Statistics

- **Statistics**

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

Introduction to Statistics

1-1 Review and Preview

1-2 Statistical Thinking

1-3 Types of Data

1-4 Critical Thinking

1-5 Collecting Sample Data

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

Data

- **Data**

  collections of observations (such as measurements, genders, survey responses)

Statistics

- **Statistics**

  is 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 individuals (scores, people, measurements, and so on) to be studied; the collection is complete in the sense that it includes *all* of the individuals to be studied

Census versus Sample

- **Census**

  Collection of data from *every* member of a population

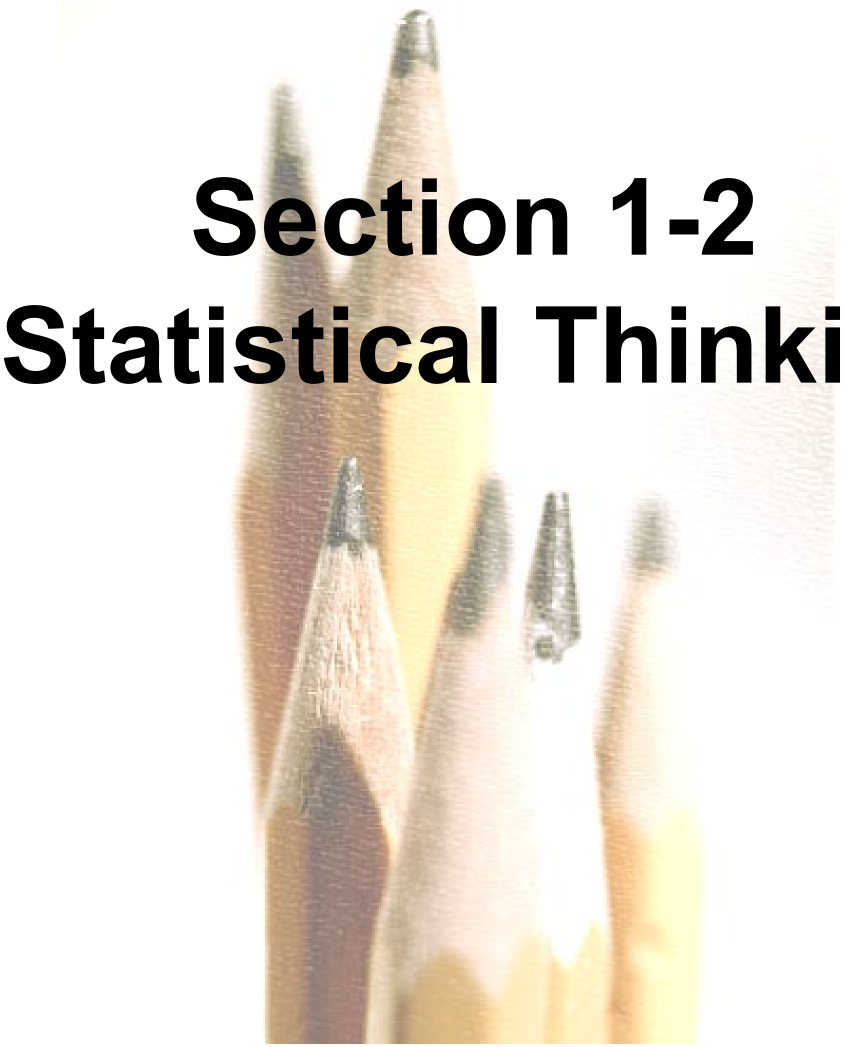
- **Sample**

  Subcollection of members selected from a population

Chapter 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.



Section 1-2
Statistical Thinking

Key Concept

This section introduces basic principles of statistical thinking used throughout this discourse. Whether conducting statistical analysis of data that we have collected, or analyzing a statistical analysis done by someone else, we should not rely on blind acceptance of mathematical calculation. We should consider these factors:

Key Concept (continued)

- Context of the data
- Source of the data
- Sampling method
- Conclusions
- Practical implications

Context

- What do the values represent?
- Where did the data come from?
- Why were they collected?
- An understanding of the context will directly affect the statistical procedure used.

Source of data

- Is the source objective?
- Is the source biased?
- Is there some incentive to distort or spin results to support some self-serving position?
- Is there something to gain or lose by distorting results?
- Be vigilant and skeptical of studies from sources that may be biased.

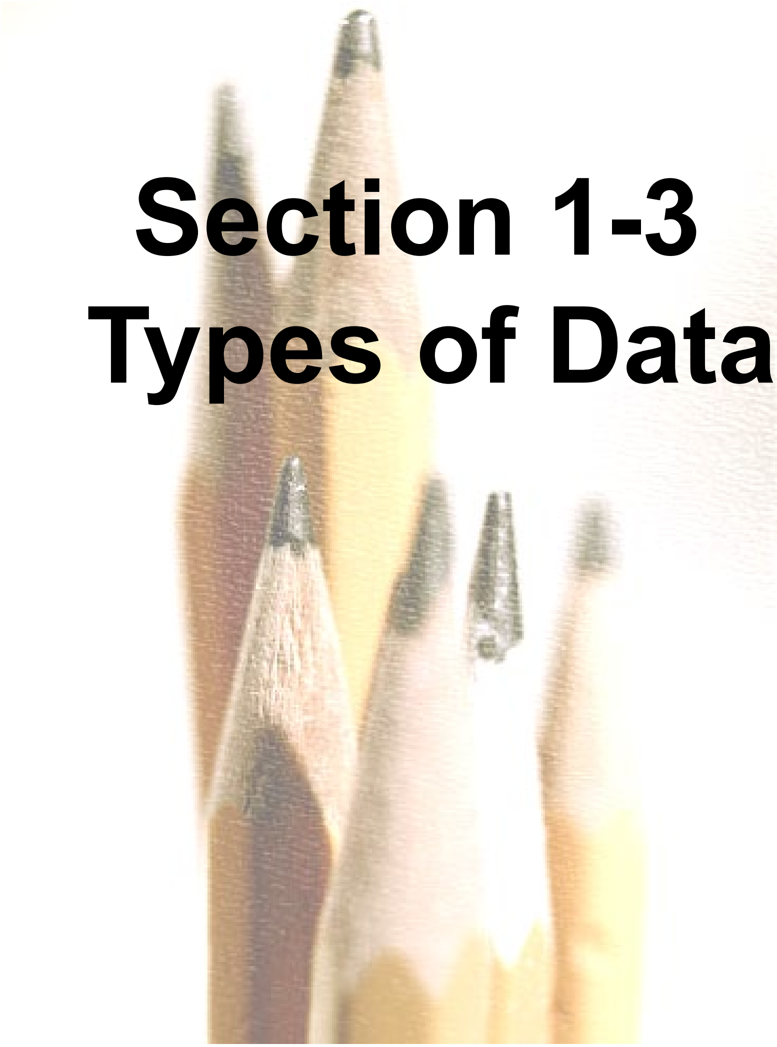
Sampling Method

- 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). These samples' results are not necessarily valid.
- Other methods are more likely to produce good results.

Conclusions

- Make statements that are clear to those without an understanding of statistics and its terminology.
- Avoid making statements not justified by the statistical analysis.

Section 1-3
Types of Data



Key Concept

The subject of statistics is largely about using sample data to make inferences (or generalizations) 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 genders (male/female) of professional athletes

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

Working with Quantitative Data

Quantitative data can further be 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. Four of these levels are discussed in the following slides.

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

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**

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

- **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 starting point
- **Ratio** - differences and a natural starting point

Recap

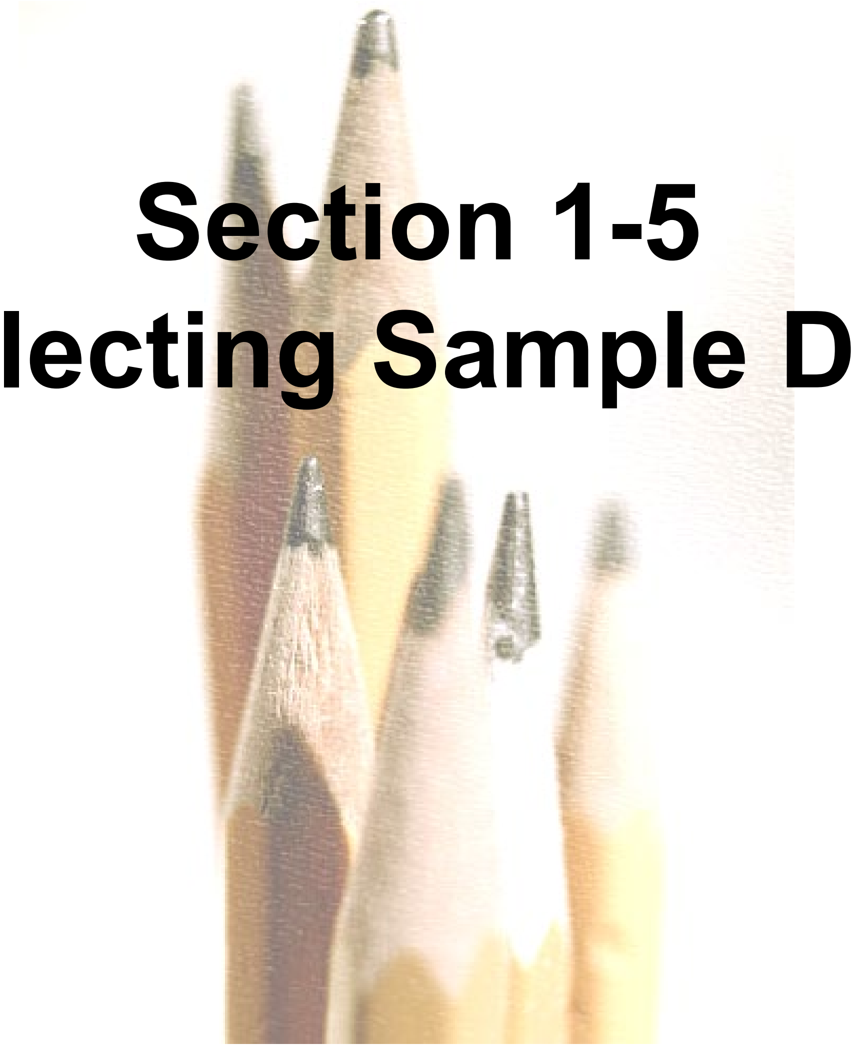
In this section we have looked at:

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

Key Concepts

- Success in the introductory statistics course typically requires more common sense than mathematical expertise.
- Improve skills in interpreting information based on data.
- This section is designed to illustrate how common sense is used when we think critically about data and statistics.
- Think carefully about the context, source, method, conclusions and practical implications.

Section 1-5
Collecting Sample Data



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.
- Method used to collect sample data influences the quality of the statistical analysis.
- Of particular importance is *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)

Simple Random Sample

- **Simple Random Sample**

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

Random & Probability Samples

- **Random Sample**

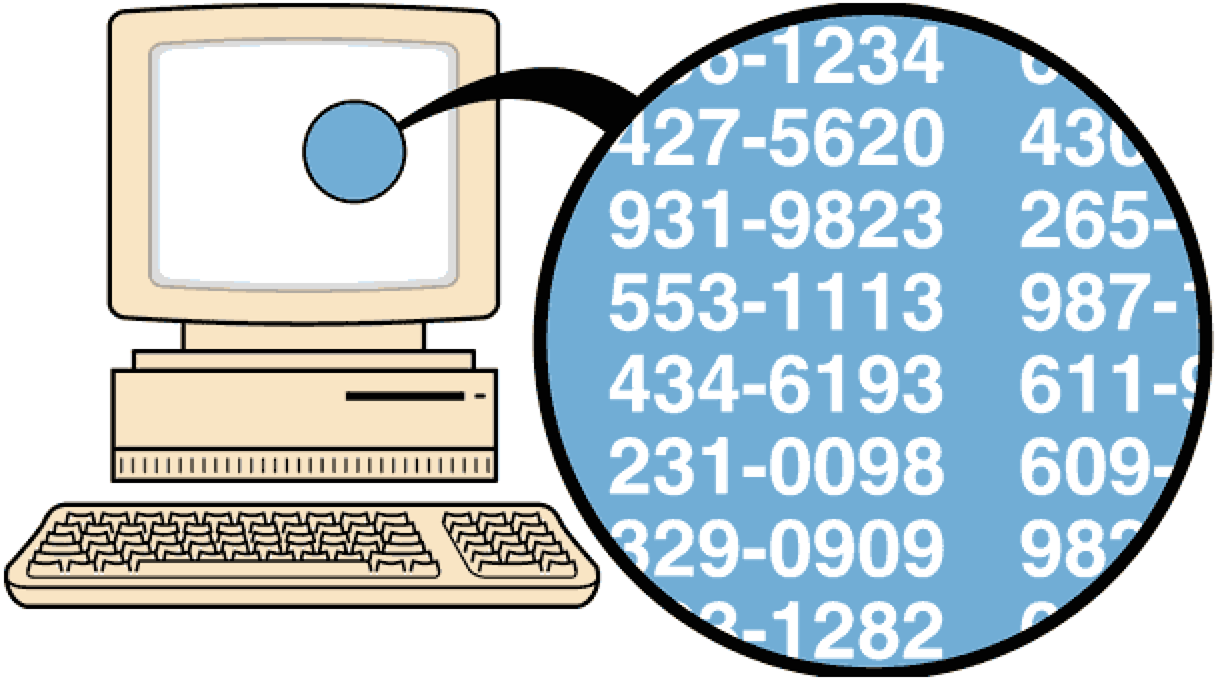
  members from the population are selected in such a way that each individual member in the population has an equal chance of being selected

