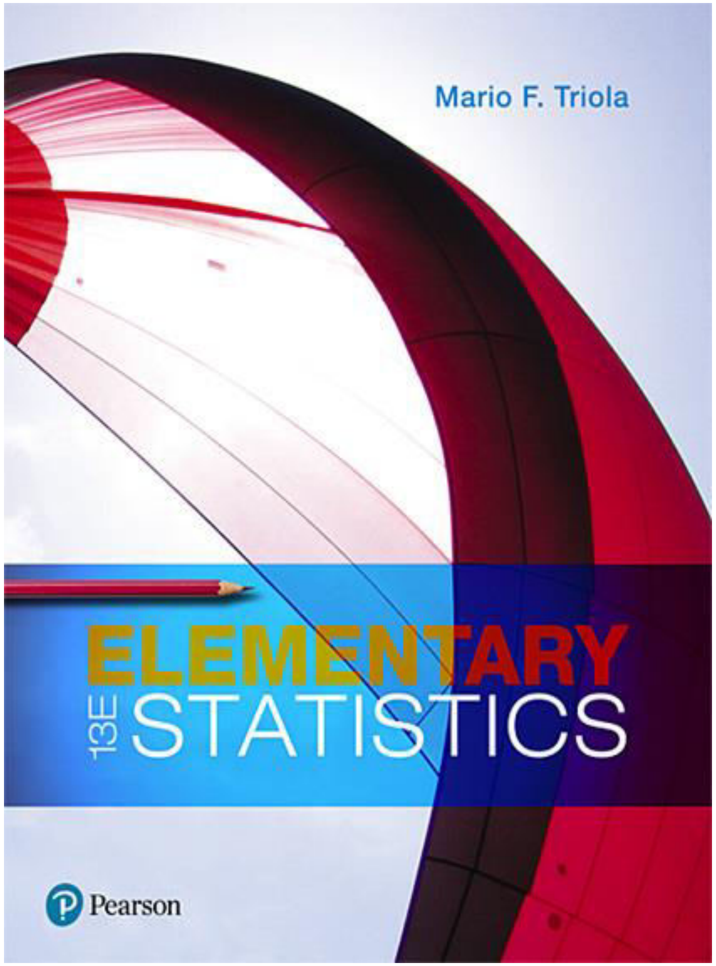
Elementary Statistics

Thirteenth Edition



Chapter 1

Introduction to Statistics

Introduction to Statistics

1-1 Statistical and Critical Thinking

1-2 Types of Data

1-3 Collecting Sample Data

Key Concept

The method used to collect sample data influences the quality of the statistical analysis.

Of particular importance is the **simple random sample**.

If sample data are not collected in an appropriate way, the data may be so utterly useless that no amount of statistical torturing can salvage them.

The Gold Standard

Randomization with placebo/treatment groups is sometimes called the “gold standard” because it is so effective. (A placebo such as a sugar pill has no medicinal effect.)

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 **experiments**.

Experiment

- Experiment
  - apply some **treatment** and then proceed to observe its effects on the individuals. (The individuals in experiments are called experimental units, and they are often called subjects when they are people.)

Observational Study

- Observational study
  - observing and measuring specific characteristics without attempting to **modify** the individuals being studied

Example: Ice Cream and Drownings (1 of 2)

- Observational Study:

  Observe past data to conclude that ice cream causes drownings (based on data showing that increases in ice cream sales are associated with increases in drownings). The mistake is to miss the lurking variable of temperature and the failure to see that as the temperature increases, ice cream sales increase and drownings increase because more people swim.

Example: Ice Cream and Drownings (2 of 2)

- Experiment:
  – Conduct an **experiment** with one group treated with ice cream while another group gets no ice cream. We would see that the rate of drowning victims is about the same in both groups, so ice cream consumption has no effect on drownings.
  – Here, the experiment is clearly better than the observational study.

Design of Experiments (1 of 4)

- Replication
  - Replication is the repetition of an experiment on more than one individual.
  - Good use of replication requires sample sizes that are large enough so that we can see effects of treatments.

Design of Experiments (2 of 4)

- Blinding
  - Blinding is a technique in which the subject doesn’t know whether he or she is receiving a treatment or a placebo.
  - Blinding is a way to get around the placebo effect, which occurs when an untreated subject reports an improvement in symptoms.

