1.1 What is Statistics?

Definition of Statistics

Statistics is the study of how to *collect*, *organize*, *analyze*, and *interpret* numerical information from data. **Descriptive statistics** involves methods of organizing, picturing and summarizing information from data. **Inferential statistics** involves methods of using information from a sample to draw conclusions about the population.

Keep in Mind:

* Statistical inferences are no more accurate than the data they are based on (weakest link).
* Statistical results should be interpreted by one who understands the methods used as well as the subject matter.

Individuals and Variables

Individuals are the people or objects included in the study. A **variable** is the characteristic of the individual to be measured or observed.

For example, if we want to do a study about the people who have climbed Mt. Everest, then the individuals in the study are the actual people who made it to the top. The variables to measure or observe might be the height, weight, race, gender, income, etc of the individuals that made it to the top of Mt. Everest.

Variables: Quantitative vs. Qualitative

A **quantitative variable** has a value or numerical measurement for which operations such as addition or averaging make sense. A **qualitative variable** describes and individual by placing the individual into a category or group such as male or female.

Example A

State whether the data is qualitative or quantitative.

1. The color of a person's eye.

2. The height of a person in inches.

3. The a/b/c/d responses on a questionnaire.

Population Data and Sample Data

In **population data**, the variable is from every individual of interest. In **sample data** the variable is only from some of the individuals of interest.

Guided Exercise 1

Television station QUE wants to know the proportion of TV owners in Virginia who watch the stations new program at least once a week. The station asked a group of 1000 TV owners in Virginia if they watch the program at least once a week.

a. Identify the individuals in the study.

b. Identify the variable.

c. Do the data comprise a sample? If so, what is the underlying population?

Yes. The implied population is the responses (watch/not watch) of all TV owners in Virginia.

d. Is the variable quantitative or qualitative?

e. Identify a quantitative variable that might of interest.

Example 1

Suppose the Hawaii Department of Agriculture wishes to conduct a study of pineapples in an experimental field.

a. What are the individuals in the study?

b. What are some possible quantitative variables of interest?

The weight of each pineapple in the experimental field.
The maximum girth of each pineapple in the experimental field.

c. What are some possible qualitative variables of interest?
The taste [color] of every pineapple in the field.

Levels of Measurement

1. **Nominal Level** (in name only): Qualities with no ranking/ordering; no numerical or quantitative value. Data consists of names, labels and categories.

a. Taos, Acoma, Zuni and Cochiti are names of four native American pueblos.

b. Car colors for a certain model are: red, silver, blue and black.

2. **Ordinal Level**: Can be arranged in some order, but the differences between the data values are meaningless.

a. Of 17 fishing reels rated: 6 were rated good quality, 4 were rated better quality, and 7 were rated best quality.

b. Out of a high school class of 319, Walter ranked 4th, June ranked 12th, and Jim ranked 20th.

3. **Interval Level**: Data values can be ranked and the differences between data values are meaningful. However, there is no intrinsic zero, or starting point, and the ratio of data values are meaningless. Note: Calendar dates and Celsius & Fahrenheit temperature readings have no meaningful zero and ratios are meaningless.

a. The years in which democrats won presidential elections.

b. Body temperature in degrees Celsius (or Fahrenheit) of trout swimming in the North River.

c. Building A was built in 1284, Building B in 1492 and Building C in 5 bce.

4. **Ratio Level**: Similar to interval, except there is a true zero, or starting point, and the ratios of data values have meaning.

a. Core temperature of stars measured in degrees Kelvin.

b. Time elapsed between the deposit of a check and the clearance of that check.

c. Length of trout in the North River.

Levels of Measurement

1. Nominal Level (in name only): Qualities with no ranking/ordering; no numerical or quantitative value.
2. Ordinal Level: Can be arranged in some numerical order, but the differences between the data values are meaningless.
3. Interval Level: Data values can be ranked and the differences between data values are meaningful. However, there is no intrinsic zero, or starting point, and the ratio of data values are meaningless.
4. Ratio Level: Similar to interval, except there is an inherent zero, or starting point, and the ratios of data values have meaning.

Guided Exercise 2

State the level of measurement for each of the following:

a. The senator's name is Sam Wilson.

b. The senator is 58 years old.

c. The senator was elected in 1963, 1969, 1981, and 1994.

d. His taxable income is $278,314.19

e. Of 1100 voters in his district: 400 strongly favor his bill; 300 favor; 200 neutral; 150 do not favor, and 50 strongly do not favor his bill.

f. The senator is married.

g. The senator had divorces in 1965 and 1982.

h. A newspaper ranked the senator 7th for his voting record on public education.

