Chapter 1: Introduction to Statistics

STATISTICS IS THE STUDY OF DATA

Things we do with data:

Gather/collect data → **Organize and Summarize** → **Analyze** → **Interpret**

Definitions:

Population	**Sample**
Example The population of ALL De Anza College students enrolled in a recent quarter	Example A sample of De Anza college students, such as 200 randomly selected students

Variable: the description in words of the characteristic of interest

Example of Variable: "the age of a De Anza College student"

Data: the information collected about the variable for individuals in the population or sample.

Example of Data: 19, 23, 42, 18, 22, 28, 31, 18, 21

- *The variable is a sentence explaining the question you are asking in order to obtain information,*
- *The data are the information obtained as answers to the question.*

After the data is collected, we use the data to calculate number that summarizes and describe the characteristic for the population or sample. This number is called a parameter or a statistic.

Parameter	**Statistic**
Examples of Parameters:	Examples of Statistics:
PROPORTION: In a recent quarter, 39% of **all** De Anza College students were over age 25	**PROPORTION*:*** 41% of a random **sample** of 200 students were over age 25.
AVERAGE*:* In a recent quarter, the average age of **all** De Anza College students was 27.1 years	**AVERAGE*:*** The average age of a random **sample** of 200 De Anza College students was 28.3 years
MEDIAN: In a recent quarter, half of **all** De Anza College students were 22 years old or younger	**MEDIAN:** In a **sample** of 200 De Anza College, 50% of the students were 22 years old or younger

Does a sample statistic always have the same numerical value as the population parameter? Why or why not?

Are sample statistics equal for all samples taken from the same population? Why or why not?

CHAPTER 1: TYPES OF VARIABLES AND DATA:

EXAMPLE 1: Suppose we are studying the commutes of De Anza College students to school from home.

What is the population?_____

Give one possible example of a sample:_____

Some examples of variables and data**:**

Variable: "the distance that a (one, individual) student commutes to De Anza College "
Data**:** 2.5 miles, 8.4 miles, 0.25 miles, 52 miles, . .

Variable: "how a De Anza student commutes to school
Data: car, car, bus, bike, walk , car, bus, bus, car , bike, car

Quantitative

- **Quantitative Discrete**

- **Quantitative Continuous**

Qualitative (also called Categorical):

LEVELS OF MEASUREMENT

Nominal: qualitative data: "how a student commutes to school" or "a student's major"

Ordinal: data that can be ranked:

- 1st place, 2nd place, 3rd place, . . . last place, such as in a race or competition
- preference ratings such as
  1 = strongly agree, 2 = agree, 3 = indifferent, 4 = disagree, 5 = strongly disagree
  where there is no set measurement associated with each of the different responses.

Interval: data has constant differences for intervals
But there is no definite natural "0" value that can be used as the base point for comparisons
Example: Temperature degrees Celsius or degrees Farenheit:
These are arbitrary scales; comparisons differ when using °C versus °F
80 degrees F = 26. 7 degrees C and 40 degrees F = 4.4 degrees C
Is it twice as hot (80 degrees F is two times 40 degrees F)
OR is it 6 times as hot (26. 7 degrees C is about 6 times 4.4 degrees C)?
Due to lack of a natural zero base point, the ratio comparisons do not make sense.

Ratio: data has constant differences for intervals
AND there is a natural "0" value that can be used as a base point for ratio comparisons
Example: distance travelled from home to school;
Asaf travels 3 miles and Ben travels 6 miles. Ben travels twice as far as Asaf
Example: It took Tran 1.5 hours to do homework and took Li 3 hours to do homework.
Tran took half as long as Li to do homework.

EXAMPLE 2: Identify the type of variable and data and identify the level of measurement:

a. Variable: "the distance that a (one, individual) student commutes to De Anza College "
Data**:** 2.5 miles, 8.4 miles, 0.25 miles, 52 miles, . .

b. Variable: "the number of days per week that a student comes to campus"
Data**:** 2, 5, 4, 3, 1, 2, 5, 4, 4, 3, 3, 1, 2, . . .

c. Variable: "how a De Anza student commutes to school
Data: car, car, car, bus, bike, bus, walk , car, bus, bus, car , bike, car, bus

d. Variable: Was your commute time this Friday (1) faster than, (2) same as, (3) slower than usual?
Data: 1, 1, 3, 2, 1, 3, 2, 2, 2, 1, 2, 1, 3, 2, 2, 3, 2 . . .