Design of Experiments (3 of 4)

- 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 (4 of 4)

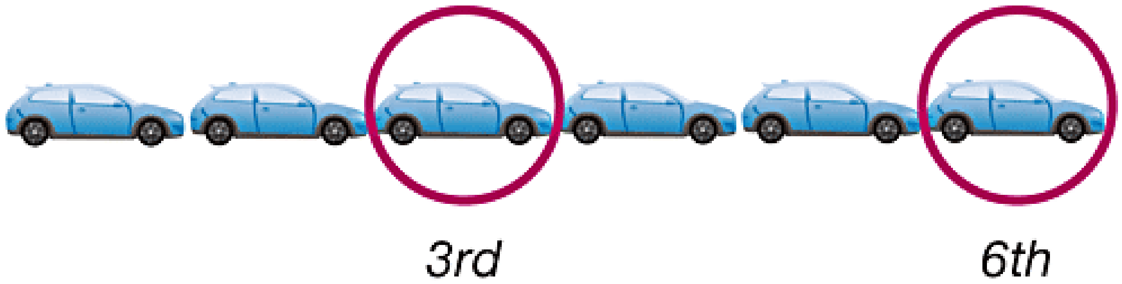
- Randomization
  - Randomization is used when subjects are assigned to different groups through a process of random selection. The logic is to use chance as a way to create two groups that are similar.

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.
  - A simple random sample is often called a random sample, but strictly speaking, a **random sample** has the weaker requirement that all members of the population have the same chance of being selected.

Systematic Sampling

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

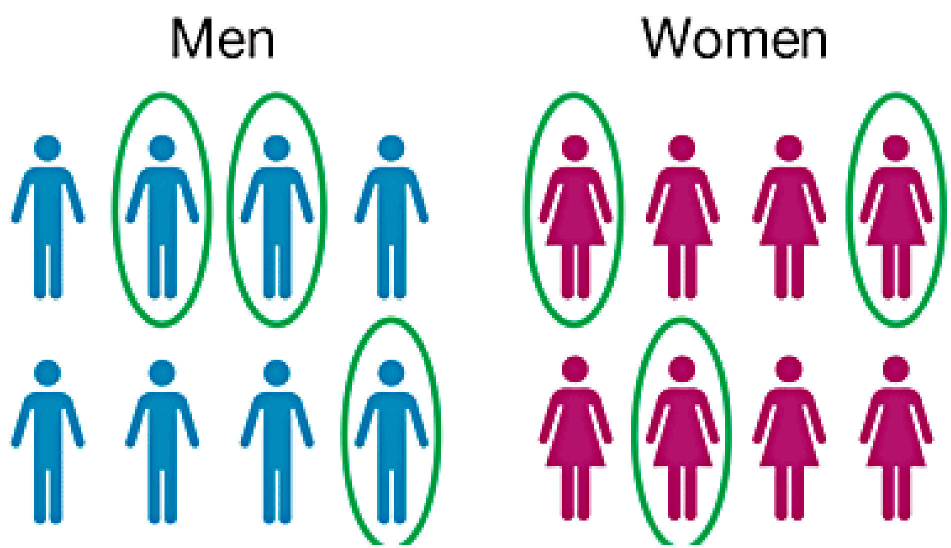


Convenience Sampling

- Use data that are very easy to get.

