we need to assume that $n \le r+b$.

Let Y denote the number of heads in a sequence of n

draws, exactly as in (3.6). The difference is that here the draws (the X_i) are not independent. The distribution of Y is called the *hypergeometric distribution*[16] with parameters (n, r, b). We will denote this distribution by $\text{Hyper}(n, r, b)$.

We already computed its mass function in (2.20).

Proposition 3.6 (Hypergeometric distribution)**.** *The hypergeometric distribution with parameters (n, r, b) has mass function*

$$f(k) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{r+b}{n}}, \quad k \in \{0, \dots, \min(n, r)\}. \tag{3.8}$$

3.5 Geometric distributions

Consider Bernoulli trials as in Section 3.3 but now assume that we toss the p-coin until it lands heads. This experiment was described in Example 1.12. Define the X_i as before, and let Y denote the number of tails until the first heads. For example, $Y(\omega) = 3$ when $\omega = \text{TTTH}$. Note that Y is a random variable on Ω.

[16] There does not seem to be a broad agreement on how to parameterize this family of distributions.

It is particularly straightforward to derive the distribution of Y. Indeed, for any integer $k \geq 0$,

$$\begin{aligned}\mathbb{P}(Y = k) &= \mathbb{P}(X_1 = 0, \dots, X_k = 0, X_{k+1} = 1) \\ &= \mathbb{P}(X_1 = 0) \times \cdots \times \mathbb{P}(X_k = 0) \times \mathbb{P}(X_{k+1} = 1) \\ &= \underbrace{(1-p) \times \cdots \times (1-p)}_{k \text{ times}} \times p = (1-p)^k p,\end{aligned}$$

using the independence of the X_i in the second line.

The distribution of Y is called the *geometric distribution*[17] with parameter p. We will denote this distribution by $\text{Geom}(p)$. It is supported on $\{0, 1, 2, \dots\}$.

Problem 3.7. Because of the Law of Total Probability,

$$\sum_{k=0}^{\infty} (1-p)^k p = 1, \quad \text{for all } p \in (0, 1). \tag{3.9}$$

Prove this directly.

Problem 3.8. Show that $\mathbb{P}(Y > k) = (1-p)^{k+1}$ for $k \geq 0$ integer.

Remark 3.9 (Law of truly large numbers)**.** In [60], Diaconis and Mosteller introduced this principle as one possible

[17] This distribution is sometimes defined a bit differently, as the number of trials, including the last one, until the first heads. This is the case, for example, in [113].

source for coincidences. In their own words, the 'law' says that

> *When enormous numbers of events and people and their interactions cumulate over time, almost any outrageous event is bound to occur.*

Related concepts include *Murphy's law*, *Littlewood's law*, and the *Infinite Monkey Theorem*. Mathematically, the principle can be formalized as the following theorem: *If a p-coin, with $p > 0$, is tossed repeatedly independently, it will land heads eventually.* This theorem is an immediate consequence of (3.9).

Problem 3.10 (Memoryless property)**.** Prove that a geometric random variable Y satisfies

$$\mathbb{P}(Y > k + t \mid Y > k) = \mathbb{P}(Y > t), \quad \text{for all } t \geq 0 \text{ and all } k \geq 0.$$

3.6 Other discrete distributions

We already saw the families of Bernoulli, of binomial, and of hypergeometric distributions. We introduce a few more.

3.6.1 Discrete uniform distributions

A *discrete uniform distribution* (on the real line) is a uniform on a finite set of points in $\mathbb{R}$. Thus the family is parameterized by finite sets of points: such a set, say $\mathcal{X} \subset \mathbb{R}$, defines the distribution with mass function

$$f(x) = \frac{\{x \in \mathcal{X}\}}{|\mathcal{X}|}, \quad x \in \mathbb{R}.$$

The subfamily corresponding to sets of the form $\mathcal{X} = \{1, \ldots, N\}$ plays a special role. This subfamily is obviously much smaller and can be parameterized by the positive integers.

3.6.2 Negative binomial distributions

Consider an experiment where we toss a p-coin repeatedly until the mth heads, where $m \geq 1$ is given. Let Y denote the number of tails until we stop. For example, if $m = 3$, then $Y(\omega) = 4$ when $\omega = \text{HTHTTTH}$. Y is clearly a random variable on the same sample space, and has the so-called *negative binomial distribution*[18] with parameters (m, p). We will denote this distribution by $\text{NegBin}(m, p)$. It is

[18] This distribution is sometimes defined a bit differently, as the number of trials, including the last one, until the mth heads. This is the case, for example, in [113].

supported on $k \in \{0, 1, 2, \dots\}$. Clearly, $\text{NegBin}(1, p) = \text{Geom}(p)$, so the negative binomial family includes the geometric family.

Proposition 3.11. *The negative binomial distribution with parameters (m, p) has mass function*

$$f(k) = \binom{m+k-1}{m-1}(1-p)^k p^m, \quad k \geq 0.$$

Problem 3.12. Prove this last proposition. The arguments are very similar to those leading to Proposition 3.5.

Proposition 3.13. *The sum of m independent random variables, each having the geometric distribution with parameter p, has the negative binomial distribution with parameters (m, p).*

Problem 3.14. Prove this last proposition.

3.6.3 Negative hypergeometric distributions

As the name indicates, this distribution arises when, instead of flipping a coin, we draw without replacement from an urn. Assume the urn contains r red balls and b blue balls. Let Y denote the number of blue balls drawn before drawing the mth red ball, where $m \leq r$. Y is a random variable on the same sample space, and has the so-called *negative hypergeometric distribution* with parameters (m, r, b).

Problem 3.15. Derive the mass function of the negative hypergeometric distribution with parameters $(m, r, b) = (3, 4, 5)$.

3.6.4 Poisson distributions

The *Poisson distribution*[19] with parameter $\lambda \geq 0$ is given by the following mass function

$$f(k) := \mathrm{e}^{-\lambda}\frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots.$$

By convention, $0^0 = 1$, so that when $\lambda = 0$, the right-hand side is 1 at $k = 0$ and 0 otherwise.

Problem 3.16. Show that this is indeed a mass function on the non-negative integers.[20]

Proposition 3.17 (Stability of the Poisson family)**.** *The sum of a finite number of independent Poisson random variables is Poisson.*

The Poisson distribution arises when counting rare events. This is partly justified by Theorem 3.18.

[19] Named after Siméon Poisson (1781 - 1840).

[20] Recall that $\mathrm{e}^x = \sum_{j\geq 0} x^j/j!$ for all x.

3.7 Law of Small Numbers

Suppose that we are counting the occurrence of a rare phenomenon. For a historical example, Gosset[21] was counting the number of yeast cells using an hemocytometer [226]. This is a microscope slide subdivided into a grid of identical units that can hold a solution. In his experiments, Gosset prepared solutions containing the cells. Each solution was well mixed and spread on the hemocytometer. He then counted the number of cells in each unit. He wanted to understand the distribution of these counts. He performed a number of experiments. One of them is shown in Table 3.1.

Mathematically, he reasoned as follows. Let N denote the number of units and n the number of cells. When the solution is well mixed and well spread out over the units, each cell can be assumed to fall in any unit with equal probability $1/N$. Under this model, the number of cells found in a given unit has the binomial distribution with parameters $(n, 1/N)$. Gosset considered the limit where n and N are both large and proved that this distribution

[21] William Sealy Gosset (1876 - 1937) was working at the Guinness brewery, which required that he publish his work anonymously so as not to disclose the fact that he was working for Guinness and that his work could be used in the beer brewing business. He chose the pseudonym 'Student'.

Table 3.1: The following is Table 2 in [226]. There were 103 units with 0 cells, 143 units with 1 cell, etc. See Problem 12.30 for a comparison with what is expected under a Poisson model.

Number of cells	0	1	2	3	4	5	6
Number of units	103	143	98	42	8	4	2

is 'close' to the Poisson distribution with parameter n/N when that number is not too large. This approximation is sometimes referred to as the *Law of Small Numbers*, and is formalized below.

Theorem 3.18 (Poisson approximation to the binomial distribution)**.** *Consider a sequence* (p_n) *with* $p_n \in [0,1]$ *and* $np_n \to \lambda$ *as* $n \to \infty$*. Then if* Y_n *has distribution* $Bin(n, p_n)$ *and* $k \geq 0$ *is an integer,*

$$\mathbb{P}(Y_n = k) \to e^{-\lambda}\frac{\lambda^k}{k!}, \quad \text{as } n \to \infty.$$

Problem 3.19. Prove Theorem 3.18 using Stirling's formula (2.21).

Bateman arrived at the same conclusion in the context of experiments conducted by Rutherford and Geiger in the early 1900's to better understand the decay of radioactive

particles. In one such experiment [201], they counted the number of particles emitted by a film of polonium in 2608 time intervals of 7.5 seconds duration. The data is reproduced here in Table 3.2.

3.8 Coupon Collector Problem

This problem arises from considering an individual collecting coupons of a certain type, say of players in a certain sports league in a certain sports season. The collector progressively completes his collection by buying envelopes, each containing an undisclosed coupon. With every purchase, the collector hopes the enclosed coupon will be new to his collection. (We assume here that the collector does not trade with others.) If there are N players in the league that season (and therefore that many coupons to collect), how many envelopes would the collector need to purchase in order to complete his collection?

In the simplest setting, an envelope is equally likely to contain any one of the N distinct coupons. In that case, the situation can be modeled as a probability experiment where balls are drawn repeatedly with replacement from an urn containing N balls, all distinct, until all the balls in the urn have been drawn at least once. For example, if $N = 10$, the sequence of draws might look like this

3 6 9 9 9 5 7 9 5 8 3 2 1 2 5 2 7 10 3 3 10 1 8 7 9 1 6 4

Let T denote the length of the resulting sequence ($T = 28$ in this particular realization of the experiment).

Problem 3.20. Write a function in R taking in N and returning a realization of the experiment. [Use the function sample and a repeat statement.] Run your function on $N = 10$ a few times to get a sense of how typical outcomes look like.

Problem 3.21. Let $T_0 = 0$, and for $i \in \{1, \ldots, N\}$, let T_i denote the number of balls needed to secure i distinct balls, so that $T = T_N$, and define $W_i = T_i - T_{i-1}$. Show that $W_1, \ldots, W_N$ are independent. Then derive the distribution of W_i.

Problem 3.22. Write a function in R taking in N and returning a realization of the experiment, but this time based on Problem 3.21. Compare this function and that of Problem 3.20 in terms of computational speed. [The function proc.time will prove useful.]

Problem 3.23. For $i \in \{1, \ldots, N\}$, let X_i denote the number of trials it takes to draw ball i. Note that $T = \max\{X_i : 1 \le i \le N\}$.

Table 3.2: The following is part of the table on page 701 of [201]. The counting of particles was done over 2608 time intervals of 7.5 seconds each. See Problem 12.31 for a comparison with what is expected under a Poisson model.

Number of particles	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15+
Number of intervals	57	203	383	525	532	408	273	139	45	27	10	4	0	1	1	0

(i) Show that

$$T \le t \iff \{X_i \le t,\ \forall i = 1, \dots, N\}.$$

(ii) What is the distribution of X_i?

(iii) For $n \ge 1$, and for $k \in \{1, \dots, N\}$ and any $1 \le i_1 < \cdots < i_k \le N$, compute

$$\mathbb{P}(X_{i_1} \ge n, \dots, X_{i_k} \ge n).$$

(iv) Use this and the inclusion-exclusion formula (1.15) to derive the mass function of T in closed form.

3.9 Additional problems

Problem 3.24. Show that for any $n \ge 1$ integer and any $p \in [0,1]$,

$$\text{Bin}(n, 1-p) \text{ coincides with } n - \text{Bin}(n,p). \tag{3.10}$$

Problem 3.25. Let Y be binomial with parameters $(n, 1/2)$. Using the symmetry (3.10), show that

$$\mathbb{P}(Y > n/2) = \mathbb{P}(Y < n/2). \tag{3.11}$$

This means that, when the coin is fair, the probability of getting strictly more heads than tails is the same as the probability of getting strictly more tails than heads. When n is odd, show that (3.11) implies that

$$\mathbb{P}(Y > n/2) = \mathbb{P}(Y < n/2) = \frac{1}{2}.$$

When n is even, show that (3.11) implies that

$$\mathbb{P}(Y > n/2) = \frac{1}{2} + \frac{1}{2}\mathbb{P}(Y = n/2).$$

Then using Stirling's formula (2.21), show that

$$\mathbb{P}(Y = n/2) \sim \sqrt{\frac{2}{\pi n}}, \quad \text{as } n \to \infty.$$

This approximation is in fact very good. Verify this numerically in R.

Problem 3.26. For $y \in \{0, \dots, n\}$, let $\mathsf{F}_{n,\theta}(y)$ denote the probability that the binomial distribution with parameters (n, θ) puts on $\{0, \dots, y\}$. Fix $y < n$ and show that

$$\theta \mapsto \mathsf{F}_{n,\theta}(y) \text{ is strictly decreasing, continuous, and one-to-one as a map of } [0,1] \text{ to itself.} \tag{3.12}$$

What happens when $y = n$?

Problem 3.27. Continuing with the setting of Problem 3.26, show that for any $y \geq 0$ integer and any $\theta \in [0,1]$,

$$n \mapsto \mathsf{F}_{n,\theta}(y) \text{ is non-increasing.} \tag{3.13}$$

In fact, if $0 < \theta < 1$, this function is decreasing once $n \geq y$.

Problem 3.28. Suppose that you have access to a computer routine that takes as input a vector of any length k of numbers in $[0,1]$, say $(q_1, \dots, q_k)$, and generates $(B_1, \dots, B_k)$ independent Bernoulli with these parameters (i.e., $B_i \sim \mathrm{Ber}(q_i)$). The question is how to use this routine to generate a random variable from a given mass function f (with finite support). Assume that f is supported on $x_1, \dots, x_N$ and that $f(x_j) = p_j$.

(i) Quickly argue that the case $N = 2$ is trivial.

(ii) Consider the case $N = 3$. Show that the following works. Assume without loss of generality that $p_1 \leq p_2 \leq p_3$. Apply the routine to $q_1 = p_1$ and $q_2 = p_2/(1-p_1)$ obtaining (B_1, B_2). If $B_1 = 1$, return x_1; if $B_1 = 0$ and $B_2 = 1$, return x_2; otherwise, return x_3.

(iii) Extend this procedure to the general case.

Problem 3.29. Show that the sum of independent binomial random variables with same probability parameter p is also binomial with probability parameter p.

Problem 3.30. Prove Proposition 3.17.

Problem 3.31. Suppose that X and Y are two independent Poisson random variables. Show that the distribution of X conditional on $X + Y = t$ is binomial and specify the parameters.

Problem 3.32. For any $p \in (0,1)$, show that there is $c_p > 0$ such that the following is a mass function on the positive integers

$$f(k) = c_p \frac{p^k}{k}, \quad k \geq 1 \text{ integer.}$$

Derive c_p in closed form. [22]

[22] Recall that $\log(1-x) = -\sum_{k \geq 1} x^k/k$ for all $x \in (0,1)$.

Problem 3.33. For what values of α can one normalize $g(k) = k^{-\alpha}$ into a mass function on the positive integers? Similarly, for what values of α and β can one normalize $g(k) = k^{-\alpha}(\log(k+1))^{-\beta}$ into a mass function on the positive integers?

Problem 3.34 (The binomial approximation to the hypergeometric distribution)**.** Problem 2.21 asks you to prove that sampling without replacement from an (r, b)-urn amounts, in the limit where $r/(r+b) \to p$, to tossing a p-coin. Argue that, therefore, the hypergeometric distribution with parameters (n, r, b) must approach the binomial distribution with parameters (n, p) when n is fixed and $r/(r+b) \to p$. [Argue in terms of mass functions.]

Problem 3.35. Continuing with the same problem, answer the same question analytically using Stirling's formula.

Problem 3.36 (Game of Googol)**.** Martin Gardner posed the following puzzle in his column in a 1960 edition of *Scientific American*: "Ask someone to take as many slips of paper as he pleases, and on each slip write a different positive number. The numbers may range from small fractions of 1 to a number the size of a googol[23] or even larger. These slips are turned face down and shuffled over the top of a table. One at a time you turn the slips face up. The aim is to stop turning when you come to the number that you guess to be the largest of the series. You cannot go back and pick a previously turned slip. If you turn over all the slips, then of course you must pick the last one turned."

Let n be the total number of slips. A natural strategy is, for a given $r \in \{1, \ldots, n\}$, to turn r slips, and then keep turning slips until either reaching the last one (in which case this is our final slip) or stop when the slip shows a number that is at least as large as the largest number among the first r slips.

(i) Compute the probability that this strategy is correct as a function of n and r.

(ii) Let r_n denote the optimal choice of r as a function of n. (If there are several optimal choices, it is the smallest.) Compute r_n using R.

(iii) Formally derive r_n to first order when $n \to \infty$.

(This problem has a long history [84].)

[23] A *googol* is defined as 10^{100}.

CHAPTER 4

DISTRIBUTIONS ON THE REAL LINE

In Chapter 3, which considered a discrete sample space (equipped, as usual, with its power set as σ-algebra), and random variables were simply defined as arbitrary real-valued functions on that space. When the sample space is not discrete, more care is needed. In fact, a proper definition necessitates the introduction of a particular σ-algebra other than the power set.

As before, we consider a probability space, $(\Omega, \Sigma, \mathbb{P})$, modeling a certain experiment.

4.1 Random variables

Consider a measurement on the outcome of the experiment. At the very minimum, we will want to compute the probability that the measurement does not exceed a certain amount. For this reason, we say that a real-valued function $X : \Omega \to \mathbb{R}$ is a *random variable* on (Ω, Σ) if

$$\{X \le x\} \in \Sigma, \quad \text{for all } x \in \mathbb{R}. \tag{4.1}$$

(Recall the notation introduced in Remark 3.1.) In particular, if X is a random variable on (Ω, Σ), and $\mathbb{P}$ is a probability distribution on Σ, then we can ask for the corresponding probability that X is bounded by x from above, in other words, $\mathbb{P}(X \leq x)$ is well-defined.

4.2 Borel σ-algebra

In Chapter 3 we saw that a discrete random variable defines a (discrete) distribution on the real line equipped with its power set. While this can be done without loss of generality in that context, beyond it is better to equip the real line with a smaller σ-algebra. (Equipping the real line with its power set would effectively exclude most distributions commonly used in practice.)

At the very minimum, because of (4.1), we require the σ-algebra (over $\mathbb{R}$) to include all sets of the form

$$(-\infty, x], \quad \text{for } x \in \mathbb{R}. \tag{4.2}$$

The *Borel σ-algebra*[24], denoted $\mathcal{B}$ henceforth, is the σ-algebra generated by these intervals, meaning, the smallest σ-algebra over $\mathbb{R}$ that contains all such intervals (Problem 1.51).

[24] Named after Émile Borel (1871 - 1956).

Proposition 4.1. *The Borel σ-algebra $\mathcal{B}$ contains all intervals, as well as all open sets and all closed sets.*

Proof. We only show that $\mathcal{B}$ contains all intervals. For example, take $a < b$. Since $\mathcal{B}$ contains $(-\infty, a]$ and $(-\infty, b]$, it must contain $(-\infty, a]^{\mathsf{c}}$ by (1.2) and also

$$(-\infty, a]^{\mathsf{c}} \cap (-\infty, b],$$

by (1.3). But this is $(a, b]$. Therefore $\mathcal{B}$ contains all intervals of the form $(a, b]$, where $a = -\infty$ and $b = \infty$ are allowed.

Take an interval of the form $(-\infty, x)$. Note that it is open on the right. Define $\mathcal{U}_n = (-\infty, x - 1/n]$. By assumption, $\mathcal{U}_n \in \mathcal{B}$ for all n. Because of (1.4), $\mathcal{B}$ must also contain their union, and we conclude with the fact that

$$\bigcup_{n \geq 1} \mathcal{U}_n = (-\infty, x).$$

Now that we know that $\mathcal{B}$ contains all intervals of the form $(-\infty, x)$, we can reason as before and show that it must contain all intervals of the form $[a, b)$, where $a = -\infty$ and $b = \infty$ are allowed.

Finally, for any $-\infty \leq a < b \leq \infty$,

$$[a, b] = [a, d) \cup (c, b], \quad \text{and} \quad (a, b) = (a, c] \cup [c, b),$$

for any c, d such that $a < c < d < b$, so that $\mathcal{B}$ must also include intervals of the form $[a, b]$ and (a, b). □

We will say that a function $g : \mathbb{R} \to \mathbb{R}$ is *measurable* if

$$g^{-1}(\mathcal{V}) \in \mathcal{B}, \quad \text{for all } \mathcal{V} \in \mathcal{B}. \tag{4.3}$$

4.3 Distributions on the real line

When considering a probability distribution on the real line, we will always assume that it is defined on the Borel σ-algebra.

The *support* of a distribution P on $(\mathbb{R}, \mathcal{B})$ is the smallest closed set $\mathcal{A}$ such that $\mathsf{P}(\mathcal{A}) = 1$.

Problem 4.2. Show that a distribution on $(\mathbb{R}, 2^{\mathbb{R}})$ with discrete support is also a distribution on $(\mathbb{R}, \mathcal{B})$.

A random variable X on a probability space $(\Omega, \Sigma, \mathbb{P})$ defines a distribution on $(\mathbb{R}, \mathcal{B})$,

$$\mathsf{P}_X(\mathcal{U}) := \mathbb{P}(X \in \mathcal{U}), \quad \text{for } \mathcal{U} \in \mathcal{B}. \tag{4.4}$$

Note that $\{X \in \mathcal{U}\}$ is sometimes denoted by $X^{-1}(\mathcal{U})$.

The *range* of a random variable X on (Ω, Σ) is defined as

$$X(\Omega) := \{X(\omega) : \omega \in \Omega\}. \tag{4.5}$$

Problem 4.3. Show that the support of P_X is included in the range of X. When is the inclusion strict?

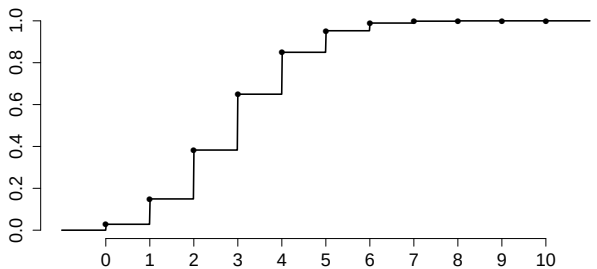
4.4 Distribution function

The *distribution function* (aka *cumulative distribution function*) of a distribution P on $(\mathbb{R}, \mathcal{B})$ is defined as

$$\mathsf{F}(x) := \mathsf{P}((-\infty, x]). \tag{4.6}$$

See Figure 4.1 for an example.

Figure 4.1: A plot of the distribution function of the binomial distribution with parameters $n = 10$ and $p = 0.3$.



Proposition 4.4. *A distribution is characterized by its distribution function in the sense that two distributions with identical distribution functions must coincide.*

Problem 4.5. Let F be the distribution function of a distribution P on $(\mathbb{R}, \mathcal{B})$.

(i) Prove that

$$\mathsf{F} \text{ is non-decreasing}, \tag{4.7}$$
$$\lim_{x\to-\infty} \mathsf{F}(x) = 0, \tag{4.8}$$
$$\lim_{x\to+\infty} \mathsf{F}(x) = 1. \tag{4.9}$$

(ii) Prove that F is continuous from the right, meaning

$$\lim_{t\searrow x} \mathsf{F}(t) = \mathsf{F}(x), \quad \text{for all } x \in \mathbb{R}. \tag{4.10}$$

[Use the 3rd probability axiom (1.7).]

It so happens that these properties above define a distribution function, in the sense that any function satisfying these properties is the distribution function of some distribution on the real line.

Theorem 4.6. *Let* $\mathsf{F} : \mathbb{R} \to [0,1]$ *satisfy* (4.7)-(4.10). *Then* F *defines a distribution*[25] P *on* $\mathcal{B}$ *via* (4.6). *In particular,*

$$\mathsf{P}((a,b]) = \mathsf{F}(b) - \mathsf{F}(a), \quad \textit{for } -\infty \le a < b \le \infty. \tag{4.11}$$

[25] The distribution P is known as the *Lebesgue–Stieltjes distribution* generated by F.

Problem 4.7. In the context of the last theorem, for $x \in \mathbb{R}$, define the *left limit* of F at x as

$$\mathsf{F}(x^-) := \lim_{t\nearrow x} \mathsf{F}(t). \tag{4.12}$$

Show that this limit is well defined. Then prove that

$$\mathsf{F}(x) - \mathsf{F}(x^-) = \mathsf{P}(\{x\}), \quad \text{for all } x \in \mathbb{R}. \tag{4.13}$$

Problem 4.8. Show that a distribution on the real line is discrete if and only if its distribution function is piecewise constant (i.e., staircase) with the set of discontinuities (i.e., jumps) corresponding to the support of the distribution.

Problem 4.9. Show that the set of points where a monotone function $\mathsf{F} : \mathbb{R} \to \mathbb{R}$ is discontinuous is countable.

The distribution function of a random variable X is simply the distribution function of its distribution P_X. It can be expressed as

$$\mathsf{F}_X(x) := \mathbb{P}(X \le x).$$

4.5 Survival function

Consider a distribution P on $(\mathbb{R}, \mathcal{B})$ with distribution function F. The *survival function* of P is defined as

$$\bar{\mathsf{F}}(x) := \mathsf{P}((x,\infty)). \tag{4.14}$$

Problem 4.10. Show that $\bar{\mathsf{F}} = 1 - \mathsf{F}$.

Problem 4.11. Show that a survival function $\bar{\mathsf{F}}$ is non-increasing, continuous from the right, and lower semi-continuous, meaning that

$$\liminf_{x \to x_0} \bar{\mathsf{F}}(x) \geq \bar{\mathsf{F}}(x_0), \quad \text{for all } x_0 \in \mathbb{R}. \tag{4.15}$$

The survival function of a random variable X is simply the survival function of its distribution P_X. It can be expressed as

$$\bar{\mathsf{F}}_X(x) := \mathbb{P}(X > x).$$

4.6 Quantile function

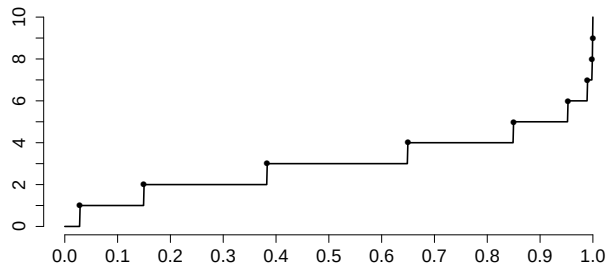
Consider a distribution P on $(\mathbb{R}, \mathcal{B})$ with distribution function F. We saw in Problem 4.8 that F may not be strictly increasing or continuous, in which case it does not admit an inverse in the usual sense. However, as a non-decreasing function, F admits the following form of pseudo-inverse [26]

$$\mathsf{F}^-(u) := \min\{x : \mathsf{F}(x) \geq u\}, \tag{4.16}$$

sometimes referred the *quantile function* of P. See Figure 4.2 for an example.

[26] That it is a minimum instead of an infimum is because of (4.24).

Figure 4.2: A plot of the distribution function of the binomial distribution with parameters $n = 10$ and $p = 0.3$.



Note that F^- is defined on $(0, 1)$, and if we allow it to return $-\infty$ and ∞ values, it can always be defined on $[0, 1]$.

Problem 4.12. Show that F^- is non-decreasing, continuous from the right, and

$$\mathsf{F}(x) \geq u \iff x \geq \mathsf{F}^-(u). \tag{4.17}$$

Problem 4.13. Show that

$$\mathsf{F}^-(u) = \sup\{x : \mathsf{F}(x) < u\}. \tag{4.18}$$

Problem 4.14. Define the following variant of the survival function

$$\tilde{\mathsf{F}}(x) = \mathsf{P}([x, \infty)). \tag{4.19}$$

Compare with (4.14), and note that the two definitions coincide when F is continuous. Show that

$$\mathsf{F}^-(u) = \inf\{x : \tilde{\mathsf{F}}(x) \le 1-u\}.$$

Deduce that

$$\tilde{\mathsf{F}}(x) \le 1-u \iff x \ge \mathsf{F}^-(u). \tag{4.20}$$

Quantiles We say that x is a *u-quantile* of P if

$$\mathsf{F}(x) \ge u \quad \text{and} \quad \tilde{\mathsf{F}}(x) \ge 1-u, \tag{4.21}$$

or equivalently, if X denotes a random variable with distribution P,

$$\mathbb{P}(X \le x) \ge u \quad \text{and} \quad \mathbb{P}(X \ge x) \ge 1-u.$$

With this definition, $x \in \mathbb{R}$ is a u-quantile for any

$$1 - \tilde{\mathsf{F}}(x) \le u \le \mathsf{F}(x).$$

Remark 4.15 (Median and other quartiles)**.** A 1/4-quantile is called a *1st quartile*, a 1/2-quantile is called *2nd quartile* or more commonly a *median*, and a 3/4-quantile is called a *3rd quartile*. The quartiles, together with other features, can be visualized using a *boxplot*.

Problem 4.16. Show that for any $u \in (0,1)$, the set of u-quantiles is either a singleton or an interval of the form $[a,b)$ for some $a < b$ that admit a simple characterization, where by convention $[a,a) = \{a\}$.

Problem 4.17. The previous problem implies that there always exists a u-quantile when $u \in (0,1)$. What happens when $u = 0$ or $u = 1$?

Problem 4.18. Show that $\mathsf{F}^-(u)$ is a u-quantile of F. Thus (reassuringly) the quantile function returns bona fide quantiles.

Remark 4.19. Other definitions of pseudo-inverse are possible, each leading to a possibly different notion of quantile. For example,

$$\mathsf{F}^{\boxminus}(x) := \sup\{x : \mathsf{F}(x) \le u\}, \tag{4.22}$$

and

$$\mathsf{F}^{\ominus}(u) := \frac{1}{2}\big(\mathsf{F}^-(u) + \mathsf{F}^{\boxminus}(u)\big). \tag{4.23}$$

Problem 4.20. Compare F^-, $\mathsf{F}^{\boxminus}$, and $\mathsf{F}^{\ominus}$. In particular, find examples of distributions where they are different, and also derive conditions under which they coincide.

4.7 Additional problems

Problem 4.21. Let F denote a distribution function.

(i) Prove that F takes values in $[0, 1]$.
(ii) Prove that F is upper semi-continuous, meaning

$$\limsup_{x \to x_0} \mathsf{F}(x) \leq \mathsf{F}(x_0), \quad \forall x_0 \in \mathbb{R}. \tag{4.24}$$

CHAPTER 5

CONTINUOUS DISTRIBUTIONS

In some areas of mathematics, physics, and elsewhere, continuous objects and structures are often motivated, or even defined, as limits of discrete objects. For example, in mathematics, the real numbers are defined as the limit of sequences of rationals, and in physics, the laws of thermodynamics arise as the number of particles in a system tends to infinity (the so-called thermodynamic or macroscopic limit).

In Chapter 3 we introduced and discussed discrete random variables and the (discrete) distributions they generate on the real line. Taking these discrete distributions to their continuous limits, which is done by letting their support size increase to infinity in a controlled manner, gives rise to continuous distributions on the real line.

In what follows, when we make probability statements, we assume that we have in the background a probability space, which by default will be denoted by $(\Omega, \Sigma, \mathbb{P})$. As usual, when Ω is discrete, Σ will be taken to be its power set. As in Chapter 4, we always equip $\mathbb{R}$ with its Borel

σ-algebra, denoted $\mathcal{B}$ and defined in Section 4.2.

5.1 From the discrete to the continuous

Some of the discrete distributions introduced in Chapter 3 have a 'natural' continuous limit when we let the size of their supports increase. We formalize this passage to the continuum by working with distribution functions. (Recall that a distribution on the real line is characterized by its distribution function.)

5.1.1 From uniform to uniform

For a positive integer N, let P_N denote the (discrete) uniform distribution on $\{1, \dots, N\}$, and let F_N denote its distribution function.

Problem 5.1. Show that, for any $x \in \mathbb{R}$,

$$\lim_{N\to\infty} \mathsf{F}_N(Nx) = \mathsf{F}(x) := \begin{cases} 0 & \text{if } x \le 0, \\ x & \text{if } 0 < x < 1, \\ 1 & \text{if } x \ge 1. \end{cases}$$

The limit function F above is continuous and satisfies the conditions of Theorem 4.6 and so defines a distribution, referred to as the *uniform distribution* on $[0,1]$; see Section 5.6 for more details.

Remark 5.2. Note that $\mathsf{F}_N(x) \to 0$ for all $x \in \mathbb{R}$, so that scaling x by N is crucial to obtain the limit above.

Remark 5.3. The family of discrete uniform distributions on the real line is much larger. It turns out that it is so large that it is in some sense dense among the class of all distributions on $(\mathbb{R}, \mathcal{B})$. You are asked to prove this in Problem 5.43.

5.1.2 From binomial to normal

The following limiting behavior of binomial distributions is one of the pillars of Probability Theory.

Theorem 5.4 (De Moivre–Laplace Theorem [27])**.** *Fix $p \in (0,1)$ and let F_n denote the distribution function of the binomial distribution with parameters (n,p). Then, for any $x \in \mathbb{R}$,*

$$\lim_{n\to\infty} \mathsf{F}_n\big(np + x\sqrt{np(1-p)}\big) = \Phi(x), \tag{5.1}$$

where

$$\Phi(x) := \int_{-\infty}^{x} \frac{e^{-t^2/2}}{\sqrt{2\pi}} \mathrm{d}t. \tag{5.2}$$

[27] Named after Abraham de Moivre (1667 - 1754) and Pierre-Simon, marquis de Laplace (1749 - 1827).

Proposition 5.5. *The function* Φ *in* (5.2) *satisfies the conditions of Theorem 4.6, in particular because*

$$\int_{-\infty}^{\infty} e^{-t^2/2}\mathrm{d}t = \sqrt{2\pi}.$$

Thus the function Φ defined in (5.2) defines a distribution, referred to as the *standard normal distribution*; see Section 5.7 for more details.

The theorem above is sometimes referred to as the *normal approximation to the binomial distribution*. See Figure 5.1 for an illustration.

As the proof of Theorem 5.4 can be relatively long, we only provide some guidance. Let $\sigma = \sqrt{p(1-p)}$.

Problem 5.6. Let $\mathsf{G}_n(x) = \mathsf{F}_n\big(np + x\sigma\sqrt{n}\big)$. Show that it suffices to prove that $\mathsf{G}_n(b) - \mathsf{G}_n(a) \to \Phi(b) - \Phi(a)$ for all $-\infty < a < b < \infty$.

Problem 5.7. Using (3.7), show that

$$\mathsf{G}_n(b) - \mathsf{G}_n(a) = \sigma\sqrt{n}\int_a^b \binom{n}{\kappa_n(t)} p^{\kappa_n(t)}(1-p)^{n-\kappa_n(t)}\mathrm{d}t + O(1/\sqrt{n}),$$

where $\kappa_n(t) := \lfloor np + t\sigma\sqrt{n}\rfloor$.

Problem 5.8. Show that

$$\sigma\sqrt{n}\binom{n}{\kappa_n(t)} p^{\kappa_n(t)}(1-p)^{n-\kappa_n(t)} \to \frac{\mathrm{e}^{-t^2/2}}{\sqrt{2\pi}}, \quad \text{as } n \to \infty,$$

uniformly in $t \in [a, b]$. The rather long, but elementary calculations are based on Stirling's formula in the form of (2.22).

5.1.3 From geometric to exponential

Let F_N denote the distribution function of the geometric distribution with parameter $(\lambda/N) \wedge 1$, where $\lambda > 0$ is fixed.

Problem 5.9. Show that, for any $x \in \mathbb{R}$,

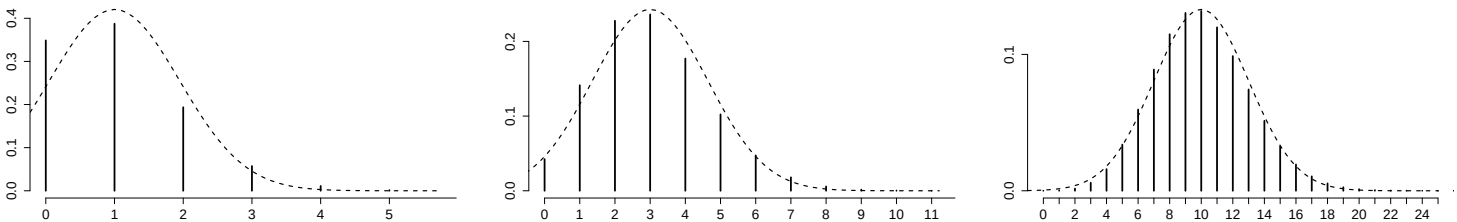
$$\lim_{N\to\infty} \mathsf{F}_N(Nx) = \mathsf{F}(x) := (1 - \mathrm{e}^{-\lambda x})\,\{x > 0\}.$$

The limit function F above satisfies the conditions of Theorem 4.6 and so defines a distribution, referred to as the *exponential distribution* with rate λ; see Section 5.8 for more details.

5.2 Continuous distributions

A distribution P on $(\mathbb{R}, \mathcal{B})$, with distribution function F, is a *continuous distribution* if F is continuous as a function.

Figure 5.1: An illustration of the normal approximation to the binomial distribution with parameters $n \in \{10, 30, 100\}$ (from left to right) and $p = 0.1$.



Problem 5.10. Show that P is continuous if and only if $\mathsf{P}(\{x\}) = 0$ for all x.

Problem 5.11. Show that $\mathsf{F} : \mathbb{R} \to \mathbb{R}$ is the distribution function of a continuous distribution if and only if it is continuous, non-decreasing, and satisfies

$$\lim_{x \to -\infty} \mathsf{F}(x) = 0, \quad \lim_{x \to \infty} \mathsf{F}(x) = 1.$$

We say that a distribution P is a mixture of distributions P_0 and P_1 if there is $b \in [0, 1]$ such that

$$\mathsf{P} = (1 - b)\mathsf{P}_0 + b\mathsf{P}_1. \tag{5.3}$$

Theorem 5.12. *Every distribution on the real line is the mixture of a discrete distribution and a continuous distribution.*

Proof. Let P be a distribution on the real line, with distribution function denoted F. Assume that P is neither discrete nor continuous, for otherwise there is nothing to prove.

Let $\mathcal{D}$ denote the set of points where F is discontinuous. By Problem 4.9 and the fact that F is non-decreasing (see (4.7)), $\mathcal{D}$ is countable, and since we have assumed that P is not continuous, $b := \mathsf{P}(\mathcal{D}) > 0$. Define

$$\mathsf{F}_1(x) = \frac{1}{b} \sum_{t \le x, t \in \mathcal{D}} \mathsf{P}(\{t\}). \tag{5.4}$$

It is easy to see that F_1 is a piecewise constant distribution function, which thus defines a discrete distribution, denoted P_1.

Define

$$\mathsf{F}_0 = \frac{1}{1-b}(\mathsf{F} - b\mathsf{F}_1). \tag{5.5}$$

It is easy to see that F_0 is a distribution function, which therefore defines a distribution, denoted P_0.

By construction, (5.3) holds, and so it remains to prove that F_0 is a continuous. Since F_0 is continuous from the right (see (4.10)), it suffices to show that it is continuous from the left as well, or equivalently, that $\mathsf{F}_0(x) - \mathsf{F}_0(x^-) = 0$ for all $x \in \mathbb{R}$. (Recall the definition (4.12).) For $x \in \mathbb{R}$, by (4.13), it suffices to establish that $\mathsf{P}_0(\{x\}) = 0$. We have

$$\mathsf{P}_0(\{x\}) = \frac{1}{1-b}(\mathsf{P}(\{x\}) - b\mathsf{P}_1(\{x\})). \tag{5.6}$$

If $x \notin \mathcal{D}$, then $\mathsf{P}(\{x\}) = 0$ and $\mathsf{P}_1(\{x\}) = 0$, while if $x \in \mathcal{D}$, $b\mathsf{P}_1(\{x\}) = \mathsf{P}(\{x\})$, so in any case $\mathsf{P}_0(\{x\}) = 0$. □

5.3 Absolutely continuous distributions

A distribution P on $(\mathbb{R}, \mathcal{B})$, with distribution function F, is *absolutely continuous* if F is absolutely continuous as a function, meaning that there is an integrable function f such that

$$\mathsf{F}(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t. \tag{5.7}$$

In that case, we say that f is a *density* of P.

Remark 5.13 (Integrable functions)**.** There are a number of notions of integral. The most natural one in the context of Probability Theory is the *Lebesgue integral*. However, the *Riemann integral* has a somewhat more elementary definition. We will only consider functions for which the two notions coincide and will call these functions *integrable*. This includes *piecewise continuous functions*.[28]

Remark 5.14 (Non-uniqueness of a density)**.** The function f in (5.7) is not uniquely determined by F. For example, if g coincides with f except on a finite number of points, then g also satisfies (5.7). Even then, it is customary to speak of 'the' density of a distribution, and we will do the same on occasion. This is particularly warranted when there is a continuous function f satisfying (5.7). In that case, it is the only one with that property and the most natural choice for the density of F. More generally, f is chosen as 'simple' possible.

Problem 5.15. Suppose that f and g both satisfy (5.7). Show that if they are both continuous they must coincide.

Problem 5.16. Show that a function f satisfying (5.7)

[28] A function f is piecewise constant if its discontinuity points are nowhere dense, or equivalently, if there is a strictly increasing sequence $(a_k : k \in \mathbb{Z})$ with $\lim_{k\to-\infty} a_k = -\infty$ and $\lim_{k\to-\infty} a_k = \infty$ such that f is continuous on (a_k, a_{k+1}).

must be non-negative at all of its continuity points.

Problem 5.17. Show that if f satisfies (5.7), then so do $\max(f, 0)$ and $|f|$. Therefore, any absolutely continuous distribution always admits a density function that is non-negative everywhere, and henceforth, we always choose to work with such a density.

Proposition 5.18. *An integrable function $f : \mathbb{R} \to [0, \infty)$ is a density of a distribution if and only if*

$$\int_{-\infty}^{\infty} f(x)\mathrm{d}x = 1. \tag{5.8}$$

In that case, it defines an absolutely continuous distribution via (5.7).

Remark 5.19. Density functions are to absolutely continuous distributions what mass functions are to discrete distributions.

5.4 Continuous random variables

We say that X is a (resp. absolutely) *continuous random variable* on a sample space if it is a random variable on that space as defined in Chapter 4 and its distribution P_X is (resp. absolutely) continuous, meaning that F_X is (resp. absolutely) continuous as a function. We let f_X denote a density of P_X when one exists.

Problem 5.20. For a continuous random variable X, verify that, for all $a < b$,

$$\begin{aligned}\mathbb{P}(X \in (a, b]) &= \mathbb{P}(X \in [a, b)) \\ &= \mathbb{P}(X \in [a, b]) = \mathbb{P}(X \in (a, b)),\end{aligned}$$

and, assuming X is absolutely continuous,

$$\begin{aligned}\mathbb{P}(X \in (a, b]) &= \mathsf{P}_X((a, b]) \\ &= \mathsf{F}_X(b) - \mathsf{F}_X(a) \\ &= \int_a^b f_X(x)\mathrm{d}x.\end{aligned}$$

Problem 5.21. Show that X is a continuous random variable if and only if

$$\mathbb{P}(X = x) = 0, \quad \text{for all } x \in \mathbb{R}.$$

(This is a bit perplexing at first.) In particular, the mass function of X is utterly useless.

Problem 5.22. Assume that X has a density f_X. Show that, for any x where f_X is continuous,

$$\mathbb{P}(X \in [x - h, x + h]) \sim 2h f_X(x), \quad \text{as } h \to 0.$$

5.5 Location/scale families of distributions

Let X be a random variable. Then the family of distributions defined by the random variables $\{X + b : b \in \mathbb{R}\}$, is the *location family of distributions* generated by X.

Similarly, the family of distributions defined by the random variables $\{aX : a > 0\}$, is the *scale family of distributions* generated by X, and the family of distributions defined by the random variables $\{aX + b : a > 0, b \in \mathbb{R}\}$, is the *location-scale family of distributions* generated by X.

Problem 5.23. Show that $aX + b$ has distribution function $\mathsf{F}_X((\cdot - b)/a)$ and density $\frac{1}{a} f_X((\cdot - b)/a)$.

5.6 Uniform distributions

The *uniform distribution* on an interval $[a, b]$ is given by the density

$$f(x) = \frac{1}{b-a}\{x \in [a, b]\}.$$

We will denote this distribution by $\mathrm{Unif}(a, b)$.

We saw in Section 5.1.1 how this sort of distribution arises as a limit of discrete uniform distributions; see also Problem 5.43.

Problem 5.24. Compute the distribution function of $\mathrm{Unif}(a, b)$.

Problem 5.25 (Location-scale family)**.** Show that the family of uniform of distributions, meaning

$$\big\{\mathrm{Unif}(a, b) : a < b\big\},$$

is a location-scale family by verifying that

$$\mathrm{Unif}(a, b) \equiv (b - a)\mathrm{Unif}(0, 1) + a.$$

Proposition 5.26. *Let U be uniform in $[0, 1]$ and let F be any distribution function with quantile function F^-. Then $\mathsf{F}^-(U)$ has distribution F.*

Problem 5.27. Prove this result, at least in the case where F is continuous and strictly increasing (in which case F^- is a true inverse).

5.7 Normal distributions

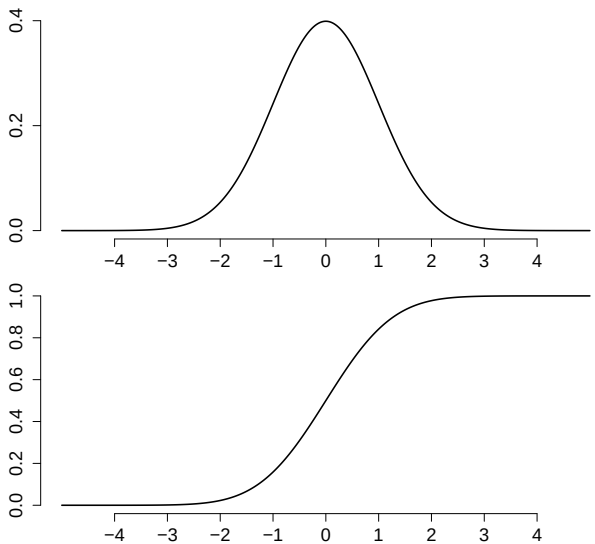
The *normal distribution* (aka *Gaussian distribution*[29]) with parameters μ and σ^2 is given by the density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big). \tag{5.9}$$

[29] Named after Carl Gauss (1777 - 1855).

(That this is a density is due to Proposition 5.5.) We will denote this distribution by $\mathcal{N}(\mu, \sigma^2)$. It so happens that μ is the mean and σ^2 the variance of $\mathcal{N}(\mu, \sigma^2)$. See Chapter 7. See Figure 5.2 for an illustration.

Figure 5.2: A plot of the standard normal density function (top) and distribution function (bottom).



We saw in Theorem 5.4 that a normal distribution arises as the limit of binomial distributions, but in fact this limiting behavior is much more general, in particular because of Theorem 8.31, which partly explains why this family is so important.

Problem 5.28 (Location-scale family)**.** Show that the family of normal distributions, meaning

$$\left\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\right\},$$

is a location-scale family by verifying that

$$\mathcal{N}(\mu, \sigma^2) \equiv \sigma\mathcal{N}(0, 1) + \mu.$$

The distribution $\mathcal{N}(0, 1)$ is often called the *standard normal distribution*.

Proposition 5.29 (Stability of the normal family)**.** *The sum of a finite number of independent normal random variables is normal.*

5.8 Exponential distributions

The *exponential distribution* with rate λ is given by the density

$$f(x) = \lambda \exp(-\lambda x)\,\{x \geq 0\}.$$

We will denote this distribution by $\text{Exp}(\lambda)$.

We saw in Section 5.1.3 how this distribution arises as a continuous limit of geometric distributions; see also Problem 8.58.

Problem 5.30 (Scale family)**.** Show that the family of exponential distributions, meaning

$$\big\{\mathrm{Exp}(\lambda) : \lambda > 0\big\},$$

is a scale family by verifying that

$$\mathrm{Exp}(\lambda) \equiv \frac{1}{\lambda}\mathrm{Exp}(1).$$

Problem 5.31. Compute the distribution function of $\mathrm{Exp}(\lambda)$.

Problem 5.32 (Memoryless property)**.** Show that any exponential distribution has the memoryless property of Problem 3.10.

5.9 Other continuous distributions

There are many other continuous distributions and families of such distributions. We introduce a few more below.

5.9.1 Gamma distributions

The *gamma distribution* with rate λ and shape parameter κ is given by the density

$$f(x) = \frac{\lambda^\kappa}{\Gamma(\kappa)} x^{\kappa-1} \exp(-\lambda x)\,\{x \geq 0\},$$

where Γ is the so-called gamma function. We will denote this distribution by $\mathrm{Gamma}(\lambda, \kappa)$.

Problem 5.33. Show that f above has finite integral if and only if $\lambda > 0$ and $\kappa > 0$.

Problem 5.34. Express the gamma function as an integral. [Use the fact that f above is a density.]

Problem 5.35 (Scale family)**.** Show that the family of gamma distributions with same shape parameter κ, meaning

$$\big\{\mathrm{Gamma}(\lambda, \kappa) : \lambda > 0\big\},$$

is a scale family.

It can be shown that a gamma distribution can arise as the continuous limit of negative binomial distributions; see Problem 5.45. The following is the analog of Proposition 3.13.

Proposition 5.36. *Consider m independent random variables having the exponential distribution with rate λ.*

Then their sum has the gamma distribution with parameters (λ, m).

5.9.2 Beta distributions

The *beta distribution* with parameters (a, b) is given by the density

$$f(x) = \frac{1}{\mathsf{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\{x \in [0,1]\},$$

where B is the beta function. It can be shown that

$$\mathsf{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

Problem 5.37. Show that f above has finite integral if and only if $\alpha > 0$ and $\beta > 0$.

Problem 5.38. Express the beta function as an integral.

Problem 5.39. Prove that this is not a location and/or scale family of distributions.

5.9.3 Families related to the normal family

A number of families are closely related to the normal family. The following are the main ones.

Chi-squared distributions The *chi-squared distribution* with parameter $m \in \mathbb{N}$ is the distribution of $Z_1^2 + \cdots + Z_m^2$ when $Z_1, \ldots, Z_m$ are independent standard normal random variables. This happens to be a subfamily of the gamma family.

Proposition 5.40. *The chi-squared distribution with parameter* m *coincides with the gamma distribution with shape* $\kappa = m/2$ *and rate* $\lambda = 1/2$.

Student distributions The *Student distribution*[30] (aka, *t-distribution*) with parameter $m \in \mathbb{N}$ is the distribution of $Z/\sqrt{W/m}$ when Z and W are independent, with Z being standard normal and W being chi-squared with parameter m.

Remark 5.41. The Student distribution with parameter $m = 1$ coincides with the *Cauchy distribution*[31], defined by its density function

$$f(x) := \frac{1}{\pi(1+x^2)}. \tag{5.10}$$

Fisher distributions[13] The *Fisher distribution* (aka *F-distribution*) with parameters (m_1, m_2) is the distribu-

[30] Named after 'Student', the pen name of Gosset[21].
[31] Named after Augustin-Louis Cauchy (1789 - 1857).

tion of $(W_1/m_1)/(W_2/m_2)$ when W_1 and W_2 are independent, with W_j being chi-squared with parameter m_j.

Problem 5.42. Relate the Fisher distribution with parameters $(1, m)$ and the Student distribution with parameter m.

5.10 Additional problems

Problem 5.43. Let F denote a continuous distribution function, with quantile function denoted F^-. For $1 \le k \le N$ both integers, define $x_{k:N} = \mathsf{F}^-(k/(N+1))$, and let P_N denote the (discrete) uniform distribution on $\{x_{1:N}, \dots, x_{N:N}\}$. If F_N denotes the corresponding distribution function, show that $\mathsf{F}_N(x) \to \mathsf{F}(x)$ as $N \to \infty$ for all $x \in \mathbb{R}$.

Problem 5.44 (Bernoulli trials and the uniform distribution)**.** Let that $(X_i : i \ge 1)$ be independent with same distribution Ber(1/2). Show that $Y := \sum_{i \ge 1} 2^{-i} X_i$ is uniform in $[0, 1]$. Conversely, let Y be uniform in $[0, 1]$, and let $\sum_{i \ge 1} 2^{-i} X_i$ be its binary expansion. Show that $(X_i : i \ge 1)$ are independent with same distribution Ber(1/2).

Problem 5.45. We saw how a sequence of geometric distributions can have as limit an exponential distribution. Show by extension how a sequence of negative binomial distributions can have as limit a gamma distribution. [There is a simple argument based on the fact that a negative binomial (resp. gamma) random variable can be expressed as a sum of independent geometric (resp. exponential) random variables. An analytic proof will resemble that of Theorem 5.4.]

Problem 5.46. Verify Theorem 5.4 by simulation in R. For each $n \in \{10, 100, 1000\}$ and each $p \in \{0.05, 0.2, 0.5\}$, generate $M = 500$ realizations from $\mathrm{Bin}(n, p)$ using the function rbinom and plot the corresponding histogram (with 50 bins) using the function hist. Overlay the graph of the standard normal density.

CHAPTER 6

MULTIVARIATE DISTRIBUTIONS

Some experiments lead to considering not one, but several random variables. For example, in the context of an experiment that consists in flipping a coin n times, we defined n random variables, one for each coin flip, according to (3.4).

In what follows, all the random variables that we consider are assumed to be defined on the same probability space,[32] denoted $(\Omega, \Sigma, \mathbb{P})$.

6.1 Random vectors

Let $X_1, \dots, X_r$ be r random variables on Ω, meaning that each X_i satisfies (4.1). Then $\boldsymbol{X} := (X_1, \dots, X_r)$ is a *random vector* on Ω, which is thus a function on Ω with values in $\mathbb{R}^r$,

$$\omega \in \Omega \longmapsto \boldsymbol{X}(\omega) = (X_1(\omega), \dots, X_r(\omega)) \in \mathbb{R}^r.$$

[32] This can be assumed without loss of generality (Section 8.2).

Problem 6.1. Show that,

$$\{\boldsymbol{X} \in \mathcal{V}\} \in \Sigma, \tag{6.1}$$

for any set $\mathcal{V}$ of the form

$$(-\infty, x_1] \times \cdots \times (-\infty, x_r], \tag{6.2}$$

where $x_1, \ldots, x_r \in \mathbb{R}$.

We define the Borel σ-algebra of $\mathbb{R}^r$, denoted $\mathcal{B}_r$, as the σ-algebra generated by all hyper-rectangles of the form (6.2). We will always equip $\mathbb{R}^r$ with its Borel σ-algebra. The following generalizes Proposition 4.1.

Proposition 6.2. *The Borel σ-algebra of $\mathbb{R}^r$ contains all hyper-rectangles, as well as all open sets and all closed sets.*

The *support* of a distribution P on $(\mathbb{R}^r, \mathcal{B}_r)$ is the smallest closed set $\mathcal{A}$ such that $\mathsf{P}(\mathcal{A}) = 1$. The *distribution function* of a distribution P on $(\mathbb{R}^r, \mathcal{B}_r)$ is defined as

$$\mathsf{F}(x_1, \ldots, x_r) := \mathsf{P}\big((-\infty, x_1] \times \cdots \times (-\infty, x_r]\big). \tag{6.3}$$

The distribution of $\boldsymbol{X}$, also referred to as the *joint distribution* of $X_1, \ldots, X_r$, is defined on the Borel sets

$$\mathsf{P}_{\boldsymbol{X}}(\mathcal{V}) := \mathbb{P}(\boldsymbol{X} \in \mathcal{V}), \quad \text{for } \mathcal{V} \in \mathcal{B}_r.$$

Note that for product sets, meaning when $\mathcal{V} = \mathcal{V}_1 \times \cdots \times \mathcal{V}_r$,

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{V}) = \mathbb{P}(X_1 \in \mathcal{V}_1) \times \cdots \times \mathbb{P}(X_r \in \mathcal{V}_r).$$

The distribution function of $\boldsymbol{X}$ is (of course) the distribution function of $\mathsf{P}_{\boldsymbol{X}}$, and can be expressed as

$$\mathsf{F}_{\boldsymbol{X}}(x_1, \ldots, x_r) := \mathbb{P}(X_1 \le x_1, \ldots, X_r \le x_r). \tag{6.4}$$

The following generalizes Proposition 4.4.

Proposition 6.3. *A distribution is characterized by its distribution function.*

The distribution of X_i, seen as the ith component of a random vector $\boldsymbol{X} = (X_1, \ldots, X_r)$, is often called the *marginal distribution* of X_i, which is nothing else but its distribution, disregarding the other variables.

6.1.1 Discrete distributions

We say that a distribution P on $\mathbb{R}^r$ is discrete if it has countable support set. For such a distribution, it is useful to consider its *mass function*, defined as

$$f(\boldsymbol{x}) := \mathsf{P}(\{\boldsymbol{x}\}), \tag{6.5}$$

or, equivalently,

$$f(x_1, \ldots, x_r) = \mathsf{P}(\{(x_1, \ldots, x_r)\}). \tag{6.6}$$

Problem 6.4. Show that a discrete distribution is characterized by it mass function.

Problem 6.5. Show that when all r variables are discrete, so is the random vector they form.

The mass function of a random vector $\boldsymbol{X}$ is the mass function of its distribution. It can be expressed as

$$f_{\boldsymbol{X}}(\boldsymbol{x}) := \mathbb{P}(\boldsymbol{X} = \boldsymbol{x}),$$

or, equivalently,

$$f_{\boldsymbol{X}}(x_1, \dots, x_r) = \mathbb{P}(X_1 = x_1, \dots, X_r = x_r). \tag{6.7}$$

Proposition 6.6. *Let $\boldsymbol{X} = (X_1, \dots, X_r)$ be a discrete random vector with support on $\mathbb{Z}^r$. Then the (marginal) mass function of X_i can be computed as follows*

$$f_{X_i}(x_i) = \sum_{j \neq i} \sum_{x_j \in \mathbb{Z}} f_{\boldsymbol{X}}(x_1, \dots, x_r), \qquad \text{for } x_i \in \mathbb{Z}.$$

For example, with two random variables, denoted X and Y, both supported on $\mathbb{Z}$,

$$\mathbb{P}(X = x) = \sum_{y \in \mathbb{Z}} \mathbb{P}(X = x, Y = y), \quad \text{for } x \in \mathbb{Z}. \tag{6.8}$$

Problem 6.7. Prove (6.8).

Binary random vectors An r-dimensional binary random vector is a random vector with values in $\{0,1\}^r$ (sometimes $\{-1,1\}^r$). Such random vectors are particularly important as they are often used to represent outcomes that are *categorical* in nature (as opposed to numerical). For example, consider an experiment where we roll a die with six faces. Assume without loss of generality that they are numbered $1, \dots, 6$. The fact that the face labels are numbers is typically not be relevant, and representing the result of rolling the die with a random variable (with support $\{1, \dots, 6\}$) could be misleading. We may instead use a binary random vector for that purpose, as follows

$$\begin{aligned} 1 &\to (1,0,0,0,0,0) \\ 2 &\to (0,1,0,0,0,0) \\ 3 &\to (0,0,1,0,0,0) \\ 4 &\to (0,0,0,1,0,0) \\ 5 &\to (0,0,0,0,1,0) \\ 6 &\to (0,0,0,0,0,1) \end{aligned}$$

This allows for the use of vector algebra, which will be done later on, in particular in Section 15.1.

Remark 6.8. In terms of coding, this is far from optimal, as discussed in Section 23.6. But the intention here is completely different: it is to facilitate algebraic manipulations of random variables tied to categorical outcomes.

6.1.2 Continuous distributions

We say that a distribution P on $\mathbb{R}^r$ is *continuous* if its distribution function F is continuous as a function on $\mathbb{R}^r$. We say that P is *absolutely continuous* if there is $f : \mathbb{R}^r \to \mathbb{R}$ integrable such that

$$\mathsf{P}(\mathcal{V}) = \int_{\mathcal{V}} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \quad \text{for all } \mathcal{V} \in \mathcal{B}_r.$$

In that case, f is a *density* of P.

Remark 6.9. Just as for distributions on the real line, a density is not unique and can be always taken to be non-negative.

Problem 6.10. Show that a distribution is characterized by any one of its density functions.

We say that the random vector $\boldsymbol{X}$ is (resp. *absolutely*) *continuous* if its distribution is (resp. absolutely) continuous, and will denote a density (when applicable) by $f_{\boldsymbol{X}}$, which in particular satisfies

$$\mathbb{P}(\boldsymbol{X} \in \mathcal{V}) = \int_{\mathcal{V}} f(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \quad \text{for all } \mathcal{V} \in \mathcal{B}_r.$$

Proposition 6.11. *Let $\boldsymbol{X} = (X_1, \dots, X_r)$ be a random vector with density f. Then X_i has a density f_i given by*

$$f_i(x_i) := \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_r) \prod_{j \neq i} \mathrm{d}x_j,$$

for $x_i \in \mathbb{R}$.

For example, with two random variables, denoted X and Y for convenience, with joint density $f_{X,Y}$,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) \mathrm{d}y, \quad \text{for } x \in \mathbb{R}.$$

Remark 6.12. Even when all the random variables are continuous with a density, the random vector they define may not have a density in the anticipated sense. This happens, for example, when one variable is a function of the others or, more generally, when the variables are tied by an equation.

Example 6.13. Let T be uniform in $[0, 2\pi]$ and define $X = \cos(T)$ and $Y = \sin(T)$. Then $X^2 + Y^2 = 1$ by construction. In fact, (X, Y) is uniformly distributed on the unit circle.

6.2 Independence

Two random variables, X and Y, are said to be *independent* if

$$\mathbb{P}(X \in \mathcal{U}, Y \in \mathcal{V}) = \mathbb{P}(X \in \mathcal{U})\, \mathbb{P}(Y \in \mathcal{V}), \quad \text{for all } \mathcal{U}, \mathcal{V} \in \mathcal{B}.$$

This is sometimes denoted by $X \perp\!\!\!\perp Y$.

Proposition 6.14. *X and Y are independent if and only if, for all $x, y \in \mathbb{R}$,*

$$\mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x)\,\mathbb{P}(Y \le y),$$

or, equivalently, for all $x, y \in \mathbb{R}$,

$$\mathsf{F}_{X,Y}(x,y) = \mathsf{F}_X(x)\mathsf{F}_Y(y).$$

Problem 6.15. Show that, if they are both discrete, X and Y are independent if and only if their joint mass function factorizes as the product of their marginal mass functions, or in formula,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y), \quad \text{for all } x, y \in \mathbb{R}.$$

Problem 6.16. Show that, if (X,Y) has a density, then X and Y are independent if and only if the product of a density for X and a density for Y is a density for (X,Y).

Remark 6.17. Problem 6.16 has the following corollary. Assume that (X,Y) has a continuous density. Then X and Y have continuous (marginal) densities, and are independent if and only if, for all $x, y \in \mathbb{R}$,

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

The random variables $X_1, \dots, X_r$ are said to be *mutually independent* if, $\mathcal{V}_1, \dots, \mathcal{V}_r \in \mathcal{B}$,

$$\mathbb{P}(X_1 \in \mathcal{V}_1, \dots, X_r \in \mathcal{V}_r) = \mathbb{P}(X_1 \in \mathcal{V}_1) \cdots \mathbb{P}(X_r \in \mathcal{V}_r).$$

Proposition 6.18. *$X_1, \dots, X_r$ are mutually independent if and only if, for all $x_1, \dots, x_r \in \mathbb{R}$,*

$$\mathbb{P}(X_1 \le x_1, \dots, X_r \le x_r) = \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_r \le x_r).$$

Problem 6.19. State and solve the analog to Problem 6.15.

Problem 6.20. State and solve the analog to Problem 6.16.

When some variables are said to be independent, what is meant by default is mutual independence.

Problem 6.21. Let $X_1, \dots, X_r$ be independent random variables. Show that $g_1(X_1), \dots, g_r(X_r)$ are independent random variables for any measurable functions, $g_1, \dots, g_r$.

6.3 Conditional distribution

6.3.1 Discrete case

Given two discrete random variables X and Y, the *conditional distribution* of X given $Y = y$ is defined as the distribution of X conditional on the event $\{Y = y\}$ following the definition given in Section 1.5.1, namely

$$\mathbb{P}(X \in \mathcal{U} \mid Y = y) = \frac{\mathbb{P}(X \in \mathcal{U}, Y = y)}{\mathbb{P}(Y = y)}. \tag{6.9}$$

For any $y \in \mathbb{R}$ such that $\mathbb{P}(Y = y) > 0$, this defines a discrete distribution, with corresponding mass function

$$f_{X|Y}(x\,|\,y) := \mathbb{P}(X = x \mid Y = y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Problem 6.22 (Law of Total Probability revisited)**.** Show the following form of the Law of Total Probability. For two discrete random variables X and Y, both supported on $\mathbb{Z}$,

$$f_X(x) = \sum_{y\in\mathbb{Z}} f_{X|Y}(x\,|\,y) f_Y(y), \quad \text{for all } x \in \mathbb{Z}.$$

Problem 6.23 (Conditional distribution and independence)**.** Show that two discrete random variables X and Y are independent if and only if the conditional distribution of X given $Y = y$ is the same for all y in the support of Y. (And vice versa, as the roles of X and Y can be interchanged.)

6.3.2 Continuous case

When Y is discrete, the distribution of X given $Y = y$ as defined in (6.9) still makes sense, even when X is continuous. This is no longer the case when Y is continuous, for in that case $\mathbb{P}(Y = y) = 0$ for all $y \in \mathbb{R}$. It is nevertheless possible to make sense of the distribution of X given $Y = y$. We consider the special case where (X, Y) has a density. In that case, the distribution of X given $Y = y$ is defined as the distribution with density

$$f_{X|Y}(x\,|\,y) := \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

with the convention that $0/0 = 0$.

Problem 6.24. Show that $f_{X|Y}(\cdot\,|\,y)$ is indeed a density function for any y such that $f_Y(y) > 0$.

Problem 6.25. Show that, for any $x \in \mathbb{R}$,

$$\mathbb{P}(X \le x \mid Y \in [y-h, y+h]) \underset{h\to 0}{\longrightarrow} \int_{-\infty}^{x} f_{X|Y}(t\,|\,y)\mathrm{d}t.$$

[For simplicity, assume that $f_{X,Y}$ is continuous and that $f_Y > 0$ everywhere.]

6.4 Additional problems

Problem 6.26. Consider an experiment where two fair dice are rolled. Let X_i denote the result of the ith die, $i \in \{1, 2\}$. Assume the variables are independent. Let $X = X_1 + X_2$. Show that X is a bona fide random variable with support $\mathcal{X} = \{0, 1, \ldots, 12\}$, and compute its mass

function and its distribution function. (You can use R for the computations. Present the solution in a table.)

Problem 6.27 (Uniform distributions)**.** For a Borel set $\mathcal{V}$ in $\mathbb{R}^d$, its volume is defined as

$$|\mathcal{V}| := \int \{\boldsymbol{x} \in \mathcal{V}\} \mathrm{d}\boldsymbol{x} = \int_{\mathcal{V}} \mathrm{d}\boldsymbol{x}. \tag{6.10}$$

When $0 < |\mathcal{V}| < \infty$, we can define the uniform distribution on $\mathcal{V}$ as the distribution with density

$$f_{\mathcal{V}}(\boldsymbol{x}) := \frac{1}{|\mathcal{V}|} \{\boldsymbol{x} \in \mathcal{V}\}.$$

Show that $\boldsymbol{X} = (X_1, \ldots, X_d)$ is uniform on $[0,1]^d$ if and only if $X_1, \ldots, X_d$ are independent and uniform in $[0,1]$.

Problem 6.28 (Convolution)**.** Suppose that X and Y have densities and are independent. Show that the distribution of $Z := X + Y$ has density

$$f_Z(z) := \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) \mathrm{d}y. \tag{6.11}$$

This is called the *convolution* of f_X and f_Y, and often denoted by $f_X * f_Y$. State and prove a similar result when X and Y are both supported on $\mathbb{Z}$.

Problem 6.29. Assume that X and Y are independent random variables, with X having a continuous distribution. Show that $\mathbb{P}(X \neq Y) = 1$.

Problem 6.30. Consider an m-by-m matrix with elements being independent continuous random variables. Show that this *random matrix* is invertible with probability 1.

CHAPTER 7

EXPECTATION AND CONCENTRATION

An expectation is simply an average and averages are at the core of Probability Theory and Statistics. While it may not have been clear why random variables were introduced in previous chapters, we will see that they are quite useful when computing expectations. Otherwise, everything we do here can be done directly with distributions instead of random variables.

7.1 Expectation

The definition and computation of an expectation are based on sums when the underlying distribution is discrete, and on integrals when the underlying distribution is continuous. We will assume the reader is familiar with the concepts of *absolute summability* and *absolute integrability*, and the *Fubini–Tonelli theorem*.

7.1.1 Discrete expectation

Let X be a discrete random variable with mass function f_X. Its *expectation* or *mean* is defined as

$$\mathbb{E}(X) := \sum_{x \in \mathbb{Z}} x\, f_X(x), \tag{7.1}$$

so long as the sum converges absolutely.

Example 7.1. When X has a uniform distribution, its expectation is simply the average of the elements belonging to its support. To be more specific, assume that X has the uniform distribution on $\{x_1, \dots, x_N\}$. Then

$$\mathbb{E}(X) = \frac{x_1 + \cdots + x_N}{N}.$$

Example 7.2. We say that X is *constant* (as a random variable) if there is $x_0 \in \mathbb{R}$ such that $\mathbb{P}(X = x_0) = 1$. In that case, X has an expectation, given by $\mathbb{E}(X) = x_0$.

Proposition 7.3 (Change of variables)**.** *Let X be a discrete random variable and $g : \mathbb{Z} \to \mathbb{R}$. Then, as long as the sum converges absolutely,*

$$\mathbb{E}(g(X)) = \sum_{x \in \mathbb{Z}} g(x)\, \mathbb{P}(X = x).$$

Problem 7.4. Prove this proposition.

Problem 7.5 (Summation by parts)**.** Let X be a discrete random variable with values in the non-negative integers and with an expectation. Show that

$$\mathbb{E}(X) = \sum_{x \geq 0} \mathbb{P}(X > x).$$

7.1.2 Continuous expectation

Let X be a continuous random variable with density f_X. Its *expectation* or *mean* is defined as

$$\mathbb{E}(X) := \int_{-\infty}^{\infty} x f_X(x) \mathrm{d}x, \tag{7.2}$$

as long as the integrand is absolutely integrable.

Problem 7.6. Show that a random variable with the uniform distribution on $[a, b]$ has mean $(a + b)/2$, the midpoint of the support interval.

Proposition 7.7 (Change of variables)**.** *Let X be a continuous random variable with density f_X and g a measurable function on $\mathbb{R}$. Then, as long as the integrand is absolutely integrable,*

$$\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) \mathrm{d}x.$$

Problem 7.8. Prove this proposition.

Problem 7.9 (Integration by parts)**.** Let X be a non-negative continuous random variable with an expectation. Show that

$$\mathbb{E}(X) = \int_0^\infty \mathbb{P}(X > x)\mathrm{d}x.$$

7.1.3 Properties

Problem 7.10. Show that if X is non-negative and such that $\mathbb{E}(X) = 0$, then X is equal to 0 with probability 1, meaning $\mathbb{P}(X = 0) = 1$.

Problem 7.11. Prove that, for a random variable X with well-defined expectation, and $a \in \mathbb{R}$,

$$\mathbb{E}(aX) = a\,\mathbb{E}(X), \text{ and } \mathbb{E}(X + a) = \mathbb{E}(X) + a. \tag{7.3}$$

Problem 7.12. Show that the expectation is monotone in the sense that, for random variables X and Y,

$$X \le Y \;\Rightarrow\; \mathbb{E}(X) \le \mathbb{E}(Y). \tag{7.4}$$

(The left-hand side is short for $\mathbb{P}(X \le Y) = 1$.)

Recall that a convex function g on an interval I is a function that satisfies

$$g(ax + (1-a)y) \le ag(x) + (1-a)g(y), \quad \text{for all } x, y \in I \text{ and all } a \in [0,1]. \tag{7.5}$$

The function g is strictly convex if the inequality is strict whenever $x \neq y$ and $a \in (0,1)$.

Theorem 7.13 (Jensen's inequality[33])**.** *Let X be a continuous random variable and g a convex function on an interval containing the support of X such that both X and $g(X)$ have expectations. Then*

$$g(\mathbb{E}(X)) \le \mathbb{E}(g(X)).$$

If g is strictly convex, the inequality is strict unless X is constant.

For example, for a random variable X with an expectation,

$$|\,\mathbb{E}(X)| \le \mathbb{E}(|X|). \tag{7.6}$$

Proof sketch. We sketch a proof for the case where the variable has finite support. Let the support be $\{x_1, \ldots, x_N\}$ and let $p_j = f_X(x_j)$. Then

$$g(\mathbb{E}(X)) = g\Big(\sum_{j=1}^N p_j x_j\Big),$$

and

$$\mathbb{E}(g(X)) = \sum_{j=1}^N p_j g(x_j).$$

[33] Named after Johan Jensen (1859 - 1925).

Thus we need to prove that

$$g\Big(\sum_{j=1}^{N} p_j x_j\Big) \le \sum_{j=1}^{N} p_j g(x_j).$$

When $N = 2$, this is a direct consequence of (7.5): simply take $x = x_1$, $y = x_2$, and $a = p_1$, so that $1 - a = p_2$. The general case is proved by induction on N (and is a well-known property on convex functions). □

Proposition 7.14. *For two random variables X and Y with expectations, $X + Y$ has an expectation, which given by*

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y).$$

Proof sketch. We prove the result when the variables are discrete. Let $g(x, y) = x + y$. Then $X + Y = g(X, Y)$. Using a the analog of Proposition 7.3 for random vectors, we have

$$\mathbb{E}(g(X, Y)) = \sum_{x\in\mathbb{Z}} \sum_{y\in\mathbb{Z}} g(x, y)\, \mathbb{P}(X = x, Y = y). \tag{7.7}$$

Thus, using that as a starting point, and then interchanging sums as needed (which is possible because of absolute summability), we derive

$$\begin{aligned}
&\mathbb{E}(X + Y)\\
&= \sum_{x\in\mathbb{Z}} \sum_{y\in\mathbb{Z}} (x + y)\, \mathbb{P}(X = x, Y = y)\\
&= \sum_{x\in\mathbb{Z}} x \sum_{y\in\mathbb{Z}} \mathbb{P}(X = x, Y = y) + \sum_{y\in\mathbb{Z}} y \sum_{x\in\mathbb{Z}} \mathbb{P}(X = x, Y = y)\\
&= \sum_{x\in\mathbb{Z}} x\mathbb{P}(X = x) + \sum_{y\in\mathbb{Z}} y\mathbb{P}(Y = y)\\
&= \mathbb{E}(X) + \mathbb{E}(Y).
\end{aligned}$$

Equation (6.8) justifies the 3rd equality. □

Problem 7.15. Prove the result when the variables are continuous.

Problem 7.16. Prove by recursion that if $X_1, \dots, X_r$ are random variables with expectations, then $X_1 + \cdots + X_r$ has an expectation, which is given by

$$\mathbb{E}(X_1 + \cdots + X_r) = \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_r). \tag{7.8}$$

Problem 7.17 (Binomial mean)**.** Show that the binomial distribution with parameters (n, p) has mean np. An easy way to do so uses the definition of $\mathrm{Bin}(n, p)$ as the sum of n independent random variables with distribution $\mathrm{Ber}(p)$, as in (3.6), and then using (7.8). A harder way uses the

expression for its mass function (3.7) and the definition of expectation (7.1), which leads one to proving that

$$\sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = np.$$

Problem 7.18 (Hypergeometric mean)**.** Show that the hypergeometric distribution with parameter (n, r, b) has mean np where $p := r/(r+b)$. The easier way, analogous to that described in Problem 7.17, is recommended.

Proposition 7.19. *For two* independent *random variables X and Y with expectations, XY has an expectation, which is given by*

$$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y).$$

Compare with Proposition 7.14, which does not require independence.

Proof sketch. We prove the result when the variables are discrete. Let $g(x, y) = xy$. Then $XY = g(X, Y)$. Using the analog of Proposition 7.3 for random vectors, as before, we have (7.7). Using that as a starting point, and then the fact that multiplication distributes over summation,

$$\begin{aligned}\mathbb{E}(XY) &= \sum_{x\in\mathbb{Z}}\sum_{y\in\mathbb{Z}} xy\,\mathbb{P}(X=x, Y=y)\\ &= \sum_{x\in\mathbb{Z}}\sum_{y\in\mathbb{Z}} xy\,\mathbb{P}(X=x)\,\mathbb{P}(Y=y)\\ &= \sum_{x\in\mathbb{Z}} x\mathbb{P}(X=x)\sum_{y\in\mathbb{Z}} y\mathbb{P}(Y=y)\\ &= \mathbb{E}(X)\,\mathbb{E}(Y).\end{aligned}$$

We used the independence of X and Y in the 2nd line and absolute convergence in the 3rd line. □

Problem 7.20. Prove the result when the variables are continuous.

7.2 Moments

For a random variable X and a non-negative integer k, define the kth *moment* of X as the expectation of X^k, if X^k has an expectation. The 1st moment of a random variable is simply its mean.

Problem 7.21 (Binomial moments)**.** Compute the first four moments of the binomial distribution with parameters (n, p).

Problem 7.22 (Geometric moments)**.** Compute the first four moments of the geometric distribution with parameter p. [Start by proving that, for any $x \in (0,1)$, $\sum_{j\geq 0} x^j = (1-x)^{-1}$. Then differentiate up to four times to derive useful identities.]

Problem 7.23 (Uniform moments)**.** Compute the kth moment of the uniform distribution on $[0,1]$. [Use a recursion on k and integration by parts.]

Problem 7.24 (Normal moments)**.** Compute the first four moments of the standard normal distribution. Verify that they are respectively equal to $0, 1, 0, 3$. Deduce the first four moments of the normal distribution with parameters (μ, σ^2), which in particular has mean μ.

Problem 7.25. Show that if X has a kth moment for some $k \geq 1$, then it has a lth moment for any $l \leq k$. [Use Jensen's inequality.]

Problem 7.26 (Symmetric distributions)**.** Suppose that X and $-X$ have the same distribution. Show that if that distribution has a kth moment and k is odd, then that moment is 0.

The following is one of the most celebrated inequalities in Probability Theory.

Theorem 7.27 (Cauchy–Schwarz inequality [34])**.** *For two independent random variables X and Y with 2nd moments,*

$$|\mathbb{E}(XY)| \leq \sqrt{\mathbb{E}(X^2)}\sqrt{\mathbb{E}(Y^2)}.$$

Moreover the inequality is strict unless there is $a \in \mathbb{R}$ such that $\mathbb{P}(X = aY) = 1$ or $\mathbb{P}(Y = aX) = 1$.

Proof. By Jensen's inequality, in particular (7.6),

$$|\mathbb{E}(XY)| \leq \mathbb{E}(|XY|) = \mathbb{E}(|X||Y|),$$

so that it suffices to prove the result when X and Y are non-negative. The assumptions imply that for any real t, $X + tY$ has a 2nd moment, and we have

$$\begin{aligned} g(t) &:= \mathbb{E}\left((X+tY)^2\right) \\ &= \mathbb{E}(X^2) + 2t\,\mathbb{E}(XY) + t^2\,\mathbb{E}(Y^2). \end{aligned}$$

Thus g is a polynomial of degree at most 2. Since it is non-negative everywhere it must be non-negative at its minimum. The minimum happening at $t = -\mathbb{E}(XY)/\mathbb{E}(Y^2)$, we thus have

$$\begin{aligned} \mathbb{E}(X^2) - 2(\mathbb{E}(XY)/\mathbb{E}(Y^2))\,\mathbb{E}(XY) \\ + (\mathbb{E}(XY)/\mathbb{E}(Y^2))^2\,\mathbb{E}(Y^2) \geq 0, \end{aligned}$$

[34] Named after Cauchy [31] and Hermann Schwarz (1843 - 1921).

which leads to the stated inequality after simplification.

That the inequality is strict unless X and Y are proportional is left as an exercise. □

Problem 7.28. In the proof, we implicitly assumed that $\mathbb{E}(Y^2) > 0$. Show that the result holds (trivially) when $\mathbb{E}(Y^2) = 0$.

7.3 Variance and standard deviation

Assume that X has a 2nd moment. We can then define its *variance* as

$$\operatorname{Var}(X) := \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2. \tag{7.9}$$

The *standard deviation* of a random variable X is the square-root of its variance.

Remark 7.29 (Central moments)**.** Transforming X into $X - \mathbb{E}(X)$ is sometimes referred to as "centering X". Then, if k is a non-negative integer, the kth *central moment* of X is the kth moment of $X - \mathbb{E}(X)$, assuming it is well-defined. We show below that the variance corresponds to the 2nd central moment. (Note that the 1st central moment is 0.)

Proposition 7.30. *For a random variable X with a 2nd moment,*

$$\operatorname{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}(X))^2\big].$$

Proof. For the sake of clarity, set $\mu := \mathbb{E}(X)$. Then

$$(X - \mathbb{E}(X))^2 = (X - \mu)^2 = X^2 - 2\mu X + \mu^2.$$

Hence,

$$\begin{aligned}\operatorname{Var}(X) &= \mathbb{E}(X^2 - 2\mu X + \mu^2)\\ &= \mathbb{E}(X^2) + \mathbb{E}(-2\mu X) + \mathbb{E}(\mu^2)\\ &= \mathbb{E}(X^2) - 2\mu\,\mathbb{E}(X) + \mu^2\\ &= \mathbb{E}(X^2) - \mu^2.\end{aligned}$$

We used Proposition 7.14, then (7.3), and the fact that the expectation of a constant is itself, and finally the fact that $\mathbb{E}(X) = \mu$ and some simplifying algebra. □

Problem 7.31. Let X be random variable with a 2nd moment. Show that $\operatorname{Var}(X) = 0$ if and only if X is constant.

Problem 7.32. Prove that for a random variable X with a 2nd moment, and $a \in \mathbb{R}$,

$$\operatorname{Var}(aX) = a^2 \operatorname{Var}(X), \quad \operatorname{Var}(X + a) = \operatorname{Var}(X). \tag{7.10}$$

Problem 7.33 (Normal variance)**.** Show that the normal distribution with parameters (μ, σ^2) has variance σ^2.

Proposition 7.34. *For two* independent *random variables X and Y with 2nd moments,*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y). \tag{7.11}$$

Compare with Proposition 7.14, which does not require independence.

Problem 7.35. Prove Proposition 7.34 using Proposition 7.14.

Problem 7.36. Extend Proposition 7.34 to more than two (independent) random variables. [There is a simple argument by induction.]

Problem 7.37 (Binomial variance)**.** Show that the binomial distribution with parameters (n,p) has variance $np(1-p)$.

7.4 Covariance and correlation

In this section, all the random variables that we consider are assumed to have a 2nd moment.

Covariance To generalize Proposition 7.34 to non-independent random variables requires the *covariance* of two random variables, which is defined by

$$\mathrm{Cov}(X,Y) := \mathbb{E}\big((X-\mathbb{E}(X))(Y-\mathbb{E}(Y))\big). \tag{7.12}$$

Note that

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X).$$

Problem 7.38. Prove that

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y).$$

Problem 7.39. Prove that

$$X \perp\!\!\!\perp Y \;\Rightarrow\; \mathrm{Cov}(X,Y) = 0.$$

Problem 7.40. For random variables X, Y, Z, and reals a, b, show that

$$\mathrm{Cov}(aX,bY) = ab\,\mathrm{Cov}(X,Y),$$

and

$$\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z).$$

We are now ready to generalize Proposition 7.34.

Proposition 7.41. *For random variables X and Y with 2nd moment,*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y).$$

Problem 7.42. More generally, prove (by induction) that for random variables $X_1, \dots, X_r$,

$$\operatorname{Var}\Big(\sum_{i=1}^{r} X_i\Big) = \sum_{i=1}^{r}\sum_{j=1}^{r} \operatorname{Cov}(X_i, X_j)$$
$$= \sum_{i=1}^{r} \operatorname{Var}(X_i) + 2\sum_{1\le i<j\le r}\sum \operatorname{Cov}(X_i, X_j).$$

Problem 7.43 (Hypergeometric variance)**.** Show that the hypergeometric distribution with parameters (n, r, b) has variance $np(1-p)\frac{r+b-n}{r+b-1}$ where $p := r/(r+b)$. The easy way described in Problem 7.17 is recommended.

Correlation The *correlation* of X and Y is defined as

$$\operatorname{Corr}(X, Y) := \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}. \tag{7.13}$$

Problem 7.44. Show that the correlation has no unit in the (usual) sense that it is invariant with respect to affine transformations, or in formula, that

$$\operatorname{Corr}(aX + b, cY + d) = \operatorname{Corr}(X, Y),$$
for all $a, c > 0$ and all $b, d \in \mathbb{R}$.

Problem 7.45. Show that

$$\operatorname{Corr}(X, Y) \in [-1, 1], \tag{7.14}$$

and equal to ± 1 if and only if there are $a, b \in \mathbb{R}$ such that $\mathbb{P}(X = aY + b) = 1$ or $\mathbb{P}(Y = aX + b) = 1$.

7.5 Conditional expectation

Consider two random variables X and Y, with X having an expectation. Then conditionally on $Y = y$, X also has an expectation.

Problem 7.46. Prove this when both variables are discrete.

When the variables are both discrete, the *conditional expectation* of X given $Y = y$ can be expressed as follows

$$\mathbb{E}(X \,|\, Y = y) = \sum_{x\in\mathbb{Z}} x\, f_{X|Y}(x \,|\, y).$$

When the variables are both continuous, it can be expressed as

$$\mathbb{E}(X \,|\, Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x \,|\, y)\mathrm{d}x.$$

Note that $\mathbb{E}(X \,|\, Y)$ is a random variable. In fact, $\mathbb{E}(X \,|\, Y) = g(Y)$ with $g(y) := \mathbb{E}(X \,|\, Y = y)$, which happens to be measurable.

Problem 7.47 (Law of Total Expectation)**.** Show that

$$\mathbb{E}(\mathbb{E}(X\,|\,Y)) = \mathbb{E}(X). \tag{7.15}$$

Conditional variance If X has a second moment, then this is also the case of $X\,|\,Y = y$, which therefore has a variance, called the *conditional variance* of X given $Y = y$ and denoted by $\mathrm{Var}(X\,|\,Y = y)$.

Note that $\mathrm{Var}(X\,|\,Y)$ is a random variable.

Problem 7.48 (Law of Total Variance)**.** Show that

$$\mathrm{Var}(X) = \mathbb{E}(\mathrm{Var}(X\,|\,Y)) + \mathrm{Var}(\mathbb{E}(X\,|\,Y)). \tag{7.16}$$

7.6 Moment generating function

The *moment generating function* of a random variable X is defined as

$$\zeta_X(t) := \mathbb{E}(\exp(tX)), \quad \text{for } t \in \mathbb{R}. \tag{7.17}$$

As a function taking values in $[0, \infty]$, it is indeed well-defined everywhere.

Problem 7.49. Show that $\{t : \zeta_X(t) < \infty\}$ is an interval (possibly a singleton). [Use Jensen's inequality.]

In the special case where X is supported on the non-negative integers,

$$\zeta_X(t) = \sum_{k\geq 0} f_X(k)\, \mathrm{e}^{tk}.$$

The moment generating function derives its name from the following.

Proposition 7.50. *Assume that ζ_X is finite in an open interval containing 0. Then ζ_X is infinitely differentiable (in fact, analytic) in that interval and*

$$\zeta_X^{(k)}(0) = \mathbb{E}(X^k), \quad \textit{for all } k \geq 0.$$

Theorem 7.51. *Two distributions whose moment generating functions are finite and coincide on an open interval containing zero must be equal.*

Remark 7.52 (Laplace transform)**.** When X has a density, its moment generating function may be expressed as

$$\zeta_X(t) = \int_{-\infty}^{\infty} f_X(x)\mathrm{e}^{tx}\mathrm{d}x.$$

This coincides with the *Laplace transform* of f_X evaluated at $-t$, and a standard proof of Theorem 7.51 relies on the fact that the Laplace transform is invertible under the stated conditions.

7.7 Probability generating function

The *probability generating function* of a non-negative random variable X is defined as

$$\gamma_X(z) := \mathbb{E}(z^X), \quad \text{for } z \in [-1,1]. \tag{7.18}$$

Note that

$$\zeta_X(t) = \gamma_X(e^t), \quad \text{for all } t \le 0.$$

In the special case where X is supported on the non-negative integers,

$$\gamma_X(z) = \sum_{k\ge 0} f_X(k) z^k.$$

The probability generating function derives its name from the following.

Proposition 7.53. *Assume X is non-negative. Then γ_X is well-defined and finite on $[-1,1]$, and infinitely differentiable (in fact, analytic) in $(-1,1)$. Moreover, if X is supported on the non-negative integers,*

$$\gamma_X^{(k)}(0) = k!\, f_X(k), \quad \textit{for all } k \ge 0. \tag{7.19}$$

Problem 7.54. Show that any distribution that is supported on the non-negative integers is characterized by its probability generating function.

7.8 Characteristic function

The *characteristic function* of a random variable X is defined as

$$\varphi_X(t) := \mathbb{E}(\exp(\imath t X)), \quad \text{for } t \in \mathbb{R}, \tag{7.20}$$

where $\imath^2 = -1$. Compare with the definition of the moment generating function in (7.17). While the moment generating function may be infinite at any $t \neq 0$, the characteristic function is always well-defined for all $t \in \mathbb{R}$ as a complex-valued function.

Problem 7.55. Show that if X and Y are independent random variables then

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t), \quad \text{for all } t \in \mathbb{R}. \tag{7.21}$$

The converse is not true, meaning that there are situations where (7.21) holds even though X and Y are not independent. Find an example of that.

The characteristic function owes its name to the following.

Theorem 7.56. *A distribution is characterized by it characteristic function. Furthermore, if X is supported on the non-negative integers, then*

$$f_X(x) = \frac{1}{2\pi}\int_0^{2\pi} \exp(-\imath t x)\varphi_X(t)\mathrm{d}t. \tag{7.22}$$

If instead X is absolutely continuous, and its characteristic function is absolutely integrable, then

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-\imath t x)\varphi_X(t)\mathrm{d}t. \tag{7.23}$$

Problem 7.57. Prove (7.22).

Remark 7.58 (Fourier transform)**.** When X has a density,

$$\varphi_X(t) = \int_{-\infty}^{\infty} \exp(\imath t x) f_X(x)\mathrm{d}x. \tag{7.24}$$

This coincides with the *Fourier transform* of f_X evaluated at $-t/2\pi$ and a standard proof of Theorem 7.56 relies on the fact that the Fourier transform is invertible.

Remark 7.59. It is possible to define the characteristic function of a random vector $\boldsymbol{X}$. If $\boldsymbol{X}$ is r-dimensional, it is defined as

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) := \mathbb{E}(\exp(\imath\langle \boldsymbol{t}, \boldsymbol{X}\rangle)), \quad \text{for } \boldsymbol{t} \in \mathbb{R}^r, \tag{7.25}$$

where $\langle \boldsymbol{u}, \boldsymbol{v}\rangle$ denotes the inner product of $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^r$. We note that an analog of Theorem 7.56 holds for random vectors.

7.9 Concentration inequalities

An important question when examining a random variable is to know how far it strays away from its mean (which we assume is well-defined whenever needed). This is a probability statement, and we present some inequalities that bound the corresponding probability.

Proposition 7.60 (Markov's inequality[35])**.** *For a non-negative random variable X with expectation μ,*

$$\mathbb{P}\big(X \geq t\mu\big) \leq 1/t, \quad \textit{for all } t > 0. \tag{7.26}$$

For example, if X is non-negative with mean μ, then $X \geq 2\mu$ with at most 50% chance, while $X \geq 10\mu$ with at most 10% chance. (This is so regardless of the distribution of X.)

Proof. We have

$$\{X \geq t\} = \{X/t \geq 1\} \leq X/t.$$

and we conclude by taking the expectation and using its monotonicity property (Problem 7.12). □

Proposition 7.61 (Chebyshev's inequality[36])**.** *For a random variable X with expectation μ and standard deviation σ,*

$$\mathbb{P}\big(|X - \mu| \geq t\sigma\big) \leq 1/t^2, \quad \textit{for all } t > 0. \tag{7.27}$$

[35] Named after Andrey Markov (1856 - 1922).
[36] Named after Pafnuty Chebyshev (1821 - 1894).

Moreover,

$$\mathbb{P}\big(X \geq \mu + t\sigma\big) \leq 1/(1+t^2), \quad \textit{for all } t \geq 0, \tag{7.28}$$

and

$$\mathbb{P}\big(X \leq \mu - t\sigma\big) \leq 1/(1+t^2), \quad \textit{for all } t \geq 0. \tag{7.29}$$

Problem 7.62. Prove these inequalities by applying Markov's inequality to carefully chosen random variables.

For example, if X has mean μ and standard deviation σ, then $|X-\mu| \geq 2\sigma$ with at most 25% chance and $|X-\mu| \geq 5\sigma$ with at most 4% chance. (This is so regardless of the distribution of X.)

Markov's and Chebyshev's inequalities are examples of *concentration inequalities*. These are inequalities that bound the probability that a random variable is away from its mean (or sometimes median) by a certain amount.

Markov's inequality gives a concentration bound with a linear decay, while Chebyshev's inequality gives a concentration bound with a quadratic decay. Even stronger concentration is possible.

Problem 7.63. Consider a random variable Y with mean μ and such that $\alpha_s := \mathbb{E}(|Y-\mu|^s) < \infty$ for some $s > 1$ (not necessarily integer). Show that

$$\mathbb{P}(|Y-\mu| \geq y) \leq \alpha_s y^{-s}, \quad \text{for all } y > 0.$$

Proposition 7.64 (Chernoff's bound[37])**.** *Consider a random variable Y such that as $\zeta(\lambda) := \mathbb{E}(\exp(\lambda Y)) < \infty$ for some $\lambda > 0$. Then*

$$\mathbb{P}(Y \geq y) \leq \zeta(\lambda)\exp(-\lambda y), \quad \textit{for all } y > 0.$$

Proof. For any $\lambda \geq 0$,

$$Y \geq y \;\Rightarrow\; \lambda Y - \lambda y \geq 0 \;\Rightarrow\; \exp(\lambda Y - \lambda y) \geq 1.$$

Thus,

$$\begin{aligned}\mathbb{P}(Y \geq y) &\leq \mathbb{P}(\exp(\lambda Y - \lambda y) \geq 1)\\ &\leq \mathbb{E}(\exp(\lambda Y - \lambda y))\\ &= \exp(-\lambda y)\zeta(\lambda)\\ &= \exp(-\lambda y + \log\zeta(\lambda)),\end{aligned}$$

using Markov's inequality. □

Chernoff's bound is particularly useful when Y is the sum of independent random variables. This is in large part because of the following.

Problem 7.65. Suppose that $Y = X_1 + \cdots + X_n$, where $X_1, \ldots, X_n$ are independent. Let ζ_i denote the moment

[37] Named after Herman Chernoff (1923 -), who attributes the result to a colleague of his, Herman Rubin.

generating function of X_i, and ζ that of Y. Show that if $\zeta_i(\lambda) < \infty$ for all i, then $\zeta(\lambda) < \infty$ and

$$\zeta(\lambda) = \prod_{i=1}^{n} \zeta_i(\lambda).$$

Binomial distribution Let $Y = X_1 + \cdots + X_n$, where the X_i are independent, each being Bernoulli with parameter p, so that Y is binomial with parameters (n, p).

Problem 7.66. Show that

$$\zeta(\lambda) = (1 - p + pe^{\lambda})^n, \quad \text{for all } \lambda \in \mathbb{R}.$$

By Chernoff's bound, for any $\lambda \geq 0$,

$$\log \mathbb{P}(Y \geq y) \leq -\lambda y + \log \zeta(\lambda).$$

Since the left-hand side does not depend on λ, to sharpen the bound we minimize the right-hand side with respect to $\lambda \geq 0$, yielding

$$\log \mathbb{P}(Y \geq y) \leq \inf_{\lambda \geq 0} \big[-\lambda y + \log \zeta(\lambda) \big] \tag{7.30}$$

$$= -\sup_{\lambda \geq 0} \big[\lambda y - \log \zeta(\lambda) \big]. \tag{7.31}$$

We turn to maximizing $g(\lambda) := \lambda y - \log \zeta(\lambda)$ over $\lambda \geq 0$. Let $b = y/n$, so that $b \in [0, 1]$ in principle; however, since we are interested in deviations from the mean np, and because the case $b = 1$ requires a special (but trivial) treatment, we assume that $p < b < 1$.

Problem 7.67. Verify that g is infinitely differentiable and that it has a unique maximizer at

$$\lambda^* = \log \left(\frac{(1-p)b}{p(1-b)} \right),$$

and that $g(\lambda^*) = n\mathsf{H}_p(b)$, where

$$\mathsf{H}_p(b) := b \log \left(\frac{b}{p} \right) + (1-b) \log \left(\frac{1-b}{1-p} \right).$$

We thus arrived at the following.

