Last lecture we reviewed data: the types of data, the concept of the variable and the forms a variable can take. This lecture focuses on typical ways to present data.

1. Frequency Tables

For nearly any variable we can produce a table of values. A frequency table adds a tally or count or frequency of occurrence of each individual value in the table. Often, when researchers are referring to a "frequency table" it also includes information on the "relative frequency" or percentage of a given value in the table. Finally, some software packages calculate a "cumulative frequency" or a "cumulative percentage" as part of a frequency table. Notice that the total frequency of a frequency table is 100%. Examine the handout for comparison.

Discrete valued quantitative values work best, but categorical variables also work well. What is desirable is a limited number of values. Continuous variables almost always need to be grouped or "collapsed" into "cells". There are problems with grouping variables – the detailed information contained in a continuous variable may be lost or obscured.

A frequency table may the best way, however, to examine categorical variables in more detail.

SIDEBAR: Definition of a percentile: a number y is the nth PERCENTILE for the data if n% of the data are less than or equal to y. The QUARTILES are the 25th, 50th, and 75th percentiles and are denoted Q1, Q2, and Q3. The 50^{th} percentile has an additional name: the median. The difference in the values of the Q3 and Q1 (that is value Q3-value Q1 is called the IQR or inter-quartile range)

2. Box Plots (avoid the book's explanation)

A way of graphically summarizing data. Looking at one from top to bottom it uses the values of: The maximum or (1.5*IQR) + value of Q3 (whichever one is closer to the Q2); the value of Q3; the value of Q2 (or the Median); the value of Q1; Q1- (1.5*IQR) or the minimum (whichever one is closer to the median). If there are outliers (extreme values) in the distribution they are identified by "*"s. An outlier in this class is a value that is either larger than (1.5*IQR) + value of Q3 but less than or equal to the maximum OR smaller than (1.5*IQR) - value of Q1 and greater than or equal to the minimum.

3. Histograms -- a way to examine distributions

A histogram of relative frequencies, is a graph that shows percentages by area for a given variable. 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:

The horizontal axis (also called the x axis) must have a scale with units. Technically, the vertical axis should have units of percentage per unit of the horizontal axis or at least percentages using bins with equally sized intervals of the horizontal axis (x axis).

For a proper histogram (a histogram of relative frequencies) the scale of the vertical axis is automatically imposed by the fact that the area of the histogram must be 100% (all the data fall somewhere on the plot).

4. Scatterplots (Chapter 2.7)

A scatterplot is a two dimensional plot of data; the horizontal dimension is called X, and the vertical dimension is called Y. Each point on a scatterplot shows two values, an X value and a Y value; each point represents a single case or observation. Explanatory variables (also known as independent variables) are by convention identified as the X value and the response variable (the outcome variable or the dependent variable) is the Y.

B. Things to think about when looking at scatterplots

Form: does it have a shape (like an egg, a circle, a cigar, a "U")

Direction: does the data have a direction (is it generally running up from the lower left to the upper right or downwards from the upper left to the lower right)

Strength: Are the points close together or scattered?

Basically, In a scatterplot of two variables, if the scatter in the values of the variable plotted on the vertical axis is smaller in narrow ranges of the variable plotted on the horizontal axis (*i.e.,* in vertical "slices") than it is overall, the two variables demonstrate association.

C. Positive and negative relationships

There is a POSITIVE relationship if above-average values of X are associated with above-average values of Y; conversely, there is a NEGATIVE relationship if above-average values of X are associated with below average values of Y.

D. Warning! Scatterplots only show association; but association is not causation (firefighters, fire damage confounding example).

5. Time Plots

A time plot is simply a plot with some variable representing time (e.g. order of observation, years, minutes) on the horizontal axis (see previous lecture handout on the unemployment rate). Points are generally connected in their time order (see the figures on pages 54-56;59). Economic data is frequently organized in times series. Time plots can help you identify patterns in the data:

A. Seasonal variation or seasonality

A pattern in a time plot that repeats itself regularly is evidence of seasonality. For example, consumer expenditures tend to be high in the month December (holiday shopping) and drop in January (bad weather and post-holiday cash problems).

B. Trend

A persistent pattern, either a long-term rise or a long-term fall.

6. Things to be aware of with respect to graphs

Center: what is the "typical" value?

Symmetry or skewness: are the data evenly divided is there a tail? Are there bumps? If the data is in groups, are they appropriately constructed?

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

Exceptions ("outliers"): are there points that do not fit in the general picture? What are the highs and lows?

Remark: are the data portrayed fairly or is it misleading? Are unusual intervals being used, are large changes being obscured? Are small changes exaggerated?

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