Chapter 6

Probability and Statistics

Probability is often defined as *long-term relative frequency*. The study of probability addresses questions like this: In the *long term* (meaning: after a certain scenario or "experiment" – for example, flipping a coin – has been repeated many, many times), what fraction of the time (that is, with what *relative frequency*) does a given outcome (for example, the coin landing heads up) result?

The theory of probability can be applied to the study of *statistics*, which may be defined as *the branch of mathematics concerned with the collection, classification, analysis, and interpretation of numerical facts, for the purposes of drawing inferences from their quantifiable probability.*

The big idea here is that natural phenomena are, in general, not completely *deterministic*. That is, they do not evolve according to precisely predictable formulas or recipes. Still, deterministic models like those examined in prior chapters often give good approximations to what happens "in real life." And we can apply statistical analyses to get a sense of *how well* these models reflect reality. And we can thereby quantify, in rigorous ways, statements like "we have this much confidence that this given situation will yield an outcome in the following prescribed range."

In this chapter, we present a sketch of the above ideas, with just enough detail that we can investigate two main (related) tools in statistical inference: *hypothesis testing* and *confidence intervals*. The reader should note, along the way, how our development of these ideas mirrors, and uses, some previously studied concepts – concepts of definite integrals, areas, Riemann sums, and the Fundamental Theorem of Calculus.

6.1 Relative frequency density

Flipping coins; the central limit theorem

Consider the experiment of flipping six coins, and recording the number of heads that come up. The *outcome* of such an experiment will, of course, be one of the integers 0, 1, 2, 3, 4, 5, 6.

A certain Calculus class at the University of Colorado Boulder repeated this experiment 3,444 times. The results of these 3,444 trials of this experiment are summarized in the following "frequency table."

Table 6.1.1. Six coins, flipped 3,444 times

Number of heads	0	1	2	3	4	5	6
Frequency	66	341	825	1048	780	333	51

We can compile this data into a *histogram*, with "frequency" on the vertical axis and "number of heads" on the horizontal. We get a figure like this:

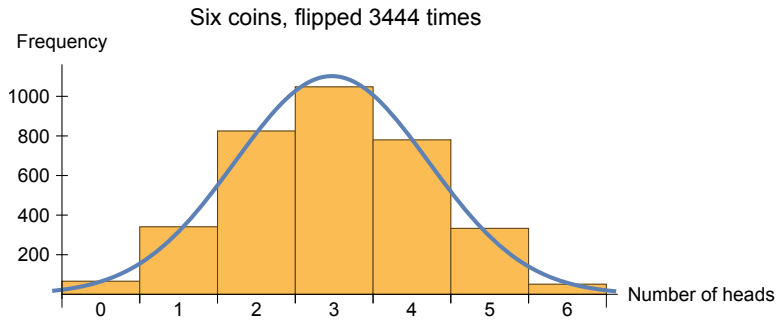

Figure 6.1. Histogram for trials of a coin-flipping experiment

We've superimposed, on the above histogram, a certain bell-shaped curve that approximates the "shape" of the data. The curve we've used is a particular kind of bell-shaped curve, known as a *normal* curve, and the normal curve we've chosen is one that is especially well-suited to the data. We'll explain the meaning of all of this in the next section.

(Note that the bars of our histogram are *contiguous*; there is no space between them. This is a convention that we will always follow, and that will be important to our interpretation, later in this section, of histograms in terms of area.)

The "normal" shape of a coin-flipping experiment becomes even more evident if each trial of the experiment entails a larger number of coins, and if many more trials are performed. For example, if an experiment comprises the flipping of 50 coins, and this experiment is repeated 300,000 times, then the result might look like this:

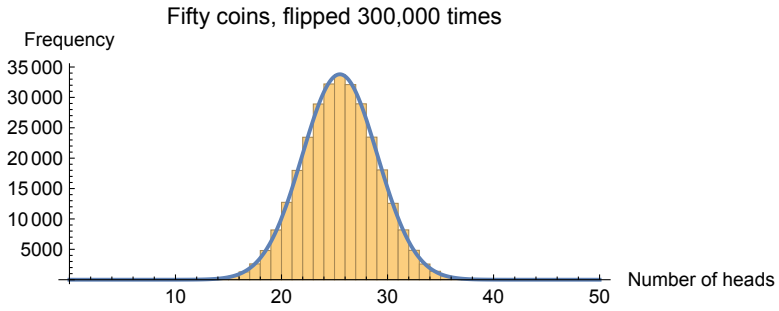

Figure 6.2. Histogram for trials of another coin-flipping experiment

(The data for the above histogram were obtained through a *simulation*. In other words, this data was *not* obtained by actually flipping 50 coins, 300,000 times. Rather, a computer mathematics package was used to pick, at random, 50 numbers, where each number could either be a zero – representing a coin turning up "heads" – or a one – representing "tails." The number of zeroes resulting from this was recorded, and this process was repeated 300,000 times.)

As before, we have fit the histogram with a certain "normal" curve that is particularly well-suited to the data. In this case, the fit is quite close.

The above discussions illustrate a *huge* result in probability, which will be central (pun intended) to our discussions of statistical inference in the next section.

If each trial of an experiment comprises many small, independent factors, and many trials are performed, then (under some mild technical conditions) the outcomes of the experiment will follow a roughly *normal* distrbution.

The Central Limit Theorem

We will not prove this result; proofs may be found in most advanced texts on probability and statistics. We do take a moment, though, to note how this theorem reflects our discussions above. Consider, in particular, the scenario encapsulated by Figure 6.2. There, the 50 coins being tossed are the "many small, independent factors" of the theorem. (They are *independent* in the sense that no one coin affects the behavior of any others. Of course, the coins might bounce against each other on the way down, but we can assume that any effects of this contact cancel each other out, in terms of the probability of any coins coming up heads. Alternatively, we can imagine that the fifty coins are flipped one at a time.) And the 300,000 repetitions are the "many trials" cited in the theorem.

We have not been specific about *how close* to normal one's distribution will be, or the manner in which this might depend on *how many factors* there are, or *how many trials* are performed. Nor will we elaborate on this much. The important idea, for our purposes, is that *more factors* and *more trials* tend to produce distributions that are *more normal*. This idea is exemplified by comparing the scenarios and histograms of Figures 6.1 and 6.2 above. (The importance of having numerous factors and numerous trails may also be appreciated by considering some rather extreme cases. Specifically, imagine flipping a *single* coin, any number of times, or a *huge* number of coins, just once. In neither case will the histogram thus obtained look at all bell-shaped!)

Mean and standard deviation

Ultimately, we wish to draw stronger connections between histograms (like the ones in Figures 6.1 and 6.2 above) and curves that "fit" them (like the ones superimposed on the histograms in the above figures). To this end, we'll need formulas for certain quantities related to the "shape" of a data set.

We begin with the following.

Definition 6.1.1. Let X be a set consisting of n numerical (not necessarily distinct) data points, labeled $x_1, x_2, x_3, \ldots, x_n$. That is,

$$X = \{x_1, x_2, \ldots, x_n\}.$$

Then:

(a) We define the *mean* $\bar{x}$ and *standard deviation* s of X by:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}, \qquad s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n-1}}.$$

(b) Suppose the data points in X take on only m **distinct** values; labeled $y_1, y_2, y_3, \ldots, y_m$. Further, let f_j denote the number of times that a given value y_j occurs in X, for $1 \leq j \leq m$. (That is: various different x_k's can have the same value y_j; we denote number of x_k's that do so by f_j. So $f_1 + f_2 + \cdots + f_m =$ the total number of data points in $X = n$.) Then the above quantities $\bar{x}$ and s can be computed using the following formulas:

$$\bar{x} = \frac{f_1 y_1 + f_2 y_2 + \ldots + f_n y_n}{n}, \qquad s = \sqrt{\frac{f_1(y_1 - \bar{x})^2 + f_2(y_2 - \bar{x})^2 + \ldots + f_n(y_n - \bar{x})^2}{n-1}}.$$