Proposition 7.68 (Chernoff's bound for the binomial distribution)**.** *For $Y \sim Bin(n, p)$, with $0 < p < 1$,*

$$\mathbb{P}(Y \geq nb) \leq \exp\big(-n\mathsf{H}_p(b) \big), \quad \textit{for all } b \in [p, 1]. \tag{7.32}$$

Problem 7.69. Verify that the bound indeed applies to the cases we left off, namely, when $b = p$ and when $b = 1$.

Remark 7.70. A bound on $\mathbb{P}(Y \leq nb)$ when $b \in [0, p]$ can be derived in a similar fashion, or using Property (3.10).

We end this section with a general exponential concentration inequality whose roots are also in Chernoff's inequality.

Theorem 7.71 (Bernstein's inequality)**.** *Suppose that $X_1, \dots, X_n$ are independent with zero mean and such that $\max_i |X_i| \le c$. Define $\sigma_i^2 = \mathrm{Var}(X_i) = \mathbb{E}(X_i^2)$. Then, for all $y \ge 0$,*

$$\mathbb{P}\Big(\sum_{i=1}^n X_i \ge y\Big) \le \exp\left(-\frac{y^2/2}{\sum_{i=1}^n \sigma_i^2 + \frac{1}{3}cy}\right).$$

Problem 7.72. Apply Bernstein's inequality to get a concentration inequality for the binomial distribution. Compare the resulting bound with the one obtained from Chernoff's inequality in (7.32).

7.10 Further topics

7.10.1 Random sums of random variables

Suppose that $\{X_i : i \ge 1\}$ are independent with the same distribution, and independent of a random variable N supported on the non-negative integers. Together, these define the following *compound sum*

$$Y = \sum_{i=1}^N X_i.$$

By convention, the sum is zero if $N = 0$. Put differently, the distribution of Y given $N = n$ is that of $\sum_{i=1}^n X_i$.

Problem 7.73. Assume that the X_i have a 2nd moment. Derive the mean and variance of Y (showing in the process that Y has a 2nd moment).

Problem 7.74. Assume that the X_i are non-negative. Compute the probability generating function of Y.

When N has a Poisson distribution, the resulting distribution is called a *compound Poisson distribution*.

The negative binomial distribution is known to have a compound Poisson representation. In detail, first define the *logarithmic distribution* via its mass function

$$f_p(k) := \frac{1}{\log(\frac{1}{1-p})}\frac{p^k}{k}, \quad k \ge 1.$$

Problem 7.75. Show that, for $p \in (0,1)$, this defines a probability distribution on the positive integers. [Use the expression of the logarithm as a power series.]

Proposition 7.76. *Let $(X_i : i \ge 1)$ be independent with the logarithmic distribution with parameter p, and let N be Poisson with parameter $m\log(\frac{1}{1-p})$. Then $\sum_{i=1}^N X_i$ is negative binomial with parameters (m, p).*

Problem 7.77. Prove this result using Problem 7.74 and Theorem 7.51.

7.10.2 Estimation from finite samples

Consider a urn containing coins. Without additional information, to compute the average value of these coins, one would have to go through all coins and sum their values. But what if an approximation is sufficient — is it possible to do that without looking at all the coins? It turns out that the answer is yes, at least under some sampling schemes, because of concentration.

More generally, suppose that a urn contains N tickets numbered $c_1, \ldots, c_N \in \mathbb{R}$, and the goal is to approximate their average, $\mu := \frac{1}{N}(c_1 + \cdots + c_N)$. We assume we have the ability to sample uniformly at random with replacement from the urn n times. We do so and let $X_1, \ldots, X_n$ denote the resulting a sample, with corresponding average $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$.

Problem 7.78. Based on Chebyshev's inequality, show that, for any $t > 0$,

$$\left|\bar{X}_n - \mu\right| \leq t\sigma/\sqrt{n} \tag{7.33}$$

with probability at least $1 - 1/t^2$, where we have denoted $\sigma^2 := \frac{1}{N}(c_1^2 + \cdots + c_N^2) - \mu^2$.

The surprising fact in the approximation bound (7.33) is that it depends on the ticket values only through μ and σ, so that N could be infinite in principle.

The fact that we can "learn" about the contents of a possibly infinite urn based a finite sample from it is at the core of Statistics. It also explains why a carefully designed and conducted poll of a few thousand individuals can yield reliable information on a population of hundreds of millions (Section 11.1).

Problem 7.79. Obtain an approximation bound based Chernoff's bound instead. Compare this bound with that obtained in (7.33) via Chebyshev's inequality.

Remark 7.80 (Estimation)**.** This sort of approximation based on a sample is often referred to as *estimation*, and will be developed in later chapters. In particular, Section 23.1 will consider the same estimation problem but under different sampling schemes.

7.10.3 Saint Petersburg Paradox

Suppose a casino offers a gambler the opportunity to play the following fictitious game, attributed to Nicolas Bernoulli (1687 - 1759). The game starts with $2 on the table. At each round a fair coin is flipped: if it lands heads, the amount is doubled and the game continues; if it lands

tails, the game ends and the player pockets whatever is on the table. The question is: how much should the gambler be willing to pay to play the game?

A paradox arises when the gambler aims at optimizing his expected return, defined as $X - c$, where X is the gain (the amount on the table at the end of the game) and c is the entry cost (the amount the gambler pays the casino to enter the game).

Problem 7.81. Show that the expected return is infinite regardless of the cost.

Thus, in principle, a rational gambler would be willing to pay any amount to enter the game. However, a gambler with common sense would only be willing to pay very little to enter the game, hence, the paradox. Indeed, although the expected return is infinite, the probability of a positive return can be quite small.

Problem 7.82. Suppose the gambler pays c dollars to enter the game. Compute his chances of ending with a positive return as a function of c.

This, and other similar considerations, have lead some commentators to argue that the expected return is not what the gambler should be optimizing. Daniel Bernoulli (1700 - 1782) proposed as a solution in [17] (translated from Latin to English in [18]) to optimize the expected log return, i.e., $\mathbb{E}(\log(X/c))$, which he argued was more natural. (Daniel and Nicolas were brothers.)

Problem 7.83. Find in closed form or numerically (in R) the amount the gambler should be willing to pay to enter the game if his goal is to optimize the expected log return.

Another possibility is to optimize the median instead of the mean.

Problem 7.84. Find in closed form or numerically (in R) the amount the gambler should be willing to pay to enter the game if his goal is to optimize the median return.

Remark 7.85. The specific form of the game provides a colorful context and may have been motivated, at least in part, by the work of (uncle) Jacob Bernoulli (1655 - 1705) on what would later be called Bernoulli trials. However, the details are clearly unimportant and all that matters is that the gain has infinite expectation.

Remark 7.86 (Pascal's wager)**.** The essential component of the paradox arises from the extremely unlikely possibility of an enormous gain. This was considered by the philosopher Pascal in questions of faith in the existence of God (in the context of his Catholic faith). As he saw it, a person had to decide whether to believe in God or

not. From his book *Pensées* (1670): "Let us weigh the gain and the loss in wagering that God is. Let us estimate these two chances. If you gain, you gain all; if you lose, you lose nothing." [38]

7.11 Additional problems

Problem 7.87. In the context of Problem 6.26, compute the expectation of X. First, do so directly using the definition of expectation (7.1) based on the distribution of X found in that problem. (You may use R for that purpose.) Then do this using Proposition 7.14.

Problem 7.88. Using R, compute the first ten moments of the random variable X defined in Problem 6.26. [Do this efficiently using vector/matrix manipulations.]

Problem 7.89 (Location/scale families)**.** Consider a location scale family as in Section 5.5, therefore, of the form

$$\mathsf{F}_{a,b} := \mathsf{F}((x-b)/a), \quad \text{for } a > 0,\ b \in \mathbb{R},$$

where F is some given distribution function. Assume that F has finite second moment and is not constant. Show that there is exactly one distribution in this family with mean 0 and variance 1. Assuming F itself is that distribution, compute the mean and variance of $\mathsf{F}_{a,b}$ in terms of (a, b).

Problem 7.90 (Coupon Collector Problem)**.** Recall the Coupon Collector Problem of Section 3.8. Compute the mean and variance of T.

Problem 7.91. Let X be a random variable with a 2nd moment. Show that $a \mapsto \mathbb{E}((X-a)^2)$ is uniquely minimized at the mean of X.

Problem 7.92. Let X be a random variable with a 1st moment. Show that $a \mapsto \mathbb{E}(|X-a|)$ is minimized at any median of X and that any minimizer is a median of X.

Problem 7.93. Compute the mean and variance of the distribution of Problem 3.32.

Problem 7.94. Compute the characteristic function of
(i) the uniform distribution on $\{1, \dots, N\}$;
(ii) the Poisson distribution with mean λ;
(iii) the geometric distribution with parameter p.
Repeat with the moment generating function, specifying where it is finite.

Problem 7.95. Compute the characteristic function of
(i) the binomial distribution with parameters n and p;
(ii) the negative binomial distribution with parameters m and p.

[38] Translation by F. Trotter

Repeat with the moment generating function, specifying where it is finite.

Problem 7.96. Compute the characteristic function of
(i) the uniform distribution on $[a, b]$;
(ii) the exponential distribution with rate λ;
(iii) the normal distribution parameters (μ, σ^2).
Repeat with the moment generating function, specifying where it is finite.

Problem 7.97. Compute the characteristic function of the normal distribution with mean μ and variance σ^2. Then combine Problem 7.55 and Theorem 7.56 to prove Proposition 5.29.

Problem 7.98. Compute the characteristic function of the Poisson distribution with mean λ. Then combine Problem 7.55 and Theorem 7.56 to prove Proposition 3.17.

Problem 7.99. Suppose that X is supported on the nonnegative integers. Show that

$$\mathsf{F}_X(x) = \frac{1}{2\pi}\int_0^{2\pi} \frac{\sin(t(x+1)/2)}{\sin(t/2)} \mathrm{e}^{-\imath t x/2}\varphi_X(t)\mathrm{d}t. \quad (7.34)$$

Problem 7.100 (Markov vs Chebyshev)**.** Evaluate the accuracy of these two inequalities for the exponential distribution with rate $\lambda = 1$. One way to do so is to draw the survival function, the bound given by the Markov inequality, and the bound given by the Chebyshev inequality (7.28). Do so in R, and start at $x = 1$ (which is the mean in this case). Put all the graphs in the same plot, in different colors identified by a legend.

Problem 7.101. Write an R function that generates k independent numbers from the compound Poisson distribution obtained when N is Poisson with parameter λ and the X_i are Bernoulli with parameter p. Perform some simulations to better understand this distribution for various choices of parameter values.

Problem 7.102 (Passphrases)**.** The article [146] advocates choosing a strong password by selecting seven words at random from a list of 7776 English words. It claims that an adversary able to try one trillion guesses per second would have to keep trying for about 27 million years before discovering the correct passphrase. (This is so even if the adversary knows how the passphrase was generated.) Perform some calculations to corroborate this claim.

Problem 7.103 (Two envelopes - randomized strategy)**.** In the Two Envelopes Problem (Section 2.5.3), it turns out that it is possible to do better than random guessing. This is possible with even less information, in a setting were we are not told anything about the amounts inside the envelopes. Cover [45] offered the following strategy,

which relies on the ability to draw a random number. Having chosen a distribution with support the positive real line, we draw a number from this distribution and, if the amount in the envelope we opened is less than that number, we switch, otherwise we keep the envelope we opened. Show that this strategy beats random guessing.

Problem 7.104 (Two envelopes - model 1)**.** A first model for the Two Envelopes Problem is the following. Suppose X is supported on the positive integers. Given $X = x$, put x in Envelope A and either $x/2$ or $2x$ in Envelope B, each with probability 1/2. We are shown the contents of Envelope A and we need to decide whether to keep the amount found there or switch for the (unknown) amount in Envelope B. Consider three strategies: *(i)* always keep A; *(ii)* always switch to B; *(iii)* random switch (50% chance of keeping A, regardless of the amount it contains). For each strategy, compute the expected gain. Then describe an optimal strategy assuming the distribution of X is known. [Consider the discrete case first, and then the absolutely continuous case.]

Problem 7.105 (Two envelopes - model 2)**.** Another model for the Two Envelopes Problem is the following. Here, given $X = x$, let the contents of the envelopes (A, B) be $(x, 2x)$, $(2x, x)$, $(x, x/2)$, $(x/2, x)$, each with probability 1/4. We are shown the contents of Envelope A (although it does not matter in this model). For each strategy described in Problem 7.104, compute the expected gain. Then derive an optimal strategy.

CHAPTER 8

CONVERGENCE OF RANDOM VARIABLES

The convergence of random variables (or, equivalently, distributions) plays an important role in Probability Theory. This is particularly true of the Law of Large Numbers, which underpins the frequentist notion of probability. Another famous convergence result is a refinement known as the Central Limit Theorem, which underpins much of the large-sample statistical theory.

8.1 Product spaces

We might want to consider an experiment where a coin is repeatedly tossed, without end, and consider how the number of heads in the first n trials behaves as n increases. The resulting sample space is the space of all infinite sequences of elements in $\{\mathrm{H}, \mathrm{T}\}$, meaning $\Omega := \{\mathrm{H}, \mathrm{T}\}^{\mathbb{N}}$. The question is how to define a distribution on Ω that models the described experiment, where we assume the tosses to be independent.

For any integer $n \geq 1$, the mass function for the first n

tosses is given in (2.13). If we define a distribution on Ω, we want it to be compatible with that. It turns out that there is such a distribution and it is uniquely defined with that property.

8.1.1 Product of measurable spaces

More generally, for each integer $i \geq 1$, let (Ω_i, Σ_i) denote a measurable space. Their *product*, denoted (Ω, Σ), is defined as follows:

- Ω is simply the Cartesian product of $\Omega_i, i \geq 1$, meaning
$$\Omega := \Omega_1 \times \Omega_2 \times \cdots = \bigtimes_{i \geq 1} \Omega_i. \tag{8.1}$$
- Σ is the σ-algebra generated by sets of Ω of the form [39]
$$\bigtimes_{i \leq m-1} \Omega_i \times \mathcal{A}_m \times \bigtimes_{i \geq m+1} \Omega_i,$$
for some $m \geq 1$ and with $\mathcal{A}_m \in \Sigma_m$.

Problem 8.1. Show that the Cartesian product of σ-algebras is in general <u>not</u> a σ-algebra by producing a simple counter-example.

[39] Sets of this form are sometimes called *cylinder sets*.

8.1.2 Kolmogorov's extension theorem

This theorem allows one to (uniquely) define a distribution on (Ω, Σ) based on its restriction to events of the form
$$\mathcal{A}_1 \times \cdots \times \mathcal{A}_m \times \bigtimes_{i \geq m+1} \Omega_i, \tag{8.2}$$
which we identify with $\mathcal{A}_1 \times \cdots \times \mathcal{A}_m$.

Below, $(\Omega_{(m)}, \Sigma_{(m)})$ will denote the product of the measurable spaces $(\Omega_i, \Sigma_i), 1 \leq i \leq m$, and $\mathcal{A}_i$ will be generic for an event in Σ_i.

Theorem 8.2 (Extension theorem)**.** *A distribution $\mathbb{P}$ on (Ω, Σ) is uniquely determined by its values on events of the form $\mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. Conversely, any sequence of distributions $(\mathbb{P}_{(m)} : m \geq 1)$, with $\mathbb{P}_{(m)}$ being a distribution on $(\Omega_{(m)}, \Sigma_{(m)})$, which is compatible in the sense that*
$$\mathbb{P}_{(m)}(\mathcal{A}_1 \times \cdots \times \mathcal{A}_{m-1} \times \Omega_m) = \mathbb{P}_{(m-1)}(\mathcal{A}_1 \times \cdots \times \mathcal{A}_{m-1}),$$
for all $m \geq 2$, defines a distribution $\mathbb{P}$ on (Ω, Σ) via
$$\mathbb{P}(\mathcal{A}_1 \times \cdots \times \mathcal{A}_m) = \mathbb{P}_{(m)}(\mathcal{A}_1 \times \cdots \times \mathcal{A}_m),$$
for all $m \geq 1$.

8.1.3 Product distributions

Let $\mathbb{P}_i$ be a distribution on (Ω_i, Σ_i). The (tensor) *product distribution* is the unique distribution $\mathbb{P}$ on (Ω, Σ) such that, for any events $\mathcal{A}_i \in \Sigma_i$,

$$\mathbb{P}(\mathcal{A}_1 \times \mathcal{A}_2 \times \cdots) = \mathbb{P}_1(\mathcal{A}_1)\,\mathbb{P}_2(\mathcal{A}_2) \cdots ,$$

meaning

$$\mathbb{P}\Big(\bigtimes_{i\geq 1} \mathcal{A}_i\Big) = \prod_{i\geq 1} \mathbb{P}_i(\mathcal{A}_i).$$

In particular,

$$\mathbb{P}(\mathcal{A}_1 \times \cdots \times \mathcal{A}_m) = \mathbb{P}_1(\mathcal{A}_1) \times \cdots \times \mathbb{P}_m(\mathcal{A}_m),$$

for all $m \geq 1$. This distribution is well-defined by Theorem 8.2 and is sometimes denoted

$$\mathbb{P} = \mathbb{P}_1 \otimes \mathbb{P}_2 \otimes \cdots = \bigotimes_{i\geq 1} \mathbb{P}_i. \tag{8.3}$$

Problem 8.3. Show that any set of events $\mathcal{A}_1, \ldots, \mathcal{A}_k$ with $\mathcal{A}_i \in \Sigma_i$ are mutually independent under the product distribution (as events in Σ).

Thus the probability space $(\Omega, \Sigma, \mathbb{P})$ models an experiment consisting of infinitely many trials, with the ith trial corresponding to $(\Omega_i, \Sigma_i, \mathbb{P}_i)$ and independent of all others.

8.2 Sequences of random variables

Suppose that we are given two random variables, X_1 on a sample space Ω_1 and X_2 on a sample space Ω_2. We might want to take their sum or product or manipulate them in any other way. However, strictly speaking, if $\Omega_1 \neq \Omega_2$, this is not possible, simply because X_1 is a function on Ω_1 and X_2 a function on Ω_2.

The situation can be remedied by considering a meta sample space and identifying each of the variables with ones defined on that space. In detail, consider the product space $\Omega_1 \times \Omega_2$, and define

$$\tilde{X}_i : (\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 \longmapsto X_i(\omega_i).$$

Now that $\tilde{X}_1$ and $\tilde{X}_2$ are random variables on the same space, we can manipulate them in any way that we can manipulate real-valued functions defined on the same space, which definitely includes taking their sum or product.

This device generalizes to infinite sequences. Suppose that X_i is a random variable on a sample space Ω_i. Let Ω denote the product sample space defined in (8.1), and define

$$\tilde{X}_i : \omega = (\omega_1, \omega_2, \ldots) \in \Omega \longmapsto X_i(\omega_i).$$

The common practice is to redefine X_i as $\tilde{X}_i$, and we do so henceforth without warning.

The moral is that we can always consider a discrete set of random variables as being defined on a common sample space.

What about probability distributions? Suppose that each Ω_i comes equipped with a σ-algebra Σ_i and a distribution $\mathbb{P}_i$.

Problem 8.4. Show that under the product distribution introduced in Section 8.1, the random variables X_i are independent.

Example 8.5 (Tossing a p-coin)**.** Consider an experiment that consists in tossing a p-coin repeatedly without end. Let $X_i = 1$ if the ith trial results in heads, and $X_i = 0$ otherwise, so that, as in (3.5),

$$\mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) = p. \tag{8.4}$$

At this point, the setting is not properly defined. We do so now (although much of what follows is typically left implicit). For each $i \geq 1$, $\Omega_i = \{\mathrm{H}, \mathrm{T}\}$ and $\mathbb{P}_i$ is the distribution on Ω_i given by $\mathbb{P}_i(\mathrm{H}) = p$. The product sample space is $\Omega = \{\mathrm{H}, \mathrm{T}\}^{\mathbb{N}}$ and for $\omega = (\omega_1, \omega_2, \dots) \in \Omega$, let $X_i(\omega) = 1$ if $\omega_i = \mathrm{H}$ and $X_i(\omega) = 0$ if $\omega_i = \mathrm{T}$.

It remains to define $\mathbb{P}$. We know that the distribution $\mathbb{P}$ in (8.4) is the product distribution if and only if the tosses are independent. However, there are other distributions on Ω that satisfy (8.4). For a simple example, define $\mathbb{P}$ on Ω by

$$\mathbb{P}(\mathrm{H}, \mathrm{H}, \mathrm{H}, \dots) = p, \quad \mathbb{P}(\mathrm{T}, \mathrm{T}, \mathrm{T}, \dots) = 1 - p.$$

We can see that (8.4) holds even though the X_i are (very) dependent.

Problem 8.6. For a more interesting example, given $h, t \in [0, 1]$, define the following function on $\{\mathrm{H}, \mathrm{T}\} \times \{\mathrm{H}, \mathrm{T}\}$

$$g(\mathrm{H} \,|\, \mathrm{H}) = h, \tag{8.5}$$
$$g(\mathrm{T} \,|\, \mathrm{H}) = 1 - h, \tag{8.6}$$
$$g(\mathrm{T} \,|\, \mathrm{T}) = t, \tag{8.7}$$
$$g(\mathrm{H} \,|\, \mathrm{T}) = 1 - t. \tag{8.8}$$

Define $f_{(n)}$ on $\{\mathrm{H}, \mathrm{T}\}^n$ as

$$f_{(n)}(\omega_1, \dots, \omega_n) = f(\omega_1) \prod_{i=2}^{n} g(\omega_i \,|\, \omega_{i-1}),$$

where $f(\mathrm{H}) = 1 - f(\mathrm{T}) = p$ is the mass function of a p-coin. Verify that $f_{(n)}$ is a mass function on $\{\mathrm{H}, \mathrm{T}\}^n$. Let $\mathbb{P}_{(n)}$ be the distribution with mass function $f_{(n)}$. Show that $(\mathbb{P}_{(n)} : n \geq 1)$ are compatible in the sense of Theorem 8.2 and let $\mathbb{P}$ denote the distribution on Ω they define. Then find the values of (h, t) that make $\mathbb{P}$ satisfy (8.4).

Remark 8.7. In the previous problem, $\mathbb{P}$ models an experiment using three coins, a p-coin, an h-coin, and a t-coin. We first toss the p-coin, then in the sequence, if the previous toss landed heads, we use the h-coin, otherwise the t-coin.

8.3 Zero-one laws

We start with the Borel–Cantelli lemmas [40], which will lead to an example of a *zero-one law*. These lemmas have to do with an infinite sequence of events and whether infinitely many events among these will happen or not.

To formalize our discussion, consider a probability space $(\Omega, \Sigma, \mathbb{P})$ and a sequence of events $\mathcal{A}_1, \mathcal{A}_2, \ldots$. The event that infinitely many such events happen is the so-called *limit supremum* of these events, defined as

$$\bar{\mathcal{A}} := \bigcap_{m\geq 1}^{\infty} \bigcup_{n\geq m}^{\infty} \mathcal{A}_n.$$

Problem 8.8 (1st Borel–Cantelli lemma)**.** Prove that

$$\sum_{n\geq 1} \mathbb{P}(\mathcal{A}_n) < \infty \;\Rightarrow\; \mathbb{P}(\bar{\mathcal{A}}) = 0.$$

[40] Named after Émile Borel (1871 - 1956) and Francesco Paolo Cantelli (1875 - 1966).

The following converse requires independence.

Problem 8.9 (2nd Borel–Cantelli lemma)**.** Assuming in addition that the events are independent, prove that

$$\sum_{n\geq 1} \mathbb{P}(\mathcal{A}_n) = \infty \;\Rightarrow\; \mathbb{P}(\bar{\mathcal{A}}) = 1.$$

Combining these two lemmas, we arrive at the following.

Proposition 8.10 (Borel–Cantelli's zero-one law)**.** *In the present context, and assuming in addition that the events are independent, we have* $\mathbb{P}(\bar{\mathcal{A}}) = 0$ *or* 1 *according the whether* $\sum_{n\geq 1} \mathbb{P}(\mathcal{A}_n) < \infty$ *or* $= \infty$*.*

Thus, in the context of this proposition, the situation is black or white: the limit supremum event has probability equal to 0 or 1. This is an example of a *zero-one law*. Another famous example is the following.

Theorem 8.11 (Kolmogorov's zero-one law)**.** *Consider an infinite sequence of independent random variables. Any event determined by this sequence, but independent of any finite subsequence, has probability zero or one.*

For example, consider a sequence (X_n) of independent random variables. The event "X_n has a finite limit", which

can also be expressed as

$$\big\{-\infty < \liminf_n X_n = \limsup_n X_n < \infty\big\},$$

is obviously determined by (X_n), and yet it is independent of $(X_1, \dots, X_k)$ for any $k \geq 1$, and thus independent of any finite subsequence. Applying Kolmogorov's zero-one law, this event has therefore probability zero or one.

Problem 8.12. Provide some other examples of such events.

Problem 8.13. Show that the assumption of independence is crucial for the result to hold in this generality, by providing a simple counter-example.

8.4 Convergence of random variables

Random variables are functions on the sample space, therefore, defining notions of convergence for random variables relies on similar notions for sequences of functions. We present the two main notions here.

8.4.1 Convergence in probability

We say that a sequence of random variables $(X_n : n \geq 1)$ *converges in probability* towards a random variable X if, for any fixed $\varepsilon > 0$,

$$\mathbb{P}(|X_n - X| \geq \varepsilon) \longrightarrow 0, \quad \text{as } n \to \infty.$$

We will denote this convergence by $X_n \to_{\mathrm{P}} X$.

Example 8.14. For a simple example, let Y be a random variable, and let $g : \mathbb{R}^2 \to \mathbb{R}$ be continuous in the second variable, and define $X_n = g(Y, 1/n)$. Then

$$X_n \xrightarrow{\mathrm{P}} X := g(Y, 0).$$

This example encompasses instances like $X_n = a_n Y + b_n$, where (a_n) and (b_n) are convergent deterministic sequences.

Problem 8.15. Show that

$$X_n \xrightarrow{\mathrm{P}} X \;\Leftrightarrow\; X_n - X \xrightarrow{\mathrm{P}} 0. \tag{8.9}$$

Problem 8.16. Show that if $X_n \geq 0$ for all n and $\mathbb{E}(X_n) \to 0$, then $X_n \to_{\mathrm{P}} 0$.

Problem 8.17. Show that if $\mathbb{E}(X_n) = 0$ for all n and $\mathrm{Var}(X_n) \to 0$, then $X_n \to_{\mathrm{P}} 0$.

Proposition 8.18 (Dominated convergence)**.** *Suppose that $X_n \to_P X$ and that $|X_n| \leq Y$ for all n, where Y has finite expectation. Then X_n, for all n, and X have an expectation, and $\mathbb{E}(X_n) \to \mathbb{E}(X)$ as $n \to \infty$.*

Problem 8.19. Prove this proposition, at least when Y is constant.

8.4.2 Convergence in distribution

We say that a sequence of distribution functions $(\mathsf{F}_n : n \geq 1)$ *converges weakly* to a distribution function F if, for any point $x \in \mathbb{R}$ where F is continuous,

$$\mathsf{F}_n(x) \longrightarrow \mathsf{F}(x), \quad \text{as } n \to \infty.$$

We will denote this by $\mathsf{F}_n \to_{\mathcal{L}} \mathsf{F}$. A sequence of random variables $(X_n : n \geq 1)$ *converges in distribution* to a random variable X if $\mathsf{F}_{X_n} \to_{\mathcal{L}} \mathsf{F}_X$. We will denote this by $X_n \to_{\mathcal{L}} X$.

Remark 8.20. Unlike convergence in probability, convergence in distribution does not require that the variables be defined on the same probability space.

Remark 8.21. The consideration of continuity points is important. As an illustration, take the simple example of constant variables, $X_n \equiv 1/n$. We anticipate that (X_n) converges weakly to $X \equiv 0$. Indeed, the distribution function of X_n is $\mathsf{F}_n(x) := \{x \leq 1/n\}$, while that of X is $\mathsf{F}(x) := \{x \leq 0\}$. Clearly, $\mathsf{F}_n(x) \to \mathsf{F}(x)$ for all $x \neq 0$, but not at 0 since $\mathsf{F}_n(0) = 0$ and $\mathsf{F}(0) = 1$. Fortunately, weak convergence only requires a pointwise convergence at the continuity points of F, which is the case here.

Problem 8.22. Prove that convergence in probability implies convergence in distribution, meaning that if $X_n \to_{\mathrm{P}} X$ then $X_n \to_{\mathcal{L}} X$. The converse is not true in general (and in fact may not be applicable in view of Remark 8.20). Indeed, let $X \sim \mathrm{Ber}(1/2)$ and define

$$X_n = \begin{cases} X & \text{if } n \text{ is odd,} \\ 1 - X & \text{if } n \text{ is even.} \end{cases}$$

Show that X_n converges weakly to X but not in probability.

Problem 8.23. Show that convergence in distribution to a constant implies convergence in probability to that constant.

More generally, we have the following, which is a corollary of Skorokhod's representation theorem. [41]

Theorem 8.24 (Representation theorem)**.** *Suppose that* (F_n) *converges weakly to* F*. Then there exist* (X_n) *and* X*, random variables defined on the same probability space, with* X_n *having distribution* F_n *and* X *having distribution* F*, and such that* $X_n \to_P X$*.*

[41] Named after Anatoliy Skorokhod (1930 - 2011).

8.5 Law of Large Numbers

Consider Bernoulli trials where a fair coin (i.e., a p-coin with $p = 1/2$) is tossed repeatedly. Common sense (based real-world experience) would lead one to anticipate that, after a large number of tosses, the proportion of heads would be close to 1/2. Thankfully, this is also the case within the theoretical framework built on Kolmogorov's axioms.

Remark 8.25 (iid sequences)**.** Random variables that are independent and have the same marginal distribution are said to be *independent and identically distributed (iid).*

Theorem 8.26 (Law of Large Numbers)**.** *Let* (X_n) *be a sequence of iid random variables with expectation* μ*. Then*

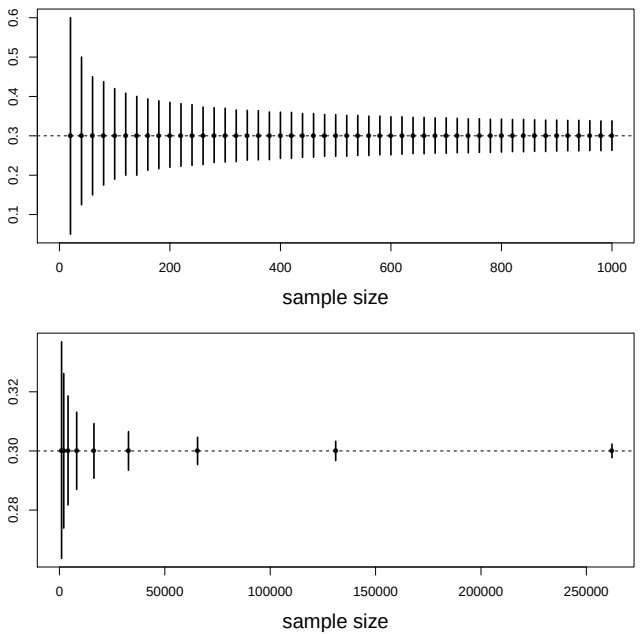
$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu, \quad \text{as } n \to \infty.$$

See Figure 8.1 for an illustration.

If the variables have a 2nd moment, then the result is easy to prove using Chebyshev's inequality.

Problem 8.27. Let $X_1, \ldots, X_n$ be random variables with the same mean μ and variances all bounded by σ^2, and assume their pairwise covariances are non-positive. Let

Figure 8.1: An illustration of the Law of Large Number in the context of Bernoulli trials. The horizontal axis represents the sample size n. The vertical segment corresponding to a sample size n is defined by the 0.005 and 0.995 quantiles of the distribution of the mean of n Bernoulli trials with parameter $p = 1/2$.



$Y_n = \sum_{i=1}^n X_i$. Show that

$$\mathbb{P}\big(|Y_n/n - \mu| \geq \varepsilon\big) \leq \frac{\sigma^2}{n\varepsilon^2}, \quad \text{for all } \varepsilon > 0. \tag{8.10}$$

Problem 8.28. Apply Problem 8.27 to the number of heads in a sequence of n tosses of a p-coin. In particular, find ε such that the probability bound is 5%. Turn this into a statement about the "typical" number of heads in 100 tosses of a fair coin.

Problem 8.29. Repeat with the number of red balls drawn without replacement n times from an urn with r red balls and b blue balls. [The pairwise covariances were computed as part of Problem 7.43.] Make the statement about the "typical" number of red balls in 100 draws from an urn with 100 red and 100 blue balls. How does your statement change when there are 1000 red and 1000 blue balls instead?

Remark 8.30. Theorem 8.26 is in fact known as the Weak Law of Large Numbers. There is indeed a Strong Law of Large Numbers, and it says that, under the same conditions, with probability one,

$$\frac{1}{n}\sum_{i=1}^n X_i \longrightarrow \mu, \quad \text{as } n \to \infty.$$

8.6 Central Limit Theorem

The bound (8.10) can be rewritten as

$$\mathbb{P}\left(\frac{|Y_n - n\mu|}{\sigma\sqrt{n}} \geq t\right) \leq \frac{1}{t^2},$$

for all $t > 0$ and all $n \geq 1$ integer. In fact, under some additional conditions, it is possible to obtain the exact limit as $n \to \infty$.

8.6.1 Central Limit Theorem

Theorem 8.31 (Central Limit Theorem)**.** *Let (X_n) be a sequence of iid random variables with mean μ and variance σ^2. Let $Y_n = \sum_{i=1}^n X_i$. Then $(Y_n - n\mu)/\sigma\sqrt{n}$ converges in distribution to the standard normal distribution, or equivalently, for all $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\frac{Y_n - n\mu}{\sigma\sqrt{n}} \leq t\right) \longrightarrow \Phi(t), \quad \text{as } n \to \infty, \tag{8.11}$$

where Φ is the distribution function of the standard normal distribution defined in (5.2).

Importantly, $n\mu$ is the mean of Y_n and $\sigma\sqrt{n}$ is its standard deviation. So the Central Limit Theorem says that

the standardized sum of iid random variables (with 2nd moment) converges to the standard normal distribution.

Problem 8.32. Show that the Central Limit Theorem encompasses the De Moivre–Laplace theorem (Theorem 5.4) as a special case. In particular, if $(X_i : i \geq 1)$ is a sequence of iid Bernoulli random variables with parameter $p \in (0,1)$, then

$$\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

A standard proof of Theorem 8.31 relies on the Fourier transform, and for that reason is rather sophisticated. So we only provide some pointers. We assume, without loss of generality, that $\mu = 0$ and $\sigma = 1$.

We focus on the case where the X_i have density f and characteristic function φ that is integrable. In that case, we have the Fourier inversion formula (7.23).

Let f_n denote the density of Y_n. Based on (6.11), we know that f_n exists and furthermore that it is the nth convolution power of f. Let $Z_n = Y_n/\sqrt{n}$, which has density $g_n(z) := \sqrt{n}\, f_n(\sqrt{n}z)$. We want to show, or at least argue, that g_n converges to the standard normal density, denoted

$$\phi(z) := \frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}}, \quad \text{for } z \in \mathbb{R}.$$

Because of (7.21), Y_n has characteristic function φ^n.

Problem 8.33. Show that, for any positive integer $n \geq 1$, φ^n is integrable when φ is integrable.

We may therefore apply (7.23) to derive

$$\begin{aligned} g_n(z) &= \frac{\sqrt{n}}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\imath t\sqrt{n}z} \varphi(t)^n \mathrm{d}t \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\imath sz} \varphi(s/\sqrt{n})^n \mathrm{d}s, \end{aligned}$$

using a simple change of variables in the 2nd line.

Problem 8.34. Recall that we assumed that f has zero mean and unit variance. Based on that, show that φ is twice continuously differentiable, with $\varphi(0) = 1$, $\varphi'(0) = 0$, and $\varphi''(0) = 1$. Deduce that

$$\begin{aligned} \varphi(s/\sqrt{n})^n &= (1 - s^2/2n + o(1/n))^n \\ &\to \mathrm{e}^{-s^2/2}, \quad \text{as } n \to \infty. \end{aligned}$$

Thus, if passing to the limit under the integral is justified (✣), we obtain

$$g_n(z) \longrightarrow \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{e}^{-\imath sz} \mathrm{e}^{-s^2/2} \mathrm{d}s, \quad \text{as } n \to \infty.$$

Problem 8.35. Prove that the limit is $\phi(z)$, either directly, or using a combination of (7.23) and the fact that

the standard normal characteristic function is $e^{-s^2/2}$ (Problem 7.96).

This completes the proof that $g_n \to \phi$ pointwise, modulo (✣) above. Even then, this does not prove Theorem 8.31, which is a result on the distribution functions rather than the densities.

8.6.2 Lindeberg's Central Limit Theorem

This well-known variant generalizes the classical version presented in Theorem 8.31. While it still requires independence, it does not require the variables to be identically distributed.

Theorem 8.36 (Lindeberg's Central Limit Theorem [42])**.** *Let $(X_i : i \geq 1)$ be independent random variables, with X_i having mean μ_i and variance σ_i^2. Define $s_n^2 = \sum_{i=1}^n \sigma_i^2$ and assume that, for any fixed $\varepsilon > 0$,*

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{i=1}^n \mathbb{E}\left((X_i - \mu_i)^2 \{|X_i - \mu_i| > \varepsilon s_n\}\right) = 0. \quad (8.12)$$

Then $s_n^{-1} \sum_{i=1}^n (X_i - \mu_i)$ converges in distribution to the standard normal distribution.

[42] Named after Jarl Lindeberg (1876 - 1932).

Problem 8.37. Verify that Theorem 8.36 implies Theorem 8.31.

Problem 8.38. Consider independent Bernoulli variables, $X_i \sim \text{Ber}(p_i)$. Assume that $p_i \leq 1/2$ for all i. Show that (8.12) holds if and only if $\sum_{i=1}^n p_i \to \infty$ as $n \to \infty$.

Problem 8.39 (Lyapunov's Central Limit Theorem [43])**.** Show that (8.12) holds when there is $\delta > 0$ such that

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}\left(|X_i - \mu_i|^{2+\delta}\right) = 0. \quad (8.13)$$

[Use Jensen's inequality.]

8.7 Extreme value theory

Extreme Value Theory is the branch of Probability that studies such things as the extrema of iid random variables. Its main results are 'universal' convergence results, the most famous of which is the following.

Theorem 8.40 (Extreme Value Theorem)**.** *Let (X_n) be iid random variables. Let $Y_n = \max_{i\leq n} X_i$. Suppose that there are sequences (a_n) and (b_n) such that $a_n Y_n + b_n \to_{\mathcal{L}} Z$*

[43] Named after Aleksandr Lyapunov (1857 - 1918).

where Z is not constant. Then Z has either a Weibull, a Gumbel, or a Fréchet distribution.

The *Weibull family*[44] with shape parameter $\kappa > 0$ is the location-scale family generated by

$$\mathsf{G}_\kappa(z) := 1 - \exp(-z^\kappa), \quad z > 0. \tag{8.14}$$

(Thus the entire Weibull family has three parameters.)

The *Gumbel family*[45], is location-scale family generated by

$$\mathsf{G}(z) := 1 - \exp(-\exp(-z)), \quad z \in \mathbb{R}. \tag{8.15}$$

(Thus the entire Gumbel family has two parameters.)

The *Fréchet family*[46] with shape parameter $\kappa > 0$ is the location-scale family generated by

$$\mathsf{G}_\kappa(z) := \exp(-z^{-\kappa}), \quad z > 0. \tag{8.16}$$

(Thus the entire Fréchet family has three parameters.)

Problem 8.41. Verify that (8.14), (8.15), and (8.16) are bona fide distribution functions.

[44] Named after Waloddi Weibull (1887 - 1979).
[45] Named after Emil Julius Gumbel (1891 - 1966).
[46] Named after Maurice René Fréchet (1878 - 1973).

Problem 8.42 (Distributions with finite support)**.** Suppose that the distribution generating the iid sequence has finite support, say $c_1 < \cdots < c_N$. Note that N is fixed. Show that

$$\mathbb{P}(Y_n = c_N) \longrightarrow 1, \quad \text{as } n \to \infty.$$

Deduce that the Extreme Value Theorem does not apply to this case.

Rather than proving the theorem, we provide some examples, one for each case. We place ourselves in the context of the theorem.

Problem 8.43. Let F denote the distribution function of the X_i. Show that the distribution function of Y_n is F^n.

Problem 8.44 (Maximum of a uniform sample)**.** Let $X_1, \ldots, X_n$ be iid uniform in $[0, 1]$. Show that, for any $z > 0$,

$$\mathbb{P}(n(1 - Y_n) \le z) \longrightarrow 1 - \exp(-z), \quad \text{as } n \to \infty.$$

Thus the limiting distribution is in the Weibull family.

Problem 8.45 (Maximum of a normal sample)**.** Let $X_1, \ldots, X_n$ be iid standard normal. Show that, for any $z \in \mathbb{R}$,

$$\mathbb{P}(a_n Y_n + b_n \le z) \longrightarrow \exp(-\exp(-z)), \quad \text{as } n \to \infty,$$

where

$$a_n := \sqrt{2\log n}, \quad b_n = -2\log n + \frac{1}{2}\log\log n + \frac{1}{2}\log(4\pi).$$

Thus the limiting distribution is in the Gumbel family. [To prove the result, use the fact that, Φ denoting the standard normal distribution function,

$$1 - \Phi(x) \sim \frac{1}{\sqrt{2\pi}x}\exp(-x^2/2), \quad \text{as } x \to \infty,$$

which can be obtained via integration by parts.]

Problem 8.46 (Maximum of a Cauchy sample)**.** Let $X_1, \ldots, X_n$ be iid from the Cauchy distribution. We saw the density in (5.10), and the corresponding distribution function is given by

$$\mathsf{F}(x) := \frac{1}{\pi}\tan^{-1}(x) + \frac{1}{2}.$$

Show that, for any $z > 0$,

$$\mathbb{P}\Big(\frac{\pi}{n}Y_n \le z\Big) \longrightarrow 1 - \exp(-1/z), \quad \text{as } n \to \infty.$$

Thus the limiting distribution is in the Fréchet family. [To prove the result, use the fact that $1 - \mathsf{F}(x) \sim 1/\pi x$ as $x \to \infty$.]

8.8 Further topics

8.8.1 Continuous mapping theorem, Slutsky's theorem, and the delta method

The following result says that applying a continuous function to a convergent sequence of random variables results in a convergent sequence of random variables, where the type of convergence remains the same. (The theorem applies to random vectors as well.)

Problem 8.47 (Continuous Mapping Theorem)**.** Let (X_n) be a sequence of random variables and let $g : \mathbb{R} \to \mathbb{R}$ be continuous. Prove that, if (X_n) converges in probability (resp. in distribution) to X, then $(g(X_n))$ converges in probability (resp. in distribution) to $g(X)$.

The following is a simple corollary.

Theorem 8.48 (Slutky's theorem)**.** *If $X_n \to_{\mathcal{L}} X$ while $A_n \to_P a$ and $B_n \to_P b$, where a and b are constants, then $A_nX_n + B_n \to_{\mathcal{L}} aX + b$.*

Problem 8.49. Prove Theorem 8.48.

The following is a refinement of the Continuous Mapping Theorem.

Problem 8.50 (Delta Method)**.** Let (Y_n) be a sequence

of random variables and (a_n) a sequence of real numbers such that $a_n \to \infty$ and $a_n Y_n \to_{\mathcal{L}} Z$, where Z is some random variable. Also, let g be any function on the real line with a derivative at 0 and such that $g(0) = 0$. Prove that $a_n g(Y_n) \to_{\mathcal{L}} g'(0)Z$.

8.8.2 Exchangeable random variables

The random variables $X_1, \ldots, X_n$ are said to be *exchangeable* if their joint distribution is invariant with respect to permutations. This means that for any permutation $(\pi_1, \ldots, \pi_n)$ of $(1, \ldots, n)$, the random vectors $(X_1, \ldots, X_n)$ and $(X_{\pi_1}, \ldots, X_{\pi_n})$ have the same distribution.

Problem 8.51. Show that $X_1, \ldots, X_n$ are exchangeable if and only if their joint distribution function is invariant with respect to permutations. Show that the same is true of the mass function (if discrete) or density (if absolutely continuous).

Problem 8.52. Show that if $X_1, \ldots, X_n$ are exchangeable then they necessarily have the same marginal distribution.

Problem 8.53. Show that independent and identically distributed random variables are exchangeable. Show that the converse is not true.

Problem 8.54. Consider drawing from an urn with red and blue balls. Let $X_i = 1$ if the ith draw is red and $X_i = 0$ if it is blue. Show that, whether the sampling is without or with replacement (Section 2.4), or follows Pólya's scheme (Section 2.4.4), $X_1, \ldots, X_n$ are exchangeable.

Problem 8.55. Let $X_1, \ldots, X_n$ and Y be random variables such that, conditionally on Y the X_i are independent and identically distributed. Show that $X_1, \ldots, X_n$ are exchangeable.

The following provides a converse for a sequence of Bernoulli random variables.

Theorem 8.56 (de Finetti's theorem)**.** *Suppose that (X_n) is an exchangeable sequence of Bernoulli random variables with same parameter. Then there is a random variable Y on $[0,1]$ such that, given $Y = y$, the X_i are iid Bernoulli with parameter y.*

8.9 Additional problems

Problem 8.57 (From discrete to continuous uniform)**.** Let X_N be uniform on $\{\frac{1}{N+1}, \frac{2}{N+1}, \ldots, \frac{N}{N+1}\}$. Show that (X_N) converges in distribution to the uniform distribution on $[0,1]$.

Problem 8.58 (From geometric to exponential)**.** Let X_N be geometric with parameter p_N, where $Np_N \to \lambda > 0$. Show that (X_N/N) converges in distribution to the exponential distribution with rate λ.

Problem 8.59. Consider a sequence of distribution functions (F_n) that converges weakly to some distribution function F. Recall the definition of pseudo-inverse defined in (4.16) and show that $\mathsf{F}_n^-(u) \to \mathsf{F}^-(u)$ at any u where F is continuous and strictly increasing.

Problem 8.60 (Coupon Collector Problem)**.** Recall the setting of Section 3.8. Show that the X_i are exchangeable but not independent.

Problem 8.61 (Coupon Collector Problem)**.** Recall the setting of Section 3.8. We denote T by T_N and let $N \to \infty$. It is known [76] that, for any $a \in \mathbb{R}$,

$$\mathbb{P}(T_N \le N \log N + aN) \to \exp(-\exp(-a)), \quad N \to \infty.$$

Re-express this statement as a convergence in distribution. Note that the limiting distribution is Gumbel. Using the function implemented in Problem 3.22, perform some simulations to confirm this mathematical result.

Problem 8.62 (Coupon Collector Problem)**.** Continuing with the setting of the previous problem, for $q \in [0,1]$, $T_{\lceil qN \rceil}$ is the number of trials needed to sample a fraction of at least q of the entire collection of N coupons. Show that, when $q \in (0,1)$ is fixed while $N \to \infty$, the limiting distribution of $T_{\lceil qN \rceil}$ is normal. [First, express $T_{\lceil qN \rceil}$ as the sum of certain W_i as in Problem 3.21. Then apply Lyapunov's CLT (Problem 8.39).]

Problem 8.63. Suppose that $X_1, \dots, X_n$ are exchangeable with continuous marginal distribution. Define $Y = \#\{i : X_i \ge X_1\}$ and show that Y has the uniform distribution on $\{1, 2, \dots, n\}$. In any case, show that

$$\mathbb{P}(Y \le y) \le y/n, \quad \text{for all } y \ge 1.$$

Problem 8.64. Suppose the distribution of the random vector $(X_1, \dots, X_n)$ is invariant with respect to some set of permutations $\mathcal{S}$, and that $\mathcal{S}$ is such that, for every pair of distinct $i, j \in \{1, \dots, n\}$, there is $\sigma = (\sigma_1, \dots, \sigma_n) \in \mathcal{S}$ such that $\sigma_i = j$. Show that the conclusions of Problem 8.63 apply to this more general situation.

Problem 8.65. Let $X_1, \dots, X_n$ be exchangeable nonnegative random variables. For any $k \le n$, compute

$$\mathbb{E}\left(\frac{\sum_{i=1}^k X_i}{\sum_{i=1}^n X_i}\right).$$

Problem 8.66. Suppose that a box contains m balls. The goal is to estimate m given the ability to sample

uniformly at random with replacement from the urn. We consider a protocol which consists in repeatedly sampling from the urn and marking the resulting ball with a unique symbol. The process stops when the ball we draw has been previously marked. [47] Let K denote the total number of draws in this process. Show that $K/\sqrt{m} \to \sqrt{\pi/2}$ in probability as $m \to \infty$.

Problem 8.67 (Tracy–Widom distribution)**.** Let $\Lambda_{m,n}$ denote the the square of the largest singular value of an m-by-n matrix with iid standard normal coefficients. Then there are deterministic sequences, $a_{m,n}$ and $b_{m,n}$ such that, as $m/n \to \gamma \in [0, \infty]$, $(\Lambda_{m,n} - a_{m,n})/b_{m,n}$ converges in distribution to the so-called *Tracy–Widom distribution* of order 1. In R, perform some numerical simulations to probe into this phenomenon. [Note that the amount of computation is nontrivial.]

[47] This can be seen as a form of capture-recapture sampling scheme (Section 23.5.1).

CHAPTER 9

STOCHASTIC PROCESSES

Stochastic processes model experiments whose outcomes are collections of variables organized in some fashion. We present some classical examples and cover some basic properties to give the reader a glimpse of how much more sophisticated probability models can be.

9.1 Poisson processes

Poisson processes are point processes that are routinely used to model temporal and spatial data.

9.1.1 Poisson process on a bounded interval

The *Poisson process* with constant intensity λ on the interval $[0,1)$ can be constructed as follows. Let N denote a Poisson random variable with parameter λ. When $N = n \geq 1$, draw n points $X_1, \ldots, X_n$ iid from the uniform distribution on $[0,1)$; when $N = 0$, return the empty set. The result is the random set of points $\{X_1, \ldots, X_N\}$ (empty when $N = 0$).

Define $N_t = \#\{i : X_i \in [0,t)\}$. Note that $N = N_1$ and, for $0 \le s < t \le 1$,

$$N_t - N_s = \#\big\{i : X_i \in [s,t)\big\}.$$

For a subset $\mathcal{A} \subset [0,1)$, define

$$\mathsf{N}(\mathcal{A}) := \#\{i : X_i \in \mathcal{A}\}.$$

Remark 9.1. Clearly, $N_t - N_s = \mathsf{N}([s,t))$ for any $0 \le s < t \le 1$, and in particular, $N_t = \mathsf{N}([0,t))$ for all t. In fact, specifying $(N_t : t \in [0,1])$ is equivalent to specifying $(\mathsf{N}(\mathcal{A}) : \mathcal{A}$ Borel set in $[0,1])$. They are both referred to the Poisson process with intensity λ on $[0,1]$. The former is a representation as a (step) function, while the latter is a representation as a so-called *counting measure*. N is often referred to as a *counting process*.

Proposition 9.2. *For any Borel set $\mathcal{A} \subset [0,1]$,*

$$\mathsf{N}(\mathcal{A}) \sim \mathit{Poisson}(|\mathcal{A}|), \tag{9.1}$$

where $|\mathcal{A}|$ denotes the Lebesgue measure of $\mathcal{A}$.

Problem 9.3. Prove this proposition. Start by showing that, given $N = n$, $\mathsf{N}(\mathcal{A})$ has the $\mathrm{Bin}(n, t-s)$ distribution. Then use the Law of Total Probability to conclude.

Proposition 9.4. *For any pairwise disjoint Borel sets $\mathcal{A}_1, \dots, \mathcal{A}_m$ in $[0,1)$, $\mathsf{N}(\mathcal{A}_1), \dots, \mathsf{N}(\mathcal{A}_m)$ are independent.*

Problem 9.5. Prove this proposition, starting with the case $m = 2$. First show that

$$\mathsf{N}(\mathcal{A}_1) + \mathsf{N}(\mathcal{A}_2) \sim \mathrm{Poisson}(\lambda(v_1 + v_2)), \tag{9.2}$$

using (9.1), where $v_i := |\mathcal{A}_i|$. Then reason by conditioning on $\mathsf{N}(\mathcal{A}_1) + \mathsf{N}(\mathcal{A}_2)$, using the Law of Total Probability, and the fact that, given $\mathsf{N}(\mathcal{A}_1) + \mathsf{N}(\mathcal{A}_2) = \ell$, $\mathsf{N}(\mathcal{A}_1)$ has the $\mathrm{Bin}(\ell, v_1/(v_1+v_2))$ distribution, as seen in Problem 3.31.

More generally, the Poisson process with constant intensity λ on the interval $[a,b)$ results from drawing $N \sim \mathrm{Poisson}(\lambda(b-a))$ and, given $N = n \ge 1$, drawing $X_1, \dots, X_n$ iid from the uniform distribution on $[a,b)$.

Problem 9.6. Generalize Proposition 9.2 and Proposition 9.4 to the case of an arbitrary interval $[a,b)$.

Remark 9.7. λ is the density of points per unit length, here meaning that λr is the mean number of points in an interval of length r (inside the interval where the process is defined).

9.1.2 Poisson process on a bounded domain

The Poisson process with intensity λ on $[0,1)^d$ can be constructed by drawing $N \sim \text{Poisson}(\lambda)$ and, given $N = n \geq 1$, drawing $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ iid from the uniform distribution on $[0,1)^d$. More generally, the Poisson process with intensity λ on $[a_1, b_1) \times \cdots \times [a_d, b_d)$ can be constructed by drawing $N \sim \text{Poisson}(\lambda \prod_j (b_j - a_j))$ and, given $N = n \geq 1$, drawing $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ iid uniform on that hyperrectangle.

Problem 9.8. Consider a Poisson process with intensity λ on $[a_1, b_1) \times \cdots \times [a_d, b_d)$. Show that its projection onto the jth coordinate is a Poisson process on $[a_j, b_j)$. What is its intensity?

More generally, for a compact set $\mathcal{D} \subset \mathbb{R}^d$ with positive volume ($|\mathcal{D}| > 0$), the Poisson process with intensity λ on $\mathcal{D}$ is defined by drawing N from the Poisson distribution with mean $\lambda|\mathcal{D}|$ and, given $N = n \geq 1$, drawing $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ iid uniform on $\mathcal{D}$.

Problem 9.9. Generalize Proposition 9.2 and Proposition 9.4 to the Poisson process with intensity λ on $\mathcal{D}$.

9.1.3 Poisson process on $[0, \infty)$

The Poisson process with intensity λ on $[0, \infty)$ can be constructed by independently generating a Poisson process with intensity λ on each interval $[k-1, k), k \geq 1$ integer.

Problem 9.10. Show that the restriction of such a process to any interval is a Poisson process on that interval with same intensity.

Problem 9.11. Show that, in the construction of the process, the partition $[0,1), [1,2), [2,3), \ldots$ can be replaced by any other partition into intervals.

Proposition 9.12. *The Poisson process with intensity λ on $[0, \infty)$ can also be constructed by generating $(W_i : i \geq 1)$ iid from the exponential distribution with rate λ and then defining $X_i = W_1 + \cdots + W_i$.*

In this construction, the X_i are ordered in the sense that $X_1 \leq X_2 \leq \cdots$. In a number of settings, X_i denotes the time of the ith event, in which case $W_i = X_i - X_{i-1}$ represents the time between the $(i-1)$th and ith events. For that reason, the W_i are often called *inter-arrival times*.

9.1.4 General Poisson process

The Poisson processes seen thus far are said to be homogenous. This is because the intensity is constant. In contrast, a general Poisson process may be inhomogeneous in that its intensity may vary over the space over which

the process is defined.

Formally, given a function $\lambda : \mathbb{R}^d \to \mathbb{R}_+$ assumed to have well-defined and finite integral over any compact set, the Poisson process with intensity function λ is a point process over $\mathbb{R}^d$ which, when defined via its counting process N, satisfies the following two properties:

(i) For every compact set $\mathcal{A}$, $\mathsf{N}(\mathcal{A})$ has the Poisson distribution with mean $\int_{\mathcal{A}} \lambda(x)\mathrm{d}x$.

(ii) For any pairwise disjoint compact sets $\mathcal{A}_1, \dots, \mathcal{A}_m$, $\mathsf{N}(\mathcal{A}_1), \dots, \mathsf{N}(\mathcal{A}_m)$ are independent.

That such a process exists is a theorem whose proof is not straightforward and will be omitted. Of course, if λ is a constant function, then we recover an homogeneous Poisson process as defined previously.

9.2 Markov chains

Some situations are poorly modeled by sequences of independent random variables. Think, for example, of the daily closing price of a stock, or the maximum daily temperature on successive days. Markov chains are some of the most natural stochastic processes for modeling dependencies. We provide a very brief introduction, focusing on the case where the observations are in a discrete space. The topic is more extensively treated in the textbook [113]. See also the lectures notes available here. [48]

9.2.1 Definition

Let $\mathcal{X}$ be a discrete space and let $f(\cdot\,|\,\cdot)$ denote a conditional mass function on $\mathcal{X} \times \mathcal{X}$, namely, for each $x_0 \in \mathcal{X}$, $x \mapsto f(x\,|\,x_0)$ is a mass function on $\mathcal{X}$. The corresponding chain, starting at $x_0 \in \mathcal{X}$, proceeds as follows:

(i) X_1 is drawn from $f(\cdot\,|\,x_0)$ resulting in x_1;

(ii) for $t \geq 1$, given $X_t = x_t$, X_{t+1} is drawn from $f(\cdot\,|\,x_t)$ resulting in x_{t+1}.

The result is the outcome $(x_1, x_2, \dots)$, or left unspecified as a sequence of random variables, it is $(X_1, X_2, \dots)$.

Remark 9.13. If in actuality $f(x\,|\,x_0)$ does not depend on x_0, in which case we write it as $f(x)$, the process generates an iid sequence from f. Hence, an iid sequence is a Markov chain.

If the chain is at state $x \in \mathcal{X}$, that state is referred to as the present state. The next state, generated using $f(\cdot\,|\,x)$, is the state one (time) step into the future. That future state is generated without any reference to previous states

[48] Lectures notes by Richard Weber at Cambridge University (statslab.cam.ac.uk/~rrw1/markov)

except for the present state. In that sense, a Markov chain only 'remembers' the present state. See Section 9.2.6 for an extension.

9.2.2 Two-state Markov chains

Let us consider the simple setting of a state space $\mathcal{X}$ with two elements. In fact, we already examined this situation in Problem 8.6. Here we take $\mathcal{X} = \{1, 2\}$ without loss of generality. Because $f(\cdot|\cdot)$ is a conditional mass function it needs to satisfy

$$a := f(1|1) = 1 - f(2|1), \tag{9.3}$$
$$b := f(2|2) = 1 - f(1|2). \tag{9.4}$$

The parameters $a, b \in [0, 1]$ are free and define the Markov chain. The conditional probabilities above may be organized in a so-called *transition matrix*, which here takes the form

$$\text{present state} \begin{Bmatrix} \\ \\ \end{Bmatrix} \overbrace{\begin{pmatrix} a & 1-a \\ 1-b & b \end{pmatrix}}^{\text{next state}} \tag{9.5}$$

Problem 9.14. Starting at $x_0 = 1$, compute the probability of observing $(x_1, x_2, x_3) = (1, 2, 1)$ as a function of (a, b).

Problem 9.15. In R, write a function taking as input the parameters of the chain (a, b), the starting state $x_0 \in \{1, 2\}$, and the number of steps t, and returning a random sequence $(x_1, \dots, x_t)$ from the corresponding process.

9.2.3 General Markov chains

Since we assume the state space $\mathcal{X}$ to be discrete, we may take it to be the positive integers, $\mathcal{X} = \{1, 2, \dots\}$, without loss of generality. For $i, j \in \mathcal{X}$, let $\theta_{ij} = f(j|i)$, which is the probability of transitioning from i to j. These are organized in a transition matrix

$$\text{present state} \begin{Bmatrix} \\ \\ \\ \\ \end{Bmatrix} \overbrace{\begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \cdots \\ \theta_{21} & \theta_{22} & \theta_{33} & \cdots \\ \theta_{31} & \theta_{32} & \theta_{33} & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}}^{\text{next state}} \tag{9.6}$$

Note that the matrix can be infinite in principle. This general case includes the case where the state space is finite. Denote the transition matrix by

$$\Theta := (\theta_{ij} : i, j \geq 1).$$

The transition matrix defines the chain, and together with the initial state, defines the distribution of the random process.

Problem 9.16. Show that, for any $t \geq 1$,

$$\mathbb{P}(X_t = j \mid X_0 = i) = \theta_{ij}^{(t)}, \tag{9.7}$$

where $\Theta^t = (\theta_{ij}^{(t)} : i, j \geq 1)$.

9.2.4 Long-term behavior

In the study of a Markov chain, quantities of interest include the long-term behavior of the chain, the average time (number of steps) it takes to visit a given state (or set of states) starting from a given state (or set of states), the average number of such visits, and more. We focus on the limiting marginal distribution.

We say that a mass function on $\mathcal{X}$, $\mathsf{q} := (q_1, q_2, \dots)$, is a *stationary distribution* of the chain with transition matrix $\Theta := (\theta_{ij} : i, j \geq 1)$ if $X_0 \sim \mathsf{q}$ and $X \mid X_0 \sim f(\cdot \mid X_0)$ yields $X \sim \mathsf{q}$.

Problem 9.17. Show that q is a stationary distribution of Θ if and only if $\mathsf{q}\Theta = \mathsf{q}$, when interpreting q as a row vector. (Note that the multiplication is on the left.)

The chain Θ is said to be *irreducible* if for any $i, j \geq 1$ there is $t = t(i, j)$ such that $\theta_{ij}^{(t)} > 0$. This means that, starting at any state, the chain can eventually reach any other state with positive probability.

The state i is said to be *aperiodic* if

$$\mathsf{gcd}(t \geq 1 : \theta_{ii}^{(t)} > 0) = 1,$$

where gcd is short for 'greatest common divisor'. To understand this, suppose that state i is such that

$$\mathsf{gcd}(t \geq 1 : \theta_{ii}^{(t)} > 0) = 2.$$

Then this would imply that the chain starting at i cannot be at i after an odd number of steps.

A chain is aperiodic if all its states are aperiodic.

Proposition 9.18. *Suppose that the chain is irreducible. If one state is aperiodic then all states are aperiodic.*

State i is *positive recurrent* if, starting at i, the expected time it takes the chain to return to i is finite. A chain is positive recurrent if all its states are positive recurrent.

Proposition 9.19. *A finite irreducible chain is positive recurrent.*

(A finite chain is a chain over a finite state space.)

Theorem 9.20. *An irreducible, aperiodic, and positively recurrent chain has a unique stationary distribution. Moreover, the chain converges weakly to the stationary distribution regardless of the initial state, meaning that, if*

X_t denotes the state the chain is at at time t, and if $\mathsf{q} = (q_1, q_2, \dots)$ denotes the stationary distribution, then

$$\lim_{t\to\infty} \mathbb{P}(X_t = j \mid X_0 = i) = q_j, \quad \textit{for any states } i \textit{ and } j.$$

Problem 9.21. Verify that a finite chain whose transition probabilities are all positive satisfies the requirements of the theorem.

Problem 9.22. Provide necessary and sufficient conditions for a two-state Markov chain to satisfy the requirements of the theorem. When these are satisfied, derive the limiting distribution.

9.2.5 Reversible Markov chains

Running a chain backward Consider a chain with transition probabilities (θ_{ij}) with unique stationary distribution $\mathsf{q} = (q_i)$. Assuming the state is i at time $t = 0$, a step forward is taken according to

$$\mathbb{P}(X_1 = j \mid X_0 = i) = \theta_{ij}, \quad \text{for all } i, j.$$

In contrast, a step backward is taken according to

$$\mathbb{P}(X_{-1} = j \mid X_0 = i) = \frac{\theta_{ij} q_j}{q_i}, \quad \text{for all } i, j. \tag{9.8}$$

Reversible chains In essence, a chain is reversible if running it forward is equivalent (in terms of distribution) to running it backward. More generally, a chain with transition probabilities (θ_{ij}) is said to be *reversible* if there is a probability mass function $\mathsf{q} = (q_i)$ such that

$$q_i \theta_{ij} = q_j \theta_{ji}, \quad \text{for all } i, j. \tag{9.9}$$

This means that, if we draw a state from q and then run the chain for one step, the probability of obtaining (i, j) is the same as that of obtaining (j, i).

Problem 9.23. Show that a distribution q satisfying (9.9) is stationary for the chain.

Problem 9.24. Show that if X_0 is sampled according to q and we run the chain for t steps resulting in $X_1, \dots, X_t$, the distribution of $(X_0, \dots, X_t)$ is the same as that of $(X_t, \dots, X_0)$. [Note that this does not imply that these variables are exchangeable, only that we can reverse the order. See Problem 9.57.]

Example 9.25 (Simple random walk on a graph)**.** A *graph* is a set of nodes and edges between some pairs of nodes. Two nodes that form an edge are said to be neighbors. Assume that each node has a finite number of neighbors. Consider the following process: starting at some node, at each step choose a node uniformly at

random among the neighbors of the present node. Then the resulting chain is reversible.

9.2.6 Extensions

We have discussed the simplest variant of Markov chain: it is called a discrete time, discrete space, time homogeneous Markov process. The definition of a continuous time Markov process requires technicalities that we will avoid. But we can elaborate on the other aspects.

General state space The state space $\mathcal{X}$ does not need to be discrete. Indeed, suppose the state space is equipped with a σ-algebra Γ. What are needed are transition probabilities, $\{\mathsf{P}(\cdot \,|\, x) : x \in \mathcal{X}\}$, on Γ. Then, given a present state $x \in \mathcal{X}$, the next state is drawn from $\mathsf{P}(\cdot \,|\, x)$.

Problem 9.26. Suppose that $(W_t : t \geq 1)$ are iid with distribution P on $(\mathbb{R}, \mathcal{B})$. Starting at $X_0 = x_0 \in \mathbb{R}$, successively define $X_t = X_{t-1} + W_t$. (Equivalently, $X_t = x_0 + W_1 + \cdots + W_t$.) What are the transition probabilities in this case?

Time-varying transitions The transition probabilities may depend on time, meaning

$$\mathbb{P}(X_{t+1} = j \mid X_t = i) = \theta_{ij}(t), \quad \text{for all } i, j,$$

where now a sequence of transition matrices, $\Theta(t) := (\theta_{ij}(t))$, defines the chain.

More memory A Markov chain only remembers the present. However, with little effort, it is possible to have it remember some of its past as well.

The number of states it remembers is the *order* of the chain. A Markov chain of order m is such that

$$\begin{aligned}&\mathbb{P}(X_t = i_t \mid X_{t-m} = i_{t-m}, \ldots, X_{t-1} = i_{t-1})\\ &= \theta(i_{t-m}, \ldots, i_t), \quad \text{for all } i_{t-m}, \ldots, i_t,\end{aligned}$$

where the $(m+1)$-dimensional array

$$(\theta(i_0, i_1, \ldots, i_m) : i_0, i_1, \ldots, i_m \geq 1)$$

now defines the chain.

In fact, a finite order chain can be seen as an order 1 chain in an enlarged state space. Indeed, consider a chain of order m on $\mathcal{X}$, and let X_t denote its state at time t. Based on this, define

$$Y_t := (X_{t-m+1}, \ldots, X_{t-1}, X_t).$$

Then $(Y_m, Y_{m+1}, \dots)$ forms a Markov chain of order 1 on $\mathcal{X}^m$, with transition probabilities

$$\begin{aligned}&\mathbb{P}(Y_t = (i_{t-m+1}, \dots, i_t) \mid Y_{t-1} = (i_{t-m}, \dots, i_{t-1}))\\ &= \theta(i_{t-m}, \dots, i_t), \quad \text{for all } i_{t-m}, \dots, i_t,\end{aligned}$$

and all other possible transitions given the probability 0.

9.3 Simple random walk

Let $(X_i : i \geq 1)$ be iid with

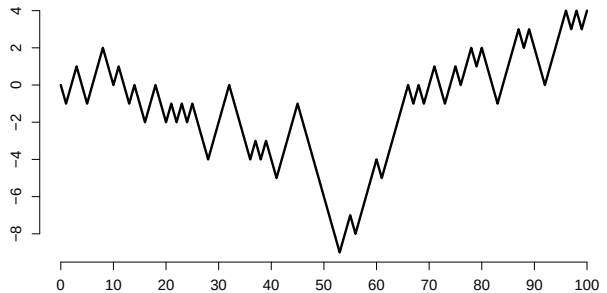
$$\mathbb{P}(X_i = 1) = p, \quad \mathbb{P}(X_i = -1) = 1 - p,$$

and define $S_n = \sum_{i=1}^n X_i$. Note that $(X_i + 1)/2 \sim \text{Ber}(p)$. Then $(S_n : n \geq 0)$, with $S_0 = 0$ by default, is a *simple random walk*.[49] See Figure 9.1 for an illustration.

This is a special case of Example 9.25, where the graph is the one-dimensional lattice. Indeed, at any stage, if the process is at $k \in \mathbb{Z}$, it moves to either $k+1$ or $k-1$, with probabilities p and $1-p$, respectively. When $p = 1/2$, the walk is said to be symmetric (for obvious reasons).

[49] The simple random walk is one of the most well-studied stochastic processes. We refer the reader to Feller's classic textbook [83] for a more thorough yet accessible exposition.

Figure 9.1: A realization of a symmetric ($p = 1/2$) simple random walk, specifically, a plot of a linear interpolation of a realization of $\{(n, S_n) : n = 0, \dots, 100\}$.



Problem 9.27. Show that a simple random walk is a Markov chain with state space $\mathbb{Z}$. Display the transition matrix. Show that the chain is irreducible and periodic. (What is the period?)

Proposition 9.28. *The simple random walk is positive recurrent if and only if it is symmetric.*

9.3.1 Gambler's Ruin

Consider a gambler than bets one dollar at every trial and doubles or looses that dollar, each with probability 1/2

independently of the other trials. Suppose the gambler starts with s dollars and that he stops gambling if that amount reaches 0 (loss) or some prescribed amount $w > s$ (win).

Problem 9.29. Leaving w implicit, let γ_s denote the probability that the gambler loses. Show that

$$\gamma_s = p\gamma_{s+1} + (1-p)\gamma_{s-1}, \quad \text{for all } s \in \{1, \dots, w-1\}.$$

Deduce that

$$\gamma_s = 1 - s/w, \quad \text{if } p = 1/2,$$

while

$$\gamma_s = \frac{1 - (\frac{p}{1-p})^{w-s}}{1 - (\frac{p}{1-p})^{w}}, \quad \text{if } p \neq 1/2.$$

The result applies to $w = \infty$, meaning to the setting where the gambler keeps on playing as long as he has money. In that case, we see that the gambler loses with probability 1 if $p \leq 1/2$ and with probability $(1/p - 1)^s$ if $p > 1/2$. In particular, if $p > 1/2$, with probability $1 - (1/p - 1)^s$, the gambler's fortune increases without bound.

9.3.2 Fluctuations

A realization of a simple random walk can be surprising. Indeed, if asked to guess at such a realization, most (untrained) people would have the tendency to make the walk fluctuate much more than it typically does.

Problem 9.30. In R write a function which plots $\{(k, S_k) : k = 0, \dots, n\}$, where $S_0, S_1, \dots, S_n$ is a realization of a simple random walk with parameter p. Try your function on a variety of combinations of (n, p).

We say that a sign change occurs at step n if $S_{n-1}S_{n+1} < 0$. Note that this implies that $S_n = 0$.

Proposition 9.31 (Sign changes)**.** *The number of sign changes in the first $2n+1$ steps of a symmetric simple random walk equals k with probability $2^{-2n}\binom{2n+1}{2k+1}$.*

9.3.3 Maximum

There is a beautifully simple argument, called the *reflection principle*, which leads to the following clear description of how the maximum of the random walk up to step n behaves

Proposition 9.32. *For a symmetric simple random walk,*

for all $n \geq 1$ *and all* $r \geq 0$,

$$\mathbb{P}\Big(\max_{k\leq n} S_k = r\Big) = \mathbb{P}(S_n = r) + \mathbb{P}(S_n = r+1).$$

Assume the walk is symmetric. Since S_n has mean 0 and standard deviation $\sqrt{n}$, it is natural to study the normalized random walk given by $(S_n/\sqrt{n} : n \geq 1)$.

Theorem 9.33 (Erdös and Rényi [51])**.** *Suppose that* $(X_i : i \geq 1)$ *are iid with mean 0 and variance 1. Define* $S_n = \sum_{i\leq k} X_i$*, as well as* $a_n = \sqrt{2\log\log n}$ *and* $b_n = \frac{1}{2}\log\log\log n$*. Then for any* $t \in \mathbb{R}$,

$$\lim_{n\to\infty} \mathbb{P}\left(\max_{k\leq n} \frac{S_k}{\sqrt{k}} \leq a_n + \frac{b_n}{a_n} + \frac{t}{a_n}\right) = \exp\big(-e^{-t}/2\sqrt{\pi}\big).$$

Thus, the maximum of a simple random walk, properly normalized, converges to a Gumbel distribution.

Problem 9.34. In the context of this theorem, show that

$$\sqrt{2\log\log n} \leq \max_{k\leq n} \frac{S_k}{\sqrt{k}} \leq \sqrt{2\log\log n} + 1,$$

with probability tending to 1 as n increases.

9.4 Galton–Watson processes

Francis Galton (1822 - 1911) and Henry William Watson (1827 - 1903) were interested in the extinction of family names. Their model assumes that the family name is passed from father to son and that each male has a number of male descendants, with the number of descendants being independent and identically distributed.

More formally, suppose that we start with one male with a certain family name. Let $X_0 = 1$. The male has $\xi_{0,1}$ male descendants. If $\xi_{0,1} = 0$, the family name dies. Otherwise, the first male descendant has $\xi_{1,1}$ male descendants, the second has $\xi_{1,2}$ male descendants, etc. In general, let $\xi_{n,j}$ be the number of male descendants of the jth male in the nth generation. The order within each generation is arbitrary and only used for identification purposes. The central assumption is that the ξ are iid. Let f denote their common mass function.

The number of male individuals in the nth generation is thus

$$X_n = \sum_{j=1}^{X_{n-1}} \xi_{n-1,j}.$$

(This is an example of a compound sum as seen in Section 7.10.1.)

Problem 9.35. Show that $(X_1, X_2, \dots)$ forms a Markov chain and give the transition probabilities in terms of f.

Problem 9.36. Provide sufficient conditions on f under which, as a Markov chain, the process is irreducible and

aperiodic.

Note that if $X_n = 0$, then the family name dies and in particular $X_m = 0$ for all $m \geq n$. Of interest, therefore, is

$$d_n := \mathbb{P}(X_n = 0).$$

Clearly, (d_n) is increasing, and being in $[0,1]$, converges to a limit d_∞, which is the probability that the male line dies out eventually. A basic problem is to determine d_∞.

Remark 9.37. In Markov chain parlance, the state 0 is *absorbing* in the sense that, once reached, the chain stays there forever after.

Problem 9.38. Show that $d_\infty > 0$ if and only if $f(0) > 0$.

Problem 9.39. Show that an irreducible chain with an absorbing state cannot be (positive) recurrent. [You can start with a two-state chain, with one state being absorbing, and then generalize from there.]

Problem 9.40 (Trivial setting)**.** The setting is trivial when f has support in $\{0,1\}$. Compute d_n in that case and show that $d_\infty = 1$ when $f(0) > 0$.

We assume henceforth that we are not in the trivial setting, meaning that $f(0) + f(1) < 1$.

Problem 9.41. Show that in that case the state space cannot be taken to be finite.

9.4.1 First-step analysis

A *first-step analysis* consists in conditioning on the value of X_1. Suppose that $X_1 = k$, thus out of the first generation are born k lines. The whole line dies by the nth generation if and only if all these k lines die by the nth generation. But the nth generation in the whole line corresponds to the $(n-1)$th generation for these lines because they started at generation 1 instead of generation 0. By the Markov property, each of these lines has the same distribution as the whole line, and therefore dies by its $(n-1)$th generation with probability d_{n-1}. Thus, by independence, they all die by their $(n-1)$th generation with probability d_{n-1}^k. Thus we proved that

$$\mathbb{P}(X_n = 0 \mid X_1 = k) = d_{n-1}^k.$$

By the Law of Total Probability, this yields

$$\begin{aligned} d_n &= \mathbb{P}(X_n = 0 \mid X_0 = 1) \\ &= \sum_{k \geq 0} \mathbb{P}(X_n = 0 \mid X_1 = k)\, \mathbb{P}(X_1 = k \mid X_0 = 1) \\ &= \sum_{k \geq 0} d_{n-1}^k f(k). \end{aligned}$$

Thus, letting γ denote the probability generating function of f, we showed that

$$d_n = \gamma(d_{n-1}), \quad \text{for all } n \geq 1.$$

Note that $d_0 = 0$ since $X_0 = 1$. Thus, by the continuity of γ on $[0,1]$, d_∞ is necessarily a fixed point of γ, meaning that $\gamma(d_\infty) = d_\infty$. There are some standard ways of dealing with this situation and we refer the reader to the textbook [113] for details.

Let μ be the (possibly infinite) mean of f. This is the mean number of male descendants of a given male. The mean plays a special role, in particular because $\gamma'(1) = \mu$.

Theorem 9.42 (Probability of extinction)**.** *The line becomes extinct with probability one (meaning $d_\infty = 1$) if and only if $\mu \le 1$. If $\mu > 1$, d_∞ is the unique fixed point of γ in $(0,1)$.*

Many more things are known about this process, such as the average growth of the population and features of the population tree.

9.4.2 Extensions

The model has many variants grouped under the umbrella of *branching process*. These processes are used to model a much wider variety of situations, particularly in genetics. While the basic model can be considered asexual (only males carry the family name), some models are bi-sexual in the sense that there are male and female individuals, and in each generation couples are formed in some prescribed (typically random) fashion and offsprings are born out of each union.

9.5 Random graph models

A *graph* is a set equipped with a relationship between pairs of its elements, and is most typically defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes a subset of elements often called *vertices* (aka *nodes*) and $\mathcal{E}$ is a subset of $\mathcal{V} \times \mathcal{V}$ indicating which pairs of elements of $\mathcal{V}$ are related. The fact that $u, v \in \mathcal{V}$ are related is expressed as $(u, v) \in \mathcal{E}$, and (u, v) is then called an *edge* in the graph, and u and v are said to be *neighbors*. A graph is said to be *undirected* if its relationship is symmetric, or said differently, if the edges are not oriented, or formally, if $(u, v) \in \mathcal{E} \iff (v, u) \in \mathcal{E}$. For an undirected graph, the *degree* of vertex is the number of neighbors it has.

A graph is often represented by an *adjacency matrix*. For a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with a finite node set $\mathcal{V}$ taken to be $\{1, \dots, n\}$ without loss of generality, its adjacency matrix is $\boldsymbol{A} = (A_{ij})$ where $A_{ij} = 1$ if $(i, j) \in \mathcal{E}$ and $A_{ij} = 0$ otherwise. Note that an adjacency matrix is binary, with coefficients in $\{0, 1\}$.

9.5.1 Erdős-Rényi and Gilbert models

Arguably the simplest models of random graphs are those introduced concurrently by Paul Erdős (1913 - 1996) and Alfréd Rényi (1921 - 1970) [75] and Edgar Nelson Gilbert (1923 - 2013) [104].

- *Erdős-Rényi model* This model is parameterized by positive integers n and m, and corresponds to the uniform distribution on undirected graphs with n vertices and m edges.
- *Gilbert model* This model is parameterized by a positive integer n and a real numbers $p \in [0, 1]$. A realization from the corresponding distribution is an undirected graph with n vertices where each pair of nodes forms an edge with probability p independently of the other pairs.

Problem 9.43. Describe the Erdős-Rényi model with parameters n and m as a distribution on binary matrices.

Problem 9.44. Describe the Gilbert model with parameters n and p as a distribution on binary matrices.

Problem 9.45. Show that a Gilbert random graph with parameters n and $0 < p < 1$ conditioned on having m vertices is an Erdős-Rényi random graph with parameters n and m.

This starts to explain why the two models are so closely related, and it goes beyond that. Henceforth, we consider the Gilbert model, which is somewhat easier to handle because the edges are present independently of each other.

Very many properties are known of this model [27]. We focus on two properties that are very well-known, both having to do with the connectivity of the graph. A *path* in a graph is a sequence of vertices $(v_1, v_2, \dots)$ such that $(v_t, v_{t+1}) \in \mathcal{E}$ for all $t \geq 1$. For an undirected graph, we say that two vertices, u and v, are *connected* if there is a path in the graph $(w_1, \dots, w_T)$ with $w_1 = u$ and $w_T = v$. A subset of vertices is said to be connected if any of its two nodes are connected. A *connected component* is a connected subset of vertices which are otherwise not connected to vertices outside that subset.

Problem 9.46. Show that a graph (in fact, its vertex set) is partitioned into its connected components.

Largest connected component Consider a Gilbert random graph $\mathcal{G}$ with parameters n and $p = \lambda/n$, where $\lambda > 0$ is fixed while $n \to \infty$. Let C_n denote the size of a largest connected component of $\mathcal{G}$. It turns out that there is phase transition at $\lambda = 1$. Indeed, C_n behaves very differently (as n increases) when $\lambda < 1$ as compared to

when $\lambda > 1$. The setting where $\lambda = 1$ is called the *critical regime*.

In the subcritical regime where $\lambda < 1$, it turns out that connected components are at most logarithmic in size.

Theorem 9.47. *Assume $\lambda < 1$ and define $I_\lambda = \lambda - 1 - \log\lambda$. As $n \to \infty$, $C_n / \log n$ converges in probability to $1/I_\lambda$.*

In contrast, in the supercritical regime where $\lambda > 1$, there is a unique largest connected of size of order n, and the remaining connected components are at most logarithmic in size.

Theorem 9.48. *Assume $\lambda > 1$ and define ζ_λ as the survival probability of a Galton–Watson process based on the Poisson distribution with parameter λ. As $n \to \infty$, C_n/n converges in probability to ζ_λ, and moreover, with probability tending to 1, there is a unique largest connected component. In addition, if $C_n^{(2)}$ denotes the size of a 2nd largest connected component, as $n \to \infty$, $C_n^{(2)} / \log n$ converges in probability to $1/I_\alpha$ where $\alpha = \lambda(1 - \zeta_\lambda)$.*

The critical regime where $\lambda = 1$ is also well-understood. In particular, C_n is of size of order $n^{2/3}$.

Connectivity There is also a phase transition in terms of the whole graph being connected or not. Note that the graph is connected exactly when the largest connected component occupies the entire graph (i.e., $C_n = n$).

Problem 9.49. Prove that $\zeta_\lambda < 1$ for any $\lambda > 0$ and use this to argue that the graph with parameters n and $p = \lambda/n$ is disconnected with probability tending to 1 when λ is fixed.

Theorem 9.50. *Assume that $\lambda = t + \log n$ where $t \in \mathbb{R}$ is fixed. As $n \to \infty$, the probability that the graph is connected converges to $\exp(-e^{-t})$.*

Problem 9.51. Show that, as $n \to \infty$, the probability that the graph is connected converges to 1 if $\lambda - \log n \to \infty$ and converges to 0 if $\lambda - \log n \to -\infty$.

This implies that if $p = a\log(n)/n$, then the probability that the graph is connected converges to 1 if $a > 1$ and converges to 0 if $a < 0$.

9.5.2 Preferential attachment models

Roughly speaking, a *preferential attachment model* is a dynamic random graph model where one node is added to the graph and connected to one or more existing nodes

with preference to nodes with larger degrees. This type of model has a long history [5].

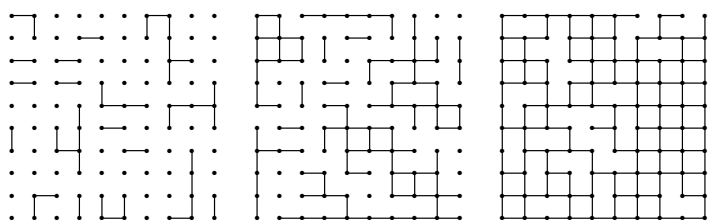
For simplicity, we restrict ourselves here to the case where each new node connects to only one existing node. In that case, the model is of the following form. Let $\mathcal{G}_0$ denote a finite graph, possibly the graph with only one vextex, and let h be a function on the non-negative integer taking non-negative values. At stage $t+1$, a new node is added to $\mathcal{G}_t$ and connected to only one node in $\mathcal{G}_t$. That node is chosen to be u with probability proportional to $h(d_u^{(t)})$, where $d_u^{(t)}$ denotes the degree of node u in $\mathcal{G}_t$. The case where $h(d) = d$ is of particular interest, and goes back to modelization of citation networks [50] in [186].

9.5.3 Percolation models

A simple model for percolation in porous materials is that of a set of *sites* representing holes in the material that may or may not be connected by *bonds* representing pathways that would allow a liquid to move from site to site. The simplest (but not simple) model is that of the square lattice in dimension 2, where the sites are represented by elements of $\mathbb{Z}^2$, thus pairs of integers of the form (i, j), and possible bonds between (i, j) and (k, l) only when $|i-k| + |j-l| = 1$.

In this context, the *bond percolation model* with parameter $p \in [0, 1]$ is the one where bonds are 'open' independently with probability p. The model dates back to Broadbent and Hammersley [31]. See Figure 9.2 for an illustration. Note that, unlike the previous random graph models introduced earlier, percolation models have a spatial interpretation.

Figure 9.2: Realizations of a bond percolation model with $p = 0.2$ (left), $p = 0.5$ (center) and $p = 0.8$ (right) on a 10-by-10 portion of the square lattice.



Although the bond percolation model is much harder to analyze than the Gilbert model, there are parallels. In particular, there is a phase transition in terms of the largest connected component in the graph where the nodes

[50] A citation network is a graph where each node represents an article and (u, v) forms an edge if the article u cites the article v. The resulting graph is thus directed.

are the sites and the bonds are the edges. A connected component is also called an *open cluster*. If we are considering the entire lattice $\mathbb{Z}^2$, the main question of interest is whether there is an open cluster of infinite size.

Problem 9.52. Argue based on Kolmogorov's 0-1 Law that this probability is either 0 or 1. Deduce the existence of a 'critical' probability p_c such that, with probability 1, if $p < p_c$, all open clusters are finite, while if $p > p_c$, there is an infinite open cluster.

Problem 9.53. Assuming that $0 < p < 1$, show that, with probability 1, for each non-negative integer k there is an open cluster of size exactly k.

Theorem 9.54 (Kesten [141])**.** *For the two-dimesionnal lattice,* $p_c = 1/2$.

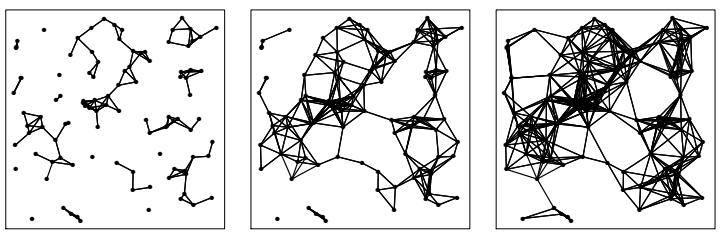
Much more is known. It turns out that, when $p < 1/2$, there is a unique closed cluster [51] of infinite size, and all open clusters are finite, while when $p > 1/2$, the opposite is true meaning that there is a unique open cluster of infinite size, and all closed clusters are finite.

[51] A *closed cluster* is a connected component in the graph where the edges are between sites where the bond is closed.

9.5.4 Random geometric graphs

A realization of the *random geometric graph* with distribution P (on $\mathbb{R}^d$), number of nodes n, and connectivity radius $r > 0$, is generated by sampling n points iid from P and then creating an edge between any two points that are within distance r of each other. These models have a spatial aspect to them, and are used to model sensor positioning systems, for example.

Figure 9.3: Realizations of the random geometric graph model with P being the uniform distribution on $[0,1]^2$, $n = 100$ and $r = 0.10$ (left), $r = 0.15$ (center) and $r = 0.20$ (right).



Gilbert was also interested in such models [105]. Although the two models are seemingly different, there is an interesting parallel. In particular, the same questions related to the size of the largest connected component and

that of connectivity of the graph can also be considered here, and surprisingly enough, the behaviors are comparable. Consider the special case where P is the uniform distribution on the unit square $[0,1]^2$ as in Figure 9.3. Let $X_1, \dots, X_n$ be sampled iid from P and, given these points, let R_n denote the smallest r such that the graph with connectivity radius r is connected.

Theorem 9.55. *As $n \to \infty$, $n\pi R_n^2 / \log n \to 1$ in probability.*

Problem 9.56. Use this result to prove the following. Consider the random geometric graph based on n points uniformly distributed in the unit square with connectivity radius $r_n = (a \log(n)/n)^{1/2}$ with $a > 0$ fixed. Then, as $n \to \infty$, the probability that the graph is connected converges to 1 (resp. 0) if $a > 1/\pi$ (resp. $a < 1/\pi$).

Compared with a Gilbert random graph model, the result is essentially the same. To see this, note that the probability that two points are connected is almost equal to $p := \pi r^2$ when r is small.

Continuum percolation Random geometric graphs are also closely related to percolation models. The main difference here is that the nodes are not on a regular grid. Specifically, consider a Poisson process on the entire plane with constant intensity equal to λ, and then connect every two points that are within distance $r = 1$. This defines a random geometric graph on the entire plane. As before, one of the main questions is whether there is a connected component of infinite size. It turns out that there is a critical $\lambda_c > 0$ such that, if $\lambda < \lambda_c$, all the connected components are finite, while if $\lambda > \lambda_c$, there is a unique infinite component.

9.6 Additional problems

Problem 9.57. Consider a Markov chain and a distribution q on the state space such that, if the chain is started at X_0 drawn from q, then $(X_0, X_1, X_2, \dots)$ are exchangeable. Show that, necessarily, the sequence is iid with distribution q.

Problem 9.58 (Three-state Markov chains)**.** Repeat Problem 9.22 with a three-state Markov chain.

Problem 9.59. Consider a Markov chain with a symmetric transition matrix. Show that the chain is reversible and that the uniform distribution is stationary.

Problem 9.60 (Random walks)**.** A sequence of iid random variables, (X_i), defines a *random walk* by taking the

partial sums, $S_n := X_1 + \cdots + X_n$ for $n \geq 1$ integer. The X_i are called the *increments*. It turns out that random walks with increments having zero mean and finite second moment behave similarly. Perform some numerical experiments to ascertain this claim.

Part B

Practical Considerations

CHAPTER 10

SAMPLING AND SIMULATION

In this chapter we introduce some tools for sampling from a distribution. We also explain how to use computer simulations to approximate probabilities and, more generally, expectations, which can allow one to circumvent complicated mathematical derivations.

10.1 Monte Carlo simulation

Consider a probability space $(\Omega, \Sigma, \mathbb{P})$ and suppose that we want to compute $\mathbb{P}(\mathcal{A})$ for a given event $\mathcal{A} \in \Sigma$. There are several avenues for that.

Analytic calculations In some situations, it might be possible to compute this probability (or at least approximate it) directly by calculations 'with pen and paper' (or use a computer to perform symbolic calculations), and possibly a simple calculator to numerically evaluate the final expression. In the pre-computer age, researchers with sophisticated mathematical skills spent a lot of effort

on such problems and, almost universally, had to rely on some form of approximation to arrive at a useful result. (We will see some example of that in later chapters.)

NUMERICAL SIMULATIONS In some situations, it might be possible to generate independent realizations from $\mathbb{P}$. By the Law of Large Numbers (Theorem 8.26), if $\omega_1, \omega_2, \dots$ are sampled iid from $\mathbb{P}$,

$$Q_m := \frac{1}{m} \sum_{i=1}^{m} \{\omega_i \in \mathcal{A}\} \xrightarrow{\text{P}} \mathbb{P}(\mathcal{A}), \quad \text{as } m \to \infty.$$

The idea, then, is to choose a large integer m, generate $\omega_1, \dots, \omega_m$ iid from $\mathbb{P}$, and then output Q_m as an approximation to $\mathbb{P}(\mathcal{A})$. The larger m, the better the approximation, and a priori the only reason to settle for a particular m are the available computational resources. This approach is often called *Monte Carlo simulation*.[52]

Applying Chebyshev's inequality (7.27), we derive

$$\left|Q_m - \mathbb{P}(\mathcal{A})\right| \le \frac{t}{2\sqrt{m}}, \tag{10.1}$$

with probability at least $1 - 1/t^2$. For example, choosing $t = 10$ we have that $\left|Q_m - \mathbb{P}(\mathcal{A})\right| \le 5/\sqrt{m}$ with probability at least 99%. Therefore, an approximation based on m Monte Carlo draws is accurate to within order $1/\sqrt{m}$.

Problem 10.1. Verify the assertions made here.

Problem 10.2. Apply Chernoff's bound for the binomial distribution (7.32) to obtain a sharper bound on the probability of (10.1).

Problem 10.3. Consider the following problem.[53] A gardener plants three maple trees, four oaks, and five birch trees in a row. He plants them in random order, each arrangement being equally likely. What is the probability that no two birch trees are next to one another? Compute this probability analytically. Then, using R, approximate this probability by Monte Carlo simulation.

Problem 10.4 (Monty Hall by simulation)**.** In [237], Andrew Vazsonyi tells us that even Paul Erdös, a prominent figure in Probability Theory and Combinatorics, was challenged by the Monty Hall problem (Example 1.33). Vazsonyi performed some computer simulations to demonstrate to Erdös that the solution was indeed 1/3 (no switch) and 2/3 (switch). Do the same in R. First, simulate the process when there is no switch. Do that many times (say $m = 10^6$) and record the fraction of successes. Repeat,

[52] See [71] for an account of early developments at Los Alamos National Laboratory in the context research in nuclear fission.

[53] This problem appeared in the AIME, 1984 edition.

this time when there is a switch.

10.2 Monte Carlo integration

Monte Carlo integration applies the principle underlying Monte Carlo simulation to the computation of expectations.

Suppose we want to integrate a function h on $[0,1]^d$. Typical numerical integration methods work by evaluating h on a grid of points and approximating the integral with a linear combination of these values. The simplest scheme of this sort is based on the definition of the Riemann integral. For example, in dimension $d = 1$,

$$\int_0^1 h(x)\mathrm{d}x \approx \frac{1}{m}\sum_{i=1}^{m} h(x_i),$$

where $x_i := i/m$. This is based on a piecewise constant approximation to h. If h is smoother, a higher-order approximation would yield a better approximation for the same value of m.

R corner. The function integrate in R uses a quadratic approximation on an adaptive grid.

Such methods work well in any dimension, but only in theory. Indeed, in practice, a grid in dimension d, even for moderately large d, is too large, even for modern computers. For instance, suppose that we want to sample the function every 1/10 along each coordinate. In dimension d, this requires a grid of size 10^d, meaning there are that many x_i. If $d \geq 10$, this number is quite large already, and if $d \geq 100$, it is beyond hope for any computer.

Monte Carlo integration provides a way to approximate the integral of h at a rate of order $1/\sqrt{m}$ regardless of the dimension d. The simplest scheme uses randomness and is motivated as before by the Law of Large Numbers. Indeed, let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ be iid uniform in $[0,1]^d$. Then $h(\boldsymbol{X}_1), \ldots, h(\boldsymbol{X}_m)$ are iid with mean

$$\mathsf{I}_h := \int_{[0,1]^d} h(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

which is the quantity of interest.

Problem 10.5. Show that $h(\boldsymbol{X}_1), \ldots, h(\boldsymbol{X}_m)$ have finite second moment if and only if h is square-integrable (meaning that h^2 is integrable) over $[0,1]^d$. Then compute their variance (denoted σ_h^2 in what follows).

Applying Chebyshev's inequality to

$$I_m := \frac{1}{m}\sum_{i=1}^{m} h(\boldsymbol{X}_i),$$

we derive

$$\left|I_m - \mathsf{I}_h\right| \le \frac{t\sigma_h}{\sqrt{m}},$$

with probability at least $1 - 1/t^2$.

Although computing σ_h is likely as hard, or harder, than computing I_h itself, it can be easily bounded when an upper bound on h is known. Indeed, if it is known that $|h| \le b$ over $[0,1]^d$, then $\sigma_h \le b$. (Note that numerically verifying that $|h| \le b$ over $[0,1]^d$ can be a challenge in high dimensions.)

Problem 10.6. Verify the assertions made here

Problem 10.7. Compute $\int_0^1 \sqrt{1-x^2}\,\mathrm{d}x$ in three ways:
(i) Analytically. [Change variables $x \to \sin(u)$.]
(ii) Numerically, using the function integrate in R.
(iii) By Monte Carlo integration, also in R.

10.3 Rejection sampling

Suppose we want to sample from the uniform distribution with support $\mathcal{A}$, a compact set in $\mathbb{R}^d$. Translating and scaling $\mathcal{A}$ as needed, we may assume without loss of generality that $\mathcal{A} \subset [0,1]^d$. Then consider the following procedure: repeatedly sample a point from $\mathrm{Unif}([0,1]^d)$ until the point belongs to $\mathcal{A}$, and return that last point. It turns out that the resulting point has the uniform distribution on $\mathcal{A}$. This comes from the following fact.

