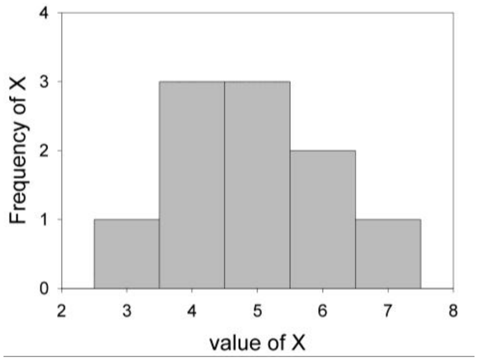
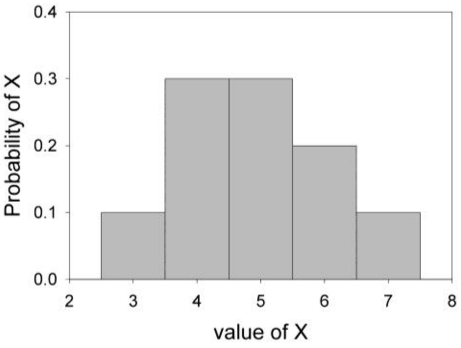
Pre-lecture Notes I.2 – Introduction to Psychological Data

Recall that psychology is an empirical science, which means that we test our theories by comparing their predictions to actual data. Now consider all of the different kinds of thing that psychologists generate theories about. These range from relatively simple perceptual or motor functions, such as the ability to discriminate the brightnesses of two different lights or the ability to press a particular button, to rather complicated functions, such as the ability to ask for something in a manner that maximizes the chances that a particular listener – with their own unique background, history, set of desires, etc. – will agree to the request. It should not be surprising, therefore – with theories that concern so many different abilities – that psychological data involve a huge variety of specific behaviors, ranging from simple yes/no responses or reaction times to whether a spoken sentence will include several dependent clauses or employ an inverted structure (as in Yoda-speak: "hard to follow, this sentence is"). To reduce the confusion that this variety of data may cause, I suggest that we focus, for now, on the commonalities. And, to ease into this area slowly, let's start with a simple situation.

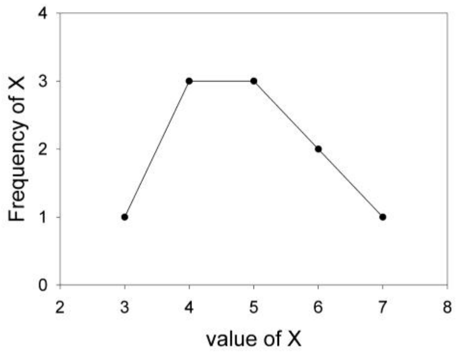
Assume that you already have a single set of numbers – i.e., one piece of data from each of many research participants. What should you do? The first step is simplify matters by coming up with a summary of all of these data that can be easily communicated to others. The most straight-forward way to do this is via a graph or a plot of some sort. The most popular type of plot to use for univariate (single-variable) psychological data is the distribution function.

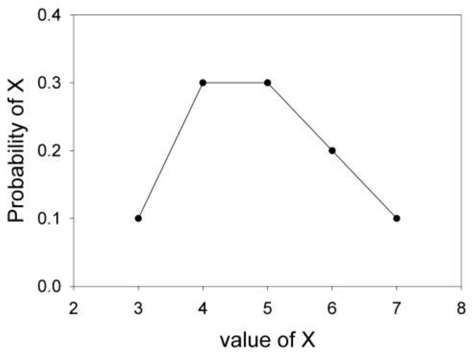
Assume that you have measured 10 people on something called X and observed the values of 4, 6, 3, 5, 7, 4, 6, 5, 4, and 5. (If you'd like a concrete example, X can be how many meals were skipped in the previous seven days.) In our plot, we're going to have the value of X on the horizontal axis and how many people produced this particular value on the vertical axis. The resulting plot – which is the distribution function for our data – is on the left side below. Note, however, that some people prefer to think of the data in terms of what percent or proportion of the subjects produced each particular value of X, instead of the number or frequency of people; the plot of this sort – which is a density function – is on the right side below.





A few more comments about these plots are in order. First, note that the X axis was extended a little bit beyond the observed data in both directions and that the Y axis went down to zero, as well as a bit above the higher point in the data. These are not required, but it is the convention, so you should follow it, too. Extending a bit beyond the data in every direction makes it easier for the viewer to get a picture of the entire set of data. Second, note that the shapes of the two different graphs are the same. This had to be true, since they both depict the same data. If you ever create two plots of the same data and they don't look the same, stop and find where you made a mistake. Finally, note that the data were plotted as a set of bars, one for each value. This created what is called a histogram (or bar-chart); that's a frequency histogram on the left and a density histogram on the right. The other way to plot data uses points with lines between them, instead. These are called frequency or density polygons and are shown below.





There are rules for when you use a polygon as opposed to a bar graph. In brief: when the data are capable of taking on any value between two end-points, you use a polygon; an example of this kind of data is response time, which can be anything from zero to infinity. When the data can only be certain values, such as how many siblings a person has (which can only be a whole number), then you use a histogram, instead, with one bar for each particular value that is possible.

Plots are a great way to summarize a set of data, but they do have one weakness: they haven't really reduced the complexity of the data, so they don't allow you to express what you've found in just a few numbers or words. This is where descriptive statistics come in. Descriptive statistics allow you to summarize any set of data by "boiling it down" to a very small set of values. A majority of lecture will cover the way that psychologists prefer to summarize data, but a few comments will be made here.

There are lots of different ways to summarize a set of data; psychologists like to use three things. We want a measure of the *center* of the data, a measure of how *spread* out the data are (around the center), and we want a measure of general *shape* of the data (i.e., what the density function looks like: are the data spread evenly or do they pile up in the middle, like the plots above; also, is the amount of spread at the lower end the same or different from the amount of spread at the upper end). There are several options for each of these and this is one place where there is some disagreement between psychologists. Before lecture, see if you can come up with a good way to summarize the center, spread, and shape of the data shown in the plots above.