

Lecture Guide and Student Notebook  
to accompany  
Introductory Statistics:  
A Problem Solving Approach

Stephen M. Kokoska

September 23, 2006



# Contents

<b>1</b>	<b>An Introduction to Statistics and Statistical Inference</b>	<b>1</b>
1.1	Statistics Today . . . . .	1
1.2	Populations, Samples, Probability, and Statistics . . . . .	3
1.3	Experiments and Random Samples . . . . .	9
<b>2</b>	<b>Tables and Graphs for Summarizing Data</b>	<b>13</b>
2.1	Types of Data . . . . .	13
2.2	Bar Charts and Pie Charts . . . . .	18
2.3	Stem-and-Leaf Plots . . . . .	25
2.4	Frequency Distributions and Histograms . . . . .	31
<b>3</b>	<b>Numerical Summary Measures</b>	<b>43</b>
3.1	Measures of Central Tendency . . . . .	43
3.2	Measures of Variability . . . . .	52
3.3	The Empirical Rule and Measures of Relative Standing . . . . .	60
3.4	Five-Number Summary and Box Plots . . . . .	70
<b>4</b>	<b>Probability</b>	<b>75</b>
4.0	Introduction . . . . .	75
4.1	Experiments, Sample Spaces, and Events . . . . .	75
4.2	An Introduction to Probability . . . . .	90
4.3	Counting Techniques . . . . .	104
4.4	Conditional Probability . . . . .	111
4.5	Independence . . . . .	118
<b>5</b>	<b>Random Variables and Discrete Probability Distributions</b>	<b>127</b>
5.0	Introduction . . . . .	127
5.1	Random Variables . . . . .	127

5.2	Probability Distributions for Discrete Random Variables . . . . .	132
5.3	Mean, Variance, and Standard Deviation for a Discrete Random Variable . . .	139
5.4	The Binomial Distribution . . . . .	147
5.5	Other Discrete Distributions . . . . .	158
<b>6</b>	<b>Continuous Probability Distributions</b>	<b>167</b>
6.1	Introduction . . . . .	167
6.2	Probability Distributions for a Continuous Random Variable . . . . .	167
6.3	The Normal Distribution . . . . .	178
6.4	Checking the Normality Assumption . . . . .	191
6.5	The Exponential Distribution . . . . .	195
<b>7</b>	<b>Sampling Distributions</b>	<b>199</b>
7.0	Introduction . . . . .	199
7.1	Statistics, Parameters, and Sampling Distributions . . . . .	199
7.2	The Sampling Distribution of the Sample Mean . . . . .	206
7.3	The Distribution of the Sample Proportion . . . . .	218
<b>8</b>	<b>Confidence Intervals Based on a Single Sample</b>	<b>223</b>
8.0	Introduction . . . . .	223
8.1	Point Estimation . . . . .	223
8.2	A Confidence Interval for a Population Mean when $\sigma$ is Known . . . . .	229
8.3	A Confidence Interval for a Population Mean when $\sigma$ is Unknown . . . . .	240
8.4	A Large-Sample Confidence Interval for a Population Proportion . . . . .	247
8.5	A Confidence Interval for a Population Variance . . . . .	251
<b>9</b>	<b>Hypothesis Tests Based on a Single Sample</b>	<b>257</b>
9.0	Introduction . . . . .	257
9.1	The Parts of a Hypothesis Test and Choosing the Alternative Hypothesis . . .	258
9.2	Hypothesis Test Errors . . . . .	263
9.3	Hypothesis Tests Concerning a Population Mean When $\sigma$ is Known . . . . .	268
9.4	$p$ Values . . . . .	275
9.5	Hypothesis Tests Concerning a Population Mean when $\sigma$ is Unknown . . . . .	280
9.6	Large-Sample Hypothesis Tests Concerning a Population Proportion . . . . .	285
9.7	Hypothesis Tests Concerning a Population Variance or Standard Deviation . .	287
<b>10</b>	<b>Confidence Intervals and Hypothesis Tests Based on Two Samples or Treatments</b>	<b>291</b>
10.0	Introduction and Notation . . . . .	291
10.1	Comparing Two Population Means, Independent Samples, Population Variances Known . . . . .	293
10.2	Comparing Two Population Means Using Independent Samples from Normal Populations . . . . .	300
10.3	Paired Data . . . . .	308

10.4	Comparing Two Population Proportions Using Large Samples . . . . .	313
10.5	Comparing Two Population Variances or Standard Deviations . . . . .	320
<b>11</b>	<b>The Analysis of Variance</b>	<b>327</b>
11.0	Introduction . . . . .	327
11.1	One-Way ANOVA . . . . .	327
11.2	Isolating Differences . . . . .	339
11.3	Two-Way ANOVA . . . . .	352
<b>12</b>	<b>Correlation and Simple Linear Regression</b>	<b>365</b>
12.0	Introduction . . . . .	365
12.1	Simple Linear Regression . . . . .	366
12.2	Hypothesis Tests and Correlation . . . . .	384
12.3	Inferences Concerning the Mean Value and an Observed Value of $Y$ for $x = x^*$	399
12.4	Regression Diagnostics . . . . .	406
12.5	Multiple Linear Regression . . . . .	412
<b>13</b>	<b>Categorical Data and Frequency Tables</b>	<b>431</b>
13.0	Introduction . . . . .	431
13.1	Univariate Categorical Data, Goodness-of-Fit Tests . . . . .	431
13.2	Bivariate Categorical Data, Tests for Homogeneity and Independence . . . . .	439
<b>14</b>	<b>Nonparametric Statistics</b>	<b>453</b>
14.0	Introduction . . . . .	453
14.1	The Sign Test . . . . .	454
14.2	The Signed-Rank Test . . . . .	460
14.3	The Rank-Sum Test . . . . .	468
14.4	The Kruskal–Wallis Test . . . . .	473
14.5	The Runs Test . . . . .	478
14.6	Spearman’s Rank Correlation . . . . .	482



# Preface

There is a diminishing interest in statistics and mathematics, and these courses often trigger strong negative emotions in students. Introductory statistics is a common requirement for many majors but is frequently thought of as a punishment required to obtain a certain undergraduate degree, rather than a valuable, lifelong tool. One reason for this prevailing sentiment is the method of presentation.

The teaching of statistics has not changed dramatically over the past 50 years. We have not embraced new teaching strategies nor effectively incorporated technology into our classrooms. I believe the presentation of statistics must change and should emphasize concepts and applications rather than memorization and techniques.

This supplement addresses the basic issue of stressing concepts and represents a simple, different, and direct solution for teaching statistics more effectively. Presently, many instructors introduce material in an unproductive manner focused on text and symbols. Theorems, definitions, and important examples are written at least four times:

1. Theorems, definitions, and examples are presented in detail in the text.
2. Instructors copy these items to their lecture notes in preparation for class presentations.
3. Instructors write these items on the board.
4. Students frantically copy these items from the board into their notebooks.

I believe this teaching style is a highly inefficient and an ineffective method for learning statistics. The amount of time wasted by rewriting results that are already in the text usually infringes upon the allotted class time. Consequently, a very limited amount of time remains for the discussion of the concepts and interpretation of related results. Students in an introductory statistics class cannot be active learners if we continue to teach in this manner.

This supplement eliminates much of the duplication of effort and, thus, allows instructors to spend more time on concepts and results, interesting applications, and diverse material.

It significantly changes the way we, as teachers, present material in the classroom and the manner in which students take class notes.

This *Lecture Guide and Student Notebook* to accompany *Introductory Statistics: A Problem Solving Approach* is just that, a notebook with prepared (outline style or paraphrased) lecture notes that correspond to the textbook. Theorems, definitions, graphs, and many examples are presented in the notebook. Other contents include notes and reasons as we might ordinarily write on the blackboard for students to copy and remarks we often fail to write on the board due to time constraints. There is appropriate space left on each page for students to add clarification on a topic, to add comments by the instructor, and to complete problems. I suggest using transparencies of the student notebook pages in classroom presentations. This allows more communication among the instructor and students, and more time for discussion of problems and problem solving.

A counter-argument to this method of presentation is that students no longer *learn* theorems and definitions by rote. However, that is precisely the pedagogy we as statisticians and mathematicians are trying to avoid. We do not want students to simply memorize a definition, theorem, or technique. We want our students to be able to think and reason, and solve problems. We would like to extend the understanding and appreciation of statistics by addressing applications and relationships of theorems and definitions, not just the theorems and definitions themselves.

Using the *Lecture Guide and Student Notebook*, material can be presented more effectively during a regular lecture period. The time spent with students in class is a valuable resource. Using this supplement, instructors no longer have to write excessive information on the chalkboard nor do students have to anxiously copy it into their notes (often incorrectly). This supplement is a simple and advantageous method for presenting statistics, and it allows instructors to encourage active learners since the class time is focused on interpretation, application of statistical concepts, and the use of technology.

The notes in the *Lecture Guide and Student Notebook* are written to the level of a student learning statistics, as if the student were taking notes during the class. The notes are presented in outline style and/or paraphrased from the textbook, along with specific teaching strategies. The treatment is not comprehensive, but covers the material necessary to present a particular topic. This permits instructors to use class time for the discussion of applications included in the supplement.

I have been using this teaching style successfully for several years. Students appreciate less writing and the chance to listen, concentrate, discuss, and comprehend the material during class. I have been able to incorporate more technology into my presentations and to increase interaction and communication. This supplement has made teaching more enjoyable and rewarding for me, and learning statistics more constructive and positive for students.



# CHAPTER 1

## An Introduction to Statistics and Statistical Inference

---

### 1.1 Statistics Today

Statistics data is everywhere: newspapers, magazines, the Internet, weather forecasts, medical studies, sports reports, factoids, survey results, food labels.

Statistics:

1. Science of collecting and interpreting data.
2. Make decisions, assess risk, draw a conclusion.

Statistics data are used by professionals in many different disciplines:

1. Insurance: actuaries collect and interpret data, calculate insurance premiums.
2. Finance: economists make policy decisions regarding budgets, spending, and interest rates.
3. Manufacturing for quality control.
4. Pharmacology to study the efficacy of new drugs, look for significant side effects.
5. Sports: batting averages, goals against average, touchdowns, etc.
6. Meteorology: 50% chance of rain?
7. Also: ecology, genetics, law, marketing, public health, safety.

**Theme: Statistical Inference**

1. Claim.
2. Experiment.
3. Likelihood.
4. Conclusion.

**Statistics in the news:**

1. **Statistical Inference:** In September 2004, the pharmaceuticals company Merck voluntarily recalled the drug Vioxx and stopped selling this anti-inflammatory medicine. A new long-term study suggested that Vioxx increased the risk of heart attacks and strokes.
2. **Summary Statistics:** The International Arid Lands Consortium reported that the average person uses 20–80 gallons of water each day in his/her home. Typical uses include drinking, cooking, bathing, washing clothes, washing dishes, watering lawns and gardens, maintaining swimming pools, and washing cars.
3. **Probability and Odds:** After monitoring seismic activity and the weakening of the lava dome, scientists predicted that there was a 70% chance of a small to medium eruption of Mount St. Helens within a month.
4. **Summary Statistics:** According to *USA Today*, researchers summarized more than 60 studies concerning caffeine withdrawal conducted over 170 years. Half of all adults who stop drinking coffee or caffeinated soda experience drug-withdrawal symptoms.
5. **Statistical Inference:** A recent study tested a new technique to clear clogged arteries using radioactive seeds. At the end of this long-term study, 8% of patients who had the new treatment developed blood clots. Less than 1% of the patients in the placebo group developed blood clots. Researchers concluded that the new treatment offers no long-term benefit.

No matter how you are employed, you will have to make decisions based on available data. Some questions you may have to consider:

1. Do we have enough data? How were the data obtained?  
If more data are necessary, how will they be obtained?
2. How are the data summarized?  
Are the graphical and/or numerical techniques appropriate?  
Does the summary accurately represent the data?
3. What is the appropriate statistical technique to analyze the data?  
Are the conclusions reasonable and reliable?

## 1.2 Populations, Samples, Probability, and Statistics

Two applications of statistics: descriptive statistics and inferential statistics.

### Definition

**Descriptive Statistics:** Graphical and numerical methods used to describe, organize, and summarize data.

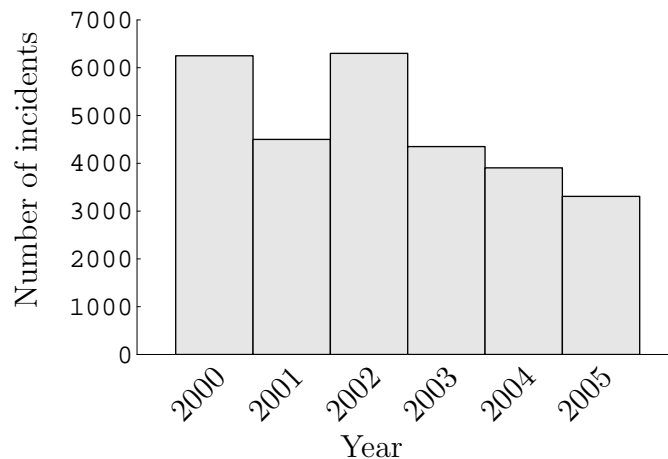
**Inferential Statistics:** Techniques and methods used to analyze a *small*, specific set of data in order to draw a conclusion about a large, more general collection of data.

**Example 1.2.1** In each of the following cases, determine whether the example involves descriptive or inferential statistics.

- (a) The following table shows the percentages of respondents that currently suffer from various health conditions. These data were reported by the National Health Interview Survey (NHIS) and Knowledge Networks (KN).

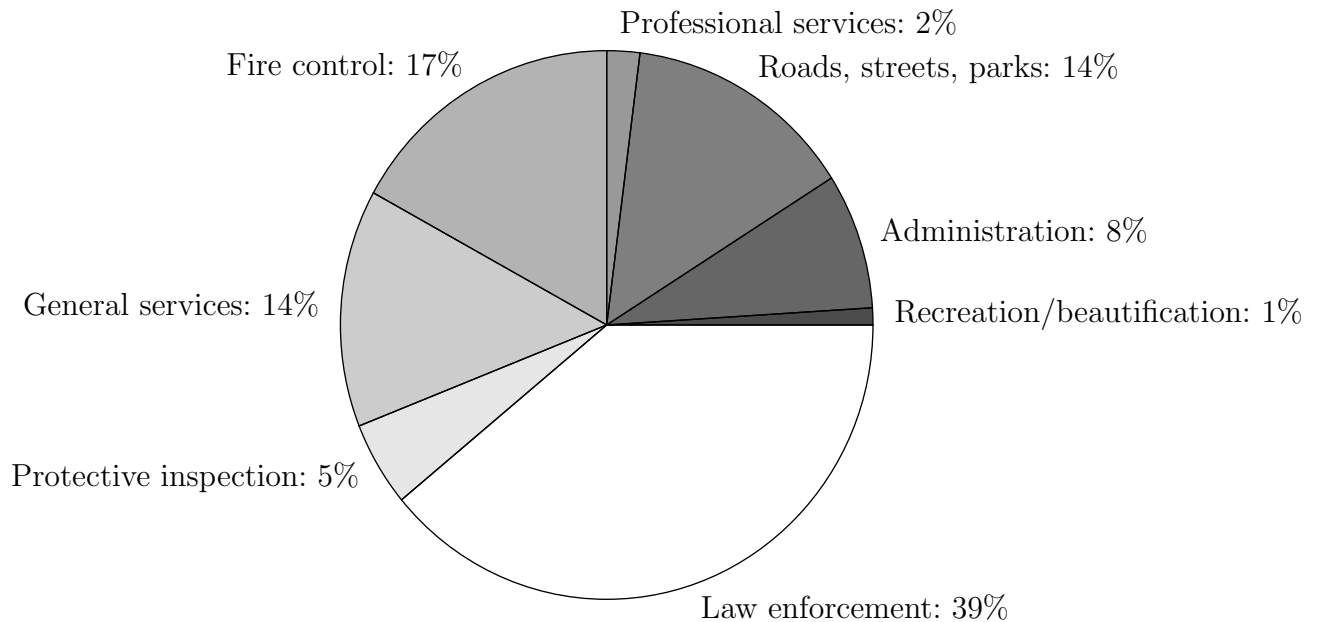
Condition	NHIS	KN
Smoking	23.3	24.7
Diabetes	6.7	7.1
Ulcer	7.3	7.1
Migraine	14.9	12.2
Stroke	2.2	1.8
Heart problems	10.1	10.1
Hypertension	22.6	16.9
Cancer	6.4	6.3

- (b) The following bar graph shows the number of police incidents in a city by year.



4 Chapter 1 An Introduction to Statistics and Statistical Inference

(c) The following pie chart shows the general fund expenditures for a small town in 2005.



(d) The World Health Organization conducted a study to determine whether SARS (severe acute respiratory syndrome) virus is spread through the air. It was determined that each patient studied infected on average three others. This number is consistent with a disease spread by direct contact with virus-laden droplets rather than with airborne particles. The World Health Organization concluded SARS is not spread through the air.

Every statistics problem involves a *population* and a *sample*.

**Definition**

A **population** is the entire collection of individuals or objects to be considered or studied.

A **sample** is a subset of the entire population, a small selection of individuals or objects taken from the entire collection.

A **variable** is a characteristic of an individual or object in a population of interest.

**Remarks**

1. A population consists of all individuals or objects of a particular type.  
Usually, there are infinitely many objects in a population, or at least so many that we cannot look at all of them.
2. A sample is (usually) a small part of a population.
3. Variable: *qualitative* (categorical) or a *quantitative* (numerical) attribute of each individual or object in a population.

**Example 1.2.2** Using a camera phone and a mobile network, a user can send pictures from his/her cell phone to a PC. Phone carriers are promoting camera phones, but there can be a high cost to download each picture. Several people who recently purchased cell phones were selected at random. Each person was classified as either having a camera phone or not having a camera phone. Describe the population, sample, and variable in this problem.

**Example 1.2.3** NCAA officials are interested in the attendance at Division II women's basketball games. Twenty-five games during the 2004–2005 season were selected at random, and the number of fans in attendance was recorded for each game. Describe the population, sample, and variable in this problem.

**Example 1.2.4** Some weather researchers believe that tornadoes are becoming more severe (determined by wind speed) in the area in the United States known as Tornado Alley. During the summer of 2005, the wind speed of five tornadoes selected at random was carefully measured. Describe the population, sample, and variable in this problem.

**Issues:**

1. Sample size.
2. Representative sample.

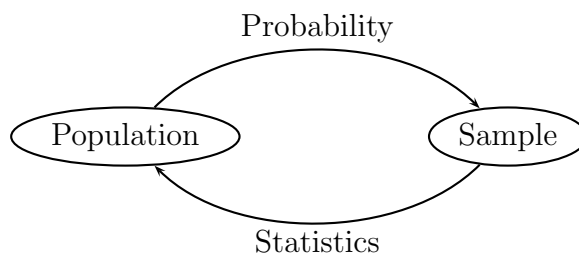
Probability and inferential statistics are both related to a population and a sample, but from different perspectives.

**Definition**

In order to solve a **probability** problem, certain characteristics of a population are assumed known. We then answer questions concerning a sample from that population.

In a **statistics** problem, we assume very little about a population. We use the information about a sample to answer questions concerning the population.

Relationships among probability, statistics, population, and sample:



**Example 1.2.5** As US Airways emerged from bankruptcy, management asked pilots whether they would accept a pay cut of approximately 18%. Consider the population consisting of the 3000 US Airways pilots and a random sample of 50 pilots from this population.

Probability question:

Suppose 60% of all pilots are in favor of accepting the pay cut.

What is the probability 35 or more (of the 50) are in favor of accepting the pay cut?

Statistics question:

Suppose 40 out of the 50 pilots in the sample are in favor of accepting the pay cut. What can we conclude about the percentage of all US Airways pilots in favor of accepting the pay cut?

**Example 1.2.6** A study was conducted to determine what one thing executives would like to change about their current job. A random sample of 30 executives was obtained, and each was asked to name one thing they would like to change.

Probability question:

Suppose 35% of all executives would like shorter workdays. What is the probability less than 10 of the 30 executives surveyed said they would like shorter workdays?

Statistics question:

Suppose 15 of the 30 executives surveyed said they would like shorter workdays. What can we conclude about the percentage of all executives who favor shorter workdays?

**Example 1.2.7** In each of the following problems, identify the population and the sample, and determine whether the question involves probability or statistics.

- (a) Thirty percent of all students ages 8–13 bring their lunch to school at least once a week. Ten students from a rural school district are selected at random. What is the probability at most two brought their lunch to school at least once in the past week?

**8** Chapter 1 An Introduction to Statistics and Statistical Inference

- (b) One hundred baby boomers were selected at random and asked to name their favorite TV show ever. Of those 100, 23 named *Laugh-In*. Estimate the true proportion of baby boomers whose favorite TV show is *Laugh-In*.
- (c) Fifty newspaper reporters were selected at random and asked to complete a short survey. Twenty of those selected were women. Is there any evidence to suggest that the proportion of newspaper reporters who are women is greater than 50%?
- (d) Recent studies suggest that as people get older they eat a more balanced diet. Suppose 70% of all people over 65 years old eat a balanced diet. Thirty people over 65 were selected at random, and each was asked to complete a questionnaire to determine whether they eat a balanced diet. Is it likely that less than 15 of the 30 eat a balanced diet?



---

## 1.3 Experiments and Random Samples

**Definition**

In an **observational study**, we observe the response for a specific variable for each individual or object.

In an **experimental study**, we investigate the effects of certain conditions on individuals or objects in the sample.

**Example 1.3.1** An observational study: Colic is a condition in newborn children characterized by prolonged, persistent, inconsolable crying. This illness begins shortly after birth and may last up to a year. The causes are unknown, but some physicians believe overfeeding, anxiety and tension in the household, and allergies to cow's milk may contribute to this ailment. A researcher decides to measure the amount of time a child has colic. A random sample of babies who had colic is obtained, and the parents are asked to report the total amount of time (in months) the child had this disorder. The data are summarized graphically and numerically.

It is important for data to be representative of the relevant population.

**Definition**

A **(simple) random sample** (SRS) of size  $n$  is a sample selected in such a way that every possible sample of size  $n$  has the same chance of being selected.

**Remarks**

1. A random sample is difficult to achieve.  
Random number tables, random number generators can be used.
2. If a sample is not random, it is *biased*.
3. Nonresponse bias: common in surveys, since most people discard the survey.
4. Self-selection bias: individuals choose to be included in the sample.
5. Infinite population: the number of possible simple random samples is infinite.  
Finite population: the number of possible simple random samples is given in Chapter 7.
6. Simple random sample: very important for a reliable inference.

**Example 1.3.2** The Human Resource (HR) director at a large company is considering assigned parking spaces in the company parking lot. She plans to select 50 employees from the 1000 that work at this location and ask each person to complete a short questionnaire. The results will be used to make a decision regarding assigned parking spaces.

The HR director would like a simple random sample of size 50. How can she select 50 people at random?

- (a) Write each person's name on a piece of paper. Select 50.
- (b) Random number table.
- (c) Random number generator.

**Example 1.3.3** An experimental study: A company markets an over-the-counter drug that is designed to fight fatigue. A researcher selects 50 adults who are suffering from fatigue and randomly assigns each person to a treatment group or a placebo group. Adults in the treatment group will take the recommended daily dose of the drug for one month, and adults in the placebo group will be given a placebo, or sugar pill, daily. Neither the subjects nor the researcher know which group is receiving the placebo and which is receiving the active drug. After one month, each subject will be asked whether they feel less fatigued. The proportion of adults feeling less fatigued in each group will be compared in order to assess the efficacy of the drug.

Confounding: several factors together contribute to an effect, but no single cause can be isolated.

### Statistical Inference Procedure

The process of checking a claim can be divided into four parts.

**Claim** This is a statement of what we assume to be true.

**Experiment** In order to check the claim, we conduct a relevant experiment.

**Likelihood** Consider the likelihood of occurrence of the observed experimental outcome, assuming the claim is true. We will use many techniques to determine whether the experimental outcome is a reasonable observation (subject to reasonable variability) or whether it is a rare occurrence.

**Conclusion** There are only two possible conclusions. (1) If the outcome is reasonable, then we cannot doubt the claim. We usually write, "There is no evidence to suggest that the claim is false." (2) If the outcome is rare, we disregard the lucky alternative, and question the claim. A rare outcome is a contradiction. It shouldn't happen (often) if the claim is true. In this case we write, "There is evidence to suggest that the claim is false."

**Example 1.3.4** A soft-drink manufacturer ships 2000 bottles of soda and claims that 1998 are sealed properly and only 2 seals are defective. When a market receives this shipment, two bottles are selected at random, and both seals are defective.

*Claim:* There are 1998 properly sealed bottles and 2 defectives.

*Experiment:* Two bottles are selected and found to be defective.

*Likelihood:* One of two things has happened.

(a) Incredibly lucky.

It is possible to select the two defective bottles, but this is very unlikely.

(b) The claim is false.

Chance of selecting the two defective bottles is very small.

More likely: the claim is false.

*Conclusion:* Discount the lucky alternative.

Selecting the two defective bottles is rare.

There is evidence to suggest that the manufacturer's claim is false.

**Example 1.3.5** The makers of Lexus automobiles claim that 80% of all owners rate the overall value of their new car as *very high* with respect to affordability, expected reliability, and resale value. An automotive consulting firm selected 100 new Lexus owners at random and asked each to rate the overall value of their new car. Seventy-eight rated the overall value as very high. Identify the claim, experiment, and likelihood, and draw a conclusion in this example.



## CHAPTER 2

# Tables and Graphs for Summarizing Data

---

## 2.1 Types of Data

Tables, charts, graphs: used to organize and summarize data.

1. Shape: skewed or symmetric.
2. Center: where is the majority of the data located?
3. Variability: spread, or dispersion, of the data; compact or spread out.

The type of table, chart, or graph, or statistical analysis depends on the type of data.

### Definition

A data set consisting of observations on only a single characteristic, or attribute, is a **univariate** data set.

If we measure, or record, two observations on each individual, the data set is **bivariate**.

If there are more than two observations on each person, the data set is **multivariate**.

Two types of univariate data:

### Example 2.1.1

- (a) What brand of graphing calculator do you use?

The response is categorical.

- (b) How much space is left on your computer's hard drive?

The response is numerical.

**Definition**

A **categorical**, or **qualitative**, univariate data set consists of non-numerical observations that may be placed in categories.

A **numerical**, or **quantitative**, univariate data set consists of observations that are numbers.

**Example 2.1.2** A set of observations is obtained as indicated below. In each case, classify the resulting data set as categorical or numerical.

- (a) The number of shares of stock purchased or sold in several transactions.
- (b) The number of rolls of film developed in one hour at several Wal-mart stores.
- (c) The type of wall covering in several hotel rooms.
- (d) The business type for several companies in a large city.
- (e) The Library of Congress classification of several books borrowed from a public library.
- (f) The actual weight of a scented candle in a jar.
- (g) The active chemical in various ice-melting products.

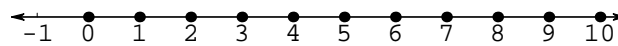
Two types of numerical data:

**Example 2.1.3** An experiment consists of recording the number of home runs by one team in a single major-league baseball game.

The possible values are  $0, 1, 2, \dots, 10$ . (The record is 10, by Toronto on 9/14/87.)

There are only a finite number of possible numerical values.

The values are discrete, isolated points on a number line.



**Example 2.1.4** An experiment consists of measuring the dissolved oxygen content in a Gulf of Mexico *dead zone*.

The dissolved oxygen is measured in parts per million (ppm) and can be any value in the continuous interval from 0 to 6 (including the endpoints).



### Definition

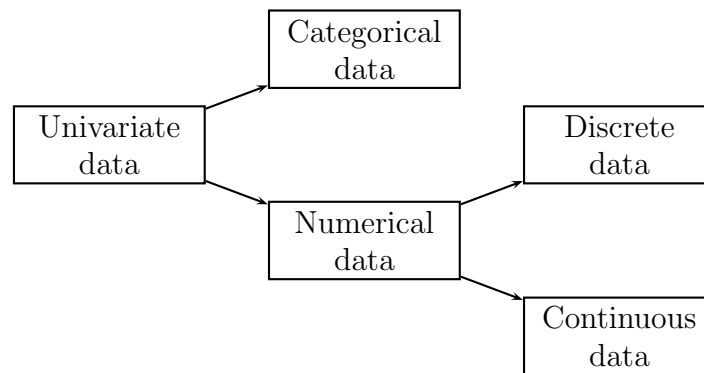
A numerical data set is **discrete** if the set of all possible values is finite, or countably infinite. Discrete data sets are usually associated with *counting*.

A numerical data set is **continuous** if the set of all possible values is an interval of numbers. Continuous data sets are usually associated with *measuring*.

### Remarks

1. Must consider *all possible* values.  
Discrete: finite or countably infinite.  
Continuous: an interval of possible values.
2. Countably infinite: infinitely many possible values, but they are countable.
3. Interval for continuous data: any interval, any length, open or closed.
4. Continuous data: any number in an interval (in theory, not in practice).

5. Classifications of univariate data:



**Example 2.1.5** A set of observations is obtained as indicated below. In each case, classify the resulting data set as categorical or numerical. If the data set is numerical, determine whether it is discrete or continuous.

(a) The number of Americans who win the Nobel Prize in Physics in specified years.

(b) The number of scheduled meetings a university president has on certain days.

(c) The first piece moved in several chess matches.

(d) The amount of omega-3 fatty acids in various cans of albacore tuna.



(e) The total number of ushers at selected weddings in June.

(f) The type of recipe featured on selected Food Channel shows.

(g) The total overtime (in minutes) worked per week by selected state troopers.

(h) The weight (in grams) of several Macintosh apples.

(i) The temperature (in °F) at several locations in the Carlsbad Caverns.

## 2.2 Bar Charts and Pie Charts

Natural summary measures for categorical data: frequency and relative frequency.

The following table summarizes the number of open contracts for certain government agencies during a month.

Class	Frequency	Relative frequency
Navy	113,424	0.3508
Army	43,308	0.1339
Air Force	45,621	0.1411
Defense Logistics Agency	112,536	0.3480
Other	8,483	0.0262
Total	323,372	1.0000

### Definition

A **frequency distribution** for categorical data is a summary table that presents categories, counts, and proportions.

1. Each unique value in a categorical data set is a label, or **class**.
2. The **frequency** is the count for each class.
3. The **relative frequency**, or sample proportion, for each class is the frequency of the class divided by the total number of observations.

**Example 2.2.1** In the frequency distribution above:

- (a) The classes: Navy, Army, Air Force, DLA, Other.
- (b) The frequency for Army contracts is 43,308.
- (c) The relative frequency for Navy contracts is  $\frac{113,424}{323,372} = 0.3508$ .

**Example 2.2.2** A random sample of people who went to a Great Lakes beach was obtained, and the destination of each is given in the following table.

Lake Shore Park	Euclid Beach	Euclid Beach	Bay Beach	Ontario Beach
Ontario Beach	Gladstone Beach	Ontario Beach	Bay Beach	Bay Beach
Ontario Beach	Gladstone Beach	Ontario Beach	Euclid Beach	Euclid Beach
Lake Shore Park	Bay Beach	Bay Beach	Gladstone Beach	Ontario Beach
Euclid Beach	Lake Shore Park	Lake Shore Park	Bay Beach	Ontario Beach

- Construct a frequency distribution to describe this data.
- What proportion of people went to Lake Shore Park or Euclid Beach?
- What proportion of people did not go to Bay Beach?

Class	Tally	Frequency	Relative frequency
Lake Shore Park			
Euclid Beach			
Bay Beach			
Ontario Beach			
Gladstone Beach			
Total			

**Example 2.2.3** A random sample of customers purchasing bread in a local supermarket was obtained. The brand purchased by each person is given in the following table using the abbreviations Wonder (W), Holsum (O), Nature’s Own (N), Pepperidge Farm (P), Home Pride (H), and Sunbeam (S).

W	W	O	H	H	S	S	W	P	P
N	N	W	P	H	S	S	N	N	W
W	N	S	P	P	S	S	W	W	N

- Construct a frequency distribution to describe this data.
- Which brand is most preferred? Justify your answer.
- What proportion of customers preferred Nature’s Own or Pepperidge Farm bread?

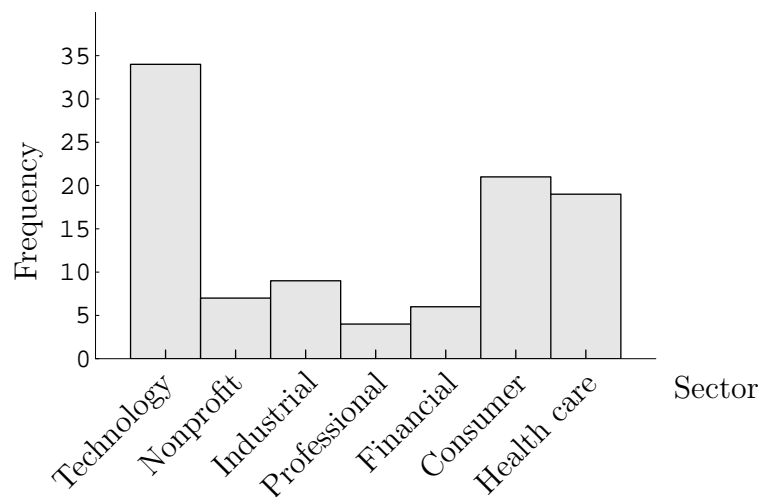
Class	Tally	Frequency	Relative frequency
Wonder			
Holsum			
Nature’s Own			
Pepperidge Farm			
Home Pride			
Sunbeam			
Total			

**Bar chart:** Graphical representation of a frequency distribution for categorical data.

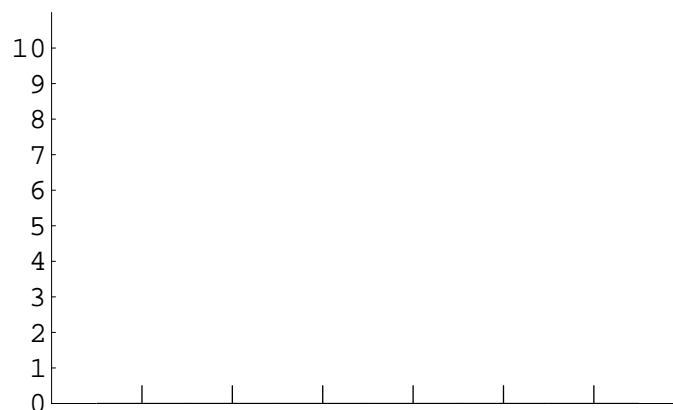
### How to Construct a Bar Chart

1. Draw a horizontal axis with equally spaced tick marks, one for each class.
2. Draw a vertical axis for the frequency (or relative frequency) and use appropriate tick marks. Label each axis.
3. Draw a rectangle centered at each tick mark (class) with height equal to, or proportional to, the frequency of each class (also called the class frequency). The bars should be of equal width, but do not necessarily have to abut one another; there can be spaces between them.

**Example 2.2.4** Here is a bar chart showing the number of businesses by sector in a city.



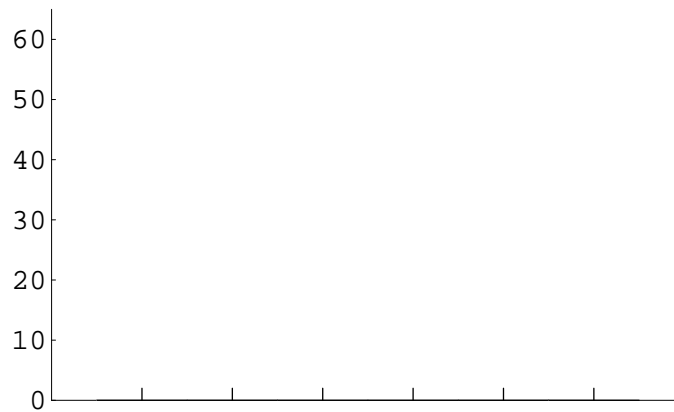
**Example 2.2.5** Construct a bar chart for the bread data in Example 2.2.3.



**Example 2.2.6** A random sample of sites in the Cannon River was selected, and the number and type of mussels found at all sites is summarized in the following table.

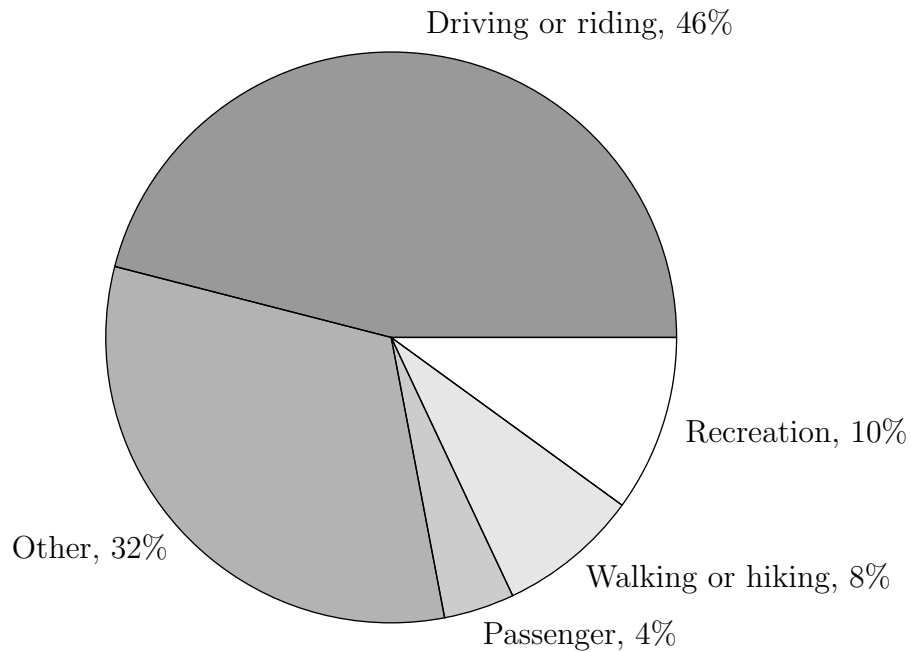
Class	Frequency	Relative frequency
L. siliquoidea	61	
P. alatus	37	
P. grandis	20	
L. complanata	7	
E. dilata	12	
L. recta	2	
Total		

Complete the frequency distribution, and construct a bar chart for this data.



**Pie chart:** another graphical representation of a frequency distribution for categorical data.

**Example 2.2.7** The following pie chart shows the activities corresponding to serious injury reports from a hospital emergency room.

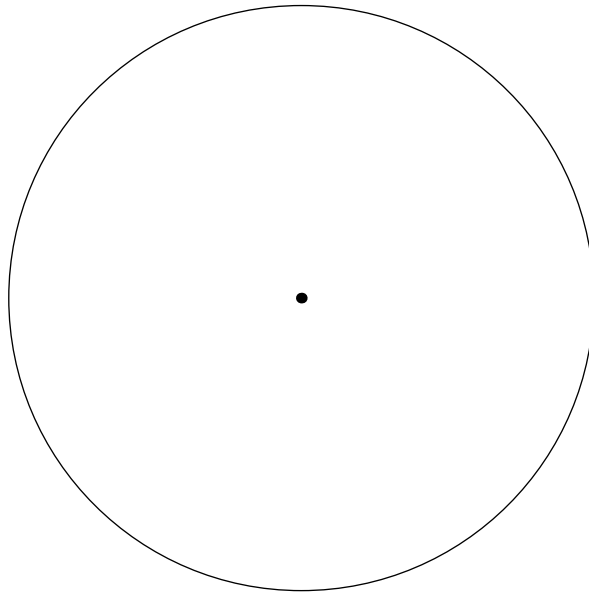


### How to Construct a Pie Chart

1. Divide a circle (or pie) into slices, or wedges, so that each slice corresponds to a class.
2. The size of each slice is measured by the angle of the slice. To compute the angle of each slice, multiply the relative frequency by  $360^\circ$  (the number of degrees in a whole, or complete, circle).
3. The first slice of a pie chart is usually drawn with an edge horizontal and to the right ( $0^\circ$ ). The angle is measured counterclockwise. Each successive slice is added counterclockwise with the appropriate angle.

**Example 2.2.8** A local bakery obtained a random sample of customers who purchased pies. Complete the following table, and construct a pie chart for this data.

Class	Frequency	Relative frequency	Angle
Apple	25		
Blueberry	5		
Cherry	16		
Pecan	27		
Pumpkin	15		
Rhubarb	9		
Total			





---

## 2.3 Stem-and-Leaf Plots

Graphical technique for describing numerical data.

### Stem-and-Leaf Plot

1. Relatively new graphical procedure. Combination of sorting and graphing.
2. Actual data is used to create the graph.
3. Can be used to describe shape, center, and variability.
  - (a) Center: a *typical* value.  
Arrange observations in increasing order.  
Approximate a middle value or range of values.
  - (b) Variability: spread or compactness.
  - (c) Look for outliers: observations far away from the rest.

#### How to Create a Stem-and-Leaf Plot

To create a stem-and-leaf plot, each observation in the data set must have at least two digits. Think of each observation as consisting of two pieces (a stem and a leaf).

1. Split each observation into a  
**Stem:** one or more of the leading, or left-hand, digits, and a  
**Leaf:** the trailing, or remaining, digit(s) to the right.  
Each observation in the data set must be split at the same place, for example between the tens place and the ones place.
2. Write a sequence of stems in a column, from the smallest occurring stem to the largest. Include all stems between the smallest and largest, even if there are no corresponding leaves.
3. List all the digits of each leaf next to its corresponding stem. It is not necessary to put the leaves in increasing order, but make sure the leaves line up vertically.
4. Indicate the units for the stems and leaves.

**Example 2.3.1** A random sample of city postal delivery routes was obtained. The number of pieces of mail ready for delivery on each route was recorded and is given in the following table.

---

669	713	715	668	684	642	701	694	715	691
682	690	705	717	695	711	692	679	722	738
697	692	702	715	702	708	664	728	703	710

---

Construct a stem-and-leaf plot for this data. What is a typical value? Are there any outliers? Justify your answer.

**Remarks**

1. Rule of thumb: try to construct a stem-and-leaf plot with 5–20 stems.  
Fewer than 5: too compact. More than 20: too spread out.
2. **Ordered** stem-and-leaf plot: leaves in increasing order.
3. Advantages: each observation is a visible part of the graph; data are sorted.  
Disadvantage: a stem-and-leaf plot can get very big, very fast.

If a data set is very large, the stems may be divided, usually in half or into fifths.

**Example 2.3.2** Consider a study in which a company that rents DVDs on-line obtained a random sample of customers and recorded the number of movies rented by each during the past year. The data are given in the following table.

33	49	41	31	31	35	27	33	48	22
14	20	38	31	49	22	24	44	33	44
38	30	18	42	21	35	29	30	50	13
36	29	33	25	27	45	50	40	34	38

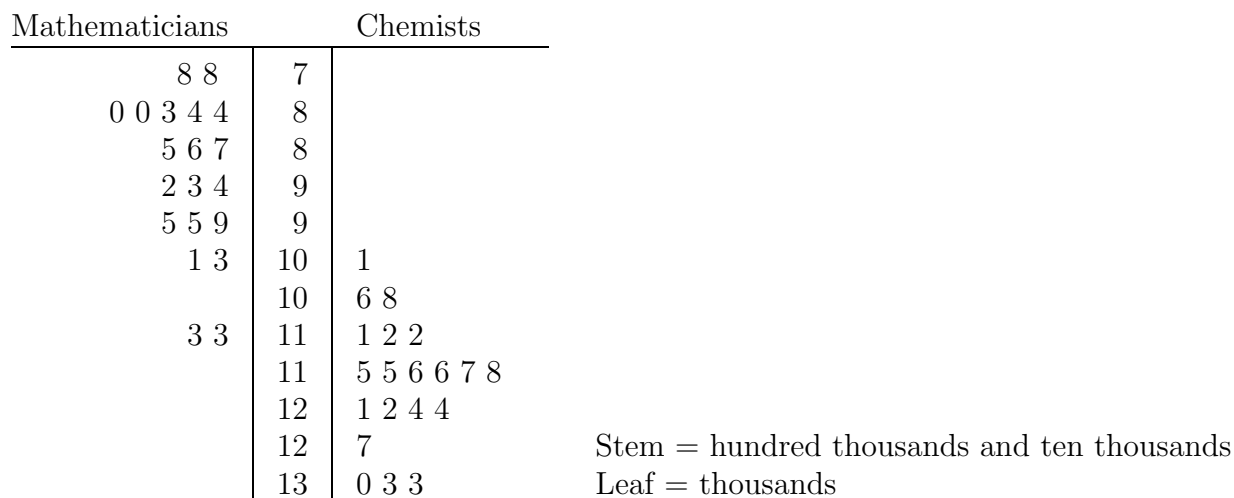
Construct a stem-and-leaf plot for this data. What is a typical value? Describe the shape of the distribution.

Two data sets can be compared graphically using a back-to-back stem-and-leaf plot.

**Example 2.3.3** A random sample of National Science Foundation grants made to chemists and mathematicians was obtained. The total amount (in thousands of dollars) awarded to each principal investigator is given in the following table.

Chemists					Mathematicians				
108	112	111	115	127	83	113	80	86	103
122	133	124	133	124	95	94	92	93	78
101	106	112	115	116	101	99	87	84	85
121	130	118	116	117	84	95	80	113	78

Here is a back-to-back stem-and-leaf plot for this data. Stems have been split in half.



Compare the two distributions using this stem-and-leaf plot.

If there are two or more digits in each leaf, the trailing digits may be truncated or rounded.

**Example 2.3.4** A random sample of the hourly pay (in dollars) for police officers with 10 years of experience was obtained. Consider the three observations 25.16, 25.62, and 25.74. Suppose each observation is split between the ones place and the tenths place.

- (a) Find the two-digit leaf for each observation.
- (b) Find the one-digit leaf for each observation if each leaf is truncated.
- (c) Find the one-digit leaf for each observation if each leaf is rounded.

**Example 2.3.5** A random sample of four-wheeled gas-powered rotary mowers was obtained, and the cutting width of each was measured (in cm). The data are given in the following table.

---

43.8	44.6	44.8	45.7	48.9	42.5	44.5	42.6	44.3	45.6
45.9	45.7	42.4	40.3	45.8	44.4	44.7	42.0	44.0	45.4
43.5	46.5	47.6	46.1	45.2	48.8	46.0	44.6	50.1	48.5

---

Construct a stem-and-leaf plot for this data. Describe the distribution.

---

## 2.4 Frequency Distributions and Histograms

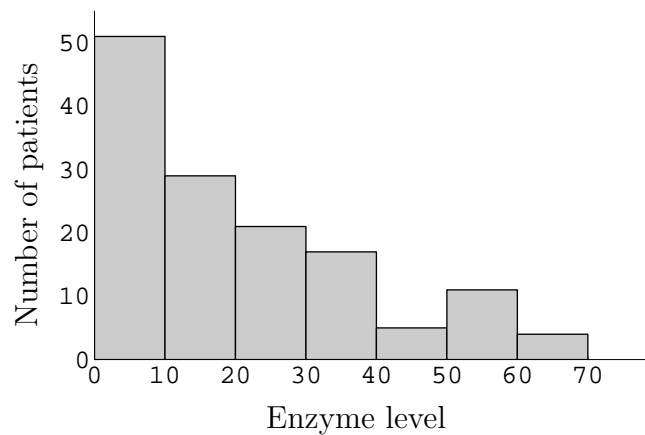
1. Numerical data set: no natural categories.

Solution: use intervals as categories, or classes.

2. Construct a frequency distribution for continuous data using intervals as classes.

Construct a graphical representation of the frequency distribution: histogram.

**Example 2.4.1** A random sample of patients with a certain thyroid disease was obtained, and a sample of blood was obtained from each person. The amount of a specific enzyme in each blood sample was measured in IU/mL. The following graph is a histogram for the data.



### Definition

A **frequency distribution** for numerical data is a summary table that displays classes, frequencies, relative frequencies, and cumulative relative frequencies.

**How to Construct a Frequency Distribution for Numerical Data**

1. Choose a range of values that captures all of the data. Divide it into non-overlapping (usually equal) intervals. Each interval is called a **class**, or **class interval**. The endpoints of each class are the *class boundaries*.
2. We use the left-endpoint convention; an observation equal to an endpoint is allocated to the class with that value as its lower endpoint. Hence, the lower class boundary is always included in the interval, and the upper class boundary is never included. This ensures that each observation falls into exactly one interval.
3. As a rule of thumb, there should be 5–20 intervals. Use *friendly* numbers, for example, 10–20, 20–30, etc., not 15.376–18.457, 18.457–21.538, etc.
4. Count the number of observations in each class interval. This count is the **class frequency** or simply the **frequency**.
5. Compute the proportion of observations in each class. This ratio, the class frequency divided by the total number of observations, is the **relative frequency**.
6. Find the **cumulative relative frequency** for each class: the sum of all the relative frequencies of classes up to and including that class. This column is a *running total* or *accumulation* of relative frequency, by row.



**Example 2.4.2** According to an article in *Food Review*, 362 billion Oreos have been sold since 1912. In 1987, Nabisco introduced the Double Stuff Oreo, with double the amount of cream in the middle. A random sample of Double Stuff Oreos was obtained, and the amount of cream (in grams) was measured for each. The data are given in the following table.

5.1	5.1	4.9	5.5	5.2	5.5	5.1	4.8	5.2	5.7
5.4	4.5	5.6	5.4	5.6	5.4	5.4	5.0	4.8	5.2
5.7	5.0	5.1	5.1	5.3	4.6	5.1	5.1	5.1	5.7

Construct a frequency distribution for this data using the class intervals 4.4–4.6, 4.6–4.8, . . . , 5.6–5.8.

Class	Frequency	Relative frequency	Cumulative relative frequency
4.4–4.6			
4.6–4.8			
4.8–5.0			
5.0–5.2			
5.2–5.4			
5.4–5.6			
5.6–5.8			
Total			

**Remarks**

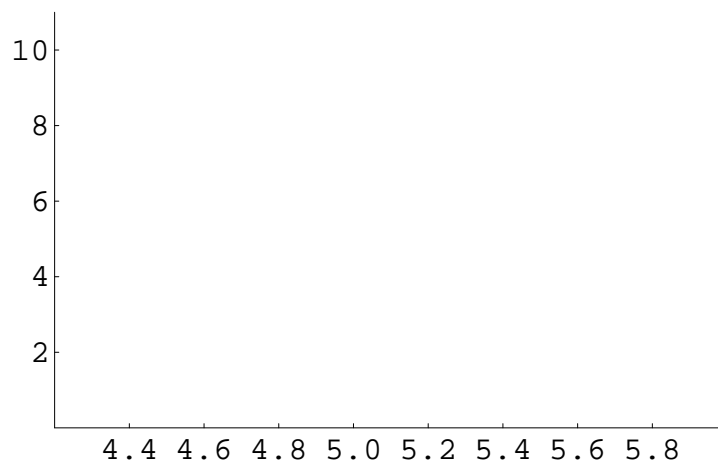
1. To find relative frequency from cumulative relative frequency (CRF):  
Take the class CRF and subtract the previous class CRF.
2. If data are discrete: Use the same procedure.  
Each value may be a class.

**Histogram:** a graphical representation of a frequency distribution, a plot of frequency versus class interval.

**How to Construct a Histogram**

1. Draw a horizontal (measurement) axis and place tick marks corresponding to the class boundaries.
2. Draw a vertical axis and place tick marks corresponding to frequency. Label each axis.
3. Draw a rectangle above each class with height equal to frequency.

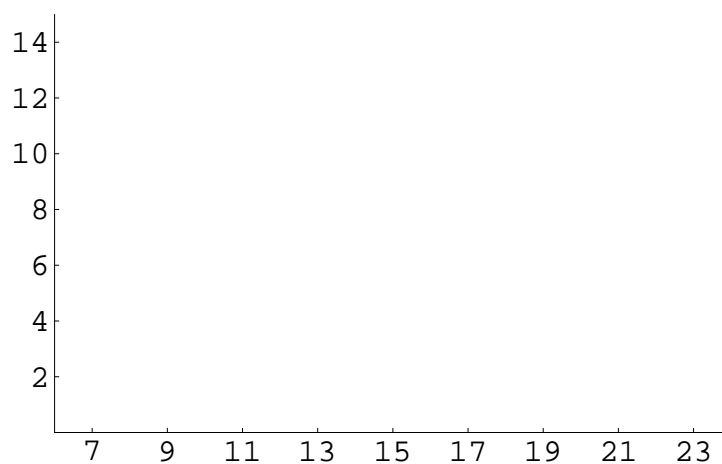
**Example 2.4.3** Construct a frequency histogram for the Oreo cookie data presented in Example 2.4.2.



**Example 2.4.4** A small town designates one day in the fall for leaf pickup. Residents must use special bags, and a town truck will collect the leaves. A random sample of leaf bags was obtained, and the weight (in pounds) of each was recorded. A partial frequency distribution for this data is given below.

Class	Frequency	Relative frequency	Cumulative relative frequency
7-9	2		
9-11	4		
11-13	5		
13-15	13		
15-17	11		
17-19	10		
19-21	3		
21-23	2		
Total			

Complete the frequency distribution, and construct a frequency histogram for this data.



**Remarks**

1. Histogram: conveys shape, center, and variability of the distribution.  
Can also identify outliers.
2. To construct a histogram by hand: need the frequency distribution.  
With software: no frequency distribution needed, only the data.
3. Relative frequency histogram: plot relative frequency versus class interval.  
Only difference between a frequency histogram and a relative frequency histogram: scale on the vertical axis.
4. Histograms should not be used for inference.  
Quick look at the distribution, suggest certain characteristics.

**Unequal Class Widths**

1. Neither frequency nor relative frequency should be used on the vertical axis.
2. Set area of each rectangle equal to relative frequency.
3. Height of each rectangle: density.

**How to Find the Density**

To find the density for each class:

1. Set the *area* of each rectangle equal to relative frequency.

The *area* of each rectangle is *height* times class *width*.

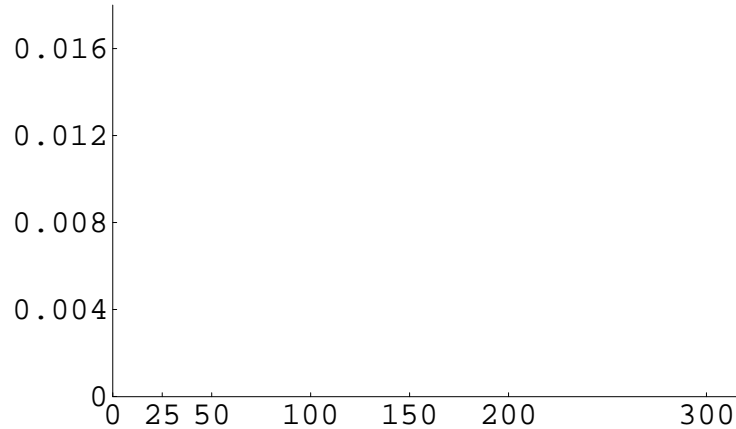
$$\begin{aligned} \text{Area of rectangle} &= \text{Relative frequency} \\ &= (\text{Height}) \times (\text{Class width}) \end{aligned}$$

2. Solve for the height.

$$\text{Density} = \text{Height} = (\text{Relative frequency}) / (\text{Class width})$$

**Example 2.4.5** A random sample of lakes in Minnesota was obtained, and the alkalinity (in mg/L) was measured for each. The data are partially summarized in the table below. Complete the table, and construct a density histogram for this data.

Class	Frequency	Relative frequency	Width	Density
0–25	100			
25–50	62			
50–100	52			
100–150	22			
150–200	8			
200–300	6			
Total				



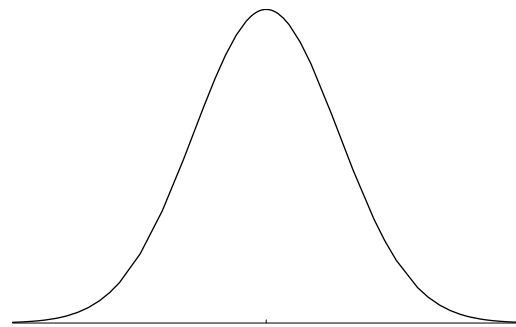
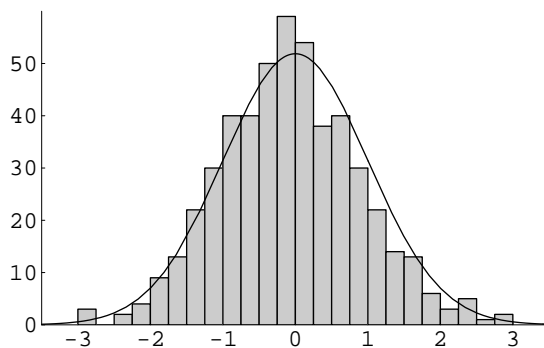
### Shape of a Distribution

1. Density histogram: relative frequency equal to the area of each rectangle.

Therefore, total area is 1.

2. Smoothed histogram: smooth curve along the tops of rectangles, captures the general nature of the distribution.

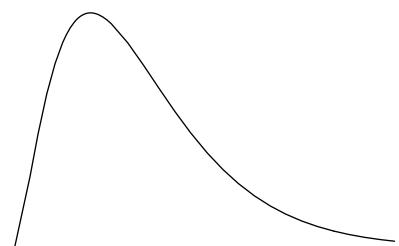
Use to help identify and describe distributions quickly.



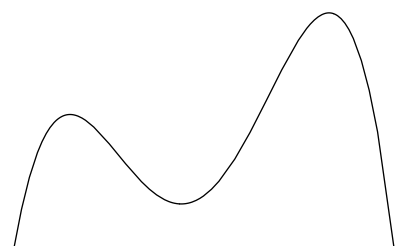
A distribution can be characterized by the number of peaks:

#### Definition

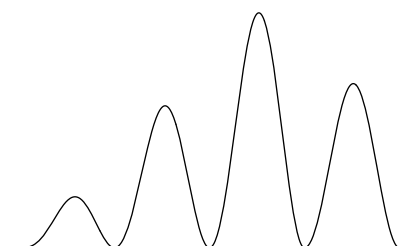
1. A **unimodal** distribution has one peak. This is very common; almost all distributions have a single peak.
2. A **bimodal** distribution has two peaks. This shape is not very common and may occur if data from two different populations are accidentally mixed.
3. A **multimodal** distribution has more than one peak. A distribution with more than two distinct peaks is very rare.



Unimodal distribution.



Bimodal distribution.

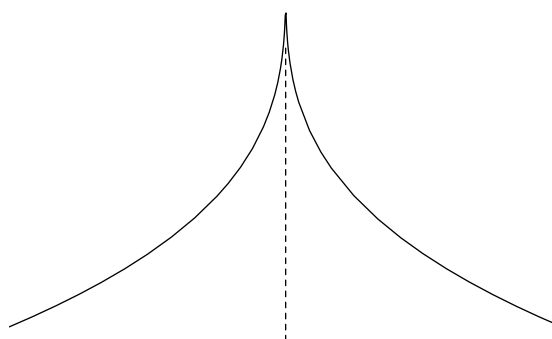
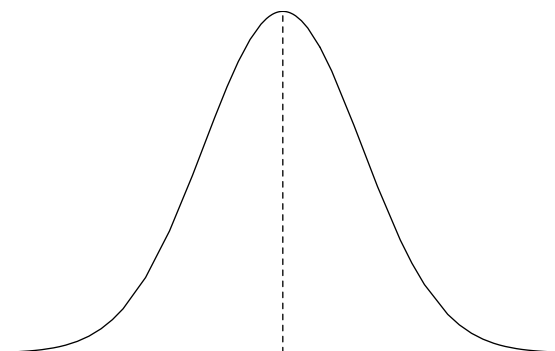


Multimodal distribution.

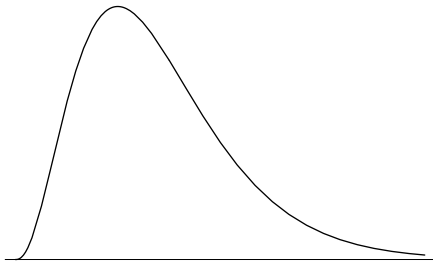
**Definition**

1. A unimodal distribution is **symmetric** if there is a vertical line of symmetry in the distribution.
2. The **lower tail** of a distribution is the leftmost portion of the distribution, and the **upper tail** is the rightmost portion of the distribution.
3. If a unimodal distribution is not symmetric, then it is **skewed**.
  - (a) In a **positively skewed** distribution, or a distribution that is **skewed to the right**, the upper tail extends farther than the lower tail.
  - (b) In a **negatively skewed** distribution, or a distribution that is **skewed to the left**, the lower tail extends farther than the upper tail.

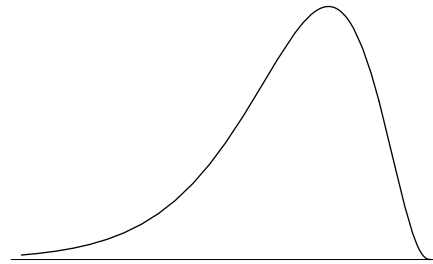
Examples of symmetric, unimodal distributions:



Examples of skewed distributions:



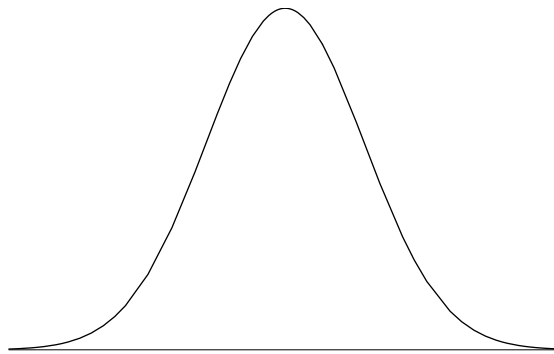
Positively skewed distribution.



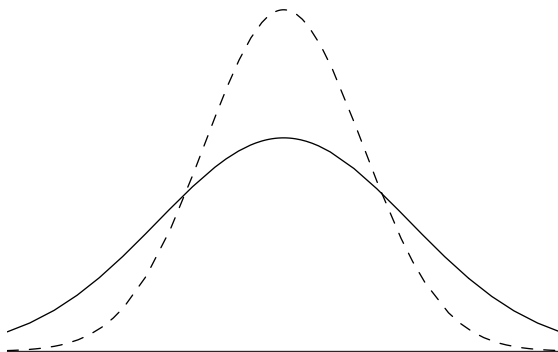
Negatively skewed distribution.

### Normal curve

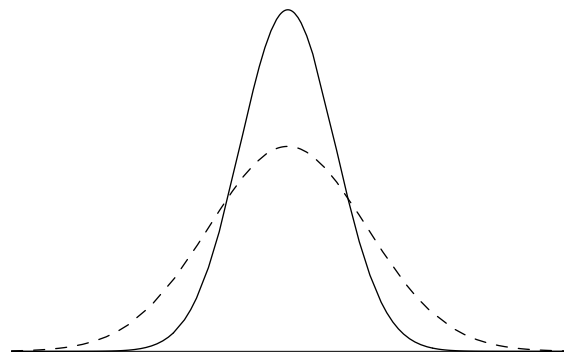
1. Most common unimodal distribution.
2. Symmetric, bell-shaped, approximates many populations.



Compare with a normal curve:



A distribution with heavy tails.



A distribution with light tails.

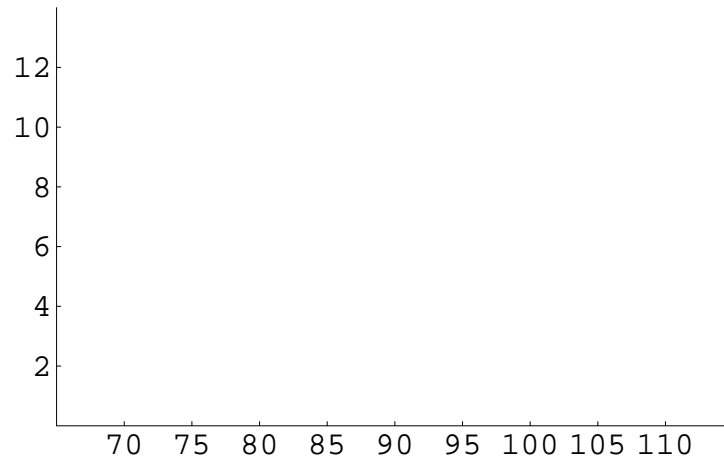


**Example 2.4.6** Nitrates are added to bacon as a preservative, to make the color more appealing, and to prevent the growth of germs. However, nitrate byproducts have been linked to higher incidence rates of cancer in animals. A random sample of uncooked, pre-sliced, thin pieces of bacon was obtained. The amount of sodium nitrate in each slice was measured (in ppm), and the data are given in the following table.

100	80	92	100	97	90	107	103	101	103
81	105	86	106	103	97	89	99	93	103
77	98	96	72	100	103	93	84	88	104
104	99	93	98	95	105	75	97	85	100

- (a) Construct a frequency distribution and a histogram for this data using the class intervals 70–75, 75–80, etc.
- (b) Describe the shape, center, and spread of the distribution.
- (c) What proportion of observations are at least 100 ppm?
- (d) What proportion of observations are less than 90 ppm?

Class	Frequency	Relative frequency	Cumulative relative frequency
70–75			
75–80			
80–85			
85–90			
90–95			
95–100			
100–105			
105–110			
Total			



## CHAPTER 3

# Numerical Summary Measures

---

### 3.1 Measures of Central Tendency

1. Graphical techniques provide useful summaries of data.
2. These techniques are not suitable for statistical inference.
3. Numerical summary measures: more precise, single numbers, used for inference.

A single number computed from a sample, conveys a specific characteristic.

4. Measures of central tendency: where the data are centered or clustered.

#### Notation

1.  $x$  : specific, fixed observation on a variable.

Lowercase letters,  $y, z$ , used to represent observations.

2.  $n$  : number of observations in a data set, the sample size.

Two data sets:  $m, n$ .      More than two data sets:  $n_1, n_2, n_3$ .

3.  $x_1, x_2, x_3, \dots, x_n$  : a set of fixed observations on a variable.

Subscripts: indicate order selected, not magnitude.

4.  $\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$  : summation notation.

$i$ : index of summation; 1: lower bound;  $n$ : upper bound.

**Example 3.1.1** Suppose  $x_1 = 5$ ,  $x_2 = -3$ ,  $x_3 = 7$ ,  $x_4 = 8$ , and  $x_5 = 2$ . Compute the following sums.

$$(a) \left( \sum_{i=1}^5 x_i \right) + 10 \quad (b) \left( \sum_{i=1}^5 x_i \right)^2 \quad (c) \sum_{i=1}^5 x_i^2 \quad (d) \sum_{i=1}^5 2x_i \quad (e) \sum_{i=1}^5 x_i^2 - \frac{1}{5} \left( \sum_{i=1}^5 x_i \right)^2$$

**Definition**

The **sample (arithmetic) mean**, denoted  $\bar{x}$ , of the  $n$  observations  $x_1, x_2, \dots, x_n$  is the sum of the observations divided by  $n$ . Written mathematically:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

**Remarks**

1. Notation: for  $x_1, x_2, \dots, x_n$ : use  $\bar{x}$ ; for  $y_1, y_2, \dots, y_n$ : use  $\bar{y}$ .
2. Population mean:  $\mu$ .

**Example 3.1.2** A vision center allows customers interested in using contact lenses the opportunity to test various types and strengths until an effective and comfortable pair is found. A random sample of customers was obtained, and the number of trial lenses needed was recorded for each. The data are given in the following table.