EXAMPLE 3: 1500 randomly selected registered voters in a large city are asked the following questions:
What is your age? What is your annual income? Do you intend to vote?
If you intend to vote, do you intend to vote in person or by absentee ballot?
A newspaper article discussing expected voter turnout reported that for this sample, the average annual income is $61,000, the average age is 44, 63% of the registered voters sampled intend to vote, 42% intend to vote by absentee ballot while 58% intend to vote in person at the polling place.

a. Describe the sample.

b. What would be the appropriate target population for this sample?

c. What are the variables? List each variable as quantitative or qualitative:

The QUANTITATIVE variables are:	The QUALITATIVE variables are:

d. Make up examples of data that might have been given by somebody in the sample.

e. Are the values reported in the newspaper statistics or parameters? Explain.

EXAMPLE 4: A weight loss clinic is studying the exercise habits of its clients.
A survey of all client records shows that 64% of all clients exercise regularly.
A survey of a random sample of 100 clients who exercise regularly shows that of those clients in the sample, the average amount of time that clients exercise each week is 2.5 hours, and that they exercise on average 4 days each week.

a. For each variable listed below, indicate if it is qualitative, quantitative discrete, or quantitative continuous?

Whether a client exercises regularly: _____

The amount of time a client exercises each week: _____

The number of days a client exercises each week: _____

b The number 64% above is a : A. data B. statistic C. parameter D. sample E. population

c. The numbers 2.5 and 4 above are: A. data B. statistics C. parameters D. sample E. population

d. A client responds that she exercises 3 days a week for 2.4 hours total; the numbers 3 and 2.4 are _____

STATISTICAL METHODS

Descriptive statistics

Inferential statistics

TYPES OF STATISTICAL STUDIES:

A. Census vs Sampling

1. Census:

2. Sampling:

B. Observational Study vs Experiment

1. Observational Study:

2. Experiment:

CHAPTER 1: TYPES OF STATISTICAL STUDIES / CRITICAL EVALUATION

Vocabulary for Experiments: Treatment Placebo Response
Control Group Blinding Double Blind

EXAMPLE 5: An example of an experiment to test a new drug

- A study is done to test a new drug that is supposed to relieve pain.
- The patients enrolled in the study are randomly divided into two groups.
- One group is given the new drug, called the _____.
- The other group is called the _____ group and is given the _____ instead of the treatment.
- The _____ of both groups is measured to determine if the drug is more effective at relieving pain than not receiving the drug.
- The study is _____ because neither the patient nor the doctor knows who is receiving the new drug and who is receiving the placebo.
- If a patient develops problems, the doctor works with the study administrator who knows who is receiving the drug and who is receiving the placebo.
- The doctor, the study administrator, and a statistician are part of a team of people who evaluate the effectiveness of the drug based on the results of the study.

How this relates to the real world: As patients, we would not want to use a drug or a medical device unless it has been shown to be safe and effective. In the Silicon Valley area, there are many biotech companies developing drugs and medical devices. They need to conduct studies of the safety and effectiveness of these drugs or devices. UC Santa Cruz Extension has a professional "clinical trials" certification program. People enrolling in this program are expected to already have college degrees in related fields and in this program they study topics such as how the law applies to medical studies, the ethics of medical studies, statistics, design of experiments and other topics.

CRITICAL EVALUATION OF STATISTICAL STUDIES AND RESULTS

IT IS VERY IMPORTANT TO THINK CRITICALLY ABOUT THE VALIDITY AND RESULTS OF STATISTICAL STUDIES RATHER THAN BLINDLY BELIEVING THE RESULTS OF ALL STUDIES.

Common Problems in Statistics to beware of:

- Problems with Samples:
  - A sample should be representative of the population.
  - A sample that is not representative of the population is called "biased".
  - Non-response or refusal of subject to participate
  - Self-Selected Samples might be biased
  - Sample Size Issues
- Collecting data or asking questions in a way that influences the response
- Causality: A relationship between two variables does not necessarily imply that one **causes** the other
  They may both be affected by some other variable.
- Self-Funded or Self-Interest Studies
- Misleading Use of Data: improperly displayed graphs, incomplete data, lack of context, not enough information given to understand the data
- Confounding: when the effects of multiple factors on a response can not be separated, it becomes difficult or impossible to draw valid conclusions about the effect of each factor.

EXAMPLE 6:

Study I: Employees of a company are randomly divided into two groups. Group A gets classroom training from an instructor who is available to help and answer questions; Group B gets training via online software with an online discussion board available to get help and answers to questions.

