9 January 2004
revised 18 January 2004

INTRODUCTION TO MATHEMATICAL MODELLING LECTURES 3-4: BASIC PROBABILITY THEORY

David A. Meyer

Project in Geometry and Physics, Department of Mathematics
University of California/San Diego, La Jolla, CA 92093-0112
http://math.ucsd.edu/~dmeyer/; dmeyer@math.ucsd.edu

Example

Suppose we observe a gambler enter a casino with \$100 in his/her pocket, and then leave a few hours later with \$87. This is a situation we might want to model, particularly if we are thinking about entering the casino ourselves. Without any additional information, *i.e.*, any additional data, there is little more that we can predict than if someone else goes into the casino with \$100 and leaves after the same amount of time, s/he will also have only \$87. This is certainly a case in which we must adjust our data collection.

So suppose we enter the casino and find that the gambler is repeatedly playing a very simple game (this is not intended to be realistic): s/he flips a coin, winning a dollar if it comes up heads, and losing a dollar if it comes up tails. After playing 100 times, the gambler leaves the casino. This gives us much more information with which to build our model, although perhaps not as much as we might like.

Probabilistic models

In principle (to the extent that physics is classical), if we could measure exactly how the coin is being flipped, exactly how the coin is shaped and weighted, exactly the gravitational acceleration, exactly how the coin bounces when it hits the table, exactly how the air currents are blowing on the coin while it is in the air, *etc.*, we could do a complicated physics calculation and determine whether the coin will land head up or tail up. In practice, of course, we cannot know most of these details, so we summarize our ignorance by saying that the coin comes up heads *with probability* p, *i.e.*, a fraction p of the time, where $0 \leq p \leq 1$. That is, probability is an accounting for the effects of parts of the world that we are not trying to model (*exogenous* variables) and about which we have only limited knowledge. There are always such effects in any real situation, which is why we are starting this course with a discussion of them, rather than pretending that we can

usually make complete mathematical models and then only discussing probabilistic effects at the end of the course. Most of the models that we discuss will not make deterministic predictions that something will certainly happen, but rather make probabilistic predictions that different outcomes will occur different fractions of the time.

The binomial distribution

If the gambler only plays once, it is easy to make a prediction:

outcome	payoff	probability
H	\$1	p
T	−\$1	$1-p$

A fraction p of the time s/he will leave the casino with \$101, and a fraction $1-p$ of the time s/he will leave the casino with \$99. Playing twice is not much harder to understand:

outcome	payoff	probability
HH	\$2	p^2
HT	\$0	$p(1-p)$
TH	\$0	$(1-p)p$
TT	−\$2	$(1-p)^2$

So a fraction p^2 of the time the gambler will leave with \$102, a fraction $2p(1-p)$ of the time with \$100, and a fraction $(1-p)^2$ of the time with \$98. Finally, if the gambler plays three times we compute:

outcome	payoff	probability
HHH	\$3	p^3
HHT	\$1	$p^2(1-p)$
HTH	\$1	$p^2(1-p)$
THH	\$1	$p^2(1-p)$
TTH	−\$1	$p(1-p)^2$
THT	−\$1	$p(1-p)^2$
HTT	−\$1	$p(1-p)^2$
TTT	−\$3	$(1-p)^3$

Now there are four results: \$103, \$101, \$99 and \$97, which we predict to occur with probabilities p^3, $3p^2(1-p)$, $3p(1-p)^2$ and $(1-p)^3$, respectively.

It would be extremely tedious to analyze the case of 100 coin flips like this. Fortunately, we can be cleverer. Notice that the payoffs depend only on how many heads there are, and that for a given total number of flips, n, the probability of a specific outcome with w heads is the same, $p^w(1-p)^{n-w}$, no matter when the heads appear. (We have made an assumption here, that the outcomes of different coin flips are independent, which we will discuss later.) To figure out how many outcomes there are with w heads, imagine that the coins are labelled from 1 to n. We can arrange them in $n(n-1)(n-2)\cdots 3\cdot 2\cdot 1 = n!$ different orders, by picking any of n for the first, any of the remaining $n-1$ for the second, *etc.* Not all of these correspond to different outcomes, however, since any ordering with the w heads in the same positions is the same outcome, no matter in what order the labels on the w heads are arranged. But these w labels can be arranged in $w!$ orders by the same argument. Similarly, the $n-w$ labels on the tail up coins can be arranged in $(n-w)!$ ways. So the total number of different outcomes with w heads is