8	3	4	5	9	2	1	6	8	6
---	---	---	---	---	---	---	---	---	---

Find the sample mean number of trial contact lenses per customer.

**Remarks**

1.  $\bar{x}$ : a sample characteristic.  
One extra decimal place to the right.  
Physically,  $\bar{x}$  is a balance point.
2. The sample mean is *an average*. There are many other averages.
3.  $\mu$ : a population characteristic.  
Usually,  $\mu$  is an unknown constant we would like to estimate.  
What if the population is of finite size  $N$ ?
4.  $\mu$  is a fixed constant.  $\bar{x}$  varies from sample to sample.

**Example 3.1.3** Modify the data in the previous example. Suppose the first customer really needed 18 trial lenses before finding a comfortable pair. The data set is now

---

18	3	4	5	9	2	1	6	8	6
----	---	---	---	---	---	---	---	---	---

---

Find the sample mean for this data set. Describe the effect the observation 18 has on the sample mean.

**Definition**

The **sample median**, denoted  $\tilde{x}$ , of the  $n$  observations  $x_1, x_2, \dots, x_n$  is the *middle number* when the observations are arranged in order from smallest to largest.

1. If  $n$  is odd, the sample median is the single middle value.
2. If  $n$  is even, the sample median is the mean of the two middle values.

**Remarks**

1. Sample median divides the data in half.
2. Only one calculation necessary (none if  $n$  is odd).  
Put the observations in ascending order, and find the middle value.
3. Notation: for  $y_1, y_2, \dots, y_n$ : use  $\tilde{y}$ .
4. Population median:  $\tilde{\mu}$ .

**Example 3.1.4** Find the sample median for each of the following data sets.

(a) 130 151 143 134 120

(b) 130 351 143 134 120

(c) 130 151 143 134 120 188

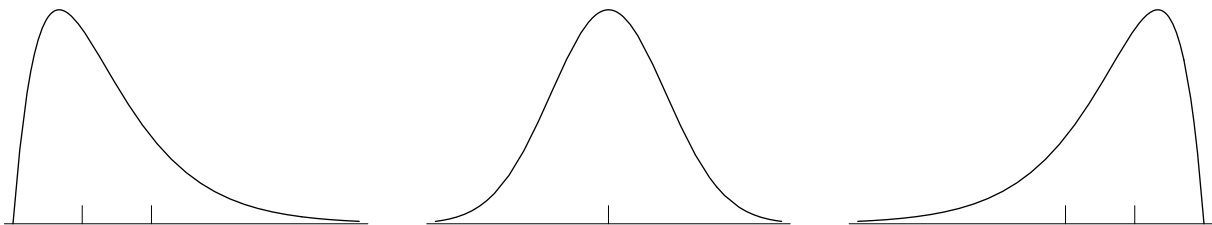
**Example 3.1.5** Commercial aircraft pilots raise their landing gear as soon as the airplane is clear of the ground and definitely airborne. A random sample of airplane takeoffs was selected, and the time (in seconds) from leaving the ground until the landing gear is fully retracted was recorded for each. The data are given in the following table.

15.9	13.7	16.8	11.4	13.8	18.6	14.9
18.4	17.1	11.2	14.7	15.3	14.9	16.0

Find the sample median time to raise the landing gear after takeoff.

### Remarks

1. In general,  $\bar{x} \neq \tilde{x}$  and  $\mu \neq \tilde{\mu}$ .
2. The relative positions of  $\bar{x}$  and  $\tilde{x}$  suggest the shape of a distribution.





**Definition**

A **100p% trimmed mean**, denoted  $\bar{x}_{\text{tr}(p)}$ , of the  $n$  observations  $x_1, x_2, \dots, x_n$  is the sample mean of the *trimmed* data set.

1. Order the observations from smallest to largest.
2. Delete, or trim, the smallest 100p% and the largest 100p% of the observations from the data set.
3. Compute the sample mean for the remaining data.

100p is the **trimming percentage**, the percentage of observations deleted from *each end* of the ordered list.

**Remarks**

1. Trimmed mean is computed by deleting the smallest and largest values—possible outliers.
2. 100p% trimmed mean: smallest 100p% and largest 100p% deleted.  
Total: 2(100p)% of the observations are deleted.
3. To select  $p$ : no set rules;  $np$  an integer.
4. Notation:  $\bar{x}_{\text{tr}(0.025)}$  is a  $(100)(0.025) = 2.5\%$  trimmed mean.  
5% of the observations are deleted.

**Example 3.1.6** The length of time it takes a detective, medical examiners, and the crime lab to arrive at a homicide scene directly affects the likelihood of solving the case. A random sample of homicide cases in a large city was obtained. The length of time (in minutes) until the first detective arrived on the scene was recorded for each. The (ordered) data are given in the following table.

8	15	16	17	17	18	19	19	19	19
20	20	22	23	24	25	25	27	33	43

Find a 5% trimmed mean.

**Definition**

The **mode**, denoted  $M$ , of the  $n$  observations  $x_1, x_2, \dots, x_n$  is the value that occurs most often, or with the greatest frequency.

If all the observations occur with the same frequency, then the mode does not exist.

If two or more observations occur with the same greatest frequency, then the mode is not unique.

**Remarks**

1. Grouped data

Observations:  $x_1, x_2, \dots, x_k$       Frequencies:  $f_1, f_2, \dots, f_k$

Total number of observations:  $n = \sum_{i=1}^k f_i$

Corresponding formulas for measures of central tendency also exist for grouped data.

2. Many other averages: weighted mean, geometric mean, harmonic mean, etc.

Natural summary measures for qualitative data: frequency and relative frequency.

**Example 3.1.7** A random sample of college students was obtained, and the type of laundry detergent used by each was recorded. The (categorical) data is summarized in the following table.

Category	Frequency	Relative frequency
Cheer	35	0.1167
Dial	55	0.1833
Era	30	0.1000
Sam's Club	45	0.1500
Dreft	40	0.1333
Ivory Snow	20	0.0667
Tide	75	0.2500

**Dichotomous or Bernoulli variable**

1. Categorical variable with only two possible responses.
2. One response called a success, denoted S; the other response called a failure, denoted F.

**Definition**

For observations on a categorical variable with only two responses, the **sample proportion of successes**, denoted  $\hat{p}$ , is the relative frequency of occurrence of successes:

$$\hat{p} = \frac{\text{number of S's in the sample}}{\text{total number of responses}} = \frac{n(S)}{n}.$$

**Remarks**

1. Population proportion of successes:  $p$ .
2. A success is not necessarily a good thing.
3.  $\hat{p}$  can be thought of as a sample mean.

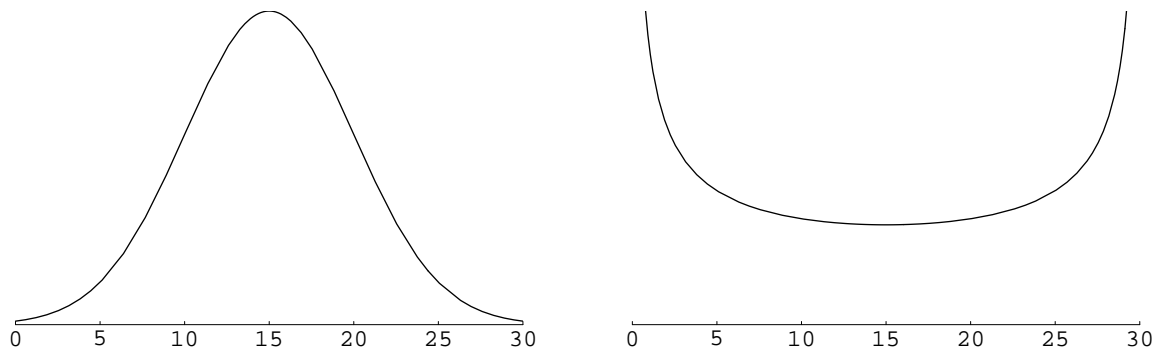
**Example 3.1.8** A random sample of adult males at least 50 years old was obtained. Each was asked whether they collected baseball cards when they were young. A success was recorded for a person who did collect baseball cards, and a failure otherwise. The observations are given in the following table.

F	F	S	F	F	S	S	S	F	F	F	S	F	S	S
S	S	F	F	F	F	S	F	S	S	S	S	F	S	S
F	F	S	F	F	F	F	F	F	S					

Find the sample proportion of adults who collected baseball cards.

## 3.2 Measures of Variability

1. Measures of central tendency are not sufficient to completely describe a sample.
2. Two different data sets can have similar measures of central tendency.



### Definition

The (**sample**) **range**, denoted  $R$ , of the  $n$  observations  $x_1, x_2, \dots, x_n$  is the largest observation minus the smallest observation. Written mathematically:

$$R = x_{\max} - x_{\min},$$

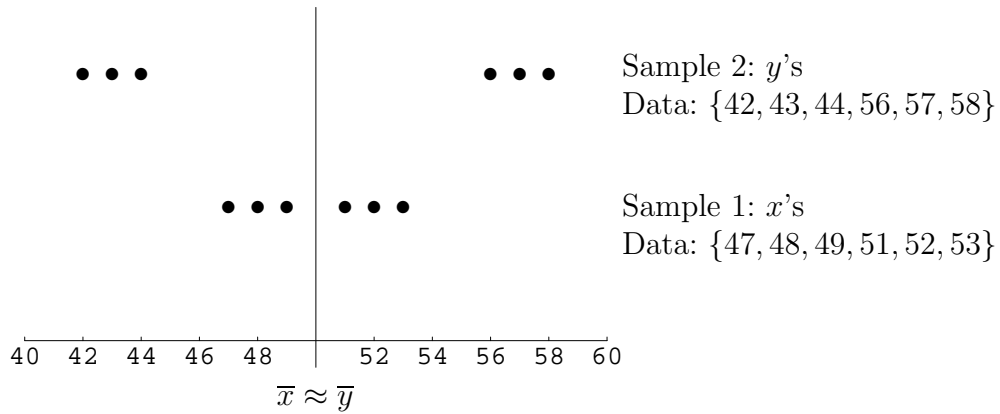
where  $x_{\max}$  denotes the maximum, or largest, observation, and  $x_{\min}$  stands for the minimum, or smallest, observation.

### Remarks

1. In theory, sample range does measure variability.  
Little variability: small range. Lots of variability: large range.
2. Used in many quality-control applications.
3. Not adequate for describing variability.

**Dot Plot**

1. To visualize the variability in a data set.
2. To suggest another measure of variability.



**Definition**

Given a set of  $n$  observations  $x_1, x_2, \dots, x_n$ , the  $i$ th deviation about the mean is  $x_i - \bar{x}$ .

**Remarks**

1. To compute  $x_i - \bar{x}$ : find the mean  $\bar{x}$ , and subtract.
2. Use all the deviations to find a measure of variability.
3. If  $x_i - \bar{x} > 0$ , then  $x_i > \bar{x}$ .      If  $x_i - \bar{x} < 0$ , then  $x_i < \bar{x}$ .

Ideas for measures of variability:

1.  $\sum_{i=1}^n (x_i - \bar{x})$
2.  $\sum_{i=1}^n |x_i - \bar{x}|$

**Definition**

The **sample variance**, denoted  $s^2$ , of the  $n$  observations  $x_1, x_2, \dots, x_n$  is the sum of the squared deviations about the mean divided by  $n - 1$ . Written mathematically:

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] \end{aligned}$$

The **sample standard deviation**, denoted  $s$ , is the positive square root of the sample variance. Written mathematically:  $s = \sqrt{s^2}$ .

**Remarks**

1. Population variance:  $\sigma^2$ . Population standard deviation:  $\sigma$ .
2.  $s^2$  alone doesn't say much about variability. Useful in comparisons.
3.  $s$  is used in many statistical inference problems.
4. Units for  $s^2$ : original data units squared. Units for  $s$ : same as for the original data.
5. Notation: for  $x_1, x_2, \dots, x_n$ : use  $s_x^2$ ; for  $y_1, y_2, \dots, y_n$ : use  $s_y^2$

**Example 3.2.1** A random sample of magazines in the waiting room of doctors' offices was obtained. The age of each magazine (in days, from the publication date) was recorded. The observations were: 12, 35, 14, 21, and 45. Find the sample variance and the sample standard deviation for this data.

**Definition**

The computational formula for the sample variance is

$$s^2 = \frac{1}{n-1} \left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right].$$

**Example 3.2.2** Use the computational formula for  $s^2$  to find the sample variance for the magazine data in the previous example.



**Remarks**

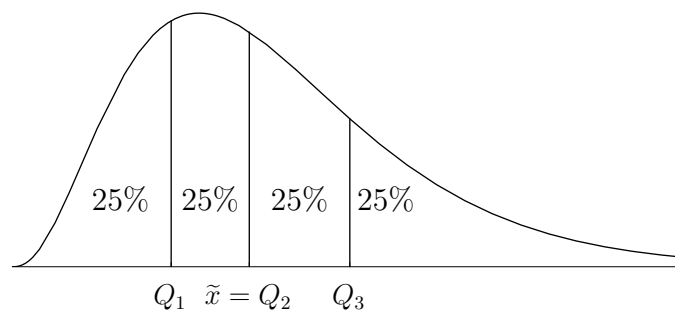
1. Computational formula for  $s^2$ : fewer calculations, more accurate.
2.  $s^2 \geq 0$ . Consider the definition.
3.  $s^2 = 0$ : all observations the same.
4.  $n = 1$ : variance is undefined.

**Definition**

Let  $x_1, x_2, \dots, x_n$  be a set of observations. The **quartiles** divide the data into four parts.

1. The **first (lower) quartile**, denoted  $Q_1$  ( $Q_L$ ) is the median of the lower half of the observations when arranged in ascending order.
2. The second quartile is the median:  $\tilde{x} = Q_2$ .
3. The **third (upper) quartile**, denoted  $Q_3$  ( $Q_U$ ) is the median of the upper half of the observations when arranged in ascending order.
4. The **interquartile range**, denoted  $IQR$ , is the difference  $IQR = Q_3 - Q_1$ .

Smoothed histogram and quartiles:



Quartiles: split the data into four parts.

$\tilde{x} = Q_2$  is the middle value.

$Q_1$ : median of the lower half.     $Q_3$ : median of the upper half.

**How To Compute Quartiles**

Suppose  $x_1, x_2, \dots, x_n$  is a set of  $n$  observations.

1. Arrange the observations in ascending order, from smallest to largest.
2. To find  $Q_1$ , compute  $d_1 = n/4$ .
  - (a) If  $d_1$  is a whole number, then the depth of  $Q_1$  (position in the ordered list) is  $d_1 + 0.5$ .  $Q_1$  is the mean of the observations in positions  $d_1$  and  $d_1 + 1$  in the ordered list.
  - (b) If  $d_1$  is not a whole number, round up to the next whole number for the depth of  $Q_1$ .
3. To find  $Q_3$ , compute  $d_3 = 3n/4$ .
  - (a) If  $d_3$  is a whole number, then the depth of  $Q_3$  is  $d_3 + 0.5$ .  $Q_3$  is the mean of the observations in positions  $d_3$  and  $d_3 + 1$  in the ordered list.
  - (b) If  $d_3$  is not a whole number, round up to the next whole number for the depth of  $Q_3$ .

**Example 3.2.3** A random sample of knives used in an industrial process was obtained. The sharpness of each was measured (in g) by push-cutting light thread. The data are given in the following table.

153	130	76	49	151	68	121	136	132	86	146	107
-----	-----	----	----	-----	----	-----	-----	-----	----	-----	-----

- (a) Find the first quartile, the third quartile, and the interquartile range.
- (b) Suppose there were 14 observations, with  $x_{13} = 45$  and  $x_{14} = 41$ . Find the first and the third quartile for this modified data set.

Example (continued)

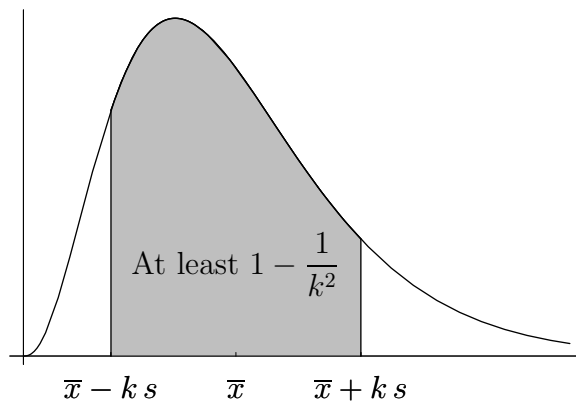
### 3.3 The Empirical Rule and Measures of Relative Standing

1. Measures of central tendency and variability: describe general nature of a data set.
2. Combine these two types of measures to describe a data set more precisely.
3. Measures of relative standing: used to compare observations in different data sets.

#### Chebyshev's Rule

Let  $k > 1$ . For *any* set of observations, the proportion of observations within  $k$  standard deviations of the mean [lying in the interval  $(\bar{x} - ks, \bar{x} + ks)$ , where  $s$  is the standard deviation] is at least  $1 - \frac{1}{k^2}$ .

Smoothed histogram and Chebyshev's Rule:



$k$	$1 - \frac{1}{k^2}$
1.2	
2.0	
2.5	
3.0	

**Remarks**

1. Used to describe a set of observations using symmetric intervals about the mean.
2. Proportion of observations outside the interval  $(\bar{x} - k s, \bar{x} + k s)$ :
3.  $k$  can be any value greater than 1. Most common values:  $k = 2, 3$ .
4. Very conservative. Applies to any set of observations.
5. May also be used to describe a population. Use  $\mu$  and  $\sigma$ .

**Example 3.3.1** Homes in a certain rural area often experience power outages during the winter months due to extreme weather conditions. A random sample of homes affected by a power outage was selected, and the time (in hours) until power was restored was recorded for each. The sample mean was  $\bar{x} = 17$  and the sample standard deviation was  $s = 4.2$ . Use Chebyshev's Rule with  $k = 2$  and  $k = 3$  to describe this distribution of time to restore power.

**Example 3.3.2** Many home gardeners add used coffee grounds to the soil for mulch and compost. A random sample of 100-gram packages of used coffee grounds was obtained, and the amount of nitrogen (in grams) was measured in each. The summary statistics were  $\bar{x} = 3.5$  and  $s = 0.4$ .

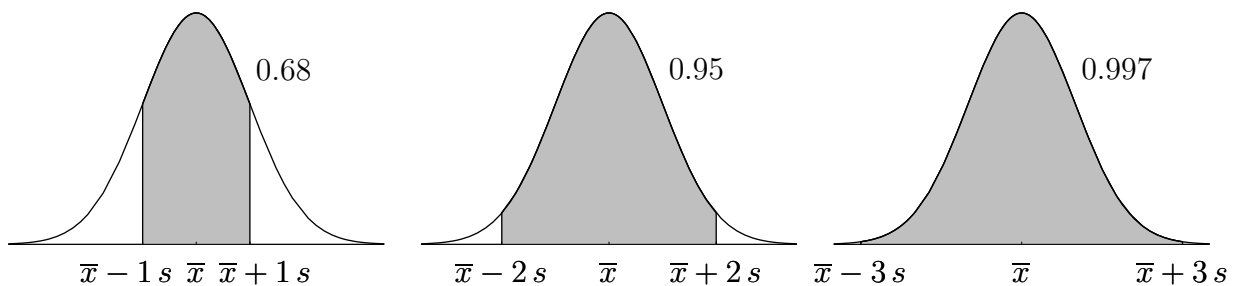
- (a) Find the approximate proportion of observations between 2.7 and 4.3.
- (b) Find the approximate proportion of observations less than 2.3 or greater than 4.7.
- (c) Approximately what proportion of observations had less than 1.9 grams of nitrogen?

**The Empirical Rule**

If the shape of the distribution of a set of observations is approximately normal, then:

1. The proportion of observations within one standard deviation of the mean is approximately 0.68.
2. The proportion of observations within two standard deviations of the mean is approximately 0.95.
3. The proportion of observations within three standard deviations of the mean is approximately 0.997.

Symmetric intervals and proportions:

**Remarks**

1. Backward Empirical Rule: used to check normality.

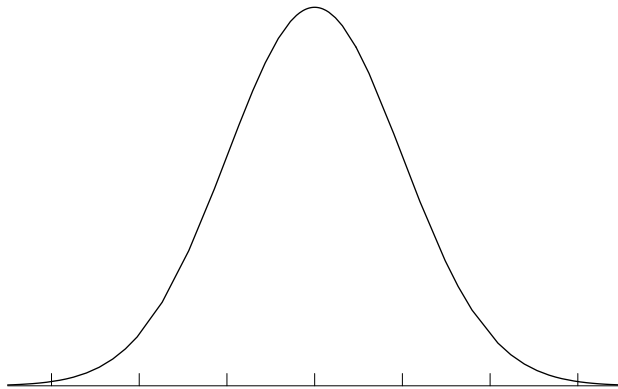
Compute the actual proportions, and compare with 0.68, 0.95, and 0.997.

2. May also be used to describe a population: use  $\mu$  and  $\sigma$ .
3. Proportion beyond three standard deviations:  $1 - 0.997 = 0.003$ .

What if an observation is more than three standard deviations from the mean?

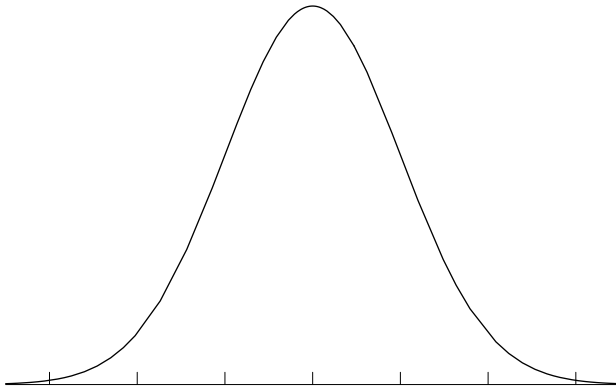
**Example 3.3.3** A random sample of tankless water heaters was obtained, and the first-hour rating (as specified by the Department of Energy) was measured for each (in gal/hr). The shape of the distribution is approximately normal with  $\bar{x} = 200$  and  $s = 25$ .

- (a) Approximately what proportion of observations is between 150 and 250?
- (b) Approximately what proportion of observations is less than 150 or greater than 250?
- (c) Approximately what proportion of observations is less than 275?
- (d) Approximately what proportion of observations is between 175 and 275?





**Example 3.3.4** A hardware manufacturer makes a certain 3-inch-diameter pipe for use in commercial buildings. The company claims that the distribution of the weight of the pipes (in pounds per foot) is approximately normal with mean  $\mu = 18.6$  and standard deviation  $\sigma = 0.25$ . A 3-inch pipe manufactured by this company is randomly selected, and the weight is 19.4 pounds per foot. Is there any evidence to refute the manufacturer's claim?



**Definition**

Suppose  $x_1, x_2, \dots, x_n$  is a set of  $n$  observations with mean  $\bar{x}$  and standard deviation  $s$ . The  **$z$ -score** corresponding to the  $i$ th observation,  $x_i$ , is given by

$$z_i = \frac{x_i - \bar{x}}{s}.$$

$z_i$  is a measure associated with  $x_i$  that indicates the distance from  $\bar{x}$  in standard deviations.

**Remarks**

1.  $z_i$  may be positive, negative, or zero.

$z_i > 0$ : observation to the right of the mean.

$z_i < 0$ : observation to the left of the mean.

2.  $z$ -score: a measure of relative standing—where the observation lies in relation to the rest.

Other methods of standardization exist.

3.  $\sum_{i=1}^n z_i =$

**Example 3.3.5** A manufacturer has assembly lines for making three different cell phones. Engineering specifications for the mean weight (in ounces) and standard deviation of the weight are given in the table below. One cell phone was randomly selected from each assembly line. The weight was carefully measured and is also given in the table below.

Assembly line	Mean	Standard deviation	Observation
1	3.52	0.20	3.66
2	4.10	0.05	4.05
3	2.95	0.03	3.05

(a) Find the  $z$ -score corresponding to each observation.

(b) Using the  $z$ -scores, do you think any assembly line should be checked? Justify your answer.

Example (continued)

**Example 3.3.6** Many residents in flood-prone areas use a sump pump to keep their basements dry and mold-free. Suppose a certain sump pump is designed to pump on average 800 gallons of water per hour with a standard deviation of 10 gallons per hour. A consumer agency tested one of these pumps and reported a pump rate of 793 gallons per hour. Is this a reasonable pump rate, or do you think the manufacturer should lower the rating for this pump? Explain.

**Definition**

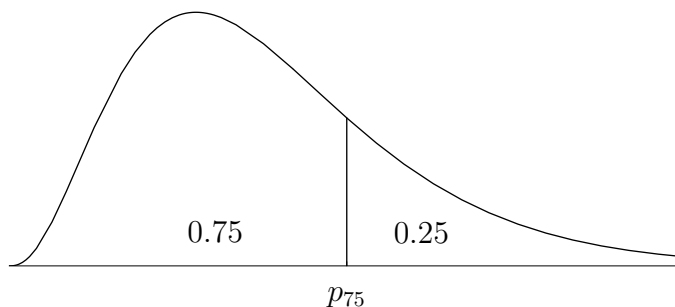
Let  $x_1, x_2, \dots, x_n$  be a set of observations. The **percentiles** divide the data set into 100 parts. For any integer  $r$  ( $0 < r < 100$ ) the  **$r$ th percentile**, denoted  $p_r$ , is a value such that  $r$  percent of the observations lie at or below  $p_r$  (and  $1 - r$  percent lie above  $p_r$ ).

**Remarks**

1. 50th percentile is the median:  $p_{50} = \tilde{x}$ .
2. 25th percentile is the first quartile, 75th percentile is the third quartile:

$$p_{25} = Q_1, p_{75} = Q_3.$$

Illustration of the 75th percentile:

**How To Compute Percentiles**

Suppose  $x_1, x_2, \dots, x_n$  is a set of  $n$  observations.

1. Arrange the observations in ascending order, from smallest to largest.
2. To find  $p_r$ , compute  $d_r = \frac{n \cdot r}{100}$ .
  - (a) If  $d_r$  is a whole number, then the depth of  $p_r$  (position in the ordered list) is  $d_r + 0.5$ .  $p_r$  is the mean of the observations in positions  $d_r$  and  $d_r + 1$  in the ordered list.
  - (b) If  $d_r$  is not a whole number, round up to the next whole number for the depth of  $p_r$ .

**Example 3.3.7** The weight of all cargo on a merchant vessel is carefully monitored by the Coast Guard and Port Authorities. A vessel arriving at the Port of Charleston reported a cargo weight of 200 tons, which lies at the 65th percentile. Interpret this value.

**Example 3.3.8** The absolute neutrophil count (ANC), a measure of the number of white blood cells, is an indication of an individual's ability to fight infection. A random sample of adults was obtained, and the ANC for each person was measured. The results are given in the following table.

3165	3446	4197	5511	4524	3940	5168	5279	3290	3749
4658	3624	3185	4709	5803	5581	4466	4155	4965	3254
4018	4237	4447	5124	4839	3109	5087	4089	5462	4267
3190	4493	3179	5517	4321	3608	2998	3076	4803	3496

Find the 70th percentile and the 33rd percentile.

### 3.4 Five-Number Summary and Box Plots

1. Box plot: compact graphical summary, conveys information about central tendency, symmetry, skewness, variability, and outliers.
2. Standard box plot versus modified box plot.

#### Definition

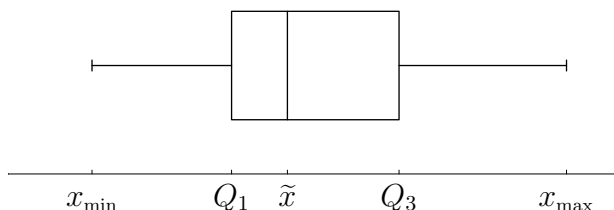
The **five-number summary** for a set of  $n$  observations  $x_1, x_2, \dots, x_n$  consists of the minimum value, the maximum value, the first and third quartiles, and the median.

#### How To Construct a Standard Box Plot

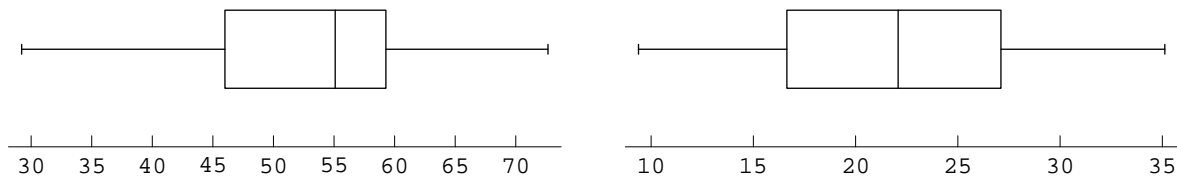
Given a set of  $n$  observations  $x_1, x_2, \dots, x_n$ :

1. Find the five-number summary:  $x_{\min}, Q_1, \tilde{x}, Q_3, x_{\max}$ .
2. Draw a (horizontal) measurement axis. Carefully sketch a box with edges at the quartiles: left edge at  $Q_1$ , right edge at  $Q_3$ . (The height of the box is irrelevant.)
3. Draw a vertical line in the box at the median.
4. Draw a horizontal line (whisker) from the left edge of the box to the minimum value (from  $Q_1$  to  $x_{\min}$ ). Draw a horizontal line (whisker) from the right edge of the box to the maximum value (from  $Q_3$  to  $x_{\max}$ ).

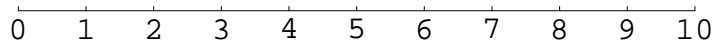
Standard box plot:



A box plot conveys symmetry or skewness.



**Example 3.4.1** A random sample of IRS-classified small corporations was obtained, and the gross receipts (in millions of dollars) was recorded for each. The five-number summary was:  $x_{\min} = 0.2$ ,  $Q_1 = 3.1$ ,  $\tilde{x} = 4.5$ ,  $Q_3 = 6.2$ ,  $x_{\max} = 9.1$ . Use the five-number summary to construct a box plot.



### Remarks

1. Only one measurement axis, may be horizontal or vertical.
2. Five-number summary usually not displayed on the graph.  
Tick marks and scale selected for convenience.

**How To Construct a Modified Box Plot**

Given a set of  $n$  observations  $x_1, x_2, \dots, x_n$ :

1. Find the quartiles, the median, and the interquartile range:

$$Q_1, \tilde{x}, Q_3, IQR = Q_3 - Q_1.$$

2. Compute the two inner *fences* (low and high) and two outer (low and high) *fences* using the following formulas:

$$IF_L = Q_1 - 1.5(IQR) \quad IF_H = Q_3 + 1.5(IQR)$$

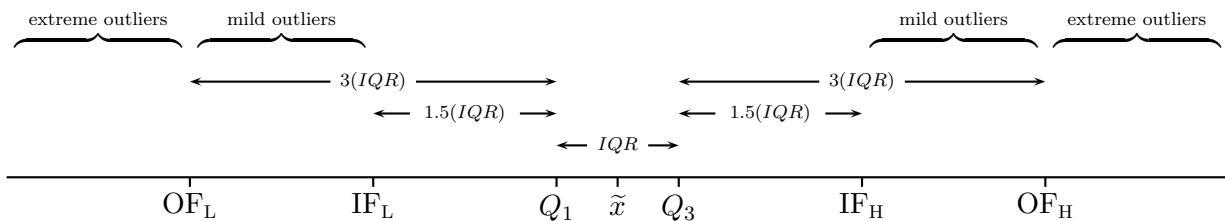
$$OF_L = Q_1 - 3(IQR) \quad OF_H = Q_3 + 3(IQR)$$

Think of the interquartile range as a *step*. The inner fences are 1.5 steps away from the quartiles, and the outer fences are 3 steps away from the quartiles.

3. Draw a (horizontal) measurement axis. Carefully sketch a box with edges at the quartiles: left edge at  $Q_1$ , right edge at  $Q_3$ . Draw a vertical line in the box at the median.
4. Draw a horizontal line (whisker) from the left edge of the box to the most extreme observation within the low inner fence. This line will extend from  $Q_1$  to at most  $IF_L$ . Draw a horizontal line (whisker) from the right edge of the box to the most extreme observation within the high inner fence. This line will extend from  $Q_3$  to at most  $IF_H$ .
5. Any observations between the inner and outer fences (between  $IF_L$  and  $OF_L$ , or between  $IF_H$  and  $OF_H$ ) are classified as *mild outliers* and are plotted separately with shaded circles.

Any observations outside the outer fences (less than  $OF_L$ , or greater than  $OF_H$ ) are classified as *extreme outliers* and are plotted separately with open circles.

Modified box plot construction points:





**Example 3.4.2** A random sample of classes in the California State University system was obtained, and the number of students in each class was recorded. The data are given in the following table.

21	25	34	32	4	12	15	21	20	6
19	88	60	85	21	14	6	10	16	2

Construct a modified box plot for this data, and use this graph to describe the distribution.



**Example 3.4.3** An experiment was conducted to measure the incubation period (in days) for hepatitis A. The data are given in the following table.

55	31	38	38	33	33	44	29	34	43	32	48	32	30	40
25	38	12	35	36	29	48	29	32	25	23	17	31	24	26
30	34	39	41	29	35	22	27	33	11	14	30	33	36	36

Construct a modified box plot for this data, and use this graph to describe the distribution.



## CHAPTER 4

# Probability

---

## 4.0 Introduction

Recall:

1. Probability and statistics are both related to a sample and a population.
2. Probability: certain properties of a population are assumed to be known.  
Calculate the probability of observing a specific outcome associated with a sample.
3. Statistics: Use the information in a sample to draw a conclusion about an entire population.
4. Need a solid background in probability in order to understand and solve statistics problems.

### Probability

1. A study of randomness and uncertainty.
2. Characterize indeterminate or indefinite events.
3. Determine the likelihood of an observed outcome—important for statistical inference.

---

## 4.1 Experiments, Sample Spaces, and Events

1. In order to understand probability, we need to think precisely about *experiments*.
2. Consider: tossing a coin, selecting a card, counting the number of unforced errors in a tennis tournament, or weighing a pumpkin.
3. In every one of these activities, there is uncertainty.  
We do not know for certain what the outcome will be.

**Definition**

An **experiment** is an activity in which there are at least two possible outcomes and the result of the activity cannot be predicted with absolute certainty.

**Example 4.1.1** Here are some examples of experiments.

- (a) Select a computer in a public library and record the number of icons on the desktop.

We cannot say for certain whether there will be 1, or 2, or . . . .  
This activity is an experiment.

- (b) Count the number of students who have brunch in the cafeteria on a Sunday morning.

Using previous information, might be able to estimate.  
But, no way of knowing the exact number.

- (c) Select a home with beachfront property, and measure the distance from the home to the water at high tide.

There may be some construction guidelines and legal restrictions.  
But, cannot predict the exact distance to the ocean.

- (d) Select two events held at a local arena, and record whether or not they were sellouts.

Might be able to guess on the basis of popularity.  
No way of knowing whether both are sellouts, one is a sellout, or neither is a sellout.

The outcome in an experiment is uncertain.

Need to consider *all* possible outcomes.

**Example 4.1.2** Suppose a resident of a large apartment complex is selected, and the last digit of their apartment number is recorded. How many possible outcomes are there, and what are they?

**Example 4.1.3** Two residents of Mesa, Arizona, are selected at random and asked whether or not they use a water filter.

- (a) How many possible outcomes are there, and what are they?
- (b) Construct a **tree diagram** for this experiment.
- (c) How many possible outcomes are there if we ask three residents whether or not they use a water filter?

**Remarks**

1. Tree diagrams can get very big, very fast.
2. Used to prove the *multiplication rule*, an arithmetic technique used to count the number of possible outcomes in certain experiments.
3. A tree diagram does not have to be symmetric.

**Example 4.1.4** A history professor is conducting research into how well citizens know and understand the Constitution. As part of this study, an experiment will be conducted in which three students will be asked whether they can recite the Preamble to the Constitution from memory. The experiment will stop if a student can recite the Preamble. How many possible outcomes are there, and what are they?

**Definition**

The **sample space** associated with an experiment is a listing of all the possible outcomes, *using set notation*. It is the collection of all outcomes written mathematically, with curly braces, and denoted by  $S$ .

**Example 4.1.5** Find the sample space for each of the four experiments above.

- (a) Last digit of the apartment number:
  
  
- (b) Water filter experiment:
  
  
- (c) Extended water filter experiment (three residents):
  
  
- (d) Recite the Preamble experiment:

**Definition**

1. An **event** is any collection (or set) of outcomes from an experiment (any subset of the sample space).
2. A **simple event** is an event consisting of exactly one outcome.
3. An event has **occurred** if the resulting outcome is contained in the event.

**Remarks**

1. An event may be given in standard set notation, or it may be defined in words.
2. Notation:
  - (a) Events are denoted with capital letters, for example,  $A, B, C, \dots$
  - (b) Simple events are often denoted by  $E_1, E_2, E_3, \dots$
3. It is possible for an event to be empty.  
An event containing no outcomes is denoted by  $\{ \}$  or  $\emptyset$  (the empty set).

**Example 4.1.6** Two universities are selected at random, and each is classified by the type of predominant network: wireless (W) or hard-wired (H). An experiment consists of recording the two classifications.

- (a) Find the sample space  $S$  for this experiment.
- (b) Identify the four simple events.
- (c) List the outcomes in each of the following events.  
 $A$  = both campuses are classified the same.  
 $B$  = at most one campus is wireless.  
 $C$  = at least one campus is hard-wired.



**Example 4.1.7** A commuter taking a train to Grand Central Station may get on at the Apple Valley (1), Noxon Road (2), Maple Knoll (3), or North Cherry Street (4) stop, and may hold a monthly rail pass (M), weekly rail pass (W), or a one-way ticket (O). An experiment consists of selecting a commuter arriving at Grand Central Station and recording the boarding stop and type of ticket.

- (a) Construct a tree diagram to illustrate this experiment.  
Find the sample space  $S$  for this experiment.
- (b) List the outcomes in each of the following events.  
 $A$  = the commuter boarded at one of the first two stops and had a one-way ticket.  
 $B$  = the commuter boarded at an even-numbered stop.  
 $C$  = the commuter had a monthly rail pass.  
 $D$  = the commuter boarded at the last stop before Grand Central Station.

**Example 4.1.8** A financial consultant may advise a client to invest in a growth fund (G), bonds (B), or the money market (M), and the stock market may rise (R), fall (F), or stay about the same (E). An experiment consists of recording a financial consultant's advice and the stock market movement over the next month.

- (a) Find the sample space  $S$  for this experiment.
- (b) List the outcomes in each of the following events.

$A$  = the financial consultant advises the client to invest in a growth fund or the money market, and the market falls.

$C$  = the financial consultant advises the client to invest in the money market.

$D$  = the market rises over the next month.

### Remarks

1. When an experiment is conducted, only one outcome can occur.
2. Given an experiment, the sample space, and some relevant events, we often combine events in various ways to create and study new events.

**Definition**

Let  $A$  and  $B$  denote two events associated with a sample space  $S$ .

1. The event  **$A$  complement**, denoted  $A'$ , consists of all outcomes in the sample space  $S$  *not* in  $A$ .
2. The event  **$A$  union  $B$** , denoted  $A \cup B$ , consists of all outcomes in  $A$  or  $B$  or both.
3. The event  **$A$  intersection  $B$** , denoted  $A \cap B$ , consists of all outcomes in both  $A$  and  $B$ .
4. If  $A$  and  $B$  have no elements in common, they are **disjoint** or **mutually exclusive**, written  $A \cap B = \{ \}$ .

**Remarks**

1. The event  $A'$  is also called **not  $A$** . The word *not* usually means complement.
2. *Or* usually means union;  $A$  or  $B$  means  $A \cup B$ .  
*And* usually means intersection;  $A$  and  $B$  means  $A \cap B$ .
3. Any outcome in *both*  $A$  and  $B$  is included only once in the event  $A \cup B$ .
4.  $A'$ ,  $A \cup B$ , and  $A \cap B$   
Traditional mathematical symbols for complement, union, and intersection.  
Could use any other symbols.
5. It is possible for one of these new events to contain all the outcomes in the sample space.

**Example 4.1.9** A beverage retailer sells three brands of energy drinks, AMP (A), Blue Ox (B), and Crunk (C). An experiment consists of recording the brand of energy drink purchased by the next two customers who buy this kind of beverage. Consider the following events.

$E$  = both customers buy the same brand.

$F$  = neither customer buys AMP.

$G$  = at least one customer buys Crunk.

$H$  = exactly one customer buys Blue Ox.

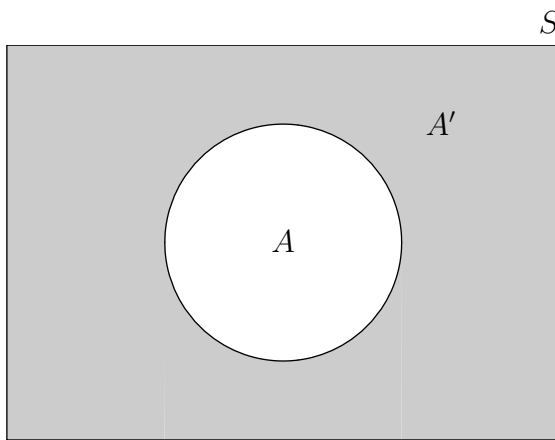
Describe the following events, and list the outcomes in each:

$E'$ ,  $F \cup G$ ,  $E \cap G$ ,  $F \cap H$ ,  $(E \cap G)'$ ,  $(G \cup H)'$

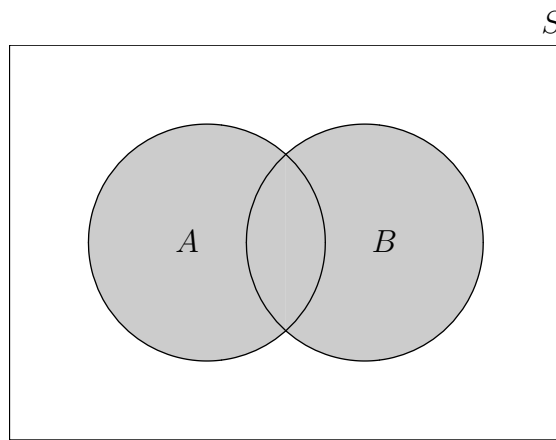
**Venn Diagram**

1. Used to visualize a sample space, events, and combinations of events.
2. Draw a rectangle to represent the sample space.
3. Figures (often circles) are drawn inside the rectangle to represent events.
4. Plane regions represent events.

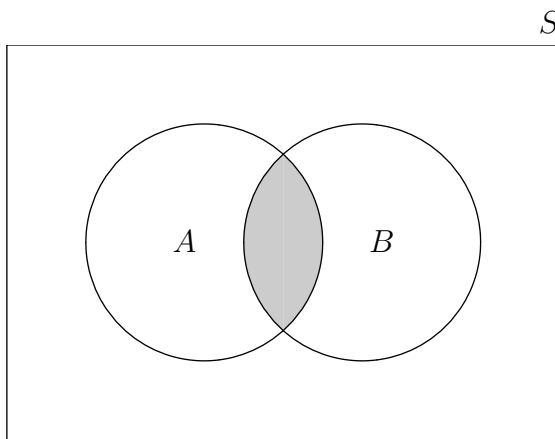
Combinations of events represented by Venn diagrams:



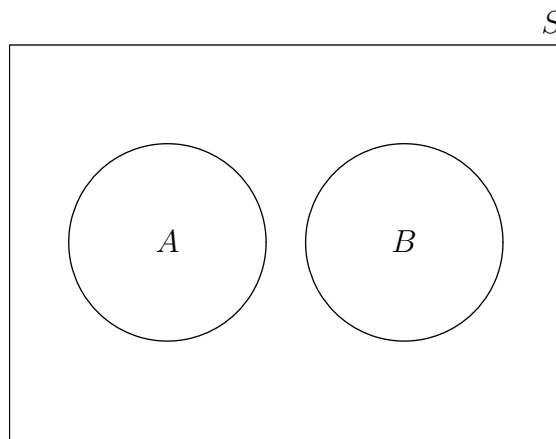
$A$  complement:  $A'$



$A$  union  $B$ :  $A \cup B$



$A$  intersection  $B$ :  $A \cap B$



$A$  and  $B$  are disjoint:  $A \cap B = \{ \}$

**Definition**

Let  $A_1, A_2, A_3, \dots, A_k$  be a collection of  $k$  events.

1. The event  $A_1 \cup A_2 \cup \dots \cup A_k$  is a **generalized union** and consists of all outcomes in at least one of the events  $A_1, A_2, A_3, \dots, A_k$ .
2. The event  $A_1 \cap A_2 \cap \dots \cap A_k$  is a **generalized intersection** and consists of all outcomes in every one of the events  $A_1, A_2, A_3, \dots, A_k$ .
3. The  $k$  events  $A_1, A_2, A_3, \dots, A_k$  are **disjoint** if no two have any outcome in common.

**Example 4.1.10** Consider an experiment with sample space  $S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$  and the events

$$A = \{0, 2, 4, 6, 8\}, \quad B = \{5, 6, 7, 8, 9\}, \quad C = \{0, 1, 8, 9\}, \quad D = \{3, 4, 5, 6, 7\}$$

- (a) Use a Venn diagram to illustrate the events  $A$ ,  $B$ , and  $C$ , and list the outcomes in the event  $A \cup B \cup C$ .
- (b) Use a Venn diagram to illustrate the events  $B$ ,  $C$ , and  $D$ , and list the outcomes in the event  $B \cup C \cup D$ .
- (c) List the outcomes in each of the following events.
  - (i)  $A \cap B \cap C$
  - (ii)  $B \cap C \cap D$
  - (iii)  $(C \cup D)'$
  - (iv)  $(A \cap B \cap D)'$

Example (continued)

**Example 4.1.11** An appliance store sells refrigerators with the freezer on the top, on the bottom, or in a side-by-side model. Each refrigerator is available in white, black, autumn wheat, or stainless steel. An experiment consists of recording the color and model type of the next refrigerator purchase. Consider the events

$A$  = the refrigerator purchased is white or black and not side-by-side.

$B$  = the refrigerator purchased is stainless steel.

$C$  = the refrigerator purchased is side-by-side.

- (a) Find the sample space  $S$  for this experiment.
- (b) List the outcomes in the events  $A$ ,  $B$ , and  $C$ .
- (c) List the outcomes in each of the following events, and illustrate with a Venn diagram.
  - (i)  $B \cap C$
  - (ii)  $A \cup B$
  - (iii)  $A \cup C$
  - (iv)  $(A \cup B \cup C)'$



Example (continued)

---

## 4.2 An Introduction to Probability

1. Given an experiment, some events are more likely to occur than others.
2. For an event  $A$ , assign a number that conveys the likelihood of occurrence of  $A$ .
3. This number is called the probability of the event  $A$ , denoted  $P(A)$ .

### Definition

The probability of an event  $A$  is a number between 0 and 1 (including those endpoints) that measures the likelihood  $A$  will occur.

1. If the probability of an event is close to 1, then the event is likely to occur.
2. If the probability of an event is close to 0, then the event is not likely to occur.

A reasonable method for assigning a probability to an event is linked to relative frequency.

### Definition

The **relative frequency of occurrence of an event** is the number of times the event occurs divided by the total number of times the experiment is conducted.

**Example 4.2.1** An American roulette wheel has 38 slots. The slots are numbered 1–36, 0, and 00. Eighteen numbers are red, eighteen are black, and 0 and 00 are green. Suppose an experiment consists of spinning the wheel, dropping a ball into the wheel, and recording the color of the slot in which the ball lands. What is the probability the ball will land in a red slot?

### Solution

- (S1) Let  $R$  be the event that the ball lands in a red slot. Want  $P(R)$ .
- (S2) To estimate the likelihood of a red slot, use the relative frequency of occurrence of a red slot.

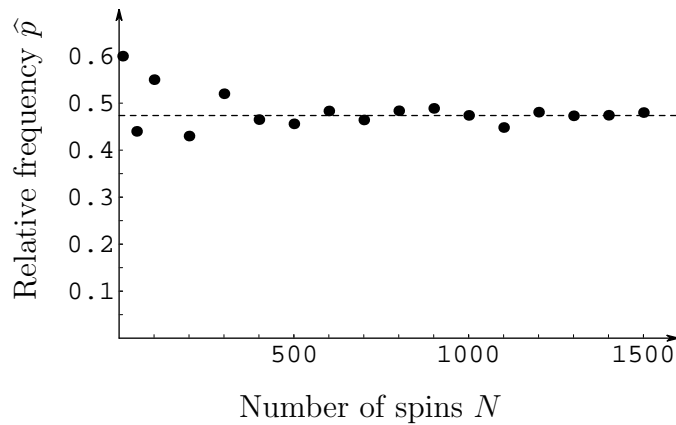
$$\text{relative frequency} = \frac{\text{number of times a red slot occurs}}{\text{total number of spins}}$$

(S3) Let  $N$  be the total number of spins and  $\hat{p}$  be the relative frequency of occurrence of a red slot.

Suppose we conduct the experiment a large number of times.

$N$	10	50	100	200	300	400	500	600	700
$\hat{p}$	0.6000	0.4400	0.5500	0.4300	0.5200	0.4650	0.4560	0.4833	0.4643
$N$	800	900	1000	1100	1200	1300	1400	1500	
$\hat{p}$	0.4837	0.4889	0.4740	0.4482	0.4808	0.4731	0.4743	0.4800	

(S4) Here is a scatter plot of relative frequency versus number of trials.



(S5) As  $N$  increases, the points are closer to the dashed line.

The relative frequencies home in on one number (around 0.47).

This relative frequency should be the probability of the event  $R$ .

(S6) In the long run, the relative frequencies tend to stabilize, or even out.

They close in on one number, the *limiting relative frequency*.

**Remarks**

1. If an experiment is conducted  $N$  times and an event occurs  $n$  times, then the probability of the event is *approximately*  $n/N$  (the relative frequency of occurrence).
2. The **probability of an event**  $A$ ,  $P(A)$ , is the *limiting* relative frequency.
3. How do we find the limiting relative frequency?  
Can we find the exact limiting relative frequency?

**Example 4.2.2** Suppose an experiment consists of tossing a fair coin and recording the side face-up. The event  $T$  is the coin landing with heads face-up. Find  $P(T)$ .

**Example 4.2.3** An experiment consists of tossing a fair six-sided die and recording the number face-up. Consider the event  $E = \{4\}$ , rolling a one. Find  $P(E)$ .

**Example 4.2.4** An experiment consists of selecting one card from a regular 52-card deck, and recording the denomination. Suppose  $A$  is the event of selecting an ace. What is  $P(A)$ ?

**Remarks**

1. These examples suggest that it is possible to find the limiting relative frequency.
2. They are special cases: all of the outcomes are *equally likely*.

**Properties of Probability**

1. For any event  $A$ ,  $0 \leq P(A) \leq 1$ .

The probability of any event is a limiting *relative frequency*, and a relative frequency is a number between 0 and 1. An event with probability close to 0 is very unlikely to occur, and an event with probability close to 1 is very likely to occur.

2. For any event  $A$ ,  $P(A)$  is the sum of the probabilities of all of the outcomes in  $A$ . To compute  $P(A)$ , just add up the probabilities of each outcome, or simple event, in  $A$ .
3. The sum of the probabilities of all possible outcomes in a sample space is 1:  $P(S) = 1$ . The sample space  $S$  is an event. If an experiment is conducted,  $S$  is guaranteed to occur.
4. The probability of the empty set is 0:  $P(\{ \}) = P(\emptyset) = 0$ . This event contains no outcomes.

**Example 4.2.5** An online retailer sells six different ink-jet cartridges for a specific printer. An experiment consists of classifying the next cartridge purchase. The probability of each simple event is given in the following table.

Simple event	Black	Photo cyan	Cyan	Photo magenta	Magenta	Yellow
Probability	0.30	0.15	0.10	0.20	0.08	0.17

Consider the following events.

$A$  = The next cartridge purchased is either photo cyan or cyan. =  $\{PC, CY\}$

$B$  = The next cartridge purchased is one of the magentas or yellow. =  $\{PM, M, Y\}$

$C$  = The next cartridge purchased is different from black. =  $\{PC, CY, PM, M, Y\}$

Find  $P(A)$ ,  $P(A \cup B)$ , and  $P(B \cap C)$ .

**Example 4.2.6** The Bureau of Labor Statistics compiles data regarding fatal occupation injuries of employees while performing the duties of their job. Recently, the number of fatal injuries and the fatality rate have declined to new lows in the United States. An experiment consists of recording the location of the next fatal occupation injury. The probability of each simple event is given in the following table.

Simple event	Grocery store G	Liquor store L	Restaurant R	Fast-food establishment F	Lounge N	Transit vehicle V	Other O
Probability	0.28	0.21	0.16	0.12	0.10	0.08	0.05

Consider the following events.

$$A = \{G, L, R\} \quad B = \{F, N, V\}$$

$$C = \{G, V, O\} \quad D = \{L, F, N, V, O\}$$

Find  $P(A)$ ,  $P(C)$ ,  $P(D)$ ,  $P(A \cup B)$ ,  $P(B \cap C)$ , and  $P(B \cap D)$ .

**Equally Likely Outcome Experiment**

1. Suppose an experiment has  $n$  equally likely outcomes,  $S = \{e_1, e_2, e_3, \dots, e_n\}$ .
2. Probability of each is  $1/n$ ;  $P(e_i) = 1/n$ . Limiting relative frequency of  $e_i$ :  $1/n$ .
3. Consider an event  $A = \{e_1, e_2, e_3, e_4, e_5\}$ .

To find  $P(A)$ , add up the probabilities of each simple event in  $A$ .

$$\begin{aligned} P(A) &= P(e_1) + P(e_2) + P(e_3) + P(e_4) + P(e_5) \\ &= \frac{1}{n} + \frac{1}{n} + \frac{1}{n} + \frac{1}{n} + \frac{1}{n} = \frac{5}{n} \\ &= \frac{\text{number of outcomes in } A}{\text{number of outcomes in the sample space } S} = \frac{N(A)}{N(S)} \end{aligned}$$

**How To Find Probabilities in an Equally Likely Outcome Experiment**

In an equally likely outcome experiment, the probability of *any* event  $A$  is the number of outcomes in  $A$  divided by the total number of outcomes in the sample space  $S$ . Finding the probability of any event, in this case, means counting the number of outcomes in  $A$ , counting the number of outcomes in the sample space  $S$ , and dividing.

$$P(A) = \frac{N(A)}{N(S)}$$

**Example 4.2.7** A software company has five people who routinely test new video games before they are marketed. Two of the five use the Xbox, and the other three use the Sony PlayStation. Suppose two testers will be selected at random to try a new game.

- (a) What is the probability both testers selected will be Xbox users?
- (b) What is the probability one Xbox user and one PlayStation user will be selected?

**Example 4.2.8** The EPA has warned travelers that the water on some airplanes may not be fit for drinking or even for hand-washing. EPA officials have decided to conduct a surprise inspection at a major airport. They will inspect two of the six planes currently being serviced. Suppose two of the six have high levels of bacteria in the drinking water.

- (a) Find the probability both planes inspected will have high levels of bacteria in the drinking water.
- (b) Find the probability exactly one plane selected will have a high level of bacteria in the drinking water.
- (c) Find the probability neither plane selected will have a high level of bacteria in the drinking water.



**Example 4.2.9** An outdoor family entertainment center has go-carts, batting cages, and two miniature golf courses called The Bay and The Fortress. The owner believes that families have no preference for either course, that is, each family that plays miniature golf selects a course at random. Four families who intend to play miniature golf are selected, and the course played is recorded.

- (a) Find the probability exactly one family plays The Bay course.
- (b) Find the probability exactly two families play The Fortress course.
- (c) Suppose all four families play The Fortress course. Is there any evidence to suggest that families prefer one course over the other?

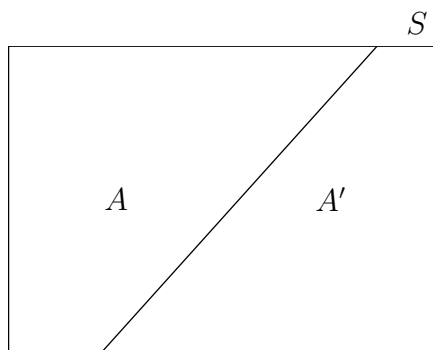
**Remarks**

1. Sometimes we can use known probabilities to find the probability of a new event.
2. We may not have to look at simple events, or count outcomes.

**The Complement Rule**

For any event  $A$ ,  $P(A) = 1 - P(A')$ .

Venn diagram for visualizing the Complement Rule:



Keywords: not, at least, at most.

**Example 4.2.10** A cafeteria worker in a middle school places a piece of fruit on every student's tray as they pass through the lunch line. She randomly selects either an apple or a banana. An experiment consists of recording the type of fruit given to the next four students.

- (a) Find the probability all four students get an apple.
- (b) Find the probability at least one student gets an apple.

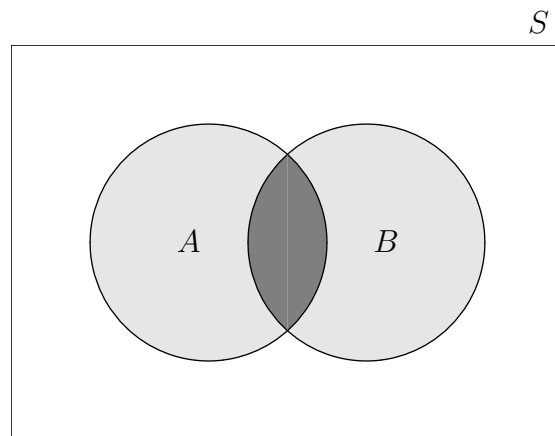
**Example 4.2.11** Suppose a customer at a newsstand is equally likely to purchase a newspaper, a magazine, or a concession candy. An experiment consists of recording the purchase of the next three customers.

- (a) Find the probability all three customers purchase different items.
- (b) Find the probability the second customer does not purchase a newspaper.
- (c) Find the probability at least one person buys a newspaper.

**The Addition Rule for Two Events**

1. For any two events  $A$  and  $B$ :  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .
2. For any two *disjoint* events  $A$  and  $B$ :  $P(A \cup B) = P(A) + P(B)$ .

Venn diagram for visualizing the Addition Rule

**Remarks**

1.  $P(A \cup B) = P(B \cup A)$ ; order doesn't matter.
2. For any three events  $A$ ,  $B$ , and  $C$ :

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

3. Let  $A_1, A_2, A_3, \dots, A_k$  be a collection of  $k$  *disjoint* events.

$$P(A_1 \cup A_2 \cup \dots \cup A_k) = P(A_1) + P(A_2) + \dots + P(A_k).$$

In words: if the events are disjoint, to find the probability of a union, add the corresponding probabilities.

**Example 4.2.12** Research by a travel agency indicates that 55% of all drivers carry a working flashlight in their car, 35% carry an emergency flare, and 20% carry both. Suppose a driver is selected at random.

- (a) Draw a Venn diagram to illustrate the events in this problem.
- (b) What is the probability the driver carries at least one of these two items?
- (c) What is the probability the driver carries neither?
- (d) What is the probability the driver carries just a flare?
- (e) What is the probability the driver carries just one of these two items?

**Example 4.2.13** A fisherman in California is required by law to have a valid fishing license and an approved life vest. Records from the Fish and Game Warden's Office indicate that 80% of all fishermen stopped for inspection have a valid fishing license, 60% have a U.S. Coast Guard-approved life vest, and 55% have both. Suppose a fisherman is randomly selected.

- (a) What is the probability the fisherman has at least one of these two requirements?
- (b) What is the probability the fisherman has neither of these two requirements and is therefore in violation of the law?
- (c) What is the probability the fisherman has only a valid license?
- (d) What is the probability the fisherman has exactly one of the two requirements?

**Example 4.2.14** Children attending a summer camp may participate in any one of six arts and crafts projects. The probability of selecting each project is given in the following table.

Craft	$C_1$	$C_2$	$C_3$	$C_4$	$C_5$	$C_6$
Probability	0.14	0.20	0.10	0.18	0.26	0.12

Find the probability of each of the following events.

- $A = \{C_1, C_3\}$  = jewelry crafts  
 $B = \{C_4, C_5, C_6\}$  = crafts requiring glue  
 $D = \{C_2, C_5\}$  = crafts using toothpicks  
 $E = \{C_2, C_6\}$  = painted crafts

Suppose a camper is randomly selected. Find the probability he/she

- made a craft requiring glue or paint.
- made a craft with toothpicks or paint,
- made a piece of jewelry, or used glue, or used paint.

## 4.3 Counting Techniques

1. In an equally likely outcome experiment:  $P(A) = \frac{N(A)}{N(S)}$
2. Certain rules can be used to count outcomes in an event and/or a sample space.

### The Multiplication Rule

Suppose an outcome in an experiment consists of an ordered list of  $k$  items selected using the following procedure:

1. There are  $n_1$  choices for the first item.
2. There are  $n_2$  choices for the second item, no matter which first item was selected.
3. The process continues until there are  $n_k$  choices for the  $k$ th item, regardless of the previous items selected.

There are  $N(S) = n_1 \cdot n_2 \cdot n_3 \cdots n_k$  outcomes in the sample space  $S$ .

### Remarks

1. You can picture this rule by drawing a tree diagram.
2. Think of each choice as a slot, or a position, to fill.

$$\begin{array}{ccccccc}
 & \text{Number of choices for each slot.} & & & & & \\
 & \downarrow & & \downarrow & & \cdots & \downarrow \\
 \frac{n_1}{\text{Item 1}} & \times & \frac{n_2}{\text{Item 2}} & \times & \cdots & \times & \frac{n_k}{\text{Item } k} \\
 & & & & & & = n_1 \cdot n_2 \cdots n_k
 \end{array}$$

**Example 4.3.1** A cell-phone account consists of a plan, a phone, and options packages. A local wireless phone company has 4 plans, 15 phones, and 5 options packages. How many possible cell-phone accounts can be created?



**Example 4.3.2** A computer login name consists of two letters (the user's initials) followed by five numbers.

- (a) How many different login names are possible?
- (b) How many login names begin with the letter K?

**Example 4.3.3** In the game of Risk, players try to occupy every territory on the board and thereby conquer the world. During a certain turn, suppose a total of four dice are rolled, three by the attacker and one by the defender.

- (a) How many different rolls are possible?
- (b) What is the probability two of the attacker's three dice will be sixes?

**Example 4.3.4** Every patient who comes to a certain city medical clinic must speak with a receptionist, a nurse, and a doctor. Suppose there are 3 receptionists, 10 nurses, and 5 doctors. An experiment consists of recording the next patient's *visit*: the receptionist, the nurse, and the doctor who speak with the patient.

- (a) How many different ways are there to visit this clinic?
- (b) Suppose there are two female doctors. What is the probability a visit includes a female doctor?
- (c) Nurse Krachet is known for her sour demeanor and terse remarks. What is the probability a visit will not include Nurse Krachet?

**Example 4.3.5** A DJ has just enough time to randomly select and play three songs from the top 10. An experiment consists of recording the selection list—the order of the songs played. For example, the outcome (4, 2, 9) means song 4 was selected first, then song 2, and finally song 9.

- (a) How many different selection lists are possible?
- (b) What is the probability the number-one hit is not played?
- (c) What is the probability song 2 or 3 will be played first?

**Definition**

For any positive whole number  $n$ , the symbol  $n!$  (read “ $n$  factorial”) is defined by

$$n! = n(n-1)(n-2) \cdots (3)(2)(1).$$

In addition,  $0! = 1$  (0 factorial is 1).

**Example 4.3.6** Find the following:

$$5! =$$

$$9! =$$

**Definition**

Given a collection of  $n$  different items, an ordered arrangement, or subset, of these items is called a **permutation**. The number of permutations of  $n$  items, taken  $r$  at a time, is given by

$${}_n P_r = n(n-1)(n-2) \cdots [n-(r-1)].$$

Using the definition of factorial,

$${}_n P_r = \frac{n!}{(n-r)!}$$

In the denominator, do the subtraction first, then the factorial.

**Remarks**

1. All  $n$  items must be different.
2. A distinguishing characteristic of a permutation: order matters.

**Example 4.3.7** A television sports announcer is preparing a short *teaser* for the upcoming full report at 11:00 pm. There are eight important sports stories to talk about, but he only has time to mention three in the teaser. How many different arrangements of sports stories are possible in the teaser?

**Example 4.3.8** A dessert stand at a county fair offers a special ice-cream cone with three different flavored scoops, stacked one on top of the other. There are twelve flavors to choose from.

- (a) How many different ice-cream cones are possible?
- (b) Suppose one of the possible flavors is chocolate. What is the probability this flavor will be in the middle?

**Definition**

Given a collection of  $n$  different items, an unordered arrangement, or subset, of these items is called a **combination**. The number of combinations of  $n$  items, taken  $r$  at a time, is given by

$${}_n C_r = \binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{{}_n P_r}{r!}.$$

**Remarks**

1. A distinguishing characteristic of a combination: order does not matter.
2. The final answer must be an integer. Why?

**Example 4.3.9** How many different ways are there to select 5 frequencies from a list of 15 for use in a police scanner?

**Example 4.3.10** A person leaving for vacation is going to select three paperback books at random from a stack of ten.

- (a) How many different combinations of books can the person select?
- (b) Suppose one of the ten books is a mystery. What is the probability none of the three books selected is the mystery novel?

**Example 4.3.11** The mayor's office has requested 4 limousines from a ride service for a special political event. Suppose there are 5 limousines in the fleet of 12 that need to be cleaned, and the manager of the ride service randomly selects the 4 limousines for the event.

- (a) What is the probability all four limousines selected will be clean?
- (b) What is the probability exactly one of limousines selected will be dirty?
- (c) What is the probability at most two of the limousines selected will be dirty?

## 4.4 Conditional Probability

1. Probability questions so far: *unconditional* probability.  
No special conditions imposed, nor any extra information given.
2. Sometimes two events are related so that the probability of one *depends* on whether the other has occurred.
3. In this case, knowing something extra may affect the probability assignment.

**Example 4.4.1** Consider the following events and probability questions.

$A$  = Tickets for the next home game of your favorite baseball team are sold out.

$B$  = Your favorite baseball team is in first place.

$C$  = Your favorite baseball team is in last place.

- (a) What is the probability tickets for the next home game of your favorite baseball team are sold out?

$$P(A) =$$

- (b) What is the probability tickets for the next home game of your favorite baseball team are sold out, given your favorite baseball team is in first place?

$$P(A | B) =$$

- (c) What is the probability tickets for the next home game of your favorite baseball team are sold out, given your favorite baseball team is in last place?

$$P(A | C) =$$

**Example 4.4.2** An experiment consists of selecting a single card at random from a regular 52-card deck and recording the denomination. Consider the following events.

