

## Histograms and the Shape of Distributions

Remember a distribution is just a collection of numbers. A histogram is a great way to get a visual image of the data which gives a lot of information about where the data are clumped, how spread out the numbers are etc. Producing a histogram can be a rather fussy and annoying process, even with a computer, but in the very common situation that we just want a quick peek at broad qualitative features, we will not need to worry about the fussiness much. We will distinguish a **presentation histogram**, where you want everything to look nice, be maximally informative and generally very detailed, from a **rough histogram**, where the purpose is simply to get an idea of the general shape of the histogram.

### Bins

No matter how you create a histogram you will need to choose the bins. The Excel template we will use, and many other statistical software tools, will choose the bin ranges for you, but they all do a pretty lousy job and come up with strange numbers for the boundaries. For all presentation histograms and sometimes for rough histograms you will need to choose the bins yourself.

Bins are the numerical ranges into which you'll group the data. **In presentation histograms your bins should all be the same size and should encompass all of the data.** In addition the boundaries should generally occur at nice round numbers. To choose the bin sizes, first find the smallest and largest data point. You can do this easily in Excel by typing  $\text{MIN}(A:A)$  and  $\text{MAX}(A:A)$  if your data is in column A. If you have the Data Analysis toolkit installed on your Excel, you can also look in the Tool menu for **Data Analysis**, and choose **Descriptive Statistics**. This gives lots of useful information including the max and min. Then lower the min a little and raise the max a little if necessary to make them nice round numbers. E.g., if your minimum value were 41.3 and your max value were 58.7, you might make your *lower boundary* 41 and your *upper boundary* 59, or to make it even rounder from 40 to 60. The difference between these two values is your *range*.

Now decide how many bins to use. This is very much a matter of judgment, but here are some guidelines. Too many bins makes the graph very uneven and 'noisy,' too few gives too little information. You can always change your mind after you look at it. Generally you should never have fewer than 5 or 6. For rough graphs 6-12 is generally good. The more data you have the more bins you should have, and some people recommend using the square root of the number of data points as your number. It is nice if the number of bins divides evenly into the range, and you can fiddle your range a little to make that true. Your *bin size* is your range divided by the number of bins. To get the actual *bin boundaries*, start from the minimum and keep adding the number *bin size* to it until you get to the

maximum. The range  $min$  to  $min + bin\ size$  is the first bin,  $min + bin\ size$  to  $min + 2\ bin\ size$  is the second and so on.

Example: You have 100 data points from 11, to 53.2. You decide 10 bins would be good. You try lowering the min to 10 and raising the max to 55, giving you a range of  $55 - 10 = 45$ , so a bin size of  $45/10 = 4.5$ . this is OK, but better would be to go from 10 to 60 for a range of 50 and a nice round bin size of 5. Your bin boundaries are 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60.

When making a rough histogram, you can be more flexible about what your range is. In particular in large data sets there are often a few data points which are much smaller or much larger than the other values. These are called *outliers*. If you follow the rules above you may find that you have a few bins at either end with one or no data points in them, taking up a lot of space without telling you much. **In a rough histogram, you can make your minimum be above the low outliers, and your maximum be below the high outliers.** Then include the outliers in the count of the smallest and largest bins. Thus you are effectively making those two bins cover more range than the others.

For example, suppose you are looking at test data in a large class, and you find that 13 people got 90-100, 18 got 80-90, 20 got 70-80, 12 got 60-70, 6 got 50-60, and the lowest two scores were 28 and 45. This is not an unusual distribution for test scores. If you made 20 the minimum and 100 the maximum, you would either have 8 bins, some empty, or you would need such big bins that you would be lumping As and Bs together. Instead we make the bins 90-100, 80-90, 70-80, 60-70, and “under 60,” the last bin containing 8 scores. This gives a nice simple picture of the data.

## Drawing the Histogram: Excel

For a quick and crude histogram, if you have fewer than 3,000 data points you can just paste your data into the first column of the “Data” tab in the Histogram template on my web page. It uses exactly 10 bins no matter what is appropriate, so it sometimes gives foolish looking results, but it makes a very good guess as to what the bin minimum and maximum should be, so you get a decent picture of the data with no further work. If you check the “Round bin size/min” it will make these reasonably round numbers without changing them too much, which is sometimes easier to see. If you check “Choose my own bin size/min” and enter values for these two in the white text boxes, it will make the histogram with your choices.

If you have a large amount of data, or if the histogram template does a lousy job on your data, or if you want to explore the data in more detail than it permits, or if you are trying to make a presentation histogram, you will have to do it yourself with the Data Analysis Add-in package. If you do not have this installed in your copy of Excel, you will need the original installation disk for Excel.

Somewhere convenient in the same workbook as your data, enter the numbers  $min + bin\ size$ ,  $min + 2\ bin\ size$ ,  $\dots$   $max-bin\ size$ . That is, all the boundaries between bins except the minimum and the maximum. Then use Tools  $\leftrightarrow$  Data Analysis  $\leftrightarrow$  Histogram, enter the data range of the data as your data, and the bin boundaries for bin. It will automatically

draw a problematic histogram. To make it less problematic, first double click on a bar. Choose **Options**, and set **Gap** width to zero (you can play around with the other choices here for graph prettification). Then re-scale the graph with the mouse as desired.