Study II: Researchers are studying whether an early retirement age increases the rate of memory problems in senior citizens. A survey of retired senior citizens showed that those who retired earlier tended to have a higher rate of memory problems after retirement than those who retired at an older age.

Study III: 300 randomly selected individuals are asked if they had been on a diet in the last 8 weeks and how much their weight has changed over the last 8 weeks. Weight change for dieters and non-dieters are compared.

Study IV: 100 individuals are put on a low fat diet, 100 on a low carb diet and 100 eat their normal diet. Their weight change over an 8 week period is recorded.

a. For each of the above, what type of study is it?

I:_____ II:_____

III:_____ IV:_____

b. What problem can you see in Study II?

c. Which weight loss study (III or IV) do you think would give the best information about the effect of diet on weight loss? Why?

EXAMPLE 7: A large city is proposing a parcel tax to support education. Each property owner would be assessed a tax of $100 per property per year. The parcel tax will be voted on by voters in the next election. It will pass if 2/3 of the voters vote in favor of the tax.

I. A group of parents and teachers supporting the parcel tax randomly select and call residents in the city. They identify themselves as members of the Parent Teachers Association for the school system and ask the person who answers the telephone call if they support the parcel tax.

II. A TV news station in the city conducts a "Facebook" survey. Viewers are asked whether they favor or oppose the tax and are given instructions to visit the TV stations Facebook page to respond about their opinion. The poll is publicized and responses are solicited by announcements on the TV station's evening news programs.

III. A professional polling organization conducts a survey by randomly calling selected residents in the city. If the resident is a registered voter, he or she is asked his/her their opinion about the proposed parcel tax. They are asked whether they favor the tax, oppose the tax, or have no opinion. These three choices are presented to the individual in random order, so that not all respondents hear the choices in the same order.

a. Which survey do you think would produce the most accurate prediction of the election results?

b. For each of the other two surveys, what problems do you think there might be with the information obtained? Explain your reasoning for your answers to these questions.

CHAPTER 1: TYPES OF SAMPLES

A sample is a part of or a subset of a population.
A sample should be representative of the population
A sample that is not representative of the population is called "biased".

Vocabulary and Concepts:

Sampling Error: Random error obtained by using part of the population to represent the whole population

Non-Sampling Error: Non-random error: improper data collection recording or sampling techniques, bias,

Random Sample: Every person has equal chance of being included in the sample

Sampling Methods:

- **simple random sample**

- **systematic sample**

- **stratified sample**

- **cluster sample**

- **convenience sample**

EXAMPLE 8: *(a -f from Example 1.6 Collaborative Statistics, Chapter 1, B. Illowsky & S. Dean;free at www.cnx.org)*

Determine the type of sampling method used:

a. To form a recreational soccer team, a soccer coach randomly selects 6 players from a group of boys ages 8 to 10, 7 players from a group of boys ages 11 to 12, and 3 players from a group of boys age 13 to 14. _____

b. For a survey of human resource (HR) personnel at high tech companies, a pollster interviews all HR personnel in each of 5 randomly selected high tech companies. _____

d. A medical researcher for a hospital interviews every third cancer patient from a list of cancer patients at that local hospital. _____

e. A high school counselor uses a computer to generate 50 random numbers and then selects students whose names correspond to the numbers. _____

f. A student interviews classmates in his algebra class to determine how many pairs of jeans a student owns, on average. _____

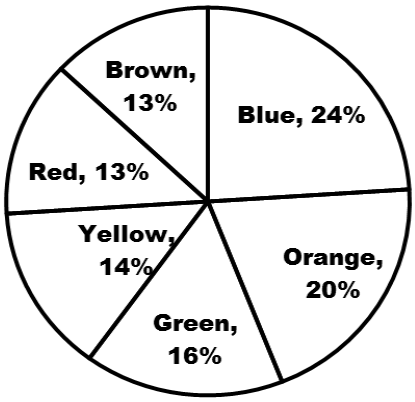
g. In a study to learn what types of after school child care are used in their district, a school district administrator randomly selects 6 classes at each school and surveys all parents with children in the selected classes. _____