$A$  = A jack is selected =  $\{J\}$

$B$  = A jack, queen, king, or ace is selected =  $\{J, Q, K, A\}$

$$P(A) = \qquad \qquad \qquad \text{(Unconditional probability.)}$$

What is the probability of selecting a jack, given I selected a jack, queen, king, or ace?

$$P(A | B) = \qquad \qquad \qquad \text{(Conditional probability, reduced sample space.)}$$

**Definition**

Suppose  $A$  and  $B$  are events with  $P(B) > 0$ . The **conditional probability of the event  $A$  given the event  $B$  has occurred**,  $P(A|B)$ , is

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

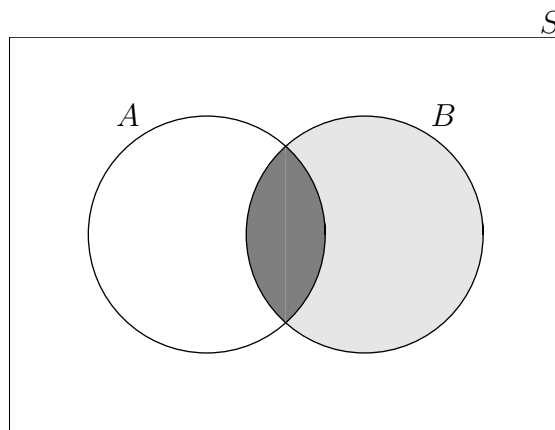
**Remarks**

1. The unconditional probability of an event  $A$  can be written as

$$P(A) = \frac{P(A)}{1} = \frac{P(A)}{P(S)} = \frac{\text{probability of the event } A}{\text{probability of the relevant sample space } S}$$

2. Given  $B$  has occurred, the relevant sample space has changed. It is *reduced* from  $S$  to  $B$ .

An illustration for calculating conditional probability:



3. Given  $B$  has occurred, the only way  $A$  can occur is if  $A \cap B$  has occurred, because the sample space has been reduced to  $B$ .
4.  $P(A|B)$  is the probability  $A$  has occurred,  $P(A \cap B)$ , divided by the probability of the relevant sample space,  $P(B)$ .



**Example 4.4.3** Use the definition of conditional probability to solve the previous problem involving selecting a card at random from a 52-card deck.

$$A = \text{A jack is selected} = \{J\}$$

$$B = \text{A jack, queen, king, or ace is selected} = \{J, Q, K, A\}$$

Find  $P(A | B)$ .

**Example 4.4.4** Which of the following equations is/are true? Justify your answer.

$$P(A \cup B) = P(B \cup A)$$

$$P(A \cap B) = P(B \cap A)$$

$$P(A | B) = P(B | A)$$

**Keywords** that suggest conditional probability: *given* and *suppose*.

**Example 4.4.5** A certain high school has a dance after every home football game. A survey at this high school indicates that 75% of all students plan to go to the football game and 35% plan to attend both the football game and the dance. Suppose a student at the high school is selected at random. If the student plans to attend the football game, what is the probability he/she plans to attend the dance after the game?

**Example 4.4.6** The manager of a shoe store asked 400 customers selected at random to identify the style of shoe they were shopping for. The results of this survey are given in the following two-way table.

		Shoe style			
		Athletic ( $A$ )	Casual ( $C$ )	Dress ( $D$ )	
Gender	Male ( $M$ )	90	75	30	195
	Female ( $F$ )	40	95	70	205
		130	170	100	400

Assume that these results are representative of all customers who shop at this store and that a customer is selected at random.

- Find the probability the customer is shopping for a dress shoe.
- Find the probability the customer is male.
- Suppose the customer selected is female, what is the probability she is shopping for a casual shoe?
- Suppose the customer is shopping for an athletic shoe, what is the probability the customer is male?

**Example 4.4.7** A psychologist has designed a questionnaire to identify the personality type of an individual. Each subject is randomly selected, asked to select a writing implement, and asked to complete the questionnaire. After extensive research, the psychologist has compiled the following joint probability table.

		Writing implement			
		Mechanical pencil ( $M$ )	Ballpoint pen ( $B$ )	Liquid ink pen ( $L$ )	
Personality type	Nurturer ( $N$ )	0.20	0.08	0.10	0.38
	Artist ( $A$ )	0.06	0.12	0.18	0.36
	Visionary ( $V$ )	0.17	0.05	0.04	0.26
		0.43	0.25	0.32	1.00

Suppose a subject is selected at random.

- Find the probability the subject is classified as an artist and uses a liquid ink pen.
- Suppose the subject is classified as a visionary, what is the probability the subject selected a mechanical pencil?
- Suppose the subject selected a ballpoint pen, what is the probability the subject is classified as a nurturer?

**Example 4.4.8** A greenhouse sells geraniums with blossoms in four colors and three pot sizes. The manager has compiled the following joint probability table for customers who purchase a geranium.

		Color				
		White	Red	Pink	Blue	
Pot diameter (inches)	5	0.07	0.05	0.07	0.05	0.24
	9	0.09	0.12	0.14	0.02	0.37
	12	0.12	0.17	0.06	0.04	0.39
		0.28	0.34	0.27	0.11	1.00

Suppose a customer who purchases a geranium is randomly selected.

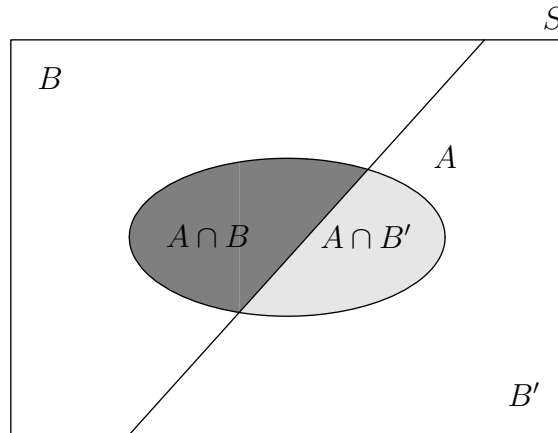
- What is the probability the blossoms are blue and the pot size is 9 inches?
- Suppose the customer purchases a geranium in a 12-inch pot, what is the probability the blossoms are red?
- Suppose the customer does not purchase a geranium with blue blossoms, what is the probability the pot size is 5 inches?
- Suppose the customer purchases a plant with red or pink blossoms, what is the probability the pot size is 9 inches?

**Remarks**

1. In general, for any two events  $A$  and  $B$ :

$$P(A) = P(A \cap B) + P(A \cap B')$$

The following Venn diagram illustrates this decomposition equation:



2. Suppose  $B_1$ ,  $B_2$ , and  $B_3$  are mutually exclusive and *exhaustive*:  $B_1 \cup B_2 \cup B_3 = S$ .  
For any other event  $A$ :

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

Venn diagram showing three mutually exclusive, exhaustive events and the decomposition of the event  $A$ :



---

## 4.5 Independence

1. Knowing extra information *may* change a probability assignment.
2. Additional information may have no effect on a probability assignment.

**Example 4.5.1** An experiment consists of tossing a fair coin and recording either a head or a tail. Suppose five heads in a row are recorded. What is the probability of a head on the sixth coin toss?

**Example 4.5.2** Suppose any letter delivered by the Postal Service addressed to *Resident* is classified as junk mail. Research has shown that receiving junk mail is not related to type of residence, for example, home, apartment, mobile home, or condominium.

Let the event  $J$  = receive junk mail. Suppose the unconditional probability a person receives junk mail on any given day is 0.35.

Suppose a person lives in an apartment. The research suggests the dwelling type has no effect on receiving junk mail. What is the probability the person receives junk mail today, given the event  $A$  = they live in an apartment?

$$P(J | A) = P(J) = 0.35$$

Knowing extra information does not change the probability assignment. Intuitively, these two events are *independent*.

### Remark:

If the occurrence or nonoccurrence of one event has no effect on the occurrence of the other, the two events are *independent*.

### Definition

Two events  $A$  and  $B$  are **independent** if and only if

$$P(A | B) = P(A).$$

If  $A$  and  $B$  are *not* independent, they are said to be **dependent** events.

**Remarks**

1. If the events  $A$  and  $B$  are independent, then

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B).$$

If either one of these equations is true, then the other is also true, and the events are independent.

2. If  $A$  and  $B$  are independent events, then so are all combinations of these two events and their complements.

Mathematical translation: If  $P(A|B) = P(A)$  then

$$P(A|B') = P(A), \quad P(A'|B) = P(A'), \quad \text{and} \quad P(A'|B') = P(A').$$

3. Independent events can *not* be shown on a Venn diagram.
4. Disjoint events are **dependent**.

**The Probability Multiplication Rule**

For any two events  $A$  and  $B$ :

$$\left. \begin{aligned} P(A \cap B) &= P(B) \cdot P(A|B) \\ &= P(A) \cdot P(B|A) \end{aligned} \right\} \text{Always true.}$$

$$= P(A) \cdot P(B) \quad \text{Only true if } A \text{ and } B \text{ are independent.}$$

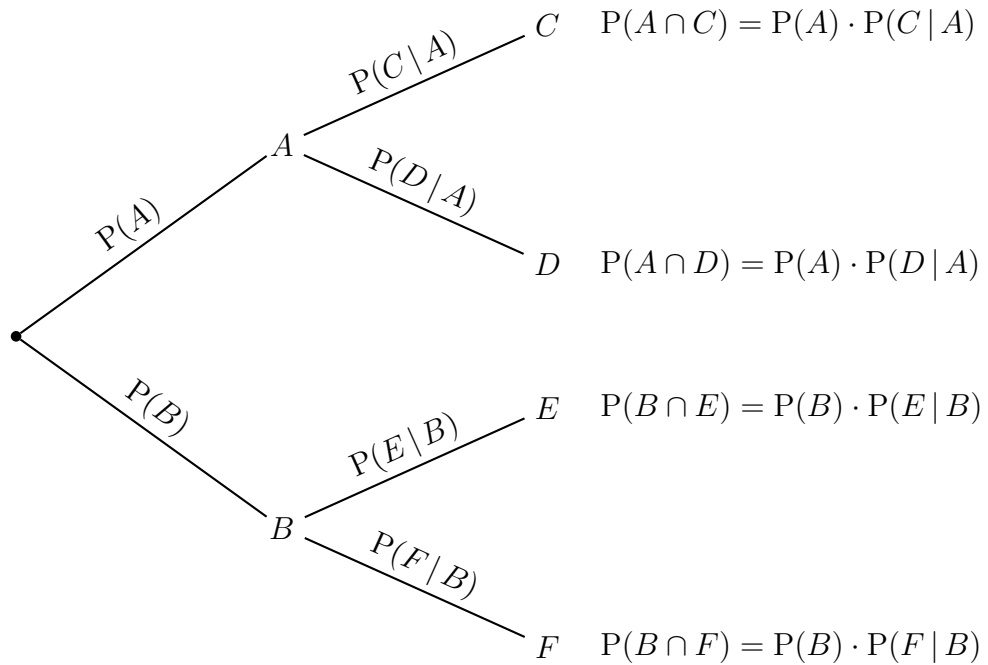
**Remarks**

1. Real skill: knowing which equality to use.

The first two equalities are always true.

The third equality is true only if  $A$  and  $B$  are independent.

2. If events are dependent, a modified tree diagram can be used to apply the Probability Multiplication Rule:



3. The Probability Multiplication Rule can be extended.

For any three events  $A$ ,  $B$ , and  $C$ :

$$P(A \cap B \cap C) = P(A) \cdot P(B | A) \cdot P(C | A \cap B).$$

4. If the events  $A_1, A_2, \dots, A_k$  are mutually independent:

$$P(A_1 \cap A_2 \cap \dots \cap A_k) = P(A_1) \cdot P(A_2) \cdot \dots \cdot P(A_k).$$



**Example 4.5.3** Census data and circulation information indicate that the probability a randomly selected adult 30–39 years old subscribes to *Time* magazine is 0.07. If an adult in this age group subscribes to *Time*, the probability he/she subscribes to *People* is 0.18. Suppose an adult in this age group is randomly selected. What is the probability he/she subscribes to *Time* and *People*?

**Example 4.5.4** A recent audit by the Justice Department indicate that only  $\frac{2}{3}$  of all foreign-language intercepts were translated within 12 hours. Suppose two foreign-language intercepts are selected at random.

- (a) What is the probability both intercepts will be translated within 12 hours?
- (b) What is the probability both intercepts will not be translated within 12 hours?
- (c) What is the probability exactly one intercept will be translated within 12 hours?

**Example 4.5.5** Research indicates that 25% of all people who participated in a high-school marching band experience some form of hearing loss by the age of 40. Suppose three people who were in a high-school marching band are selected at random.

- (a) What is the probability all three will experience hearing loss by the age of 40?
- (b) What is the probability none of the three will experience hearing loss by the age of 40?
- (c) Find the probability exactly one of the three will experience hearing loss by the age of 40.

**Example 4.5.6** Most hospitals still enforce a ban on cell-phones due to the potential interference with delicate medical equipment. Suppose the probability of a hospital in the United States enforcing this ban is 0.915. Four U.S. hospitals are selected at random.

- (a) What is the probability all four hospitals have a ban on cell-phones?
- (b) What is the probability exactly one hospital has a ban on cell-phones?
- (c) What is the probability at least one hospital has a ban on cell-phones?

**Example 4.5.7** Three carpenters are employed by the same contractor and work as needed. The probability that Mike works more than 30 hours per week is 0.85, for Robbie it is 0.70, and for Chip 0.65. Suppose work hours are determined independently, and a week is selected at random.

- (a) Find the probability all three carpenters work more than 30 hours during the week.
- (b) Find the probability exactly one of the three carpenters works more than 30 hours during the week.
- (c) Find the probability at least one of the three carpenters works more than 30 hours during the week.

**Example 4.5.8** A traveling salesman routinely takes a morning flight from New York to Washington, DC. The probability he flies on Continental is 0.45, on Delta it is 0.35, and on United 0.20. The salesman prefers an aisle seat, but cannot always get one. Ten percent of the time he is able to sit on the aisle while flying Continental, 12% of the time with Delta, and 25% of the time with United. Suppose the salesman travels from New York to Washington, DC.

- (a) Find the probability the salesman flies on United and gets an aisle seat.
- (b) Find the probability the salesman gets an aisle seat.
- (c) Suppose the salesman got an aisle set, what is the probability he is flying on United?



## CHAPTER 5

# Random Variables and Discrete Probability Distributions

---

## 5.0 Introduction

1. The outcomes in an experiment may not be numerical.
2. In order to analyze an experiment, to perform statistical inference, must convert each outcome to a number.
3. Each outcome is *assigned* a number.

### Random variable

1. The bridge between the experimenter's world and the statistician's world.
  2. The connection between experimental outcomes and the number associated with each outcome.
- 

## 5.1 Random Variables

Recall: a function  $f$  is a rule that takes an input value and computes an output value.

Consider  $f(x) = 3x - 5$ . Find  $f(2)$  and  $f(10)$ .

**Definition**

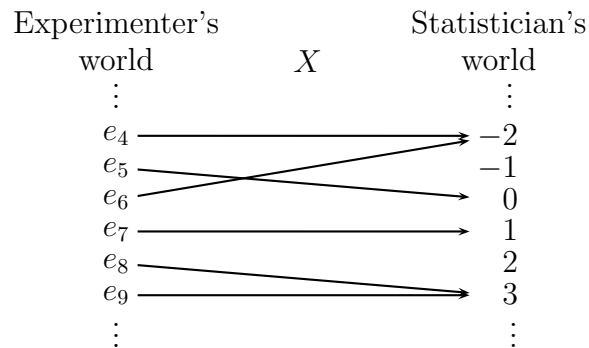
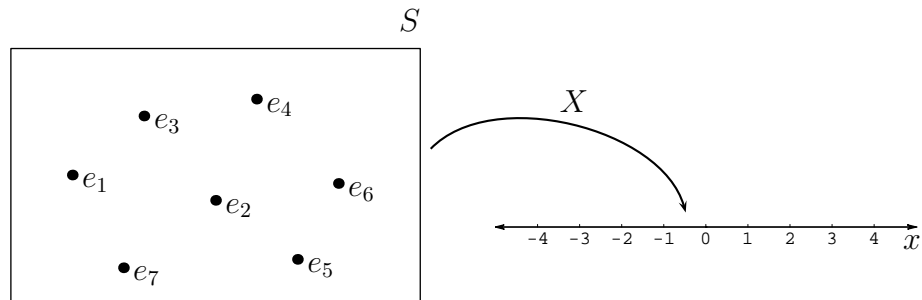
A **random variable** is a function that assigns a unique numerical value to each outcome in a sample space.

**Remarks**

1. Called random variable because the value can not be predicted with certainty.
2. Capital letters, like  $X$  and  $Y$ , are used to represent random variables.
3. A rule for assigning each outcome in a sample space to a unique real number.

$$X: S \rightarrow \mathbf{R}$$

4. Suppose  $e$  is an experimental outcome and  $x$  is a real number:  $X(e) = x$ .
5. Some visualizations of the definition of a random variable:





**Example 5.1.1** Since 1980, there has been a dramatic rise in the number of premature births in the United States. Some causes for this alarming trend include the increasing number of multiple births, stress, and diet. Suppose three births at a local hospital are selected at random, and each is classified as normal (N) or premature (E, for early). Let the random variable  $X$  be the number of premature births.

- (a) Find the sample space for this experiment.
- (b) Find the value of the random variable associated with each outcome.
- (c) Describe how to find the probability the random variable  $X$  takes on the value 2.

**Definition**

A random variable is **discrete** if the set of all possible values is finite, or countably infinite.

A random variable is **continuous** if the set of all possible values is an interval of numbers.

Discrete random variable: counting; finite or countably infinite number of values.

Continuous random variable: measuring; any interval of possible values.

In theory, a continuous random variable may take on any value in some interval, but not in reality.

**Example 5.1.2** Consider each experiment and determine whether the associated random variable is discrete or continuous.

- (a) Let the random variable  $X$  be the amount of carbohydrates in an 8-ounce glass of orange juice.
  
  
  
  
  
  
  
  
  
  
- (b) A stock broker is going to begin calling clients in order to find someone interested in buying shares in a small company. Let  $Y$  be the number of clients he must call until a buyer is found.
  
  
  
  
  
  
  
  
  
  
- (c) Let  $V$  be the number of people standing on a randomly selected city bus.

- (d) An experiment consists of selecting 20 sports bars in Boston and recording whether each is smoke-free or not smoke-free. Let the random variable  $W$  be the number of smoke-free bars.
- (e) A chain of hotels recently started to offer in-room Internet service. Let the random variable  $X$  be the amount of time connected to the Internet by a randomly selected guest.
- (f) Let  $Y$  be the number of new jobs created in the United States during a randomly selected month.
- (g) Let  $W$  be the amount of time a randomly selected NASA astronaut has spent in space.

---

## 5.2 Probability Distributions for Discrete Random Variables

1. Complete description of a discrete random variable includes:
  - (a) all the values the random variable can take on;
  - (b) all the associated probabilities.
2. An outcome and its probability: both associated with the same value of the random variable.

This connection determines probability assignments for a random variable.

### Definition

The **probability distribution for a discrete random variable  $X$**  is a method for specifying *all* of the possible values of  $X$  and the probability associated with each value.

### Remarks

1. Probability distribution for a discrete random variable: an itemized listing, a table, a graph, or a function.
2. Probability mass function (pmf), denoted  $p$ : probability that a discrete random variable is equal to some specific value.

In symbols,  $p(x) = \underbrace{P(X = x)}_{\text{Rule}}$ .

The function  $p$  and its probability rule are used interchangeably.

Suppose  $X$  is a discrete random variable.  $p(7) = P(X = 7)$ .

**Example 5.2.1** This example illustrates various methods for representing a probability distribution for a discrete random variable  $Y$ . A listing of possible values and probabilities is given below.

$$P(Y = 1) = \frac{1}{17} \quad P(Y = 2) = \frac{4}{17} \quad P(Y = 3) = \frac{7}{17} \quad P(Y = 4) = \frac{4}{17} \quad P(Y = 5) = \frac{1}{17}$$

- (a) Give the probability distribution as a table of values and probabilities.
- (b) Construct a probability histogram.
- (c) Construct a point representation.
- (d) Find a formula for the probability mass function.

Example (continued)

**Remarks**

1. Constructing a probability distribution:

To find the probability that  $X$  takes on the value  $x$ , look back at the experiment, and find all the outcomes that are mapped to  $x$ .

*Drag along* these probabilities and sum them.

2. The probability distribution for a random variable  $X$  is a reference for use in answering probability questions about the random variable.

In the expression  $P(X = 2)$ , think of  $X = 2$  as an *event* stated in terms of a random variable.

**Example 5.2.2** Suppose an experiment has 12 possible outcomes, each denoted by a sequence of 3 letters, each an M, O, or F. The probability of each outcome is given in the following table.

Outcome	MOM	MOO	MOF	MFM	MFO	MFF	OOM	OOO	OOF	OFM	OFO	OFF
Probability	0.105	0.085	0.070	0.075	0.055	0.080	0.072	0.090	0.123	0.045	0.090	0.110

The random variable  $X$  is defined to be the number of O's in an outcome. Find the probability distribution for  $X$ .

**Example 5.2.3** Suppose research suggests that a cup of kava tea one hour before bedtime will induce dreams in 40% of all people. Suppose a group of four adults are selected at random, and each has a cup of kava tea one hour before bedtime. Each person is recorded as having a dream (D) or no dream (N). Let the random variable  $X$  be the number of people who experience a dream.

- Find the probability distribution for  $X$ .
- Find the probability at most two people experience a dream.
- Suppose a second group of four people is randomly selected, and each person also has a cup of kava tea one hour before bedtime. What is the probability exactly one person in each group experiences a dream?

Outcome	Probability	$X$ Value	Outcome	Probability	$X$ Value
NNNN	0.1296		NDDN	0.0576	
NNND	0.0864		DNDN	0.0576	
NNDN	0.0864		DDNN	0.0576	
NDNN	0.0864		NDDD	0.0384	
DNNN	0.0864		DNDD	0.0384	
NNDD	0.0576		DDND	0.0384	
NDND	0.0576		DDDN	0.0384	
DNND	0.0576		DDDD	0.0256	



**Example 5.2.4** The manufacturer of a new drug claims that it relieves headaches within one hour in 90% of all patients. Suppose three people suffering from a headache are selected at random. Each is given the new medication and their condition after one hour is recorded. Let  $Y$  be the number of people who experience relief from their headache.

- (a) Find the probability distribution for  $Y$ .
- (b) Suppose none of the three people experience relief from their headache. Do you believe the manufacturer's claim? Justify your answer.

**Properties of a Valid Probability Distribution for a Discrete Random Variable**

1.  $0 \leq p(x) \leq 1$

The probability that  $X$  takes on any value,  $p(x) = P(X = x)$ , must be between 0 and 1.

2.  $\sum_{\text{all } x} p(x) = 1$

The sum of all the probabilities in a probability distribution for a discrete random variable must equal 1.

**Example 5.2.5** A random variable  $X$  has the following probability distribution:

$x$	10	20	30	40	50	60	70
$p(x)$	0.15	0.18	0.08	0.28	0.16	?	0.10

- (a) Find  $p(60)$ .
- (b) Find  $P(40 \leq X \leq 70)$  and  $P(40 < X \leq 70)$ .
- (c) Find  $P(X = 10 | X \leq 40)$ .

---

## 5.3 Mean, Variance, and Standard Deviation for a Discrete Random Variable

1. Descriptive measures of a sample:  $\bar{x}$ ,  $s^2$ ,  $s$ .

Descriptive measures of a population:  $\mu$ ,  $\sigma^2$ ,  $\sigma$ .

2. A random variable can model a population.

The descriptive measures of the population are conveyed by the probability distribution.

3. Learn methods for computing the mean, variance, and standard deviation of a discrete random variable.

**Example 5.3.1** The first morning train leaves a certain Metro station for Washington, DC, at 4:30 a.m. On five days (Monday–Friday), there are exactly 15 people who board the train at this station. On the remaining two days, there are exactly 8 people who board the train at this station. How many people board this train at this station per day on *average*? Or, in the long run, how many people board this train at this station each day?

**Definition**

Let  $X$  be a discrete random variable with probability mass function  $p(x)$ . The **mean**, or **expected value**, of  $X$  is

$$\underbrace{E(X) = \mu = \mu_X}_{\text{Notation}} = \underbrace{\sum_{\text{all } x} [x \cdot p(x)]}_{\text{Calculation}}.$$

**Remarks**

1. The capital E stands for *expected value*, a function.

E accepts as an input any *function* of a random variable.

Suppose  $f(X)$  is a function of a discrete random variable  $X$ .

The expected value of  $f(X)$  is  $E[f(X)] = \sum_{\text{all } x} [f(x) \cdot p(x)]$ .

2.  $\mu$  is the mean, or expected value, of a random variable (population).

If necessary, use a subscript for identification, for example  $\mu_X$  or  $\mu_Y$ .

3. Computation: Multiply each value of the random variable by its corresponding probability, and add the products.
4. The mean of a random variable is a *weighted average* and is only what happens on average. The mean may not be any of the possible values of the random variable.

**Example 5.3.2** Suppose  $X$  is a discrete random variable with probability distribution given in the following table.

$x$	1	3	6	10
$p(x)$	0.3	0.2	0.4	0.1

Find the mean of  $X$ .

**Example 5.3.3** A certain car audio system has six preset buttons such that the user can assign one radio station to each button. Suppose  $X$  is a random variable that represents the number of preset buttons used, or programmed, by the listener. Extensive research was used to construct the probability distribution for  $X$  in the following table.

$x$	0	1	2	3	4	5	6
$p(x)$	0.05	0.19	0.21	0.35	0.12	0.05	0.03

Find the expected number of preset buttons programmed.

**Example 5.3.4** A farm store sells striped sunflower seeds in 5, 10, 25, 50, and 100-pound bags. Let the random variable  $Y$  be the number of pounds of the bag purchased by the next customer. Store records over the past two years were used to construct the probability distribution of  $Y$  given in the following table.

$y$	5	10	25	50	100
$p(y)$	0.15	0.35	0.25	0.20	0.05

Find the expected number of pounds of sunflower seed purchased.

**Definition**

Let  $X$  be a discrete random variable with probability mass function  $p(x)$ . The **variance** of  $X$  is

$$\underbrace{\text{Var}(X) = \sigma^2 = \sigma_X^2}_{\text{Notation}} = \underbrace{\sum_{\text{all } x} [(x - \mu)^2 \cdot p(x)]}_{\text{Calculation}} = \underbrace{\text{E}[(X - \mu)^2]}_{\text{Definition in terms of expected value.}}$$

The **standard deviation** of  $X$  is the positive square root of the variance:

$$\underbrace{\sigma = \sigma_X}_{\text{Notation}} = \underbrace{\sqrt{\sigma^2}}_{\text{Calculation}}$$

**Remarks**

1. In words: variance is the expected value of the *squared deviations about the mean*.
2. The symbol Var stands for variance, a function.
3. To calculate the variance using the definition:
  - (a) Find the mean,  $\mu$ , of  $X$ .
  - (b) Find each difference:  $(x - \mu)$ .
  - (c) Square each difference:  $(x - \mu)^2$ .
  - (d) Multiply each squared difference by the associated probability.
  - (e) Sum the products.

**Computational Formula for  $\sigma^2$** 

$$\sigma^2 = \text{E}(X^2) - \text{E}(X)^2 = \text{E}(X^2) - \mu^2.$$

**Example 5.3.5** Suppose  $X$  is a discrete random variable with probability distribution given in the following table.

$x$	2	3	5	8	13
$p(x)$	0.05	0.20	0.20	0.25	0.30

- (a) Find the expected value, variance, and standard deviation of  $X$ .  
Use the definition of the variance, and verify the result using the computational formula.
- (b) Find the probability the random variable  $X$  takes on a value less than one standard deviation from the mean.

Example (continued)



**Remarks**

1. Computational formula for variance: quicker, produces less round-off error.
2. Use the Empirical Rule only when the distribution is (approximately) normal.  
     Don't use Chebyshev's Rule if you know the exact distribution.
3. Technology: no built-in functions to compute the mean, variance, and standard deviation.

**Example 5.3.6** Sophisticated bread machines make it easy to produce freshly baked bread with the touch of a button. For most recipes, you simply combine all of the ingredients, select the appropriate settings, and the machine does the rest: kneading, rising, and baking. Let  $X$  be the amount of yeast (in teaspoons) for a randomly selected bread recipe in the booklet that accompanies a certain machine. The probability distribution for  $X$  is given in the following table.

$x$	0.25	0.50	1.00	1.50	2.00
$p(x)$	0.075	0.155	0.675	0.090	0.005

Find the expected amount of yeast required in a bread recipe, and the variance and the standard deviation of the amount of yeast required.

**Example 5.3.7** The owner of a bed and breakfast has three unique rooms for rent. Let  $X$  be the number of rooms booked on a randomly selected night. The probability distribution for  $X$  is given in the following table.

$x$	0	1	2	3
$p(x)$	0.05	0.20	0.60	0.15

- Find the mean, variance, and standard deviation of  $X$ .
- Suppose each room rents for \$125 per night. Find the expected revenue, and the variance and standard deviation of the revenue.

---

## 5.4 The Binomial Distribution

1. Binomial random variable: common, used to model many real-world populations, used in inference.
2. Binomial random variable is related to a certain kind of experiment.

Consider the following experiments and look for similarities.

- (a) Select six soft drinks, and check the inside of each cap to see whether you have won a prize. Count the number of instant winners.
- (b) On a cold winter day in North Dakota, select 100 cars at random. Count the number of cars that start on the first attempt.
- (c) Select 25 workers at a steel mill, and record whether or not each worked any overtime during the past week.
- (d) Five hundred students applying to a certain college are selected at random. Each is classified according to whether they took an Advanced Placement (AP) exam or not. Count the number of applicants who took an AP exam.

Common properties: use to describe a *binomial experiment*.

Binomial experiment leads to a *binomial random variable*.

### Properties of a Binomial Experiment

1. The experiment consists of  $n$  identical trials.
2. Each trial can result in only one of two possible (mutually exclusive) outcomes. One outcome is usually designated a success (S) and the other a failure (F).
3. The outcomes of the trials are independent.
4. The probability of a success,  $p$ , is constant from trial to trial.

**Remarks**

1. Trial: a small part of the larger experiment.

A trial results in a single occurrence of either a success or a failure.

2. A success does not have to be a good thing.

Example: a success might be that an animal has mad cow disease.

3. Independent trials: whatever happens on one trial has no effect on any other trial.

4. Probability of a success is the same on every trial.

5. In a binomial experiment: outcomes consist of a sequence of S's and F's.

Example: Suppose  $n = 10$ , a typical outcome: SSFFSFSFFF

**The Binomial Random Variable**

The **binomial random variable** maps each outcome in a binomial experiment to a real number, and is defined to be the *number of successes* in  $n$  trials.

**Notation:**

1.  $p$  = the probability of a success:  $P(S) = p$  and  $P(F) = 1 - p = q$ .
2. Binomial random variable  $X$ : *completely determined* by the number of trials  $n$  and the probability of a success  $p$ .

Shorthand notation:  $X \sim B(n, p)$

$X$  is (distributed as) a binomial random variable with  $n$  trials and probability of a success  $p$ .

Example,  $X \sim B(50, 0.25)$ :  $X$  is a binomial random variable with 50 trials and probability of success 0.25.

**Goals:**

1. Find the probability distribution for a binomial random variable.
2. If  $X \sim B(n, p)$ , find  $P(X = x) = p(x)$ .

Example: If  $X \sim B(10, 0.35)$ , find  $P(X = 7)$ .

**Example 5.4.1** Consider a binomial random variable with  $n = 4$  trials and probability of a success  $p$ .

- (a) A typical outcome: FSFS. Find the probability of this outcome.
- (b) Another possible outcome: FSSF. Find the probability of this outcome.
- (c) Compare the results in (a) and (b).
- (d) Find the probability of  $X = 2$  successes.

**Example 5.4.2** Suppose  $X \sim B(n, p)$ .

- Find the probability of a single outcome with  $x$  successes.
- Find the probability of obtaining  $x$  successes in  $n$  trials.

Recall:

- Factorial:  $n! = n(n-1)(n-2)\cdots(3)(2)(1)$  and  $0! = 1$ .

$$5! =$$

$$10! =$$

- Given a collection of  $n$  items, the number of combinations of size  $x$  is given by

$${}_n C_x = \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

$$\binom{7}{3} =$$

$$\binom{10}{5} =$$

Final piece: The number of outcomes with  $x$  successes is  $\binom{n}{x}$ .

**The Binomial Probability Distribution**

Suppose  $X$  is a binomial random variable with  $n$  trials and probability of a success  $p$ :  $X \sim B(n, p)$ . Then

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, 3, \dots, n.$$

Number of outcomes
Probability of  $x$  successes and  $n - x$   
with  $x$  successes
failures in any single outcome

**Example 5.4.3** The old tradition of throwing rice at a departing wedding couple has been replaced in many places by tossing birdseed or confetti, or ringing bells. Suppose, however, rice is still used at 20% of all weddings, and 10 weddings are selected at random.

- (a) Find the probability exactly four weddings use rice.
- (b) Find the probability at most three weddings use rice.

**Remarks**

1. Even for small  $n$ , computing these probabilities is tedious.

Table 1: cumulative probabilities for a binomial random variable.

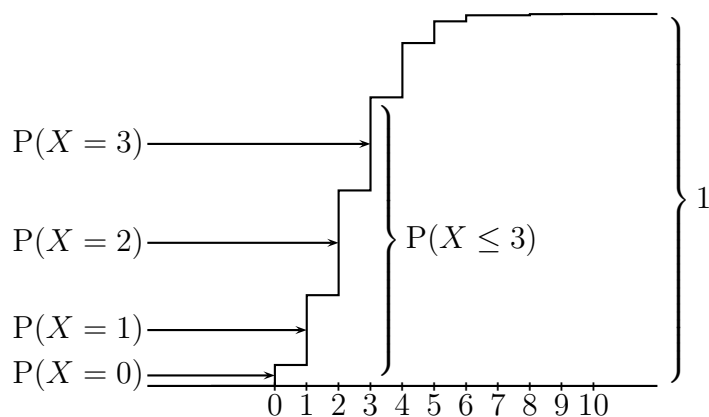
Technology: probability of a single value and cumulative probability.

2. Suppose  $X \sim B(n, p)$ .

Cumulative probability: the probability  $X$  takes on a value less than or equal to  $x$ .

$$\begin{aligned} P(X \leq x) &= \sum_{k=0}^x P(X = k) \\ &= P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = x) \end{aligned}$$

3. Illustration of cumulative probability:

**Note:**

1. Every probability question about a binomial random variable can be answered using cumulative probability.
2. There may be other, faster techniques.



**Example 5.4.4** Seventy-five percent of all professional painters prefer to use a roller for applying paint to a wall or ceiling. Suppose 15 professional painters are selected at random.

- (a) Find the probability at most 12 prefer a roller.
- (b) Find the probability exactly 10 prefer a roller.
- (c) Find the probability at least 9 prefer a roller.
- (d) Find the probability between 8 and 13 (inclusive) prefer a roller.

**Example 5.4.5** According to the National Health and Nutrition Examination Survey, approximately 60% of Americans ages 65–74 have full or partial dentures. Suppose 30 Americans ages 65–74 are selected at random.

- (a) Find the probability more than 20 have full or partial dentures.
- (b) Find the probability at least 22 but less than 27 have full or partial dentures.
- (c) Find the probability exactly 10 have full or partial dentures.

**Example 5.4.6** A recent medical report claimed that 25% of all cardiac surgery patients received unnecessary units of blood. Suppose 30 cardiac surgery patients are selected at random, and the number of patients who received unnecessary units of blood is recorded.

- (a) Find the probability that at least 10 patients received unnecessary units of blood.
- (b) Suppose two patients received unnecessary units of blood. Is there any evidence to suggest that the claim is wrong? Justify your answer.

**Remarks**

1. Random variable: often described by its mean and variance (or standard deviation),  $\mu$ ,  $\sigma^2$ , (or  $\sigma$ ).
2. If we know  $\mu$  and  $\sigma$ , we know the most likely values the random variable takes on.

Useful for inference. We can determine the likelihood of an experimental outcome.

3. Mean and variance of a binomial random variable: mathematical definitions.

Intuition: suppose  $X \sim B(50, 0.5)$   $\mu =$

**Definition**

If  $X$  is a binomial random variable with  $n$  trials and probability of a success  $p$ ,  $X \sim B(n, p)$ , then

$$\mu = np, \quad \sigma^2 = np(1 - p), \quad \text{and} \quad \sigma = \sqrt{np(1 - p)}.$$

**Example 5.4.7** A consumer group claims that 60% of all new-car purchases are made by women. Suppose 50 new-car purchases are selected at random.

- (a) Find the mean, variance, and standard deviation of the number of new-car purchases made by women.
- (b) Find the probability the number of new-car purchases made by women is more than two standard deviations from the mean.

**Example 5.4.8** A food manufacturer claims that 25% of all Americans are supertasters, i.e., those who perceive sweet, sour, bitter, and salty tastes more intensely than others. Suppose 100 Americans are selected at random, and the number of supertasters is recorded.

- (a) Find the mean, variance, and standard deviation of the number of supertasters.
- (b) Suppose 31 of the selected Americans are supertasters. Is there any evidence to suggest that the manufacturer's claim is false? Justify your answer.

---

## 5.5 Other Discrete Distributions

To solve problems involving these common discrete distributions:

1. Define a random variable and identify its probability distribution. (Distribution statement)
2. Translate the words into a probability question; the event is stated in terms of the random variable. (Probability statement)
3. Use cumulative probability if necessary.

1. Geometric distribution: Related to the binomial distribution.
2. Binomial distribution:  $n$  fixed, and the number of successes varies.
3. Geometric distribution: number of successes fixed at 1, number of trials varies.

### Properties of a Geometric Experiment

1. The experiment consists of identical trials.
2. Each trial can result in only one of two possible outcomes: a success (S) or a failure (F).
3. The trials are independent.
4. The probability of a success,  $p$ , is constant from trial to trial.

**Note:** The experiment ends when the first success is obtained.

### The Geometric Random Variable

The **geometric random variable** is the number of trials necessary to realize the first success.

**Example 5.5.1** Let  $X$  be a geometric random variable with probability of a success  $p$ . Find  $P(X = x)$ .

**The Geometric Probability Distribution**

Suppose  $X$  is a geometric random variable with probability of a success  $p$ . Then

$$p(x) = P(X = x) = (1 - p)^{x-1}p, \quad x = 1, 2, 3, \dots$$

$$\mu = \frac{1}{p} \quad \text{and} \quad \sigma^2 = \frac{1-p}{p^2}.$$

**Remarks**

1. The geometric random variable is discrete.

The number of possible values is countably infinite:  $1, 2, 3, \dots$

2. The geometric distribution is completely characterized, or defined, by one parameter,  $p$ .
3. Do not need a table to find cumulative probabilities associated with a geometric random variable.

$$P(X \leq x) = 1 - (1 - p)^x$$

Example: Suppose  $X$  is a geometric random variable with  $p = 0.35$ . Find  $P(X \leq 10)$ .

4. This is a valid probability distribution.

Each probability is between 0 and 1.

Sum of all the probabilities is an *infinite series*.

The sum 
$$\sum_{x=1}^{\infty} P(X = x) = \sum_{x=1}^{\infty} (1 - p)^{x-1}p$$

is called a *geometric series* and it does sum to 1!



**Example 5.5.2** The manager of a Starbucks claims that 4 out of 10 people purchase a pastry with their cup of coffee. Suppose the manager selects customers who purchase coffee at random and checks to see whether they purchase a pastry or just coffee.

- (a) What is the probability the fifth coffee buyer will be the first to also purchase a pastry?
- (b) What is the probability it will take at most four coffee buyers before one person purchases a pastry?
- (c) What is the probability it will take at least 10 coffee buyers before one person purchases a pastry?

**Poisson Distribution:**

1. Often associated with rare events.
2. A count of the number of occurrences of a certain event in a given unit of time, space, volume, distance, etc.
3. Properties: Poisson process, difficult to verify.

**Properties of a Poisson Experiment**

1. The probability that a specific event occurs in a given interval (of time, length, volume, etc.) is the same for all intervals.
2. The number of events that occur in any interval is independent of the number that occur in any other interval.

**The Poisson Random Variable**

The **Poisson random variable** is a count of the number of times the specific event occurs during a given interval.

**The Poisson Probability Distribution**

Suppose  $X$  is a Poisson random variable with mean  $\lambda$ . Then

$$p(x) = P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, 3, \dots$$

$$\mu = \lambda, \quad \text{and} \quad \sigma^2 = \lambda.$$

**Remarks**

1. The Poisson random variable is discrete: number of possible values is countably infinite:  $0, 1, 2, 3, \dots$
2. Poisson distribution: completely characterized by only one parameter,  $\lambda$  (usually small, for rare events).

The mean and the variance are both equal to the same value,  $\lambda$ .

3. All of the probabilities are between 0 and 1.

Sum of all the probabilities is 1 (another infinite series).

4.  $e \approx 2.71828$ : the base of the natural logarithm.
5. The denominator contains  $x!$  ( $x$  factorial).

Recall:  $x! = x(x-1)(x-2) \cdots (3)(2)(1)$  and  $0! = 1$ .

6. Table 2:  $P(X \leq x)$  (cumulative probability) for various values of  $\lambda$ .

**Example 5.5.3** During daylight, birds often see reflections (like trees) in windows of tall buildings, and fly into them. The artificial light at nighttime also attracts the birds. In Toronto, the mean number of bird deaths per day due to a collision with a skyscraper is four. Suppose a day is selected at random.

- (a) Find the probability exactly one bird will collide with a skyscraper and die.
- (b) Find the probability at most six birds will collide with a skyscraper and die.
- (c) Suppose at least three birds collide with a skyscraper and die. What is the probability at least seven birds collide with a skyscraper and die?

Example (continued)

**Hypergeometric Probability Distribution:**

1. Sampling without replacement from a finite population.
2. Each element in the population: a success or a failure.
3. Count the number of successes.

**Properties of a Hypergeometric Experiment**

1. The population consists of  $N$  objects, of which  $M$  are successes and  $N - M$  are failures.
2. A sample of  $n$  objects is selected *without* replacement.
3. Each sample of size  $n$  is equally likely.

### The Hypergeometric Random Variable

The **hypergeometric random variable** is a count of the number of successes in a random sample of size  $n$ .

### The Hypergeometric Probability Distribution

Suppose  $X$  is a hypergeometric random variable characterized by sample size  $n$ , population size  $N$ , and number of successes  $M$ . Then

$$p(x) = P(X = x) = \frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}, \quad \max(0, n - N + M) \leq x \leq \min(n, M)$$

$$\mu = n \frac{M}{N} \quad \text{and} \quad \sigma^2 = \left(\frac{N-n}{N-1}\right) n \frac{M}{N} \left(1 - \frac{M}{N}\right)$$

### Remarks

1. Restriction on the possible values for the random variable  $X$ :

$\max(0, n - N + M) \leq x$ :  $x$  must be at least 0 or  $n - N + M$ , whichever is bigger.

If  $n - N + M$  is positive, it is impossible to obtain fewer than  $n - N + M$  successes.

$x \leq \min(n, M)$ :  $x$  can be at most  $n$  or  $M$ , whichever is smaller.

The greatest number of successes possible is either  $n$  or the total number of successes in the population.

Suppose  $n = 7$ ,  $N = 18$ , and  $M = 8$ .

$$\max(0, n - N + M) =$$

$$\min(n, M) =$$

2. The hypergeometric random variable is discrete.

All of the probabilities are between 0 and 1.

Probabilities do sum to 1.

3. Recall:  ${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

**Example 5.5.4** Twelve out of the 20 residents in a small development subscribe to the local newspaper. Three of the 20 will be selected at random and asked to participate in a survey.

- (a) What is the probability exactly two of the three selected subscribe to the local newspaper?
- (b) What is the probability at most one of the three selected subscribes to the local newspaper?

## CHAPTER 6

# Continuous Probability Distributions

---

## 6.1 Introduction

Discrete probability distributions

1. Completely describe a discrete random variable.
2. Techniques for computing probabilities, cumulative probability.
3. Mean and variance tell us the most likely values of the random variable.

This chapter: similar methods for continuous random variables.

Continuous random variable

1. Usually associated with measurement.
2. Can take on any value in some interval.

---

## 6.2 Probability Distributions for a Continuous Random Variable

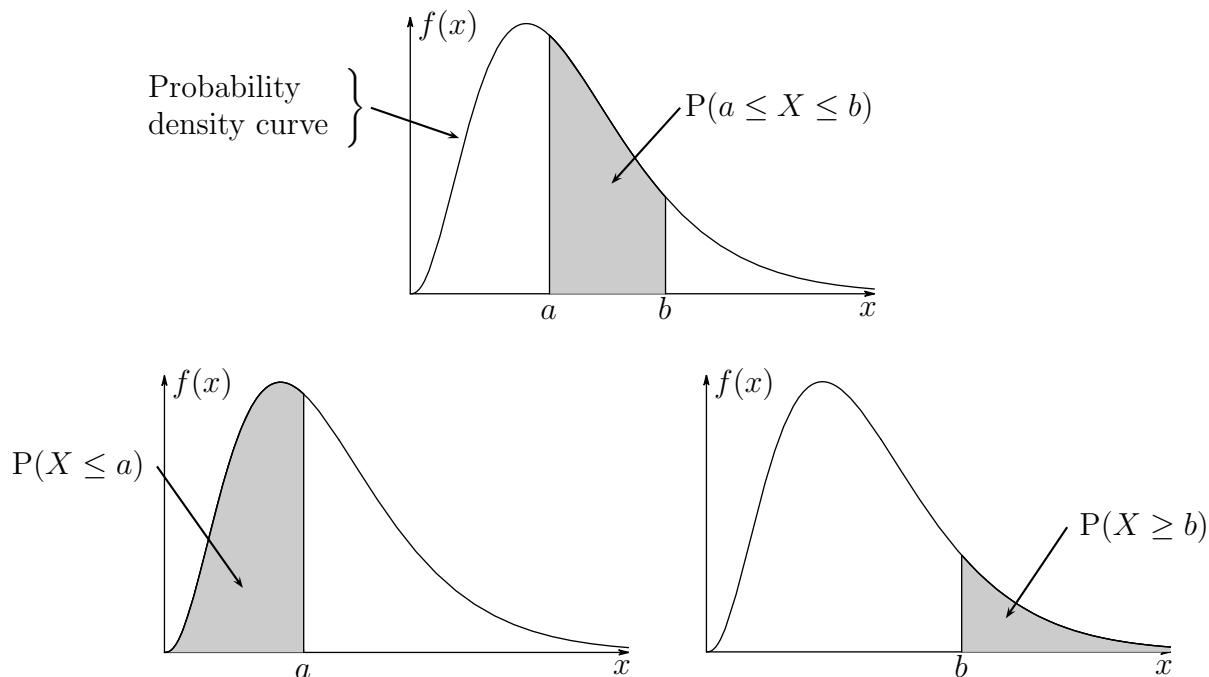
**Continuous probability distribution:** completely describes the random variable and is used to compute probabilities associated with the random variable.

### Definition

A **probability distribution for a continuous random variable  $X$**  is given by a smooth curve called a **density curve**, or **probability density function (pdf)**. The curve is defined so that the probability  $X$  takes on a value between  $a$  and  $b$  ( $a < b$ ) is the area under the curve between  $a$  and  $b$ .

**Remarks**

1. Probability in a continuous world is *area under a curve*.



2. The density curve, or probability density function: denoted by  $f$ .

It is a *function*, defined for *all* real numbers.

$f(x)$  is *not* the probability the random variable  $X$  equals the specific value  $x$ .

The function  $f$  leads to, or conveys, probability through area.

3. Shape of the graph of  $f$  can vary considerably.

Density function must satisfy the following two properties:

- (a)  $f$  must be defined so that the total area under the curve (from  $-\infty$  to  $\infty$ ) is 1.

$f(x)$ , a specific value of the density function, *may* be greater than 1 (while the total area under the curve is still exactly 1).

- (b)  $f(x) \geq 0$  for all  $x$ .

4.  $P(X = a) = 0$  for any  $a$ .

Since there is no probability associated with a single point, the following four probabilities are all the same:

$$P(a \leq X \leq b) = P(a < X \leq b) = P(a \leq X < b) = P(a < X < b).$$



5. To find *area under a curve*: calculus problem.

Common geometric figures: rectangle, triangle, and trapezoid.

Tables and technology for the regions with more complicated shapes.

### Definition

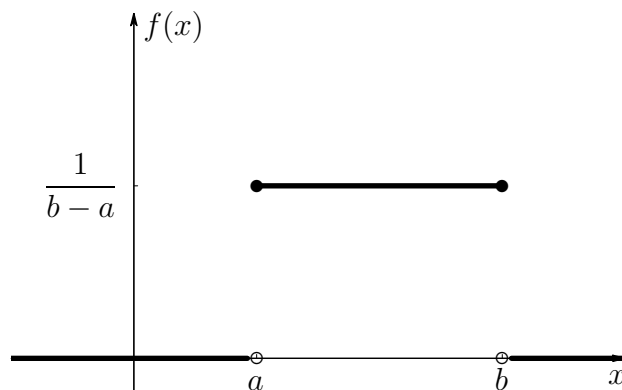
The random variable  $X$  has a **uniform distribution** on the interval  $[a, b]$  if

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \quad -\infty < a < b < \infty$$

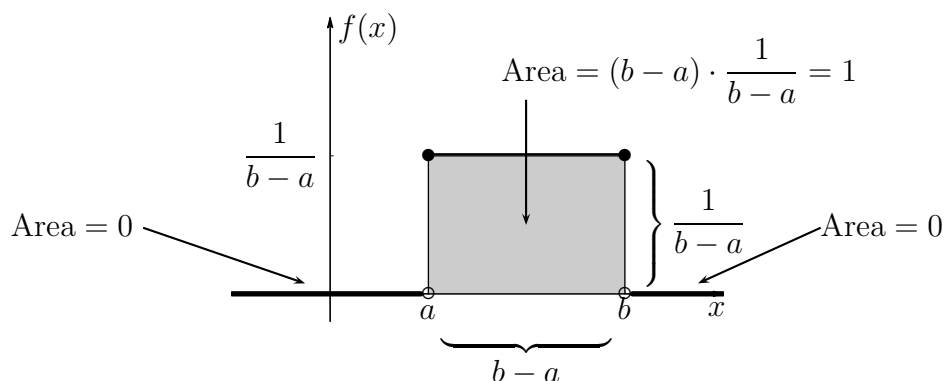
$$\mu = \frac{a+b}{2} \quad \text{and} \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

### Remarks

1.  $a, b$ : any real numbers,  $a < b$
2. A graph of the probability density function:



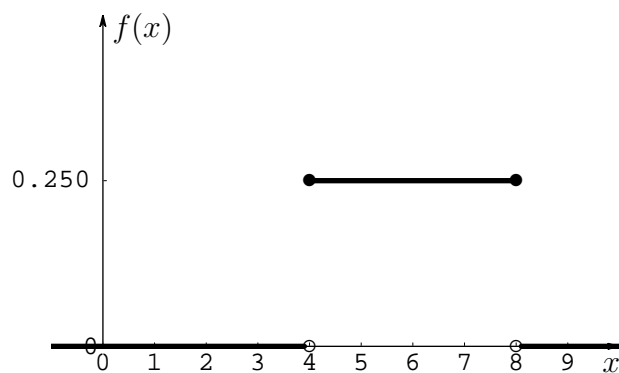
3. Valid probability density function:
  - (a)  $f(x) \geq 0$  for all  $x$ .
  - (b) The total area under the curve is 1.

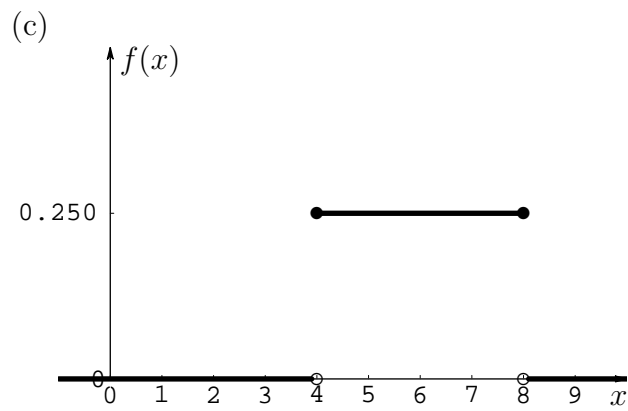
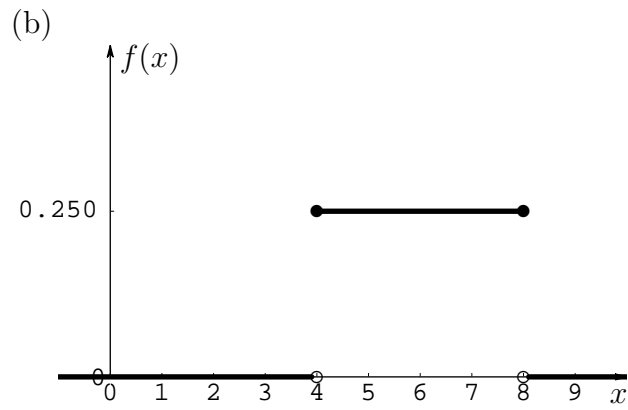


**Example 6.2.1** The time (in minutes) to preheat a gas oven to  $350^\circ\text{F}$  has a uniform distribution between 4 and 8. Suppose a gas oven is selected at random and set to preheat to  $350^\circ\text{F}$ .

- Carefully sketch a graph of the probability density function.
- Find the probability it takes at most  $5\frac{1}{2}$  minutes to preheat to  $350^\circ\text{F}$ .
- Find the probability it takes between 6 and 8 minutes to preheat to  $350^\circ\text{F}$ .
- Find the mean time it takes to preheat to  $350^\circ\text{F}$ , and the variance and standard deviation.

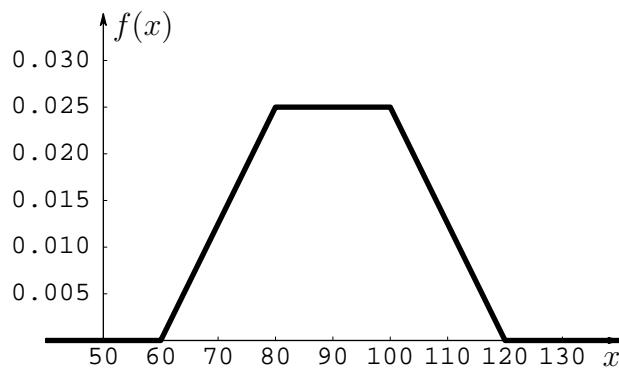
(a)





(d)

**Example 6.2.2** Amplifiers along cable television lines are used to compensate for the power loss in the transmission cable. The distance (in meters) between adjacent amplifiers is a random variable,  $X$ , with probability density function given below.



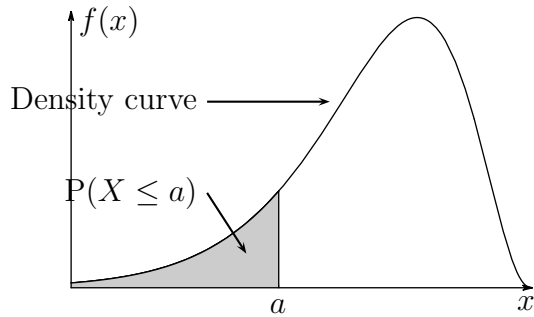
$$f(x) = \begin{cases} 0.00125x - 0.075 & \text{for } 60 \leq x < 80 \\ 0.025 & \text{for } 80 \leq x \leq 100 \\ -0.00125x + 0.150 & \text{for } 100 < x \leq 120 \\ 0 & \text{otherwise} \end{cases}$$

- Verify that  $f$  is a valid probability density function.
- Find the probability the distance between two adjacent amplifiers is less than 80 meters.
- Find the probability the distance between two adjacent amplifiers is greater than 95 meters.

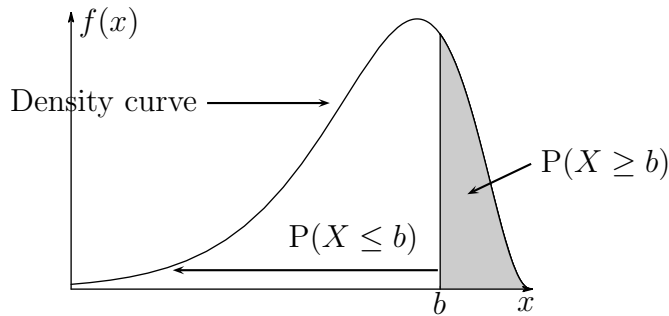
Cumulative probability:  $P(X \leq x)$

1. Accumulate probability up to and including a fixed value.
2. Find all the area under the density curve to the left of the fixed value.

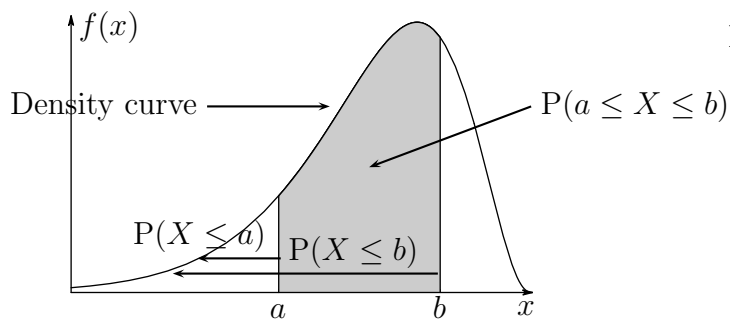
Using cumulative probability:



$$P(X \leq a)$$

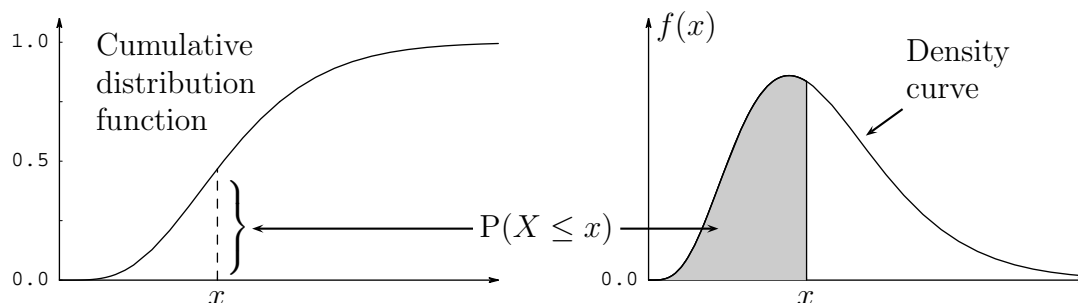


$$\begin{aligned} P(X \geq b) &= 1 - P(X < b) \\ &= 1 - \underbrace{P(X \leq b)}_{\text{cumulative probability}} \end{aligned}$$



$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X < a) \\ &= P(X \leq b) - P(X \leq a) \end{aligned}$$

Another way to illustrate cumulative probability:



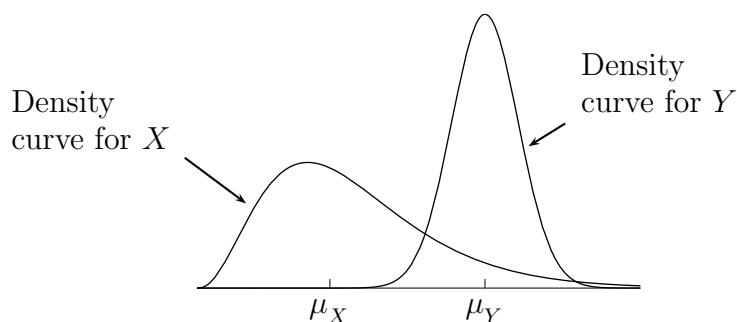
### Remarks

1. The drawing on the left is a graph of a *cumulative distribution function*.
2. The mean,  $\mu$ , and the variance,  $\sigma^2$ , for a continuous random variable are computed using calculus.

Interpretations are the same.

$\mu$ : a measure of the *center* of the distribution.

$\sigma^2$  (or  $\sigma$ ): a measure of the spread, or *variability*, of the distribution.



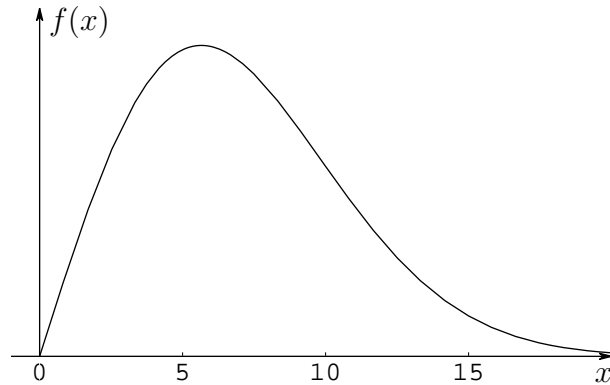
- (a)  $\mu_X < \mu_Y$ :

*Center* of the distribution of  $X$  is to the left of the *center* of the distribution of  $Y$ .

- (b)  $\sigma_X > \sigma_Y$ :

Distribution of  $X$  is more spread out, and thus has more variability, than the distribution of  $Y$ .

**Example 6.2.3** The owner's manual for a new Whirlpool washer contains a toll-free customer support number. An owner with a question about the operation of the washer can call this number and leave a message. A customer support technician will call the customer back as soon as possible. The customer support waiting time for a callback (in minutes) is a random variable,  $X$ , with probability density function shown in the figure below.



The cumulative probability is given by

$$P(X \leq x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-x^2/64} & x \geq 0 \end{cases}$$

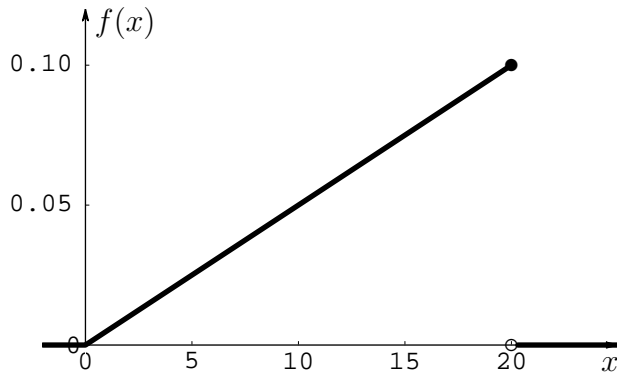
Suppose a customer call is randomly selected.

- (a) What is the probability the callback time is less than 5 minutes?
- (b) What is the probability the callback time is more than 15 minutes?
- (c) What is the probability the callback time is between 2 and 6 minutes?
- (d) Find the median callback time.

Example (continued)



**Example 6.2.4** A student in a certain veterinary college may borrow up to \$20,000 each year. The amount borrowed by a first-year student (in thousands of dollars) is a random variable,  $X$ , with probability density function given below.



$$f(x) = \begin{cases} 0.005x & \text{for } 0 \leq x \leq 20 \\ 0 & \text{otherwise} \end{cases}$$

Suppose a first-year student at this college is selected at random.

- Find the probability the student borrows less than \$10,000. Why isn't this answer 0.5?
- Find an expression for the cumulative distribution function.
- Find the amount  $d$  such that 90% of all first-year students borrow less than  $d$ .

## 6.3 The Normal Distribution

1. Normal distribution: very common and the most important distribution in statistics.
2. Used to model many natural phenomena. Use extensively in inference.
3. Completely characterized by its mean  $\mu$  and variance  $\sigma^2$  (or  $\mu$  and  $\sigma$ ).

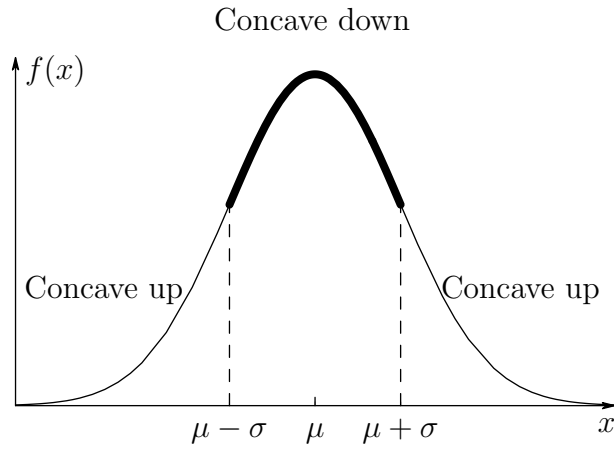
### The Normal Probability Distribution

Suppose  $X$  is a normal random variable with mean  $\mu$  and variance  $\sigma^2$ . The probability density function is given by

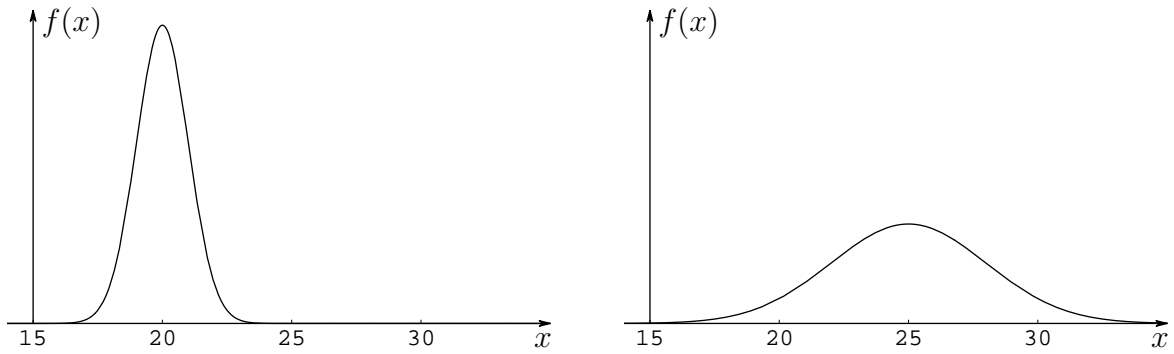
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad \text{and} \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma^2 > 0.$$

### Remarks

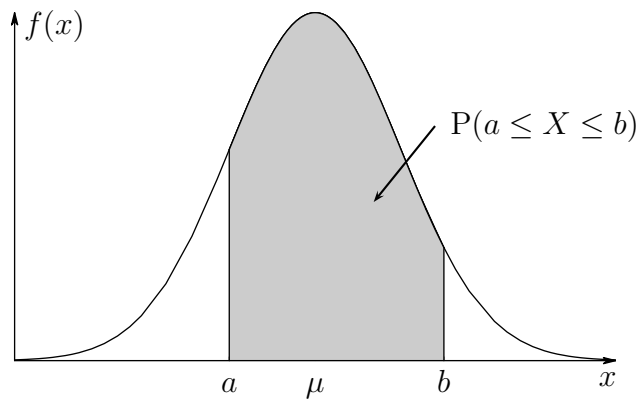
1.  $e \approx 2.71828$ : base of the natural logarithm.  
 $\pi \approx 3.14159$ : commonly used in trigonometry.
2. Shorthand notation:  $X \sim N(\mu, \sigma^2)$ .  
 $X$  is distributed as a normal random variable with mean  $\mu$  and variance  $\sigma^2$ .  
 Example:  $X \sim N(57, 4)$
3. Density curve continues forever in both directions:  $x$  can be any real number.  
 $\mu$  can be any real number.  
 $\sigma^2$  must be positive.
4. For any  $\mu$  and  $\sigma^2$ : density curve is symmetric about the mean, unimodal, and bell-shaped.  
 Density curve changes concavity at  $x = \mu - \sigma$  and  $x = \mu + \sigma$ .  
 $\mu = \tilde{\mu}$   
 Total area under the density curve is 1.



5.  $\mu$ : location parameter.  $\sigma^2$ : determines the spread of the distribution.



Problem: Suppose  $X \sim N(\mu, \sigma^2)$ . Find  $P(a \leq X \leq b)$ .



Finding  $P(a \leq X \leq b)$ :

1. Shaded region is not a rectangle, triangle, or trapezoid; it's bounded by a curve.
2. No nice formula. Calculus?
3. A probability statement involving *any* normal random variable can be *transformed* into an equivalent expression involving a *standard normal random variable*.
4. Cumulative probabilities for the standard normal random variable: Table 3.

