1 Introduction and Data Collection

The information we gather with *experiments* and with *surveys* is collectively called **data**.

Statistics is the art and science of designing studies and analyzing the data that those studies produce. Its ultimate goal is translating data into knowledge and understanding of the world around us. In short, **statistics** is the art and science of learning from data.

Why study statistics?

According to Mark Twain (1835-1910, Samuel Clemens), who incorrectly attributed the quote to British Prime Minister Benjamin Disraeli (1804-1881): "There are three kinds of lies: lies, damned lies, and statistics."

Example: First (Persian) Gulf War - Economic Sanctions and 567,000 Additional Child Deaths (*Significance, Royal Statistical Society and American Statistical Association*, 2010, vol. 7, #3)

1. 1990 August 2: Iraq, under the Presidency of Saddam Hussein, invaded Kuwait.
2. 1990 August 6: Economic sanctions were imposed on Iraq.
3. 1991 January 17: The United States invaded Iraq, officially beginning the Gulf War.
4. 1991 February 28: The war ended, and President George Bush declared victory.
5. 1995: *United Nations Food and Agriculture Organization* AND *Iraq's Ministry of Agriculture and Nutrition Research Institute* (FAO-NRI) interviewed households in Baghdad.

6. 1996: Lesley Stahl (CBS newscaster, winning an Emmy for this interview regarding **economic sanctions**): "We have heard that a **half million children** [under the age of 5] have died. I mean, that's more children than died in Hiroshima. And, you know, is the price worth it?"

   Madeleine Albright (U.S. Secretary of of State): "I think this is a very hard choice, but the price - we think the price is worth it."

7. 1997: Conclusions of FAO-NRI study were withdrawn.
8. 2003: Economic sanctions were lifted with the beginning of the Second Gulf War.
9. 2010 January: Tony Blair (British Prime Minister 1997-2007) told an official British panel investigating his Iraq policy: "50,000 young people, children [survived, and] that's the result that getting rid of Saddam makes."

Did economic sanctions really cause **567,000** Iraqi children to die between 1991 and 1995?

1. war, two uprisings, and mass migration
2. fear
3. extrapolation
4. World Health Organization

□

Example: Autism vs. Vaccines. In 1998 Dr. Andrew Wakefield published in *Lancet* his belief that the MMR (measles, mumps, and rubella) vaccine causes autism. Jenny McCarthy (a celebrity) for years publicly claimed that the MMR vaccine caused autism in her son.

□

Example: The Freshman Fifteen. TRUE or FALSE: *The average weight gain of college students during their freshman year is 15 pounds.*

Where else is statistics used in the real world?

Example: In World War II, Japan attacked Midway (North Pacific Ocean) on June 4, 1942.

Example: Satellite imagery.

Example: Drug development, approval, and safety.

Example: Assessing disease risk. Based on history, environment or behavior, how great is the risk for an individual for cancer, heart attack or stroke?

Example: Health policy - track the nation's health care system.

Example: Economic productivity - monitor trade deficit, gross national product, consumer price index, and unemployment rate; software / web development; test marketing.

Example: Environmental monitoring - pollution regulation vs. environmental health, climate change, monitor natural resources.

Example: Energy policy - track energy production and consumption, energy efficiency, projecting future energy supply and demand, model effects of policy interventions.

Example: Sports and gambling? MOSTLY JUST ENTERTAINMENT.

□

The **population** is the total set of subjects in which we are interested.

A **sample** is the subset of the population for whom we have (or plan to have) data.

A **parameter** is a numerical summary of the *population*.

A **statistic** is a numerical summary of a *sample* taken from the population.

Variables may be **numerical (quantitative)** or **categorical (qualitative)**.

Examples of **numerical** variables are:

Examples of **categorical** variables are:

Two types of **numerical** variables are **discrete** and **continuous**.

1. **Discrete variable** takes values which are distinct numbers with gaps.

2. **Continuous variable** takes any value in an interval.

Levels of Measurement and Measurement Scales

Categorical data

"A **nominal scale** classifies data into distinct categories in which no ranking is implied."

"An **ordinal scale** classifies data into distinct categories in which ranking is implied."

Numerical data

"An **interval scale** is an ordered scale in which the difference between measurements is a meaningful quantity but **does not involve a true zero point**."

"A **ratio scale** is an ordered scale in which the difference between the measurements involves a true zero point."

Read pp. 16–21, Appendix E1: Introduction to Microsoft Excel.