$$\frac{\text{no. orders of } n \text{ coins}}{(\text{no. orders of } w \text{ heads})(\text{no. orders of } n-w \text{ tails})} = \frac{n!}{w!(n-w)!} = \binom{n}{w},$$

where the last symbol is pronounced "n choose w", and is called a *binomial coefficient*. Multiplying the number of different outcomes with w heads by the probability of a specific outcome with w heads gives the probability that the gambler will win w times out of n:

$$\binom{n}{w} p^w (1-p)^{n-w}.$$

Random variables

This is an example of a *probability function*: We say that the number of heads, W, is a *random variable* and the *probability* that $W = w$,

$$\text{prob}(W = w) = \binom{n}{w} p^w (1-p)^{n-w}.$$

For any probability function, if we add the probabilities of every possible outcome we must get 1; in this case:

$$1 = \sum_{w=0}^{n} \text{prob}(W = w) = \sum_{w=0}^{n} \binom{n}{w} p^w (1-p)^{n-w}. \tag{3.1}$$

Homework: Read Larsen & Marx [1], p. 135–136.
Show algebraically that the sum on the right of eq. 3.1 equals 1.

The *expectation value* of a random variable, $E[W]$, is the probability weighted average of its possible values; in this case:

$$\begin{aligned}
E[W] &= \sum_{w=0}^{n} w\,\mathrm{prob}(W=w) \\
&= \sum_{w=0}^{n} w\binom{n}{w}p^w(1-p)^{n-w} \\
&= \sum_{w=0}^{n} w\frac{n!}{w!(n-w)!}p^w(1-p)^{n-w} \\
&= \sum_{w=1}^{n} w\frac{n!}{w!(n-w)!}p^w(1-p)^{n-w} \quad \text{(since the } w=0 \text{ term in the sum is 0)} \\
&= \sum_{w=1}^{n} \frac{n!}{(w-1)!(n-w)!}p^w(1-p)^{n-w} \\
&= np\sum_{w=1}^{n} \frac{(n-1)!}{(w-1)!\big((n-1)-(w-1)\big)!}p^{w-1}(1-p)^{(n-1)-(w-1)} \\
&\qquad\qquad \text{(since } (n-1)-(w-1)=n-w\text{)} \\
&= np\sum_{v=0}^{n-1} \frac{(n-1)!}{v!\big((n-1)-v\big)!}p^v(1-p)^{(n-1)-v} \qquad \text{(letting } v=w-1\text{)} \\
&= np, \qquad \text{(using eq. 3.1 with } n \text{ replaced by } n-1\text{)}
\end{aligned}$$

which is what you most likely expected. That is, for a fair coin ($p=\frac{1}{2}$), the expected number of times the gambler who plays 100 times will win is 50, so his/her expected payoff is \$50 − \$50 = \$0. Of course, this does not happen every time; there is some variation in the outcomes.

The *variance* of a random variable, $Var[W]$, is the probability weighted average of the squared differences of its possible values from $E[W]$:

$$Var[W] = \sum_{w}(w-E[W])^2\,\mathrm{prob}(W=w).$$

Evaluating the variance of the binomial distribution is substantially more complicated than evaluating the expectation value. But we can avoid doing the algebra by learning a little more about basic probability theory, which will also show us how to compute the expectation value much more easily than the calculation above.

Sums of random variables

Notice that $W = X_1 + \cdots + X_n$, where X_i is a random variable that can take two values:

$$X_i = \begin{cases} 1 & \text{if the } i^{\text{th}} \text{ coin flipped lands head up;} \\ 0 & \text{if the } i^{\text{th}} \text{ coin flipped lands tail up.} \end{cases}$$

That is, the number of wins can be computed by adding 1 for each coin that lands head up, and 0 for each coin that lands tail up. It is easy to calculate the expectation value of X_i. From the definition as the probability-weighted sum of the possible outcomes we have:

$$E[X_i] = p \cdot 1 + (1-p) \cdot 0 = p.$$

Now consider the case $n = 2$, *i.e.*, $W = X_1 + X_2$.

$$\begin{aligned} E[X_1 + X_2] &= \sum_{x_1,x_2} (x_1 + x_2)\text{prob}(X_1 = x_1 \wedge X_2 = x_2) \\ &= \sum_{x_1,x_2} x_1 \text{prob}(X_1 = x_1 \wedge X_2 = x_2) + \sum_{x_1,x_2} x_2 \text{prob}(X_1 = x_1 \wedge X_2 = x_2) \\ &= \sum_{x_1} x_1 \sum_{x_2} \text{prob}(X_1 = x_1 \wedge X_2 = x_2) + \sum_{x_2} x_2 \sum_{x_1} \text{prob}(X_1 = x_1 \wedge X_2 = x_2) \\ &= \sum_{x_1} x_1 \text{prob}(X_1 = x_1) + \sum_{x_2} x_2 \text{prob}(X_2 = x_2) \\ &= E[X_1] + E[X_2], \end{aligned}$$

where $x_i \in \{0, 1\}$ and $\wedge$ means 'and'. The penultimate equality follows from the fact that for any random variables X and Y, $\text{prob}(X = x) = \sum_y \text{prob}(X = x \wedge Y = y)$. Thus, in general, the expectation value of the sum of random variables is the sum of their expectation values. In particular,

$$E[W] = E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] = np,$$

which is what we computed previously, with considerably more effort.

Homework: Use mathematical induction to prove the middle equality in this equation, for all $n \in \mathbb{N}$.

So we have showed that expectation values of random variables add. We might ask, "Do they multiply?". The answer is, "Sometimes.". They do if an important property holds: Two random variables X and Y are *independent* if and only if

$$\text{prob}(X = x \wedge Y = y) = \text{prob}(X = x)\text{prob}(Y = y).$$

In this case we can compute the expectation value of their product:

$$\begin{aligned} E[XY] &= \sum_{x,y} xy \,\text{prob}(X = x \wedge Y = y) \\ &= \sum_{x,y} xy \,\text{prob}(X = x)\text{prob}(Y = y) \\ &= \sum_{x} x \,\text{prob}(X = x) \sum_{y} y \,\text{prob}(Y = y) \\ &= E[X]E[Y]. \end{aligned}$$

We use this fact to study the variance of a sum of random variables:

$$\begin{aligned}
Var[X+Y] &= \sum_{x,y}(x+y-E[X]-E[Y])^2\,\mathrm{prob}(X=x\wedge Y=y)\\
&= \sum_{x,y}\big((x-E[X])+(y-E[Y])\big)^2\,\mathrm{prob}(X=x\wedge Y=y)\\
&= \sum_{x,y}\big((x-E[X])^2+2(x-E[X])(y-E[Y])+(y-E[Y])^2\big)\cdot\\
&\qquad\cdot\,\mathrm{prob}(X=x\wedge Y=y)\\
&= \sum_{x,y}(x-E[X])^2\,\mathrm{prob}(X=x\wedge Y=y)\\
&\qquad+2\sum_{x,y}(x-E[X])(y-E[Y])\,\mathrm{prob}(X=x\wedge Y=y)\\
&\qquad+\sum_{x,y}(y-E[Y])^2\,\mathrm{prob}(X=x\wedge Y=y)\\
&= \sum_{x}(x-E[X])^2\,\mathrm{prob}(X=x)\\
&\qquad+2\sum_{x,y}(x-E[X])(y-E[Y])\,\mathrm{prob}(X=x)\mathrm{prob}(Y=y)\\
&\qquad+\sum_{y}(y-E[Y])^2\,\mathrm{prob}(Y=y)\\
&= Var[X]+Var[Y]\\
&\qquad+2\sum_{x}(x-E[X])\,\mathrm{prob}(X=x)\sum_{y}(y-E[Y])\,\mathrm{prob}(Y=y)\\
&= Var[X]+Var[Y],
\end{aligned}$$

where the last equality follows from

$$\begin{aligned}
\sum_x(x-E[X])\mathrm{prob}(X=x) &= \sum_x x\,\mathrm{prob}(X=x)-E[X]\sum_x\mathrm{prob}(X=x)\\
&= E[X]-E[X]\cdot 1\\
&= 0.
\end{aligned}$$

For the coin flipping game, the outcomes of different coin flips are assumed to be independent, and

$$Var[X_i] = p\cdot(1-p)^2+(1-p)\cdot(0-p)^2 = p(1-p),$$

so

$$Var[W] = Var[X_1]+\cdots+Var[X_n] = np(1-p).$$

Now that we have computed the expectation value and the variance of the binomial distribution, we can investigate how close it is to a normal distribution with the same mean and variance, just as we did with the height data in Lecture 2 [2]. Figure 3.1 shows the results for two binomial distributions with $n = 20$. The central one (in red) has $p = 1/2$ and is very well approximated by the normal distribution with the same mean and variance (in blue). The left one (in green) has $p = 1/4$ and is slightly less well approximated by the corresponding normal distribution (in blue). But the observation that each is well approximated by a normal distribution is correct, and is a consequence of the following theorem.

