Outlier

Section Summary | Statistical Thinking

Introduction to Statistics

1. What is statistics?
    a. Statistics is a branch of mathematics focused on collecting, analyzing, and interpreting data
    b. It is the science of learning from data
    c. It tries to model variation in the world to make predictions and make decisions
    d. It answers complex questions with numerical results that are easy to interpret
    e. We just need to learn how to properly collect, analyze, and interpret data
2. Why is statistics important?
    a. Statistics is used virtually everywhere from the government to sports to business
    b. You need to understand statistics to be a more informed citizen
    c. Statistics can be used in your own life to make decisions, or it can be used to critique the statistics you see in the media
3. What will you take away?
    a. There is far more to statistics than just crunching numbers
    b. We will focus on understanding and interpreting results
    c. You will have a working knowledge of statistical methods that you can apply to the real world
    d. You will be able to:
        i. Summarize data
        ii. Use probabilities to make decisions
        iii. Gamble effectively
        iv. Standardize complex datasets
        v. Predict the future

The Statistical Process

1. Keywords in statistics
    a. Data: characteristics or information that are collected through observation
        i. Measurements
        ii. Survey responses
        iii. Ratings
    b. Statistics: the science of collecting, analyzing, and interpreting data

- c. Descriptive statistics: methods for organizing and summarizing information
  - i. Graphs
  - ii. Charts
  - iii. Measures of center and spread
- d. Inferential statistics: methods for drawing conclusions about a population based on information from a sample
- e. Population: complete collection of all individuals or items being considered
- f. Sample: a subcollection of members selected from a population

2. Steps to a statistical analysis
   - a. Identify a question
     - i. What is the question I want to answer?
     - ii. What are my hypotheses?
   - b. Collect the data
     - i. Can I use data that already exists?
     - ii. Do I need to collect data myself?
   - c. Analyze the data
     - i. How should I graph and explore the data?
     - ii. What statistical method is appropriate?
   - d. Draw conclusions
     - i. Do my results have statistical significance?
     - ii. How can I accurately report my findings?

Data + Sampling

1. Types of data
    a. Parameter vs. Statistic
        i. A parameter describes some measurement of a population
        ii. A statistic describes some measurement of a sample
    b. Quantitative vs. Qualitative Data
        i. Quantitative data is numerical data that consists of numbers representing counts or measurements
        ii. Qualitative data is categorical data that consists of names or labels
    c. Discrete vs. Continuous Data
        i. Discrete data can only take on certain values, and the number of allowed values is countable
        ii. Continuous data can take on any value within a range, and the values are uncountably infinite
        iii. Continuous data usually requires a measuring device to collect
    d. Missing Data
        i. Missing at random: the value is just as likely to be missing as any other value (a random mistake has occurred)
        ii. Missing not at random: there is some underlying reason for the value to be missing
        iii. When a data point has a value missing at random, we can usually just remove that data point from the sample without affecting the results
        iv. When a data point has a value missing not at random, we cannot remove that data point because the underlying reason for the missing data will skew our results
        v. We can usually fill in the missing data through a process called “Imputing the missing value”
2. Collecting samples from a population
    a. Simple random sample: any combination of the same number of members has the same chance of being selected from the population
    b. Systematic sample: using a consistent method/procedure to select members for your sample
    c. Stratified sample: separate the population into groups (called strata) based on certain characteristics and then randomly select members from each group

d. Cluster sample: partition the population into different clusters, randomly select certain clusters, then take all the members from those selected clusters into the sample

e. Convenience sample: collect members based on convenience (this will usually not be totally random because the convenience may contain some underlying factor that skews the sample)

f. Sampling error

  i. Even though you use a random sampling method, there may still be discrepancy between your sample result and the actual population result

  ii. This occurs because of chance fluctuations in the sample selection

  iii. Combat this error by taking larger samples

Experimental Design + Ethics

1. Ways to collect data
    a. Observation: observe what’s already true
        i. Cannot control/isolate variables
        ii. Difficult to establish cause and effect relationships
    b. Experiment: impose a change on the subject and observe the results
        i. Typically more effective than observation
        ii. You can control variables
        iii. Once you isolate the variable of interest, you can determine cause and effect relationships
2. 3 elements of a good experiment
    a. Randomization: subjects are assigned to groups randomly
        i. Avoid bias between groups by using good random methods
        ii. The characteristics between groups should be consistent
    b. Replication: experiment is repeated on different subjects
        i. Ensure a large sample size
        ii. Larger sample sizes allow us to observe the full range of results
    c. Blinding: subjects don’t know what group they’re in
        i. This prevents the Placebo Effect
        ii. Placebo Effect: untreated subjects report a change
        iii. Double-Blind Experiment: both the subjects and the researchers don’t know who’s in which group
3. Ethics
    a. Health and safety of human subjects
        i. Subjects are fully informed of elements of the experiment and associated risks
        ii. Subjects must give full consent to participate
    b. Sampling bias
        i. If the sampling method is not random, the data can be biased
        ii. Reporting this biased data would be an ethics violation
        iii. Avoid bias with good random sampling methods
    c. Data collection and analysis
        i. Using inappropriate methods will result in illegitimate data
        ii. Falsifying results to support the desired conclusion will also produce illegitimate data
        iii. Reporting such illegitimate data is an ethics violation
4. How to be an informed reader

a. Questions to ask about the source of the study:
   i. Who is reporting this data?
   ii. Are there conflicts of interest?
   iii. Who is funding the study?

b. Questions to ask about the statistics:
   i. Was there a large enough unbiased sample?
   ii. Is the data new and relevant or old and outdated?
   iii. Are complete details of the methods and assumptions provided?
   iv. Was the study peer-reviewed by experts before publication?

5. How to ensure all your reports are ethical
   a. Report all data
   b. Be honest in your reporting
   c. Report your methods and assumptions completely
   d. Report any conflicts of interest
   e. Collaborate with peers and experts in the field
   f. Cite any contributors to your study appropriately