1. About Data

Data, in this class, is information on people, places, companies, almost anything. Some vocabulary.

VARIABLE: a characteristic of a person, animal, place, thing that can be expressed as a number.

VALUE: A value is the actual number associated with the variable

TYPES OF VARIABLES (1) Quantitative - have a "natural" ordering which may be discrete like the spots on a single die. They may also be continuous like time, temperature. (2) Qualitative - do not have a "natural" ordering like major, occupation, name brand.

The Distribution (pattern) of Data

Statisticians love examining distributions of variables. And a graphical representation of a distribution can answer questions like how many are large, how many are small, how many fall between two numbers, what is the most common number?

2. The Histogram -- a way to examine distributions

A histogram is a way of examining distributions. It can quickly summarize an enormous amount of information on a single variable and it makes use of your natural ability to recognize patterns. Examples are on pages 30 & 31.

A HISTOGRAM is a graph that shows percentages by area. The rectangles are called "bins." The key to a histogram is that it is the AREA of the bin, not the height of the bin, that is important. The area of the bin is proportional to the relative frequency of observations in the bin:

- A histogram represents data observations by AREA for different class intervals, not height. The area of each "block" is proportional to the number of observations in the class interval. The total area MUST be 100%. Specifically:

(#observations in the bin)/(total number of observations).

- It has class intervals on its horizontal axis. This axis must be scaled.
- It does not require a scale on its vertical axis. The vertical axis is scaled in percent per unit of the horizontal axis, and a scale is automatically imposed by the fact that the area of the histogram **must** be 100%

We also need an "endpoint convention" to be able to draw a histogram: if an observation falls on the boundary between two class intervals, to which one should we associate it? The two standard choices are always to include the left boundary and exclude the right, except for the rightmost bin, or always to include the right boundary and exclude the left, except for the leftmost bin.

4. Building a Histogram

To plot a histogram, we first need to sort the data into increasing order, and pick the class intervals. We then count the number of data that fall in each class interval, and plot rectangles with the areas proportional to the relative frequencies with which the data fall in each class interval.

There are no standard rules for determining appropriate class intervals, and the impression one gets of how the data are distributed depends on the number and location of the intervals.

Histograms are usually used for large datasets.

5. Things to be aware of with respect to histograms

Center: what is the "typical" value?

Symmetry or skewness: are the data evenly divided is there a tail? Are there bumps?

Spread: are the data near to each other or far apart?

Exceptions ("outliers"): are there points that don't fit the general distribution?

Remark: the histogram can take all kinds of shapes.

The whole point of this exercise is really to help you learn how to convey information in a meaningful way using graphics.