### The Standard Normal Distribution

The normal distribution with  $\mu = 0$  and  $\sigma^2 = 1$  (and  $\sigma = 1$ ) is called the **standard normal distribution**. A random variable that has a standard normal distribution is called a **standard normal random variable**, usually denoted  $Z$ . The probability density function for  $Z$  is given by

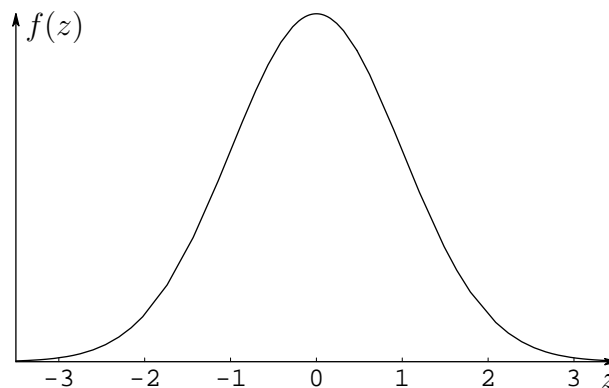
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

### Remarks

1. Independent variable ( $z$ ) is a placeholder; could be any letter.

Standard normal random variable is usually denoted by  $Z$ . ( $Z$  world)

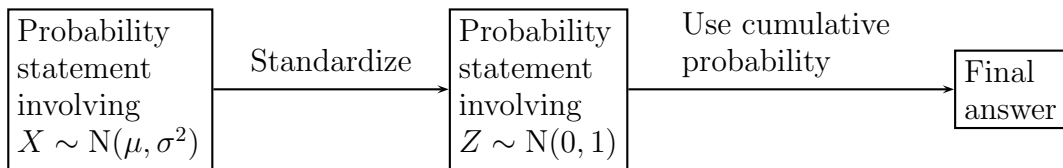
2. Let  $\mu = 0$  and  $\sigma = 1$ .  $Z \sim N(0, 1)$ .



3. Standard normal distribution: not common, but used as a reference distribution.

Standardize:

1. Any probability statement involving any normal random variable can be transformed into an equivalent expression involving a  $Z$  random variable.
2. Need to understand how to compute probabilities in a  $Z$  world.
3. Probabilities associated with  $Z$ : use cumulative probability.
4. Strategy for computing probability associated with any normal random variable:



Cumulative probability for a standard normal random variable:

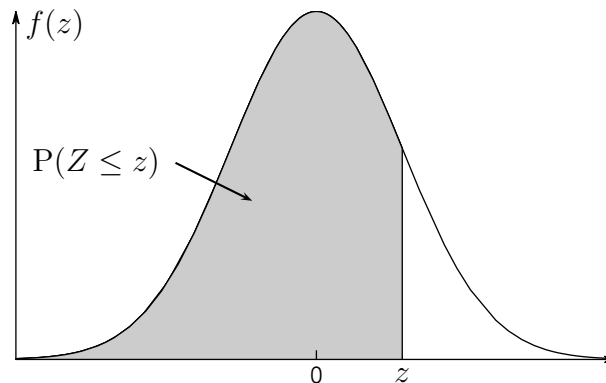
1.  $P(Z \leq z)$ : Table 3

Units and tenths digits in  $z$  along the left side of the table.

Hundredths digit in  $z$  across the top row.

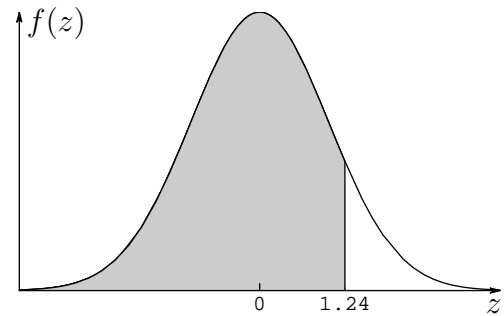
Intersection of this row and column: cumulative probability.

2.  $P(Z \leq z)$ : illustration.

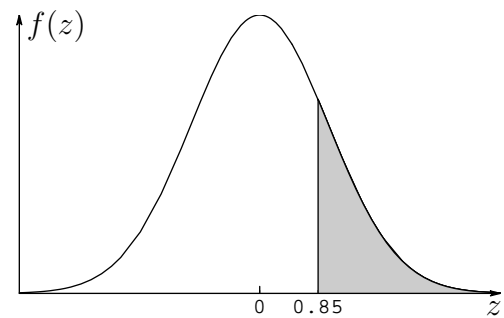


**Example 6.3.1** Find each probability associated with the standard normal distribution.

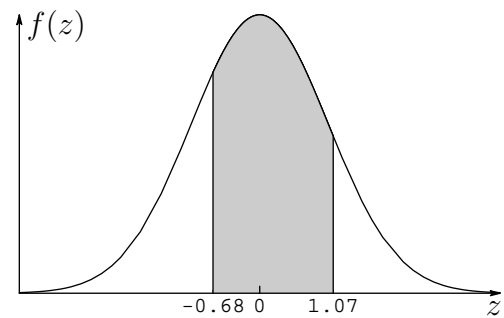
(a)  $P(Z \leq 1.24) =$



(b)  $P(Z \geq 0.85) =$

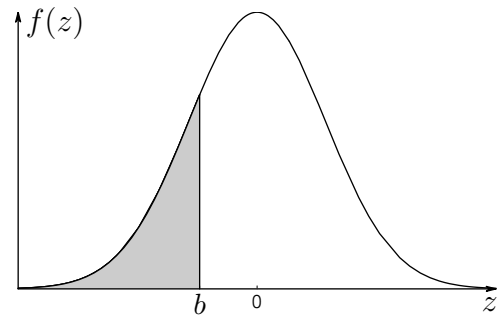


(c)  $P(-0.68 \leq Z \leq 1.07) =$



**Example 6.3.2** Find the value  $b$  such that  $P(Z \leq b) = 0.20$ .

$$P(Z \leq b) = 0.20$$



### Standardization Rule

If  $X$  is a normal random variable with mean  $\mu$  and variance  $\sigma^2$ , then a standard normal random variable is given by

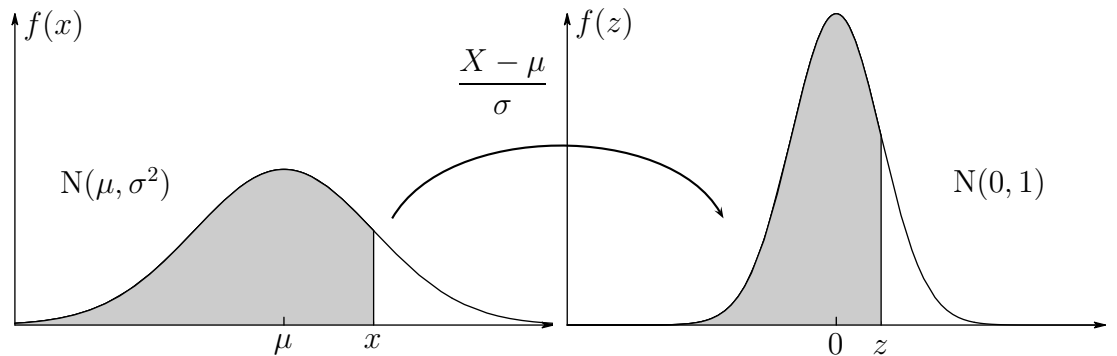
$$Z = \frac{X - \mu}{\sigma}$$

### Remarks

1. Process of converting from  $X$  to  $Z$ : standardization.

$Z$ : standardized random variable.

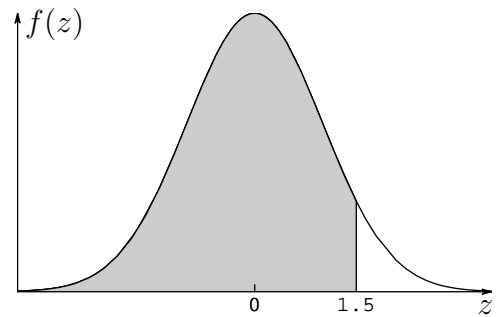
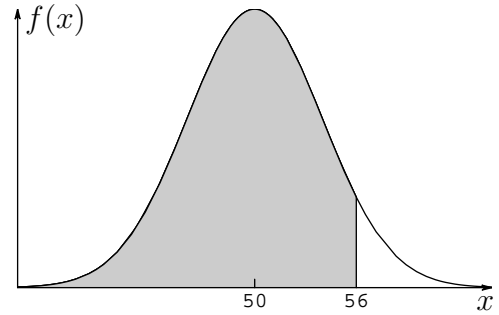
2. Illustration:



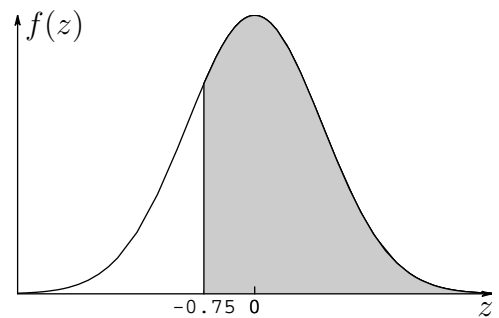
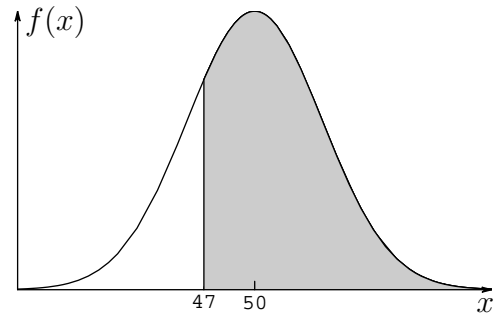
$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = P(Z \leq z)$$

**Example 6.3.3** Suppose  $X$  is a normal random variable with mean 50 and variance 16:  $X \sim N(50, 16)$ , and  $\sigma = 4$ .

(a)  $P(X \leq 56) =$

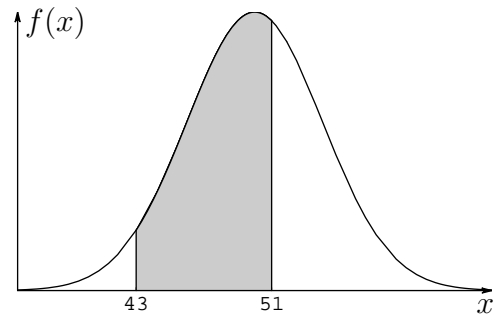


(b)  $P(X > 47) =$

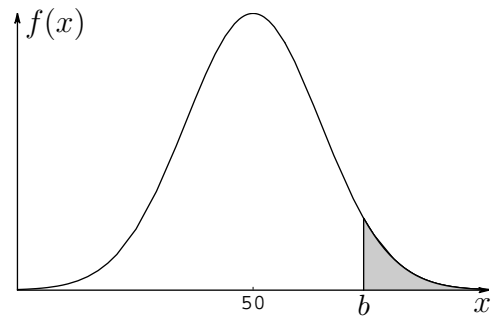




(c)  $P(43 \leq X \leq 51) =$



(d) Find a value  $b$  such that  $P(X > b) = 0.05$ .



**Example 6.3.4** Mildura is a small city 550 km northwest of Melbourne, Australia. The annual rainfall is 292 mm, and the month of October is usually the year's wettest. Suppose the amount of rainfall in October is normally distributed with mean 31 mm and standard deviation 2.3 mm.

- (a) For a randomly selected October, find the probability the amount of rainfall is less than 34 mm.
- (b) For a randomly selected October, find the probability the amount of rainfall is between 25 and 30 mm.
- (c) If the amount of rain in October is over 38 mm, the Murray River will flood. Find the probability the Murray River will flood in October.

**Example 6.3.5** Suppose the area (in square feet) of an office cubicle is normally distributed with mean 220 and standard deviation 15.5.

- (a) Find the probability a randomly selected office cubicle has an area of more than 250 square feet.
- (b) Find the probability the area of a randomly selected office cubicle is between 215 and 222 square feet.

**Example 6.3.6** There are many naturally occurring radioactive elements in air, water, and soil, for example in rocks and oceans. In the United States, an adult receives on average approximately 360 millirem of radiation per year from natural sources. Suppose the amount of radiation an adult receives during a year is normally distributed with mean 360 millirem and standard deviation 28.8 millirem.

- (a) Find the probability a randomly selected adult in the United States is exposed to less than 300 millirem of radiation during a year.
- (b) Find the probability a randomly selected adult in the United States is exposed to between 370 and 390 millirem of radiation during a year.

**Example 6.3.7** The capacity (in cubic feet) of refrigerators with either a top or bottom freezer is normally distributed with mean 22.5 and standard deviation 3.5. Find the value  $c$  such that 10% of all refrigerators of this type have a capacity greater than  $c$ .

**Example 6.3.8** A manufacturing plant produces various chemicals and compounds. The amount (in pounds) of polyvinyl chloride (PVC) produced during a randomly selected workday is normally distributed with mean 1100 and standard deviation 65. Find a symmetric interval about the mean such that on 95% of all workdays the amount of PVC produced lies in this interval.

---

## 6.4 Checking the Normality Assumption

1. Almost every inferential statistics procedure requires certain assumptions.
2. Many statistical techniques are valid only if the observations are from a normal distribution.
3. Therefore: check for normality.
4.  $\bar{x}$  and  $s$ : estimates for  $\mu$  and  $\sigma$ .

But we still need to check the distribution assumption: normality.

5. Methods presented are used to check for any evidence of non-normality: not bell-shaped, skewed, or heavy tails.

### Methods:

1. Graphs

Histogram, stem-and-leaf plot, dot plot.

Consider the shape of the distribution for indications the distribution is not bell-shaped and symmetric.

2. Backwards Empirical Rule

Find  $\bar{x}$  and  $s$ , and the intervals  $(\bar{x} - ks, \bar{x} + ks)$ ,  $k = 1, 2, 3$ .

Compute the actual proportion of observations in each interval and compare with 0.68, 0.95, and 0.997.

3.  $IQR/s$

Find the interquartile range divided by the sample standard deviation.

This ratio should be close to 1.3 if the distribution is approximately normal.

4. Normal Probability Plot

A scatter plot of each observation versus its corresponding standardized normal score.

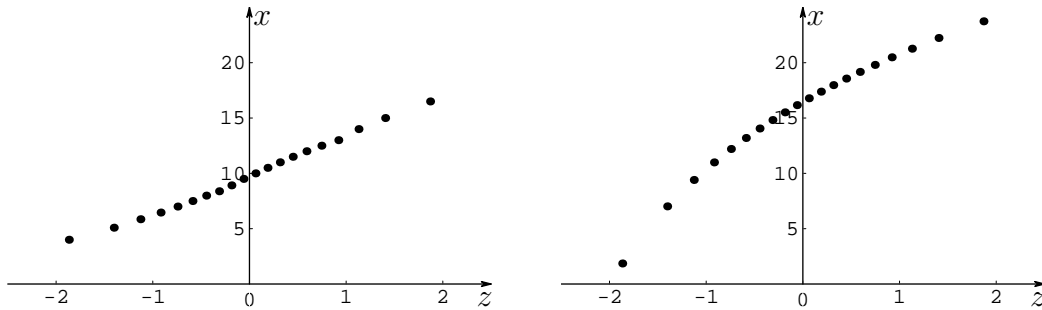
For a normal distribution, points fall along a straight line.

### How to Construct a Normal Probability Plot

Suppose  $x_1, x_2, \dots, x_n$  is a set of observations.

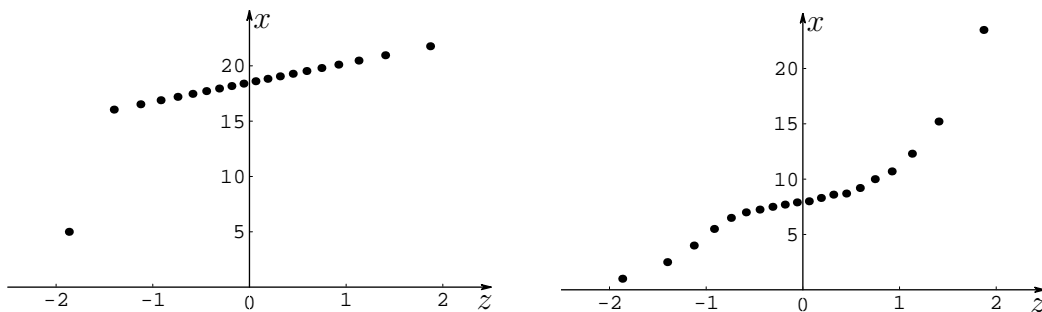
1. Order the observations from smallest to largest, and let  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  represent the set of ordered observations.
2. Find the standardized normal scores for a sample of size  $n$  in Table 4:  $z_1, z_2, \dots, z_n$ .
3. Plot the ordered pairs  $(z_i, x_{(i)})$ .

Examples of normal probability plots (the second graph is an example of non-normality):



### Remarks

1. If the scatter plot is nonlinear: evidence to suggest the data did not come from a normal distribution.
2. Data axis can be horizontal or vertical.
3. Interpretation is very subjective. Look for the points to lie along a straight line.
4. Examples of non-normality:





**Example 6.4.1** Agricultural and storm-water runoffs have carried many different pollutants to the receiving bodies of water. The sources include fertilizers, insecticides, detergents, and motor vehicles. A random sample of surface water along Pacific coast beaches was obtained, and the amount of phosphorus was measured (in mg/L) for each. The 20 observations are given in the following table.

---

2.10	2.39	2.51	2.76	3.21	3.27	3.50	3.81	3.92	4.06
4.23	4.23	4.26	4.48	4.59	4.60	4.87	5.01	5.35	5.91

---

Is there any evidence to suggest this distribution is not normally distributed?

**Example 6.4.2** The amount of explosives used for hard rock extraction at quarries is carefully monitored. A random sample of explosions was obtained, and the amount of explosives (in kg) was recorded for each. The data are given in the following table.

---

1080	1086	1092	1098	1111	1122	1141	1152	1176	1204
1211	1218	1223	1223	1223	1241	1245	1249	1262	1263
1282	1317	1334	1341	1377	1420	1434	1465	1465	1478

---

Is there any evidence to suggest this distribution is not normally distributed?

## 6.5 The Exponential Distribution

1. Many other common continuous distributions:  $t$ , chi-square,  $F$ , etc.
2. Exponential distribution: models the time to failure.

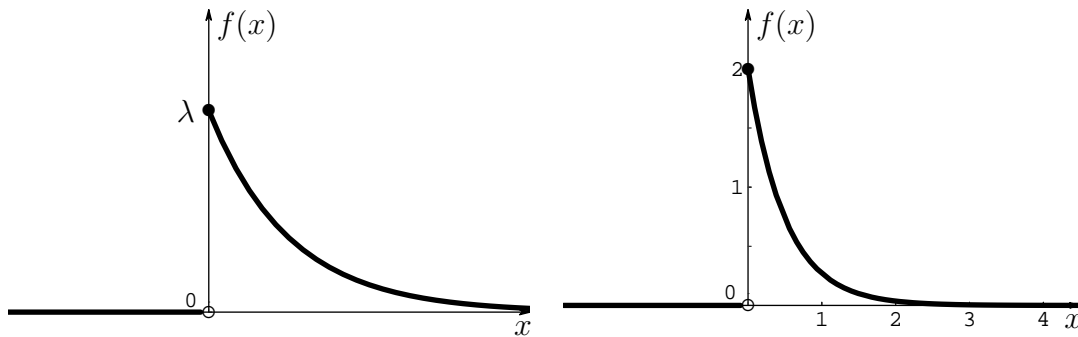
### The Exponential Probability Distribution

Suppose  $X$  is an exponential random variable with parameter  $\lambda$  (with  $\lambda > 0$ ). The probability density function is given by

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

### Remarks

1. Exponential distribution completely characterized by one parameter:  $\lambda$ .  
 $\lambda$ : represents the failure rate.
2.  $e \approx 2.71827$ : base of the natural logarithm.
3. Exponential distribution examples:



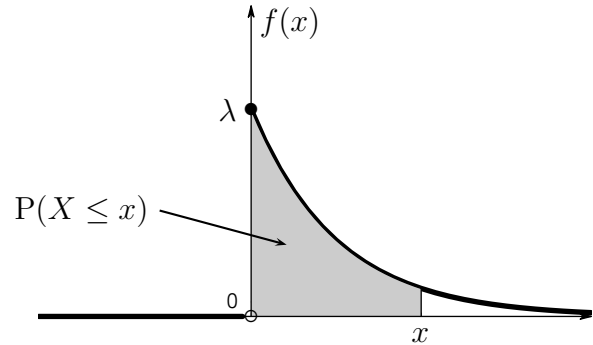
4. If  $X$  is an exponential random variable with parameter  $\lambda$ :

$$\mu = \frac{1}{\lambda} \quad \text{and} \quad \sigma^2 = \frac{1}{\lambda^2}.$$

Probability calculations associated with an exponential random variable with parameter  $\lambda$ :

Use cumulative probability:

$$P(X \leq x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$$



Note: for a right-tail probability,

$$P(X > x) = 1 - P(X \leq x) = 1 - (1 - e^{-\lambda x}) = 1 - 1 + e^{-\lambda x} = e^{-\lambda x}$$

**Example 6.5.1** A manufacturer makes a special vertical turbine pump for seawater applications. Once the pump is installed, the length of time (in years) until the pump fails is a random variable,  $X$ , that has an exponential distribution with parameter  $\lambda = 0.2$  per year.

- Find the mean, the variance, and the standard deviation of the length of time until pump failure.
- Find the probability a randomly selected pump fails after 10 years.

**Example 6.5.2** Certain chip solder joints in mainframe computers were studied in the temperature range  $100\text{--}140^\circ\text{C}$  with current densities of  $1.90\text{--}2.75 \times 10^4$  A/cm<sup>2</sup>. The time (in thousands of hours) until a solder joint fails is a random variable with an exponential distribution. Suppose the mean time until failure is 40 (thousand hours).

- (a) Find the value of the parameter  $\lambda$  that characterizes this exponential distribution.
- (b) Carefully sketch a graph of the probability density function for this exponential distribution.
- (c) Find the probability a randomly selected solder joint fails after between 30 and 50 thousand hours.

**Example 6.5.3** Commercial airplane window glass lasts, on average, approximately 14 years. The lifetime (in years) of a randomly selected airplane windowpane can be modeled by an exponential random variable with  $\lambda = 0.072$  per year. Suppose a windowpane is selected at random.

- (a) What is the probability the windowpane will last for at least four years?
- (b) Suppose the windowpane lasts for four years. What is the probability it will last for at least another four years?

## CHAPTER 7

# Sampling Distributions

---

## 7.0 Introduction

Numbers represented by  $\mu$ ,  $\sigma^2$ ,  $p$ : population characteristics.

These values are usually unknown.

Often we want to estimate population characteristics, or draw conclusions about them.

It seems reasonable to use  $\bar{x}$  to estimate  $\mu$ .

**Key:** the value of  $\bar{x}$  dances around the true population mean  $\mu$ .

Need to understand this variability.

---

## 7.1 Statistics, Parameters, and Sampling Distributions

### Definition

A **parameter** is a numerical descriptive measure of a population.

A **statistic** is any quantity computed from values in a sample.

### Remarks

1. Parameter: a population quantity.

Used to describe some characteristic of a population.

Usually, cannot measure a parameter; it is an unknown constant we would like to estimate.

2. Statistic: any sample quantity.

We could compute infinitely many quantities using the data in a sample.

Parameters describe populations; statistics describe samples.

We use statistics to make inferences about parameters.

Therefore, the properties of a statistic are important.

**Example 7.1.1** In each of the following statements, identify the **boldface** number as the value of a population parameter or a sample statistic.

- (a) In a survey of first-year college students living on campus, **67%** said they were happy with their roommate.
  
- (b) A hospital official reported that the mean length of coma for all head-trauma patients is **36.2** days.
  
- (c) In a study of driving habits, **32%** of the participants had some kind of bumper sticker.
  
- (d) The mean attendance for all NFL football games during the 2005 season was **35,756.3**.
  
- (e) In a survey conducted by the U.S. Treasury, **59%** of the Americans questioned said we should keep the penny in circulation.
  
- (f) The mean tax on a package of cigarettes for all 50 states is **\$2.46**.



Statistics are random variables

1. Consider two samples of size  $n$  from the same population with means  $\bar{x}_1$  and  $\bar{x}_2$ .

$\bar{x}_1$  and  $\bar{x}_2$  should be close, **but different!**

2. Sample mean will differ from sample to sample: sampling variability.
3.  $\bar{X}$  is a random variable, with a mean, variance, and standard deviation.

The probability distribution is called the sampling distribution.

4. Any statistic is a random variable: differs from sample to sample.

We need to know the properties of the distribution of the statistic.

### Definition

The **sampling distribution** of a statistic is the probability distribution of the statistic.

### Remarks

1. Sampling distribution: describes the long-run behavior of the statistic.
2. Finding a sampling distribution:
  - (a) Construct a histogram (or stem-and-leaf plot) using values from the population: approximate the distribution.
  - (b) Exact sampling distribution may be obtained in some cases.

**Example 7.1.2** Every weeknight, there are six different television news shows a viewer may choose from. On a recent Wednesday evening, the number of viewers for each network news show (in millions) were 13.2, 10.9, 11.6, 15.7, 20.6, and 12.1. Suppose three of these news shows are selected at random, and the sample median number of viewers is computed. Find the sampling distribution for the sample median.

$$\binom{6}{3} =$$

Sample	$\tilde{x}$	Probability	Sample	$\tilde{x}$	Probability
13.2, 10.9, 11.6			13.2, 10.9, 15.7		
13.2, 10.9, 20.6			13.2, 10.9, 12.1		
13.2, 11.6, 15.7			13.2, 11.6, 20.6		
13.2, 11.6, 12.1			13.2, 15.7, 20.6		
13.2, 15.7, 12.1			13.2, 20.6, 12.1		
10.9, 11.6, 15.7			10.9, 11.6, 20.6		
10.9, 11.6, 12.1			10.9, 15.7, 20.6		
10.9, 15.7, 12.1			10.9, 20.6, 12.1		
11.6, 15.7, 20.6			11.6, 20.6, 12.1		
11.6, 15.7, 12.1			15.7, 20.6, 12.1		

**Remarks**

1. Find the mean, variance, and standard deviation for  $\tilde{X}$  in Example 7.1.2.
2. Sampling with replacement changes the probability distribution.

**Example 7.1.3** A solar physicist checks the sun every day and records the number of visible sunspots. Years of research indicate that the probability distribution for  $X$ , the number of sunspots visible on a randomly selected day, is given by

$x$	0	1	2	3
$p(x)$	0.3	0.2	0.4	0.1

Two days are selected at random, and the number of sunspots on each day is recorded. Consider the statistic  $M$ , the minimum number of sunspots visible for the two days. Find the probability distribution for  $M$ .

Sample	$m$	Probability	Sample	$m$	Probability
0, 0			2, 0		
0, 1			2, 1		
0, 2			2, 2		
0, 3			2, 3		
1, 0			3, 0		
1, 1			3, 1		
1, 2			3, 2		
1, 3			3, 3		

Observational studies: data obtained from a simple random sample.

1. Usually: without replacement; individual not placed back into the population.
2. Therefore, individual responses are dependent.
3. If the population is large enough and the sample size is small relative to the population: responses are almost independent.
4. Calculate probabilities as though responses are independent: little loss of accuracy.
5. Rule of Thumb: if the sample size is at most 5% of the total population, successive observations can be considered independent.

Recall:

**Definition**

A **(simple) random sample** (SRS) of size  $n$  is a sample selected in such a way that every possible sample of size  $n$  has the same chance of being selected.

**Remarks**

1. Finite population of size  $N$  and sample of size  $n$ .

Number of possible simple random samples is  $\binom{N}{n}$ .

2. We often refer to values of the variable as the random sample rather than the individuals.
3. Unless otherwise stated: all data are obtained from a simple random sample.

**Example 7.1.4** A state health official monitors the operations of 30 industrial plants. The total amount of toxic chemicals released into the air (in pounds) for each facility last year is given in the following table.

164	136	179	104	124	155	133	104	132	180
188	197	136	105	108	165	111	108	184	111
192	101	100	176	133	102	163	142	168	128

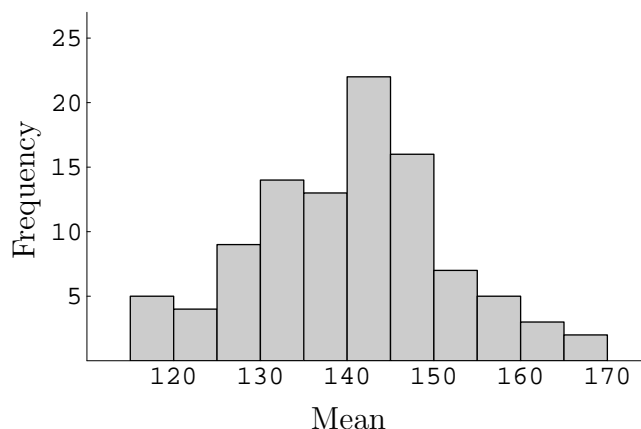
Find an approximate sampling distribution for the sample mean of 6 observations from this population.

Number of possible samples of size 6:

Select 100 random samples of size 6, and compute the mean for each sample.

For example: 184, 132, 164, 124, 188, 163 :  $\bar{x} = 159.2$ .

Here is a histogram of the sample means for 100 samples of size 6.



Observations:

1. Shape of the sampling distribution is approximately normal!
2. Center of the sampling distribution is approximately the population mean ( $\mu = 141.0$ ).
3. Less variability in the sampling distribution than in the population distribution.

## 7.2 The Sampling Distribution of the Sample Mean

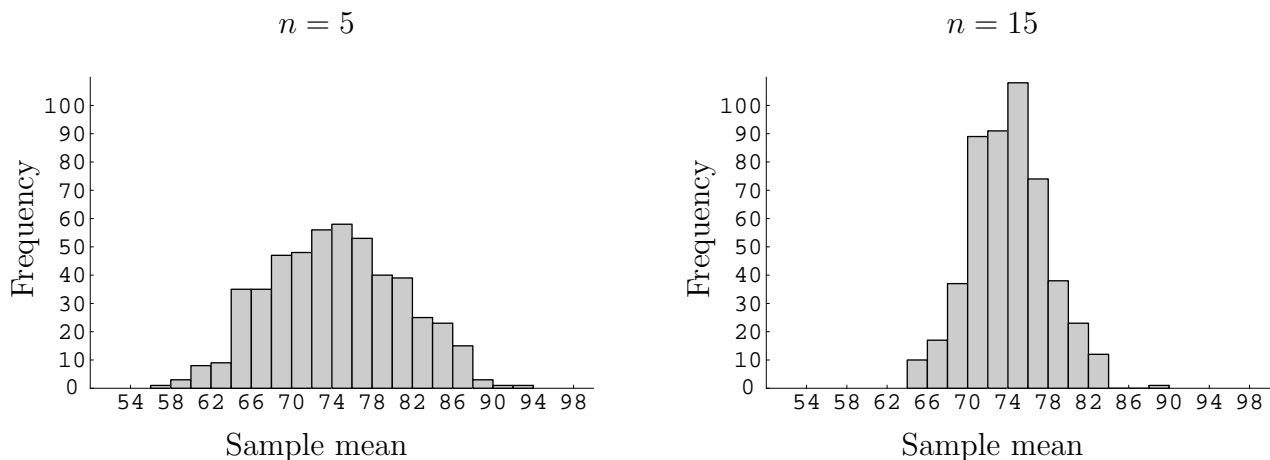
1. Reasonable to use  $\bar{x}$  to estimate  $\mu$ .
2. Sampling variability makes it difficult to know how far a specific  $\bar{x}$  is from  $\mu$ .
3. We'll find the exact probability distribution of  $\bar{X}$ .

**Example 7.2.1** Consider a population consisting of the numbers 50, 52, 54, 56, 58, 60,  $\dots$ , 98. The population mean is  $\mu = (50 + 52 + \dots + 98)/25 = 74$ . Use frequency histograms to approximate the distribution of the mean,  $\bar{X}$ .

Consider a random sample of  $n$  observations selected with replacement.

For  $n = 5$  and for  $n = 15$ , 500 random samples were selected, and the sample mean was computed for each..

Here are the resulting histograms.



Observations:

1. Each distribution is centered near the population mean, 74.
2. The shape of each distribution is approximately normal!
3. Sampling variability decreases as  $n$  increases.

**NOTE:** The distribution of the sample mean is approximately normal, with mean equal to the underlying, original, population mean, and variance related to the sample size.

**Example 7.2.2** A large hotel has a special parking area near the front door for taxis waiting for riders. Past research indicates that the number of taxis waiting at 6:00 a.m. is a random variable,  $X$ , with probability distribution given in the following table.

$x$	0	1	2	3
$p(x)$	0.1	0.2	0.6	0.1

Suppose two days are selected at random.

- (a) Find the sampling distribution of  $\bar{X}$ , the sample mean number of taxis waiting.
- (b) Find the mean and the variance of the random variable  $X$ .
- (c) Find the mean and the variance of the random variable  $\bar{X}$ .
- (d) How do the results from (b) compare with the results from (c)?

There are  $4 \times 4 = 16$  possible samples.

The probability of each sample is computed using independence.

Use the following table.

Sample	$\bar{x}$	Probability	Sample	$\bar{x}$	Probability
0, 0	0.0	0.01	0, 1	0.5	0.02
0, 2	1.0	0.06	0, 3	1.5	0.01
1, 0	0.5	0.02	1, 1	1.0	0.04
1, 2	1.5	0.12	1, 3	2.0	0.02
2, 0	1.0	0.06	2, 1	1.5	0.12
2, 2	2.0	0.36	2, 3	2.5	0.06
3, 0	1.5	0.01	3, 1	2.0	0.02
3, 2	2.5	0.06	3, 3	3.0	0.01

$\bar{x}$	0.0	0.5	1.0	1.5	2.0	2.5	3.0
$p(\bar{x})$							

Example (continued)



Another illustration to show the connections between the distribution of the sample mean and the distribution of the original population:

1. Three distributions:
  - (a) Normal distribution with mean 0 and standard deviation 1.
  - (b) Uniform distribution with  $a = 0$  and  $b = 1$ .
  - (c) Exponential distribution with  $\lambda = 0.5$ .
2. Select 500 samples of size  $n = 2$ .

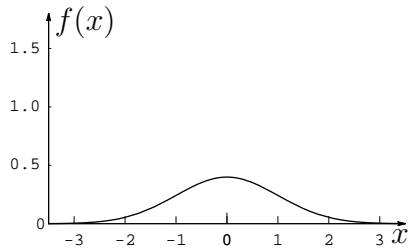
Compute the mean for each sample.

Construct a (smoothed) histogram of the sample means.
3. Repeat this procedure for  $n = 5, 10,$  and  $20$ .

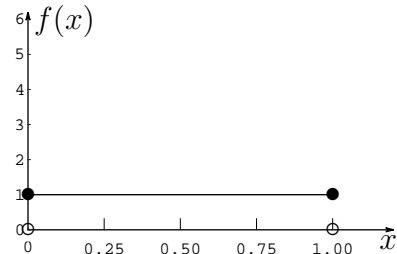
Observations:

1. If the underlying population is normal, the distribution of the sample mean appears to be normal, regardless of the sample size.
2. Even if the underlying population is *not* normal, the distribution of the sample mean becomes *more normal* as  $n$  increases.
3. The sampling distribution of the mean is centered at the mean of the underlying population.
4. As the sample size,  $n$ , increases, the variance of the distribution of the sample mean decreases.

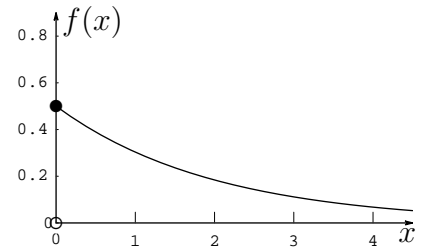
Normal distribution  
 $\mu = 0, \sigma^2 = 1$



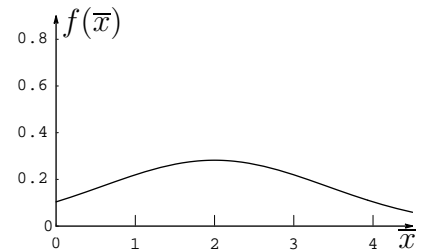
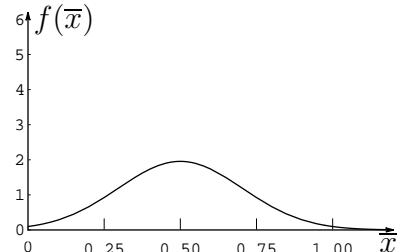
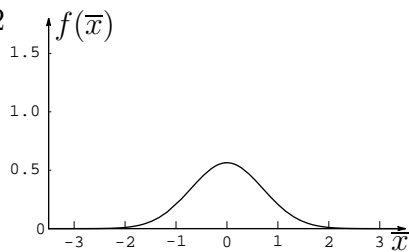
Uniform distribution  
 $a = 0, b = 1, \mu = 0.5, \sigma^2 = 0.0833$



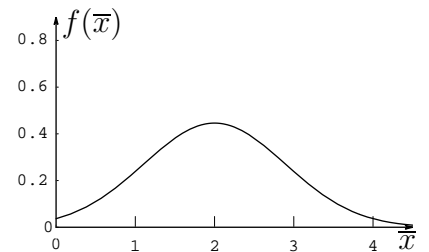
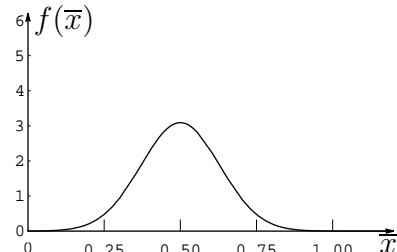
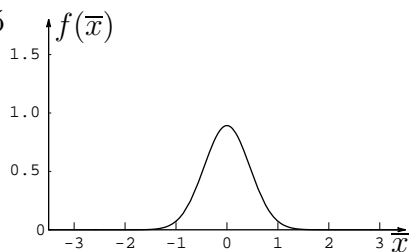
Exponential distribution  
 $\lambda = 0.5, \mu = 2, \sigma^2 = 4$



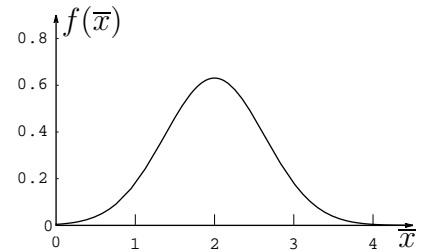
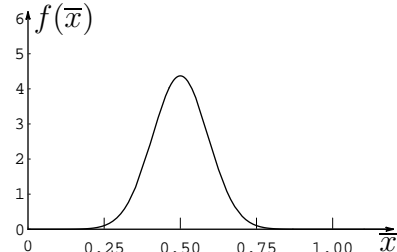
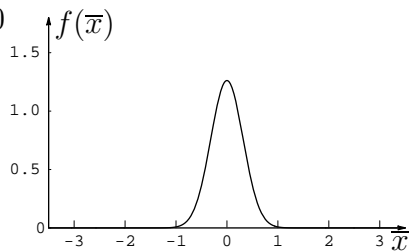
$n = 2$



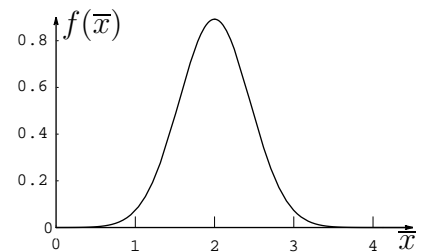
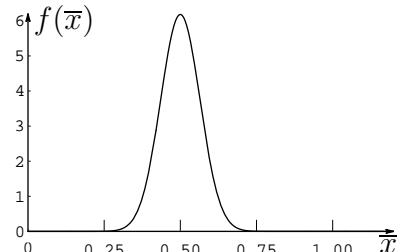
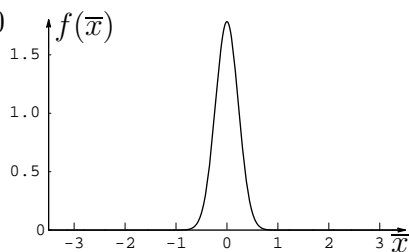
$n = 5$



$n = 10$



$n = 20$



**Properties of the Sample Mean**

Let  $\bar{X}$  be the mean of observations in a random sample of size  $n$  drawn from a population with mean  $\mu$  and variance  $\sigma^2$ .

1. The mean of  $\bar{X}$  is equal to the mean of the underlying population.

In symbols:  $\mu_{\bar{X}} = \mu$ .

2. The variance of  $\bar{X}$  is equal to the variance of the underlying population divided by the sample size.

In symbols:  $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$ .

The standard deviation of  $\bar{X}$  is  $\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ .

3. If the underlying population is distributed normally, then the distribution of  $\bar{X}$  is also *exactly* normal for any sample size.

In symbols: If  $X \sim N(\mu, \sigma^2)$ , then  $\bar{X} \sim N(\mu, \sigma^2/n)$

Note:

1. The sample mean is an unbiased estimator of the population mean  $\mu$ .
2. Even if the underlying distribution is not normal, the distribution of  $\bar{X}$  becomes more normal as the sample size increases.

This remarkable result: **The Central Limit Theorem**

**Central Limit Theorem (CLT)**

Let  $\bar{X}$  be the mean of observations in a random sample of size  $n$  drawn from a population with mean  $\mu$  and finite variance  $\sigma^2$ . As the sample size  $n$  increases, the sampling distribution of  $\bar{X}$  will increasingly approximate a normal distribution, with mean  $\mu$  and variance  $\sigma^2/n$ , regardless of the shape of the underlying population distribution.

In symbols:  $\bar{X} \rightsquigarrow N(\mu, \sigma^2/n)$ .

**Remarks**

1. A better name: the Normal Convergence Theorem.

The distribution of  $\bar{X}$  *converges* to, or gets closer and closer to, a normal distribution.

2. If the original population is normally distributed, then the distribution of  $\bar{X}$  is normal, no matter how large or small the sample size ( $n$ ).
3. If the original population is *not* normal, the CLT says the distribution of  $\bar{X}$  approaches a normal distribution as  $n$  increases.

The approximation improves as  $n$  increases.

The approximation gets better and better as the sample size  $n$  gets bigger and bigger.

No magical threshold value for  $n$ .

In most cases: if  $n \geq 30$ , then the approximation is pretty good.

In some cases, for  $n$  as little as 5 the approximation will be excellent. In others,  $n$  might have to be at least 25 before the approximation is good.

4. To compute probability involving  $\bar{X}$ : just like for any other normal random variable.

Standardize, use cumulative probability where appropriate.

Same method even if  $\bar{X}$  is only approximately normal.

5. As  $n$  increases, the distribution of  $\bar{X}$  becomes more compact.

Consider  $\sigma_{\bar{X}}^2 = \sigma^2/n$ .

6. General version of the CLT: statement about the sum of independent observations,  $T$ .

If  $n$  is sufficiently large, the distribution of  $T$  approaches a normal distribution with mean  $n\mu$  and variance  $n\sigma^2$ .

In symbols:  $T \overset{\circ}{\sim} N(n\mu, n\sigma^2)$ .

**Example 7.2.3** Certain federal judges are appointed and preside over a wide variety of cases. Although most of these judges are formally reappointed every 4 or 8 years, they may serve as long as they want. A recent government survey reported that the mean number of years of service for a current federal judge is 22.5 and the standard deviation is  $\sigma = 5.7$ . Suppose the distribution of length of service is normally distributed. A random sample of 18 current federal judges is obtained, and the length of service for each is recorded.

- (a) Find the probability the sample mean length of service will be less than 20 years.
- (b) Find the probability the sample mean will be more than 23 years.
- (c) Find the probability the sample mean will be within 2 years of the population mean, 22.5.

**Example 7.2.4** Mount McKinley in Alaska is one of the world's most dangerous mountain peaks for climbers. On average, it takes 408 hours to traverse the 20,320-foot peak, with standard deviation 96 hours. Suppose 38 climbers are selected at random.

- (a) Find the probability the sample mean time to climb this mountain is greater than 430 hours.
- (b) Find a value  $c$  such that the probability the sample mean is less than  $c$  is 0.05.

**Example 7.2.5** Albumin is a protein found in egg whites and milk that increases the body's ability to fight disease and infection. The mean amount of albumin in a healthy adult's blood is reported to be 44 g/L and the standard deviation, is  $\sigma = 0.5$ . Suppose 40 adults are selected at random, and the albumin level in each patient's blood is measured.

- (a) Find the probability the sample mean albumin level is between 44.1 and 44.2 g/L.
- (b) Find the probability the sample mean albumin level is greater than 44.5 g/L.
- (c) Find a symmetric interval about the population mean, 44, such that the probability the sample mean lies in this interval is 0.99.

**Example 7.2.6** Biotin (vitamin H) is important for the synthesis of proteins and fatty acids, and has been linked to hair loss in men. Suppose the mean amount of biotin ingested per day by adult men is 50 mcg (assume  $\sigma = 25$ ). A sample of 32 men with hair loss was randomly selected. A nutrition survey was used to determine the amount of biotin (in mcg) ingested per day by each person. The sample mean was  $\bar{x} = 39.8$ . Is there any evidence to suggest that the population mean biotin intake per day is less than 50 mcg for men with hair loss?



**Example 7.2.7** The Sanctuary resort and spa on Kiawah Island, South Carolina, has 255 rooms and a mission to provide the finest service for every guest. The manager claims that the mean time for room service to deliver any order to any room is 12 minutes (assume  $\sigma = 2.5$ ). A random sample of 36 room-service orders was obtained, and the time to delivery was recorded for each. The sample mean was  $\bar{x} = 12.9$  minutes. Is there any evidence to suggest that the mean time to deliver a room-service order is greater than 12 minutes?

## 7.3 The Distribution of the Sample Proportion

1. Use the sample proportion,  $\hat{p}$ , to estimate the population proportion  $p$ .
2. Consider a sample of  $n$  individuals. Let  $X$  be the number of successes in the sample.

The sample proportion (the random variable) is defined to be

$$\hat{P} = \frac{X}{n} = \frac{\text{The number of successes in the sample}}{\text{The sample size}}.$$

3. We need to know the distribution of  $\hat{P}$  in order to answer probability questions concerning the sample proportion.

### The Sampling Distribution of $\hat{P}$

Let  $\hat{P}$  be the sample proportion of successes in a sample of size  $n$  from a population with a true proportion of success  $p$ .

1. The mean of  $\hat{P}$  is the true population proportion.

In symbols:  $\mu_{\hat{P}} = p$ .

2. The variance of  $\hat{P}$  is  $\sigma_{\hat{P}}^2 = \frac{p(1-p)}{n}$ .

The standard deviation of  $\hat{P}$  is  $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$ .

3. If  $n$  is large and both  $np \geq 5$  and  $n(1-p) \geq 5$ , then the distribution of  $\hat{P}$  is approximately normal.

In symbols:  $\hat{P} \dot{\sim} N(p, p(1-p)/n)$ .

**Remarks**

1. As  $n$  increases, the distribution of  $\hat{P}$  approaches a normal distribution. There is no threshold value for  $n$ . The larger the value of  $n$  and the closer  $p$  is to 0.5, the better the approximation.
2.  $\hat{P}$  is approximately normal. Sample proportion can be written as a sample mean. How?
3.  $\hat{P}$  is an unbiased estimator for  $p$ .
4. A large sample isn't enough for normality.

Non-skewness criteria:  $np \geq 5$  and  $n(1 - p) \geq 5$ .

5. Probabilities involving  $\hat{P}$ : standardize, use cumulative probability where appropriate.

**Example 7.3.1** In a recent study of retired Americans, it was reported that 33% of all retired women had been forced to retire. Suppose 140 retired women are selected at random.

- (a) Find the probability distribution of the sample proportion,  $\hat{P}$ , of women who were forced to retire.
- (b) What is the probability the sample proportion of women forced to retire is greater than 0.40?

**Example 7.3.2** In Florida, 10.5% of all Medicare beneficiaries underwent echocardiography (a diagnostic procedure that provides information about the structure and functioning of the heart) last year. A random sample of 220 Medicare beneficiaries in Florida was obtained, and each was asked whether they had an echocardiogram last year.

- (a) What is the probability the sample proportion of Medicare beneficiaries who had an echocardiogram is greater than 0.13?
- (b) Find a value  $c$  such that the probability the sample proportion is less than  $c$  is 0.90.

**Example 7.3.3** The National Center for Chronic Disease Prevention and Health Promotion reported that 77% of all employed people are protected by smoking policies at their workplace. A random sample of 250 California workers was obtained, and each was asked whether their employer had a smoking policy to protect workers.

- (a) What is the probability the sample proportion of workers protected by smoking policies is between 0.75 and 0.85?
- (b) Suppose the sample proportion for the 250 workers is 0.70. Is there any evidence to suggest that the true proportion of workers protected by smoking policies is less than 0.77?



## CHAPTER 8

# Confidence Intervals Based on a Single Sample

---

## 8.0 Introduction

A single value of a statistic computed from a sample conveys little information about confidence and reliability.

Alternative method: use a single value to construct an interval in which we are fairly certain the true value lies.

What makes a point estimator *good*?

Methods for constructing confidence intervals.

---

## 8.1 Point Estimation

Point estimate of a population parameter: a single number computed from a sample which serves a best guess for the parameter.

1. *Estimator*: a statistic of interest, a random variable.

An estimator has a distribution, a mean, a variance, and a standard deviation.

2. *Estimate*: a specific value of an estimator.

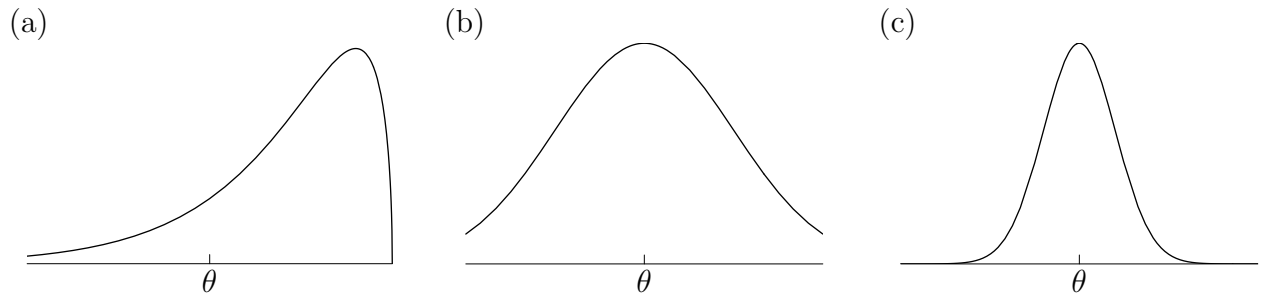
### Definition

An **estimator** (statistic) is a rule used to produce a point estimate of a population parameter.

Problem: Estimate a population parameter  $\theta$ .

There are many different statistics available.

Which statistic should we use?



(a) Unlikely to produce a value close to  $\theta$ .

Sampling distribution is skewed to the left.

Most of the values of the statistic are to the right of  $\theta$ .

(b) Statistic in (b) is centered at  $\theta$ .

On average (in the long run), this statistic will produce  $\theta$ .

This statistic has large variance.

Even though the sampling distribution is centered at the true value of the population parameter, specific estimates will probably be *far away* from  $\theta$ .

(c) Statistic in (c) exhibits two very desirable properties.

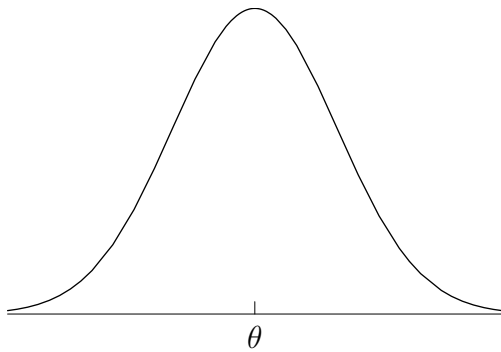
Centered at the true value of the population parameter, and has small variance.



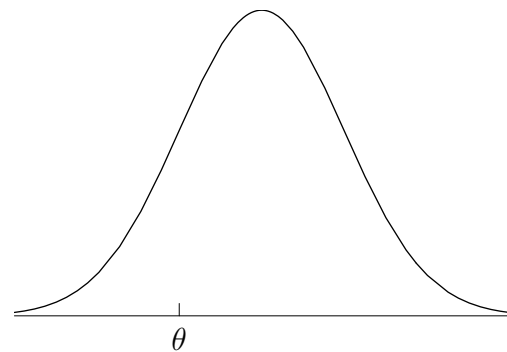
**Definition**

A statistic  $\hat{\theta}$  is an **unbiased estimator** of a population parameter  $\theta$  if  $E(\hat{\theta}) = \theta$ , the mean of  $\hat{\theta}$  is  $\theta$ .

If  $E(\hat{\theta}) \neq \theta$ , then the statistic  $\hat{\theta}$  is a **biased estimator** of  $\theta$ .



An unbiased estimator for  $\theta$ .



A biased estimator for  $\theta$ .

Some unbiased estimators:

1. Sample mean,  $\bar{X}$ , is an unbiased statistic for estimating the population mean  $\mu$ .

$$E(\bar{X}) = \mu.$$

2. Sample proportion,  $\hat{P}$ , is an unbiased statistic for estimating the population proportion  $p$ .

$$E(\hat{P}) = p.$$

3. Sample variance,  $S^2$ , is an unbiased statistic for estimating the population variance  $\sigma^2$ .

$$E(S^2) = \sigma^2.$$

Note: The sample standard deviation  $S$  is a biased estimator for the population standard deviation  $\sigma$ .

$$E(S) = E(\sqrt{S^2}) \neq \sqrt{E(S^2)} = \sqrt{\sigma^2} = \sigma$$

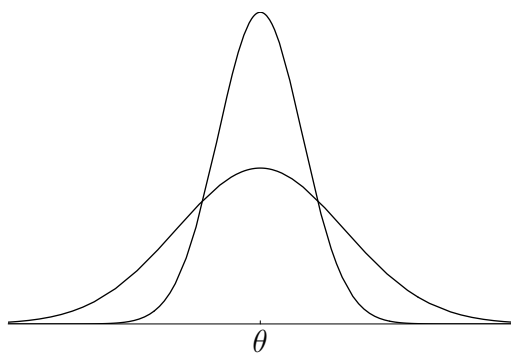
Even though  $S$  is biased, it is still important in statistical inference.

Second rule for choosing a statistic:

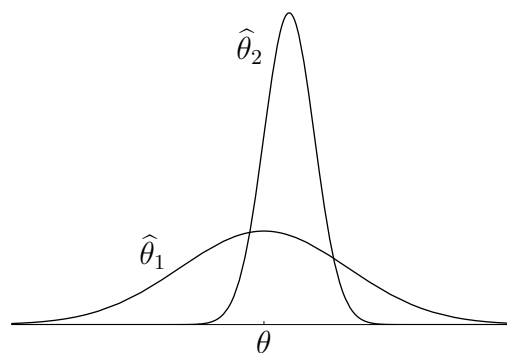
Of all unbiased statistics, use the one with the smallest variance.

This statistic will, on average, be close to the true value of the population parameter.

Illustration:



Two unbiased statistics for estimating  $\theta$ .  
Use the statistic with smaller variance.



Sampling distributions of an unbiased statistic  $\theta_1$  (with large variance) and of a biased statistic  $\theta_2$  (with small variance).

Problem: Two statistics to choose from:  $\hat{\theta}_1$  and  $\hat{\theta}_2$ .

Statistic  $\hat{\theta}_1$ : unbiased, but has large variance;

Statistic  $\hat{\theta}_2$ : slightly biased, but has small variance.

The choice of an estimator is a difficult decision, and there is no definitive answer.

### Remarks

1. If several unbiased statistics from which to choose:

If one of these statistics has the smallest possible variance, it is called the minimum-variance unbiased estimator (MVUE).

2. If the underlying population is normal,  $\bar{X}$  is the MVUE for estimating  $\mu$ .

So, if the population is normal, the sample mean is a *really good* statistic to use for estimating  $\mu$ .

$\bar{X}$  is unbiased, and it has the smallest variance of all possible unbiased estimators for  $\mu$ .

**Example 8.1.1** Movies are a very popular form of entertainment. Whether you like action, horror, drama, romance, or comedy, the local cinema usually has something you'll enjoy. One problem: the price of snacks at a movie theater is sometimes outrageous. A random sample of movie theaters in a large city was obtained, and the price of a small box of popcorn (in dollars) at each is given in the following table.

5.50	2.50	2.80	5.90	1.30	3.20	2.75	5.75	3.25	4.10
3.20	5.25	5.85	6.40	6.10	5.20				

Find point estimates for the population mean price of a small box of popcorn and for the population median price of a small box of popcorn at movie theaters in this city.

Here are the observations in order:

1.30	2.50	2.75	2.80	3.20	3.20	3.25	4.10	5.20	5.25
5.50	5.75	5.85	5.90	6.10	6.40				

**Example 8.1.2** As foreign investment and private ownership have increased in China, there are more buildings over 20 stories high in large cities. Random samples of buildings in Beijing and in Shanghai were obtained. The data are given in the following table.

City	Number of buildings sampled	Number of buildings over 20 stories
Beijing	120	42
Shanghai	108	27

- Find a point estimate for the proportion of buildings in Beijing over 20 stories high.
- Find a point estimate for the proportion of buildings in Shanghai over 20 stories high.
- Find a point estimate for the difference in the proportions of buildings over 20 stories high in Beijing and Shanghai.

---

## 8.2 A Confidence Interval for a Population Mean when $\sigma$ is Known

1. A good estimator: unbiased, small variance.
2. Use the estimate to produce a confidence interval.

### Definition

A **confidence interval** (CI) for a population parameter is an interval of values constructed so that, with a specified degree of confidence, the value of the population parameter lies in this interval.

The **confidence coefficient** is the probability the CI encloses the population parameter in repeated samplings.

The **confidence level** is the confidence coefficient expressed as a percentage.

### Remarks

1. Confidence interval is usually expressed as an *open* interval.

Example:  $(12.8, 32.6)$

12.8 is the left endpoint, or lowerbound, and 32.6 is the right endpoint, or upperbound.

The interval extends all the way to, but does not include, the endpoints.

2. Typical confidence coefficients are 0.95 and 0.99.
3. Typical confidence levels are, therefore, 95% and 99%.

Constructing a rough 95% confidence interval for a population mean  $\mu$ .

1. Suppose either (a) the underlying population is normal, or (b) the sample size  $n$  is large, or both, and the population standard deviation  $\sigma$  is known.
2. Use the properties of  $\bar{X}$  and the CLT (if necessary):

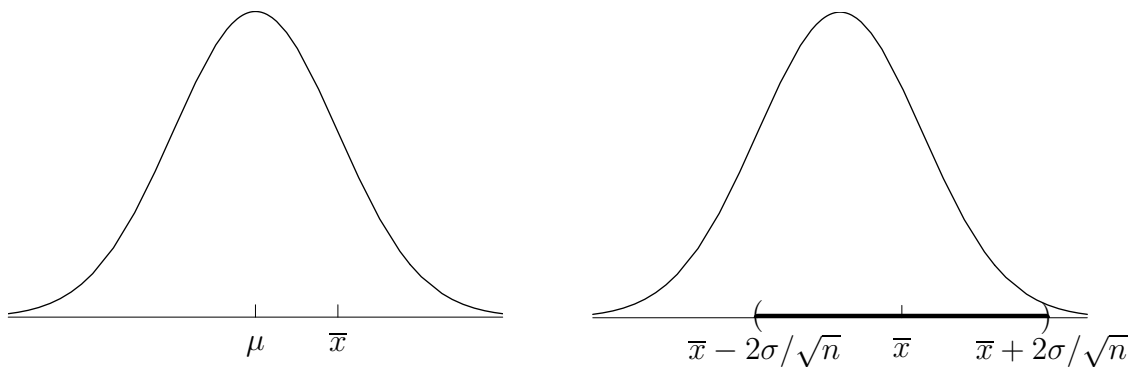
The sample mean  $\bar{X}$  is (approximately) normal with mean  $\mu$  and variance  $\sigma^2/n$ .

In symbols,  $\bar{X} \sim N(\mu, \sigma^2/n)$ .

3. By the Empirical Rule, approximately 95% of all values of the sample mean lie within two standard deviations of the mean.
4. Even though we know the distribution of  $\bar{X}$  is centered at  $\mu$ , we do not know the true value of  $\mu$ .

In order to *capture*  $\mu$  it seems reasonable to *step* two standard deviations from an estimate  $\bar{x}$  in both directions.

The resulting (rough) 95% confidence interval is  $(\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n})$ .



Constructing a more accurate 95% confidence interval for a population mean  $\mu$ .

$$1. \bar{X} \sim N(\mu, \sigma^2/n) \longrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

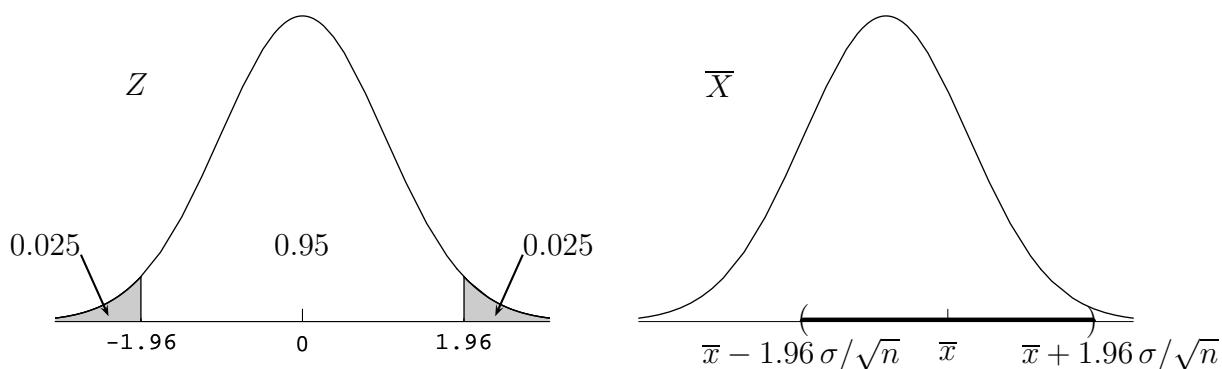
2. Find a symmetric interval about 0 such that the probability  $Z$  lies in this interval is 0.95.

$$P(-1.96 < Z < 1.96) = 0.95$$

3. Substitute for  $Z$  and manipulate the interval inside the probability statement so that  $\mu$  is *caught* in the middle.

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = 0.95$$



### Remarks

1. An exact 95% confidence interval for  $\mu$ : step 1.96 standard deviations from a specific value  $\bar{x}$  in both directions.
2. Generalization: find a  $100(1 - \alpha)\%$  confidence interval for  $\mu$ .
3. Usually,  $\alpha$  is small.

If  $\alpha = 0.05$ , the confidence level is:

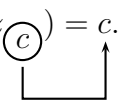
**Definition**

$z_{\alpha/2}$  is a **critical value**. It is a value on the measurement axis in a **standard normal distribution** such that  $P(Z \geq z_{\alpha/2}) = \alpha/2$ .

**Remarks**

1. The subscript on  $z$  could be *any variable*, or letter.

For example,  $P(Z \geq z_{(c)}) = c$ .



2.  $z_{\alpha/2}$ : a  $z$  value such that there is  $\alpha/2$  of the area (probability) to the right of  $z_{\alpha/2}$ .

$-z_{\alpha/2}$ : the negative critical value.

3. Critical values are *always* defined in terms of right-tail probability.
4.  $z$  critical values are easy to find by using the Complement Rule and working backward.

$$P(Z \geq z_{\alpha/2}) = \alpha/2 \quad \text{Definition of critical value.}$$

$$P(Z \leq z_{\alpha/2}) = 1 - \alpha/2 \quad \text{The Complement Rule.}$$

Work backward in Table 3 to find  $z_{\alpha/2}$ .

**Example 8.2.1** Find each of the following critical values: (a)  $z_{0.025}$ ; (b)  $z_{0.001}$



Constructing a general  $100(1 - \alpha)\%$  confidence interval for a population mean  $\mu$ .

1. Find a symmetric interval about 0 such that the probability  $Z$  lies in this interval is  $1 - \alpha$ .

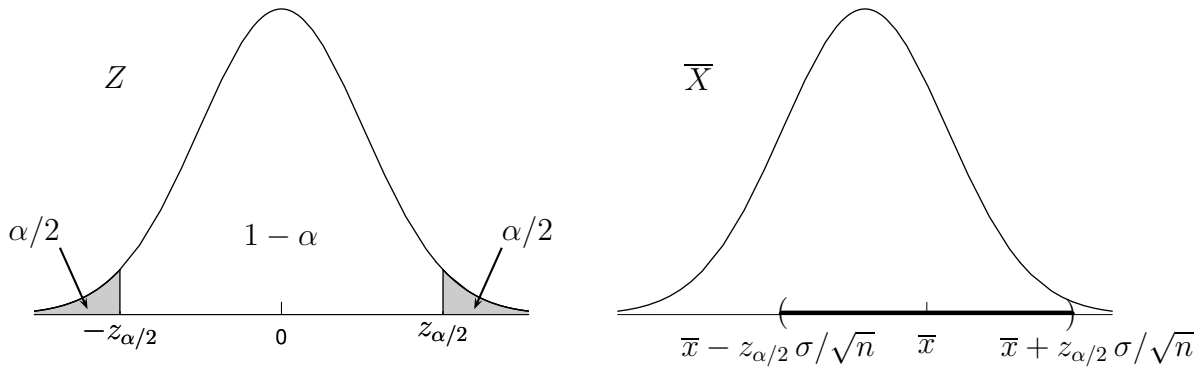
$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

2. Substitute for  $Z$ .

$$P\left(-z_{\alpha/2} < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

3. Manipulate this equation to obtain the probability statement

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$



**How to find a  $100(1 - \alpha)\%$  Confidence Interval for a Population Mean when  $\sigma$  is Known**

Given a random sample of size  $n$  from a population with mean  $\mu$ , if

1. the underlying population distribution is normal and/or  $n$  is large, and
2. the population standard deviation  $\sigma$  is known, then

a  $100(1 - \alpha)\%$  confidence interval for  $\mu$  has as endpoints the values

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

**Remarks**

1. This CI for  $\mu$  can only be used if  $\sigma$  is known.
2. If  $n$  large and  $\sigma$  unknown: some statisticians substitute  $s$  for  $\sigma$ .

This produces an approximate confidence interval.

Next section: an exact confidence interval for  $\mu$  when  $\sigma$  is unknown.

3. As confidence coefficient increases ( $n, \sigma$  constant): CI is wider.

**Example 8.2.2** Most digital video cameras are designed to operate using a battery or an AC adapter connected to an electrical outlet. Suppose the power (in volts) necessary to operate a digital video camera is normally distributed with standard deviation  $\sigma = 2.3$  volts. In a random sample of 16 digital video cameras, the sample mean power required was  $\bar{x} = 8.6$  volts. Find a 95% confidence interval for the true mean voltage necessary to operate a digital video camera.

Concepts to remember:

1. The population parameter,  $\mu$ , is *fixed*.

The confidence interval *varies* from sample to sample.

Correct statement: We are 95% confident the interval *captures* the true mean  $\mu$ .