If you want the histogram to look nicer, go to the table next to your histogram and change the first column. The bins are all labeled by their top boundary (except the last which is labeled by ‘More’), which is pretty confusing. Better is to label them by the range. So if the first column says, say, 60, 70, 80, 90, More, you can replace it with 60 – 70, 70 – 80, 80 – 90, 90 – 100. If you do, use two hyphens in a row, otherwise Excel sometimes thinks you are typing a date.

Don’t hesitate to play around with this. If you click on the chart to select it, you can double click on all sorts of things and get options for changing colors, patterns, labels etc. Excel does not make producing histograms easy, but you can make them as beautiful as you have the patience for.

## Drawing a Histogram: By Hand

To draw a histogram by hand, you still need to choose the bin sizes and the cutoffs between bins. Now make a table, with one row for each bin, and in the first column write the range, perhaps in the form “60-90.” Now go through your data one at a time, and for each observation put a tick mark next to the bin in which it falls. So an observation of 71 would go in the “70-80” bin, and an observation of 90 would go in the “90-100” bin. Traditionally, observations exactly on the boundary get put in the upper bin (so 90 goes in 90-100 rather than 80-90), though Excel uses the opposite convention for some reason. It rarely matters, but you should be aware that the decision needs to be made. When you’re done, count the tick marks for each bin. Now draw a horizontal number line going from your lower boundary to your upper boundary, with the bin cutoffs marked. Draw a vertical number line starting at the horizontal line and going up to the largest number of ticks you got in any bin (or a little higher). Now draw a rectangle above each bin, whose width is the width of the bin and whose height is the number of tick marks in the bin. Color it in if desired.

## 1 The Shape of a Histogram

The shape of the histogram tells us several important things about the distribution. Generally, **when we speak of the shape of the histogram, we are only interested in features of the picture that we would not expect to change if there were more data, or if we changed the number of bins or the size of the bins.** If what we are looking at is a sample from some larger population (and it almost always will be) and it is small (i.e., it only has a few individuals in it), then chance variations will usually make the histogram wobble up and down a lot, and these wobbles look different depending on how you picked the bin boundaries. We would like to separate off this unimportant “noise” as best we can from the important features that would still hold if we picked a large sample.

In fact it is helpful to imagine a histogram for a huge amount of data with many small bins. If you imagine such a histogram, the tops of the rectangles begin to look almost like a smooth curve, and we will generally draw histograms when thinking abstractly by just drawing the smooth curve, allowing us to focus our thinking on the important features of the histogram.

The most common situation will be that most of our data is clumped in one area, with “tails” of larger and smaller data stretched out to either side. If you picture the histogram you will see that it looks like a rounded hill or mound. We will call such data *unimodular* or “mounded,” although it is such a common situation that people rarely use any term for it.

If the left and right sides of the histogram look the same, like a mirror image, we say the data is *symmetric*. The best situation to be in is when your data is unimodular *and* symmetric, in which case we say it is *bell-shaped*. This happens very generally when the variation you are seeing is built up of many independent small causes of variation, something we will explore in more depth later in the semester. The vast majority of the techniques we will learn this semester are accurate only when you have a large sample or your data is bell-shaped.

If the data is not symmetric it is often either *left skewed* or *right skewed*. Left skewed means that the data contains a long left tail and a short right tail. The only place left-skewed data is commonly seen is in test grades, where typically scores cluster in the 70s and 80s, with a short tail above (it can’t go any higher than 100) and a long tail below (there are often a few students who never showed up to class and bomb the test). Right skewed means the data contains a long right tail and a short left tail. This is actually quite common. It arises, for example, if you look at distributions of incomes, or housing prices, or indeed most data involving dollar value. For example in incomes most people are clumped around a “typical” income, and since no one has an income less than 0, there is a limit to how long the left tail can be, but a few people will have incomes much larger than typical, going up to some with incomes hundreds of times more than most other people.

When the data is not unimodular, it might be *bimodular*, which means it has two distinct mounds (like a camel, or a dromedary, whichever a unimodular distribution is not like!). This typically indicates your population consists of two distinct groups with rather different features, such as men and women. Small samples may appear bimodular just by chance, don’t jump to conclusions here. The data can also be *uniform*, meaning the distribution looks like a rectangle with a flat top. This happens in situations that are truly random, so that all outcomes are equally likely. For example, if you roll a die a thousand times and make a histogram of the resulting numbers, you will get a uniform distribution with each of the six possible values occurring equally often (approximately).

You should be able to look at a histogram and decide which of the above terms apply. You should also be able to think about the population and variable from which the data came and try to explain why you are seeing these qualitative features. By the end of the semester you will often be able to predict the shape of a histogram just knowing the population and the variable in question!