Cheat Sheet: Thinking Statistically and Collecting Data

Essential Concepts

- Statistics is a science that deals with the collection, analysis, interpretation, and presentation of data.
- The goal of statistics is to make an inference (a logical conclusion or guess) about the population based on a sample.
- The statistical process is an investigative process; it is also a repetitive or cyclical process.
    - Ask a question that can be answered by collecting data
    - Define the population and design a study
    - Collect data from a sample (or samples) of the population
    - Summarize and analyze data
    - Interpret results and draw a conclusion
- A statistical investigative question is required for any statistical study. A good statistical question will anticipate variability, will require data collection, and will not have a single definitive answer that can be easily looked up.
- The population is the group of individuals or entities that our research question pertains to, and a parameter is a numerical summary measure that summarizes that population (e.g., the proportion who use social media).
- A sample is a group of individuals or entities on which we collect data, and a statistic is a numerical summary measure of a sample.
- We often collect data on different variables using survey questions, data can be quantitative or qualitative (also called categorical) depending on how it can be analyzed.

Generally, quantitative data is the result of counting or measuring and qualitative data is the result of categorizing or describing.

- Simple random sampling assigns a number to every member of the population, then uses a random number generator to select a sample.
- Other sampling techniques include:
    - Systematic sampling assigns a number to every member of the population, then chooses individuals/entities from the population at regular intervals (e.g. every 4th individual from a randomly selected starting point).
    - Stratified sampling divides a population into groups via some criterion, then uses simple random selection or systematic selection to collect a sample from each group.
    - Cluster sampling divides a population into groups via some criterion, then uses simple random selection or systematic selection to select one or more groups as the sample.
    - Convenience sampling selects a sample most accessible to the researcher.
- A sampling method is unbiased if, on average, it results in a representative sample of the population. A sampling method is biased if it has a tendency to produce samples that are not representative of the population.
- Here are the four main sources of bias to consider when sampling from a population:
    - Undercoverage occurs when some groups of the population are left out of the sampling process and the individuals in these groups do not have an equal chance of being selected for the sample.
    - Non-response bias occurs when an individual chosen for a sample cannot be contacted or decides to not participate in the study or research. This type of bias occurs after the sample has been selected and can create potential bias in the data collected.
    - Response bias is defined as a systemic pattern of inaccurate responses to questions. This type of bias can occur when a person does not understand a question or feels influenced to respond to a question in a certain way. Response bias can also occur as a result of the wording of questions that are of a sensitive nature.
    - A voluntary response bias is another form of bias because the sample is not random or representative of the population. The people who volunteer for a study or survey may be more inclined to respond to questions or report certain behaviors.

Glossary

biased

samples that are not representative of the population

cluster sampling

divides a population into groups via some criterion, then uses simple random selection or systematic selection to select one or more groups as the sample.

convenience sampling

a sample of individuals who are most accessible to the researcher. A convenience sample is usually not random or representative of the population

data

factual information about a group of individuals, animals, or objects

informed consent

risks of participation must be clearly explained to the subjects of the study

non-response bias

when an individual chosen for a sample cannot be contacted or decides to not participate in the study or research

observational units

the group of individuals, animals, or objects in the study

parameter

a numerical summary measure that summarizes that population

population

an entire group of people, objects, or animals; usually a large group

response bias

a systemic pattern of inaccurate responses to questions

sample

a randomly selected subset or subgroup of a population

sampling bias

when a sample is collected from a population and some members of the population are not as likely to be chosen as others

simple random sample

a random mechanism to choose a sample, without replacement, from the population so that every sample of a given size has the same chance of being selected

statistic

a numerical summary measure of a sample

statistical investigative question

a question that can be used as the starting point for an investigation that involves data collection and data analysis

stratified sampling

a population is divided into two or more groups (called strata) according to some criterion, and a sample is selected from each strata using simple random sampling or systematic sampling.

survey question

questions researchers ask in order to collect data, which is expected to vary from individual to individual

systematic sampling

every individual in the population is given a number and individuals/entities are chosen at regular intervals, with a random starting point

qualitative data, categorical data

categorizing or describing attributes of a population

quantitative continuous data

data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers

quantitative data

counting or measuring attributes of a population

quantitative discrete data

data that can take on only certain numerical values

unbiased

a representative sample of the population

undercoverage

when some groups of the population are left out of the sampling process and the individuals in these groups do not have an equal chance of being selected for the sample

variables

the characteristics of observational units

variability

the variance between data points

voluntary response bias

people who volunteer for a study or survey may be more inclined to respond to questions or report certain behaviors