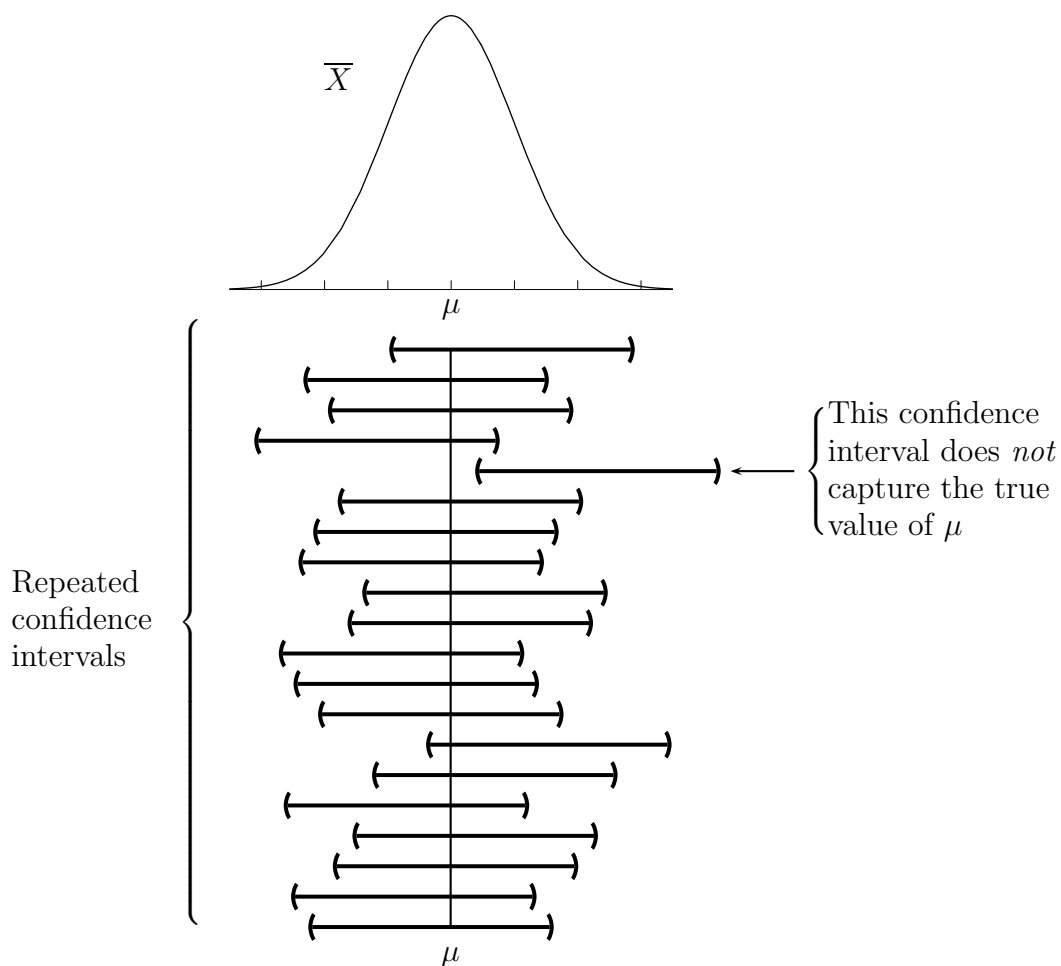
Incorrect statement: We are 95% confident  $\mu$  lies in the interval.

2. Confidence coefficient: a probability, a long-run limiting relative frequency.

In repeated samples, the proportion of confidence intervals that capture the true value of  $\mu$  approaches the confidence coefficient, in this case 0.95.

Cannot be certain about any one specific confidence interval.

The confidence is in the long-run process.



**Example 8.2.3** A company sells healing crystals designed to be worn in a pouch near your heart so that the crystal vibrations cure imbalances. Suppose the weight (in grams) of small healing crystals is normally distributed with standard deviation 12.5. A random sample of 26 small healing crystals was obtained, and the sample mean was  $\bar{x} = 203.6$ . Find a 99% confidence interval for the true mean weight of small healing crystals.

**Example 8.2.4** The English Soccer League has 20 teams and is known for its boisterous, rowdy fans. A random sample of 17 games played in this league was obtained, and the total number of goals scored was recorded for each. The sample mean was  $\bar{x} = 2.87$ . Assume the distribution of total number of goals is normal and  $\sigma = 0.68$ .

- (a) Find a 98% confidence interval for the true mean number of goals scored in an English Soccer League game.
- (b) Using the confidence interval in part (a), is there any evidence to suggest that the mean number of goals scored per game is more than 2.4?

**Example 8.2.5** Hot water for a conference center is generated in a boiler room located in the basement of the facility. During the summer, the temperature in the boiler room climbs very high and can be dangerous for employees. A random sample of summer days was obtained, and the temperature (in  $^{\circ}\text{F}$ ) was measured at noon on each day. The data are given in the following table.

114	113	117	106	111	104	118	101	116	111	106
115	105	109	100	108	109	107	108	101	104	

Assume the distribution of summer temperatures in the boiler room is normal with  $\sigma = 5.7$ .

- Find a 95% confidence interval for the true mean summer temperature at noon in the boiler room.
- If the true mean temperature is  $115^{\circ}\text{F}$  or greater, conditions are considered dangerous and a special air conditioning system will be installed. Using the confidence interval in part (a), is there any evidence to suggest that the true mean temperature is less than  $115^{\circ}\text{F}$ ?

General form of a confidence interval:

Suppose  $\hat{\theta}$  is used to estimate the population parameter  $\theta$ .

$$(\text{point estimate using } \hat{\theta}) \pm (\text{critical value}) \cdot [(\text{estimate of}) \text{ std. dev. of } \hat{\theta}].$$

Problem:

1. Usually, no control over the sample size.
2. Given summary statistics. Construct a confidence interval.
3. However, accuracy can be expressed by the width of the confidence interval.
4. Given  $\sigma$  and confidence level, find a sample size  $n$  such that the resulting CI has the desired width.

Sample size calculation:

1. Suppose  $n$  is large (unknown),  $\sigma$  is known, confidence level is  $100(1 - \alpha)\%$ .
2. Desired width  $W$ .  $B = W/2$ : bound on the error of estimation.

$B$  is half the width of the confidence interval.

3. Confidence interval endpoints:  $\bar{x} \pm \underbrace{z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}_{B = \text{bound}}$

4. Let  $B = z_{\alpha/2} (\sigma/\sqrt{n})$ , and solve for  $n$ .

The resulting formula for  $n$  is given by  $n = \left[ \frac{\sigma z_{\alpha/2}}{B} \right]^2$ .

**Remarks**

1. Consider the effects on  $n$  as one value changes (others constant):

$\sigma$  increases:

$z_{\alpha/2}$  increases:

$B$  decreases:

2. It is likely  $n$  will not be an integer. Always round up.

3. It is unlikely  $\sigma$  is known. Guess  $\sigma$  from previous experience, preliminary study.

**Example 8.2.6** The added weight of Americans has caused airlines to spend more money on fuel. An airline official would like to find a 95% confidence interval for the mean amount of fuel (in gallons) a 747-400 full of passengers needs to fly from New York to Los Angeles. The bound on the error of estimation should be 400, and assume the standard deviation is approximately 1250. How large a sample is necessary in order to achieve this accuracy?

### 8.3 A Confidence Interval for a Population Mean when $\sigma$ is Unknown

1. Confidence interval for  $\mu$  based on  $Z$ : valid only when  $\sigma$  is known (unrealistic).
2. If  $\sigma$  is known, confidence interval derived from the expression

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}.$$

Only  $\bar{X}$  contributes to the variability in this expression.

3. If  $\sigma$  is unknown, confidence interval based on a similar standardization:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}.$$

Two sources of variability:  $\bar{X}$  and  $S$ .

This random variable has a  $t$  distribution. Related to the normal distribution.

#### Properties of a $t$ Distribution

1. A  $t$  distribution is completely determined (characterized) by only one parameter,  $\nu$ , called the number of degrees of freedom (df).  $\nu$  must be a positive integer ( $\nu = 1, 2, 3, 4, \dots$ ) and there is a different  $t$  distribution corresponding to each value of  $\nu$ .
2. If  $T$  (a random variable) has a  $t$  distribution with  $\nu$  degrees of freedom, denoted  $T \sim t_\nu$ , then

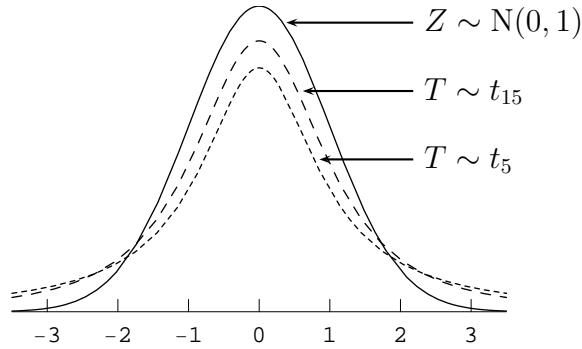
$$\mu_T = 0 \quad \text{and} \quad \sigma_T^2 = \frac{\nu}{\nu - 2}, \quad (\nu \geq 3).$$

3. Suppose  $T \sim t_\nu$ . The density curve for  $T$  is bell-shaped and centered at 0 but more spread out than the density curve for a standard normal random variable,  $Z$ . As  $\nu$  increases, the density curve for  $T$  becomes more compact and closer to the density curve for  $Z$ .



**Remarks**

1. The  $t$  distribution was derived by William Gosset in 1908.
2. A comparison of density curves:



**Definition**

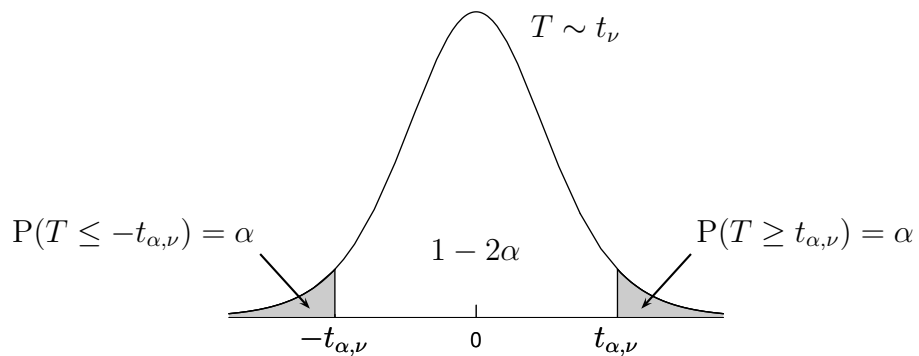
$t_{\alpha,\nu}$  is a **critical value** related to a  **$t$  distribution** with  $\nu$  degrees of freedom. If  $T \sim t_\nu$ , then  $P(T \geq t_{\alpha,\nu}) = \alpha$ .

**Remarks**

1.  $t_{\alpha,\nu}$ : a  $t$  value such that there is  $\alpha$  of the area to the right of  $t_{\alpha,\nu}$ .

$-t_{\alpha,\nu}$ : the negative critical value.

Because the  $t$  distribution is symmetric,  $P(T \leq -t_{\alpha,\nu}) = P(T \geq t_{\alpha,\nu}) = \alpha$ .

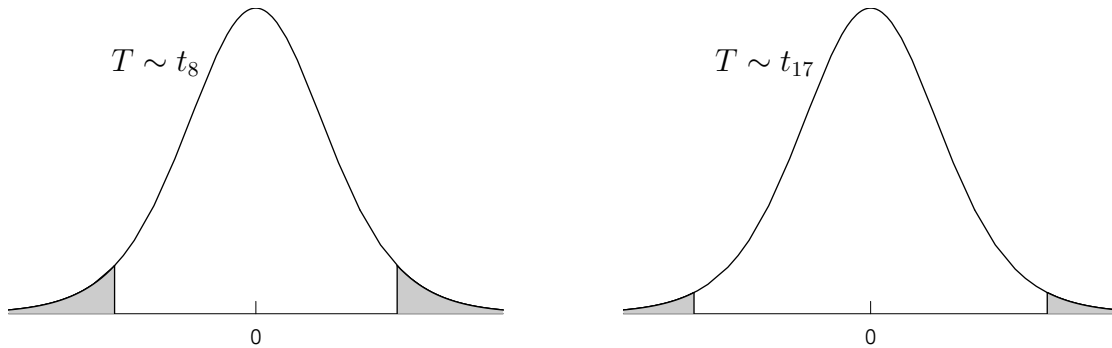


2. Remember: the  $\alpha$  in  $t_{\alpha,\nu}$  is a *placeholder*.

Any symbol could be used here to represent a right-tail probability.  $P(T \geq t_{(c),\nu}) = c$

3.  $t$  critical values can be found in Table 5.

**Example 8.3.1** Find each critical value: (a)  $t_{0.025,8}$ , (b)  $t_{0.005,17}$ .



### Remarks

1. Table 5 is very limited. Use technology where appropriate.
2. As  $\nu$  increases,  $t$  critical values approach corresponding  $Z$  critical values.

Confidence interval based on the following result.

#### 8.1 Theorem

Let  $\bar{X}$  be the mean of a random sample of size  $n$  from a normal distribution with mean  $\mu$ . The random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a  $t$  distribution with  $n - 1$  degrees of freedom.

Constructing a general  $100(1 - \alpha)\%$  confidence interval for a population mean  $\mu$  when  $\sigma$  is unknown.

1. Start with a symmetric interval about 0 such that the probability  $T$  lies in this interval is  $1 - \alpha$ .

$$P(-t_{\alpha/2, n-1} < T < t_{\alpha/2, n-1}) = 1 - \alpha$$

2. Substitute for  $T$ .

$$P\left(-t_{\alpha/2, n-1} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2, n-1}\right) = 1 - \alpha$$

3. Manipulate this equation to obtain the probability statement

$$P\left(\bar{X} - t_{\alpha/2, n-1} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2, n-1} \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

**How to find a  $100(1 - \alpha)\%$  Confidence Interval for a Population Mean when  $\sigma$  is Unknown**

Given a random sample of size  $n$  with sample standard deviation  $s$  from a population with mean  $\mu$ ; if the underlying population distribution is normal, a  $100(1 - \alpha)\%$  confidence interval for  $\mu$  has as endpoints the values

$$\bar{x} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}.$$

**Remarks**

1. This CI can be used with any sample size  $n$  ( $\geq 2$ ).

This produces an exact CI for  $\mu$ .

2. This CI for  $\mu$  is valid only if the underlying population is normal.

**Example 8.3.2** A jai alai ball, called a pelota, is the hardest and fastest ball used in any sport. All pelotas are handmade at the jai alai arena and covered with a heavy-duty nylon. Suppose the weight of pelotas is normally distributed. A random sample of 12 pelotas was obtained, and each was carefully weighed. The sample mean was  $\bar{x} = 145.7$  grams with standard deviation  $s = 1.57$  grams. Find a 95% confidence interval for the true mean weight of a pelota.

**Example 8.3.3** In a study concerning infant nourishment conducted by doctors at a large city hospital, 21 one-month-old babies were selected at random. The Ponderal Index (PI,  $\text{g}/\text{cm}^3$ ), a measure of soft-tissue development, was computed for each child. The mean was 2.62 and the standard deviation was 0.29. Assume the underlying distribution is normal, and find a 99% confidence interval for the true mean PI for one-month-old babies born at this hospital.

**Example 8.3.4** Hershey's Kisses are milk chocolate chunks wrapped in silver foil with a thin white flag. A consumer group is concerned that Hershey's has decreased the amount of chocolate in each Kiss but is still charging the same price. Suppose the weight of Hershey's Kisses is normally distributed. A random sample of Hershey's Kisses was obtained, and the weight (in grams) of each is given in the following table.

---

4.84	4.25	4.39	4.28	4.97	5.20	5.17	4.35	4.67
------	------	------	------	------	------	------	------	------

---

- (a) Find a 95% confidence interval for the true mean weight of a Hershey Kiss.
- (b) Ten years ago, Hershey's claimed that the mean weight of a Kiss was 4.75 grams. Using the confidence interval in part (a), is there any evidence to suggest that the mean has decreased?

**Example 8.3.5** A random sample of Low Earth Orbit (LEO) satellites was obtained, and the distance above the Earth (in miles) for each is given in the following table.

---

265.3	261.2	250.5	224.1	229.8	266.9	251.9	285.4	300.9	247.8
219.7	253.7	241.0	242.1	257.5	257.8	253.8	248.6	249.0	281.2
259.6	269.1	229.0							

---

Assume the underlying distribution is normal, and find a 99% confidence interval for the true mean distance above the Earth for LEO satellites.

### Remarks

1. Table 5,  $\nu$  and/or  $\alpha$  not listed: linear interpolation.
2. Sample size calculations much more complicated.

$\sigma$  is unknown,  $t_{\alpha/2, n-1}$  depends on  $n$ .

## 8.4 A Large-Sample Confidence Interval for a Population Proportion

1. Let  $p$  = true population proportion of a success.

Use the sample proportion,  $\hat{p}$ , to construct a CI for  $p$ .

2. Sample of  $n$  individuals:

$X$  = number of individuals with the characteristic, or number of successes.

3. Recall: sample proportion is a relative frequency.

$$\hat{P} = \frac{X}{n} = \frac{\text{The number of individuals with the characteristic}}{\text{The sample size}}$$

If  $n$  is large and both  $np \geq 5$  and  $n(1 - p) \geq 5$ , then  $\hat{P} \overset{\bullet}{\sim} N(p, p(1 - p)/n)$ .

4. Standardize:  $Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1 - p)}{n}}} \overset{\bullet}{\sim} N(0, 1)$

5. A symmetric interval about 0 such that the probability  $Z$  lies in this interval is  $1 - \alpha$ :

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Substitute for  $Z$ :

$$P\left(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

Problem:

$p$  appears in both numerator and denominator: sandwiching  $p$  is tricky.

Solution:

Use the sample proportion  $\hat{p}$  as a good estimator of  $p$ , only in the denominator.

**How to find a Large-Sample  $100(1 - \alpha)\%$  Confidence Interval for a Population Proportion**

Given a random sample of size  $n$ , if  $n$  is large and both  $n\hat{p} \geq 5$  and  $n(1 - \hat{p}) \geq 5$ ; a large-sample  $100(1 - \alpha)\%$  confidence interval for  $p$ , the true population proportion, has as endpoints the values

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Note: Use  $\hat{p}$  as an estimate of  $p$  to check the *non-skewness criteria*.

**Example 8.4.1** Adults with diabetes are encouraged to have their eyes examined once a year. In a random sample of 270 adults with diabetes, 215 had their eyes examined last year. Find a 95% confidence interval for the true proportion of adults with diabetes who had their eyes examined last year.



**Example 8.4.2** The composition of families has changed dramatically over the past decade. In a random sample of 510 families living in a large city, 122 were classified as single-parent.

- (a) Find a 98% confidence interval for the true proportion of single-parent families living in this city.
- (b) A conservative group claims that the proportion of single-parent families in the city is 0.30. Using the confidence interval in part (a), is there any evidence to suggest that this claim is false?

Sample size calculation:

1. Suppose  $n$  is large, confidence level is  $100(1 - \alpha)\%$ , bound on the error of estimation is  $B$ .
2. The bound on the error of estimation is the step in each direction.

$$B = z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

3. Solve for  $n$ :  $n = \hat{p}(1 - \hat{p}) \left[ \frac{z_{\alpha/2}}{B} \right]^2$

Problem:  $\hat{p}$  is unknown. Don't know  $\hat{p}$  until we have  $n$ .

Solutions:

1. Use a reasonable estimate for  $\hat{p}$  from previous experience.
2. If no prior information is available: use  $\hat{p} = 0.5$ .

This produces a very conservative, large value of  $n$ .

Remember: If  $n$  is not an integer, always round up.

**Example 8.4.3** Many Americans with monthly car-loan payments are unable to purchase any other big-ticket item. A study will be conducted to estimate the proportion of those with car-loan payments who classify these loans as a major burden. A 95% confidence interval for  $p$  with bound on the error of estimation 0.03 is needed. How large a sample size is necessary in each of the following cases?

- (a) Prior experience suggests  $\hat{p} \approx 0.17$ .
- (b) There is no prior information regarding the proportion of Americans with car-loan payments who classify these loans as a major burden.

## 8.5 A Confidence Interval for a Population Variance

1. Seems reasonable to use  $S^2$  as an estimator for  $\sigma^2$ .
2. A CI for  $\sigma^2$  is based on a new standardization and a chi-square distribution.
3. Chi-square ( $\chi^2$ ) distribution:
  - (a) Positive probability only for non-negative values.
  - (b) Focus on the properties and a method for finding critical values.

### Properties of a Chi-Square Distribution

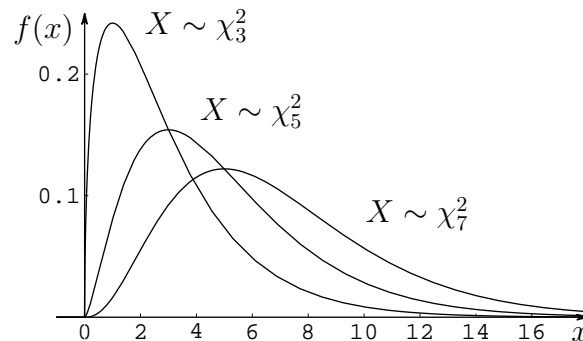
1. A chi-square distribution is completely determined by one parameter,  $\nu$ , the number of degrees of freedom, a positive integer ( $\nu = 1, 2, 3, 4, \dots$ ). There is a different chi-square distribution corresponding to each value of  $\nu$ .
2. If  $X$  has a chi-square distribution with  $\nu$  degrees of freedom, denoted  $X \sim \chi_\nu^2$ , then

$$\mu_X = \nu \quad \text{and} \quad \sigma_X^2 = 2\nu.$$

The mean of  $X$  is  $\nu$ , the number of degrees of freedom, and the variance is  $2\nu$ , twice the number of degrees of freedom.

3. Suppose  $X \sim \chi_\nu^2$ . The density curve for  $X$  is positively skewed (*not* symmetric), and as  $x$  increases it gets closer and closer to the  $x$ -axis but never touches it. As  $\nu$  increases, the density curve becomes flatter and actually looks more normal.

Density curves for several chi-square distributions:



Definition and notation for a chi-square critical value.

**Definition**

$\chi_{\alpha,\nu}^2$  is a **critical value** related to a **chi-square distribution** with  $\nu$  degrees of freedom. If  $X \sim \chi_{\nu}^2$ , then  $P(X \geq \chi_{\alpha,\nu}^2) = \alpha$ .

**Remarks**

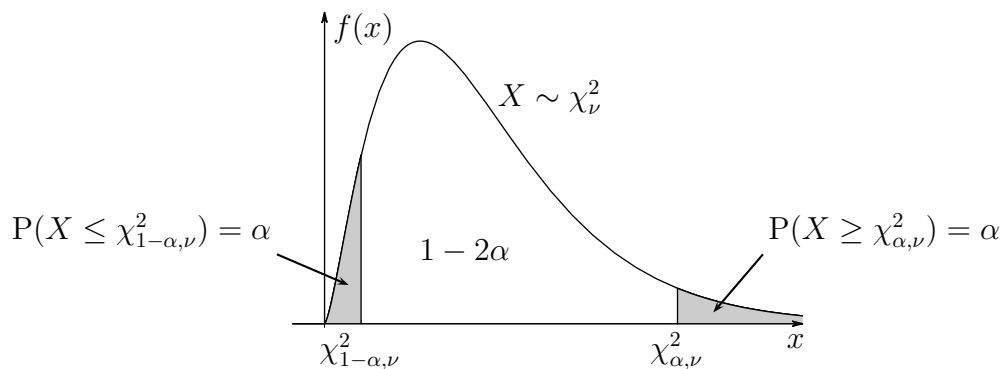
1.  $\chi_{\alpha,\nu}^2$ : a value on the measurement axis in a chi-square world with  $\nu$  degrees of freedom such that there is  $\alpha$  of the area (probability) to the right of  $\chi_{\alpha,\nu}^2$ .

There is *no* symmetry in chi-square critical values.

2. It will be necessary to find critical values denoted  $\chi_{1-\alpha,\nu}^2$  (with *large* values for  $1 - \alpha$ ).

$$P(X \geq \chi_{1-\alpha,\nu}^2) = 1 - \alpha, \quad P(X \leq \chi_{1-\alpha,\nu}^2) = \alpha.$$

Illustration:



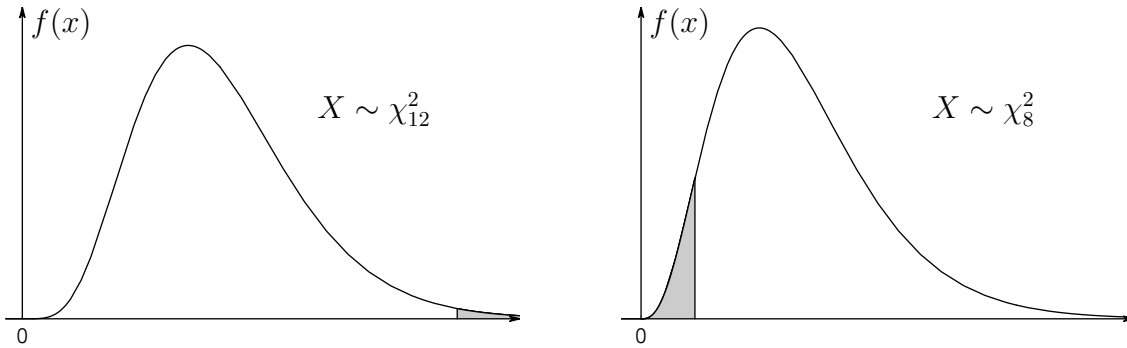
3. Table 6: selected critical values associated with various chi-square distributions.

Left-tail critical values correspond to large right-tail probabilities.

Right-tail critical values correspond to small right-tail probabilities.

4. Table 6 is very limited. Use technology where appropriate.

**Example 8.5.1** Find each critical value: (a)  $\chi_{0.01,12}^2$ , (b)  $\chi_{0.95,8}^2$ .



Confidence interval for a population variance based on the following result.

### 8.2 Theorem

Let  $S^2$  be the sample variance of a random sample of size  $n$  from a normal distribution with variance  $\sigma^2$ . The random variable

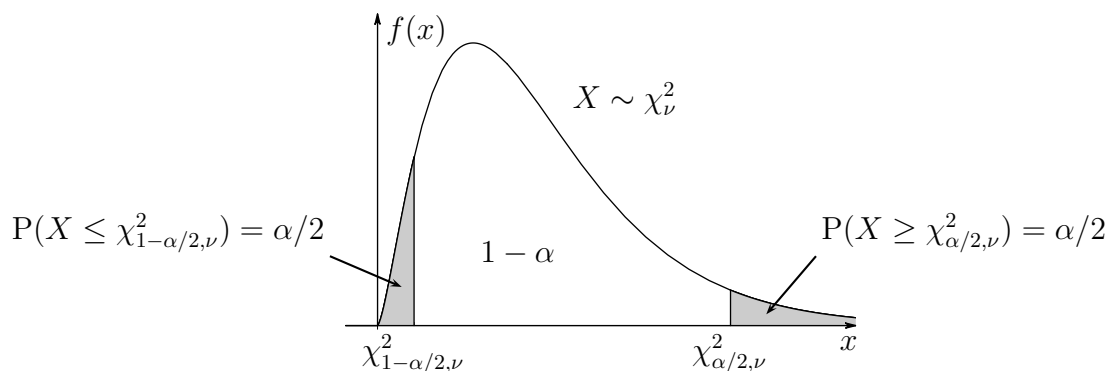
$$X = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-square distribution with  $n - 1$  degrees of freedom.

Constructing a confidence interval for a population variance  $\sigma^2$ .

1. Let  $X \sim \chi_{n-1}^2$ .

Start with an interval that captures  $1 - \alpha$  in the middle of this chi-square distribution.



$$P(\chi_{1-\alpha/2, n-1}^2 < X < \chi_{\alpha/2, n-1}^2) = 1 - \alpha$$

2. Substitute for the random variable  $X$ .

$$P\left(\chi_{1-\alpha/2, n-1}^2 < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2, n-1}^2\right) = 1 - \alpha$$

3. Manipulate this equation to sandwich  $\sigma^2$ .

$$P\left(\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2}\right) = 1 - \alpha$$

### How to find a $100(1 - \alpha)\%$ Confidence Interval for a Population Variance

Given a random sample of size  $n$  from a population with variance  $\sigma^2$ ; if the underlying population is normal, a  $100(1 - \alpha)\%$  confidence interval for  $\sigma^2$  is given by

$$\left(\frac{(n-1)s^2}{\chi_{\alpha/2, n-1}^2}, \frac{(n-1)s^2}{\chi_{1-\alpha/2, n-1}^2}\right).$$

**Remarks**

1. This CI for  $\sigma^2$  is valid only if the underlying population is normal.
2. Take the square root of each endpoint to obtain a  $100(1 - \alpha)\%$  CI for  $\sigma$ .

**Example 8.5.2** Some people believe that the length of the brown stripe in the coat of the woolly bear caterpillar can be used to predict the severity of the coming winter. A random sample of 20 woolly bear caterpillars was obtained, and the length of the brown stripe was measured (in cm) for each. The sample variance was  $0.088 \text{ cm}^2$ . Assume the underlying distribution is normal, and find a 95% confidence interval for the true population variance in the length of the brown stripe on a woolly bear caterpillar.

**Example 8.5.3** One nonsurgical treatment for carpal tunnel syndrome involves injecting the patient's wrist with a steroid, for example methylprednisolone. Patients who received this treatment were randomly selected, and the concentration of the steroid used (in mg per mL) for each is given in the following table.

---

34.2	42.6	38.2	39.8	45.0	37.2	26.1	42.1	46.0	38.0
28.2	36.4	33.5	38.0	44.5	45.6	45.3	35.3	36.8	39.2
41.2	37.8	46.1	35.1	34.4					

---

- (a) Find a 99% confidence interval for the true population variance in the concentration of steroid used in this nonsurgical procedure.
- (b) A physicians' group has established  $30 \text{ (mg per mL)}^2$  as the maximum variance in dosage. Using the confidence interval in part (a), is there any evidence that the true population variance is greater than  $30 \text{ (mg per mL)}^2$ ?



## CHAPTER 9

# Hypothesis Tests Based on a Single Sample

---

### 9.0 Introduction

Many real-world studies and experiments require a decision.

For example:

1. Is the mean amount of sodium in 1 ounce of plain potato chips more than 186 mg?
2. Is the mean diameter of a certain pipe less than 7 mm?

Focus in this chapter: the decision process.

Inference procedure: Claim, Experiment, Likelihood, and Conclusion.

Translate this into statistical terms.

**Goal:** use available information to make a decision about a parameter.

1. Yes, there is evidence to suggest that the claim is false.
2. No, there is no evidence to suggest that the claim is false.

## 9.1 The Parts of a Hypothesis Test and Choosing the Alternative Hypothesis

### Definition

In statistics, a **hypothesis** is a declaration, or claim, in the form of a mathematical statement, about the value of a specific population parameter (or about the values of several population characteristics).

**Example 9.1.1** Some examples of statistical hypotheses.

(a)  $\mu = 37$

where  $\mu$  is the population mean amount of coarse salt (in grains) on a Philly soft pretzel.

(b)  $p < 0.12$

where  $p$  is the population proportion of couples who are married by a Justice of the Peace.

(c)  $\sigma \neq 1000$

where  $\sigma$  is the population standard deviation in the amount (in dollars) of travel insurance purchased by people who plan a trip to Europe.

(d)  $\sigma^2 > 5$

where  $\sigma^2$  is the population variance in the amount (in mL) of milk found in a coconut.

**Note:**

1. A hypothesis is a claim about a population parameter, not about a sample statistic.
2.  $\mu = 10$  and  $p = 0.37$  are valid hypotheses.

$\bar{x} = 23.6$  and  $s = 17.8$  are not.

### The Four Parts of a Hypothesis Test

1. The **null hypothesis**, denoted  $H_0$ .

This is the claim (about a population parameter) assumed to be true, what is believed to be true, or the hypothesis to be tested. Sometimes referred to as the *no-change hypothesis*, this claim usually represents the status quo or existing state. There is an implied inequality in  $H_0$ ; however, the null hypothesis is written in terms of a single value (with an equal sign), for example,  $\theta = 5$ . Although it may seem strange, we usually try to *reject* the null hypothesis.

2. The **alternative hypothesis**, denoted  $H_a$ .

This statement identifies other possible values of the population parameter, or simply, a possibility not included in the null hypothesis.  $H_a$  indicates the possible values of the parameter if  $H_0$  is false. Experiments are often designed to determine whether there is evidence in favor of  $H_a$ . The alternative hypothesis represents change in the current standard or existing state.

3. The **test statistic**, denoted TS.

This statistic is a rule, related to the null hypothesis, involving the information in a sample. The *value* of the test statistic will be used to determine which hypothesis is more likely to be true,  $H_0$  or  $H_a$ .

4. The **rejection region** or **critical region**, denoted RR or CR.

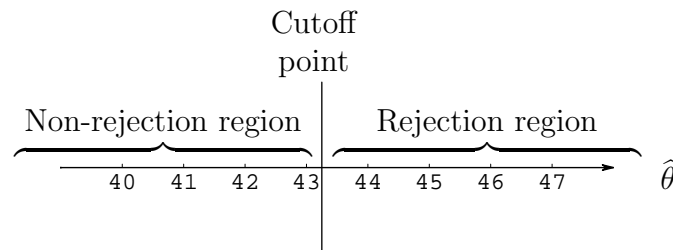
This is an interval or set of numbers specified such that if the value of the test statistic lies in the rejection region, then the null hypothesis is rejected. There is also a corresponding *non-rejection region*; if the value of the test statistic lies in this set, then we *cannot reject*  $H_0$ .

### Remarks

1. Used to decide whether there is evidence to suggest that the alternative hypothesis is true.

Use the information in the sample to determine which hypothesis is more likely:  $H_0$  or  $H_a$ .

2. The rejection region and the non-rejection region divide the world (values of the test statistic) into parts.



The value of  $\theta$  must lie in one of the regions.

3. The hypothesis test procedure is very prescriptive.

Identify the four parts, use the sample data to compute a value of the test statistic.

There are only two possible conclusions.

- (a) If the value of the test statistic lies in the rejection region, then we reject  $H_0$ .

There is evidence to suggest that the alternative hypothesis is true.

- (b) If the value of the test statistic does not lie in the rejection region, then we cannot reject  $H_0$ .

There is no evidence to suggest that the alternative hypothesis is true.

A hypothesis test is designed to prove the alternative hypothesis.

If there is no evidence in favor of  $H_a$ , this does *not* imply  $H_0$  is true.

4. A hypothesis test can only provide support in favor of  $H_a$ .

If the value of the test statistic lies in the rejection region, reject the null hypothesis. There is evidence to suggest that  $H_a$  is true.

If the value of the test statistic does not lie in the rejection region, do not reject  $H_0$ . There is no evidence to suggest that  $H_a$  is true.

We **never** *accept* the null hypothesis.

Rather, we say that we *do not reject*  $H_0$ .

5. Formal hypothesis test procedure: analogous to the four-step inference procedure.

*Claim* corresponds to  $H_0$ , a claim about a population parameter.

*Experiment* is equivalent to a value of the test statistic.

*Likelihood* is expressed in terms of the non-rejection region (likely values of the test statistic) and the rejection region (unlikely values of the test statistic).

The *Conclusion* is completely determined by the region in which the value of the test statistic lies.

If the value is in the rejection region, we reject  $H_0$ ; otherwise, we cannot reject  $H_0$ .

Consider a hypothesis test concerning  $\theta$ .

Let  $\theta_0$  be a specific value of  $\theta$ .

Null hypothesis: always stated in terms of a single value.

Only three possible alternative hypotheses.

$$\begin{array}{l} H_0 : \theta = \theta_0 \\ H_a : \theta > \theta_0 \\ \quad \theta < \theta_0 \end{array} \left. \vphantom{\begin{array}{l} H_0 \\ H_a \end{array}} \right\} \text{one-sided alternatives.}$$

$$\left. \begin{array}{l} \theta > \theta_0 \\ \theta < \theta_0 \end{array} \right\} \text{two-sided alternative.}$$

Only one alternative is selected. What are you trying to prove?

Must have a statement similar to  $H_0$  and one of the three alternatives.

**Example 9.1.2** A company sells white cranberry juice advertised to contain on average 120 calories per serving. The company recently changed the formula for this juice with the hope of decreasing the mean number of calories per serving. An experiment is conducted to determine whether the new juice has a smaller mean number of calories per serving. What null and alternative hypotheses should be used?

**Example 9.1.3** A construction company claims that the mean solar reflectance of its roof coating after three years is 0.744. A consumer group is concerned that this value is actually much lower. A long-term experiment is conducted to determine whether there is any evidence the company's claim is false. State the null and alternative hypotheses.

**Example 9.1.4** A manufacturer of bedsprings claims that the proportion of defective springs is 0.03. A company receives a large order of these springs but will send the entire shipment back if there is any evidence the proportion of defectives is greater than 0.03. A random sample of bed springs is selected, and each is carefully tested. What null and alternative hypotheses should be tested?

**Example 9.1.5** A new computer-controlled heating system has been installed in a large office building. The purpose of this system is to maintain a more constant temperature, without many fluctuations. The manufacturer claims that the variance in temperature (measured in °F) is no greater than 4. A random sample of the temperatures in this office building is obtained, and the data will be used to check the manufacturer's claim. State the null and alternative hypotheses.

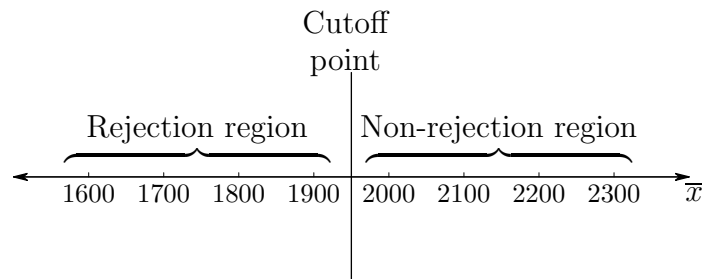
## 9.2 Hypothesis Test Errors

1. Hypothesis test: use the information in a sample to draw a conclusion about the value of a population parameter.
2. Sample data  $\implies$  test statistic value  $\implies$  decision.
3. Always a chance of making a mistake (the wrong decision). Why?
4. Need to examine what could go wrong in a hypothesis test.

**Example 9.2.1** The efficiency of night vision devices is often measured in photoresponse (in  $\mu\text{A}/\text{lm}$ ). A larger photoresponse is better. A company has designed a new night vision device and claims that the photoresponse is 2000. A consumer group selects 20 devices at random, and the photoresponse is measured for each. The sample mean is used to test the hypothesis

$$H_0: \mu = 2000 \quad \text{versus} \quad H_a: \mu < 2000$$

The consumer group has decided to use a cutoff point of 1950. A sample mean of  $\bar{x} \leq 1950$  is far enough away from 2000 that it cannot be attributed to ordinary variation about the population mean. Here is an illustration of the rejection region and the non-rejection region.



Here is what can happen when the sample data are collected and the hypothesis is tested.

1. Suppose the true population mean is 2000 or greater.
  - (a) If  $\bar{x} > 1950$ : no evidence to reject  $H_0$ . Conclusion correct.
  - (b) If  $\bar{x} \leq 1950$ : evidence to reject  $H_0$ . Conclusion incorrect.

The device is really performing as advertised.

2. Suppose the true population mean is less than 2000.
- (a) If  $\bar{x} > 1950$ : no evidence to reject  $H_0$ . Conclusion incorrect.

The device is not really performing as advertised.

- (b) If  $\bar{x} < 1950$ : evidence to reject  $H_0$ . Conclusion correct.

### Definition

1. The value of the test statistic may lie in the rejection region, but the null hypothesis is true. If we reject  $H_0$  when  $H_0$  is really true, this is called a **type I error**. The probability of a type I error is called the *significance level* of the hypothesis test and is denoted by  $\alpha$ :  $P(\text{type I error}) = \alpha$ .
2. The value of the test statistic may not lie in the rejection region, but the alternative hypothesis is true. If we do not reject the null hypothesis when  $H_a$  is really true, this is called a **type II error**. The probability of a type II error is denoted by  $\beta$ :  $P(\text{type II error}) = \beta$ .

Illustration:

		Decision	
		Reject $H_0$	Do not reject $H_0$
Truth	$H_0$	Type I error	Correct decision
	$H_a$	Correct decision	Type II error



**Example 9.2.2** A farm store sells corn seeds to growers and claims that only 3% of all seeds do not germinate. A random sample of corn seeds is obtained and tested for germination. Let  $p$  be the true population proportion of seeds that do not germinate. The information in the sample will be used to test the hypothesis

$$H_0: p = 0.03 \quad \text{versus} \quad H_a: p > 0.03.$$

If  $H_0$  is rejected, growers will not order seeds from this farm store. Describe the consequences of the decision to reject or not to reject for each truth assumption.

**Example 9.2.3** A new dehumidifier has an advertised rating of 42 pints per day. That is, the unit can remove on average 42 pints of water from the air per day. A random sample of these dehumidifiers is obtained, each unit is test, and the rating is recorded. The information in the sample will be used to test the hypothesis

$$H_0: \mu = 42 \quad \text{versus} \quad H_a: \mu < 42$$

where  $\mu$  is the mean rating of the dehumidifier. If  $H_0$  is rejected, the group conducting the test will issue a warning to consumers. Discuss the consequences of the decision to reject or not to reject for each truth assumption.

**Example 9.2.4** Hospital patients who require maintenance fluids are often given a saline or Ringer's lactated solution in various concentrations. An insurance company is investigating the concentration of saline used in adults, with a focus on the variance in concentration. A random sample of adult patients who received a saline solution at various hospitals is obtained, and the concentration of each solution is recorded (in ml/kg/hour). The sample variance is used to test the hypothesis

$$H_0: \sigma^2 = 4 \quad \text{versus} \quad H_a: \sigma^2 > 4$$

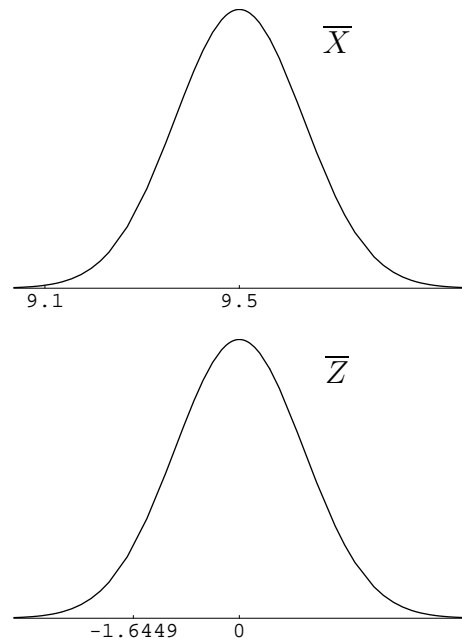
where  $\sigma^2$  is the population variance in concentration of saline. If  $H_0$  is rejected, the company will raise the malpractice insurance rates for the hospital. Discuss the consequences of the decision to reject or not to reject for each truth assumption.

### 9.3 Hypothesis Tests Concerning a Population Mean When $\sigma$ is Known

Recall:

1. Every hypothesis test has four parts.
2. Null hypothesis: stated in terms of a population parameter, represents the current state.

**Example 9.3.1** A wheelbarrow used by contractors includes a solid steel axle, extra-strong aluminum and steel frame, and a brick-carrier adapter. The polyethylene tray is designed to have a mean thickness of 9.5 mm (with standard deviation  $\sigma = 0.75$ ). If the mean thickness of the tray is less than 9.5, the wheelbarrow will not last long, and contractors will be disenfranchised. Thirty-two wheelbarrows were randomly selected, and the thickness of each tray was carefully measured (in mm). The sample mean thickness was  $\bar{x} = 9.1$ . Conduct a hypothesis test to determine whether there is any evidence to suggest that the mean thickness of the wheelbarrow tray is less than 9.5. Use  $\alpha = 0.05$ .



**Remarks** In any hypothesis test: assume  $H_0$  is true and consider the likelihood of the sample outcome (expressed as a single value of the test statistic).

1. If the value of the test statistic is reasonable: cannot reject  $H_0$ .
2. If the value of the test statistic is unlikely under  $H_0$ , reject  $H_0$ .

**Hypothesis Tests Concerning a Population Mean When  $\sigma$  is Known**

Given a random sample of size  $n$  from a population with mean  $\mu$ , assume

1. the underlying population is normal and/or  $n$  is large, and
2. the population standard deviation  $\sigma$  is known.

A hypothesis test about the population mean  $\mu$  with significance level  $\alpha$  has the form:

$$H_0: \mu = \mu_0$$

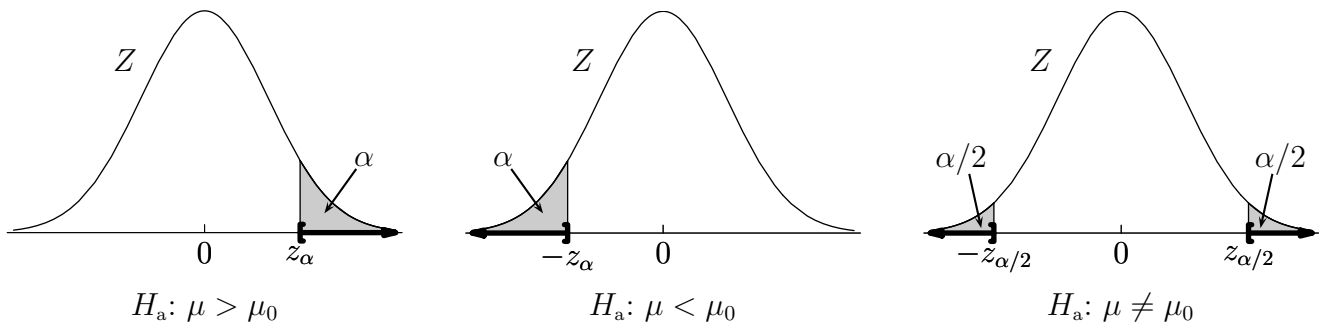
$$H_a: \mu > \mu_0, \quad \mu < \mu_0, \quad \text{or} \quad \mu \neq \mu_0$$

$$\text{TS: } Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

$$\text{RR: } Z \geq z_\alpha, \quad Z \leq -z_\alpha, \quad \text{or} \quad |Z| \geq z_{\alpha/2}$$

**Remarks**

1.  $\mu_0$ : fixed, hypothesized value of the population mean  $\mu$ .
2. Use only one alternative hypothesis and the corresponding rejection region.



3. For a two-sided alternative hypothesis:  $|Z| \geq z_{\alpha/2}$  is short for  $Z \geq z_{\alpha/2}$  or  $Z \leq -z_{\alpha/2}$ .
4. For a given significance level, corresponding critical value is found using Table 3, backward.

Common values for  $\alpha$  are 0.05, 0.025, and 0.01.

5. This test procedure can be used *only* if  $\sigma$  is known.

If  $n$  is large and  $\sigma$  is unknown, some statisticians use  $s$  in place of  $\sigma$ .

This produces an *approximate* test statistic.

Section 9.5 presents an *exact* test procedure concerning  $\mu$  when  $\sigma$  is unknown.

**Example 9.3.2** A company sells a strong commercial floor cleaner and claims that the flashpoint (the lowest temperature at which the vapor of a combustible liquid can be ignited in air) is 200 °F. A random sample of cleaner was obtained, and the flashpoint of each was measured (in °F). The sample mean was  $\bar{x} = 197.2$ . Assume the distribution of flashpoints is normal and  $\sigma = 10$ . Conduct a hypothesis test to determine whether there is any evidence the mean flashpoint is less than 200. Use  $\alpha = 0.01$ .

**Example 9.3.3** A handcrafted 18ct gold wedding ring is designed to weigh 3.1 grams. If it weighs less than 3.1 grams, the customer is paying too much, and if it weighs more than 3.1 grams, the jeweler is losing money. A random sample of these wedding rings was obtained, and the weight of each (in grams) is given in the following table.

3.18	3.30	3.12	3.17	3.22	3.22	3.13	3.45	3.45	2.92
2.98	3.32	2.82	3.17	3.36	3.13	3.00	3.08	3.13	2.78

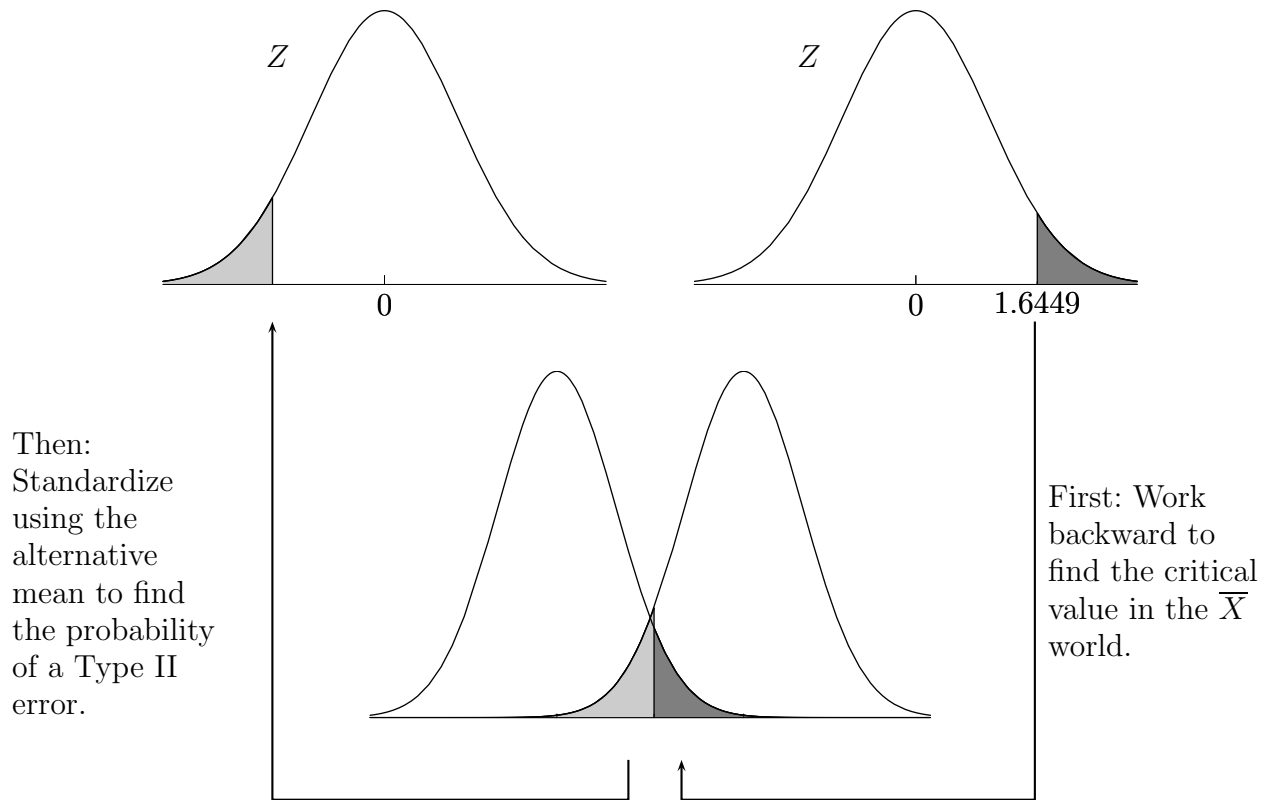
Assume the distribution of weights is normal and  $\sigma = 0.2$ . Conduct a hypothesis test to determine whether there is any evidence the mean weight of this type of wedding ring is different from 3.1. Use  $\alpha = 0.01$ .

**Example 9.3.4** In order for a clothes washer to qualify for ENERGY STAR, it must have a Modified Energy Factor (MEF, a dimensionless quantity) of 1.42 or greater. Thirty-eight Speed Queen commercial washers were selected at random, and the MEF for each was measured. The sample mean was  $\bar{x} = 1.228$ . Assume the distribution of MEF is normal and  $\sigma = 0.5$ . Conduct a hypothesis test to determine whether there is any evidence the mean MEF of this Speed Queen washer is less than 1.42. Use  $\alpha = 0.01$ .



**Example 9.3.5** Technical specifications for an architectural glazing resin indicate that the impact strength at  $0^\circ\text{C}$  is  $107\text{ J/m}$ . A new additive is developed, designed to increase this impact strength. A quality-control inspector selects 40 cans of the new resin at random. A hypothesis test is conducted to determine whether there is any evidence that the mean impact strength of this resin is greater than  $107\text{ J/m}$ . Let  $\alpha = 0.05$ , assume  $\sigma = 8\text{ J/m}$ , and find the probability of a type II error if the true mean impact strength is  $111\text{ J/m}$ .

Example (continued)



Then:  
Standardize  
using the  
alternative  
mean to find  
the probability  
of a Type II  
error.

First: Work  
backward to  
find the critical  
value in the  $\bar{X}$   
world.

**Remarks**

1. Inverse relationship between  $\alpha$  and  $\beta$ .
2. Focus on the method for finding a type II error.

Don't try to memorize a formula.

---

## 9.4 $p$ Values

- Using a value of the test statistic and a rejection region:

one way to determine the likelihood of an observation, how far in the tail of the distribution it is.

Using a fixed rejection region can lead to a peculiar dilemma.

- $p$  value is another way to convey likelihood.

**Example 9.4.1** Good water quality in a swimming pool is often measured by the free available chlorine (FAC). A pool and spa manager claims that the FAC should be 2.0 ppm. A random sample of 36 residential pools was obtained, and the FAC was measured in each. The sample mean was  $\bar{x} = 1.72$ , and assume  $\sigma = 0.95$ . Is there any evidence to suggest that the mean FAC is less than 2.0 ppm?

$\alpha$	Rejection region	Conclusion
0.10	$Z \leq -z_{0.10} = -1.2816$	
0.05	$Z \leq -z_{0.05} = -1.6449$	
0.025	$Z \leq -z_{0.025} = -1.9600$	
0.01	$Z \leq -z_{0.01} = -2.3263$	

---

**Definition**

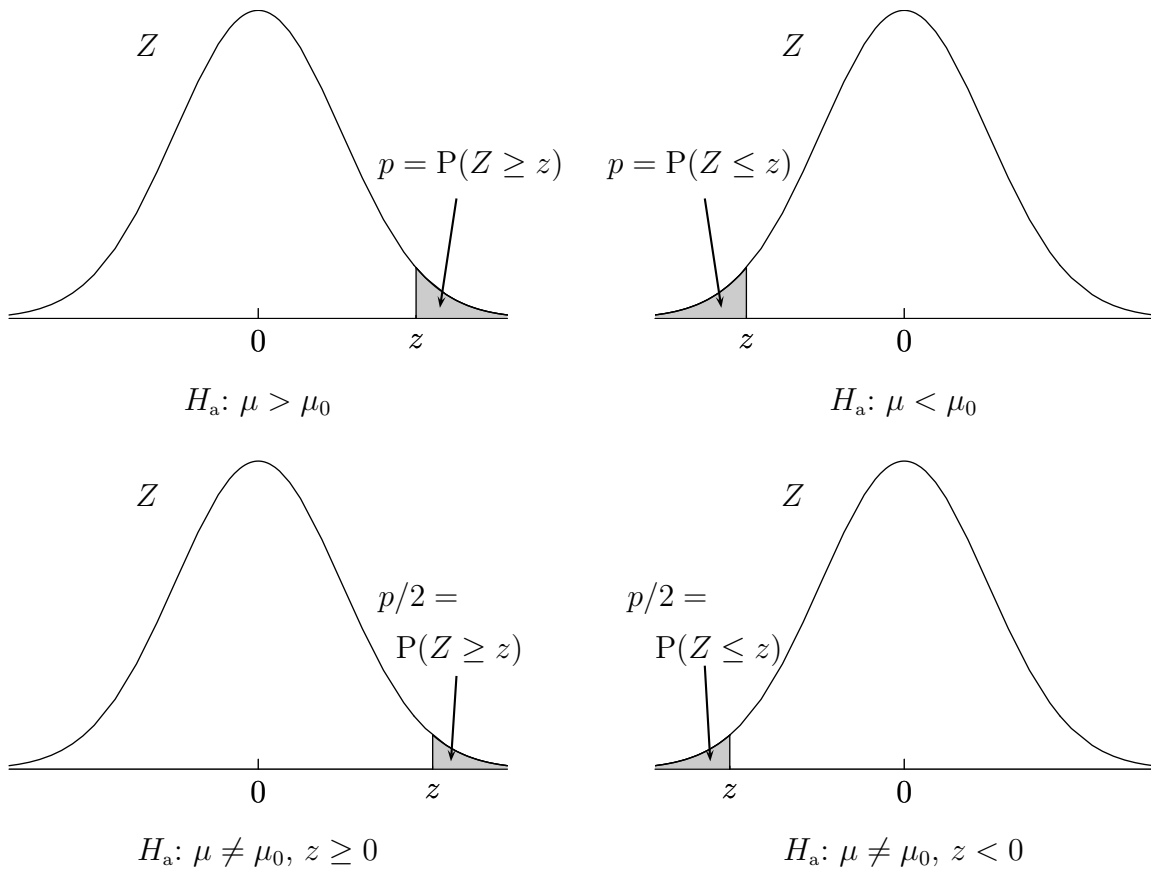
The ***p* value**, denoted  $p$ , for a hypothesis test is the smallest significance level (value of  $\alpha$ ) for which the null hypothesis,  $H_0$ , can be rejected.

*p* value: a tail probability that conveys likelihood.

The tail is determined by the alternative hypothesis.

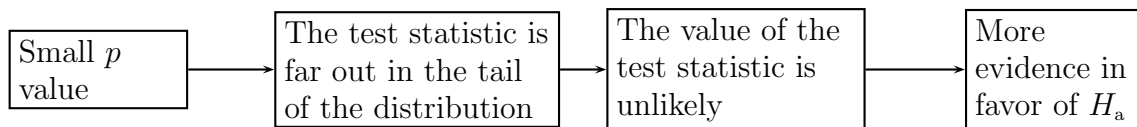
Consider a hypothesis test concerning a population mean  $\mu$  when  $\sigma$  is known.

Alternative hypothesis	Probability definition
$H_a: \mu > \mu_0$	$p = P(Z \geq z)$
$H_a: \mu < \mu_0$	$p = P(Z \leq z)$
$H_a: \mu \neq \mu_0$	$p/2 = P(Z \geq z)$ if $z \geq 0$ , $p/2 = P(Z \leq z)$ if $z < 0$



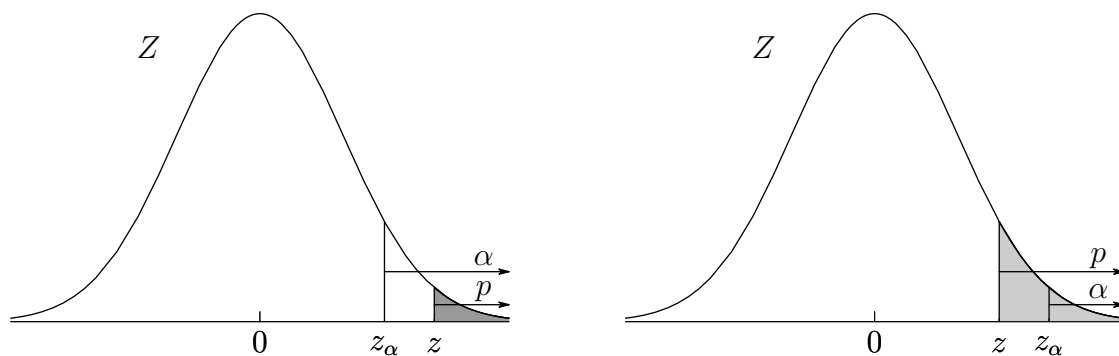
**Remarks**

1.  $p$  value conveys the strength of the evidence in favor of  $H_a$ .
2. The smaller  $p$  is, the farther out in the tail the test statistic is, the more unlikely the value of the test statistic is, and the more evidence there is in favor of  $H_a$ .



3. *Small* usually means  $p \leq 0.05$ .
4. Suppose  $\alpha$  is the significance level.
  - (a)  $p \leq \alpha$ : Reject  $H_0$ . Evidence to suggest that  $H_a$  is true.
  - (b)  $p > \alpha$ : Do not reject  $H_0$ . No evidence to suggest that  $H_0$  is false.

Illustration:



$p$  value: the smallest significance level for which  $H_0$  can be rejected.

**Example 9.4.2** The mean inside perimeter (in mm) for a police handcuff should be 185. If the inside perimeter is larger, then a suspect may be able to slip his/her hands out of the handcuffs. A random sample of 20 police handcuffs was obtained, and the inside perimeter of each was carefully measured. The sample mean was  $\bar{x} = 182.1$ . Assume the underlying distribution is normal and  $\sigma = 4.56$ . Is there any evidence to suggest the true mean inside perimeter is greater than 185 mm? Find the  $p$  value associated with this hypothesis test.

**Example 9.4.3** Some natural-food stores claim that borage oil helps in treating arthritis. The medicinal benefits are attributed to the fat gamma linolenic acid (GLA) in borage oil. A manufacturer of a 500-mg borage oil tablet claims that the mean amount of GLA per tablet is 115 mg. A random sample of 500-mg borage oil tablets was obtained, and the amount of GLA in each was measured. The data (in mg) are given in the following table.

---

110.2	120.8	121.3	127.2	88.8	102.7	111.4	106.1	118.2	107.8	115.2
115.8	104.7	121.8	122.2	111.1	112.3	105.1	106.4	100.7	121.2	124.1
106.0	120.1	119.9	123.7	98.2	106.9	120.6	102.8	105.5	106.5	114.5

---

Assume  $\sigma = 10$ . Is there any evidence to suggest the true mean amount of GLA is different from 115 mg? Use  $\alpha = 0.01$ . Find the  $p$  value associated with this hypothesis test.

## 9.5 Hypothesis Tests Concerning a Population Mean when $\sigma$ is Unknown

Hypothesis test procedure based on the following result.

### 9.1 Theorem

Given a random sample of size  $n$  from a normal distribution with mean  $\mu$ , the random variable

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

has a  $t$  distribution with  $n - 1$  degrees of freedom.

### Hypothesis Tests Concerning a Population Mean when $\sigma$ is Unknown

Given a random sample of size  $n$  from a normal population with mean  $\mu$ , a hypothesis test concerning the population mean  $\mu$  with significance level  $\alpha$  has the form:

$$H_0: \mu = \mu_0$$

$$H_a: \mu > \mu_0, \quad \mu < \mu_0, \quad \text{or} \quad \mu \neq \mu_0$$

$$\text{TS: } T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$\text{RR: } T \geq t_{\alpha, n-1}, \quad T \leq -t_{\alpha, n-1}, \quad \text{or} \quad |T| \geq t_{\alpha/2, n-1}$$

### Remarks

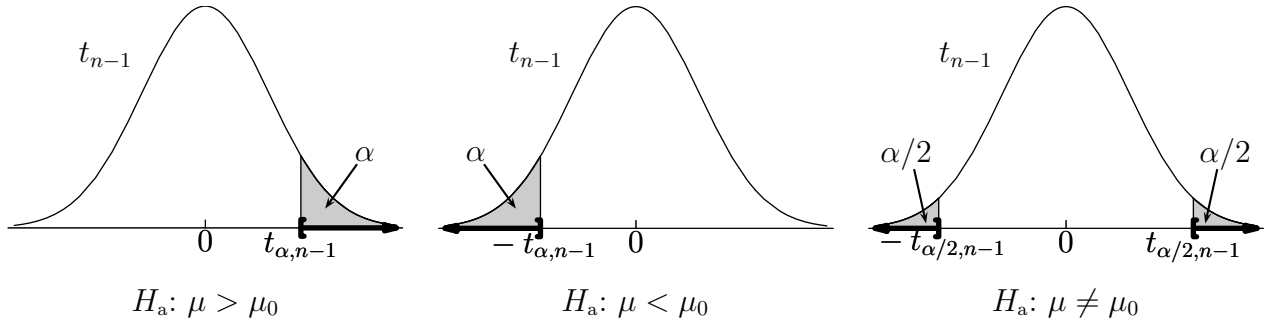
1. Called a small-sample test, or  $t$  test.

Valid for any sample size as long as underlying population is normal.

2. Table 5: selected critical values associated with various  $t$  distributions.



3. Rejection regions:



**Example 9.5.1** A company that sells a mesh motorcycle ramp for loading a bike onto a truck bed claims that the mean capacity is 1000 pounds. A random sample of 14 motorcycle ramps was obtained, and a destructive test was used to determine the capacity of each (in pounds). The summary statistics were  $\bar{x} = 989.2$  and  $s = 15.8$ . Is there any evidence to suggest that the true mean capacity of this type of motorcycle ramp is less than 1000 pounds? Assume the underlying population is normal, and use  $\alpha = 0.05$ .

**Example 9.5.2** The manufacturer of a miter saw claims that it has a new quick-stop feature. That is, once the arm angle reaches 45 degrees, the blade will stop turning within 7 seconds. A random sample of 10 miter saws was obtained, and the quick-stop feature was tested in each. The summary statistics (in seconds) were  $\bar{x} = 7.22$  and  $s = 0.58$ . Is there any evidence to suggest that the mean time for the blade to stop turning is greater than 7 seconds? Assume the underlying population is normal, and use  $\alpha = 0.01$ .

**Example 9.5.3** A 100% recycled, heavy-duty, deluxe parking block is designed to weigh 26 pounds. If it weighs more than 26 pounds, the block is very difficult to install, and if it weighs less than 26 pounds, the block does not perform as expected. A random sample of these parking blocks was obtained, and each was carefully weighed. The data (in pounds) are given in the following table.

26.12	26.51	26.59	26.43	25.99	26.93
-------	-------	-------	-------	-------	-------

Is there any evidence to suggest that the true mean weight is different from 26 pounds? Assume the underlying population is normal, and use  $\alpha = 0.05$ .

**Note:** Using Table 5, we can only bound the  $p$  value associated with a  $t$  test.

### How To Bound the $p$ Value for a $t$ Test

Suppose  $t$  is the value of the test statistic in a one-sided hypothesis test.

1. Select the row in Table 5 corresponding to  $n - 1$ , the number of degrees of freedom associated with the test.
2. Place  $|t|$  in this ordered list of critical values.
3. To compute  $p$ :
  - (a) If  $|t|$  is between two critical values in the  $n - 1$  row, then the  $p$  value is bounded by the corresponding significance levels.
  - (b) If  $|t|$  is greater than the largest critical value in the  $n - 1$  row, then  $p < 0.0001$  (the smallest significance level in the table).
  - (c) If  $|t|$  is less than the smallest critical value in the  $n - 1$  row, then  $p > 0.20$  (the largest significance level in the table).

If  $t < 0$  for a right-tailed test, or  $t > 0$  for a left-tailed test, then  $p > 0.5$ . If the hypothesis test is two-sided, this method produces a bound on  $p/2$ .

**Example 9.5.4** A certain surfboard is designed to be 114 inches long. A random sample of 21 surfboards was obtained, and the length of each was measured (in inches). The summary statistics were  $\bar{x} = 114.8$  and  $s = 1.822$ . Assume the underlying population is normal.

- (a) Is there any evidence to suggest that the mean length of these surfboards is greater than 114 inches? Use  $\alpha = 0.05$ .
- (b) Find bounds on the  $p$  value associated with this hypothesis test.

**Example 9.5.5** The mean sheen rating for a satin-finish paint used on interior walls in a government building is specified to be 27.5 units. A random sample of government buildings was obtained, and the sheen rating on a first-floor wall was measured for each. The data are given in the following table.

27.1	32.4	27.6	25.6	30.2	34.8	33.6	26.9	30.9
29.3	32.2	32.8	30.3	34.6	23.9	26.5	28.8	34.6

Assume the underlying population is normal.

