Chapter 1
Introduction to Statistics

1-1 Overview
1-2 Types of Data
1-3 Critical Thinking
1-4 Design of Experiments

Overview

A common goal of surveys and other data collecting tools is to collect data from a smaller part of a larger group so we can learn something about the larger group.

In this section we will look at some of ways to describe data.

Definitions

Data

observations (such as measurements, genders, survey responses) that have been collected.

Statistics

a collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.

Population

the complete collection of all elements (scores, people, measurements, and so on) to be studied. The collection is complete in the sense that it includes all subjects to be studied.

Census

the collection of data from every member of the population.

Sample

a sub-collection of elements drawn from a population.
Key Concepts

- Sample data must be collected in an appropriate way, such as through a process of random selection.
- If sample data are not collected in an appropriate way, the data may be so completely useless that no amount of statistical torturing can salvage them.

Definitions

- **Parameter**: a numerical measurement describing some characteristic of a population.

Definitions

- **Statistic**: a numerical measurement describing some characteristic of a sample.

Definitions

- **Quantitative data**: numbers representing counts or measurements.
  Example: weights of supermodels.

Definitions

- **Qualitative (or categorical or attribute) data**: can be separated into different categories that are distinguished by some nonnumeric characteristics.
  Example: genders (male/female) of professional athletes.
Quantitative data can further be distinguished between discrete and continuous types.

Discrete data result when the number of possible values is either a finite number or a ‘countable’ number of possible values. 0, 1, 2, 3, . . .
Example: The number of eggs that hens lay.

Continuous data result from infinitely many possible values that correspond to some continuous scale that covers a range of values without gaps, interruptions, or jumps.

Example: The amount of milk that a cow produces; e.g. 2.343115 gallons per day.

Another way to classify data is to use levels of measurement. Four of these levels are discussed in the following slides.

Nominal level of measurement characterized by data that consist of names, labels, or categories only. The data cannot be arranged in a meaningful ordering scheme (such as low to high)
Example: survey responses yes, no, undecided

Ordinal level of measurement involves data that may be arranged in some order, but differences between data values either cannot be determined or are meaningless
Example: Course grades A, A-, B+, B, B-, etc.
Definitions

- **interval level of measurement**
  like the ordinal level, with the additional property that the difference between any two data values is meaningful. However, there is no natural zero starting point (where none of the quantity is present).
  
  Example: Years 1000, 2000, 1776, and 1492

- **ratio level of measurement**
  the interval level modified to include the natural zero starting point (where zero indicates that none of the quantity is present). For values at this level, differences and ratios are meaningful.
  
  Example: Prices of college textbooks ($0 represents no cost)

Definitions

Summary - Levels of Measurement

- **Nominal** - categories only
- **Ordinal** - categories with some order
- **Interval** - differences but no natural starting point
- **Ratio** - differences and a natural starting point

Recap

In Sections 1-1 and 1-2 we have looked at:

- Basic definitions and terms describing data
- Parameters versus statistics
- Types of data (quantitative and qualitative)
- Levels of measurement

Success in Statistics

- Success in the introductory statistics course typically requires more common sense than mathematical expertise.
- This section is designed to illustrate how common sense is used when we think critically about data and statistics.

Section 1-3

Critical Thinking

Created by Tom Wegleitner, Centreville, Virginia
Definitions

- Voluntary response sample
  (or self-selected survey)
  one in which the respondents themselves decide whether to be included.
  In this case, valid conclusions can be made only about the specific group of people who agree to participate.

Misuses of Statistics

- Voluntary response sample
- Small Samples
- Misleading Graphs

Conclusions should not be based on samples which are far too small.

Bar graphs and pie charts can be used to exaggerate or understate the true nature of data.

To correctly interpret a graph, we should analyze the numerical information given in the graph instead of being mislead by its general shape.

Drawing of objects, called Pictographs, may also be misleading.
Double the length, width, and height of a cube, and the volume increases by a factor of eight

Figure 1-2

Misuses of Statistics

- Bad Samples
- Small Samples
- Misleading Graphs
- Pictographs
- Distorted Percentages
- Loaded Questions

Survey questions can be "loaded" or intentionally worded to elicit a desired response.

Survey questions are often used.

97% yes: “Should the President have the line item veto to eliminate waste?”

57% yes: “Should the President have the line item veto, or not?”

Misuses of Statistics

- Small Samples
- Misleading Graphs
- Pictographs
- Distorted Percentages
- Loaded Questions

Sometimes survey questions can be unintentionally loaded by presenting certain order of question prompting questions to elicit a desired response.

97% yes: “Should the President have the line item veto to eliminate waste?”

57% yes: “Should the President have the line item veto, or not?”

Recap

In this section we have:

- Reviewed 13 misuses of statistics.
- Illustrated how common sense can play a big role in interpreting data and statistics.

Section 1-4: Design of Experiments

Created by Tom Wegleitner, Centreville, Virginia
Broadly, there could be two kinds of study in Statistics

- **Observational Study**
  - observing and measuring specific characteristics without attempting to modify the subjects being studied

- **Experiment**
  - apply some treatment and then observe its effects on the subjects

Different types of Observational Studies

- **Cross Sectional Study**
  - Data are observed, measured, and collected at one point in time.

- **Retrospective (or Case Control) Study**
  - Data are collected from the past by going back in time.

- **Prospective (or Longitudinal or Cohort) Study**
  - Data are collected in the future from groups (called cohorts) sharing common factors.

**Random Sample**
- members of the population are selected in such a way that each individual member has an equal chance of being selected

**Simple Random Sample** (of size \( n \))
- subjects selected in such a way that every possible sample of the same size \( n \) has the same chance of being chosen

**Systematic Sampling**
- Select some starting point and then select every \( K \)th element in the population
Convenience Sampling
use results that are easy to get

Stratified Sampling
subdivide the population into at least two different subgroups that share the same characteristics, then draw a sample from each subgroup (or stratum)

Cluster Sampling
divide the population into sections (or clusters); randomly select some of those clusters; choose all members from selected clusters

Methods of Sampling
- Random
- Systematic
- Convenience
- Stratified
- Cluster

Definitions
- Sampling Error
  the difference between a sample result and the true population result; such an error results from chance sample fluctuations
- Nonsampling Error
  sample data that are incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly)