Collecting Data in Reasonable Ways

STAT 250

Lecture 1A

Science of Statistics (3 Steps)

1. Collecting Data
2. Describing Data
3. Inference

Collecting Data

- Types of Statistical Studies
- Planning an Observational Study
- Designing an Experiment
- The Role of Random Selection and Random Assignment

Population vs. Sample

- Native American proverb –
- Population –
- Sample -

Representative Data

Sample data can be used to make inferences about a much larger group if the data can be considered to be representative with regard to the question(s) of interest.

Are SDSU students representative of college students in general?

What about USD students?

- For ethnicity?
- For time spent online?
- For family income?
- For religious affiliation?
- For time spent outdoors?

Observational study – a study in which the person conducting the study

Experiment - a study in which the person conducting the study

Key Difference

- in an experiment, _____ _____ determines who will be in what experimental groups and what the experimental conditions will be

- in an observational study, the person carrying out the study _____ determine who will be in what groups

Does calcium reduce blood pressure? Identify the type of study.

- Patients fill out questionnaires about their calcium intake. Medical practitioners take their blood pressure. The correlation between calcium intake and blood pressure is calculated to determine if people with higher calcium intake have lower blood pressure.

- Medical practitioners take patients' blood pressure. Patients prescribed a multivitamin with or without calcium. After six months, blood pressure is measured again. The change in blood pressure is compared for the calcium group and the no calcium group.

Observational Studies

- Purpose is to collect data that will allow you to learn about a single population or about how two or more populations differ
- Allows you to answer questions like:

- But not questions like:

  – An experiment MUST be used.

Census versus Sample

- Measurements that require destroying the item
- Difficult to find entire population
- Limited resources

When you ask questions like “What is the proportion of . . .?” or “What is the average of . . .?”, you are interested in the **population characteristic**.

- A **population characteristic** is a number that

- A **statistic** is a number that

Methods for Sampling

- Simple Random Sample
- Stratified
- Cluster
- Systematic

- Convenience
- Voluntary Response

Simple Random Sampling

A sample of size *n* is selected from the population in a way that ensures that every

_____

has the same chance of being selected.

Selecting an SRS

- You need:
  1. A list of units in the population
  2. A source of random numbers
- Give each unit in your list an ID number and generate n random numbers from your source. The units with ID numbers that match the random numbers generated become the sample.

A simple random sample of 2 students from a group of 5 students (A, B, C, D, E) is to be chosen to represent the group at a conference.

- List the possible samples. How many are there?

- What is the probability that A and B are chosen to be the sample?

- What is the probability that A is chosen to be in the sample?

Replacement

- **Sampling with replacement** - Sampling in which an individual or object, once selected, is put back into the population before the next selection. This allows an object or individual to be selected more than once for a sample.
- **Sampling without replacement** - Sampling in which an individual or object, once selected, is **NOT** put back into the population before the next selection; once an individual or object is selected, they are not replaced and cannot be selected again.
- Although sampling with and without replacement are different, they can be treated as the same when the sample size n is relatively small compared to the population size (no more than 10% of the population).

Stratified Random Sample

- The population is first divided into non-overlapping subgroups (called strata).
- Then separate simple random samples are selected from each subgroup (stratum).

- Illustration

Identify how you would break the population into subgroups.

- A food services director wants to learn about on-campus dining habits and expenditures of students enrolled on your campus.

- A human resources manager wants to find out how employees feel about proposed changes to their benefits structure (retirement, health insurance, vacation/sick leave).

- County government needs to estimate average housing costs (rents and home values).

Cluster Sampling

- Sometimes it is easier to select groups of individuals from a population than it is to select individuals themselves.
- Cluster sampling divides the population of interest into nonoverlapping subgroups, called clusters. Clusters are then selected at random, and ALL individuals in the selected clusters are included in the sample.
- The ideal situation for cluster sampling is when each cluster mirrors the characteristics of the population.

Identify how you would carry out cluster sampling.

- A regional scout administrator wants to know how many badges scouts have earned this year.

- A medical researcher is interested in how well diabetes patients in her area are complying with their treatment plan.

- A tour company wants to give a customer satisfaction survey to clients.

1 in *k* Systematic Sampling

- Selects an ordered arrangement from a population by
  - first choosing a starting point at random from the first *k* individuals
  - then every *k* th individual after that
- Suppose you wish to select a sample of faculty members from the faculty phone directory. You would first randomly select a faculty from the first 20 (*k* = 20) faculty listed in the directory. Then select every 20^{th} faculty after that on the list.

Convenience Sampling

- Selecting individuals or objects that are easy or convenient to sample.
- Suppose your statistics professor asked you to gather a sample of 20 students from your college. You survey 20 students in your next class which is music theory.
  - Will this sample be representative of the population of all students at your college? Why or why not?

Voluntary Response

- **Voluntary response** is a type of convenience sampling which relies solely on individuals volunteering to be part of the study.
- People who are motivated to volunteer responses often **hold strong opinions.** It is extremely unlikely that they are representative of the population!

Identify the sampling design

- The Educational Testing Service (ETS) needed a sample of colleges. ETS first divided all colleges into 6 subgroups of similar types (small public, small private, medium public, medium private, large public, and large private). Then they randomly selected 3 colleges from each group.

