David Youngberg
BSAD 210—Montgomery College

LECTURE 06: OF DATA AND DISPLAYS II

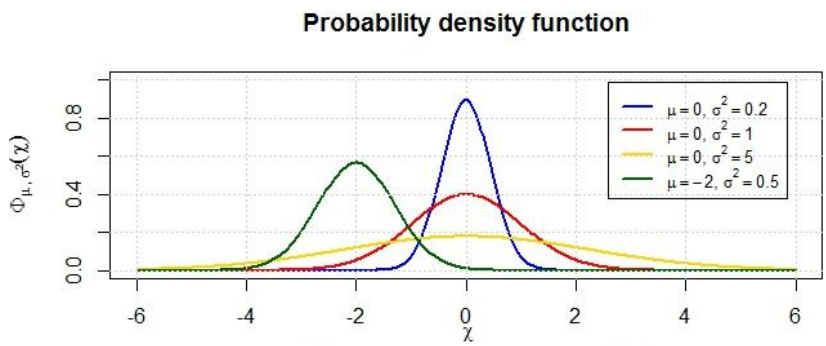
I. Scatterplot
   a. The first step in any research project is finding data (this sometimes occurs even before you know what you want to investigate).
   b. The second step is determining your approach to the data.
   c. A *scatter diagram* indicates how two (or more, if you are feeling adventurous) values relate to each other.
   d. Gapminder (www.gapminder.org) is an excellent resource to explore relations between different variables. The website employs data from all over the world to various sophisticated scatter plots. The raw data are available in Excel format.
   e. You'll notice on Gapminder that you can express a variable on a linear (lin) or logarithmic (log) scale.
      i. A linear scale means each unit is some previous unit plus a fixed value. For example: 10; 20; 30; 40; 50; etc.
      ii. A logarithmic scale means each unit is some previous unit *times* a fixed value. For example: 10; 100; 1,000; 10,000; etc
      iii. For values with a wide range (especially ones skewed right) logarithmic scales are a better visual choice.

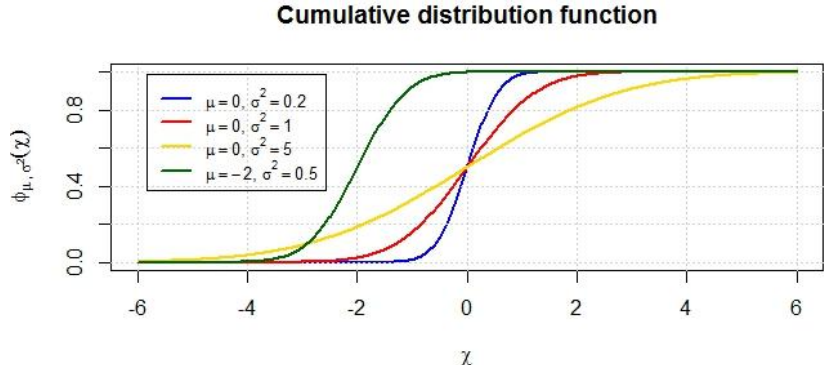
II. CDFs and PDFs
   a. Where a discrete distribution has "chunky" values (1, 2, 3, 4, etc.), a continuous distribution has infinitely divisible values. For a limited range (e.g. 1 through 6), there are limited values (e.g. 6) for a discrete distribution. For continuous distribution, there are an infinite (or just very large) number of values in a limited range.
   b. When illustrating a continuous distribution, it's typical to use a *probability density function* (PDF), where the frequency of each value is indicated on the y-axis.
   c. You can also illustrate a distribution with a *cumulative distribution function* (CDF), where the total instances of each value thus far accumulated, starting from the lowest value, is indicated on a y-axis. CDFs have the following attributes.
      i. The farthest right point is always at 100% (or whatever the total number of observations is).
      ii. In a continuous distribution, the farthest left point is at 0% (or zero observations).

iii. It never slopes down; either the number of observations is increasing (some additional observations) or stays the same (no additional observations).

iv. The slope of the line indicates how quickly observations are being added in that range. Steeper lines means more observations added.

d. For example, here's a PDF of what are called normal distributions:



The red curve is the standard normal distribution

e. And here's the CDF of those same distributions: