

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)

	Company	Stock price¹	Industry
Observations	Apple	\$103.13	Consumer Electronics
	Microsoft	\$55.22	Software
	Wal-Mart	\$62.86	Retail

Diagram annotations: An arrow labeled "Element" points to the "Company" header. An arrow labeled "Observations" points to the first three rows. A bracket labeled "Variables" spans the "Stock price" and "Industry" columns.

- 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 be meaningless.
- c. Elements can result into two different kinds of data: cross-sectional or time series.
- i. *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.
 - ii. *Time series data* are collected across time; the element is always time in some fashion: months, days, years, etc.

¹ As of January 5, 2016 2:55 pm