Problem 10.8. Let $\mathcal{A} \subset \mathcal{B}$, where both $\mathcal{A}$ and $\mathcal{B}$ are compact subsets of $\mathbb{R}^d$. Let $X \sim \mathrm{Unif}(\mathcal{B})$. Show that, conditional on $X \in \mathcal{A}$, X is uniform in $\mathcal{A}$.

Problem 10.9. In R, implement this procedure for sampling from the lozenge in the plane with vertices $(1,0), (0,1), (-1,0), (0,-1)$.

This is arguably the most basic example of *rejection sampling*. The name comes from the fact that draws are repeatedly rejected unless a prescribed condition is met.

Problem 10.10. In R, implement a rejection sampling algorithm for 'estimating' the number π based on the fact that the unit disc (centered at the origin and of radius one) has surface area equal to π. How many samples should be generated to estimate π with this method to within precision ε with probability at least $1 - \delta$? Here $\varepsilon > 0$ and $\delta > 0$ are given.

In general, suppose that we want to sample from a distribution with density f on $\mathbb{R}^d$. Let f_0 be another density on $\mathbb{R}^d$ such that

$$f(x) \le c f_0(x), \quad \text{for all } x \text{ in the support of } f, \qquad (10.2)$$

for some constant c. The density f_0 plays the role of *proposal distribution*. Besides (10.2), the other requirement is that we need to be able to sample from f_0. Assuming this is the case, consider Algorithm 1.

Algorithm 1 Basic rejection sampling

Input: target density f, proposal density f_0, constant c satisfying (10.2).
Output: one realization from f

Repeat: generate y from f_0 and u from Unif($[0,1]$), independently
Until $u \le f(y)/cf_0(y)$
Return the last y

To see that Algorithm 1 outputs a realization from f, let $Y \sim f_0$ and $U \sim \text{Unif}(0,1)$ be independent, and define the event $\mathcal{V} := \{U \le f(Y)/cf_0(Y)\}$. If X denotes the output of Algorithm 1, then X has the distribution of $Y \,|\, \mathcal{V}$. Thus, for any Borel set $\mathcal{A}$,

$$\begin{aligned}\mathbb{P}(X \in \mathcal{A}) &= \mathbb{P}(Y \in \mathcal{A} \,|\, \mathcal{V}) \\ &= \frac{\mathbb{P}(Y \in \mathcal{A} \text{ and } \mathcal{V})}{\mathbb{P}(\mathcal{V})},\end{aligned}$$

with

$$\mathbb{P}(Y \in \mathcal{A} \text{ and } \mathcal{V}) = \int_{\mathcal{A}} \mathbb{P}(\mathcal{V} \,|\, Y = y) f_0(y) \mathrm{d}y \tag{10.3}$$

$$= \int_{\mathcal{A}} \frac{f(y)}{cf_0(y)} f_0(y) \mathrm{d}y \tag{10.4}$$

$$= \frac{1}{c} \int_{\mathcal{A}} f(y) \mathrm{d}y, \tag{10.5}$$

where in the 2nd line we used the fact that

$$\mathbb{P}(\mathcal{V} \,|\, Y = y) = \mathbb{P}(U \le f(y)/cf_0(y) \,|\, Y = y) \tag{10.6}$$

$$= \mathbb{P}(U \le f(y)/cf_0(y)) \tag{10.7}$$

$$= f(y)/cf_0(y), \tag{10.8}$$

since U is independent of Y and uniform in $[0,1]$. By taking $\mathcal{A} = \mathbb{R}^d$, this gives

$$\mathbb{P}(\mathcal{V}) = \mathbb{P}(Y \in \mathbb{R}^d \text{ and } \mathcal{V}) = \frac{1}{c} \int_{\mathbb{R}} f(y) \mathrm{d}y = \frac{1}{c}.$$

Hence,

$$\mathbb{P}(X \in \mathcal{A}) = \int_{\mathcal{A}} f(y) \mathrm{d}y,$$

and this being valid for any Borel set $\mathcal{A}$, we have established that X has density f, as desired.

Problem 10.11. Let S be the number of samples generated by Algorithm 1. Show that $\mathbb{E}(S) = c$. What is the distribution of S?

Problem 10.12. From the previous problem, we see that the algorithm is more efficient the smaller c is. Show that $c \geq 1$ with equality if and only if f and f_0 are densities for the same distribution.

The *ratio of uniforms* is another rejection sampling method proposed by Kinderman and Monahan [142]. It is based on the following.

Problem 10.13. Suppose that g is non-negative and integrable over the real line with integral b, and define $\mathcal{A} := \{(u,v) : 0 < v < \sqrt{g(u/v)}\}$. Assuming that $\mathcal{A}$ has finite area (i.e., Lebesgue measure), show that if (U,V) is uniform in $\mathcal{A}$, then $X := U/V$ has distribution $f := g/b$.

Problem 10.14. Implement the method in R for the special case where g is supported on $[0,1]$.

10.4 Markov chain Monte Carlo (MCMC)

Markov chains can be used to sample from a distribution when doing so 'directly' is not available. In discrete settings, this may be the case because the space is too large and there is no simple way of enumerating the elements in the space. We consider such a setting in what follows, in particular since we only discussed Markov chains over discrete state spaces. Let $\mathsf{q} = (q_i)$ be a mass function on a discrete space from which we want to sample. The idea is to construct a chain, meaning devise a transition matrix $\Theta = (\theta_{ij})$, such that the reversibility condition (9.9) holds. If in addition the chain satisfies the requirements of Theorem 9.20, then the chain converges in distribution to q. Thus a possible method for generating an observation from q, at least approximately, is as in Algorithm 2. Obviously, we need to be able to sample from the distribution q_0.

Algorithm 2 Basic MCMC sampling

Input: chain Θ, initial distribution q_0, total number of steps t
Output: one state

Initialize: draw a state according to q_0
Run the chain Θ for t steps
Return the last state

Remark 10.15. If q_0 coincides with q, then the method is exact since q is stationary. (In the present context, choosing q_0 equal to q is of course not an option.) More generally, the closer q_0 is to q, the more accurate the method is (for a given number of steps t).

10.4.1 Binary matrices with given row and column sums

We are tasked with sampling uniformly at random from the set of $m \times n$ matrices with entries in $\{0, 1\}$ with given row sums, $r_1, \ldots, r_m$, and given column sums, $c_1, \ldots, c_n$. Let that set be denoted by $\mathcal{M}(\boldsymbol{r}, \boldsymbol{c})$, where $\boldsymbol{r} := (r_1, \ldots, r_m)$ and $\boldsymbol{c} = (c_1, \ldots, c_n)$. Importantly, we assume that we already have in our possession one such matrix, which has the added benefit of guarantying that this set is non-empty. This setting is motivated by applications in Psychometry and Ecology (Section 22.1).

The space $\mathcal{M}(\boldsymbol{r}, \boldsymbol{c})$ is typically gigantic and there is no known way to enumerate it to enable drawing from the uniform distribution directly. [54] However, an MCMC approach is viable. The following is based on the work of Besag and Clifford [20].

The chain is defined as follows. At each step, choose two rows and two columns uniformly at random. If the resulting submatrix is of the form

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

then switch one for the other. If the resulting submatrix is not of this form, then stay put.

Problem 10.16. Show that this chain is indeed a reversible chain on $\mathcal{M}(\boldsymbol{r}, \boldsymbol{c})$ and that the uniform distribution is stationary. To complete the picture, show that the chain satisfies the requirements of Theorem 9.20. [The only real difficulty is proving irreducibility.]

Problem 10.17. Staying put may seem like a waste of time (i.e., computational resources). Show that skipping that compromises the algorithm in that the uniform distribution may not be stationary anymore. [For example, examine the case of 3-by-3 binary matrices with row and column sums equal to $(2, 1, 1)$.]

10.4.2 Generating a sample

Typically it is desired to generate not one but several independent samples from a given distribution q. Doing so (approximately) using MCMC is possible if we already have a Markov chain with q as limiting distribution. An obvious procedure is to repeat the process, say Algorithm 2, the desired number of times n to obtain an iid sample of size n from a distribution that approximates the target distribution q.

[54] In fact, merely computing the cardinality of $\mathcal{M}(\boldsymbol{r}, \boldsymbol{c})$ is difficult enough [59].

This process is deemed wasteful in situations where the chain converges slowly to its stationary distribution. Indeed, in such circumstances, the number of steps t in Algorithm 2 to generate a single draw can be quite large, and if $(x_0, \dots, x_t)$ represents a realization of the chain, then $x_0, \dots, x_{t-1}$ are discarded and only x_t is returned by the algorithm. This would be repeated n times, thus generating $n(t+1)$ states to only keep n of them.

Various methods and heuristics exist to attempt to make better use of the states computed along the way. The main issue is that states generated in sequence by a single run of the chain are dependent. Nevertheless, the following generalization of the Law of Large Numbers is true. [55]

Theorem 10.18 (Ergodic theorem)**.** *Consider a Markov chain on a discrete state space $\mathcal{X}$. Assume the chain is irreducible and positive recurrent, and with stationary distribution q. Let X_0 have any distribution on $\mathcal{X}$ and start the chain at that state, resulting in $X_1, X_2, \dots$. Then for any bounded function h,*

$$\frac{1}{t}\sum_{s=1}^{t} h(X_s) \xrightarrow{P} \sum_{x\in\mathcal{X}} \mathsf{q}(x)h(x), \quad \text{as } t \to \infty.$$

[55] A Central Limit Theorem also holds under some additional conditions.

10.5 Metropolis–Hastings algorithm

This algorithm offers a method for constructing a reversible Markov chain for MCMC sampling. It is closely related to rejection sampling, as we shall see. We consider the discrete case, although the same procedure applies more generally almost verbatim.

Suppose we want to sample from a distribution with mass function q. The algorithm seeks to express the transition probability $\mathsf{p}(\cdot\,|\,\cdot)$ as follows

$$\mathsf{p}(x\,|\,x_0) = \mathsf{p}_0(x\,|\,x_0)a(x\,|\,x_0), \tag{10.9}$$

where $\mathsf{p}_0(\cdot\,|\,\cdot)$ is the proposal conditional mass function and $\mathsf{a}(\cdot\,|\,\cdot)$ is the acceptance probability function. The transition probability $\mathsf{p}(\cdot\,|\,\cdot)$ is reversible with stationary distribution q if

$$\mathsf{p}(x\,|\,x_0)\mathsf{q}(x_0) = \mathsf{p}(x_0\,|\,x)\mathsf{q}(x), \quad \text{for all } x, x_0.$$

When p is as in (10.9), this condition is equivalent to

$$\frac{\mathsf{a}(x\,|\,x_0)}{\mathsf{a}(x_0\,|\,x)} = \frac{\mathsf{p}_0(x_0\,|\,x)\mathsf{q}(x)}{\mathsf{p}_0(x\,|\,x_0)\mathsf{q}(x_0)}, \quad \text{for all } x, x_0.$$

Problem 10.19. Prove that

$$\mathsf{a}(x\,|\,x_0) := 1 \wedge \frac{\mathsf{p}_0(x_0\,|\,x)\mathsf{q}(x)}{\mathsf{p}_0(x\,|\,x_0)\mathsf{q}(x_0)} \tag{10.10}$$

satisfies this condition.

The *Metropolis–Hastings algorithm*[56] is an MCMC algorithm with Markov chain of the form (10.9), with p_0 chosen by the user and a as in (10.10). A detailed description is given in Algorithm 3.

Algorithm 3 Metropolis–Hastings sampling

Input: target q, proposal p_0, initial distribution q_0, number of steps t

Output: one state

Initialize: draw x_0 from q_0
For $s = 1, \dots, t$
  draw x from $\mathsf{p}_0(\cdot \,|\, x_{s-1})$
  draw u from $\mathrm{Unif}(0,1)$ and set $x_s = x$ if $u \le \mathsf{a}(x \,|\, x_{t-1})$ and $x_s = x_{s-1}$ otherwise
Return the last state x_t

Remark 10.20. Importantly, we only need to be able to compute $\mathsf{q}(x)/\mathsf{q}(x_0)$ for two states x_0, x. This makes the method applicable in settings where $\mathsf{q} = c\tilde{\mathsf{q}}$ with c a normalizing constant that is hard to compute while $\tilde{\mathsf{q}}$ a function that is relatively easy to evaluate.

[56] Named after Nicholas Metropolis (1915 - 1999) and Wilfred Keith Hastings (1930 - 2016).

Example 10.21 (Ising model [57])**.** The *Ising model* is a model of ferromagnetism where the (iron) atoms are organized in a regular lattice and each atom has a spin which is either − or +. We consider such a model in dimension two. Let $x_{ij} \in \{-1, +1\}$ denote the spin of the atom at position (i, j) in the m-by-n rectangular lattice $\{1, \dots, m\} \times \{1, \dots, n\}$. In its simplest form, the Ising model presumes that the set of random variables $\boldsymbol{X} = (X_{ij})$ has a distribution of the form

$$\mathbb{P}(\boldsymbol{X} = \boldsymbol{x}) = \mathsf{C}(u, v) \exp(u\xi(\boldsymbol{x}) + v\zeta(\boldsymbol{x})), \tag{10.11}$$

where $u, v \in \mathbb{R}$ are parameters, $\boldsymbol{x} = (x_{ij})$ with $x_{ij} \in \{-1, +1\}$, and

$$\xi(\boldsymbol{x}) := \sum_{(i,j)} x_{ij},$$

and

$$\zeta(\boldsymbol{x}) := \frac{1}{2} \sum_{(i,j)} \sum_{\leftrightarrow (k,l)} x_{ij} x_{kl},$$

with $(i, j) \leftrightarrow (k, l)$ if and only if $i = k$ and $|j - l| = 1$ or $|i - k| = 1$ and $j = l$.

The normalization constant $\mathsf{C}(\mathrm{u,v})$ may be difficult to compute in general, as in principle it involves summing

[57] Named after Ernst Ising (1900 - 1998).

over the whole state space, which is of size 2^{mn}. However, the functions ξ and ζ are rather easy to evaluate, which makes a Metropolis–Hastings approach particularly attractive. We present a simple variant. We say that $\boldsymbol{x}$ and $\boldsymbol{x}'$ are neighbors, denoted $\boldsymbol{x} \leftrightarrow \boldsymbol{x}'$, if they differ in exactly one node of the lattice. We choose as $\mathsf{p}_0(\cdot \,|\, \boldsymbol{x}')$ the uniform distribution over the neighbors of $\boldsymbol{x}'$. Then the acceptance probability takes the following simple form

$$\begin{aligned}\mathsf{a}(\boldsymbol{x} \,|\, \boldsymbol{x}') &= 1 \wedge \frac{\mathsf{q}(\boldsymbol{x})}{\mathsf{q}(\boldsymbol{x}')} \\ &= 1 \wedge \exp\bigl[u(\xi(\boldsymbol{x}) - \xi(\boldsymbol{x}')) + v(\zeta(\boldsymbol{x}) - \zeta(\boldsymbol{x}'))\bigr].\end{aligned}$$

Note that, if $\boldsymbol{x}$ and $\boldsymbol{x}'$ differ at (i, j), then

$$\xi(\boldsymbol{x}) - \xi(\boldsymbol{x}') = x_{ij} - x'_{ij},$$

while

$$\zeta(\boldsymbol{x}) - \zeta(\boldsymbol{x}') = \sum_{(k,l)\leftrightarrow(i,j)} (x_{ij}x_{kl} - x'_{ij}x'_{kl}),$$

where the last sum is only over the neighbors of (i, j) (and there are at most 4 of them).

Problem 10.22. In R, simulate realizations of such an Ising model using the Metropolis–Hastings algorithm just described. Do so for $m = 100$ and $n = 200$, and various choices of parameters a and b, chosen carefully to exhibit different regimes. [The realizations can be visualized using the function image.]

10.6 Pseudo-random numbers

We have assumed in several places that we have the ability to generate random numbers, at least from simple distributions, such as the uniform distribution on $[0, 1]$. Doing so, in fact, presents quite a conundrum since the computer is a deterministic machine. The conundrum is solved by the use of a *pseudo-random number generator*, which is a program that outputs a sequence of numbers that are not random but designed to behave as if they were random.

10.6.1 Number π

To take a familiar example, the digits in the representation of π come close to achieving this. Here [58] are the first 100 digits of π (in base 10)

3 1 4 1 5 9 2 6 5 3 5 8 9 7 9 3 2 3 8 4
6 2 6 4 3 3 8 3 2 7 9 5 0 2 8 8 4 1 9 7
1 6 9 3 9 9 3 7 5 1 0 5 8 2 0 9 7 4 9 4

[58] Taken from the On-Line Encyclopedia of Integer Sequences (oeis.org/A000796/b000796.txt).

4 5 9 2 3 0 7 8 1 6 4 0 6 2 8 6 2 0 8 9
9 8 6 2 8 0 3 4 8 2 5 3 4 2 1 1 7 0 6 7

Although obviously deterministic, the sequence of digits defining π behaves very much like a sequence of iid random variables from the uniform distribution on $\{0, \dots, 9\}$.

Problem 10.23. Suppose that you have access to Routine A, which provides the ability to generate an iid sequence of random variables from the uniform distribution on $\{0, \dots, 9\}$ of any prescribed length n. Explain how you would use Routine A to implement Routine B, the one described in Problem 2.41. With Remark 2.17 and Problem 3.28, explain how you would use Routine B to approximately sample from any prescribed distribution with finite support.

However, it turns out that π is not as 'random' as one would want, and besides that, it is not clear how one would use it to draw digits.

10.6.2 Linear congruential generators

These generators produce sequences (x_n) of the form

$$x_n = (ax_{n-1} + c) \mod m,$$

where a, c, m are given integers chosen appropriately. The starting value x_0 needs to be provided and is called the *seed*.

The sequence (x_n) is in $\{0, \dots, m-1\}$ and designed to behave like an iid sequence from the uniform distribution on that set.

R corner. The default generator in R is the Mersenne–Twister algorithm [164]. We refer the reader to [69] for more details, as well as a comprehensive discussion of pseudo-random number generators in R.

CHAPTER 11

DATA COLLECTION

Statistics is the science of data collection and data analysis. We only provide, in this chapter, a brief introduction to principles and techniques for data collection, traditionally divided into survey sampling and experimental design.

While most of this book is on mathematical theory, covering aspects of Probability Theory and Statistics, the collection of data is, by nature, much more practical, and often requires domain-specific knowledge.

Example 11.1 (Collection of data in ESP experiments)**.** In [191], magician and paranormal investigator James 'The Amazing' Randi relates the story of how scientists at the Stanford Research Institute (SRI) were investigating a person claiming to have psychic abilities. The scientists were apparently fooled by relatively standard magic tricks into believing that this person was indeed psychic. This has led Randi, and others such as Diaconis [58], to strongly recommend that a person competent in magic or deception be present during an ESP experiment or be consulted during the planning phase of the experiment.

Even though we will spend much more time on data analysis (Part C), careful data collection is of paramount importance. Indeed, data that were improperly collected can be completely useless and unsalvageable by any technique of analysis. And it is worth keeping in mind that the collection phase is typically much more expensive that the analysis phase that ensues (e.g., clinical trials, car crash tests, etc). Thus the collection of data should be carefully planned according to well-established protocols or with expert advice.

11.1 Survey sampling

Survey Sampling is the process of sampling a population to determine characteristics of that population. The population is most often modeled as an urn, sometimes idealized to be infinite. The type of surveys that we will consider are those that involve sampling only a (usually small) fraction of the population.

Remark 11.2 (Census)**.** A census aims for an exhaustive survey of the population. Any statistical analysis of census data is necessarily descriptive since (at least in principle) the entire population is revealed. Some adjustments, based on complex statistical modeling, may be performed to attempt to palliate some deficiencies having to do with the undercounting of some subpopulations. See [97] for a relatively nontechnical introduction to the census as conducted by the US Census Bureau and a critique of such adjustments.

We present in this section some essentials and refer the reader to Chapter 19 in [92] or Chapter 4 in [236] for more comprehensive, yet gentle introductions.

11.1.1 Survey sampling as an urn model

Consider polling a given population. Suppose the poll consists in one multiple-choice question with s possible choices. Let's say that n people are polled and asked to answer the question (with a single choice). Then the survey can be modeled as sampling from an urn (the population) made of balls with s possible colors.

Note that the possible choices usually include one or several options like "I do not know", "I am not aware of the issue", etc, for individuals that do not have an opinion on the topic, or are unaware of the issue, or are undecided in other ways. See Table 11.1 for an example. Special care may be warranted to deal with nonrespondents and people that did not properly fill the questionnaire.

Table 11.1: New York Times - CBS News poll of 1022 adults in the US (March 11-15, 2016). "Do you think re-establishing relations between the US and Cuba will be mostly good for the US or mostly bad for the US?"

mostly good	mostly bad	unsure/no answer
62%	24%	15%

11.1.2 Simple random sampling

It is often desirable to sample uniformly at random from the population. If this is achieved, the experiment can be modeled as sampling from an urn where at each stage each ball has the same probability of being drawn. The sampling is typically done without replacement. (This is the case in standard polls, where a person is only interviewed once.) We studied the resulting probability model in Section 2.4.2.

It turns out that sampling uniformly at random from a population of interest is rather difficult to do in practice. Modern sampling of human populations is often done by phone based on a method for sampling phone numbers that has been designed with great care.

11.1.3 Bias in survey sampling

When simple random sampling is desired but the actual survey results in a different sampling distribution, it is said that the sampling is *biased*. Such a sample may be said to not representative of the underlying population.

There are a number of factors that could lead to a biased sample, including the following:

- *Self-selection bias* This may occur when people can volunteer to take the poll.
- *Non-response bias* This occurs when the nonrespondents differ in opinion on the question of interest from the respondents.
- *Response bias* This may occur, for example, when the way the question is presented has an unintended (and often unanticipated) influence on the response.

Self-selection bias and non-response bias are closely related. See Section 11.1.4 for an example where they may have played out. The following provides an example where response bias might have influence the outcome of a US presidential election.

Example 11.3 (2000 US presidential election)**.** In 2000, George W. Bush won the presidential election by an extremely small margin against his main opponent, Al Gore.

Indeed, Bush won the deciding state, Florida, by a mere 537 votes. [59] It has been argued that this very slim advantage might have been reversed if not for some difficulties with some ballot designs used in the state, in particular, the "butterfly ballot" used in the county of Palm Beach, which may have lead some voters to vote for another candidate, Pat Buchanan, instead of Gore [2, 215, 245].

The following types of sampling are generally known to generate biased samples:

- *Quota sampling* In this scheme, each interviewer is required to interview a certain number of people with certain characteristics (e.g., socio-economic). Within these quotas, the interviewer is left to decide which persons to interview. The natural inclination (typically unconscious) is to reach out to people that are more accessible and seem more likely to agree to be interviewed. This generates a bias.
- *Sampling volunteers* This is when individuals are given the opportunity to volunteer to take the poll. A prototypical example is a television poll where viewers are asked a question and are given the opportunity to answer that question by calling or texting a given phone number. This sampling scheme leads, almost by definition, to self-selection bias.
- *Convenience sampling* The last two schemes above are examples of convenience sampling. A prototypical example is an individual asking his friends who they will vote for in an upcoming election. Almost by construction, there is no hope that this will result in a sample that is representative of the population of interest (presumably, all eligible voters).

[59] This margin was in fact so small that it required a recount (by state law). However, in a controversial (and 5-4 split) decision, the US Supreme Court halted the recount, in the process overruling the Florida Supreme Court.

Remark 11.4 (Coverage error)**.** Bias typically leads to some members of the population being sampled with a higher probability than other members of the population. This is problematic if the intent is to sample the population uniformly at random. On the other hand, this is fine if the resulting sampling distribution is known, as there are ways to deal with non-uniform distributions. See Remark 11.6 and Section 23.1.2.

11.1.4 A large biased sample is still biased

A simple random sample is typically much smaller than the entire population it is generated from. Even then, as long as the sample size is sufficiently large, the sample

is representative of the population. As we shall see, this happens rather quickly, even if the underlying population is very large or even infinite. By contrast, a biased sample cannot be guarantied to be representative of the population, even if the sample size is comparable to the size of the entire population. We can thus state the following (echoed in [92, Ch 19, Sec 10]).

Proverb. *Large samples offer no protection against bias.*

Literary Digest Poll An emblematic example of this is a Literary Digest poll of the 1936 US presidential election. The main contenders that year were Franklin Roosevelt (D) and Alf Landon (R). The Digest (a US magazine at the time) mailed 10 million questionnaires resulting in 2.3 million returned. The poll predicted an overwhelming victory for Landon, with about 57% of the respondents in his favor. On election day, however, Roosevelt won by a landslide with 62% of the vote. [60]

Problem 11.5. That year, about 44.5 million people voted. Suppose for a moment that the sample collected by the Digest was unbiased. If so, as in Problem 8.29, derive a typical range for the proportion favoring Roosevelt in such a poll.

What happened? For one thing, the response rate (about 23%) was rather small, so that any non-response bias could be substantial. Also, and perhaps more importantly, the sampling favored more affluent people. Indeed, the list of recipients was compiled from a variety of sources, including car and telephone owners, club memberships, and their own readers, and in the 1930's, a person on that list would likely be more affluent (and therefore more supportive of the Republican candidate) than the typical voter.

By comparison, George Gallup — who went on to found the renown Gallup polling company — accurately predicted the result of the election with a sample of size 50,000. In fact, modern polls are typically even smaller, based on samples of size in the few thousands.

The moral of this story is that a smaller, but less biased sample may be preferable to a larger, but more biased sample. (This statement assumes that there is no possibility of correcting for the bias.) The story itself is worth reading in more detail, for example, in [92, 221, 236].

[60] The numbers are taken from [92]. They are a little bit different in [221].

11.1.5 Examples of sampling plans

We briefly describe (mostly by example) a number of sampling plans that are in use in various settings.

The following sampling designs are typically used for reasons of efficiency, cost, or (limited) resources.

- *Systematic sampling* An example of such a sampling plan would be, in the context of an election poll conducted in a residential suburb, to interview every tenth household. Here one relies implicitly on the assumption that the residents are (already) distributed in a way that is independent of the question of interest.
- *Cluster sampling* An example of such a sampling plan would be, in the same context, to interview every household on several blocks chosen in some random fashion among all blocks in the suburb. For a *two-stage* variant, households could be sampled at random within each selected block. (This is an example of *multi-stage sampling*.)
- *Network sampling* An example of that would be, in the context of surveying a population of drug addicts, to ask a person all the addicts he knows, which are then interviewed in turn, or otherwise observed or counted. This type of sampling is indeed popular when surveying hard-to-reach populations. There are many variants, known under various names such as *chain-referral sampling*, *respondent-driven sampling*, or *snowball sampling*. These are related to *web crawls* performed in the World Wide Web. Some of these designs may be considered to be of convenience, particularly when they do not involve randomization.

The following sampling designs are meant to improve on simple random sampling.

- *Stratified sampling* An example would be, in the context of estimating the average household income in a city, to divide the city into socio-economic neighborhoods (which play the role of strata here and would have to be known in advance) and then sample at random housing units in each neighborhood. In general, stratified sampling improves on simple random sampling when the strata are more homogeneous in terms of the quantity or characteristic of interest.

Remark 11.6 (Post-stratification)**.** A *stratification* is sometimes done *after* collecting the sample. This can be used, in some circumstances, to correct a possible bias in the sample.

Example 11.7 (Polling gamers)**.** For an example of post-stratification, consider the polling scheme described

in [246]. Quoting from the text, this was "an opt-in poll which was available continuously on the Xbox gaming platform during the 45 days preceding the 2012 US presidential election. Each day, three to five questions were posted, one of which gauged voter intention. [...] The respondents were allowed to answer at most once per day. The first time they participated in an Xbox poll, respondents were also asked to provide basic demographic information about themselves, including their sex, race, age, education, state [of residence], party [affiliation], political ideology, and who they voted for in the 2008 presidential election".

The survey resulted in a very large sample. However, as the authors warn, "the pool of Xbox respondents is far from being representative of the voting population".

An (apparently successful) attempt to correct for the obvious bias was made using post-stratification based on the side information collected on each respondent. In the authors' own words, "the central idea is to partition the data into thousands of demographic cells, estimate voter intent at the cell level [...], and finally aggregate the cell-level estimates in accordance with the target population's demographic composition".

11.2 Experimental design

> *To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*
>
> Ronald A. Fisher

An experiment is designed with a particular purpose in mind. An example is that of comparing treatments for a certain medical condition. Another example is that of comparing NPK fertilizer mixtures to optimize the yield of a certain crop.

In this section we present some essential elements of experimental design and then describe some classical designs. For a book-length exposition, we recommend the book by Oehlert [178].

A good design follows some fundamental principles proven to lead to correct analyses (avoiding systematic bias) with good power (due to increased precision). Some of these principles include

- *Randomization* To avoid systematic bias.
- *Replication* To increase power.

- *Blocking* To better control variation.

Importantly, the design needs to be chosen with care *before* the collection of data starts.

11.2.1 Setting

The setting of an experiment consists of experimental units, a set of treatments assigned to the experimental units, and the response(s) measured on the experimental units. For example, in a medical experiment comparing two or more treatments for a certain disease, the experimental units are the human subjects and the response is a measure of the severity of the symptoms associated with the disease. (Such an experiment is often referred to as a *clinical trial*.) In an agricultural experiment where several fertilizer mixtures are compared in terms of yield for a certain crop, the experimental units may be the plots of land, the treatments are the fertilizer mixtures, and the response is the yield. (The plants may be referred to as measurement units.)

The way the experimental units are chosen, the way the treatments are assigned to the experimental units, and the way the response is measured, are all part of the design.

There may be several (real-valued) response measurements that are collected in a single experiment. Even then, there is typically one primary response, the other responses being secondary. We will focus on the primary response (henceforth simply called 'response').

Remark 11.8. Although not strictly part of the design, the method of analysis should be decided beforehand. Some of the dangers of not doing so are discussed in Section 23.8.3.

11.2.2 Enrollment (Sampling)

The inference drawn from the experiment (e.g., 'treatment A shows a statistically significant improvement over treatment B') applies, strictly speaking, to the experimental units themselves. For the inference to apply more broadly — which is typically desired — additional conditions on the design need to be fulfilled in addition to randomization (Section 11.2.4).

For example, in medicine, a clinical trial is typically set up to compare the efficacy of two or more treatments for a particular disease or set of symptoms. Human subjects are enrolled in the trial and their response to one (or several) of the treatment(s) is recorded. Based on this limited information, the goal is often to draw conclusions on the relative effectiveness of the treatments when given to members of a certain population. (This is invariably

true of clinical trials for drug development.) For this to be possible, as we have already saw in Section 11.1, the sample of subjects in the trial needs to be representative of the population.

Example 11.9 (Psychology experiments on campuses)**.** Psychology experiments carried out on academic campuses have been criticized on that basis, that the experiments are often conducted on students while the conclusions that are reached are, at least implicitly, supposed to generalize to a much larger population [123, 205]. Undeniably, such samples are of convenience and generalization to a larger population, although potentially valid, is far from automatic.

11.2.3 Power calculations

In addition to a protocol for enrolling subjects that will (hopefully) yield a sample that is representative of the population of interest, the sample size needs to be large enough that the experiment, properly conducted and analyzed, would lead to detecting an effect (typically the difference in efficacy between treatments) of a certain prescribed size if such an effect is present. The target effect size is typically chosen based on scientific, societal, or commercial importance. For instance, for an experiment comparing treatments A and B, the investigators might want to calculate a minimum sample size n that would enable them to detect a potential 10% difference in response between the two treatments.

Such *power calculations* are important and we will come back to them in Section 23.8. Indeed, simply put, if the sample size is too small to detect what would be considered an interesting or important effect, then what is the point of conducting the experiment? The issue is exacerbated by the fact that conducting an experiment typically requires a substantial amount of resources.

11.2.4 Randomization

Randomization consists in assigning treatments to experimental units following a pre-specified random mechanism. This process is typically performed on a computer. We already saw the central role that randomness plays in survey sampling. The same is true in experimental design, where randomization helps avoid systematic bias.

Confounding Systematic bias may be due to *confounding*, which happens when the effect of a factor (i.e., a characteristic of the experimental units) is related both to the received treatment and to the measured response. Importantly, a factor may go unaccounted for.

Example 11.10 (Prostate cancer)**.** Staging in surgical patients with prostate cancer includes the dissection of lymph nodes. If the nodes are found to be cancerous, it is accepted that the disease has spread over the body and a prostatectomy (the removal of the prostate) is not performed. Such staging is not part of other approaches to prostate cancer such as radiotherapy and, as argued in [155], can lead to a comparison that unfairly favors surgery. In such comparisons, the survival time is often the response, and in the context of comparing the effectiveness surgery with (say) radiotherapy in prostate cancer, the grade (i.e., severity) of the cancer is a likely confounder.

Problem 11.11. Name a potential confounder identified by the meta-analysis described in Example 20.39.

Randomization offers some protection against confounding, and an experiment where randomization is employed may enable causal inference (Section 22.2).

BLINDING Just as in survey sampling where little or no freedom of decision is given to the surveyor, randomization is done using a computer to prevent an experimenter (a human administering the treatments) from introducing bias. However, an experimenter has been known to bias the results in other, sometimes quite subtle ways. To minimize bias of any kind, the experimenter is often blinded to the treatment he is administering to a given unit.

In addition to that, in particular in experiments involving human subjects, the subjects are blinded to the treatment they are receiving. This is done, for example, to minimize non-compliance.

A clinical trial where both the experimenter and the subjects are blind to the administered treatment is said to be *double-blind*. This is the gold standard and has motivated the routine use of a *placebo* when no other control is available.[61] See [54] for a historical perspective.

Example 11.12 (Placebo effect)**.** As a treatment is often costly and may induce undesirable side effects, it is important that it perform better than a placebo. As defined in [139], "placebo effects are improvements in patients' symptoms that are attributable to their participation in the therapeutic encounter, with its rituals, symbols, and

[61] In fact, placebos are often the only control even when another competing treatment is available. Indeed, according to [222]: "The FDA requires developers of new treatments to demonstrate that they are safe and effective in order to receive approval for market entry, but the agency demands proof of superiority to existing products only when it is patently unethical to withhold treatment from study patients, as in the cases of therapies for AIDS and cancer. Many new drugs are approved on the basis of demonstrated superiority to placebo. Even less is required for many new medical devices."

interactions". These effects can be substantial. For example, as argued in [244], "for psychological disorders, particularly depression, it has been shown that pill placebos are nearly as effective as active medications".

Example 11.13 (Sham surgery)**.** A double-blind design is not always possible. This is particularly true in surgery as, almost invariably, the operating surgeon knows whether he is performing a real or a sham procedure. For example, in [210], arthroscopic partial meniscectomy (which amounts to trimming the meniscus) is compared with sham surgery (which serves as placebo) to relieve symptoms attributed to meniscal tear. (Incidentally, this is another example where the placebo is shown to be as effective as the active procedure.)

To further avoid bias, this time at the level of the analysis, it is sometimes recommended that the analyst(s) also be blinded to the detailed results of the experiments, for example, by anonymizing the treatments. This blinding is, in principle, unnecessary if the analysis is decided before the experiment starts.

Remark 11.14 (Keeping a trial blind)**.** Humans are curious creatures, and in a long clinical trial, with some lasting a decade or more, it becomes tempting to guess what treatment is given to whom and which treatment is most effective. It turns out that keeping a clinical trial blind is a nontrivial aspect of its design, and for which some techniques have been specifically developed (see, e.g., [99, Ch 7]).

Inference based on randomization Although the main purpose of randomization is to offer some protection against confounding, it also allows for a rather natural form of inference (Section 22.1.1). In a nutshell, assume the goal is to compare treatments. If in reality there is no difference between treatments, then the responses that result from the experiment, understood as random variables, are exchangeable (Section 8.8.2) with respect to the randomization. This invariance can be utilized for inferential purposes.

11.2.5 Some classical designs

Completely randomized designs A completely randomized design simply assigns the treatments at random with the only constraint being on the treatment group sizes, which are chosen beforehand. In detail, assume there are n experimental units and g treatments. Suppose we decide that Treatment j is to be assigned to a total of n_j units, where $n_1 + \cdots + n_g = n$. Then n_1 units are chosen uniformly at random and given Treatment 1; n_2 units are

chosen uniformly at random and given Treatment 2; etc. (The sampling is without replacement.)

Example 11.15 (Back pain)**.** In [3], the efficacy of a device for managing back pain is examined. In this randomized, double-blind trial, the treatment consisted of six sessions of transcutaneous electrical nerve stimulation (TENS) combined with mechanical pressure, while the placebo excluded the electrical stimulation.

Remark 11.16 (Balanced designs)**.** To optimize the power of the subsequent analysis, it is usually preferable that the design be *balanced* as much as possible, meaning that the treatment group sizes be as close to equal as possible. In a perfectly balanced design, $r := n/g$ is an integer representing the number of units per treatment group.

Remark 11.17 (Sequential designs)**.** This is perhaps the simplest of all designs for comparing several treatments. Some sequential variants are often employed in clinical trials where subjects enter the trial over time. The simplest consists in assigning Treatment j with probability p_j to a subject entering the trial, where $p_1, \ldots, p_g$ are chosen beforehand. *A/B testing* is the term most often used in the tech industry. For example, the administrators of a website may want to try different configurations to maximize revenue (add, sales, etc). [62]

Complete block designs Blocking is used to reduce the variance. Blocks are defined with the goal of forming more homogeneous groups of units. In its simplest variant, called a *randomized complete block design*, each block is randomized as in a completely randomized design, and the randomization is independent from block to block. When there is at least one unit for each treatment within each block, the design is complete.

In a balanced design, if there is exactly one unit within each block assigned to each treatment, we have $n = gb$, where b denotes the number of blocks. Further replication can be realized when, for example, each experimental unit contains several measurement units.

Blocking is often done based on one or several characteristics (aka factors) of the units. In clinical trials, common blocking variables include gender and age group.

Example 11.18 (Carbon sequestration)**.** A randomized complete block design is used in [159] to study the effects of different cultural practices on carbon sequestration in soil planted with switchgrass.

[62] For example, the company VWO (vwo.com) offers to run such experiments for websites and other entities.

Incomplete block designs These are designs that involve blocking but in which the blocks have fewer units than treatments. The simplest variant is the *balanced incomplete block design*. Suppose, as before, that there are g treatments and that n experimental units are available for the experiments, and let b denote the number of blocks. Such a design is structured so that each block has the same number of units (say k) and each treatment is assigned to the same number of units (say r). This is referred to as the *first-order balance* and requires $n = kb = gr$. The *second-order balance* refers to each pair of treatments appearing together in the same number of blocks (say λ) and requires that $\lambda := r(k-1)/(g-1)$. See Table 11.2 for an illustration.

Split plots As the name indicates, this design comes from agricultural experiments and is useful in situations where a factor is 'hard to vary'. To paraphrase an example given in [178], suppose that some land is available for an experiment meant to compare the productivity of 3 varieties of rice (A, B, C). We want to control the effect of irrigation and consider 2 levels of irrigation, say high and low. Irrigation may be hard and/or costly to control spatially. An option is to consider plots (called *whole plots*) that are sufficiently separated so that their irrigation can be done independently of one another. These plots are then subdivided into plots (called *split plots*), each planted with one of the varieties of rice under consideration. In such a design, irrigation is randomized at the whole plot level, while seed variety is randomized at the split plot level within each whole plot.

Table 11.2: Example of a balanced incomplete block design with $g = 5$ treatments (labeled A, B, C, D, E), $b = 5$ blocks each with $k = 4$ units, $r = 4$ replicates per treatment, resulting in each pair of treatments appearing together in $\lambda = 3$ blocks.

Block 1	C	A	D	B
Block 2	A	E	B	C
Block 3	B	D	E	C
Block 4	E	C	A	D
Block 5	B	D	E	A

Group testing Robert Dorfman [65] proposed an experimental design during World War II to minimize the number of blood tests required to test for syphilis in soldiers. His idea was to test blood samples from a number

of soldiers simultaneously, repeating the process a few times. Group testing has been used in many other settings including quality control in product manufacturing and cDNA library screening.

More generally, consider a setting where we are testing n individuals for a disease based on blood samples. We consider the case where the design is set beforehand, as opposed to sequential. In that case, the design can be represented by an n-by-t binary matrix $\boldsymbol{X} = (x_{i,j})$ where $x_{i,j} = 1$ if blood from the ith individual is in the jth testing pool, and $x_{i,j} = 0$ otherwise. In particular, t denotes the total number of testing pools. The result of each test is either positive $\oplus$ or negative $\ominus$. We assume for simplicity that the test is perfectly accurate in that it comes up positive when applied to a testing pool if and only if the pool includes the blood sample of at least one affected individual.

The design is *d-disjunct* if the sum of any of its d rows does not contain any other row [67], where in the present context we say that a row vector $u = (u_j)$ contains a row vector $v = (v_j)$ if $u_j \geq v_j$ for all j.

Problem 11.19. If the design is d-disjunct, and there are at most d diseased individuals in total, then each non-diseased individual will appear in at least one pool with no diseased individual. Deduce from this property a simple procedure for identifying the diseased individuals.

Thus a design that is d-disjunct allows the experimenter to identify the diseased individuals as long as there are at most d of them. Note that this property is sufficient for that, but not necessary, although it offers the advantage of a particularly simple identification procedure.

A number of ways have been proposed for constructing disjunct designs, the goal being to achieve a d-disjunct design with a minimum number of testing pools t for a given number of subjects n. In particular, there is a parallel with the construction of codes in the area of Information Theory (Section 23.6). We content ourselves with constructions that rely on randomness. In the simplest such construction, the elements of the design matrix, meaning the $x_{i,j}$, are iid Bernoulli with parameter p.

Proposition 11.20. *The probability that a random n-by-t design with Bernoulli parameter p is d-disjunct is at least*

$$1 - (d+1)\binom{n}{d+1}\left[1 - p(1-p)^d\right]^t.$$

The proof is in fact elementary and relies on the union bound. For details, see the proof of [67, Thm 8.1.3].

Problem 11.21. For which value of p is the design most likely to be d-disjunct? For that value of p, deduce that

there is a numeric constant C_0 such that this random group design is d-disjunct with probability at least $1 - \delta$ when

$$t \geq C_0\big[d^2 \log(n/d) + d \log(d/\delta)\big].$$

Repeated measures This is a type of design that is commonly used in longitudinal studies. For example, some human subjects are 'followed' over time to better understand how a certain condition evolves depending on a number of factors.

Example 11.22 (Neck pain)**.** In [32], 191 patients with chronic neck pain were randomized to one of 3 treatments: (A) spinal manipulation combined with rehabilitative neck exercise, (B) MedX rehabilitative neck exercise, or (C) spinal manipulation alone. After 20 sessions, the patients were assessed 3, 6, and 12 months afterward for self-rated neck pain, neck disability, functional health status, global improvement, satisfaction with care, and medication use, as well as range of motion, muscle strength, and muscle endurance.

Such a design looks like a split plot design, with subjects as whole plots and the successive evaluations as split plots. The main difference is that there is no randomization at the split plot level.

Crossover design In a clinical trial with a crossover design, each subject is given each one of the treatments that are being compared and the relevant response(s) to each treatment is(are) measured. There is generally a *washout* period between treatments in an attempt to isolate the effect of each treatment or, said differently, to minimize the *residual effect* (aka *carryover effect*) of the preceding treatment. In addition, the order in which a subject receives the treatments needs to be randomized to avoid any systematic bias. First-order balance further imposes that the number of subjects that receive treatment X as their jth treatment not depend on X or j. When this is the case, the design is called a *Latin square design*. See Table 11.3 for an illustration.

Table 11.3: Example of a crossover design with 4 subjects, each receiving 4 treatments in sequence based on a Latin square design, with each treatment order appearing only once.

Subject 1	A	B	C	D
Subject 2	C	D	A	B
Subject 3	B	D	E	C
Subject 4	D	C	B	A

Example 11.23 (Medical cannabis)**.** The study [74] evaluates the potential of cannabis for pain management in 34 HIV patients suffering from neuropathic pain in a crossover trial where the treatments being compared are placebo (without THC) and active (with THC) cannabis cigarettes.

Example 11.24 (Gluten intolerance)**.** The study [57] is on gluten intolerance. It involves 61 adult subjects without celiac disease or any other (formally diagnosed) wheat allergy, who nevertheless believe that ingesting gluten causes them some digestive problems. The design is a crossover design comparing rice starch (which serves as placebo) and actual gluten.

Second-order balance further imposes that each treatment follow every other treatment an equal number of times.

Problem 11.25. Determine whether the design in Table 11.3 is second-order balanced. If not, find such a design.

Matched-pairs design In a study that adopts this design to compare two treatments, subjects that are similar in key attributes (i.e., possible confounders) are matched and the randomization to treatment happens within each pair. (Thus this is a special case of a complete block design where each pair forms a block.)

Example 11.26 (Cognitive therapy for PTSD)**.** In [160], which took place in Dresden, Germany, 42 motor vehicle accident survivors with post-traumatic syndrome were recruited to be enrolled in a study designed to examine the efficacy of cognitive behavioral treatment (CBT) protocols and methods. Subjects were matched after an initial assessment and then randomized to either CBT or control (which consisted of no treatment).

11.3 Observational studies

Let us start with an example.

Example 11.27 (Role of anxiety in postoperative pain)**.** In [138], 53 women who underwent an hysterectomy where assessed for anxiety, coping style, and perceived stress two weeks prior to the intervention. This was repeated several times during the recovery period. Analgesic consumption and level of perceived pain were also measured.

This study shares some of the attributes of a repeated measures design or a crossover design, yet there was no randomization involved. Broadly speaking, the ability

to implement randomization is what distinguishes experimental designs and observational studies.

11.3.1 Why observational studies?

There is wide agreement that randomization should be a central part of a study whenever possible, because it offers some protection against confounding. As discussed in [185, 214, 255], there are quite a few examples of observational studies that were later refuted by randomized trials.

However, there are situations where researchers resort to an observational study. We follow Nick Black [23], who argues that experiments and observational studies are complementary. Although conceding that clinical trials are the gold standard, he says that "observational methods are needed [when] experimentation [is] unnecessary, inappropriate, impossible, or inadequate".

- *Experimentation may be unnecessary* when the effect size is very large. (Black cites the obvious effectiveness of penicillin for treating bacterial infections.)
- *Experimentation may be inappropriate* in situations where the sample size needed to detect a rare side effect of a drug, for example, far exceeds the size of a feasible clinical trial. Another example is the detection of long-term adverse effects, as this would require a study that exceeds in length that of most typical clinical trials.
- *Experimentation may be impossible* for a number of reasons. One reason could be *ethics*. For example, randomizing pregnant women to smoking and non-smoking groups is not ethical in large part because we know that smoking carries substantial negative effects for both the mother and the fetus/baby. Another reason could be lack of *control*. For example, randomizing US cities to a minimum wage of (say) \$15 per hour versus no minimum wage is difficult; [63] similarly, setting the temperature in large geographical regions to better understand the effects of climate change is not an option.
- *Experimentation may be inadequate* when it comes to generalizing the findings to a broader population and to how medicine is practiced in that population on a day-to-day basis. While observational studies, almost by definition, take stock of how medicine is practiced in real life, clinical trials occur in more controlled settings, often take place in university hospitals or clinics, and may attract particular types of subjects.

[63] An example of large-scale experimentation with policy includes Finland's Design for Government project, and an experiment with the universal income in Canada (Mincome) [35].

Example 11.28 (Effects of smoking)**.** In the study of how smoking affects lung cancer and other ailments, for ethical reasons, researchers have had to rely on observational studies and experiments on animals. These have provided very strong circumstantial evidence that tobacco consumption is associated with the onset and development of various pulmonary and cardiovascular diseases. In the meantime, the tobacco industry has insisted that these are only associations and that no causation has been established [167]. (The story might be repeating itself with the consumption of sugar [158].)

Example 11.29 (Human role in climate change)**.** There is a parallel in the question of climate change, which is another area where experimentation is hardly possible. To determine the role of human activity in climate change, scientists have had to rely on indirect evidence such as the known warming effects of carbon dioxide, methane, and other 'greenhouse' gases, combined with the fact that the increased presence of these gases in the atmosphere is due to human activity. Scientists have also been able to rely computer models. Here too, the sum total evidence is overwhelming, and scientific consensus is essentially unanimous, yet the fossil fuel industry and others still claim all this does not prove that human activity is a substantial contributor to climate change [179].

11.3.2 Types of observational studies

Problem 11.30. In the examples given below, identify as many obstacles to randomization as you can among the ones listed above, and possibly others.

Cohort study A cohort in this context is a group of individuals sharing one or several characteristics of interest. For example, a birth cohort is made of subjects that were born at about the same time.

Example 11.31 (Obesity in children)**.** In context of a large longitudinal study, the Avon Longitudinal Study of Parents and Children, the goal of researchers in [193] was to "identify risk factors in early life (up to 3 years of age) for obesity in children (by age 7) in the United Kingdom".

Another example would be people that smoke, that are then followed to understand the implications of smoking, in which case another cohort of non-smokers may be used to serve as control. In general, such a study follows subjects with a certain condition with the goal of drawing associations with particular outcomes.

Example 11.32 (Head injuries)**.** The study [231] followed almost 3000 individuals that were admitted to one of a number of hospitals in the Glasgow area, Scotland, after a head injury. The stated goal was to "determine

the frequency of disability in young people and adults admitted to hospital with a head injury and to estimate the annual incidence in the community".

A matched-pairs cohort study[64] is a variant where subjects are matched and then followed as a cohort. This leads to analyzing the data using methods for paired data.

Example 11.33 (Helmet use and death in motorcycle crashes)**.** The study [176] is based on the "Fatality Analysis Reporting System data, from 1980 through 1998, for motorcycles that crashed with two riders and either the driver or the passenger, or both, died". Matching was (obviously) by motorcycle. To quote the authors of the study, "by estimating effects among naturally matched-pairs on the same motorcycle, one can account for potential confounding by motorcycle characteristics, crash characteristics, and other factors that may influence the outcome".

Case-control study In a case-control study, subjects with a certain condition (typically a disease) of interest are identified and included in the study to serve as cases. At the same time, subjects not experiencing that condition (without the disease or with the disease but of lower severity) are identified and included in the study to serve as controls.

Example 11.34 (Lipoprotein(a) and coronary heart disease)**.** The study [195] involves a sample of men aged 50 at the start of the study (therefore, a birth cohort) from Gothenburg, Sweden. At baseline, a blood sample was taken from each subject and frozen. After six years, the concentration of lipoprotein(a) was measured in men having suffered a myocardial infarction or died of coronary heart disease. For each of these men — which represent the cases in this study — four men were sampled at random among the remaining ones to serve as controls and their blood concentration of lipoprotein(a) was measured. The goal was to "examine the association between the serum lipoprotein(a) concentration and subsequent coronary heart disease".

Example 11.35 (Venous thromboembolism and hormone replacement therapy)**.** Based on the General Practice Research Database (United Kingdom), in [115], "a cohort of 347,253 women aged 50 to 79 without major risk factors for venous thromboembolism was identified. Cases were 292 women admitted to hospital for a first episode of pulmonary embolism or deep venous thrombosis; 10,000 controls were randomly selected from the source cohort."

[64] Compare with the randomized matched-pairs design.

(The cohort here is a birth cohort and not based on a particular risk factor.) The goal was to "evaluate the association between use of hormone replacement therapy and the risk of idiopathic venous thromboembolism".

Remark 11.36 (Cohort vs case-control)**.** A cohort study starts with a possible risk factor (e.g., smoking) and aims at discovering the diseases it is associated with. A case-control study, on the other hand, starts with the disease and aims at discovering risk factors associated with it.[65]

Problem 11.37 (Rare diseases)**.** A case-control study is often more suitable when studying a rare disease, which would otherwise require following a very large cohort. Consider a disease affecting one out of 100,000 people in a certain large population (many millions). How large would a sample need to be in order to include 10 cases with probability at least 99%?

Cross-sectional study While a cohort study and case-control study both follow a certain sample of subjects over time, a cross-sectional study examines a sample of individuals at a specific point in time. In particular, associations are comparatively harder to interpret in cross-sectional studies. The main advantage is simply cost, as such studies can be based on data collected for other purposes.

Example 11.38 (Green tea and heart disease)**.** In [130], "1,371 men aged over 40 years residing in Yoshimi [were] surveyed on their living habits including daily consumption of green tea. Their peripheral blood samples were subjected to several biochemical assays." The goal was to "investigate the association between consumption of green tea and various serum markers in a Japanese population, with special reference to preventive effects of green tea against cardiovascular disease and disorders of the liver."

11.3.3 Matching

While in an observational study randomization cannot be purposely implemented, a number of techniques have been proposed to at least minimize the influence of other factors [42].

Matching is an umbrella name for a variety of techniques that aim at matching cases with controls in order to have a treatment group and a control group that look alike in important aspects. These important attributes are typically chosen for their potential to affect the response under consideration, that is, they are believed to be risk factors.

[65] For an illustration, compare Figures 3.5 and 3.6 in [28].

Example 11.39 (Effects of coaching on SAT)**.** The study [184] attempted to measure the effect of attending an SAT coaching program on the resulting test score. (The paper focused on *SAT I: Reasoning Test Scores.*) These programs are offered by commercial test preparation companies that claim a certain level of success. The data were based on a survey of about 4,000 SAT takers in 1995-96, out of which about 12% had attended a coaching program. Besides the response (the SAT score itself), some 27 variables were measured on each test taker, including coaching status (the main factor of interest), racial and socioeconomic indicators (e.g., father's education), various measures of academic preparation (e.g., math grade), etc. The idea, of course, is to isolate the effect of coaching from these other factors, some of which are undoubtedly important. The authors applied a number of techniques including a variant of matching. Other variants are applied to the same data in [117].

The intention behind matching is to control for confounding by balancing possible confounding factors with the intention of canceling out their confounding effect. We formalize this in Section 22.2.3, where we show that matching works under some conditions, the most important one being that there are no unmeasured variables that confound the effect. The beauty (and usefulness) of randomization is that it achieves this balancing automatically (although only on average) without a priori knowledge of any confounding variable. This is shown formally in Section 22.2.2.

11.3.4 Natural experiments

As we mentioned above, and as will be detailed in Section 22.2, a proper use of randomization allows for causal inference. However, randomization is only possible in controlled experiments. In observational studies, matching can allow for causal inference, but only if one can guaranty that there are no confounders.

In general, the issue of drawing causal inferences from observational studies is complex, and in fact remains controversial, so we will keep the discussion simple. Essentially, in the context of an observational study, causal inference is possible if one can successfully argue that the treatment assignment (which was done by Nature) was done independently of the response to treatment. Doing this successfully often requires domain-specific expertise. Freedman speaks of a shoe leather approach to data analysis [94]. In these circumstances, the observational study is often qualified as being a *natural experiment* [47, 148].

Example 11.40 (Snow's discovery of cholera)**.** A clas-

sical example of a natural experiment is that of John Snow in mid-1800 London, who in the process discovered that cholera could be transmitted by the consumption of contaminated water [217]. The account is worth reading in more detail, but in a nutshell, Snow suspected that the consumption of water had something to do with the spread of cholera, and to confirm his hypothesis, he examined the houses in London served by one of two water companies, Southwark & Vauxhall and Lambeth, and found that the death rate in the houses served by the former was several times higher. He then explained this by the fact that, although both companies sourced their water from the Thames River, Lambeth was taping the river upstream of the main sewage discharge points, while Southwark & Vauxhall was getting its water downstream. Although this is an observational study, a case for 'natural' randomization can be argued, as Snow did, on the basis that the companies served the same neighborhoods. In his own words: "Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies."

Example 11.41 (The Oregon Experiment on Medicaid)**.** In 2008, the state of Oregon in the US decided to expand a joint federal and state health insurance program for people with low-income known as Medicaid. As described in [30]: "Correctly anticipating excess demand for the available new enrollment slots, the state conducted a lottery, randomly selecting individuals from a list of those who signed up in early 2008. Lottery winners and members of their households were able to apply for Medicaid. Applicants who met the eligibility requirements were then enrolled in [the] Oregon Health Plan Standard." In the end, 29,834 individuals won the lottery out of 74,922 individuals who participated. Both groups have nevertheless been followed and compared on various metrics (e.g., health care utilization). Note that this is an ongoing study.

Natural experiments are relatively rare, but some situations have been identified where they arise routinely. For example, in Genetics, pairs of twins[66] are sought after and examined to better understand how much a particular behavior is due to "nature" versus "nurture".

Regression discontinuity design This is an area of statistics specializing in situations where an intervention is applied or not based on some sort of score being above or below some threshold. If the threshold is arbitrary to some degree, it may justifiable to compare, in terms of

[66] There is an entire journal dedicated to the study of twin pairs: *Twin Research and Human Genetics*.

outcome, cases with a score right above the threshold with cases with a score right below the threshold.

Example 11.42 (Medicare)**.** In the US, most people become eligible at age 65 to enroll in a federal health insurance program called Medicare. This has lead researchers, as in [37], to examine the effects of access to this program on various health outcomes.

Example 11.43 (Elite secondary schools in Kenya)**.** Students from elite schools tend to perform better, but is this due to elite schools being truly superior to other schools, or simply a result of attracting the best and brightest students? This question is examined in [156] in the context of secondary schools in Kenya, employing a regression discontinuity design approach that takes advantage of "the random variation generated by the centralized school admissions process".

Remark 11.44. In a highly cited paper [125], Bradford Hill proposes nine aspects to pay attention to when attempting to draw causal inferences from observational studies. One of the main aspects is *specificity*, which is akin to identifying a natural experiment. Another important aspect, when present, is that of *experimental* evidence, which is akin to identifying a discontinuity.

Part C

Elements of Statistical Inference

CHAPTER 12

MODELS, ESTIMATORS, AND TESTS

A prototypical (although somewhat idealized) workflow in any scientific investigation starts with the design of the experiment to probe a question or hypothesis of interest. The experiment is modeled using several plausible mechanisms. The experiment is conducted and the data are collected. These data are finally analyzed to identify the most adequate mechanism, meaning the one among those considered that best explains the data.

Although an experiment is supposed to be repeatable, this is not always possible, particularly if the system under study is chaotic or random in nature. When this is the case, the mechanisms above are expressed as probability distributions. We then talk about *probabilistic modeling*, albeit here there are several probability distributions under consideration. It is as if we contemplate several probability experiments (in the sense of Chapter 1), and the goal of *statistical inference* is to decide on the most plausible in view of the collected data.

To illustrate the various concepts that introduced in this chapter, we use Bernoulli trials as our running example of an experiment, which includes as special case the model where we sample with replacement from an urn (Remark 2.17). Although basic, this model is relevant in a number of important real-life situations (e.g., election polls, clinical trials, etc). We will study Bernoulli trials in more detail in Section 14.1.

12.1 Statistical models

A *statistical model* for a given experiment is of the form $(\Omega, \Sigma, \mathcal{P})$ where Ω is the sample space containing all possible outcomes, Σ is the class of events of interest, and $\mathcal{P}$ is a *family* of distributions on Σ.[67] Modeling the experiment with $(\Omega, \Sigma, \mathcal{P})$ postulates that the outcome of the experiment $\omega \in \Omega$ was generated from a distribution $\mathbb{P} \in \mathcal{P}$. The goal, then, is to determine which $\mathbb{P}$ best explains the data ω.

We follow the tradition of parameterizing the family $\mathcal{P}$. This is natural in some contexts and can always be done without loss of generality (since any set can be parameterized by itself). By default, the *parameter* will be denoted by θ and the *parameter space* (where θ belongs) by Θ, so that

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}. \tag{12.1}$$

Remark 12.1 (Identifiability)**.** Unless otherwise specified, we will assume that the model (12.1) is *identifiable*, meaning that the parameterization is one-to-one, or in formula, $\mathbb{P}_\theta \neq \mathbb{P}_{\theta'}$ when $\theta \neq \theta'$.

Example 12.2 (Binomial experiment)**.** Suppose we model an experiment as a sequence of Bernoulli trials with probability parameter θ and predetermined length n. This assumes that each trial results in one of two possible values. Labeling these values as H and T, which we can always do at least formally, the sample space is the set of all sequences of length n with values in $\{\mathrm{H}, \mathrm{T}\}$, or in formula

$$\Omega = \{\mathrm{H}, \mathrm{T}\}^n.$$

(We already saw this in Example 1.8.) The statistical model also assumes that the observed sequence was generated by one of the Bernoulli distributions, namely

$$\mathbb{P}_\theta(\{\omega\}) = \theta^{Y(\omega)}(1-\theta)^{n-Y(\omega)}, \tag{12.2}$$

where $Y(\omega)$ is the number of heads in ω. (We already saw that in (2.13).) Therefore, the family of distributions

[67] Compare with a probability space, which only includes one distribution. A statistical model includes several distributions to model situations where the mechanism driving the experimental results is not perfectly known.

under consideration is

$$\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}, \quad \text{where } \Theta := [0,1].$$

Note that the dependency in n is left implicit as n is given and not a parameter of the family. We call this a *binomial experiment* because of the central role that Y plays and we know that Y has the binomial distribution with parameters (n, θ).

Remark 12.3 (Correctness of the model)**.** As is the custom, we will proceed assuming that the model is correct, that indeed one of the distributions in the family $\mathcal{P}$ generated the data. This depends, in large part, on how the data were generated or collected (Chapter 11). For example, a (binary) poll that successfully implements simple random sampling from a very large population can be accurately modeled as a binomial experiment. When the model is correct, we will sometimes use θ_* to denote the true value of the parameter. In practice, a model is rarely strictly correct. When the model is only approximate, the resulting inference will necessarily be approximate also.

12.2 Statistics and estimators

A *statistic* is a random variable on the sample space Ω. It is meant to summarize the data in a way that is useful for the purpose of drawing inferences from the data.

Let φ be a function defined on the parameter space Θ representing a feature of interest (e.g., the mean). Note that φ is often used to denote $\varphi(\theta)$ (a clear abuse of notation) when confusion is unlikely. We will adopt this common practice. It is often the case that θ itself is the feature of interest, in which case $\varphi(\theta) = \theta$.

An *estimator* for $\varphi(\theta)$ is a statistic, say S, chosen for the purpose of approximating it. Note that while φ is defined on the parameter space (Θ), S is defined on the sample space (Ω).

Remark 12.4. The problem of estimating $\varphi(\theta)$ is well-defined if $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ implies that $\varphi(\theta) = \varphi(\theta')$, which is always the case if the model is identifiable.

Remark 12.5 (Estimators and estimates)**.** An estimator is thus a statistic. The value that an estimator takes in a given experiment is often called an *estimate*. For example, if S is an estimator, and ω denotes the data, then $S(\omega)$ is an estimate.

Desiderata. A good estimator is one which returns an estimate that is 'close' to the true value of the quantity of interest.

12.2.1 Measures of performance

Quantifying closeness is not completely trivial as we are talking about estimators, whose output is random by definition. (An estimator is a function of the data and the data are assumed to be random.) We detail the situation in the context of a parametric model (12.1) where $\Theta \subset \mathbb{R}$ and consider estimating θ itself.

Mean squared error A popular measure of performance for an estimator S is the *mean squared error (MSE)*, defined as

$$\mathsf{mse}_\theta(S) := \mathbb{E}_\theta\big[(S-\theta)^2\big], \tag{12.3}$$

where $\mathbb{E}_\theta$ denotes the expectation with respect to $\mathbb{P}_\theta$.

Problem 12.6 (Squared bias + variance)**.** Assume that an estimator S for θ has a 2nd moment. Show that

$$\mathsf{mse}_\theta(S) = \underbrace{\big(\mathbb{E}_\theta(S)-\theta\big)^2}_{\text{squared bias}} + \underbrace{\mathrm{Var}_\theta(S)}_{\text{variance}}. \tag{12.4}$$

(Var_θ denotes the variance under $\mathbb{P}_\theta$.)

Mean absolute error Another popular measure of performance for an estimator S is the *mean absolute error (MAE)*, defined as

$$\mathsf{mae}_\theta(S) := \mathbb{E}_\theta\big[|S-\theta|\big]. \tag{12.5}$$

Other loss functions In general, let $\mathcal{L}(\theta',\theta)$ be a function measuring the discrepancy between θ' and θ. This is called a *loss function* as it is meant to quantify the loss incurred when the true parameter is θ and our estimate is θ'. A popular choice is

$$\mathcal{L}(\theta',\theta) = |\theta'-\theta|^\gamma,$$

for some pre-specified $\gamma > 0$. This includes the MSE and the MAE as special cases.

Having chosen a loss function we define the *risk* of an estimator as its expected loss, which for a loss function $\mathcal{L}$ and an estimator S may be expressed as

$$\mathcal{R}_\theta(S) := \mathbb{E}_\theta\big[\mathcal{L}(S,\theta)\big]. \tag{12.6}$$

Remark 12.7 (Frequentist interpretation)**.** Let θ denote the true value of the parameter and let S denote an estimator. Suppose the experiment is repeated independently m times and let ω_j denote the outcome of the jth experiment. Compute the average loss over these m experiments, meaning

$$L_m := \frac{1}{m}\sum_{j=1}^{m} \mathcal{L}(S(\omega_j),\theta).$$

By the Law of Large Numbers (which, as we saw, is at the core of the frequentist interpretation of probability),

$$L_m \xrightarrow{\mathrm{P}} \mathcal{R}_\theta(S), \quad \text{as } m \to \infty.$$

12.2.2 Maximum likelihood estimator

As the name indicates, this estimator returns a distribution that maximizes, among those in the family, the chances that the experiment would result in the observed data. Assume that the statistical model is discrete in the sense that the sample space is discrete. See Section 12.5.1 for other situations.

Denoting the data by $\omega \in \Omega$ as before, the *likelihood function* is defined as

$$\mathsf{lik}(\theta) = \mathbb{P}_\theta(\{\omega\}). \tag{12.7}$$

(Note that this function also depends on the data, but this dependency is traditionally left implicit.) Assuming that, for all possible outcomes, this function has a unique maximizer, the *maximum likelihood estimator (MLE)* is defined as that maximizer.

Remark 12.8. When the likelihood admits several maximizers, one of them can be chosen according to some criterion.

Example 12.9 (Binomial experiment)**.** In the setting of Example 12.2, the MLE is found by maximizing the likelihood (12.2) with respect to $\theta \in [0,1]$. To simplify the expression a little bit, let $y = Y(\omega)$, which is the number of heads in the sequence ω. We then have the following expression for the likelihood

$$\mathsf{lik}(\theta) := \theta^y (1-\theta)^{n-y}.$$

This is a polynomial in θ and therefore differentiable. To maximize the likelihood we thus look for critical points. First assume that $1 \le y \le n-1$. In that case, setting the derivative to 0, we obtain the equation

$$\theta^y (1-\theta)^{n-y} \big(y(1-\theta) - (n-y)\theta \big) = 0.$$

The solutions are $\theta = 0$, $\theta = 1$, and $\theta = y/n$. Since the likelihood is zero at $\theta = 0$ or $\theta = 1$, and strictly positive at $\theta = y/n$, we conclude that the maximizer is unique and equal to y/n. If $y = 0$, the likelihood is easily seen to have a unique maximizer at $\theta = 0$. If $y = 1$, the likelihood is easily seen to have a unique maximizer at $\theta = 1$. Thus, in any case, the maximizer is unique and given by y/n. We conclude that the MLE is well-defined and equal to Y/n, that is, the proportion of heads in the data.

12.3 Confidence intervals

An estimator, as presented above, provides a way to obtain an informed guess (the estimate) for the true value of the parameter. In addition to that, it is often quite useful to

know how far we can expect the estimate to be from the true value of the parameter.

When the parameter is real-valued (as we assume it to be by default), this is typically done via an interval whose bounds are random variables. We say that such an interval, denoted $I(\omega)$, is a $(1-\alpha)$-level *confidence interval* for θ if

$$\mathbb{P}_\theta(\theta \in I) \geq 1-\alpha, \quad \text{for all } \theta \in \Theta. \tag{12.8}$$

For example, $\alpha = 0.10$ gives a 90% confidence interval. See Figure 12.1 for an illustration.

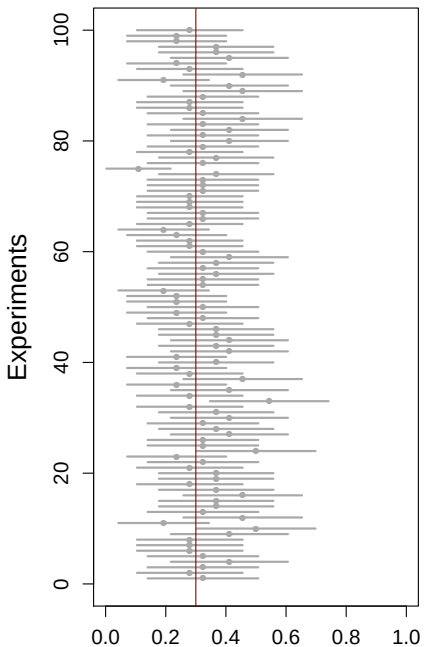
Desiderata. A good confidence interval is one that has the prescribed level of confidence and is relatively short (compared to other confidence intervals having the same level of confidence).

12.3.1 Using Chebyshev's inequality to construct a confidence interval

Confidence intervals are often constructed based on an estimator, here denoted S. Suppose first that the estimator is *unbiased*, meaning

$$\mathbb{E}_\theta(S) = \theta, \quad \text{for all } \theta \in \Theta. \tag{12.9}$$

Figure 12.1: An illustration of the concept of confidence interval. We consider a binomial experiment with parameters $n = 10$ and $\theta = 0.3$. We repeat the experiment 100 times, each time computing the Clopper–Pearson two-sided 90% confidence interval for θ specified in (14.6). The vertical line is at the true value of θ.



Assume furthermore that S as a finite 2nd moment under any $\theta \in \Theta$, with uniformly bounded variance in the sense that

$$\sigma_\theta^2 := \mathrm{Var}_\theta(S) \le \sigma^2, \tag{12.10}$$

for some positive constant σ. Importantly, we assume that such a constant is available (meaning known to the analyst).

Chebyshev's inequality (7.27) can then be applied to construct a confidence interval. Indeed, the inequality gives

$$\mathbb{P}_\theta\big(|S-\theta| < c\sigma_\theta\big) \ge 1 - 1/c^2, \quad \text{for all } \theta \in \Theta,$$

for any $c > 0$. We then have

$$|S-\theta| < c\sigma_\theta \;\Rightarrow\; |S-\theta| < c\sigma \tag{12.11}$$

$$\Leftrightarrow\; S - \sigma c < \theta < S + \sigma c. \tag{12.12}$$

Thus, if we define $I_c := (S - \sigma c, S + \sigma c)$, we have that

$$\mathbb{P}_\theta(\theta \in I_c) \ge 1 - 1/c^2.$$

Given α, we choose c such that $1/c^2 = \alpha$, that is, $c = 1/\sqrt{\alpha}$. Then the resulting interval I_c is a $(1-\alpha)$-confidence interval for θ.

Problem 12.10 (Binomial experiment)**.** In the setting of Example 12.2, apply the procedure above to derive a $(1-\alpha)$-confidence interval for θ. (The upper bound on the standard deviation, σ above, should be explicit.)

We note that a refinement is possible when σ_θ is known in closed-form. See Problem 14.4.

12.4 Testing statistical hypotheses

Suppose we do not need to estimate a function of the parameter, but rather only need to know whether the parameter value satisfies a given property. In what follows, we take the default hypothesis, called the *null hypothesis* and often denoted $\mathcal{H}_0$, to be that the parameter value satisfies the property. Deciding whether the data is congruent with this hypothesis is called *testing* the hypothesis.

Let $\Theta_0 \subset \Theta$ denote the subset of parameter values that satisfy the property. We will call Θ_0 the *null set*. The null hypothesis can be expressed as

$$\mathcal{H}_0 : \theta_* \in \Theta_0.$$

(Recall that θ_* denotes the true value of the parameter, assuming the statistical model is correct.) The complement of Θ_0 in Θ,

$$\Theta_1 := \Theta \setminus \Theta_0, \tag{12.13}$$

is often called the *alternative set*, and

$$\mathcal{H}_1 : \theta_* \in \Theta_1$$

is often called the *alternative hypothesis*.

Example (Binomial experiment)**.** In the setting of Example 12.2, we may want to know whether the parameter value is below some given θ_0. This corresponds to testing the null hypothesis

$$\mathcal{H}_0 : \theta_* \leq \theta_0, \tag{12.14}$$

which corresponds to the following null set

$$\Theta_0 = \{\theta \in [0,1] : \theta \leq \theta_0\} = [0, \theta_0].$$

12.4.1 Test statistics

A *test statistic* is used to decide whether the null hypothesis is reasonable. If, based on a chosen test statistic, the hypothesis is found to be 'substantially incompatible' with the data, it is *rejected*.

Although the goal may not be that of estimation, good estimators typically make good test statistics. Indeed, given an estimator S, one could think of rejecting $\mathcal{H}_0$ when $S(\omega) \notin \Theta_0$. Tempting as it is, this is typically too harsh, as it does not properly account for the randomness in the estimator. Instead, it is generally better to reject $\mathcal{H}_0$ if $S(\omega)$ is 'far enough' from Θ_0.

Desiderata. A good test statistic is one that behaves differently according to whether the null hypothesis is true or not.

Example (Binomial experiment)**.** Consider the problem of testing the hypothesis $\mathcal{H}_0$ of (12.14). As a test statistic, let us use the maximum likelihood estimator, $S := Y/n$. In that case, it is tempting to reject $\mathcal{H}_0$ when $S > \theta_0$. However, doing so would lead us to reject by mistake quite often if θ_* is in the null set yet close to the alternative set: in the most extreme case where $\theta_* = \theta_0$, the probability of rejection approaches $1/2$ as the sample size n increases.

Problem 12.11. Prove the last claim using the Central Limit Theorem. Then examine the situation where $\theta_* < \theta_0$.

Remark 12.12 (Equivalent test statistics)**.** We say that two test statistics, S and T, are equivalent if there is a strictly monotone function g such that $T = g(S)$. Clearly, equivalent statistics provide the same amount of evidence against the null hypothesis, since we can recover one from the other.

12.4.2 Likelihood ratio

The *likelihood ratio (LR)* is to hypothesis testing what the maximum likelihood estimator is to parameter estimation. It presents a general procedure for deriving a test statistic. In the setting of Section 12.2.2, having observed ω, the likelihood ratio is defined as [68]

$$\frac{\max_{\theta\in\Theta_1} \mathsf{lik}(\theta)}{\max_{\theta\in\Theta_0} \mathsf{lik}(\theta)}. \tag{12.15}$$

By construction, a large value of that statistic provides evidence against the null hypothesis.

Remark 12.13 (Variants)**.** The LR is sometimes defined differently, for example,

$$\frac{\max_{\theta\in\Theta} \mathsf{lik}(\theta)}{\max_{\theta\in\Theta_0} \mathsf{lik}(\theta)}. \tag{12.16}$$

or its inverse (in which case small values of the statistic weigh against the null hypothesis). However, all these variants are strictly monotonic functions of each other and are therefore equivalent for testing purposes (Remark 12.12).

Example 12.14 (Binomial experiment)**.** Consider the problem of testing the hypothesis $\mathcal{H}_0$ of (12.14). Recall the MLE is $S := Y/n$. With $y = Y(\omega)$ and $\hat\theta := y/n$ (which is the estimate), we compute

$$\begin{aligned}\max_{\theta\le\theta_0} \mathsf{lik}(\theta) &= \max_{\theta\le\theta_0} \theta^y(1-\theta)^{n-y} \\ &= \min(\hat\theta,\theta_0)^y(1-\min(\hat\theta,\theta_0))^{n-y},\end{aligned}$$

and

$$\begin{aligned}\max_{\theta\in[0,1]} \mathsf{lik}(\theta) &= \max_{\theta\in[0,1]} \theta^y(1-\theta)^{n-y} \\ &= \hat\theta^y(1-\hat\theta)^{n-y}.\end{aligned}$$

Hence, the variant (12.16) of the LR is equal to 1 (the minimum possible value) if $\hat\theta \le \theta_0$, and

$$\left(\frac{\hat\theta}{\theta_0}\right)^y\left(\frac{1-\hat\theta}{1-\theta_0}\right)^{n-y},$$

otherwise. Taking the logarithm and multiplying by $1/n$ yields an equivalent statistic equal to 0 if $\hat\theta \le \theta_0$, and

$$\hat\theta\log\left(\frac{\hat\theta}{\theta_0}\right) + (1-\hat\theta)\log\left(\frac{1-\hat\theta}{1-\theta_0}\right),$$

otherwise. Thus the LR is a function of the MLE. However, the monotonicity is not strict, and therefore the MLE and the LR are not equivalent test statistics. That said, they yield the same inference in most cases of interest (Problem 12.19).

[68] This statistic is sometimes referred to as the generalized likelihood ratio.

12.4.3 p-values

Given a test statistic, we need to decide what values of the statistic provide evidence against the null hypothesis. In other words, we need to decide what values of the statistic are 'unusual' or 'extreme' under the null, in the sense of being unlikely if the null (hypothesis) were true.

Suppose we decide that large values of a test statistic S are evidence that the null is not true. Let ω denote the observed data and let $s = S(\omega)$ denote the observed value of the statistic. In this context, we define the *p-value* as

$$\mathsf{pv}_S(s) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(S \geq s). \tag{12.17}$$

To be sure,

$$\mathbb{P}_\theta(S \geq s) \text{ is shorthand for } \mathbb{P}_\theta\big(\{\omega' : S(\omega') \geq S(\omega)\}\big).$$

In words, this is the supremum probability under any null distribution of observing a value of the (chosen) test statistic as extreme as the one that was observed. A small p-value is evidence that the null hypothesis is false.

Note that a p-value is associated with a particular test statistic: different test statistics lead to different p-values, in general.

Remark 12.15 (Replication interpretation of the p–value)**.** The definition itself lends us to believe that replications are needed to compute the p-value. Such replications may be out of the question, however. In many cases, the experiment has been performed and the data have been collected, and inference needs to be performed based on these data alone. This is where assuming a model is crucial. Indeed, if we are able to derive (or approximate) the distribution of the test statistic under any null distribution, then we can compute (or approximate) the p-value, at least in principle, without having to repeat the experiment.

Proposition 12.16 (Semi-continuity of the p-value)**.** *The function defined in* (12.17) *is lower semi-continuous, meaning*

$$\liminf_{s \to s_0} \mathsf{pv}_S(s) \geq \mathsf{pv}_S(s_0), \quad \textit{for all } s_0. \tag{12.18}$$

Proof sketch. For any random variable S on any probability space, $s \mapsto \mathbb{P}(S \geq s)$ is lower semi-continuous. This comes from the fact that $\mathbb{P}(S \leq s)$ is upper semi-continuous (Problem 4.5). We then use the fact that the supremum of any collection of lower semi-continuous functions is lower semi-continuous. □

Proposition 12.17 (Monotone transformation)**.** *Consider a test statistic S and that large values of S provide*

evidence against the null hypothesis. Let g be strictly increasing (resp. decreasing). Then large (resp. small) values of $g(S)$ provide evidence against the null hypothesis and the resulting p-value is equal to that based on S.

Proof. Suppose, for instance, that g is strictly increasing and let $T = g(S)$. Suppose that the experiment resulted in ω. Then S is observed to be $s := S(\omega)$ and T is observed to be $t := T(\omega)$. Noting that $t = g(s)$, for any possible outcome ω', we have

$$T(\omega') \geq t \iff g(S(\omega')) \geq g(s) \iff S(\omega') \geq s.$$

From this, $\mathsf{pv}_T(t) = \mathsf{pv}_S(s)$ follows, which is what we needed to prove. □

Problem 12.18. Show that the different variants of likelihood ratio test statistic, (12.15) and (12.16), lead to the same p-value.

Example (Binomial experiment)**.** Consider the problem of testing the hypothesis $\mathcal{H}_0$ of (12.14). As a test statistic, let us use the maximum likelihood estimator, $S := Y/n$. Here large values of S weigh against the null hypothesis. If the data are ω, $s := S(\omega)$ is the observed value of the test statistic, and the resulting p-value is given by

$$\sup_{\theta \leq \theta_0} \mathbb{P}_\theta(S \geq s).$$

By a monotonicity argument (Problem 3.26), it turns out that the supremum is always achieved at $\theta = \theta_0$, right at the boundary of the null and alternative sets. In terms of the number of heads Y, the p-value is thus equal to

$$\mathbb{P}_{\theta_0}(Y \geq y),$$

where $y := Y(\omega)$. Under $\theta = \theta_0$, Y has distribution $\mathrm{Bin}(n, \theta_0)$, and therefore this p-value can be computed by reference to that distribution.

Problem 12.19 (Near equivalence of the MLE and LR)**.** We saw in Example 12.14 that the MLE and the LR were not, strictly speaking, equivalent. Although they may yield different p-values, show that these coincide when they are smaller than $\mathbb{P}_{\theta_0}(S \geq \theta_0)$ (which is close to 0.5).

12.4.4 Tests

A test statistic yields a p-value that is used to quantify the amount of evidence in the data against the postulated null hypothesis. Sometimes, though, the end goal is actual decision: *reject or not reject* is the question. Tests formalize such decisions.

Formally, a *test* is a statistic with values in $\{0, 1\}$, returning 1 if the null hypothesis is to be rejected and 0 otherwise. If large values of a test statistic S provide

evidence against the null, then a test based on S will be of the form

$$\phi(\omega) = \{S(\omega) \geq c\}. \tag{12.19}$$

The subset $\{S \geq c\} \equiv \{\omega : S(\omega) \geq c\}$ is called the *rejection region* of the test. The threshold c is typically referred to as the *critical value*.

Test errors When applying a test to data, two types of error are possible:

- A *Type I error* or *false positive* happens when the test rejects even though the null hypothesis is true.
- A *Type II error* or *false negative* happens when the test does not reject even though the null hypothesis is false.

Table 12.1 illustrates the situation.

Table 12.1: Types of error that a test can make.

	null is true	null is false
rejection	type I	correct
no rejection	correct	type II

For a test that rejects for large values of a statistic, the choice of critical value drives the probabilities of Type I and Type II errors. Qualitatively speaking, increasing the critical value makes the test reject less often, which decreases the probability of Type I error and increases the probability of Type II error. Of course, decreasing the critical value has the reverse effect.

12.4.5 Level

The *size* of a test ϕ is the maximum probability of Type I error,

$$\text{size}(\phi) := \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\phi = 1). \tag{12.20}$$

Let $\alpha \in [0, 1]$ denote the desired control on the probability of Type I error, called the *significance level*. A test ϕ is said to have level α if its size is bounded by α,

$$\text{size}(\phi) \leq \alpha.$$

Given a test statistic S whose large values weigh against the null hypothesis, the corresponding test has level α if the critical value c satisfies

$$\mathsf{pv}_S(c) \leq \alpha.$$

In order to minimize the probability of Type II error, we want to choose the smallest c that satisfies this requirement. Let c_α denote that value, or in formula

$$c_\alpha := \min\left\{c : \mathsf{pv}_S(c) \leq \alpha\right\}. \tag{12.21}$$

(The minimum is attained because of Proposition 12.16.) The resulting test has rejection region $\{S \geq c_\alpha\}$.

Remark 12.20. The rejection region can also be expressed directly in terms of the p-value as $\{\mathsf{pv}_S(S) \leq \alpha\}$, so that the test rejects at level α if the associated p-value is less than or equal to α.

The following shows that this test controls the probability of Type I error at the desired level α.

Proposition 12.21. *For any* $\alpha \in [0,1]$,

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta\big(\mathsf{pv}_S(S) \leq \alpha\big) \leq \alpha. \tag{12.22}$$

Proof. Let pv be shorthand for pv_S. As in Problem 4.14, define $\tilde{\mathsf{F}}_\theta(s) = \mathbb{P}_\theta(S \geq s)$, so that

$$\mathsf{pv}(s) = \sup_{\theta \in \Theta_0} \tilde{\mathsf{F}}_\theta(s).$$

In particular, for any $\theta \in \Theta_0$,

$$\mathsf{pv}(s) \leq \alpha \;\Rightarrow\; \tilde{\mathsf{F}}_\theta(s) \leq \alpha. \tag{12.23}$$

Fix such a θ. Let F_θ^- denote the quantile function of S under θ as defined in Section 4.6. Then

$$\begin{aligned}
\mathbb{P}_\theta\big(\mathsf{pv}(S) \leq \alpha\big) &\leq \mathbb{P}_\theta\big(\tilde{\mathsf{F}}_\theta(S) \leq \alpha\big) && (12.24)\\
&= \mathbb{P}_\theta\big(S \geq \mathsf{F}_\theta^-(1-\alpha)\big) && (12.25)\\
&= \tilde{\mathsf{F}}_\theta(\mathsf{F}_\theta^-(1-\alpha)) && (12.26)\\
&\leq \alpha. && (12.27)
\end{aligned}$$

In the 1st line we used (12.23), in the 2nd we used (4.20), in the 3rd we used the definition of $\tilde{\mathsf{F}}_\theta$, and in the 4th we used (4.20) again. Since $\theta \in \Theta_0$ is arbitrary, the proof is complete. □

Example (Binomial experiment)**.** Continuing with the same setting, and recalling that we use as test statistic the MLE, $S = Y/n$, and that we reject for large values of that statistic, the critical value for the level α is given by

$$\begin{aligned}
c_\alpha &= \min\big\{c : \sup_{\theta \leq \theta_0} \mathbb{P}_\theta(S \geq c) \leq \alpha\big\} && (12.28)\\
&= \min\big\{c : \mathbb{P}_{\theta_0}(S \geq c) \leq \alpha\big\}, && (12.29)
\end{aligned}$$

where we used (3.12) in the second line. Equivalently, $c_\alpha = b_\alpha/n$ where b_α is the $(1-\alpha)$-quantile of $\mathrm{Bin}(n, \theta_0)$.

Controlling the level is equivalent to controlling the number of false alarms, which is crucial in applications.[69]

[69] The way an alarm system is designed plays an important role too, as briefly discussed here.

Example 12.22 (Security at Y-12)**.** We learn in [203] that, in 2012, three activists (including an eighty two year old nun) broke into the Y-12 National Security Complex in Oak Ridge, Tennessee. "Y-12 is the only industrial complex in the United States devoted to the fabrication and storage of weapons-grade uranium. Every nuclear warhead and bomb in the American arsenal contains uranium from Y-12." This is a highly guarded complex and the activists did set up an alarm, but there were several hundred false alarms per month. ([257] claims there were upward of 2,000 alarms per day.) This reminds one of the allegory of the boy who cried wolf...

12.4.6 Power

The *power* of a test ϕ at $\theta \in \Theta$ is

$$\mathsf{pwr}_\phi(\theta) := \mathbb{P}_\theta(\phi = 1).$$

If the test is of the form $\phi = \{S \geq c\}$, then

$$\mathsf{pwr}_\phi(\theta) = \mathbb{P}_\theta(S \geq c).$$

Remark 12.23. The test ϕ has level α if and only if

$$\mathsf{pwr}_\phi(\theta) \leq \alpha, \quad \text{for all } \theta \in \Theta_0.$$

Problem 12.24 (Binomial experiment)**.** Consider the problem of testing the hypothesis $\mathcal{H}_0$ of (12.14) and continue to use the test derived from the MLE. Set the level at $\alpha = 0.01$ and, in R, plot the power as a function of $\theta \in [0,1]$. Do this for $n \in \{10, 20, 50, 100, 200, 500, 1000\}$. Repeat, now setting the level at $\alpha = 0.10$ instead.

Desiderata. A good test has large power against alternatives of interest when compared to other tests at the same significance level.

From confidence intervals to tests and back There is an equivalence between confidence intervals and tests of hypotheses.

12.4.7 A confidence interval gives a test

Suppose that I is a $(1-\alpha)$-confidence interval for θ. Define $\phi = \{\Theta_0 \cap I = \varnothing\}$, which is clearly a test. Moreover, it has level α. To see this, take $\theta \in \Theta_0$ and derive

$$\phi = 1 \iff \Theta_0 \cap I = \varnothing \implies \theta \notin I,$$

so that

$$\mathbb{P}_\theta(\phi = 1) \leq \mathbb{P}_\theta(\theta \notin I) \leq \alpha,$$

where the last inequality is due to the fact that I is a $(1-\alpha)$-confidence interval for θ.

12.4.8 A family of confidence intervals gives a p-value

Assume a family of confidence intervals for θ, denoted $\{I_\gamma : \gamma \in [0,1]\}$, such that I_γ has confidence level γ and $I_\gamma \subset I_{\gamma'}$ when $\gamma \le \gamma'$. Then define

$$P := \sup\big\{\alpha : \Theta_0 \cap I_{1-\alpha} \neq \varnothing\big\}.$$

This is the random variable (with values in $[0,1]$) when seen the intervals as random.

Proposition 12.25. *This quantity is a valid p-value in the sense that it satisfies* (12.22)*, meaning*

$$\sup_{\theta\in\Theta_0} \mathbb{P}_\theta(P \le \alpha) \le \alpha, \quad \textit{for all } \alpha \in [0,1].$$

Proof. Take any $\theta \in \Theta_0$ and any $\alpha \in [0,1]$. We need to prove that

$$\mathbb{P}_\theta(P \le \alpha) \le \alpha. \tag{12.30}$$

By definition of P, we have $\Theta_0 \cap I_{1-u} = \varnothing$ for any $u > P$. In particular, for any $u > \alpha$,

$$\begin{aligned}\mathbb{P}_\theta(P \le \alpha) &\le \mathbb{P}_\theta(\Theta_0 \cap I_{1-u} = \varnothing)\\ &\le \mathbb{P}_\theta(\theta \notin I_{1-u})\\ &\le 1 - (1-u) = u.\end{aligned}$$

This being true for all $u > \alpha$, we obtain (12.30). □

12.4.9 A family of tests gives a confidence interval

For each $\theta \in \Theta$, suppose we have available a level α test denoted ϕ_θ, for the null hypothesis that $\theta_* = \theta$. Thus ϕ_θ is testing the null hypothesis that the true value of the parameter is θ. Define

$$I(\omega) = \{\theta : \phi_\theta(\omega) = 0\}, \tag{12.31}$$

which is the set of θ whose associated test does not reject. This is an interval in many classical situations where $\Theta \subset \mathbb{R}$. We assume this is the case, although what follows does not rely on that assumption. Then I is level $(1-\alpha)$-confidence interval for θ, meaning it satisfies (12.8). To see that, take any $\theta \in \Theta$. By definition of I and the fact that each ϕ_θ has level α, we get

$$\mathbb{P}_\theta(\theta \notin I) = \mathbb{P}_\theta(\phi_\theta = 1) \le \alpha.$$

Example. Consider the setting of Example 12.2. Suppose we use the MLE (still denoted $S = Y/n$) as test statistic. In line with how we tested (12.14), let ϕ_θ denote the level α test based on S for $\theta_* \le \theta$. The test ϕ_θ can be used for testing $\theta_* = \theta$ since this implies $\theta_* \le \theta$. Let F_θ and F_θ^- denote the distribution and quantile functions of $\mathrm{Bin}(n,\theta)$. Following the same steps that lead to (12.28),

and working directly with the number of heads Y, we derive $\phi_\theta = \{Y \geq c_{\alpha,\theta}\}$, where $c_{\alpha,\theta} := \mathsf{F}_\theta^-(1-\alpha)$. Then

$$\phi_\theta = 0 \;\Leftrightarrow\; Y < c_{\alpha,\theta} \;\Leftrightarrow\; \mathsf{F}_\theta(Y) < 1-\alpha, \tag{12.32}$$

where the 2nd equivalence is due to (4.17). Hence,

$$I = \{\theta : \mathsf{F}_\theta(Y) < 1-\alpha\}.$$

Define

$$\boldsymbol{S} := \inf\{\theta : \mathsf{F}_\theta(Y) < 1-\alpha\},$$

with the convention that $\boldsymbol{S} = 1$ if this set is empty (which only happens when $Y = n$). Then, by (3.12),

$$I = (\boldsymbol{S}, 1]. \tag{12.33}$$

Problem 12.26 (Binomial one-sided confidence interval)**.** Write an R function that computes this interval. A simple grid search works: at each θ in a grid of values, one checks whether

$$\mathsf{F}_\theta(y) < 1-\alpha. \tag{12.34}$$

However, taking advantage of the monotonicity (3.12), a bisection search is applicable, which is much more efficient.

12.5 Further topics

12.5.1 Likelihood methods when the model is continuous

In our exposition of likelihood methods — maximum likelihood estimation and likelihood ratio testing — we have assumed that the statistical model was discrete in the sense that the sample space Ω was discrete.

Consider the common situation where the experiment results in a d-dimensional random vector $\boldsymbol{X}(\omega)$ and that our inference is based on that random vector. Assume that $\boldsymbol{X}$ has a continuous distribution and let f_θ denote its density under $\mathbb{P}_\theta$.

The likelihood methods are defined as before with the density replacing the mass function. This is (at least morally) justified by the fact that, in a continuous setting, a density plays the role of mass function. In more detail, letting $\boldsymbol{x} = \boldsymbol{X}(\omega)$, the *likelihood function* is now defined as

$$\mathsf{lik}(\theta) := f_\theta(\boldsymbol{x}).$$

Based on this new definition of the likelihood, the maximum likelihood estimator and the likelihood ratio are defined as they were before.

12.5.2 Two-sided p-value

It is sometimes the case that large and small values of a test statistic provide evidence against the null hypothesis. For example, in the binomial experiment of Example 12.2, consider testing

$$\mathcal{H}_0 : \theta_* = \theta_0.$$

Problem 12.27. Show that the LR is an increasing function of

$$\hat{\theta} \log\left(\frac{\hat{\theta}}{\theta_0}\right) + (1-\hat{\theta}) \log\left(\frac{1-\hat{\theta}}{1-\theta_0}\right),$$

where $\hat{\theta}$ denotes the maximum likelihood estimate as in Example 12.14.

Thus the application of the likelihood ratio procedure is as straightforward as it was in the one-sided situation considered earlier.

However, let's look directly at the MLE. If $\hat{\theta}$ is quite large, or if it is quite small, compared to θ_0, this is evidence against the null. In such a situation, the p-value can be defined in a number of ways. A popular way is based on the minimum of the two one-sided p-values, namely

$$2 \min\left\{\mathbb{P}_{\theta_0}(Y \geq y), \mathbb{P}_{\theta_0}(Y \leq y)\right\},$$

where Y is the total number of heads in the sequence and $y = Y(\omega)$ is the observed value of Y, as before.

Problem 12.28. Compare this p-value with the p-value resulting from using the LR.

12.6 Additional problems

Problem 12.29 (German Tank Problem[70])**.** Suppose we have an iid sample from the uniform distribution on $\{1, \ldots, \theta\}$, where $\theta \in \mathbb{N}$ is unknown. Derive the maximum likelihood estimator for θ.

Problem 12.30 (Gosset's experiment)**.** Fit a Poisson model to the data of Table 3.1 by maximum likelihood. Add a row to the table to display the corresponding expected counts so as to easily compare them with the actual counts. In R, do this visually by drawing side-by-side bar plots with different colors and a legend.

Problem 12.31 (Rutherford's experiment)**.** Same as in Problem 12.30, but with the data of Table 3.2.

[70] The name of the problem comes from World War II, where the Western Allies wanted to estimate the total number of German tanks in operation (θ above) based on the serial numbers of captured or destroyed German tanks. In such a setting, is the assumption of iid-ness reasonable?

CHAPTER 13

PROPERTIES OF ESTIMATORS AND TESTS

In this chapter we introduce and briefly discuss some properties of estimators and tests.

13.1 Sufficiency

We have referred to the setting of Example 12.2 as a binomial experiment. The main reason is that the binomial distribution is at the very center of the resulting statistical inference. In what we did, this was a consequence of us relying on the number of heads, Y, which has the binomial distribution with parameters (n, θ). Surely, we could have based our inference on a different statistic. However, there is a fundamental reason that inference should be based on Y: because Y contains all the information about θ that we can extract from the data. This is rather intuitive because in going from the sequence of tosses ω to the number of heads $Y(\omega)$, all that is lost is the position of the heads in the sequence, that is, the order. But because the tosses are assumed iid, the order cannot provide any

information on the parameter θ.

More formally, assume a statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$. We say that a k-dimensional vector-valued statistic $\boldsymbol{Y}$ is *sufficient* for this family if

$$\text{For any event } \mathcal{A} \text{ and any } \boldsymbol{y} \in \mathbb{R}^k,$$
$$\mathbb{P}_\theta(\mathcal{A} \mid \boldsymbol{Y} = \boldsymbol{y}) \text{ does not depend on } \theta.$$

If $\boldsymbol{Y} = (Y_1, \dots, Y_k)$, we say that the statistics $Y_1, \dots, Y_k$ are *jointly sufficient*.

Thus, intuitively, a statistic $\boldsymbol{Y}$ is sufficient if the randomness left after conditioning on $\boldsymbol{Y}$ does not depend on the value of θ, so that this leftover randomness cannot be used to improve the inference.