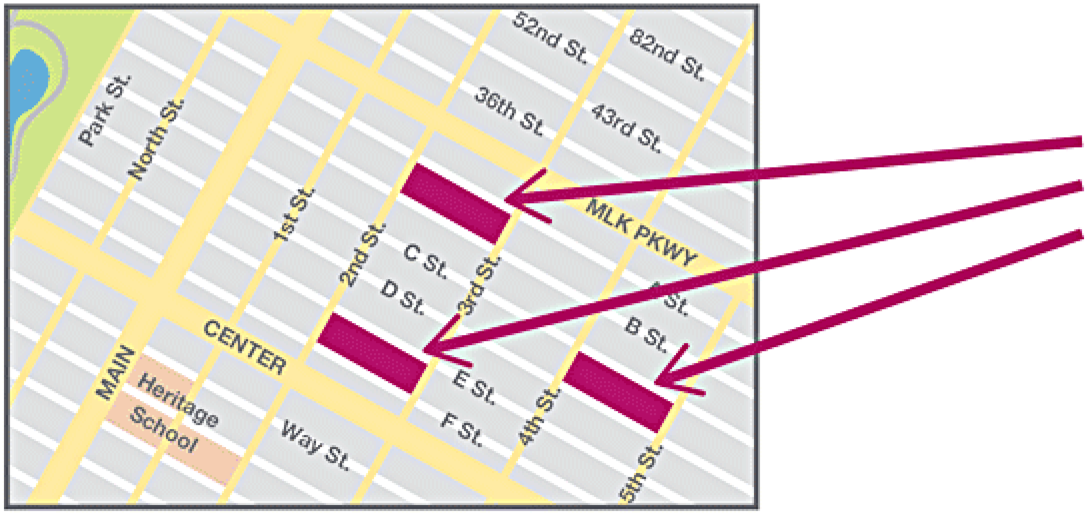
Stratified Sampling

Subdivide the population into at least two different subgroups (or strata) so that the subjects within the same subgroup 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, and choose **all** the members from those 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.

Observational Studies

Observe and measure, but do not modify.

Types of Observational Studies

- Cross-sectional study
  - Data are observed, measured, and collected at one point in time, not over a period of time.
- Retrospective (or case control) study
  - Data are collected from a past time period by going back in time (through examination of records, interviews, and so on).
- Prospective (or longitudinal or cohort) study
  - Data are collected in the future from groups sharing common factors (called **cohorts**).

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 (1 of 2)

- 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 (2 of 2)

- Matched Pairs Design
  - Compare two treatment groups by 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.

Sampling Errors (1 of 2)

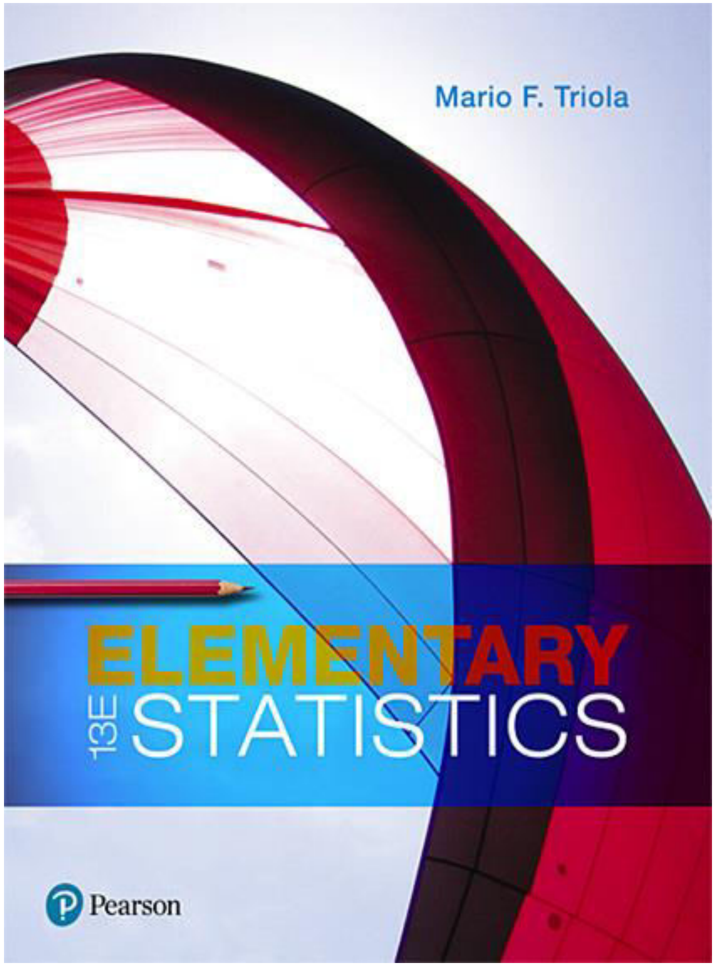
- No matter how well you plan and execute the sample collection process, there is likely to be some error in the results.
- Sampling error (or random sampling error)
  - occurs when the sample has been selected with a random method, but there is a discrepancy between a sample result and the true population result; such an error results from chance sample fluctuations.