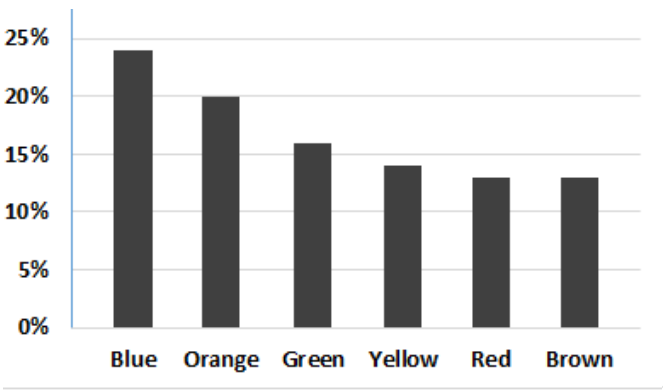
EXAMPLE 9: The Mars Candy company used to post the distribution of colors of their candies on their website. This example is based on the distribution of colors for milk chocolate m&m candies that appeared on their site.

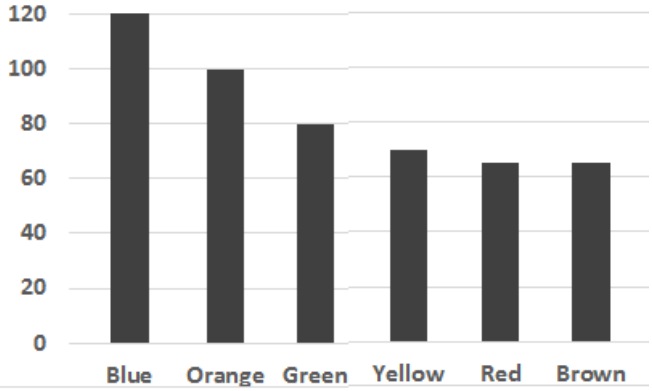
Suppose that a sample of 500 m&m's was categorized by color.
The variable is "the color of one m&m candy"
Since the data are colors, the data are qualitative.
A list of the data might read "red, red, brown, blue, yellow, blue, orange, green, red, . . . , green, brown"

To organize qualitative data, we can count the number of in each category (and use a table to display counts and/or percents, and we can use bar or pie charts to display the data graphically.

m&m candies	Number of each color	Percent
Blue	120	24%
Orange	100	20%
Green	80	16%
Yellow	70	14%
Red	65	13%
Brown	65	13%
Total	500 Candies	100%







It is usually helpful to create several types of graphs and examine them to choose those that best display the information for the situation being analyzed.

Chapter 1 in the textbook has examples of graphical displays of qualitative data and discusses some of the pros and cons of the various types of graphs, as well as how to organize data to best display the information.

CHAPTER 1: NUMERICAL SUMMARIES OF QUANTITATIVE DATA: FREQUENCY DISTRIBUTIONS

Quantitative (numerical) data can be summarized in a frequency distribution table. The data may be presented individually (ungrouped) or grouped into intervals.

- A frequency distribution counts the number of data items that fall into each interval. **Frequency = count = number of data values that lie in the interval.**
- A relative frequency distribution shows the proportion (fraction or percent) of data items in each interval. **Relative Frequency = proportion of data values that lie in the interval = $\frac{\text{Frequency}}{\text{Number of Observations}}$.**
- **Cumulative Relative Frequency = sum of relative frequencies for intervals up to and including current interval**

EXAMPLE 10:
Individual Data Values
(ungrouped data)

Plants are being studied in a lab experiment. For a sample of 16 plants, the number of flowers on a plant are:
2,5,3,1,2,4,1,2,3,1,1,2,7,4,2,3

X = Number of Flowers	Frequency	Relative Frequency	Cumulative Relative Frequency
1			
2			
3			
4			
5			
6			
7			

a. What percent of plants had exactly 3 flowers?

b. What percent of plants had <u>at most</u> 3 flowers?

c. What percent of plants had <u>more than</u> 3 flowers?

d. What percent of plants had <u>at least</u> 5 flowers?

EXAMPLE 10: Grouped Data Values

Birthweights, in grams, for a sample of 400 newborn babies born at a hospital.

Babies with low birth weight often have serious health problems.
What percent of babies born at this hospital had low birth weights of under 2500 g.?

Weight (grams) Interval Class Limits	Frequency	Relative Frequency	Cumulative Relative Frequency
500-999	3	0.0075	0.0075
1000-1499	3	0.0075	0.0150
1500-1999	7	0.0175	0.0325
2000-2499	21	0.0525	0.0850
2500-2999	78	0.1950	0.2800
3000-3499	131	0.3275	0.6075
3500-3999	116	0.2900	0.8975
4000-4499	37	0.0925	0.9900
4500-4999	4	0.0100	1.0000