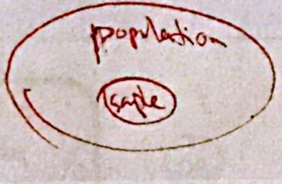
Chapter 1: Introduction to Statistics

1.1 Statistical and Critical Thinking

The SITUATION	The PROBLEM	The (IMPERFECT) SOLUTION
We want to know *SOMETHING* about a *CERTAIN GROUP*	Most of the time the *GROUP* we're interested in is TOO BIG! .	Don't ask *EVERYONE* about *WHAT YOU'RE INTERESTED IN*, just ask some of them and infer.
Ex: Of UCLA undergraduate students, what proportion likes poke bowls?	Ex: As of 2017, UCLA had about ≈ 31,000 undergraduates.	Ex: Ask 500 UCLA undergraduates if they like poke bowls and use the collected DATA to infer about the total proportion of undergraduates who like poke bowls.

Def Statistics science of planning studies and experiments; obtaining data; organizing data; summarizing data; presenting data; & interpreting those data & drawing conclusions!

POPULATION: the entire group to be studied.



SAMPLE: a subset of the population that is being studied.

Ex: *You are walking down the street and notice that a person walking in front of you drops $100. Nobody notices the $100 except you. Since you could keep the money without anyone knowing, would you keep the money or return it to the owner?*

Let's say you want to do a study to gauge the morality of the students at Pasadena City College by determining the percent of students who would return the money. You survey fifty students, and thirty-four of them say they would return the money.

In this example, what is our population and what is the sample?

Population: all PCC students Sample: 50PCC students surveyed Data: $\frac{34}{50}$ = 68% would return money

Descriptive Statistics: organizing and summarizing the data.

Ex: "68% of PCC students would return the money"

(Chapters 1- 6)

Inferential Statistics: uses methods that take results from a sample, extend it to the population, and measure the reliability of the result.

Ex: "I am 95% confident that between 64% and 72% of all PCC students would return the money"

(Chapters 7 - 12)

• just focus on statement, not how to get it • we'll study this later

We said statistics begins with wanting to know *SOMETHING* about a *CERTAIN GROUP*. Well, the *CERTAIN GROUP* is our POPULATION and the *SOMETHING* is the VARIABLE.

Def Variable the characteristic of the individual to be measured or observed.

Ex: The Gallup Organization contacts 1028 teenagers who are 13 to 17 years of age and live in the United States and asks whether or not they had been prescribed medications for any mental disorders, such as depression.

Population: all US teenagers ages 13 to 17 Sample: 1028 teenagers ages 13 to 17 that live in US Variable: whether they take prescribed medication or not

1.2 Types of Data

Def **Data** are collections of observations.

Def A **parameter** is a numerical measurement describing some characteristic of a population. "PP" population parameter
Number is FIXED based on the entire population.

Def A **statistic** is a numerical measurement describing some characteristic of a sample. "SS" sample statistic
Number varies based on the sample.

EX: **Identify whether the underlined value is a parameter or a statistic.**

a) Following the 2018 national midterm election, <u>23.4%</u> of the representatives in the U.S. House of Representatives are female. population parameter	**b)** In a 2015 national survey of high school students (grades 9 to 12), <u>15.5%</u> of the respondents reported that they had been cyber-bullied. ↳ sample statistic
c) Only 12 people have walked on the moon. The average time these people spent on the moon was <u>43.92 hours</u>. parameter	**d)** A study of 6076 adults in public restrooms (in Atlanta, Chicago, New York City, and San Francisco) found that <u>23%</u> did not wash their hands before exiting. statistic

QUALITATIVE QL
-Data consists of names or labels.
* categorical
* can't do arithmetic w/ it!

OR

QUANTITATIVE QN
- Data consists of numbers representing counts or measurements.

DISCRETE "count"
- has a countable (*finite*) number of values
1, 2, 3, etc

CONTINUOUS "measure"
- infinitely many possible values
1, 1.1, 1.01, 1.0995

EX hair color
ice cream flavor
yes/no Qs

EX: **Determine whether the following variables are qualitative or quantitative.** QL QN

a) Gender QL

b) Temperature QN, continuous

c) Number of days in the past week that you studied
QN, discrete

d) Zip code
QL

EX: **Determine whether the quantitative variables are discrete or continuous.**

a) The number of heads obtained after flipping a coin five times.
↳ count discrete

b) The number of cars that arrive at McDonald's drive-thru between 12:00 pm and 1:00pm
↳ count discrete

c) The distance a 2014 Toyota Prius can travel in city driving conditions with a full tank of gas.
↳ measure continuous

d) The average test score on the first Statistics exam in a class of 35 students.
↳ measure continuous

/2

***Statistically Significant** is achieved in a study when we get a result that is very unlikely to occur by chance. rule: " less than 5% "	***Practical Significance*** looks at whether the difference is large enough to be of value in a practical sense.