1.2 Random Samples

Simple Random Sample

A **simple random sample** of *n* measurements from a population is one selected in such a manner that

1. every sample of size *n* from the population has equal probability of being selected, and
2. every member of the population has equal probability of being included in the sample.

Example C

Consider the population of all coyotes in the western U.S. The sample of that population that ranchers observe is largely the coyotes that prefer to live near a ranch; they also like to eat lamb. The ranchers concluded that all coyotes are dangerous and got the government to assist in distributing a poison bait to reduce the overall coyote population. The overall population of coyotes was reduced, but the ranchers still lost almost as many sheep as before. Why?

a. Is the sample the ranchers observed a random sample?

b. If not, is it safe to use the results to describe the entire population?

c. Is the idea of reducing the size of the entire population justified based on the ranchers experience?

Note: Further study has shown that coyotes who eat sheep are consistent in their preference for sheep, whereas the majority of coyotes in the wild stick to foods found in the wild.

Example D

Do the following procedures give a random sample for the entire population of New York City? Why / Why not?

a. Select every third woman entering a beauty shop.

b. Select every third person coming out of a boxing match at Madison Square Garden.

Example 6: Random Number Table

To determine if the latest shipment of 500 Toyotas meets emission standards, a random sample of 30 is chosen and tested. How can you make sure the sample of 30 chosen is a random sample? See the random number table in Appendix II, page A9.

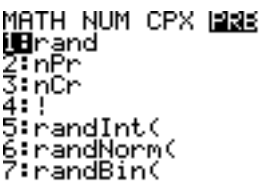
92630	78240	19267	95457	53497	23894	33708
79445	78735	71549	44843	26104	67318	00701
28703	51709	94456	95761	33393	99411	25270
17138	14280	51333	51287	99281	81017	55137
23501	85846	88472	64688	37818	96593	28661

Simulation

A **simulation** is a numerical facsimile or representation of a real-world phenomenon.

Random Number Generator on the TI-83

1. MATH / PRB / 1:rand
2. MATH / PRB / 5:randInt(min, max , number of integers)



Example E

Use a random number table to simulate each of the following.

a. Choose the numbers for the next lottery. That is, randomly choose six numbers from 1 to 52.

92630	78240	19267	95457	53497	23894	33708
79445	78735	71549	44843	26104	67318	00701
28703	51709	94456	95761	33393	99411	25270
17138	14280	51333	51287	99281	81017	55137
23501	85846	88472	64688	37818	96593	28661

b. The outcomes of tossing a die 20 times.

92630	78240	19267	95457	53497	23894	33708
79445	78735	71549	44843	26104	67318	00701
28703	51709	94456	95761	33393	99411	25270
17138	14280	51333	51287	99281	81017	55137
23501	85846	88472	64688	37818	96593	28661

c. Assign the letters a, b, c, d, or e as the correct response on a 10-question multiple-choice exam.

92630	78240	19267	95457	53497	23894	33708
79445	78735	71549	44843	26104	67318	00701
28703	51709	94456	95761	33393	99411	25270
17138	14280	51333	51287	99281	81017	55137
23501	85846	88472	64688	37818	96593	28661

Other Sampling Techniques

1. **Stratified Sampling**

Stratify (divide) the population by a common characteristic (such as age, class, gender, etc.) and take a random sample of each stratum [often in accordance to their percent of the population].

e.g. divide the population by race and survey a number of randomly selected individuals from each race (often in the same proportion they occur in the overall population).

2. **Systematic Sampling**

Used when the elements of the population are arranged in a natural sequential order.

e.g. take every 5th person coming through a cafeteria line.

3. **Cluster Sampling**

Used extensively by government and research organizations. Randomly select a sample of pre-existing sections or clusters (often geographic sections). Every member of the cluster is included in the sample/survey.

e.g. randomly select 30 schools and survey every student in each school.

4. **Convenience Sampling**

Uses data that are conveniently and readily obtained.

e.g. walk outside and survey the first 100 people that will talk to you.