The mean $\bar{x}$ is considered a measure of the *average*, or *center*, or *central tendency*, of the data in the set X. To see that this is a reasonable way to think of the mean, let's consider the formula for $\bar{x}$ given in part (a) of the above definition What this formula tells us is this: Suppose you have n perhaps unequal parts, of sizes $x_1, x_2, \ldots, x_n$, which constitute a whole of size $x_1 + x_2 + \ldots + x_n$. If you want to divide this whole into n *equal* parts, then each part must have size $\bar{x}$.

Similarly, the standard deviation s may be considered a measure of the *spread* of the data in X. The rationale for this way of thinking comes essentially from the quantities $(x_k - \bar{x})^2$ appearing in the above (first) definition of s. The idea here is that the magnitude of $x_k - \bar{x}$ tells us how far the data point x_k is from the center $\bar{x}$ of the data; so adding up all the $(x_k - \bar{x})^2$'s gives us a sense of how far the data is *collectively* from this center.

The squaring of each $x_k - \bar{x}$, in our definition of s, ensures that all of our summands are positive, so that negative terms won't cancel out positive ones. We divide by $n - 1$ to "level the playing field" among data sets of different sizes, so that adding data points to a set does not automatically increase its standard deviation. (In some definitions of s, the division is by n rather than $n - 1$. This is for technical reasons that we will not discuss here.) And finally, we take the square root at the end to compensate, in some sense, for the squaring that we applied to each $x_k - \bar{x}$.

There is another measure of spread in a data set called *mean absolute deviation.* This definition looks like the above definition of s, except that $(x_k - \bar{x})^2$ is replaced by $|x_k - \bar{x}|$ for each k, and no square root is taken at the end. The primary advantage of standard deviation over mean absolute deviation is that the latter entails absolute values, which are not locally linear everywhere. (See Section 2.2.) This makes calculus much harder to apply, and makes mean absolute deviation unwieldy from a mathematical perspective.

To highlight the above definitions, and in particular, how they work for data that's "grouped" into distinct values, let's return to our "six coins, flipped 3,444 times" data. The data set X in this case has size $n = 3{,}444$; a data point x_k tell us how many heads were observed on a particular trial (say, the kth trial) of the flipping experiment.

Of course, each x_k must be an integer from 0 to 6. That is, our data groups into distinct values $y_1 = 0$, $y_2 = 1$, $y_3 = 2$, and so on. The frequency f_j with which each of these y_j's occurs is given by Table 6.1.1. We may therefore use the formulas from part (b) of Definition 6.1.1 to compute $\bar{x}$ and s, as follows:

$$\bar{x} = \frac{(66 \times 0) + (341 \times 1) + \cdots + (333 \times 5) + (51 \times 6)}{3444} = 2.96922,$$

$$s = \sqrt{\frac{66(0 - \bar{x})^2 + 341(1 - \bar{x})^2 + \cdots + 333(5 - \bar{x})^2 + 51(6 - \bar{x})^2}{3443}} = 1.24663.$$

It's not surprising that our computed value of $\bar{x}$ is close to 3. The mean measures central tendency, or average, and if we flip six (fair) coins, then we would expect that, on the "average" flip, half of those coins (that is, three of them) should come up heads.

There is no equally simple, intuitive interpretation of standard deviation. However, there is a good "reality check" on the number we obtained for s above. Namely: it's known that, if data has an approximate normal distribution, then all or nearly all of that data should fall *within three standard deviations of the mean*. In mathematical terms this means that, for such a distribution, most or all of the data lies in the interval $(\bar{x} - 3s, \bar{x} + 3s)$.