- Is there any evidence to suggest that the mean sheen rating is different from 27.5? Use  $\alpha = 0.01$ .
- Find bounds on the  $p$  value associated with this hypothesis test.

## 9.6 Large-Sample Hypothesis Tests Concerning a Population Proportion

1. Many experiments and studies concerning a population proportion.
2. Use the sample proportion,  $\hat{p}$ , as an estimate of the population proportion  $p$ .
3. Recall: If

$$\hat{P} = \frac{X}{n} = \frac{\text{The number of individuals with the characteristic}}{\text{The sample size}}$$

and if  $n$  is large and both  $np \geq 5$  and  $n(1 - p) \geq 5$ , then

$$\hat{P} \rightsquigarrow N(p, p(1 - p)/n).$$

### Large-Sample Hypothesis Tests Concerning a Population Proportion

Given a random sample of size  $n$ , a large-sample hypothesis test concerning the population proportion  $p$  with significance level  $\alpha$  has the form:

$$H_0: p = p_0$$

$$H_a: p > p_0, \quad p < p_0, \quad \text{or} \quad p \neq p_0$$

$$\text{TS: } Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$\text{RR: } Z \geq z_\alpha, \quad Z \leq -z_\alpha, \quad \text{or} \quad |Z| \geq z_{\alpha/2}$$

### Remarks

1. Valid test as long as  $np_0 \geq 5$  and  $n(1 - p_0) \geq 5$ .
2. Critical values from the standard normal distribution,  $Z$ .

**Example 9.6.1** A study was conducted to examine the proportion of children in grades K–4 who suffer from at least one sleep-related problem. A random sample of 494 children in grades K–4 was obtained, and interviews with parents, teachers, and the children were used to determine that 183 suffered from at least one sleep-related problem. Is there any evidence to suggest the true proportion of children in grades K–4 who suffer from at least one sleep-related problem is greater than 0.32? Use  $\alpha = 0.01$ .

**Example 9.6.2** The fixation rate for a company is defined to be the proportion of university graduates who stay with the company for more than five years. A random sample of 542 employees (university graduates) from U.S. companies was obtained, and 461 had remained with the company for more than five years. Conduct a hypothesis test to determine whether the true fixation rate of U.S. companies is different from 0.82, the reported fixation rate for Japanese corporations. Find the  $p$  value associated with this test.

## 9.7 Hypothesis Tests Concerning a Population Variance or Standard Deviation

1. Many real-world, practical decisions involve variability, or a population variance.
2.  $S^2$  is used as an estimator for  $\sigma^2$ .
3. Recall:  $\frac{(n - 1)S^2}{\sigma^2}$  has a chi-square distribution with  $n - 1$  degrees of freedom.

### Hypothesis Test Concerning a Population Variance

Given a random sample of size  $n$  from a normal population with variance  $\sigma^2$ , a hypothesis test concerning the population variance  $\sigma^2$  with significance level  $\alpha$  has the form:

$$H_0: \sigma^2 = \sigma_0^2$$

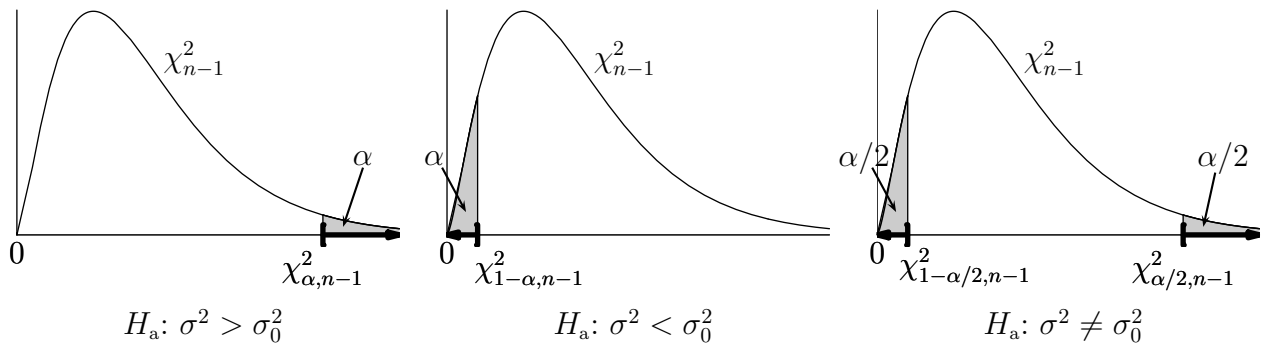
$$H_a: \sigma^2 > \sigma_0^2, \quad \sigma^2 < \sigma_0^2, \quad \text{or} \quad \sigma^2 \neq \sigma_0^2$$

$$\text{TS: } X^2 = \frac{(n - 1)S^2}{\sigma_0^2}$$

$$\text{RR: } X^2 \geq \chi_{\alpha, n-1}^2, \quad X^2 \leq \chi_{1-\alpha, n-1}^2, \quad \text{or} \quad X^2 \leq \chi_{1-\alpha/2, n-1}^2 \quad \text{or} \quad X^2 \geq \chi_{\alpha/2, n-1}^2$$

### Remarks

1.  $X$  (a Greek capital chi) is a random variable.  $\chi$  (a Greek small chi) is a specific value.
2. Test valid for any sample size, as long as the underlying population is normal.
3. Table 6: selected critical values associated with various chi-square distributions.
4. Rejection regions:



**Example 9.7.1** The cooling system in a certain automobile is designed to hold 7 quarts of antifreeze and water. A random sample of these cars was obtained, and the amount of coolant in each was measured (in quarts). The sample mean was  $\bar{x} = 7.1$  and the sample variance was  $s^2 = 0.42$ . Is there any evidence to suggest that the population variance is greater than 0.30? Assume the underlying population is normal, and use  $\alpha = 0.05$ .



**Example 9.7.2** Coal used by electric power plants contains various amounts of mercury. A random sample of coal used by a power plant in the Northeast was obtained. The amount of mercury was measured (in pounds per trillion Btu) for each. The data are given in the following table.

---

12.0	6.2	1.9	13.7	8.7	6.2	9.7	4.9	9.5	6.1
15.3	9.8	12.8	8.3	3.4	6.9	5.0	15.0	22.8	9.8
3.2	10.7	0.6	9.4	7.6					

---

Assume the underlying population is normal.

- (a) Is there any evidence to suggest that the true population variance is greater than 16? Use  $\alpha = 0.01$ .
- (b) Find bounds on the  $p$  value associated with this test.



## CHAPTER 10

# Confidence Intervals and Hypothesis Tests Based on Two Samples or Treatments

---

### 10.0 Introduction and Notation

1. Many studies to compare two population parameters.
2. Modify the single-sample procedures.

Notation:

	Population parameters			
	Mean	Variance	Standard deviation	Proportion
Population 1	$\mu_1$	$\sigma_1^2$	$\sigma_1$	$p_1$
Population 2	$\mu_2$	$\sigma_2^2$	$\sigma_2$	$p_2$

	Sample statistics				
	Sample size	Mean	Variance	Standard deviation	Proportion
Sample from population 1	$n_1$	$\bar{x}_1$	$s_1^2$	$s_1$	$\hat{p}_1$
Sample from population 2	$n_2$	$\bar{x}_2$	$s_2^2$	$s_2$	$\hat{p}_2$

To compare two population parameters, we often consider a difference.

1. To compare  $\mu_1$  and  $\mu_2$ : consider the difference  $\mu_1 - \mu_2$ .
2. To compare  $p_1$  and  $p_2$ : consider the difference  $p_1 - p_2$ .

Why consider a difference?

1. A typical relationship between two population parameters can be written in terms of a difference.

Example:

Standard notation		Difference notation
$\mu_1 = \mu_2$	is equivalent to	$\mu_1 - \mu_2 = 0$
$\mu_1 > \mu_2$	is equivalent to	$\mu_1 - \mu_2 > 0$
$\mu_1 < \mu_2$	is equivalent to	$\mu_1 - \mu_2 < 0$

$H_0: \mu_1 - \mu_2 = 0$  corresponds to a test of  $H_0: \mu_1 = \mu_2$ .

$H_a: \mu_1 - \mu_2 > 0$  is equivalent to  $H_a: \mu_1 > \mu_2$ .

Hypothesized difference between the two means may be nonzero.

$H_0: \mu_1 = \mu_2 + 5$  is equivalent to  $H_0: \mu_1 - \mu_2 = 5$ .

2. A difference like  $\mu_1 - \mu_2$  is a *single* population parameter.

Use  $\bar{X}_1 - \bar{X}_2$  to estimate  $\mu_1 - \mu_2$ .

New assumptions in this case.

### Definition

1. Two samples are **independent** if the process of selecting individuals or objects in sample 1 has no effect on, or no relation to, the selection of individuals or objects in sample 2. If the samples are not independent, they are **dependent**.
2. A **paired** data set is the result of matching each individual or object in sample 1 with a *similar* individual or object in sample 2. The most common experiment in which paired data is obtained involves a *before* and an *after* measurement on each individual or object. Each *before* observation is matched, or paired, with an *after* observation.

## 10.1 Comparing Two Population Means using Independent Samples when Population Variances are Known

1.  $\bar{X}_1$  is a good estimator for  $\mu_1$ .  $\bar{X}_2$  is a good estimator for  $\mu_2$ .

Reasonable to use  $\bar{X}_1 - \bar{X}_2$  to estimate the parameter  $\mu_1 - \mu_2$ .

2. Need properties of the estimator, or the distribution of the random variable,  $\bar{X}_1 - \bar{X}_2$ .

Two-sample  $Z$  Test Assumptions

Suppose

1.  $\bar{X}_1$  is the mean of a random sample of size  $n_1$  from a normal population with mean  $\mu_1$  and variance  $\sigma_1^2$ .
2.  $\bar{X}_2$  is the mean of a random sample of size  $n_2$  from a normal population with mean  $\mu_2$  and variance  $\sigma_2^2$ .
3. The samples are independent.

### Properties of $\bar{X}_1 - \bar{X}_2$

If the two-sample  $Z$  test assumptions are true, then the random variable  $\bar{X}_1 - \bar{X}_2$  has the following properties.

1.  $E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

$\bar{X}_1 - \bar{X}_2$  is an unbiased estimator of the parameter  $\mu_1 - \mu_2$ . The distribution is centered at  $\mu_1 - \mu_2$ .

2.  $\text{Var}(\bar{X}_1 - \bar{X}_2) = \sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$  and the standard deviation is

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

3. The distribution of  $\bar{X}_1 - \bar{X}_2$  is normal.

If the underlying distributions are not known, but both  $n_1$  and  $n_2$  are large, then  $\bar{X}_1 - \bar{X}_2$  is approximately normal (by the Central Limit Theorem).

### Hypothesis Tests Concerning Two Population Means when Population Variances are Known

Given two independent random samples, the first of size  $n_1$  from a population with mean  $\mu_1$  and the second of size  $n_2$  from a population with mean  $\mu_2$ , assume

1. the underlying populations are normal and/or both sample sizes are large, and
2. the population variances,  $\sigma_1^2$  and  $\sigma_2^2$ , are known.

A hypothesis test concerning two population means, in terms of the difference in means  $\mu_1 - \mu_2$ , with significance level  $\alpha$ , has the form:

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

$$H_a: \mu_1 - \mu_2 > \Delta_0, \quad \mu_1 - \mu_2 < \Delta_0, \quad \text{or} \quad \mu_1 - \mu_2 \neq \Delta_0$$

$$\text{TS: } Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$\text{RR: } Z \geq z_\alpha, \quad Z \leq -z_\alpha, \quad \text{or} \quad |Z| \geq z_{\alpha/2}$$

#### Remarks

1.  $\Delta_0$ : the fixed, hypothesized difference in means. Usually  $\Delta_0 = 0$ .

$H_0: \mu_1 - \mu_2 = 0$  is equivalent to  $H_0: \mu_1 = \mu_2$ .

$\Delta_0$  may be some nonzero value.

2. Remember: use only one alternative hypothesis and the corresponding rejection region.

$z$  critical values are from the standard normal distribution.

3. This two-sample  $Z$  test can be used *only* if both population variances are known.

If  $\sigma_1^2, \sigma_2^2$  unknown, sample sizes large: some statisticians substitute  $s_1^2$  for  $\sigma_1^2$  and  $s_2^2$  for  $\sigma_2^2$ .

This produces an *approximate* test statistic.

**Example 10.1.1** Steel tie-bolts that join aircraft wheels are routinely inspected and replaced if a crack is detected. The length and depth of each crack in a tie-bolt is always measured and recorded. Independent random samples of cracked tie-bolts from two types of airplanes were obtained, and the length of each crack was recorded (in mm). The summary statistics and known variances are given in the following table.

Airplane	Sample size	Sample mean	Population variance
Airbus A380	21	1.241	0.21
Boeing 747	17	1.393	0.08

Is there any evidence to suggest that the true mean length of a crack found in tie-bolts is different in Airbus A380's and Boeing 747's? Use  $\alpha = 0.05$  and assume that each underlying distribution is normal.

**Example 10.1.2** A company claims that a standard fluorescent light with an electronic ballast lasts longer than one with a magnetic ballast. Independent random samples of both types of fluorescent light fixtures were obtained, and the lifetime (in hours) of each light was measured. The summary statistics and known standard deviations are given in the following table.

Ballast	Sample size	Sample mean	Population standard deviation
Electronic	36	796.1	23.5
Magnetic	42	783.9	30.7

- (a) Is there any evidence to suggest that the true mean lifetime is greater for fluorescent lights with an electronic ballast than for those with a magnetic ballast? Use  $\alpha = 0.01$ .
- (b) Find the  $p$  value associated with this experiment.



**Example 10.1.3** Activated charcoal (Type 1), currently used in water purification systems to absorb impurities and chlorine, has a mean surface area of  $1000 \text{ m}^2/\text{g}$ . A company claims that a new manufacturing technique results in more porous charcoal (Type 2), which therefore has a larger mean surface area. Independent random samples of both types of activated charcoal were obtained, and the surface area of each was measured (in  $\text{m}^2/\text{g}$ ). The data are given in the following table.

Type 1 activated charcoal

1003	1044	1003	1015	976	965	939	970	994
1110	944	889						

Type 2 activated charcoal

983	1097	1021	1119	1009	1065	1067	977	1129
1032	986	1146	989	1068	1080			

Assume both underlying populations are normal,  $\sigma_1 = 50$  and  $\sigma_2 = 55$ . Is there any evidence to suggest that Type 2 activated charcoal has a larger mean surface area than Type 1? Use  $\alpha = 0.01$ .

Confidence interval for  $\mu_1 - \mu_2$ :

Given: the two-sample  $Z$  test assumptions and the properties of  $\bar{X}_1 - \bar{X}_2$ .

Consider a symmetric interval about 0:

$$P\left(-z_{\alpha/2} < \underbrace{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}_Z < z_{\alpha/2}\right) = 1 - \alpha$$

Manipulate the inequality, sandwich  $\mu_1 - \mu_2$ .

$$P\left[(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] = 1 - \alpha$$

**How to Find a  $100(1 - \alpha)\%$  Confidence Interval for  $\mu_1 - \mu_2$  when Population Variances are Known**

Given the two-sample  $Z$  test assumptions, a  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  has as endpoints the values

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

**Example 10.1.4** Steam irons are lightweight, have large-capacity water tanks, and are available in cordless models. The power of these devices is often measured by the steam pressure. Independent random samples of Bosch and Rowenta steam irons were obtained, and the steam pressure was measured in each (in psi). The summary statistics and known variances are given in the following table.

Steam iron	Sample size	Sample mean	Population variance
Bosch	45	40.5	9.61
Rowenta	36	41.2	20.25

- Find a 95% confidence interval for the difference in mean steam pressures.
- Use the confidence interval in part (a) to determine whether there is any evidence to suggest that the mean steam pressure in Bosch irons is different from the mean steam pressure in Rowenta irons.

## 10.2 Comparing Two Population Means Using Independent Samples from Normal Populations

1. Unrealistic to assume the population variances are known.
2. Assume underlying populations are normal.  
Additional assumption regarding the population variances.

Two-sample  $t$  test assumptions:

1.  $\bar{X}_1$  is the mean of a random sample of size  $n_1$  from a normal population with mean  $\mu_1$ .
2.  $\bar{X}_2$  is the mean of a random sample of size  $n_2$  from a normal population with mean  $\mu_2$ .
3. The samples are independent.
4. The two population variances are *unknown* but *equal*.

The common variance is denoted  $\sigma^2$  ( $= \sigma_1^2 = \sigma_2^2$ ).

### Properties of $\bar{X}_1 - \bar{X}_2$

If the two-sample  $t$  test assumptions are true, then the estimator  $\bar{X}_1 - \bar{X}_2$  has the following properties.

1.  $E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$

$\bar{X}_1 - \bar{X}_2$  is still an unbiased estimator of the parameter  $\mu_1 - \mu_2$ .

2.  $\text{Var}(\bar{X}_1 - \bar{X}_2) = \sigma_{\bar{X}_1 - \bar{X}_2}^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)$

and the standard deviation is  $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$ .

3. Since both underlying populations are normal, the distribution of  $\bar{X}_1 - \bar{X}_2$  is also normal.

**Remarks**

1. An estimate of the common variance is necessary.
2. Appropriate standardization results in a  $t$  distribution.
3. An estimate of the common variance uses the information in both samples.

Give more weight to the larger sample.

**Definition**

The **pooled estimator** for the common variance  $\sigma^2$ , denoted  $S_p^2$ , is

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \\ &= \left( \frac{n_1 - 1}{n_1 + n_2 - 2} \right) S_1^2 + \left( \frac{n_2 - 1}{n_1 + n_2 - 2} \right) S_2^2. \end{aligned}$$

The pooled estimator for the common standard deviation  $\sigma$  is  $S_p = \sqrt{S_p^2}$ .

**Remarks**

1.  $S_p^2$  is indeed a weighted average:  $S_p^2 = \lambda S_1^2 + (1 - \lambda)S_2^2$  where  $0 \leq \lambda \leq 1$ .
2. The constants in the definition are related to degrees of freedom.

**Theorem**

If the two-sample  $t$  test assumptions are true, then the random variable

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

has a  $t$  distribution with  $n_1 + n_2 - 2$  degrees of freedom.

### Hypothesis Tests Concerning Two Population Means when Population Variances are Unknown but Equal

Given the two-sample  $t$  test assumptions, a hypothesis test concerning two population means in terms of the difference in means  $\mu_1 - \mu_2$ , with significance level  $\alpha$ , has the form:

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

$$H_a: \mu_1 - \mu_2 > \Delta_0, \quad \mu_1 - \mu_2 < \Delta_0, \quad \text{or} \quad \mu_1 - \mu_2 \neq \Delta_0$$

$$\text{TS: } T = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{S_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\text{RR: } T \geq t_{\alpha, n_1+n_2-2}, \quad T \leq -t_{\alpha, n_1+n_2-2}, \quad \text{or} \quad |T| \geq t_{\alpha/2, n_1+n_2-2}$$

**Example 10.2.1** The value of fine china is often measured by the dielectric strength of the porcelain enamel coating (measured in volts/mil). Independent random samples of Lenox and Noritake oval dinner plates were obtained, and the dielectric strength was measured in each. The resulting summary statistics are given in the following table.

China	Sample size	Sample mean	Sample standard deviation
Lenox	14	399.9	16.2
Noritake	18	380.1	24.6

Is there any evidence to suggest that Lenox oval dinner plates have a higher mean dielectric strength than Noritake oval dinner plates? Use  $\alpha = 0.01$  and assume that the underlying populations are normal, with equal variances.

**Example 10.2.2** High-heeled women's shoes have been linked to hip, back, and toe compression injuries. Nevertheless, manufacturers still market a wide variety of evening shoes with high heels. Independent random samples of women's evening shoes from two manufacturers were obtained, and the heel height was measured (in inches) on each. The resulting summary statistics are given in the following table.

Shoe manufacturer	Sample size	Sample mean	Sample standard deviation
DKNY	24	1.992	0.600
Van Eli	26	1.687	0.497

Is there any evidence to suggest that the true mean heel heights are different for DKNY and Van Eli evening shoes? Assume that the populations are normal, with equal variances, and use  $\alpha = 0.05$ . Find bounds on the  $p$  value associated with this test.

**Example 10.2.3** Sweet onions have low amounts of sulfur and can be eaten raw. Independent random samples of Vidalia and Texas Supersweet onions were obtained, and each was measured for sweetness using the sensory rating scale. The scores for each onion are given in the following table.

Vidalia									
2.16	3.40	4.53	2.90	2.10	4.67	3.57	4.07	4.79	3.82
Texas Supersweet									
2.16	3.66	1.95	1.99	2.31	2.31	1.82	2.68		

Assume the populations are normal, with equal variances.

- (a) Is there any evidence to suggest that the true mean sensory rating is different in Vidalia and Texas Supersweet onions? Use  $\alpha = 0.01$ .
- (b) Find bounds on the  $p$  value associated with this test.



Confidence interval for  $\mu_1 - \mu_2$ :

1. Symmetric interval about 0, probability T in this interval  $1 - \alpha$ .
2. Sandwich the parameter  $\mu_1 - \mu_2$ .

**How to Find a  $100(1 - \alpha)\%$  Confidence Interval for  $\mu_1 - \mu_2$  when Population Variances are Unknown but Equal**

Given the two-sample  $t$  test assumptions, a  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  has as endpoints the values

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} \sqrt{s_p^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}$$

**Example 10.2.4** Active telephone users can purchase wireless headsets that can be used to answer a call by using a voice command or pressing a single button. An experiment was conducted to compare the range of two types of headsets. Independent random samples of comparable Sony and Nokia headsets were obtained, and the range of each was tested and measured (in meters). The summary statistics are given in the following table.

Headset manufacturer	Sample size	Sample mean	Sample standard deviation
Sony	13	9.007	0.518
Nokia	12	9.153	0.630

Assume the populations are normal and the variances are equal. Find a 95% confidence interval for the difference in population mean ranges.

Note:

1. Two-sample  $t$  test and confidence interval: robust.
2. If population variances are unequal, no nice procedure.

### Hypothesis Tests and Confidence Interval Concerning Two Population Means when Population Variances are Unknown and Unequal

Given the *modified* two-sample  $t$  test assumptions (population variances unknown and assumed unequal), an *approximate* hypothesis test concerning two population means in terms of the difference,  $\mu_1 - \mu_2$ , with significance level  $\alpha$ , has the form:

$$H_0: \mu_1 - \mu_2 = \Delta_0$$

$$H_a: \mu_1 - \mu_2 > \Delta_0, \quad \mu_1 - \mu_2 < \Delta_0, \quad \text{or} \quad \mu_1 - \mu_2 \neq \Delta_0$$

$$\text{TS: } T' = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

$$\text{RR: } T' \geq t_{\alpha, \nu}, \quad T' \leq -t_{\alpha, \nu}, \quad \text{or} \quad |T'| \geq t_{\alpha/2, \nu}, \quad \text{where} \quad \nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$$

An approximate  $100(1 - \alpha)\%$  confidence interval for  $\mu_1 - \mu_2$  has as endpoints the values

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

### Remarks

1.  $T'$  has an approximate  $t$  distribution with  $\nu$  degrees of freedom.
2.  $\nu$  will probably *not* be an integer.

In order to be conservative, always round down (to the nearest integer).

3. Test for equality of population variances: presented in Section 10.5.

This hypothesis test is often used to determine whether equal population variances is a reasonable assumption.

**Example 10.2.5** Bonemeal is an excellent natural source of phosphorus for plants. Independent random samples of powdered bonemeal from two garden stores were obtained, and the amount of phosphorus per teaspoon was measured (in mg). The summary statistics are given in the following table.

Garden store	Sample size	Sample mean	Sample standard deviation
Extremely Green	15	503.9	21.5
Backyard Gardener	18	495.7	12.0

Is there any evidence to suggest that the population mean amount of phosphorus in bonemeal from Extremely Green is different from that in bonemeal from Backyard Gardener? Assume both populations are normal and use  $\alpha = 0.05$ .

---

## 10.3 Paired Data

1. Previous section: samples were obtained independently.
2. Many experiments involve only  $n$  individuals.

Two observations are made of each individual.

3. Paired observations are dependent.
4. Still want to consider the difference  $\mu_1 - \mu_2$ .

$\bar{X}_1$  and  $\bar{X}_2$  are not independent. Previous standardizations are not applicable.

Two-sample paired  $t$  test assumptions:

1. There are  $n$  individuals or objects, or  $n$  pairs of individuals or objects, that are related in an important way or share a common characteristic.
2. There are two observations of each *individual*. The population of first observations is normal, and the population of second observations is also normal.

Consequences:

1.  $X_1$ : a randomly selected first observation.  
 $X_2$ : second observation on the same individual.
2. Let  $D = X_1 - X_2$ , the difference in the observations.  
 $n$  observed differences:  $d_i = (x_1)_i - (x_2)_i$ ,  $i = 1, 2, \dots, n$ .
3. Since  $X_1$  and  $X_2$  are both normal,  $D$  is also normal.

The differences are independent!

4. Hypothesis test concerning  $\mu_1 - \mu_2$  is based on the sample mean of the differences  $\bar{D}$ .

**Properties of  $\bar{D}$** 

1.  $E(\bar{D}) = \mu_1 - \mu_2$

$\bar{D}$  is an unbiased estimator for the difference in means  $\mu_1 - \mu_2$ .

2. The variance of  $\bar{D}$  is unknown, but can be estimated using the sample variance of the differences.

3. Since both underlying populations are normal,  $D$  is normal, and hence,  $\bar{D}$  is also normal.

Interpretation:

1. To compare population means,  $\mu_1$  and  $\mu_2$ , when the data are paired, we focus on the difference  $\mu_1 - \mu_2$ .
2.  $H_0: \mu_1 = \mu_2$  is equivalent to  $H_0: \mu_1 - \mu_2 = 0$ .
3. A test to determine whether the underlying population means of two paired samples are equal is equivalent to a test to determine whether the population mean of the paired differences is zero.
4. Compute the differences,  $d_1, d_2, \dots, d_n$ , and conduct a one-sample  $t$  test (with  $n - 1$  degrees of freedom) using the differences.

**Hypothesis Tests Concerning Two Population Means when Data are Paired**

Given the two-sample paired  $t$  test assumptions, a hypothesis test concerning the two population means in terms of the difference  $\mu_D = \mu_1 - \mu_2$ , with significance level  $\alpha$ , has the form:

$$H_0: \mu_D = \mu_1 - \mu_2 = \Delta_0$$

$$H_a: \mu_D > \Delta_0, \quad \mu_D < \Delta_0, \quad \text{or} \quad \mu_D \neq \Delta_0$$

$$\text{TS: } T = \frac{\bar{D} - \Delta_0}{S_D / \sqrt{n}}$$

where  $S_D$  is the sample standard deviation of the differences.

$$\text{RR: } T \geq t_{\alpha, n-1}, \quad T \leq -t_{\alpha, n-1}, \quad \text{or} \quad |T| \geq t_{\alpha/2, n-1}$$

**Remarks**

1. Usually  $\Delta_0 = 0$ : the null hypothesis is that the two population means are equal.
2. A paired  $t$  test is valid even if the underlying population variances are unequal.
3. If a paired  $t$  test is appropriate, the test statistic is based on  $n - 1$  degrees of freedom.

A two-sample  $t$  test (incorrect here) would be based on a test statistic with  $n+n-2 = 2n-2$  degrees of freedom.

Correct analysis is based on a distribution with greater variability and is more conservative.

**Example 10.3.1** A new drug has been developed to ease the pain caused by rheumatoid arthritis. Eight patients with this disease were selected at random. An initial T-cell count (per cubic millimeter) was obtained, each person was given the new drug, and 48 hours later a second T-cell count was recorded. The data are given in the following table.

Subject	1	2	3	4	5	6	7	8
Initial count	1159	1058	1310	1265	1280	1308	1388	1266
Final count	1154	1065	1107	1085	1200	1375	1182	988
Difference								

Is there any evidence to suggest that the new drug reduced the mean T-cell count? Assume the underlying distributions of initial and final T-cell count are approximately normal, and use  $\alpha = 0.05$ .

**Example 10.3.2** A company has developed a new controller for vertical machine centers and claims that it improves accuracy and reduces the cycle time in engineering processes. A random sample of drawings of three-dimensional objects was obtained. Each drawing was subjected to a slight change and the cycle time (in minutes) using each controller was recorded. The data are given in the following table.

Drawing	1	2	3	4	5	6	7	8	9	10
Current controller	22.1	27.8	18.2	19.5	29.9	35.3	32.4	21.9	36.8	24.1
New controller	25.3	23.3	22.7	27.3	35.2	22.0	20.7	23.3	33.9	23.5

Is there any evidence to suggest the new controller reduces the mean cycle time? Assume the underlying distributions are normal, use  $\alpha = 0.01$ , and find bounds on the  $p$  value associated with this hypothesis test.

**How to Find a  $100(1 - \alpha)\%$  Confidence Interval for  $\mu_D$** 

Given the paired  $t$  test assumptions, a  $100(1 - \alpha)\%$  confidence interval for  $\mu_D$  has as endpoints the values

$$\bar{d} \pm t_{\alpha/2, n-1} \frac{s_D}{\sqrt{n}}.$$

**Example 10.3.3** High levels of homocysteine (an amino acid) have been linked to increased risk of coronary artery disease. A researcher believes that the combination of folic acid, vitamin B6, and vitamin B12 will lower homocysteine levels. Twenty adults were randomly selected, and the homocysteine level in each was measured (in  $\mu\text{mol/L}$ ). All adults were then given a supplement containing folic acid, vitamin B6, and vitamin B12 to take once a day, for one month. After a month, the homocysteine level in each adult was measured again. The summary statistics for the differences (before-supplement homocysteine level – after-supplement homocysteine level) were:  $\bar{d} = 0.435$ ,  $s_D = 3.466$ . Assuming normality, find a 99% confidence interval for the true difference in mean homocysteine level.



## 10.4 Comparing Two Population Proportions Using Large Samples

Notation:

Population proportion of successes:	$p_1, p_2$
Sample size:	$n_1, n_2$
Number of successes:	$x_1, x_2$
Corresponding random variables:	$X_1, X_2$
Sample proportion of successes:	$\hat{p}_1 = x_1/n_1, \hat{p}_2 = x_2/n_2$
Corresponding random variables:	$\hat{P}_1 = X_1/n_1, \hat{P}_2 = X_2/n_2$

Null hypothesis is  $H_0: p_1 - p_2 = \Delta_0$ .

Two cases to consider: (1)  $\Delta_0 = 0$  and (2)  $\Delta_0 \neq 0$ .

In both cases, a reasonable estimator for  $p_1 - p_2$  is the difference in sample proportions,  $\hat{P}_1 - \hat{P}_2$ .

### Properties of the Sampling Distribution of $\hat{P}_1 - \hat{P}_2$

1. The mean of  $\hat{P}_1 - \hat{P}_2$  is the true difference in population proportions  $p_1 - p_2$ .

In symbols:  $\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$ .

2. The variance of  $\hat{P}_1 - \hat{P}_2$  is  $\sigma_{\hat{P}_1 - \hat{P}_2}^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$ .

The standard deviation of  $\hat{P}_1 - \hat{P}_2$  is  $\sigma_{\hat{P}_1 - \hat{P}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$ .

3. If (a) both  $n_1$  and  $n_2$  are large, (b)  $n_1 p_1 \geq 5$  and  $n_1(1-p_1) \geq 5$ , and (c)  $n_2 p_2 \geq 5$  and  $n_2(1-p_2) \geq 5$ , then the distribution of  $\hat{P}_1 - \hat{P}_2$  is approximately normal.

In symbols:  $\hat{P}_1 - \hat{P}_2 \overset{\circ}{\sim} N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$ .

**Case 1:**  $H_0: p_1 - p_2 = 0$ , or  $p_1 = p_2$  ( $\Delta_0 = 0$ )

If  $H_0$  is true, one common value for the two population proportions, denoted  $p$  ( $= p_1 = p_2$ ).

The variance of  $\hat{P}_1 - \hat{P}_2$  is

$$\sigma_{\hat{P}_1 - \hat{P}_2}^2 = \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2} = p(1-p) \left( \frac{1}{n_1} + \frac{1}{n_2} \right).$$

Using the properties of  $\hat{P}_1 - \hat{P}_2$ , the random variable

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - 0}{\sqrt{p(1-p) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

is approximately standard normal.

### Definition

The pooled or **combined estimate of the common population proportion** is

$$\hat{P}_c = \frac{X_1 + X_2}{n_1 + n_2} = \left( \frac{n_1}{n_1 + n_2} \right) \hat{P}_1 + \left( \frac{n_2}{n_1 + n_2} \right) \hat{P}_2.$$

### Hypothesis Tests Concerning Two Population Proportions when $\Delta_0 = 0$

Given two random samples of sizes  $n_1$  and  $n_2$ , a large-sample hypothesis test concerning two population proportions in terms of the difference  $p_1 - p_2$  (with  $\Delta_0 = 0$ ) with significance level  $\alpha$ , has the form:

$$H_0: p_1 - p_2 = 0$$

$$H_a: p_1 - p_2 > 0, \quad p_1 - p_2 < 0, \quad \text{or} \quad p_1 - p_2 \neq 0$$

$$\text{TS: } Z = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\hat{P}_c(1 - \hat{P}_c) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

$$\text{RR: } Z \geq z_{\alpha}, \quad Z \leq -z_{\alpha}, \quad \text{or} \quad |Z| \geq z_{\alpha/2}$$

### Remarks

1. No confidence interval in this case. Assumed  $p_1 = p_2$ .
2. Non-skewness criteria must hold for both samples. Use  $\hat{p}_1$  and  $\hat{p}_2$  to check.

**Example 10.4.1** Football tends to be more popular in the United States, while soccer is more popular in Europe. In a random sample of 325 households in Sweden, 198 had a soccer ball, and in a random sample of 344 households in Germany, 202 had a soccer ball. Is there any evidence to suggest that the true proportion of households with a soccer ball is different in Sweden and in Germany? Use  $\alpha = 0.05$ .

**Example 10.4.2** There is a growing concern that CEOs of large corporations are paid too much. In a random sample of 275 female MBA students, 168 said CEOs are paid too much, and in a random sample of 308 male MBA students, 153 said CEOs are paid too much.

- (a) Is there any evidence to suggest that the true proportion of female MBA students who think CEOs are paid too much is greater than the true proportion of male MBA students who think CEOs are paid too much? Use  $\alpha = 0.05$ .
- (b) Find the  $p$  value associated with this hypothesis test.

**Case 2:**  $H_0: p_1 - p_2 = \Delta_0 \neq 0$

$\Delta_0 \neq 0$  is less common.

Since  $p_1$  and  $p_2$  are assumed unequal, there is no common population proportion.

Hypothesis test follows from the properties of  $\hat{P}_1 - \hat{P}_2$ .

### Hypothesis Tests Concerning Two Population Proportions when $\Delta_0 \neq 0$

Given two random samples of sizes  $n_1$  and  $n_2$ , a large-sample hypothesis test concerning two population proportions in terms of the difference  $p_1 - p_2$  (with  $\Delta_0 \neq 0$ ), with significance level  $\alpha$ , has the form:

$$H_0: p_1 - p_2 = \Delta_0$$

$$H_a: p_1 - p_2 > \Delta_0, \quad p_1 - p_2 < \Delta_0, \quad \text{or} \quad p_1 - p_2 \neq \Delta_0$$

$$\text{TS: } Z = \frac{(\hat{P}_1 - \hat{P}_2) - \Delta_0}{\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}}$$

$$\text{RR: } Z \geq z_\alpha, \quad Z \leq -z_\alpha, \quad \text{or} \quad |Z| \geq z_{\alpha/2}$$

Confidence interval derived in the usual way.

### How to Find a $100(1 - \alpha)\%$ Confidence Interval for $p_1 - p_2$

Given two (large) random samples of sizes  $n_1$  and  $n_2$ , a  $100(1 - \alpha)\%$  confidence interval for  $p_1 - p_2$  has as endpoints the values

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2(1 - \hat{p}_2)}{n_2}}.$$

**Example 10.4.3** Government officials near a nuclear power plant have been working on an emergency preparedness program. In 2004, a random sample of 405 community residents was obtained, and 275 said they were confident they would be notified quickly of a radioactive incident. In 2005, after the program started, a random sample of 388 community residents was obtained, and 310 said they were confident they would be notified quickly of a radioactive incident. Conduct the appropriate hypothesis test to determine whether there is evidence that the true proportion of residents who were confident they would be notified quickly of a radioactive incident in 2005 is more than 0.10 greater than in 2004.

**Example 10.4.4** Independent random samples of adults living in Boston and New York City were obtained, and each person was asked whether they regularly listen to classical music on NPR. The data are given in the following table.

	Boston	New York City
Sample size	$n_1 = 256$	$n_2 = 284$
Number who listen to classical music on NPR	$x_1 = 136$	$x_2 = 117$

Construct a 99% confidence interval for the true difference in proportions of adults who regularly listen to classical music on NPR.

## 10.5 Comparing Two Population Variances or Standard Deviations

1.  $S_1^2$  and  $S_2^2$ : good (unbiased) estimators for the population variances  $\sigma_1^2$  and  $\sigma_2^2$ .
2. Hypothesis test for comparing  $\sigma_1^2$  and  $\sigma_2^2$ : based on a new standardization.
3.  $F$  distribution: positive probability only for non-negative values.

Probability density function for an  $F$  random variable is 0 for  $x < 0$ .

Focus on the properties of an  $F$  distribution and the method for finding critical values.

### Properties of an $F$ Distribution

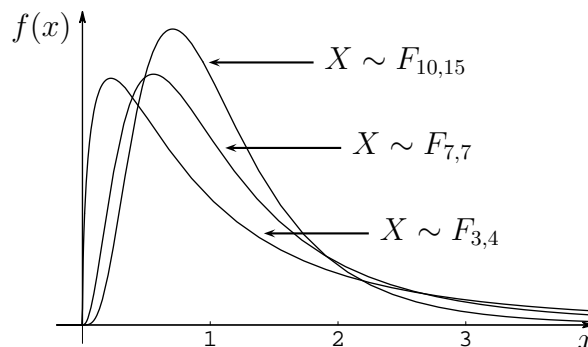
1. An  $F$  distribution is completely determined by two parameters,  $\nu_1$  and  $\nu_2$ , both positive integers (1, 2, 3, ...).  $\nu_1$  is the number of degrees of freedom in the numerator, and  $\nu_2$  is the number of degrees of freedom in the denominator. There is, of course, a different  $F$  distribution for every combination of  $\nu_1$  and  $\nu_2$ .

2. If  $X$  has an  $F$  distribution with  $\nu_1$  and  $\nu_2$  degrees of freedom, denoted  $X \sim F_{\nu_1, \nu_2}$ , then

$$\mu_X = \frac{\nu_2}{\nu_2 - 2}, \quad \nu_2 \geq 3; \quad \sigma_X^2 = \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, \quad \nu_2 \geq 5.$$

3. Suppose  $X \sim F_{\nu_1, \nu_2}$ . The density curve for  $X$  is positively skewed (*not* symmetric), and gets closer and closer to the  $x$ -axis but never touches it. As both degrees of freedom increase, the density curve becomes taller and more compact.

Density curves for several  $F$  distributions:





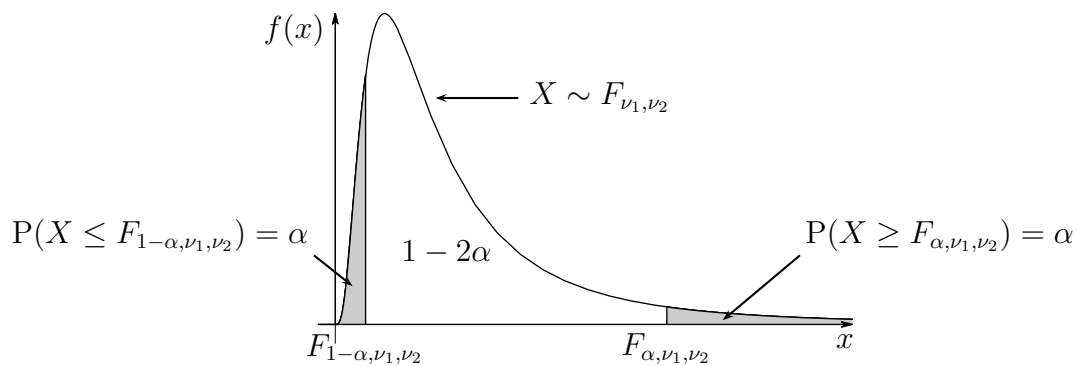
**Definition**

$F_{\alpha, \nu_1, \nu_2}$  is a critical value related to an  $F$  distribution with  $\nu_1$  and  $\nu_2$  degrees of freedom. If  $X \sim F_{\nu_1, \nu_2}$ , then  $P(X \geq F_{\alpha, \nu_1, \nu_2}) = \alpha$ .

**Remarks**

1.  $F_{\alpha, \nu_1, \nu_2}$  is a value on the measurement axis such that there is  $\alpha$  of the area (probability) to the right of  $F_{\alpha, \nu_1, \nu_2}$ . No symmetry.
2. Need to find critical values denoted  $F_{1-\alpha, \nu_1, \nu_2}$ , where  $1 - \alpha$  is large.

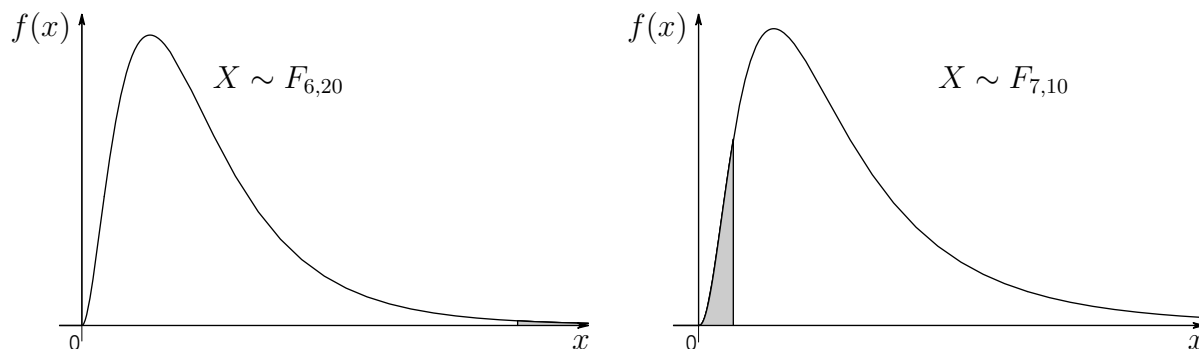
$P(X \geq F_{1-\alpha, \nu_1, \nu_2}) = 1 - \alpha$ , and by the Complement Rule,  $P(X \leq F_{1-\alpha, \nu_1, \nu_2}) = \alpha$ .



3.  $F$  critical values are related according to the following equation:  $F_{1-\alpha, \nu_1, \nu_2} = \frac{1}{F_{\alpha, \nu_2, \nu_1}}$ .

4. Table 7: selected critical values associated with various  $F$  distributions.

**Example 10.5.1** Find each critical value: (a)  $F_{0.01,6,20}$ , (b)  $F_{0.95,7,10}$ .



### Remarks

1. Table 7: very limited. Three values of  $\alpha$ , limited values of  $\nu_1$  and  $\nu_2$ .
2. Use technology to find critical values if necessary.

Two-sample  $F$  test assumptions:

1.  $S_1^2$  is the sample variance of a random sample of size  $n_1$  from a normal distribution with variance  $\sigma_1^2$ .
2.  $S_2^2$  is the sample variance of a random sample of size  $n_2$  from a normal distribution with variance  $\sigma_2^2$ .
3. The samples are independent.

Consequences:

1. The random variable  $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$  has an  $F$  distribution

with  $n_1 - 1$  (from the numerator) and  $n_2 - 1$  (from the denominator) degrees of freedom.

2. If  $H_0: \sigma_1^2 = \sigma_2^2$ , then the random variable simplifies to  $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2}$ .

**Hypothesis Tests Concerning Two Population Variances**

Given the two-sample  $F$  test assumptions, a hypothesis test concerning two population variances with significance level  $\alpha$  has the form:

$H_0: \sigma_1^2 = \sigma_2^2$

$H_a: \sigma_1^2 > \sigma_2^2, \quad \sigma_1^2 < \sigma_2^2, \quad \text{or} \quad \sigma_1^2 \neq \sigma_2^2$

TS:  $F = \frac{S_1^2}{S_2^2}$

RR:  $F \geq F_{\alpha, n_1-1, n_2-1}, \quad F \leq F_{1-\alpha, n_1-1, n_2-1}, \quad \text{or}$

$F \leq F_{1-\frac{\alpha}{2}, n_1-1, n_2-1} \quad \text{or} \quad F \geq F_{\frac{\alpha}{2}, n_1-1, n_2-1}$

Usual technique to find a confidence interval.

**How to Find a 100(1 -  $\alpha$ )% Confidence Interval for the Ratio of Two Population Variances**

Given the two-sample  $F$  test assumptions, a 100(1 -  $\alpha$ )% confidence interval for  $\sigma_1^2/\sigma_2^2$  is given by

$$\left( \frac{s_1^2}{s_2^2} \frac{1}{F_{\frac{\alpha}{2}, n_1-1, n_2-1}}, \frac{s_1^2}{s_2^2} \frac{1}{F_{1-\frac{\alpha}{2}, n_1-1, n_2-1}} \right).$$

We can simplify the quotient in the upperbound, and the confidence interval can be written as

$$\left( \frac{s_1^2}{s_2^2} \frac{1}{F_{\frac{\alpha}{2}, n_1-1, n_2-1}}, \frac{s_1^2}{s_2^2} F_{\frac{\alpha}{2}, n_2-1, n_1-1} \right).$$

**Example 10.5.2** A jeweler is interested in comparing the variability in ring-finger sizes between men and women. Independent random samples of men and women were obtained, and the circumference of each person's left hand ring-finger was measured (in cm). The data are given in the following table.

Women							
3.43	4.74	7.30	6.54	6.49	5.01	7.25	
Men							
5.26	4.60	5.88	5.93	6.31	4.36	5.87	5.49

- (a) Conduct the appropriate hypothesis test to determine whether there is any evidence that the population variance in circumference of the ring finger is different in women and men. Use  $\alpha = 0.10$ .
- (b) Find bounds on the  $p$  value associated with this hypothesis test.

**Example 10.5.3** Researchers have concluded that genetically modified (GM) corn, resistant to certain insect pests and drought, is safe to eat. Independent random samples of GM and non-GM modified mature yellow corn plants were obtained. The height of each stalk was measured (in inches) and the data are summarized in the following table.

Genetically modified corn:	$n_1 = 21$	$s_1^2 = 24.6$
Non-genetically modified corn:	$n_2 = 31$	$s_2^2 = 32.8$

Construct a 98% confidence interval for the ratio of population variances in the height of GM and non-GM corn stalks.



## CHAPTER 11

# The Analysis of Variance

---

## 11.0 Introduction

1. Chapter 10: Compared *two* population parameters.
2. Here: compare  $k$  ( $> 2$ ) population means—analysis of variance, ANOVA.
3. ANOVA: compare means by dividing total variation into appropriate pieces.
4. Two kinds of ANOVA tests.
  - (a) One-way ANOVA: assess the effect of a single factor on  $k$  population means.

Example: Compare the mean weight of  $k = 5$  different kinds of gourmet chocolate chip cookies.

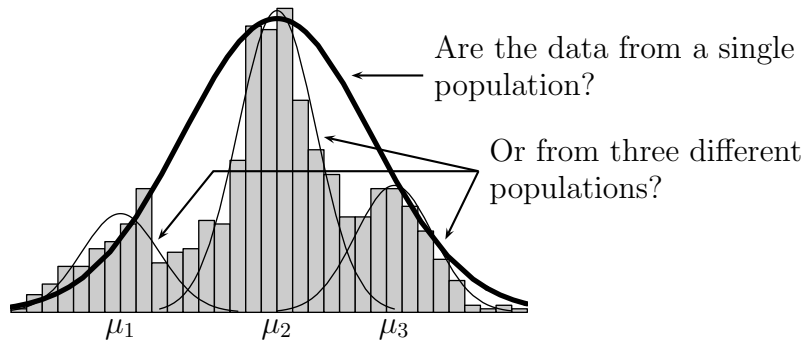
- (b) Two-way ANOVA: Assess the effect of two factors on a response variable.

Example: Compare the mean weight of  $5 \times 7$  handmade carpets with respect to material and maker.

---

## 11.1 One-Way ANOVA

1. One-way, or single-factor, ANOVA: the analysis of data sampled from more than two populations.
2. The only difference among the populations is a single factor.
3. Example: Compare the mean signal strength of four different makes of TV remote controls.
4. Suppose three random samples are obtained. Consider the following histogram.



ANOVA is used to determine whether the data came from a single population or whether at least two samples came from populations with different means.

**ANOVA Notation**

$k$  = the number of populations under investigation.

Population	1	2	...	$i$	...	$k$
Population mean	$\mu_1$	$\mu_2$	...	$\mu_i$	...	$\mu_k$
Population variance	$\sigma_1^2$	$\sigma_2^2$	...	$\sigma_i^2$	...	$\sigma_k^2$
Sample size	$n_1$	$n_2$	...	$n_i$	...	$n_k$
Sample mean	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_i$	...	$\bar{x}_k$
Sample variance	$s_1^2$	$s_2^2$	...	$s_i^2$	...	$s_k^2$

$$n = n_1 + n_2 + \dots + n_k$$

= the total number of observations in the *entire* data set.

Null and alternative hypotheses stated in terms of the population means.

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k \quad (\text{all } k \text{ population means are equal})$$

$$H_a: \mu_i \neq \mu_j \text{ for some } i \neq j \quad (\text{at least two of the } k \text{ population means differ})$$

One-way ANOVA assumptions:

1. The  $k$  population distributions are normal.
2. The  $k$  population variances are equal. That is,  $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$ .
3. The samples are selected randomly and independently from the respective populations.



Subscript notation:

To denote observations: single letter with two subscripts.

First subscript: sample number.

Second subscript: observation number within the sample.

$x_{ij}$  = the  $j$ th measurement taken from the  $i$ th population.

$X_{ij}$  = the corresponding random variable.

Use a comma if there is ambiguity:

$x_{2,34}$ : the 34th observation in the 2nd sample.

$x_{23,4}$ : the 4th observation in the 23rd sample.

Dot notation:

1. A dot in the second subscript indicates a sum over that subscript, while the other is held fixed.

Mean of the observations in the  $i$ th sample:

$$\bar{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n_i} (x_{i1} + x_{i2} + \cdots + x_{in_i})$$

2. **Grand mean:** sum of all the observations divided by  $n$ :

$$\bar{x}_{..} = \frac{1}{n} \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}$$

3. Other (dot) notation to make some calculations easier.

$$t_{i.} = \sum_{j=1}^{n_i} x_{ij} \quad = \text{sum of the observations in the } i\text{th sample.}$$

$$t_{..} = \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij} = \text{sum of all the observations.}$$

Analysis of *variance*:

Total variation in the data: total sum of squares.

Variability of individual observations from the grand mean.

Total sum of squares is decomposed into a sum of:

1. Between-samples variation: sum of squares due to the factor.

Variability in the sample means; how different the sample means are from each other.

2. Within-samples variation: sum of squares due to error.

Variability of the observations from their sample mean.

Related to the sample variance.

### One-Way ANOVA Identity

Let SST = total sum of squares, SSA = sum of squares due to factor, and SSE = sum of squares due to error.

$$\underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^k n_i (\bar{x}_{i.} - \bar{x}_{..})^2}_{\text{SSA}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}_{\text{SSE}}$$

### Computational Formulas

$$\text{SST} = \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{..})^2}_{\text{definition}} = \underbrace{\left( \sum_{i=1}^k \sum_{j=1}^{n_i} x_{ij}^2 \right) - \frac{t_{..}^2}{n}}_{\text{computational formula}}$$

$$\text{SSA} = \underbrace{\sum_{i=1}^k n_i (\bar{x}_{i.} - \bar{x}_{..})^2}_{\text{definition}} = \underbrace{\left( \sum_{i=1}^k \frac{t_{i.}^2}{n_i} \right) - \frac{t_{..}^2}{n}}_{\text{computational formula}}$$

$$\text{SSE} = \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i.})^2}_{\text{definition}} = \underbrace{\text{SST} - \text{SSA}}_{\text{computational formula}}$$

**Remarks**

1. SSA is used instead of SSF: in a two-way ANOVA there is a factor A and a factor B.
2. Sample size is used as a *weight* in the expression for SSA.

The test statistic:

1. If  $H_0$  is true, each observation comes from the same population, mean  $\mu$ , variance  $\sigma^2$ .
2. Sample means,  $\bar{x}_i$ 's, should all be about the same, and close to the grand mean,  $\bar{x}_{..}$ .
3. If at least two population means differ:  
at least two  $\bar{x}_i$ 's should be different, and these values will be far from  $\bar{x}_{..}$ .
4. Test statistic based on two separate estimates for  $\sigma^2$ .

**Definition**

The **mean square due to factor**, **MSA**, is SSA divided by  $k - 1$ :  $MSA = \frac{SSA}{k - 1}$ .

The **mean square due to error**, **MSE**, is SSE divided by  $n - k$ :  $MSE = \frac{SSE}{n - k}$ .

5. If  $H_0$  is true: MSA is an unbiased estimator of  $\sigma^2$ .  
If  $H_a$  is true: MSA tends to overestimate  $\sigma^2$ .
6. MSE is an unbiased estimator of  $\sigma^2$  whether  $H_0$  or  $H_a$  is true.
7. Consider  $F = \frac{MSA}{MSE}$ .
  - (a) If  $F$  is close to 1: two estimates of  $\sigma^2$  approximately the same.  
No evidence to suggest the population means are different.
  - (b) If  $F$  is much greater than 1: variation *between* samples is greater than variation *within* samples.  
Suggests the alternative hypothesis is true.

8. If the one-way ANOVA assumptions are satisfied and  $H_0$  is true:

$F = \frac{MSA}{MSE}$  has an  $F$  distribution with  $\nu_1 = k - 1$  and  $\nu_2 = n - k$  degrees of freedom.

Large values of  $F$  suggest  $H_a$  is true: rejection region is only in the right tail.

### One-way ANOVA Test Procedure

Given the one-way ANOVA assumptions, the test procedure with significance level  $\alpha$  is:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

$$H_a: \mu_i \neq \mu_j \text{ for some } i \neq j$$

$$\text{TS: } F = \frac{MSA}{MSE}$$

$$\text{RR: } F \geq F_{\alpha, k-1, n-k}$$

### Remarks

1. If  $F$  is less than the critical value: no evidence to reject  $H_0$ .
2. If  $F$  is in the rejection region: there is a statistically significant difference among the population means.
3. Recall: can also use the  $p$  value.

If  $p \leq \alpha$  then we reject  $H_0$ .

One-way ANOVA calculations are often presented in an **analysis of variance table**, or ANOVA table.

One-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor	SSA	$k - 1$	$MSA = \frac{SSA}{k - 1}$	$\frac{MSA}{MSE}$	$p$
Error	SSE	$n - k$	$MSE = \frac{SSE}{n - k}$		
Total	SST	$n - 1$			

**Example 11.1.1** Down comforters are often rated by fill power, the amount of space a single ounce of goose down occupies. Higher fill-power ratings indicate a fluffier comforter offering more warmth and less weight. Independent samples of four different brands of luxury down comforters were obtained, and the fill power of each was measured (in cubic inches per ounce). The data are given in the following table.

Sample	Observations				
Brand 1	518	562	519	564	561
Brand 2	560	536	594	569	577
Brand 3	500	529	538	493	547
Brand 4	550	573	619	599	612

Is there any evidence to suggest that at least two of the population mean fill-power ratings are different? Use  $\alpha = 0.05$ .

Sample	Sample size	Sample total	Sample mean	Sample variance
Brand 1				
Brand 2				
Brand 3				
Brand 4				

Example (continued)

One-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor					
Error					
Total					

**Example 11.1.2** Olives were first brought to California by Franciscan missionaries from Mexico and are now grown on over 33 thousand acres in the San Joaquin Valley and Northern Sacramento Valley. California olives contain potassium and calcium, and even small amounts of boron (which can be toxic in concentrations above 185 ppm). Independent random samples of olives from five California cultivators were obtained, and the concentration of boron in each olive (in ppm) was measured. The data are given in the following table.

Brand (factor)		Observations						
Ascolano	(1)	100	112	116	82	94	104	127
Barouni	(2)	80	84	127	90	52	89	
Manzanillo	(3)	95	121	107	114	132	81	
Mission	(4)	99	96	84	89	100	75	95
Sevillano	(5)	102	103	99	106	89	122	136

Is there any evidence to suggest a difference in the mean amount of boron in olives from the five cultivators? Use  $\alpha = 0.01$ .

Example (continued)

One-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor					
Error					
Total					



**Example 11.1.3** For a discerning chess player, the weight of each piece is an important characteristic. Independent random samples of wooden chess sets were obtained, and the king in each set was carefully weighed (in ounces). The data are given in the following table.

Drueke		Kolobob		Staunton		Westminster	
4.3	4.0	3.9	4.2	4.1	3.9	4.5	4.6
4.4	3.7	4.3	4.4	3.6	3.9	4.2	4.5
4.3	4.2	4.3	4.3	3.8	4.1	3.8	4.3
3.6	4.3	4.3	4.5	3.9	3.8	4.5	4.3
3.7	4.2	4.3	4.4	4.0	3.8	3.9	4.4

Is there any evidence to suggest a difference in the mean weight of kings from the four chess set manufacturers? Use  $\alpha = 0.05$ .

Example (continued)

One-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor					
Error					
Total					

---

## 11.2 Isolating Differences

1. Suppose we fail to reject  $H_0$  in a one-way ANOVA.

There is no evidence to suggest any difference among population means.

The statistical analysis stops.

2. Suppose we reject  $H_0$  in a one-way ANOVA.

There is evidence to suggest an overall difference among population means.

Next step: try and isolate the difference(s).

Find the pair(s) of means contributing to the overall significant difference.

Use a multiple comparison procedure.

Multiple comparison procedures:

1. Comparing two means:  $t$  test (or  $Z$  test) is appropriate.
2. Comparing three or more means: analysis is a little trickier.

Conduct a test on every possible pair of means?

However, cannot set the significance level in each *individual* test.

The probability of a Type I error is set under the assumption that *only one* test is conducted per experiment.

The more tests, the greater the chance of making an error.

3. We want to control the overall probability of making at least one mistake.

Typically, set this overall error probability and work backward to compute individual error probabilities.

We cap the probability of making at least one mistake in all of the comparisons.

4. Instead of hypothesis tests, we usually construct multiple confidence intervals for the difference between population means.

Recall: if a CI for  $\mu_1 - \mu_2$  contains 0, there is no evidence to suggest that the population means are different.

If 0 is not included in the CI, there is evidence to suggest that the two population means are different.

We want a  $100(1 - \alpha)\%$  CI for *all* possible paired comparisons.

**Example 11.2.1** Consider a one-way ANOVA with samples from three populations, in which we reject  $H_0$ . Suppose a multiple comparison procedure produces the following confidence intervals.

Difference	Confidence interval
$\mu_1 - \mu_2$	( 3.62, 7.11 )
$\mu_1 - \mu_3$	(-5.62, -1.25 )
$\mu_2 - \mu_3$	(-2.37, 3.22 )

Identify the pair(s) of means contributing to the overall difference.

Bonferroni confidence intervals:

1. Each is similar to a CI for the difference between two means based on a  $t$  distribution.
2. Use a pooled estimate of the common variance: MSE.
3.  $t$  critical value results in a simultaneous, or familywise, confidence level of  $100(1 - \alpha)\%$ .

### The Bonferroni Multiple Comparison Procedure

In a one-way analysis of variance, suppose there are  $k$  populations,  $n = n_1 + n_2 + \cdots + n_k$  total observations, and  $H_0$  is rejected.

1. There are  $c = \binom{k}{2} = \frac{k(k-1)}{2}$  pairs of population means to compare.
2. The  $c$  simultaneous  $100(1 - \alpha)\%$  **Bonferroni confidence intervals** have as endpoints the values

$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/(2c), n-k} \sqrt{\text{MSE}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad \text{for all } i \neq j.$$

**Example 11.2.2** A study was conducted to compare the amount of camphor (in grams) in 4-gram sticks of medicated lip balm. Independent random samples of 3 brands of lip balm were obtained, with 8 observations for each brand. The resulting sample means and the ANOVA table are shown below.

Group number	1	2	3
Factor	Blistex	Neutrogena	ChapStick
Sample mean	0.125	0.051	0.074

One-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor	0.0233	2	0.0117	4.33	0.0266
Error	0.0574	21	0.0027		
Total	0.0807	23			

The ANOVA test is significant at the  $p = 0.0266$  level. There is evidence to suggest that at least one pair of population means is different (an overall difference). Construct the Bonferroni 95% confidence intervals and use them to isolate the pair(s) of means contributing to this overall experiment difference.

Example (continued)

**Example 11.2.3** The weight of a canoe is an important consideration in the event the canoe must be carried around falls or in shallow water. Independent random samples of 4 brands of 17-foot canoes were obtained, and the weight of each canoe was recorded (in pounds). The summary statistics and the ANOVA table are shown below.

Group number	1	2	3	4
Factor	Clearwater	Dagger	Old Town	Wenonah
Sample size	10	11	12	11
Sample mean	61.74	60.79	69.03	65.86

One-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor	493.23	3	164.41	8.07	0.0002
Error	815.35	40	20.38		
Total	1308.58	43			

The ANOVA test is significant at the  $p = 0.0002$  level. There is evidence to suggest that at least one pair of population means is different (an overall difference). Construct the Bonferroni 99% confidence intervals and use them to isolate the pair(s) of means contributing to this overall experiment difference.



Example (continued)

Graphical method to summarize the results of a multiple comparison procedure:

1. Write the sample means in order from smallest to largest.
2. Use the results from a multiple comparison procedure to draw a horizontal line under the groups of means that are *not* significantly different.

**Example 11.2.4** Present a graphical summary of the multiple comparison procedure in Example 11.2.2.

**Example 11.2.5** Present a graphical summary of the multiple comparison procedure in Example 11.2.3.

The Tukey multiple comparison procedure.

1. Form of the CIs is similar to Bonferroni CIs.
2. Based on a  $Q$  critical value from the **Studentized range distribution**.

This distribution is characterized by two parameters.

Degrees of freedom in the numerator and denominator,  $m$  and  $\nu$ .

$Q_{\alpha,m,\nu}$ : right-tail critical value.

3. Table 8: selected critical values associated with various Studentized range distributions.

Top row: degrees of freedom in the numerator.

Left column: degrees of freedom in the denominator.

Body of the table:  $Q_{\alpha,m,\nu}$  at the intersection of column  $m$  and row  $\nu$ .

### The Tukey Multiple Comparison Procedure

In a one-way analysis of variance, suppose there are  $k$  populations,  $n = n_1 + n_2 + \cdots + n_k$  total observations, and  $H_0$  is rejected. The set of  $c = \binom{k}{2}$  simultaneous  $100(1 - \alpha)\%$  **Tukey confidence intervals** have as endpoints the values

$$(\bar{x}_{i.} - \bar{x}_{j.}) \pm \frac{1}{\sqrt{2}} Q_{\alpha,k,n-k} \sqrt{\text{MSE}} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \quad \text{for all } i \neq j.$$

### Remarks

1. If all pairwise comparisons are considered, the Bonferroni procedure produces wider confidence intervals than the Tukey procedure.
2. If only a subset of all pairwise comparisons is needed, then the Bonferroni method may be better.
3. There are also other methods for comparing population means following an ANOVA test. No single comparison method is uniformly best.

**Example 11.2.6** A pocket bike is a sort of mini-motorcycle with an aluminum frame and a style similar to a racing bike. The performance and speed of a pocket bike depend on the model and the engine power output. Independent random samples of three pocket bike models were selected, and the engine power output was measured (in hp) for each. The summary statistics and the ANOVA table are shown below.

Group number	1	2	3
Factor	Blade	Raptor	Tornado
Sample size	7	8	9
Sample mean	4.37	4.57	4.43

One-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor	0.1668	2	0.0834	6.67	0.0057
Error	0.2615	21	0.0125		
Total	0.4283	23			

The ANOVA test is significant at the  $p = 0.0057$  level. There is evidence to suggest that at least one pair of population means is different (an overall difference). Construct the Tukey 95% confidence intervals and use them to isolate the pair(s) of means contributing to this overall experiment difference.

Example (continued)

**Example 11.2.7** The U.S. Army is investigating the performance of several parachutes for troops. Independent random samples of 4 types of parachutes were obtained, and the rate of descent at 90 kg all-up suspended weight (AUW) was measured (in m/s) for each. The summary statistics and the ANOVA table are shown below.

Group number	1	2	3	4
Factor	CT-1	T-108	T-10C	T-10D
Sample size	10	10	12	12
Sample mean	4.91	5.10	5.32	4.95

One-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor	1.1596	3	0.3865	6.13	0.0014
Error	2.5185	40	0.0630		
Total	3.6781	43			

The ANOVA test is significant at the  $p = 0.0014$  level. There is evidence to suggest that at least one pair of population means is different (an overall difference). Construct the Tukey 99% confidence intervals and use them to isolate the pair(s) of means contributing to this overall experiment difference.

Example (continued)

## 11.3 Two-Way ANOVA

1. Two-way ANOVA: compare the means of populations that can be classified in two different ways.

2. Consider the breaking strength of the cords used in bungee jumping.

Breaking strength (in pounds) may vary by: type of cord and/or the diameter of the cord.

3. In general:  $a$  levels of factor A,  $b$  levels of factor B,

$n$  observations for each combination of levels, total of  $abn$  observations.

4. Bungee-cord example: there could be

$a = 4$  levels of factor A: 4 different types of cords.

$b = 3$  levels of factor B: 3 different diameters.

$n = 6$  observations for each combination of type and diameter.

$abn = (4)(3)(6) = 72$  total observations.

5.  $x_{ijk}$ : the  $k$ th observation for the  $i$ th level of factor A and the  $j$ th level of factor B.

6. Presentation of data in a two-way ANOVA: the bungee-cord example.

		Factor B								
		1			2			3		
Factor A	1	1471	1576	1697	1344	1506	1563	1499	1591	1535
		1406	1470	1314	1497	1464	1336	1470	1540	1510
	2	1708	1570	1426	1656	1458	1471	1617	1640	1560
		1492	1515	1462	1484	1666	1514	1351	1428	1490
	3	1426	1719	1387	1412	1463	1672	1537	1379	1398
		1477	1533	1453	1474	1507	1505	1637	1364	1403
	4	1523	1453	1532	1730	1339	1599	1563	1531	1502
		1453	1718	1348	1418	1465	1510	1398	1380	1538



Two-way ANOVA assumptions:

1. The  $ab$  population distributions are normal.
2. The  $ab$  population variances are equal.
3. The samples are selected randomly and independently from the respective populations.

Similar dot notation:

Dots in the subscript of  $\bar{x}$  and  $t$  indicate the mean and the sum of  $x_{ijk}$ , respectively, over the appropriate subscripts.

For example:

$$\bar{x}_{ij.} =$$

$$\bar{x}_{i..} =$$

$$t_{.j.} =$$

$$\bar{x}_{...} =$$

$$t_{...} =$$

Total variation in the data: total sum of squares decomposed into:

1. The sum of squares due to factor A.
2. The sum of squares due to factor B.
3. The sum of squares due to interaction between the factors.
4. The sum of squares due to error.