Sampling Errors (2 of 2)

- Nonsampling error
  - Nonsampling is the result of human error, including such factors as wrong data entries, computing errors, questions with biased wording, false data provided by respondents, forming biased conclusions, or applying statistical methods that are not appropriate for the circumstances.
- Nonrandom sampling error
  - Nonrandom sampling error is the result of using a sampling method that is not random, such as using a convenience sample or a voluntary response sample.

Elementary Statistics

Thirteenth Edition



Chapter 1

Introduction to Statistics

Introduction to Statistics

1-1 Statistical and Critical Thinking

1-2 Types of Data

1-3 Collecting Sample Data

Key Concept

A major use of statistics is to collect and use sample data to make conclusions about populations.

Parameter

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

Statistic

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

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 (not numbers that represent counts or measurements).

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 data values are quantitative and the number of values is finite, or "countable."

Example: The number of tosses of a coin before getting tails

Continuous Data

- Continuous (numerical) data
  - result from infinitely many possible quantitative values, where the collection of values is not countable.

Example: The lengths of distances from 0 cm to 12 cm

Levels of Measurement

- Another way of classifying data is to use four levels of measurement: nominal, ordinal, interval, and ratio.

Nominal Level

- Nominal level of measurement
  - characterized by data that consist of names, labels, or categories only, and the data cannot be arranged in some order (such as low to high).

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

Ordinal Level

- Ordinal level of measurement
  - involves data that can be arranged in some order, but differences (obtained by subtraction) 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 differences between data values can be found and are meaningful. However, there is no **natural** zero starting point at which none of the quantity is present.

Example: Years 1000, 2000, 1776, and 1492

Ratio Level

- Ratio level of measurement
  - data can be arranged in order, differences can be found and are meaningful, and there is a natural zero starting point (where zero indicates that none of the quantity is present). Differences and ratios are both meaningful.

Example: Class times of 50 minutes and 100 minutes

Summary - Levels of Measurement

- **Nominal** - categories only
- **Ordinal** - categories with some order
- **Interval** - differences but no natural zero point
- **Ratio** - differences <u>and</u> a natural zero point

Big Data

- Big data
  - refers to data sets so large and so complex that their analysis is beyond the capabilities of traditional software tools. Analysis of big data may require software simultaneously running in parallel on many different computers.
- Data science
  - involves applications of statistics, computer science, and software engineering, along with some other relevant fields (such as sociology or finance).

Missing Data

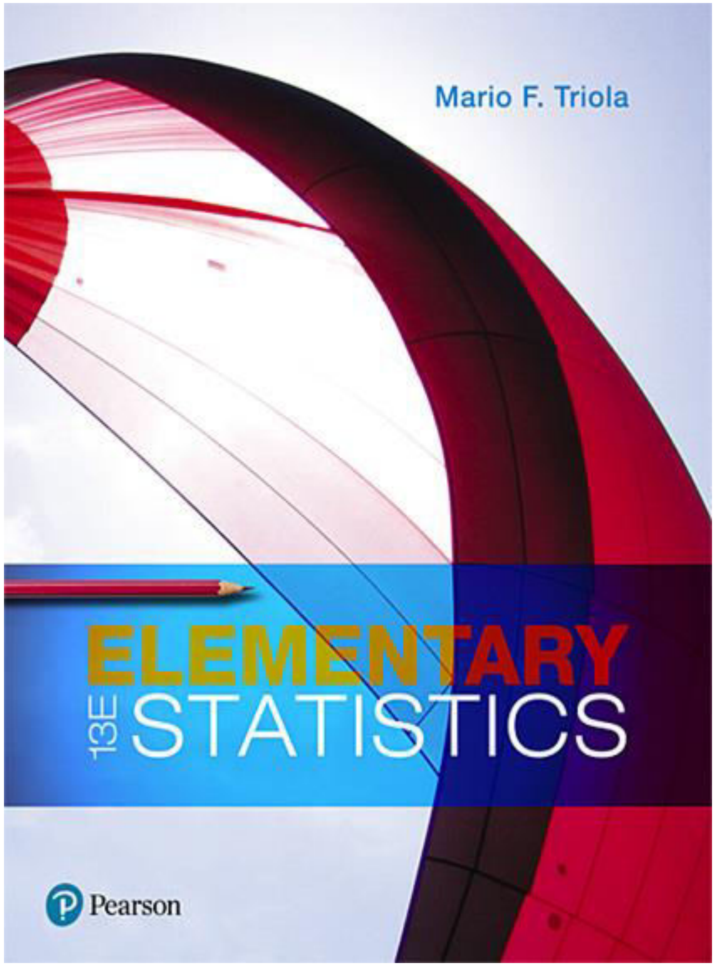
- A data value is **missing completely at random** if the likelihood of its being missing is independent of its value or any of the other values in the data set. That is, any data value is just as likely to be missing as any other data value.
- A data value is **missing not at random** if the missing value is related to the reason that it is missing.