Is this the case for our coin-flip data set X? Here, we have

$$(\bar{x} - 3s, \bar{x} + 3s) = (2.96922 - 3 \times 1.24663, 2.96922 + 3 \times 1.24663) = (-0.77067, 6.70911).$$

Since all data values lie between 0 and 6 inclusive, this interval *does*, in fact, capture all of the data – and does so without too much room to spare. That is, the interval does not overshoot the actual range of data values by much. All of this tells us that our computed value of s is at least in the right ballpark.

A different kind of histogram

A cornerstone of probability theory is the interpretation of probability as an *area under a graph*. Such an interpretation allows the full force of Calculus – Riemann sums, antiderivatives, the Fundamental Theorem, and so on – to be brought to bear on the study of probability.

Histograms, as considered above, are a first step towards realizing this interpretation. To go further, we will next need to *rescale*, or *renormalize*, these histograms, through the concept of *relative frequency density* (also called *probability density*). Here's the definition.

Definition 6.1.2. Let X be a data set, consisting of n real number data points. Consider any *bin*, meaning simply an interval of real numbers. Let F denote the number of data points in X

that lie in the bin, and let B denote the length of the bin. Then we define the *relative frequency density* of the bin, denoted RFD, by the formula

$$\text{RFD} = \frac{F}{B \times n}. \tag{6.1.1}$$

In short, the relative frequency density of a bin is the number of data points in that bin, divided by the product of the length of the bin and the size of the data set. (Strictly speaking RFD, for a given data set X, is a function of bin b: $\text{RFD} = f(b)$.)

Before discussing the importance of relative frequency density, we make the definition concrete by means of an example.

Example 6.1.1. Consider a data set X of $n = 68$ exam scores, distributed as follows.

Bin	F	RFD $= F/(B \times 68)$
[40,60)	11	$11/(20 \times 68) = 0.0081$
[60,70)	16	$16/(10 \times 68) = 0.0235$
[70,80)	14	$14/(10 \times 68) = 0.0206$
[80,85)	13	$13/(5 \times 68) = 0.0382$
[85,90)	9	$9/(5 \times 68) = 0.0265$
[90,100)	5	$5/(10 \times 68) = 0.0074$

We can draw a *relative frequency density histogram*, which is like the histograms drawn above, but now, the vertical axis denotes relative frequency density.

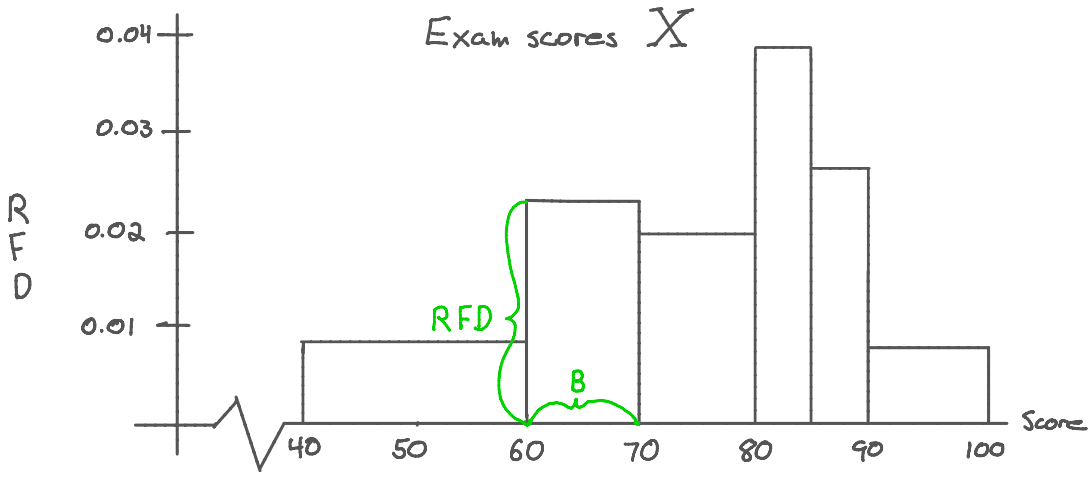

Figure 6.3. A relative frequency density histogram

Let's now see why relative frequency density is such a useful construct. To do this, we take the above definition (6.1.1) of RFD, and multiply both sides by B, to get

$$\text{RFD} \times B = \frac{F}{n}. \tag{6.1.2}$$

Let's look carefully at both sides of equation (6.1.2). The left-hand side gives the *height* times the *baselength* – that is, the *area* – of the RFD histogram "bar" that lies over the range in question. See Figure 6.3 above.

The quantity F/n on the right-hand side of (6.1.2) gives the number of points in X that lie in the given bin, divided by the *total* number of points in X. This quotient is just the *fraction*, or *proportion*, of the data that lies in the given bin. Or, put differently: F/n is the *probability* that a data point in X, chosen at random, lies in the given range.

Let's write $P(a \leq x < b)$ to denote the probability that a data point in X, chosen at random, lies in the interval $[a,b)$. Since the two sides of (6.1.2) are, in fact, equal, we then have the following conclusion.

In a relative frequency density histogram, the area of a bar over an interval $[a,b)$ equals $P(a \leq x < b)$.

Relationship between area and probability, in an RFD histogram

The bottom line is this: by plotting relative frequency density (rather than just frequency) on the vertical axis, we obtain histograms where area represents probability. This is a powerful idea, which we will exploit more fully in the next section.

In the meantime, we note that this idea can be applied "several bars at a time." For example, using the data and histogram from Example 6.1.1 above, we can compute that

$$\begin{aligned} P(60 \leq x < 85) &= \text{area enclosed by the bars covering the range } [60,85) \\ &= \text{sum of areas of bars over } [60,70),\ [70,80),\ \text{and } [80,85) \\ &= \text{sum of (height times baselength) of these bars} \\ &= \text{sum of (RFD times baselength) of these bars} \\ &= (0.0235 \times 10) + (0.0206 \times 10) + (0.0382 \times 5) = 0.6320. \end{aligned}$$

That is, 63.2% of the exam scores lie in the interval [60,85).

Of course, we could have argued more simply. Specifically, we could have used the relative frequency density table above to conclude that the proportion of data in [60,85) is $(16+14+13)/68 = 0.632353$. Example 6.1.1 helps to illustrate the mechanics of Definition 6.1.1, but the real *value* of relative frequency density will not be seen until the next section, where we use this construct in contexts where we can't simply "count data points."

In any case, our computation of $P(60 \leq x < 85)$ has taken advantage of the fact that the interval [60,85) is precisely spanned by three of our given bins. Were this not the case, we might only be able to *approximate* probabilities. For example, given the above information concerning our set X of exam scores, we might estimate that

$$\begin{aligned} P(63 \leq x < 82) &\approx \text{area enclosed by the bars covering the range [63,82)} \\ &= \text{sum of areas of bars over [63,70), [70,80), and [80,82)} \\ &= \text{sum of (height times baselength) of these bars} \\ &= \text{sum of (RFD times baselength) of these bars} \\ &= (0.0235 \times 7) + (0.0206 \times 10) + (0.0382 \times 2) = 0.44.69. \end{aligned}$$

So approximately 44.69% of the exam scores lie in the interval [63,82).

The above answer is approximate because we don't know that the exam scores are evenly distributed across each bin. For example, knowing that 16 data points lie in the interval [60,70), of length 10, clearly does not imply that $0.7 \times 16 = 11.2$ data points lie in the interval [63,70), of length 7.

We could get better estimates if we had narrower bins. In the next section, we'll consider bins that can, at least in theory, be made arbitrarily thin. This will lead us to the study of *probability density functions*, which are central to probability theory and statistical inference.