1 Overview and Descriptive Statistics

Introduction

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7. 1997: Conclusions of FAO-NRI study were withdrawn.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.*

Example: TRUE, FALSE, or DUBIOUS: *Half of all marriages end in divorce.*

Where else are 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.

□

Statistics consists of

- **Design:** Planning how to obtain data to answer the questions of interest.
- **Description:** Summarizing the data that are obtained.
- **Inference:** Making decisions and predictions based on the data.

1.1 Populations, Samples, and Processes

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.

Variables may be **univariate**, **bivariate**, or **multivariate**.

For **univariate** data, observations are made on just one variable. For example, income.

For **bivariate** data, observations are made on two variables. For example, (1) whether or not an individual smoked, and (2) whether or not the individual gets lung cancer.

For **multivariate** data, observations are made on at least two variables. For example, (1) blood pressure (systolic and diastolic), (2) cholesterol level, (3) incidence of heart attack.

Branches of Statistics

The field of *statistics* may be divided into the two branches:

- **descriptive statistics:** Summarizing data via graphs and numbers (such as averages and percentages).

- **inferential statistics:** Making decisions or predictions about a population based on a sample. Hence, the population is unknown, but a sample is known.

Example of inferential statistics:

In Math 318 we study both *statistics* and *probability*.

Regarding **probability**, the *population* is assumed known, and statements are made regarding the likelihood of obtaining certain data values.

Example of probability:

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

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

Randomness and Variability

Random sampling implies that data are selected for the sample at random from the population.

Enumerative Versus Analytic Studies

Definition: Enumerative studies are based on a **sampling frame**, which is a **non-hypothetical** population from which the sample is taken.

Example: If we randomly select 5 stocks on the New York Stock Exchange from January 1, 2008, what is the probability that at least 2 of these stocks exceeded \$20/share on January 1, 2008?

Example: What proportion of Americans *currently* earn over \$40,000?

Definition: Analytic studies are based on a **hypothetical** population.

Example: What proportion of stocks on the New York Stock Exchange will exceed \$20/share **next year**?

Example: What is the chance of rain tomorrow?

Example: Who wrote the unsigned *Federalist Papers*?

Example: What is the likelihood that the United States will win at least one gold medal in figure skating in the next Olympics?

Example: What is the likelihood that human life on earth resulted from evolution?

Collecting Data via Sample Surveys

Definition: A **simple random sample** of n subjects from a population is one in which each possible sample of that size has the same chance of being selected.

Example: Suppose we want to sample 80 students out of 7,000. Should we select first 80 names on the list?

□

Definition: A **stratified random sample** divides the population into groups called **strata**, and then selects a simple random sample from each stratum.

Example: Suppose that a university is known to be 60% female and 40% male, and a survey is to be conducted related to the abortion issue. Enough funding (or time) is available to sample 100 students.

(a) How would a *simple random sample* be taken?

(b) How would a *stratified random sample* be taken?

(c) Which sample is better and why?

Definition: Divide the population into a large number of **clusters**. Select a simple random sample of the clusters. All elements within each *sampled* cluster are sampled, to form a **cluster random sample**.

Example: You have one week to estimate the average annual church donation of Baptists members in Rhode Island.

(a) How would a *simple random sample* be taken?

(b) How would a *cluster random sample* be taken?

(c) Which sample is better and why?

Example: The *Literary Digest* Poll *(Know this example in detail, although you need NOT memorize the numbers.)*

Franklin Roosevelt vs. Alfred Landon, Election of 1936.

Since 1916, the *Literary Digest* correctly picked the Presidents.

Digest mailed questionnaires to 10 million people, whose names were from country club membership lists, phone books, and automobile registrations.

George Gallup, polling 50,000 people, predicted *Digest's* results in advance.

3rd party candidates were excluded in the numbers below.

	Roosevelt's percentage
The election result	62
Digest's prediction	43
Gallup's prediction of *Digest*	44
Gallup's prediction of election	56

□

In general, taking larger samples increases **precision** but does not reduce **bias**.

Collecting Data via The Design of Experiments

A properly performed **experiment** requires **control**, **randomization**, and **replication**. A bonus is **double-blindedness**.

Example: Salk Vaccine Field Trial *(Know this example in detail, although you need NOT memorize the numbers.)*

In 1916 polio epidemic in United States.

In 1950s Jonas Salk had a promising "vaccine," which worked well in laboratory (i.e., the vaccine seemed safe and produced antibodies against polio).

What now? Test whether or not vaccine works.

(a) (hypothetical) Test vaccine on a small sample (e.g., 10) of children.

If successful on them, mass distribute the vaccine.

(b) (hypothetical) Offer vaccine to a large number of children.

Typically, not everyone will accept the vaccine.

We have two groups: **treatment** (those who accepted the vaccine) and **control** (those who declined the vaccine).