- A county commissioner wants to survey people in her district to determine their opinions on a particular law up for adoption. She decides to randomly select blocks in her district and then survey all who live on those blocks.

- A local restaurant manager wants to survey customers about the service they receive. Each night the manager randomly chooses a number between 1 & 10. He then gives a survey to that customer, and to every 10^{th} customer after them, to fill it out before they leave.

Sources of Bias

- Sample statistics do not typically match population characteristics exactly.
- Error that comes from drawing a sample instead of taking a census is called **sampling error**.
- Other errors or biases, can be caused by the data collection process

Selection bias

- Occurs when the way the sample is selected systematically excludes some part of the population of interest

- May also occur if only volunteers or self-selected individuals are used in a study

Measurement or Response bias

- Occurs when the method of observation tends to produce values that systematically differ from the true value in some way
  - Improperly calibrated scale is used to weigh items
  - Tendency of people not to be completely honest when asked about illegal behavior or unpopular beliefs
  - Appearance or behavior of the person asking the questions
- Questions on a survey are worded in a way that tends to influence the response

Nonresponse Bias

- Occurs when responses are not obtained from all individuals selected for inclusion in the sample
- To minimize nonresponse bias, it is critical that a serious effort be made to follow up with individuals who did not respond to the initial request for information

- Will increasing the sample size reduce the effects of bias in the study?

Identify the type of bias the survey is most likely to suffer from

- An elected official is interested in residents' opinions of a proposed rezoning law. His office conducts phone interviews with a random sample of 500 people who are listed in the phone book as living in his district.

- A telephone survey asks participants if they voted in the last election.

- A talk radio show asks listeners to respond by phone or text to a question about whether gay marriage should be legal.

- An email survey of all of a department's alumni from the last 10 years asks about employment and income.

Experimental Design

- In a **simple comparative experiment**, the value of some **response variable** is measured under different **experimental conditions (treatments)**. **Experimental units** are the smallest unit to which a treatment is applied.
- The goal of a simple comparative experiment is to determine the effects of the treatment on the response variable.

Isolating the treatment effect

We must consider and manage other potential sources of variability in the response

- Eliminate them (direct control) – Holding a potential source of variability constant at some fixed level, or
- Ensure they produce chance-like variability . When we cannot control directly or haven't even thought of other variables, **random assignment** should evenly spread variables into treatment groups. We expect these variables to affect all the experimental groups in the same way; therefore, their effects are not confounding.

Confounding

- When you cannot distinguish between the effects of two variables on the response, the two variables are **confounded**

- A researcher wishes to study the effect of class size on final exam performance in a multi-section course using historical data. What are some potential confounding variables?

Types of Experimental Design

- Completely Randomized Design
- Randomized Block Design
  - Matched Pairs Design

Completely Randomized Design

- An experiment in which experimental units are randomly assigned to treatments
- Can moving their hands help children learn math?
  - An experiment using 128 students was conducted to compare two different methods for teaching children how to solve math problems
  - Diagram

Randomized Block Design

- An experiment that incorporates blocking by dividing the experimental units into blocks of similar units and then randomly assigns the individuals within each block to treatments.
- Suppose that you were worried that gender might also be related to performance in the math experiment.
  - Direct control of gender – use only boys or only girls. Conclusions can ONLY be generalized to the group that was used.
  - Blocking by gender (diagram)

More Experiment Concepts

Control group

- allows the experimenter to assess how the response variable behaves when the treatment is not used.
- provides a baseline against which the treatment groups can be compared to determine whether the treatment had an effect.
- Consider Anna, a waitress. She decides to perform an experiment to determine if writing **"Thank you"** on the receipt increases her tip percentage. She plans on having two groups. On one group she will write **"Thank you"** on the receipt and on the other group she will not write "Thank you" on the receipt.

Placebos

A placebo is something that is identical (in appearance, taste, feel, etc) to the control group, except that it contains no active ingredients.

- People respond to the power of suggestion.
- To measure a true treatment effect, it is better to compare a treatment to a placebo than to a control

Blinding

- Single Blind – either participant or evaluator does not know which treatment they are receiving
- Double Blind – both participant and evaluator do not know which treatment they are receiving

Using Volunteers in an Experiment

- Remember – Using volunteers in observational studies is never a good idea!
- However – It is common practice to use volunteers as subjects in an experiment.
  - Random assignment of the volunteers to treatments allows for cause-and-effect conclusions
  - But, limits the ability to generalize to population

Example

You wish to test whether epicatechin, an antioxidant found in green tea and dark chocolate, can help people recover from a workout. 40 college student volunteers are available along with equipment and a technician to measure F2-isoprostanes which assess the oxidative stress level of the subjects.

- Describe how this experiment could be conducted using a simple comparative experiment, a block design, and matched pairs design.
- Also discuss how blinding should be employed in the study.

Drawing Conclusions from Statistical Studies

Common Mistakes

- Drawing a cause-and-effect conclusion from an observational study.
- Generalizing results of an experiment that uses volunteers as subjects.
- Generalizing conclusions based on data from a poorly designed observational study.
- Generalizing conclusions based on an observational study that used voluntary response or convenience sampling to a larger population.