Ex: In a study of the Gender Aide method of gender selection used to increase the likelihood of a baby being born a girl, 2000 users of the method gave birth to 980 boys and 1020 girls. Would you pay $50,000 to use this method?

chance of boy/girl ≈ 50%

Gender Aide:

Boys: $\frac{980}{2000} = 49\%$

Girls: $\frac{1020}{2000} = 51\%$

No! It's too close to a 50% chance of each sex

$50,000 for a 1% increase? No way!

Levels of Measurement

(LOM) → helps decide which procedure to use

key don't use statistical methods that are not appropriate for the data

Def **Nominal** Level of Measurement - Data that consists of names, labels, or categories only. However, data cannot be arranged in an ordering scheme or hierarchy.

ie don't compute average of ice cream flavors

Ex eye colors, yes/no Qs,

Def **Ordinal** Level of Measurement - Categorical data that can be arranged in some order, but differences cannot be determined or are meaningless.

Ex rankings of colleges, course grades (ABCD, but "A-B" meaningless)

Def **Interval** Level of Measurement - Numerical data in which the difference between any two data values is meaningful. However, there is no natural zero starting point and ratios are meaningless.

Ex temperatures, years

Def **Ratio** Level of Measurement - Numerical data with a natural zero starting point and ratios are meaningful. Zero indicates that none of the quantity is present.

"zero = no quantity present"

Ex heights, lengths, distances, ~~temperatures~~

very similar: to distinguish: "ratio test" (if can say "twice" as much → ratio not interval); "true zero" (zero = no quantity present) why temperature fails

Ex: **Determine the *level of measurement* of each variable.**

a) Nation of origin

nominal

b) Movie ratings of one star through five stars

ordinal

c) Volume of water used by a household in a day

ratio

d) Year of birth of college students

interval

e) Highest degree conferred (high school, bachelor's, and so on)

ordinal

f) Eye Color

nominal

g) Assessed value of a house

ratio

(use "ratio test": double the value makes sense)

h) Time of day measured in military time

ratio

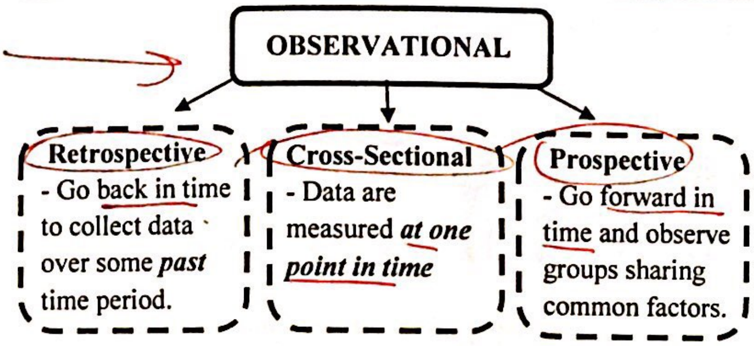
(use "true zero" test)

0 time is absence of time elapsed in that day

/3

1.3 Collecting Sample Data

OBSERVATIONAL STUDY	DESIGNED EXPERIMENTAL STUDY
*No intervention (Researcher **has no control**)	*Intervention (Researcher **has control**)
-Study that looks at data that has already been collected or as it occurs naturally	-Study that applies some treatment and then proceeds to observe its effects on the individual.
Ex: survey	Ex: split into 2 groups (placebo & pill)
(+) Less Expensive, (-) Can't claim causation, only association.	(+) Controls unknown variables (-) More costly and sometimes unethical



OBSERVATIONAL

Retrospective - Go back in time to collect data over some *past* time period.

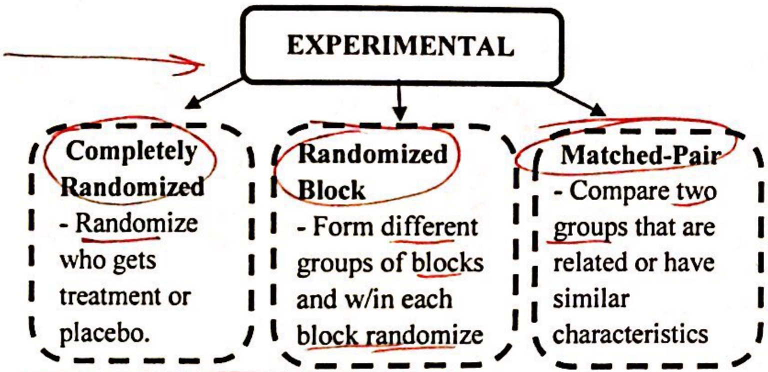
Cross-Sectional - Data are measured ***at one point in time***

Prospective - Go forward in time and observe groups sharing common factors.