**Two-Way ANOVA Identity**

Let SST = total sum of squares,  
 SSA = sum of squares due to factor A,  
 SSB = sum of squares due to factor B,  
 SS(AB) = sum of squares due to interaction, and  
 SSE = sum of squares due to error.

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SS(AB)} + \text{SSE}$$

$$\text{SST} = \underbrace{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x}_{...})^2}_{\text{definition}} = \underbrace{\left( \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n x_{ijk}^2 \right) - \frac{t_{...}^2}{abn}}_{\text{computational formula}}$$

$$\text{SSA} = \underbrace{bn \sum_{i=1}^a (\bar{x}_{i..} - \bar{x}_{...})^2}_{\text{definition}} = \underbrace{\frac{\sum_{i=1}^a t_{i..}^2}{bn} - \frac{t_{...}^2}{abn}}_{\text{computational formula}}$$

$$\text{SSB} = \underbrace{an \sum_{j=1}^b (\bar{x}_{.j.} - \bar{x}_{...})^2}_{\text{definition}} = \underbrace{\frac{\sum_{j=1}^b t_{.j.}^2}{an} - \frac{t_{...}^2}{abn}}_{\text{computational formula}}$$

$$\text{SS(AB)} = \underbrace{n \sum_{i=1}^a \sum_{j=1}^b (\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{x}_{...})^2}_{\text{definition}} = \underbrace{\frac{\sum_{i=1}^a \sum_{j=1}^b t_{ij.}^2}{n} - \frac{\sum_{i=1}^a t_{i..}^2}{bn} - \frac{\sum_{j=1}^b t_{.j.}^2}{an} + \frac{t_{...}^2}{abn}}_{\text{computational formula}}$$

$$\text{SSE} = \underbrace{\sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (x_{ijk} - \bar{x}_{ij.})^2}_{\text{definition}} = \underbrace{\text{SST} - \text{SSA} - \text{SSB} - \text{SS(AB)}}_{\text{computational formula}}$$

Two-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor A	SSA	$a - 1$	$MSA = \frac{SSA}{a - 1}$	$F_A = \frac{MSA}{MSE}$	$p_A$
Factor B	SSB	$b - 1$	$MSB = \frac{SSB}{b - 1}$	$F_B = \frac{MSB}{MSE}$	$p_B$
Interaction	SS(AB)	$(a - 1)(b - 1)$	$MS(AB) = \frac{SS(AB)}{(a - 1)(b - 1)}$	$F_{AB} = \frac{MS(AB)}{MSE}$	$p_{AB}$
Error	SSE	$ab(n - 1)$	$MSE = \frac{SSE}{ab(n - 1)}$		
Total	SST	$abn - 1$			

### Two-Way ANOVA Tests

Test 1. Test for an interaction effect.

$H_0$ : There is no interaction effect.

$H_a$ : There is an effect due to interaction.

$$\text{TS: } F_{AB} = \frac{MS(AB)}{MSE}$$

$$\text{RR: } F_{AB} \geq F_{\alpha, (a-1)(b-1), ab(n-1)}$$

Test 2. Test for an effect due to factor A.

$H_0$ : There is no effect due to factor A.

$H_a$ : There is an effect due to factor A.

$$\text{TS: } F_A = \frac{MSA}{MSE}$$

$$\text{RR: } F_A \geq F_{\alpha, a-1, ab(n-1)}$$

Test 3. Test for an effect due to factor B.

$H_0$ : There is no effect due to factor B.

$H_a$ : There is an effect due to factor B.

$$\text{TS: } F_B = \frac{MSB}{MSE}$$

$$\text{RR: } F_B \geq F_{\alpha, b-1, ab(n-1)}$$

Two-way ANOVA procedure:

The hypothesis test for an interaction effect is usually considered first.

Case 1.

If the null hypothesis is *not* rejected, then the other two hypothesis tests can be conducted as usual, to see whether there are effects due to either (or both) factors.

Case 2.

If the null hypothesis is rejected, then there is evidence of a significant interaction.

1. If we reject a null hypothesis of no effect due to factor A (and/or factor B), then the effect due to factor A (and/or factor B) is probably significant.
2. If we do not reject a null hypothesis of no effect due to factor A (and/or factor B), then the test for an effect due to factor A (and/or factor B) is inconclusive.

**Example 11.3.1** The weight of a snowshoe is an important consideration for hunters, trackers, and hikers. A study was conducted to determine whether the weight of a snowshoe is related to the type of frame and/or the type of deck. Twenty-two-inch snowshoes were considered, and for each combination of frame (factor A) and deck (factor B), a random sample of snowshoes was obtained. The weight of each was measured (in pounds) and the data are given in the following table.

		Deck type											
		1				2				3			
Frame type	1	2.48	2.40	2.48	2.52	2.49	2.69	2.81	2.45	2.21	2.49	1.92	2.22
	2	2.14	2.21	2.39	1.54	3.18	2.67	2.93	2.44	2.38	2.63	2.72	2.29
	3	1.70	2.94	2.81	1.95	2.62	2.42	2.89	2.57	2.04	3.06	2.22	1.88
	4	2.50	1.72	2.45	2.31	2.63	2.10	3.00	2.04	2.05	2.28	2.51	2.26

Conduct a two-way analysis of variance to determine whether snowshoe weight is affected by frame type and/or deck type. Use  $\alpha = 0.05$ .

Sample totals:

		Deck type		
		1	2	3
Frame type	1			
	2			
	3			
	4			

Example (continued)

Example (continued)

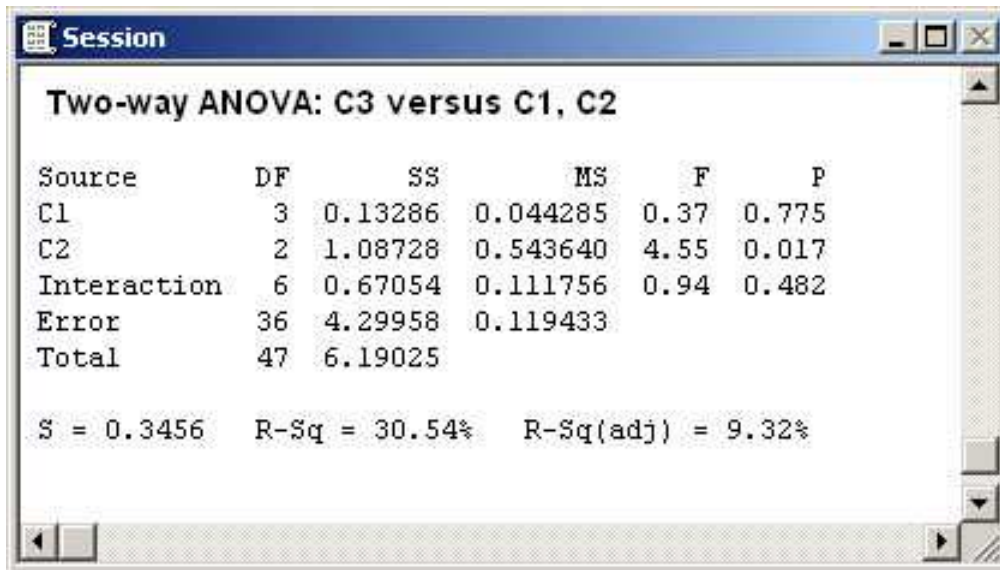
Example (continued)

Two-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor A					
Factor B					
Interaction					
Error					
Total					



Minitab output:



The image shows a screenshot of a Minitab Session window. The title bar reads "Session". The main content area displays the results of a Two-way ANOVA for C3 versus C1, C2. The results are presented in a table with columns for Source, DF, SS, MS, F, and P. Below the table, summary statistics are provided: S = 0.3456, R-Sq = 30.54%, and R-Sq(adj) = 9.32%.

Source	DF	SS	MS	F	P
C1	3	0.13286	0.044285	0.37	0.775
C2	2	1.08728	0.543640	4.55	0.017
Interaction	6	0.67054	0.111756	0.94	0.482
Error	36	4.29958	0.119433		
Total	47	6.19025			

S = 0.3456    R-Sq = 30.54%    R-Sq(adj) = 9.32%

**Example 11.3.2** Kerosene and oil lamps are decorative and, for some, essential for home lighting. A study was conducted to determine whether the light output is related to the type of oil used (factor A) and/or the size of the wick (factor B). For each of five oils ( $a = 5$ ) and three wick sizes ( $b = 3$ ), independent random samples of size  $n = 6$  were obtained. The light output for each lamp was measured in candlepower. The following ANOVA summary table was obtained (from Minitab).

Two-way ANOVA summary table

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Factor A	120.72	4	30.18	5.12	0.0011
Factor B	59.38	2	29.69	5.04	0.0088
Interaction	82.64	8	10.33	1.75	0.1008
Error	442.04	75	5.89		
Total	704.78	89			

- (a) Is there any evidence of interaction? Use  $\alpha = 0.05$ .
- (b) Is there any evidence that oil type and/or wick size affects the light output? Use  $\alpha = 0.05$ .

Example (continued)



## CHAPTER 12

# Correlation and Simple Linear Regression

---

## 12.0 Introduction

1. Two variables may be related in a variety of ways.
2. In this chapter, we consider two variables that are linearly related.

(a) Values of  $X$  and  $Y$  tend to *move* together.

Large values of  $X$  are associated with large values of  $Y$ , and small values of  $X$  are associated with small values of  $Y$ .

This is a positive linear relationship between  $X$  and  $Y$ .

Alternative explanation: as the values of  $X$  increase, so do the values of  $Y$ .

(b) Values of  $X$  and  $Y$  tend to *move* in opposite directions.

Small values of  $X$  are associated with large values of  $Y$ , and large values of  $X$  are associated with small values of  $Y$ .

This is a negative linear relationship between  $X$  and  $Y$ .

Alternative explanation: as the values of  $X$  decrease, the values of  $Y$  increase.

Suppose  $X$  and  $Y$  are linearly related.

1. We may be able to use a value  $x$  to predict the value of  $Y$ : regression analysis.
2. A scatter plot can be used to visualize the relationship between two variables.
3. We can measure linear association by computing the correlation.

This value is a measure of the strength of the linear association.

## 12.1 Simple Linear Regression

Deterministic relationship between  $x$  and  $y$ .

1.  $y$  is completely determined by the value of  $x$ .
2. Example:  $y = f(x)$ ,  $y = x^2 - 3x + 5$ .
3. Independent variable:  $x$     Dependent variable:  $y$

Linear (deterministic) relationship between  $x$  and  $y$ :  $y = \alpha + \beta x$ .

1. Slope:  $\beta$      $y$ -intercept:  $\alpha$ . Remember  $y = mx + b$ ?
2. Example:  $y = 5 + 3x$

Probabilistic model.

1. For a fixed value  $x$ , the value of the second variable is randomly distributed.
2. Example: Consider the relationship between the gross receipts and the number of people who dine at a restaurant.

If  $x = 62$  people, the amount of money spent is a random variable  $Y$ .

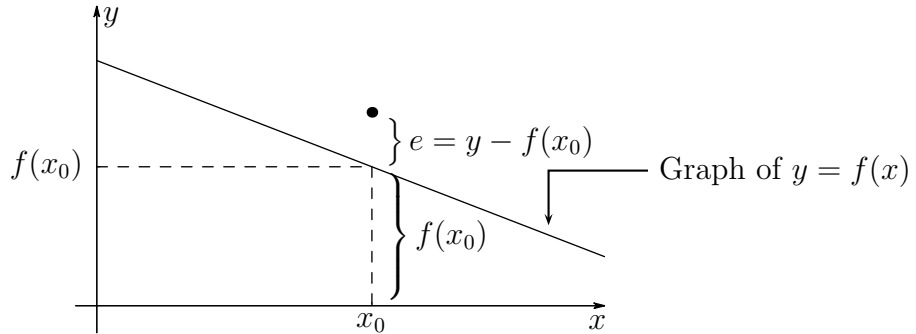
3. Independent variable:  $x$ .
4. Dependent variable:  $Y$ . An observed value of  $Y$  is denoted  $y$ .
5. Additive probabilistic model: a deterministic part and a random part.

The model can be written as

$$\begin{aligned} Y &= (\text{deterministic function of } x) + (\text{random deviation}) \\ &= f(x) + E \end{aligned}$$

where  $E$  is a random variable, called the random error.

An illustration of a probabilistic model:



Notation: Suppose there are  $n$  observations on fixed values of the independent variable.

1. The observed values of the independent variable are denoted  $x_1, x_2, \dots, x_n$ .
2.  $Y_i$  and  $y_i$  are the random variable and the observed value of the random variable associated with  $x_i$ , for  $i = 1, 2, \dots, n$ .

For each  $x_i$ , there is a corresponding random variable  $Y_i$ .

There are really  $n$  random variables in these problems.

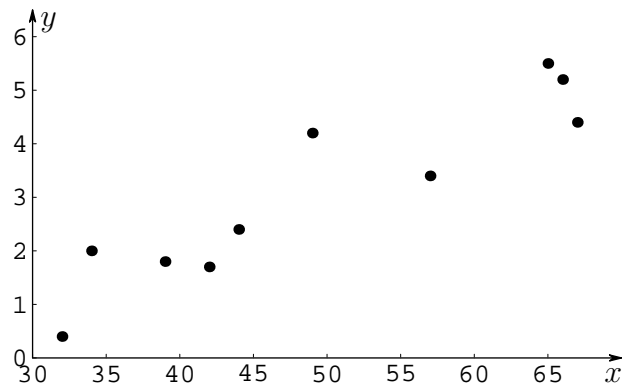
3. The data set consists of  $n$  ordered pairs:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .

**Example 12.1.1** Government officials in Nebraska have become concerned about chemical runoff from farms into rivers and streams. A study was conducted to examine the relationship between the percentage of land used for farming and the nitrate concentration in nearby streams. A random sample of streams was selected, the percentage of land used for farming in each drainage basin was recorded, and the nitrate concentration in each stream was measured (in milligrams per liter). The data are given in the following table.

Percentage, $x$	34	57	67	66	49	42	65	32	39	44
Concentration, $y$	2.0	3.4	4.4	5.2	4.2	1.7	5.5	0.4	1.8	2.4

- (a) Identify the independent and the dependent variables.
- (b) List the ordered pairs in the data set.
- (c) Construct a scatter plot for this data. What does the plot suggest about the relationship between the variables?

Example (continued)

**Remarks**

1. Simple linear regression model: the deterministic function  $f(x)$  is assumed to be linear:  
 $f(x) = \alpha + \beta x$ .
2. Add a possible deterministic straight line to the scatter plot above.

Data points lie close to the line.

Vertical distance between a point and the line depends on the value of the random error.



**Simple Linear Regression Model**

Let  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  be  $n$  pairs of observations such that  $y_i$  is an observed value of the random variable  $Y_i$ . We assume that there exist constants  $\alpha$  and  $\beta$  such that

$$Y_i = \alpha + \beta x_i + E_i$$

where  $E_1, E_2, \dots, E_n$  are independent, normal random variables with mean 0 and variance  $\sigma^2$ . That is,

1. The  $E_i$ 's are normally distributed (which implies that the  $Y_i$ 's are normally distributed).
2. The expected value of  $E_i$  is 0 (which implies that  $E(Y_i) = \alpha + \beta x_i$ ).
3.  $\text{Var}(E_i) = \sigma^2$  (which implies that  $\text{Var}(Y_i) = \sigma^2$ ).
4. The  $E_i$ 's are independent (which implies that the  $Y_i$ 's are independent).

**Remarks**

1.  $E_i$ 's: the **random deviations** or **random error terms**.
2.  $y = \alpha + \beta x$ : the **true regression line**.

Each point,  $(x_i, y_i)$ , lies *near* the true regression line, its distance from the line depending upon the value of the random error term,  $e_i$ .

3. The four assumptions in the simple linear regression model stated compactly:

$$E_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2).$$

More notation.

1. Consider each random variable  $Y_i = Y | x_i$ .

$\mu_{Y|x_i} = E(Y|x_i)$ : the expected value of  $Y$  for a fixed value  $x_i$

$\sigma_{Y|x_i}^2$ : the variance of  $Y$  for a fixed value  $x_i$

2. The simple linear regression model assumptions imply:

(a)  $\mu_{Y|x_i} = E(\alpha + \beta x_i + E_i) = \alpha + \beta x_i + E(E_i) = \alpha + \beta x_i$

(b)  $\sigma_{Y|x_i}^2 = \text{Var}(\alpha + \beta x_i + E_i) = \sigma^2$

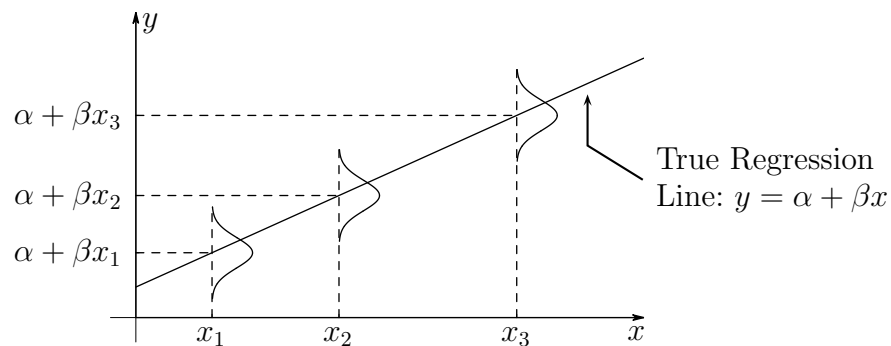
(c)  $Y|x_i$  is normal

3. Therefore, the mean value of  $Y$  is a linear function of  $x$ .

The true regression line passes through the *line of mean values*.

The variability in the distribution of  $Y$  is the *same* for every value of  $x$  (homogeneity of variance).

An illustration of the simple linear regression model assumptions and the resulting properties:



**Example 12.1.2** Agricultural studies suggest the antioxidant enzyme level in apples ( $y$ , superoxide dismutase activity, in mg) is related to cumulative temperature ( $x$ , degree-days greater than  $10^\circ\text{C}$  in the growing season). Suppose the true regression line is  $y = -1233 + 1.25x$ .

- (a) Find the expected antioxidant activity when the cumulative temperature is  $1200^\circ\text{C}$ .
- (b) How much change in antioxidant activity is expected if the cumulative temperature decreases by  $25^\circ\text{C}$ ? Increases by  $100^\circ\text{C}$ ?
- (c) Suppose  $\sigma = 30$  mg. Find the probability an observed value of antioxidant activity is less than 500 mg when the cumulative temperature is  $1400^\circ\text{C}$ .

**Example 12.1.3** Medical researchers believe the amount of fruits and vegetables in a person's diet ( $x$ , a score between 0 and 100 based on the number of servings per day) is linearly related to their systolic blood pressure ( $y$ ). For men between the ages of 35 and 50, suppose the true regression line is  $y = 162 - 0.7x$ .

- (a) Find the expected systolic blood pressure when the fruits and vegetables score is 45.
- (b) How much change in systolic blood pressure is expected if the fruits and vegetables score increases by 10? Decreases by 30?
- (c) Suppose  $\sigma = 5$ . Find the probability the systolic blood pressure is between 110 and 120 if the fruits and vegetables score is 68.

The line of best fit, or estimated regression line.

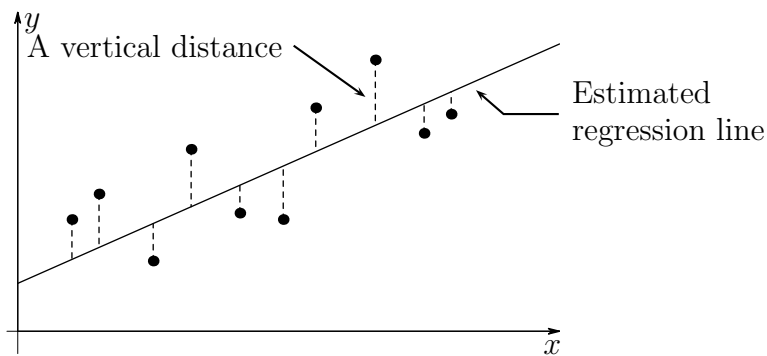
1. Assume that the observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  are independent.

Use the sample data to estimate the model parameters  $\alpha$  and  $\beta$ .

2. The line of best fit is obtained using the **principle of least squares**.

Minimize the sum of the squared deviations, or vertical distances from the observed points to the line.

The principle of least squares produces an estimated regression line such that the sum of all squared vertical distances is a minimum.



### Least-Squares Estimates

The least-squares estimates of the  $y$ -intercept ( $\alpha$ ) and the slope ( $\beta$ ) of the true regression line are

$$b = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2}$$

$$a = \frac{\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i}{n} = \bar{y} - b\bar{x}$$

The estimated regression line is  $y = a + bx$ .

**Remarks**

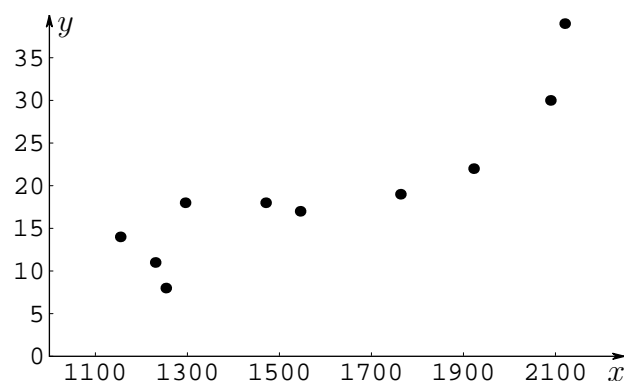
1. Always consider a scatter plot to make sure a linear model is reasonable.
2. If  $x^*$  is a specific value of the independent, or *predictor*, variable  $x$ , let  $y^* = a + bx^*$ .
  - (a)  $y^*$  is an estimate of the *mean* value of  $Y$  for  $x = x^*$ , denoted  $\mu_{Y|x^*}$ .
  - (b)  $y^*$  is also an estimate of an *observed* value of  $Y$  for  $x = x^*$ .

**Example 12.1.4** The offset crash test (40% frontal offset crash at 64 km/h into a deformable barrier) is used to determine an automobile's safety rating. Some researchers believe the offset score is linearly related to (and can be predicted by) the mass of the automobile. A random sample of automobiles was obtained, and the mass (in kg) and offset score for each are given in the following table.

Mass	1154	2089	1922	1230	1545	1763	1470	1295	1253	2120
Offset score	14	30	22	11	17	19	18	18	8	39

- (a) Find the estimated regression line.
- (b) Estimate the true mean offset score for an automobile with mass 1750 kg.

Here is a scatter plot of the data.



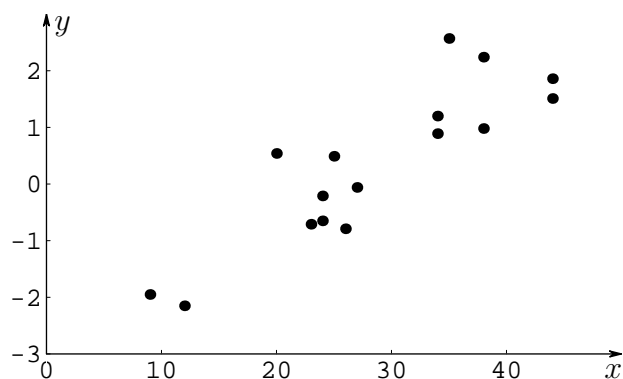
Example (continued)

**Example 12.1.5** Some medical researchers believe the amount of exercise a 12–18-year-old girl gets is linearly related to the density and strength of her hip bone, and is an important factor in preventing hip fractures later on in life. A random sample of 12–18-year-old girls was obtained, and the sports-exercise score ( $x$ , a unitless quantity) and bone mineral density ( $y$ , a T-score) were measured for each. The data are given in the following table.

$x$	24	27	44	44	35	25	24	20
$y$	-0.65	-0.06	1.51	1.86	2.57	0.49	-0.21	0.54
$x$	26	34	23	38	38	9	34	12
$y$	-0.79	1.20	-0.71	0.98	2.24	-1.95	0.89	-2.15

- Find the estimated regression line.
- Estimate the true mean bone mineral density for a sports-exercise score of 31.

Here is a scatter plot of the data.





Example (continued)

The variance  $\sigma^2$ .

1. A measure of the underlying variability in the simple linear regression model.
2. An estimate of  $\sigma^2$  is used to conduct hypothesis tests and to construct CIs related to simple linear regression.

More notation:

$$S_{xx} = \underbrace{\sum_{i=1}^n (x_i - \bar{x})^2}_{\text{definition}} = \underbrace{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2}_{\text{computational formula}}$$

$$S_{yy} = \underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{definition}} = \underbrace{\sum_{i=1}^n y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n y_i \right)^2}_{\text{computational formula}}$$

$$S_{xy} = \underbrace{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}_{\text{definition}} = \underbrace{\sum_{i=1}^n x_i y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}_{\text{computational formula}}$$

The *i*th predicted, or fitted, value, denoted  $\hat{y}_i$ , is  $\hat{y}_i = a + bx_i$ .

The *estimated* regression line evaluated at  $x_i$ .

The *i*th residual is  $y_i - \hat{y}_i$ .

A measure of how far away the observed value of  $Y$  is from the estimated value of  $Y$ .

The total variation in the data (**total sum of squares, SST**) is decomposed into:

1. A sum of the variation explained by the model (**sum of squares due to regression, SSR**) and
2. The variation about the regression line (the **sum of squares due to error, SSE**).

**Sum of Squares**

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{SSE}}$$

Here are the *computational formulas* for these sums of squares:

$$\text{SST} = S_{yy}, \quad \text{SSR} = bS_{xy}, \quad \text{SSE} = \text{SST} - \text{SSR}.$$

ANOVA table.

1. Summary of regression computations.
2. Mean squares: corresponding sums of squares divided by the associated degrees of freedom.
3. *F* test for a significant regression and *p* value included.

ANOVA summary table for simple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	<i>F</i>	<i>p</i> value
Regression	SSR	1	$\text{MSR} = \frac{\text{SSR}}{1}$	$\frac{\text{MSR}}{\text{MSE}}$	<i>p</i>
Error	SSE	<i>n</i> - 2	$\text{MSE} = \frac{\text{SSE}}{n - 2}$		
Total	SST	<i>n</i> - 1			

**Coefficient of Variation**

The **coefficient of variation**, denoted *r*<sup>2</sup>, is a measure of the proportion of the variation in the data that is explained by the regression model, and is defined by

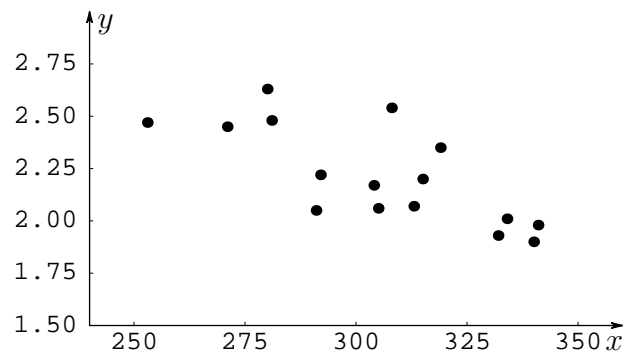
$$r^2 = \text{SSR}/\text{SST}.$$

**Example 12.1.6** A recent medical study examined the relationship between the voluntary running distance and the weight of laboratory rats. A random sample of animals was obtained. The animals were weighed (in grams) and placed in individual cages with running wheels. The running distance during the first 24 hours in the cage was measured (in km). The data are given in the following table.

Weight, $x$	341	319	313	253	340	315	308	304
Distance, $y$	1.98	2.35	2.07	2.47	1.90	2.20	2.54	2.17
Weight, $x$	334	291	271	292	305	280	281	332
Distance, $y$	2.01	2.05	2.45	2.22	2.06	2.63	2.48	1.93

- (a) Find the estimated regression line.  
 (b) Complete the ANOVA table (without the  $p$  value), and find the coefficient of variation.

Here is a scatter plot of the data.



Example (continued)

ANOVA summary table for simple linear regression

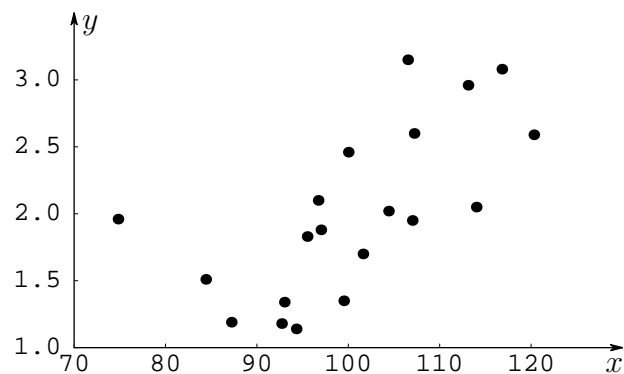
Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Regression					—
Error					
Total					

**Example 12.1.7** Pantothenic acid is a B-complex vitamin and is essential for human growth. A study was conducted to examine the relationship between the amount of nitrogen in the top two feet of soil ( $x$ , measured in pounds per acre) at the base of a tree and the mean amount of pantothenic acid ( $y$ , measured in mg) in avocados harvested from that tree. A random sample of avocado trees was obtained, and the data are given in the following table.

Nitrogen	93.0	74.8	84.4	100.0	104.4	107.2	92.7	107.0	96.7	101.6
Acid	1.34	1.96	1.51	2.46	2.02	2.60	1.18	1.95	2.10	1.70
Nitrogen	114.0	99.5	87.2	106.5	95.5	120.3	94.3	97.0	113.1	116.8
Acid	2.05	1.35	1.19	3.15	1.83	2.59	1.14	1.88	2.96	3.08

- (a) Find the estimated regression line.  
 (b) Complete the ANOVA table (without the  $p$  value), and find the coefficient of variation.

Here is a scatter plot of the data.



Example (continued)

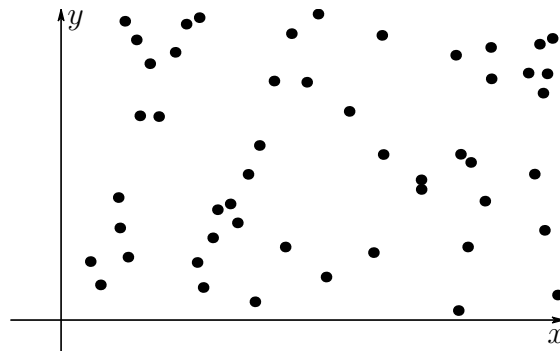
ANOVA summary table for simple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Regression					—
Error					
Total					

## 12.2 Hypothesis Tests and Correlation

1. The mean squares (in the ANOVA table) are used to determine whether the linear relationship is statistically significant.
2. Null hypothesis: variation in  $Y$  completely random, independent of the value of  $x$ .

Scatter plot would have a shotgun appearance.



3. Simple linear regression: test of significance equivalent to testing  $H_0: \beta = 0$ .

If  $H_0$  true, mean value of  $Y$  for any value of  $x$  is the same.

### Hypothesis Test for a Significant Regression

A hypothesis test for a significant regression with significance level  $\alpha$  has the form:

$H_0$ : There is no significant linear relationship ( $\beta = 0$ )

$H_a$ : There is a significant linear relationship ( $\beta \neq 0$ )

$$\text{TS: } F = \frac{\text{MSR}}{\text{MSE}}$$

$$\text{RR: } F \geq F_{\alpha, 1, n-2}$$

### Remarks

1. The null hypothesis is rejected only for large values of the test statistic.
2. Often called a *model utility test*. In general, if  $H_0$  is rejected, then  $r^2$  is usually large.
3. Alternatively, we can conduct this hypothesis test by using the  $p$  value.



Consider the random variable  $B$ , an estimator for  $\beta$ , and

$S^2 = \text{MSE} = \text{SSE}/(n - 2)$ , an estimator for the underlying variance  $\sigma^2$ .

If the simple linear regression assumptions are true, then  $S^2$  is an unbiased estimator for  $\sigma^2$ , and the estimator  $B$  has the following properties.

1.  $B$  is an unbiased estimator for  $\beta$ :  $E(B) = \mu_B = \beta$ .
2. The variance of  $B$  is  $\text{Var}(B) = \sigma_B^2 = \sigma^2/S_{xx}$ .

If we use  $s^2$  as an estimate of  $\sigma^2$ , then an estimate of the variance of  $B$  is  $s_B^2 = s^2/S_{xx}$ .

3. The random variable  $B$  has a normal distribution.

### Theorem

If the simple linear regression assumptions are true, then the random variable

$$T = \frac{B - \beta}{S/\sqrt{S_{xx}}} = \frac{B - \beta}{S_B}$$

has a  $t$  distribution with  $n - 2$  degrees of freedom.

For simple linear regression, the following hypothesis test (with  $\beta_0 = 0$ ) is equivalent to an  $F$  test for a significant regression with significance level  $\alpha$ .

### Hypothesis Test and Confidence Interval Concerning $\beta$

$$H_0: \beta = \beta_0$$

$$H_a: \beta > \beta_0, \quad \beta < \beta_0, \quad \text{or} \quad \beta \neq \beta_0$$

$$\text{TS: } T = \frac{B - \beta_0}{S_B}$$

$$\text{RR: } T \geq t_{\alpha, n-2}, \quad T \leq -t_{\alpha, n-2}, \quad \text{or} \quad |T| \geq t_{\alpha/2, n-2}$$

A  $100(1 - \alpha)\%$  confidence interval for  $\beta$  has as endpoints the values

$$b \pm t_{\alpha/2, n-2} s_B.$$

Similar properties, a hypothesis test procedure, and confidence interval concerning the simple linear regression parameter  $\alpha$ .

The estimator  $A$  has the following properties.

1.  $A$  is an unbiased estimator for  $\alpha$ :  $E(A) = \mu_A = \alpha$ .

2. The variance of  $A$  is  $\text{Var}(A) = \sigma_A^2 = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{nS_{xx}}$ .

If we use  $s^2$  as an estimate of  $\sigma^2$ , then an estimate of the variance of  $A$  is  $s_A^2 = \frac{s^2 \sum_{i=1}^n x_i^2}{nS_{xx}}$ .

3. The random variable  $A$  has a normal distribution.

### 12.1 Theorem

If the simple linear regression assumptions are true, then the random variable

$$T = \frac{A - \alpha}{S \sqrt{\sum_{i=1}^n x_i^2 / nS_{xx}}} = \frac{A - \alpha}{S_A}$$

has a  $t$  distribution with  $n - 2$  degrees of freedom.

### Hypothesis Test and Confidence Interval Concerning $\alpha$

$$H_0: \alpha = \alpha_0$$

$$H_a: \alpha > \alpha_0, \quad \alpha < \alpha_0, \quad \text{or} \quad \alpha \neq \alpha_0$$

$$\text{TS: } T = \frac{A - \alpha_0}{S_A}$$

$$\text{RR: } T \geq t_{\alpha, n-2}, \quad T \leq -t_{\alpha, n-2}, \quad \text{or} \quad |T| \geq t_{\alpha/2, n-2}$$

A  $100(1 - \alpha)\%$  confidence interval for  $\alpha$  has as endpoints

$$a \pm t_{\alpha/2, n-2} s_A.$$

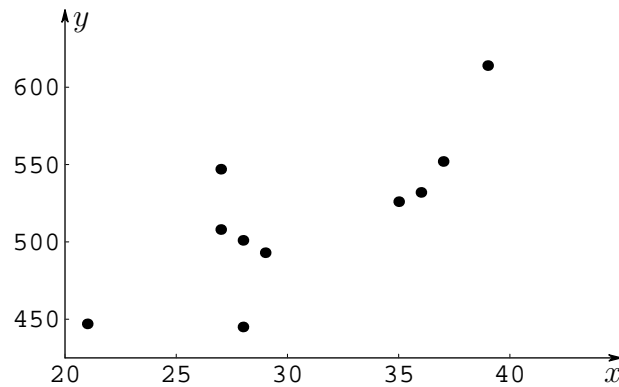
*Note:* the symbol  $\alpha$  is being used in two different ways here—as the significance level and as a parameter of the regression line.

**Example 12.2.1** The manufacturer of an artificial sweetener is investigating the relationship between the amount of time the mixture is steam-heated ( $x$ , measured in minutes) and the sweetness compared to regular sugar ( $y$ ). A random sample of batches was obtained, and the steam-heating time and the sweetness for each are given in the following table.

Time	36	27	29	28	37	35	39	21	28	27
Sweetness	532	508	493	501	552	526	614	447	445	547

- Find the estimated regression line.
- Complete the ANOVA table and conduct an  $F$  test for a significant regression. Use a significance level of 0.05.
- Conduct a  $t$  test (concerning  $\beta$ ) for a significant regression. Use a significance level of 0.05.

Here is a scatter plot of the data.



Example (continued)

ANOVA summary table for simple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Regression					
Error					
Total					

Example (continued)

**Example 12.2.2** Some medical researchers believe that the heat from a laptop computer—when used on the lap—can increase body temperature. Twenty-nine men were randomly selected, and the amount of time each person used the laptop ( $x$ , in minutes) and body temperature at the end of usage ( $y$ , in °F) was measured for each. The summary statistics are given below.

$$\begin{array}{lll} \sum_{i=1}^{29} x_i = 1556.0 & \sum_{i=1}^{29} y_i = 2962.7 & \sum_{i=1}^{29} x_i y_i = 160,387.0 \\ \sum_{i=1}^{29} x_i^2 = 102,556.0 & \sum_{i=1}^{29} y_i^2 = 303,045.0 & \end{array}$$

- Find the estimated regression line.
- Complete the ANOVA table and conduct an  $F$  test for a significant regression. Use a significance level of 0.01.
- Conduct a  $t$  test (concerning  $\beta$ ) for a significant regression. Use a significance level of 0.01.
- Conduct the hypothesis test  $H_0: \alpha = 98.6$  versus  $H_a: \alpha \neq 98.6$  with significance level 0.0001. Is there any evidence to suggest that the value of  $\alpha$  is different from 98.6?

Example (continued)

ANOVA summary table for simple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Regression					
Error					
Total					

Example (continued)



Correlation: a statistical term indicating a relationship between two variables.

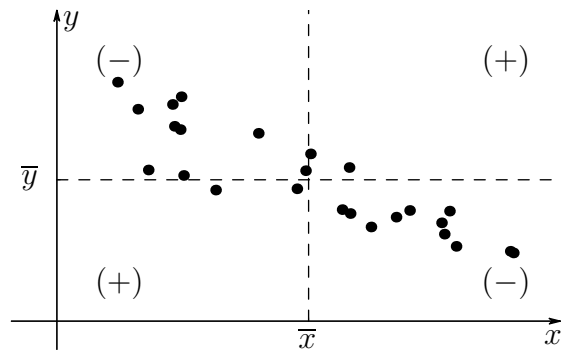
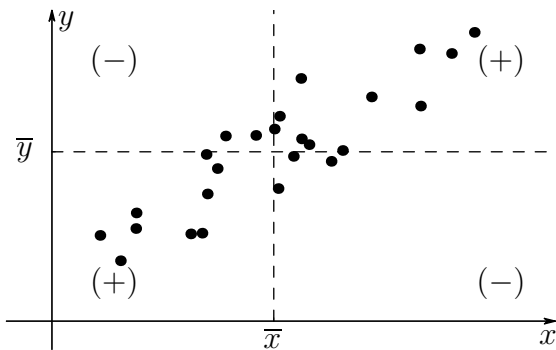
Examples:

1. Exercise is correlated with the risk of a heart attack.
2. The height of waves onshore at high tide is correlated with the wind speed.

Sample correlation coefficient.

1. A measure of the strength of a linear relationship between two continuous variables.
2. Suppose there are  $n$  pairs of observations:  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ .
3. If large values of  $x$  are associated with large values of  $y$ , then  $x$  and  $y$  are positively related.
4. If small values of  $x$  are associated with large values of  $y$ , then  $x$  and  $y$  are negatively related.

Consider the following scatter plots and the quantity  $S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$ .



The magnitude of  $S_{xy}$  depends on the units of  $x$  and  $y$ .

The sample correlation coefficient adjusts  $S_{xy}$  so that it is unit-independent.

**Sample Correlation Coefficient**

Suppose there are  $n$  pairs of observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . The **sample correlation coefficient** for these  $n$  pairs is

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{\sqrt{\left[ \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right] \left[ \sum_{i=1}^n y_i^2 - \frac{1}{n} \left( \sum_{i=1}^n y_i \right)^2 \right]}}$$

**Remarks**

1. The value of  $r$  does not depend upon the order of the variables and is independent of units.
2.  $-1 \leq r \leq +1$

$r = +1$  if all of the ordered pairs lie on a straight line with positive slope.

$r = -1$  if all of the ordered pairs lie on a straight line with negative slope.

3. The square of the sample correlation coefficient is the coefficient of variation in a simple linear regression model.

Since  $-1 \leq r \leq +1$ ,  $0 \leq r^2 \leq 1$ .

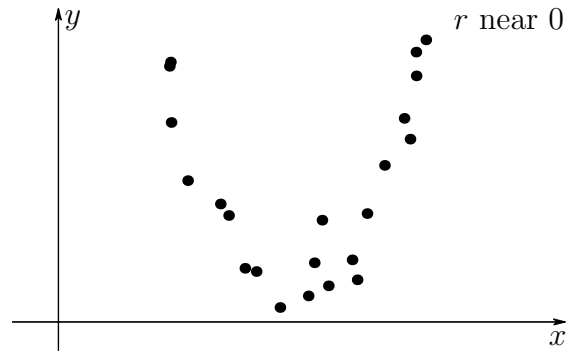
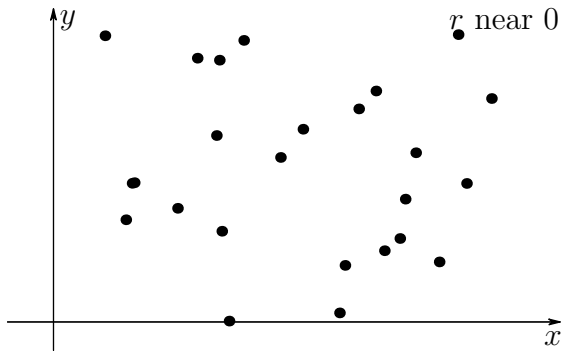
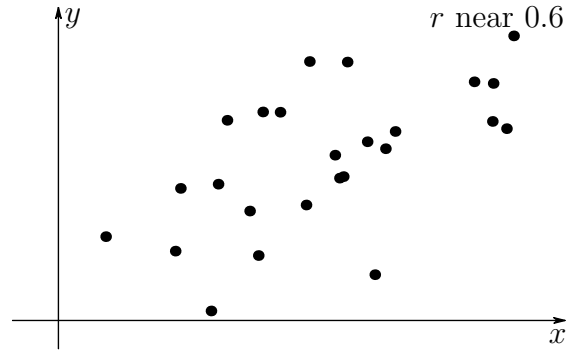
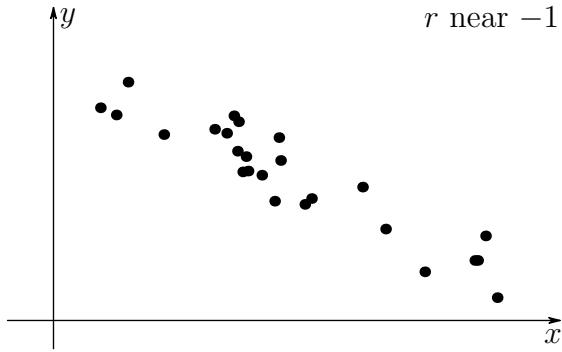
4.  $r$  is a measure of the strength of a *linear* relationship.

If  $r$  is near 0, there is no evidence of a linear relationship, but  $x$  and  $y$  may be related in another way.

The following general *rule* is used to describe the linear relationship between two variables, based on the value of the sample correlation coefficient.

1. If  $0 \leq |r| \leq 0.5$ , then there is a *weak* linear relationship.
2. If  $0.5 < |r| \leq 0.8$ , then there is a *moderate* linear relationship.
3. If  $|r| > 0.8$ , then there is a *strong* linear relationship.

Illustrations of the sample correlation coefficient:



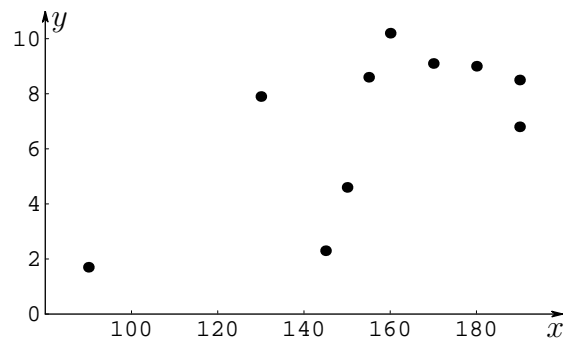
**Example 12.2.3** A company is testing a new seasonal allergy medicine for sneezing and itchy, watery eyes. A random sample of adults with allergies was obtained, and each was given a certain dose of the medicine (in mg). The length of time (in hours) that each person experienced allergy relief was recorded, and the data are given in the following table.

Dose, $x$	150	190	160	155	145	130	90	190	180	170
Relief time, $y$	4.6	8.5	10.2	8.6	2.3	7.9	1.7	6.8	9.0	9.1

Find the sample correlation coefficient between dose and relief time, and interpret this value.

Example (continued)

Scatter plot of the data:



**Example 12.2.4** An ornithologist believes that there is a relationship between the height a bald eagle flies and the air temperature. Bald eagles from various parts of the United States were randomly selected. The ground-level temperature (in °F) at noon and the mean height (in feet) during a flight around that time was recorded for each bird. The data are given in the following table.

Temperature	51	44	43	76	76	80	71	72	39	50
Height	1292	1672	3190	2161	2442	2944	3177	1814	3333	3108
Temperature	60	64	42	70	69	76	42	58	49	23
Height	1918	2338	2434	3104	2110	3039	2204	1627	3111	2904

Find the sample correlation coefficient between temperature and height, and interpret this value.

Example (continued)

## 12.3 Inferences Concerning the Mean Value and an Observed Value of $Y$ for $x = x^*$

Suppose  $x^*$  is a specific value of the independent variable  $x$  and  $y = a + bx$  is the estimated regression line. The value  $y^* = a + bx^*$  is

1. an estimate of the *mean* value of  $Y$  for  $x = x^*$ , and
2. an estimate of an *observed* value of  $Y$  for  $x = x^*$ .

The error in estimating the *mean* value of  $Y$  is less than the error in estimating an *observed* value of  $Y$ .

Consider

1. A hypothesis test and confidence interval concerning the mean value of  $Y$  for  $x = x^*$ .
2. A *prediction interval* for an observed value of  $Y$  if  $x = x^*$ .

Suppose the simple linear regression assumptions are true.

For  $x = x^*$ , the random variable  $A + Bx^*$  has the following properties.

1. It has a normal distribution.
2. The expected value is  $E(A + Bx^*) = \alpha + \beta x^*$
3. The variance is  $\text{Var}(A + Bx^*) = \sigma^2 \left[ \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$

The standard deviation is the square root of this expression.

An estimate of the standard deviation is obtained by using  $s$  as an estimate for  $\sigma$ .

Note: The variance of  $A + Bx^*$  is smallest when  $x = \bar{x}$ .

### Theorem

If the simple linear regression assumptions are true, then the random variable

$$T = \frac{(A + Bx^*) - (\alpha + \beta x^*)}{S \sqrt{(1/n) + [(x^* - \bar{x})^2/S_{xx}]}}$$

has a  $t$  distribution with  $n - 2$  degrees of freedom.

**Hypothesis Test and Confidence Interval Concerning the Mean Value of  $Y$  for  $x = x^*$** 

$$H_0: y^* = y_0^*$$

$$H_a: y^* > y_0^*, \quad y^* < y_0^*, \quad \text{or} \quad y^* \neq y_0^*$$

$$\text{TS: } T = \frac{(A + Bx^*) - y_0^*}{S\sqrt{(1/n) + [(x^* - \bar{x})^2/S_{xx}]}}$$

$$\text{RR: } T \geq t_{\alpha, n-2}, \quad T \leq -t_{\alpha, n-2}, \quad \text{or} \quad |T| \geq t_{\alpha/2, n-2}$$

A  $100(1 - \alpha)\%$  confidence interval for  $\mu_{Y|x^*}$ , the mean value of  $Y$  for  $x = x^*$ , has as endpoints the values

$$(a + bx^*) \pm t_{\alpha/2, n-2} s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

**Example 12.3.1** A transportation worker is testing various road salt products for use during the winter months. A random sample of products is obtained, and the percentage ( $x$ , by weight) of salt in each is measured. The road salt is then tested on a 1-inch slab of ice, and the time to melt a hole completely through the ice is recorded ( $y$ , in minutes). The data are given in the following table.

$x$	28.5	25.0	24.6	17.8	26.4	27.3	26.0	17.9	22.1	19.2	21.4	24.6
$y$	16.4	20.6	28.7	30.7	17.1	15.6	18.9	30.4	21.0	27.1	17.3	24.0

The estimated regression line is  $y = 49.902 - 1.1789x$ , and the summary ANOVA table is given below.

ANOVA summary table for simple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Regression	204.80	1	204.80	14.01	0.0038
Error	146.14	10	14.61		
Total	350.94	11			

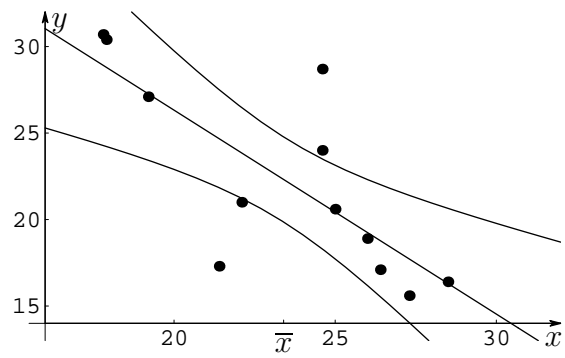
- For  $x = 25\%$ , conduct a hypothesis test to determine whether there is any evidence that the mean time to melt the ice is less than 20 minutes. Use a significance level of 0.05.
- Construct a 95% confidence interval for the true mean time to melt the ice if the salt percentage is 40.



Example (continued)

Example (continued)

Illustration of the 95% confidence bands:



Constructing an interval of possible values for an *observed* value of  $Y$  if  $x = x^*$ .

1. An observed value of  $Y$  (for  $x = x^*$ ) is a value of a random variable, not a fixed parameter.

The error of estimation for an observed value is larger than the error of estimation for a single mean value of  $Y$ .

2. The interval of possible values is called a **prediction interval**.
3. The random variable  $(A + Bx^*) - (\alpha + \beta x^* + E^*)$  is used in order to derive a prediction interval for  $Y$ .

**Prediction Interval**

A  $100(1 - \alpha)\%$  prediction interval for an observed value of  $Y$  when  $x = x^*$  has as endpoints the values

$$(a + bx^*) \pm t_{\alpha/2, n-2} s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}$$

**Example 12.3.2** Cookware coated with Teflon can give off toxic gases when heated for only a short time on a conventional stove. A random sample of Teflon-coated pans was obtained, and the thickness of the coating on the cooking surface was measured ( $x$ , in inches) on each. The Teflon-coated pans were then heated from a cold start on a stove. The time ( $y$ , in minutes) required until toxic gases were emitted was recorded for each pan. The data are given in the following table.

x	0.00181	0.00127	0.00080	0.00117	0.00128	0.00156	0.00183	0.00163
y	5.1	2.1	3.1	3.1	3.0	5.0	4.4	6.5
x	0.00092	0.00164	0.00097	0.00167	0.00164	0.00039	0.00066	
y	2.7	1.2	4.5	3.7	2.2	1.1	1.6	

- (a) Find the estimated regression line, and complete the ANOVA summary table.
- (b) Suppose a single Teflon pan with a coating 0.00125 inches thick is selected at random. Find a 95% prediction interval for the heating time required for the pan to emit toxic gases.

Example (continued)

ANOVA summary table for simple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Regression					
Error					
Total					

Example (continued)

## 12.4 Regression Diagnostics

Simple linear regression model assumptions:  $E_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$ .

If the true regression line were known, the set of actual random errors,

$$\begin{aligned} e_1 &= y_1 - (\alpha + \beta x_1) \\ e_2 &= y_2 - (\alpha + \beta x_2) \\ &\vdots \\ e_n &= y_n - (\alpha + \beta x_n) \end{aligned}$$

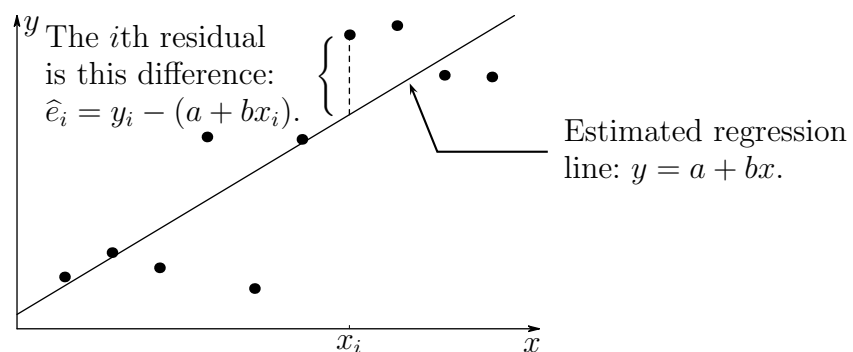
could be computed and used to check the assumptions.

Don't know the values for  $\alpha$  and  $\beta$ , the **residuals**, or deviations from the estimated regression line,

$$\begin{aligned} \hat{e}_1 &= y_1 - (a + bx_1) \\ \hat{e}_2 &= y_2 - (a + bx_2) \\ &\vdots \\ \hat{e}_n &= y_n - (a + bx_n) \end{aligned}$$

are used to check for assumption violations.

Illustration of a residual:



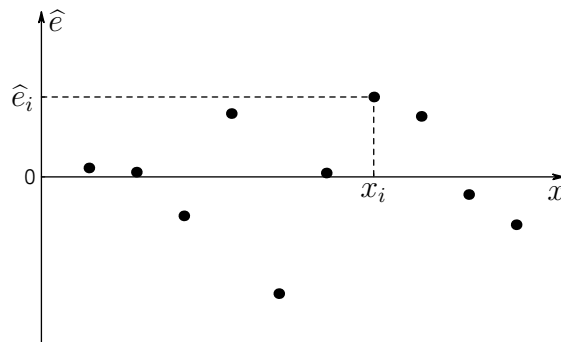
These estimates of the random errors are used in a variety of diagnostic checks.

This section presents several preliminary graphical procedures used to reveal assumption violations.

Checking for violations in assumptions:

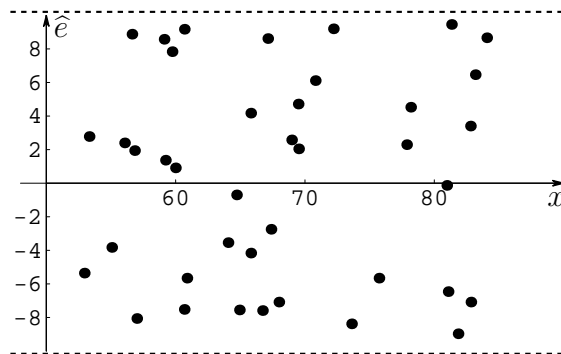
1. Normal probability plot of the residuals: check the normality assumption.
2. Histogram, stem-and-leaf plot of residuals: check the normality assumption.
3. Scatter plot of residuals versus the independent variable values: ordered pairs  $(x_i, \hat{e}_i)$ .

Illustration:



4. If no violation in assumptions: scatter plot should look like a horizontal band around zero with randomly distributed points.

Illustration:



Patterns in a residual plot that indicate a possible violation in assumptions:

1. A distinct *curve* in the plot, either mound- or bowl-shaped (parabolic).

An additional or different variable may be necessary.

A *linear* model is not appropriate.

2. A *nonconstant* spread.

Suggests that the variance is not constant.

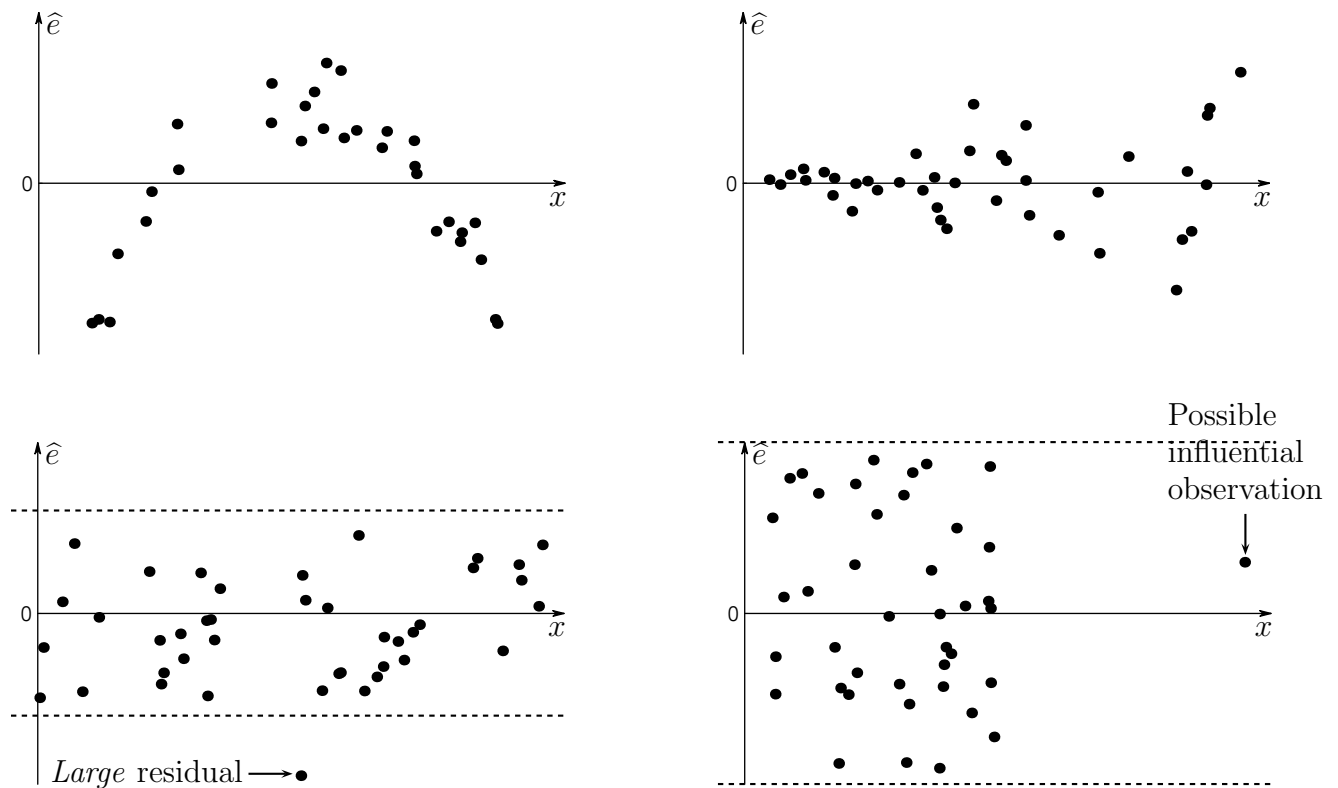
3. Any unusually large (in magnitude) residual.

The data may have been recorded or entered incorrectly.

4. Any *outliers*.

One observation has an unusually large influence on the estimated regression line.

Illustrations:





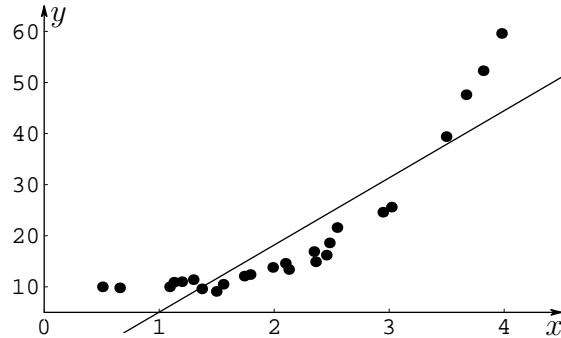
**Example 12.4.1** A grocery-store manager is investigating the relationship between the price of chicken and the demand for beef. He believes that as the price of chicken increases, consumers buy more beef. A sample of days was randomly selected, and the price of chicken ( $x$ , dollars per pound for a whole chicken) and the number of pounds of hamburger sold ( $y$ ) were recorded for each day. The data are given in the following table.

$x$	0.94	1.07	1.12	1.17	0.84	0.66	1.10	1.25	0.85	0.72
$y$	333	383	452	574	302	353	407	536	422	345
$x$	0.87	0.93	1.29	1.02	1.25	1.07	1.07	1.09	0.73	1.21
$y$	364	502	565	501	566	518	480	539	293	621

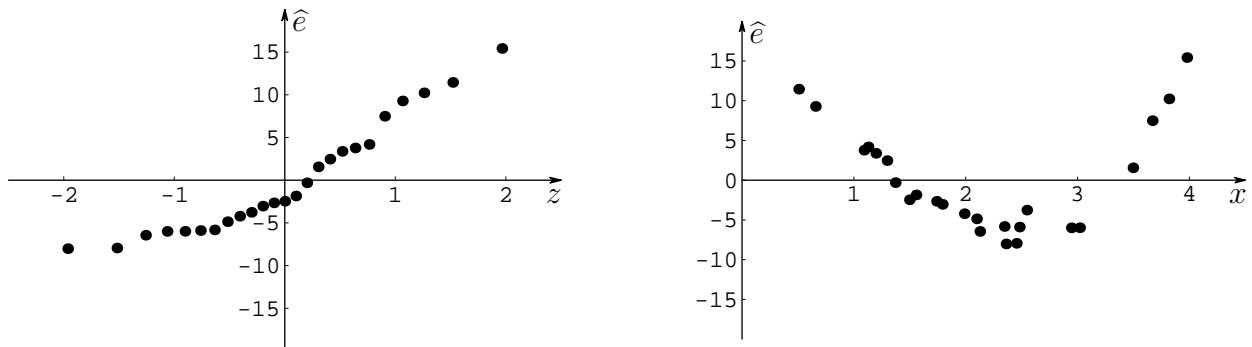
The estimated regression line is  $y = 5.75 + 441.53x$ . Compute the residuals, construct a normal probability plot of the residuals, and carefully sketch a graph of the residuals versus the independent (predictor) variable values. Is there any evidence of a violation in the assumptions for a simple linear regression model?

Example (continued)

**Example 12.4.2** A study was conducted to examine the relationship between the heat dissipated by an office printer and the number of pages printed. A sample of hour-long intervals during several workdays was randomly selected. The number of pages printed ( $x$ , in thousands of pages) and the dissipated heat ( $y$ , in kBTU) were recorded for each hour. The estimated regression line is  $y = -8.103 + 13.145x$ . The following scatter plot of the data includes a graph of the estimated regression line.



Here are a normal probability plot of the residuals and a plot of the residuals versus the independent (predictor) variable values.



Is there any evidence of a violation in the assumptions for a simple linear regression model? Justify your answer. How would you improve the regression model?

## 12.5 Multiple Linear Regression

Many real-world problems involve a model with a dependent variable  $Y$  and at least two independent variables,  $x_1, x_2, \dots, x_k$ .

1. Example: First-year college GPA might be predicted by high school GPA, total SAT scores, class rank, and quality of letters of recommendation.

Example: The yield on a 30-year treasury bond might be predicted by the prime rate, the unemployment rate, the consumer price index, retail sales, and the M1 money supply.

2. Purpose of this section: extend the simple linear regression model to  $k$  ( $\geq 2$ ) predictor variables.

### Multiple Linear Regression Model

Let  $(x_{11}, x_{21}, \dots, x_{k1}, y_1), (x_{12}, x_{22}, \dots, x_{k2}, y_2), \dots, (x_{1n}, x_{2n}, \dots, x_{kn}, y_n)$  be  $n$  sets of observations such that  $y_i$  is an observed value of the random variable  $Y_i$ . We assume that there exist constants  $\beta_0, \beta_1, \dots, \beta_k$  such that

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + E_i$$

where  $E_1, E_2, \dots, E_n$  are independent, normal random variables with mean 0 and variance  $\sigma^2$ . That is,

1. The  $E_i$ 's are normally distributed (which means that the  $Y_i$ 's are normally distributed).
2. The expected value of  $E_i$  is 0 (which implies that  $E(Y_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$ ).
3.  $\text{Var}(E_i) = \sigma^2$  (which implies that  $\text{Var}(Y_i) = \sigma^2$ ).
4. The  $E_i$ 's are independent (which implies that the  $Y_i$ 's are independent).

### Remarks

1. Double subscript on  $x$ : indicates both the variable and the observation.

$x_{42}$ : value of the variable  $x_4$  that corresponds to the observed value of  $y_2$ .

$x_{24}$ : value of the variable  $x_2$  that corresponds to the observed value of  $y_4$ .

2.  $E_i$ 's: **random deviations** or **random error terms**.
3. **True regression equation:**  $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_kx_k$ .

This equation is a *linear* function of the unknown parameters  $\beta_0, \beta_1, \dots, \beta_k$ .

The graph of the true regression line is, in general, a surface.

4.  $\beta_0, \beta_1, \dots, \beta_k$ : partial regression coefficients.

$\beta_i$  represents the mean change in  $y$  for every increase of one unit in  $x_i$  if the values of all other predictor variables are held fixed.

In this section

1. Focus on the method for finding the best deterministic linear model.

Find estimates of the unknown parameters  $\beta_0, \beta_1, \dots, \beta_k$ .

2. Hypothesis tests:

(a) Does the overall model explain a significant amount of the variability in the dependent variable?

(b) Evaluate the contribution of each independent variable.

3. Principle of least squares is used again to minimize the sum of squares due to error.

4. Use technology to compute the estimates  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  for the true regression parameters  $\beta_0, \beta_1, \dots, \beta_k$ .

5. Notation:

For specific values of the independent variables, let  $(x_1^*, x_2^*, \dots, x_k^*) = \mathbf{x}^*$ .

$$y^* = \hat{\beta}_0 + \hat{\beta}_1x_1^* + \hat{\beta}_2x_2^* + \cdots + \hat{\beta}_kx_k^*.$$

$y^*$  is an estimate of the mean value of  $Y$  for  $\mathbf{x} = \mathbf{x}^*$ , denoted  $\mu_{Y|\mathbf{x}^*}$ .

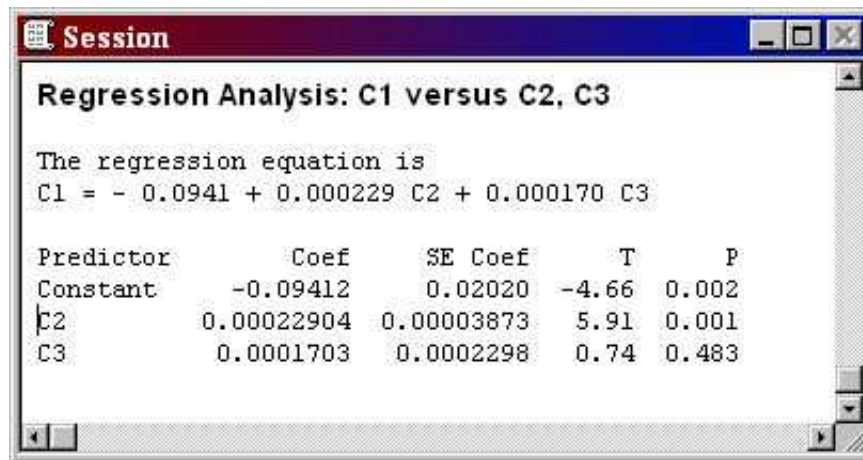
Also an estimate of an observed value of  $Y$  for  $\mathbf{x} = \mathbf{x}^*$ .