Theorem 13.1 (Factorization criterion)**.** *Consider a family of either mass functions or densities, $\{f_\theta : \theta \in \Theta\}$, over a sample space Ω. Then a statistic $\boldsymbol{Y}$ is sufficient for this model if and only if there are functions $\{g_\theta : \theta \in \Theta\}$ and h such that, for all $\theta \in \Theta$,*

$$f_\theta(\omega) = g_\theta(\boldsymbol{Y}(\omega))h(\omega), \quad \textit{for all } \omega \in \Omega.$$

Problem 13.2. In the binomial experiment (Example 12.2), show that the number of heads is sufficient.

Problem 13.3. In the German tank problem (Problem 12.29), show that the maximum of the observed (serial) numbers is sufficient.

13.2 Consistency

The notion of consistency is best understood in an asymptotic model where the sample size becomes large. This can be confusing, as in a given experiment, we only have access to a finite sample, which is fixed. We adopt here a rather formal stance for the sake of clarity.

Consider a sequence of statistical models, $(\Omega_n, \Sigma_n, \mathcal{P}_n)$, where $\mathcal{P}_n = \{\mathbb{P}_{n,\theta} : \theta \in \Theta\}$. Note that the parameter space does not vary with n. An important special case is that of product spaces where $\Omega_n = \Omega_\star^n$ for some set $\Omega_\star$, meaning that $\omega \in \Omega_n$ can be written as $\omega = (\omega_1, \dots, \omega_n)$ with $\omega_i \in \Omega_\star$. In that case, n typically represents the sample size.

A *statistical procedure* is a sequence of statistics, therefore of the form $S = (S_n)$ with S_n being a statistic defined on Ω_n. An example of procedure in the context of estimation is the maximum likelihood method. An example of procedure in the context of testing is the likelihood ratio method.

Remark 13.4. Asymptotic results, meaning those that describe a situation as $n \to \infty$, can be difficult to interpret. Indeed, in most real-life situations, a sample is collected and the subsequent analysis is necessarily based on that sample alone, as no additional observations are collected.

That being said, such results do provide some theoretical foundation for Statistics not unlike the Law of Large Numbers does for Probability Theory.

13.2.1 Consistent estimators

We say that an procedure $S = (S_n)$ is *consistent* for estimating $\varphi(\theta)$ if, for any $\varepsilon > 0$,

$$\mathbb{P}_{n,\theta}(|S_n - \varphi(\theta)| \geq \varepsilon) \to 0, \quad \text{as } n \to \infty.$$

Problem 13.5 (Binomial experiment)**.** In Example 12.2, consider the task of estimating the parameter θ. Show that the maximum likelihood method yields an estimation procedure that is consistent. [This is a simple consequence of the Law of Large Numbers.]

In fact, the MLE is consistent under fairly broad assumptions. If there are multiple values of the parameter that maximize the likelihood, we assume that one such value is chosen to define the MLE.

Proposition 13.6 (Consistency of the MLE)**.** *Consider an identifiable family of densities or mass functions* $\{f_\theta : \theta \in \Theta\}$ *having same support* $\mathcal{X}$*. Assume that* Θ *is a compact subset of some Euclidean space; that* $f_\theta(x) > 0$ *and that* $\theta \mapsto f_\theta(x)$ *is continuous, for all* $x \in \mathcal{X}$*; and that* $\sup_{\theta\in\Theta} |f_\theta(x)|$ *is integrable with respect to* f_{θ^*}*, where* θ^* *denotes the true value of the parameter. Then, as a procedure, the MLE is well defined and consistent.*

Problem 13.7. Prove this proposition when Θ is a compact interval of the real line. [The main ingredients are Jensen's inequality and the uniform law of large numbers as stated in Problem 16.101.]

13.2.2 Consistent tests

We say that a procedure $S = (S_n)$ is *consistent* for testing $\mathcal{H}_0 : \theta \in \Theta_0$, if there is a sequence of critical values (c_n) such that

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta}\big(S_n \geq c_n\big) = 0, \quad \text{for all } \theta \in \Theta_0, \tag{13.1}$$

$$\lim_{n\to\infty} \mathbb{P}_{n,\theta}\big(S_n \geq c_n\big) = 1, \quad \text{for all } \theta \notin \Theta_0. \tag{13.2}$$

(We have implicitly assumed that large values of S_n weigh against the null hypothesis.)

Problem 13.8 (Binomial experiment)**.** In Example 12.2, consider the task of testing $\mathcal{H}_0 : \theta_* \leq \theta_0$. Show that likelihood ratio method yields a procedure that is consistent for $\mathcal{H}_0$.

The LR test is consistent under fairly broad assumptions.

Problem 13.9. State and prove a consistency result for the LR test under conditions similar to those in Proposition 13.6.

13.3 Notions of optimality for estimators

Given a loss function $\mathcal{L}$, the risk of an estimator S for θ is defined in (12.6). Obviously, the smaller the risk the better the estimator. However, this presents a difficulty since the risk is a function of θ rather than just a number.

13.3.1 Maximum risk and average risk

We present two ways of reducing the risk function to a number.

Maximum risk The first avenue is to consider the *maximum risk*, namely the maximum of the risk function,

$$\mathcal{R}_{\max}(S) := \sup_{\theta\in\Theta} \mathcal{R}_\theta(S).$$

An estimator that minimizes the maximum risk, if one exists, is said to be *minimax*, and its risk is called the *minimax risk*, denoted $\mathcal{R}^*_{\max}$.

Average risk The second avenue is to consider the *average risk* (aka *Bayes risk*). To do so, we need to choose a distribution on the parameter space. (A distribution on the parameter space is often called a *prior*.) Assuming that Θ is a subset of some Euclidean space, let λ be a density supported on Θ. We can then consider the average risk with respect to λ,

$$\mathcal{R}_\lambda(S) := \int_\Theta \mathcal{R}_\theta(S)\lambda(\theta)\mathrm{d}\theta.$$

An estimator that minimizes this average risk is called a *Bayes estimator* (with respect to λ) and the minimum average risk, denoted $\mathcal{R}^*_\lambda$.

A Bayes estimator is often derived as follows. For simplicity, assume a family of densities $\{f_\theta : \theta \in \Theta\}$ on some Euclidean sample space Ω and that we are estimating θ itself. Applying the Fubini–Tonelli Theorem, we have

$$\begin{aligned}\mathcal{R}_\lambda(S) &= \int_\Theta \int_\Omega \mathcal{L}(S(\omega),\theta) f_\theta(\omega)\mathrm{d}\omega\lambda(\theta)\mathrm{d}\theta \\ &= \int_\Omega \int_\Theta \mathcal{L}(S(\omega),\theta) f_\theta(\omega)\lambda(\theta)\mathrm{d}\theta\mathrm{d}\omega.\end{aligned}$$

Thus, if the following is well-defined,

$$S_\lambda(\omega) := \arg\min_{s\in\Theta} \int_\Theta \mathcal{L}(s,\theta) f_\theta(\omega)\lambda(\theta)\mathrm{d}\theta,$$

it is readily seen to minimize the λ-average risk.

Problem 13.10 (Binomial experiment)**.** Consider a binomial experiment as in Example 12.9 and the estimation of θ under squared error loss. For the MLE:

(i) Compute its maximum risk.

(ii) Compute its average risk with respect to the uniform distribution on $[0, 1]$.

Maximum risk optimality and average risk optimality are intricately connected. For example, for any prior λ,

$$\mathcal{R}_\lambda(S) \le \mathcal{R}_{\max}(S), \quad \text{for all } S,$$

and this immediately implies that

$$\mathcal{R}^*_\lambda \le \mathcal{R}^*_{\max}.$$

Problem 13.11. Show that an estimator S is minimax when there is a sequence of priors (λ_k) such that $\liminf_{k\to\infty} \mathcal{R}^*_{\lambda_k} \ge \mathcal{R}_{\max}(S)$.

Problem 13.12. Use the previous problem to show that a Bayes estimator with constant risk function is necessarily minimax.

Problem 13.13. In a binomial experiment, derive a minimax estimator. To do so, find a prior in the Beta family such that the resulting Bayes estimator has constant risk function. Using R, produce a graph comparing the risk functions of this estimator and that of the MLE. Do so for various values of n.

13.3.2 Admissibility

We say that an estimator S is *inadmissible* if there is an estimator T such that

$$\mathcal{R}_\theta(T) \le \mathcal{R}_\theta(S), \quad \text{for all } \theta \in \Theta,$$

and the inequality is strict for at least one $\theta \in \Theta$. Otherwise we say that the estimator S is *admissible*.

At least in theory, if an estimator is inadmissible, it can be replaced by another estimator that is uniformly better in terms of risk. However, even then, there might be other reasons for using an inadmissible estimator, such as simplicity or ease of computation.

Admissibility is interrelated with maximum risk and average risk optimality.

Problem 13.14. Show that an estimator that is unique Bayes for some prior is necessarily admissible.

Problem 13.15. Show that an estimator that is unique minimax is necessarily admissible.

13.3.3 Risk unbiasedness

We say here that an estimator S is *risk unbiased* if

$$\mathbb{E}_\theta[\mathcal{L}(S,\theta)] \le \mathbb{E}_\theta[\mathcal{L}(S,\theta')]. \tag{13.3}$$

In words, this means that S is on average as close (as measured by the loss function) to the true value of the parameter as any other value.

We assume that the feature of interest, $\varphi(\theta)$, is real.

Mean unbiased estimators Suppose that $\mathcal{L}$ is the squared error loss, meaning $\mathcal{L}(s,\theta) = (s-\varphi(\theta))^2$.

Problem 13.16. Show that S is risk unbiased if and only if it is unbiased in the sense of (12.9).

Such estimators are said to be *mean-unbiased*, or more commonly, simply *unbiased*.

From the bias-variance decomposition (12.4), an estimator with small MSE has necessarily small bias (and also small variance). Therefore, a small bias is desirable. However, strict unbiasedness is not necessarily desirable, for a small MSE does not imply unbiasedness. In fact, unbiased estimators may not even exist.

Problem 13.17. Consider a binomial experiment as in Example 12.9. Show that there is an unbiased estimator of $\varphi(\theta)$ if and only if φ is a polynomial of degree at most n.

Median unbiased estimators Suppose that $\mathcal{L}$ is the absolute loss, meaning $\mathcal{L}(s,\theta) = |s-\varphi(\theta)|$.

Problem 13.18. Show that S is risk unbiased if and only if, for all $\theta \in \Theta$, $\varphi(\theta)$ is a median of S under $\mathbb{P}_\theta$.

Such estimators are said to be *median-unbiased*.

Problem 13.19. Show that if S is median-unbiased for $\varphi(\theta)$, then for any strictly monotone function $g:\mathbb{R}\to\mathbb{R}$, $g(S)$ is median-unbiased for $g(\varphi(\theta))$. Show by exhibiting a counter-example that this is no longer true if 'median' is replaced with 'mean'.

13.4 Notions of optimality for tests

Our discussion of optimality for estimators has parallels for tests. Indeed, once the level is under control, the larger the power (i.e., the more the test rejects) the better. However, this is quantified by a power function and not a simple number.

We consider a statistical model as in (12.1) and consider testing $\mathcal{H}_0 : \theta \in \Theta_0$. Unless otherwise specified, Θ_1 is the complement of Θ_0 in Θ, as in (12.13).

Remark 13.20 (Level requirement)**.** We assume that all tests that appear below have level α. It is crucial that the test being evaluated satisfy the prescribed level, for otherwise the evaluation is not accurate and a comparison with other tests that do satisfy the level requirement is not fair.

13.4.1 Minimum power and average power

There are various ways of reducing the power function to a single number.

Minimum power A first avenue is to consider the *minimum power*,

$$\inf_{\theta \in \Theta_1} \mathbb{P}_\theta(\phi = 1).$$

In a number of classical models, Θ is a domain of a Euclidean space and the power function of any test is continuous over Θ. In such a setting, if there is no 'separation' between Θ_0 and Θ_1, then any test has minimum power bounded from above by its size, which is not very interesting. The consideration of minimum power is thus only relevant when there is a 'separation' between the null and alternative sets.

Average power A second avenue is to consider the *average power*. Let λ be a density on Θ_1 (assuming Θ is a subset of a Euclidean space). We can then average the power with respect to λ,

$$\int_{\Theta_1} \mathbb{P}_\theta(\phi = 1)\lambda(\theta)\mathrm{d}\theta.$$

Problem 13.21 (Binomial experiment)**.** Consider a binomial experiment as in Example 12.14 and the testing of $\Theta_0 := [0, \theta_0]$, where θ_0 is given. For the level α test based on rejecting for large values of Y:

(i) Compute the minimum power when $\Theta = \Theta_0 \cup \Theta_1$ where $\Theta_1 := [\theta_1, 1]$ for $\theta_1 > \theta_0$ given.

(ii) Compute the average power with respect to the uniform distribution on Θ_1.

Answer these questions analytically and also numerically, using R.

13.4.2 Admissibility

We say that a test ϕ is *inadmissible* if there is a test ψ such that

$$\mathbb{P}_\theta(\phi = 1) \le \mathbb{P}_\theta(\psi = 1), \quad \text{for all } \theta \in \Theta_1,$$

and the inequality is strict for at least one $\theta \in \Theta_1$.

At least in theory, if a test is inadmissible, it can be replaced by another test that is uniformly better in terms of power. However, even then, there might be other reasons for using an inadmissible test, such as simplicity or ease of computation.

13.4.3 Uniformly most powerful tests

Consider testing $\theta \in \Theta_0$. A test ϕ is said to be *uniformly most powerful (UMP)* among level α tests if ϕ itself has level α and is at least as powerful as any other level α test, meaning that for any other test ψ with level α,

$$\mathbb{P}_\theta(\phi = 1) \geq \mathbb{P}_\theta(\psi = 1), \quad \text{for all } \theta \in \Theta_1.$$

This is clearly the best one can hope for. However, a UMP test seldom exists.

Simple vs simple A very particular case where a UMP test exists is when both the null set and alternative set are singletons. We say that a hypothesis is *simple* if the corresponding parameter subset is a singleton; it is said to be *composite* otherwise. Suppose therefore that $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, and let $\mathbb{Q}_j$ be short for $\mathbb{P}_{\theta_j}$.

The following is a consequence of the *Neyman–Pearson Lemma*[71], one of the most celebrated results in the theory of tests. Recall that a likelihood ratio test is any test that rejects for large values of the likelihood ratio.

Theorem 13.22. *In the present context, any LR test is UMP at level α equal to its size.*

To understand where the result comes from, suppose we want to test at a prescribed level α in a situation where the sample space is discrete. In that case, we want to solve the following optimization problem

$$\begin{aligned} &\text{maximize} \sum_{\omega \in \mathcal{R}} \mathbb{Q}_1(\omega) \\ &\text{subject to} \sum_{\omega \in \mathcal{R}} \mathbb{Q}_0(\omega) \leq \alpha. \end{aligned}$$

The optimization is over subsets $\mathcal{R} \subset \Omega$, which represent candidate rejection regions. In that case, it makes sense to rank $\omega \in \Omega$ according to its likelihood ratio value $L(\omega) := \mathbb{Q}_1(\omega)/\mathbb{Q}_0(\omega)$. If we denote by $\omega_1, \omega_2, \dots$ the elements of Ω in decreasing order, meaning that $L(\omega_1) \geq L(\omega_2) \geq \cdots$, and define the regions $\mathcal{R}_k = \{\omega_1, \dots, \omega_k\}$, then it makes

[71] Named after Jerzy Neyman (1894 - 1981) and Egon Pearson (1895 - 1980).

intuitive sense to choose $\mathcal{R}_{k_\alpha}$, where

$$k_\alpha := \max\{k : \mathbb{Q}_0(\omega_1) + \cdots + \mathbb{Q}_0(\omega_k) \le \alpha\}.$$

This is correct when the level can be exactly achieved in this fashion. Otherwise, the optimization is more complex, leading to a linear program.

In a different setting where the sample space is a subset of some Euclidean space, suppose that $\mathbb{Q}_1$ has density f_1 and that $\mathbb{Q}_0$ has density f_0. In that case, the likelihood ratio is defined as $L := f_1/f_0$. If L has continuous distribution under $\mathbb{Q}_0$, by choosing the critical value c appropriately, any prescribed level α can be attained. As a consequence, if c is chosen such that $\mathbb{Q}_0(L \ge c) = \alpha$, then the test with rejection region $\{L \ge c\}$ is UMP at level α. The same cannot always be done if L does not have a continuous distribution under the null.

Monotone likelihood ratio property As in Section 12.2.2, assume a discrete model for concreteness, although what follows applies more broadly. We consider a situation where Θ is an interval of $\mathbb{R}$.

The family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is said to have the *monotone likelihood ratio (MLR)* property in T if the model is identifiable and, for any $\theta < \theta'$, $\mathbb{P}_{\theta'}/\mathbb{P}_\theta$ is monotone increasing in T, meaning there is a non-decreasing function $g_{\theta,\theta'} : \mathbb{R} \to \mathbb{R}$ satisfying

$$\frac{\mathbb{P}_{\theta'}(\omega)}{\mathbb{P}_\theta(\omega)} = g_{\theta,\theta'}(T(\omega)), \quad \text{for all } \omega \in \Omega.$$

Theorem 13.23. *Assume that the family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ has the MLR property in T and that the null hypothesis is of the form $\Theta_0 = \{\theta \in \Theta : \theta \le \theta_0\}$. Then any test of the form $\{T \ge t\}$ has size $\mathbb{P}_{\theta_0}(T \ge t)$, and is UMP at level α equal to its size.*

Problem 13.24. Show that in the context of Theorem 13.23, the LR is a non-decreasing function of T and that, therefore, an LR test is UMP at level α equal to its size.

Problem 13.25 (Binomial experiment)**.** Show that the MLR property holds in a binomial experiment, and that in particular, for testing $\mathcal{H}_0 : \theta_* \le \theta_0$, a LR test is UMP at its size. Show that the same is true of a test based on the maximum likelihood estimator.

13.4.4 Unbiased tests

As we said above, a UMP test rarely exists. In particular, it does not exist in the most popular two-sided situations, including if the MLR property holds.

Consider a setting where Θ is an interval of the real line and the null hypothesis to be tested is $\mathcal{H}_0 : \theta_* = \theta_0$ for some given θ_0 in the interior of Θ. (If θ_0 is one of the boundary points, the situation is one-sided.) In such a situation, a UMP test would have to be at least as good as a UMP test for the one-sided null $\mathcal{H}_0^{\leq} : \theta_* \leq \theta_0$ and at least as good as a UMP test for the one-sided null $\mathcal{H}_0^{\geq} : \theta_* \geq \theta_0$. In most cases, this proves impossible.

In some sense this competition is unfair, because a test for the one-sided null such as $\mathcal{H}_0^{\leq}$ is ill-suited for the two-sided null $\mathcal{H}_0$. Indeed, if a test for $\mathcal{H}_0^{\leq}$ has level α, then the probability that it rejects when $\theta_* \leq \theta_0$ is bounded from above by α.

To prevent one-sided tests from competing in two-sided testing problems, one may restrict attention to so-called unbiased tests. A test is said to be *unbiased* at level α if it has level α, and the probability of rejecting at any $\theta \in \Theta_1$ is bounded from below by α.

While the number of situations where a UMP test exists is rather limited, there are many more situations where there exists a test that is UMP among unbiased tests. Such a test is said to be *UMPU*.

An important class of situations where this occurs includes the case of *general exponential families*, where we work with a family of densities $\{f_\theta : \theta \in \Theta\}$ of the form

$$f_\theta(\omega) = A(\theta) \exp(\varphi(\theta) T(\omega))\, h(\omega),$$

where, in addition, φ is strictly increasing on Θ, assumed to be an open interval. Suppose in this setting that the null set Θ_0 is a closed subinterval of Θ, possibly a singleton.

Theorem 13.26. *In the present setting, any test of the form $\{T \leq t_1\} \cup \{T \geq t_2\}$, with $t_1 < t_2$, is UMPU at its size.*

Problem 13.27 (Binomial experiment)**.** Show that a binomial experiment leads to a general exponential family.

Remark 13.28. In a binomial experiment, an equal-tailed two-sided LR test is approximately UMPU for $\theta_* = \theta_0$ as long as θ_0 is not too close to 0 or 1. (The larger the number of trials n, the closer θ_0 can be to these extremes.)

13.5 Additional problems

Problem 13.29. Consider a binomial experiment as before and the estimation of θ. Is the MLE median-unbiased?

CHAPTER 14

ONE PROPORTION

Estimating a proportion is one of the most basic problems in statistics. Although basic, it arises in a number of important real-life situations, for example:

- *Election polls* are conducted to estimate the proportion of people that will vote for a particular candidate.
- In *quality control*, the proportion of defective items manufactured at a particular plant or assembly line needs to be monitored, and one may resort to statistical inference to avoid having to check every single item.
- *Clinical trials* are conducted in part to estimate the proportion of people that would benefit (or suffer serious side effects) from receiving a particular treatment.

The situation is commonly modeled as sampling from an urn. The resulting distribution, as we know, depends on the contents of the urn and on how the sampling is done. This, of course, changes how statistical inference is

performed.

14.1 Binomial experiments

We start with the binomial experiment of Example 12.2, which has served us as our running example in the last two chapters. Remember that this is an experiment where a θ-coin is tossed a predetermined number of times n, and the goal is to infer the value of θ based on the outcome of the experiment (the data).

To summarize what we have learned about this model so far, in Chapter 12 we derived the maximum likelihood estimator, which is the sample proportion of heads, that is, $S = Y/n$. We then focused on testing null hypotheses of the form $\theta_* \le \theta_0$ for a given θ_0. We considered tests which reject for large values of the MLE, meaning with rejection regions of the form $\{S \ge c\}$. Using the correspondence between tests and confidence intervals, we derived the (one-sided) confidence interval (12.33).

Problem 14.1. Adapt the arguments to testing null hypotheses of the form $\theta_* \ge \theta_0$ for a given θ_0 and derive the corresponding confidence interval at level $1 - \alpha$. The interval will be of the form

$$I = [0, \bar{S}). \tag{14.1}$$

14.1.1 Two-sided tests and confidence intervals

We now consider the two-sided situation. In brief, we study the problem of testing null hypotheses of the form $\mathcal{H}_0 : \theta_* = \theta_0$ for a given $\theta_0 \in (0, 1)$ and then derive a confidence interval as in Section 12.4.9. (If $\theta_0 = 0$ or $= 1$, the problem is one-sided.) We still use the MLE as test statistic.

Tests For the null $\theta_* = \theta_0$, both large and small values of Y (the number of heads) are evidence against the null hypothesis. This leads us to consider tests of the form

$$\phi = \{Y \le a \text{ or } Y \ge b\}. \tag{14.2}$$

The critical values, a and b, are chosen to control the level at some prescribed α, meaning

$$\mathbb{P}_{\theta_0}(\phi = 1) \le \alpha.$$

Equivalently,

$$\mathbb{P}_{\theta_0}(Y \le a) + \mathbb{P}_{\theta_0}(Y \ge b) \le \alpha. \tag{14.3}$$

Note that a number of choices are possible. Two natural ones are

- *Equal tail.* This choice corresponds to choosing

$$a = \max\{a' : \mathbb{P}_{\theta_0}(Y \le a') \le \alpha/2\}, \tag{14.4}$$
$$b = \min\{b' : \mathbb{P}_{\theta_0}(Y \ge b') \le \alpha/2\}, \tag{14.5}$$

  where the maximization and minimization are over integers.

- *Minimum length.* This choice corresponds to minimizing $b - a$ subject to (14.3).

Problem 14.2. Derive the LR test in the present context. You will find it is of the form (14.2) for particular critical values a and b (to be derived explicitly). How does the LR test compare with the equal-tail and minimum-length tests above?

Confidence intervals Let ϕ_θ be a test for $\theta_* = \theta$ as constructed above, meaning with rejection region of the form $\{Y \le a_\theta\} \cup \{Y \ge b_\theta\}$, where

$$\mathbb{P}_\theta(Y \le a_\theta) + \mathbb{P}_\theta(Y \ge b_\theta) \le \alpha.$$

As in Section 12.4.9, based on this family of tests we can obtain a $(1-\alpha)$-confidence interval of the form

$$I = (\boldsymbol{S}, \bar{S}). \tag{14.6}$$

For building a confidence interval, the minimum-length choice for a_θ and b_θ is particularly appealing.

Problem 14.3. Derive this confidence interval.

The one-sided intervals (12.33) and (14.1), and the two-sided interval (14.6) with the equal-tail construction, are due to Clopper and Pearson [41]. The construction yields an exact interval in the sense that the desired confidence level is achieved.

R corner. The Clopper–Pearson interval (one-sided or two-sided), and the related test, can be computed in R using the function binom.test.

We describe below other traditional ways of constructing confidence intervals. [72] Compared to the Clopper–Pearson construction, they are less labor intensive although they are not as precise. They were particularly useful in the pre-computer age, and some of them are still in use.

14.1.2 Chebyshev's confidence interval

We saw in Problem 12.10 how to compute a confidence interval using Chebyshev's inequality. Since, for all θ,

$$\mathrm{Var}_\theta(S) = \mathrm{Var}_\theta(Y/n) = \frac{n\theta(1-\theta)}{n^2} \le \frac{1}{4n},$$

[72] We focus here on confidence intervals, from which we know tests can be derived as in Section 12.4.7.

(Problem 7.37), the resulting interval, at confidence level $1-\alpha$, is

$$\Big(S \pm \frac{1}{\sqrt{4\alpha n}}\Big). \tag{14.7}$$

This interval is rather simple to derive and does not require heavy computations that would necessitate the use of a computer. However, it is very conservative, meaning quite wide compared to the Clopper–Pearson interval.

As a possible refinement, one can avoid using an upper bound on the variance. Indeed, Chebyshev's inequality tells us that

$$\frac{|S-\theta|}{\sqrt{\theta(1-\theta)/n}} < z, \tag{14.8}$$

with probability at least $1-1/z^2$.

Problem 14.4. Prove that (14.8) is equivalent to

$$\theta \in I_z := \Big(S_z \pm z\frac{\sigma_z}{\sqrt{n}}\Big), \tag{14.9}$$

where

$$S_z := \frac{S+z^2/2n}{1+z^2/n}, \quad \sigma_z^2 := \frac{S(1-S)+z^2/4}{(1+z^2/n)^2}.$$

In particular I_z defined in (14.9) is a $(1-1/z^2)$-confidence interval for θ.

14.1.3 Confidence intervals based on the normal approximation

The Chebyshev's confidence interval is commonly believed to be too conservative, and practitioners have instead relied on the normal approximation to the binomial distribution instead of Chebyshev's inequality, which is deemed too crude. Let Φ denote the distribution function of the standard normal distribution (which we saw in (5.2)). Then, using the Central Limit Theorem,

$$\mathbb{P}_\theta\Big(\frac{|S-\theta|}{\sqrt{\theta(1-\theta)/n}} < z\Big) \approx 2\Phi(z)-1, \tag{14.10}$$

for all $z>0$ as long as the sample size n is large enough. (More formally, the left-hand side converges to the right-hand side as $n\to\infty$.)

Wilson's normal interval The construction of this interval, proposed by Wilson [254], is based on the large-sample approximation (14.10) and the derivations of Problem 14.4, which together imply that the interval defined in (14.9) has approximate confidence level $2\Phi(z)-1$.

R corner. Wilson's interval (one-sided or two-sided) and the related test, can be computed in R using the function prop.test.

The simpler variants that follow are also based on the approximation (14.10). However, they offer no advantage compared to Wilson's interval, except for simplicity if calculations must be done by hand. These constructions start by noticing that (14.10) implies

$$\mathbb{P}_\theta(\theta \in J_\theta) \approx 1 - \alpha, \tag{14.11}$$

where

$$J_\theta := \left(S_n \pm z_{1-\alpha/2} \frac{\sigma_\theta}{\sqrt{n}}\right), \tag{14.12}$$

where $z_u := \Phi^{-1}(u)$ and $\sigma_\theta^2 := \theta(1-\theta)$. At this point, J_θ is not a confidence interval as its computation depends on θ, which is unknown.

Conservative normal interval In our derivation of the interval (14.7), we used the fact that $\theta \in [0,1] \mapsto \theta(1-\theta)$ is maximized at $\theta = 1/2$. Therefore,

$$J_\theta \subset J_{1/2} = \left(S \pm \frac{z_{1-\alpha/2}}{2\sqrt{n}}\right).$$

$J_{1/2}$ can be computed without knowledge of θ and, because of (14.11), it achieves a confidence level of at least $1-\alpha$ in the large-sample limit. Unless the true value of θ happens to be equal to $1/2$, this interval will be conservative in large samples.

Plug-in normal interval It is very tempting to replace θ in (14.12) with S. After all, S is a consistent estimator of θ. The resulting interval is

$$J_S = \left(S \pm z_{1-\alpha/2}\, \sigma_S\right).$$

J_S is a bona fide confidence interval. Moreover, (14.11), coupled with Slutsky's theorem (Theorem 8.48), implies that J_S achieves a confidence level of $1-\alpha$ in the large-sample limit. Note that this construction relies on two approximations.

Problem 14.5. Verify the claims made here.

14.2 Hypergeometric experiments

Consider an experiment where balls are repeatedly drawn from an urn containing r red balls and b blue balls a predetermined number of times n. The total number of balls in the urn, $v := r + b$, is assumed known. The goal is to infer the proportion of red balls in the urn, namely, r/v. If the draws are with replacement, this is a binomial experiment with probability parameter $\theta = r/v$, a case that was treated in Section 14.1. We assume here that the draws are without replacement.

Let Y denote the number of red balls that are drawn. Assume that $n < v$, for otherwise the experiment reveals

the contents of the urn and there is no inference left to do. We call the resulting experiment a *hypergeometric experiment* because Y is hypergeometric and sufficient for this experiment.

Problem 14.6 (Sufficiency)**.** Prove that Y is indeed sufficient in this model.

14.2.1 Maximum likelihood estimator

Let y denote the realization of Y, meaning $y = Y(\omega)$ with ω denoting the observed outcome of the experiment. Recalling the definition of falling factorials (2.14), the likelihood is given by (see (2.17))

$$\mathsf{lik}(r) = \frac{(r)_y\,(v-r)_{n-y}}{(v)_n},$$

where we used the fact that $b = v - r$. (As we saw before, although the likelihood is a function of y also, this dependency is left implicit to focus on the parameter r.)

Although it may look intimidating, this is a tame function. It suffices to consider r in the range $y \le r \le v - n + y$, for otherwise the likelihood is zero. (This is congruent with the fact that, having drawn y red balls and $n-y$ blue balls, we know that there were that many red and blue balls in the urn to start with.) For $r < v - n + y$, we have

$$\frac{\mathsf{lik}(r+1)}{\mathsf{lik}(r)} = \frac{(r+1)(v-r-n+y)}{(r-y+1)(v-r)},$$

so that

$$\mathsf{lik}(r+1) \le \mathsf{lik}(r) \;\Leftrightarrow\; r \ge \frac{yv-n+y}{n}.$$

Similarly, for $r > y$, we have

$$\mathsf{lik}(r-1) \le \mathsf{lik}(r) \;\Leftrightarrow\; r \le \frac{yv+y}{n}.$$

We conclude that any r that maximizes the likelihood satisfies

$$-\frac{n-y}{nv} \le \frac{r}{v} - \frac{y}{n} \le \frac{y}{nv}. \tag{14.13}$$

More than necessary, this is also sufficient for r to maximize the likelihood.

Problem 14.7. Verify that r (integer) satisfies this condition if and only if r/v is closest to y/n. Show that there is only one such r, except when $y/n = k/(v+1)$ for some integer k, in which case there are two such r. Conclude that, in any case, any r satisfying (14.13) maximizes the likelihood.

In the situation where there are two maximizers of the likelihood, we let the MLE be their average. Note that r/v is the proportion of red balls in the urn, and thus the MLE is in essence the same as in a binomial experiment with $\theta = r/v$ as parameter, except that the proportion of reds is here an integer multiple of $1/v$.

14.2.2 Confidence intervals

The various constructions of a confidence interval that we presented in the context of a binomial experiment apply almost verbatim in the context of a hypergeometric experiment. This is because the same normal approximation that applies to the binomial distribution with parameters (n, θ) also applies to the hypergeometric distribution with parameters $(n, r, v - r)$, with r/v in place of θ, if

$$v \to \infty, \quad r \to \infty, \quad n \to \infty, \tag{14.14}$$

$$\text{with } r/v \to \theta \in (0, 1), \quad n/r \to 0. \tag{14.15}$$

Problem 14.8. Prove this under the more stringent condition that $n/\sqrt{r} \to 0$. [Use Problem 2.36 and the normal approximation to the binomial.]

Problem 14.9. Derive the exact (Clopper–Pearson) one-sided and then two-sided confidence intervals for a hypergeometric experiment. Then implement this as a function in R.

14.2.3 Comparison with a binomial experiment

We already argued that a hypergeometric experiment with parameters $(n, r, v - r)$ is very similar to a binomial experiment with parameters (n, θ) with $\theta = r/v$. In fact, the two are essentially equivalent when the size of the urn increases and (14.14) holds.

In finer detail, however, it would seem that the former, where sampling is without replacement, allows for more precise inference compared to the latter, where sampling is with replacement and therefore seemingly wasteful to a certain degree. Indeed, if the balls are numbered (which we can always assume, at least as a thought experiment) and we have already drawn ball number i, then drawing it again does not provide any additional information on the contents of the urn.

Sampling without replacement is indeed preferable, because the resulting confidence intervals are narrower.

Problem 14.10. Verify numerically that the Clopper–Pearson two-sided interval is narrower in a hypergeometric experiment compared to the corresponding binomial experiment.

14.3 Negative binomial and negative hypergeometric experiments

Negative binomial experiments Consider an experiment that consists in tossing a θ-coin until a predetermined number of heads, m, has been observed. Thus the number of trials is not set in advance, in contrast with a binomial experiment. The goal, as before, is to infer the value of θ based on the result of such an experiment. In practice, such a design might be appropriate in situations where θ is believed to be small.

Thus let $(X_i : i \geq 1)$ denote the Bernoulli trials with parameter θ and let N denote the number of tails until m heads are observed. We call the resulting experiment a *negative binomial experiment* because N is negative binomial with parameters (m, θ) and sufficient for this experiment.

Problem 14.11 (Sufficiency)**.** Prove that N is indeed sufficient.

Problem 14.12 (Maximum likelihood)**.** Prove that the MLE for θ is $m/(m + N)$. Note that this is still the observed proportion of heads in the sequence, just as in a binomial experiment.

Problem 14.13. Derive the exact (Clopper–Pearson) one-sided and then two-sided confidence intervals for a negative binomial experiment. (These intervals have a simple closed form when $m = 1$, which could be called a *geometric experiment*.) Then implement this as a function in R.

Negative hypergeometric experiments When the experiment consists in repeatedly sampling without replacement from an urn, with r red and b blue balls, until m red balls are collected, we talk of a *negative hypergeometric experiment*, in particular because the key distribution in this case is the negative hypergeometric distribution.

Problem 14.14. Consider and solve the previous three problems in the present context.

14.4 Sequential experiments

We present here another classical experimental design where the number of trials is not set in advance of conducting the experiment, that may be appropriate in surveillance applications (e.g., epidemiological monitoring, quality control, etc). The setting is again that of a θ-coin being repeatedly tossed, resulting in Bernoulli trials $(X_i : i \geq 1)$. As before, we let $Y_n = \sum_{i=1}^{n} X_i$ and $S_n = Y_n/n$, which are

the number of heads and the proportion of heads in the first n tosses, respectively.

14.4.1 Sequential probability ratio test

Suppose we want to decide between two hypotheses

$$\mathcal{H}_0^{\leq} : \theta_* \leq \theta_0, \quad \text{versus} \quad \mathcal{H}_1^{\geq} : \theta_* \geq \theta_1, \tag{14.16}$$

where $0 \leq \theta_0 < \theta_1 \leq 1$ are given.

Example 14.15 (Multistage testing)**.** Such designs are used in mastery tests where a human subject's knowledge and command of some material or topic is tested on a computer. In such a context, $X_i = 1$ if the ith question is answered correctly, and $= 0$ otherwise, and θ_1 and θ_0 are the thresholds for Pass/Fail, respectively.

The *sequential probability ratio test (SPRT)* (aka *sequential likelihood ratio test*) was proposed by Abraham Wald (1902 - 1950) for this situation [242], except that he originally considered testing

$$\mathcal{H}_0^{=} : \theta_* = \theta_0, \quad \text{versus} \quad \mathcal{H}_1^{=} : \theta_* = \theta_1, \tag{14.17}$$

However, the same test can be applied verbatim to (14.16), which is more general. The procedure is based on the sequence of likelihood ratio test statistics (as the number of trials increases). The method is general and is here specialized to the case of Bernoulli trials.

The likelihood ratio statistic for $\mathcal{H}_0^{=}$ versus $\mathcal{H}_1^{=}$ is

$$L_n := \left(\frac{\theta_1}{\theta_0}\right)^{Y_n} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{n-Y_n}.$$

The test makes a decision in favor of $\mathcal{H}_0^{=}$ (resp. $\mathcal{H}_1^{=}$) if $L_n \leq c_0$ (resp. $L_n \geq c_1$), where the thresholds $c_0 < c_1$ are predetermined based on the desired level and power. (These depend in principle on n, but this is left implicit in what follows.) This testing procedure amounts to stopping the trials when there is enough evidence against either $\mathcal{H}_0^{=}$ or against $\mathcal{H}_1^{=}$.

More specifically, given $\alpha_0, \alpha_1 \in (0,1)$, c_0 and c_1 are chosen so that

$$s_0 := \mathbb{P}_{\theta_0}(L_n \geq c_1) \leq \alpha_0, \tag{14.18}$$

$$s_1 := \mathbb{P}_{\theta_1}(L_n \leq c_0) \leq \alpha_1. \tag{14.19}$$

These thresholds can be determined numerically in an efficient manner using a bisection search.

Problem 14.16. In R, write a function taking as input $(\theta_0, \theta_1, \alpha_0, \alpha_1)$ and returning (approximate) values for c_0 and c_1. Try your function on simulated data.

Proposition 14.17. *With* $c_0 = \alpha_1/(1-\alpha_0)$ *and* $c_1 = (1-\alpha_1)/\alpha_0$*, it holds that* $s_0 + s_1 \le \alpha_0 + \alpha_1$.

The choice of (c_0, c_1) considered in this proposition is known to be very accurate in most practical situations. It allows to control $s_0 + s_1$, which is the probability that the test makes a mistake.

Problem 14.18. Show that, although the procedure is designed for testing $\mathcal{H}_0^=$ versus $\mathcal{H}_1^=$, it applies to testing $\mathcal{H}_0^\le$ versus $\mathcal{H}_1^\ge$ above, meaning that if $s_0 \le \alpha_0$ and $s_1 \le \alpha_1$, then it also holds that

$$\mathbb{P}_\theta(L_n \ge c_1) \le \alpha_0, \quad \text{for all } \theta \le \theta_0,$$
$$\mathbb{P}_\theta(L_n \le c_0) \le \alpha_1, \quad \text{for all } \theta \ge \theta_1.$$

14.4.2 Experiments with optional stopping

In [190], Randi debunks a number of experiments claimed to exhibit paranormal effects. He recounts experiments where a self-proclaimed psychic tries to influence the outcome of a computer-generated sequence of coin tosses. Randi questions the validity of these experiments because the subject had the option of stopping or continuing the experiment at will.

Much more disturbing is the fact that such strategies are commonly employed by scientists. Indeed, in [137, 256], we learn that psychologists doing academic research routinely stop or continue to collect data based on the data collected up to that point. (It is reasonable to assume that psychologists are not unique in this habit and that this issue concerns all sciences, as we discuss further in Section 23.8.)

Below we argue that if optional stopping is allowed and not taken into account in the inference, then the resulting inference can be grossly incorrect. Thus, although lacking any formal training in statistics, Randi shows good statistical sense.

This is not the end of the story, however. If we know that optional stopping was allowed, it is still possible to make sensible use of the data. We discuss below a principled way to account for optional stopping. [73]

Experiment We place ourselves in the context of Bernoulli trials, although the same qualitative conclusions hold in general. Let $(X_i : i \ge 1)$ be iid Bernoulli with parameter θ. Suppose we want to test

$$\mathcal{H}_0 : \theta_* \le \theta_0.$$

[73] The setting is definitely non-standard, and therefore not typically discussed in textbooks. The solution expounded here may not be optimal in any way, but it is at least based on sound principles.

As opposed to a binomial experiment where the sample size is set beforehand, here we allow the stopping of the trials based on the trials observed so far. It is helpful here to imagine an experimenter who wants to provide the most evidence against the null hypothesis.

We say that the experiment includes *optional stopping* when the experimenter can stop the experiment at any moment and make that decision based on the result of previous tosses. The experimenter's strategy is modeled by a collection of functions

$$S_n : \{0,1\}^n \to \{\mathsf{stop}, \mathsf{continue}\}.$$

After observing the first n tosses, $x_1, \dots, x_n$, the experimenter decides to stop if $S_n(x_1, \dots, x_n) = \mathsf{stop}$, and otherwise continues. Let N denote the number of tosses until the trials are stopped,

$$N(\boldsymbol{x}) := \inf\big\{n : S_n(x_1, \dots, x_n) = \mathsf{stop}\big\},$$

where $\boldsymbol{x} := (x_i : i \geq 1)$.

When optional stopping is ignored We say that optional stopping has not been taken into account in the statistical analysis if, when $N = n$, the statistical analysis is performed based on the binomial experiment with sample size n. In effect, this means that the inference is performed as if the sample size had been predetermined.

Assume that the same test statistic considered earlier is used, meaning the total number of heads, $Y_n := \sum_{i=1}^n X_i$. When $N = n$, based on $x_1, \dots, x_n$ and $y_n := \sum_{i=1}^n x_i$, and optional stopping is not taken into account, the reported 'p-value' is

$$g_n(y_n) := \mathbb{P}_{\theta_0}(Y_n \geq y_n). \tag{14.20}$$

Although this would be referred to as a 'p-value' in a report describing the experiment, it is not a bona fide p-value in general, in the sense that it does not satisfy (12.22), as we argue below.

Remember that we have in mind an experimenter that wants to provide strong evidence against the null hypothesis. If a p-value below some predetermined $\alpha > 0$ is deemed sufficiently small for that purpose, the experimenter can simply stop when this happens. Formally, this corresponds to using the strategy

$$S_n(x_1, \dots, x_n) = \begin{cases} \mathsf{stop} & \text{if } g_n(x_1 + \cdots + x_n) \leq \alpha, \\ \mathsf{continue} & \text{otherwise.} \end{cases}$$

With this strategy, 'significance' can be achieved at any prescribed α, and so regardless of whether the null hypothesis is true or not. Indeed, suppose that $\theta_* = \theta_0$, so

the null hypothesis is true and $(X_i : i \geq 1)$ are iid Bernoulli with parameter θ_0. Then

$$\min_{k \leq n} g_k(Y_k) \xrightarrow{\mathrm{P}} 0, \quad \text{as } n \to \infty. \tag{14.21}$$

Problem 14.19. Show that (14.21) holds when, in probability,

$$\limsup_{n \to \infty} \frac{Y_n - n\theta_0}{\sqrt{n}} = \infty. \tag{14.22}$$

[Apply Chebyshev's inequality.] In turn, show that (14.22) holds using Problem 9.34.

We conclude that, with a proper choice of optional stopping strategy, the experimenter can make the 'p-value' (14.20) as small as desired, regardless of whether the null hypothesis is true or not, as long as he can continue the experiment at will. (This is clearly problematic.)

Taking optional stopping into account Suppose we are not willing to assume anything about the optional stopping strategy. We construct a test statistic that yields a valid p-value regardless of the strategy being used. Crucially, we assume that the data have not been tampered with, so that the sequence is a genuinely realization of Bernoulli trials. We can then ask the question: if the null hypothesis were true, how surprising would it be if a significance of $\alpha_n := g_n(y_n)$ were achieved after $N = n$ trials, given that the experimenter had the option of stopping the process at any point before that?

Suppose that we know the significance level, $\alpha \in (0, 1)$, that the experimenter wants to achieve. The idea, then, is to consider the following test statistic

$$T(\boldsymbol{x}) := \inf \big\{ k : g_k(x_1 + \cdots + x_k) \leq \alpha \big\}.$$

Noting that small values of T weigh against the null, and assuming that n trials were performed before stopping, the resulting p-value is

$$\mathbb{P}_{\theta_0}(T \leq n). \tag{14.23}$$

Problem 14.20. Show that this is indeed a valid p-value (in the sense of (12.22)) regardless of what optional stopping strategy was employed.

Problem 14.21. Suppose that α was set at 1%. How small would n have to be in order for the p-value (14.23) to be below 2%? (Assume $\theta_0 = 1/2$.) Perform some numerical experiments in R to answer the question.

14.5 Additional problems

Problem 14.22 (Proportion test)**.** In the context of the binomial experiment of Example 12.2, a *proportion test* can be defined based on the fact that the sample proportion is approximately normal (Theorem 5.4). Based on this, obtain a p-value for testing $\theta_* \leq \theta_0$. (Note that the p-value will only be valid in the large-sample limit.) [In R, this test is implemented in the prop.test function.]

Problem 14.23 (A comparison of confidence intervals)**.** A numerical comparison of various confidence intervals for a proportion is presented in [175]. Perform simulations to reproduce Table I, rows 1, 3, and 5, in the article. [Of course, due to randomness, the numbers resulting from your numerical simulations will be a little different.]

Problem 14.24 (ESP experiments)**.** Suppose that a person claiming to have psychic abilities is studied by some scientist. The person claims to be able to make a coin land heads more often than it would under normal circumstances without even touching it. The scientist builds a machine that can toss a coin any number of times. The mechanical system has been properly tested beforehand to ascertain that the coin lands heads with probability sufficiently close to 1/2. This system is kept out of reach of the subject at all times.

Suppose that it is agreed beforehand that the coin would be tossed 200 times. It is usually much safer to stick to the design that was chosen before the experiment begins. That said, it is still possible to perform a valid statistical analysis even if the design is changed in the course of the experiment. [74]

Below are a few situations. For each situation, explain how the scientist could accommodate the subject's request and still perform a sensible statistical analysis.

(i) In the course of the experiment, the subject insists on stopping after 130 tosses, claiming to be tired and unable to continue.

(ii) In the course of the experiment, the subject insists on continuing past 200 tosses, claiming that the first few dozen tosses only served as warm up.

(iii) In the course of the experiment, multiple times, the subject insists on not counting a particular trial claiming that he was not 'feeling it'.

[The first two situations can be handled as in Section 14.4.2. The last situation is somewhat different, but the same kind of reasoning will prove fruitful.]

Problem 14.25 (More on ESP experiments)**.** Detail the

[74] Doing so is sometimes necessary, for example, in clinical trials, although a well-designed trial will include a protocol for early termination.

calculations (implicitly) done in the "Feedback Experiments" section of the paper [58].

CHAPTER 15

MULTIPLE PROPORTIONS

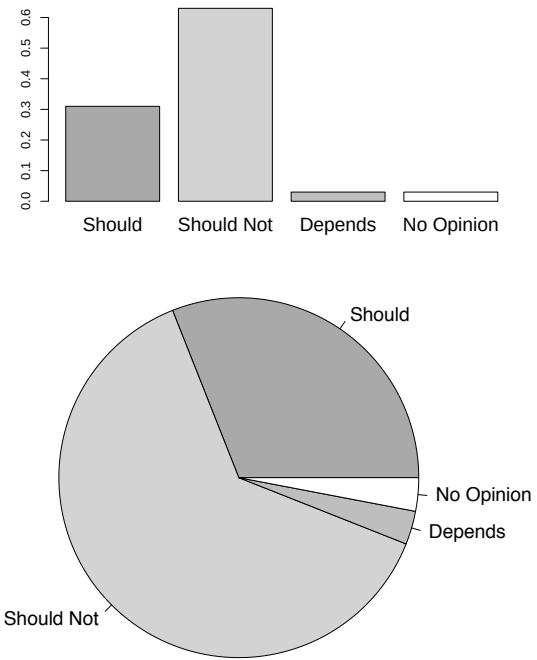
When a die with m faces is rolled, the result of each trial can take one of m possible values. The same is true in the context of an urn experiment, when the balls in the urn are of m different colors. Such models are broadly applicable. Indeed, even 'yes/no' polls almost always include at least one other option like 'not sure' or 'no opinion'. See Table 15.1 for an example. These data can be plotted, for instance, as a *bar chart* or a *pie chart*, as shown in Figure 15.1.

Table 15.1: Washington Post - ABC News poll of 1003 adults in the US (March 7-10, 2012). "Do you think a political leader should or should not rely on his or her religious beliefs in making policy decisions?"

Should	Should Not	Depends	No Opinion
31%	63%	3%	3%

Another situation where discrete variables arise is when two or more coins are compared in terms of their chances

Figure 15.1: A bar chart and a pie chart of the data appearing in Table 15.1.



of landing heads, or more generally, when two or more (otherwise identical) dice are compared in terms of their chances of landing on a particular face. In terms of urn experiments, the analog is a situation where balls are drawn from multiple urns. This sort of experiments can be used to model clinical trials where several treatments are compared and the outcome is dichotomous. See Table 15.2 for an example. These data can be plotted, for instance, as a *segmented bar chart*, as shown in Figure 15.2.

Table 15.2: The study [189] examined the impact of supplementing newborn infants with vitamin A on early infant mortality. This was a randomized, double-blind trial, performed in two rural districts of Tamil Nadu, India, where newborn infants (11619 in total) received either vitamin A or a placebo. The primary response was mortality at 6 months.

	Death	No Death
Placebo	188	5645
Vitamin A	146	5640

When the coins are tossed together, or when the dice are rolled together, we might want to test for independence. Although not immediately apparent, we will see that, depending on the design of an experiment, the same

question can be modeled as an experiment comparing several dice or testing for their independence.

15.1 Multinomial distributions

In this entire chapter we will talk about an experiment where one or several dice are rolled. We consider that a die can have any number $m \geq 2$ of faces, thus generalizing a coin. Just like a binomial distribution arises when a coin is tossed a predetermined number of times, the multinomial distribution arises when a die is rolled a predetermined number of times, say n. We assume that the die has faces with distinct labels, say $1, \ldots, m$, and for $s \in \{1, \ldots, m\}$, we let θ_s denote the probability that in a given trial the die lands on s. The outcome of the experiment is of the form $\omega = (\omega_1, \ldots, \omega_n)$, where $\omega_i = s$ if the ith roll resulted in face s. We assume that the rolls are independent.

Let $Y_1, \ldots, Y_m$ denote the *counts*

$$Y_s(\omega) := \#\big\{i \in \{1, \ldots, n\} : \omega_i = s\big\}. \tag{15.1}$$

Note that $Y_s \sim \mathrm{Bin}(n, \theta_s)$. Under the stated circumstances, the random vector of counts $(Y_1, \ldots, Y_m)$ is said to have the *multinomial distribution* with parameters $(n, \theta_1, \ldots, \theta_m)$.

Remark 15.1. There is some redundancy in the vector of counts, since $Y_1 + \cdots + Y_m = n$, and also in the parameterization, since $\theta_1 + \cdots + \theta_m = 1$. Except for that redundancy, the multinomial distribution with parameters (n, θ_1, θ_2) (where necessarily $\theta_2 = 1 - \theta_1$) corresponds to the binomial distribution with parameters (n, θ_1).

Proposition 15.2. *The multinomial distribution with parameters $(n, \boldsymbol{\theta})$, where $\boldsymbol{\theta} := (\theta_1, \ldots, \theta_m)$, has probability mass function*

$$f_{\boldsymbol{\theta}}(y_1, \ldots, y_m) := \frac{n!}{y_1! \cdots y_m!} \theta_1^{y_1} \cdots \theta_m^{y_m}, \tag{15.2}$$

supported on the m-tuples of integers $y_1, \ldots, y_m \geq 0$ satisfying $y_1 + \cdots + y_m = n$.

Problem 15.3. Suppose that there are n balls, with y_s balls of color s, and that except for their color the balls are indistinguishable. The balls are to be placed in bins numbered $1, \ldots, n$. Show that there are

$$\frac{(y_1 + \cdots + y_m)!}{y_1! \cdots y_m!}$$

different ways of doing so. Then use this combinatorial result to prove Proposition 15.2.

Problem 15.4. Show that $(Y_1, \dots, Y_m)$ is sufficient for $(\theta_1, \dots, \theta_m)$. Note that, when focusing the counts rather than the trials themselves, all that is lost is the order of the trials, which at least intuitively is not informative since the trials are assumed to be iid.

Problem 15.5. Show that $(Y_1/n, \dots, Y_m/n)$ is the maximum likelihood estimator for $(\theta_1, \dots, \theta_m)$.

In what follows, we will let $y_s = Y_s(\omega)$, and $\hat{\theta}_s := y_s/n$, which are the observed counts and observed averages, respectively.

15.2 One-sample goodness-of-fit testing

Various questions may arise regarding the parameter vector $\boldsymbol{\theta}$. These can be recast in the context of multiple testing (Chapter 20). We present here a more classical treatment focusing on *goodness-of-fit testing*, where the central question is whether the underlying distribution is a given distribution or, said differently, how well a given distribution fits the data. In detail, given $\boldsymbol{\theta}_0 = (\theta_{0,1}, \dots, \theta_{0,m})$, we are interested in testing

$$\mathcal{H}_0 : \boldsymbol{\theta}^* = \boldsymbol{\theta}_0,$$

where, as before, $\boldsymbol{\theta}^* = (\theta_1^*, \dots, \theta_m^*)$ denotes the true value of the parameter.

We first try a likelihood approach. In the variant (12.16), the likelihood ratio is here given by

$$\frac{\hat{\theta}_1^{y_1} \cdots \hat{\theta}_m^{y_m}}{\theta_{0,1}^{y_1} \cdots \theta_{0,m}^{y_m}},$$

and taking the logarithm, this becomes

$$\sum_{s=1}^{m} y_s \log\left(\frac{y_s}{n\theta_{0,s}}\right), \tag{15.3}$$

and dividing by n, this becomes

$$\sum_{s=1}^{m} \hat{\theta}_s \log\left(\frac{\hat{\theta}_s}{\theta_{0,s}}\right). \tag{15.4}$$

Remark 15.6 (Observed and expected counts)**.** While $y_1, \dots, y_m$ are the *observed counts*, $n\theta_{0,1}, \dots, n\theta_{0,m}$ are often referred to as the *expected counts*. This is because $\mathbb{E}_{\boldsymbol{\theta}_0}(Y_s) = n\theta_{0,s}$.

15.2.1 Monte Carlo p-value

In general, if T denotes a test statistic whose large values weigh against the null hypothesis, after observing a value t of this statistic, the p-value is $\mathbb{P}_{\boldsymbol{\theta}_0}(T \geq t)$. This p-value can be estimated by Monte Carlo simulation on a computer

(as seen in Section 10.1). The general procedure is detailed in Algorithm 4. (A variant of the algorithm consists in directly generating values of the test statistic under its null distribution.)

Algorithm 4 Monte Carlo p-value

Input: data ω, test statistic T, null distribution $\mathbb{P}_0$, number of Monte Carlo samples B

Output: an estimate of the p-value

Compute $t = T(\omega)$
For $b = 1, \dots, B$
  generate ω_b from $\mathbb{P}_0$
  compute $t_b = T(\omega_b)$
Return

$$\widehat{\mathsf{pv}}_{\mathrm{MC}} := \frac{\#\{b : t_b \geq t\} + 1}{B + 1}. \tag{15.5}$$

Proposition 15.7. *The Monte Carlo p-value* (15.5) *is itself a valid p-value in the sense of* (12.22).

Problem 15.8. Prove this result using the conclusions of Problem 8.63.

Problem 15.9. In R write a function that takes as input the vector of observed counts, the null parameter vector, and a number of Monte Carlo replicates, and returns an estimate of the p-value above. [Use the function rmultinom to generate Monte Carlo counts under the null.] Compare your function with the built-in chisq.test function.

15.3 Multi-sample goodness-of-fit testing

In Section 15.2 we assumed that we were provided with a null distribution and tasked with determining how well that distribution fits the data.

In some other situations, a null distribution is not available, as in clinical trials where the efficacy of a treatment is compared to an existing treatment or a placebo, such as in the example of Table 15.2. In other situations, more than two groups are to be compared. A basic question then is whether these groups of observations were generated by the same distribution. The difference here with the setting of Section 15.2 is that this hypothesized common distribution is not given.

An abstract model for this setting is that of an experiment involving g dice, with the jth die rolled n_j times, all rolls being independent. As before, each die has m faces labeled $1, \dots, m$. Let $\boldsymbol{\theta}_j$ denote the probability vector of the jth die. Our goal is to test the following null

hypothesis:

$$\mathcal{H}_0 : \boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_g \ .$$

This is sometimes referred to as *testing for homogeneity*.

Let $\omega_{ij} = s$ if the ith roll of the jth die results in s, so that $\omega = (\omega_{ij})$ are the data, and define the (per group) counts

$$Y_{sj}(\omega) = \#\{i : \omega_{ij} = s\}.$$

Problem 15.10. Show that these counts are jointly sufficient.

Problem 15.11. Show that (Y_{sj}/n_j) is the maximum likelihood estimator for (θ_{sj}).

The total sample size (meaning the total number of rolls) is $n := \sum_{j=1}^{s} n_j$, and the total counts are defined as

$$Y_s := \sum_{j=1}^{g} Y_{sj}. \tag{15.6}$$

In what follows, we will let $y_{sj} := Y_{sj}(\omega)$ and $\hat{\theta}_{sj} := y_{sj}/n_j$, as well as $y_s := Y_s(\omega) = \sum_j y_{sj}$ and $\hat{\theta}_s := y_s/n$, and $\hat{\boldsymbol{\theta}}_j := (\hat{\theta}_{1j}, \ldots, \hat{\theta}_{mj})$ and $\hat{\boldsymbol{\theta}} := (\hat{\theta}_1, \ldots, \hat{\theta}_m)$.

15.3.1 Likelihood ratio

Because of independence, the likelihood of all the observations combined is just the product of the likelihoods, one for each die (see (15.2)).

Problem 15.12.

(i) Prove that, without any constraints on the parameters, the likelihood is maximized at $(\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_g)$.

(ii) Prove that, under the constraint that $\boldsymbol{\theta}_1 = \cdots = \boldsymbol{\theta}_g$, this is maximized at $(\hat{\boldsymbol{\theta}}, \ldots, \hat{\boldsymbol{\theta}})$.

(iii) Deduce that the likelihood ratio is given by (after some simplifications)

$$\frac{\prod_{j=1}^{g} \prod_{s=1}^{m} \hat{\theta}_{sj}^{y_{sj}}}{\prod_{s=1}^{m} \hat{\theta}_s^{y_s}} = \prod_{j=1}^{g} \prod_{s=1}^{m} \left(\frac{\hat{\theta}_{sj}}{\hat{\theta}_s} \right)^{y_{sj}} .$$

Taking the logarithm, this becomes

$$\sum_{j=1}^{g} \sum_{s=1}^{m} y_{sj} \log \left(\frac{y_{sj}}{n_j y_s / n} \right), \tag{15.7}$$

and dividing by n, this becomes

$$\sum_{j=1}^{g} \sum_{s=1}^{m} \hat{\theta}_{sj} \log \left(\frac{\hat{\theta}_{sj}}{\hat{\theta}_s} \right). \tag{15.8}$$

Remark 15.13 (Estimated expected counts)**.** The (y_{sj}) are the observed counts. Expected counts are not available since the common null distribution is not given. Nonetheless, it can be estimated. Indeed, under the null hypothesis,

$$\mathbb{E}_{\boldsymbol{\theta}}(Y_{sj}) = n_j \theta_s,$$

and we can estimate this by plugging in $\hat{\theta}_s$ in place of θ_s, leading to the following *estimated expected counts*

$$\mathbb{E}_{\boldsymbol{\theta}}(Y_{sj}) \approx n_j \hat{\theta}_s.$$

15.3.2 Bootstrap p-value

Now that we have derived the LR, we need to compute or estimate the corresponding p-value. In Section 15.2 this was done by Monte Carlo simulation, made possible by the fact that the null distribution was provided. Here the null distribution (that generated all the samples) is not provided.

We already encountered this issue in Remark 15.13, where it was noted that the expected counts were not available. This was addressed by simply replacing them by estimates. In the same way, we can estimate the (entire) null distribution, by again plugging in $\hat{\boldsymbol{\theta}}$ in place of the unknown $\boldsymbol{\theta}$ (which parameterizes the null distribution). The idea then is to estimate the p-value by Monte Carlo simulation, as before, but now using the estimated null distribution to generate the samples. This process, of performing Monte Carlo simulations based on an estimated distribution, is generally called a *bootstrap*. The resulting samples are typically called *bootstrap samples*.

Remark 15.14. Because it relies on an estimate for the null distribution, as opposed to the exact null distribution needed for Monte Carlo simulation, the bootstrap p-value is not exactly valid. That said, it is exactly valid in the large-sample limit.

Problem 15.15. In R write a function that takes as input the list of observed counts as a g-by-m matrix of counts and the number of bootstrap samples to be drawn, and returns an estimate of the p-value. Apply your function to the dataset of Table 15.2.

15.4 Completely randomized experiments

The data presented in Table 15.2 can be analyzed using the methodology for comparing groups presented in Section 15.3, as done in Problem 15.15. We present a different perspective which leads to a different analysis. It is reassuring to know that the two analyses will yield similar results as long as the group sizes are not too small.

The typical null hypothesis in such a situation is that the treatment and the placebo are equally effective. We describe a model where inference can only be done for the group of subjects — while a generalization to a larger population would be contingent on this sample of individuals being representative of the said population.

Suppose that the group sizes are n_1 and n_2, for a total sample size of $n = n_1 + n_2$. The result of the experiment can be summarized in a table of counts, Table 15.3, often called a *contingency table*.

If $z = z_1 + z_2$ denotes the total number of successes, in the present model it is assumed to be deterministic since we are drawing inferences on the group of subjects in the study. What is random is the group labeling, and if treatment and placebo truly have the same effect on these individuals, then the group labeling is completely arbitrary. Thus this is the null hypothesis to be tested.

There is no model for the alternative, so we cannot derive the LR, for example. However, it is fairly clear what kind of test statistic we should be using. In fact, a good option is the same statistic (15.7), which in the context of Table 15.3 takes the form

$$z_1 \log\left(\frac{nz_1}{n_1 z}\right) + (n_1 - z_1)\log\left(\frac{n(n_1 - z_1)}{n_1(n-z)}\right) + z_2 \log\left(\frac{nz_2}{n_2 z}\right) + (n_2 - z_2)\log\left(\frac{n(n_2 - z_2)}{n_2(n-z)}\right).$$

Another option is the odds ratio, which after applying a log transformation takes the form

$$\log\left(\frac{z_1}{n_1 - z_1}\right) - \log\left(\frac{z_2}{n_2 - z_2}\right).$$

More generally, consider a randomized experiment where the subjects are assigned to one of g treatment groups and the response can take m possible values. With the notation of Section 15.3, the jth group is of size n_j, for

Table 15.3: A prototypical contingency table summarizing the result of a completely randomized experiment with two groups and two possible outcomes.

	Success	Failure	Total
Group 1	z_1	$n_1 - z_1$	n_1
Group 2	z_2	$n_2 - z_2$	n_2
Total	z	$n - z$	n

a total sample size of $n := n_1 + \cdots + n_g$. In the null model, the total counts (15.6) are here taken to be deterministic (while there are random in Section 15.3), while the group labeling is random. The test statistic of choice remains (15.7).

Problem 15.16. Write down the contingency table (in general form) using the notation of Section 15.3.

15.4.1 Permutation p-value

Regardless of the test statistic that is chosen, the corresponding p-value is obtained under the null model. Since the group sizes are set, the null model amounts to permuting the labels. This procedure is an example of rerandomization testing, developed further in Section 22.1.1.

Let Π denote the set of all permutations of the labels. There are

$$|\Pi| = \frac{n!}{n_1! \cdots n_g!}$$

such permutations in the present setting. Importantly, we permute the labels placed on the rolls, which then yield new counts. (We do not permute the counts.) Let T be a test statistic whose large values provide evidence against the null hypothesis. We let t denote the observed value of T and, for a permutation $\pi \in \Pi$, we let t_π denote the value of T applied to the corresponding permuted data. Then the *permutation p-value* is defined as

$$\mathsf{pv}_{\text{perm}} := \frac{\#\{\pi : t_\pi \geq t\}}{|\Pi|}. \tag{15.9}$$

Proposition 15.17. *The permutation p-value* (15.9) *is a valid p-value in the sense of* (12.22).

Problem 15.18. Prove this result using the conclusions of Problem 8.63.

Monte Carlo estimation Unless the group sizes are very small, $|\Pi|$ is impractically large, and this leads one to estimate the p-value by sampling permutations uniformly at random from Π. This may be called a *Monte Carlo permutation p-value*. The general procedure is detailed in Algorithm 5.

Proposition 15.19. *The Monte Carlo permutation p-value* (15.10) *is a valid p-value in the sense of* (12.22).

Problem 15.20. Prove this result using the conclusions of Problem 8.63.

Problem 15.21. In R, write a function that implements Algorithm 5. Apply your function to the data in Table 15.2.

Algorithm 5 Monte Carlo Permutation p-value

Input: data ω, test statistic T, group Π of permutations that leave the null invariant, number of Monte Carlo samples B

Output: an estimate of the p-value

Compute $t = T(\omega)$
For $b = 1, \dots, B$
  draw π_b uniformly at random from Π
  permute ω according to π_b to get ω_b
  compute $t_b = T(\omega_b)$
Return

$$\widehat{\mathsf{pv}}_{\text{perm}} := \frac{\#\{b : t_b \geq t\} + 1}{B + 1}. \tag{15.10}$$

Remark 15.22 (Conditional inference)**.** This proposition holds true also in the setting of Section 15.3. The use of a permutation p-value there is an example of conditional inference, which is discussed in Section 22.1.

15.5 Matched-pairs experiments

Consider a randomized matched-pairs design where two treatments are compared (Section 11.2.5). The outcome is binary ('success' or 'failure'). The sample is of the form $(\omega_{11}, \omega_{21}), \dots, (\omega_{n1}, \omega_{n2})$, where ω_{ij} is the response for the subject in pair i that received Treatment j, with $\omega_{ij} = 1$ indicating success. If no other information on the subjects is taken into account in the analysis, the data can be summarized in a 2-by-2 table contingency table (Table 15.4) displaying the counts where $y_{st} := \#\{i : (\omega_{i1}, \omega_{i2}) = (s, t)\}$.

The iid assumption At this point we cannot claim that the counts are jointly sufficient. This is the case, however, in a situation where the pairs can be assumed to be sampled uniformly at random from a population. In that case, the pairs can be taken to be iid and we may define θ_{st} as the probability of observing the pair (s, t). The treatment (one-sided) effect is defined as $\theta_{10} - \theta_{01}$.

Table 15.4: A prototypical contingency table summarizing the result of a matched-pairs experiment with two possible outcomes.

		Treatment B	
		Success	Failure
Treatment A	Success	y_{11}	y_{10}
	Failure	y_{01}	y_{00}

The null hypothesis of no treatment effect is $\theta_{10} = \theta_{01}$.

It is rather natural to ignore the pairs where the two subjects responded in the same way and base the inference on the pairs where the subjects responded differently, which leads to rejecting for large values of $Y_{10} - Y_{01}$ while conditioning on (Y_{11}, Y_{00}). After all, $(Y_{10} - Y_{01})/n$ is unbiased for $\theta_{10} - \theta_{01}$. This is the *McNemar test* [165] and it is known to be uniformly most powerful among unbiased tests (UMPU) in this situation.

Problem 15.23. Argue that rejecting for large values of $Y_{10} - Y_{01}$ given (Y_{11}, Y_{00}) is equivalent to rejecting for large values of Y_{10} given (Y_{11}, Y_{00}). Further, show that given $(Y_{11}, Y_{00}) = (y_{11}, y_{00})$, Y_{10} has the binomial distribution with parameters $k := n - y_{11} - y_{00}$ and $p := \theta_{10}/(\theta_{10} + \theta_{01})$. Conclude that the McNemar test reduces to testing $p = 1/2$ in a binomial experiment with parameters (k, p).

Problem 15.24. Is the McNemar test the likelihood ratio test in the present context?

Remark 15.25 (Observational studies)**.** Although we worked in the context of a randomized experiment, the test may be applied in the context of an observational study with the caveat that the conclusion is conservatively understood as being in terms of association instead of causality.

Beyond the iid assumption In some situations it may not be realistic to assume that the pairs constitute a representative sample from a population. Even then, a permutation approach remains valid due to the initial randomization. The key observation is that, if there is no treatment effect, then ω_{i1} and ω_{i2} are exchangeable by design. The idea then is to condition on the observed ω_{ij} and permute within each pair. Then, under the null hypothesis, any such permutation is (conditionally) equally likely, and this is exploited in the derivation of a p-value. This procedure is again an example of re-randomization testing (Section 22.1.1).

In more detail, a permutation in the present context transforms the (observed) data,

$$(\omega_{11}, \omega_{12}), \ldots, (\omega_{n1}, \omega_{n2}),$$

into

$$(\omega_{1\pi_1(1)}, \omega_{1\pi_1(2)}), \ldots, (\omega_{n\pi_n(1)}, \omega_{n\pi_n(2)}),$$

with $(\pi_i(1), \pi_i(2)) = (1, 2)$ or $= (2, 1)$. Let Π denote the class of $\pi = (\pi_1, \ldots, \pi_n)$ where each π_i is a permutation of $\{1, 2\}$. Note that $|\Pi| = 2^n$.

Suppose we still reject for large values of $T := Y_{10} - Y_{01}$, which after all remains appropriate. The observed value of this statistic is $t := y_{10} - y_{01}$. Let t_π denote the value of

this statistic computed on the data permuted by applying $\pi \in \Pi$. Then the *permutation p-value* is defined as

$$\mathsf{pv} := \frac{\#\{\pi \in \Pi : t_\pi \geq t\}}{|\Pi|}.$$

This is a valid p-value in the sense of (12.22).

Computing this p-value may be challenging, as there are many possible permutations and one would in principle have to consider every single one of them. Luckily, this is not necessary.

Problem 15.26. Find an efficient way of computing the p-value.

Problem 15.27. Show that, in fact, this permutation test is equivalent to the McNemar test.

15.6 Fisher's exact test

Fisher [13] describes in [86] a now famous experiment meant to illustrate his concept of null hypothesis. The setting is that of a lady who claims to be able to distinguish whether milk or tea was added to the cup first. To test her professed ability, she is given 8 cups of tea, in four of which milk was added first. With full information on how the experiment is being conducted, she is asked to choose 4 cups where she believes milk was poured first. The resulting counts, reproduced from Fisher's original account, are displayed in Table 15.5.

Table 15.5: The columns are labeled by the liquid that was poured first (milk or tea) and the rows are labeled by the lady's guesses.

	Truth	
Lady's guess	Milk	Tea
Milk	3	1
Tea	1	3

The null hypothesis is that of no association between the true order of pouring and the woman's guess, while the alternative is that of a positive association. [75] In particular, under the null, the lady is purely guessing and the lady's guesses are independent of the truth. This gives the null distribution, which leads to a permutation p-value.

Remark 15.28. Thus a p-value is obtained by permutation, exactly as in Section 15.4, because here too the total counts are all fixed. However, the situation is not exactly

[75] Some statisticians might prefer an alternative of no association, which would lead to a two-sided test.

the same. The difference is subtle: here, all the totals are fixed by design; there, the totals are fixed because of the focus on the subjects in the study.

In such a 2×2 setting, it is possible to test for a positive association. Indeed, first note that, since the margin totals are fixed (all equal to 4), the number in the top-left corner ('milk', 'milk') determines all the others, so it suffices to consider this number (which is obviously sufficient). Now, clearly, a large value of that number indicates a positive association. This leads us to rejecting for large values of this statistic.

Consider a general 2×2 setting, where there are n cups, with k of them receiving milk first. The lady is to select k cups that she believes received the milk first. Then the contingency table would look something like this

	Truth		
Lady's guess	Milk	Tea	Total
Milk	y_{11}	y_{12}	k
Tea	y_{21}	y_{22}	$n-k$
Total	k	$n-k$	n

The test statistic is Y_{11}, whose large values weigh against the null, and the resulting test is often referred to as *Fisher's exact test*.

Problem 15.29. Prove that, under the null, Y_{11} is hypergeometric with parameters $(k, k, n-k)$.

Thus computing the p-value can be done exactly, without having to enumerate all possible permutations.

Problem 15.30. Derive the p-value for the original experiment in closed form, and confirm your answer numerically using R.

R corner. Fisher's exact test is implemented in the function fisher.test.

Remark 15.31. Fisher's exact test is applicable in the context of Section 15.4, and doing so is an example of conditional inference (Section 22.1).

15.7 Association in observational studies

Consider the poll summarized in Table 15.6 and depicted in Figure 15.2. The description of the polling procedure that appears on the website, and additional practical considerations, lead one to believe that the interviewees were selected without a priori knowledge of their political party affiliation. We will assume this is the case.

It is safe to guess that one of the main reasons for collecting these data is to determine whether there is an *association* between party affiliation and views on climate

Table 15.6: New York Times - CBS News poll of 1030 adults in the US (November 18-22, 2015). "Do you think the US should or should not join an international treaty requiring America to reduce emissions in an effort to fight global warming?"

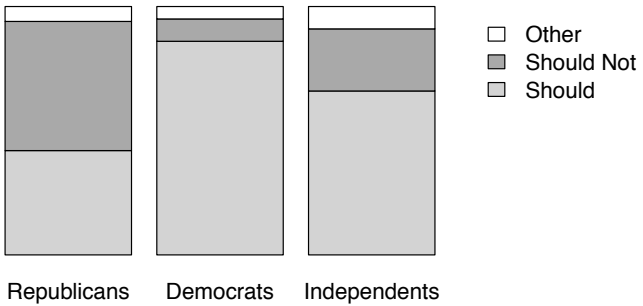
	should	should not	other
Republicans	42%	52%	6%
Democrats	86%	9%	5%
Independents	66%	25%	9%

change, and the steps that the US Government should take to address this issue if any. (The entire poll questionnaire, not shown here, includes other questions related to climate change.)

Formally, a lack of association in such a setting is modeled as *independence*. The variables here are A for 'party membership', equal to either 'Republican', 'Democrat', or 'Independent'; and B for 'opinion', equal to either 'should', 'should not', or 'other'.

Remark 15.32 (Factors)**.** In statistics, a categorical variable is often called a *factor* and the values it takes are called *levels*. Thus A is a factor with levels {'Republican', 'Democrat', 'Independent'}.

Figure 15.2: A segmented bar chart of the data appearing in Table 15.6.



The raw data here is of the form $\{(a_i, b_i) : i = 1, \ldots, n\}$, where $n = 1030$ is the sample size. Table 15.6 provides the percentages. We are told that there were 254 republicans, 304 democrats, and 472 independents. With this information, we can recover the table of counts up to rounding error. See Table 15.7.

An abstract model for the present setting is that of an experiment where two dice are rolled *together* n times. Die A has faces labeled $1, \ldots, m_a$, while Die B has faces labeled $1, \ldots, m_b$. (As before, the labeling by integers is for convenience, as the faces could be labeled by any other symbols.) The outcome of the experiment is $\omega =$

Table 15.7: The table of counts corresponding to the poll summarized in Table 15.6.

	should	should not	other
Republicans	107	132	15
Democrats	261	27	15
Independents	312	118	42

$((a_1, b_1), \ldots, (a_n, b_n))$. Define the cross-counts as

$$y_{st} := Y_{st}(\omega) := \#\big\{i : (a_i, b_i) = (s, t)\big\}.$$

Assuming the rolls are iid, which we do, these counts are jointly sufficient, and have the multinomial distribution with parameters n and (θ_{st}), where θ_{st} is the probability that a roll results in (s, t). These counts are often displayed in a contingency table.

If (A, B) denotes the result of a roll, the task is *testing for the independence* of A and B. Under (θ_{st}), A has (marginally) the multinomial distribution with parameters n and (θ^a_s), while B has (marginally) the multinomial distribution with parameters n and (θ^b_s), where

$$\theta^a_s := \sum_{t=1}^{m_b} \theta_{st} \ , \quad \theta^b_t := \sum_{s=1}^{m_a} \theta_{st} \ .$$

The null hypothesis of independence can be formulated as

$$\mathcal{H}_0 : \theta_{st} = \theta^a_s \theta^b_t,$$
$$\text{for all } s = 1, \ldots, m_a \text{ and all } t = 1, \ldots, m_b.$$

Remark 15.33. Contrast the present situation with that of Section 15.4, where only the null distribution is modeled.

15.7.1 Likelihood ratio

Recall that (y_{st}) denotes the observed counts. Based on these, define the proportions $\hat{\theta}_{st} = y_{st}/n$, the marginal counts

$$y^a_s = \sum_{t=1}^{m_b} y_{st}, \quad y^b_t = \sum_{s=1}^{m_a} y_{st}, \tag{15.11}$$

and the corresponding marginal proportions $\hat{\theta}^a_s = y^a_s/n$ and $\hat{\theta}^b_t = y^b_t/n$.

Problem 15.34. Show that the LR in the variant (12.16) is equal to

$$\prod_{s=1}^{m_a} \prod_{t=1}^{m_b} \left(\frac{\hat{\theta}_{st}}{\hat{\theta}^a_s \hat{\theta}^b_t} \right)^{y_{st}},$$

Taking the logarithm, this becomes

$$\sum_{s=1}^{m_a} \sum_{t=1}^{m_b} y_{st} \log \left(\frac{y_{st}}{y^a_s y^b_t / n} \right), \tag{15.12}$$

and dividing by n, this becomes

$$\sum_{s=1}^{m_a}\sum_{t=1}^{m_b}\hat{\theta}_{st}\log\left(\frac{\hat{\theta}_{st}}{\hat{\theta}^a_s\hat{\theta}^b_t}\right).$$

Problem 15.35 (Estimated expected counts)**.** Justify calling $y^a_s y^b_t/n$ the *estimated expected countestimated expected counts* for (s,t).

15.7.2 Deterministic or random group sizes

Compare (15.7) and (15.12). They are *identical* as functions of the counts. This is rather surprising, perhaps shocking, given that the statistic is rather peculiar and the counts are, at first glance, quite different. Although in both cases the counts are organized in a table, in Section 15.3 they are indexed by (value, group), while in the present section they are indexed by (A value, B value).

This can be explained by viewing 'group' in the setting of Section 15.3 as a variable. Indeed, from this perspective the only difference between the two settings is that, in Section 15.3, the group sizes are predetermined, while here they are random. The fact that the likelihood ratios coincide can be explained by the fact that the group sizes are not informative.

However, despite the fact that the likelihood ratios are the same, and that the same test statistics can be used in both settings, the p-value is derived in (slightly) different ways.

15.7.3 Bootstrap p-value

In Section 15.3.2, we presented a form of bootstrap tailored to the situation there. The situation is a little different here since the group sizes are not predetermined and another form of bootstrap is more appropriate. In the end, however, these two methods will yield similar p-values as long as the sample sizes are not too small.

The motivation for using a bootstrap approach is the same. Indeed, if we were given the marginal distributions, (θ^a_s) and (θ^b_t), we would simply sample their product, as this is the null distribution in the present situation. However, these distributions are not available to us, but we have estimates, $(\hat{\theta}^a_s)$ and $(\hat{\theta}^b_t)$, and the bootstrap method consists in estimating the p-value by Monte Carlo simulation by repeatedly sampling from the corresponding product distribution.

Problem 15.36. In R, write a function which takes as input the matrix of observed counts and the number of bootstrap samples to be drawn, and returns an estimate of the p-value. Apply your function to the dataset of Table 15.7.

15.7.4 Extension to several variables

The narrative above focused on two discrete variables. It extends without conceptual difficulty to any number of discrete random variables. When there are k factors with $m_1, \ldots, m_k$ levels respectively, the counts are organized in a $m_1 \times \cdots \times m_k$ array.

In particular, the methodology developed here applies to testing whether the variables are mutually independent. Beyond that, when there are $k \geq 3$ variables, other questions can be considered, for example, whether the two variables are independent conditional on the remaining variables.

15.7.5 Simpson's paradox

In Problem 2.42 we saw that inequalities involving probabilities could be reversed when conditioning. This is a relatively common phenomenon in real life.

Example 15.37 (Berkeley admissions)**.** Consider the situation discussed in [22]. In 1973, the Graduate Division at the University of California, Berkeley, received a number of applications. Ignoring incomplete applications, there were 8442 male applicants of whom 3738 were admitted (44%), compared with 4321 female applicants of whom 1494 were admitted (35%). The difference is not only statistically highly significant but also substantial with a difference of almost 10% in the admission rate when comparing men and women. This appeared to be strong and damning evidence of gender bias on the part of the admission office(s) of this university.

However, a breakdown of admission rates by department [76] revealed a much more nuanced picture, where in fact few departments showed any significant difference in admission rates when comparing men and women, and among these departments about as many favored men as women. In the end, the authors of [22] concluded that there was little evidence for gender bias, and that the numbers could be explained by the fact that women tended to apply to departments with lower admission rates. [77]

Example 15.38 (Race and death-penalty)**.** Consider the following table, taken from [188], which examined the issue of race in death penalty sentences in murder cases in the state of Florida from 1976 through 1987:

[76] Graduate admissions at University of California, Berkeley rest with each department.

[77] The authors did not publish the entire dataset for reasons of privacy and university policy, but the published data are available in R as UCBAdmissions.

Defendant	Death (yes/no)
Caucasian	53/430
African-American	15/176

Thus, in the aggregate, Caucasians were more likely to be sentenced to the death penalty. However, after stratifying by the race of the victim, the table becomes:

Victim	Defendant	Death (yes/no)
Caucasian	Caucasian	53/414
Caucasian	African-American	11/37
African-American	Caucasian	0/16
African-American	African-American	4/139

Thus, in the disaggregate, African-Americans were more likely to be sentenced to the death penalty, particularly in cases where the victim was Caucasian.

Thus the analyst needs to use extreme caution when deriving causal inferences based on observational data, or any other situation where randomization was not properly applied, as this can easily lead to confounding. (In Example 15.37, a confounder is the department, while in Example 15.38, a confounder is the victim's race.)

15.8 Tests of randomness

Consider an experiment where the outcome is a sequence of symbols of length n denoted $\omega = (\omega_1, \ldots, \omega_n)$. In Section 15.2, we focused on the situation where this sequence is the result of drawing repeatedly from an unknown distribution which was the object of our interest.

We now turn to the question of whether the sequence was generated iid from some distribution. When this is the case, the sequence is said to be random, and procedures addressing this question are called *tests of randomness*.

Formally, suppose we know before observing the outcome that each ω_i belongs to some space $\Omega_\star$, so that $\omega \in \Omega := \Omega_\star^n$, which is the sample space. The question of interest is whether the distribution that generated ω, denoted $\mathbb{P}$, is the product of its marginals, and whether these marginals are all the same. Let $\mathcal{P}_0$ denote the class of iid distributions on Ω, meaning that $\mathbb{P} \in \mathcal{P}_0$ is of the form $\mathbb{P}_\star^{\otimes n}$ for some distribution $\mathbb{P}_\star$ on $\Omega_\star$. We want to test the null hypothesis that $\mathbb{P} \in \mathcal{P}_0$.

Example 15.39 (Binary setting)**.** In a binary setting where $\Omega_\star$ has cardinality 2 and thus can be taken to be $\{0, 1\}$ without loss of generality, $\mathcal{P}_0$ is the family of Bernoulli trials of length n.

IID-NESS VS INDEPENDENCE We are testing iid-ness and not independence, as we presuppose that the marginal distributions are all the same. In fact, testing for independence is ill-posed without further restricting the model. To see this, suppose the outcome is $\omega = (\omega_1, \ldots, \omega_n)$. This could have been the result of sampling from the point mass at ω, for which independence trivially holds, and the available data, ω, are clearly not enough to discard this possibility.

IID-NESS VS EXCHANGEABILITY $\mathbb{P}$ is exchangeable if it is invariant with respect to permutation, meaning that, for any permutation $\pi = (\pi_1, \ldots, \pi_n)$ of $(1, \ldots, n)$, $\mathbb{P}(\omega) = \mathbb{P}(\omega_\pi)$, where $\omega_\pi := (\omega_{\pi_1}, \ldots, \omega_{\pi_n})$. As we know, this property is more general than iid-ness, but it turns out that the available data, ω, are not sufficient to tell the two apart. This can be explained by de Finetti's theorem (Theorem 8.56). To directly argue this point, though, we place ourself in the binary setting of Example 15.39. Within that setting, consider the distribution $\mathbb{P}$ where, with probability 1/2, we generate an iid sequence of length n from Ber(1/4), while with probability 1/2, we generate an iid sequence of length n from Ber(3/4). Thus $\mathbb{P}$ here is not an iid distribution. However, the outcome will be a realization of an iid distribution — either $\text{Ber}(1/4)^{\otimes n}$ or $\text{Ber}(3/4)^{\otimes n}$.

Thus, what we can hope to test is *exchangeability*. (That said, to adhere to tradition, we will use 'randomness' in place of 'exchangeability'.) The tests that follow all take a conditional inference approach by conditioning on the values of the ω_i without regard to their order. This leads to obtaining their p-values by permutation.

There are many tests of randomness. We present a few here that are tailored to the discrete setting. Such tests are important for evaluating the accuracy of a generator of pseudo-random numbers. Examples include the *DieHard* suite of George Marsaglia, included and expanded in the *DieHarder* suite of Robert Brown [78], and some tests developed by the US National Institute of Standards and Technology (NIST) for the binary case [200]. Some of these tests are available in the RDieHarder package in R.

15.8.1 Tests based on runs

Some tests of randomness are based on runs, where a *run* is a sequence of identical symbols. Take the binary setting and consider the following outcome sequence (of length

[78] webhome.phy.duke.edu/ rgb/General/dieharder.php

$n = 20$ here) where there are 9 runs total:

$$\underbrace{1}\ \underbrace{000}\ \underbrace{1}\ \underbrace{0000}\ \underbrace{1111}\ \underbrace{0}\ \underbrace{1}\ \underbrace{00}\ \underbrace{111}$$

NUMBER OF RUNS TEST This test rejects for small values of the total number of runs. Intuitively, a small number of runs indicates less 'mixing'. The test dates back to Wald and Wolfowitz [243], who proposed the test for the purpose of two-sample goodness-of-fit testing (Section 17.3.5).

Problem 15.40. The conditional null distribution of this statistic is known in closed form in the binary setting. To derive this distribution, first consider the number of 0-runs. (There are 4 such runs in the sequence displayed above.) Derive its null distribution. Then use that to derive the null distribution of the total number of runs.

Problem 15.41. What kinds of alternatives do you expect this test to be powerful against?

LONGEST RUN TEST This test rejects for large values of the length of the longest run (equal to 4 in the sequence displayed above). Intuitively, the presence of a long run indicates less 'mixing'. The test is due to Mosteller [171].

Remark 15.42 (Erdős–Rényi Law)**.** The conditional null distribution of this statistic, denoted L_n, is not known in an amiable form, although it can be estimated by Monte Carlo permutation in practice. However, the asymptotic behavior of L_n under the unconditional null is well understood, at least in the binary setting. The first-order behavior was derived by Erdös and Rényi [77], which in the context of Bernoulli trials with parameter θ is given by

$$\frac{L_n}{\log n} \xrightarrow{\text{P}} \frac{1}{\log(1/\theta)}, \quad \text{as } n \to \infty.$$

(Although L_n does not have a limiting distribution, it has a family of limiting distributions [8].)

Problem 15.43. What kinds of alternatives do you expect this test to be powerful against?

15.8.2 TESTS BASED ON TRANSITIONS

The following class of tests are designed with Markov chain alternatives in mind. The simplest such test is based on counting transitions $a \to b$, meaning instances where $(\omega_i, \omega_{i+1}) = (a, b)$, where $a, b \in \Omega_\star$. The test consists in computing a test statistic for independence applied to the pairs

$$(\omega_1, \omega_2), (\omega_2, \omega_3), \ldots, (\omega_{n-1}, \omega_n),$$

and then obtaining a p-value by permutation (which is typically estimated by Monte Carlo, as usual).

Problem 15.44. In R, write a function that takes in the observed sequence and a number of Monte Carlo replicates, and outputs the p-value just described. Compare this procedure with the number-of-runs test in the binary setting.

Problem 15.45. In the binary setting, perform some numerical simulations to evaluate the power of this procedure against a distribution corresponding to starting at 0 or 1 with probability 1/2 each (which is the stationary distribution) and then running the Markov chain with the following transition matrix $n-1$ times

$$\begin{pmatrix} q & 1-q \\ 1-q & q \end{pmatrix}$$

(i) Show that the resulting distribution is exchangeable if and only if $q = 1/2$. (In fact, in that case the distribution is iid.)

(ii) Evaluate the power by applying the procedure of the previous problem to various settings: try $n \in \{10, 100, 1000\}$ and for each n choose a set of q in $[1/2, 1]$ that reveal a transition from powerless to powerful as q decreases towards 1. Repeat each setting 1000 times. Draw a power curve for each n. (Note that, as q approaches 0, the sequence is more and more mixed.)

Problem 15.46. The test procedures described here are based on first-order transitions, meaning of the form $a \to b$. How would test procedures based on second-order transitions look like? Implement such a procedure in R, and apply it to the setting of the previous problem.

Problem 15.47. Continuing with Problem 15.60, apply all the tests of randomness introduced in this section to test the exchangeability of the first 20000 digits of the number π.

15.9 Further topics

15.9.1 Pearson's approximations

Before the advent of computers, computing logarithms was not trivial. Karl Pearson [79] (1857 - 1936) suggested an approximation to the likelihood ratios (15.4) and (15.8) that can be computed using simpler calculations [180].

Take (15.4) for simplicity. Pearson's approximation is based on two facts:

[79] This is the father of the Egon Pearson [71].

(i) Under the null, $\hat{\theta}_s \to_{\mathrm{P}} \theta_{0,s}$ as the sample size increases (due to the Law of Large Numbers), and this is true for all s.

(ii) Based on a Taylor development of the logarithm,

$$x \log(x/x_0) = x - x_0 + \frac{(x-x_0)^2}{2x_0} + O(x-x_0)^3,$$

when $x \to x_0 > 0$.

Problem 15.48. Use these facts to show that, under the null,

$$\sum_{s=1}^{m} y_s \log\left(\frac{y_s}{n\theta_{0,s}}\right) = (1/2 + R_n) \sum_{s=1}^{m} \frac{(y_s - n\theta_{0,s})^2}{n\theta_{0,s}},$$

as the sample size increases to infinity, where R_n is an unspecified term that converges to 0 in probability as $n \to \infty$. The sum of the right-hand side is Pearson's statistic.

15.9.2 Sampling without replacement

Our discussion was so far based on rolling dice. The same concepts and methods extend to experiments with urns. As we know, if the sampling of balls is done with replacement, then the setting is equivalent to that of rolling dice. Therefore, we assume that the sampling is done without replacement.

Multivariate hypergeometric distributions We sample without replacement from an urn containing v_s balls of color $s \in \{1, \dots, m\}$. This is done n times and, for this to be possible, we assume that $n \le v := v_1 + \cdots + v_m$. Let $\omega_i = s$ if the ith draw resulted in a ball with color s, and let $(Y_1, \dots, Y_m)$ be the counts, as before. We say that this vector of counts has the *multivariate hypergeometric distribution* with parameters $(n, v_1, \dots, v_m)$.

Problem 15.49. Argue as simply as you can that Y_s has (marginally) the hypergeometric distribution with parameters $(n, v_s, v - v_s)$.

Proposition 15.50. *The multivariate hypergeometric distribution with parameters $(n, v_1, \dots, v_m)$ has probability mass function*

$$f(y_1, \dots, y_m) := \frac{\binom{v_1}{y_1} \cdots \binom{v_m}{y_m}}{\binom{v}{n}}, \tag{15.13}$$

supported on m-tuples of integers $y_1, \dots, y_m \ge 0$ such that $y_1 + \cdots + y_m = n$.

Problem 15.51. Prove Proposition 15.50.

Problem 15.52 (Maximum likelihood)**.** Assume that the total number of balls in the urn (denoted v above) is

known. Can you derive the maximum likelihood estimator for $(v_1, \ldots, v_m)$?

Problem 15.53 (Sufficiency)**.** Show that $(Y_1, \ldots, Y_m)$ is sufficient for $(v_1, \ldots, v_m)$.

Goodness-of-fit testing In principle, the likelihood ratio statistic can be derived under the multivariate hypergeometric model in the various settings considered in the present chapter under the multinomial model. An approach that is often good enough, though, is to use the test statistic obtained under the multinomial model and obtain the corresponding p-value under the multivariate hypergeometric model.

15.10 Additional problems

Problem 15.54. For $Y = (Y_1, \ldots, Y_m)$ multinomial with parameters $(n, \theta_1, \ldots, \theta_m)$, compute $\mathrm{Cov}(Y_s, Y_t)$ for $s \neq t$.

Problem 15.55 (Daily 3 lottery)**.** The *Daily 3* is a lottery game run by the State of California. Each day of the year, 3 digits are drawn independently and uniformly at random from $\{0, \ldots, 9\}$. Note that the order matters. According the website "The draws are conducted using an Automated Draw Machine, which is a state-of-the-art computer used to draw winning numbers." A goodness-of-fit test could be used to test the machine. Before continuing, download all past winning numbers from the website. [80]

- *Independent digits.* Suppose we are willing to assume that the digits are generated independently. The question that remains, then, is whether they are generated uniformly at random. Take $m = 10$ and $\theta_{0,s} = 1/m$ for all s, and perform the test using the function of Problem 15.9.
- *Independent daily draws.* Suppose we are not willing to assume a priori that the digits are generated independently, but we are willing to assume that the daily draws (each consisting of 3 digits) are generated independently. There are $10 \times 10 \times 10 = 1000$ possible draws (since the order matters), [81] so that here $m = 1000$ and $\theta_{0,s} = 1/m$ for all s. Perform the test using the function of Problem 15.9.

Problem 15.56. (continued) Although the second test makes fewer assumptions, it is not necessarily better than the first test in terms of power. Indeed, while by construction the first test is insensitive to any dependency

[80] calottery.com/play/draw-games/daily-3

[81] The number of possible values, $m = 1000$, is rather large. See Problem 15.58. However, this is compensated by the fact that the sample size is quite large, as the 16000th draw was on June 19, 2019.

between the digits in a draw, it is more powerful than the second test if there is no such dependency. Perform some numerical simulations to probe this claim.

Problem 15.57. Another reasonable way to obtain a test statistic in the context of Section 15.2 is to come up with an estimator $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$ (for example the MLE) and use as test statistic $\mathcal{L}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0)$, where $\mathcal{L}$ is some predetermined loss function, e.g., $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|$, where $\|\cdot\|$ denotes here the Euclidean norm. In R, perform some simulations to compare the resulting test with the Pearson test of Section 15.9.1.

Problem 15.58. Consider a goodness-of-fit situation with m possible values for each trial and assume that n trials are performed. Suppose we want to test

$$\mathcal{H}_0 : \text{the distribution is uniform,}$$

versus

$$\mathcal{H}_1 : \text{the distribution has support of size } \lfloor m/2 \rfloor.$$

These hypotheses are seemingly quite 'far apart', but in fact this really depends on how large m is compared to n.

(i) Show that, if $m, n \to \infty$ with $n \ll \sqrt{m}$ then no test has any power in the limit meaning that any test at level α has limiting power α.

(ii) Confirm this with numerical experiments. In R, perform some simulations, evaluating the power of the LR test. Set the level at $\alpha = 0.01$. Try $m \in \{100, 1000, 10000\}$ and $n = \lfloor \sqrt{m} \rfloor$.

Problem 15.59 (Group sizes and power)**.** Consider a simple setting where we want to compare two coins in terms of their chances of landing heads. Coin j is a θ_j-coin and is tossed n_j times, for $j \in \{1, 2\}$. Fix the total sample size at $n = n_1 + n_2 = 100$. Also, fix θ_1 at $1/2$. For $n_2 \in \{10, 20, 30, 40, 50\}$, evaluate the power of the LR test as a function of θ_2 carefully chosen on a grid that changes with n_2. A possible way to present the results is to set the level at $\alpha = 0.10$ and draw the (estimated) power curve as a function of θ_2 for each n_2.

Problem 15.60 (The number π)**.** We mentioned in Section 10.6.1 that the digits defining π in base 10 behave very much like a sequence of iid random variables from the uniform distribution on $\{0, \ldots, 9\}$. Consider the first $n = 20000$ digits. [82] Ignoring the order, could these numbers be construed as a sample from the uniform distribution?

Problem 15.61 (Gauss-Kuzmin distribution)**.** Let X be uniform in $[0, 1]$ and let (K_m) denote the coefficients in

[82] Available at http://oeis.org/A000796/b000796.txt

the continued fraction expansion of X, meaning that

$$X = \cfrac{1}{K_1 + \cfrac{1}{K_2 + \ldots}}.$$

Then (K_m) converges weakly to the *Gauss-Kuzmin distribution*, defined by its mass function

$$f(k) := -\log_2\left(1 - 1/(k+1)^2\right), \quad \text{for } k \geq 1.$$

($\log_2$ denotes the logarithm in base 2.) Perform some simulations in R to numerically corroborate this statement.

Problem 15.62 (Racial discrimination in the labor market)**.** The article [19] describes an experiment where resumés are sent in response to real job ads in Boston and Chicago with randomly assigned African-American- or White- sounding names. Look at the data summarized in Table 1 of that paper. Identify and then apply the most relevant testing procedure introduced in the present chapter.