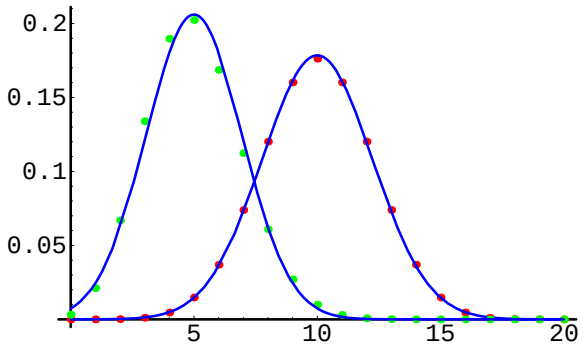

Figure 3.1. Binomial distributions for $n = 20$ and $p = 1/2$ (central distribution, red), $p = 1/4$ (left distribution, green), and normal distributions with the same mean and variance, respectively.

DeMoivre-Laplace Limit Theorem. *Let W be a binomial random variable describing the number of successes in n trials, each of which succeeds with probability p. For all $a \leq b \in \mathbb{R}$,*

$$\lim_{n\to\infty} \text{prob}\Big(a \leq \frac{W - np}{\sqrt{np(1-p)}} \leq b\Big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-w^2/2} \mathrm{d}w.$$

We could equally well write this as:

$$\lim_{n\to\infty} \text{prob}(a \leq W \leq b) = \frac{1}{\sqrt{2\pi np(1-p)}} \int_a^b e^{-(w-np)^2/2np(1-p)} \mathrm{d}w.$$

Both statements say that the area under either of the blue curves in Figure 3.1, between $w = a$ and $w = b$, is approximately equal to the sum of the probabilities at the points with $a \leq w \leq b$ for the corresponding binomial probability function. The approximation is better for larger n, and for p further from 0 or 1.

Remember that we used the formula $W = X_1 + \cdots + X_n$ to calculate the expectation value and variance of W. We can imagine a random variable that is the sum $Y_1 + \cdots + Y_n$ for a sequence of random variables Y_i that are not simply binary valued like X_i. Even for this more general situation, the same result holds:

Central Limit Theorem. *Let $Y_1, Y_2, \ldots$ be an infinite sequence of independent random variables, each having the same distribution, with $E[Y_i] = \mu < \infty$ and $\text{Var}[Y_i] = \sigma^2 < \infty$. For all $a \leq b \in \mathbb{R}$,*

$$\lim_{n\to\infty} \text{prob}\Big(a \leq \frac{Y_1 + \cdots + Y_n - n\mu}{\sqrt{n}\sigma} \leq b\Big) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-w^2/2} \mathrm{d}w.$$

For proofs of these theorems, see any standard probability/statistics text, *e.g.*, Larsen & Marx [1]. The Central Limit Theorem is most interesting for us because it suggests a characteristic of mathematical models that would produce a normal distribution for some variable—the model likely includes several or many independent variables with similar distributions that are summed to give the variable with the normal distribution. In the case of human height, the data we have already seen indicate that male/female is an important factor, and we would expect that nutrition is also important. The observation that the distribution for each sex is separately approximately normal suggests further that there might be a genetic model that involves the action of multiple genes, each of which can contribute to larger or smaller height. This is in contrast to the famous pea plants originally studied by Mendel [3], which are only tall or short, *i.e.*, have a binary distribution, not a normal one. From what we have learned in this lecture, we would expect pea plant height to be controlled by only one gene; this is true, and the action of that gene (*Le*) is now understood [4].

References

[1] R. J. Larsen and M. L. Marx, *An Introduction to Mathematical Statistics and Its Applications* (Upper Saddle River, NJ: Prentice Hall 2001).

[2] D. A. Meyer, "Introduction to Mathematical Modelling: Data collection", http://math.ucsd.edu/~dmeyer/teaching/111winter04/IMM040107.pdf.

[3] G. Mendel, (1866). "*Versuche über Pflanzen-hybriden*", *Verh. Natforsch. Ver. Brünn* **4** (1866) 3–47; http://www.mendelweb.org/MWGerText.html;
English transl. by C. T. Druery and W. Bateson, "Experiments in plant hybridization", http://www.mendelweb.org/Mendel.html.

[4] D. R. Lester, J. J. Ross, P. J. Davies and J. B. Reid, "Mendel's stem length gene (*Le*) encodes a gibberellin 3β-hydroxylase", *Plant Cell* **9** (1997) 1435–1443.