Lecture Slides

Essentials of Statistics
5th Edition

and the Triola Statistics Series

by Mario F. Triola

Preview

Polls, studies, surveys and other data collecting tools collect data from a small part of a larger group so that we can learn something about the larger group.

This is a common and important goal of statistics: Learn about a large group by examining data from some of its members.

Preview

In this context, the terms *sample* and *population* have special meaning. Formal definitions for these and other basic terms will be given here.

In this chapter, we will look at some of the ways to describe data.

11

Chapter 1
Introduction to Statistics

Data

- Data -

  Collections of observations, such as measurements, genders, or survey responses

Statistics

- Statistics -

  The science of planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data

Population

- **Population -**

  The complete collection of *all* measurements or data that are being considered

Census versus Sample

- **Census -**

  Collection of data from ***every*** member of a population

- **Sample -**

  Subcollection of members selected from a population

Example

- The Gallup corporation collected data from 1013 adults in the United States. Results showed that 66% of the respondents worried about identity theft.
- The population consists of all 241,472,385 adults in the United States.
- The sample consists of the 1013 polled adults.
- The objective is to use the sample data as a basis for drawing a conclusion about the whole population.

22

Chapter 1
Introduction to Statistics

Key Concept

This section provides an overview of the process involved in conducting a statistical study:

•Prepare

•Analyze

•Conclude

Prepare - Context

- What do the data mean?
- What is the goal of the study?

Prepare - Source of the Data

- Is the source objective?
- Is the source biased?
- Be vigilant and skeptical of studies from sources that may be biased.

Prepare - Sampling Method

- Does the method chosen greatly influence the validity of the conclusion?
- Voluntary response (or self-selected) samples often have bias (those with special interest are more likely to participate).
- Other methods are more likely to produce good results.

Analyze – Graph and Explore

- Every analysis should begin with appropriate graphs (Chapter 2).

33

Analyze – Apply Statistical Methods

- Later chapters describe important statistical methods.
- With technology, good analysis does not require strong computational skills, but it does require using common sense and paying attention to sound statistical methods.

Conclude – Statistical Significance

- *Statistical significance* is achieved in a study when we get a result that is very unlikely to occur by chance.

Conclude - Practical Significance

- State practical implications of the results.
- Common sense might suggest that the finding does not make enough of a difference to justify its use or to be practical.

Example

- In a test of the Atkins weight loss program, 40 subjects had a mean weight loss of 4.6 pounds after one year.
- Using formal methods of statistical analysis, we can conclude the diet appears to be effective.

Example - continued

- However, although 4.6 pounds is statistically significant, using common sense, it does not seem very worthwhile.

Potential Pitfalls – Misleading Conclusions

- Concluding that one variable *causes* the other variable when in fact the variables are only correlated or associated together.

  Two variables that may seemed linked, are smoking and pulse rate.

  We cannot conclude the one causes the other. *Correlation does not imply causality*.

Potential Pitfalls - Small Samples

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

Example: Basing a school suspension rate on a sample of only three students

Potential Pitfalls - Loaded Questions

If survey questions are not worded carefully, the results of a study can be misleading.

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?”

Potential Pitfalls - Order of Questions

Questions are unintentionally loaded by such factors as the order of the items being considered.

Would you say traffic contributes more or less to air pollution than industry? Results: traffic - 45%; industry - 27%

When order reversed.
Results: industry - 57%; traffic - 24%

Potential Pitfalls - Nonresponse

Occurs when someone either refuses to respond to a survey question or is unavailable.

People who refuse to talk to pollsters have a view of the world around them that is markedly different than those who will let pollsters into their homes.

Potential Pitfalls - Missing Data

Can dramatically affect results.

Subjects may drop out for reasons unrelated to the study.

Example - People with low incomes are less likely to report their incomes.

Example – U.S. Census suffers from missing people (tend to be homeless or low income).

Potential Pitfalls - Precise Numbers

Because as a figure is precise, many people incorrectly assume that it is also *accurate*.

A precise number can be an estimate, and it should be referred to that way.

Potential Pitfalls - Percentages

Misleading or unclear percentages are sometimes used.

Example – Continental Airlines ran an ad claiming “We’ve already improved 100% in the last six months” with respect to lost baggage.

Does this mean Continental made no mistakes?

Chapter 1
Introduction to Statistics

Key Concept

The subject of statistics is largely about using sample data to make inferences about an entire population.

It is essential to know and understand the definitions that follow.

Parameter

- **Parameter**

  a numerical measurement describing some characteristic of a population.

population

parameter

Statistic

- **Statistic**

  a numerical measurement describing some characteristic of a sample.

sample

statistic

Quantitative Data

- **Quantitative (or numerical) data**

  consists of *numbers* representing counts or measurements.

  Example: The weights of supermodels

  Example: The ages of respondents

Categorical Data

- **Categorical (or qualitative or attribute) data**

  consists of names or labels (representing categories).

Example: The gender (male/female) of professional athletes

Example: Shirt numbers on professional athletes uniforms - substitutes for names.

Working with Quantitative Data

Quantitative data can be further described by distinguishing between *discrete* and *continuous* types.

Discrete Data

- **Discrete data**

  result when the number of possible values is either a finite number or a 'countable' number (i.e. the number of possible values is 0, 1, 2, 3, . . .).

  Example: The number of eggs that a hen lays

Continuous Data

❖ Continuous (numerical) 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

Levels of Measurement

Another way to classify data is to use levels of measurement.

Nominal Level

❖ Nominal level of measurement

characterized by data that consist of names, labels, or categories only, and the data cannot be arranged in an ordering scheme (such as low to high).

Example: Survey responses *yes*, *no*, *undecided*

77

Ordinal Level

❖ Ordinal level of measurement

involves data that can be arranged in some order, but differences between data values either cannot be determined or are meaningless.

Example: Course grades A, B, C, D, or F

Interval Level

❖ Interval level of measurement

involves data that can be arranged in order and 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

❖ Ratio level of measurement

the interval level with the additional property that there is also a 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, a $100 book costs twice as much as a $50 book)

Summary - Levels of Measurement

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

Chapter 1
Introduction to Statistics

Key Concept

- 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.
- The method used to collect sample data influences the quality of the statistical analysis.
- Of particular importance is the *simple random sample*.

Basics of Collecting Data

Statistical methods are driven by the data that we collect. We typically obtain data from two distinct sources: *observational studies* and *experiment*.

Observational Study

- Observational study

  observing and measuring specific characteristics without attempting to *modify* the subjects being studied.

Experiment

- Experiment

  apply some *treatment* and then observe its effects on the subjects (subjects in experiments are called experimental units)

Example

- The Pew Research Center surveyed 2252 adults and found that 59% of them go online wirelessly.
- This an observational study because the adults had no treatment applied to them.

Example

- In the largest public health experiment ever conducted, 200,745 children were given the Salk vaccine, while another 201,229 children were given a placebo.
- The vaccine injections constitute a treatment that modified the subjects, so this is an example of an experiment.

Simple Random Sample

- Simple Random Sample

  A sample of *n* subjects is selected in such a way that every possible *sample of the same size n* has the same chance of being chosen.

99

Random Sample

- Random Sample

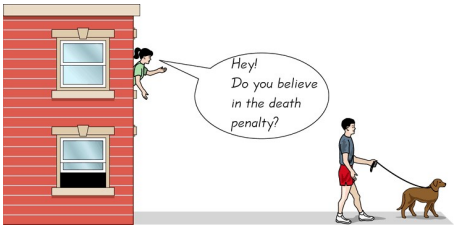
  Members from the population are selected in such a way that each individual member in the population has an equal chance of being selected.

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.

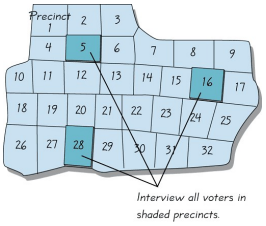


Stratified Sampling

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 area into sections (or clusters). Then randomly select some of those clusters. Now choose *all* members from selected clusters.



Multistage Sampling

Collect data by using some combination of the basic sampling methods.

In a multistage sample design, pollsters select a sample in different stages, and each stage might use different methods of sampling.

100

Methods of Sampling - Summary

- Random
- Systematic
- Convenience
- Stratified
- Cluster
- Multistage

Beyond the Basics of Collecting Data

Different types of observational studies and experiment design.

Types of 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 (examine records, interviews, and so on …).

- Prospective (or longitudinal or cohort) study

  Data are collected in the future from groups sharing common factors (called *cohorts*).

Design of Experiments

- **Randomization**

  is used when subjects are assigned to different groups through a process of random selection. The logic is to use chance as a way to create two groups that are similar.

Design of Experiments

- **Replication**

  is the repetition of an experiment on more than one subject.

  Samples should be large enough so that the erratic behavior that is characteristic of very small samples will not disguise the true effects of different treatments.

  It is used effectively when there are enough subjects to recognize the differences from different treatments.

Design of Experiments

- **Replication**

Use a sample size that is large enough to let us see the true nature of any effects, and obtain the sample using an appropriate method, such as one based on *randomness*.

Design of Experiments

- **Blinding**

  is a technique in which the subject doesn't know whether he or she is receiving a treatment or a placebo.

  Blinding allows us to determine whether the treatment effect is significantly different from a placebo effect, which occurs when an untreated subject reports improvement in symptoms.

Design of Experiments

- **Double-Blind**

  Blinding occurs at two levels:

  (1) The subject doesn't know whether he or she is receiving the treatment or a placebo.

  (2) The experimenter does not know whether he or she is administering the treatment or placebo.

Design of Experiments

- **Confounding**

  occurs in an experiment when the experimenter is not able to distinguish between the effects of different factors.

  Try to plan the experiment so that confounding does not occur.

Controlling Effects of Variables

- Completely Randomized Experimental Design
  assign subjects to different treatment groups through a process of *random selection*.

- Randomized Block Design
  a block is a group of subjects that are similar, but blocks differ in ways that might affect the outcome of the experiment.

Controlling Effects of Variables

- Matched Pairs Design
  compare exactly two treatment groups using subjects matched in pairs that are somehow related or have similar characteristics.

- Rigorously Controlled Design
  carefully assign subjects to different treatment groups, so that those given each treatment are similar in ways that are important to the experiment.

Summary

Three very important considerations in the design of experiments are the following:

1. Use *randomization* to assign subjects to different groups.
2. Use replication by repeating the experiment on enough subjects so that effects of treatment or other factors can be clearly seen.
3. *Control the effects of variables* by using such techniques as blinding and a completely randomized experimental design.

Errors

No matter how well you plan and execute the sample collection process, there is likely to be some error in the results.

- 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 incorrectly collected, recorded, or analyzed (such as by selecting a biased sample, using a defective instrument, or copying the data incorrectly).

Errors

No matter how well you plan and execute the sample collection process, there is likely to be some error in the results.

- Nonrandom sampling error
  result of using a sampling method that is not random, such as using a convenience sample or a voluntary response sample.