Problem 15.63 (Proportions test)**.** The authors of [19] used a procedure not introduced in the present chapter called the *proportions test*, which is based on the fact that the difference of the sample proportions is approximately normal. In general, consider an experiment as in Table 15.3. Define $\bar{Z}_j = Z_j/n_j$ and $\bar{Z} = (Z_1 + Z_2)/(n_1 + n_2)$. Show that, under the null hypothesis,

$$\frac{\bar{Z}_1 - \bar{Z}_2}{\sqrt{\bar{Z}(1-\bar{Z})(1/n_1 + 1/n_2)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1),$$

in the large-sample limit where $n_1 \wedge n_2 \to \infty$. Based on this, obtain a p-value. (Note that the p-value will only be valid in the large-sample limit.) Apply the resulting testing procedure to the data of Table 1 in [19]. [In R, this test is implemented in the prop.test function.]

CHAPTER 16

ONE NUMERICAL SAMPLE

We consider an experiment that yields, as data, a sample of independent and identically distributed (real-valued) random variables, $X_1, \dots, X_n$, with common distribution denoted P, having distribution function F, and density f (when absolutely continuous). We will let $x_1, \dots, x_n$ denote a realization of $X_1, \dots, X_n$, and denote $\boldsymbol{X} = \boldsymbol{X}_n = (X_1, \dots, X_n)$ and $\boldsymbol{x} = \boldsymbol{x}_n = (x_1, \dots, x_n)$. Throughout, we will let X denote a generic random variable with distribution P. The distribution P is assumed to belong to some class of distributions on the real line, which will be taken to be all such distributions whenever that class is not specified. The goal, as usual, is to infer this distribution, or some of its features, from the observed data.

Example 16.1 (Exoplanets)**.** The Extrasolar Planets Encyclopaedia offers a catalog of confirmed exoplanets together with some characteristics of these planets including mass, which is available for hundreds of these planets.

Remark 16.2. There is a statistical model in the background, $(\Omega, \Sigma, \mathcal{P})$, which will be left implicit, except that $\mathbb{P} \in \mathcal{P}$ will be used on occasion to denote the (true) underlying distribution.

16.1 Order statistics

Order $X_1, \ldots, X_n$ to get the so-called *order statistics*, typically denoted

$$X_{(1)} \le \cdots \le X_{(n)}.$$

Each $X_{(k)}$ is indeed a bona fide statistic. In particular, $X_{(1)} = \min(X_1, \ldots, X_n)$ and $X_{(n)} = \max(X_1, \ldots, X_n)$.

Problem 16.3. Derive the distribution of $(X_1, \ldots, X_n)$ given $(X_{(1)}, \ldots, X_{(n)}) = (y_1, \ldots, y_n)$, where $y_1 \le \cdots \le y_n$. [This distribution only depends on $(y_1, \ldots, y_n)$.]

This proves that the orders statistics are jointly sufficient, regardless of the assumed statistical model. That the order statistics are sufficient is intuitively clear since when reducing the sample to the order statistics all that is lost is the ordering of $X_1, \ldots, X_n$, and that ordering does not carry any information on the underlying distribution because the sample is assumed to be iid.

Problem 16.4. Show that, if $X_1, \ldots, X_n$ are iid from a continuous distribution, then they are all distinct with probability 1, implying in particular that, with probability 1,

$$X_{(1)} < \cdots < X_{(n)}. \tag{16.1}$$

16.2 Empirical distribution

The *empirical distribution* is defined as the uniform distribution on $\{x_1, \ldots, x_n\}$, meaning

$$\widehat{\mathsf{P}}_{\boldsymbol{x}}(\mathcal{A}) := \frac{\#\{i : x_i \in \mathcal{A}\}}{n}, \quad \text{for } \mathcal{A} \subset \mathbb{R}. \tag{16.2}$$

As a function of $X_1, \ldots, X_n$, $\widehat{\mathsf{P}}_{\boldsymbol{X}}$ is a random distribution on the real line. (We sometimes drop the subscript in what follows.)

Problem 16.5 (Consistency of the empirical distribution)**.** Using the Law of Large Numbers, show that, for any Borel set $\mathcal{A} \subset \mathbb{R}$,

$$\widehat{\mathsf{P}}_{\boldsymbol{X}_n}(\mathcal{A}) \xrightarrow{\mathrm{P}} \mathsf{P}(\mathcal{A}), \quad \text{as } n \to \infty. \tag{16.3}$$

16.2.1 Empirical distribution function

The *empirical distribution function* is the distribution function of the empirical distribution defined in (16.2).

Problem 16.6. Show that the empirical distribution function is given by

$$\widehat{\mathsf{F}}_{\boldsymbol{x}}(x) := \frac{1}{n}\sum_{i=1}^{n}\{x_i \le x\}.$$

Seen as a function of $X_1, \dots, X_n$, $\widehat{\mathsf{F}}_{\boldsymbol{X}}$ is a random distribution function. (We sometimes drop the subscript in what follows.)

When the observations are all distinct (which happens with probability one when the distribution is continuous, as seen in Problem 16.4), $\widehat{\mathsf{F}}_{\boldsymbol{x}}$ is a step function jumping an amount of $1/n$ at each x_i, and

$$\widehat{\mathsf{F}}_{\boldsymbol{x}}(x_{(i)}) = \frac{i}{n}, \quad \text{for all } i = 1, \dots, n.$$

See Figure 16.1 for an illustration.

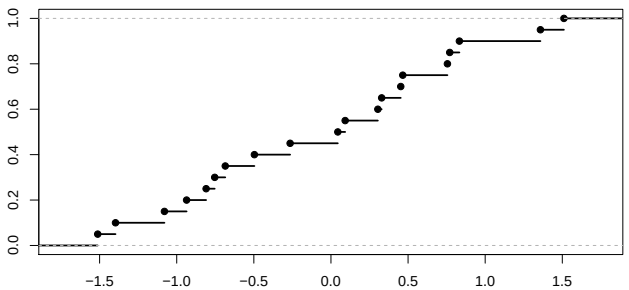
In general, if $x_{(i-1)} < x_{(i)} = \cdots = x_{(i+k-1)} < x_{(i+k)}$, then $\widehat{\mathsf{F}}_{\boldsymbol{x}}$ jumps an amount of k/n at $x_{(i)}$

R corner. The function `ecdf` takes a sample (in the form of a numerical vector) and returns the empirical distribution function.

Problem 16.7 (Consistency of the empirical distribution function)**.** Using the Law of Large Numbers, show that, for any $x \in \mathbb{R}$,

$$\widehat{\mathsf{F}}_{\boldsymbol{X}_n}(x) \xrightarrow{\mathrm{P}} \mathsf{F}(x), \quad \text{as } n \to \infty. \tag{16.4}$$

Figure 16.1: A plot of the empirical distribution function of a sample of size $n = 20$ drawn from the standard normal distribution.



Thus the empirical distribution function is a pointwise consistent estimator of the distribution function. In fact, the convergence is uniform over the whole real line.

Theorem 16.8 (Glivenko–Cantelli[83])**.** *In the present context of an empirical distribution function based on an iid sample of size n with distribution function F, $\boldsymbol{X}_n$,*

$$\sup_{x\in\mathbb{R}} \left|\widehat{\mathsf{F}}_{\boldsymbol{X}_n}(x) - \mathsf{F}(x)\right| \xrightarrow{P} 0, \quad \textit{as } n \to \infty.$$

In fact, the convergence happens at the $\sqrt{n}$ rate.

[83] Named after Valery Glivenko (1897 - 1940) and Cantelli[40].

Theorem 16.9 (Dvoretzky–Kiefer–Wolfowitz [70])**.** *In the context of Theorem 16.8, assuming that* F *is continuous, for all* $t \geq 0$*,*

$$\mathsf{P}\Big(\sup_{x \in \mathbb{R}} \big| \widehat{\mathsf{F}}_{\boldsymbol{X}_n}(x) - \mathsf{F}(x) \big| \geq t/\sqrt{n} \Big) \leq 2 \exp(-2t).$$

Remark 16.10 (Continuous interpolation)**.** Even if the underlying distribution function is continuous, its empirical counterpart is a step function. For this reason, it is sometimes preferred to use a continuous variant of the empirical distribution function. A popular one is the function that linearly interpolates the points

$$(x_{(1)}, 1/n), (x_{(2)}, 2/n), \ldots, (x_{(n)}, 1).$$

(This assumes the observations are distinct.) The function is defined to take the value 1 at $x > x_{(n)}$, but it is not clear how to define this function at $x < x_{(1)}$. An option is to define it as 0 there, but in case the resulting function is discontinuous at $x_{(1)}$. If the underlying distribution is known to be supported on the positive real line, for example, then the function can be made to linearly interpolate $(0, 0)$ and $(x_{(1)}, 1/n)$, and take the value 0 at $x < 0$.

16.2.2 Empirical quantile function

The *empirical quantile function* is simply the quantile function of the empirical distribution, or equivalently, the pseudo-inverse defined in (4.16) of the empirical distribution function. (If one prefers the variant of the empirical distribution function defined in Remark 16.10, then its pseudo-inverse should be preferred also.)

Problem 16.11. The function quantile in R computes quantiles in a number of ways. In fact, the method offers no fewer than 9 ways of doing so.

(i) What type of quantile corresponds to our definition above (based on (4.16))?

(ii) What type of quantile corresponds to the pseudo-inverse (4.23) of the empirical distribution function?

(iii) What type of quantile corresponds to the pseudo-inverse of the piecewise linear variant of the empirical distribution function defined in Remark 16.10?

Remark 16.12. When the observations are distinct, according to the definition given in (4.21), $x_{(i)}$ is a u-quantile of the empirical distribution for any $(i-1)/n \leq u \leq i/n$.

Problem 16.13 (Consistency of the empirical quantile function)**.** Show that, at any point u where F is continuous and strictly increasing,

$$\widehat{\mathsf{F}}^{-}_{\boldsymbol{X}_n}(u) \xrightarrow{\mathrm{P}} \mathsf{F}^{-}(u), \quad \text{as } n \to \infty,$$

[This is based on (16.4) and arguments similar to those underlying Problem 8.59.]

16.2.3 Histogram

Assume that F has a density, denoted f. Is there an empirical equivalent to f? $\widehat{\mathsf{F}}_{\boldsymbol{X}}$, being discrete, does not have a density but rather a mass function, and it is clear that a mass function cannot provide a good approximation to a density.

The idea behind the construction of a *histogram* is to consider averages of the mass function over short intervals so that it provides a piecewise constant approximation to the underlying density. This approximation turns out to be pointwise consistent under some conditions. See Figure 16.2 for an illustration.

Consider a strictly increasing sequence $(a_k : k \in \mathbb{Z})$, which defines a partition of the real line into the intervals $\{(a_{k-1}, a_k] : k \in \mathbb{Z}\}$, often called *bins* in the present context. Define the corresponding *counts*, also called *frequencies*, as

$$y_k := \#\{i : x_i \in (a_{k-1}, a_k]\}, \quad k \in \mathbb{Z},$$

with the corresponding random variables being denoted $(Y_k : k \in \mathbb{Z})$. We have that $Y_k \sim \mathrm{Bin}(n, p_k)$ with

$$p_k := \mathsf{F}(a_k) - \mathsf{F}(a_{k-1}),$$

which can therefore be estimated by $\hat{p}_k := y_k/n$.

Assume that f can be taken to be continuous. (The discussion that follows extends without difficulty to the case where f has a finite number of discontinuities.) If $a_k - a_{k-1}$ is small, then

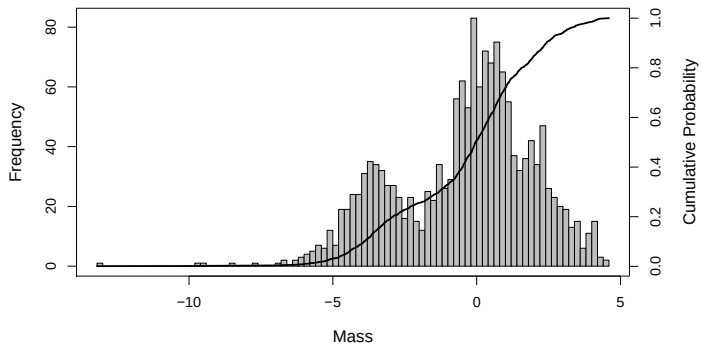
$$p_k = \mathsf{F}(a_k) - \mathsf{F}(a_{k-1}) \approx (a_k - a_{k-1}) f(a_k). \tag{16.5}$$

The approximation is in fact exact to first order.

The histogram with bins defined by (a_k) is the piecewise constant function

$$\hat{f}_{\boldsymbol{x}_n}(x) := \frac{y_k}{n(a_k - a_{k-1})}, \quad \text{when } x \in (a_{k-1}, a_k].$$

Figure 16.2: A histogram and a plot of the distribution function of the data described in Example 16.1, meaning of the mass of 1582 exoplanets discovered in or before 2017. The mass is measured in Jupiter mass ($\mathrm{M_{Jup}}$), presented here in logarithmic scale for clarity.



For $x \in (a_{k-1}, a_k]$,

$$\hat{f}_{\boldsymbol{X}_n}(x) \xrightarrow{\mathrm{P}} \frac{p_k}{(a_k - a_{k-1})} \approx f(a_k), \quad \text{as } n \to \infty, \tag{16.6}$$

where the approximation is accurate when $a_k - a_{k-1}$ is small, as seen in (16.5).

For $\hat{f}_{\boldsymbol{X}_n}$ to be consistent for f, it is thus necessary that the bins become smaller and smaller as the sample size increases. Below we let the bins depend on n and denote $(a_{k,n})$ the sequence defining the bins and, for clarity, we let $\hat{f}_n$ denote the resulting histogram based on $\boldsymbol{X}_n$ and these bins.

Problem 16.14. Suppose that

$$\begin{aligned} &\max_k (a_{k,n} - a_{k-1,n}) \to 0, \\ &\text{with } \min_k (a_{k,n} - a_{k-1,n}) \gg 1/n. \end{aligned} \tag{16.7}$$

Show that, at any point $x \in \mathbb{R}$ where f is continuous,

$$\hat{f}_n(x) \xrightarrow{\mathrm{P}} f(x), \quad \text{as } n \to \infty.$$

To better appreciate the crucial role that (16.7) plays, consider the regular grid $a_{k,n} = k/n$, so that all bins have size $1/n$. In particular, this sequence does not satisfy (16.7). In fact, in that case, show that, at any point x,

$$\mathbb{P}(\hat{f}_n(x) = 0) \to 1/\mathrm{e}, \quad \text{as } n \to \infty.$$

Remark 16.15 (Choice of bins)**.** Choosing the bins automatically is in general a complex task. Often, a regular partition is chosen, for example $a_k = kh$, and even then, the choice of $h > 0$ is nontrivial. It is known that, if the function has a bounded first derivative, a bin size of order $h \propto n^{-1/3}$ is best. Although this can provide some guidance, the best choice depends on the underlying density, resulting in a chicken-and-egg problem.[84] See Figure 16.3 for an illustration.

R corner. The function hist computes and (by default) plots a histogram based on the data. The function offers the possibility of manually choosing the bins as well as three methods for choosing the bins automatically.

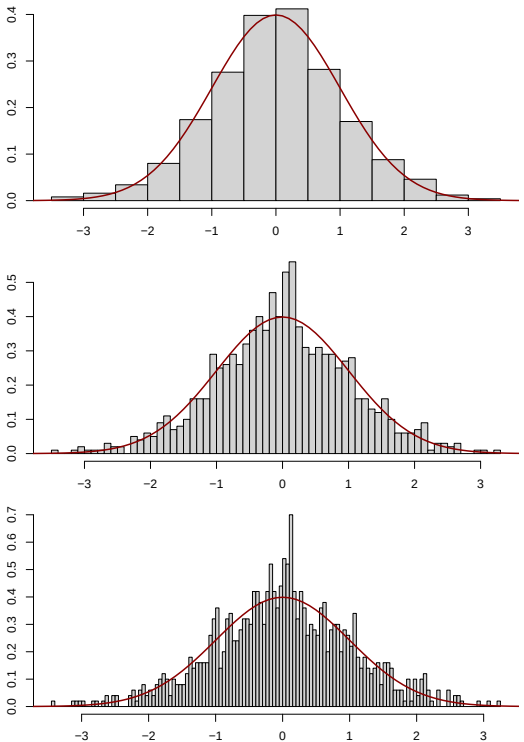
16.3 Inference about the median

Recall the definition of a u-quantile of P or, equivalently, F, given in (4.21). We called *median* any 1/2-quantile. In particular, by definition, x is a median of F if (recall (4.19))

$$\mathsf{F}(x) \geq 1/2 \text{ and } \tilde{\mathsf{F}}(x) \geq 1/2,$$

[84] It is possible to choose the bin size h by cross-validation, as Rudemo proposes in [199]. We provide some details in the closely related context of kernel density estimation in Section 16.10.6.

Figure 16.3: Histograms of a sample of size $n = 1000$ drawn from the standard normal distribution with different number of bins. (The bins themselves where automatically chosen by the function hist.)



or, equivalently in terms of $X \sim \mathsf{P}$,

$$\mathbb{P}(X \le x) \ge 1/2 \text{ and } \mathbb{P}(X \ge x) \ge 1/2.$$

Problem 16.16. Show that these inequalities are in fact equalities when F is continuous at any of its median points.

Problem 16.17. Show that the set of medians is the interval $[a, b)$ where $a := \inf\{x : \mathsf{F}(x) \ge 1/2\}$ and $b := \sup\{x : \mathsf{F}(x) = \mathsf{F}(a)\}$, where by convention $[a, a) = \{a\}$. Conclude that there is a unique median if and only if the distribution function is strictly increasing at any of its median points.

In what follows, to ease the exposition, we consider the case where there is a unique median (denoted μ). The reader is invited to examine how what follows generalizes to the case where the median is not unique.

16.3.1 Sample median

We already have an estimator of the median, namely, $\widehat{\mathsf{F}}^{-}_{\boldsymbol{X}}(1/2)$, which is consistent if F is strictly increasing and continuous at μ (Problem 16.13). Any such estimator for the median can be called a *sample median*. Any reasonable definition leads to a consistent estimator.

R corner. In R, the function median computes the median of a sample based on a different definition of pseudo-

inverse, specifically $\widehat{\mathsf{F}}_{\boldsymbol{x}}^{\ominus}$ as defined in (4.23). In particular, if all the data points are distinct,

$$\widehat{\mathsf{F}}_{\boldsymbol{x}}^{\ominus}(1/2) = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) & \text{if } n \text{ is even.} \end{cases}$$

16.3.2 Confidence interval

A (good) confidence interval can be built for the median without really any assumption on the underlying distribution. The interval is of the form $[X_{(k)}, X_{(l)}]$ for some $k \le l$ chosen as functions of the desired level of confidence.

We start with

$$\mathbb{P}(X_{(k)} \le \mu \le X_{(l)}) = \mathbb{P}(X_{(k)} \le \mu) - \mathbb{P}(X_{(l)} < \mu).$$

Let

$$q_k := \text{Prob}\big\{\text{Bin}(n, 1/2) \ge k\big\}.$$

We have

$$\mathbb{P}(X_{(k)} \le \mu) = \mathbb{P}\big(\#\{i : X_i \le \mu\} \ge k\big) \ge q_k,$$

because $\#\{i : X_i \le \mu\}$ is binomial with probability of success $\mathbb{P}(X \le \mu) \ge 1/2$, and similarly,

$$\mathbb{P}(X_{(l)} < \mu) = \mathbb{P}\big(\#\{i : X_i < \mu\} \ge l\big) \le q_l,$$

because $\mathbb{P}(X < \mu) \le 1/2$. Thus,

$$\mathbb{P}(X_{(k)} \le \mu \le X_{(l)}) \ge q_k - q_l.$$

Hence, $[X_{(k)}, X_{(l)}]$ is a confidence interval for μ at level $q_k - q_l$. Choosing k as the largest integer such that $q_k \ge 1 - \alpha/2$ and l the smallest integer such that $q_l \le \alpha/2$, we obtain a $(1-\alpha)$-confidence interval for μ.

Problem 16.18 (One-sided interval)**.** Derive a one-sided $(1-\alpha)$-confidence interval for the median following the same reasoning.

Problem 16.19. In R, write a function that takes as input the data points, the desired confidence level, and the type of interval, and returns the corresponding confidence interval for the median. Try your function on a sample of size $n \in \{10, 20, 30, \ldots, 100\}$ from the exponential distribution with rate 1. Repeat each setting $N = 200$ times and plot the average length of the confidence interval as a function of n.

16.3.3 Sign test

Suppose we want to test $\mathcal{H}_0 : \mu = \mu_0$. We saw in Section 12.4.6 how to derive a p-value from a procedure for constructing confidence intervals.

Problem 16.20. In R, write a function that takes as input the data points and μ_0, and returns the p-value based on the above procedure for building a confidence interval for the median.

Depending on what variant of the median is used, and on whether there are observed values at the median, the resulting test procedure coincides with, or is very close to, the following test known as the *sign test*. Let

$$Y_+ = \#\{i : X_i > \mu_0\}, \quad Y_- = \#\{i : X_i < \mu_0\},$$

and

$$Y_0 = \#\{i : X_i = \mu_0\}.$$

The test rejects for large values of $S := \max(Y_+, Y_-)$, with the p-value computed conditional on Y_0.

Remark 16.21. The name of the test comes from looking at the sign of $X_i - \mu_0$ and counting how many are positive, negative, or zero.

Problem 16.22. Show that, if the underlying distribution has median μ_0,

$$\mathbb{P}(Y_+ \geq k \mid Y_0 = y_0) \leq \text{Prob}\big\{\text{Bin}(n - y_0, 1/2) \geq k\big\},$$
$$\mathbb{P}(Y_- \geq k \mid Y_0 = y_0) \leq \text{Prob}\big\{\text{Bin}(n - y_0, 1/2) \geq k\big\},$$

and deduce that

$$\mathbb{P}(S \geq k \mid Y_0 = y_0) \leq 2\,\text{Prob}\big\{\text{Bin}(n - y_0, 1/2) \geq k\big\}. \quad (16.8)$$

[This upper bound can be used as a (conservative) p-value.]

Problem 16.23. Derive the sign test and its (conservative) p-value for testing $\mathcal{H}_0 : \mu \leq \mu_0$.

Problem 16.24. In R, write a function that takes in the data points and μ_0, and the type of alternative, and returns the (conservative) p-value for the corresponding sign test.

16.3.4 Inference about a quantile

Whatever was said thus far about estimating or testing about the median can be extended to any u-quantile with $0 < u < 1$.

Problem 16.25. Repeat for the 1st quartile what was done for the median.

Estimating the 0-quantile or the 1-quantile amounts to estimating the boundary points of the support of the distribution.

16.4 Possible difficulties

We consider some emblematic situations where estimating the median is difficult. We do the same for the mean, as a prelude to studying its inference. In the process, we provide some insights into why it is much more complicated than inference for the median.

16.4.1 Difficulties with the median

A difficult situation for inference about the median is when the underlying distribution is flat at the median. For $\theta \in [0,1]$, consider the following density

$$f_\theta(x) = (1-\theta)\,\{x \in [0,1]\} + \theta\,\{x \in [2,3]\},$$

and let F_θ denote the corresponding distribution function.

Problem 16.26. Show that sampling from f_θ amounts to drawing $\xi \sim \text{Ber}(\theta)$ and then drawing $X \sim \text{Unif}(0,1)$ if $\xi = 0$, and $X \sim \text{Unif}(2,3)$ if $\xi = 1$.

Problem 16.27. Show that F_θ is flat at its median if and only if $\theta = 1/2$. When this is the case, show that any point in $[1,2]$ is a median. When this is not the case, meaning when $\theta \neq 1/2$, show that the median is unique and derive it as a function of θ.

Assume we have an iid sample from f_θ of size n and that our goal is to draw inference about 'the' median.

Problem 16.28. Show that, when $\theta = 1/2$ and n is odd, the sample median belongs to $[0,1]$ with probability $1/2$ and belongs to $[2,3]$ with probability $1/2$.

The difficulty is only in appearance, however. Indeed, the sample median will converge, as the sample size increases, to a median of the underlying distribution and, more importantly, the confidence interval of Section 16.3.2 has the desired confidence no matter what, although it can be quite wide.

Problem 16.29. In R, generate a sample from f_θ of size $n = 101$ (so it is odd) and produce a 95% confidence interval for the median using the function of Problem 16.19. Do that for $\theta \in \{0.2, 0.4, 0.45, 0.49, 0.5, 0.51, 0.55, 0.6, 0.8\}$. Repeat each setting a few times to get a feel for the randomness.

Remark 16.30. The situation is qualitatively the same when

$$f_\theta(x) = (1-\theta)\,f_0(x) + \theta\,f_1(x),$$

with f_0 and f_1 being densities with disjoint supports.

16.4.2 Difficulties with the mean

While estimating the median does not pose particular difficulties despite some cases where it is 'unstable', estimating the mean poses very real difficulties, to the point that the problem is almost ill-posed.

For a prototypical example, consider the family of densities

$$f_\theta(x) = (1-\theta)\,\{x \in [0,1]\} + \theta\,\{[h(\theta), h(\theta)+1]\},$$

parameterized by $\theta \geq 0$, where $h : \mathbb{R}_+ \to \mathbb{R}_+$ is some function.

Problem 16.31. Show that f_θ has mean $1/2 + \theta h(\theta)$.

As before, sampling from f_θ amounts to generating $\xi \sim \text{Ber}(\theta)$, and then drawing $X \sim \text{Unif}([0,1])$ if $\xi = 0$, and $X \sim \text{Unif}([h(\theta), h(\theta)+1])$ if $\xi = 1$.

Problem 16.32. In a sample of size n from f_θ, show that the number of points generated from $\text{Unif}([0,1])$ is binomial with parameters $(n, 1-\theta)$.

Consider the situation where h is such that $\theta h(\theta) \to \infty$ as $\theta \to 0$, and choose $\theta = \theta_n$ such that $n\theta_n \to 0$. In particular, f_{θ_n} has mean $1/2 + \theta_n h(\theta_n) \to \infty$ as $n \to \infty$, while with probability tending to 1, the entire sample is drawn from $\text{Unif}([0,1])$, which has mean $1/2$.

Remark 16.33. A very similar difficulty is at the core of the Saint Petersburg Paradox discussed in Section 7.10.3. We saw there that considering the median instead of the mean offers an attractive way of solving the apparent paradox.

16.5 Bootstrap

Inference about the mean will rely on the bootstrap. The reasoning is the same as in Section 15.3.2 and Section 15.7.3. It goes as follows. If we could sample from P at will, we would be able to estimate any feature of P (including mean, median, quantiles, etc) by Monte Carlo simulation, and the accuracy of our inference would only be limited by the amount of computational resources at our disposal. Doing this is not possible since P is unknown, but we can estimate it by the empirical distribution. This is justified by the fact that the empirical distribution is a consistent estimator, as seen in (16.3).

Remark 16.34 (Bootstrap world)**.** The *bootstrap world* is the fictitious world where we pretend that the empirical distribution is the underlying distribution. In that world, we know everything, at least in principle, since we know the empirical distribution, and in particular we can use Monte Carlo simulations to compute any feature

of interest of that distribution. An asterisk $*$ next of a symbol representing some quantity is often used to denote the corresponding quantity in the bootstrap world. For example, if μ denotes the median, then μ^* will denote the empirical median. In particular, we will use P^* in place of $\widehat{\mathsf{P}}_{\boldsymbol{x}}$ in what follows to denote the empirical distribution.

16.5.1 Bootstrap sample

Sampling from P^* is relatively straightforward, since it is the uniform distribution on $\{x_1, \ldots, x_n\}$. A sample (of same size n) drawn from the empirical distribution is called a *bootstrap sample* and denoted $X_1^*, \ldots, X_n^*$.

> A bootstrap sample is generated by sampling *with replacement* n times from $\{x_1, \ldots, x_n\}$.

R corner. In R, the function sample can be used to sample (with or without replacement) from a finite set, where a finite set is represented by a vector.

Remark 16.35 (Ties in the bootstrap sample)**.** Even when all the observations are distinct, by construction, a bootstrap sample may include some ties.

Problem 16.36. Compute the probability that there are no ties in a bootstrap sample when $x_1, \ldots, x_n$ are all distinct.

16.5.2 Bootstrap distribution

Let T be a statistic of interest, and let P_T denote its distribution (meaning, the distribution of $T(X_1, \ldots, X_n)$). Having observed $x_1, \ldots, x_n$, resulting in the empirical distribution P^*, the *bootstrap distribution* of T is the distribution of $T(X_1^*, \ldots, X_n^*)$. We denote this distribution by P_T^*. It is used to estimate the distribution of T.

In practice, only rarely can we obtain P_T^* in closed form. Instead, P_T^* is typically estimated by Monte Carlo simulation, which is available to us since we may sample from F^* at will.

Problem 16.37. In R, generate a sample of size $n \in \{5, 10, 20, 50\}$ from the exponential distribution with rate 1. Let T be the sample mean and estimate its bootstrap distribution by Monte Carlo using $B = 10^4$ replicates. For each n, draw a histogram of this estimate, overlay the density given by the normal approximation, and overlay the density of the gamma distribution with shape parameter n and rate n, which is the distribution of T (meaning P_T) in this case.

16.5.3 Bootstrap estimate for a bias

Suppose a particular statistic T is meant to estimate a feature of the underlying distribution, denoted $\varphi(\mathsf{P})$. For that purpose, its bias is defined as

$$b := \mathbb{E}(T) - \varphi,$$

where $\mathbb{E}(T)$ is shorthand for $\mathbb{E}(T(X_1, \ldots, X_n))$ where $X_1, \ldots, X_n$ are iid from P, and φ is shorthand for $\varphi(\mathsf{P})$.

It turns out that it can be estimated by bootstrap. Indeed, in the bootstrap world the corresponding quantity is

$$b^* := \mathbb{E}^*(T) - \varphi^*,$$

where $\mathbb{E}^*(T)$ is shorthand for $\mathbb{E}(T(X_1^*, \ldots, X_n^*))$ where $X_1^*, \ldots, X_n^*$ are iid from P^*, and φ^* is shorthand for $\varphi(\mathsf{P}^*)$. (We assume that φ applies to discrete distributions such as the empirical distribution. This is the case, for example, if φ is a moment or quantile.)

In the bootstrap world, we know P^*, and therefore b^*, at least in principle. In practice, though, b^* is typically estimated by Monte Carlo simulation by repeatedly sampling from P^*. This MC estimate for b^* serves as an estimate for b, the bias of T in the 'real' world.

16.5.4 Bootstrap estimate for a variance

In addition to the bias, the variance can also be estimated by bootstrap. Suppose that we are interested in estimating the variance of a given statistic T. The bootstrap estimate is simply its variance in the bootstrap world, namely, its variance under P^*, or in formula

$$\mathrm{Var}^*(T) = \mathbb{E}^*(T^2) - (\mathbb{E}^*(T))^2.$$

As before, this is estimated by Monte Carlo simulation based on repeatedly sampling from P^*.

16.5.5 What makes the bootstrap work

We say that the bootstrap works when it provides consistent estimates as the sample size $n \to \infty$.

The bootstrap tends to work when the statistic of interest is a 'smooth' function of the data. Mere continuity is not enough, as the next example shows.

Problem 16.38. Consider[85] the case where $X_1, \ldots, X_n$ are iid uniform in $[0, \theta]$ and we want to estimate θ.

(i) Show that $X_{(n)} = \max(X_1, \ldots, X_n)$ is the MLE for θ. (Note that this statistic is continuous in the sample.)

[85] This example appears in [247] under a different light.

(ii) Assume henceforth that $\theta = 1$, which is really without loss of generality since we are dealing with a scale family. Let $Y_n := n(1 - X_{(n)})$. Derive the distribution function of Y_n in closed form and then its limit as $n \to \infty$. Plot this limiting distribution function as a dashed line.

(iii) In R, generate a sample of size $n = 10^6$ and estimate the bootstrap distribution of Y_n using $B = 10^4$ Monte Carlo samples.[86] Add the corresponding distribution function to the same plot as a solid line.

16.6 Inference about the mean

While the inference about the median can be performed sensibly without really any assumption on the underlying distribution, the same cannot be said of the mean. The reason is that the mean is not as well-behaved as the median, as we saw in Section 16.4. Thus, it might be preferable to focus on the median rather than the mean whenever possible. However, some situations call for inference about the mean.

[86] Of course, the larger n and B, the better, but with finite computational resources, we might have to choose. Why is it more important in this particular setting to have n large rather than B large? (Of course, in practice n is the sample size, and therefore set once the data are collected.)

16.6.1 Sample mean

We assume in this section that P has a mean, which we denote by μ. A statistic of choice here is the *sample mean* defined as the average of $x_1, \ldots, x_n$, namely $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$. This is also the mean of the empirical distribution. We will denote the corresponding random variable $\bar{X}$, or $\bar{X}_n$ if we want to emphasize that it was computed from a sample of size n.

The Law of Large Numbers implies that the sample mean is consistent for the mean, that is

$$\bar{X}_n \xrightarrow{\mathrm{P}} \mu, \quad \text{as } n \to \infty.$$

16.6.2 Normal confidence interval

Assume that P has variance σ^2. The Central Limit Theorem implies that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

This in turn implies that, if z_u denotes the u-quantile of the standard normal distribution, we have

$$\mu \in \left[\bar{X}_n - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{16.9}$$

with probability converging to $1-\alpha$ as $n \to \infty$.

If σ^2 is known, then the interval in (16.9) is a bona fide confidence interval and its level is $1-\alpha$ in the large-sample limit, although the confidence level at a given sample size n will depend on the underlying distribution.

Problem 16.39. Bound the confidence level from below using Chebyshev's inequality.

If σ^2 is unknown, we estimate it using the *sample variance*, which may be defined as

$$S^2 := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{16.10}$$

This is the variance of the empirical distribution. By Slutsky's theorem, in conjunction with the Central Limit Theorem, it holds that

$$\frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad \text{as } n \to \infty,$$

which in turn implies that

$$\mu \in \left[\bar{X}_n - z_{1-\alpha/2}\frac{S_n}{\sqrt{n}}, \bar{X}_n - z_{\alpha/2}\frac{S_n}{\sqrt{n}}\right] \tag{16.11}$$

with probability converging to $1-\alpha$ as $n \to \infty$.

Remark 16.40 (Unbiased sample variance)**.** The following variant is sometimes used instead in place of (16.10)

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2. \tag{16.12}$$

This is what the R function var computes. This variant happens to be unbiased (Problem 16.92). In practice, the two variants are of course very close to each other, unless the sample size is quite small.

STUDENT CONFIDENCE INTERVAL When $X_1, \ldots, X_n$ are iid from a normal distribution,

$$\frac{\bar{X} - \mu}{S/\sqrt{n-1}}$$

has the Student distribution with parameter $n-1$. In particular, if t_u^n denotes the u-quantile of this distribution, then

$$\mu \in \left[\bar{X}_n - t_{1-\alpha/2}^{n-1}\frac{S_n}{\sqrt{n-1}}, \bar{X}_n - t_{\alpha/2}^{n-1}\frac{S_n}{\sqrt{n-1}}\right] \tag{16.13}$$

with probability $1-\alpha$ when the underlying distribution is normal. In general, this is only true in the large-sample limit.

Problem 16.41. Show that the Student distribution with n degrees of freedom converges to the standard normal distribution as n increases. Deduce that, for any $u \in (0,1)$, $t_u^n \to z_u$ as $n \to \infty$.

Remark 16.42. The Student confidence interval (16.13) appears to be more popular than the normal confidence interval (16.11).

R corner. The family of Student distributions is available via the functions `dt` (density), `pt` (distribution function), `qt` (quantile function), and `rt` (pseudo-random number generator). The Student confidence intervals and the corresponding tests can be computed using the function `t.test`.

16.6.3 Bootstrap confidence interval

When σ is known, the confidence interval in (16.9) relies on the fact that the distribution of $\bar{X} - \mu$ is approximately normal with mean 0 and variance σ^2/n (that is, if n is large enough).

Remark 16.43. The random variable $\bar{X} - \mu$ is often called a *pivot*. It is not a statistic, as it cannot be computed from the data alone (since μ is unknown).

Let us carefully examine the process of deriving this confidence interval. Let Q denote the distribution of $\bar{X} - \mu$. Let q_u denote a u-quantile of Q. By (4.17) and (4.20),

$$\mathbb{P}(\bar{X} - \mu \le q_{1-\alpha/2}) \ge 1 - \alpha/2, \tag{16.14}$$

$$\mathbb{P}(\bar{X} - \mu < q_{\alpha/2}) \le \alpha/2, \tag{16.15}$$

so that

$$\mathbb{P}(q_{\alpha/2} \le \bar{X} - \mu \le q_{1-\alpha/2}) \ge 1 - \alpha,$$

or, equivalently,

$$\mu \in \left[\bar{X} - q_{1-\alpha/2}, \bar{X} - q_{\alpha/2}\right], \tag{16.16}$$

with probability at least $1 - \alpha$. The issue here, of course, is that this construction relies on Q, which is unknown, so that this interval is not a bona fide confidence interval. In (16.9), Q is approximated by a normal distribution. Here we estimate Q by bootstrap instead.

The bootstrap estimation of Q is done, as usual, by going to the bootstrap world. Suppose we have observed $x_1, \dots, x_n$. In the bootstrap world, the equivalent of $\bar{X} - \mu$ is $\bar{X}^* - \mu^*$, where $\bar{X}^*$ is the average of a bootstrap sample of size n and μ^* is the mean of P^*, so that $\mu^* = \bar{x}$. Let Q^* denote the distribution function of $\bar{X}^* - \mu^*$. This is the bootstrap estimate for Q. The hope, as usual, is that the sample is large enough that Q^* is close to Q.

Remark 16.44. In practice, Q^* is itself estimated by Monte Carlo simulation from P^*, and that estimate is our estimate for Q. Below we reason as if we knew Q^*, or equivalently, as if we had infinite computational power and we had the luxury of drawing an infinite number of Monte Carlo samples. (This allows us to separate statistical issues from computational issues.)

Having computed Q^*, a confidence interval is built as done above. Let q_u^* denote the u-quantile of Q^*. The bootstrap $(1-\alpha)$-confidence interval for μ is obtained by plugging q^* in place of q in (16.16), resulting in

$$\left[\bar{X} - q^*_{1-\alpha/2}, \bar{X} - q^*_{\alpha/2}\right]. \tag{16.17}$$

(Note that it is $\bar{X}$ and not $\bar{X}^*$. The latter does not really have a meaning since it denotes the average of a generic bootstrap sample and the procedure is based on drawing many such samples.)

Problem 16.45. In R, write a function that takes in the data, the desired confidence level, and a number of Monte Carlo replicates, and returns the bootstrap confidence interval (16.17).

16.6.4 Bootstrap Studentized confidence interval

Instead of using $\bar{X} - \mu$ as pivot as we did in Section 16.6.3, we now use $(\bar{X}-\mu)/S$. The process of deriving a bootstrap confidence interval is then completely parallel. [87]

Redefine Q as the distribution of $(\bar{X}-\mu)/S$, and redefine q_u as the u-quantile of Q. Then

$$\mu \in \left[\bar{X} - Sq_{1-\alpha/2}, \bar{X} - Sq_{\alpha/2}\right], \tag{16.18}$$

with probability at least $1-\alpha$. In (16.11), Q is approximated by a normal distribution. Here we estimate Q by bootstrap instead.

Suppose we have observed $x_1, \ldots, x_n$. In the bootstrap world, the equivalent of $(\bar{X}-\mu)/S$ is $(\bar{X}^*-\mu^*)/S^*$, where S^* is the sample standard deviation of a bootstrap sample of size n. Let Q^* denote the distribution function of $(\bar{X}^*-\mu^*)/S^*$ and let q_u^* denote its u-quantile. The bootstrap Studentized $(1-\alpha)$-confidence interval for μ is obtained by plugging q^* in place of q in (16.18), resulting in

$$\left[\bar{X} - Sq^*_{1-\alpha/2}, \bar{X} - Sq^*_{\alpha/2}\right]. \tag{16.19}$$

(Note that it is $\bar{X}$ and S, and not $\bar{X}^*$ and S^*.)

[87] We only repeat it for the reader's convenience, however the reader is invited to anticipate what follows.

Problem 16.46. In R, write a function that takes in the data, the desired confidence level, and number of Monte Carlo replicates, and returns the bootstrap confidence interval (16.19).

Remark 16.47 (Comparison)**.** The Studentized version is typically more accurate. You are asked to probe this in Problem 16.95.

16.6.5 Bootstrap tests

Suppose we want to test $\mathcal{H}_0 : \mu = \mu_0$. We saw in Section 12.4.6 how to derive a p-value from a confidence interval procedure.

Problem 16.48. In R, write a function that takes as input the data points and μ_0 and returns the p-value based on the bootstrap Studentized confidence interval.

16.6.6 Inference about other moments and beyond

The bootstrap approach to drawing inference about the mean generalizes to other moments, and more generally, to any expectation such as $\mu_\psi := \mathbb{E}(\psi(X))$, where ψ is a given function (e.g., $\psi(x) = x^k$ gives the kth moment). This is because, if we define $Y_i = \psi(X_i)$, then $Y_1, \ldots, Y_n$ are iid with mean μ_ψ.

Problem 16.49. Define a bootstrap Studentized confidence interval for the kth moment.

16.7 Inference about the variance and beyond

A confidence interval for the variance σ^2 can be obtained by computing a confidence interval for the mean and a confidence interval for the second moment, and combining these to get a confidence interval for the variance, since

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2.$$

However, there is a more direct route, which is typically preferred.

In general, consider a feature of interest $\varphi(\mathsf{P})$. We assume that φ applies to discrete distributions. In that case, a natural estimator is the plug-in estimator $\varphi(\mathsf{P}^*)$. A bootstrap confidence interval can then be derived as we did for the mean in Section 16.6.3, using $\varphi(\mathsf{P}^*) - \varphi(\mathsf{P})$ as pivot.

Indeed, let v_u denote the u-quantile of its distribution. Then

$$\varphi(\mathsf{P}) \in \left[\varphi(\mathsf{P}^*) - v_{\alpha/2}, \varphi(\mathsf{P}^*) - v_{1-\alpha/2}\right]$$

with probability at least $1-\alpha$. Since v_u is not available, we go to the bootstrap world. There the corresponding object is $\varphi(\mathsf{P}^{**}) - \varphi(\mathsf{P}^*)$, where P^{**} is the empirical distribution function of a sample of size n from P^*. We then estimate v_u by v_u^*, the u-quantile of the bootstrap distribution of $\varphi(\mathsf{P}^{**}) - \varphi(\mathsf{P}^*)$, resulting in

$$\varphi(\mathsf{P}) \in \left[\varphi(\mathsf{P}^*) - v^*_{\alpha/2}, \varphi(\mathsf{P}^*) - v^*_{1-\alpha/2}\right]$$

with approximate probability $1-\alpha$ under suitable circumstances (see Section 16.5.5).

Remark 16.50. A bootstrap Studentized confidence interval can also be constructed based on $(\varphi(\mathsf{P}^*) - \varphi(\mathsf{P}))/D$ as pivot, where D is an estimate for the standard deviation of $\varphi(\mathsf{P}^*)$. Unless a simpler estimator is available, we can always use the bootstrap estimate for the standard deviation. If this is our D, then the construction of the Studentized confidence interval requires the computation of a bootstrap estimate for the variance in the bootstrap world, meaning that such an estimate is computed for each bootstrap sample. A direct implementation of this requires a loop within a loop, and is therefore computationally intensive.

16.8 Goodness-of-fit testing and confidence bands

Suppose we want to test whether the underlying distribution is a given distribution, often called the *null distribution* and denoted P_0 henceforth, with distribution function F_0 and (when applicable) density f_0. Recall that we considered this problem in the discrete setting in Section 15.2.

If this happens in the context of a family of distributions that admits a simple parameterization, the hypothesis testing problem will likely be approachable via a standard test (e.g., the likelihood ratio test) on the underlying parameter. This is the case, for example, when we assume that the underlying distribution is in $\{\text{Beta}(\theta, 1) : \theta > 0\}$, and we are testing whether the distribution is uniform distribution on $[0, 1]$, which is equivalent in this context to testing whether $\theta = 1$.

We place ourselves in a setting where no simple parameterization is available. In that case, a plugin approach points to comparing the empirical distribution with the null distribution (since the empirical distribution is always available as an estimate of the underlying distribution). We discuss two approaches for doing so: one based on comparing distribution functions and another one based

on comparing densities.

Remark 16.51 (Tests for uniformity)**.** Under the null distribution, the transformed data $U_1, \ldots, U_n$, with $U_i = \mathsf{F}_0(X_i)$, are iid uniform in $[0,1]$. In principle, therefore, testing for a particular null distribution can be reduced to *testing for uniformity*, that is, the special case where the null distribution is $\mathsf{P}_0 = \mathrm{Unif}(0,1)$. For pedagogical reasons, we chose not to work with this reduction in what follows. (See Section 16.10.1 for additional details.)

16.8.1 Tests based on the distribution function

A goodness-of-fit test based on comparing distribution functions is generally based on a statistic of the form

$$\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}}, \mathsf{F}_0),$$

where Δ is a measure of dissimilarity between distribution functions. There are many such dissimilarities and we present a few classical examples. In each case, large values of the statistic weigh against the null hypothesis.

Kolmogorov–Smirnov test[88] This test uses the supremum norm as a measure of dissimilarity, namely

$$\Delta(\mathsf{F}, \mathsf{G}) := \sup_{x \in \mathbb{R}} \big|\mathsf{F}(x) - \mathsf{G}(x)\big|. \tag{16.20}$$

Problem 16.52. Show that

$$\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}}, \mathsf{F}_0) = \max_{i=1,\ldots,n} \max\left\{\frac{i}{n} - \mathsf{F}_0(X_{(i)}), \mathsf{F}_0(X_{(i)}) - \frac{i-1}{n}\right\}.$$

Calibration is (obviously) by Monte Carlo under F_0. This calibration by Monte Carlo necessitates the use of a computer, yet the method was in use before the advent of computers. What made this possible is the following.

Proposition 16.53. *The distribution of* $\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}}, \mathsf{F}_0)$ *under* F_0 *does not depend on* F_0 *as long as* F_0 *is continuous.*

Proof. We prove the result in the special case where F_0 is strictly increasing on the real line. The key point is that, under F_0, $\mathsf{F}_0(X) \sim \mathrm{Unif}(0,1)$. Let $U_i = \mathsf{F}_0(X_i)$ and let $\widehat{\mathsf{G}}$ denote the empirical distribution function of $U_1, \ldots, U_n$. Also, let $\widehat{\mathsf{F}}$ be shorthand for $\widehat{\mathsf{F}}_{\boldsymbol{X}}$. For $x \in \mathbb{R}$, let $u = \mathsf{F}_0(x)$, and derive

$$\begin{aligned}\widehat{\mathsf{F}}(x) - \mathsf{F}_0(x) &= \widehat{\mathsf{F}}(\mathsf{F}_0^{-1}(u)) - \mathsf{F}_0(\mathsf{F}_0^{-1}(u)) \\ &= \widehat{\mathsf{G}}(u) - u.\end{aligned}$$

[88] Named after Kolmogorov [1] and Nikolai Smirnov (1900 - 1966).

Thus, because $\mathsf{F}_0 : \mathbb{R} \to (0,1)$ is one-to-one, we have

$$\sup_{x \in \mathbb{R}} \left| \widehat{\mathsf{F}}(x) - \mathsf{F}_0(x) \right| = \sup_{u \in (0,1)} \left| \widehat{\mathsf{G}}(u) - u \right|.$$

Although the computation of $\widehat{\mathsf{G}}$ surely depends on F_0 (since F_0 is used to define the U_i), clearly its distribution under F_0 does not. Indeed, it is simply the empirical distribution function of an iid sample of size n from Unif$(0,1)$. Note that the function $u \mapsto u$ coincides with the distribution function of Unif$(0,1)$ on the interval $(0,1)$. □

Remark 16.54. In the pre-computer days, the distribution of $\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}}, \mathsf{F}_0)$ under F_0 was obtained in the special case where F_0 is the distribution function of Unif$(0,1)$. There are recursion formulas for the exact computation of the p-value. The large-sample limiting distribution, known as the *Kolmogorov distribution*, was derived by Kolmogorov [143] and tabulated by Smirnov [213], and used for larger sample sizes. Details are provided in Problem 16.93.

R corner. In R, the test is implemented in ks.test, which returns a warning if there are ties in the data. The Kolmogorov distribution function is available in the package kolmim.

Remark 16.55. The recursion formulas just mentioned are only valid when there are no ties in the data, a condition which is satisfied (with probability one) in the context of Proposition 16.53, as F_0 is assumed there to be continuous. Although some strategies are available for handling ties, a calibration by Monte Carlo simulation is always available and accurate.

Problem 16.56. In R, write a function that mimics ks.test but instead returns a Monte Carlo p-value based on a specified number of replicates.

Cramér–von Mises test [89] This test uses the following dissimilarity measure

$$\Delta(\mathsf{F}, \mathsf{G})^2 := \mathbb{E}\left[(\mathsf{F}(X) - \mathsf{G}(X))^2 \right], \quad X \sim \mathsf{G}. \tag{16.21}$$

Problem 16.57. Show that

$$\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}}, \mathsf{F}_0)^2 = \frac{1}{12n^2} + \frac{1}{n} \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - \mathsf{F}_0(X_{(i)}) \right]^2.$$

As before, calibration is done by Monte Carlo simulation under F_0.

Problem 16.58. Show that Proposition 16.53 applies.

[89] Named after Harald Cramér (1893 - 1985) and Richard von Mises (1883 - 1953).

Anderson–Darling tests[90] These tests use dissimilarities of the form

$$\Delta(\mathsf{F},\mathsf{G}) := \sup_{x\in\mathbb{R}} w(x)\big|\mathsf{F}(x)-\mathsf{G}(x)\big|, \tag{16.22}$$

or of the form

$$\Delta(\mathsf{F},\mathsf{G})^2 := \mathbb{E}\big[w(x)^2(\mathsf{F}(X)-\mathsf{G}(X))^2\big], \quad X\sim\mathsf{G}, \tag{16.23}$$

where w is a (non-negative) weight function. In both cases, a common choice is

$$w(x) := [\mathsf{G}(x)(1-\mathsf{G}(x))]^{-1/2}.$$

This is motivated by the fact that, under the null, for any $x\in\mathbb{R}$, $\widehat{\mathsf{F}}_{\boldsymbol{X}}(x)-\mathsf{F}_0(x)$ has mean 0 and variance $\mathsf{F}_0(x)(1-\mathsf{F}_0(x))$.

Problem 16.59. Prove this assertion.

16.8.2 Confidence bands

Confidence bands are the equivalent of confidence intervals when the object to be estimated is a function rather than a real number. A confidence band can be obtained by inverting a test based on a measure of dissimilarity Δ, for example (16.20), following the process described in Section 12.4.6. In what follows, we assume that the underlying distribution, F, is continuous and that Δ is such that Proposition 16.53 applies.

Let δ_u denote the u-quantile of $\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}},\mathsf{F})$, which does not depend on F by Proposition 16.53. Within the space of continuous distribution functions, define the following (random) subset

$$\mathcal{B} := \big\{\mathsf{G} : \Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}},\mathsf{G}) \le \delta_{1-\alpha}\big\}.$$

This is the acceptance region for the level α test defined by Δ. This is thus the analog of (12.31), and following the arguments provided in Section 12.4.9, we obtain

$$\mathbb{P}_{\mathsf{F}}(\mathsf{F}\in\mathcal{B}) = \mathbb{P}_{\mathsf{F}}\big(\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}},\mathsf{F}) \le \delta_{1-\alpha}\big) = 1-\alpha.$$

($\mathbb{P}_{\mathsf{F}}$ denotes the distribution under F, meaning when $X_1,\dots,X_n$ are iid from F.)

Remark 16.60. The region $\mathcal{B}$ is called a 'confidence ban' because, if all the distribution functions in $\mathcal{B}$ are plotted, it yields a band as a subset of the plane (at least this is the case for the most popular measures of dissimilarity).

Problem 16.61. In the particular case of the supremum norm (16.20), the band is particularly easy to compute or

[90] Named after Theodore Anderson (1918 - 2016) and Donald Darling (1915 - 2014).

draw, because it can be defined pointwise. Indeed, show that in this case the band (as a subset of $\mathbb{R}^2$) is defined as

$$\{(x,p) : |p - \widehat{\mathsf{F}}(x)| \leq \delta_{1-\alpha}\}.$$

In R, write a function that takes in the data and the desired confidence level, and plots the empirical distribution function as a solid black line and the corresponding band in grey. [The function polygon can be used to draw the band.] Try out your function on simulated data from the standard normal distribution, with sample sizes $n \in \{10, 100, 1000\}$. Each time, overlay the real distribution function plotted as a red line. See Figure 16.4 for an illustration.

16.8.3 Tests based on the density

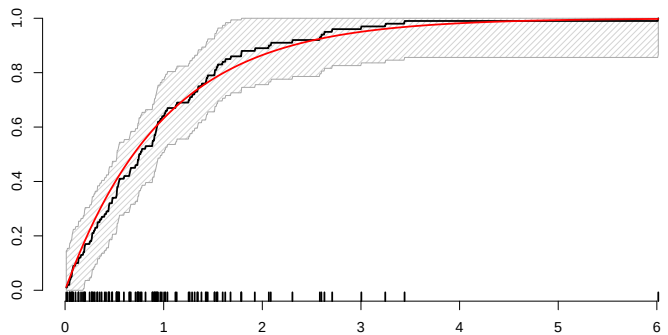
A goodness-of-fit test based on comparing densities rejects for large values of a test statistic of the form

$$\Delta(\hat{f}_{\boldsymbol{X}}, f_0),$$

where $\hat{f}_{\boldsymbol{X}}$ is an estimator for a density of the underlying distribution, f_0 is a density of F_0, and Δ is a measure of dissimilarity between densities such as the *total variation distance*,

$$\Delta(f,g) := \int_{\mathbb{R}} |f(x) - g(x)| \mathrm{d}x,$$

Figure 16.4: A 95% confidence band for the distribution function based on a sample of size $n = 1000$ from the exponential distribution with rate $\lambda = 1$. In black is the empirical distribution function, in red is the underlying distribution function, and in grey is the confidence band. The marks identify the sample points.



or the *Kullback–Leibler divergence*,

$$\Delta(f,g) := \int_{\mathbb{R}} f(x) \log(f(x)/g(x)) \mathrm{d}x. \tag{16.24}$$

Remark 16.62. If a histogram is used as a density estimator, the procedure is similar to binning the data and then using a goodness-of-fit test for discrete data (Sec-

tion 15.2). (Note that this approach is possible even if the null distribution does not have a density.)

Problem 16.63. In R, write a function that implements the procedure of Remark 16.62 using the likelihood ratio goodness-of-fit test detailed in Section 15.2. Use the function hist to bin the data and let it choose the bins automatically. The function returns a Monte Carlo p-value based on a specified number of replicates.

16.9 Censored observations

In some settings, the observations are censored. An emblematic example is that of clinical trials where patient survival is the primary outcome. In such a setting, patients might be lost in the course of the study for other reasons, such as the patient moving too far away from any participating center, or simply by the termination of the study. Other fields where censoring is common include, for example, quality control where the reliability of some manufactured item is examined. The study of such settings is called *Survival Analysis*.

We consider a model of independent right-censoring. In the context of a clinical trial, let T_i denote the time to event (say death) for Subject i, with $T_1, \dots, T_n$ assumed iid from some distribution H. These are not observed directly as they may be subject to censoring. Let $C_1, \dots, C_n$ denote the censoring times, assumed to be iid from some distribution G. The actual observations are $(X_1, \delta_1), \dots, (X_n, \delta_n)$, where

$$X_i = \min(T_i, C_i), \quad \delta_i = \{X_i = T_i\},$$

so that $\delta_i = 1$ indicates that the ith case was observed uncensored. The goal is to infer the underlying distribution H, or some features of H such as its median.

Problem 16.64. Recall the definition of the survival function given in (4.14). Show that $X_1, \dots, X_n$ are iid with survival function $\bar{\mathsf{F}}(x) := \bar{\mathsf{H}}(x)\bar{\mathsf{G}}(x)$.

16.9.1 Kaplan-Meier estimator

We consider the case where the variables are discrete. We assume they are supported on the positive integers without loss of generality. In a survival context, this is the case, for example, when survival time is the number days to death since the subject entered the study. In that special case, it is possible to derive the maximum likelihood estimator for H. Denote the corresponding survival and mass functions by $\bar{\mathsf{H}}(t) = 1 - \mathsf{H}(t)$ and $h(t) = \mathsf{H}(t) - \mathsf{H}(t-1)$, respectively. Let $I = \{i : \delta_i = 1\}$, which indexes the uncensored survival times.

Problem 16.65. Show that the likelihood is

$$\mathsf{lik}(\mathsf{H}) := \prod_{i \in I} h(X_i) \times \prod_{i \notin I} \bar{\mathsf{H}}(X_i). \tag{16.25}$$

As usual, we need to maximize this with respect to H. For a positive integer t, define the hazard rate

$$\begin{aligned} \lambda(t) &= \mathbb{P}(T = t \mid T > t-1) \\ &= 1 - \mathbb{P}(T > t \mid T > t-1) \\ &= h(t)/\bar{\mathsf{H}}(t-1). \end{aligned}$$

By the Law of Multiplication (1.21),

$$\bar{\mathsf{H}}(t) = \mathbb{P}(T > t) = \prod_{k=1}^{t} \mathbb{P}(T > k \mid T > k-1) \tag{16.26}$$

$$= \prod_{k=1}^{t} (1 - \lambda(k)), \tag{16.27}$$

so that

$$\begin{aligned} h(t) &= \bar{\mathsf{H}}(t-1) - \bar{\mathsf{H}}(t) \\ &= \lambda(t) \prod_{k=1}^{t-1} (1 - \lambda(k)). \end{aligned}$$

In particular, λ determines h, and therefore H.

Problem 16.66. Deduce that the likelihood, as a function of λ, is given by

$$\mathsf{lik}(\lambda) = \prod_{t \geq 1} \lambda(t)^{D_t} (1 - \lambda(t))^{Y_t - D_t}.$$

where D_t is the number of events at time t, meaning

$$D_t := \#\{i : X_i = t, \delta_i = 1\},$$

and Y_t is the number of subjects still alive (said to be *at risk*) at time t, meaning

$$Y_t := \#\{i : X_i \geq t\}.$$

Then show that the maximizer is $\hat{\lambda}$ given by

$$\hat{\lambda}(t) := D_t / Y_t.$$

The corresponding estimate for $\bar{\mathsf{H}}$ is obtained by plugging $\hat{\lambda}$ in (16.27), resulting in the MLE for $\bar{\mathsf{H}}$ being

$$\bar{\mathsf{H}}^{\mathrm{KM}}(t) := \prod_{k=1}^{t} (1 - D_k / Y_k),$$

known as the *Kaplan-Meier estimator*.

Remark 16.67. Bootstrap confidence bands for this estimator are discussed in [4]. The situation is a bit complex and we omit details.

16.10 Further topics

16.10.1 Reduction to a uniform sample

Assume that $X_1, \ldots, X_n$ are iid from a continuous distribution F. If we define $U_i := \mathsf{F}(X_i)$, $U_1, \ldots, U_n$ are iid Unif$(0,1)$. Thus the study of the order statistics $(X_{(1)}, \ldots, X_{(n)})$ reduces to the study of the uniform orders statistics $(U_{(1)}, \ldots, U_{(n)})$, which is the object of an area known as *Empirical Process Theory*. To emphasize the total sample size n, the notation $(U_{(1:n)}, \ldots, U_{(n:n)})$ is often used.

It is known, for example, that $U_{(i:n)}$ has the beta distribution with parameters $(i, n+1-i)$, which in particular implies that

$$\mathbb{E}(U_{(i:n)}) = \frac{i}{n+1},$$

and

$$\mathrm{Var}(U_{(i:n)}) = \frac{i(n+1-i)}{(n+1)^2(n+2)} \leq \frac{1}{4n+8}.$$

Problem 16.68. Prove that $U_{(i:n)}$ is concentrated near its mean by showing that, for all $t > 0$,

$$\mathbb{P}\big(|U_{(i:n)} - i/(n+1)| \geq t/\sqrt{n}\big) \leq 1/4t^2.$$

(See also Problem 16.96.)

It is also known that $U_{(i:n)}$ is approximately normal when n is large and i/n is not too close to 0 or 1. In particular, this implies the following (via the delta method).

Proposition 16.69 (Asymptotic normality of the sample median)**.** *Assume that $(X_i : i \geq 1)$ are iid with median θ and distribution function F having a strictly positive derivative at θ (denoted $f(\theta)$). Then*

$$\sqrt{n}\big(Med(X_1, \ldots, X_n) - \theta\big) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{1}{4\,f(\theta)^2}\Big),$$

as $n \to \infty$.

As mentioned in Remark 16.51, when testing for a particular null distribution F_0, we can work with the transformed sample $U_i := \mathsf{F}_0(X_i)$ and test for uniformity. The test that Berk and Jones [15] proposed directly exploits the fact that distribution under the null of each order statistic is known. Specifically, the test rejects for small values of $\min(S^-, S+)$, where $S^+ := \min_i \mathsf{G}_i(U_{(i)})$ and $S^- := \min_i(1 - \mathsf{G}_i(U_{(i)}))$, where G_i is the distribution function of Beta$(i, n+1-i)$. The asymptotic distribution of this test statistic under the null hypothesis is derived in [170].

16.10.2 Other forms of bootstrap

There are other forms of bootstrap besides the one that relies on the empirical distribution (often called the *empirical bootstrap*). We briefly describe a few.

Smooth bootstrap The empirical distribution is always available as an estimate of the underlying distribution, but being discrete, it is not 'smooth'. A smooth bootstrap is based on a smoother estimate of the underlying distribution, for example, the piecewise linear empirical distribution function (Remark 16.10) or a kernel density estimate (Section 16.10.6).

A common way to implement the latter is as follows. Let $x_1, \dots, x_n$ denote the observations. Given a distribution with density K, generate a (smooth) bootstrap sample by first generating an iid sample from K, denoted $W_1, \dots, W_n$, and then adding that to the observations, resulting in the bootstrap sample $X_1^*, \dots, X_n^*$ with $X_i^* := x_i + W_i$. By (6.11) and Section 16.10.6, conditional on the observations, $X_1^*, \dots, X_n^*$ are iid with density

$$f^*(x) := \frac{1}{n} \sum_{i=1}^{n} \mathsf{K}(x_i - x),$$

which is a kernel density estimate.

Remark 16.70. When the task is to estimate a parameter of interest, $\varphi(\mathsf{F})$, with φ not defined on discrete distributions, the empirical bootstrap is not applicable, but a smooth bootstrap might.

Parametric bootstrap If we know (or rather, if we are willing to assume) that F is in some parametric family of distributions, say $\{\mathsf{F}_\theta : \theta \in \Theta\}$, then a possible approach is to estimate F with $\mathsf{F}_{\hat{\theta}}$, where $\hat{\theta}$ is an estimator for θ, for example, the MLE. (This is in fact what we did in Section 15.3.2 in the context of discrete distributions.)

16.10.3 Method of moments

Suppose an experiment results in an iid sample, denoted $X_1, \dots, X_n$, having distribution P_θ on $\mathbb{R}$, where $\theta \in \Theta$ is unknown and needs to be estimated. This is for example the case of the binomial experiment of Example 12.2 if we define $X_i = 1$ when the ith trial results in heads and $X_i = 0$ otherwise. Let $\widehat{\mathsf{P}}_{\boldsymbol{X}}$ denote the empirical distribution.

We already saw maximum likelihood estimation. A competing approach is the *method of moments*. Having chosen some distributional features of interest (e.g., various moments), the idea is to find a value of θ such that, in terms these features, P_θ is close to the empirical distribution.

A feature here is a real-valued function on distributions on $\mathbb{R}$. Let $\Lambda_1, \ldots, \Lambda_k$ denote k such features. Based on these, define the following estimator

$$S := \arg\min_{\theta \in \Theta} \sum_{j=1}^{k} \left[\Lambda_j(\mathsf{P}_\theta) - \Lambda_j(\widehat{\mathsf{P}}_{\boldsymbol{X}})\right]^2.$$

(This assumes the minimization problem has a unique solution.)

Remark 16.71 (Classical method of moments)**.** The features are traditionally chosen to be moments, meaning, $\Lambda_j(\mathsf{P}) = \mathbb{E}_{\mathsf{P}}(X^j)$, where $\mathbb{E}_{\mathsf{P}}$ denotes the expectation under $X \sim \mathsf{P}$.

In some classical settings, it is possible to find $\theta \in \Theta$ such that $\Lambda_j(\mathsf{P}_\theta) = \Lambda_j(\widehat{\mathsf{P}}_{\boldsymbol{X}})$ for all $j = 1, \ldots, k$.

Problem 16.72 (Binomial experiment)**.** Consider the binomial experiment of Example 12.2. Apply the method of moments estimator based solely on the 1st moment, meaning that $k = 1$ and $\Lambda_1(\mathsf{P}) = \mathsf{mean}(\mathsf{P})$, and show that the resulting estimator coincides with the maximum likelihood estimator.

Problem 16.73. More generally, assume that $\Theta \subset \mathbb{R}$ and that θ is the mean of P_θ. Show that the classical method of moments estimate for θ, with $k = 1$, is the sample mean. Extend this to the case where $\Theta = \mathbb{R} \times \mathbb{R}_+$ and $\theta = (\mu, \sigma^2)$, with μ being the mean and σ^2 the variance of P_θ.

16.10.4 Prediction intervals

In Section 16.3.2 and in Section 16.6 our goal was to construct a confidence interval for the location parameter of interest, respectively the median and the mean. Consider instead the problem of constructing an interval for a new observation sampled from the same underlying distribution. Such an interval is called a *prediction interval*.

In what follows, we let $X_1, \ldots, X_n, X_{n+1}$ be iid from a distribution P on the real line, where $\boldsymbol{X}_n := (X_1, \ldots, X_n)$ plays the role of the available sample, while X_{n+1} plays the role of a new datum.

Problem 16.74. Suppose that P is the normal distribution with unknown mean μ and unknown variance σ^2. Let $\bar{X}_n$ and S_n denote the sample mean and standard deviation based on $\boldsymbol{X}_n$. Show that

$$X_{n+1} \in \left[\bar{X}_n - t^{n-1}_{1-\alpha/2} S_n \sqrt{\tfrac{n-1}{n+1}}, \bar{X}_n - t^{n-1}_{\alpha/2} S_n \sqrt{\tfrac{n-1}{n+1}}\right]$$

with probability $1 - \alpha$.

Compared to the confidence interval for the mean given in (16.13), the prediction interval above is much wider. In

fact, its half-width converges (in probability) to $\sigma z_{1-\alpha/2}$, and therefore does not converge to zero as $n \to \infty$. (As before, z_u denotes the u-quantile of the standard normal distribution.) This is to be expected. Indeed, even when μ and σ are known, we cannot do better asymptotically.

Problem 16.75. Show that $[\mu - z_{1-\alpha/2}\sigma, \mu - z_{\alpha/2}\sigma]$ is the shortest interval I such that

$$\mathbb{P}(X \in I) \geq 1 - \alpha, \tag{16.28}$$

when X is sampled from $\mathcal{N}(\mu, \sigma^2)$.

When no parametric family is assumed, one can rely on the empirical distribution to obtain a prediction interval. One way to do so is via the appropriate sample quantiles.

Problem 16.76. Let $\widehat{\mathsf{F}}_n^-$ denote the empirical quantile function based on $\boldsymbol{X}_n$. Prove that

$$X_{n+1} \in [\widehat{\mathsf{F}}_n^-(\alpha/2), \widehat{\mathsf{F}}_n^-(1-\alpha/2)] \tag{16.29}$$

with probability tending to $1 - \alpha$ as $n \to \infty$. Propose another variant such that, asymptotically, the interval is shortest (at the same confidence level $1 - \alpha$).

The prediction interval (16.29) has asymptotically the prescribed confidence level. If the level must be guaranteed in finite samples, one can use the Dvoretzky–Kiefer–Wolfowitz bound.

Problem 16.77. Derive a prediction interval based on Theorem 16.9 that satisfies the prescribed level of confidence in finite samples. [The shorter the better.] How does the interval behave in the large-sample limit?

The following may provide a more satisfying option.

Proposition 16.78. *With $X_{(1)} \leq \cdots \leq X_{(n)}$ denoting the ordered sample, it holds that*

$$X_{n+1} \in \left[X_{(\lfloor n\alpha/2 \rfloor)}, X_{(\lceil n(1-\alpha/2) \rceil)}\right] \tag{16.30}$$

with probability at least $1 - \alpha$.

Proof. For a sample $x_1, \ldots, x_n \in \mathbb{R}$, let

$$I(x_1, \ldots, x_n) = \left[x_{(\lfloor n\alpha/2 \rfloor)}, x_{(\lceil n(1-\alpha/2) \rceil)}\right],$$

where $x_{(1)} \leq \cdots \leq x_{(n)}$ denote the ordered sample. For $i \in \{1, \ldots, n+1\}$, let

$$Y_i = \begin{cases} 1 & \text{if } X_i \in I(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{n+1}); \\ 0 & \text{otherwise.} \end{cases}$$

Because the event (16.30) can be equivalently stated as $Y_{n+1} = 1$, it suffices to prove that $\mathbb{P}(Y_{n+1} = 1) \geq 1 - \alpha$. Since Y_{n+1} takes values in $\{0, 1\}$, $\mathbb{P}(Y_{n+1} = 1) = \mathbb{E}(Y_{n+1})$.

Although $Y_1, \dots, Y_{n+1}$ are not independent, they are exchangeable, and in particular they have the same expectation, so that

$$\mathbb{E}(Y_{n+1}) = \mathbb{E}\left(\frac{Y_1 + \cdots + Y_{n+1}}{n+1}\right).$$

We then conclude with the following problem. □

Problem 16.79. For $x_1, \dots, x_{n+1} \in \mathbb{R}$, define $y_i = 1$ if $x_i \in I(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$, and $y_i = 0$ otherwise. Show that

$$\frac{y_1 + \cdots + y_{n+1}}{n+1} \geq 1 - \alpha.$$

Remark 16.80 (Conformal prediction)**.** The approach underlying the construction of the prediction interval (16.30) can be seen as an example of *conformal prediction* [208, 239], which may be seen as a general approach based on inverting a permutation test for goodness-of-fit comparing samples $\{X_1, \dots, X_n\}$ and $\{X_{n+1}\}$ using an approach similar to that of Section 12.4.6 for inverting a test to obtain a confidence interval.

16.10.5 Testing against a family of distributions

Section 16.8 dealt with a situation where there is a single null distribution to test against. We saw in Remark 16.51 that, as long as the null distribution is continuous, we can assume it to be the uniform distribution on $[0, 1]$. In fact, it can be assumed to be any other continuous distribution, e.g., the standard normal distribution.

We now consider a situation where there is a null family of distributions to test against. For example, we may want to know whether the sample was generated by the uniform distribution on an unspecified interval (Problem 22.9). Or we may need to decide whether the sample comes from a normal distribution, often referred to as *testing for normality*.

Suppose, therefore, that we have access to a sample, $\boldsymbol{X} = (X_1, \dots, X_n)$, assumed iid from a continuous distribution, and our task is to test whether the underlying distribution is in some given (null) family of distributions, $\{\mathsf{F}_\theta : \theta \in \Theta\}$, against the alternative that is not in that family.

Assuming we have an estimator for θ, denoted $S(\boldsymbol{X})$, and working with distribution functions as in Section 16.8.1, a plug-in approach leads to using a test statistic of the form

$$\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}}, \mathsf{F}_{S(\boldsymbol{X})}). \tag{16.31}$$

We still reject for large values of this statistic, although the main difference here is that a p-value is obtained by bootstrap, since the null distribution needs to be estimated.

LILLIEFORS TEST This is a test for normality that derives from using the supremum norm as a measure of dissimilarity (16.20). The parameters (mean and variance) are obtained by maximum likelihood or the method of moments.

Although the p-value is in principle obtained by bootstrap as explained above, here the test statistic has the same distribution under the null hypothesis regardless of the sample values. This is because the normal family is a location-scale family, and the supremum norm is invariant with respect to affine transformations.

Problem 16.81. Prove that the statistic (16.31), with Δ denoting the supremum norm (16.20), has the same distribution under any normal distribution.

Problem 16.82. Compare the Lilliefors test to the Kolmogorov–Smirnov test against the estimated null distribution.

R corner. In R, the test is implemented in the function lillie.test in the package nortest.

16.10.6 KERNEL DENSITY ESTIMATION

Consider a sample, $X_1, \dots, X_n$, drawn iid from a density f that we want to estimate. Let K be a function on $\mathbb{R}$, which here plays the role of *kernel function*, and for $a > 0$ define

$$\mathsf{K}_a(x) = a^{-1}\mathsf{K}(x/a). \tag{16.32}$$

The key is the following result.

Proposition 16.83. *Suppose that* $\int_{\mathbb{R}} \mathsf{K}(x)\mathrm{d}x = 1$ *and* f *is continuous at* x *and bounded on* $\mathbb{R}$. *Then*

$$\int_{\mathbb{R}} \mathsf{K}_a(z-x)f(z)\mathrm{d}z \to f(x), \quad \text{as } a \to 0.$$

Since,

$$\int_{\mathbb{R}} \mathsf{K}_a(z-x)f(z)\mathrm{d}z = \mathbb{E}_f(\mathsf{K}_a(X-x)),$$

where $\mathbb{E}_f$ denotes the expectation with respect to $X \sim f$, this suggests estimating f with

$$\hat{f}_a(x) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{K}_a(x_i - x). \tag{16.33}$$

The method has one parameter $a > 0$, called the *bandwidth*, that plays the exact same role as the bin size in the construction of a histogram (with bins of equal size). In fact, when K is the so-called rectangular (aka flat) kernel, namely, $\mathsf{K}(x) = \{x \in [-1/2, 1/2]\}$, the estimate is close to a histogram with bin size h.

R corner. The function density in R offers a number of choices for the kernel K. The default is the Gaussian kernel $\mathsf{K}(x) := \frac{1}{\sqrt{2\pi}} \exp(-x^2/2)$.

A kernel density estimate is at least as smooth the kernel used to define it. In particular, a kernel density estimate with the Gaussian kernel is infinitely differentiable.

Desiderata. For reasons of approximation accuracy, it is desired that the kernel is most typically chosen even and either compactly supported or fast-decaying.

CHOICE OF BANDWIDTH The choice of bandwidth a has generated a lot of proposals. It is directly related to the choice of bin size in the construction of a histogram, and the same chicken-and-egg situation arises.

For example, here too, the optimal choice for a is of order $n^{-1/3}$ when f has bounded slope. To see why, suppose that f is bounded by c_0 and has slope bounded by c_1 everywhere, meaning

$$f(x) \le c_0, \quad \text{for all } x \in \mathbb{R}, \tag{16.34}$$

$$|f(x) - f(z)| \le c_1 |x - z|, \quad \text{for all } x, z \in \mathbb{R}. \tag{16.35}$$

Assume that the kernel is non-negative, has support in $[-1/2, 1/2]$, and is bounded by some $c_2 > 0$ in absolute value.

The mean squared error at $x \in \mathbb{R}$ is

$$\mathsf{mse}_a(x) := \mathbb{E}\big[(\hat{f}_a(x) - f(x))^2\big],$$

where the expectation is with respect to the sample defining $\hat{f}_a$.

Problem 16.84. Derive the following *bias-variance decomposition*

$$\mathsf{mse}_a(x) = \big(\mathbb{E}[\hat{f}_a(x)] - f(x)\big)^2 + \mathrm{Var}[\hat{f}_a(x)].$$

For the mean, by (16.33),

$$\begin{aligned}
\mathbb{E}[\hat{f}_a(x)] - f(x) &= \mathbb{E}[\mathsf{K}_a(X - x)] - f(x) \\
&= \int_{\mathbb{R}} \mathsf{K}_a(z - x) f(z) \mathrm{d}z - f(x) \\
&= \int_{\mathbb{R}} \mathsf{K}_a(z - x)\big(f(z) - f(x)\big) \mathrm{d}z,
\end{aligned}$$

using the fact that K_a integrates to 1. Hence, using Jensen's inequality,

$$\begin{aligned}
\big|\mathbb{E}[\hat{f}_a(x)] - f(x)\big| &\le \int_{\mathbb{R}} \mathsf{K}_a(z - x)\big|f(z) - f(x)\big| \mathrm{d}z \\
&\le \int_{\mathbb{R}} \mathsf{K}_a(z - x) c_1 |z - x| \mathrm{d}z \\
&\le c_1 a \int_{[-a/2, a/2]} \mathsf{K}_a(z) \mathrm{d}z \\
&= c_1 a,
\end{aligned}$$

using the bound on the slope of f, the fact that the kernel is non-negative, supported on $[-1/2, 1/2]$, and integrates to 1.

For the variance, by (16.33) and independence,

$$\mathrm{Var}[\hat{f}_a(x)] = \frac{1}{n}\mathrm{Var}[\mathsf{K}_a(X-x)],$$

with

$$\begin{aligned}\mathrm{Var}[\mathsf{K}_a(X-x)] &\le \mathbb{E}[\mathsf{K}_a(X-x)^2] \\ &= \int_{\mathbb{R}} \mathsf{K}_a(z-x)^2 f(z)\mathrm{d}z \\ &\le \frac{c_2}{a} c_0 \int_{\mathbb{R}} \mathsf{K}_a(z-x)\mathrm{d}z \\ &= c_2 c_0 / a,\end{aligned}$$

using the bound on K and the bound on f, and the fact that the kernel integrates to 1.

Thus,

$$\mathsf{mse}_a(x) \le (c_1 a)^2 + \frac{c_2 c_0}{na},$$

and the right-hand side is minimized at $c_3 n^{-1/3}$ where $c_3 := (c_0 c_2 / 2c_1^2)^{1/3}$.

These calculations are crude and refinements are definitely possible. However, the order of magnitude of the optimal bandwidth is known to be $\propto n^{-1/3}$, with a multiplicative constant that depends on the (unknown) density. That constant is important in practice and makes it necessary to choose the bandwidth based on the data.

Cross-validation A popular way to choose the bandwidth is by *cross-validation (CV)*. We present a particular variant called *leave-one-out cross-validation*, proposed by Rudemo in [199]. The idea is to choose a to minimize the following *risk*

$$\mathcal{R}(a) := \mathbb{E}\Big[\int (\hat{f}_a(x) - f(x))^2 \mathrm{d}x\Big],$$

where the expectation is with respect to the sample that underlies the estimate $\hat{f}_a$. Based on this risk, Rudemo proposes the following choice of bandwidth

$$\hat{a} := \arg\min_{a>0} Q(a),$$

$$Q(a) := \int \hat{f}_a(x)^2 \mathrm{d}x - \frac{2}{n-1}\sum_{i=1}^{n} \hat{f}_a(X_i).$$

Problem 16.85. Relate $\mathbb{E}[Q(a)]$ to $\mathcal{R}(a)$.

R corner. This is very close to how the R function bw.ucv selects the bandwidth. A more faithful implementation may be found in the function h.ucv in the kedd package.

16.10.7 Monotonic density estimation

Kernel density estimation is, as we saw, founded on the implicit assumptions that the underlying density has a certain degree of smoothness. An alternative is to assume that the density has a certain *shape*. We present the simplest, and most famous example, where the underlying density is supported on $\mathbb{R}_+$ and assumed to be monotone (and therefore non-increasing).

It so happens that there is a maximum likelihood estimator for this model, proposed by Ulf Grenander (1923 - 2016) [112]. The likelihood is here defined as

$$\arg\max_{f \in \mathcal{F}} \prod_{i=1}^{n} f(x_i),$$

where $x_1, \ldots, x_n$ denote the observations and $\mathcal{F}$ denotes the class of monotone densities on $\mathbb{R}_+$. It turns out to be maximized by the first derivative of the least concave majorant of the empirical distribution function. It is a decreasing piecewise-constant function.

R corner. The grenander function in the fdrtool package computes this estimator.

16.11 Additional problems

Problem 16.86. In R, for $n \in \{10, 100, 1000\}$, generate n points from the uniform distribution on $[0, 1]$. Draw the empirical distribution function (solid line) and overlay the actual distribution function (dashed line). Do it several times for each n to get a feel for the randomness. Repeat with the standard normal distribution.

Problem 16.87. In R, write a function which behaves as ecdf but returns the piecewise linear variant of Remark 16.10 instead. In addition, write another function which behaves as plot.ecdf. To each plot of Problem 16.86, add (in a different color) a graph of this variant.

Problem 16.88. In R, generate a sample of size $n = 10$ from the standard normal distribution. Plot the quantile function, type 1, 4, 5, in red, green, and blue, respectively. Add the underlying quantile function (given by qnorm). Make sure to use a fine grid, say, 1000 points covering $[0, 1]$, for otherwise visual artifacts will result. Add dotted vertical lines at k/n for $k = 0, 1, \ldots, n$, and dotted horizontal lines at the data points. Repeat with $n = 100$.

Problem 16.89. In R, generate a sample of size $n \in \{100, 1000\}$ from the standard normal distribution. Plot the histogram in various ways, each time overlaying the

density. Do it several times in each setting to get a feel for the randomness. Try different methods for choosing the bins.

Problem 16.90. Suppose that P_θ is the exponential distribution with rate $\theta > 0$, and that we have an iid sample of size n, $X_1, \dots, X_n$, from it. Suppose we want to estimate $\varphi := \theta^{1/2}$.

(i) Show that the MLE for θ is $\bar{X}^{-1}$, where $\bar{X}$ is the sample mean. We thus use $\bar{X}^{-1/2}$ to estimate φ.

(ii) Compute its bias by numerical integration. [Note that the family is a scale family, so that the bias under θ can be obtained from the bias under $\theta = 1$.]

(iii) In R, for $n \in \{10, 20, 30, 50\}$, under $\theta = 1$, estimate that quantity by Monte Carlo, using $B = 10^4$ replicates.

(iv) Now estimate that quantity by bootstrap based on $B = 10^4$ replicates. Do this for various choices of n. Repeat a few times to get a feel for the randomness and compare with the value obtained by numerical integration.

Problem 16.91. Repeat Problem 16.90 with the variance in place of the bias.

Problem 16.92. Show that the variant (16.12) is unbiased. However, 'unbiased' does not mean 'better', and indeed, show that for the estimation of the variance in the normal location-scale family, (16.10) has smaller mean squared error than (16.12).

Problem 16.93 (Kolmogorov distribution)**.** In [143], Kolmogorov derived the null distribution of the test statistic (16.20). He showed that, based on an iid sample of size n from a continuous distribution F, the empirical distribution function $\widehat{\mathsf{F}}_n$ as a random function satisfies

$$\lim_{n\to\infty} \mathsf{P}\Big(\sup_{x\in\mathbb{R}} \big|\widehat{\mathsf{F}}_n(x) - \mathsf{F}(x)\big| \geq t/\sqrt{n} \Big) = 2 \sum_{k\geq 1} (-1)^{k-1} \exp(-2k^2 t), \quad (16.36)$$

for all $t \geq 0$. In R, perform simulations to probe the accuracy of this limit for various choices of sample size n.

Problem 16.94 (A failure of the bootstrap)**.** Consider the situation where $X_1, \dots, X_n$ are iid normal with mean θ and variance 1. It is desired to provide a confidence interval for $|\theta|$. A natural estimator is the plug-in estimator $|\bar{X}|$. The bootstrap confidence interval of Section 16.6.3 is based on estimating the distribution of $|\bar{X}| - |\theta|$ by the bootstrap distribution of $|\bar{X}^*| - |\bar{X}|$. It happens to fail when $\theta = 0$, because the absolute value, as a function, is not smooth at the origin.

(i) Show that the MLE for $|\theta|$ is $|\bar{X}|$.

(ii) Compute the distribution function of $\sqrt{n}(|\bar{X}| - |\theta|)$. Specialize to the case where $\theta = 0$ and draw it.

(iii) In R, generate a sample of size $n = 10^6$ and estimate the bootstrap distribution $\sqrt{n}(|\bar{X}^*| - |\bar{X}|)$ using $B = 1000$ Monte Carlo replicates. Add the resulting empirical distribution function to the plot.

Problem 16.95. In R, do the following. Generate a sample of size $n \in \{10, 20, 50, 100, 200, 500, 1000\}$ from the standard normal distribution (so that $\mu = 0$ and $\sigma = 1$). Set the confidence level at 0.95.

(i) Compute the Student confidence interval. This is the gold standard if the sample is known to be normal and the variance is unknown.

(ii) Compute the bootstrap confidence interval using the function implemented in Problem 16.45.

(iii) Compute the bootstrap Studentized confidence interval using the function implemented in Problem 16.46.

Each time, record whether the true mean is in the interval and measure the length of the interval. After doing that 200 times, for each of the three confidence interval constructions, display the fraction of times the interval contained the true mean and plot a histogram of its length.

Problem 16.96. Derive a bound that is sub-exponential in t for the probability that appears in Problem 16.68.

Problem 16.97 (Student test)**.** Consider a normal experiment where we observe a realization of $X_1, \dots, X_n$ assumed an iid sample from $\mathcal{N}(\mu, \sigma^2)$. Both parameters are unknown. Our goal is to test $\mu = \mu_0$ for some given $\mu_0 \in \mathbb{R}$. The *Student test* (aka *t-test*) rejects for large values of $|T|$ where $T := (\bar{X} - \mu_0)/S$, with $\bar{X}$ being the sample mean and S being the sample standard deviation.

(i) Show that there is a constant c_n such that, under the null hypothesis, $c_n T$ has the Student distribution with $n - 1$ degrees of freedom.

(ii) Show that this test corresponds to the likelihood ratio test under the present model.

Problem 16.98. In R, do the following. Generate a sample of size n from some exponential distribution (say with rate $\lambda = 1$, although this is inconsequential). Plot a histogram. Compute the maximum likelihood estimator and overlay the corresponding density. Then compute the Grenander's estimator and overlay the corresponding density (in a different color). Repeat several times for several choice of sample size n.

Problem 16.99 (Log-concave densities)**.** A function $f : \mathbb{R} \to \mathbb{R}_+$ is said to be *log-concave* if $\log f$ is concave. Any normal distribution is log-concave, for example. Just as for monotonic densities, the class of log-concave densities

admits a maximum likelihood estimator, which also turns out to be piecewise linear. This estimator is available in the package longcondens. In R, do the following. Generate a sample of size n from some normal distribution (say, standard normal, although this is inconsequential). Plot a histogram. Compute the maximum likelihood estimator among normal distributions and overlay the corresponding density. Then compute the maximum likelihood estimator among log-concave distributions and overlay the corresponding density (in a different color). Repeat several times for several choice of sample size n.

Problem 16.100. Propose and study, analytically or via computer simulations, a test for goodness-of-fit based on the characteristic function, that is, based on a test statistic comparing the sample characteristic function (i.e., the characteristic function of the empirical distribution) and the null characteristic function (i.e., the characteristic function of the null distribution).

Problem 16.101 (A uniform law of large numbers)**.** The Glivenko–Cantelli Theorem is an example of a uniform law of large numbers. Here is another example, due to Jennrich [136]. Assume that X is a random variable on some probability space, and that Θ is a compact parameter space. Consider a (measurable) function $g : \mathbb{R} \times \Theta \to \mathbb{R}$ such that $\theta \mapsto g(x, \theta)$ is continuous for all $x \in \mathbb{R}$; and $|g(x, \theta)| \leq h(x)$ for all x and all θ, with $\mathbb{E}[h(X)] < \infty$. Then, if (X_i) are iid copies of X,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta) - \mathbb{E}[g(X, \theta)] \right| \xrightarrow{\mathrm{P}} 0, \quad \text{as } n \to \infty.$$

Prove this result.

Problem 16.102. Verify as many statements in [52] as you can, in particular those in dimension one.

CHAPTER 17

MULTIPLE NUMERICAL SAMPLES

In the previous chapter we considered an experiment resulting in an iid real-valued sample from an unknown distribution and the task was to infer this distribution or some of its features. Here we consider a setting where either two or more real-valued samples are observed, and the goal is to compare the underlying distributions that generated the samples. Randomized clinical trials are a rich source of examples (Section 11.2.5). Although it is somewhat more complicated, the methodology parallels that of Chapter 16. Permutation tests play a central role in goodness-of-fit testing.

Example 17.1 (Weight loss maintenance)**.** In [227], a "two-phase trial in which 1032 overweight or obese adults with hypertension, dyslipidemia, or both, who had lost at least 4 kg during a 6-month weight loss program (Phase 1) were randomized to a weight-loss maintenance intervention (Phase 2)." There were 3 intervention groups: monthly personal contact, unlimited access to an interactive technology, or self-directed (which served as control).

Subjects were followed for a total of 30 months.

Example 17.2 (Manual vs. automated blood pressure measurement)**.** The paper [172] reports on a clinical trial where the objective was to "compare the quality and accuracy of manual office blood pressure and automated office blood pressure". It was a cluster randomized trial where 67 medical practices in eastern Canada were randomized to measuring blood pressure either manually or using a BpTRU device. (The awake ambulatory blood pressure was used as gold standard.)

We start with two groups to ease the presentation, and then extend the narrative to multiple groups. We typically leave the dependence on the sample sizes implicit to lighten up the notation.

17.1 Inference about the difference in means

We have two samples, $X_{1,1}, \ldots, X_{n_1,1}$ assumed to be iid from F_1 and, independently, $X_{1,2}, \ldots, X_{n_2,2}$ assumed to be iid from F_2. The sample sizes are therefore n_1 and n_2, respectively, and we let $n := n_1 + n_2$ denote the total sample size.

Assume that F_1 has mean μ_1 and that F_2 has mean μ_2. We are interested in the difference

$$\delta := \mu_1 - \mu_2.$$

The plugin estimator is the difference of the sample means, or in formula

$$D := \bar{X}_1 - \bar{X}_2,$$

where

$$\bar{X}_j := \frac{1}{n_j} \sum_{i=1}^{n_j} X_{i,j}.$$

In particular, D is unbiased for δ.

In the following, we present various ways of building a confidence interval based on that estimator.

17.1.1 Normal confidence interval

Assume that F_1 has variance σ_1^2 and F_2 has variance σ_2^2. Note that D has variance

$$\gamma^2 := \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Problem 17.3. Prove, using the Central Limit Theorem, that $(D - \delta)/\gamma$ is asymptotically standard normal as $n_1, n_2 \to \infty$. Is this still true if one of the sample sizes remains bounded?

This normal limit implies that, if z_u denotes the u-quantile of the standard normal distribution,

$$\delta \in \Big[D - z_{1-\alpha/2}\,\gamma, D - z_{\alpha/2}\,\gamma\Big], \tag{17.1}$$

with probability converging to $1-\alpha$ as $n_1, n_2 \to \infty$.

Typically, σ_1^2 and σ_2^2 are unknown, in which case this interval is not a confidence interval. We obtain a bona fide confidence interval by plugging the sample variances, S_1^2 and S_2^2, in place of the variances, where

$$S_j^2 := \frac{1}{n_j}\sum_{i=1}^{n_j}(X_{i,j} - \bar{X}_j)^2, \tag{17.2}$$

so that γ is estimated by

$$G := \big[S_1^2/n_1 + S_2^2/n_2\big]^{1/2}.$$

Then by Slutsky's theorem, $(D-\delta)/G$ is asymptotically standard normal as the sample sizes diverge to infinity, and this implies that

$$\delta \in \Big[D - z_{1-\alpha/2}G, D - z_{\alpha/2}G\Big], \tag{17.3}$$

with probability converging to $1-\alpha$ as $n_1, n_2 \to \infty$. (This interval is a confidence interval since it can be computed from the data.)

Student confidence interval As in the one-sample setting, using quantiles from the Student distribution is the common practice, despite the fact that even with normal samples the Student distribution is only an approximation. (In that case, the sample variance is defined as in(16.12).) The computation of the number of degrees of freedom identifying the Student distribution is a bit involved. This derivation is due to Welch [251] and for this reason the test sometimes bares his name. A conservative choice is $\min(n_1, n_2) - 1$.

R corner. These Student-Welch confidence intervals, as well as the corresponding tests, can be computed using the function `t.test`.

17.1.2 Bootstrap Studentized confidence interval

There are analogs to the bootstrap confidence intervals presented in Section 16.6.3 and Section 16.6.4 in the setting of one sample. The Studentized variant is typically preferred as it tends to me more accurate, so this is the one we focus on.

Problem 17.4. Before or after reading this subsection, derive the analog to the bootstrap confidence interval of Section 16.6.3 in the present setting of two samples.

The idea is to estimate the distribution of $T := (D-\delta)/G$ by bootstrap instead of relying on a normal approximation, as done in (17.3). For this we go to the bootstrap world. Assume that the data, $(x_{i,j})$, have been observed. As usual, a star indicates a quantity in the bootstrap world. In particular, F_j^* denotes the empirical distribution for Group j. A bootstrap sample is thus $(X_{i,j}^*)$, where

$$X_{i,j}^* \sim \mathsf{F}_j^*, \quad \text{independent.} \tag{17.4}$$

Its sample mean and variance will be denoted by $\bar{X}_j^*$ and S_j^*. In the bootstrap world we can define the analog of the various quantities that are needed for the analysis:

$$\mu_j^* := \bar{x}_j, \qquad \delta^* := \mu_1^* - \mu_2^* = \bar{x}_1 - \bar{x}_2,$$

and

$$D^* := \bar{X}_1^* - \bar{X}_2^*, \qquad G^* := \left[S_1^{*2}/n_1 + S_2^{*2}/n_2\right]^{1/2},$$

and $T^* := (D^* - \delta^*)/G^*$. The bootstrap distribution of T^* is used as an estimate of the distribution of T. Let t_u^* be the u-quantile of the bootstrap distribution of T^*. Then

$$\delta \in \left[D - t_{1-\alpha/2}^* G, D - t_{\alpha/2}^* G\right], \tag{17.5}$$

with approximate probability $1-\alpha$.

As usual, an analytically derivation of the bootstrap distribution of T^* is impractical and one resorts to Monte Carlo simulation to estimate it.

Problem 17.5. In R, write a function that takes in the data, the confidence level, and number of Monte Carlo replicates, and returns the bootstrap confidence interval (17.5).

17.2 Inference about a parameter

In this section we consider the task of comparing a parameter of interest, denoted φ (e.g., standard deviation). Let $\varphi_j = \varphi(\mathsf{F}_j)$. We are interested in building a confidence interval for

$$\delta := \varphi_1 - \varphi_2.$$

Remark 17.6. One might be interested, instead, in the ratio φ_1/φ_2. Almost invariably, φ is a non-negative parameter in this case, and if so, one can simply take the logarithm and the problem becomes that of estimating the difference $\log\varphi_1 - \log\varphi_2$.

Problem 17.7. Suppose that $I = [A, B]$ is a $(1-\alpha)$-confidence interval for $\log\varphi_1 - \log\varphi_2$. Turn that into a $(1-\alpha)$-confidence interval for φ_1/φ_2.

17.2.1 Naive approach

A naive, yet very reasonable approach consists in building a confidence interval for each parameter based on the corresponding sample, for example, using a bootstrap approach as described in Section 16.7, and then combining these intervals to obtain a confidence interval for the difference.

In detail, let $I_1 = [A_1, B_1]$ be a $(1-\alpha/2)$-confidence interval for φ_1 based on the 1st sample, and let $I_2 = [A_2, B_2]$ be a $(1-\alpha/2)$-confidence interval for φ_2 based on the 2nd sample

Problem 17.8. Show that $[A_1 - B_2, B_1 - A_2]$ is a $(1-\alpha)$-confidence interval for δ.

Although this approach is clearly valid, it can be conservative, in the usual sense that, although the level of confidence may be valid, the resulting interval is relatively wide.

17.2.2 Bootstrap confidence interval

We now present the two-sample analog of the approach described in Section 16.6.3. We assume that φ can be defined for a discrete distribution so that we may use the empirical bootstrap. A reasonable estimator for δ is the plugin estimator, which may be written as

$$D := \varphi(\widehat{\mathsf{F}}_1) - \varphi(\widehat{\mathsf{F}}_2),$$

where $\widehat{\mathsf{F}}_j$ denotes the empirical distribution for Group j (seen as a random distribution). Beyond mere estimation, the construction of a confidence interval necessitates knowledge of the distribution of $D-\delta$. As in Section 16.6.3, we estimate this distribution by bootstrap.

Having observed the data $(x_{i,j})$, we travel to the bootstrap world. There, bootstrap samples are generated as in (17.4), and the quantities of interest are

$$\delta^* := \varphi(\mathsf{F}_1^*) - \varphi(\mathsf{F}_2^*),$$

and

$$D^* := \varphi(\mathsf{F}_1^{**}) - \varphi(\mathsf{F}_2^{**}),$$

where F_j^{**} is the empirical distribution of $X_{1,j}^*, \ldots, X_{n_j,j}^*$.

We estimate the distribution of $D-\delta$ by the bootstrap distribution of $D^*-\delta^*$. As usual, computing this bootstrap distribution in closed form is practically impossible and we resorts to Monte Carlo simulation to estimate it.

