§1.2 Types of Data

Goals: Tell the difference between a

Parameter vs. Statistic

Quantitative vs. Categorical

Discrete vs. Continuous

Levels of measurement

Parameter is a Numerical Summary describing the entire population. A census gives

$p = .488$ → 48.8% of all people in US are women

$\mu = 64.3$in → The mean height of all women is 64.3 inches

N = population size

Statistic is a Numerical Summary of a Set of Sample Data

$\hat{p} = \frac{14}{35} = .40 = \frac{x}{n} = \frac{\#\text{women}}{\#\text{in sample}}$

$\hat{p} = \frac{x}{n}$

$x = \hat{p}n$ must be an integer

$\bar{x}$ = Sample mean

n = Sample size

§1.2 Types of Data
§1.3 Collecting Data

Ex Parameter or Statistic

1) In the 2016 census, it was found that the mean age of californians was 41.6 years.

$\mu = 41.6_{\text{years}}$ is population parameters because a census means every californian was included.

2) In a sample from a health club the mean heart rate was 71 beats per minute.

$\bar{x} = 71_{\text{bpm}}$ is a sample Statistic → Quantitative

Data = 69, 73, 71, 75, 64, ...

3) If we ask the gender of everyone in class and find proportion of females

$x = 14$ females $\quad n = 31$ students

Data M F F F F M F ... is categorical

$$\text{Proportion of females} = p = \frac{x}{n} = \frac{14}{31} = .4516$$

Population = class then p is a population parameter

Using the class as a Sample of All Statistic Students @ SRJC then $\hat{p} = .452$ is a Sample Statistic.

§1.2 Types of Data

Categorical

Words

Right, Left

Male, Female

Blond, Brown, Red, Black

$\hat{p} = \frac{x}{n}$ = Summary Statistic

Quantitative

Numbers

Height

weight

Temperature

# of course Unit

$\bar{x} = \frac{\Sigma x}{n}$

Which is an example of quantitative data?

A. Weights of high school students

B. Genders of actors and actresses

C. Colors of the rainbow

D. Consumer ratings of a particular automobile (below average, average, and above average)

Levels of Measurement for Quantitative Data

Nominal - Sports Jersey → Average does Not Make Sense

Categorical - Just a name

Ordinal - Ordered but Difference Don't Matter

1st, 2nd, 3rd in a race

Interval - Differences Meaningful

ratios are Meaningless

- Temperature 30° is Not twice 15°
- Date → 2000 is Not Twice 1000

No Natural Zero

Ratio - Ratios are Meaningful

Age

Height

# of units

Questions on a survey are scored with integers 1 thru 5 with 1 representing Strongly Disagree and 5 Strongly Agree. This is an example of what kind of measurement?

A. Nominal

B. Ratio

C. Ordinal

D. Interval

Data Set 3: Body Temperatures (in degrees Fahrenheit) of Healthy Adults *(continued)*

				Temperature Day 1		Temperature Day 2	
Subject	Age	Sex	Smoke	8 AM	12 AM	8 AM	12 AM
50	31	M	Y	99.0	99.0	—	98.6
51	26	M	N	—	98.0	—	98.6
52	18	M	N	—	—	—	97.8
53	23	M	N	—	99.4	—	99.0
54	28	M	Y	—	—	—	96.5
55	19	M	Y	—	97.8	—	97.6
56	21	M	N	—	—	—	98.0
57	27	M	Y	—	98.2	—	96.9
58	29	M	Y	—	99.2	—	97.6
59	38	M	N	—	99.0	—	97.1
60	29	F	Y	—	97.7	—	97.9
61	22	M	Y	—	98.2	—	98.4
62	22	M	Y	—	98.2	—	97.3
63	26	M	Y	—	98.8	—	98.0

Subject # → Quantitative – Nominal – NA

Age " Ratio Continuous

Sex Categorical Nominal NA

Smoke Categorical Nominal NA

Temperature Quantitative Interval Continuous

Word Count Quant Ratio Discrete

Volume Q Ratio Cts

Quantitative vs Categorical

Numbers	Level	Cts or Discrete	
Weights	R	C	
height	R	C	
Age	R	C	
Income	R	D	(often treated as cts)
BMI	R	C	
Temperature	I	C	
IQ	I	D	
# of hours worked	R	D or C	Depends on System
# of courses	R	D	
# of units	R	D	
Place in a race	O	D	Not Applicable
License Plate #	N	Not Applicable	

Categories

Names
M/F
Hair Color
eye color
Ethnicity
Occupation

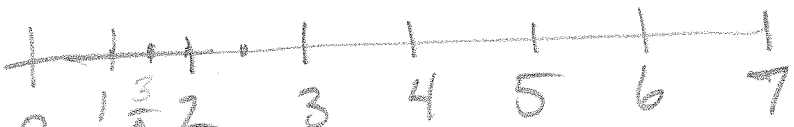
Continuous (CTS) ★ vs Discrete ★

Measured

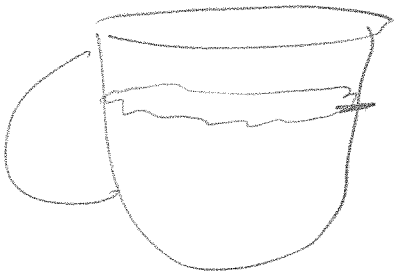
Counted

Number line

↑
Data value can fall anywher



Amount of milk



1,3 ℓ
2,1 ℓ

# of Eggs

3
5
1

Quantitative Data

Discrete

Counted

# of Students in class

# of cars

Money $20.53

vs.

Continuous

Measured

Amt of Milk

Height

Weight

Temperature

Distance

Which is not an example of continuous data?

A. Temperature on a thermometer

B. Number of students in an algebra class

C. Mean weight of 100 flour sacks

D. Amount of water pumped from a pond per day.

Parameter vs Statistic

Describes Whole population

Census is a Survey of the Whole Population

$\mu = 64.3\text{in}$ = Mean height all women

$p = .10$ = Proportion of Lefties

Describe the Sample

$\bar{x} = \frac{63+64+62}{3} = 63$

$\hat{p} = \frac{3}{33} = .09$

$\hat{p} = \frac{16}{30} = .533$

= prop. of women in class

§1.3 Collecting Data

Sampling Techniques

Simple Random Sample - Best

Every subgroup of size n is equally likely to be chosen.

- Names in hat
- Give everyone a Number
  Have a random # generater pick n

Random Sample - Good - Works

Every individual is equally likely to be included in sample

Sampling Methods

Systematic - Pick every k^{th} person
Start with a random person

Stratified - Separate population into Groups → With Different Caracteristics like Race, Gender, political
- Survey correct proportion of each
By randomly selecting Individuals in each group.

Cluster - Population in Groups - with Same characteristics

Randomly Select Groups and interview all in that group

Convience - Ask everyone around you

Statistics Math 15 Jones Name _____

use as examples

Chapter 1 Worksheet

1. Label each of the following as a parameter or a statistic.

   (a) All of the people in a South Dakota county are polled, and it is discovered that their average height is 5 feet, 7 inches.

   (b) Among a sample of 75 people, it is determined that the average systolic blood pressure is 125.2.

   (c) In a presidential election, 42.3% of the 20,087,232 total votes cast were cast in favor of candidate A.

2. For each of the following examples, indicate the type of sampling which is being used: stratified, convenience, systematic, random, or cluster

   3 4 5 1 2

   (a) The Gallup Organization plans to conduct a poll of New York City residents with the "212" area code. Computers are used to randomly generate telephone numbers that are automatically called. (Triola)

   (b) A marketing expert for MTV is planning a survey in which 500 people will be randomly selected from each age group of 10-19, 20-29, and so on. (Triola)

   (c) A Johns Hopkins University researcher surveys all cardiac patients in each of 30 randomly selected hospitals. (Triola)

   (d) The Dutchess County Commissioner of Jurors obtains a list of 42,763 car owners and constructs a pool of jurors by selecting every 100th name on that list. (Triola)

   (e) A lobbyist for the tobacco industry obtains a sample of members of Congress by writing 535 names on individual index cards, putting them in a box, mixing them, then selecting different names. (Triola)

   (f) An economist is studying the effect of education on salary and conducts a survey of 150 randomly selected workers from each of these categories: less than a high school diploma, high school diploma, more than a high school diploma. (Triola)

   (g) CNN is planning an exit poll in which 100 polling stations will be randomly selected and all voters will be interviewed as they leave the premises. (Triola)

3. You need to conduct a study of longevity for people who were born after the end of World War II in 1945. If you were to visit graveyards and use the birth and death rates listed on tombstones, would you get good results? Why or why not? (Triola)

4. "The Swiss physician H.C. Lombard once compiled longevity data for different professions. He used death certificates that included name, age at death, and profession. He then proceeded to compute the average (mean) length of life for the different professions, and he found that students were the lowest with a mean of 20.7 years!" (Taken from Triola) Does this mean that being a student is the most dangerous profession?

e) Name from hat is a SRS individuals

f) Stratified wheth 150 ~~from~~ randomly from 3 groups

g) Cluster when Randomly Selecting the 100 polling station

§1.3 Collecting Data

Type of Study

Observational Study vs. Experiment

No Treatment

Treatment Applied

Ex Do a Survey Ask two Questions

Do you ~~exercise~~ Do Music? Yes No

What is your GPA? $\bar{x} = 3.6$ $\bar{x} = 2.1$

Observational → No Treatment

Can't prove → We can see a correlation

Other Variables - Money, helicopter Mom parent ∴ time

To prove Need an experimet

Give random group Music lessons

Compare to group with No Music

See if GPA is different

Sampling Error

Sampling error is the expected difference between the population parameter and the Sample Statistic.

$$|\mu - \bar{x}| = |64.3 - 63| = 1.3$$

$$\text{or } |p - \hat{p}| = |.10 - .09| = .01$$

Non Sampling Error → We Made a Mistake

Error in a Sample Statistic that results from a

- Bad Sampling Method - **Bias**
- Math Error

Sampling Error - Expected Difference Between the Sample Statistic and population parameter

$|\mu - \bar{x}|$ → Pop. Mean & Sample Mean

$|p - \hat{p}|$ Pop. prop. & Sample prop.

Decrease as n increases in a Predictable way. → As long as we have a **GOOD** Sample.

Nonsampling Error - An error in our Sample Statistic due to a Mistake in Math or Sampling Method

Data is Not a good representation of the population we are trying to describe

Bias in Samples

Bad Frame → Group Surveyed does Not include all Outcomes

Loaded Question →

Biased Interviewer

Biasing

Volontary Responce

Sampling error - The expected variation in the difference between the Sample statistic and population Parameter. So

Sample error = $|p - \hat{p}| < E$ ← Marginal Error

or $|\mu - \bar{x}|$

Nonsampling error - is the error in our sample statistic that is introduced by bias in our sampling technique.

Voluntary responce sample from Ann landers

30% would have kids again

True parameter 90% so

.60 = .9 - .3 = Non Sampling error

↑ Avoid this