Empirical Distribution Functions

Supplemental Reading for MA206 Probability and Statistics

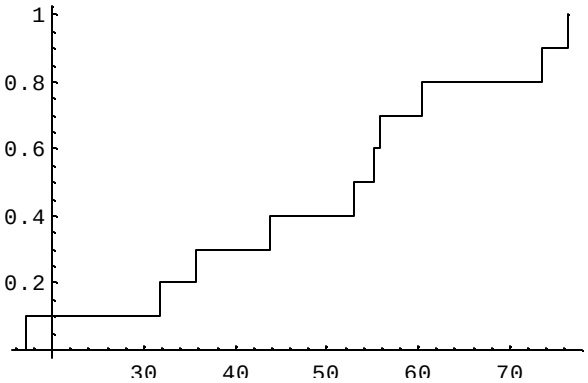

Figure 1: Example Empirical Distribution Function

Background and Purpose

So far in the course, we have looked at various ways of displaying data we have collected, or important characteristics of that data.

The empirical distribution function (EDF) is one way of displaying the data — one that emphasizes the *distribution* of the data, in the sense of how the probability of the events is distributed over the range of the random phenomenon we are studying.

One of the reasons for doing this is to let us create models of the data -- simplification of the real data that represent it well enough for the particular use we have for it.

Support - The Range of Values for the Variable

An important characteristic that helps us choose among various models of the actual data is the set of possible values it can take on—the *range* of the variable of interest, which is also the *domain* of the EDF. Reall that the *domain* of a function is the set of "input" values, while the *range* is the set of resulting "output" values. So the "inputs" to the EDF—plotted on the horizontal axis—are the set of possible values the phenomenom we are studying can assume. The "outputs" of the EDF—plotted on the vertical axis—are *probabilities*, specifically the probability that the variable of interest was observed to be less than or equal to the particular value

There are several different cases that are interesting:

- The range of the variable is the whole real line

- The range of the variable is the positive real line
- The range of the variable is an interval on the real line
- The range of the variable is a discrete set of values

For example, in Figure 2 we can see that the possible values for the varaible we are studying all lie between 5 and 10 (*e.g.* a interval on the real line). This figure is an EDF from a sample of size 250, which should give a good picture of the behavior of the variable. Most of the figures in this paper are plotted with large samples to make the characteristics of the distribution easier to see; in an actual analysis we usually are not that lucky.

Figure 2: EDF from a variable that appears to be distributed uniformly between 5 and 10

For comparison, look at Figure 3. Notice that, while the values that were actually observed are (of course) bounded by a finite interval, the EDF appears to approach one asymtotically as the variable gets larger and larger, and zero as the variable gets smaller and smaller. The variable being studied here *can* take on very small (*i.e.* negative) or very large values. But it's most likely to take on values somewhere around 10; as we get farther and farther away from 10, it's increasingly unlikely to observe the variable.

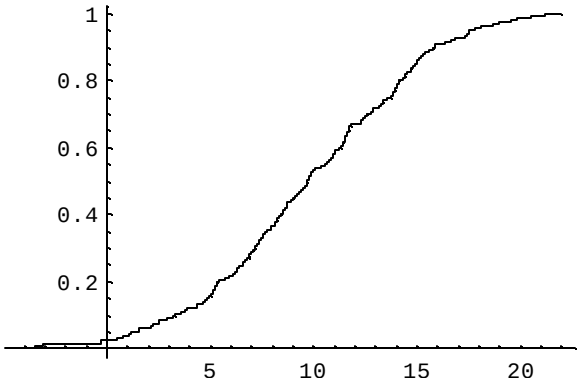

Figure 3: EDF from a variable whose possible values are the entire real line

Symmetry and Shape

The shape of the EDF can also tell us about the way the variable behaves. For example, if the EDF seems to be sloping up from zero to one in more or less a straight line (as in Figure 2), this indicates that the variable can assume any value within its range with approximately equal probability. Figure 3, on the other hand, shows the EDF of a variable that is most likely to be near the a particular value (10, in this case) and is less and less likely to appear as we move away from that central value.

Both Figures 2 and 3 share an important characteristic. If you compare the "upper half" of the graph (the part that corresponds to probabilities between 0.5 and 1) to the lower half (probabilities between 0 and 0.5), you can see that the distribution of probability is *symmetric*. Imagine pinning the graph where it crosses the probability = 0.5 point and rotating the upper half counterclockwise so that the line for probability = 1 overlays the probability = 0 line. The graph of the EDF from the upper half will lay on top of the graph of the lower half.

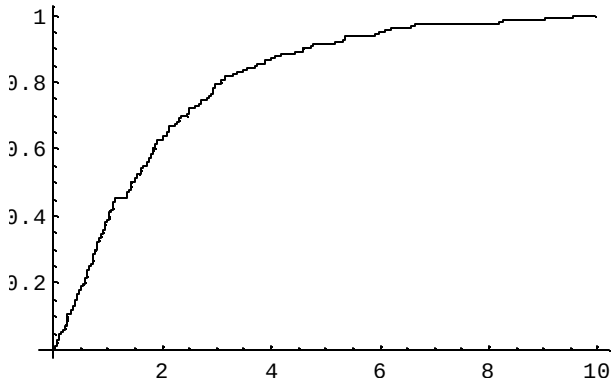

Figure 4: EDF of a variable with a strongly asymmetric distribution

In contrast to Figures 2 and 3, look at Figure 4. The "top half" of this distribution looks very different than the "bottom half"—in fact the variable seems to be able to take on very large values (albeit with low probability), but cannot take on negative values. This characteristic is *asymmetry*—the opposite of symmetry.

Figure 4 is also an example of a variable whose range is the *positive real line*, incidentally.