Experimental Design Problems

General

Question 1

What is a confounding variable?

Answer

Is one that affects the response variable and is related to the explanatory variable. (ex. people given leeches produce magical tears that heal wounds, the tears would be confounding variable if testing leeches effects on wound healing)

Question 2

What is survivorship bias?

Answer

Survivorship bias or survival bias is the logical error of concentrating on entities that passed a selection process while overlooking those that did not. (ex. cannot test on fatal shots to planes)

Question 3

Where can data come from?

Answer

observational studies, experiments, and synthetic

Question 4

What are the 3 main requirements in designing an experiment?

Answer

Can be replicated, randomly selected subjects, control group

Question 5

What are the types of obsservational studies

Answer

- **Cross sectional:** Look at data a a single point in time (*Present*)
- **Retrospective studies:** Looking at studies of events in the past (*Past*)
- **Prospective studies:** Researchers follow and observe groups closely (*Future*)

Question 6

What is synthetic data, where does it come from?

Answer

Data created by you, typically through simulation

Question 7

What are ways to control data

Answer

- **Blinding**: participants are unaware of the kind of treatment they are recieving, if any at all.
- **Placebo effect**
- **Blocking:** arranging experimental units into similar groups (based on treatment applied.

Question 8

What are the different sampling techniques

Answer

- Systematic Sampling
  - sampling method where researchers select members of the population at a regular interval
- Stratified Sampling
  - in a stratified sample, researchers divide a population into homogeneous subpopulations, called strata, based on specific characteristics (ex. race, gender identity, location, etc.)
- Cluster Sampling
  - in cluster sampling, researchers randomly divide a population into smaller groups known as clusters. They then randomly select among these clusters to form a sample.
- Multistage Sampling
  - in multistage sampling you draw a sample from a population using smaller and smaller groups at each stage.
- Convenience Sampling
  - method of collecting samples by taking samples that are conveniently located around a location or Internet service.

Question 9

Stratified vs blocking

Answer

Stratification groups subjects based on characteristics which the experimenter cannot control (ex. eye color). Blocking groups subjects based on variables the experimenter can control such as the treatments.

Question 10

Cluster vs stratification

Answer

In clustering subjects are grouped randomly, while in stratification they are grouped based on shared characteristics.

Question 11

What are the types of error in experimentation?

Answer

- **Sampling Error**
  - Unrepresentative sample taken
- **Non-Sampling Error**
  - Errors due to sample data that are incorrectly collected, recorded, or analyzed

Question 12

What are the types of errors associated with surveys.

Answer

- Wording of questions
- Ordering of questions (planting ideas)
- Convenience samples
- Desire of respondents to please
- Non-response bias
- Lizardman constant (around 3% of respondents just messing around)

Design an Experiment

Question 1

You are a data scientist at facebook, you have a theory that memes get less engagement. design an experiment to test this theory. (on exam such a question would have a major issue)

Answer

1. give some users with regular weight (control) - 99% and downweight (treatment) - 1%
2. stratified sampling to select users

Question 2

Design an experiment to detremine if a new drug is effective in curing migranes. Population: adults who are experiencing migranes.

Answer

1. control, with drug
2. blind, placebo, blocking if necessary
3. stratified sampling

Question 3

In what scenario would you use:
- stratified sampling
- systematic
- cluster
- mutlistage
- convenience

Answer

- stratified sampling
  - separating people based on age and gender during drug trial
- systematic
  - surveying every other customers at a store about how frequently they come to that store
- cluster
  - typically used when respondents spread out over large geographic location. Example: want to know favorite soda of inhabitants of a city.
- mutlistage
  - The Census Bureau: The U.S. Census Bureau uses multistage sampling by first taking a simple random sample of counties in each state, then

taking another simple random sample of households in each county and collecting data on those households.
    - mutistage sampling is a type of cluster sampling where you have various layers of grouping
- convenience
    - when you want respondents who are passionate about the subject of the survey. so in cases where you want feedback on a product.

Problems

Issue with Experiment

Question 1

You want to see if a new drug can reduce the risk of stroke in adults. To test this you provide one group the drug and the other group no drug. From the control and trail groups you select subjects using systematic sampling. What is wrong with this experiment?

Answer

- not blind
- no placebo
- stratified sampling more appropriate

Question 3

You want to know what the most popular convenience store brand is among college students in the US compared to the rest of the adult population. You want to test if college students typically visit different stores than other populations. For your experiment you survey people from NYU, Boston College, and University of Chicago regions as they walk into Target, taking a systematic sample. What is wrong with this experiment?

Answer

1. issue is that respondents sampled are not representative of the entire college student population in the US. Many of the students surveyed for

this experiment are from universities in the city.
2. Also a confounding variable to consider is that the most popular convenience store in a region may be limited to the location, certain stores like chains may be more popular across the board because they have more locations.
3. they survey took place in a Target, so the respondents answers may be skewed in favor of target.

Confounding Variables

Question 1

What are possible confounding variables in this experiment:
You collect data on sunburns and ice cream consumption. You find that higher ice cream consumption is associated with a higher probability of sunburn. Does that mean ice cream consumption causes sunburn?

Answer

The confounding vairiable here is that people tend to eat more ice cream when the weather is hot

Question 2

What are possible confounding variables in this experiment:
Suppose we wanted to measure the effects of caloric intake (IV) on weight (DV).

Answer

- metabolic rate
- age
- physical activity
- height

Question 3

What are possible confounding variables in this experiment:
Low blood pressure is associated with a higher risk of mortality.

Answer

This association is non-causal; it is due to the confounding effect of heart disease.