What is the **explanatory variable**?

What is the **response variable**?

Would this study be considered a valid experiment?

(c) (real data) The National Foundation for Infantile Paralysis (NFIP) proposed vaccinating all grade 2 children (if consent was given), and leaving grades 1 and 3 for control.

Would this study be considered a valid experiment?

Again, this creates a **bias against** the vaccine, since the study likely would conclude that the vaccine does not work as well as the vaccine really does.

(d) (real data) Randomized control.

Offer many children the ability to participate in the experiment, but do not tell them if they are given treatment or placebo.

Also, do not tell doctors or nurses.

How should it be decided as to who goes into which group? Income? Health?

Suppose the doctors offered the drug to **sickest** patients.

Statistics tells us that with large enough samples, **randomized, controlled** experiments determine whether or not a **treatment** (e.g., drug, vaccine) works.

If the treatment group has a higher success rate than the placebo group, we need to decide if this was due to chance or due to a successful treatment.

Typically, we require overwhelming evidence that the treatment was successful before marketing the new vaccine.

Salk vaccine trial of 1954

Rate of polio per 100,000

Randomized controlled double-blinded experiment			The NFIP study		
	size	rate		size	rate
treatment (high hyg.)	200,000	28	Grade 2 (vaccine, high hyg.)	225,000	25
control (high hyg.)	200,000	71	Grades 1 & 3 (control, average hyg.)	725,000	54
no consent (low hyg.)	350,000	46	Grade 2 (no consent, low hyg.)	125,000	44

□

1.2 Pictorial and Tabular Methods in Descriptive Statistics

Graphs for Quantitative Variables

Stem-and-Leaf Plots

Example: Each stem represents tens, and each leaf represents ones, in this example.

Stems	Leaves
0	5 5 6
1	0 0 3 7
2	1 4 5 6 9
3	0 3 4 8
4	1 3 6 6
5	7 8
6	2
7	
8	
9	5

Stem-and-leaf display can be used to observe the shape (and location and spread) of the data or detect *outliers*.

An **outlier** is an observation that falls well above or well below the overall bulk of the data.

Example: In a random sample of heights of 100 women, a 6-foot-8-inch-tall woman is recorded.

Dot Plots

Example: Construct the *dot plot* for the following data on personal income (in thousands of dollars): 35, 49, 70, 21, 49, 80, 57, 160.

Histogram

A **histogram** is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

Discrete case with a small number of possible outcomes:

Example: Construct a relative frequency histogram for data on household sizes.

# of people	frequency
1	34
2	51
3	42
4	30
5	20
6	13
7	4
8	6

Discrete case with a large number of possible outcomes, OR *continuous* case:

Example: Construct a relative frequency histogram for income (in terms of hourly wage).

income level	# of people
[\$0, \$5)	35
[\$5, \$10)	50
[\$10, \$15)	70
[\$15, \$20)	115
[\$20, \$30)	100
[\$30, \$50)	130

Note: The area of a **sample histogram** for continuous data is one.

Sample histograms and population histograms

Compare and contrast sample histograms with population histograms.

The shape of a distribution

A histogram might be described as

1. unimodal
2. bimodal
3. multimodal

A **sample or population histogram** might be described as

1. symmetric
2. skewed to the right
3. skewed to the left

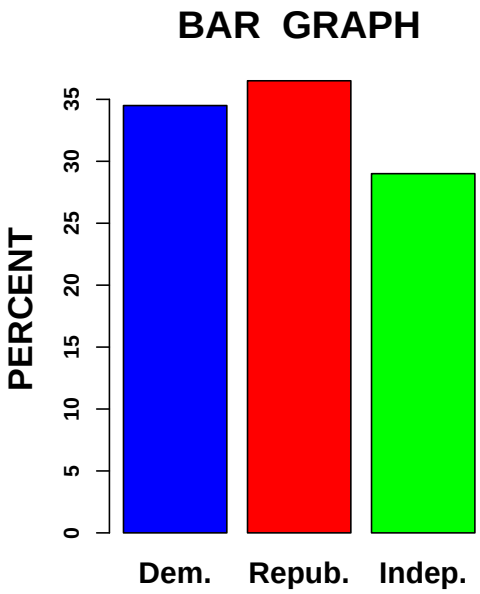
Graphs for Qualitative Data

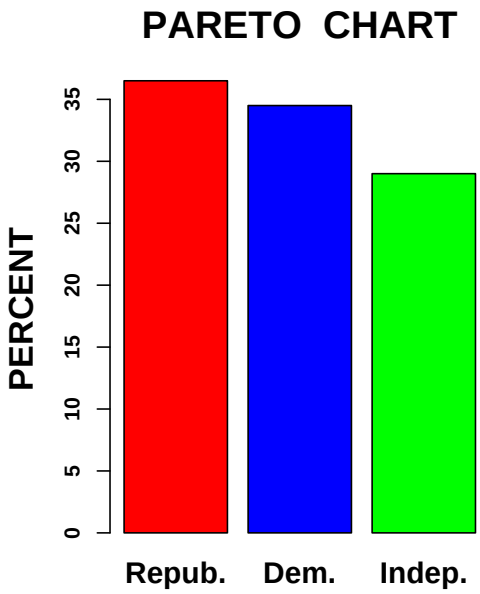
Bar graphs, Pareto charts, and **pie charts** graph the relative frequency of categorical data.

