Part I: Gathering Data

Basic Definitions

You'll want to be able to differentiate between the following four definitions: population, sample, parameter and statistic. If given an example of a statistical study, could you identify each part?

- **Population:** the entire group of individuals that we want information about. Any numeric value from the population is a **parameter.**
- **Sample:** subset of the population that we actually examine to gather information about the population. Any numeric value from the sample is a **statistic.**
    - *Survey, sample and estimate are all keywords signaling a statistic.*

Example Questions:

1.) In a survey of 598 Internet users, 68% said they have a wireless network in their home. The value 68% is a:

A. Sample B. Population C. Parameter D. Statistic

Hint:
Since 68% is a numeric value, it must be a ***parameter*** *or a* ***statistic****.*

2.) In a sample of 700 LA County residents, 61% of them supported a particular state ballot measure. When the election happened, the ballot measure passed with 64% voting for. Identify the population, sample, parameter and statistic in this scenario.

Types of Data

Be able to identify if a variable is **quantitative** (numerical, takes a quantity) or **categorical** (divides people into categories). Remember, not all numerical variables are quantitative – your Red ID, SSN or zip code are all numerical variables that divide people into categories.

Example Questions:

For each of the following, determine if the given variable is **categorical** or **quantitative:**

1.) SAT score
2.) Gender
3.) Eye color
4.) Age
5.) Social security number
6.) Weight
7.) Number of people in a household
8.) Make of car
9.) Systolic blood pressure
10.) Distance to city center
11.) College major
12.) Military rank
13.) Zip code
14.) Shoe size
15.) Medications taken
16.) Number of alcoholic beverages consumed per day

Hint:
Numeric variables are usually ***quantitative.*** *If a numeric variable is* ***categorical****, it won't have any units!*

If you increase your income by 1, that's 1 dollar per year. If you increase your area code, that's 1 what?

Sampling Designs

Make sure you have the big 4 memorized: SRS, Systematic, Cluster and Stratified. Remember, most students struggle with telling stratified from cluster samples.
See the table below for the key differences:

Stratified	Cluster
Subjects within group are SIMILAR for some characteristic or set of characteristics	Subjects within group are DISSIMILAR
We choose a SAMPLE from EACH group	We choose ENTIRE group(s) at random

- **Simple Random Sampling (SRS):** We choose people at random (i.e. picking names out of a hat). Each member of the population has an equal chance of being included.
- **Systematic Sampling:** We choose every **nth** item.
- **Cluster Sampling:** The population is divided into groups that are **dissimilar** on characteristics. We choose **entire** clusters at random and combine 1 or more clusters to get our overall sample.
- **Stratified sampling:** The population is divided into groups that are **similar** on some characteristic or set of characteristics, we then choose people at random (by SRS) from each group (**we sample from every group rather than taking a few entire groups!**) and combine those samples into our overall sample.
- **Multistage Random Sampling:** Any combination of a variety of sampling methods
- V**oluntary Response Sampling:** Sample choose themselves. (Web survey, call-in polls)
- **Convenience Sampling:** We choose people that are easiest to reach.

If each member does not have an equal chance of being selected, then the sample is **biased**!!

Example Questions:

For each of the following, determine the type of sampling method employed:

1.) To determine the impact of the Educational Opportunity Programs on SDSU freshman, a researcher obtains a list of those students enrolled in the EOP and randomly selects 100 of them using a random number generator.

2.) Of 25 school districts, 5 are selected and all households in the district are contacted to answer a questionnaire.

Hint:
*If there's any mention of selecting EVERYONE in a group, it is a **cluster** sample.*

3.) The Safari Park Half Marathon posted a poll about the raising the race fees. 76% of respondents oppose a hike in fees.

4.) Mike, a student in a business ethics course, conducted a survey on student's attitudes about copyright infringement. He stopped by the campus student union and randomly sampled 50 students. He tabulated the results and included them in his final project.

5.) In order to determine demographics of listeners to various local radio stations, a polling firm divided the county into 15 regions and randomly selected 30 people from each region to ask which radio stations they listen to.

6.) USD wanted to determine the proportion of undergraduates who have traveled to Mexico. They divide the population by major and randomly select 10 students per major.

We also had a good example in the lecture notes about a teacher interested in sampling 200 students from his school:

7.) Using an official school roster of the 2000 students, pick every 10^{th} name.

8.) Separate the students by class (freshman, sophomore, junior and senior). Pick a simple random sample of 50 from each of the classes, then combine these into a single group for your sample.

9.) Using a program, select 200 names from the list of 2000 students registered at the school.

10.) Pick a random sample of 8 of the homerooms. If a homeroom is selected, all 25 students in that homeroom are included in the sample.

We could write a similar series of examples for a farmer interested in seeing his crop yield:

11.) Divide the farm into 100 1-acre plots, randomly select 5 acres and measure the yield for all plants on those 5 acres.

12.) He could randomly choose 500 plants to measure yield from.

13.) For each of the 100 1-acre plots on the farm, randomly select 5 plants to measure the yield from.

14.) Walking along his farm, he can pick every 50^{th} plant to record yield from.

Types of Bias

Any question about biases will ask for the **most prevalent** type of bias. Even if you can argue that a few of the types listed below are answers (they usually will be), be on the lookout for those keys to let you know what the problem being tested is!

- **Undercoverage:** Entire population targeted is not included in the design of the sample.
  - Be on the lookout for any mention of a certain group in the sample design (if they mention they only sampled females, or people of a certain age group, or people from a certain region) and check that the group mentioned is the same as the population they are interested in. If not, you have undercoverage!
- **Non-response:** An individual selected into the sample cannot be contacted or refuses to cooperate.
  - Any mention of choosing people into the sample and people not responding – whether not being home for interview or phone call, refusing to participate, etc…
- **Response Bias:** Interviewee's responses are influenced by the interviewer.
  - Any mention of inappropriate behavior on the interviewer or interviewee, if you can see any reason the interviewee may lie - the way the question was asked, who was asking the question, etc…
- **Voluntary Response Bias:** The type of bias associated with voluntary response samples.
  - Web surveys, mail in surveys, call in surveys…

Example Questions:

For each of the following scenarios, identify the most prominent source of bias:

1.) In a sample of 700 LA County residents, 61% of them supported a particular state ballot measure.

Hint:
If you don't see an obvious answer at first, figure out the population and the sample – see if they are different!

2.) A business magazine mailed a questionnaire of the HR directors of all the Fortune 500 companies, and received response from 23% of them. Those responding reported that they did not that such surveys intruded significantly on their workday.

3.) Suppose you are conducting a survey regarding illicit drug use among sophomores in the Baltimore school district. You obtain a random sample of 12 schools within the district and have teachers verbally administer the surveys to all sophomores.

4.) A polling organization is going to conduct a study to estimate the percentage of households that speak a foreign language as a primary language. It mails a questionnaire to 1,000 households, 546 questionnaires are returned.

5.) In order to determine the percent of Californians that will vote for a balanced budget amendment, 900 college students are sampled and asked their opinion.

6.) A school district was thinking about uniforms in the schools. A polling firm called households with children in the school district to determine feelings towards mandatory uniforms between the hours of 10am and 2pm.

7.) A rescue group is interested in the percent of potential adopters who would consider declawing their cats. They hand out pamphlets and explain that declawing is amputation before asking if they would consider declawing as an option.

8.) Suppose a local radio station is interested in seeing what proportion of people agree with the suspension of a local football coach. Of the people that call in, 90% oppose the suspension.

Experiments & Observational Studies

Be able to tell the difference between an observational study and an experiment. If we assign individuals to a treatment group (this can be as innocuous as us asking them to eat a serving of vegetables or watch a commercial), it's an experiment. If we just ask them about or observe their behavior, then it's an observational study.

Example Questions:

For each of the following, determine if the scenario is an experiment or an observational study:

1.) To determine if physical activity helps with depression, a group conducts a study where they ask subjects with depressive symptoms to engage in 30 minutes of physical activity per day and then record their depressive symptoms after three months.

2.) To rate customer opinion, a focus group watches two advertisements and then fills out a brief questionnaire.

3.) In order to determine possible causes for premature births, expectant mothers are followed during their pregnancy and have their behavior and vitals recorded every month at their regular checkups.

4.) Looking at risk factors for stroke, a researcher asks recent stroke victims questions about their prior medical history.

5.) To see the proportion of California drivers use their cell phones while driving, a group puts up stations at several high traffic areas and does a visual count of drivers using their cell phones.

Observational Studies

Observational studies require that we only observe! We cannot apply a treatment to the subjects.

- **Retrospective study:** looks **backward** in time, we ask subjects about what happened in the past to try to determine possible risk factors.
- **Prospective study:** looks **forward** in time, normally we follow subjects to see if an outcome occurs.

Example Questions:

1.) To determine the cause of an E.coli outbreak, an epidemiologist goes through the food receipts and past purchases for residents in a certain county.

2.) The study I work for researches aging by giving subjects tests for mental and physical functioning every 5 years allowing them to determine risk factors for depression, cognitive impairment, and other factors associated with aging.

3.) To look into the mechanics behind childhood obesity, a medical group followed children from birth to age 10, monitoring physical activity and eating habits.

Principles of Experimental Design

You may be asked questions regarding the basic principles of a good experiment, so make sure you know these 3!

- **Control:** The experimental conditions for all treatment groups to assure that lurking variables do not bias the results, part of this is including a control (nontreatment) group.
- **Randomization:** The experimental units must be **RANDOMLY** assigned to treatments
- **Replication:** Replications of our experiment must be used to reduce chance variation in the results.

All of the above requirements are important because they help control for lurking variables.