Correcting for Missing Data

1. **Delete Cases:** One very common method for dealing with missing data is to delete all subjects having any missing values.
2. **Impute Missing Values:** We “impute” missing data values when we substitute values for them.

Elementary Statistics

Thirteenth Edition



Chapter 1

Introduction to Statistics

Introduction to Statistics

1-1 Statistical and Critical Thinking

1-2 Types of Data

1-3 Collecting Sample Data

Key Concept

The process involved in conducting a statistical study consists of “prepare, analyze, and conclude.”

Statistical thinking involves critical thinking and the ability to make sense of results. Statistical thinking demands so much more than the ability to execute complicated calculations.

Data

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

Statistics

- Statistics
  - The science of planning studies and experiments, obtaining data, and organizing, summarizing, presenting, analyzing, and interpreting those data and then drawing conclusions based on them.

Population

- Population
  - The complete collection of **all** measurements or data that are being considered. Typically, a population is the complete collection of data that we would like to make inferences about.

Census versus Sample

- Census
  - The collection of data from **every** member of a population
- Sample
  - A **subcollection** of members selected from a population

Example: Residential Carbon Monoxide Detectors (1 of 2)

In the journal article "Residential Carbon Monoxide Detector Failure Rates in the United States", it was stated that there are 38 million carbon monoxide detectors installed in the United States. When 30 of them were randomly selected and tested, it was found that 12 of them failed to provide an alarm in hazardous carbon monoxide conditions.

Example: Residential Carbon Monoxide Detectors (2 of 2)

In this case, the population and sample are as follows:

- **Population:** All 38 million carbon monoxide detectors in the United States
- **Sample:** The 30 carbon monoxide detectors that were selected and tested

The objective is to use the sample data as a basis for drawing a conclusion about the population of all carbon monoxide detectors, and methods of statistics are helpful in drawing such conclusions.

Statistical and Critical Thinking

Prepare

1. **Context**
   - What do the data represent?
   - What is the goal of study?
2. **Source of the Data**
   - Are the data from a source with a special interest so that there is pressure to obtain results that are favorable to the source?
3. **Sampling Method**
   - Were the data collected in a way that is unbiased, or were the data collected in a way that is biased (such as a procedure in which respondents volunteer to participate)?

↓

Analyze

1. **Graph the Data**
2. **Explore the Data**
   - Are there any outliers (numbers very far away from almost all of the other data)?
   - What important statistics summarize the data (such as the mean and standard deviation described in Chapter 3)?
   - How are the data distributed?
   - Are there missing data?
   - Did many selected subjects refuse to respond?
3. **Apply Statistical Methods**
   - Use technology to obtain results.

↓

Conclude

1. **Significance**
   - Do the results have statistical significance?
   - Do the results have practical significance?

Prepare (1 of 2)

Pleasure Boats and Manatee Fatalities from Boat Encounters

Pleasure Boats (ten of thousands)	99	99	97	95	90	90	87	90	90
Manatee Fatalities	92	73	90	97	83	88	81	73	68

- **Context**
  - The table includes the number of registered pleasure boats in Florida (tens of thousands) and the number of manatee fatalities from encounters with boats in Florida for each of several recent years.
  - The format of the table suggests the following goal: Determine whether there is a **relationship** between numbers of boats and numbers of manatee deaths from boats.

Prepare (2 of 2)

Pleasure Boats and Manatee Fatalities from Boat Encounters

Pleasure Boats (ten of thousands)	99	99	97	95	90	90	87	90	90
Manatee Fatalities	92	73	90	97	83	88	81	73	68

- **Source of the Data**
  - The data in the table are from the Florida Department of Highway Safety and Motor Vehicles and the Florida Marine Research Institute. The sources certainly appear to be reputable.
- **Sampling Method**
  - The data were obtained from official government records known to be reliable. The sampling method appears to be sound.