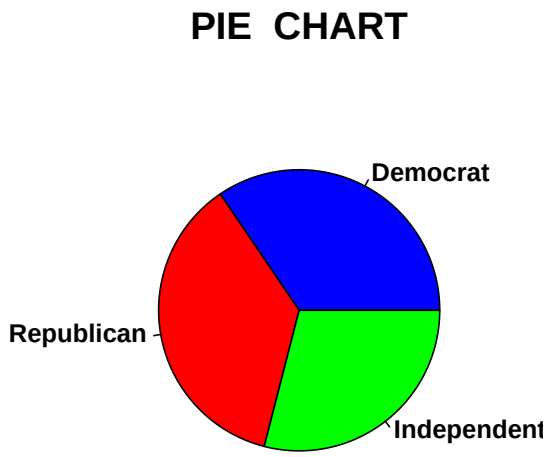
relative frequency$= \dfrac{\text{frequency in the category}}{\text{total \# of observations}}$

Example: Poll of 400 students at a university, we have 138 Democrats, 146 Republicans, and 116 Independents.

Construct the **bar graph, Pareto chart,** and **pie chart.**







What is the **mode** in the above graph?

1.3 Measures of Location

Summarize the data by taking an average.

Example: Suppose that a student has the following 6 quiz grades: 95, 100, 85, 89, 10, 97.

In general, the *sample mean* of n observations on x is denoted

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

The **sample median**, denoted $\tilde{x}$, is the middle value when the measurements are arranged from smallest to largest. For an even number of data, $\tilde{x}$ is the average of the two middle values.

When outliers exist or at least one tail of the distribution is heavy, then the **sample median** typically is preferred over the **sample mean**, as a measurement of center. Otherwise, the **sample mean** typically has less variability and is preferred over the **sample median**.

Example: Explain how to estimate the average income, μ, in Virginia.

For **symmetric** distributions (with a finite mean, μ), the population mean and median are equal.

For **right** skewed distributions, the population mean is greater than the population median.

For **left** skewed distributions, the population mean is less than the population median.

Trimmed Means

Here, we average, say, the middle 80% of the ordered observations; i.e., remove the smallest 10% of the observations and the largest 10% of the observations and take an average.

This average is called a **10% trimmed mean**.

A **trimmed mean** is _____ sensitive to outliers than the sample **mean** but _____ sensitive to outliers than the sample **median**.

Example: Suppose in a sporting event 10 judges record an individual's speed in a race. The scores (in seconds) are the following: {56.2, 56.3, 56.4, 56.4, 56.5, 56.5, 56.6, 56.6, 56.8, 57.4}. *Note that the observations are already ordered.* Are there any outliers?

Compute the 10% trimmed mean.

□

Categorical Data and Sample Proportions

Example: Suppose the governor is interested in proportion of Virginian adults who approve of his budget.

Let p be the population proportion of Virginian adults who approve of his budget.

A **sample proportion** is the **sample mean** of zeros and ones, where "one" indicates success and "zero" indicates failure.

The Mode of a Data Set

Exercise 1.72, p. 49: The **mode** of a numerical data set is the value that occurs most frequently in the set.

(a) Determine the mode for the data given in exercise 1.71. The ordered data are {0.78, 0.81, 0.81, 0.85, 0.85, 0.86, 0.92, 0.92, 0.93, 0.93, 0.93, 0.93, 0.95, 0.95, 0.96, 0.96, 1.00, 1.05, 1.06, 1.06}.

(b) For a **categorical** sample, how would you define the modal category?

Example: List additional *specific* examples where the **modal category** might be of interest.

□

Statistics joke: What happens when a JMU student transfers to UVa?

1.4 Measures of Variability

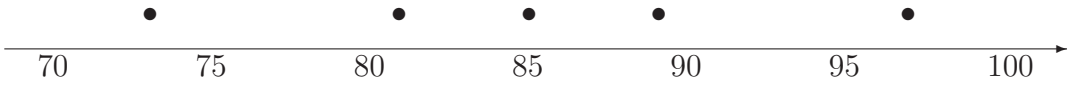
The **sample range** is the maximum value minus the minimum value; this measure is reasonable for small data sets, but not large ones.