Problem 17.9. In R, write a function that takes in the two samples as numerical vectors, the desired confidence level, and a number of Monte Carlo replicates, and returns an appropriate bootstrap confidence interval for the difference in medians.

Remark 17.10. A bootstrap Studentized confidence interval can also be derived, but the method is in general more complex and computationally more intensive.

Remark 17.11. In the two-sample setting, there does not seem to be a distribution-free confidence interval for the difference in medians that would mimic the one developed in Section 16.3.2 in the one-sample setting. (That is, unless one is willing to assume that the two underlying distributions are translates of each other, as in (17.8).)

17.3 Goodness-of-fit testing

Consider now the problem of goodness-of-fit testing, meaning that we want to test

$$\mathcal{H}_0 : \mathsf{F}_1 = \mathsf{F}_2. \tag{17.6}$$

We considered this problem in the discrete setting in Section 15.3.

Problem 17.12 (Naive approach)**.** A naive, although reasonable approach, consists in computing a $(1-\alpha/2)$-confidence band for F_j as in Section 16.8.2, denoted $\mathcal{B}_j$, and then rejecting the null if $\mathsf{F}_1 \notin \mathcal{B}_2$ or $\mathsf{F}_2 \notin \mathcal{B}_1$. Show that this yields a test at level α.

17.3.1 Kolmogorov–Smirnov test

We present a more direct approach based on rejecting for large values of $\Delta(\widehat{\mathsf{F}}_1, \widehat{\mathsf{F}}_2)$, where Δ is a measure of dissimilarity between distribution functions. We saw some examples in Section 16.8.1. We focus on the supremum norm (16.20), which is one of the most popular.

Proposition 17.13. *The distribution of $\Delta(\widehat{\mathsf{F}}_1, \widehat{\mathsf{F}}_2)$ when both samples are drawn from the same distribution, say F_0, does not depend on F_0 as long as it is continuous. (The distribution does depend on the sample sizes n_1, n_2, left implicit here.)*

Problem 17.14. Prove this proposition.

There are recursive formulas for computing that distribution, which are valid when there are no ties in the data.

R corner. Such recursive formulas are implemented in the function `ks.test`, although for larger sample sizes, the function relies on the asymptotic distribution, which is also known in closed form.

When there are no ties in the data, the p-value may be obtained by Monte Carlo simulation, which consists in drawing the two samples independently from the uniform

distribution in $[0,1]$ (or any other continuous distribution). Alternatively, this can be done by permutation (Section 17.3.2).

Problem 17.15 (Random walk)**.** There is a close connection with the simple random walk process of Example 9.25, particularly when the sample sizes are equal ($n_1 = n_2$). In that case, provide an interpretation of $\Delta(\widehat{\mathsf{F}}_1, \widehat{\mathsf{F}}_2)$ in terms of a simple random walk.

17.3.2 Permutation distribution

Define the concatenated sample

$$X_i = \begin{cases} X_{i,1} & \text{if } i \le n_1; \\ X_{i-n_1,2} & \text{if } i > n_1. \end{cases}$$

Thus, if the samples are $\boldsymbol{x}_1 = (x_{i,1} : i = 1, \ldots, n_1)$ and $\boldsymbol{x}_2 = (x_{i,2} : i = 1, \ldots, n_2)$, the concatenated sample is $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)$ and is of length $n_1 + n_2 = n$.

Let Π denote the group of permutations of $\{1, \ldots, n\}$. For a permutation π, let $\boldsymbol{x}_\pi$ denote the corresponding vector, meaning $\boldsymbol{x}_\pi = (x_{\pi_1}, \ldots, x_{\pi_n})$ if $\pi = (\pi_1, \ldots, \pi_n)$. The *permutation distribution* of a statistic T conditional on $\boldsymbol{x}$ is the uniform distribution on $(T(\boldsymbol{x}_\pi) : \pi \in \Pi)$. If a test rejects for large values of T, the corresponding p-value is computed as in (15.9).

Problem 17.16. Assume that the samples come from the same continuous distribution. Show that the distribution of the Kolmogorov–Smirnov statistic coincides with its permutation distribution. Show that the same is true of any test statistic based on a test of randomness (after reading Section 15.8).

17.3.3 General permutation tests

Any test statistic $T(\boldsymbol{z})$ can be used to test for goodness-of-fit, with the corresponding p-value being obtained by permutation as in (15.9). Although this results in a valid p-value in the sense of (12.22), the resulting testing procedure will have more or less power depending on the choice of test statistic T and on the particular alternative.

Remark 17.17 (Conditional inference)**.** In the context of a randomized trial, where (say) individuals are assigned to either of two groups at random to receive one of two treatments, under the null distribution, all the permutations are equally likely, and obtaining the p-value by permutation is an example of re-randomization testing (Section 22.1.1). That said, the permutation p-value is valid regardless, and can be motivated by conditional inference (Section 22.1)

In the present setting, there is no uniformly most pow-

erful test (Problem 17.43). The choice of test statistic is rather guided by alternatives of particular interest. For example, if a difference in means is practically important, it makes sense to use a statistic like the difference in sample means.

Remark 17.18. In general, let φ be a parameter, such as the mean. If there is evidence against $\varphi(\mathsf{F}_1) = \varphi(\mathsf{F}_2)$, then this is obviously evidence against $\mathsf{F}_1 = \mathsf{F}_2$. To test $\varphi(\mathsf{F}_1) = \varphi(\mathsf{F}_2)$ it is natural to base the inference on the empirical equivalent, $\varphi(\widehat{\mathsf{F}}_1) - \varphi(\widehat{\mathsf{F}}_2)$, for which we can obtain a bootstrap p-value. (All this was detailed in Section 17.1 and Section 17.2.) However, when the null hypothesis we are truly interested in is $\mathsf{F}_1 = \mathsf{F}_2$, then obtaining a p-value by permutation is also an option (and typically preferred). See Problem 17.44 for a numerical comparison of these two options.

Remark 17.19 (Re-randomization and conditional inference)**.** When an experiment is based on a completely randomized design, relying on the permutation p-value is a form of re-randomization testing (Section 22.1.1). More generally, the permutation distribution is the distribution of T under the null conditional on the observed values. The resulting inference is thus a form of conditional inference (Section 22.1).

17.3.4 Rank tests

Let $r_{i,j}$ denote the *rank* of $x_{i,j}$ in increasing order when the two samples are combined. Ties, if present, can be broken in any number of ways, for example by giving to all the tied observations their average rank, or by breaking them them at random.

R corner. The function `rank` offers a number of ways for breaking ties, including these two.

Problem 17.20. Prove that

$$\sum_{j=1}^{2}\sum_{i=1}^{n_j} r_{i,j} = n(n+1)/2,$$

so that the set of ranks for Group 1 determines the set of ranks for Group 2.

A rank statistic is any statistic that can be computed based on the ranks.

Problem 17.21. Show that the Kolmogorov–Smirnov statistic is a rank statistic.

Problem 17.22 (Rank tests are permutation tests)**.** Show that under the null hypothesis where both samples were generated by the same distribution, if that distribution is continuous or the ranks are broken at random, the concatenated vector ranks, $(R_{i,j})$, is uniformly distributed

over all permutations of $\{1, \ldots, n_1 + n_2\}$. In particular, the p-value is obtained by permutation.

WILCOXON RANK-SUM TEST This (popular) test is based on the sum of the ranks from Group 1, meaning

$$r_1 := \sum_{i=1}^{n_1} r_{i,1}. \tag{17.7}$$

In the two-sided setting of (17.6), we reject for large and small values of this test statistic. Equivalently, the test rejects for large values of the difference in the sum of ranks, $r_1 - r_2$, where r_2 is the sum of ranks for Group 2.

Problem 17.23. Show that the null distribution of R_1 is symmetric about its mean, namely, $n_1(n+1)/2$. Deduce that the two-sided rank-sum test corresponds to rejecting for large value of $\big(r_1 - n_1(n+1)/2\big)^2$.

Remark 17.24. There are advantages and disadvantages to replacing the observations by their ranks and working with the latter. The main disadvantage is a loss in sensitivity, which may result in a loss of power. This loss is typically quite mild. The two main advantages are in terms of computation and robustness.

- *Computation and tabulation.* Assuming that ties are broken at random, the null distribution of R_1 only depends on (n_1, n_2). This was particularly important in the pre-computer age as the null distribution could be tabulated once for each (n_1, n_2). (This was done using recursive formulas for smaller sample sizes and asymptotic calculations for larger sample sizes.)
- *Robustness.* Using ranks offers some protection against gross outliers. For example, consider the permutation test based on the difference in sample means. Assume that one value has been corrupted and is now larger than the sum of the absolute values of all the other observations. Then the p-value will be approximately 1/2, whether the null is true or not. If one uses ranks, leading to the Wilcoxon rank-sum test, the influence of this corrupted observation remains minimal. The most it can do is change the rank of that observation (before being corrupted) to one of the extreme ranks (1 or n). The overall impact will be quite minimal.

Problem 17.25. Substantiate the claims above.

Remark 17.26. The Wilcoxon rank-sum test is sometimes presented as a test for comparing medians. This is not correct in general, the reason being that, as any other permutation test (Problem 17.22), the rank-sum test is a goodness-of-fit test. By this we mean that it

results in a valid p-value (in the sense of (12.22)) when the distributions are the same under the null. When only the medians are the same under the null, the p-value is not guarantied to be valid. That being said, testing for the equality of medians is equivalent to goodness-of-fit testing when the underlying distributions are assumed to be translates of each other, meaning

$$\mathsf{F}_1 = \mathsf{F}_2(\cdot - \mu) \text{ for some } \mu. \tag{17.8}$$

In that case, obviously, the distributions are the same if and only if their medians are the same.

17.3.5 Patterns and tests of randomness

A *pattern* is obtained by ordering the observations in the combined sample and labeling each observation according to the group it belongs to. For example, the data

Group 1:	2.5	0.9	-1.0	-1.6	0.7
Group 2:	-1.2	-0.5	-1.4		

results in the following pattern 12212111.

Once the pattern is computed on a particular dataset, tests of randomness (Section 15.8) can be used for goodness-of-fit testing.

Problem 17.27. Suppose that F_1 and F_2 are continuous. Show that any pattern is equally likely if and only if $\mathsf{F}_1 = \mathsf{F}_2$.

Therefore, when ties are broken at random, we can rely on the p-value returned by the test of randomness applied to the pattern. (In any case, we can rely on the permutation p-value.)

Remark 17.28. A pattern provides the same information as the ranks modulo the ordering within each group, but this within-group ordering is irrelevant for inference because each group is assumed to be iid. Because of this, any reasonable test statistic based on the ranks can be computed based on the pattern.

17.4 Multiple samples

We now consider the more general situation where g groups of observations are available and need to be compared. The observations from Group j are denoted $(X_{i,j} : i = 1, \ldots, n_j)$, so that n_j is the sample size for Group j. We assume that these are iid from some distribution, F_j, and that the samples are independent of each other. We let $n := n_1 + \cdots + n_g$ denote the total sample size.

Remark 17.29 (Format on a computer)**.** While a two-

sample method can be implemented as a function taking the two samples as two vectors (plus additional parameters), a method for multiple samples requires a different data format as it needs to be able to handle any number of groups. One way to format the data is as a two-column array, with one column listing the values and the other column listing the corresponding group index. For example, the data

Group 1:	2.2	0.8	1.0
Group 2:	1.1	0.3	1.6
Group 3:	0.3	0.8	

results in the array (here horizontal)

Values:	2.2	0.8	1.0	1.1	0.3	1.6	0.3	0.8
Group:	1	1	1	2	2	2	3	3

17.4.1 All pairwise comparisons

A reasonable approach is to perform all pairwise comparisons using a method for the comparison of two groups. If we are performing a test, for example, this leads us to do so for every pair, so that $\binom{g}{2}$ tests are performed in total. Such *multiple testing* situations are discussed in more detail in Chapter 20.

R corner. The function pairwise.t.test performs all the pairwise Student-Welch tests, while pairwise.wilcox.test performs all the pairwise Wilcoxon rank-sum tests. (A number of ways to correct for multiple testing are offered.)

Remark 17.30 (Many-to-one comparisons)**.** Another approach consists in performing a *many-to-one comparison*, where all the 'treatment' groups are compared to a 'control' group, the latter serving as benchmark. In particular, treatment groups are not compared to each other. (For more on this, see the classic book Miller [169].)

In the remainder of this section, we focus on various forms of *global testing*, by which we mean testing whether there is any difference between the groups. (We develop this further in Section 20.2.)

17.4.2 Testing for a difference in means

Let μ_j denote the mean of F_j. We first focus on testing for the equality in means

$$\mathcal{H}_0 : \mu_1 = \cdots = \mu_g.$$

This generalizes the testing problem of Section 17.1, where we considered the case $g = 2$. The methods presented there have analogs.

F-TEST This test was proposed by Fisher in the 1920's and later modified by Welch [250] to settings where the groups do not necessarily have the same variance. This test, which we will call the Fisher-Welch test, generalizes the Student-Welch test and, in particular, also relies on the Central Limit Theorem. Its form and derivation are rather complex and will not be given here.

R corner. This test is implemented in the R function oneway.test.

BOOTSTRAP TEST The bootstrap procedure is analogous to the one presented in Section 17.1.2 in that the resampling is within groups.

As for the choice of a test statistic, one possibility is to choose the *treatment sum-of-squares*, defined as

$$\sum_{j=1}^{p} n_j(\bar{x}_j - \bar{x})^2, \tag{17.9}$$

where $\bar{x}_j$ is the average for Group j and $\bar{x}$ is the overall average. Another option is to use the Fisher-Welch test statistic, which leads to an extension of the bootstrap Studentized procedure.

We assume that we reject for large values of a test statistic T. Importantly, we need to place ourselves under the null distribution before bootstrapping. We do this by centering each group, which effectively makes the groups have the same mean (equal to 0). Having observed the data, let $\mathsf{F}_j^{\circledast}$ denote the empirical distribution for Group j after centering. A bootstrap sample is thus $\boldsymbol{X}^* = (X_{i,j}^*)$, where

$$X_{i,j}^* \sim \mathsf{F}_j^{\circledast}, \quad \text{independent.}$$

Then the bootstrap p-value is, as usual, the probability that $T^* := T(\boldsymbol{X}^*) \geq t := T(\boldsymbol{x})$. (Note that t is the observed value of the test statistic.) This p-value is typically estimated by Monte Carlo simulation.

Problem 17.31. In R, write a function that takes in the values and group labels, and the number of bootstrap samples to be generated, and returns the Monte Carlo estimate for the bootstrap p-value of the Fisher-Welch test. (Although we did not provide an analytic form for this statistic, it can be computed using the function oneway.test.)

17.4.3 GOODNESS-OF-FIT TESTING

As in Section 17.3, suppose we are interested in comparing the distributions that generated the groups in the sense of testing

$$\mathcal{H}_0 : \mathsf{F}_1 = \cdots = \mathsf{F}_g.$$

PERMUTATION TESTS As in Section 17.3.3, a calibration by permutation is particularly appealing. (In particular, Remark 17.19 applies.) Suppose we decide to reject for large values of a test statistic T. This could be the treatment sum-of-squares or the Fisher-Welch statistic, or any other. We concatenate the groups as we did in Section 17.3.2, obtaining $\boldsymbol{x} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_g)$, where $\boldsymbol{x}_j := (x_{1,j}, \dots, x_{n_j,j})$ are the observations from Group j. Let Π be the group of permutations of $\{1, \dots, n\}$ and for $\pi \in \Pi$, let $\boldsymbol{x}_\pi$ denote the dataset permuted according to π. The corresponding p-value is computed as in (15.9), and it is a valid p-value under the null hypothesis $\mathcal{H}_0$ above.

Problem 17.32. In R, write a function that takes in the values and group labels and the number of permutations to be generated, and returns a Monte Carlo permutation p-value for the Fisher-Welch test statistic.

RANK TESTS As in Section 17.3.4, replacing the observations by their ranks is a viable option. Let $r_{i,j}$ be the rank of $x_{i,j}$ in increasing order in the combined sample. A direct extension of the Wilcoxon rank-sum test, in particular in view of Problem 17.23, consists in rejecting for large values of

$$\sum_{j=1}^{g} \big(r_j - n_j(n+1)/2\big)^2, \tag{17.10}$$

where $r_j := \sum_{i=1}^{n_j} r_{i,j}$ is the rank sum for Group j.

Problem 17.33. Show that, indeed, this reduces to the two-sided rank-sum test when $g = 2$.

Problem 17.34. In R, write a function that implements the test based on (17.10). The function takes as input the data and the number of permutations to be drawn and returns the corresponding estimated p-value.

The most rank test in the present setting, however, is the *Kruskal–Wallis test*, which is based instead on rejecting for large values of the treatment sum-of-squares (17.9) computed on the ranks, namely

$$\sum_{j=1}^{g} \frac{1}{n_j}\big(r_j - n_j(n+1)/2\big)^2. \tag{17.11}$$

Problem 17.35. Show that the Kruskal–Wallis test equivalently rejects for large value of $\sum_{j=1}^{g} r_j^2/n_j$. [Use the fact that $r_1 + \cdots + r_g = n(n+1)/2$.]

Remark 17.36. The actual Kruskal–Wallis test statistic involves a standardization that makes the resulting

statistic have, under the null hypothesis, the chi-squared distribution with $g-1$ degrees of freedom in the large-sample limit where $\min_j n_j \to \infty$.

R corner. This test is implemented in the function kruskal.test, which returns a p-value based on the limiting distribution.

Problem 17.37. In R, write a function that takes in the data and a number of Monte Carlo replicates, and returns a Monte Carlo estimate of the permutation p-value for the Kruskal–Wallis test. (Being a rank test, the permutation p-value is the exact p-value, at least when the ties are broken at random.)

17.5 Further topics

17.5.1 Two-sample median test

Despite its name, the *median test* is for goodness-of-fit, meaning a test for (17.6). It works as follows. We consider two groups, of sizes n_1 and n_2, as before. Let M denote the sample median of all the observations combined. Let T denote the number of observations from Group 1 that exceed M. The two-sided variant of the test rejects for large and small values of T.

Problem 17.38. Assume that ties are broken at random.

(i) Assume the $n = n_1 + n_2$ is even. Show that, under the null hypothesis (17.6), T has the hypergeometric distribution with parameters $(n_1, n/2, n/2)$.

(ii) Assume that n is odd. What is the distribution of T under the null?

Problem 17.39. In R, write a function that implements this test.

17.5.2 Consistency

Consider the two-sample setting. We say that a testing procedure is universally consistent if, when $\mathsf{F}_1 \neq \mathsf{F}_2$, at any level $\alpha > 0$ the corresponding test has power converging to 1 has $n_1 \wedge n_2 \to \infty$.

Problem 17.40. Show that the Kolmogorov–Smirnov test procedure is universally consistent. [Use the fact that the empirical distribution function is consistent for the underlying distribution function.]

Proposition 17.41. *The Wilcoxon rank-sum test procedure is not universally consistent.*

The Wilcoxon rank-sum test is, however, consistent in a shift model (17.8), and in fact, in that model, it tends to be substantially more powerful that the Kolmogorov–Smirnov

test. The two tests are compared in Problem 17.43.

Remark 17.42 (Stochastic dominance)**.** More generally, the rank-sum test is consistent when, under the alternative, F_1 stochastically dominates F_2, or vice versa (assuming the two-sided version of the test). We say that F *stochastically dominates* G if

$$\bar{\mathsf{F}}(t) \geq \bar{\mathsf{G}}(t), \quad \text{for all } t \in \mathbb{R}.$$

This means that, if X has distribution F and Y has distribution G, then $\mathbb{P}(X > t) \geq \mathbb{P}(Y > t)$ for all t.

17.5.3 Inverting a permutation test

All the goodness-of-fit tests that we saw, including the rank tests, are permutation tests. And a permutation test, strictly speaking, is a goodness-of-fit test. If, however, we are willing to assume that the distributions are in the same location family, that is

$$\mathsf{F}_j(x) = \mathsf{F}(x - \mu_j), \quad \text{for all } j = 1, \ldots, g, \tag{17.12}$$

for some distribution F and some shifts μ_j, then a goodness-of-fit test can be used to compare the means or medians. In the two-sample setting, when this assumption is in place, a permutation test can be used to build a confidence interval for the difference in means, as seen in Section 12.4.9. This will be a valid confidence interval with the desired level of confidence, as long as all the assumptions are valid, including (17.12).

17.6 Additional problems

Problem 17.43 (Kolmogorow-Smirnov vs Wilcoxon)**.** In R, perform some simulations to compare the power of the Kolmogorov–Smirnov test and that of the Wilcoxon test. Consider the case where both samples are of same size $m \in \{20, 50, 100\}$. The first group comes from $\mathcal{N}(0, 1)$. In the first situation, the second group comes from $\mathcal{N}(\theta, 1)$. In the second situation, the second group comes from $\mathcal{N}(0, 1+\theta)$. In each case $\theta > 0$ is chosen carefully on a grid to make the setting interesting, showing the power going from $\alpha := 0.01$ to close to 1. (This interesting range for θ will depend on m.) Repeat each setting (situation, m, θ) $B = 1000$ times. For each (situation, m), in the same plot draw the power curve for each test as a function of θ. Use different colors and add a legend.

Problem 17.44 (Bootstrap goodness-of-fit tests)**.** When testing for goodness-of-fit, permutation is typically considered the calibration of choice, in large part because the permutation p-value is valid, regardless of the group sizes. However, a bootstrap approach is also possible. Explain,

when testing whether two groups come from the same distribution, how you would obtain a p-value by bootstrap. [This bootstrap is different from that used for comparing features (Section 17.2.2).]

Problem 17.45 (Permutation vs rank tests)**.** In R, perform some simulations to compare the power of the permutation test based on the difference in sample means and the corresponding rank test, which is none other than the rank-sum test. Specifically, consider the case where both samples are of same size $m \in \{20, 50, 100\}$. The first group comes from $\mathcal{N}(0, 1)$, while the second group comes from $\mathcal{N}(\theta, 1)$. In each case $\theta > 0$ is chosen carefully to make the setting interesting, showing the power going from $\alpha := 0.01$ to close to 1. Repeat each setting (situation, m, θ) $B = 1000$ times. For each (situation, m), in the same plot, draw the power curve for each test as a function of θ.

Problem 17.46. Strictly speaking, a permutation p-value is only valid for the null hypothesis that the underlying distributions are the same. How does it behave when it is used when testing for a parameter? Consider a two sample setting where we test for a difference in means. We choose as test statistic the difference in sample means, which we calibrate by permutation. The first group comes from $\mathcal{N}(0, 1)$, while the second group comes from $\mathcal{N}(0, 3)$. Clearly, we are under the null hypothesis (same means), and the distributions are different. Assume both groups are of same size m, and the permutation p-value is based on a number B of Monte Carlo replicates. By varying m and B, assess the accuracy of the permutation p-value. Offer some brief comments, and possibly some elements of explanation for that behavior.

Problem 17.47 (Mann–Whitney test)**.** This test is based on the test statistic

$$U := \sum_{j=1}^{2} \sum_{i=1}^{n_j} \{X_{i,1} > X_{i,2}\}.$$

Show that, when there are no ties, $U = W - n_1(n_1 + 1)/2$, where W is the Wilcoxon rank-sum statistic (17.7).

Problem 17.48. The two tests, based on (17.10) and (17.11) respectively, coincide when the design is balanced in the sense that the group sizes are identical. How do they compare when the group sizes are not the same? Perform some simulations to investigate that.

Problem 17.49 (Student test)**.** Consider a normal experiment where we observe independent realizations of $X_1, \ldots, X_m$, assumed an iid sample from $\mathcal{N}(\mu, \sigma^2)$, and of $Y_1, \ldots, Y_n$, assumed an iid sample from $\mathcal{N}(\xi, \sigma^2)$. Impor-

tantly, the two normal distributions are assumed to have the *same variance*. All three parameters are unknown. Our goal is to test the null hypothesis that the means are equal, $\mu = \xi$. The *Student test* rejects for large values of $|T|$ where

$$T := \frac{\bar{X} - \bar{Y}}{S}, \tag{17.13}$$

with $\bar{X}$ and $\bar{Y}$ being the sample means and S being the pooled sample standard deviation.

(i) Show that there is a constant $c_{m,n}$ such that the distribution of $c_{m,n}T$ under the null hypothesis is the Student distribution with $m + n - 2$ degrees of freedom.

(ii) Show that this test corresponds to the likelihood ratio test under the present model.

Problem 17.50 (Welch test)**.** Consider a normal experiment where we observe independent realizations of $X_1, \dots, X_m$, assumed an iid sample from $\mathcal{N}(\mu, \sigma^2)$, and of $Y_1, \dots, Y_n$, assumed an iid sample from $\mathcal{N}(\xi, \tau^2)$. All four parameters are unknown. Our goal is to test $\mu = \xi$. The *Welch test* rejects for large values of $|T|$ where

$$T := \frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/m + S_Y^2/n}}, \tag{17.14}$$

with $\bar{X}$ and $\bar{Y}$ being the sample means and S_X and S_Y being the sample standard deviations. It turns out that T does not have a Student distribution under the null hypothesis. Does this test correspond to the likelihood ratio test under the present model?

Problem 17.51 (Fisher test)**.** Consider a normal experiment where we observe independent realizations of $(X_{i,j} : i = 1, \dots, n_j)$ assumed an iid sample from $\mathcal{N}(\mu_j, \sigma^2)$, for $j = 1, \dots, g$. Importantly, the normal distributions are assumed to have the same variance. All $g + 1$ parameters are unknown. Our goal is to test the null hypothesis that $\mu_1 = \cdots = \mu_g$. The *F-test* rejects for large values of F where

$$F := \frac{\sum_{j=1}^{g} n_j (\bar{X}_j - \bar{X})^2}{S^2},$$

with $\bar{X}_j$ being the sample mean for the group j, $\bar{X}$ the pooled sample mean, and S the pooled sample standard deviation. This is the multiple-sample analog of (17.13).

(i) Show that there is a constant $c_{m,n}$ such that the distribution of $c_{m,n}F$ under the null hypothesis is the Fisher distribution with $g - 1$ and $n - g$ degrees of freedom, where n is the total sample size.

(ii) Show that this test corresponds to the likelihood ratio test under the present model.

(There is a Welch version of this test which is the multi-sample analog of (17.14). The exact form of the test statistic is rather complicated [250].)

Problem 17.52 (Energy statistics)**.** Cramér[89] proposed in [48] the following dissimilarity for comparing two distribution functions

$$\Delta(\mathsf{F},\mathsf{G})^2 := \int_{-\infty}^{\infty} \big(\mathsf{F}(x)-\mathsf{G}(x)\big)^2 \mathrm{d}x.$$

(Note the difference with the Cramér–von Mises dissimilarity defined in (16.21).)

(i) Letting X, X' be iid from F and (independently) Y, Y' be iid from G, show that

$$\Delta(\mathsf{F},\mathsf{G})^2 = \mathbb{E}[|X-Y|] - \frac{1}{2}\big(\mathbb{E}[|X-X'|]+\mathbb{E}[|Y-Y'|]\big).$$

(ii) Use this to provide an explicit (and as simple as possible) expression for the sample equivalent, meaning when applying the dissimilarity to the empirical distribution functions of the two samples under consideration.[91]

(iii) Implement this as a test in R. (Calibration is by permutation based on a specified number of MC replicates.)

(iv) Is the test distribution-free?

Problem 17.53. The test of Problem 17.52 was expressed there in terms of distribution functions. However, it turns out it also has a relatively simple expression in terms of characteristic functions. Indeed, show that

$$\Delta(\mathsf{F},\mathsf{G})^2 = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{|\varphi_{\mathsf{F}}(t)-\varphi_{\mathsf{G}}(t)|^2}{t^2}\mathrm{d}t,$$

where φ_{F} and φ_{G} are the characteristic functions of F and G, respectively.

[91] The resulting statistic is an example of what Székely and collaborators call *energy statistics*. See [228]. for a survey and historical perspective, which includes the early proposal by Cramér.

CHAPTER 18

MULTIPLE PAIRED NUMERICAL SAMPLES

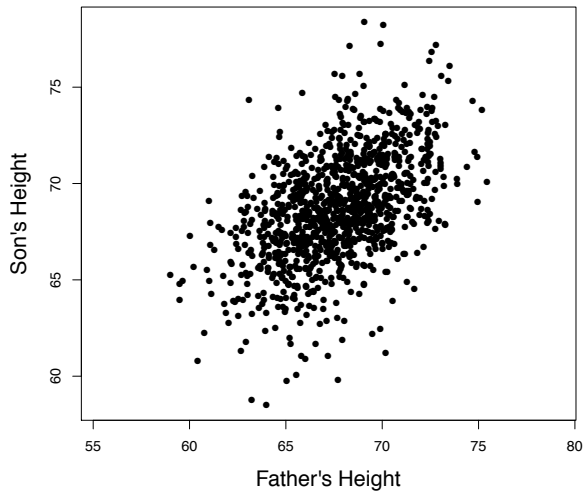
We consider in this chapter experiments where the variables of interest are paired. Importantly, we assume that these variables are directly comparable (in contrast with the following two chapters).

Crossover trials are important examples of such experiments. Other examples include the following.

Example 18.1 (Judge panel)**.** In the food industry in particular, it is common to ask individuals to rate the taste of different products, typically of the same type. For example, in [40], 12 experienced wine tasters were asked to rate 78 wines on a variety of characteristics.

Example 18.2 (Father-son heights)**.** Karl Pearson collected data on the heights of 1078 fathers and their (adult) sons [181]. In that case x_i = father's height and y_i = son's height (both in inches) for the ith pair. (This dataset is discussed at length in [92], and in fact a scatterplot of the data is on the cover of that book, and reproduced in Figure 18.1.)

Figure 18.1: A scatterplot of the data described in Example 18.2.



18.1 Two paired variables

We start with a setting where we observe a bivariate numerical sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. The pairs are assumed iid from some unknown distribution. We assume that the X and Y variables can be compared directly.

Taking the example of a crossover trial comparing two treatments (one of them could be a placebo), when there is no difference between treatments it is assumed that X and Y are exchangeable. When testing for a difference in treatment, we are thus testing the following null hypothesis

$$\mathcal{H}_{0,*} : (X, Y) \sim (Y, X).$$

There is a one-to-one correspondence

$$(X, Y) \leftrightarrow (U, Z) := (X + Y, X - Y),$$

and the null hypothesis can be equivalently expressed as follows

$$\mathcal{H}_{0,*} : (U, Z) \sim (U, -Z).$$

This invites[92] the drop of U and the focus on Z, where the null hypothesis becomes that of symmetry, meaning

$$\mathcal{H}_0 : Z \sim -Z. \tag{18.1}$$

The rest of this section is dedicated to testing this hypothesis based on an iid sample $Z_1, \ldots, Z_n$. (Unless restricted further, the underlying distribution is simply assumed to be in the family of all distributions on the real line.)

[92] This can be justified based on invariance considerations [147, Sec 6.8], although the dropping of U could a priori lead to some information loss.

18.1.1 Symmetric distributions

A random variable Z is said to be *symmetric* about μ if

$$\mathbb{P}(Z < z) = \mathbb{P}(Z > \mu - z), \quad \text{for all } z \in \mathbb{R}.$$

Problem 18.3. Assuming that Z has continuous distribution function F, show that this is equivalent to

$$\mathsf{F}(z) = 1 - \mathsf{F}(\mu - z), \quad \text{for all } z \in \mathbb{R}.$$

In particular, F is symmetric about 0 if

$$\mathsf{F}(z) = 1 - \mathsf{F}(-z),$$

that is, if $z \mapsto \mathsf{F}(z) - 1/2$ is odd.

Problem 18.4. Suppose that F has a piecewise continuous density f. Show that F is symmetric about μ if and only if

$$f(z) = f(\mu - z), \text{ at any continuity point } z.$$

In particular, if f is continuous, f is symmetric about 0 if and only f is even. (In any case, f is essentially even.)

Problem 18.5. Show that if F is symmetric about μ, then μ is necessarily a median of F, and also its mean if it has a mean.

Although symmetry can be about any point on the real line, in what follows we assume that point to be the origin. This is the most important case, in part because it is motivated by (18.1), and can be considered without loss of generality. In particular, in what follows, by 'symmetric' we mean 'symmetric about 0'.

Problem 18.5 justifies the application of tests for the median, such as the sign test (Section 16.3.3), as well as tests for the mean, for example a bootstrap test (Section 16.6.5). If such a test rejects, it is evidence against the (null) hypothesis of symmetry. However, such a test cannot be universally consistent since a distribution can be asymmetrical and yet have zero median, or zero mean, or both.

Problem 18.6. Construct a distribution that is asymmetrical and has median and mean both equal to 0. One avenue is to consider a Gaussian mixture of the form $p\mathcal{N}(a, 1) + (1 - p)\mathcal{N}(b, \sigma^2)$, where $p \in [0, 1]$, $a, b \in \mathbb{R}$, and $\sigma^2 > 0$ are chosen to satisfy the requirements. Another avenue is to consider a distribution with finite support. In that case, what is the minimum support size needed to satisfy the requirements?

The procedures presented below are based on the following characterizations of symmetry.

Problem 18.7. Let Z be a random variable. Show that the following assertions are equivalent: *(i)* Z is symmetric; *(ii)* $\text{sign}(Z)$ is symmetric and independent of $|Z|$; *(iii)* $\mathbb{P}(Z > 0) = \mathbb{P}(Z < 0)$ and $Z \mid Z > 0$ and $-Z \mid Z < 0$ share the same distribution.

Problem 18.7 can be used to motivate the comparison of the positive part of the sample, meaning $\{Z_i : Z_i > 0\}$, with the negative part of the sample, meaning $\{-Z_i : Z_i < 0\}$, as one would for two different groups. The techniques developed in Section 17.3 for that purpose are particularly relevant. Following this logic leads to two well-known methods for testing for symmetry that we present next.

Remark 18.8 (Zero values)**.** Since the values that are exactly 0 do not carry any information on the asymmetry of the underlying distribution, it is common to simply drop these values before applying a procedure. (This is an example of conditional inference.) This is what we do in what follows, and although it changes the sample size, we redefine n as the sample size after removing these observations.

18.1.2 A test based on sign flips

The following test comes from applying a permutation test for two-sample goodness-of-fit testing, as discussed in Section 17.3.3, to compare the distributions of $Z \mid Z > 0$ and $-Z \mid Z < 0$.

We let $\boldsymbol{z} = (z_1, \dots, z_n)$ denote the observed sample, and for $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n) \in \{-1, 1\}^n$, we let $\boldsymbol{z}_{\boldsymbol{\varepsilon}} = (\varepsilon_1 z_1, \dots, \varepsilon_n z_n)$. Suppose that we reject for large values of a test statistic $T(\boldsymbol{z})$. A popular choice is

$$T(\boldsymbol{z}) = \Big| \sum_{i=1}^{n} z_i \Big|. \tag{18.2}$$

Having observed $\boldsymbol{Z} = \boldsymbol{z}$, the p-value is

$$\mathsf{pv}_T(\boldsymbol{z}) = \frac{\#\{\boldsymbol{\varepsilon} : T(\boldsymbol{z}_\varepsilon) \geq T(\boldsymbol{z})\}}{2^n}. \tag{18.3}$$

The denominator is the cardinality of the set of sign vectors $\{-1, 1\}^n$. Thus this is the proportion of sign vectors that lead to a value of the test statistic that is at least as extreme as the one observed.

Proposition 18.9. *This quantity is a valid p-value in the sense that it satisfies* (12.22)*, meaning that*

$$\mathbb{P}(\mathsf{pv}_T(\boldsymbol{Z}) \leq \alpha) \leq \alpha, \quad \textit{for all } \alpha \in [0, 1],$$

when the underlying distribution is symmetric.

In practice, computing the p-value (18.3) is quickly intractable. This is because the number of sign vectors

of interest, 2^n, grows exponentially ($> 10^{30}$ when $n = 100$). As usual, one resorts to Monte Carlo simulation to estimate this p-value, by repeatedly drawing sign vectors uniformly at random.

Problem 18.10. In R, write a function that takes as input the sample and a number of Monte Carlo replicates, and returns the p-value estimated by Monte Carlo for the test statistic (18.2).

Remark 18.11. Although the test is built on a permutation test, it is not a permutation test per se. Nothing is being permuted. However, its construction can be motivated by conditional inference (Section 22.1): we condition on the absolute values, $|Z_1|, \dots, |Z_n|$, which a priori do not carry any information on whether the underlying distribution is symmetric.

Problem 18.12. How is the test above, based on the statistic (18.2) and returning the p-value (18.3) different from the permutation test, based on the absolute value of the difference in sample means, applied to compare the positive and negative samples? [The two tests are almost the same, but not quite identical.]

18.1.3 Wilcoxon signed-rank test

We now turn to the test that results from comparing the distributions of $Z \mid Z > 0$ and $-Z \mid Z < 0$ using the Wilcoxon rank-sum test presented in Section 17.3.4. This leads one to use the test statistic

$$\sum_{i=1}^{n} r_i \{z_i > 0\},$$

where r_i is the rank of $|z_i|$ among $|z_1|, \dots, |z_n|$. In the two-sided situation, we reject for large and small values of this statistic, as we did for the two-sided rank-sum test.

Problem 18.13. Verify that this is indeed the resulting test statistic when using the rank-sum test to compare the distributions of $Z \mid Z > 0$ and $-Z \mid Z < 0$.

R corner. The function wilcox.test computes the signed-rank test when provided with a numerical vector, and the rank-sum test when provided with two numerical vectors.

Problem 18.14. Show that it is equivalent to use the following statistic

$$\sum_{i=1}^{n} r_i \operatorname{sign}(z_i) = \sum_{z_i>0} r_i - \sum_{z_i<0} r_i. \tag{18.4}$$

(In this form, it is clear that this is the rank variant of the test of Section 18.1.2.)

Problem 18.15. When the underlying distribution is symmetric, and assuming in addition that it is continuous or that ties among ranks are broken at random, show that the signed-rank statistic in the form of (18.4) has the distribution of $\sum_{i=1}^{n} i\,\varepsilon_i$ with[93] $\varepsilon_1, \ldots, \varepsilon_n$ iid uniform in $\{-1, 1\}$.

Whether the underlying distribution is continuous or not, and whether ties are broken at random or not, an approach by conditional inference remains available: it consists in computing a p-value by fixing the ranks while sampling sign vectors uniformly at random.

Problem 18.16. In R, write a function that takes in the data and a number of Monte Carlo replicates, and returns the Monte Carlo estimate for this p-value. Compare your function with wilcox.test in simulations.

Sign pattern We saw in Section 17.3.5 that any rank test for goodness-of-fit in a two-sample setting is based on the pattern defined by the two samples combined. The situation is analogous here. Indeed, any test of symmetry based on the ranks and signs is based on the *sign pattern* given by ordering the absolute values and then listing the signs in that order. For example, the following observations

1.3 2.1 2.5 -1.4 1.0 0.4 -3.5 -1.0 0.2

yield the following sign sequence

$$+ \quad + \quad + \quad - \quad + \quad - \quad - \quad + \quad +$$

[93] The uniform distribution on $\{-1, 1\}$ is sometimes called the *Rademacher distribution*.

18.2 Multiple paired variables

We now consider the more general case of p paired variables. The sample is of size n, and denoted

$$(X_{1,1}, \ldots, X_{1,p}), \ldots, (X_{n,1}, \ldots, X_{n,p}).$$

Let $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,p})$ denote the ith observation, which is a vector of length p here. We assume these n observations to be iid from some unknown distribution, and we also assume that all the variables are directly comparable. This is the case Example 18.1, where $X_{i,j}$ is the rating of ith wine by the jth judge.

These observations are typically gathered in a n-by-p data matrix, $(X_{i,j})$, where a rows correspond to observations. See Table 18.1.

Table 18.1: Prototypical data matrix in the context of a crossover clinical trial.

	Treatment 1	Treatment 2	⋯	Treatment p
Subject 1	$X_{1,1}$	$X_{1,2}$	⋯	$X_{1,p}$
Subject 2	$X_{2,1}$	$X_{2,2}$	⋯	$X_{2,p}$
⋮	⋮	⋮		⋮
Subject n	$X_{n,1}$	$X_{n,2}$	⋯	$X_{n,p}$

Taking the example of a crossover clinical trial, the goal is to assess whether the treatments are different. This is again modeled by testing

$$\mathcal{H}_0 : (X_1, \dots, X_p) \text{ are exchangeable.}$$

18.2.1 Permutation tests

A calibration by permutation is particularly attractive in the present context, since permutations are at the core of the definition of exchangeability. In fact, in the context of a crossover trial, this corresponds to re-randomization testing (Section 22.1.1). Compared to a completely randomized design, the permutation in the context of a crossover trial or any other repeated measures design is done differently. Indeed, here the permutation is within subject. In particular, permuting across subjects (as is done in the context of a completely randomized design) is not appropriate since doing so does not preserve the null hypothesis.

Let T be a test statistic whose large values are evidence against the null. Having observed $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, a calibration by permutation is done as follows. Let Π_0 be the group of permutations of $\{1, \dots, p\}$ and let $\Pi := \Pi_0^{\times n}$, which is itself of group. The permutations in Π are the valid permutations in the present context.

Problem 18.17. Show that $|\Pi| = (p!)^n$.

Let t denote the observed value of the statistics, meaning $t = T(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$. For $\boldsymbol{\pi} = (\pi_1, \dots, \pi_n) \in \Pi$, with $\pi_i := (\pi_{i,1}, \dots, \pi_{i,p}) \in \Pi_0$, let t_π denote the value of the statistic when it is applied to the same data except permuted by $\boldsymbol{\pi}$, meaning $t_{\boldsymbol{\pi}} = T(\pi_1(\boldsymbol{x}_1), \dots, \pi_n(\boldsymbol{x}_n))$ where $\pi_i(\boldsymbol{x}_i) := (x_{\pi_{i,1}}, \dots, x_{\pi_{i,p}})$. The permutation p-value is then defined as usual

$$\frac{\#\{\pi \in \Pi : t_\pi \geq t\}}{|\Pi|}.$$

In practice, it is typically estimated by Monte Carlo simulation based on a number of permutations that are sampled independently and uniformly at random from Π.

Problem 18.18. Show that the permutation p-value and its Monte Carlo estimate, seen as random variables, are

both valid in the sense of (12.22). [Use the conclusions of Problem 8.63.]

Problem 18.19. In R, write a function that takes in as input the data matrix and a number of Monte Carlo replicates, and returns the estimated permutation p-value for the treatment sum-of-squares defined in (17.9). [Note that there are p groups here, with sample size $n_j = n$ for all j.]

18.2.2 Rank tests

Using ranks is particularly appealing in settings where the measurements across subjects are not easy to compare. This is for example the case when the measurements are subjective evaluations. A prototypical example is that of a judge panel experiment (Example 18.1) as the judges may be more or less liberal in the use of the full appraisal scale. Another important example is that of crossover trials where what is measured is the improvement of symptoms on a visual analog scale (VAS). Such subjective evaluations are notoriously difficult to compare. The use of ranks disregards the scale implicitly (and often unconsciously) used by a subject, and focuses on the subject's ranking instead. Thus ranks are computed within subjects.

In detail, let $r_{i,j}$ be the rank of $x_{i,j}$ among $x_{1,j}, \ldots, x_{n,j}$ and let $r_j = \sum_{i=1}^{n} r_{i,j}$ be the rank sum for Treatment j. (Although we are using the same notation, these ranks are defined differently than in Section 17.4.3.)

Remark 18.20. Rank tests are special cases of permutation tests, as their null distribution is the permutation distribution (at least when the ties are broken at random).

Friedman test Like the Kruskal–Wallis test, this test uses as test statistic the treatment sum-of-squares applied to the ranks defined above. It was proposed by Milton Friedman (1912 - 2006) in [100].

Problem 18.21. Show that using the treatment sum-of-squares is equivalent to using $\sum_{j=1}^{p} r_j^2$.

Remark 18.22. The actual Friedman test statistic involves a standardization that makes the resulting statistic have, under the null hypothesis, the chi-squared distribution with $p-1$ degrees of freedom in the large-sample limit where $n \to \infty$.

R corner. The function friedman.test, which implements that test, uses the limiting distribution to compute the p-value.

18.3 Additional problems

Problem 18.23. Consider a strictly increasing and continuous distribution function F. Derive a necessary and sufficient condition on the corresponding quantile function F^{-1} (a true inverse in this case) for F to be symmetric about a given $\mu \in \mathbb{R}$.

Problem 18.24 (Sign-flip vs signed-rank)**.** Perform some simulations to compare the power of the sign-flip test of Section 18.1.2 and the Wilcoxon signed-rank test of Section 18.1.3.

Problem 18.25 (Smirnov test)**.** The *Smirnov test for symmetry* is based on comparing the distributions of $Z \mid Z > 0$ and $-Z \mid Z < 0$ using the two-sample Kolmogorov–Smirnov test.

(i) Write down an expression, as simple as possible, for the corresponding test statistic.

(ii) In R, write a function that takes in the data and a number of Monte Carlo replicates, and returns the Monte Carlo estimate for the p-value (18.3).

CHAPTER 19

CORRELATION ANALYSIS

We consider an experiment resulting in paired numerical variables (X, Y). The general goal addressed in this chapter is that of quantifying the strength of association between these two variables. By association we mean dependence. We have an iid sample, $(X_1, Y_1), \dots, (X_n, Y_n)$, from the underlying distribution. Contrary to the setting of Section 18.1, here X and Y can be measurements of completely different kinds.

Example 19.1. The study described in [140] evaluates the impact of the time in detention on the mental health of asylum seekers in the US in terms of (self-reported) symptoms of anxiety, depression, and post-traumatic stress disorder. Focusing on just one symptom, say anxiety, the data would look like $(T_1, A_1), \dots, (T_n, A_n)$, where T_i denotes the time spent in detention and A_i the level of anxiety for Individual i. (There were $n = 70$ individuals interviewed for this study.)

A *correlation analysis* amounts to quantifying the

strength of association between X and Y. A more detailed description of this association is the goal of *regression analysis*, which is the topic of Chapter 21.

Remark 19.2. We assume throughout that neither X nor Y are constant random variables, for otherwise a correlation analysis is not relevant.

19.1 Testing for independence

Suppose we want to test

$$\mathcal{H}_0 : X \text{ and } Y \text{ are independent.} \tag{19.1}$$

This is in some sense the most extreme form of non-association between two variables. Most of this chapter is dedicated to testing for independence.

19.1.1 Tests based on binning

We saw in Section 15.7 how to test for independence between (paired) discrete variables. In principle, the tools developed there could be used when the variables are numerical (which we assume here), but only after binning.

Problem 19.3 (Independence tests based on binning)**.** Based on the tools introduced in Section 15.7 for discrete variables, propose at least one test for independence that is applicable to numerical variables. The general idea is to first bin the numerical variables, thus obtaining discrete variables, and then apply an independence test for discrete variables. Implement that test procedure in R.

(Notice the parallel with Problem 16.63.)

19.1.2 Permutation tests

In the present context, a permutation consists in permuting one coordinate, say, the Y coordinate (without loss of generality), while leaving the other variable fixed. Doing this leaves a null distribution unchanged while breaking any dependence under an alternative.

Suppose we reject for large values of a statistic T. Let t denote the observed value of the statistic, meaning $t = T((x_1, y_1), \ldots, (x_n, y_n))$. Letting Π be the group of permutations of $\{1, \ldots, n\}$, for $\pi = (\pi_1, \ldots, \pi_n) \in \Pi$, let t_π denote the value of the statistic when applied to the data permuted by π, meaning $t_\pi = T((x_1, y_{\pi_1}), \ldots, (x_n, y_{\pi_n}))$. Then the permutation p-value is defined as usual

$$\frac{\#\{\pi \in \Pi : t_\pi \geq t\}}{|\Pi|}.$$

Since $|\Pi| = n!$ can be quite large, this p-value is typically estimated by Monte Carlo simulation.

Problem 19.4. Show that the permutation p-value and its Monte Carlo estimate, seen as random variables, are both valid in the sense of (12.22). [Use the conclusions of Problem 8.63.]

Problem 19.5 (Bootstrap tests)**.** Although a calibration by permutation is favored, in large part because the permutation p-value is valid regardless of the sample size, a calibration by bootstrap is also possible (and reasonable). Propose a way to do so in the present context.

In the remainder of this chapter, we give several examples of test statistics that are commonly used for the purpose of testing for independence. The alternative set is very large, comprising all distributions on $\mathbb{R}^2$ that are not the product of their marginals, and there is no test that is uniformly best. Instead, each of the following tests is designed for certain alternatives.

19.2 Affine association

The variables X and Y are in perfect *affine association* if one of them is an affine function of the other, for example,

$$Y = aX + b, \text{ for some } a, b \in \mathbb{R}. \tag{19.2}$$

Such a perfect association is extremely rare in real applications. Even in settings governed by the laws of physics the relation is not exact, for example, because of various factors including measurement precision and error.

Example 19.6 (Boiling temperature in the Himalayas)**.** In [88], James Forbes reports data collected by Joseph Hooker on the boiling temperature of water at different elevations in the Himalayas. (Part of the dataset is available in R as Hooker in the alr4 package.) The variables are boiling temperature (degrees Fahrenheit) and barometric pressure (inches of mercury). Although the laws of physics predict an affine relationship, this is not exactly the case in this dataset, although it is an excellent model.

19.2.1 Pearson correlation

The correlation, defined in (7.13), was seen to measure affine association between paired random variables. This motivates the use of the sample correlation, defined as the correlation of the empirical distribution.

Problem 19.7. Show that the sample correlation is given by

$$r := \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

This is often called the *Pearson sample correlation*.

Problem 19.8. Show that $r \in [-1,1]$, and $|r| = 1$ if and only if there are $a, b \in \mathbb{R}$ such that $x_i = ay_i + b$ for all i or $y_i = ax_i + b$ for all i.

Problem 19.9 (Consistency of the sample correlation)**.** Let $\{(X_i, Y_i) : i \geq 1\}$ be iid bivariate numerical with correlation ρ. Let R_n denote the sample correlation of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Show that $R_n \to_{\mathrm{P}} \rho$ as $n \to \infty$.

The Pearson correlation test, which in its two-sided variant rejects for large values of $|R|$, is not universally consistent, essentially because there are bivariate distributions with $\rho = 0$ that are not the product of their marginals.

Problem 19.10. Suppose that X is uniform in $[-1,1]$ and define $Y = X^2$. Show that the distribution of (X, Y) has correlation $\rho = 0$. Generalize this result as much as you can (within reason).

Proposition 19.11. *If X and Y are independent and normal,*

$$T := \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

has the Student distribution with $n-2$ degrees of freedom. In general, if X and Y are independent and have finite second moments,

$$T = T_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

R corner. The function cor computes, by default, the Pearson correlation, while the function cor.test implements, by default, the Pearson correlation test, albeit returning a p-value computed based on Proposition 19.11.

Problem 19.12. Detail how Proposition 19.11 is used to produce a p-value. (Note that the null hypothesis, in this case, can be that the variables are independent, or more generally, that they have zero Pearson correlation. In either case, the p-value is approximate, except in the exceedingly rare situation where the underlying distribution is known to be bivariate normal.)

Problem 19.13. In R, write a function that takes in the dataset and a number of Monte Carlo replicates, and returns the estimated permutation p-value for the Pearson correlation.

19.3 Monotonic association

The variables X and Y are in perfect *monotonic association* if one of them is a monotonic function of the other,

for example,

$$Y = g(X), \text{ for some monotone function } g. \tag{19.3}$$

As before, and for similar reasons, perfect monotonic association is extremely rare in practice.

Example 19.14 (Antoine's equation)**.** For a pure liquid, the vapor pressure (p) and temperature (t) are related, to first order, by *Antoine's equation*

$$\log(p/p_0) = t_0/t,$$

where p_0 and t_0 are constants. Note that p is a monotone function of t in this model. Actual data does not fit the equation perfectly, but comes very close to that. (See [129] for more details in the case of mercury, including a discussion of more refined equations.)

Rank pattern The two most popular tests for monotonic association, which we introduce below, are based on the *rank pattern*, which is given by ranking the X_i among themselves, and then listing these ranks according to increasing values of the Y_i. For example, the following data

X:	-1.0	-1.3	0.8	1.1	-0.4	-0.3	0.9	-0.8
Y:	0.2	0.0	1.6	0.1	0.6	-0.6	1.0	-0.7

yield the following rank pattern

3 5 1 8 2 4 7 6

Problem 19.15. Suppose that ties are broken at random. Prove that the rank pattern is uniformly distributed among the permutations of $(1, \ldots, n)$ when X and Y are independent.

19.3.1 Spearman correlation

The Spearman correlation is the rank variant of the Pearson correlation. We start with the sample version. Let a_i denote the rank of x_i within $x_1, \ldots, x_n$ and b_i denote the rank of y_i within $y_1, \ldots, y_n$. The *Spearman sample correlation* is the Pearson sample correlation of $(a_1, b_1), \ldots, (a_n, b_n)$.

Problem 19.16. Show that this is a rank statistic.

Problem 19.17. Show that the Spearman sample correlation can be written as

$$r_S = 1 - \frac{6}{n^3 - n} \sum_{i=1}^{n} (a_i - b_i)^2.$$

Problem 19.18. Show that $r_S \in [-1, 1]$ and equal to 1 (resp. -1) if and only if there is a non-decreasing (resp. non-increasing) function g such that $y_i = g(x_i)$ for all i.

We have first defined the sample version of the Spearman correlation, the reason being that it is easy to motivate. One may wonder if there is a corresponding feature of the underlying distribution. Coming from another angle, is there a result analogous to Problem 19.9 here? The answer, encapsulated in the following problem, is Yes.

Proposition 19.19. *Let $\{(X_i, Y_i) : i \geq 1\}$ be iid bivariate numerical. Let $R_{S,n}$ denote the Spearman sample correlation of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then $R_{S,n} \to_P \rho_S$ as $n \to \infty$, where*

$$\rho_S := 3\,\mathbb{E}\big[\operatorname{sign}\big((X_1 - X_2)(Y_1 - Y_3)\big)\big].$$

(ρ_S is sometimes called *Spearman's ρ*.)

The Spearman correlation test, which in its two-sided variant rejects for large values of $|R_S|$, is not universally consistent, essentially because there are bivariate distributions with $\rho_S = 0$ that are not the product of their marginals.

Problem 19.20. Show that the distribution(s) of Problem 19.10 are examples of such distributions.

R corner. The function cor can be used to compute the Spearman correlation, while the function cor.test can be used to perform the Spearman correlation test. The p-value is computed analytically up to a certain sample size, and after that the large-sample null distribution is used. (It turns out that the second part of Proposition 19.11 applies to R_S.)

19.3.2 Kendall correlation

The *Kendall sample correlation* is defined as

$$r_K := \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \operatorname{sign}\big((x_j - x_i)(y_j - y_i)\big).$$

Problem 19.21. Show that this is a rank statistic.

Problem 19.22. Show that $r_K \in [-1, 1]$ and equal to 1 (resp. -1) if and only if there is a non-decreasing (resp. non-increasing) function g such that $y_i = g(x_i)$ for all i.

Here too, this statistic estimates a feature of the underlying distribution.

Proposition 19.23. *Let $\{(X_i, Y_i) : i \geq 1\}$ be iid bivariate numerical. Let $R_{K,n}$ denote the Kendall sample correlation of $(X_1, Y_1), \ldots, (X_n, Y_n)$. Then $R_{K,n} \to_P \rho_K$ as $n \to \infty$, where*

$$\rho_K := \mathbb{E}\big[\operatorname{sign}\big((X_1 - X_2)(Y_1 - Y_2)\big)\big].$$

(ρ_K is sometimes denoted by τ and called *Kendall's* τ.)

The Kendall correlation test, which in its two-sided variant rejects for large values of $|R_K|$, is not universally consistent, essentially because there are bivariate distributions with $\rho_K = 0$ that are not the product of their marginals.

Problem 19.24. Show that the distribution(s) of Problem 19.10 are examples of such distributions.

R corner. The function cor can be used to compute the Kendall correlation, while the function cor.test can be used to perform the Kendall correlation test. The p-value is computed analytically up to a certain sample size, and after that the large-sample null distribution is used. (It turns out that R_K is asymptotically normal.)

Problem 19.25. Show that, under the null hypothesis of independence, R_K has mean zero and variance given by $(4n+10)/9n(n-1)$.

19.4 Universal tests for independence

We saw that none of the correlation tests is universally consistent. This is because they focus on features that are not characteristic of independence. We present below approaches that can lead to universally consistent tests, which do so by looking at the entire distribution through its distribution, density, and characteristic function.

19.4.1 Tests based on the distribution function

Recall the definition of the distribution function of a random vector given in (6.4). For (X, Y) bivariate numerical, it is defined as

$$\mathsf{F}_{X,Y}(x,y) = \mathbb{P}(X \le x, Y \le y).$$

We saw in Proposition 6.3 that it characterizes the underlying distribution. In particular, the following is true.

Problem 19.26. X and Y are independent if and only if

$$\mathsf{F}_{X,Y}(x,y) = \mathsf{F}_X(x)\mathsf{F}_Y(y), \quad \text{for all } x, y \in \mathbb{R}.$$

Prove this claim.

In view of this, it becomes natural to consider test statistics of the form

$$\Delta\big(\widehat{\mathsf{F}}_{\boldsymbol{X},\boldsymbol{Y}}, \widehat{\mathsf{F}}_{\boldsymbol{X}} \otimes \widehat{\mathsf{F}}_{\boldsymbol{Y}}\big), \tag{19.4}$$

where Δ denotes a measure of dissimilarity between distribution functions as considered in Section 16.8.1, while

$\widehat{\mathsf{F}}_{\boldsymbol{X}}$, $\widehat{\mathsf{F}}_{\boldsymbol{Y}}$, and $\widehat{\mathsf{F}}_{\boldsymbol{X},\boldsymbol{Y}}$ denote the empirical distribution functions of the X, Y, and (X,Y) samples, respectively. In particular,

$$\widehat{\mathsf{F}}_{\boldsymbol{X},\boldsymbol{Y}}(x,y) := \frac{1}{n}\sum_{i=1}^{n}\{X_i \le x, Y_i \le y\}.$$

For example, the analogue of the Kolmogorov–Smirnov test rejects for large values of

$$\sup_{x,y\in\mathbb{R}} \left|\widehat{\mathsf{F}}_{\boldsymbol{X},\boldsymbol{Y}}(x,y) - \widehat{\mathsf{F}}_{\boldsymbol{X}}(x)\widehat{\mathsf{F}}_{\boldsymbol{Y}}(y)\right|. \tag{19.5}$$

Problem 19.27. Show that this statistic is a function of the ranks (so that the resulting test is a rank test).

Problem 19.28. Argue that this test is universally consistent.

Hoeffding [126] proposed, instead, the analogue of the Cramér–von Mises test, except in reverse, as it rejects which rejects for large values of $\Delta(\widehat{\mathsf{F}}_{\boldsymbol{X}} \otimes \widehat{\mathsf{F}}_{\boldsymbol{Y}}, \widehat{\mathsf{F}}_{\boldsymbol{X},\boldsymbol{Y}})$, with the Δ defined in (16.21).

Problem 19.29. The statistic (19.4), with the same Δ, appears to be an equally fine choice. Perform some numerical experiments to compare these two choices.

19.4.2 Tests based on the density

Tests for independence based on binning the observations, as studied in Problem 19.3, can interpreted as tests based on the density function.

Problem 19.30. In parallel with Section 19.4.1, but this time in analogy with Section 16.8.3, propose a class of tests for independence based on the density function. Speculate on whether the tests you propose are universally consistent, or not. Implement your favorite test among these in R, and perform some simulations to assess its power.

19.4.3 Tests based on the characteristic function

As we saw in Remark 7.59, the characteristic function of a random vector, (X,Y), is defined as

$$\varphi_{X,Y}(s,t) := \mathbb{E}[\exp(\imath(sX+tY))].$$

We also saw there that a distribution on $\mathbb{R}^2$ is characterized by its characteristic function.

Problem 19.31. Show that X and Y are independent if and only if $\varphi_{X,Y}(s,t) = \varphi_X(s)\varphi_Y(t)$ for all $s,t \in \mathbb{R}$.

Problem 19.32. Propose a class of tests for independence based on the characteristic function. Speculate on

whether the tests you propose are universally consistent, or not. Implement your favorite test among these in R, and perform some simulations to assess its power.

DISTANCE COVARIANCE In [229], Székely, Rizzo and Bakirov propose a test for independence based on pairwise distances, which in fact turns out to be based on the characteristic function.

Based on data $(x_1, y_1), \dots, (x_n, y_n)$, define

$$a_{ij} = |x_i - x_j|, \quad a_i = \frac{1}{n}\sum_{j=1}^{n} a_{ij}, \quad a = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij},$$

and

$$u_{ij} = a_{ij} - a_i - a_j + a.$$

Similarly, define

$$b_{ij} = |y_i - y_j|, \quad b_i = \frac{1}{n}\sum_{j=1}^{n} b_{ij}, \quad b = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij},$$

and

$$v_{ij} = b_{ij} - b_i - b_j + b.$$

The test rejects for large values of the sample *distance covariance* defined as

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} u_{ij} v_{ij}. \tag{19.6}$$

Being a test for independence, a p-value is typically obtained by (Monte Carlo) permutation.

To see how the test is based on the characteristic function, let $\widehat{\varphi}_{\boldsymbol{x}}$ denote the empirical characteristic function based on the sample $\boldsymbol{x}$, meaning the characteristic function of $\widehat{\mathsf{F}}_{\boldsymbol{x}}$, and define $\widehat{\varphi}_{\boldsymbol{y}}$ as well as $\widehat{\varphi}_{\boldsymbol{x},\boldsymbol{y}}$ analogously.

Problem 19.33. Show that, when computed on $\boldsymbol{x} = (x_1, \dots, x_n)$,

$$\widehat{\varphi}_{\boldsymbol{x}}(s) = \frac{1}{n}\sum_{j=1}^{n} \exp(\imath s x_j). \tag{19.7}$$

Prove that the sample characteristic function is pointwise consistent for the characteristic function, meaning that $\widehat{\varphi}_{\boldsymbol{X}_n}(s) \to_{\mathrm{P}} \varphi_X(s)$ as $n \to \infty$, for all $s \in \mathbb{R}$, where $\boldsymbol{X}_n = (X_1, \dots, X_n)$ are iid copies of a random variable X.

Repeat with the joint characteristic function, $\widehat{\varphi}_{\boldsymbol{x},\boldsymbol{y}}$.

Proposition 19.34. *The statistic* (19.6) *is equal to*

$$\frac{1}{\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{\left|\widehat{\varphi}_{\boldsymbol{X},\boldsymbol{Y}}(s,t) - \widehat{\varphi}_{\boldsymbol{X}}(s)\widehat{\varphi}_{\boldsymbol{Y}}(t)\right|^2}{s^2 t^2}\,\mathrm{d}s\,\mathrm{d}t$$

In view of this result, it is not too hard to believe that the sample distance covariance is consistent for the

distance covariance of (X, Y), defined as

$$\frac{1}{\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\left|\varphi_{X,Y}(s,t)-\varphi_X(s)\varphi_Y(t)\right|^2}{s^2t^2}\mathrm{d}s\,\mathrm{d}t.$$

And from this it is not too hard to argue that the distance covariance test is universally consistent.

Problem 19.35. The distance covariance is intimately related to the energy statistic of Problem 17.52. Can you see that? [This is analogous to how the Hoeffding test is related to the Cramér–von Mises test.]

19.5 Further topics

19.5.1 When one variable is categorical

In this chapter we have focused on the situation where both variables are numerical. And in Section 15.7 we addressed the situation where both are categorical (or, more generally, discrete). Suppose now that one of the variables, say X, is categorical while the other variable, Y, is numerical.

To derive tests for independence, we make a connection with goodness-of-fit testing. Testing for the independence of X and Y is equivalent to testing

$$Y \mid X = x \text{ is distributed as } Y, \quad \text{for all } x.$$

This is a problem of goodness-of-fit testing where the groups are the ones defined by the values of X. The only difference with the setting of Section 17.3 and Section 17.4.3 is that here the group sizes are random. However, conditional on $X_1, \ldots, X_n$, the setting is exactly that of goodness-of-fit testing, and the methods presented there are applicable. Note that this is an example of conditional inference (Section 22.1).

CHAPTER 20

MULTIPLE TESTING

In a wide range of real-life situations, not one but several, even many hypotheses are to be tested.

Example 20.1 (Genetics)**.** In genetics an important line of research revolves around discovering how an individual's genetic material influences his/her health. In particular, biologists have developed ways to measure how 'expressed' a gene is, and a typical experiment for understanding what genes are at play in a given disease can be described as follows. A number of subjects with the disease, and a number of subjects without the disease, are recruited. For each individual in the study, the expression levels of certain genes (m of them) are measured. For each gene, a test comparing the two groups is performed, so that m tests are performed in total [50]. In practice, for human subjects, m is on the order of 10,000. Experiments focusing on single nucleotide polymorphisms (SNP's) instead of genes result in an even larger number of tests, on the order of 100,000.

Example 20.2 (Surveillance)**.** In a surveillance setting, a signal is observed over time and the task is to detect a change in the signal of particular relevance. The signal can be almost anything and the change is typically in terms of features that are deemed important for the task at hand. Practical examples include the detection of fires from satellite images [149] and the detection of epidemics (aka *syndromic surveillance*) based on a variety of data such as transcripts from hospital emergency visits and pharmacy sales of over-the-counter drugs [121, 122]. In such settings, a test is applied at every location/time point.

Example 20.3 (Functional MRI)**.** Functional magnetic resonance imaging (fMRI) can be used as a non-invasive technique for understanding what regions of the human brain are active when performing a certain task [154]. In an experiment involving a single subject, a person's brain is observed over time while the individual is put under two or more conditions, where a condition might consist in performing a given task. The goal is then to identify which parts of the brain are most active under a condition relative to the other conditions. The identification is typically done by performing a test for each voxel, where a voxel represents a small unit of space (typically, a $3 \times 3 \times 3$ millimeter cube). There are on the order of 1,000,000 voxels.

For now, consider a simplified situation where m null hypotheses, $\mathcal{H}_1, \ldots, \mathcal{H}_m$, need to be tested. (Note that, in the present setting, $\mathcal{H}_1$ is a null hypothesis and not an alternative hypothesis.) We apply a test to each null hypothesis $\mathcal{H}_j$, resulting in a p-value denoted P_j. Assume for simplicity that the p-values are independent and that each P_j is uniform in $[0,1]$ under $\mathcal{H}_j$, that is,

$$\begin{gathered} P_1, \ldots, P_m \text{ are independent, with} \\ P_j \sim \text{Unif}(0,1) \text{ under } \mathcal{H}_j, \text{ for } j = 1, \ldots, m. \end{gathered} \tag{20.1}$$

Here are two aspects of the situation that illustrate the underlying difficulties:

- Suppose that we proceed as usual, choosing a level $\alpha \in (0,1)$ and rejecting $\mathcal{H}_j$ if $P_j \leq \alpha$. Then, even if all the hypotheses are true, on average there are αm rejections (all incorrect). In settings where very many tests are performed, meaning that m is large, choosing α to be the usual 0.05 or 0.01 leads to an impractically large number of rejections. Take Example 20.1, where (say) $m = 10,000$ tests are performed. Then rejecting at level $\alpha = 0.05$ leads to 500 rejections on average, even when all the hypotheses are true (meaning that no gene is truly differentially expressed when comparing the two conditions).

- The smallest p-value can be quite small even if all the hypotheses are true. Indeed, $\min_j P_j$ has expectation $1/(m+1)$ in that case.

When confronted with the task of testing a number of (null) hypotheses we talk of *multiple testing*.

20.1 Setting

We postulate a statistical model $(\Omega, \Sigma, \mathcal{P})$ as in Chapter 12, where Ω is the sample space containing all possible outcomes, Σ is the class of events of interest, and $\mathcal{P}$ is a family of distributions on Σ. We assume as before that $\mathcal{P}$ is parameterized as $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$.

Within this framework, we consider a situation where m null hypotheses need to be tested, with the jth null hypothesis being

$$\mathcal{H}_j : \theta_* \in \Theta_j,$$

for some given $\Theta_j \subset \Theta$. (Recall that θ_* denotes the true value of the parameter.) The alternative to $\mathcal{H}_j$ will simply the negation of $\mathcal{H}_j$ and denoted $\mathcal{H}_j^{\mathsf{c}}$ (which makes sense since it is the complement of $\mathcal{H}_j$).

Recall that a test is applied to each null hypothesis, resulting in a total of m p-values, denoted $P_1, \ldots, P_m$. What tests are used obviously depends on the situation. In Example 20.1, for instance, the rank-sum test could be applied to each hypothesis. We always assume that each p-value P_j is valid in the usual sense that it satisfies (12.22), meaning here that

$$\text{Under } \mathcal{H}_j : \quad \mathbb{P}(P_j \le \alpha) \le \alpha, \ \forall \alpha \in (0,1). \tag{20.2}$$

Remark 20.4. For the sake of conciseness, we focus on methods for multiple testing that are based on the p-values. Such methods are all based on the ordered p-values, denoted

$$p_{(1)} \le \cdots \le p_{(m)}. \tag{20.3}$$

For future reference, we note that ordering the p-values can be done in $O(m \log m)$ basic operations using a good sorting algorithm. We will let $\mathcal{H}_{(j)}$ denote the hypothesis associated with $p_{(j)}$.

20.1.1 Normal sequence model

The *normal sequence model* provides a stylized mathematical framework within which methods can be studied. Although it is too simple to accurately model real-life situations, it is nevertheless relevant, in part because common test statistics used in practice are approximately normal in large samples.

The model is as follows. We observe $Y_1, \dots, Y_m$, independent, with $Y_j \sim \mathcal{N}(\theta_j, 1)$. We are interested in testing $\mathcal{H}_1, \dots, \mathcal{H}_m$, where

$$\mathcal{H}_j : \theta_j = 0.$$

The problem can be consider one-sided, in which case $\mathcal{H}_j^{\mathsf{c}} : \theta_j > 0$; or two-sided, in which case $\mathcal{H}_j^{\mathsf{c}} : \theta_j \neq 0$.

Assuming the one-sided setting, it makes sense to reject $\mathcal{H}_j$ for large values of Y_j, since doing so is optimal for that particular hypothesis (Theorem 13.23). The corresponding p-value is $P_j := 1 - \Phi(Y_j)$, where Φ denotes the standard normal distribution function.

20.2 Global null hypothesis

The *global null hypothesis* (aka *complete null hypothesis*) is defined as

$$\mathcal{H}_0 : \text{“the hypotheses } \mathcal{H}_1, \dots, \mathcal{H}_m \text{ are all true”},$$

or, equivalently,

$$\mathcal{H}_0 : \theta_* \in \Theta_0 := \bigcap_{j=1}^{m} \Theta_j.$$

In most situations, a null hypothesis represents “business as usual”. We will assume this is the case throughout. Then the global null hypothesis represents “there is nothing at all going on”. Although one is typically interested in identifying the false hypotheses, testing the global null hypothesis might be relevant in some applications, for example, in surveillance settings (Example 20.2).

The global null hypothesis is just a null hypothesis. We present below some commonly used tests, all based on the available p-values. Such tests are sometimes called *combination tests*.

Remark 20.5. Recall a small p-value provides evidence against the hypothesis it is associated with. This explains why all the tests below are all one-sided.

Fisher test This test rejects for large values of

$$T(p_1, \dots, p_m) := -2 \sum_{j=1}^{m} \log p_j.$$

The test statistic was designed that way because its null distribution is stochastically dominated by the chi-squared distribution with $2m$ degrees of freedom. (This explains the presence of the factor 2.)

Problem 20.6. Assume (20.1). Show that, under the global null, T has the chi-squared distribution with $2m$ degrees of freedom.

LIPTAK–STOUFFER TEST This test rejects [151] for large values of

$$T(p_1,\dots,p_m) := \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \Phi^{-1}(1-p_j). \tag{20.4}$$

The test statistic was designed that way because its null distribution is stochastically dominated by the standard normal distribution.

Problem 20.7. Assume (20.1). Show that, under the global null, T has the standard normal distribution.

Problem 20.8. Consider the normal model of Section 20.1.1. First, express the test statistic as a function of $y_1,\dots,y_m$. Then, setting the level to some given $\alpha \in (0,1)$, provide a sufficient condition for the test to have power tending to 1. [Use Chebyshev's inequality.]

TIPPETT–ŠIDÁK TEST This test [209] rejects for small values of

$$T(p_1,\dots,p_m) := \min_{j=1,\dots,m} p_j. \tag{20.5}$$

Problem 20.9. Assume (20.1). Derive the distribution of T under the global null.

Problem 20.10. Repeat Problem 20.8 with this test. [This time, use Boole's inequality together with the fact that $1-\Phi(x) \le \phi(x)/x$, where ϕ denotes the density of the standard normal distribution.]

SIMES TEST This test [211] rejects for small values of

$$T(p_1,\dots,p_m) := \min_{j=1,\dots,m} m\,p_{(j)}/j.$$

Proposition 20.11. *Assuming that (20.1) holds, T has the uniform distribution in $[0,1]$ under the global null.*

Problem 20.12. Prove this proposition, and perform some simulations in R to numerically confirm it.

TUKEY TEST Better known as the *higher criticism* test, it comes from applying the one-sided Anderson–Darling procedure (16.22) to test the hypothesis that the p-values are iid uniform in $[0,1]$ — that is, the global null under (20.1).

Problem 20.13. Show that the test rejects for large values of

$$T(p_1,\dots,p_m) := \max_{j=1,\dots,m} \frac{j/m - p_{(j)}}{\sqrt{p_{(j)}(1-p_{(j)})}}.$$

Under the global null, T has a complicated distribution, but asymptotically ($m \to \infty$) it becomes of Gumbel type (after a proper standardization) [134].

Problem 20.14. In R, write a function that computes the statistic T above based on the p-values. Then, using that function, write another one hc.test that returns a p-value for the test based on a specified number of Monte Carlo replicates.

In general, any test for uniformity in its proper one-sided version is appropriate.

Problem 20.15. Show that the relevant form of the Berk–Jones test (Section 16.10.1) for the present setting rejects for small values of

$$T(p_1,\dots,p_m) := \min_{j=1,\dots,m} \text{Prob}\big(\text{Beta}(j, m-j+1) \le p_{(j)}\big).$$

Problem 20.16. Repeat Problem 20.14 with this test. [Change hc into bj.]

20.3 Multiple tests

Testing the global null amounts to weighing the evidence that one or several hypotheses are false. However, even if we reject, we do not know what hypotheses are doubtful. We now turn to the more ambitious goal of identifying the false hypotheses (if there are any). We will call a procedure for this task a *multiple test*.

While a test is a function of the data with values in $\{0,1\}$, with '1' indicating a rejection, a multiple test (Remark 20.4) is a function of the p-values with values in $\{0,1\}^m$, with '1' in the jth component indicating a rejection of $\mathcal{H}_j$. Thus a multiple test is of the form

$$\begin{array}{rccc} \varphi\colon & [0,1]^m & \to & \{0,1\}^m \\ & \boldsymbol{p} := (p_1,\dots,p_m) & \mapsto & (\varphi_1(\boldsymbol{p}),\dots,\varphi_m(\boldsymbol{p})) \end{array}$$

Seen as a function on $[0,1]^m$, φ_j is a test for $\mathcal{H}_j$, but possibly based on all the p-values instead of just p_j.

For $\theta \in \Theta$, let $h_j(\theta) = \{\theta \in \Theta_j\}$, so that $h_j(\theta) = 0$ if $\mathcal{H}_j$ is true, and $= 1$ if it is false. Also, let $m_0(\theta) = \#\{j : \theta \in \Theta_j\}$, which is the number of true hypotheses, and $m_1(\theta) = \#\{j : \theta \notin \Theta_j\}$, which is the number of false hypotheses. Note that $m_0(\theta) + m_1(\theta) = m$. For a given multiple test φ, define the following quantities:[94]

$$N_{0|0}(\varphi,\theta) = \#\{j : \varphi_j = 0 \text{ and } h_j(\theta) = 0\}, \tag{20.6}$$
$$N_{1|0}(\varphi,\theta) = \#\{j : \varphi_j = 1 \text{ and } h_j(\theta) = 0\}, \tag{20.7}$$
$$N_{0|1}(\varphi,\theta) = \#\{j : \varphi_j = 0 \text{ and } h_j(\theta) = 1\}, \tag{20.8}$$
$$N_{1|1}(\varphi,\theta) = \#\{j : \varphi_j = 1 \text{ and } h_j(\theta) = 1\}. \tag{20.9}$$

[94] A different notation is typically used in the literature, stemming from the influential paper [12], but that choice of notation is not particularly mnemonic. Instead, we follow [103].

(These are summarized in Table 20.1.) In particular, $N_{1|0}(\varphi,\theta)$ is the number of Type I errors and $N_{0|1}(\varphi,\theta)$ the number of Type II errors made by the multiple test φ when the true value of the parameter is θ. In particular, the total number of errors made by the multiple test is given by

$$N_{1|0}(\varphi,\theta) + N_{0|1}(\varphi,\theta) = \#\big\{j : \varphi_j \neq h_j(\theta)\big\},$$

and the total number of rejections is given by

$$R(\varphi) = N_{1|0}(\varphi,\theta) + N_{1|1}(\varphi,\theta) = \#\{j : \varphi_j = 1\}.$$

These counts are all functions of the p-values and, with the exception of $R(\varphi)$, of the true value of the parameter. This is left implicit.

For a single hypothesis, the accepted modus operandi is to control the level and, within that constraint, design a test that maximizes the power (as much as possible). We introduce some notion of level and power for multiple tests below. These apply to a given multiple test φ which is left implicit in places.

20.3.1 Notions of level for multiple tests

Family-wise error rate (FWER) For a long time, this was the main notion of level for multiple tests. It is

Table 20.1: The counts below summarize the result of applying a multiple test φ when the true value of the parameter is θ.

	No Rejection	Rejection	Total
True Null	$N_{0\|0}(\varphi,\theta)$	$N_{1\|0}(\varphi,\theta)$	$m_0(\theta)$
False Null	$N_{0\|1}(\varphi,\theta)$	$N_{1\|1}(\varphi,\theta)$	$m_1(\theta)$
Total	$m - R(\varphi)$	$R(\varphi)$	m

defined as the probability of making at least one Type I error or, using the notation of Table 20.1,

$$\text{FWER}(\varphi) = \sup_{\theta\in\Theta} \mathbb{P}_\theta\big(N_{1|0}(\varphi,\theta) \geq 1\big).$$

A multiple test φ controls the FWER at α if

$$\text{FWER}(\varphi) \leq \alpha.$$

Problem 20.17. Assume (20.1). Derive the FWER for the multiple test defined by

$$\varphi_j = \{P_j \leq \alpha\}. \tag{20.10}$$

False discovery rate (FDR) This notion is more recent. It was suggested in the mid 1990's by Benjamini

and Hochberg [12]. It is now the main notion of level used in large-scale multiple testing problems. It is defined as the expected proportion of incorrect rejections among all rejections or, using the notation of Table 20.1,

$$\mathrm{FDR}(\varphi) = \sup_{\theta \in \Theta} \mathbb{E}_\theta \left(\frac{N_{1|0}(\varphi, \theta)}{R(\varphi) \vee 1} \right).$$

A multiple test φ controls the FDR at α if

$$\mathrm{FDR}(\varphi) \leq \alpha.$$

(The name comes from the fact that, in most settings, a rejection indicates a *discovery*.)

Problem 20.18. As notions of level for multiple tests, the FDR is always less severe then the FWER. Indeed, show that in any situation and any multiple test φ,

$$\mathrm{FDR}(\varphi) \leq \mathrm{FWER}(\varphi).$$

20.3.2 Notions of power for multiple tests

Notions of power for multiple tests can be defined by analogy to the notions of level presented above, by having Type II errors play the role of Type I errors.

Problem 20.19. Define the power equivalent of FWER. (This quantity does not seem to have a name in the literature.)

False non-discovery rate (FNR) To define the power equivalent of FDR, one possibility is [103]

$$\mathrm{FNR}(\varphi) = \sup_{\theta \in \Theta} \mathbb{E}_\theta \left(\frac{N_{0|1}(\varphi, \theta)}{(m - R(\varphi)) \vee 1} \right).$$

This definition leads to a quantity that could look artificially small when m_0/m is close to 1 (meaning the vast majority of the hypotheses are true), which is common in practice, in which case the following variant might be preferred:

$$\mathrm{FNR}(\varphi) = \sup_{\theta \in \Theta} \frac{\mathbb{E}_\theta(N_{0|1}(\varphi, \theta))}{m_1(\theta) \vee 1}.$$

20.4 Methods for FWER control

For a given $\theta \in \Theta$, define

$$\mathcal{T}_\theta = \{j : \theta \in \Theta_j\}, \tag{20.11}$$

which is the subset of indices corresponding to true null hypotheses.

Tippett multiple test This multiple test $\mathcal{H}_j$ if

$$p_j \leq c_\alpha,$$

where c_α is such that

$$\sup_{\theta\in\Theta} \mathbb{P}_\theta\Big(\min_{j\in\mathcal{T}_\theta} P_j \le c_\alpha\Big) \le \alpha. \tag{20.12}$$

Note that c_α is a valid critical value for the Tippett test for the global null $\mathcal{H}_0$.

Proposition 20.20. *The Tippett multiple test controls the FWER at α.*

Proof. A Type I error occurs if the multiple test rejects some $\mathcal{H}_j$ with $j \in \mathcal{T}_\theta$. This happens with probability

$$\mathbb{P}_\theta\big(\exists j \in \mathcal{T}_\theta : \varphi_j = 1\big) = \mathbb{P}_\theta\Big(\min_{j\in\mathcal{T}_\theta} P_j \le c_\alpha\Big) \le \alpha,$$

using (20.12) at the end. □

Problem 20.21 (Šidák multiple test)**.** Show that, under (20.1), the inequality (20.12) holds with $c_\alpha = 1 - (1-\alpha)^m$.

Problem 20.22 (Bonferroni multiple test)**.** Show that, under all circumstances, the inequality (20.12) holds with $c_\alpha = \alpha/m$.

Holm multiple test This multiple test [127] rejects $\mathcal{H}_{(j)}$ if

$$p_{(k)} \le \alpha/(m-k+1) \text{ for all } k \le j.$$

Problem 20.23. Show that a brute force implementation based on this description requires on the order of $O(m^2)$ basic operations. In fact, the method can be implemented in order $O(m)$ basic operations after ordering the p-values. Describe such an implementation.

Remark 20.24 (Step down methods)**.** This is a *step-down* procedure as it moves from the most significant to the least significant p-value.

Proposition 20.25. *The Holm multiple test above controls the FWER at α.*

Proof. Recall (20.11) and let $j_0 = \arg\min_{j\in\mathcal{T}_\theta} p_{(j)}$. Note that j_0 is a function of the p-values. Since $|\mathcal{T}_\theta| = m_0(\theta)$, necessarily,

$$j_0 \le m - m_0(\theta) + 1. \tag{20.13}$$

The procedure makes an incorrect rejection if and only if it rejects $\mathcal{H}_{(j_0)}$, which happens exactly when

$$p_{(j)} \le \alpha/(m-j+1), \quad \text{for all } j \le j_0,$$

which in particular implies that

$$p_{(j_0)} \le \alpha/(m-j_0+1) \le \alpha/m_0(\theta), \tag{20.14}$$

by (20.13). But

$$\begin{aligned}\mathbb{P}_\theta(P_{(j_0)} \le \alpha/m_0(\theta)) &= \mathbb{P}_\theta\big(\min_{j\in\mathcal{T}_\theta} P_j \le \alpha/m_0(\theta)\big)\\ &\le \sum_{j\in\mathcal{T}_\theta} \mathbb{P}_\theta(P_j \le \alpha/m_0(\theta))\\ &\le \sum_{j\in\mathcal{T}_\theta} \alpha/m_0(\theta)\\ &= \alpha,\end{aligned}$$

using the union bound and then the fact that the p-values are valid in the sense of (20.2). □

Problem 20.26. Prove (20.13).

Problem 20.27. How would you change Holm's procedure if you knew that the p-values where independent?

In practice, Holm's procedure is thus always preferred over Bonferroni's.

Problem 20.28 (Holm vs Bonferroni)**.** Show that Holm's procedure is always preferable to Bonferroni's, in the (strongest possible) sense that any hypothesis that Bonferroni's rejects Holm's also rejects.

Hochberg multiple test This multiple test rejects $\mathcal{H}_{(j)}$ if

$$p_{(k)} \le \alpha/(m-k+1) \text{ for some } k \ge j.$$

Problem 20.29. Repeat Problem 20.23 but for the Hochberg multiple test.

Remark 20.30 (Step up methods)**.** This is a *step-up* procedure as it moves from the least significant to the most significant p-value.

Proposition 20.31. *The Hochberg multiple test above controls the FWER at α when the p-values are independent.*

Problem 20.32 (Hochberg vs Holm)**.** Prove that, if the p-values are independent, Hochberg's multiple test is more powerful than Holm's in the (strongest possible) sense that any hypothesis that Holm's rejects Hochberg's also rejects.

Remark 20.33 (Hommel multiple test)**.** There is another procedure, Hommel's, that is more powerful than Hochberg's. However, it is a bit complicated to describe, and we do not detail it here.

20.5 Methods for FDR control

Benjamini–Hochberg multiple test This multiple rejects $\mathcal{H}_{(j)}$ if

$$p_{(k)} \le k\alpha/m \text{ for some } k \ge j.$$

It was the first (and still the main) method guarantied to control the FDR at the desired level under appropriate conditions.

Problem 20.34. Let r denote the number of rejections when this method is applied to a particular situation. Show that the method rejects $\mathcal{H}_j$ if and only if $p_j \le r\alpha/m$.

Proposition 20.35. *The Benjamini–Hochberg multiple test above controls the FDR at αC_m, in general, where $C_m := 1 + 1/2 + \cdots + 1/m$, and at α if the p-values are independent.*

Proof sketch. We only prove[95] the first part, and only when each p-value is uniform in $[0,1]$ under its respective null. Let φ denote the multiple test and let $R = R(\varphi)$ denote the number of rejections when applied to a particular situation. Define $A_l = ((l-1)\alpha/m, l\alpha/m]$. We have

$$\frac{\{\varphi_j = 1\}}{R \vee 1} = \sum_{r=1}^{m} \{p_j \le r\tfrac{\alpha}{m}\} \frac{\{R(\varphi) = r\}}{r} \tag{20.15}$$

$$= \sum_{r=1}^{m} \sum_{l=1}^{r} \{p_j \in A_l\} \frac{\{R(\varphi) = r\}}{r} \tag{20.16}$$

$$= \sum_{l=1}^{m} \{p_j \in A_l\} \frac{\{R(\varphi) \ge l\}}{R(\varphi)} \tag{20.17}$$

$$\le \sum_{l=1}^{m} \frac{\{p_j \in A_l\}}{l}. \tag{20.18}$$

Thus,

$$\mathbb{E}_\theta\left(\frac{N_{1|0}(\varphi,\theta)}{R(\varphi) \vee 1}\right) = \mathbb{E}_\theta\left(\frac{\sum_{j\in\mathcal{T}_\theta}\{\varphi_j = 1\}}{R(\varphi) \vee 1}\right) \tag{20.19}$$

$$\le \sum_{j\in\mathcal{T}_\theta} \sum_{l=1}^{m} \frac{1}{l} \mathbb{P}_\theta(P_j \in A_l) \tag{20.20}$$

$$= \sum_{j\in\mathcal{T}_\theta} \sum_{l=1}^{m} \frac{1}{l}\frac{\alpha}{m} \tag{20.21}$$

$$\le \alpha C_m. \tag{20.22}$$

In the last line we used the fact that $|\mathcal{T}_\theta| = m_0(\theta) \le m$. □

Problem 20.36. Show that $C_m < \log m + 1$. [Start by showing that $1/k \le \int_{k-1}^{k} \mathrm{d}x/x$.]

[95] We learned of this proof from Emmanuel Candès.

One way to arrive at the Benjamini–Hochberg procedure is as follows. Consider the multiple test that rejects $\mathcal{H}_j$ when $p_j \le t$. With some abuse of notation, let $N_{1|0}(t)$ and $R(t)$ denote the corresponding number of Type I errors and total number of rejections, and define $F_t = N_{1|0}(t)/(R(t) \vee 1)$. Ideally, we would like to choose t largest such that $F_t \le \alpha$. However, $N_{1|0}(t)$ cannot be computed solely based on the p-values as it depends on knowing which hypotheses are true. The idea is to replace it by an estimate. Since we assume the p-values to be valid (20.2), we have

$$\mathbb{E}_\theta(N_{1|0}(t)) \le m_0(\theta)t \le mt.$$

If we replace $N_{1|0}(t)$ by mt, we effectively estimate F_t by $\hat{F}_t := mt/(R(t) \vee 1)$.

Problem 20.37. Let $\hat{t} = \max\{t : \hat{F}_t \le \alpha\}$. Show that the multiple test that rejects $\mathcal{H}_j$ when $p_j \le \hat{t}$ is Benjamini–Hochberg's.

20.6 Meta-analysis

Meta-analysis is a branch of Statistics/Epidemiology that focuses on combining multiple studies in order to reach stronger conclusions on a particular issue.

Example 20.38 (Bone density and fractures)**.** The study [162] is a "meta-analysis of prospective cohort studies published between 1985 and end of 1994 with a baseline measurement of bone density in women and subsequent follow up for fractures". The stated purpose of this analysis was to "determine the ability of measurements of bone density in women to predict later fractures". Combined, the studies comprised "eleven separate study populations with about 90,000 person years of observation time and over 2,000 fractures".

Example 20.39 (Alcohol consumption)**.** The paper [223] presents a meta-analysis of 87 studies on the relationship between alcohol consumption and all-cause mortality. Some previous studies had concluded that consuming a small amount of alcohol (1-2 drinks per day) was associated with a slightly longer lifespan. The authors here argue that this association can be explained in great part by the classification of former drinkers (who might have stopped drinking because of health issues) as abstainers.

Remark 20.40. As one would expect, there are meta-analyses of meta-analyses [56].

It is often the case that several studies examine the same effect, and it is rather tempting to use all this information combined to boost the power of the statistical

inference. This is possible under restrictive assumptions. In particular, the studies have to be comparable.

The bulk of the effort in a meta-analysis goes, in fact, to deciding which studies to include. There are a number of criteria for that, which by nature are ad hoc, although there are some guidelines [124, Ch 5]. Importantly, the studies need to be comparable, and judging of that often requires domain-specific knowledge.

Otherwise, in terms of methods for inference, the meta-analysist makes use of various tests, including combination tests (Section 20.2). Some, mostly ad hoc, methods for detecting the presence of *publication bias* (introduced in Section 23.8.2) have also been developed. Most, like the popular *funnel plot*, are based on the so-called *small study effect*, which is the empirically-observed fact that small studies are more prone to publication bias compared to larger studies, presumably because large studies require more funding and cannot remain unpublished as easily.

20.6.1 Cochran–Mantel–Haenszel test

It is not uncommon for studies in the medical field (e.g., clinical trials) to result in a 2-by-2 table. This happens, for example, with a completely randomized design on two treatments and a binary outcome of 'success' or 'failure'. The *Cochran–Mantel–Haenszel (CMH) test* is applied when examining a number of such studies. The goal is to determine whether there is a treatment effect or not, and (optionally) to specify the direction of the effect when it is determined that there is an effect.

Example 20.41 (Low protein diets in chronic renal insufficiency)**.** In [90], a meta-analysis is undertaken to better assess the impact that low protein diets have on chronic renal insufficiency. A total of 46 clinical trials where examined, from which 6 were selected (5 European and 1 Australian, between 1982 and 1991). This amounted to a combined sample of size 890 subjects with mild to severe chronic renal failure. Among these, 450 patients received a low protein diet (treatment) and 440 a control diet. Assignment was at random in all trials. Each subject was followed for at least one year. The main outcome was renal death (start of dialysis or death) during the study. Table 20.2 provides a summary in the form of six contingency tables (one for each study).

Simplest approach is arguably to *collapse* of all these 2-by-2 tables into a single 2-by-2 table, followed by applying one of the tests seen in Section 15.3, or Section 15.4, or Section 15.6. Some meta-analysts, however, are reluctant to do that because, although the studies are supposed to be comparable, they are invariably performed on samples

Table 20.2: The following is taken from Table 3 in [90]. See description in Example 20.41.

	Diet	Survived	Died
Study 1	Control	95	15
	Treatment	110	8
Study 2	Control	2	7
	Treatment	5	5
Study 3	Control	194	32
	Treatment	209	21
Study 4	Control	8	17
	Treatment	14	11
Study 5	Control	25	13
	Treatment	30	4
Study 6	Control	21	11
	Treatment	21	12

from different populations, and collapsing the tables could dilute the strength of association in some of the studies. (Remember Simpson's paradox.) The CMH test does not combine tables.

For notation, assume there are m studies and let the contingency table resulting from jth study be as follows

	Success	Failure
Treatment	a_j	b_j
Control	c_j	d_j

The one-sided CMH test is based on rejecting for large values of the total number of successes in the treatment group, namely $\sum_{j=1}^{m} a_j$.

Problem 20.42. In which direction is the test one-sided? How would you define a two-sided CMH test?

Problem 20.43 (Normal approximation)**.** The classical version of the test relies on a normal approximation for calibration. Specify this normal approximation. Can you justify this normal approximation when the results from the studies are independent?

Problem 20.44. How would you calibrate the test by Monte Carlo simulation?

Problem 20.45. Compare this with the test that collapses the tables into a single table. (This test is based on the same statistic. What distinguishes the tests is in how the p-value is computed.)

It is rather natural to approach this problem from a multiple testing perspective. After all, we are testing a

global null.

Problem 20.46. Can you propose a combination test based on Fisher's exact test (Section 15.6) applied to each table? [Recall Remark 15.31.]

20.6.2 File drawer problem

Rosenthal [197] refers to publication bias (Section 23.8.2) as the *file drawer problem*. This was the 1970's, so that manuscripts were written on paper and stored in a file drawer. Unpublished papers would remain hidden from view in such file drawers.

Rosenthal considers a setting where some published papers address the same general question (formalized as a null hypothesis) and report on a p-value obtained from a test of significance performed on independently collected data. Having access to all these published papers, Rosenthal asks the question: How many papers addressing the same question (and each performing a test on a separate dataset) would have to be left unpublished, stored away in a file drawer, to offset the combined significance resulting from the published papers?

Assume that we have available m studies testing for the presence of the same effect (e.g., effectiveness of a particular drug compared to a placebo), with Study j resulting in a p-value denoted p_j. The Liptak-Stouffer (LS) combination test seen in (20.4) rejects for large values of $y := \sum_{j=1}^{m} z_j$, where $z_j := \Phi^{-1}(1-p_j)$ is the z-score associated with p_j. (Φ denotes the standard normal distribution function.) The test is significant at level γ if $y/\sqrt{m} \geq \Phi^{-1}(1-\gamma)$.

Rosenthal's method Suppose the LS test is significant at level γ. In an effort to answer his own question, Rosenthal then computes the *fail-safe number* [96], defined as

$$\hat{n} := \min\left\{n : y/\sqrt{m+n} < \Phi^{-1}(1-\gamma)\right\}.$$

To motivate this definition, let $z_{m+1}, \ldots, z_{m+n}$ denote the z-scores corresponding to the studies that have remained unpublished. Let $y' = \sum_{j=m+1}^{m+n} z_j$. If we had access to all studies, published and unpublished, we would base our inference on $\sum_{j=1}^{m+n} z_j = y + y'$. Specifically, we would fail to reject at level γ when

$$(y+y')/\sqrt{m+n} < \Phi^{-1}(1-\gamma).$$

[96] Rosenthal does not make a connection with the Liptak-Stouffer test in his original paper. Also, we are working with p-values and transforming them into z-scores. Rosenthal works directly with z-scores and assumes a two-sided situation. This leads to a different definition for the fail-safe number.

In light of this, the fail-safe number is based on replacing the unobservable y' with 0, which is the mean of the z_j when there is no effect and no selection.

In a variant, Iyengar and Greenhouse [133] suggest replacing 0 above with the expected value of a standard normal conditional on not exceeding $\Phi^{-1}(1-\alpha)$, where α is the common level of significance a study is typically required to achieve to be published ($\alpha = 0.05$ is common). (Note that α may be different from γ.)

Gleser and Olkin's method Gleser and Olkin [107] take a more principled approach, based on a worst-case scenario in which, out of a total of $m+n$ studies, each yielding a p-value as above, we only get to observe the smallest m. The goal remains to estimate n assuming that there is no effect. Assume that all these p-values are independent and that each p-value is uniformly distributed in $[0,1]$ under its associated null hypothesis. Thus, denoting $p_{(1)} \le \cdots \le p_{(m+n)}$ the ordered p-values, we only get to observe $p_{(1)}, \dots, p_{(m)}$.

We have a bona fide probability model, with likelihood

$$\frac{(m+n)!}{n!}(1-q_m)^n \{0 \le q_1 \le \cdots \le q_m \le 1\},$$

where we wrote q_j in place of $p_{(j)}$ for clarity. (This plays the role of null model in the present context.)

Problem 20.47. Show that $p_{(m)}$ is sufficient for n and derive its distribution.

Problem 20.48. Show that the maximum likelihood estimator is

$$\hat{n}_{\text{MLE}} := \lfloor m(1-p_{(m)})/p_{(m)} \rfloor.$$

In fact, Gleser and Olkin prefer to use an unbiased estimator.