Experiment Terminology

Of all the experiment terminology, it's most likely you'll be asked about number of treatments or to list all the treatments. Remember that treatments include **all possible combinations** of the levels of each factor. If given a description of an experiment, you should be able to identify each of the following:

- **Experimental units/Subjects:** Individuals you are studying in the experiment.
- **Response variable:** Outcome or dependent variable. This is what we are ultimately measuring or interested in (our y variable in a regression context).
- **Factors:** The **explanatory (independent, x) variables** that are thought to influence the response variable studied. Combine specific values (**levels**) of each factor to form a treatment.
- **Treatment:** A specific condition applied to the subjects. (A combination consisting of a level of each factor.)

A good experiment will usually employ the following to help **CONTROL** for lurking variables. Know these definitions:

- **Placebo:** A dummy treatment. Used to control for the placebo effect.
- **Blinding:** A **double-blind** experiment is one where neither the participant nor the researcher taking the measurements knows who had which treatment. A **single-blind** experiment is one where the participants do not know which treatment they have been assigned. **Helps reduce bias!**

A last definition associated with experiments that you may be asked about:

- **Statistically Significant:** An observed effect so large that it would rarely occur by chance.

Experimental Designs "How did we determine who got what treatment?"

If you're asked a question about experimental design, cross out anything that isn't what you're being asked for! Some experimental designs have methods almost identical to our sampling designs (SRS v. completely randomized and stratified v. block design). These will distract us if we aren't careful.

- **Completely Randomized Design:** Similar to an SRS setup, each subject is equally likely to get assigned to any treatment.
- **Matched Pairs:** Subjects are paired according to variables that affect the response and then randomly assigned with one to the treatment and the other to the control group.
    - Keep an eye out for before and after studies. They are a matched pair design where each subject is their own control! If it seems like every got the treatment, check to see if you've been given a before and after study.
- **Block Design:** Blocks of similar subjects are formed and within each block, they are randomly assigned to treatment groups.
    - Similar to a stratified set up, if we first **divide** our subjects into groups that are **similar**, then randomly assign from **each** group into our treatment groups, we have a block design.

Example Questions:

1.) In order to test a new drug for heartburn, 1000 heartburn sufferers were divided into groups by age, 20-29, 30-39, 40-49 and 50-59. These patients were then assigned from each group into one of three treatment groups: placebo, 150mg or 300mg. What type of experimental design is being employed?

A. Stratified B. Completely Randomized C. Matched Pairs D. Block

2.) To determine the effects of a drug on cholesterol levels, a medical group took cholesterol levels on subjects randomly assigned subjects to take either the drug or a placebo. Additionally, to control for diet on cholesterol, they assigned the patients into one of three groups – low fat, moderate fat or high fat.

How many possible treatments are in this study?

A. 2 B. 3 C. 5 D. 6

3.) An advertising firm is interested in seeing how an advertisement affects people's opinion of a certain brand. They ask 100 randomly selected people to fill out a brief questionnaire indicating how they feel about the brand. They then view a 45 second advertisement and fill out the questionnaire again. What type of experiment did the advertising group use?

A. Stratified B. Completely Randomized C. Matched Pairs D. SRS

4.) To determine how to maximize the yield on his crops, a farmer decided to try three different levels of fertilizer and two different watering methods – a larger amount once a day or smaller amounts multiple times a day. He divided the crops into two strains and randomly assigned them to the treatment groups.

Determine the type of experimental design.

What is the response variable?

What are the factors (or explanatory variables)?

What are the treatments?

<u>Confounding v. Lurking Variables</u>

Remember the drawings from class. Confounding variables add another relationship with the response variable (in addition to the one with the existing explanatory variable) while lurking variables have a relationship with both existing explanatory and response variables (which is the entire cause of the relationship we saw between the explanatory and response variable).

- Two variables are **confounded** when you can't tell which of them (or whether it's the combination) had an affect on the response variable.
    - Example: You might want to test if a fertilizer helps your garden produce more tomatoes. Suppose you spread it on half the plants and record the number of tomatoes they produce. But you spread it on the sunny half, leaving the shady half unfertilized! Now, even if you have more tomatoes on your fertilized plants, you don't know whether that's because of fertilizer or sunshine (or the two together, which is actually the most likely case).

- A **lurking variable** is sometimes referred to as common response. It's a variable that drives two other variables, creating the impression of an association between them while in reality the two variables are BOTH response variables.
    - Example: Suppose a researcher finds a strong association between the number of computers per capita and life expectancy - countries with fewer computers have lower life expectancy. Do we think that the computers affect life expectancy in some way? NO! The general socioeconomic status (which could be measured in something like gross domestic product (GDP)) is likely causing both the number of computers and the life expectancy to rise.

Example Questions:

1.) Foods rich in omega-3 fatty acids are rumored to help with allergies. Suppose if to study this, a matched pairs experiment is run where 100 allergy sufferers record their allergy symptoms and then change their diet to increase the amount of food containing omega-3 fatty acids. After a week, they record their symptoms again and find that there's a statistically significant decrease in symptoms. However, someone also took pollen readings and found that the pollen levels dropped from the beginning of the week to the end of the week. Pollen levels would be an example of a:

A. Lurking variable B. Confounding variable C. All of the above. D. None of the above.

2.) In our crop yield experiment on the previous page, list a possible lurking variable.