Review

Arnav Patri

May 4, 2022

Contents

Topic 1: Sampling Techniques and Sources of Bias

1. Know and understand the difference between a *population* and *sample*
   - How is each one measured? Use the proper *vocab term*!

     A population is measured via a census while a sample is measured via a survey/experiment.
   - Why do we often measure samples instead of populations?

     It is often infeasible to collect data from every member of a population.
2. Know the different types of *bias* and how to spot them in different situations
   - What is the difference between *sampling error* and *sampling bias*?

     Sampling error is unrepresentativeness as a result of the fact that a sample is being taken rather than a census. As the sample is never going to exactly represent the population, it is unavoidable, but it can be mitigated by a larger sample or a more representative sampling technique. Sampling bias is the systematic skewing of results away from the true population results as a result of fault inherent to the sampling procedure.
   - How can a small sample size affect the validity of the sample? (*this is related to sampling error rather than bias*)

     The smaller the sample, the greater the likelihood of it misrepresenting the full population, simply because it is difficult to get a full picture from only a few data points. It is also more likely for the sample to be unrepresentative of the population due to pure chance.

Define the types of **sampling bias** (a bias in *who* was in the sample)	Define the types of **response bias** (a bias in *what* the sample is saying).
Under-coverage bias is bias caused by certain subsets of the population being underrepresented in the sample. **Nonresponse bias** is bias as a result in certain subsets of the population choosing not to participate in the sample despite being selected. **Voluntary response bias** is bias as a result of only those choosing to participate in a survey doing so.	**Loaded questions** may be difficult for respondents to answer fully or truthfully, introducing bias. **False answers** may be incentivized by the sample, resulting in the results being skewed.

Simple random sample: Every member in a population should be equally likely to be chosen for the sample for it to be a simple random sample.	**Stratified random sample:** The population of interest should be divided into strata such that the strata are heterogenous without but homogenous within. Simple random samples should then be taken from each stratum. **Stratifying will **reduce variability** of possible sample results!*
Systematic random sample: Every member in a population should be assigned a number and a number. A random starting number should be selected and every k^{th} person (such that $k = N/n$) should be chosen.	**Cluster sample:** A simple random sample should be taken and those close in proximity to those selected should also be included in the sample.
Multistage sample: A simple random sample of large groups should be taken and simple random samples of those selected should be used in the sample.	**Convenience sample:** Those that are convenient to sample should be sampled.

Example: A principal wants to create an advisory committee of 20 randomly-selected students out of the 1,800 students their high school. Describe how he could do so using a...

Simple random sample: Every student at the high school could be assigned a number from 0 to 1799, assigned alphabetically by last name. 20 non-repeating integers between 0 and 1799 (inclusive) could then be randomly generated using software. Those whose could then be assigned to the committee.	**Systematic random sample:** Every student at the high school could be assigned a number from 0 to 1799, assigned alphabetically by last name. Software could then be used to generate a random number from 0 to 89 (inclusive). In addition to the person whose number was initially selected, every 20^{th} person could be selected for the committee.
Stratified random sample: Within each grade, each student could be assigned a number from 0 to $n-1$ (where n is the number of students in the grade). Software could then be used to randomly select 5 numbers between 0 and $n-1$ (inclusive), and those whose numbers were selected could be added to the committee. This process could be repeated for each grade.	**Cluster sample:**
Multistage sample:	**Convenience sample:** The first 20 students that the principal sees could be selected for the committee.

Topic 2: Experimental Design

1. Know the vocabulary of experiments and experimental design

   - What is the difference between an Experiment and an Observational Study? Which one lets us establish cause-and-effect relationships? ***HINT:*** *There is one* dead giveaway *keyword when identifying an experiment. It starts with the letter A*

     An experiment deliberately attempts to alter the results of those being experimented on while an observational study simply aims to observe. The former enables causality to be justified.

   - Define *Treatment* –

     A treatment is something imposed by those performing the experiment onto those being experimented on in an attempt to affect the results.

   - Define *Experimental Units* (*Subjects* when human) –

     The experimental units are the things that the data is being collected on.

2. Know the four principles of a good experiment

   - Why is it important to *randomize* the assignment of treatments?

     If treatments are not randomly assigned, causality cannot be justified, as there may be some other variable resulting in the responses.

   - Why is *comparison* (either with a control group OR a second treatment group) important?

     Without a comparison, there is no way of knowing whether or not the data is significant.

   - In experiments, it is important to *control* for outside factors or variables. What does this mean?

     Any variables that may affect results should be either mitigated or accounted for in the assignment process.

   - What is *replication*, and how can we make sure our sample has it?

     For an experiment to be replicable, repeated trials must be possible and produce the same results from the same initial conditions. If an experiment is not replicable, its claims cannot be corroborated, making it more likely that the results were due to sheer chance.

3. Know methods for **controlling** an experiment to prevent confounding

   - Control group (what is it, and what does it allow us to do?)

     *(**NOTE:** A control group is NOT mandatory; it is just one way to get comparison, which IS mandatory)*

     A control group is a group given the "default" treatment, often a placebo, so that the results can be tested against those of the groups given other treatments.

   - Placebo effect –

     The placebo effect refers to the phenomenon of a treatment working simply because those taking it believe that it will work.

- Blind study –

  A blind study is one in which the subjects are unaware of which treatment they are being given.

- Double-blind study –

  A double-blind study is one in which both those giving and receiving the treatment area unaware of which treatment it is.

4. Know the different types of experimental design and how to identify which one is being used (as well as the *advantages* and *disadvantages* of each)

   - Completely Randomized Design

     Every subject has an equal chance of receiving any given treatment.

   - Randomized Block Design ("Blocking")

     The subjects are grouped by their attributes that may affect the response into blocks. Within each block, treatments are then randomly assigned. This mitigates confounding variables and makes comparison between treatments easier.

   - Matched Pairs Design

     The data is paired, measuring some change in results from the same initial conditions. This ensures that any change is due only to the treatment.

Example: *A researcher studied a random sample of 100 teens in Oklahoma. To which populations will the results of this researcher's findings be generalizable?* (Circle ALL that apply)

(A.) The 100 Oklahoma teens in the study
(B.) All teens in Oklahoma
C. All teens
D. All Oklahomans

Topic 3: Analyzing Data

1. The 5 things you should discuss when analyzing a **distribution** of data:

   •

   NOTE: *if asked to compare data sets, make sure you explicitly compare them, not just describe both of them (For example, "The first distribution has a greater mean that than the second distribution, while the second distribution has a greater spread than the first."*

2. **Center**

   What does it tell us about our data?

   The center identifies where the data is centered, giving us the expected value of the variable.

Measure	*How to find it*	*Resistant to the effects of outliers?*
Mean *Population:* μ *Sample:* $\bar{x}$	$\bar{x} = \frac{\sum x_i}{n} \quad \mu = \frac{\sum x_i}{N}$	No

Topic 4: Normal Distributions and Z-Scores

1.
   - *THEORETICAL* distribution (in reality, we consider data to be approximately normal)
   -