Example: Determine which type of observational study is shown below:

Samples of subjects with and without heart disease were selected, then researchers looked at what the subject did ten years ago to determine whether they took aspirin on a regular basis.

observational
↳ retrospective



EXPERIMENTAL

Completely Randomized - Randomize who gets treatment or placebo.

Randomized Block - Form different groups of blocks and w/in each block randomize

Matched-Pair - Compare two groups that are related or have similar characteristics

Example: Determine which type of experimental study is shown below:

A clinical trial of Lipitor treatments is being planned to determine whether its effects on diastolic blood pressure are different for men and women.

experimental
↳ randomized block
(split into groups)

Random Sampling Methods

Regardless of whether or not a researcher decides to use an observational study or a designed experiment, a sample group needs to be chosen to best represent the population. The ***BEST*** way to choose a sample is to randomly select individuals to stay unbiased.

Def **Random** Sample - Members from the population are selected in such a way that each individual member in the population has an equal chance of being selected.

Ex: putting names in a hat (same sized paper) & drawing out names

Def **Simple Random** Sample - A sample of n subjects is selected in such a way that every possible sample of the same size n has the same chance of being chosen.

Ex:

* randomization is hard to achieve

/4

Google "random # generator)

Ex: Randomly sample six of the following companies for a survey on profit margins.

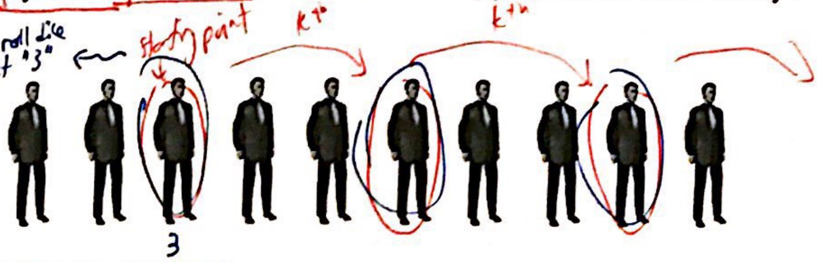
- 1 Alaskan Airlines
- 2 Akoa
- 3 Ashland
- 4 Bank of America
- 5 BellSouth
- 6 Chevron
- 7 Citigroup
- 8 Delta Airlines
- 9 Disney
- 10 DuPont
- 11 ExxonMobil
- 12 General Dynamics
- 13 General Electric GE
- 14 Clorox

Using a random number generator, the six randomly sampled companies are: Google : 13, 8, 9, 10, 5, 3 ← answers will vary!

GE, Delta, Disney, DuPont, BellSouth, Ashland

→ *Def* **Systematic** Sampling - Select some starting point and then select every kth element in the population.

Ex: roll dice get "3" ← starting point



→ *Def* **Stratified** Sampling - Subdivide the population into at least two different subgroups that share the same characteristics, then draw a random sample from each subgroup, proportional to the population (or stratum).

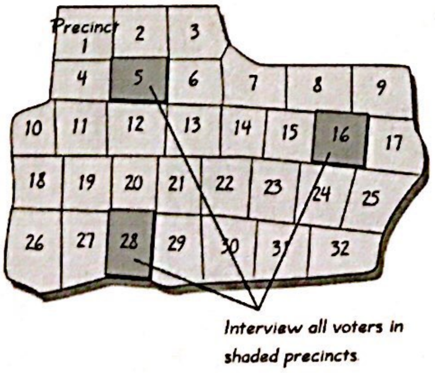
Ex:

Female ← split → Male

→ *Def* **Cluster** Sampling - Divide the population area into sections (or clusters). Then randomly select some of those clusters. Now choose all members from selected clusters.

Ex:

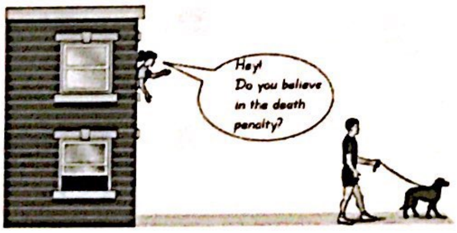
separate state into counties & survey everyone from 3 counties selected



→ *Def* **Convenience** Sampling **(non-random)** - Use results that are easy to get.

Ex:

- chose person next to you family/friends
- stand in front of grocery store & ask customers their favorite brand of milk



/5