- **Probability Sample**

  selecting members from a population in such a way that each member of the population has a known (but not necessarily the same) chance of being selected

Random Sampling

selection so that each individual member has an equal chance of being selected

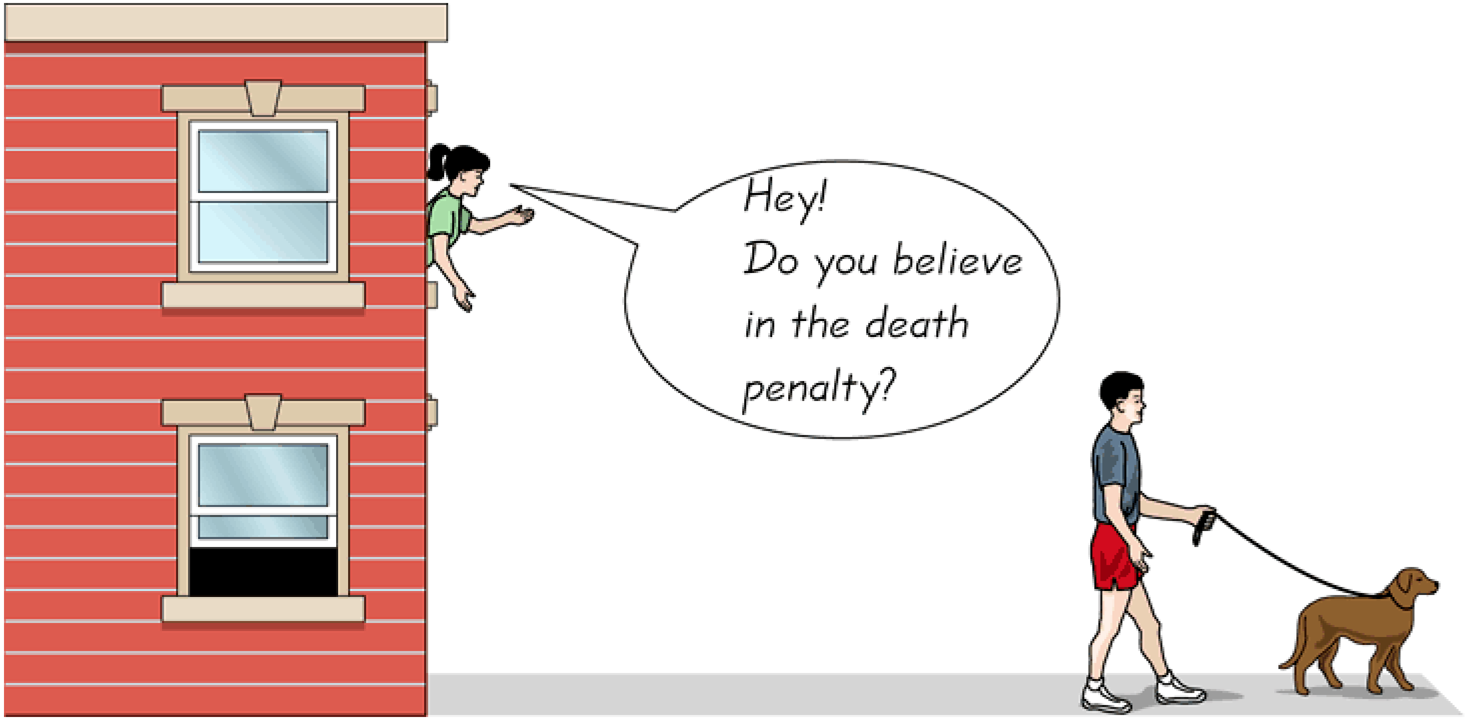


Systematic Sampling

Select some starting point and then select every *k*th element in the population

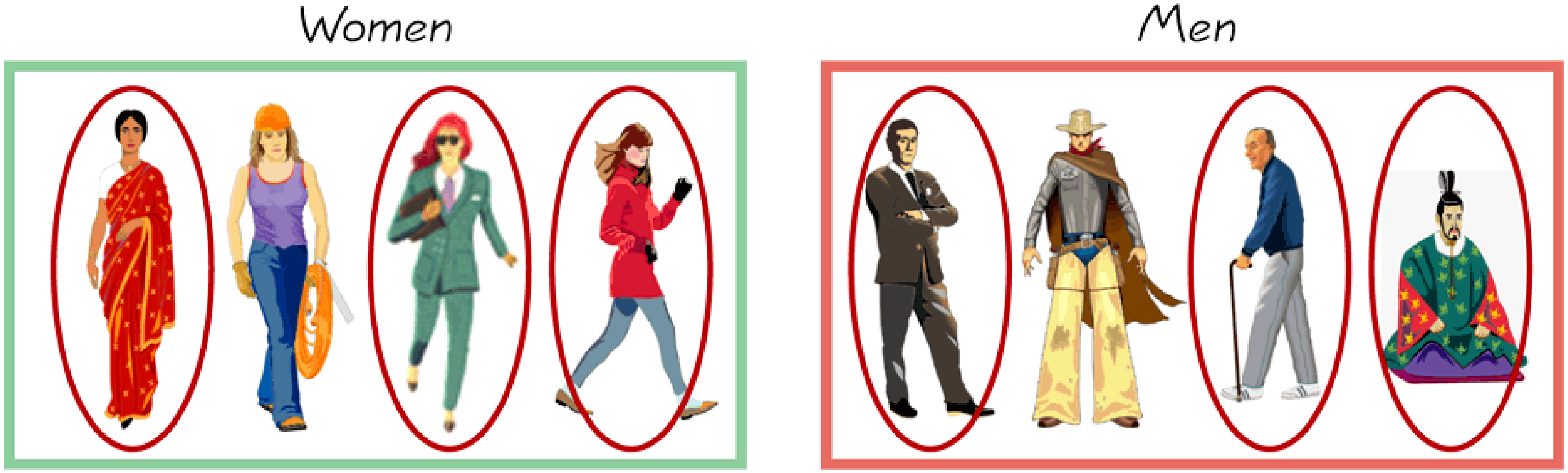
Convenience Sampling

use results that are easy to get



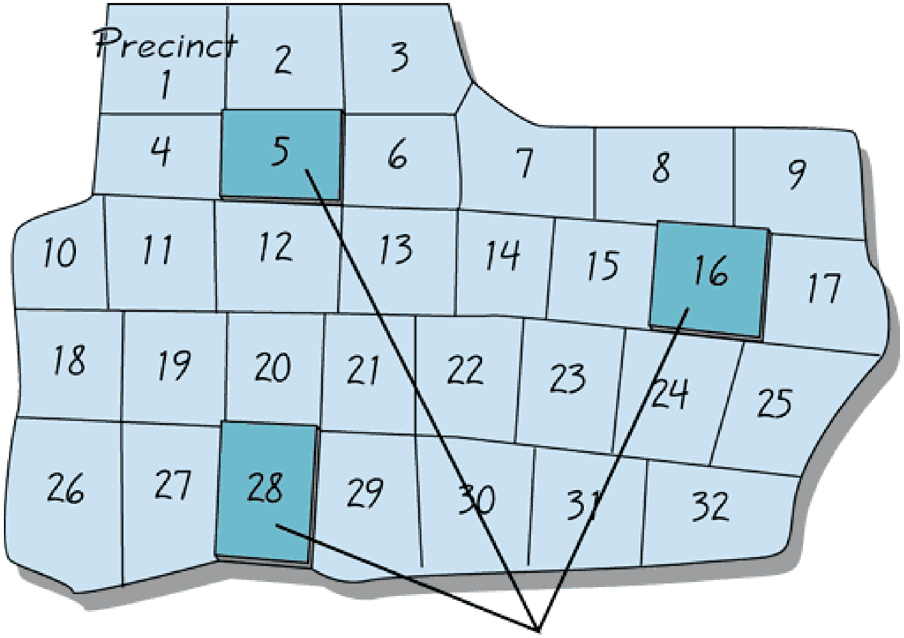
Stratified Sampling

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); randomly select some of those clusters; choose all members from selected clusters



Interview all voters in shaded precincts.

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

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, …)

- **Prospective (or longitudinal or cohort) study**

  data are collected in the future from groups sharing common factors (called cohorts)

Randomization

- **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.

Replication

- **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.

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*.

Blinding

- **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.

Double Blind

- **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

Confounding

- **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

- **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

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

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)

Recap

In this section we have looked at:

- Types of studies and experiments
- Controlling the effects of variables
- Randomization
- Types of sampling
- Sampling errors