Example: Suppose a sample of 100 incomes among employed Virginians might

have the smallest value of \$10,000 and the largest value of \$300,000.

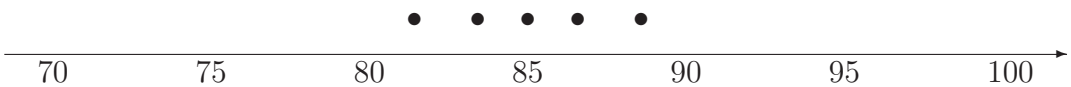
Consider examples with light-tailed distributions and no outliers in the data set.

Example: Grades by instructor **Jill** for 5 students on a chemistry exam are {81, 85, 97, 73, 89}.



Dot plot of grades for Jill's class

Example: Grades by instructor **Susan** for 5 students on a chemistry exam are {84, 88, 85, 86, 82}.



Dot plot of grades for Susan's class

Which instructor do you prefer?

Let $S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$.

$$\text{sample variance} = s^2 = \frac{S_{xx}}{n-1} = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$\text{sample standard deviation} = s = \sqrt{s^2}$$

Example: Jill's class

Recall: $\bar{X}$ "gets close" to μ as n gets large.

Likewise, s^2 "gets close" to the population variance, σ^2, as n gets large.

Similarly, s "gets close" to the population standard deviation, σ, as n gets large.

Remark: Adding a constant to the data does not affect s or s^2. For example, adding 10 points to everyone's grade increases $\bar{X}$ and the sample median by 10 points but does not affect s, the spread.

Example: Everyone at a company receives a \$1000 raise. Compute the **new** sample standard deviation, in comparison to the **old** sample standard deviation.

□

Remark: Multiplying data by a constant c changes s by a factor of $|c|$.

Example: At a company the salaries are \$50,000, \$55,000, and \$60,000.

Example: Everyone at a company receives a raise of 10%. Compute the **new** sample standard deviation, in comparison to the **old** sample standard deviation.

□

Example: Consider the data {8, 8, 8, 8, 8, 8}.

Remark: *Later in the textbook:* If the distribution is approximately **normal** (bell-shaped), then approximately **68%** of the observations lie between $(\mu - \sigma)$ and $(\mu + \sigma)$.

Prove the shortcut formula: $s^2 = \frac{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}{n-1}$.

□

Fourth Spread

To determine the **fourth spread**, denoted f_s, first determine the **lower fourth** and the **upper fourth**.

The **lower fourth** (first quartile) of a data set is the median of the ordered observations that lie

(a) **below** the position of the overall median if n is **even**,

(b) **below** and **including** the position of the overall median if n is **odd**.

The **upper fourth** (third quartile) of a data set is the median of the ordered observations that lie

(a) **above** the position of the overall median if n is **even**,

(b) **above** and **including** the position of the overall median if n is **odd**.

The **fourth spread** is the **upper fourth minus lower fourth**.

The **5-number summary** is {minimum, lower fourth, median, upper fourth, maximum}.

The **5-number summary** divides the data into four (roughly) equal sections (fourths).

Example: Consider the following 12 observations:

{45, 48, 53, 103, 160, 10, 68, 70, 55, 58, 75, 77}.

How do the above results change if we replace 160 by 1,000,000?

□

Identifying potential outliers using the fourth-spread criterion:

An observation is an **outlier** if it is at least $1.5f_s$ from its nearer fourth.

An observation is an **extreme outlier** if it is at least $3f_s$ from its nearer fourth.

An outlier which is NOT extreme is a **mild outlier**.

Box Plots That Show Outliers

Procedure:

1. Draw rectangle with edges at lower and upper fourths.
2. Draw a line through the box at the sample median.
3. Draw **whiskers**; i.e., lines from the edge of the box to the most extreme observation which is not an outlier.
4. Draw **closed circles** to represent **mild outliers**, and draw **open circles** to represent **extreme outliers**.

Previous example: Consider the following 12 observations:

{45, 48, 53, 103, 160, 10, 68, 70, 55, 58, 75, 77}.

Check for outliers on the **left**.

Check for outliers on the **right**.

□

Which measures of center and spread should one use?

Suppose a data set has outliers (or the distribution has at least one heavy tail).

Suppose a data set seems to be from a distribution which is normal (or approximately normal).

Exercise 1.65, p. 48: The following data on HC (hydrocarbon) and CO (carbon monoxide) emissions for one particular vehicle is given below.

HC	*(gm/mi)*	13.8	18.3	32.2	32.5
CO	*(gm/mi)*	118	149	232	236

Let X be the *HC* population, and let Y be the *CO* population.

(a) Compute the sample means and sample standard deviations.

(b) Compute the **sample coefficient of variation**, which is the *sample standard deviation* divided by the *sample mean*.