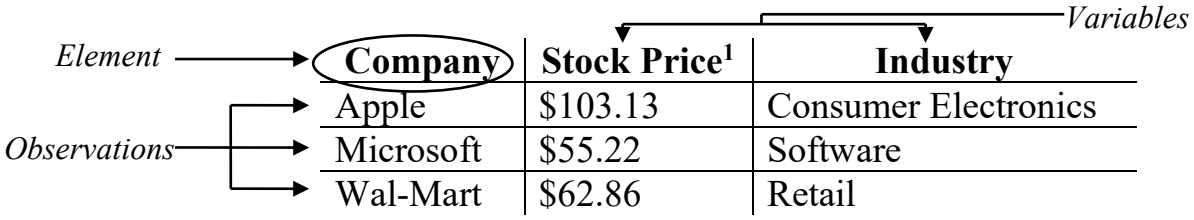
David Youngberg
BSAD 210—Montgomery College

LECTURE 01: INTRODUCTION TO STATISTICS

I. What is statistics?
   a. *Statistics* is the application of mathematics to understanding and collecting of sample data.
      i. Some of this application is through a controlled experiment. For example, determining if a focus group really likes a new product more than the current one, or if their higher opinion is just due to random chance.
      ii. But statistics is also useful when you can't run a controlled experiment. Some of the most important questions in the social sciences—from business to economics to psychology—are addressed using statistical techniques we'll explore.
   b. Consider one of the most important questions in economics: How do turn poor countries into wealthy ones?
      i. Economists have lots of ideas, but we don't know what will actually work. For example, does giving the country's government a bunch of money help?
      ii. Ideally the World Bank or IMF would randomly select half of the poor countries and then give that half some large amount of money (adjusted by population). Then we can look at the results.
      iii. But that's not an option and not just because each country is so different. There are ethical and legal constraints. The struggling countries that didn't get anything would wonder why they are left out. And, by chance, some of that money would go to countries that we know are corrupt. Even if we learn a lot, it would be a short-term disaster.

II. Types of Data
   a. Data are (note: "data" is plural of datum; datum is a single piece of information) the pieces of information to be analyzed. Within a data set there are:
      i. *Element*—on what the data are collected (e.g. companies)
      ii. *Variable*—a notable characteristic of the element (e.g. stock price)
      iii. *Observation*—a particular element in the data set (e.g. Apple)



a. Elements can result into two different kinds of data: cross-sectional or time series.
    iv. *Cross-sectional data* are collected at the same point in time; the possible elements are made up of different kinds of things: countries, firms, U.S. states, etc. The dataset you'll be working with for your memo will be cross-sectional data.
    v. *Time series data* are collected across time; the element is always time in some fashion: months, days, years, etc.
b. Note the type of variables here are very different. Stock Price uses a number while Industry refers to a category.
    i. *Quantitative data*—made up of numerical values, resulting in clear implications between observations. For example, Apple's stock price is almost twice as much as Microsoft's.
    ii. *Categorical data*—made up of names of categories. Sometimes it can be meaningfully translated into numbers and has clear comparison implications (e.g. a grade which can be transformed into a numerical 4.0 scale). But it usually cannot be. There is no inherently quantifiable difference between "Software" and "Consumer Electronics." Saying one is "twice as much as the other" would meaningless.

III. Abbreviations
a. Quantitative data is sometimes expressed with abbreviations, such as in thousands or in millions, to save space. It's important to be comfortable with these abbreviations so let's do a quick refresher.
b. There are different practices on how to represent "in thousands", etc. with a single letter. The standard I'll use for this class is:
    i. K = thousands (50K = 50,000)
    ii. M = millions (118.1M = 118,100,000)
    iii. T = trillions (76.56T = 76,560,000,000)
c. In homework, I will often ask for values in thousands or millions or whatever. This is partly to save you time, and partly so you don't get

[1] As of January 5, 2016 2:55 pm

an answer wrong because you left out or added an extra zero. *I'm trying to make things easier for you*.

d. But some people will not read instructions. They will write 3,400,000 when the question specified to put it "in millions." Even if 3,400,000 is "the answer," that's not how you should put the answer in. How should you put in that value?
    i. When you convert one abbreviation into another, you move the decimal point to the right or left in groups of three.

    ii. In other words, the student should write 3.4, not 3,400,000. 3,400,000 would be 3,400,000 million, or 3.4 trillion.

IV. Example: Data Set 0

a. Data Set 0 summarizes 2014 earnings and popularity based on different college majors, as reported by Georgetown University's Center on Education and the Workforce in 2015. Let's apply some of what we've learned.

b. What's the element of this data?
    i. Major. While the element is often the most-left column of a data set, that is not always the case. You can tell which is the element based on why values in variables are what they are. For example, in row 3 the reason why it says "Business" under Group is because business is the group for the accounting major. Similarly, the reason why Majors per 10K Graduates says "463" is because that's the number of accounting majors per 10,000 graduates.

c. What kind of data set is this data?
    i. Cross sectional. All variables represent the same time—2014.

d. What kind of variable is Group? What about Majors per 10K Graduates?
    i. Categorical for Group. There are clean lines between Art and Business—these are names of categories, not numbers.
    ii. Quantitative for Majors per 10K Graduates. These are numbers, and these numbers are not stand-ins for categories. If a major has twice as many graduates, it is twice as popular.