Problem 20.49. Show that the following estimator is unbiased

$$\hat{n} := \frac{m(1-p_{(m)})-1}{p_{(m)}}.$$

(This estimator happens to be the only unbiased estimator of n based on $p_{(m)}$.)

20.7 Further topics

20.7.1 Positive dependence

We say that a subset $\mathcal{Q} \subset \mathbb{R}^m$ is an increasing set if $\boldsymbol{u} = (u_1, \dots, u_m) \in \mathcal{Q} \Rightarrow \boldsymbol{v} = (v_1, \dots, v_m) \in \mathcal{Q}$ whenever $v_j \ge u_j$ for all j. We say that a random vector $\boldsymbol{Z} = (Z_1, \dots, Z_m)$ is *PRDS* [97] on $\mathcal{J} \subset \{1, \dots, m\}$ if for any $j \in \mathcal{J}$ and any

[97] This stands for *positive regression dependency on each one from a subset.*

increasing set $\mathcal{Q} \subset \mathbb{R}^m$, $z \mapsto \mathbb{P}(\boldsymbol{Z} \in \mathcal{Q} \mid Z_j = z)$ is non-decreasing in z.

Problem 20.50. Show that the PRDS property is invariant with respect to monotone transformations in the sense that, if $(Z_1, \ldots, Z_m)$ is PRDS on $\mathcal{J}$ and $f_1, \ldots, f_m$ are non-decreasing functions, then $(f_1(Z_1), \ldots, f_m(Z_m))$ is also PRDS on $\mathcal{J}$.

Theorem 20.51. *The Benjamini–Hochberg multiple test controls the FDR at the desired level when the p-values $(P_1, \ldots, P_m)$ are PRDS on the set of true nulls* (20.11).

20.7.2 Adjusted p-values

An *adjusted p-value* is such that, when it is below α (the desired FWER or FDR level) the corresponding null hypothesis is rejected. Take, for example, Bonferroni's multiple test meant to control the FWER at α. The corresponding Bonferroni adjusted p-values are defined as

$$p_j^{\text{Bonf}} := (m\, p_j) \wedge 1.$$

And indeed, the Bonferroni multiple test with parameter α rejects $\mathcal{H}_j$ if and only if $p_j^{\text{Bonf}} \leq \alpha$.

Problem 20.52. Show that the Holm adjusted p-value for $\mathcal{H}_j$ can be defined as

$$p_j^{\text{Holm}} := \max\left\{(m-k+1)p_k : k \leq j\right\} \wedge 1,$$

in the sense that the Holm multiple test with parameter α rejects $\mathcal{H}_j$ if and only if $p_j^{\text{Holm}} \leq \alpha$.

Problem 20.53. Derive the Hochberg adjusted p-values.

Problem 20.54. Derive the Benjamini–Hochberg adjusted p-values (called *q-values* in [225]).

R corner. The multiple tests presented here are implemented in the function `p.adjust`, which returns the adjusted p-values. (The default multiple test is Holm's, which is the safest since it applies regardless of the dependence structure of the p-values.)

20.8 Additional problems

Problem 20.55 (Comparing global tests)**.** Perform some numerical experiments to compare the tests for the global null presented in Section 20.2 in the normal sequence model of Section 20.1.1.

Problem 20.56 (Comparing multiple tests for FWER control)**.** Perform some numerical experiments to compare

the multiple tests available in p.adjust for FWER control. Do this in the setting of the normal sequence model of Section 20.1.1.

Problem 20.57 (k-FWER)**.** For a multiple test φ, and for $k \geq 1$ integer, the *k-FWER* is defined as

$$\mathrm{FWER}_k(\varphi) = \sup_{\theta \in \Theta} \mathbb{P}_\theta\big(N_{1|0}(\varphi, \theta) \geq k\big).$$

Note that the 1-FWER coincides with the FWER. In general, k stands for the number of Type I errors that the researcher is willing to tolerate.

(i) (Bonferroni) Show that the multiple test that rejects $\mathcal{H}_j$ if $P_j \leq \alpha k/m$ controls the k-FWER at α.

(ii) (Tippett) More generally, how would you change the definition of c_α given in (20.12) to control the k-FWER at the desired level?

Problem 20.58 (marginal FDR)**.** For a multiple test φ, the *marginal false discovery rate (mFDR)* is defined as

$$\mathrm{mFDR}(\varphi) = \sup_{\theta \in \Theta} \frac{\mathbb{E}_\theta[N_{1|0}(\varphi, \theta)]}{\mathbb{E}_\theta[R(\varphi)]}.$$

(i) Show that the mFDR is the probability that a null hypothesis, chosen uniformly at random among those that were rejected, is true. (Thus, in looser terms, the mFDR is the probability that a claimed discovery is actually false.)

(ii) Show that, in general, the mFDR cannot be controlled at any level strictly less than 1.

CHAPTER 21

REGRESSION ANALYSIS

Beyond quantifying the amount of association between (necessarily paired) variables, as was the goal in Chapter 19, *regression analysis* aims at describing that association. Partly because the task is so much more ambitious, the literature on the topic is vast. Our treatment in this chapter is necessarily very limited in scope but provides some essentials. For more on regression analysis, we recommend [120], or the lighter version [135].

We assume the experiment of interest results in paired observations, assumed to be iid from an underlying unknown distribution and denoted

$$\boldsymbol{D} := \{(X_1, Y_1), \dots, (X_n, Y_n)\}. \tag{21.1}$$

Throughout

$$\boldsymbol{d} := \{(x_1, y_1), \dots, (x_n, y_n)\} \tag{21.2}$$

will denote a realization of $\boldsymbol{D}$. We also let $|\boldsymbol{d}|$ denote the size of the sample $\boldsymbol{d}$, also denoted by n above and in what follows.

At this stage, the variables x and y can be of any type. Suppose we set as a general goal that of predicting y as a function of x. In that case, x be will called the *predictor variable* and y will called the *response variable*.

There are two main motives for performing a regression analysis.

- *Modeling* The main purpose is to build a model that describes how the response variable varies as a function of a value of the predictor variable. A simple, parsimonious model is often sought to ease interpretability. Finding such a model is desirable, for example, in fundamental sciences like physics, chemistry, or biology, where gaining a functional understanding is important. An example of that is relating the vapor pressure of a pure liquid to its temperature (Example 19.14).
- *Prediction* The main purpose is to predict the value of the response variable given the predictor variable. Examples of applications where this is needed abound in engineering and a broad range of industries (insurance, finance, marketing, etc). For example, in the insurance industry, when pricing a policy, the predictor variable encapsulates the available information about what is being insured, and the response variable is a measure of risk that the insurance company would take if underwriting the policy. In this context, a procedure is solely evaluated based on its performance at predicting that risk, and can otherwise be very complicated and have no simple interpretation.

We focus here on the goal of prediction because it is simpler, its scope is broader in terms of applications, and it is easier to formalize mathematically.

Example 21.1 (Real estate prices)**.** A number of real estate websites, besides listing properties that are currently on the market (for which the asking price is set by the sellers), also estimate the price of properties that are not currently for sale, using proprietary regression models that take in all the available information on these properties (prediction variable) and returns an estimated value (response). For a residential property, the prediction variable may include square footage, number of bedrooms, number of bathrooms, location, etc.

Example 21.2 (MNIST dataset)**.** The special case where the response is categorical is most often called *classification* instead of regression. The MNIST dataset is a dataset that researchers have used for many years for comparing procedures for classification. Each observation is a 28×28 grey level pixel image of a handwritten digit, which is labeled accordingly. The main goal is to recognize the

digits. This can be cast as a classification task where each observation is of the form (x, y) with $x \in \mathbb{R}^p$ with $p = 28 \times 28 = 784$ and $y \in \{0, 1, \ldots, 9\}$. Importantly, y is categorical here. Indeed, the fact that y is a digit, and therefore a number, is irrelevant, as the order between the digits is not pertinent to the classification task.

21.1 Prediction

Stating a prediction problem amounts to specifying the class of possible distributions for (X, Y), as well as a functional quantifying the error made by a procedure.

21.1.1 Loss and risk

Assume that X takes values in $\mathcal{X}$ and Y takes values in $\mathcal{Y}$. Most of the time, both $\mathcal{X}$ and $\mathcal{Y}$ are subsets of Euclidean spaces. Choose a function $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ meant to measure dissimilarity. This function $\mathcal{L}$ is referred to as the *loss function*. For a function $f : \mathcal{X} \to \mathcal{Y}$, define its *risk* (aka *expected loss* or *prediction error*) as

$$\mathcal{R}(f) = \mathbb{E}\big[\mathcal{L}(Y, f(X))\big], \tag{21.3}$$

where the expectation is with respect to (X, Y). Thus $\mathcal{R}(f)$ quantifies the average loss, measured in terms of $\mathcal{L}$, when predicting Y by $f(X)$.

Note that f has to be measurable. Henceforth, we let $\mathcal{M}$ denote the class of measurable functions from $\mathcal{X}$ to $\mathcal{Y}$. (As usual, $\mathcal{X}$ and $\mathcal{Y}$ are implicitly equipped with σ-algebras.)

Example 21.3 (Numerical response)**.** In the important case where $\mathcal{Y} = \mathbb{R}$, $\mathcal{L}$ is very often chosen of the form $\mathcal{L}(y, y') = |y - y'|^\gamma$ for some $\gamma > 0$. Popular choices in that family of losses include

$$\begin{aligned} \textit{squared error loss} \qquad & \mathcal{L}(y, y') = (y - y')^2, \\ \textit{absolute loss} \qquad & \mathcal{L}(y, y') = |y - y'|. \end{aligned}$$

Example 21.4 (Categorical response)**.** Another important example is where $\mathcal{Y}$ is a discrete set, which arises when the response is categorical, i.e., in a classification setting. A popular choice of loss function is

$$\textit{0-1 loss} \qquad \mathcal{L}(y, y') = \{y \neq y'\}. \tag{21.4}$$

21.1.2 Regression estimators

Having chosen a loss function, we turn to finding a function with relatively low risk as defined in (21.3) based on the

available data. This is the role of a *regression estimator*, which is of the form

$$\begin{aligned} \hat{f} : \mathcal{D} &\longrightarrow \mathcal{M} \\ \boldsymbol{d} &\longmapsto \hat{f}_{\boldsymbol{d}} \end{aligned}$$

where $\mathcal{D} := \bigcup_{n\geq 1}(\mathcal{X}\times\mathcal{Y})^n$ is where the data (21.2) resides.

The action of applying $\hat{f}$ to data $\boldsymbol{d}$ to obtain $\hat{f}_{\boldsymbol{d}}$ is referred to as *fitting* or *training* the regression estimator $\hat{f}$ on the data $\boldsymbol{d}$, and the resulting estimate, $\hat{f}_{\boldsymbol{d}}$, is then referred to as the fitted or trained estimator. The result is a (measurable) function from $\mathcal{X}$ to $\mathcal{Y}$, which is meant to predict Y from future observations of X.

An estimator being a random function, we use its *expected risk*, or *generalization error*, to quantify its performance, which for an estimator $\hat{f}$ is defined as

$$\bar{\mathcal{R}}_n(\hat{f}) = \mathbb{E}[\mathcal{R}(\hat{f}_{\boldsymbol{D}})], \tag{21.5}$$

where the expectation is with respect to a dataset $\boldsymbol{D}$ of size n.

21.1.3 Regression functions

A more ambitious goal than just finding a function with low risk is to approach a minimizer, meaning an element of

$$\mathcal{F}_* := \arg\min_f \mathcal{R}(f), \tag{21.6}$$

when this set is not empty. Any element of $\mathcal{F}_*$ is called a *regression function*. In many cases of interest, $\mathcal{F}_*$ is (essentially) a singleton.

It helps to work conditional on X, because of the following.

Problem 21.5 (Conditioning on X)**.** Show that

$$\inf_f \mathbb{E}\big[\mathcal{L}(Y, f(X))\big] \geq \mathbb{E}\Big(\inf_{y'\in\mathcal{Y}} \mathbb{E}\big[\mathcal{L}(Y, y') \mid X\big]\Big).$$

Deduce that any function f satisfying

$$f(x) \in \arg\min_{y'\in\mathcal{Y}} \mathbb{E}\big[\mathcal{L}(Y, y') \mid X = x\big], \quad \text{for all } x, \tag{21.7}$$

minimizes the risk (21.3). (This assumes that the minimum above is attained for any x.)

Problem 21.6 (Mean regression)**.** Consider a setting where $\mathcal{Y} = \mathbb{R}$. Assume that Y has a 2nd moment and take $\mathcal{L}$ to be the squared error loss. Show that the minimum in (21.7) is uniquely attained at $y' = \mathbb{E}[Y \mid X = x]$, so that the risk (21.3) is minimized by

$$f_*(x) := \mathbb{E}[Y \mid X = x]. \tag{21.8}$$

[Use Problem 7.91.]

Problem 21.7 (Median regression)**.** Consider a setting where $\mathcal{Y} = \mathbb{R}$. Assume that Y has a 1st moment and take $\mathcal{L}$ to be the absolute loss. Show that the minimum in (21.7) is attained at any median of $Y \,|\, X = x$, and only there. In the special case where, for all x, $Y \,|\, X = x$ has a unique median, the risk (21.3) is thus minimized by

$$f_*(x) := \mathrm{Med}(Y \,|\, X = x). \tag{21.9}$$

[Use Problem 7.92.]

Problem 21.8 (Classification with 0-1 loss)**.** Consider a classification setting with 0-1 loss. Show that the minimum in (21.7) is attained at any y' maximizing $y \mapsto \mathbb{P}(Y = y \,|\, X = x)$, so that the risk (21.3) is minimized by any function f_* satisfying

$$f_*(x) \in \arg\max_{y \in \mathcal{Y}} \mathbb{P}(Y = y \,|\, X = x). \tag{21.10}$$

Such a function is called a *Bayes classifier*.

21.2 Local methods

The methods that follow are said to be *local*, in the sense that the value of the estimated function at some point $x \in \mathcal{X}$ is computed based on the observations (x_i, y_i) with x_i in a neighborhood of x.

Let δ denote a dissimilarity on $\mathcal{X}$, so that $\delta(x, x')$ is a measure of how dissimilar $x, x' \in \mathcal{X}$ are. 'Local' is henceforth understood in the context of $\mathcal{X}$ equipped with the dissimilarity δ.

Example 21.9 (Euclidean metric)**.** When $\mathcal{X}$ is a Euclidean space, it is most common to use the Euclidean metric, meaning that $\delta(x, x') = \|x - x'\|$, with $\|\cdot\|$ denoting the Euclidean norm.

There are two main types of neighborhoods used in practice:

- *Ball neighbors* For $x \in \mathcal{X}$ and $h > 0$, its h-ball neighbors are indexed by

$$I_{\boldsymbol{d}}^h(x) := \big\{i : x_i \in \mathcal{B}_h(x)\big\}, \tag{21.11}$$

where

$$\mathcal{B}_h(x) := \big\{x' \in \mathcal{X} : \delta(x', x) \le h\big\}.$$

(This is the ball centered at x and of radius h defined by the dissimilarity δ.)

- *Nearest neighbors* For $x \in \mathcal{X}$ and $k \ge 1$ integer, its k-nearest neighbors are indexed by

$$J_{\boldsymbol{d}}^k(x) := \big\{i : \delta(x_i, x) \text{ is among the } k \text{ smallest in } \delta(x_1, x), \dots, \delta(x_n, x)\big\}. \tag{21.12}$$

We mostly work with ball neighbors for concreteness.

We assume throughout that the distribution of X has support $\mathcal{X}$, in the sense that $\mathbb{P}(X \in \mathcal{B}_h(x)) > 0$ for every $x \in \mathcal{X}$ and every $h > 0$.

Problem 21.10. In that case, show that for every $x \in \mathcal{X}$,

$$\min_{i=1,\dots,n} \delta(X_i, x) \xrightarrow{\mathrm{P}} 0, \quad \text{as } n \to \infty. \tag{21.13}$$

21.2.1 Local methods for regression

Consider the setting where $\mathcal{Y} = \mathbb{R}$ and the loss is the squared error loss, so that the regression function is the conditional expectation f_* given in (21.8). The methods that we present below aim directly at estimating f_*.

Local average Computing the regression function as given in (21.8) is impossible without access to the distribution of $Y \,|\, X$, which is unknown. A local average approach attempts to estimate this function by making two approximations:

- *Conditioning on a neighborhood* While in (21.8) the conditioning is on $X = x$, we approximate this by conditioning on a neighborhood. Using a ball neighborhood, the approximation is

$$\mathbb{E}[Y \,|\, X \in \mathcal{B}_h(x)] \approx \mathbb{E}[Y \,|\, X = x]. \tag{21.14}$$

  when h is small. This approximation is reasonable when the regression function f_* is continuous, and can indeed be shown to be valid under additional mild assumptions (Problem 21.47).

- *Averaging* As we often do, we estimate an expectation with an average, yielding the approximation

$$\frac{1}{|I^h_{\boldsymbol{D}}(x)|} \sum_{i \in I^h_{\boldsymbol{D}}(x)} Y_i \approx \mathbb{E}[Y \,|\, X \in \mathcal{B}_h(x)].$$

  By the Law of Large Numbers, the approximation is valid when $|I^h_{\boldsymbol{D}}(x)|$ is large.

The *local average* estimator combines these two approximations to take the form

$$\hat{f}^h_{\boldsymbol{d}}(x) := \frac{1}{|I^h_{\boldsymbol{d}}(x)|} \sum_{i \in I^h_{\boldsymbol{d}}(x)} y_i. \tag{21.15}$$

The tuning parameter h is often called the *bandwidth*.

Kernel regression Kernel regression is a form of weighted local average, and as such includes (21.15) as a special case. Choose a non-increasing function $\mathsf{Q} : \mathbb{R}_+ \to \mathbb{R}_+$, and for $h > 0$, define

$$\mathsf{K}_h(x', x) = \mathsf{Q}(\delta(x', x)/h). \tag{21.16}$$

The function Q is sometimes referred to as the kernel function, and most often chosen compactly supported or fast-decaying.

The *Nadaraya–Watson estimate*[98] is defined as

$$\hat{f}^h_{\boldsymbol{d}}(x) := \sum_{i=1}^n w_{i,h}(x) y_i, \tag{21.17}$$

where

$$w_{i,h}(x) := \frac{\mathsf{K}_h(x_i, x)}{\sum_{j=1}^n \mathsf{K}_h(x_j, x)}.$$

R corner. The Nadaraya–Watson kernel regression estimate can be computed using the function `ksmooth`. Several choices of kernel function are offered.

Remark 21.11. Kernel regression is analogous to kernel density estimation (Section 16.10.6).

LOCAL LINEAR REGRESSION A kernel regression estimate is built by fitting a constant locally (Problem 21.51). *Local linear regression* is based on fitting an affine function locally. For this to make sense, we need to assume that $\mathcal{X}$ is a Euclidean space, and we assume that δ is a norm for concreteness.

[98] Named after Èlizbar Nadaraya (1936 -) and Geoffrey Stuart Watson (1921 - 1998).

Assuming the regression function f_* is differentiable, we have the Taylor expansion

$$f_*(x') \approx f_*(x) + \nabla f_*(x)^\top (x' - x), \tag{21.18}$$

the approximation being accurate to first order when $\delta(x', x)$ is small. Having noticed that $x' \mapsto f_*(x) + \nabla f_*(x)^\top (x' - x)$ is an affine function, its coefficients are estimated in a neighborhood of x (since the approximation is only valid near x).

In more detail, having chosen a kernel function Q and a bandwidth $h > 0$, for $x \in \mathcal{X}$, define $(a^h_{\boldsymbol{d}}(x), b^h_{\boldsymbol{d}}(x))$ to be the solution to

$$\min_{(a,b)} \sum_{i=1}^n \mathsf{K}_h(x_i, x)\big(y_i - a - b^\top (x_i - x)\big)^2.$$

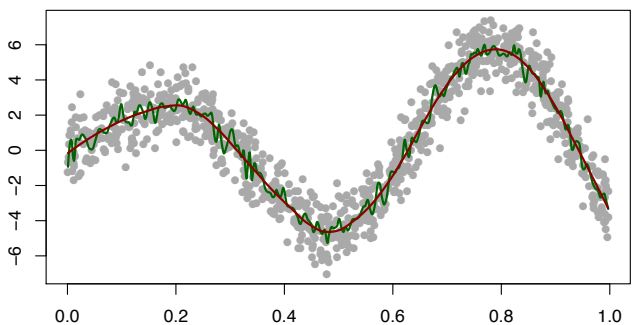
The intercept, $a^h_{\boldsymbol{d}}(x)$, is meant to estimate $f_*(x)$, while the slope, $b^h_{\boldsymbol{d}}(x)$, is meant to estimate $\nabla f_*(x)$. The local linear regression estimate is simply the intercept, namely

$$\hat{f}^h_{\boldsymbol{d}}(x) = a^h_{\boldsymbol{d}}(x), \text{ as computed above.}$$

Figure 21.1 illustrates an application of local linear regression to synthetic data.

R corner. The function `loess` implements local linear regression.

Figure 21.1: An example of application of local linear regression. The model that was generated is the following: $Y_i = f_*(X_i) + Z_i$, with $X_1, \dots, X_n$ iid uniform in $[0,1]$, $f_*(x) = (1+10x-5x^2)\sin(10x)$, and (independently) $Z_1, \dots, Z_n$ iid standard normal. Local linear regression was applied with two different values of h, resulting in a rough (green) curve and a smooth (red) curve, with the latter coming very close to the function f_*.



Remark 21.12. Loader [153] proposes a local linear density estimation method based on a Taylor expansion of the logarithm of the density.

21.2.2 Local methods for classification

Consider the setting where $\mathcal{Y}$ is discrete and the loss is the 0-1 loss, so that the regression function is the Bayes classifier f_* given in (21.10). The methods that we present below aim directly at estimating f_*.

Local majority vote The arguments that lead to the local average of (21.15) can be adapted to the present setting, starting from (21.10) instead of (21.8). The end result is the following classifier

$$\hat{f}_{\boldsymbol{d}}^h(x) \in \arg\max_{y \in \mathcal{Y}} \sum_{i \in I_{\boldsymbol{d}}^h(x)} \{y_i = y\}. \tag{21.19}$$

In words, the classifier, at a given x, returns the most common category in the neighborhood of x.

Problem 21.13. Detail the arguments leading to (21.19) following those that lead to (21.15).

The expected risk of a classifier $\hat{f}$ at a point $x \in \mathcal{X}$ when fitted to data $\boldsymbol{D}$ is

$$\mathbb{P}(Y \neq \hat{f}_{\boldsymbol{D}}(x) \mid X = x),$$

where the expectation with respect to $Y \mid X = x$ and the data $\boldsymbol{D}$. We saw in Problem 21.8 that this is bounded from below by the risk of the Bayes classifier (21.10), which at x is equal to

$$1 - \max_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x). \tag{21.20}$$

NEAREST-NEIGHBOR CLASSIFIER Local majority vote based on nearest neighbors has some universality property, in the sense that its risk comes close to that of the Bayes classifier under mild assumptions.

Proposition 21.14 (Nearest neighbor classifier)**.** *In the present setting, assume that for all $y \in \mathcal{Y}$, the function $x \mapsto g(y \mid x) := \mathbb{P}(Y = y \mid X = x)$ is continuous on $\mathcal{X}$. Then, as the sample size increases, the limiting expected risk of the nearest neighbor classifier is, pointwise, at most twice the risk of the Bayes classifier.*

Proof sketch. Fix $x \in \mathcal{X}$, and with some abuse of notation let $\hat{f}_n$ denote the nearest neighbor classifier based on a sample of size n. Specifically, $\hat{f}_n(x) = Y_{i_n(x)}$, where $i_n(x) \in \{1, \dots, n\}$ indexes one the the data points closest to x (with ties, if any, broken in a systematic way). Then (21.13) implies that

$$\delta(X_{i_n(x)}, x) \xrightarrow{\mathrm{P}} 0, \quad \text{as } n \to \infty. \tag{21.21}$$

The expected risk of $\hat{f}_n$ at x is

$$\begin{aligned}
&\mathbb{P}(Y \neq Y_{i_n(x)} \mid X = x) \\
&= 1 - \mathbb{P}(Y = Y_{i_n(x)} \mid X = x) \\
&= 1 - \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y, Y_{i_n(x)} = y \mid X = x) \\
&= 1 - \sum_{y \in \mathcal{Y}} \mathbb{P}(Y = y \mid X = x)\, \mathbb{P}(Y_{i_n(x)} = y),
\end{aligned}$$

using the independence of the generic observation (X, Y) and the data (and the fact that $Y_{i_n(x)}$ is a function of the data). We also have

$$\begin{aligned}
\mathbb{P}(Y_{i_n(x)} = y) &= \mathbb{E}\left[\mathbb{P}(Y_{i_n(x)} = y \mid X_{i_n(x)})\right] \\
&= \mathbb{E}\left[g(y \mid X_{i_n(x)})\right] \\
&\to g(y \mid x), \quad \text{as } n \to \infty,
\end{aligned} \tag{21.22}$$

by (21.21) combined with the continuity of $x \mapsto g(y \mid x)$ and dominated convergence (Proposition 8.18). Hence,

$$\mathbb{P}(Y \neq \hat{f}_n(x) \mid X = x) \to 1 - \sum_{y \in \mathcal{Y}} g(y \mid x)^2, \quad \text{as } n \to \infty.$$

We then conclude with Problem 21.16. □

Problem 21.15. Prove the convergence (21.22).

Problem 21.16. For any probability vector (p_j), show that

$$1 - \sum_j p_j^2 \le 2(1 - \max_j p_j).$$

This is enough to complete the proof of Proposition 21.14, but can you sharpen the bound?

21.2.3 Curse of dimensionality

We assume in this section that $\mathcal{X}$ is Euclidean and that the dissimilarity δ derives from a norm on $\mathcal{X}$. In this context, the local methods presented in Section 21.2 are better suited for when $\mathcal{X}$ has small dimension. (In fact, a dimension as low as $\dim(\mathcal{X}) = 5$ is already a stretch in practice.) This is because the space is mostly empty of data points unless the sample size is exponential in the dimension. This phenomenon, in regression, is called the *curse of dimensionality*.

For a concrete example, take $\mathcal{X} = [0,1]^p$, which is a 'nice' compact domain of $\mathbb{R}^p$. Assume furthermore that X has the uniform distribution on $\mathcal{X}$. Fix $h \in (0, 1/2)$. Then, in the setting where the data are as in (21.1), the chances that a Euclidean ball centered at $x \in \mathcal{X}$ of radius h is empty of data points are given by

$$\begin{aligned}
\mathbb{P}\big(\forall i : X_i \notin \mathcal{B}_h(x)\big) &= \mathbb{P}\big(X_1 \notin \mathcal{B}_h(x)\big)^n && (21.23)\\
&= \big[1 - \mathbb{P}\big(X_1 \in \mathcal{B}_h(x)\big)\big]^n && (21.24)\\
&\ge (1 - (2h)^p)^n && (21.25)\\
&\to 1, \text{ when } n(2h)^p \to 0. && (21.26)
\end{aligned}$$

(Note that the inequality is very conservative, but has the benefit of applying to balls based on other norms.) Taking h to be fixed, the condition on n and p holds, for example, when $p \gg \log n$.

We conclude that, when the dimension is a little more than logarithmic in the sample size, the ball neighborhood of any given point is very likely empty of data points, which is of course very problematic for any local method.

Problem 21.17. In the same setting, compute as precisely as you can the minimum sample size n such that the probability that a Euclidean ball of radius h is empty of data points is at most $1/2$. Do this for $p = 1, \dots, 10$. [Calculations may be done using a computer.]

21.3 Empirical risk minimization

The empirical risk is the risk computed on the empirical distribution. It can be used to produce an estimator,

by minimizing it over an appropriately chosen function class. Although such an estimator is typically less 'local', it may nevertheless suffer from the curse of dimensionality (Problem 21.53). This is not the case for estimators based on linear models, an important family of function classes.

21.3.1 Empirical risk

Given data (21.1), the *empirical risk* of a function f is defined as

$$\widehat{\mathcal{R}}_{\boldsymbol{D}}(f) := \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(Y_i, f(X_i)). \tag{21.27}$$

Problem 21.18. Show that the empirical risk is an unbiased and consistent estimate for the risk, in the sense that, for any function $f \in \mathcal{M}$, $\mathbb{E}[\widehat{\mathcal{R}}_{\boldsymbol{D}}(f)] = \mathcal{R}(f)$, and

$$\widehat{\mathcal{R}}_{\boldsymbol{D}}(f) \xrightarrow{\text{P}} \mathcal{R}(f), \quad \text{as } |\boldsymbol{D}| \to \infty. \tag{21.28}$$

21.3.2 Empirical risk minimization

With the empirical risk estimating the risk, it is rather natural to aim at minimizing the empirical risk. This is done over a carefully chosen subclass $\mathcal{F} \subset \mathcal{M}$.

Assuming a minimizer over $\mathcal{F}$ exists, and that some other measurability issues are taken care of, *empirical risk minimization (ERM)* over the class $\mathcal{F}$ amounts to returning a minimizer of the empirical risk over $\mathcal{F}$, namely

$$\hat{f}^{\mathcal{F}}_{\boldsymbol{D}} \in \arg\min_{f\in\mathcal{F}} \widehat{\mathcal{R}}_{\boldsymbol{D}}(f). \tag{21.29}$$

Thus a function class $\mathcal{F} \subset \mathcal{M}$ defines an estimator via ERM minimization, namely $\hat{f}^{\mathcal{F}}$.

We say that ERM is *consistent*[99] for the class $\mathcal{F}$ when

$$\mathcal{R}(\hat{f}^{\mathcal{F}}_{\boldsymbol{D}}) \xrightarrow{\text{P}} \inf_{f\in\mathcal{F}} \mathcal{R}(f), \quad \text{as } |\boldsymbol{D}| \to \infty. \tag{21.30}$$

Importantly, this consistency is not implied by (21.28). Instead, what is needed is a uniform consistency over $\mathcal{F}$.

Problem 21.19. Show that (21.30) holds when

$$\sup_{f\in\mathcal{F}} \left|\widehat{\mathcal{R}}_{\boldsymbol{D}}(f) - \mathcal{R}(f)\right| \xrightarrow{\text{P}} 0, \quad \text{as } |\boldsymbol{D}| \to \infty. \tag{21.31}$$

21.3.3 Interpolation and inconsistency

Consider the case where $\mathcal{X} = \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$. For simplicity, assume that X has a continuous distribution, so that

[99] This notion of consistency is with respect to the underlying distribution. If it holds regardless of the underlying distributions, then ERM is said to be *universally consistent*.

the X_i are distinct with probability one. We work with squared error loss, meaning $\mathcal{L}(y, y') = (y - y')^2$.

Having observed the data, for a function f, the empirical risk (21.27) takes the form

$$\widehat{\mathcal{R}}_{\boldsymbol{d}}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2.$$

Clearly, this risk is non-negative and equal to 0 if and only if $y_i = f(x_i)$ for all $i = 1, \ldots, n$, meaning that the function f *interpolates* the data points. Consequently, if for any sample size n there is a function in $\mathcal{F}$ that interpolates the data points, then $\widehat{\mathcal{R}}_{\boldsymbol{d}}(\hat{f}_{\boldsymbol{d}}^{\mathcal{F}}) = 0$. If, at the same time, $\inf_{f\in\mathcal{F}} \mathcal{R}(f) > 0$, then ERM cannot be consistent.

Problem 21.20. Suppose, without loss of generality, that Y as support $\mathcal{Y}$. Take a loss $\mathcal{L}$ such that $\mathcal{L}(y, y') = 0$ if and only if $y = y'$. Show that $\inf_{f\in\mathcal{F}} \mathcal{R}(f) = 0$ if and only if

$$\mathbb{P}(Y = f(X)) = 1, \quad \text{for some } f \in \mathcal{F}. \tag{21.32}$$

Overfitting We find it desirable to choose a function class for which ERM is consistent, for otherwise it is difficult to know what ERM does. When ERM is not consistent, we say that it *overfits*, and from our discussion above, we know that this happens, for example, when the function class is so 'expressive' that interpolation is possible.

Problem 21.21. In R, generate data according to the model of Figure 21.1. In an effort to perform ERM on the class of all polynomials, interpolate the data points by Lagrange interpolation, which is available via the package polynom. Produce a scatterplot with the fitted polynomial overlaid (as done in that same figure for local linear regression). Repeat for increasing values of n to get a sense of how (wildly) the estimated function behaves.

21.3.4 Linear models

Linear function classes, that is, classes of functions which have the structure of a linear space, have been popular for decades. This is because of their simplicity, their expressive power, and the fact that ERM is relatively easy to compute (or at least approximate). Throughout, we assume that the linear class has finite dimension.

Linear regression Assume that the response is numerical, meaning that $\mathcal{Y} = \mathbb{R}$. Given a set of functions, $f_1, \ldots, f_m : \mathcal{X} \to \mathbb{R}$, we may consider linear combinations, meaning functions of the form

$$f(x) = a_1 f_1(x) + \cdots + a_m f_m(x), \tag{21.33}$$

for arbitrary reals $a_1, \dots, a_m$.

Example 21.22 (Polynomial regression)**.** Polynomials of degree at most k form such a class. In dimension one, meaning when $\mathcal{X} = \mathbb{R}$, we can choose $f_j(x) = x^{j-1}$ for $j = 1, \dots, k+1$.

Proposition 21.23. *Assume that the loss is as in Example 21.3, and that each f_j has finite risk. Then ERM is consistent for the linear class defined by $f_1, \dots, f_m$.*

Problem 21.24. Prove this proposition, possibly under some additional assumptions, based on Problem 16.101.

Problem 21.25 (Least squares)**.** ERM with the square error loss is implemented via the *method of least squares* , defined by the following optimization problem

$$\text{minimize} \sum_{i=1}^{n} \big(y_i - a_1 f_1(x_i) - \cdots - a_m f_m(x_i)\big)^2 \quad \text{over } a_1, \dots, a_m \in \mathbb{R}.$$

Show that this optimization problem can be reduced to solving an $m \times m$ linear system.

Linear classification Assume that the response is binary, so that we may take $\mathcal{Y}$ to be $\{-1, 1\}$ without loss of generality. Given a set of functions, $g_1, \dots, g_m : \mathcal{X} \to \mathbb{R}$, we may consider the class of functions of the form

$$f(x) = \operatorname{sign}(a_1 g_1(x) + \cdots + a_m g_m(x)),$$

for arbitrary reals $a_1, \dots, a_m$.

Proposition 21.26. *Working with the 0-1 loss, ERM is consistent for any such class.*

It turns out that implementing ERM with the 0-1 loss is in general quite difficult. For this reason, a *surrogate loss* is sometimes chosen. This loss is defined not on $\mathcal{Y} = \{-1, 1\}$, but on $\mathbb{R}$. Let $\mathcal{S} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ be such a loss function. In general, a class $\mathcal{G}$ of functions $g : \mathcal{X} \to \mathbb{R}$ defines a class $\mathcal{F}$ of functions $f : \mathcal{X} \to \{-1, 1\}$ of the form $f(x) = \operatorname{sign}(g(x))$, for some $g \in \mathcal{G}$. ERM for such a class $\mathcal{F}$ with the surrogate loss $\mathcal{S}$ proceeds by first minimizing the empirical risk over $\mathcal{G}$, yielding

$$\hat{g}_{\boldsymbol{d}} \in \arg\min_{g \in \mathcal{G}} \sum_{i=1}^{n} \mathcal{S}(y_i, g(x_i)),$$

and then returning $\hat{f}_{\boldsymbol{d}} = \operatorname{sign}(\hat{g}_{\boldsymbol{d}})$.

Examples of surrogate losses used in practice include

$$\begin{aligned} \textit{exponential loss} \quad & \mathcal{S}(y,z) = \exp(-yz), \\ \textit{logistic loss} \quad & \mathcal{S}(y,z) = \log(1+\exp(-yz)), \\ \textit{hinge loss} \quad & \mathcal{S}(y,z) = (1-yz) \vee 0. \end{aligned}$$

These are all convex and lead to a convex optimization problem when applying ERM to a linear class.

Remark 21.27. ERM with the logistic loss is also called *logistic regression*, and ERM with the hinge loss corresponds to *support vector machines*.

Under some conditions, it turns out that ERM with one of these surrogates losses leads to consistency with respect to the 0-1 loss. This is the case, for example, when the class is linear and, importantly, when the minimum risk is achieved over that class (i.e., when the linear class contains a Bayes classifier) [29, Sec 4.2].

21.4 Selection

The local methods presented in Section 21.2 depend on a choice of bandwidth, while empirical risk minimization depends on a choice of function class. In general, when having to choose among various regression estimators, one would want to compare their expected risk (21.5). However, this is not an option, as the expected risk is based on the underlying distribution, which is known. Instead, we substitute the expected risk with an estimate.

Remark 21.28 (Beyond the empirical risk)**.** Even when consistent for the risk (21.30), the empirical risk is typically not useful for comparing various regression estimators. This is because the empirical risk favors expressiveness or richness, and as a consequence leads to choosing an estimate that interpolates the data points when this is possible.

Problem 21.29. Consider kernel regression with a given kernel function, and let $\hat{r}_h$ denote the empirical risk at bandwidth h. Show that $\hat{r}_h \le \hat{r}_{h'}$ whenever $h \le h'$. Assuming that all the x_i are distinct, show that $\hat{r}_h = 0$ (interpolation) when h is small enough. Describe in a similar way what happens for local linear regression.

Problem 21.30. For a function class $\mathcal{F}$, let $\hat{r}_{\mathcal{F}}$ denote the empirical risk of the estimate resulting from ERM over $\mathcal{F}$. Show that $\hat{r}_{\mathcal{F}} \le \hat{r}_{\mathcal{F}'}$ whenever $\mathcal{F} \supset \mathcal{F}'$. Assuming that all the x_i are distinct, show that $\hat{r}_{\mathcal{F}} = 0$ (interpolation) when $\mathcal{F}$ is 'rich' enough.

21.4.1 Data splitting

The main motive for splitting the data is to separate the two operations of fitting and risk assessment, by having them use disjoint parts of the data.

Recall that $\boldsymbol{d}$ denotes the data as in (21.2). Let $\boldsymbol{t} \subset \boldsymbol{d}$ and $\boldsymbol{v} \subset \boldsymbol{d}$ denote the training and validation sets, respectively. These are chosen disjoint, namely $\boldsymbol{t} \cap \boldsymbol{v} = \varnothing$. Let $\mathcal{C}$ denote a set of estimators to be compared. For each estimator in that set, $\hat{f} \in \mathcal{C}$, we do the following:

- *Fitting* Fit the estimator on the training set, obtaining $\hat{f}_{\boldsymbol{t}}$.
- *Assessment* Compute the average loss of the fitted estimator on the validation set, obtaining $\widehat{\mathcal{R}}_{\boldsymbol{v}}(\hat{f}_{\boldsymbol{t}})$.

Having done this, we choose the estimator among those in $\mathcal{C}$ that has the smallest estimated prediction error, obtaining

$$\hat{f}^{\mathcal{C}} := \arg\min_{\hat{f} \in \mathcal{C}} \widehat{\mathcal{R}}_{\boldsymbol{v}}(\hat{f}_{\boldsymbol{t}}). \tag{21.34}$$

We call this estimator the *selected estimator*. (The selection process was based on $\boldsymbol{t}$ and $\boldsymbol{v}$, but we leave this implicit.)

Remark 21.31. The selected estimator is typically fitted on the entire dataset, resulting in $\hat{f}^{\mathcal{C}}_{\boldsymbol{d}}$, which is in turn used for prediction.

Problem 21.32. Show that, for a given estimator $\hat{f}$, we have $\mathbb{E}[\widehat{\mathcal{R}}_{\boldsymbol{V}}(\hat{f}_{\boldsymbol{T}})] = \bar{\mathcal{R}}_m(\hat{f})$ if the training set is of size m. How does this compares with $\bar{\mathcal{R}}_n$ (which is arguably what we would like to estimate)?

Test set The use of a set separate from the training and validation sets becomes necessary if it is of interest to estimate the prediction error of the selected estimator, namely $\hat{f}^{\mathcal{C}}$ of (21.34). This set is called the *test set*. The reason the training and validation sets cannot be used for that purpose is because they were used to arrive at $\hat{f}^{\mathcal{C}}$.

Let $\boldsymbol{s} \subset \boldsymbol{d}$ denote the test set. It is disjoint from the training and validation sets, meaning that $\boldsymbol{s} \cap (\boldsymbol{t} \cup \boldsymbol{v}) = \varnothing$. The risk estimate for the selected estimator is obtained by fitting the selected estimator on the training and validation sets, and then computing the average loss on the test set, obtaining $\widehat{\mathcal{R}}_{\boldsymbol{s}}(\hat{f}^{\mathcal{C}}_{\boldsymbol{t}\cup\boldsymbol{v}})$.

21.4.2 Cross-validation

Data splitting is often seen as being wasteful in the sense that the dataset is subdivided into even smaller subsets (training, validation, and possibly test sets) with each subset playing only one role. The methods we present next mimic data splitting while attempting to make better

use of the available data. These are all variants of *cross-validation (CV)*, arguably the most popular approach for comparing estimators in the context of regression.

k-FOLD CROSS-VALIDATION Let $k \in \{2, \dots, n\}$ and partition the dataset $\boldsymbol{d}$ into k subsets of (roughly) equal size, denoted $\boldsymbol{d}_1, \dots, \boldsymbol{d}_k$. In a nutshell, in the jth round, the jth subset plays the role of validation set while the others together play the role of training set, resulting in a risk estimate. The final risk estimate is the average of these k estimates. See Table 21.1 for an illustration.

In more detail, take one of the estimators to be compared, $\hat{f} \in \mathcal{C}$. We first fit the estimator on $\boldsymbol{t}_l := \boldsymbol{d} \setminus \boldsymbol{d}_l$, obtaining $\hat{f}_{\boldsymbol{t}_l}$. Then we compute the average loss on $\boldsymbol{d}_l$. Finally, we average these k risk estimates, obtaining

$$\frac{1}{k} \sum_{l=1}^{k} \widehat{\mathcal{R}}_{\boldsymbol{d}_l}(\hat{f}_{\boldsymbol{t}_l}). \tag{21.35}$$

Remark 21.33. The choices $k = 5$ and $k = 10$ appear to be among the most popular in practice.

Problem 21.34. Compute the expectation of the risk estimate (21.35).

LEAVE-k-OUT CROSS-VALIDATION Pushing the rationale behind CV to its extreme leads to having each subset of observations of size k play the role of validation set in turn.

Problem 21.35. Write down the leave-k-out CV risk estimate directly in formula, taking care of defining any mathematical symbols that you use.

Because there are $\binom{n}{k}$ subsets of size k out of n observations total, this procedure is computationally prohibitive for almost all choices of k — even with $k = 2$ or $k = 3$ the procedure may be too costly.

Although leave-k-out CV is typically not computationally tractable, Monte Carlo simulation is possible. Indeed, each subset of size k yields a risk estimate and the leave-k-out CV risk estimate is the average of all these risk estimates. The Monte Carlo approach consists in drawing a certain number of subsets of size k at random, computing their associated risk estimates, and returning their average.

The special choice $k = 1$, *leave-one-out cross-validation*, is computationally more feasible, although it still requires fitting the estimator n times, at least in principle. Note that leave-one-out CV is equivalent to n-fold CV, and the corresponding risk estimate is also known as *prediction residual error sum-of-squares (PRESS)*.

For a comprehensive review of cross-validation (both

Table 21.1: 5-fold cross-validation for a given estimator $\hat{f}$. The jth row illustrates the jth round where the jth data block (highlighted) plays the role of validation set. Each round results in a risk estimate and the average of all these risk estimates (5 of them here) is the cross-validation risk estimate for $\hat{f}$.

	Subset 1	Subset 2	Subset 3	Subset 4	Subset 5
Round 1	**Validate**	Train	Train	Train	Train
Round 2	Train	**Validate**	Train	Train	Train
Round 3	Train	Train	**Validate**	Train	Train
Round 4	Train	Train	Train	**Validate**	Train
Round 5	Train	Train	Train	Train	**Validate**

for regression and density estimation), we refer the reader to the survey by Arlot and Celisse [7].

21.5 Further topics

21.5.1 Signal and image denoising

Denoising is an important 'low-level' task in the context of signal and image processing. (Object recognition is an example of 'high-level' task.) It is important in a wide array of contexts including astrophysics, satellite imagery, various forms of microscopy, as well as various types of medical imaging.

Signal or image denoising can be seen as a special case of regression analysis, the main specificity being that X is typically not random; rather, the signal or image is sampled on a regular grid. For example, in dimension 1, this could be $x_i = i/n$ for a signal supported on $[0, 1]$.

In the signal and image denoising literature, kernel regression is known as *linear filtering*.

R corner. In the context of signal or image processing, the function kernel provides access to a number of well-known kernel functions. Having chosen such a kernel, the function kernapply computes the corresponding kernel regression estimate.

21.5.2 Additive models

Additive models are an alternative to linear models. Although nonparametric, they do not suffer from the curse of dimensionality.

We assume throughout that $\mathcal{X} = \mathbb{R}^p$. We say that $f : \mathbb{R}^p \to \mathbb{R}$ is an *additive function* if it is of the form

$$f(x) = \sum_{j=1}^{p} f_j(x_j), \quad \text{for } x = (x_1, \ldots, x_p). \tag{21.36}$$

Additive models for regression Let $\mathcal{F}_o$ be any model class of univariate functions. The corresponding model class of additive functions of p variables is

$$\mathcal{F} := \{f \text{ as in (21.36) with } f_j \in \mathcal{F}_o\}.$$

Such models do not tend to suffer from the curse of dimensionality because, in essence, all happens on the axes.

ERM over an additive class $\mathcal{F}$ can be done via *backfitting*, described in Algorithm 6, which is based on being able to perform ERM over the class of univariate functions $\mathcal{F}_o$ defining $\mathcal{F}$.

Problem 21.36. Define a backfitting procedure based on kernel regression. In R, write a function taking in the data and a bandwidth, and fitting an additive model based on kernel regression with the corresponding bandwidth and the Gaussian kernel.

Algorithm 6 Backfitting Algorithm

Input: data $\boldsymbol{d} = \{(x_i, y_i)\}$ with $x_i = (x_{i1}, \ldots, x_{ip})$, univariate model $\mathcal{F}_o$

Output: fitted additive model

Initialize: $\hat{f}_j \equiv 0$ for all j

Repeat until convergence:

For $j = 1, \ldots, p$

(i) Compute the residuals $r_i \leftarrow y_i - \sum_{k \neq j} \hat{f}_k(x_{ik})$

(ii) Compute the ERM estimate (21.29) for $\mathcal{F}_o$ based on $\{(x_{ij}, r_i)\}$ and update $\hat{f}_j$

Additive models for classification With an additive function class, we can obtain a class of classifiers as described in Section 21.3.4. In particular, when using the logistic loss, this is sometimes called *additive logistic regression*.

21.5.3 Classification based on density estimation

A less direct way of approximating the Bayes classifier (21.10) is via the estimation of the class proportions, $\pi_y := \mathbb{P}(Y = y)$, and the class densities, ϕ_y. Note that π defines the marginal of Y and ϕ_y is the conditional density of $X \,|\, Y = y$. In terms of these quantities, the Bayes classifier can be expressed as

$$f_*(x) \in \arg\max_{y \in \mathcal{Y}} \pi_y \phi_y(x).$$

Problem 21.37. Show this using the Bayes formula.

Thus, if we have estimates, $\hat{\pi}_y$ and $\hat{\phi}_y$ for all $y \in \mathcal{Y}$, we can estimate the Bayes classifier by plugging them in, to obtain

$$\hat{f}(x) \in \arg\max_{y \in \mathcal{Y}} \hat{\pi}_y \hat{\phi}_y(x).$$

The class proportions are typically estimated by the sample class proportions, namely

$$\hat{\pi}_y = \frac{\#\{i : y_i = y\}}{n}.$$

The class densities can be estimated by applying any procedure for density estimation to each sample class.

Problem 21.38. Derive the classifier that results from applying a kernel density estimation procedure (Section 16.10.6) to each sample class. Compare with a local majority vote (with same bandwidth).

Discriminant analysis When the class distributions are modeled as Gaussian and fitted by maximum likelihood, the resulting procedure is called *quadratic discriminant analysis (QDA)*.

Problem 21.39. Show that for QDA the *classification boundaries*, meaning the sets separating the classes, are quadratic surfaces.

If in addition to being modeled as Gaussian, the class distributions are assumed to share the same covariance matrix, then procedure is called *linear discriminant analysis (LDA)*.

Problem 21.40. Show that for LDA the classification boundaries are affine surfaces.

Naive Bayes Density estimation by local methods (e.g., kernel density estimation), also suffers from a curse of dimensionality. For this reason, some structural assumptions are sometimes made.

A *naive Bayes* approach is analogous additive modeling (Section 21.5.2). The corresponding assumption here is that each class density ϕ_y is the product of its marginals, in the sense that

$$\phi_y(x) = \prod_{j=1}^{p} \phi_{j,y}(x_j),$$

for $x = (x_1, \ldots, x_p)$, where if $X = (X_1, \ldots, X_p)$, then $\phi_{j,y}$ is the density of $X_j \,|\, Y = y$. This leads to estimating, for each class, the marginal densities separately and then taking the product to obtain an estimate for the class density.

21.5.4 Isotonic regression

Assume that $\mathcal{X} = \mathbb{R}$ and that $\mathcal{Y} = \mathbb{R}$. Define

$$\mathcal{F} = \big\{ f : \mathbb{R} \to \mathbb{R} \text{ non-decreasing} \big\}. \tag{21.37}$$

ERM based on this class is called *isotonic regression.*[100]

Proposition 21.41. *ERM is consistent for the class* (21.37).

[100] One can as easily work with the class of non-increasing functions, and ERM based on this class is called *antitonic regression.*

The *pooled adjacent violators algorithm (PAVA)* computes the isotonic regression estimate. The paper [55] describes PAVA (and also presents an alternative convex optimization formulation).

R corner. Isotonic regression is available from the package isotone.

Remark 21.42. Notice the parallel with density estimation based on an assumption of monotonicity (Section 16.10.7).

21.6 Additional problems

Problem 21.43. Consider a setting where the response is categorical and $\mathcal{Y}$ is finite of size $q \geq 2$. Explain how a loss function in this context can be represented by a matrix. What matrix corresponds to the 0-1 loss?

Problem 21.44. Consider the case where X and Y are independent. Let $\xi \in \mathcal{Y}$ be such that

$$\xi \in \arg\min_{y' \in \mathcal{Y}} \mathbb{E}\big[\mathcal{L}(Y, y')\big].$$

Show that the constant function $f \equiv \xi$ minimizes the risk (21.3).

Problem 21.45. In the context of Problem 21.6, show that the converse of Problem 21.44 is also true, in the sense that if a constant function minimizes the risk, then X and Y must be independent.

Problem 21.46. Let g be a bounded and continuous function on $\mathbb{R}^p$. Let $\mathcal{B}_h(x_0)$ denote the ball centered at x_0 of radius $h > 0$ in $\mathbb{R}^p$ with respect to some norm. Show that

$$\frac{1}{|\mathcal{B}_h(x_0)|} \int_{\mathcal{B}_h(x_0)} g(x)\mathrm{d}x \to g(x_0), \quad \text{as } h \to 0.$$

Problem 21.47. Assume that X has a density ϕ on $\mathbb{R}^p$ and that (X, Y) have a joint density ψ on $\mathbb{R}^p \times \mathbb{R}$. Show that

$$\mathbb{E}[Y \,|\, X \in \mathcal{B}_h(x_0)] = \frac{\int_{\mathcal{B}_h(x_0)} \int_{\mathbb{R}} y\psi(x, y)\mathrm{d}y\mathrm{d}x}{\int_{\mathcal{B}_h(x_0)} \phi(x)\mathrm{d}x}.$$

Using Problem 21.46 and assuming continuity as needed, show that

$$\mathbb{E}[Y \,|\, X \in \mathcal{B}_h(x_0)] \to \mathbb{E}[Y \,|\, X = x_0], \quad \text{as } h \to 0,$$

thus justifying the approximation (21.14).

Problem 21.48. Bound the mean-squared error of the local average (21.15), adapting the arguments given in Section 16.10.6. Do the same for the Nadaraya–Watson method (21.17) if you can. The analysis should provide some insights on how to choose the bandwidth h (at least in theory).

Problem 21.49. Adapt the discussion of local methods for regression under squared error loss (Section 21.2.1) to the setting where the loss is the absolute loss instead.

Problem 21.50 (Local polynomial regression)**.** In the context of Section 21.2.1, if we assume that the regression function is m times differentiable, it becomes reasonable to locally estimate its Taylor expansion of order m. This results in the so-called *local polynomial regression* estimator of order m. Define this estimator in analogy with the local linear regression estimator.

Problem 21.51. Local polynomial regression of order $m = 0$ amounts to fitting a constant locally. Compare that with kernel regression (with the same kernel function and the same bandwidth).

Problem 21.52. In parallel to Problem 21.48, provide an analysis of local polynomial regression of order m when X is uniform on some interval, say the unit interval, and the regression function is $m + 1$ times continuously differentiable on that interval.

Problem 21.53 (ERM over Lipschitz classes)**.** Consider the case where $\mathcal{X} = [0,1]^p$ and $\mathcal{Y} = \mathbb{R}$. Assume that X has a continuous distribution. For $f : [0,1]^p \to \mathbb{R}$, let

$$|f|_\infty := \sup_x |f(x)|, \qquad L(f) := \sup_{x \neq x'} \frac{|f(x) - f(x')|}{\|x - x'\|}.$$

For $c_0, c_1 > 0$, define

$$\mathcal{F}_{c_0,c_1} := \big\{ f : |f|_\infty \leq c_0, L(f) \leq c_1 \big\}. \tag{21.38}$$

It is known that ERM is consistent for the class $\mathcal{F}_{c_1,c_2}$, for any given constants c_1, c_2.

(i) Show that, unless Y is a deterministic function of X, this is <u>not</u> the case when c_1 is unspecified, meaning that ERM is inconsistent for the class

$$\mathcal{F}_{c_0,*} := \big\{ f : |f|_\infty \leq c_0 \vee L(f) < \infty \big\}.$$

(ii) Argue that, even when c_0, c_1 are given, ERM over the class $\mathcal{F}_{c_1,c_2}$ suffers from the curse of dimensionality.

Problem 21.54. Define a backfitting method based on kernel regression. Implement that method in R.

Problem 21.55. Define an additive model where each component is monotonic (i.e., either non-decreasing or non-increasing). Propose a way to fit such a model, and implement that method in R.

Problem 21.56 (Kernel classification)**.** Local majority vote (21.19) is the analog in classification to the local average (21.15) in regression. Define the analog in classification to the Nadaraya–Watson estimator (21.17) in regression.

CHAPTER 22

FOUNDATIONAL ISSUES

Randomization was presented in Chapter 11 as an essential ingredient in the collection of data, both in survey sampling and in experimental design. We argue here that randomization is the essential foundation of statistical inference: It leads to *conditional inference* in an almost canonical way, and allows for *causal inference*.

22.1 Conditional inference

We already saw in previous chapters a number of situations where inference is performed conditional on some statistics. Invariably, these statistics are not informative. This includes testing for independence as discussed in Chapter 15 and Chapter 19, as well as all other situations where permutation tests are applicable.

22.1.1 Re-randomization tests

Consider an experiment that was designed to compare a number of treatments, and in which randomization (Sec-

tion 11.2.4) was used to assign treatments to experimental units (or said differently, to assign units to treatment groups). The design could be one of the classical designs presented in Section 11.2.5, or any other design that utilizes randomization. Suppose we are interested in testing the null hypothesis that the treatments are equally effective. As we saw in Chapter 17, this can be formalized as a goodness-of-fit testing problem: the null hypothesis is that the joint distribution of the response variables is exchangeable with respect to re-randomization of the treatment group labels.

More formally, suppose there are g treatments and n experimental units (i.e., human subjects in clinical trials), and let Π denote the possible treatment assignments under the randomization scheme employed in the experiment. (Note that Π often depends on characteristics of the experimental units, such as gender or age in clinical trials involving humans.) Let $\boldsymbol{\pi}^0 = (\pi_1^0, \dots, \pi_n^0) \in \Pi$ denote the assignment used in the experiment, where $\pi_i^0 \in \{1, \dots, g\}$ denotes the treatment assigned to unit $i \in \{1, \dots, n\}$. The experiment results in response y_i for unit i. We know that we can organize the data as $(y_1, \pi_1^0), \dots, (y_n, \pi_n^0)$, written henceforth as $(\boldsymbol{y}, \boldsymbol{\pi}^0)$ and seen as a two-column array. For example, in a completely randomized design, Π is the set of all permutations of $\{1, \dots, n\}$. In general, though, Π can be quite complicated.

Let T be a test statistic, with large values weighing against the null hypothesis. For example, T could be the treatment sum-of-squares. In the present context, the *randomization p-value* is defined as

$$\frac{\#\{\pi \in \Pi : T(\boldsymbol{y}, \boldsymbol{\pi}) \geq T(\boldsymbol{y}, \boldsymbol{\pi}^0)\}}{|\Pi|}. \tag{22.1}$$

This is an example of conditional inference, where the conditioning is on the responses.

In most experimental designs, if not all of them, any $\pi \in \Pi$ is a permutation of π^0. If thus seen as a subset of permutations Π forms a subgroup, in the sense that it is stable by composition, the quantity defined in (22.1) is a valid p-value in the sense of (12.22). This is a consequence of Proposition 22.3 below.

As usual, this p-value may be difficult to compute exactly as the number of possible treatment assignments (which varies according to the design) tends to be large. In such situations, one typically resorts to estimating the p-value by Monte Carlo simulation. All that is required is the ability to sample uniformly at random from Π, which is already required in order to perform the initial randomization (meaning the generation of $\boldsymbol{\pi}^0$).

Problem 22.1. Verify that the re-randomization p-value corresponds to the permutation p-value in the settings previously encountered.

Remark 22.2. Re-randomization is quite natural, as the randomness is present by design and exploiting that randomness is a rather safe approach to inference. This strategy is quite old and already mentioned in the pioneering works of Ronald Fisher and Edwin Pitman in the 1930's. However, at the time the Monte Carlo approach outlined above was impractical as there were no computers and a normal approximation was used instead. Over the years, this normal approximation became canon and is, to this day, better known than the re-randomization approach. (This normal approximation is at the foundation of the Student test, for example.)

22.1.2 Randomization p-value

Consider a general statistical model $(\Omega, \Sigma, \mathcal{P})$, where $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$. When needed, we will let $\theta_* \in \Theta$ denote the true value of the parameter. Our goal is to test a null hypothesis $\mathcal{H}_0 : \theta_* \in \Theta_0$, for some given $\Theta_0 \subset \Theta$.

Let Π denote a finite set of one-to-one [101] transformations $\pi : \Omega \to \Omega$ such that

$$\mathbb{P}_\theta(\pi^{-1}(\mathcal{A})) = \mathbb{P}_\theta(\mathcal{A}), \quad \text{for all } \mathcal{A} \in \Sigma, \quad \text{for all } \theta \in \Theta.$$

Crucially, we work under the assumption that Π forms a *group*, meaning that $\mathrm{id} \in \Pi$; that if $\pi \in \Pi$ then also $\pi^{-1} \in \Pi$; and that if $\pi_1, \pi_2 \in \Pi$ then $\pi_1 \circ \pi_2 \in \Pi$. (id is the identity transformation $\mathrm{id}: \omega \mapsto \omega$.)

Under these circumstances, Π can be used to obtain a p-value for a given test statistic T. Assuming that large values of T are evidence against the null, the following is the *randomization p-value* for T with respect to the action of Π:

$$\mathsf{pv}(\omega) := \frac{\#\{\pi \in \Pi : T(\pi(\omega)) \geq T(\omega)\}}{|\Pi|}, \tag{22.2}$$

where ω represents the observed data. This is another instance of conditional inference where the conditioning is on $\{\pi(\omega) : \pi \in \Pi\}$ (called the *orbit* of ω under the action of Π).

Proposition 22.3. *In the present context, the quantity defined in* (22.2) *is a valid p-value in the sense of* (12.22).

This proposition implies that (22.1) is a valid p-value when Π corresponds to a subgroup of permutations. In

[101] We also require that any $\pi \in \Pi$ be bi-measurable, meaning that both π and π^{-1} are measurable.

particular, this applies to the permutation goodness-of-fit tests, as well as the permutation tests for independence and the tests for symmetry seen in previous chapters.

In order to prove Proposition 22.3, we use the following result, which implies that any finite group is isomorphic to a group of permutations.

Theorem 22.4 (Cayley)**.** *Suppose that* Π *is a finite group with distinct elements denoted* $\pi_1, \dots, \pi_l$*. For each* $\pi \in \Pi$*,* $\{\pi_1 \circ \pi, \dots, \pi_l \circ \pi\}$ *is a reordering of* $\{\pi_1, \dots, \pi_l\}$*. Let* σ_π *denote the corresponding permutation of* $\{1, \dots, l\}$*. Then* $\mathcal{S}_\Pi := \{\sigma_\pi : \pi \in \Pi\}$ *is a group of permutations. Moreover,* $\pi \mapsto \sigma_\pi$ *is an isomorphism from* Π *to* $\mathcal{S}_\Pi$*.*

Proof of Proposition 22.3. We use the notation of Theorem 22.4. Let $\mathcal{S}$ be short for $\mathcal{S}_\Pi$ and, for each j, let σ_j be short for σ_{π_j}.

Assume without loss of generality that $\pi_1 = \mathrm{id}$. Define $T_j = T \circ \pi_j$ and $\boldsymbol{T} = (T_1, \dots, T_l)$. Note that (22.2) can be written $\#\{j : T_j \geq T_1\}/l$.

We consider a null distribution, left implicit. For any j, we have $\boldsymbol{T}_{\sigma_j} = \boldsymbol{T} \circ \pi_j$, which has same distribution as $\boldsymbol{T}$, by assumption. Moreover, for any j and k distinct, let m be such that $\pi_m = \pi_j^{-1} \circ \pi_k$. Since $\pi_j \circ \pi_m = \pi_k$, by definition of σ_m in Theorem 22.4, we have $\sigma_m(j) = k$. We then conclude using Problem 8.64. □

Remark 22.5 (Balanced permutations)**.** The group structure is needed besides being used in the proof above. To illustrate that, consider the case of two treatments being compared in a completely randomized design. Assume the group sizes are the same (which is often desirable). Suppose we want to calibrate a test statistic by permutation. Based on power considerations, it is rather tempting to consider permutations that move half of each group over to the other group. These are called *balanced permutations* in [219]. However, one should resist the temptation because, although power is indeed improved, the level is not guarantied to be controlled, as shown in [219].

Problem 22.6. Show that the set of balanced permutations of $\{1, \dots, 2k\}$ is not a group unless $k = 1$.

MONTE CARLO ESTIMATION In many instances, the randomization p-value (22.2) cannot be computed exactly because the group of transformations Π is too large. In that case, one can resort to an estimation by Monte Carlo simulation, which as usual requires the ability to sample from the uniform distribution over Π. In detail, in the same setting as before, we sample $\pi_1, \dots, \pi_B$ iid from the uniform distribution on Π and return

$$\widehat{\mathsf{pv}}(\omega) := \frac{\#\{b : T(\pi_b(\omega)) \geq T(\omega)\} + 1}{B + 1}.$$

Proposition 22.7. *In the context of Proposition 22.3, this Monte Carlo p-value is a valid p-value in the sense of* (12.22).

Problem 22.8. Prove this proposition by following the proof of Proposition 22.3.

22.1.3 Goodness-of-fit testing

Besides re-randomization testing, conditional inference may also be used for goodness-of-fit testing. Suppose our goal is to test for a null model of distributions. A general approach consists in conditioning on a statistic that is sufficient for that model, and then examining the remaining randomness.

In detail, suppose that we have data $\omega \in \Omega$ and want to test whether ω was generated from a distribution in a given family of distribution $\{\mathbb{P}_\theta : \theta \in \Theta\}$. Each $\mathbb{P}_\theta$ is a distribution on (Ω, Σ), where Σ is some σ-algebra on Ω. Based on particular alternatives we have in mind, we decide to reject for large values of some statistic S. But how can we obtain a p-value for S?

Suppose that T is sufficient statistic for this family. Let $t = T(\omega)$ denote the observed value of that statistic. If the null hypothesis is true, meaning if ω was generated from a distribution in that family, then the conditional distribution of S given $T = t$ is independent of $\theta \in \Theta$, and is therefore known (at least in principle). Letting $s = S(\omega)$ denote the observed value of the test statistic, we may thus define a conditional p-value for S as follows

$$\mathsf{pv}(s \,|\, t) := \mathbb{P}_{\theta_0}(S \geq s \,|\, T = t),$$

where θ_0 is an arbitrary element of Θ.

For example, a p-value for a test of randomness (Section 15.8) is, in the discrete setting, obtained by conditioning on the counts, and the resulting distribution is simply the permutation distribution.

Problem 22.9. Consider an experiment yielding a sample of size n, $X_1, \ldots, X_n$. We want to test whether the sample was generated iid from the uniform distribution on $[0, \theta]$ for some (unknown) $\theta > 0$. We choose to reject for large values of S, the largest spacing defined as $S = \max_i(X_{(i+1)} - X_{(i)})$, where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics. Describe how you would obtain, via Monte Carlo simulations, a p-value for that statistic based on the approach described above. (A more direct approach, not based on computer simulations, is proposed in [249].)

We provide further examples below. The last three are quite similar, even though they are motivated by completely different applications.

22.1.4 Genetic selection

Consider n individuals taken from a population where a particular gene is undergoing neutral selection. Let the number of alleles (variants of that gene) that appear j times be denoted M_j. Under some simplifying assumptions, and assuming a mutation rate $\theta > 0$, Warren Ewens in [80] derived

$$\mathbb{P}_\theta(M_1 = m_1, \dots, M_n = m_n) = \frac{n!}{\theta(\theta+1)\cdots(\theta+n-1)} \prod_{j=1}^{n} \frac{\theta^{m_j}}{j^{m_j} m_j!},$$

under the constraint that $m_1, \dots, m_n$ are non-negative integers such that $\sum_{j=1}^n j m_j = n$.

Problem 22.10. Show that, when $\theta = 0$, all alleles represented in the sample are the same with probability one. Show that, when $\theta \to \infty$, they are all distinct with probability tending to 1.

Problem 22.11. Show that the number of distinct alleles represented in the sample, namely $K := M_1 + \cdots + M_n$, is sufficient for this model.

Watterson [248] suggests to test for neutral selection based on

$$S(m_1, \dots, m_n) := m_1^2 + \cdots + m_n^2.$$

A test based on S can be one-sided or two-sided.

Problem 22.12. Show that S is maximum when all alleles represented in the sample are distinct, and minimum when they are all the same.

If k denotes the observed number of distinct alleles in the sample, a p-value is then obtained based on the distribution of S given $K = k$.

For more on Ewens' formula, we refer the reader to [49].

22.1.5 Rasch model

The *Rasch model* [192] was proposed in the context of Psychometry. A prototypical situation that arises in educational research is that of a exam which consists of multiple questions, taken by a number of individuals. If Individual i answers Question j correctly, set $x_{ij} = 1$, otherwise set $x_{ij} = 0$. Assuming there are m individuals and n questions, the data are organized in the m-by-n data matrix $\boldsymbol{x} := (x_{ij})$. The Rasch model presumes that, as random variables (meaning before the test is taken), the X_{ij} are independent with

$$\mathbb{P}(X_{ij} = 1) = \frac{\exp(a_i - b_j)}{1 + \exp(a_i - b_j)}, \tag{22.3}$$

where a_i is the ability of Individual i and b_j is the difficulty of Question j. These are the parameters of the model. Let $\boldsymbol{X} = (X_{ij})$ denote the (random) data matrix.

Problem 22.13. Show that this model is not identifiable. Then show that fixing the average subject ability $\frac{1}{m}\sum_{i=1}^{m} a_i$ or the average problem difficulty $\frac{1}{n}\sum_{j=1}^{n} b_j$ makes the model identifiable. (Identifiability is not of concern in what follows.)

Let $R_i = \sum_{j=1}^{n} X_{ij}$ denote the row sum for Individual i (which corresponds to the number of questions that individual answered correctly) and let $C_j = \sum_{i=1}^{m} X_{ij}$ denote the column sum for Question j (which corresponds to the number of individuals that answered that question correctly). Set $\boldsymbol{R} = (R_1, \ldots, R_m)$ and $\boldsymbol{C} = (C_1, \ldots, C_n)$.

Proposition 22.14. *The row and column sums are jointly sufficient for the individual ability and question difficulty parameters and, conditioning on these, the data matrix is uniformly distributed in the set of binary matrices with these row and column sums.*

Problem 22.15. Prove this proposition.

Suppose that we simply want to know whether the data are compatible with such a model, which we formalize as testing the null hypothesis that the data matrix was generated from an (unspecified) Rasch model. Based on the alternatives we have in mind, we choose to work with a test statistic, denoted T, whose large values provide evidence against the null hypothesis. Having observed the data matrix, $\boldsymbol{x} = (x_{ij})$, we are left with the problem of obtaining a p-value for $T(\boldsymbol{x})$.

Inspired by Proposition 22.14, we fix the margins. Let $\boldsymbol{r} = (r_1, \ldots, r_m)$ and $\boldsymbol{c} = (c_1, \ldots, c_n)$ denote the vectors of observed row and column sums. If the null hypothesis is true, then conditional on $\boldsymbol{R} = \boldsymbol{r}$ and $\boldsymbol{C} = \boldsymbol{c}$, the data matrix $\boldsymbol{X}$ is uniformly distributed in the set, denoted $\mathcal{X}(\boldsymbol{r}, \boldsymbol{c})$, of binary matrices with row and column sums given by $\boldsymbol{r}$ and $\boldsymbol{c}$. The p-value conditional on the row and column sums is consequently obtained as follows

$$\mathsf{pv}(\boldsymbol{x}) := \frac{\#\{\boldsymbol{x}' \in \mathcal{X}(\boldsymbol{r}, \boldsymbol{c}) : T(\boldsymbol{x}') \geq T(\boldsymbol{x})\}}{|\mathcal{X}(\boldsymbol{r}, \boldsymbol{c})|}. \tag{22.4}$$

Problem 22.16. Show that (22.4) is a valid p-value in the sense of (12.22).

The p-value (22.4) is hard to compute in general due to the fact that the set $\mathcal{X}(\boldsymbol{r}, \boldsymbol{c})$ can be very large and even difficult to enumerate. In fact, the mere computation of the cardinality of $\mathcal{X}(\boldsymbol{r}, \boldsymbol{c})$ is challenging [59].

22.1.6 Species co-occurrence

In Ecology, co-occurrence analysis refers to the study of how different species populate some geographical sites. [102] Various species (of interest) are observed, or not, in some geographical sites (of interest). These findings are stored in a so-called *presence-absence* matrix where rows represent species and columns represent sites, and the (i, j) entry is 1 or 0 according to whether Species i is found in Site j or not. We adopt the notation of Section 22.1.5. This time $x_{ij} = 1$ if Species i is present in Site j, and $x_{ij} = 0$ otherwise, and the row sum r_i corresponds to the total number of sites that Species i inhabits, while the column sum c_j corresponds to the total number of species that inhabit Site j.

A longstanding controversy and source of conflict in the Ecology community has surrounded the analysis and interpretation of such data. In the 1970's, Diamond [61] collected presence-absence data for various species of birds in the Bismarck Archipelago (where each island was considered a site). Based on these data, he formulated a number of 'assembly rules' having to do with competition for resources (e.g., food, shelter, breeding grounds, etc) and implying that some pairs of species would not inhabit the same site. However, Connor and Simberloff [44] questioned the basis upon which these rules where formulated. They claimed that the presence-absence patterns that Diamond attributed to his assembly rules could in fact be attributed to 'chance'. [103] An important part of the resulting (ongoing?) controversy has to do with how to interpret 'chance', meaning, what probability model to use for statistical inference.

A simple version of the original null model of Connor and Simberloff [44] amounts to the uniform distribution after conditioning on the margins. This is exactly the model we discussed in Section 22.1.5, and a test statistic of interest (perhaps chosen to test the validity of some assembly rule) is calibrated based on this model. Note that here other null models are possible and, in fact, the relevance of this model in the present situation is part of the controversy.

[102] Related concepts of co-occurrence exist in other areas such as in Linguistics and the analysis of textures in Image Processing.

[103] Part of their criticism involved questions of how the data were handled and analyzed. In essence, they claimed that Diamond had simply selected some pairs of species to support his theory. We will not elaborate on these rules or the controversy surrounding them as these are domain specific. We refer the curious reader to [111, Ch 7] for further details.

22.1.7 Ising model

Recall the Ising model described in Example 10.21.

Problem 22.17. Show that $\xi(\boldsymbol{X})$ and $\zeta(\boldsymbol{X})$ are jointly sufficient and, further, that conditional on $\xi(\boldsymbol{X}) = s$ and $\zeta(\boldsymbol{X}) = t$, $\boldsymbol{X}$ is uniformly distributed among m-by-n spin matrices satisfying these constraints.

As before, suppose that we simply want to test the null hypothesis that the data matrix was generated from an (unspecified) Ising model, and that we choose a test statistic T whose large values provide evidence against this hypothesis.

Problem 22.18. Based on Problem 22.17 and Section 22.1.5, propose a way to obtain a p-value for T. As before, implementing the method will involve serious computational challenges; specify these challenges.

22.1.8 MCMC p-value

Besag and Clifford [20] propose a Markov chain Monte Carlo (MCMC) approach to generate samples from the null distribution in the context of testing for the Rasch model — which we saw coincides with the null model of Connor and Simberloff used in the species co-occurrence problem — and in the context of testing for the Ising model. They then build on that to propose a way to obtain a valid p-value for a given test statistic.

Rasch model Recall that we want to sample from the uniform distribution on the set of binary matrices with row sums $r_1, \ldots, r_m$ and column sums $c_1, \ldots, c_n$. This model was already considered in Section 10.4.1, and we saw there how to design an Markov chain (on the set of such matrices) with stationary distribution the uniform distribution.

Remark 22.19. The Ecology community struggled for years to design a method for sampling from Connor and Simberloff's null model. The original algorithm of Connor and Simberloff [44] was quite ad hoc and inaccurate. Manly [161] used the work of Besag and Clifford [20] but mistakenly forced the chain to move at every step (Problem 10.17), a flaw that was left unnoticed for some years and upon which others built [110]. The error was apparently only discovered almost ten years later [168]. Some other efforts to sample from this null model are reviewed in [258], which goes on to propose a weighted average approach based on the ergodic theorem (Theorem 10.18).

Ising model Recall that we want to sample from the uniform distribution on the set of spin matrices with given

ξ and ζ. For this model, consider the following Markov chain. Given such a matrix, choose a pair of distinct sites (i_1, j_1) and (i_2, j_2) uniformly at random and switch their spins if it preserves ζ; otherwise, stay put. (Such a switch always preserves ξ.)

Problem 22.20. Show that this chain has stationary distribution the uniform distribution. [See Problem 10.16.]

Obtaining a p-value Suppose the data are denoted $\boldsymbol{X}$ as above, which under the null hypothesis of interest is uniformly distributed on a finite set denoted $\mathcal{X}$. Having observed $\boldsymbol{X} = \boldsymbol{x}$, the p-value corresponding to a test statistic T is defined as

$$\mathsf{pv}(\boldsymbol{x}) := \frac{\#\{\boldsymbol{x}' \in \mathcal{X} : T(\boldsymbol{x}') \geq T(\boldsymbol{x})\}}{|\mathcal{X}|}. \tag{22.5}$$

Assume that a Markov chain on $\mathcal{X}$ is available with stationary distribution the uniform distribution. The ergodic theorem could be invoked to justify running a chain from any $\boldsymbol{x}_1 \in \mathcal{X}$, obtaining $\boldsymbol{x}_2, \dots, \boldsymbol{x}_B$, and then estimating the p-value using

$$\frac{\#b : T(\boldsymbol{x}_b) \geq T(\boldsymbol{x})\}}{M}. \tag{22.6}$$

Problem 22.21. Show that this is indeed a consistent estimator of (22.5) for any choice of $\boldsymbol{x}_1 \in \mathcal{X}$, where consistency is as $B \to \infty$.

Alternatively, we can obtain a valid p-value as follows. Let K be uniformly distributed in $\{1, \dots, B\}$ and independent of the data. Assuming $K = k$, do the following: [104]

- if $k = 1$, let $\boldsymbol{x}_1 = \boldsymbol{x}$ and run the chain $B - 1$ steps (forward) from $\boldsymbol{x}_1$, obtaining $\boldsymbol{x}_2, \dots, \boldsymbol{x}_B$;
- if $k = B$, let $\boldsymbol{x}_B = \boldsymbol{x}$ and run the chain $B - 1$ steps (backward) from $\boldsymbol{x}_B$, obtaining $\boldsymbol{x}_{B-1}, \dots, \boldsymbol{x}_1$;
- if $1 < k < B$, let $\boldsymbol{x}_k = \boldsymbol{x}$, run the chain $B - k$ steps (forward) from $\boldsymbol{x}_k$, obtaining $\boldsymbol{x}_{k+1}, \dots, \boldsymbol{x}_B$, and run the chain $k - 1$ steps (backward) from $\boldsymbol{x}_k$, obtaining $\boldsymbol{x}_{k-1}, \dots, \boldsymbol{x}_1$.

Having done this, estimate the p-value as in (22.6).

Proposition 22.22. *The resulting p-value is a valid p-value in the sense of* (12.22).

Problem 22.23. Prove this proposition as follows. Show that, under the null hypothesis (where $\boldsymbol{X}$ has the uniform

[104] We saw how to run a chain backward in Section 9.2.5. Note that the two chains described in this section are reversible so that running the chain backward or forward is equivalent.

distribution on $\mathcal{X}$), and independently of K, the resulting random variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_B$ are distributed as if the first state were drawn from the uniform distribution and the chain were run from there for $B-1$ steps.

Problem 22.24. Prove the proposition using the conclusions of Problem 8.64.

Problem 22.25. The method for obtaining a valid p-value described here is the 'serial method' introduced in [20]. Read enough of this paper to understand the other method, called the 'parallel method', and prove the analog of Proposition 22.22 for that method.

22.2 Causal inference

> *I have no wish, nor the skill, to embark upon philosophical discussion of the meaning of 'causation'.*
>
> Sir A. Bradford Hill [125]

The concept of *causality* has been, and continues to be, a contentious area in Philosophy. Yet, at a very practical level, establishing cause-and-effect relationships is central to, and in many cases the ultimate goal of, most sciences. For example, in the context of Epidemiology, according to [106], "causal inference is implicitly and sometimes explicitly embedded in public health practice and policy formulation".

22.2.1 Association vs causation

It is widely accepted that properly designed experiments (that invariably use some form of randomization) can allow for causal inference. We elaborate on that in Section 22.2.2. In contrast, drawing causal inferences from observational studies is not possible in general, unless one is able to convincingly argue that there are no unmeasured confounders. We study this situation in Section 22.2.3, where we examine how matching attempts to mimic what randomization does automatically.

In general, though, observational studies can only lead to inferences about *association*. Indeed, take Example 15.37 on graduate admissions at UC Berkeley in the 1970's. Surely, there is a clear association between gender and overall admission rate. (The statistical significance is overwhelming.) However, inferring causation (gender bias) would be misleading.

In conclusion, we warn the reader that drawing causal inferences from observational studies is fraught with pitfalls and remains controversial [6, 9, 95, 98]. We adopt

this prudent stance in our discussion of causation, which can be encapsulated in the following.

Proverb. *Association is not causation.*

22.2.2 Randomization

> *It may be said that the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged.*
>
> Ronald A. Fisher [86]

We describe a simple model for causal inference called the *counterfactual model*, attributed to Neyman [105] [220] and Rubin [198]. Within this model, randomization allows for causal inference.

We describe a simple setting where two treatments are compared based on n subjects sampled uniformly at random from a large population. Each subject receives only one of the treatments. Let R_{ij} denote the response of Subject i to Treatment j. The sampling justifies our working with the assumption that the R_{i1} are iid (with distribution that of R_1) and, similarly, that the R_{i2} are iid (with distribution that of R_2). The rub is that the experiment results in observing a realization of either R_{i1} or R_{i2}, but not both since Subject i receives only one of the two treatments.

[105] Neyman's original paper dates back to 1923 and was written in Polish.

Comparing the means We discuss this model in the context of comparing the mean response to the two treatments, called the *average causal effect* and defined as

$$\theta := \mathbb{E}[R_2] - \mathbb{E}[R_1].$$

We interpret $\theta \neq 0$ as a causal effect: the change in treatment causes a change in average response in the entire population. Our immediate interest is to learn about θ from the experiment.

Let $X_i = j$ if Subject i receives Treatment j and let Y_i denote the response observed on Subject i, so that $Y_i = R_{ij}$ if $X_i = j$. We observe a realization of $(X_1, Y_1), \ldots, (X_n, Y_n)$. The *association* between treatment and response is defined as

$$\lambda := \mathbb{E}[Y \,|\, X = 2] - \mathbb{E}[Y \,|\, X = 1]. \tag{22.7}$$

There is a natural estimator for λ, namely the difference in sample means

$$D := \bar{Y}_2 - \bar{Y}_1, \tag{22.8}$$

where $\bar{Y}_j$ is the average of $\{Y_i : X_i = j\}$.

Problem 22.26. Assume that the X_i are iid with distribution X, where $\mathbb{P}(X = 1) > 0$ and $\mathbb{P}(X = 2) > 0$. Show that D is a consistent estimator for λ in that case.

In causal inference, however, our target is θ and not λ. But the two coincide when treatment assignment (namely X) is independent of the response to treatment (namely R_1 and R_2). This comes from writing

$$\lambda = \mathbb{E}[R_2 \,|\, X = 2] - \mathbb{E}[R_1 \,|\, X = 1],$$

and corresponds to an ideal situation where there is no confounding between assignment to treatment and response to treatment.

Problem 22.27. Prove that a completely randomized block design fulfills this condition and thus allows for causal inference. In this case, show that D is unbiased for θ and compute its variance.

Problem 22.28. Show that, in general, one cannot infer causation from association, by providing an example where D is bound to be a terrible estimate for θ.

22.2.3 Matching

Continuing with the same notation, assume now that another variable is available, denoted Z, and may be a confounder.

Problem 22.29. Argue that randomization allows us to effectively ignore Z in the sense that what is said in Section 22.2.2 remains applicable.

Here we want to examine whether we can do away with randomization, and in particular if matching allows us to do that. We use matching on Z with the intent of removing any confounding it might induce. To simplify the discussion, we interpret matching as simply conditioning on Z in addition to conditioning on X. (See Remark 22.33 below.)

The punchline is that matching works as intended if the dependency of Y on (X, Z) is properly modeled and there are no other (unmeasured) confounding variables at play. Both conditions are highly nontrivial. To avoid modeling issues, we assume that Z has a finite support, denoted $\mathcal{Z}$ below.

Our access to Z allows us to consider a refinement of λ above, namely

$$\lambda(z) := \mathbb{E}[Y \,|\, X = 2, Z = z] - \mathbb{E}[Y \,|\, X = 1, Z = z].$$

Note that λ in (22.7) is the expectation of $\lambda(Z)$.

Problem 22.30. Provide a consistent estimator for $\lambda(z)$ when the (X_i, Z_i) are iid from a distribution where $\mathbb{P}(X = j, Z = z) > 0$ for any $j \in \{1, 2\}$ and any $z \in \mathcal{Z}$.

In order to be able to estimate θ, we require that, conditional on Z, R_1 and R_2 be independent of X .

Problem 22.31. Under this assumption, show that

$$\theta = \sum_{z \in \mathcal{Z}} \lambda(z)\, \mathbb{P}(Z = z).$$

Problem 22.32. For any $z \in \mathcal{Z}$, propose a consistent estimator for $\mathbb{P}(Z = z)$.

Remark 22.33. Estimating $\lambda(z)$ accurately for each $z \in \mathcal{Z}$ requires, at the very least, a sample of size proportional to that of $\mathcal{Z}$. When $\mathcal{Z}$ has small cardinality, $\lambda(z)$ is estimated based on $\{(X_i, Z_i) : Z_i = z\}$. When $\mathcal{Z}$ is large, or even infinite, this simple approach may not be feasible. In such situations, one can simply stratify Z, which amounts to binning the Z_i into a few bins, in essence reducing the situation to the case where $\mathcal{Z}$ is of small cardinality. Another approach consists in modeling Y as a function of (X, Z). The validity of the causal inference in that case rests squarely on whether the assumed model is accurate, which may be hard to verify in practice [96, 150].

CHAPTER 23

SPECIALIZED TOPICS

The present chapter presents a collection of more specialized topics that offer a wide variety of more sophisticated examples of statistical inference, beyond the classical settings presented earlier in the book, as well as connections to related topics (Information Theory and Randomized Algorithms). The chapter also includes some pitfalls of scientific publishing, and how the practice of statistics is shaped in response to them.

23.1 Inference for discrete populations

We saw a number of sampling plans in Section 11.1. The prototypical situation is that of a large, possibly infinite (discrete) population where each element is marked. Without loss of generality, we index the population by the positive integers, and let x_k denote the mark on the element k. Here we assume that the mark is a numerical value, and take as our goal the estimation of the popula-

tion total, which we assume is finite:

$$\lambda = \sum_{k=1}^{\infty} x_k. \tag{23.1}$$

(If it turns out that the population is finite, say indexed by $\{1, \dots, N\}$, then we simply set $x_k = 0$ for $k > N$.)

23.1.1 Sampling with replacement: the Hansen–Hurwitz estimator

Assume that n elements are sampled *with replacement* from the population, with the element k having probability p_k of being selected with each draw. Here $\mathsf{p} := (p_k)$ is a probability vector that depends on the sampling plan. We assume it is known and satisfies $p_k > 0$ for all k.

Problem 23.1. Argue that unless $p_k > 0$ for all k, it is not possible to estimate λ. The problem is indeed ill-posed in this case.

If $(K_1, \dots, K_n)$ denotes the resulting sample of elements, the K_i are iid with distribution p. Note that $(x_{K_1}, \dots, x_{K_n})$ is the corresponding sample of values.

Problem 23.2. Show that the sample sum, $\sum_{i=1}^{n} x_{K_i}$, is not, in general, a 'suitable' estimator for λ.

As a (viable) alternative, the *Hansen–Hurwitz estimator* for λ is given by

$$L_n := \frac{1}{n} \sum_{i=1}^{n} \frac{x_{K_i}}{p_{K_i}}.$$

Problem 23.3. Show that this estimator is unbiased for λ and that it has variance

$$\mathrm{Var}(L_n) = \tau^2/n, \quad \tau^2 := \sum_{k=1}^{\infty} p_k \Big(\frac{x_k}{p_k} - \lambda \Big)^2.$$

Below we use this estimator to build various confidence intervals for λ. We follow Section 14.1.3 very closely.

Problem 23.4. Use Chebyshev's inequality to obtain a confidence interval for λ. [This requires bounding τ from above.]

This interval is often deemed too conservative and it is common (but not necessarily good) practice to rely on a normal approximation to refine it.

Problem 23.5. Show that

$$\frac{L_n - \lambda}{\tau/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1), \quad \text{as } n \to \infty.$$

This result is not enough, in itself, to yield a confidence interval, simply because τ is unknown. A reasonable estimate is the following.

Problem 23.6. Show that

$$T_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}\Big(\frac{x_{K_i}}{p_{K_i}} - L_n\Big)^2$$

is unbiased and consistent for τ^2.

Combining the last two results, we obtain the following approximate $(1-\alpha)$-confidence interval

$$\Big[L_n \pm z_{\alpha/2}\, T_n/\sqrt{n}\Big],$$

where z_α is the standard normal α-quantile. The confidence level is exact in the large-sample limit.

23.1.2 Horvitz–Thompson estimator

Assume now that m elements are sampled *without replacement* from the population.[106] Let q_k denote the probability that the element k is in the sample. If $(K_1, \dots, K_m)$ denotes the resulting sample, we assume the K_j are exchangeable.

The *Horvitz–Thompson estimator* for λ is defined as

$$L_m := \sum_{i=1}^{m}\frac{x_{K_i}}{q_{K_i}}.$$

[106] Otherwise, we simply ignore multiplicity.

Problem 23.7. Show that this estimator is unbiased for λ and that it has variance

$$\operatorname{Var}(L_m) = \sum_{k=1}^{\infty}\Big(\frac{1-q_k}{q_k}\Big)x_k^2 + \sum_{k=1}^{\infty}\sum_{l\neq k}\Big(\frac{q_{kl}-q_kq_l}{q_kq_l}\Big)x_kx_l,$$

where q_{kl} is the probability that both k and l are in the sample.

Problem 23.8. Assuming that $q_{kl} > 0$ for all $k \neq l$, show that

$$\sum_{i=1}^{m}\Big(\frac{1-q_i}{q_i}\Big)x_{K_i}^2 + \sum_{i=1}^{m}\sum_{j\neq i}\Big(\frac{q_{ij}-q_iq_j}{q_iq_j}\Big)\frac{x_{K_i}x_{K_j}}{q_{ij}}$$

is an unbiased estimator for the variance of L_m.

The Horvitz–Thompson estimator is applicable as long as one is able to compute $\mathsf{q} := (q_k)$, which may be challenging for some sampling plans. (The computation of the q_{kl} are also needed if one is to use the variance estimator above.)

Problem 23.9. Consider the setting of Section 23.1.1. Let M denote the number of distinct elements in the sample. Relate M to the Coupon Collector Problem of Section 3.8. Then express q as a function of n and p.

Problem 23.10. In the setting of Section 23.1.1, compare the Hansen–Hurwitz estimator and the Horvitz–Thompson estimator. Do so when the population is finite

and p is the uniform distribution. [Since these two estimators are both unbiased, the comparison can be in terms of variance.]

23.1.3 Example: line-intercept sampling

Line-intercept sampling is a sampling method used in spatial settings. Consider a setting where a finite population, located in a region $\mathcal{D} \subset \mathbb{R}^2$, is to be surveyed. Each specimen of the population, say i, occupies a certain amount of space, say $\mathcal{S}_i \subset \mathcal{D}$. (These may overlap.) Also, a specimen i comes with a feature of interest or mark, denoted x_i.

For example, in Forestry, a population of a certain species of trees might be of interest. Each tree occupies a certain amount of space, say, measured in terms of the projection of its canopy onto the forest floor (as if seen from an airplane).

A baseline consisting of a line segment is chosen and a point on the baseline is picked uniformly at random. The line passing through that point and perpendicular to the baseline, called transect and denoted L, is followed through the region $\mathcal{D}$, and each specimen i such that $\mathcal{S}_i \cap L \neq \varnothing$ is included in the sample, which effectively means that x_i is recorded. If x_i is numerical as before, and if q_i denotes the probability that the specimen i is sampled, the Horvitz–Thompson estimate is $\sum_{i \in \mathcal{I}} x_i/q_i$, where $\mathcal{I}$ indexes the sampled specimens.

Problem 23.11. Let $\mathcal{D} = [0,1]^2$ and consider a disc inside $\mathcal{D}$ with radius r and center z. Let the baseline be a line segment of the form $[A, A+h] \times \{0\}$, where $h > 0$ is fixed and A is uniformly distributed on $[0, 1-h]$. Compute as a function of (r, z, h) the probability, q, that the transect line (chosen as explained above) intersects the disc. Compute q to first order when r/h is small.

Remark 23.12. In a sense, line-intercept sampling is a form of 2D-from-1D *stereology* (while the usual setup is 3D-from-2D.)

23.2 Detection problems: the scan statistic

Detection problems abound, spanning a wide array of applications ranging from tumor detection from MRI images, to fire monitoring from satellite images, to disease outbreak detection (aka syndromic surveillance). Within the field of statistics, detection is a core problem in subfields such as spatial statistics and environmental statistics. In engineering, detection problems are particularly abundant in signal/image/video processing applications. In surveil-

lance applications, the main goal is often the detection of anomalies.

When mere detection is the most immediate goal, the problem is modeled in statistics as a hypothesis testing problem. After a detection has been made, it is often of interest to identify/locate the anomaly, and this problem is in turn modeled as an estimation problem.

In what follows, we focus attention on the detection of high-concentration regions.

23.2.1 Testing for clustering and the detection of clusters

Besag and Newell in [21] distinguish between *testing for clustering* and the *detection of clusters*. The difference between the two is that, in the former setting there is an apparent cluster whose (statistical) significance needs to be assessed, while in the latter setting the task is to discover statistically significant clusters.

Testing for clustering According to [21], the purpose of testing for clustering is to "investigate whether an observed pattern of cases in one or more geographical regions could reasonably have arisen by chance alone, bearing in mind the variation in background population density".

The main challenge is how to quantify the amount of evidence based on what is necessarily a post hoc analysis.

Example 23.13 (Cancer clusters)**.** A large incidence of childhood cancers in Toms Rivers, New Jersey, lead to a battle in court against a company operating a chemical plant in the area [81].

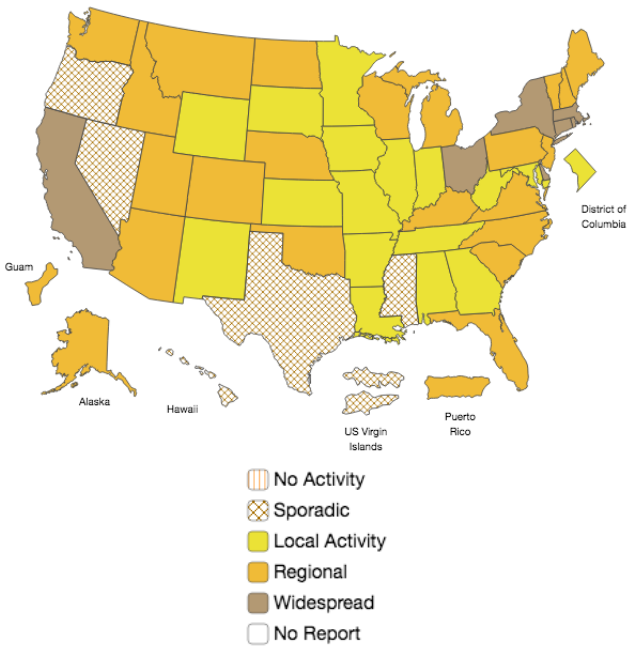
Detection of clusters According to [21], the detection of clusters consists in "screening a large region for evidence of individual 'hot spots' of disease but without any preconception about their likely locations". (In general, replace 'disease' by 'anomaly'.) See Figure 23.1 for an illustration.

We will focus on the detection of clusters problem. We will distinguish between purely spatial settings as in [145] and spatiotemporal settings as in [144]. Although time can sometimes be considered an additional dimension, most often it plays a special role. So as to keep the exposition simple, we will concern ourselves with the spatial (aka static) setting.

23.2.2 Detection in point clouds

In this setting, we observe a *point cloud*, meaning points in space, and 'look' for high-density regions.

Figure 23.1: Influenza activity estimates reported for the 14th week of 2018. Data from the Center of Disease Control (CDC).



Detecting a high-density interval Consider the emblematic (and also classical [174]) example of detecting an interval with a high-concentration of points. More formally, we observe a realization of $X_1, \ldots, X_n$, assumed to be generated iid from an unknown distribution F on $[0,1]$, and consider testing

$$\mathcal{H}_0 : \mathsf{F} \text{ is the uniform distribution on } [0,1]. \qquad (23.2)$$

(We will assume that F is continuous for concreteness.) In this generality, we would simply apply one of the goodness-of-fit tests introduced in Section 16.8.1. However, here we have in mind alternatives where there is one interval where the points are more likely to concentrate.

In fact, this is always the case, for if F is not the uniform distribution on $[0,1]$ then necessarily there is an interval $[a,b]$ such that $\mathsf{F}(b) - \mathsf{F}(a) > b - a$, and if n is large enough, such an interval will be seen to contain an unusually large number of points. ('Usual' is with respect to the null model.)

For an interval $\mathcal{I}$, let

$$N(\mathcal{I}) = \#\{i : X_i \in \mathcal{I}\}.$$

Thus N is the *counting process* in this situation.

Problem 23.14. Show that, under the null hypothesis, $N(\mathcal{I}) \sim \mathrm{Bin}(n, |\mathcal{I}|)$ for any interval $\mathcal{I} \subset [0,1]$.

When the length is known Suppose that, when present, the anomalous interval (the one with higher concentration) is of known length $\delta > 0$. In that case, it is natural to reject for large values of

$$\max_{|\mathcal{I}|=\delta} N(\mathcal{I}). \tag{23.3}$$

This is sometimes referred to as the *scan statistic*. (Here the scanning is over intervals of length δ.)

Problem 23.15. Compare the test based on (23.3) with the likelihood ratio test.

The corresponding p-value is obtained by Monte Carlo simulation, specifically, by repeatedly sampling from the uniform distribution. (We are in the setting of Section 16.8.1 after all.)

Problem 23.16. In R, write an function `scan` that takes as input the data points, an interval length, and a number of Monte Carlo replicates, and returns the scan statistic (23.3) as well as the corresponding p-value. (Optionally, have the function return the maximizing interval too, possibly together with a plot showing the points and the maximizing interval in some other color.)

When the length is unknown: normalization When the length is unknown, one cannot simply use the maximum count over all intervals, $\max_{\mathcal{I}} N(\mathcal{I})$, since this is always equal to n (the total sample size).

One possibility is to use, instead, the maximum *standardized* count, where the standardization is there to ensure that the counts over intervals of different lengths are comparable. A natural choice is

$$\max_{\mathcal{I}} \frac{N(\mathcal{I}) - |\mathcal{I}|}{\sqrt{|\mathcal{I}|(1-|\mathcal{I}|)}}. \tag{23.4}$$

This particular standardization ensures that each of the statistics we are maximizing over has zero mean and unit variance under the null hypothesis.

Problem 23.17. Compare this test with the Anderson–Darling test.

Problem 23.18. In R, write a function that computes (23.4), this time maximizing over intervals of length at least δ and at most 1/2. Follow the template of Problem 23.16. [Unless δ is quite large and/or the sample size is quite small, with high probability the maximum will be achieved by an interval with data points as end points.]

When the length is unknown: multiple testing Another possibility for dealing with the case where the anomalous interval, when present, is of unknown length, is a multiple testing approach.

Based on the count $N(\mathcal{I})$, we define the p-value associated with $\mathcal{I}$ as

$$\mathsf{pv}(\mathcal{I}) := \text{Prob}\Big\{\text{Bin}(n, |\mathcal{I}|) \geq N(\mathcal{I})\Big\}.$$

Assuming this p-value is computed for each interval, we are left dealing with a multiple testing problem. Compared to what we saw in Chapter 20, however, the setting here is unusual for two reasons: the p-values are potentially highly dependent on each other and there are infinitely many intervals to consider.

Problem 23.19. Although computing all these p-values is obviously impractical, if one is interested in using the Tippett test statistic (20.5), there are only finitely many intervals to consider. Indeed, show that the minimum p-value is attained at an interval with data points as end points.

This leads to rejecting for small values of

$$\min_{\mathcal{I}} \mathsf{pv}(\mathcal{I}) = \min_{i<j} \mathsf{pv}([X_{(i)}, X_{(j)}]),$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the ordered statistics.

Problem 23.20. In R, write a function that computes that statistic and returns the corresponding p-value. Follow the template of Problem 23.16.

Detecting a high-density region Given a compact domain $\mathcal{D} \subset \mathbb{R}^d$ and a sample of n points in $\mathcal{D}$, consider testing whether these points were generated iid from the uniform distribution on $\mathcal{D}$. The alternatives of interest are again those where there is a subregion with higher density. This, as before, calls for 'scanning' for regions with an unusually high number of points.

Problem 23.21. Generalize our treatment of the problem of detecting a high-density interval to the problem of detecting a high-density rectangle, where now the points are located in the unit square, meaning $\mathcal{D} = [0, 1]^2$.

23.2.3 Detection in discrete arrays

Suppose we observe a numerical array. As before, we 'look' for high-intensity regions within the array.

Remark 23.22 (From point clouds to arrays)**.** It is not uncommon in practice to reduce a point cloud to an array. This is usually done by binning the points as in the construction of histograms.

Detecting a high-intensity interval The prototypical example is that of a 1D array. We take the array to be $\{1, \ldots, n\}$ seen as a subset of $\mathbb{R}$, and let the intensities be denoted $X_1, \ldots, X_n$. The 'business-as-usual' scenario

is the following

$$\mathcal{H}_0 : X_1, \ldots, X_n \text{ are iid.}$$

We have in mind alternatives where large values are more likely on an (unknown) interval. This leads to 'scanning' for intervals with large average intensity. For an interval $\mathcal{I}$, which here is of the form $\mathcal{I} = \{a, \ldots, b\}$ for some $1 \le a \le b \le n$, define

$$Y[\mathcal{I}] = \sum_{i \in \mathcal{I}} X_i.$$

When the length is known When the length of the anomalous interval, when present, is known to be k, we are lead to rejecting for large values of the following scan statistic

$$\max_{|\mathcal{I}|=k} Y[\mathcal{I}], \tag{23.5}$$

where $|\mathcal{I}|$ is the cardinality of $\mathcal{I}$ and $\mathcal{I}$ is constrained to be an interval. Calibration is most naturally done by permutation.

Problem 23.23. In R, write an function perm.scan that takes as input the array, an interval length, and a number of Monte Carlo replicates, and returns the scan statistic (23.5) as well as the corresponding permutation p-value obtained by Monte Carlo. (Optionally, have the function return the maximizing interval too, possibly together with a plot of the array and the maximizing interval in some other color.)

When the length is unknown When the length of the anomalous interval, when present, is unknown, one can either take the maximum of the standardized sums, or take a multiple testing approach.

Problem 23.24. In analogy with Section 23.2.2, propose: *(i)* a scan statistic based on standardized sums, and *(ii)* a multiple testing approach.

Detecting a high-intensity region Applications in image processing and video involve (pixel) arrays in dimension 2 (still images) or 3 (video).

Problem 23.25. Formalize such a setting in analogy with Section 23.2.2. In particular, focus on the case of a 2D array, say $\{1, \ldots, n\} \times \{1, \ldots, n\}$, where an anomaly comes in the form of a rectangle (meaning, a subset of the form $\{a_1, \ldots, b_1\} \times \{a_2, \ldots, b_2\}$) with high average intensity.

23.2.4 Detection in data matrices

By data matrix we specifically mean a 2D array where the proximity between columns (resp. rows) has no meaning:

for example, two adjacent columns (resp. rows) are not, a priori, more or less related than two columns (resp. rows) that are positioned farther apart. This is in contrast with 2D pixel images, where neighboring pixels tend to be (highly) related.

An emblematic example of a detection problem in this setting is that of detecting a submatrix with unusually high intensity. Let the data matrix be denoted $(X_{ij} : i = 1, \dots, m; j = 1, \dots, n)$. 'Business-as-usual' is modeled as follows

$$\mathcal{H}_0 : \text{the matrix entries are iid.}$$

Here we have in mind alternatives where there is a submatrix with higher-than-usual intensity. This motivates 'scanning' for such submatrices. For a submatrix $\mathcal{I} \times \mathcal{J} \subset \{1, \dots, m\} \times \{1, \dots, n\}$, its total intensity is given by

$$Y[\mathcal{I}, \mathcal{J}] := \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} X_{ij}.$$

Formally, the situation parallels that of detecting a rectangle in a 2D array. In particular, if the size of an anomalous submatrix is known to be $k \times l$, then we want to reject for large values of

$$\max_{|\mathcal{I}|=k} \max_{|\mathcal{J}|=l} Y[\mathcal{I}, \mathcal{J}]. \tag{23.6}$$

Calibration is again by permutation. However, here neither $\mathcal{I}$ nor $\mathcal{J}$ is constrained to be an interval, and this makes the computation of (23.6) much harder by comparison.

Problem 23.26. Compute the number of $k \times l$ rectangles in a $m \times n$ array. Then compute the number of $k \times l$ submatrices within a $m \times n$ matrix.

For a heuristic approach, consider the procedure described in Algorithm 7. This is an *alternating maximization* procedure that consists (after some initialization) in alternating between maximizing over $\mathcal{J}$ with $\mathcal{I}$ fixed and maximizing over $\mathcal{I}$ with $\mathcal{J}$ fixed.

Algorithm 7 Alternating Maximization [207]

Input: data matrix (x_{ij}), submatrix size (k, l)
Output: submatrix $(\mathcal{I}, \mathcal{J})$

Initialize: draw $\mathcal{I}$ uniformly at random among all the row subsets of size k
Repeat until convergence:
1: Compute $\sum_{i \in \mathcal{I}} x_{ij}$ for each j and let $\mathcal{J}$ index the l largest
2: Compute $\sum_{j \in \mathcal{J}} x_{ij}$ for each i and let $\mathcal{I}$ index the k largest

Problem 23.27. Show that Algorithm 7 always terminates in a finite number of iterations. (This does not mean that one should wait until this happens. In principle, a maximum number of iterations could be specified.)

Problem 23.28. In R, write a function that implements Algorithm 7.

Remark 23.29. If the goal remains that of testing for the presence of an anomalous submatrix, or (almost equivalently) that of ascribing a significance to a submatrix, one needs to calibrate by permutation, which in principle involves computing the scan statistic (23.6) not once but $B+1$ times, where B is the number of permutations that are drawn at random.

23.3 Measurement error and deconvolution

Measurement error models presume the observations are corrupted. One of the popular models is the additive measurement error model where one observes $Y_1, \dots, Y_n$ and presupposes that

$$Y_i = X_i + \varepsilon_i,$$

with $X_1, \dots, X_n$, iid from a density f_X, independent of $\varepsilon_1, \dots, \varepsilon_n$, iid from a density f_ε. The density f_X is the object of interest and in its most ambitious form the problem is that of estimating it.

If $Y = X + \varepsilon$ denotes a generic observation, then from (6.11) we know it has density f_Y, given by the convolution of f_X and f_ε,

$$f_Y = f_X * f_\varepsilon.$$

Thus the problem of recovering f_X amounts to a *deconvolution* problem. In other words, the problem is that of undoing the convolution with f_ε. In its main variant, which is also one of its simplest, the problem setup includes knowledge of f_ε.

A popular approach then is via the characteristic function. Indeed, by the fact that X and ε are independent, we know from (7.21) that the characteristic function of Y is the product of the characteristic function of X and the characteristic function of ε, or in formula

$$\varphi_Y = \varphi_X \, \varphi_\varepsilon. \tag{23.7}$$

Thus $\varphi_X = \varphi_Y / \varphi_\varepsilon$, and using the inversion formula (7.23), we get (under some mild conditions)

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-\imath t x) \frac{\varphi_Y(t)}{\varphi_\varepsilon(t)} \, \mathrm{d}t.$$

If f_ε is known, so is φ_ε. A plug-in approach thus consists in plugging an estimator $\widehat{\varphi_Y}$ for φ_Y in the last display.

The empirical characteristic function defined in (19.7) is such an estimator. However, this estimator is typically not suitable because it performs poorly when $|t|$ is large.

Another way of obtaining an estimator for φ_Y is via an estimator, $\hat{f}_Y$, for its density f_Y, which based on (7.24) would be given by

$$\widehat{\varphi}_{\boldsymbol{Y}}(t) = \int_{-\infty}^{\infty} \exp(\imath t y)\hat{f}_Y(y)\mathrm{d}y.$$

Problem 23.30. For example, one can use a kernel density estimator for f_Y as in (16.33) based on a kernel K. Derive the resulting estimator for f_X. It happens to be a kernel density estimator. Specify the kernel in terms of f_ε and K.

23.4 Wicksell's corpuscle problem

The field of Stereology is concerned with inferencing properties of a material or tissue from one or several 2D slices.[107] Sven Dag Wicksell (1890 - 1939) was a statistician who got interested in Stereology [252] and made pioneering contributions. He was initially interested in the statistics of corpuscles present in the tissue of various human organs.

Consider the stylistic setting of a 3D medium in which a number of balls of varying radii are embedded. From a (single) 2D slide through the medium it is of interest to estimate the distribution of the radii of these balls. Note that any ball that the slice traverses leaves a disc trace on the slice.

Problem 23.31. Express the disc's radius as a function of the ball's radius and the distance of the ball's center to the slicing plane.

More formally, we assume that the ball centers are generated according to a homogeneous Poisson in $\mathbb{R}^3$ and that the radii are independent of the centers and iid with density f (and distribution function F) having mean μ. The slice can be any fixed plane because the homogeneous Poisson process is invariant with respect to orthogonal transformations.

Problem 23.32. Prove that the balls that are sliced have radii with density $xf(x)/\mu$.

Problem 23.33. Prove that, given that a ball of radius x is sliced, the radius y of the corresponding disc on the

[107] It is related but distinct from Tomography, where the goal is to reconstruct a 3D object from multiple 1D and/or 2D sections of the object.

slice has (conditional) density

$$g(y\,|\,x) := \frac{y}{x\sqrt{x^2-y^2}}, \quad \text{for } 0 < y < x.$$

Proposition 23.34. *The density of the radii of the discs that appear on the slice is given by*

$$g(y) := \frac{y}{\mu}\int_y^\infty \frac{f(x)\mathrm{d}x}{\sqrt{x^2-y^2}}. \tag{23.8}$$

Problem 23.35. Prove this proposition based on the previous two problems.

Recall that our target is f (or F) and that we can measure the disc radii on the slice so that we observe a sample from g. To see what is possible, suppose we have an infinite sample from g, so that we know g perfectly. It happens then that f is uniquely defined: indeed, the formula in (23.8) is a so-called *Abel integral* and can be inverted as follows

$$f(x) = -\frac{2\mu}{\pi}q'(x), \text{ where } q(x) := \int_x^\infty \frac{g(y)\mathrm{d}y}{\sqrt{y^2-x^2}}. \tag{23.9}$$

(In particular, the model is identifiable.)

Many approaches have been suggested to estimate f based on a finite sample (of size n) from g, denoted $Y_1, \ldots, Y_n$. In particular, there is an MLE for the distribution function. Arguably, a plug-in approach is more natural, and this is what we detail next. We follow the exposition of Groeneboom and Jongbloed in [114].

From (23.9) we obtain the following expression for the distribution function [108]

$$\mathsf{F}(x) = 1 - \frac{q(x)}{q(0)}. \tag{23.10}$$

Given an estimator $\hat{q}$ for q, a plug-in approach yields the following estimate for F:

$$\hat{\mathsf{F}}(x) := 1 - \frac{\hat{q}(x)}{\hat{q}(0)}. \tag{23.11}$$

NAIVE PLUG-IN The simplest estimator for q may well be the following plug-in estimator

$$\hat{q}_{\text{naive}}(x) := \frac{1}{n}\sum_{i=1}^n \frac{1}{\sqrt{Y_i^2-x^2}}\{Y_i > x\},$$

where we effectively estimated the distribution of Y by the empirical distribution of $Y_1, \ldots, Y_n$. The resulting

[108] Although we have assumed the existence of a density f, implying that F is absolutely continuous, this was for expository purposes. Indeed, one can work with the distribution function directly from (23.10) without assuming that F has a derivative.

estimator for the distribution function,

$$\hat{\mathsf{F}}_{\text{naive}}(x) := 1 - \frac{\hat{q}_{\text{naive}}(x)}{\hat{q}_{\text{naive}}(0)},$$

is (pointwise) consistent and achieves the 'right' convergence rate, but has unpleasant features.

Problem 23.36. Show that $\hat{\mathsf{F}}_{\text{naive}}$ is not (at all) a distribution function.

Isotonic plug-in It is possible to build on this naive estimator to obtain an estimator that is guarantied to be a bona fide distribution function. The basic idea is straightforward: Enforce properties that guaranty that the resulting estimator for F will indeed be a distribution function (Section 4.4 and Section 4.5).

Problem 23.37. Show that $q : \mathbb{R}_+ \to \mathbb{R}_+$ is bounded, non-increasing with $\lim_{x\to\infty} q(x) = 0$, and continuous from the right.

Let $\mathcal{Q}$ denote the class of functions satisfying these properties. The idea then is to find a function in $\mathcal{Q}$ that is 'close' to the naive estimator $\hat{q}_{\text{naive}}$. A popular choice for measuring 'closeness' is the L^2 metric, which leads to considering the following optimization problem

$$\text{minimize} \int_0^\infty (\tilde{q}(x) - \hat{q}_{\text{naive}}(x))^2 \, \mathrm{d}x \quad \text{over } \tilde{q} \in \mathcal{Q}.$$

This problem is solved in [114], and if $\hat{q}$ denotes the solution (shown to be unique), then the proposed estimator is given by (23.11) and by construction is a bona fide distribution function. In addition, its convergence properties are better than those of the naive estimator $\hat{\mathsf{F}}_{\text{naive}}$ as shown in the same paper.

23.5 Number of species and missing mass

The problem of estimating the number of species has a wide array of applications such as estimating biodiversity in Biology and Ecology [43], vocabulary size in Linguistics [73], the size of coinage in [78], and more [34].

We briefly discuss a few variants of this problem, including the estimation of the total number of species, the estimation of the number of species to be discovered on a second expedition, and the estimation of the missing mass.

23.5.1 Sampling models

Sampling animal species in the wild can be challenging and a variety of sampling schemes are in use. We consider a few simple sampling models where we draw from an urn in different ways. In Ecology, the draws correspond to captures or observations, the balls correspond to animals, and the labels correspond to species, and the way we draw from the urn is meant to model how animals are captured or observed.

Multinomial model As in Chapter 15, in this model we consider an experiment that consists in drawing from an urn with replacement n times. Each ball has a unique label, but what distinguishes the present situation from that of Chapter 15 is that we do not know what the labels are beforehand. This is the model we will focus on.

Hypergeometric model As in Section 15.9.2, in this model we consider an experiment that consists in drawing n balls without replacement from an urn containing balls with different labels. As before, the set of possible labels is unknown.

Bernoulli-product model In this model the experiment consists in collecting a sample in r (independent) batches of unique labels, where each batch contains label x independently with (unknown) probability π_x.

Problem 23.38. In the Bernoulli-product model, the total sample size is random. Assume that the labels are in fact positive integers. Derive a sufficient condition on $(\pi_1, \pi_2, \dots)$ for the total sample size to be finite with probability one. [An obvious condition is that $\pi_x = 0$ except for finitely many x. Find a more general condition.]

Problem 23.39. In the Bernoulli-product model, compute the expected number of labels that are not represented in the sample as a function of r and $(\pi_1, \pi_2, \dots)$.

Capture-recapture model In this model, the sampling is in two stages. In Stage 1, we draw from an urn without replacement a certain number of balls, x_1. After recording their labels, these balls are tagged and then returned to the urn. In Stage 2, we draw from an urn without replacement a certain number of balls, x_2. Among them, we find that a certain number, X_{12}, that were tagged in Stage 1.

Problem 23.40. Derive the maximum likelihood estimator for the total number of balls in the urn. (Note that x_1 and x_2 are deterministic.)

Problem 23.41. In R, write a function that takes in

(x_1, x_2, x_{12}) and a confidence level, and returns the corresponding confidence interval for the total number of balls in the urn based on inverting the likelihood ratio test (Section 12.4.9).

Problem 23.42. Consider the *Lincoln–Petersen estimator*, defined as $x_1 x_2 / X_{12}$. Derive an approximate confidence interval based on a normal approximation for this statistic in an asymptotic setting where x_1 and x_2 increase as the urn size increases.

23.5.2 Reduction by sufficiency and invariance

We consider the multinomial model above[109] The main difference with situation in Chapter 15 is that here we do not know what the labels are in advance. Nevertheless, it is always possible to reduce the situation to where the labels are the positive integers by, say, numbering the labels in the order in which they appear in the sample. For example,

sample	✣	✺	✣	✷	✱	☆	✱	★	✦	★
numbering	1	2	1	3	4	5	4	6	7	6

[109] We could as easily consider the roll of a die instead of the sampling of balls from an urn, where the only relevant distinction is that the sampling probability is not constrained to be rational.

Although this can be done, the situation is not the same as the one where we know beforehand that the labels are in $\{1, 2, \dots\}$. Indeed, in the present situation it does not make sense to talk about the probability that of drawing a ball with a certain label, since we do not know the labels beforehand.

That point being made, it is natural to model the situation as follows. We do assume that the labels are in $\{1, 2, \dots\}$ (so that we know them in advance) but only allow ourselves to consider features of the urn distribution that are invariant with respect to a permutation of the labels. An example of such a feature is the total number of distinct labels.

In the multinomial model, we thus observe an iid sample, $X_1, \dots, X_n$, from an unknown distribution f on the positive integers, where $f(x)$ is the probability that a draw from the urn result in a ball labeled $x \in \{1, 2, \dots\}$. We know that the counts

$$Y_x := \#\{i : X_i = x\}$$

are jointly sufficient. A reduction by sufficiency allows us to focus on these counts.

Invariance and sufficiency considerations, combined, lead us to base our inference not on the counts themselves as we did in Chapter 15, but on the *counts of counts* [sic],

defined as

$$\Phi_y := \#\{x : Y_x = y\}.$$

Indeed, just like knowledge of the Y_x is equivalent to knowledge of the X_i up to permutation (of the i), it is also the case that knowledge of the Φ_y is equivalent to knowledge of the Y_x up to permutation (of the x).

23.5.3 Total number of species

Consider estimating the total number of distinct labels, N, which is also the size of the support of f. This feature is indeed invariant with respect to a permutation of the labels, and so fair game in our framework.

Example 23.43 (Shakespeare's vocabulary)**.** Efron and Thisted [73] consider estimating the number of (English) words Shakespeare knew based on the published works attributed to him.

An obvious lower bound for N is the number of distinct values in the sample, denoted D. Finding an upper bound for N is however ill-posed in general. This is because the number of labels could be very large, with some labels being necessarily very minimally represented.

Problem 23.44. Fix $N \geq 1$, $n \geq 1$, and $\delta \in (0,1)$. Find f such that $f(x) > 0$ for all $x \in \{1, \dots, N\}$ and yet, with probability at least $1 - \delta$ the only label represented in a sample of size n is $x = 1$. [Note that f will depend on (N, n, δ).] Extend this to the case where $N = \infty$.

To make the problem of providing an upper bound well-posed, one typically assumes that $f(x) \geq p_{\min}$ whenever $f(x) > 0$, where $p_{\min}$ is known.[110] This immediately implies the upper bound $N \leq 1/p_{\min}$, which combined with the lower bound above yields the trivial (100%) confidence interval $D \leq N \leq 1/p_{\min}$. The goal, of course, is to improve on this, but the formulation of the problem requires some care as $D = D_n$ converges to N as n increases.

Problem 23.45. Drawing a parallel with the Coupon Collector Problem, obtain a rate of convergence for $D = D_n$. [Reduce to the worst possible case, which is when $f(x) = p_{\min}$ for all x such that $f(x) > 0$, and analyze that case.]

Taking a different route, Etsy in [79] assumes that the underlying distribution is the uniform distribution on some (unknown) subset of $\{1, 2, \dots\}$. In that case, the problem is seen to be well-posed without requiring a lower bound on f over its support set.

Problem 23.46. Show that, in this model, sufficiency

[110] If $p_{\min}$ is unknown, the assumption is vacuous.

and invariance considerations lead one to base the inference on D. In fact, show that the MLE for N is D.

It is worth comparing with the German Tank Problem (Problem 12.29). As is the case there, in the light of D being biased for N, it is tempting to consider a method of moments estimator. In view of Problem 23.46, such an estimator should be based on D.

Problem 23.47. Show that $\mathbb{E}_N(D) = N(1-(1-1/N)^n)$ and, based on that, suggest a method of moments estimator for N.

23.5.4 Number of species on a second expedition

Let us start with a classical example that lead to Fisher's involvement [87]. We follow the account given in [72, Sec 11.5]. Alexander Steven Corbet, a naturalist from the British museum, was collecting butterflies in the Malay Peninsula in the early 1940's. After two years, and perhaps considering whether it was worth staying longer, he asked Fisher how many species he would likely find if he were to continue his endeavor for an additional year. The counts of counts [sic] from his two year effort are summarized in Table 23.1.

We may formalize and generalize this problem as follows. The formalization is in terms of the number of butterflies that are captured in two consecutive expeditions instead of the time spans of these two expeditions. Focusing on the multinomial model, we assume that we have drawn n balls from the urn and ask how many new labels we are likely to draw if we were to draw another m balls from the same urn. In formula, as before, we let $X_1, \dots, X_n$ denote the available sample (of size n) and, in addition to that, we let $X_{n+1}, \dots, X_{n+m}$ denote another independent sample (of size m). Thus the X_i are iid from an unknown distribution f on $\{1, 2, \dots\}$ and the question is about the expected cardinality of $\{X_{n+1}, \dots, X_{n+m}\} \setminus \{X_1, \dots, X_n\}$.

Note that the answer to this question (after observing the second sample) does not depend on what the labels are, that is, it is invariant with respect to a permutation of the labels, and so is fair game in our framework. We also note that the underlying feature of interest depends on the urn distribution as well as on the sample sizes (n, m).

As before, sufficiency and invariance considerations leads one to focus on the counts of counts.

Problem 23.48. Show that

$$\mathbb{E}(\Phi_y) = \binom{n}{y} \sum_{x \geq 1} f(x)^y (1 - f(x))^{n-y}.$$

Table 23.1: Counts of counts summary of Corbet's expedition. Reproduced from [87, Tab 2].

y	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Φ_y	118	74	44	24	29	22	20	19	20	15	12	14	6	12	6	9	9	6	10	10	11	5	3	3

Fisher [87] approached the problem from a parametric perspective. We refer the reader to [72, Sec 11.5] for more details. The resulting estimate is a linear combination of the counts of counts, meaning of the form

$$\sum_{y\geq 1} a_y \Phi_y.$$

The coefficients (a_y) define the estimate.

Good and Toulmin [108] approached the problem directly via the method of moments and used a Poisson approximation to the binomial distribution (Theorem 3.18) to simplify some calculations. The resulting estimate is

$$\sum_{y\geq 1} (-1)^{y+1} (m/n)^y \Phi_y. \tag{23.12}$$

This estimator is almost unbiased but can have large variance, particularly when $m > n$.

23.5.5 Missing mass

In the multinomial model, given a sample of size n, the missing mass is the probability of observing a new label when drawing from the urn one more time. In formula, given the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, the missing mass is defined as

$$M_0 := \sum_{x \notin \boldsymbol{X}} f(x) = \sum_{x \geq 1} f(x)\{Y_x = 0\}.$$

The goal is to estimate the expected missing mass, $\mathbb{E}(M_0)$, which is a function of the urn distribution as well as the sample size n.

Example 23.49 (Breaking the Enigma)**.** In World War II, I.J. Good (1916 - 2009) and Alan Turing (1912 - 1954), and others at Bletchley Park, were trying to break the Enigma, a machine used by the Germans to encode messages. To use an Enigma machine, an operator had to choose a key in a key book to set up the machine before sending messages. The team of cryptanalysts at Bletchley Park wanted to estimate the frequency with which each key was used. We refer the reader to Good's own account in [109] for more details.

Problem 23.50. Show that

$$\mathbb{E}(M_0) = \sum_{x\geq 1} f(x)(1-f(x))^n.$$

More generally, define the mass of all labels represented y times in the sample as

$$M_y = \sum_{x\geq 1} f(x)\{Y_x = y\},$$

and show that

$$\mathbb{E}(M_y) = \binom{n}{y} \sum_{x\geq 1} f(x)^{y+1}(1-f(x))^{n-y}.$$

It so happens that there is no unbiased estimator of $\mathbb{E}(M_0)$. However, a method of moments approach can be based on Φ_1, which according to Problem 23.48 has expectation

$$\mathbb{E}(\Phi_1) = n \sum_{x\geq 1} f(x)(1-f(x))^{n-1},$$

leading to the estimator

$$\hat{M}_0 := \Phi_1/n.$$

Problem 23.51. Show that

$$\mathbb{E}(\hat{M}_0) - \mathbb{E}(M_0) = \frac{1}{n}\mathbb{E}(M_1),$$

and deduce that $0 \leq \mathbb{E}(\hat{M}_0 - M_0) \leq 1/n$.

Remark 23.52. Note that this estimator is the first term in the expansion defining the Good-Toulmin estimator (23.12) (here $m = 1$). The latter is in fact a refinement of the former, with smaller bias but possibly larger variance.

23.6 Information Theory

Information Theory is the mathematical science of codes and communication. It is intimately related to Probability Theory and Statistics, and we present some basics in this section. For a comprehensive introduction, we refer the reader to the classical book by Cover and Thomas [46].

Consider sending a message consisting in a sequence of symbols taken from a given discrete set over a binary communication channel. A possible way to do so is to use a pre-specified *codebook* where each symbol is coded as a binary string. For example, suppose we want to send a message originally written in English. One way to do so is to code each of the 26 letters of the English alphabet as a binary string, add one binary string to represent 'space' (to separate words), and add a binary string for each punctuation mark.

Remark 23.53. An elaboration of that is in fact implemented in the American Standard Code for Information

Interchange (ASCII). The Morse code is structured differently [46, Example 5.1.3].

Each binary symbol, 0 or 1, counts as one *bit*. The cost of sending one bit depends on the channel. With n bits one can create a codebook of size 2^n made of all possible binary sequences of length n.

Let us formalize the concept of *code*. A (binary) code for a random variable X with values in $\mathcal{X}$ (assumed discrete) is a function $\kappa : \mathcal{X} \to \bigcup_{n\geq 1}\{0,1\}^n$. We will refer to $\kappa(x)$ as the *codeword* associated with $x \in \mathcal{X}$ and the collection of all codewords is often called a *codebook*.

If f denotes the probability mass function of X then the expected length of a codeword based on the code κ is

$$\mathbb{E}(|\kappa(X)|) = \sum_{x\in\mathcal{X}} |\kappa(x)|\, f(x).$$

Desiderata. Given a random variable X, we consider the goal of designing a 'good' code, meaning a code with low expected length.

Remark 23.54. Noise and corruption complicates matters. We work with clean channels so as to keep things simple.

23.6.1 Entropy

Continuing with the same notation, the entropy of X is defined as

$$H[X] := -\sum_{x\in\mathcal{X}} f(x)\log_2 f(x),$$

where $\log_2$ denotes the logarithm in base 2. Note that $H[X] \geq 0$.

Problem 23.55. Compute the entropy of the Bernoulli distribution with parameter θ.

Proposition 23.56. *For any random variable X with finite support $\mathcal{X}$,*

$$H[X] \leq \log_2(|\mathcal{X}|),$$

with equality if and only if X is uniformly distributed on $\mathcal{X}$.

Problem 23.57. Show that if $X_1, \ldots, X_n$ are independent discrete random variables and $\boldsymbol{X} = (X_1, \ldots, X_n)$ then

$$H[\boldsymbol{X}] = H[X_1] + \cdots + H[X_n].$$

[Use a recursion on n. Note that $\boldsymbol{X}$ is here simply viewed as a discrete random variable.]

In fact, the following stronger result holds.

Proposition 23.58. *If $X_1, \dots, X_n$ are discrete random variables and $\boldsymbol{X} = (X_1, \dots, X_n)$, then*

$$H[\boldsymbol{X}] \le \sum_{i=1}^{n} H[X_i],$$

with equality if and only if $X_1, \dots, X_n$ are independent.

23.6.2 Optimal codes

A code is said to be *instantaneous* if no codeword is the prefix of another codeword. Such a code as the practical feature of being decodable sequentially, in the sense that a message made of a sequence of (code)words can be decoded as one reads the message. This is made possible by the fact that the end of a codeword is immediately recognizable without looking at the entire message.

Theorem 23.59. *The expected length of any instantaneous binary code κ for a random variable X is no less than its entropy, meaning*

$$\mathbb{E}[|\kappa(X)|] \ge H[X].$$

The lower bound can be attained if and only if X has a mass function that only takes dyadic values.

A Shannon code[111] is a code κ satisfying

$$|\kappa(x)| = \lceil \log_2(1/f(x)) \rceil.$$

The existence of a such a code relies on *Kraft's inequality* [46, Thm 5.2.2].

Problem 23.60. Show that a Shannon code is suboptimal by at most one bit per symbol, meaning that for any such code κ,

$$\mathbb{E}[|\kappa(X)|] < H[X] + 1.$$

Thus the entropy is intimately related to the expected length of an optimal code. In view of Proposition 23.56, the uniform distribution is the hardest distribution to code (which makes intuitive sense).

Problem 23.61. Choose a long text written in English with words numbering in the thousands (at least). The following operations should be done in R.

(i) Ignoring punctuation, case, hyphenation, etc, compute the distribution of the letters and the space symbol, and derive the entropy of that distribution. This gives (within one bit) the expected length of an optimal code based on coding the letters and the space symbol.

[111] Named after Claude Shannon (1916 - 2001).

(ii) Similarly, compute the distribution of the words and derive the entropy of that distribution. This gives (within one bit) the expected length of an optimal code based on coding the words. (Note that this code has a much larger codebook.)

(For more on the entropy of the English language, see [46, Sec 6.4].)

23.6.3 Quantization

Consider the problem of sending information over a discrete channel. More precisely, suppose we need to send real numbers and each number needs to be encoded with n bits. The process of reducing a number to a limited number of bits is called *quantization*.

Such a communication system is composed of two parts:

- The *encoder* takes a real number $x \in \mathbb{R}$ and returns a codeword of n bits, $\kappa : x \mapsto w$, where $w \in \{0,1\}^n$.
- The *decoder* takes the codeword w and returns a number, $\zeta : w \mapsto y$, where $y \in \mathbb{R}$.

We define the *distortion* at $x \in \mathbb{R}$ as

$$(x - \zeta \circ \kappa(x))^2.$$

Although the situation arises more generally, suppose that we want to optimize the system to send numbers that are generated iid from a distribution on the real line. Let X denote a random variable with that distribution. We then consider designing a system that has low expected distortion with respect to that distribution, meaning low

$$\mathbb{E}\left[(X - \zeta \circ \kappa(X))^2\right].$$

With $m = 2^n$, the problem is that of choosing code points $c_1, \ldots, c_m \in \mathbb{R}$ that minimize

$$\mathbb{E}\left[\min_j (X - c_j)^2\right].$$

Indeed, let $w_1, \ldots, w_m$ be codewords representing the code points $c_1, \ldots, c_m$. Then the encoding stage consists in assigning x to the codeword w_j if c_j is closest to x, and the decoding stage consists in returning the code point c_j associated with w_j.

In general, this optimization problem (over m variables) can be complicated. We present a simple heuristic method, Lloyd's algorithm, described in Algorithm 8.

R corner. Lloy's algorithm is implemented in the function `kmeans` with the option `algorithm = 'Lloyd'`. The function applies to a sample and the initialization consists in choosing m data points uniformly at random without replacement.

Algorithm 8 Lloyd's algorithm [152]

Input: distribution X, number of code points m
Output: code points $c_1, \dots, c_m$

Initialize: set c_j at the $(\frac{j}{m+1})$-quantile of X; set $a_0 = -\infty$ and $a_m = \infty$
Repeat:
(i) $a_j \leftarrow (c_j + c_{j+1})/2$ for $j = 1, \dots, m-1$
(ii) $c_j \leftarrow \mathbb{E}(X \mid X \in (a_{j-1}, a_j])$ for $j = 1, \dots, m$
Until convergence
Return the last $c_1, \dots, c_m$

Problem 23.62. The episode On Average from the podcast *99% Invisible* tells the story of how the US military designed equipment for the 'average' soldier when in fact very few soldiers are average. If you were to design military pants (say) and could do so in five waist sizes, which waist sizes would you choose? Assume you have access to the waist sizes of all soldiers.

23.7 Randomized algorithms

Randomized algorithms use (exogenous) randomness. One of the reasons for doing so is that, in some situations, it is possible to improve on what deterministic (i.e., non-randomized) algorithms can achieve, although only on average or with high probability.

23.7.1 Verifying matrix multiplication

Consider a situation where we are given three n-by-n matrices, $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}$, and the goal is to verify whether $\boldsymbol{AB} = \boldsymbol{C}$. The obvious approach — computing $\boldsymbol{AB}$ and comparing it with $\boldsymbol{C}$ — requires on the order of $O(n^\omega)$ basic operations, where $2 \le \omega < 2.373$ is the (still unknown) matrix-multiplication exponent. This method is obviously exact (up to machine precision). However, if one is willing to tolerate the possibility of making a mistake with some (small) probability, then it is possible to improve on this running time.

The basic idea is to choose a set of vectors, $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$, and simply check that $\boldsymbol{ABu}_j = \boldsymbol{Cu}_j$ for each j. For a given set of k vectors, this can be done in $O(kn^2)$ basic operations, using the fact that $\boldsymbol{ABu} = \boldsymbol{A}(\boldsymbol{Bu})$. In principle, one would want to choose $k = n$ and $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$ linearly independent. This would yield an exact method, but it would require $O(n^3)$ basic operations. The key idea is to choose $\boldsymbol{u}_1, \dots, \boldsymbol{u}_k$ at random. One way to do so is to sample these vectors independently and uniformly from the set of binary vectors of length n. It turns out that generating such vectors is not too costly, and that the

resulting procedure, *Freivalds' algorithm*, only requires $O(kn^2)$ basic operations.

Proposition 23.63. *The algorithm is always correct when $\boldsymbol{AB} = \boldsymbol{C}$, and when this is not the case, the probability that it makes a mistake is at most $1/2^k$.*

Problem 23.64. Let $\boldsymbol{D}$ be an n-by-n matrix, different from the zero matrix, and let $\boldsymbol{u}$ be sampled uniformly at random from the set of binary vectors of length n. Show that $\mathrm{Prob}(\boldsymbol{Du} = 0) \leq 1/2$. Using this, prove the proposition.

23.7.2 Randomized numerical linear algebra

There is now a substantial literature on randomized algorithms for (numerical) linear algebra [66]. This is in large part due to the fact that matrices that arise in practice are very large, and the usual ways of manipulating them do not scale too well.

One of the most important primitives is matrix multiplication. Suppose we are given an m-by-n matrix $\boldsymbol{A}$ and an n-by-p matrix $\boldsymbol{B}$, and the goal is to compute $\boldsymbol{C} := \boldsymbol{AB}$. In principle, doing this takes on the order of $O(mnp)$ basic operations. However, this running time can be substantially improved if one is willing to tolerate some random error.

The key idea is to see $\boldsymbol{C} = \boldsymbol{AB}$ as an average. Indeed, if $\boldsymbol{a}_{*k}$ denotes the kth column of $\boldsymbol{A}$ and $\boldsymbol{b}_{k*}^\top$ denotes the kth row of $\boldsymbol{B}$, then

$$\frac{1}{n}\boldsymbol{C} = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{a}_{*k}\boldsymbol{b}_{k*}^\top.$$

From the perspective of Section 7.10.2, this is the average of the population $\{\boldsymbol{a}_{*k}\boldsymbol{b}_{k*}^\top : k = 1, \ldots, n\}$. It is then tempting to sample from this population and estimate this population average with the resulting sample average. One way to do so is to sample $k_1, \ldots, k_r$ independently and uniformly from $\{1, \ldots, n\}$, and compute

$$\hat{\boldsymbol{C}} := \frac{n}{r}\sum_{s=1}^{r} \boldsymbol{a}_{*k_s}\boldsymbol{b}_{k_s*}^\top \ .$$

Computing $\hat{\boldsymbol{C}}$ can be done with $O(mrp)$ basic operations.

Problem 23.65. Show that $\hat{\boldsymbol{C}}$ is unbiased for $\boldsymbol{C}$.

Problem 23.66. For a matrix $\boldsymbol{M} = (m_{ij})$, we let $\|\boldsymbol{M}\|$ denote its Frobenius norm, defined by $\|\boldsymbol{M}\|^2 = \sum_{i,j} m_{ij}^2$. Show that

$$\mathbb{E}\left(\|\boldsymbol{C} - \hat{\boldsymbol{C}}\|^2\right) \leq \frac{n}{r}\sum_{k=1}^{n} \|\boldsymbol{a}_{*k}\|^2\|\boldsymbol{b}_{k*}\|^2.$$

Remark 23.67. It turns out that one can do better than sampling uniformly at random from $\{1, \dots, n\}$, in that there are other distributions that yield a better bound on the expected Frobenius norm above [66, Thm 22].

23.7.3 Comparing two vectors

Suppose that two binary strings $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^d$, held in memory in two distinct servers, need to be compared by a processing center while limiting the amount of information being exchanged. The straightforward method requires each server to transmit its string to the center, and this requires on the order of d bits. However, if we are willing to tolerate some probability that the strings are declared to match when in fact then do not, say $\varepsilon > 0$, then the following randomized algorithm is useful.

First, the center chooses a prime number, p, uniformly random among those not exceeding $m := 2(d/\varepsilon)\log(d/\varepsilon)$, where c is a constant to be defined below, and sends that to the servers. This requires sending $O(\log m)$ bits, because of the following variant of the Prime Number Theorem.

Lemma 23.68. *For m large enough, there are between $m/\log m$ and $2m/\log m$ prime numbers bounded by m.*

Each server computes the integer whose binary expansion corresponds to its string, computes the remainder modulo p, and sends that to the center: if $\boldsymbol{x} = (x_1, \dots, x_d)$, then this is the binary expansion of $\breve{x} := x_1 + 2x_2 + 4x_3 + \cdots + 2^{d-1}x_d$, and the corresponding server sends $\breve{x} \mod p$. Similarly, the other server sends $\breve{y} \mod p$. This requires sending $O(\log p)$ bits. The center then compares these two numbers. So the entire protocol requires the exchange of $O(\log m)$ bits. Note that $\log m = O(\log d)$ if $\varepsilon > 0$ is taken to be fixed. Below we assume that m is large enough that Lemma 23.68 applies.

Proposition 23.69. *When $\boldsymbol{x} = \boldsymbol{y}$, the algorithm is always correct, while when $\boldsymbol{x} \neq \boldsymbol{y}$, the algorithm is correct with probability at least $1 - \varepsilon$.*

Problem 23.70. Prove this proposition. [There is a short and simple proof via Lemma 23.68.]

23.7.4 Nearest neighbor search

The search for points in a database that are 'close' to a query point is a fundamental primitive in the manipulation of databases. It is also a necessary step in local methods for regression (Section 21.2).

There is an entire mathematically sophisticated literature on the topic. For a brief introduction, we focus

on a simple instance of this problem. Define the d-dimensional *Hamming space* as the space of binary strings of length d, that is $\{0,1\}^d$, equipped with the ℓ_1 metric, i.e., $\delta(\boldsymbol{x},\boldsymbol{y}) := \sum_{j=1}^{d} \{x_j \neq y_j\}$, for $\boldsymbol{x} = (x_1,\dots,x_d)$ and $\boldsymbol{y} = (y_1,\dots,y_d)$. Our goal is the following: given data points $\boldsymbol{y}_1,\dots,\boldsymbol{y}_n$, and a query point $\boldsymbol{x}$, find (if it exists) i such that

$$\delta(\boldsymbol{x},\boldsymbol{y}_i) \le r, \tag{23.13}$$

where $r > 0$ is also given.

It turns out that solving this problem exactly requires either an amount of computation that is linear in d and n, specifically, on the order of $O(dn)$ basic operations, or an amount of memory storage that is exponential in d.

Problem 23.71. Propose a simple approach with $O(dn)$ query time and $O(dn)$ storage, and then another simple approach with $O(d)$ query time and $O(2^d)$ storage. Note that the pre-processing time for the 2nd approach will also be quite large.

That said, if one is willing to tolerate some degree of approximation, and also some chance that the algorithm will not succeed, then the following simple randomized algorithm can be useful.

The idea is to subsample the coordinates and select the data points that agree with the query point on these coordinates, and then repeat this routine a number of times to achieve the desired probability of success. In detail, suppose we are willing to accept a data point $\boldsymbol{y}_i$ such that

$$\delta(\boldsymbol{x},\boldsymbol{y}_i) < cr \tag{23.14}$$

while tolerating a probability of failure of ε or smaller, for some $c > 1$ and $\varepsilon > 0$.

Here is a simple randomized algorithm for this task. Draw $\mathcal{J}_1,\dots,\mathcal{J}_m$ uniformly at random and without replacement from the class of subsets of $\{1,\dots,d\}$ of size k. For each $s \in \{1,\dots,m\}$, identify the data points that agree with the query point on the coordinates indexed by $\mathcal{J}_s$, and among these points search for one that is within distance cr of the query point. Stop the moment one such data point is found. If no such data point is found, return a symbol indicating that.

Proposition 23.72. *Suppose* $(\log n)^2 \le r \le d/\log n$. *By choosing* $k = \lceil (d/cr)\log n \rceil$ *and* $m \ge \log(1/\varepsilon) n^{1/c}$, *when* n *is large enough, with probability at least* $1-\varepsilon$ *the algorithm succeeds in finding a data point satisfying* (23.14) *when there is a data point satisfying* (23.13).

Proof. Assume for convenience that cr is an integer. Let $\mathcal{J}$ denote one of the random subsets of k coordinates.

First, let $\boldsymbol{y}$ be one of the data points. If $\delta(\boldsymbol{x}, \boldsymbol{y}) \geq cr$, meaning that $\boldsymbol{y}$ fails to satisfy (23.14), then

$$\mathbb{P}(x_j = y_j, \forall j \in \mathcal{J}) = \mathbb{P}(\mathcal{J} \subset \{j : x_j = y_j\}) \leq \frac{\binom{d-cr}{k}}{\binom{d}{k}} \leq \Big(1 - \frac{cr}{d}\Big)^k.$$

The first inequality comes from the fact that $\{j : x_j = y_j\}$ is of size at most $d - cr$. With k as defined, the expected number of such inadequate data points is at most 1.

Next, let $\boldsymbol{y}^*$ be a data point satisfying (23.13). Then the probability that it agrees with the query point on $\mathcal{J}$ is given by

$$\mathbb{P}(x_j = y_j^*, \forall j \in \mathcal{J}) = \mathbb{P}(\mathcal{J} \subset \{j : x_j = y_j^*\}) \geq \frac{\binom{d-r}{k}}{\binom{d}{k}} \geq \Big(1 - \frac{r}{d-k+1}\Big)^k.$$

By plugging in the expression for k, and some easy calculations, we find that the last expression is lower bounded by $n^{-1/c}$.

All this is for a single subset. For m subsets, we find that the expected number of inadequate points is bounded by m, and the probability of success (when possible) is not worse than if the subsets were chosen independently of each other, and for that case we find that the union bound gives a probability of success bounded from below by $1-(1-n^{-1/c})^m$, which is at least $1-\varepsilon$ when m is chosen as prescribed. □

Remark 23.73. With a use of a hash table, this approach can be implemented with a $O(dm)$ expected query time and $O(dn + mn)$ storage space.

23.7.5 Compressed sensing

Consider a setting where the state of a system is represented by a vector and our goal is to 'learn' this vector based on relatively few linear measurements. More formally, the vector $\boldsymbol{x}_* \in \mathbb{R}^d$ is the object of interest, and although it is not directly available, we have at our disposal m linear measurements, that is $\boldsymbol{a}_i^\top \boldsymbol{x}_*$, $i \in \{1, \dots, m\}$. The task consists in recovering $\boldsymbol{x}_*$ based on these inner products, ideally with m substantially smaller than n. Let $\boldsymbol{A}$ denote the matrix with row vectors $\boldsymbol{a}_1^\top, \dots, \boldsymbol{a}_m^\top$. We are provided with $\boldsymbol{y} := \boldsymbol{A}\boldsymbol{x}_*$ and our goal is to recover $\boldsymbol{x}_*$.

We consider the situation where the measurement vectors, $\boldsymbol{a}_1, \dots, \boldsymbol{a}_m \in \mathbb{R}^d$, need to be chosen beforehand. In that case, the situation is hopeless without additional structure. It turns out that assuming that $\boldsymbol{x}_*$ is a *sparse vector*, meaning a vector with few relatively large coordi-

nates, leads to a fruitful theory, and is also motivated by a number of applications. It so happens that a random design (where the choice of measurements vectors is random) is very effective. We will not elaborate on why this is so here, but rather refer the reader to [36, 233], as well as to [157] for applications to magnetic resonance imaging (MRI) and to [68] for the design of a single-pixel camera based on these principles. There is also an entire book on the topic [89].

In its simplest form, the assumption of sparsity takes the form of a vector $\boldsymbol{x}_*$ with only s non-zero coordinates. In that case, it is tempting to try to recover $\boldsymbol{x}_*$ by solving the following optimization problem:

$$\begin{aligned} &\text{minimize } \|\boldsymbol{x}\|_0 \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}, \end{aligned} \tag{23.15}$$

where $\|\boldsymbol{x}\|_0$ is the number of coordinates of $\boldsymbol{x}$ that are nonzero.

Problem 23.74. Prove that $\boldsymbol{x}_*$ is the unique solution to (23.15) if and only if $m > s$.

This is very good ($s+1$ such measurements suffice!), except that solving the problem (23.15) is computationally very difficult because of the discrete and unstable nature of $\|\cdot\|_0$. One way around this is a convex relaxation, which consists in replacing the $\|\cdot\|_0$ with $\|\cdot\|_1$, where $\|\boldsymbol{x}\|_1 := \sum_{j=1}^d |x_j|$, which is a true norm and is therefore convex:

$$\begin{aligned} &\text{minimize } \|\boldsymbol{x}\|_1 \\ &\text{subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}. \end{aligned} \tag{23.16}$$

This can be cast as linear program and is therefore more amenable to computation.

A central tenet of this theory is that, under some conditions on the design, the solution to the easier problem, (23.16), coincides with that of the harder problem, (23.15). We consider the random design where the measurement vectors are independent standard normal (in dimension d). This is equivalent to $\boldsymbol{A}$ having its entries drawn independently from the standard normal distribution.

Theorem 23.75. *Under the present conditions, assume that $m > 2s(1+\log(d/s))$. Then with probability tending to 1 as $m \to \infty$, $\boldsymbol{x}_*$ is the unique solution to (23.16).*

Note that the theorem includes settings where $s \to \infty$ or $d \to \infty$. (The result is taken from [233], which references [63], among other works.)

23.8 Statistical malpractice

In the late 1990's and early 2000's, there was an increasing concern with the fact that important findings reported in the scientific literature could not be replicated in subsequent studies. To quote Wacholder et al [241]:

> *The high chance that an initial 'statistically significant' finding will turn out to be a false-positive finding, even for large, well-designed, and well-conducted studies, is one symptom of the problem we face.*

This was later echoed by Ioannidis in [131], who then went on, with his colleagues, to identify similar issues in other scientific areas. This lack of replicability is now understood to affect many, if not all, sciences. In Oncology, researchers at a large biotech company (Amgen) were only able to replicate 6 out of 53 landmark cancer studies [11]; see also [187]. In Psychology, a group of 270 researchers joined forces to replicate 100 experiments reported in three high-ranking journals in the year 2008, and were only able to replicate less than half of them [1].

This issue has received a substantial amount of attention in the last ten years. A good overall discussion appears in [194]. In this section, we draw attention to a number of aspects of this multi-faceted issue, in particular the fact that editorial and funding policies drive researchers to ignore serious issues of multiple testing [82].

23.8.1 An examination of the Base Rate Fallacy

Both Wacholder et al [241] and Ioannidis [131], and others before [33, 218], base their discussion on a idealized situation where, in the course of research, a number of tests are performed, each having size α and power $1-\beta$ for what they are testing. Letting π denote the proportion of false null hypotheses, and applying the Bayes formula, gives (1.26) as the probability that a test chosen among these uniformly at random rejects correctly, or equivalently, gives the probability that a randomly chosen test rejects incorrectly as

$$\mathsf{B} := \frac{\alpha(1-\pi)}{(1-\beta)\pi + \alpha(1-\pi)}. \tag{23.17}$$

In particular, this is not necessarily bounded by α. (Believing so is an example of the Base Rate Fallacy.) The quantity in (23.17) is referred to as the *false positive report probability* in [241], and to the *positive predictive value* in [131].

Remark 23.76. Although simplistic, the calculations generalize to a somewhat more realistic setting as seen in Problem 1.52.

We see in (23.17) that this base rate is increasing in the level, decreasing in the power, and decreasing in the proportion of false null hypotheses. Although most researchers understand that the smaller the level the stronger the protection against false discoveries, the impact of the power and that of the proportion of false null hypotheses is not as well appreciated. Following [131], we encapsulate these impacts in the following maxims.

Proverb. *The higher the number of underpowered studies, the larger the proportion of false discoveries.*

Lack of power often comes from a sample that is too small for the sought-after effect size. (This is why power calculations, as presented in Section 11.2.3, are important, and often required when applying for research funding.)

Proverb. *The higher the number of hypotheses that are tested, the larger the proportion of false discoveries.*

This is particularly relevant in gene association studies and also in the neurosciences, where a single study can generate tens or even hundreds of thousands of null hypotheses, and it is generally accepted that among these only a tiny fraction are false.

23.8.2 Publication bias

Dickersin and Min [62] define *publication bias* as follows

> *Publication bias is any tendency on the parts of investigators or editors to fail to publish study results on the basis of the direction or strength of the study findings.*

In that paper they provide strong evidence that publication bias is pervasive in psychological and medical research over several decades, but this is true in all the sciences. Indeed, the vast majority of research articles report positive findings. This stems from a strong preference from journal editors (and the scientific community at large, as well as funding agencies and other sponsors) for publishing novel findings. This in turn has lead researchers to practice *selective reporting*, meaning, to mostly submit experiments resulting in significance. [112]

Publication bias can, and does, have important (negative) consequences, for example, it can over-inflate the

[112] For a satirical cartoon illustration, see xkcd.com/882

efficacy of certain drugs or procedures for a given ailment. To quote [234]:

> *Evidence-based medicine is valuable to the extent that the evidence base is complete and unbiased. Selective publication of clinical trials — and the outcomes within those trials — can lead to unrealistic estimates of drug effectiveness and alter the apparent risk-benefit ratio.*

Example 23.77 (FDA-registered anti-depressant studies)**.** The authors in [234] obtained "reviews from the US Food and Drug Administration (FDA) for studies of 12 antidepressant agents involving 12564 patients." There were 74 FDA-registered studies in total. Among the 38 studies viewed by the FDA as having positive results, all but one were published. Among the 36 studies viewed by the FDA as having negative or questionable results, 22 were not published and 11 were published, but presented in a in a way that conveyed a positive outcome.

Over-inflation of effects In a different paper, Ioannidis [132] reports on findings in the biomedical literature of over-inflated effects. This is in part due to the fact that there is a significance threshold to achieve for publication, which then triggers a *winner's curse* phenomenon.

Problem 23.78. As an illustration, consider a crossover experiment comparing some drug to a placebo. In the analysis, this comparison is done by testing $\mu \leq 0$, where μ is the median of the difference (treatment minus placebo), using the sign test (Section 16.3.3). Assume the experiment involves n subjects. The experiment is repeated until rejecting $\mu \leq 0$ at the 5% level. Based solely on the last experiment (the first one achieving significance), a one-sided 95% confidence interval for μ is produced. This interval tends to be biased upward and in particular does not have the right confidence level. In R, perform some simulations to verify this assertion.

Problem 23.79. Suppose you receive a promotional email from a investment manager offering his predictions on a particular stock, specifically whether the price of the stock will increase or decrease by the end of the week. For ten weeks straight, his predictions are correct. If this were a scam, explain how it would work. (Adapted from [119].)

Importance of negative findings A first step towards solving the issue of publication bias is to recognize that negative results are an important part of the scientific discovery process. To quote [163] (which provides a wealth of interesting, if not shocking examples of bias towards positive results):

> *Negative findings are fundamental to science: they encourage good scientific practice, teach us to critically analyze our pre-existing thoughts and direct new avenues of research.*

Some journals have been created for the sole purpose of publishing negative findings. [113]

Example 23.80 (Psychics)**.** In [190], Randi recounts his observation of a 'psychic' supposedly 'reading' members of an audience. We are told that the audience was left quite impressed with the performance. Randi tells us that "A few days later, I invited two of these persons [from the audience] to my home to put on tape their accounts of his performance. I then played for them the tape recording of the broadcast that I had made, and we discovered by actual count that this so-called psychic had, on the average, been correct in one out of fourteen of his statements! [...] To the dismay of my visitors, their accounts had been far from accurate. Selective thinking had led them to dismiss all the apparent misses and the obviously wrong guesses and remember only the 'hits'."

Preregistration Another step being taken, or at least seriously considered and discussed, to mitigate publication bias is that of requiring that experiments be preregistered [16]. This has already been implemented in the context of clinical trials [53, 204]. The US National Institutes of Health (NIH) provide a website for registering trials (`ClinicalTrials.gov`), which also accepts the registration of observational studies [253]. For an example of a preregistered trial, see [128]. A promising two-step submission process is described in [216].

[113] For example, the *Journal of Negative Results in BioMedicine*.

23.8.3 Pre-selection bias

Pre-selection bias occurs when a hypothesis to be tested (or confidence interval to be built), or the method of analysis, is selected based on the data. When this is not taken into account, it can lead to severe bias [206].

Taking a peak at the data in that way is sometimes called *data snooping*. The process of looking for a hypothesis and/or a methodology that yields sufficient statistical significance is known as *significance chasing* or *p-hacking* or *data-dredging*. Such practices may stem from a lack of understanding of multiple testing, or may be a result of the systemic pressure of achieving a certain level of statistical significance as discussed above.

An analysis performed after looking at the data is sometimes called a *post hoc analysis*.

MULTIPLE END POINTS In clinical trials in particular, the term *end point* is used to denote an outcome (e.g., death, death from a particular disease, the contracting of a particular disease, presence of particular symptoms, blood pressure, cholesterol level, level of anxiety, etc). There is a *multiple end points* problem when the outcome(s) of interest was(were) not clearly specified before the start of the clinical trial. In well conducted clinical trials, a primary and possibly several secondary outcomes are specified in the planning phase. Changing the primary outcome midway (or at the end) is only done under dire circumstances. Of course, this is an issue in all experiments, clinical and other. For example, it is a recognized issue in ESP research [58]. (The problem of stopping a trial early is a variant of the problem of multiple end points.)

Example 23.81 (Chocolate helps weight loss - really?)**.** In a study [26] on the effects of dark chocolate, subjects were randomized to three group. The first group was put on a low-carb diet with a daily serving of 42 grams of dark chocolate. The second group was put on the same low-carb diet but without the daily serving of dark chocolate. And the third group was told to continue to eat their regular diet. Among other 'benefits', the chocolate group experienced the most weight loss. This may be surprising. It turns out the study was a hoax! John Bohannon, a science journalist, was the instigator. He wanted to show how easy it was to publish grossly flawed diet research. Of his own account [25]:

> *The study was 100 percent authentic. My colleagues and I recruited actual human subjects in Germany. We ran an actual clinical trial, with subjects randomly assigned to different diet regimes. And the statistically significant benefits of chocolate that we reported are based on the actual data.*

So is the result of the study simply due to chance variation? Yes, but the authors could not afford to repeat the experiment many times to obtain significance. Instead, they allowed themselves to choose the end points after seeing the data. As Bohannon explains:

> *Here's a dirty little science secret: If you measure a large number of things about a small number of people, you are almost guaranteed to get a 'statistically significant' result.*

It is particularly revealing that, except for issues of multiple end points and small sample size, there were no other obvious flaws in the design or execution of the study. The authors sent the paper to multiple 'fake' journals and was

accepted overnight by several journals [114], and the press picked up the story from there.

FLEXIBILITY IN DATA ANALYSIS Not only do we need to specify, before the experiment begins, which outcomes are of interest and what we want to test about them, but the same is true of the method of analysis. The *researcher degrees of freedom* [212] refers to the fact that a researcher can choose the method of analysis based on the data (although he/she should not). Others speak of a *garden of forking paths* [102]. It is a pervasive issue. For example, a meta-analysis of fMRI studies reported [38]:

> *Across 241 studies, 223 unique combinations of analytic techniques were observed. In other words, there were nearly as many unique analysis pipelines as studies in the sample.*

CONFIDENCE INTERVAL FOR SELECTED PARAMETERS In Section 23.8.1 the discussion focused on tests of significance. We found that, although we may be testing at the α level, the (expected) proportion of incorrectly rejected hypotheses can be much higher than α. A similar phenomenon arises in the context of confidence intervals.

[114] Bohannon has written about this issue [24].

Problem 23.82. Suppose that a number of $(1-\alpha)$-confidence intervals are computed, each for a different target parameter. Prove that the expected proportion of these intervals that contain their target parameter is $1-\alpha$.

It is not uncommon, however, to first perform a test to determine whether a parameter is of interest. This *selection* phase is then followed by producing a $(1-\alpha)$-confidence interval for each selected parameter. Sorić in [218] argues that the expected proportion of these intervals that do not contain their target parameter may be much larger than α.

Problem 23.83. Argue that this is indeed the case by following the arguments of Section 23.8.1.

Benjamini and Yekutieli in [13] propose a method for controlling what they call the *false coverage rate*, defined as the expected proportion of confidence intervals for selected parameters that miss their target. (In their work, the selection method is general.)

POST SELECTION INFERENCE Regression analyses (which most of the time amount to fitting a parametric model to data) are routinely performed based on multiple looks at the data. In particular, if the predictor variable is one-dimensional, it is hard not to look at a scatterplot be-

fore choosing a model. In fact, doing this is recommended in a number of textbooks.

It is also quite standard to perform model selection (using, e.g., cross-validation) to arrive at the final model. Then inference is performed about the final model. (For example, if it is a linear model, then confidence intervals may be produced for several of the coefficients.) The whole process that leads to the final model is rarely taken into account in the inference, and this makes the inference potentially very biased. (For more details, see [14].)

23.8.4 Replicability and reproducibility

Experimental results are deemed *replicable* when subsequent experiments carried out under the same conditions replicate the result, meaning that they yield congruent results. Replicability is the golden standard by which empirical research findings are engraved as 'real phenomena' or 'laws of nature'.

Utts in [235] argues, in the context of replicability in parapsychology research, that judging how well some experimental results replicate is not intuitive at first.

Problem 23.84. Consider a binomial experiment based on k tossed of a θ-coin. Suppose the hypothesis that $\theta \leq 1/2$ is rejected at the 0.05 level. (The most powerful test for that purpose is used throughout.)

(i) Argue in words (with no mathematical formula) that the probability that an identical experiment also rejects this hypothesis at the 0.05 level can be as low as 0.05.

(ii) Compute this probability as a function of θ and k.

There are many possible reasons an experimental result may fail to replicate. A survey [10] of more than 1500 scientists from various fields asked about the main causes for this. Their answers can be organized as follows:

- *Cultural or systemic:* selective reporting; pressure to publish; not replicated enough in the original lab; insufficient oversight/mentoring; fraud; insufficient peer review.
- *Flawed methodology:* low statistical power or poor analysis; poor experimental design.
- *Missing details or data:* methods or code unavailable; raw data not available.

This is echoed in [224] with concrete examples.

The first set of issues has to do with various pressures on researchers. The second set has to do with an incorrect use of statistics and other methods for data analysis. The third set of issues falls under the umbrella of *computational*

reproducibility [64, 182] and has to do with being able to reproduce an analysis carried out in a scientific article.

Example 23.85 (Mendel's experiments)**.** Mendel's experiments on plant hybridization [166] and his theory of inheritance are important achievements in Genetics. Although this has been widely recognized for quite some time, there has been a longstanding controversy surrounding the data published in Mendel's original paper, at least since Fisher raised concerns that the data appeared too good to be true [85]. There is an entire book on the topic [91]. One of the main points of contention is how the experiments were actually performed. (Fisher based his calculations on some assumptions that have been contested since then [118].)

Open Science Some initiatives aimed at promoting a more open culture of scientific research are getting some traction. [115] Also, some recommendations have been made for the reporting of studies, particularly in the medical research literature. [116] Some general recommendations are given in [177].

[115] Examples include the *Open Science Framework* (`osf.io`), the *Open Scholarship Initiative* (`osinitiative.org`), the *Peer Reviewers' Openness Initiative* (`opennessinitiative.org`).

[116] These include clinical trials (`consort-statement.org`), observational studies (`strobe-statement.org`), and meta-analyses (`prisma-statement.org`). These and others are listed by the *Equator Network* (`equator-network.org`).

BIBLIOGRAPHY

[1] A. Aarts, J. Anderson, C. Anderson, P. Attridge, A. Attwood, and A. Fedor. Estimating the reproducibility of psychological science. *Science*, 349(6251):1–8, 2015.

[2] A. Agresti and B. Presnell. Misvotes, undervotes and overvotes: The 2000 presidential election in Florida. *Statistical Science*, 17(4):436–440, 2002.

[3] F. M. Aguilar, J. Nijs, Y. Gidron, N. Roussel, R. Vanderstraeten, D. Van Dyck, E. Huysmans, and M. De Kooning. Auto-targeted neurostimulation is not superior to placebo in chronic low back pain: A fourfold blind randomized clinical trial. *Pain Physician*, 19(5):E707, 2016.

[4] M. G. Akritas. Bootstrapping the Kaplan–Meier estimator. *Journal of the American Statistical Association*, 81(396):1032–1038, 1986.

[5] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47, 2002.

[6] K. Arceneaux, A. S. Gerber, and D. P. Green. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis*, 14(1):37–62, 2005.

[7] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

[8] R. Arratia, L. Goldstein, and L. Gordon. Two moments suffice for Poisson approximations: the Chen–Stein method. *The Annals of Probability*, pages 9–25, 1989.

[9] P. C. Austin. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Statistics in Medicine*, 27(12):2037–2049, 2008.

[10] M. Baker. 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454, 2016.

[11] C. G. Begley and L. M. Ellis. Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391):531–533, 2012.

[12] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 51(1):289–300, 1995.

[13] Y. Benjamini and D. Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

[14] R. Berk, L. Brown, and L. Zhao. Statistical inference after

model selection. *Journal of Quantitative Criminology*, 26(2):217–236, 2010.

[15] R. H. Berk and D. H. Jones. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 47(1):47–59, 1979.

[16] J. A. Berlin and D. Ghersi. Preventing publication bias: Registries and prospective meta-analysis. In H. R. Rothstein, A. J. Sutton, and M. Borenstein, editors, *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*, pages 35–48. Wiley, Chichester, UK, 2005.

[17] D. Bernoulli. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 5:175–192, 1738.

[18] D. Bernoulli. Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36, 1954.

[19] M. Bertrand and S. Mullainathan. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013, 2004.

[20] J. Besag and P. Clifford. Generalized monte carlo significance tests. *Biometrika*, 76(4):633–642, 1989.

[21] J. Besag and J. Newell. The detection of clusters in rare diseases. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, pages 143–155, 1991.

[22] P. J. Bickel, E. A. Hammel, and J. W. O'Connell. Sex bias in graduate admissions: data from Berkeley. *Science*, 187(4175):398–404, 1975.

[23] N. Black. Why we need observational studies to evaluate the effectiveness of health care. *British Medical Journal*, 312(7040):1215–1218, 1996.

[24] J. Bohannon. Who's afraid of peer review? *Science*, 342(6154):60–65, 2013.

[25] J. Bohannon. I fooled millions into thinking chocolate helps weight loss. Here's how. Gizmodo, May 27th 2015. tinyurl.com/y9g9e9ly.

[26] J. Bohannon, D. Koch, P. Homm, and A. Driehaus. Chocolate with high cocoa content as a weight-loss accelerator. *Global Journal of Medical Research*, 2015.

[27] B. Bollobás. *Random Graphs*. Cambridge University Press, 2001.

[28] R. Bonita, R. Beaglehole, and T. Kjellström. *Basic Epidemiology*. World Health Organization, 2006.

[29] S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

[30] J.-P. P. Briefcase. Insuring the uninsured. Poverty Action Lab, 2014. povertyactionlab.org/node/9604.

[31] S. R. Broadbent and J. M. Hammersley. Percolation processes: I. Crystals and mazes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 53, pages 629–641. Cambridge University Press, 1957.

[32] G. Bronfort, R. Evans, B. Nelson, P. D. Aker, C. H. Goldsmith, and H. Vernon. A randomized clinical trial of exercise and spinal manipulation for patients with chronic neck pain. *Spine*, 26(7):788–797, 2001.

[33] W. S. Browner and T. B. Newman. Are all significant p-values created equal?: The analogy between diagnostic tests and clinical research. *Journal of the American Medical Association*, 257(18):2459–2463, 1987.

[34] J. Bunge and M. Fitzpatrick. Estimating the number of species: review. *Journal of the American Statistical Association*, 88(421):364–373, 1993.

[35] D. Calnitsky. "More normal than welfare": The Mincome experiment, stigma, and community experience. *Canadian Review of Sociology/Revue canadienne de sociologie*, 53(1):26–71, 2016.

[36] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

[37] D. Card, C. Dobkin, and N. Maestas. Does Medicare save lives? *The Quarterly Journal of Economics*, 124(2):597–636, 2009.

[38] J. Carp. The secret lives of experiments: Methods reporting in the fMRI literature. *Neuroimage*, 63(1):289–300, 2012.

[39] D. L. Chen, T. J. Moskowitz, and K. Shue. Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, 131(3):1181–1242, 2016.

[40] M. A. Cliff, M. C. King, and J. Schlosser. Anthocyanin, phenolic composition, colour measurement and sensory analysis of bc commercial red wines. *Food Research International*, 40(1):92–100, 2007.

[41] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[42] W. G. Cochran and S. P. Chambers. The planning of observational studies of human populations. *Journal of the Royal Statistical Society: Series A (General)*, 128(2):234–266, 1965.

[43] R. K. Colwell, A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, 5(1):3–21, 2012.

[44] E. F. Connor and D. Simberloff. The assembly of species communities: Chance or competition? *Ecology*, 60(6):1132–1140, 1979.

[45] T. M. Cover. Pick the largest number. In *Open Problems in Communication and Computation*, pages 152–152. Springer, 1987.

[46] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

[47] P. Craig, C. Cooper, D. Gunnell, S. Haw, K. Lawson, S. Macintyre, D. Ogilvie, M. Petticrew, B. Reeves, M. Sutton, and S. Thompson. Using natural experiments to evaluate population health interventions: New medical research council guidance. *Journal of Epidemiology and Community Health*, pages 1182–1186, 2012.

[48] H. Cramér. On the composition of elementary errors: Statistical applications. *Scandinavian Actuarial Journal*,

1928(1):141–180, 1928.

[49] H. Crane. The ubiquitous Ewens sampling formula. *Statistical Science*, 31(1):1–19, 2016.

[50] X. Cui and G. A. Churchill. Statistical tests for differential expression in cdna microarray experiments. *Genome biology*, 4(4):210, 2003.

[51] D. A. Darling and P. Erdős. A limit theorem for the maximum of normalized sums of independent random variables. *Duke Mathematical Journal*, 23(1):143–155, 1956.

[52] A. Dasgupta. Right or wrong, our confidence intervals. *IMS Bulletin*, Dec. 2012.

[53] C. De Angelis, J. M. Drazen, F. A. Frizelle, C. Haug, J. Hoey, R. Horton, S. Kotzin, C. Laine, A. Marusic, and A. J. P. Overbeke. Clinical trial registration: A statement from the international committee of medical journal editors. *The Lancet*, 364(9438):911–912, 2004.

[54] A. J. De Craen, T. J. Kaptchuk, J. G. Tijssen, and J. Kleijnen. Placebos and placebo effects in medicine: Historical overview. *Journal of the Royal Society of Medicine*, 92(10):511–515, 1999.

[55] J. de Leeuw, K. Hornik, and P. Mair. Isotone optimization in R: Pool-adjacent-violators algorithm (pava) and active set methods. *Journal of Statistical Software*, 32(5):1–24, 2009.

[56] A. Dechartres, L. Trinquart, I. Boutron, and P. Ravaud. Influence of trial sample size on treatment effect estimates: Meta-epidemiological study. *British Medical Journal*, 346:f2304, 2013.

[57] A. Di Sabatino, U. Volta, C. Salvatore, P. Biancheri, G. Caio, R. De Giorgio, M. Di Stefano, and G. R. Corazza. Small amounts of gluten in subjects with suspected non-celiac gluten sensitivity: a randomized, double-blind, placebo-controlled, cross-over trial. *Clinical Gastroenterology and Hepatology*, 13(9):1604–1612, 2015.

[58] P. Diaconis. Statistical problems in ESP research. *Science*, 201(4351):131–136, 1978.

[59] P. Diaconis and A. Gangolli. Rectangular arrays with fixed margins. *IMA Volumes in Mathematics and its Applications*, 72:15–15, 1995.

[60] P. Diaconis and F. Mosteller. Methods for studying coincidences. *Journal of the American Statistical Association*, 84(408):853–861, 1989.

[61] J. Diamond. Assembly of species communities. In M. Cody and J. Diamond, editors, *Ecology and Evolution of Communities*, pages 342–444. Harvard Univ. Press, 1975.

[62] K. Dickersin and Y.-I. MIN. Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences*, 703(1):135–148, 1993.

[63] D. Donoho and J. Tanner. Counting faces of randomly projected polytopes when the projection radically lowers dimension. *Journal of the American Mathematical Society*, 22(1):1–53, 2009.

[64] D. L. Donoho. An invitation to reproducible computational research. *Biostatistics*, 11(3):385–388, 2010.

[65] R. Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.

[66] P. Drineas and M. W. Mahoney. Lectures on randomized numerical linear algebra. *arXiv preprint arXiv:1712.08880*, 2017.

[67] D. Du, F. K. Hwang, and F. Hwang. *Combinatorial Group Testing and its Applications*. World Scientific, 2000.

[68] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):83–91, 2008.

[69] C. Dutang and D. Wuertz. A note on random number generation, 2009. Vignette for the randtoolbox package.

[70] A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

[71] R. Eckhardt. Stan Ulam, John von Neumann, and the Monte Carlo method. *Los Alamos Science, Special Issue*, 15:131–137, 1987.

[72] B. Efron. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2013.

[73] B. Efron and R. Thisted. Estimating the number of unseen species: How many words did shakespeare know? *Biometrika*, 63(3):435–447, 1976.

[74] R. J. Ellis, W. Toperoff, F. Vaida, G. Van Den Brande, J. Gonzales, B. Gouaux, H. Bentley, and J. H. Atkinson. Smoked medicinal cannabis for neuropathic pain in HIV: a randomized, crossover clinical trial. *Neuropsychopharmacology*, 34(3):672–680, 2009.

[75] P. Erdős and A. Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

[76] P. Erdős and A. Rényi. On a classical problem of probability theory. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei*, 6:215–220, 1961.

[77] P. Erdős and A. Rényi. On a new law of large numbers. *Journal d'Analyse Mathématique*, 23(1):103–111, 1970.

[78] W. W. Esty. Estimation of the size of a coinage: a survey and comparison of methods. *The Numismatic Chronicle*, pages 185–215, 1986.

[79] W. W. Esty et al. The efficiency of Good's nonparametric coverage estimator. *The Annals of Statistics*, 14(3):1257–1260, 1986.

[80] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theoretical population biology*, 3(1):87–112, 1972.

[81] D. Fagin. *Toms River: a Story of Science and Salvation*. Bantam, 2013.

[82] D. Fanelli. Do pressures to publish increase scientists' bias? An empirical support from US States data. *PloS One*, 5(4):e10271, 2010.

[83] W. Feller. *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley & Sons, 3rd edition, 1968.

[84] T. S. Ferguson. Who solved the secretary problem? *Statistical science*, 4(3):282–289, 1989.

[85] R. A. Fisher. Has Mendel's work been rediscovered? *Annals of Science*, 1(2):115–137, 1936.

[86] R. A. Fisher. *The Design of Experiments*. Oliver And Boyd; Edinburgh; London, 2nd edition, 1937.

[87] R. A. Fisher, A. S. Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, 12(1):42–58, 1943.

[88] J. D. Forbes. XIV.—Further experiments and remarks on the measurement of heights by the boiling point of water. *Transactions of the Royal Society of Edinburgh*, 21(02):235–243, 1857.

[89] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013.

[90] D. Fouque, M. Laville, J. Boissel, R. Chifflet, M. Labeeuw, and P. Zech. Controlled low protein diets in chronic renal insufficiency: Meta-analysis. *British Medical Journal*, 304(6821):216–220, 1992.

[91] A. Franklin, A. Edwards, D. J. Fairbanks, and D. L. Hartl. *Ending the Mendel–Fisher Controversy*. University of Pittsburgh Press, 2008.

[92] D. Freedman, R. Pisani, and R. Purves. *Statistics*. WW Norton & Company, 4th edition, 2007.

[93] D. Freedman and P. Stark. What is the chance of an earthquake? *NATO Science Series IV: Earth and Environmental Sciences*, 32:201–213, 2003.

[94] D. A. Freedman. Statistical models and shoe leather. *Sociological methodology*, pages 291–313, 1991.

[95] D. A. Freedman and R. A. Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409, 2008.

[96] D. A. Freedman, D. Collier, J. S. Sekhon, and P. B. Stark. *Statistical Models and Causal Inference: a Dialogue with the Social Sciences*. Cambridge University Press, 2010.

[97] D. A. Freedman and K. W. Wachter. Census adjustment: Statistical promise or illusion? *Society*, 39(1):26–33, 2001.

[98] N. Freemantle, L. Marston, K. Walters, J. Wood, M. R. Reynolds, and I. Petersen. Making inferences on treatment effects from real world data: Propensity scores, confounding by indication, and other perils for the unwary in observational research. *British Medical Journal*, 347:f6409, 2013.

[99] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger. *Fundamentals of Clinical Trials*. Springer, 5th edition, 2015.

[100] M. Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32(200):675–701, 1937.

[101] M. Gardner. *The Second Scientific American Book of Mathematical Puzzles and Diversions*. Simon and Schuster, New York, 1961.

[102] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when

there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Technical report, Columbia University, 2013.

[103] C. Genovese and L. Wasserman. Operating characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):499–517, 2002.

[104] E. N. Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

[105] E. N. Gilbert. Random plane networks. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):533–543, 1961.

[106] T. A. Glass, S. N. Goodman, M. A. Hernán, and J. M. Samet. Causal inference in public health. *Annual Review of Public Health*, 34:61–75, 2013.

[107] L. J. Gleser and I. Olkin. Models for estimating the number of unpublished studies. *Statistics in Medicine*, 15(23):2493–2507, 1996.

[108] I. Good and G. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

[109] I. J. Good. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation*, 66(2):101–111, 2000.

[110] N. J. Gotelli and G. L. Entsminger. Swap and fill algorithms in null model analysis: Rethinking the knight's tour. *Oecologia*, 129(2):281–291, 2001.

[111] N. J. Gotelli and G. R. Graves. *Null Models in Ecology*. Smithsonian Institution Press, 1996.

[112] U. Grenander. On the theory of mortality measurement: Part II. *Scandinavian Actuarial Journal*, 1956(2):125–153, 1956.

[113] C. M. Grinstead and J. L. Snell. *Introduction to Probability*. American Mathematical Society, 2nd edition, 1997. Available online on the authors' website.

[114] P. Groeneboom and G. Jongbloed. Isotonic estimation and rates of convergence in wicksell's problem. *The Annals of Statistics*, 23(5):1518–1542, 1995.

[115] S. P. Gutthann, L. A. G. Rodriguez, J. Castellsague, and A. D. Oliart. Hormone replacement therapy and risk of venous thromboembolism: Population based case-control study. *British Medical Journal*, 314(7083):796, 1997.

[116] P. Halpern. Isaac Newton vs Las Vegas: How physicists used science to beat the odds at roulette. *Forbes Magazine*, May 2017.

[117] B. B. Hansen. Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, 99(467):609–618, 2004.

[118] D. L. Hartl and D. J. Fairbanks. Mud sticks: On the alleged falsification of Mendel's data. *Genetics*, 175(3):975–979, 2007.

[119] C. R. Harvey and Y. Liu. Evaluating trading strategies. *The Journal of Portfolio Management*, 40(5):108–118, 2014.

[120] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements*

of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, 2nd edition, 2009.

[121] R. Heffernan, F. Mostashari, D. Das, A. Karpati, M. Kulldorff, and D. Weiss. Syndromic surveillance in public health practice, New York City. *Emerging Infectious Diseases*, 10(5), 2004.

[122] K. J. Henning. What is syndromic surveillance? *Morbidity and Mortality Weekly Report*, 53:7–11, 2004.

[123] P. J. Henry. College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, 19(2):49–71, 2008.

[124] J. P. Higgins and S. Green, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. John Wiley & Sons, 2011. Available from handbook.cochrane.org.

[125] A. B. Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58(5):295–300, 1965.

[126] W. Hoeffding. A non-parametric test of independence. *The Annals of Mathematical Statistics*, 19(4):546–557, 1948.

[127] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[128] P. Honigmann, H. Fischer, A. Kurmann, L. Audigé, G. Schüpfer, and J. Metzger. Investigating the effect of intra-operative infiltration with local anaesthesia on the development of chronic postoperative pain after inguinal hernia repair: A randomized placebo controlled triple blinded and group sequential study design. *BMC Surgery*, 7(1):22, 2007.

[129] M. L. Huber, A. Laesecke, and D. G. Friend. Correlation for the vapor pressure of mercury. *Industrial & Engineering Chemistry Research*, 45(21):7351–7361, 2006.

[130] K. Imai and K. Nakachi. Cross sectional study of effects of drinking green tea on cardiovascular and liver diseases. *British Medical Journal*, 310(6981):693–696, 1995.

[131] J. P. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

[132] J. P. Ioannidis. Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648, 2008.

[133] S. Iyengar and J. B. Greenhouse. Selection models and the file drawer problem. *Statistical Science*, 3(1):109–117, 1988.

[134] D. Jaeschke. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *The Annals of Statistics*, 7(1):108–115, 1979.

[135] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning*, volume 112. Springer, 2013.

[136] R. I. Jennrich. Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643, 1969.

[137] L. K. John, G. Loewenstein, and D. Prelec. Measuring

the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532, 2012.

[138] Z. N. Kain, F. Sevarino, G. M. Alexander, S. Pincus, and L. C. Mayes. Preoperative anxiety and postoperative pain in women undergoing hysterectomy: A repeated-measures design. *Journal of Psychosomatic Research*, 49(6):417–422, 2000.

[139] T. J. Kaptchuk and F. G. Miller. Placebo effects in medicine. *New England Journal of Medicine*, 373(1):8–9, 2015.

[140] A. S. Keller, B. Rosenfeld, C. Trinh-Shevrin, C. Meserve, E. Sachs, J. A. Leviss, E. Singer, H. Smith, J. Wilkinson, and G. Kim. Mental health of detained asylum seekers. *The Lancet*, 362(9397):1721–1723, 2003.

[141] H. Kesten. *Percolation Theory for Mathematicians*. Springer, 1982.

[142] A. J. Kinderman and J. F. Monahan. Computer generation of random variables using the ratio of uniform deviates. *ACM Transactions on Mathematical Software*, 3(3):257–260, 1977.

[143] A. Kolmogorov. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, 4:89–91, 1933.

[144] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari. A space-time permutation scan statistic for disease outbreak detection. *PLoS Medicine*, 2(3):e59, 2005.

[145] M. Kulldorff and N. Nagarwalla. Spatial disease clusters: detection and inference. *Statistics in Medicine*, 14(8):799–810, 1995.

[146] M. Lee. Passphrases that you can memorize — but that even the NSA can't guess. *The Intercept*, March 26th 2015.

[147] E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. Springer, 2005.

[148] S. D. Levitt and J. A. List. Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1):1–18, 2009.

[149] Z. Li, S. Nadon, and J. Cihlar. Satellite-based detection of canadian boreal forest fires: Development and application of the algorithm. *International Journal of Remote Sensing*, 21(16):3057–3069, 2000.

[150] S. Lieberson and F. B. Lynn. Barking up the wrong branch: Scientific alternatives to the current model of sociological science. *Annual Review of Sociology*, 28(1):1–19, 2002.

[151] T. Liptak. On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3:171–197, 1958.

[152] S. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[153] C. R. Loader. Local likelihood density estimation. *The Annals of Statistics*, 24(4):1602–1618, 1996.

[154] N. K. Logothetis. What we can do and what we cannot do with fMRI. *Nature*, 453(7197):869, 2008.

[155] G. L. Lu-Yao and S.-L. Yao. Population-based study

of long-term survival in patients with clinically localised prostate cancer. *The Lancet*, 349(9056):906–910, 1997.

[156] A. M. Lucas and I. M. Mbiti. Effects of school quality on student achievement: Discontinuity evidence from kenya. *American Economic Journal: Applied Economics*, 6(3):234–63, 2014.

[157] M. Lustig, D. Donoho, and J. M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magnetic Resonance in Medicine*, 58(6):1182–1195, 2007.

[158] R. H. Lustig, L. A. Schmidt, and C. D. Brindis. Public health: The toxic truth about sugar. *Nature*, 482(7383):27, 2012.

[159] Z. Ma, C. Wood, and D. Bransby. Soil management impacts on soil carbon sequestration by switchgrass. *Biomass and Bioenergy*, 18(6):469–477, 2000.

[160] A. Maercker, T. Zöllner, H. Menning, S. Rabe, and A. Karl. Dresden PTSD treatment study: Randomized controlled trial of motor vehicle accident survivors. *BMC Psychiatry*, 6(1):29, 2006.

[161] B. F. Manly. A note on the analysis of species co-occurrences. *Ecology*, 76(4):1109–1115, 1995.

[162] D. Marshall, O. Johnell, and H. Wedel. Meta-analysis of how well measures of bone mineral density predict occurrence of osteoporotic fractures. *British Medical Journal*, 312(7041):1254–1259, 1996.

[163] N. Matosin, E. Frank, M. Engel, J. S. Lum, and K. A. Newell. Negativity towards negative results: A discussion of the disconnect between scientific worth and scientific culture. *Disease Models & Mechanisms*, 7(2):171, 2014.

[164] M. Matsumoto and T. Nishimura. Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8(1):3–30, 1998.

[165] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[166] G. Mendel. Versuche über pflanzen-hybriden (experiments in plant hybridization). In *Verhandlungen des Naturforschenden Vereines in Brünn (Natural History Society of Brünn)*, pages 3–47, 1866.

[167] D. Michaels. *Doubt is their Product: How Industry's Assault on Science Threatens your Health*. Oxford University Press, 2008.

[168] I. Miklós and J. Podani. Randomization of presence-absence matrices: Comments and new algorithms. *Ecology*, 85(1):86–92, 2004.

[169] R. G. Miller. *Simultaneous Statistical Inference*. Springer, 1981.

[170] A. Moscovich, B. Nadler, and C. Spiegelman. On the exact Berk–Jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354, 2016.

[171] F. Mosteller. Note on an application of runs to quality control charts. *The Annals of Mathematical Statistics*, 12(2):228–232, 1941.

[172] M. G. Myers, M. Godwin, M. Dawes, A. Kiss, S. W. Tobe, F. C. Grant, and J. Kaczorowski. Conventional versus

automated measurement of blood pressure in primary care patients with systolic hypertension: randomised parallel design controlled trial. *British Medical Journal*, 342:d286, 2011.

[173] B. Nalebuff. Puzzles: The other person's envelope is always greener. *The Journal of Economic Perspectives*, 3(1):171–181, 1989.

[174] J. I. Naus. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60(310):532–538, 1965.

[175] R. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17(8):857, 1998.

[176] D. C. Norvell and P. Cummings. Association of helmet use with death in motorcycle crashes: a matched-pair cohort study. *American Journal of Epidemiology*, 156(5):483–487, 2002.

[177] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. Bowman, S. Breckler, S. Buck, C. D. Chambers, G. Chin, and G. Christensen. Promoting an open research culture. *Science*, 348(6242):1422–1425, 2015.

[178] G. W. Oehlert. *A First Course in Design and Analysis of Experiments*. 2010. Available online on the author's website.

[179] N. Oreskes and E. M. Conway. *Merchants of Doubt*. New York: Bloomsbury Press, 2010.

[180] K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

[181] K. Pearson and A. Lee. On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika*, 2(4):357–462, 1903.

[182] R. D. Peng. Reproducible research in computational science. *Science*, 334(6060):1226–1227, 2011.

[183] C. D. Pilcher, B. Louie, S. Facente, S. Keating, J. Hackett Jr, A. Vallari, C. Hall, T. Dowling, M. P. Busch, and J. D. Klausner. Performance of rapid point-of-care and laboratory tests for acute and established HIV infection in San Francisco. *PloS ONE*, 8(12):e80629, 2013.

[184] D. E. Powers and D. A. Rock. Effects of coaching on sat i: Reasoning test scores. *Journal of Educational Measurement*, 36(2):93–118, 1999.

[185] V. K. Prasad and A. S. Cifu. *Ending Medical Reversal: Improving Outcomes, Saving Lives*. Johns Hopkins University Press, 2015.

[186] D. D. S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

[187] F. Prinz, T. Schlange, and K. Asadullah. Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug discovery*, 10(9):712–712, 2011.

[188] M. L. Radelet and G. L. Pierce. Choosing those who will die: Race and the death penalty in Florida. *Florida Law*

Review, 43:1, 1991.

[189] L. Rahmathullah, J. M. Tielsch, R. Thulasiraj, J. Katz, C. Coles, S. Devi, R. John, K. Prakash, A. Sadanand, and N. Edwin. Impact of supplementing newborn infants with vitamin A on early infant mortality: Community based randomised trial in southern India. *British Medical Journal*, 327(7409):254, 2003.

[190] J. Randi. *Flim-Flam!: Psychics, ESP, Unicorns, and other Delusions*, volume 342. Prometheus Books Buffalo, NY, 1982.

[191] J. Randi. *The Truth about Uri Geller*. Pyr Books, 1982.

[192] G. Rash. Probabilistic models for some intelligence and attainment tests. *Copenhagen: Danish Institute for Educational Research*, 1960.

[193] J. J. Reilly, J. Armstrong, A. R. Dorosty, P. M. Emmett, A. Ness, I. Rogers, C. Steer, and A. Sherriff. Early life risk factors for obesity in childhood: cohort study. *British Medical Journal*, 330(7504):1357, 2005.

[194] A. Reinhart. *Statistics Done Wrong: The Woefully Complete Guide*. No Starch Press, 2015. Available online at https://www.statisticsdonewrong.com.

[195] A. Rosengren, L. Wilhelmsen, E. Eriksson, B. Risberg, and H. Wedel. Lipoprotein(a) and coronary heart disease: A prospective case-control study in a general population sample of middle aged men. *British Medical Journal*, 301(6763):1248–1251, 1990.

[196] J. Rosenhouse. *The Monty Hall Problem: The Remarkable Story of Math's Most Contentious Brain Teaser*. Oxford University Press, 2009.

[197] R. Rosenthal. The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638, 1979.

[198] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[199] M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, 9(2):65–78, 1982.

[200] A. Rukhin, J. Soto, J. Nechvatal, E. Barker, S. Leigh, M. Levenson, D. Banks, A. Heckert, J. Dray, and S. Vo. Statistical test suite for random and pseudorandom number generators for cryptographic applications, NIST special publication. Technical Report (SP 800-22), National Institute of Standards and Technology (NIST), 2010. Revised by L. Bassham III.

[201] E. Rutherford, H. Geiger, and H. Bateman. LXXVI. The probability variations in the distribution of α particles. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 20(118):698–707, 1910.

[202] M. Sageman. *Misunderstanding Terrorism*. University of Pennsylvania Press, 2016.

[203] E. Schlosser. Break-In at Y-12: How a handful of pacifists and nuns exposed the vulnerability of America's nuclear-weapons sites. *The New Yorker*, March 9th 2015.

[204] K. F. Schulz, D. G. Altman, and D. Moher. Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, 8(1):18, 2010.

[205] D. O. Sears. College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, 51(3):515, 1986.

[206] H. C. Selvin and A. Stuart. Data-dredging procedures in survey analysis. *The American Statistician*, 20(3):20–23, 1966.

[207] A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, 3(3):985–1012, 2009.

[208] G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

[209] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[210] R. Sihvonen, M. Paavola, A. Malmivaara, A. Itälä, A. Joukainen, H. Nurmi, J. Kalske, and T. L. Järvinen. Arthroscopic partial meniscectomy versus sham surgery for a degenerative meniscal tear. *New England Journal of Medicine*, 369(26):2515–2524, 2013.

[211] R. J. Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[212] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.

[213] N. Smirnov. Table for estimating the goodness of fit of empirical distributions. *The Annals of Mathematical Statistics*, 19(2):279–281, 1948.

[214] G. D. Smith and S. Ebrahim. Data dredging, bias, or confounding: They can all get you into the BMJ and the Friday papers. *British Medical Journal*, 325(7378):1437, 2002.

[215] R. L. Smith. A statistical assessment of Buchanan's vote in Palm Beach county. *Statistical Science*, pages 441–457, 2002.

[216] Y. M. Smulders. A two-step manuscript submission process can reduce publication bias. *Journal of Clinical Epidemiology*, 66(9):946–947, 2013.

[217] J. Snow. *On the Mode of Communication of Cholera*. John Churchill, 1855.

[218] B. Sorić. Statistical "discoveries" and effect-size estimation. *Journal of the American Statistical Association*, 84(406):608–610, 1989.

[219] L. K. Southworth, S. K. Kim, and A. B. Owen. Properties of balanced permutations. *Journal of Computational Biology*, 16(4):625–638, 2009.

[220] J. Splawa-Neyman, D. M. Dabrowska, and T. Speed. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–472, 1990.

[221] P. Squire. Why the 1936 literary digest poll failed. *Public Opinion Quarterly*, 52(1):125–133, 1988.

[222] R. S. Stafford, T. H. Wagner, and P. W. Lavori. New, but

not improved? incorporating comparative-effectiveness information into fda labeling. *New England Journal of Medicine*, 361(13):1230–1233, 2009.

[223] T. Stockwell, J. Zhao, S. Panwar, A. Roemer, T. Naimi, and T. Chikritzhs. Do "moderate" drinkers have reduced mortality risk? a systematic review and meta-analysis of alcohol consumption and all-cause mortality. *Journal of Studies on Alcohol and Drugs*, 77(2):185–198, 2016.

[224] V. Stodden. Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2:1–19, 2015.

[225] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[226] Student. On the error of counting with a haemacytometer. *Biometrika*, pages 351–360, 1907.

[227] L. P. Svetkey, V. J. Stevens, P. J. Brantley, L. J. Appel, J. F. Hollis, C. M. Loria, W. M. Vollmer, C. M. Gullion, K. Funk, and P. Smith. Comparison of strategies for sustaining weight loss: The weight loss maintenance randomized controlled trial. *Journal of the American Medical Association*, 299(10):1139–1148, 2008.

[228] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, 143(8):1249–1272, 2013.

[229] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007.

[230] W. C. Thompson and E. L. Schumann. Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. *Law and Human Behavior*, 11(3):167, 1987.

[231] S. Thornhill, G. M. Teasdale, G. D. Murray, J. McEwen, C. W. Roy, and K. I. Penny. Disability in young people and adults one year after head injury: prospective cohort study. *British Medical Journal*, 320(7250):1631–1635, 2000.

[232] J. Tierney. Behind monty hall's doors: Puzzle, debate and answer? *The New York Times*, 1991.

[233] J. A. Tropp. Book review of *a mathematical introduction to compressive sensing*. *Bulletin of the American Mathematical Society*, 54(1):151–165, 2017.

[234] E. H. Turner, A. M. Matthews, E. Linardatos, R. A. Tell, and R. Rosenthal. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, 358(3):252–260, 2008.

[235] J. Utts. Replication and meta-analysis in parapsychology. *Statistical Science*, pages 363–378, 1991.

[236] J. M. Utts. *Seeing through Statistics*. Brooks/Cole, 3rd edition, 2005.

[237] A. Vazsonyi. Which door has the cadillac. *Decision Line*, 30(1):17–19, 1999.

[238] R. Von Mises. *Probability, Statistics, and Truth*. Courier Corporation, 1957.

[239] V. Vovk, A. Gammerman, and G. Shafer. *Conformal prediction*. Springer, 2005.

[240] V. G. Vovk and G. Shafer. Kolmogorov's contributions to

the foundations of probability. *Problems of Information Transmission*, 39(1):21–31, 2003.

[241] S. Wacholder, S. Chanock, M. Garcia-Closas, L. El Ghormli, and N. Rothman. Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, 96(6):434–442, 2004.

[242] A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

[243] A. Wald and J. Wolfowitz. On a test whether two samples are from the same population. *The Annals of Mathematical Statistics*, 11(2):147–162, 1940.

[244] B. E. Wampold, T. Minami, S. C. Tierney, T. W. Baskin, and K. S. Bhati. The placebo is powerful: Estimating placebo effects in medicine and psychotherapy from randomized clinical trials. *Journal of Clinical Psychology*, 61(7):835–854, 2005.

[245] J. N. Wand, K. W. Shotts, J. S. Sekhon, W. R. Mebane, M. C. Herron, and H. E. Brady. The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida. *American Political Science Review*, 95(4):793–810, 2001.

[246] W. Wang, D. Rothschild, S. Goel, and A. Gelman. Forecasting elections with non-representative polls. *International Journal of Forecasting*, 31(3):980–991, 2015.

[247] L. Wasserman. *All of Nonparametric Statistics*. Springer, New York, 2006.

[248] G. Watterson. Heterosis or neutrality? *Genetics*, 85(4):789–814, 1977.

[249] L. Weiss. A test of fit based on the largest sample spacing. *Journal of the Society for Industrial and Applied Mathematics*, 8(2):295–299, 1960.

[250] B. Welch. On the comparison of several mean values: An alternative approach. *Biometrika*, 38(3/4):330–336, 1951.

[251] B. L. Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1-2):28–35, 1947.

[252] S. Wicksell. The corpuscle problem. a mathematical study of a biometric problem. *Biometrika*, 17(1-2):84–99, 1925.

[253] R. J. Williams, T. Tse, W. R. Harlan, and D. A. Zarin. Registration of observational studies: Is it time? *Canadian Medical Association Journal*, 182(15):1638–1642, 2010.

[254] E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[255] S. S. Young and A. Karr. Deming, data and observational studies. *Significance*, 8(3):116–120, 2011.

[256] E. C. Yu, A. M. Sprenger, R. P. Thomas, and M. R. Dougherty. When decision heuristics and science collide. *Psychonomic Bulletin & Review*, 21(2):268–282, 2014.

[257] D. Zak. *Almighty: Courage, Resistance, and Existential Peril in the Nuclear Age*. Penguin Publishing Group, 2017.

[258] A. Zaman and D. Simberloff. Random binary matrices in biogeographical ecology: Instituting a good neighbor

policy. *Environmental and Ecological Statistics*, 9(4):405–421, 2002.

INDEX