1.3 Experimental Design

Basic guidelines for planning a statistical study

1. Identify the individuals or objects of interest.
2. Specify the variables to measure or observe.
3. Determine if you will use the population or a representative sample. Decides on a viable sampling method.
4. Collect the data.
5. Use the appropriate descriptive statistics methods (Chapters 2, 3 and 10) and make decisions using the appropriate inferential statistics methods (Chapters 8-12).
6. Note any concern you might have about your data collection methods and list any recommendations for future studies.

Ways to Produce Data

1. **Census**: Measurements or observations of the entire population.
2. **Sampling**: Measurements or observations from a representative part of the population – i.e. simple random sample.
3. **Simulation**: Numerical modeling of real-world phenomena.
2. **Experiment**: Impose a *treatment* and measure/observe the change in the variable of interest.
3. **Observational Study**: Observations and measurements are made in a way that does not change the response or the variable being measured.

Example F

Which data gathering technique might be best suited for each of the following situations?

a. Study the effect of stopping the cooling process in a nuclear reactor.

b. Study the amount of time college students taking a full course load would spend watching TV.

c. Study the effect of a calcium supplement given to young girls on bone mass.

d. Study the credit hour load of each student enrolled at Palomar at the end of the add/drop period.

Experiment and Observational Study

In an **observational study**, observations and measurements of individuals are conducted in a way that does not change the response or the variable being measured. In an **experiment**, a treatment is deliberately imposed on the individuals in order to observe a possible change in the response or variable being measured. In a study, often the individuals are divided into two groups – a control group and an experimental group. The **experimental group** is actually given the treatment and the **control group** is not given the treatment.

Example 8

In 1778 Captain James Cook introduced goats to the Hawaiian Islands. It was later observed that the Silver Sword plant appeared to be less and less common. Botanists suspected the goats to be the cause and conducted a statistical study. They set up stations around the islands with similar climate and soil conditions. Each station consisted of two plots of land, one with a fence around it to keep the goats out. Identify the

a. treatment

b. experimental group

c. control group

Randomized 2-Treatment Experiments – Placebo Effect

The **placebo effect** occurs when a subject in the *control group* receives no treatment, but believes she is in fact receiving treatment and responds favorably.

Example 9

For more than a decade doctors have been using lasers to drill holes in the heart to reduce angina (chest pain). Many patients reported lasting and significant relief from the procedure. To test if the relief is due to the placebo effect or not a randomized two-treatment experiment was completed. A group of 298 volunteers with severe, untreatable chest pain were randomly assigned to get the laser treatment or not. The patients were sedated but awake. They could hear the doctors discuss the laser process. Each patient thought he or she was receiving the treatment.

Conclusion: The laser patients did well. But shockingly, the placebo patients showed more improvement in pain relief. The medical impacts of this study are still being investigated.

The study of example 9 has many features of good experimental design. There is a **control group** who received the dummy treatment and the **experimental group** who received the actual treatment. The control group is used to account for influence of **lurking** or **confounding** variables that might account for some of the changes observed.

Randomization is used to assign individuals to the two treatment groups. This helps prevent bias in selecting members for each group.

Replication of the experiment on many patients reduces the possibility that the differences in pain relief for the two groups occurred by chance alone.

Double-Blind Experiment

A **double-blind experiment** is one in which neither the patients, nor the observers know which subjects are receiving the treatment. They help control the biases of doctors and researchers.

Survey Questions and their Pitfalls

Gathering data by simply asking people questions is the essence of a survey.

Cautions with Survey Questions

1. Can you convert the data to numbers?
2. Is the wording of the question unbiased?
3. Voluntary responses often over-represent strong [negative] opinions.
4. Can you expect a truthful response?
5. Is the sample representative of the population?
6. Hidden Bias: How would you design/administer a survey of Palomar College Students?
7. Other Variables may establish a Cause-Effect Relationship. Since, in general, events with higher ticket prices have higher attendance, should you raise ticket prices to increase revenue
8. Over Generalizing Results
   Results of drug experiments on lab rats cannot be generalized to other animals.
9. How do you handle a significant proportion of non-responses.
10. Poll “registered” versus “likely” voters.

Guided Exercise 4

Comment on the usefulness of data collected as described:

a. A uniformed law officer asks a group of college freshman their name and if he/she has used drugs in the last month.

b. Jane saw data that show that cities with more homeless people have more low-income housing. Does building low-income housing produce more homelessness?

c. A survey about food in the cafeteria was conducted by having survey forms available at the register. A drop box for the forms is outside the cafeteria.

d. Extensive studies on coronary problems were conducted using men over age 50 as the subjects.