7 October 2013

INTRODUCTION TO MATHEMATICAL MODELLING
NORMAL DISTRIBUTIONS

David A. Meyer

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

Summary statistics

To describe a set of real-valued data, $\{h_i\}$, we use *summary statistics*: a smaller set of numbers that somehow characterizes the data. The most basic is just the size of the set, *i.e.*, the number of data points, N. This describes the sample size, but tells us nothing about the distribution of values. The most familiar statistic is the *average* (or *mean*):

$$\bar{h} = \frac{1}{N}\sum_{i=1}^{N} h_i,$$

where h_i are the observed values—heights in the following example. Another important feature of the data is how diverse it is: Are all the values exactly the average, or are they spread out around it? A statistic that describes this feature of the data is the *sample variance*:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(h_i - \bar{h})^2,$$

which, despite the division by $N-1$ rather than N, you should think of as the average of the square of the deviation of the data points from the mean. It is also useful to have a measure of the spread that has the same units as the data, which we get by taking the square root of the sample variance. This is called the *standard deviation*:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(h_i - \bar{h})^2}.$$

Example

Figure 1 shows data from the 1960–1962 U.S. National Health Examination Survey (NHES) [1], with population fraction plotted as a function of height (in inches), for a sample of 746 men and 675 women, ages 25–34. (The data for men and women have been weighted to reflect their relative numbers in the general population.)

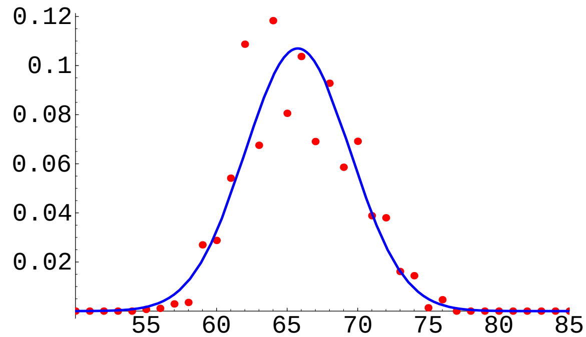

Figure 1. Population fraction by height from the 1960–1962 Health Survey.

Figure 1 shows a peculiar alternation of higher and lower values at successive numbers of inches in the NHES data.

Homework: Explain this observation. [Hint: The heights were measured in centimeters.]

Neglecting this observation, which is clearly a systematic error in the way the data were collected/reported, the smooth blue curve shown in Figure 1 seems like a reasonable approximation to the data. This curve is a graph of the *normal distribution* with mean μ and variance σ^2 (numerical values computed from the data):

$$f(h) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(h-\mu)^2/2\sigma^2}.$$

This function has the property that it is symmetric about its average, μ, takes its maximum value there, and decreases rapidly toward 0 for larger and smaller arguments. $f(h)$ is a *probability density*, which means that it takes non-negative values and integrates to 1. The integral

$$\int_a^b f(h)\mathrm{d}h$$

is the fraction of the area under the curve that lies between $h = a$ and $h = b$, and is thus approximately the fraction of the population with heights between a and b.

We can improve our understanding of the data by separating out the data for men and for women. Figure 2 shows the fractions of men and women at each height, in red and green, respectively. The two normal distributions, with mean and variance computed from each set of data, are shown as the blue curves; notice that the means and variances are different for the heights of men and women. Finally, in Figure 3 I've again plotted the weighted sum of the two data sets, and the curve that is the weighted sum of the two normal distributions. As long as we neglect the high/low alternation in the data, this curve seems to approximate the data quite well, better than the curve in Figure 1.

Homework: Read Bender, Appendix A.

To what extent have we created a mathematical model for human heights? We have collected heights for many people, then recognized that we should also keep track of which

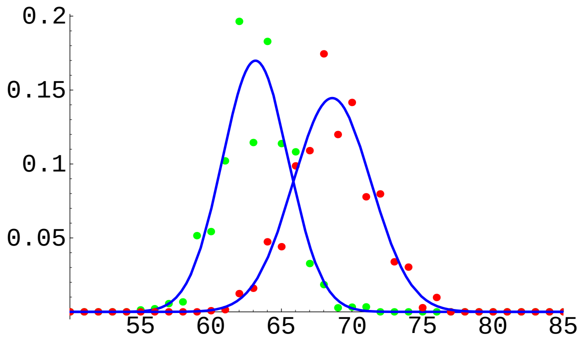

Figure 2. Fraction of men (red) and women (green) as a function of height. The curves are normal distributions.

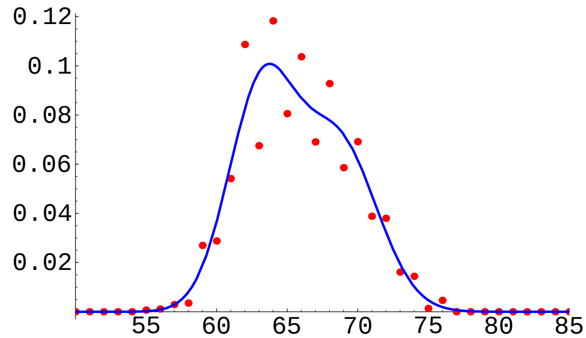

Figure 3. Total population fraction by height with a weighted sum of the normal distributions for men and women.

heights are for men and which are for women, *i.e.*, we adjusted our data collection. Figure 3 summarizes our best description of the distribution of heights. This allows for limited *predictions*: we can say what fraction of the time a randomly selected woman will be shorter than 5 ft, say. But we have no *understanding* of why the distribution of human height is the way it is, other than the difference between men and women. And we have no ability to *control* the distribution; we cannot produce 7 ft tall players for a basketball team, for example. A mathematical model of human heights would have much more to it than a probability density for heights. It would presumably involve some biological facts that we have not yet incorporated.

Reference

[1] H. W. Stoudt, A. Damon, R. McFarland and J. Roberts, "Weight, height, and selected body dimensions of adults: United States 1960–1962", *Vital and Health Statistics*, Public Health Service Publication No. 1000, Series 11, No. 8 (Washington, DC: U.S. Government Printing Office 1965);
http://www.cdc.gov/nchs/data/series/sr_11/sr11_008.pdf.