Voluntary Response Sample (1 of 2)

- Voluntary Response Sample
  - **Voluntary Response Sample** or **Self-Selected Sample** is one in which the respondents themselves decide whether to be included.

Voluntary Response Sample (2 of 2)

The following types of polls are common examples of voluntary response samples. By their very nature, all are seriously flawed because we should not make conclusions about a population on the basis of samples with a strong possibility of bias:

- Internet polls, in which people online can decide whether to respond
- Mail-in polls, in which people can decide whether to reply
- Telephone call-in polls, in which newspaper, radio, or television announcements ask that you voluntarily call a special number to register your opinion

Example: Voluntary Response Sample (1 of 2)

Nightline asked viewers to call with their opinion about whether the UN headquarters should remain in the United States. Viewers then decided themselves whether to call with their opinions, and 67% of 186,000 respondents said that the UN should be moved out of the United States.

In a separate, independent survey, 500 respondents were randomly selected and surveyed, and 38% of this group wanted the UN to move out of the United States.

Example: Voluntary Response Sample (2 of 2)

The two polls produced dramatically different results. Even though the **Nightline** poll involved 186,000 volunteer respondents, the much smaller poll of 500 randomly selected respondents is more likely to provide better results because of the far superior sampling method.

Analyze

After completing our preparation by considering the context, source, and sampling method, we begin to **analyze** the data.

- **Graph and Explore**
  - An analysis should begin with appropriate graphs and explorations of the data.
- **Apply Statistical Methods**
  - A good statistical analysis **does not** require strong computational skills. A good statistical analysis **does** require using common sense and paying careful attention to sound statistical methods.

Conclude (1 of 2)

The final step in our statistical process involves conclusions, and we should develop an ability to distinguish between statistical significance and practical significance.

- **Statistical Significance**
  - **Statistical significance** is achieved in a study if the likelihood of an event occurring by chance is 5% or less.
    - Getting 98 girls in 100 random births *is* statistically significant because such an extreme outcome is not likely to result from random chance.
    - Getting 52 girls in 100 births *is not* statistically significant because that event could easily occur with random chance.

Conclude (2 of 2)

- **Practical Significance**
  - It is possible that some treatment or finding is effective, but common sense might suggest that the treatment or finding does not make enough of a difference to justify its use or to be practical.

Example Statistical Significance Versus Practical Significance (1 of 2)

ProCare Industries once supplied a product named Gender Choice that supposedly increased the chance of a couple having a baby with the gender that they desired. In the absence of any evidence of its effectiveness, the product was banned by the Food and Drug Administration (FDA) as a "gross deception of the consumer."

- Suppose that the product was tested with 10,000 couples who wanted to have baby girls, and the results consist of 5200 baby girls born in the 10,000 births. This result is statistically significant because the likelihood of it happening due to chance is only 0.003%, so chance doesn't seem like a feasible explanation.

Example Statistical Significance Versus Practical Significance (2 of 2)

- That 52% rate of girls is statistically significant, but it lacks practical significance because 52% is only slightly above 50%. Couples would not want to spend the time and money to increase the likelihood of a girl from 50% to 52%. (**Note**: In reality, the likelihood of a baby being a girl is about 48.8%, not 50%.)

Analyzing Data: Potential Pitfalls (1 of 2)

- **Misleading Conclusions**
  - When forming a conclusion based on a statistical analysis, we should make statements that are clear even to those who have no understanding of statistics and its terminology.
- **Sample Data Reported Instead of Measured**
  - When collecting data from people, it is better to take measurements yourself instead of asking subjects to **report** results.
- **Loaded Questions**
  - If survey results are not worded carefully, the results of a study can be misleading.

Analyzing Data: Potential Pitfalls (2 of 2)

- **Order of Questions**
  - Sometimes survey questions are unintentionally loaded by the order of the items being considered.
- **Nonresponse**
  - A nonresponse occurs when someone either refuses to respond or is unavailable.
- **Percentages**
  - Some studies cite misleading percentages. Note that 100% of some quantity is **all** of it, but if there are references made to percentages that exceed 100%, such references are often not justified.