**Example 12.5.1** Transpiration rates, the evaporation of water from plants, is an important measure in order to understand the relationship between water, carbon dioxide, and energy in the atmosphere. Tree transpiration rates are often assessed by measuring the sap velocity. Researchers believe this velocity is affected by the wood density and the ambient temperature. A random sample of conifers, type *Dacrydium cupressinum*, were obtained. The sap velocity (in mm per second), the wood density (in kg per cubic meter), and the temperature (in degrees Fahrenheit) were obtained for each. The data are given in the following table.

Velocity	0.0025	0.0108	0.0078	0.0016	0.0207	0.0269	0.0113	0.0457	0.0147	0.0355
Density	422	401	420	401	457	453	403	562	458	507
Temperature	68	66	56	78	64	73	59	77	51	64

- Find the estimated regression equation.
- Estimate the true mean sap velocity for a wood density of 550 kg/m and a temperature of 64°F.

The Minitab output is shown below.



**Session**

**Regression Analysis: C1 versus C2, C3**

The regression equation is  
 $C1 = -0.0941 + 0.000229 C2 + 0.000170 C3$

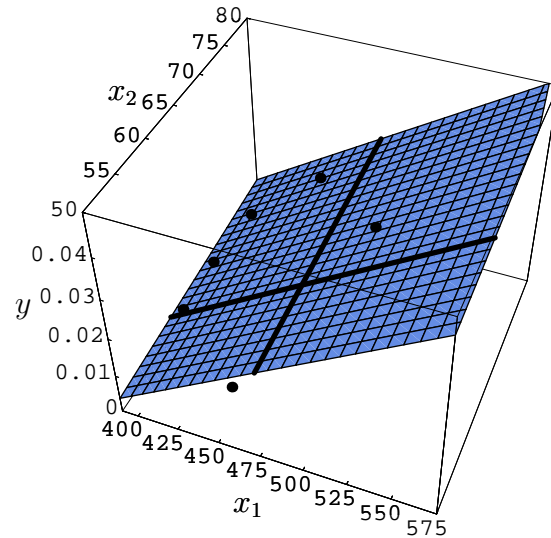
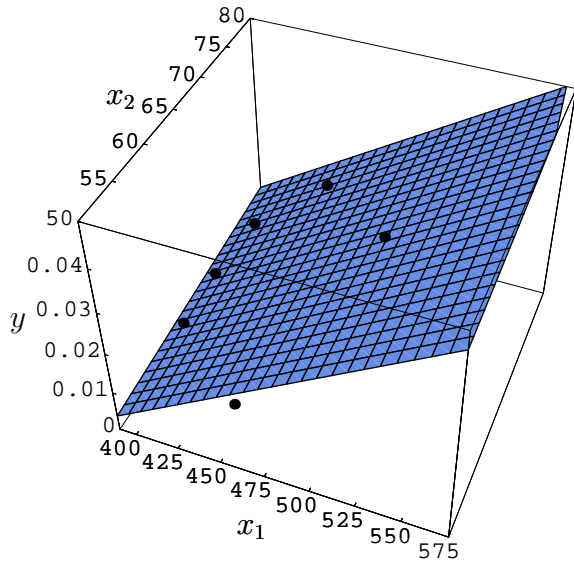
Predictor	Coef	SE Coef	T	P
Constant	-0.09412	0.02020	-4.66	0.002
C2	0.00022904	0.00003873	5.91	0.001
C3	0.0001703	0.0002298	0.74	0.483

C1: sap velocity  
 C2: wood density  
 C3: temperature

Example (continued))

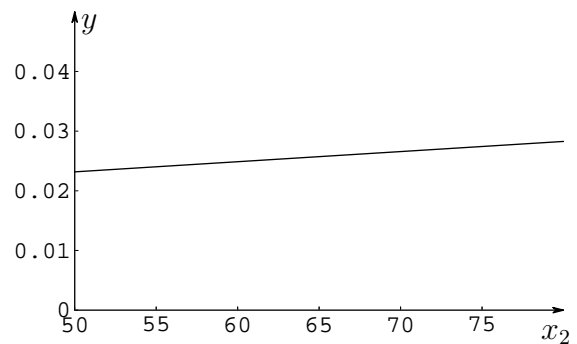
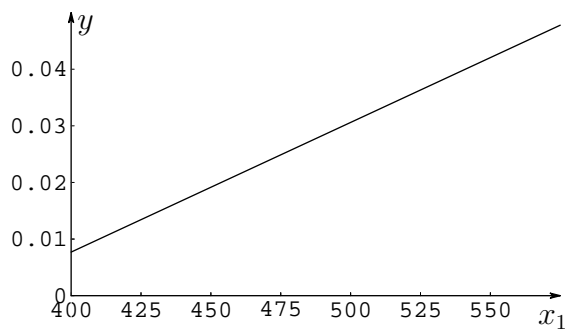
The graph of  $y = -0.0941 + 0.000229x_1 + 0.000172x_2$  is a plane in three dimensions.

Here is a three-dimensional scatter plot of the data and the graph of the estimated regression line, and other graphs to illustrate how  $y$  depends on each predictor variable separately.



A scatter plot of the data and the graph of the estimated regression equation.

A scatter plot of the data, the graph of the estimated regression equation, and two lines; one for  $x_1 = 475$  held constant; one for  $x_2 = 60$  held constant.



A graph of  $y$  versus  $x_1$  when  $x_2 = 60$ .

A graph of  $y$  versus  $x_2$  when  $x_1 = 475$ .



Multiple regression concepts

1. The variance  $\sigma^2$  is a measure of the underlying variability in the model.

An estimate of  $\sigma^2$  is used to conduct hypothesis tests and to construct confidence intervals related to multiple linear regression.

2. The  $i$ th predicted value is  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_k x_{ki}$ .
3. The  $i$ th residual is  $y_i - \hat{y}_i$ .
4. The total sum of squares, the sum of squares due to regression, and the sum of squares due to error have the same definitions, and the sum of squares identity is also true.

The total variation in the data (SST) is decomposed into a sum of the variation explained by the model (SSR) and the variation about the regression equation (SSE).

5. ANOVA table:

ANOVA summary table for multiple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Regression	SSR	$k$	$MSR = \frac{SSR}{k}$	$\frac{MSR}{MSE}$	$p$
Error	SSE	$n - k - 1$	$MSE = \frac{SSE}{n - k - 1}$		
Total	SST	$n - 1$			

6.  $r^2 = SSR/SST$ : the coefficient of variation, a measure of the proportion of variation in the data that is explained by the regression model.
7. A test of a significant regression is equivalent to testing the hypothesis that all regression coefficients (except the constant term) are zero.

$H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$ , none of the predictor variables helps to explain any variation in the dependent variable.

A test for a significant multiple linear regression model is based on the ratio of the mean square due to regression and the mean square due to error.

**Hypothesis Test for a Significant Multiple Linear Regression**

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

(None of the predictor variables helps to predict  $y$ .)

$$H_a: \beta_i \neq 0 \text{ for at least one } i$$

(At least one predictor variable helps to predict  $y$ .)

$$\text{TS: } F = \frac{\text{MSR}}{\text{MSE}}$$

$$\text{RR: } F \geq F_{\alpha, k, n-k-1}$$

**Example 12.5.2** In a paper by Lin et al, the affects of daily calcium ( $x_1$ , in mg/Kcal), vitamin A ( $x_2$ , in IU), and potassium ( $x_3$ , in mg) intake on the change in body weight ( $y$ , in kg) for normal weight young women, 18 to 31 years of age were studied (Source: Journal of the American College of Nutrition, Vol. 19. No. 6, 754–760, 2000). Suppose data from 54 randomly selected women were used to produce the following multiple linear regression equation

$$y = 0.30 + 3.90x_1 - 0.000301x_2 + 0.000436x_3.$$

In addition,  $\text{SSR} = 61.239$  and  $\text{SSE} = 189.448$ .

- Complete the summary ANOVA table and conduct an  $F$  test for a significant regression. Use a significance level of 0.05.
- Compute  $r^2$  and interpret this value.

ANOVA summary table for multiple linear regression

Source of variation	Sum of squares	Degrees of freedom	Mean square	$F$	$p$ value
Regression					
Error					
Total					

Example (continued)

Suppose the overall multiple linear regression  $F$  test is significant.

1. There is evidence to suggest that at least one of the independent variables can be used to predict the value of  $Y$ .
2. Hypothesis tests, or confidence intervals, can be used to determine whether  $x_i$  helps to predict the value of  $Y \mid \mathbf{x}$ , equivalently, whether  $\beta_i \neq 0$ .
3. The random variable  $B_i$ : an estimator for  $\beta_i$ .
4.  $S^2 = \text{MSE} = \text{SSE}/(n - k - 1)$ : an estimator for  $\sigma^2$ .
5. If the multiple linear regression assumptions are true,  $S^2$  is an unbiased estimator for  $\sigma^2$ .

#### Hypothesis Test and Confidence Interval Concerning $\beta_i$

$$H_0: \beta_i = \beta_{i0}$$

$$H_a: \beta_i > \beta_{i0}, \quad \beta_i < \beta_{i0}, \quad \text{or} \quad \beta_i \neq \beta_{i0}$$

$$\text{TS: } T = \frac{B_i - \beta_{i0}}{S_{B_i}}$$

$$\text{RR: } T \geq t_{\alpha, n-k-1}, \quad T \leq -t_{\alpha, n-k-1}, \quad \text{or} \quad |T| \geq t_{\alpha/2, n-k-1}$$

A  $100(1 - \alpha)\%$  confidence interval for  $\beta_i$  has as endpoints the values

$$\hat{\beta}_i \pm t_{\alpha/2, n-k-1} s_{B_i}. \quad (12.1)$$

Software packages:

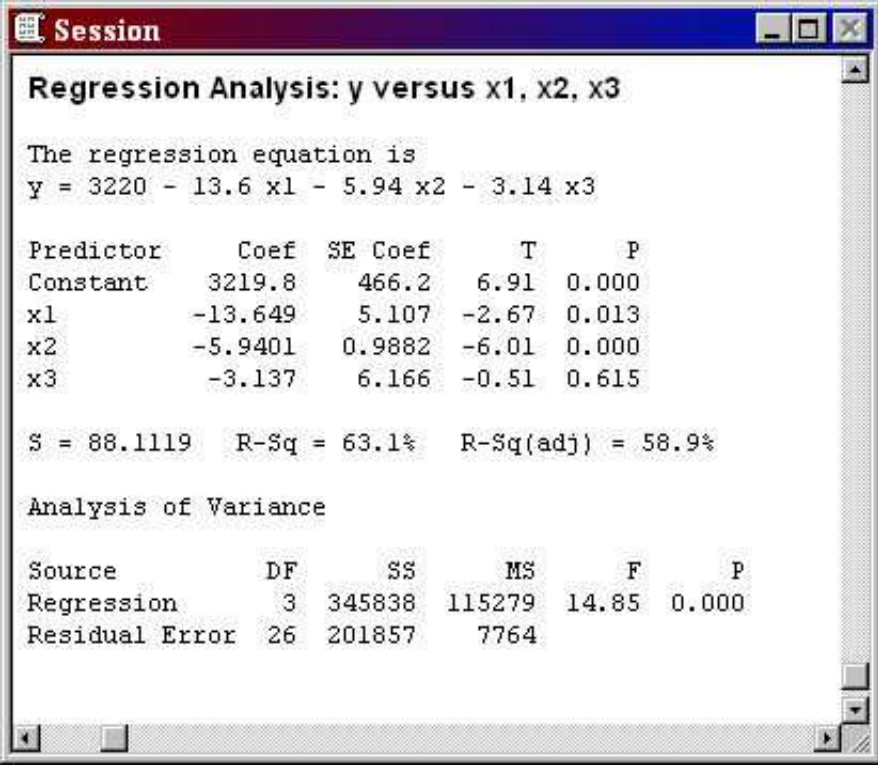
1.  $s_{B_i}$ : an estimate for the standard deviation of  $B_i$ .
2.  $k + 1$  hypothesis tests concerning  $\beta_0, \beta_1, \dots, \beta_k$ .

Null hypothesis in each case is  $H_0: \beta_i = 0$ .

Associated test statistic and the  $p$  value are usually given.

**Example 12.5.3** Research suggests that spore abundance is related to toxic elements in the soil, for example heavy metals. A random sample of locations in Arizona was selected and the spore abundance was measured for each ( $y$ , in spores per kg). In addition, the amount of copper ( $x_1$ , in mg/kg), zinc ( $x_2$ , in mg/kg), and phosphorus ( $x_3$ , in mg/kg) were also measured. Multiple linear regression was used to investigate the affects of these three metals on spore abundance. The resulting Minitab output is shown below.

- Verify that the multiple linear regression is significant at the  $\alpha = 0.01$  level.
- Conduct separate hypothesis tests to determine whether each predictor variable contributes to the overall significant regression. Use  $\alpha = 0.05$  in each test.



**Session**

**Regression Analysis: y versus x1, x2, x3**

The regression equation is  
 $y = 3220 - 13.6 x_1 - 5.94 x_2 - 3.14 x_3$

Predictor	Coef	SE Coef	T	P
Constant	3219.8	466.2	6.91	0.000
x1	-13.649	5.107	-2.67	0.013
x2	-5.9401	0.9882	-6.01	0.000
x3	-3.137	6.166	-0.51	0.615

S = 88.1119    R-Sq = 63.1%    R-Sq(adj) = 58.9%

**Analysis of Variance**

Source	DF	SS	MS	F	P
Regression	3	345838	115279	14.85	0.000
Residual Error	26	201857	7764		

Example (continued)

**Remarks**

1. Test concerning the constant term: if we fail to reject, a model without a constant term might be more appropriate.
2. If the model utility test is significant, then there are  $k \geq 2$  hypothesis tests to consider in order to isolate those variables contributing to the overall effect.

To control the probability of making at least one mistake: Bonferroni technique.

(a) If  $k$  hypothesis tests are conducted, then the significance level in each case is  $\alpha/k$ .

(b) The  $k$  simultaneous  $100(1 - \alpha)\%$  confidence intervals have as endpoints the values  $\hat{\beta}_i \pm t_{\alpha/(2k), n-k-1} S_{B_i}$ .

3. There are many different statistical procedures to select the *best* regression model.

The most reasonable method is to simply keep only those variables in the model that have regression coefficients significantly different from 0. Eliminate the others, and calculate a new, *reduced* model for prediction.

Two useful inferences in multiple linear regression:

1. Estimate of the *mean* value of  $Y$  for  $\mathbf{x} = \mathbf{x}^*$ .
2. Estimate of an *observed* value of  $Y$  for  $\mathbf{x} = \mathbf{x}^*$ .

**Hypothesis Test and Confidence Interval Concerning the Mean Value of  $Y$  for  $\mathbf{x} = \mathbf{x}^*$**

$$H_0: y^* = y_0^*$$

$$H_a: y^* > y_0^*, \quad y^* < y_0^*, \quad \text{or} \quad y^* \neq y_0^*$$

$$\text{TS: } T = \frac{(B_0 + B_1x_1^* + \cdots + B_kx_k^*) - y_0^*}{S_{Y^*}}$$

$$\text{RR: } T \geq t_{\alpha, n-k-1}, \quad \text{or} \quad T \leq -t_{\alpha, n-k-1}, \quad |T| \geq t_{\alpha/2, n-k-1}$$

A  $100(1 - \alpha)\%$  confidence interval for  $\mu_{Y|\mathbf{x}^*}$ , the mean value of  $Y$  for  $\mathbf{x} = \mathbf{x}^*$ , has as endpoints the values

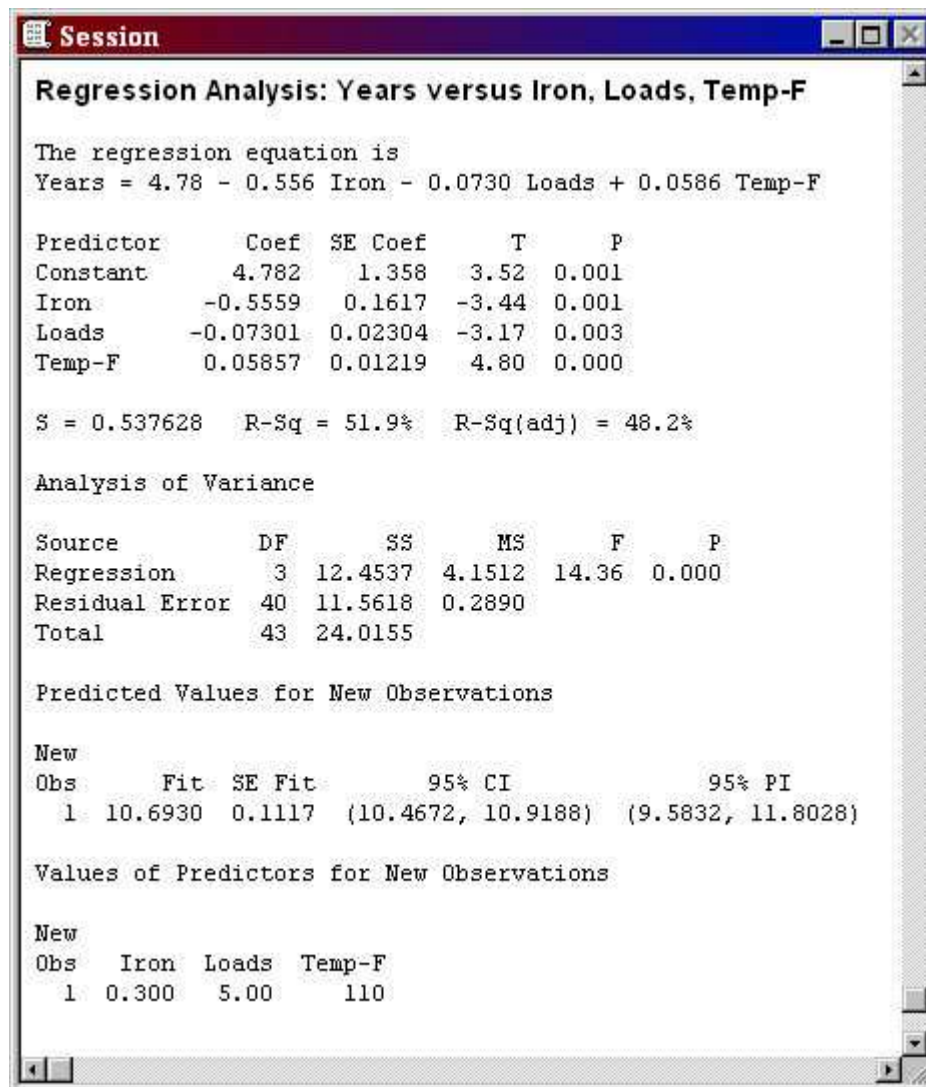
$$(\hat{\beta}_0 + \hat{\beta}_1x_1^* + \cdots + \hat{\beta}_kx_k^*) \pm t_{\alpha/2, n-k-1} S_{Y^*}$$

**Prediction Interval**

A  $100(1 - \alpha)\%$  prediction interval for an observed value of  $Y$  when  $\mathbf{x} = \mathbf{x}^*$  has as endpoints the values

$$(\hat{\beta}_0 + \hat{\beta}_1 x_1^* + \cdots + \hat{\beta}_k x_k^*) \pm t_{\alpha/2, n-k-1} \sqrt{s^2 + s_{Y^*}^2}.$$

**Example 12.5.4** Maytag recently conducted a study to determine some of the factors that affect the lifetime of a clothes washer. A random sample of households replacing washers was selected and the lifetime of each washer being replaced was recorded ( $y$ , in years). In addition, the amount of iron in the water ( $x_1$ , in mg/l), the number of loads per week ( $x_2$ ), and the temperature of the hot water at the point entering the clothes washer ( $x_3$ , in °F) were also measured. The data obtained were used to fit the model  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ . The Minitab output is shown below.





- (a) Construct a 95% confidence interval for the mean number of years a clothes washer lasts when  $x_1 = 0.3$ ,  $x_2 = 5$ , and  $x_3 = 110$ . Use this confidence interval to determine whether there is any evidence to suggest that the mean number of years a clothes washer lasts for these values is less than 11 years?
- (b) Construct a 95% confidence interval for an observed value of the number of years a clothes washer lasts when  $x_1 = 0.3$ ,  $x_2 = 5$ , and  $x_3 = 110$ .

Residual analysis

- $\hat{e}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}), i = 1, 2, \dots, n$

Estimates of the random errors, used to check the regression assumptions.

- Graphical procedures using the residuals to check the model assumptions.
  - Construct a histogram, stem-and-leaf plot, scatter plot and/or normal probability plot of the residuals. These graphs are all used to check the normality assumption.
  - Construct a scatter plot of the residuals versus *each* independent variable.

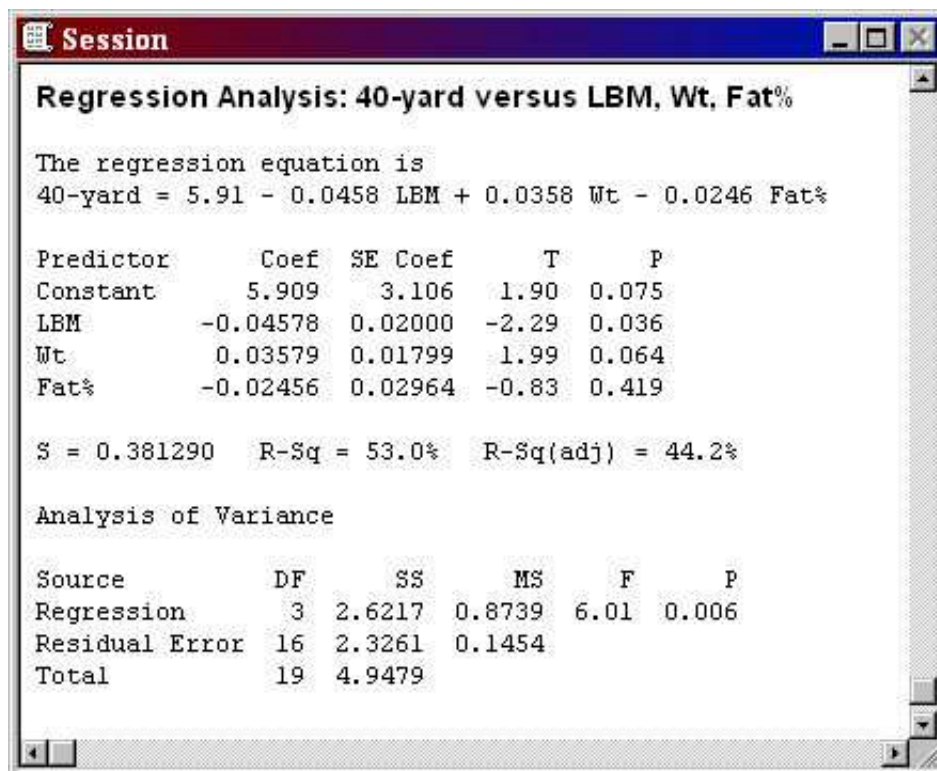
For example, The first scatter plot has the ordered pairs are  $(x_{1i}, \hat{e}_i)$ , the second has the ordered pairs  $(x_{2i}, \hat{e}_i)$ , etc.

If there are no violations in assumptions, each scatter plot should appear as a horizontal band around 0. There should be no recognizable pattern.

**Example 12.5.5** In a study by Swindler (IAHPERD Journal, 1999), variables associated with the 40-yard dash time in college football players were studied. Suppose 20 backs were selected at random. Each 40-yard dash time ( $y$ , in seconds), lean body mass ( $x_1$ , in kg), weight ( $x_2$ , in kg), and percentage of body fat ( $x_3$ ) is given in the table below.

$y_i$	$x_{1i}$	$x_{2i}$	$x_{3i}$
5.01	89.9	102.4	10.3
4.86	89.2	101.7	17.7
5.17	90.1	109.1	18.8
5.92	79.7	104.1	11.1
5.90	84.2	102.1	10.7
4.85	90.3	103.4	12.9
4.34	86.8	96.3	13.8
4.94	94.7	98.4	10.0
5.88	78.6	113.7	12.4
5.75	82.5	103.8	12.8
4.92	97.6	87.1	13.2
5.95	86.3	108.0	14.5
4.78	89.8	101.0	10.8
5.97	87.7	112.3	17.4
4.86	96.8	105.0	12.1
5.59	86.3	102.6	11.0
4.85	90.1	102.2	11.3
4.89	88.2	99.6	15.9
5.19	82.7	101.3	20.2
4.75	91.0	100.8	14.1

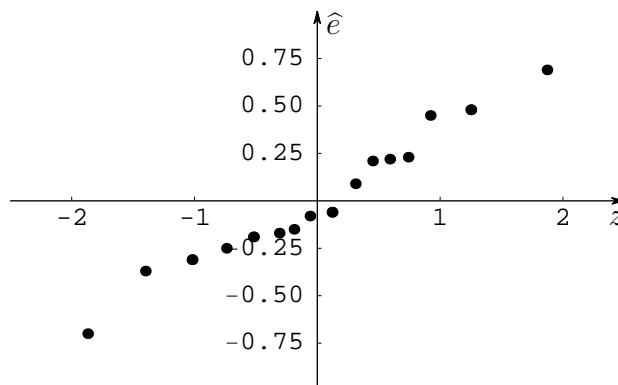
The estimated regression equation is  $y = 5.91 - 0.0458x_1 + 0.0358x_2 - 0.0246x_3$  as shown in the Minitab output. Compute the residuals and construct a normal probability plot of the residuals. Sketch a graph of the residuals versus each predictor variable and discuss any indication of violations in regression assumptions.



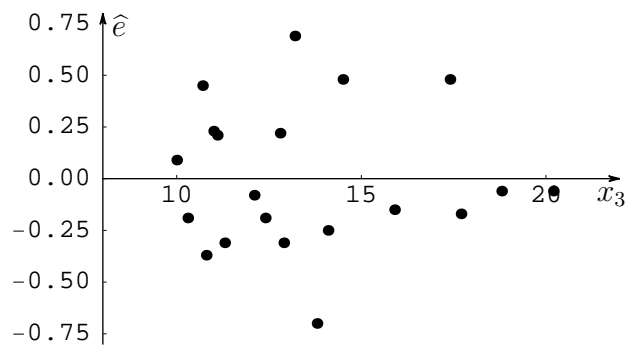
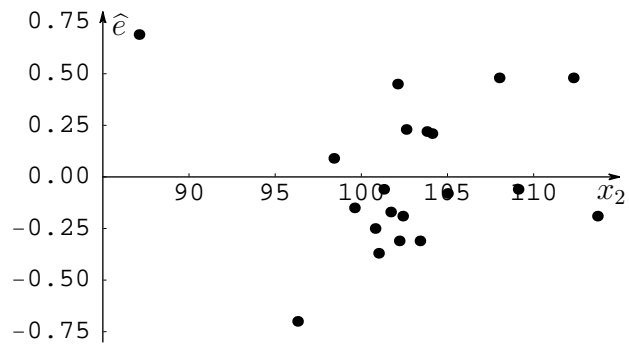
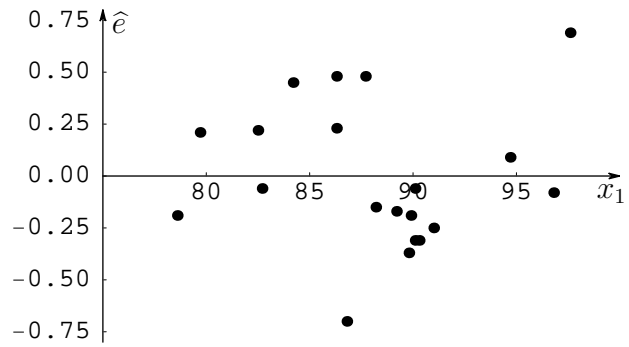
the following table includes each predicted value, residual, and normal score (associated with the residual).

$y_i$	$x_{1i}$	$x_{2i}$	$x_{3i}$	$\hat{y}_i$	$\hat{e}_i$	Normal score
5.01	89.9	102.4	10.3	5.20	-0.19	-0.52
4.86	89.2	101.7	17.7	5.03	-0.17	-0.31
5.17	90.1	109.1	18.8	5.23	-0.06	0.12
5.92	79.7	104.1	11.1	5.71	0.21	0.45
5.90	84.2	102.1	10.7	5.45	0.45	0.92
4.85	90.3	103.4	12.9	5.16	-0.31	-1.02
4.34	86.8	96.3	13.8	5.04	-0.70	-1.87
4.94	94.7	98.4	10.0	4.85	0.09	0.31
5.88	78.6	113.7	12.4	6.07	-0.19	-0.52
5.75	82.5	103.8	12.8	5.53	0.22	0.59
4.92	97.6	87.1	13.2	4.23	0.69	1.87
5.95	86.3	108.0	14.5	5.47	0.48	1.25
4.78	89.8	101.0	10.8	5.15	-0.37	-1.40
5.97	87.7	112.3	17.4	5.49	0.48	1.25
4.86	96.8	105.0	12.1	4.94	-0.08	-0.06
5.59	86.3	102.6	11.0	5.36	0.23	0.74
4.85	90.1	102.2	11.3	5.16	-0.31	-1.02
4.89	88.2	99.6	15.9	5.04	-0.15	-0.19
5.19	82.7	101.3	20.2	5.25	-0.06	0.12
4.75	91.0	100.8	14.1	5.00	-0.25	-0.74

The normal probability plot of the residuals:



Scatter plots of the residuals versus each predictor variable.



Other *linear* models

1. Quadratic model with one predictor:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + E_i$$

2. A more general  $k$ th degree polynomial model with one predictor:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + E_i$$

3. Two predictor variables and an interaction term:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + E_i$$

*Intrinsically* linear models: These models do not appear to be linear, but can be transformed into linear models.

Example: Exponential model  $Y_i = \beta_0 e^{\beta_1 x_i} + E_i$

Some models cannot be *made into* a linear model.

Example: General growth model  $Y_i = \beta_0 + \beta_1 * e^{\beta_2 x_i} + E_i$

## CHAPTER 13

# Categorical Data and Frequency Tables

---

## 13.0 Introduction

1. In this chapter, we'll consider categorical data, univariate and bivariate.

2. Categorical data: non-numerical observations that fall into categories.

Example: What beverage do you drink in the morning with breakfast?

Possible responses: orange juice, milk, coffee, tea, etc.

3. Bivariate categorical data: two non-numerical observations on each individual or object.

Example: a random sample of people is asked to name their favorite dipping chip and their favorite dip flavor.

4. Natural summary measures: frequency and relative frequency.

---

## 13.1 Univariate Categorical Data, Goodness-of-Fit Tests

1. Categorical data are often displayed in a frequency distribution.

Here, focus on the number of observations in each category.

Example: an insurance company selected a random sample of customers with extra coverage for valuables.

There were four possible responses.

Total for each response given in a one-way frequency table.

Valuable	Paintings	Sculptures	Jewelry	Religious items
Frequency	32	43	55	16

- Hypothesis test: designed to compare a set of *hypothesized* proportions with a set of *true* proportions, to check the *goodness of fit*.
- Example: Is there any evidence that the true proportions of extra insurance policies are different from 0.25, 0.25, 0.40, and 0.10?
- Notation:

	True proportion	Hypothesized proportion
Category 1	$p_1$	$p_{10}$
Category 2	$p_2$	$p_{20}$
$\vdots$	$\vdots$	$\vdots$
Category $i$	$p_i$	$p_{i0}$
$\vdots$	$\vdots$	$\vdots$
Category $k$	$p_k$	$p_{k0}$

$$p_{10} + p_{20} + \cdots + p_{k0} = 1.$$

- Goodness-of-fit test: Is there any evidence that the true population proportions differ from the hypothesized population proportions.
- The null hypothesis and the alternative hypothesis are stated in terms of the true and hypothesized category proportions.

$H_0$ :  $p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$   
 (Each true category proportion is equal to a specified hypothesized value.)

$H_a$ :  $p_i \neq p_{i0}$  for at least one  $i$ .  
 (There is at least one true category proportion that is not equal to the corresponding specified hypothesized value.)

The test statistic:

- Consider a random sample of size  $n$ .

Let  $n_i$  = the number of observations in each category ( $i = 1, 2, \dots, k$ ).

- Observed cell counts:  $n_i, i = 1, 2, \dots, k$ .

Expected cell counts:  $e_i = np_{i0}, i = 1, 2, \dots, k$ .



3. The test statistic: a measure of how far away the observed cell counts are from the expected cell counts.

If  $H_0$  is true,

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}} = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}$$

has approximately a chi-square distribution with  $k - 1$  degrees of freedom.

Approximation good if  $e_i = np_{i0} \geq 5$  for all  $i$ .

**Goodness-of-Fit Test**

Let  $n_i$  be the number of observations falling into the  $i$ th category ( $i = 1, 2, \dots, k$ ), and let  $n = n_1 + n_2 + \dots + n_k$ . A hypothesis test about the true category population proportions with significance level  $\alpha$  has the form:

$H_0: p_1 = p_{10}, p_2 = p_{20}, \dots, p_k = p_{k0}$

$H_a: p_i \neq p_{i0}$  for at least one  $i$ .

TS:  $X^2 = \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}$  where  $e_i = np_{i0}$

RR:  $X^2 \geq \chi_{\alpha, k-1}^2$

**Example 13.1.1** A random sample of American adults was obtained, and each was asked to name the most important issue facing the country. Use the following one-way frequency table to test the hypothesis that the five possible responses are of equal importance. Use  $\alpha = 0.05$ .

Issue	Terrorism	Education	Health care	Economy	Taxes
Frequency	92	80	98	79	83

Example (continued)

Cell	Category	Observed cell count	Expected cell count
1	Terrorism	92	
2	Education	80	
3	Health care	98	
4	Economy	79	
5	Taxes	83	

**Example 13.1.2** A study was conducted to determine the prevailing wind directions in order to position the major runways for a new airport. A random sample of days was selected, and the prevailing wind direction was recorded on each day. The data are given in the following table.

Direction	North	South	East	West	Northeast	Southeast	Southwest	Northwest
Frequency	32	15	52	23	16	14	10	38

Is there evidence to suggest that any of the true cell proportions differ from the following historical proportions: 0.20, 0.05, 0.20, 0.10, 0.10, 0.10, 0.05, 0.20. Use  $\alpha = 0.01$ . Find bounds on the  $p$  value associated with this test.

Cell	Category	Observed cell count	Expected cell count
1	North	32	
2	South	15	
3	East	52	
4	West	23	
5	Northeast	16	
6	Southeast	14	
7	Southwest	10	
8	Northwest	38	

Example (continued)

**Example 13.1.3** A random sample of adults was obtained, and each was asked which of the top five 77 national historic sites they would most like to visit. The data are given in the following table.

Site	Fort Point	San Juan National Historic Site	Martin Luther King Jr. birthplace and church	Salem Maritime	Fort Vancouver
Frequency	171	91	105	120	75

Is there evidence to suggest that any of the true cell proportions differ from the following historical proportions: 0.3, 0.2, 0.2, 0.2, 0.1. Use  $\alpha = 0.05$ . Find bounds on the  $p$  value associated with this test.

Cell	Category	Observed cell count	Expected cell count
1	Fort Point	171	
2	SJ Site	91	
3	MLK Jr.	105	
4	Salem	120	
5	Vancouver	75	

Example (continued)

## 13.2 Bivariate Categorical Data, Tests for Homogeneity and Independence

Two common types of bivariate categorical data:

1. Random samples are obtained from two or more populations, and each individual is classified by values of a categorical variable.

Test for homogeneity applies in this case.

2. Suppose there are two categorical variables of interest. In a (single) random sample, a value of each variable is recorded for each individual.

Test for independence applies in this case.

Data from two or more populations

**Example 13.2.1** A large company has three locations in the United States: northeast, South, and West. Every employee of the company selects one of four retirement options: ING, Galic, Schwab, or Spider. A random sample of employees from each site was selected and the retirement option of each employee was recorded.

Populations: company locations.

Categorical variable: retirement option.

Natural summary measure: number of observations in each category combination; two-way frequency table.

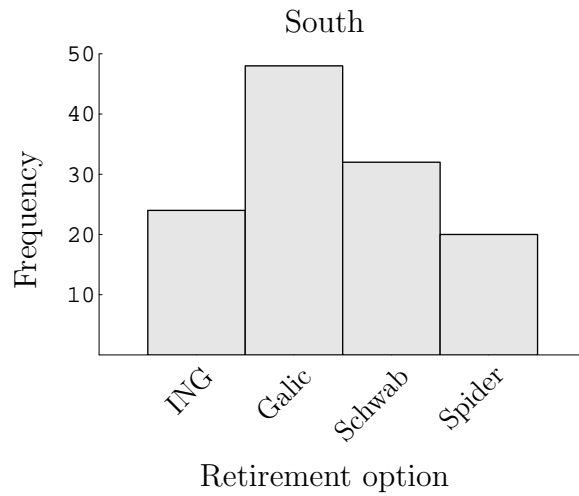
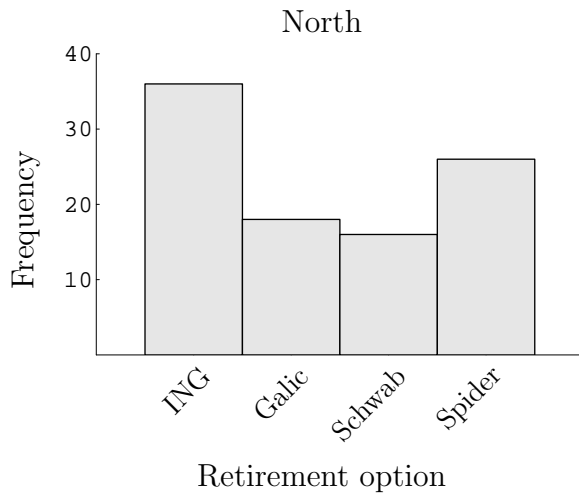
		Retirement option			
		ING	Galic	Schwab	Spider
Location	North	36	18	16	26
	South	24	48	32	20
	West	35	28	32	30

Rows: locations;    Columns: retirement options.

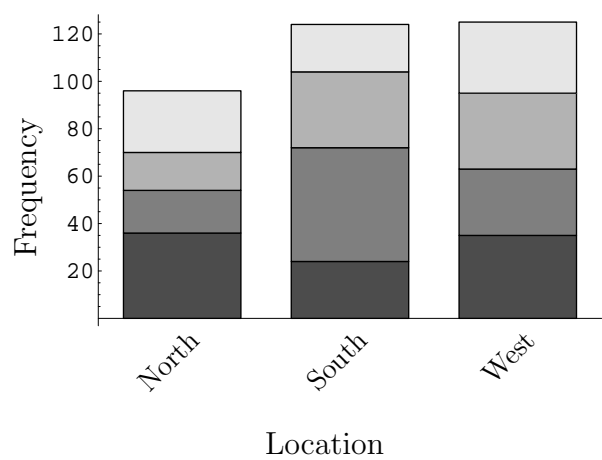
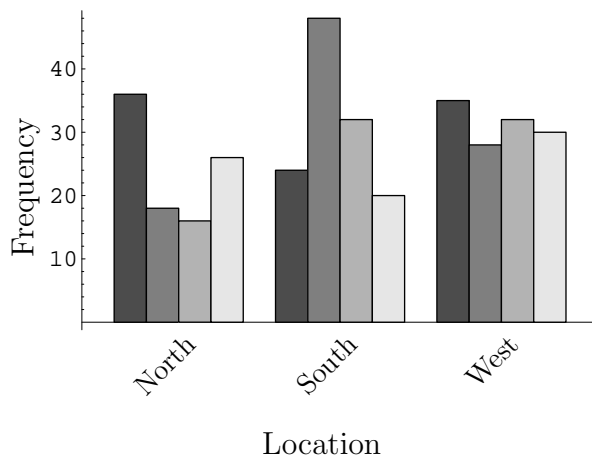
Each cell contains a frequency, or observed count.

Typical graphical summaries:

Consider a single company location (population):



Side-by-side or stacked bar chart.



Test for homogeneity of populations:

Are all of the true category proportions the same for each population.

Is the proportion of employees selecting each retirement option the same at each company location?



The test procedure:

Suppose there are  $I$  rows and  $J$  columns in a two-way frequency table.

$n_{ij}$  = the observed cell count, or frequency, in the  $(ij)$  cell (the intersection of the  $i$ th row and the  $j$ th column)

$n_{i.} = \sum_{j=1}^J n_{ij}$   
 = the  $i$ th row total, the sum of the cell counts, or observed frequencies, in the  $i$ th row.

$n_{.j} = \sum_{i=1}^I n_{ij}$   
 = the  $j$ th column total, the sum of the cell counts, or observed frequencies, in the  $j$ th column.

$n = \sum_{i=1}^I \sum_{j=1}^J n_{ij}$   
 = the grand total, the total of all cell counts, or observed frequencies.

Two-way frequency table with notation:

		Category						Row total
		1	2	...	$j$	...	$J$	
Population	1	$n_{11}$	$n_{12}$	...	$n_{1j}$	...	$n_{1J}$	$n_{1.}$
	2	$n_{21}$	$n_{22}$	...	$n_{2j}$	...	$n_{2J}$	$n_{2.}$
	:	:	:	:	:	:	:	:
	$i$	$n_{i1}$	$n_{i2}$	...	$n_{ij}$	...	$n_{iJ}$	$n_{i.}$
	:	:	:	:	:	:	:	:
	I	$n_{I1}$	$n_{I2}$	...	$n_{Ij}$	...	$n_{IJ}$	$n_{I.}$
Column total		$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.J}$	$n$

Expected frequencies:  $e_{ij} = \frac{(\text{ith row total})(\text{jth column total})}{\text{grand total}} = \frac{n_{i.} \times n_{.j}}{n}$

Test statistic: a measure of how far away the observed cell counts are from the expected cell counts.

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed cell count} - \text{expected cell count})^2}{\text{expected cell count}} = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

If there is no difference in category proportions among populations, this random variable has approximately a chi-square distribution with  $(I - 1)(J - 1)$  degrees of freedom.

Approximation good if  $e_{ij} \geq 5$  for all  $i$  and  $j$ .

### Test for Homogeneity of Populations

In an  $I \times J$  two-way frequency table, let  $n_{ij}$  be the observed count in the  $(ij)$  cell and let  $e_{ij}$  be the expected count in the  $(ij)$  cell. A hypothesis test for homogeneity of populations with significance level  $\alpha$  has the form:

$H_0$ : The true category proportions are the same for all populations (homogeneity of populations).

$H_a$ : The true category proportions are not the same for all populations.

$$\text{TS: } X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

$$\text{where } e_{ij} = \frac{(\text{ith row total})(\text{jth column total})}{\text{grand total}} = \frac{n_{i.} \times n_{.j}}{n}$$

$$\text{RR: } X^2 \geq \chi_{\alpha, (I-1)(J-1)}^2$$

### Remark

This procedure is called a test of homogeneity.

However, this property is stated in the null hypothesis, so that we are really testing for evidence of inhomogeneity.

We cannot prove homogeneity, we can only test for evidence of inhomogeneity.

**Example 13.2.2** Committees in five cities are preparing bids for the 2012 Summer Olympics. In order to help assess each bid, the International Olympic Committee obtained a random sample of adults in each city and asked each person their opinion on hosting the Summer Olympics. The observed frequencies are given in the following table.

		Opinion		
		In favor	Against	No opinion
City	London	171	282	266
	Madrid	195	237	208
	Moscow	202	247	298
	New York	191	206	195
	Paris	184	199	198

Is there evidence to suggest the true category proportions of opinions are different for cities? Use  $\alpha = 0.05$ .

		Birthday card type			Row total
		In Favor	Against	No Opinion	
City	London	171	282	266	
	Madrid	195	237	208	
	Moscow	202	247	298	
	New York	191	206	195	
	Paris	184	199	198	
Column total					

Example (continued)

**Example 13.2.3** The federal government conducted a survey concerning the creation of new jobs in various parts of the country. A random sample of new jobs was selected from the North, South, Midwest, and West, and each was classified according to economic sector. The observed frequencies are given in the following table.

		Job type			
		Manufacturing	Construction	Service	Retail
Region	North	368	395	470	300
	South	342	340	464	280
	Midwest	398	320	442	327
	West	376	410	457	352

Is there evidence to suggest the true category proportions of job type are different for regions? Use  $\alpha = 0.05$ . Find bounds on the  $p$  value associated with this test.

		Job type				Row total
		Manufacturing	Construction	Service	Retail	
Region	North	368	395	470	300	
	South	342	340	464	280	
	Midwest	398	320	442	327	
	West	376	410	457	352	
Column total						

Example (continued)

Single sample, two categorical variables recorded.

Examples:

1. A random sample of people who make purchases in a jewelry store was obtained. Each person was classified according to the type of jewelry purchased (ring, watch, pendant, etc.) and by the recipient (self, spouse, other relative, etc.).
2. A random sample of loans made by a bank was obtained. Each loan was classified according to the type (home, car, personal) and payment schedule (monthly, biweekly, weekly).

Question: Are the two (categorical) variables independent?

1. Same notation:  $I \times J$  two-way table.
2. Test based on observed and expected counts.
3. Suppose the two categorical variables are independent.

$$P[\text{an individual falls into the } (ij) \text{ cell}] = \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n}$$

Since there is a total of  $n$  individuals, the expected count, or frequency, in the  $(ij)$  cell is

$$e_{ij} = n \cdot \left( \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} \right) = \frac{n_{i.} \times n_{.j}}{n} = \frac{(\text{ith row total})(\text{jth column total})}{\text{grand total}}$$

### Test for Independence of Two Categorical Variables

In a random sample of  $n$  individuals, suppose the values of two categorical variables are recorded. In the resulting  $I \times J$  two-way frequency table, let  $n_{ij}$  be the observed count in the  $(ij)$  cell and let  $e_{ij}$  be the expected count in the  $(ij)$  cell. A hypothesis test for independence of the two categorical variables with significance level  $\alpha$  has the form:

$H_0$ : The two variables are independent.

$H_a$ : The two variables are dependent.

$$\text{TS: } X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

$$\text{where } e_{ij} = \frac{(\text{ith row total})(\text{jth column total})}{\text{grand total}} = \frac{n_{i.} \times n_{.j}}{n}$$

$$\text{RR: } X^2 \geq \chi_{\alpha, (I-1)(J-1)}^2$$

**Example 13.2.4** A random sample of estuaries from around the world was obtained, and each was classified by geologic features and water circulation. The observed frequencies are given in the following table.

		Geologic features			
		Coastal plain	Tectonic	Bar-built	Fjord
Water circulation	Salt wedge	24	29	26	10
	Partially mixed	27	11	20	24
	Well mixed	20	12	27	20

Is there any evidence to suggest that geologic features and water circulation dependent? Use  $\alpha = 0.01$ . Find bounds on the  $p$  value associated with this test.

		Geologic features				Row total
		Coastal plain	Tectonic	Bar-built	Fjord	
Water circulation	Salt wedge	24	29	26	10	
	Partially mixed	27	11	20	24	
	Well mixed	20	12	27	20	
	Column total					



Example (continued)

**Example 13.2.5** A random sample of card games was obtained, and each was classified by objective and origin (or where it is played). The observed frequencies are given in the following table.

		Objective			
		Capturing cards	Shedding or accumulating cards	Forming combinations of cards	Comparing cards
Origin	Europe	44	39	27	26
	Asia	42	45	31	28
	Africa	52	36	26	17
	North America	44	51	21	22
	Caribbean	31	25	40	26
	South America	17	30	12	13

Is there any evidence to suggest that objective and origin are dependent? Use  $\alpha = 0.05$ .

Example (continued)

	Objective				Row total
	Capturing cards	Shedding or accumulating cards	Forming combinations of cards	Comparing cards	
Europe	44	39	27	26	
Asia	42	45	31	28	
Africa	52	36	26	17	
North America	44	51	21	22	
Caribbean	31	25	40	26	
South America	17	30	12	13	
Column total					

Example (continued)

## CHAPTER 14

# Nonparametric Statistics

---

### 14.0 Introduction

1. Recall: Hypothesis tests based on a set of assumptions.

If any assumptions are violated, the conclusions may be invalid.

2. **Parametric methods:** statistical procedures based on a normality assumption.

3. **Nonparametric or distribution-free procedures:** techniques that require very few assumptions.

Nonparametric methods

1. Very few assumptions necessary in order to be valid.

2. Test statistics tend to be more intuitive and easier to apply.

3. Appropriate for data arranged according to rank.

4. Some nonparametric tests may be used to analyze qualitative data.

5. Disadvantages:

These tests usually do not utilize all of the information in the sample.

Therefore, there is a greater chance of making an error.

If both parametric and nonparametric can be used: use parametric procedure.

## 14.1 The Sign Test

Consider a random sample from a continuous (non-normal) distribution and a test concerning the population median,  $H_0: \tilde{\mu} = \tilde{\mu}_0$ .

If  $H_0$  is true, then approximately half of the observations should lie above  $\tilde{\mu}_0$ , and the other half should fall below  $\tilde{\mu}_0$ .

Replace each observation above  $\tilde{\mu}_0$  with a plus sign and each observation below  $\tilde{\mu}_0$  with a minus sign.

If  $H_0$  is true, then the number of plus signs and the number of minus signs should be about the same.

Any observations equal to  $\tilde{\mu}_0$  are excluded from the analysis.

The test statistic,  $X$ , is a count of the number of plus signs.

If the null hypothesis is true, then the probability of a plus sign is  $1/2$ .

If  $H_0$  is true,  $X$  has a binomial distribution with number of trials equal to  $n$  and  $p = 1/2$ :

We should reject the null hypothesis for very large or very small values of  $X$ .

### The Sign Test Concerning a Population Median

Suppose a random sample is obtained from a continuous distribution. A hypothesis test concerning a population median  $\tilde{\mu}$  with significance level  $\alpha$  has the form:

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

$$H_a: \tilde{\mu} > \tilde{\mu}_0, \quad \tilde{\mu} < \tilde{\mu}_0, \quad \text{or} \quad \tilde{\mu} \neq \tilde{\mu}_0$$

TS:  $X =$  the number of observations greater than  $\tilde{\mu}_0$

$$\text{RR: } X \geq c_1, \quad X \leq c_2, \quad X \geq c \quad \text{or} \quad X \leq n - c$$

The critical values  $c_1$ ,  $c_2$ , and  $c$  are obtained from Table 1 (Binomial Distribution Cumulative Probabilities) with parameters  $n$  and  $p = 0.5$  to yield a significance level of approximately  $\alpha$ . That is, so that  $P(X \geq c_1) \leq \alpha$ ,  $P(X \leq c_2) \leq \alpha$ , and  $P(X \geq c) \leq \alpha/2$ .

Observations equal to  $\tilde{\mu}_0$  are excluded from the analysis, and the sample size is reduced accordingly.

**Remarks**

1. If the underlying distribution is symmetric:  $\mu = \tilde{\mu}$ , the sign test can be used to test a hypothesis about  $\mu$ .
2. Discard any observations equal to  $\tilde{\mu}_0$ , count the number of observations greater than  $\tilde{\mu}_0$ .
3. Usually cannot find critical values to yield an exact level- $\alpha$  test. Use critical values such that the significance level is as close to  $\alpha$  as possible but not greater than  $\alpha$ .

**Example 14.1.1** Aluminum doors in certain commercial buildings have steel tension rods that run the full width of the top and bottom rails. A random sample of rods was obtained, and the diameter of each was carefully measured (in inches). The data are given in the following table.

0.390	0.350	0.392	0.321	0.374	0.360	0.355	0.382
0.421	0.357	0.378	0.372	0.358	0.373	0.372	

The underlying distribution is not normal. Is there any evidence to suggest that the median diameter of these steel tension rods is less than 0.375? Use a significance level of  $\alpha = 0.05$ .

**Example 14.1.2** A residential roof de-icing system is designed to produce 28 watts per square foot. A random sample of these units was obtained, and the power output per square foot of each was measured. The data are given in the following table.

28.1	28.5	28.2	28.5	28.5	27.8	28.1	28.3	27.9
28.6	28.3	28.5	28.1	28.1	28.3	28.4	28.1	28.3
27.5	28.0	28.0	27.9	28.4	28.7	27.8	28.5	28.1

Previous research suggests the underlying distribution is not normal. Is there any evidence to suggest that the median power per square foot produced is different from 28? Use  $\alpha = 0.01$ .



Paired data:

1. Compute each pairwise difference, consider the sign of the difference.
2. Positive differences: plus sign. Negative differences: negative sign.
3. Use these signs and the previous procedure to test  $H_0: \tilde{\mu}_1 = \tilde{\mu}_2$ .

### The Sign Test to Compare Two Medians

Suppose there are  $n$  independent pairs of observations such that the population of first observations is continuous and the population of second observations is also continuous. A hypothesis test concerning the two population medians in terms of the difference,  $\tilde{\mu}_D = \tilde{\mu}_1 - \tilde{\mu}_2$ , with significance level  $\alpha$  has the form:

$$H_0: \tilde{\mu}_D = \Delta_0$$

$$H_a: \tilde{\mu}_D > \Delta_0, \quad \tilde{\mu}_D < \Delta_0, \quad \text{or} \quad \tilde{\mu}_D \neq \Delta_0$$

TS:  $X =$  the number of pairwise differences greater than  $\Delta_0$

$$\text{RR: } X \geq c_1, \quad X \leq c_2, \quad X \geq c \text{ or } x \leq n - c$$

The critical values  $c_1$ ,  $c_2$ , and  $c$  are obtained from Table 1 (Binomial Distribution Cumulative Probabilities) with parameters  $n$  and  $p = 0.5$  to yield a significance level of approximately  $\alpha$ . That is, so that  $P(X \geq c_1) \leq \alpha$ ,  $P(X \leq c_2) \leq \alpha$ , and  $P(X \geq c) \leq \alpha/2$ .

Differences equal to  $\Delta_0$  are excluded from the analysis, and the sample size is reduced accordingly.

**Example 14.1.3** In order to study the possibility of global warming, a random sample of glacier sites in Greenland was selected. The ice sheet thickness was measured (in m) at each site in 2000 and again in 2005. The data are given in the following table.

Site	1	2	3	4	5	6	7	8
2000	566	471	622	559	680	599	645	605
2005	747	426	620	579	518	525	603	655
Site	9	10	11	12	13	14	15	
2000	630	631	599	579	548	631	550	
2005	518	530	548	539	609	467	660	

The underlying distributions are not assumed to be normal. Is there any evidence to suggest the median ice sheet thickness is less in 2005 than in 2000? Use  $\alpha = 0.05$ .

**Example 14.1.4** A new type of asphalt has been developed to reduce traffic noise. A random sample of highway locations was obtained, and the noise level was measured (in dB) at each site. The new asphalt was installed and the noise level was measured again at each site. The data are given in the following table.

Location	1	2	3	4	5	6	7	8	9
Old	59.4	68.1	63.0	66.1	73.9	65.2	68.3	75.5	79.8
New	65.6	62.9	62.2	66.4	65.6	64.5	59.5	66.6	67.7
Location	10	11	12	13	14	15	16	17	18
Old	67.8	74.8	68.2	62.5	68.9	77.7	70.6	75.2	73.9
New	62.6	73.8	69.2	58.3	64.9	62.2	66.2	64.7	69.9

The underlying distributions are not assumed to be normal. Is there any evidence to suggest that the median noise level is reduced due to the new asphalt? Use  $\alpha = 0.05$ . Find the  $p$  value associated with this test.

---

## 14.2 The Signed-Rank Test

1. Another test concerning a population median or to compare two population medians.
2. Test statistic: uses the magnitude of the relevant differences and ranks.

Consider a random sample of size  $n$  from a continuous, symmetric distribution, and  $H_0: \tilde{\mu} = \tilde{\mu}_0$ .

1. Compute the differences  $x_1 - \tilde{\mu}_0, x_2 - \tilde{\mu}_0, \dots, x_n - \tilde{\mu}_0$ .
2. Consider the magnitude, or absolute value, of each difference;  
compute  $|x_1 - \tilde{\mu}_0|, |x_2 - \tilde{\mu}_0|, \dots, |x_n - \tilde{\mu}_0|$ .
3. Place the absolute values in increasing order, and assign a rank to each from smallest (rank 1) to largest (rank  $n$ ).
4. Equal absolute values are assigned the mean rank of their positions in the ordered list.
5. Add the ranks associated with the positive differences.

If  $H_0$  is true:

Approximately half of the observations should be above the median and approximately half below the median (symmetry).

The sum of the ranks associated with the positive differences should be approximately equal to the sum of the ranks associated with the negative differences.

If  $H_0$  is not true:

The sum of the ranks associated with the positive differences should be very large or very small.

This suggests that the population median is different from  $\tilde{\mu}_0$ .

### The Wilcoxon Signed-Rank Test

Suppose a random sample is obtained from a continuous, symmetric distribution. A hypothesis test concerning a population median  $\tilde{\mu}$  with significance level  $\alpha$  has the form:

$$H_0: \tilde{\mu} = \tilde{\mu}_0$$

$$H_a: \tilde{\mu} > \tilde{\mu}_0, \quad \tilde{\mu} < \tilde{\mu}_0, \quad \text{or} \quad \tilde{\mu} \neq \tilde{\mu}_0$$

Rank the absolute differences  $|x_1 - \tilde{\mu}_0|, |x_2 - \tilde{\mu}_0|, \dots, |x_n - \tilde{\mu}_0|$ . Equal absolute values are assigned the mean rank for their positions.

TS:  $T_+$  = the sum of the ranks corresponding to the positive differences  $x_i - \tilde{\mu}_0$ .

$$\text{RR: } T_+ \geq c_1, \quad T_+ \leq c_2, \quad T_+ \geq c \quad \text{or} \quad T_+ \leq n(n+1) - c$$

The critical values  $c_1, c_2$ , and  $c$  are obtained from Table 9 such that  $P(T_+ \geq c_1) \approx \alpha$ ,  $P(T_+ \leq c_2) \approx \alpha$ , and  $P(T_+ \geq c) \approx \alpha/2$ .

Differences equal to 0 ( $x_i - \tilde{\mu}_0 = 0$ ) are excluded from the analysis, and the sample size is reduced accordingly.

*The Normal Approximation:* As  $n$  increases ( $n \geq 20$ ), the statistic  $T_+$  approaches a normal distribution with

$$\mu_{T_+} = \frac{n(n+1)}{4} \quad \text{and} \quad \sigma_{T_+}^2 = \frac{n(n+1)(2n+1)}{24}.$$

Therefore, the random variable  $Z = \frac{T_+ - \mu_{T_+}}{\sigma_{T_+}}$  has approximately a standard normal distribution. In this case ( $n \geq 20$ ),  $Z$  is the test statistic and the rejection region is

$$\text{RR: } Z \geq z_\alpha, \quad Z \leq -z_\alpha, \quad \text{or} \quad |Z| \geq z_{\alpha/2}$$

### Remark

Since we assume the underlying population is symmetric, the median is equal to the mean.

Therefore, the test procedure above can be used to test a hypothesis concerning a population mean (when the underlying population is not normal but is symmetric).

**Example 14.2.1** Moving-Pods are a relatively new method for moving personal and business possessions. A customer has a Pod delivered to a site, packs it at their leisure, and the container is then picked up and can be moved coast to coast. A random sample of packed 12-foot Pods was obtained, and the weight (in pounds) of each was measured. The data are given in the following table.

---

948 928 961 932 884 893 976 934 866 921 988 884 895 949 936

---

Assume that the underlying distribution of the weights of packed 12-foot Pods is continuous and symmetric. Use a signed-rank test to determine whether there is any evidence that the median weight is greater than 900 pounds. Use  $\alpha = 0.05$ .

Observation	Difference	Absolute difference	Rank	Signed rank
948				
928				
961				
932				
884				
893				
976				
934				
866				
921				
988				
884				
895				
949				
936				

---

Example(continued)

**Example 14.2.2** NCAA regulations require the goal posts on a water polo goal to be 3 meters apart. A random sample of goals from division I water polo pools was obtained, and the distance between goal posts was carefully measured. The data are given in the following table.

---

3.09	2.95	2.98	3.10	2.89	2.99	3.12	2.96	2.87	2.91	3.32	3.23	2.96
3.04	3.16	3.31	2.86	3.07	2.98	3.20	3.01	2.89	3.11	2.84	2.96	

---

The underlying distance distribution is assumed to be continuous and symmetric. Use a signed-rank test to determine whether there is any evidence that the median distance between goal posts is different from 3 meters. Use  $\alpha = 0.01$  and find the  $p$  value associated with this test.



Example (continued)

Observation	Difference	Absolute difference	Rank	Signed rank
3.09				
2.95				
2.98				
3.10				
2.89				
2.99				
3.12				
2.96				
2.87				
2.91				
3.32				
3.23				
2.96				
3.04				
3.16				
3.31				
2.86				
3.07				
2.98				
3.20				
3.01				
2.89				
3.11				
2.84				
2.96				

Paired data:  $H_0: \tilde{\mu}_1 - \tilde{\mu}_2 = \tilde{\mu}_D = \Delta_0$

1. Assume the distribution of pairwise differences is continuous and symmetric.
2. Each pair of values is independent of all other pairs.
3. Apply the one-sample procedure to the pairwise differences.
  - (a) Subtract  $\Delta_0$  from each pairwise difference: compute  $d_i - \Delta_0$  ( $i = 1, 2, \dots, n$ ).
  - (b) Rank the absolute differences,  $|d_i - \Delta_0|$  ( $i = 1, 2, \dots, n$ ).
  - (c) Find the sum of the ranks associated with the positive differences,  $t_+$ .

**Example 14.2.3** Tight hamstrings can cause lower back pain. A random sample of adult men suffering from lower back pain was obtained. An initial hamstring evaluation was conducted using the following stretching exercise. The patient lies on his back and elevates the right leg as far as possible. The angle between the leg and the floor is measured (in degrees). Following the initial evaluation, each patient participated in a six-week program of stretching exercises. At the end of the six weeks, each patient was evaluated again. The data are given in the following table.

Before	39	50	36	56	52	62	49	67	58	46
After	42	51	42	66	54	63	54	60	59	49

Assume the underlying distribution of the pairwise differences is continuous and symmetric. Use the Wilcoxon signed-rank test to determine whether there is any evidence the median angle increased following the stretching exercises. Use  $\alpha = 0.05$ .

Example (continued)

Before	After	Pairwise Difference	Absolute difference	Rank	Signed rank
39	42				
50	51				
36	42				
56	66				
52	54				
62	63				
49	54				
67	60				
58	59				
46	49				

## 14.3 The Rank-Sum Test

1. Used to compare two population medians.
2. Assumptions:
  - (a) Independent random samples from continuous, non-normal distributions.
  - (b) First sample size =  $n_1$ , second sample size =  $n_2$ , and  $n_1 + 1 \leq n_2$ .
3. Rank the combined  $n_1 + n_2$  observations.
4. Test statistic: the sum the ranks associated with the first sample.

### The Wilcoxon Rank-Sum Test

Suppose two independent random samples of sizes  $n_1$  and  $n_2$  ( $n_1 \leq n_2$ ) are obtained from continuous distributions. A hypothesis test concerning the two population medians with significance level  $\alpha$  has the form:

$$H_0: \tilde{\mu}_1 - \tilde{\mu}_2 = \Delta_0.$$

$$H_a: \tilde{\mu}_1 - \tilde{\mu}_2 > \Delta_0, \quad \tilde{\mu}_1 - \tilde{\mu}_2 < \Delta_0, \quad \text{or} \quad \tilde{\mu}_1 - \tilde{\mu}_2 \neq \Delta_0.$$

Subtract  $\Delta_0$  from each observation in the first sample. Combine these differences and the observations in the second sample, and rank all of these values. Equal values are assigned the mean rank for their positions.

TS:  $W$  = the sum of the ranks corresponding to the differences.

$$\text{RR: } W \geq c_1, \quad W \leq c_2, \quad W \geq c \quad \text{or} \quad W \leq n_1(n_1 + n_2 + 1) - c$$

The critical values  $c_1$ ,  $c_2$ , and  $c$  are obtained from Table 10 such that  $P(W \geq c_1) \approx \alpha$ ,  $P(W \leq c_2) \approx \alpha$ , and  $P(W \geq c) \approx \alpha/2$ .

*The Normal Approximation:* As  $n_1$  and  $n_2$  increase, the statistic  $W$  approaches a normal distribution with

$$\mu_W = \frac{n_1(n_1 + n_2 + 1)}{2} \quad \text{and} \quad \sigma_W^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

Therefore, the random variable  $Z = \frac{W - \mu_W}{\sigma_W}$  has approximately a standard normal distribution. The normal approximation is good when both  $n_1$  and  $n_2$  are greater than 8. In this case,  $Z$  is the test statistic and the rejection region is

$$\text{RR: } Z \geq z_{\alpha}, \quad Z \leq -z_{\alpha}, \quad \text{or} \quad |Z| \geq z_{\alpha/2}$$

**Remarks**

1.  $\Delta_0$  can be any hypothesized difference. Usually,  $\Delta_0 = 0$ , the two population medians are equal.
2. If both underlying populations are symmetric, this test can be used to compare population means.

If the underlying populations are also normal, consider the two-sample  $t$  test.

3. Sometimes, a slightly different test statistic is used.

The Mann–Whitney  $U$  statistic is a function of  $W$  and also approaches a normal distribution as  $n_1$  and  $n_2$  increase.

**Example 14.3.1** The percentage of light transmission can be used to determine the efficiency of glass for solar energy applications. Independent random samples of glass, 3 mm thick, designed for residential solariums were obtained, and the light transmission percentage was measured for each. The data are given in the following table.

Corning	88	88	88	90	91	95		
Erie	91	95	92	94	92	91	96	91

Assume the underlying distributions of light transmission percentage are continuous. Conduct a Wilcoxon rank-sum test to determine whether there is any evidence to suggest a difference in the population light transmission percentage medians. Use  $\alpha = 0.05$ .



**Example 14.3.2** Two high-rise apartment buildings have laundry rooms in the basement with an array of washers and dryers. A random sample of washers from each building was obtained, and the total water consumption for a medium load in each was measured (in gallons). The data are given in the following table.

Apartment A							
22.3	23.7	36.8	29.3	24.5	15.0	23.5	21.7
20.9	24.9						

Apartment B							
19.7	21.9	25.2	20.1	24.0	18.7	19.1	16.1
23.8	19.8	16.7	21.4				

Assume the underlying distributions of water consumptions are continuous. Use a Wilcoxon rank-sum test to determine whether there is any evidence to suggest a difference in the population median water consumptions. Use  $\alpha = 0.01$  and find the  $p$  value associated with this test.

Example (continued)

Ordered observations:

Company	A	B	B	B	B	B	B	B	A	B	A
Observation	15.0	16.1	16.7	18.7	19.1	19.7	19.8	20.1	20.9	21.4	21.7

Rank

Company	B	A	A	A	B	B	A	A	B	A	A
Observation	21.9	22.3	23.5	23.7	23.8	24.0	24.5	24.9	25.2	29.3	36.8

Ranks



## 14.4 The Kruskal–Wallis Test

1. A nonparametric analysis of variance of ranks.
2. Alternative to the ANOVA  $F$  test.  
No assumptions concerning normality or equal population variances.

General procedure:

1.  $k > 2$  independent random samples from continuous distributions:  
sample sizes  $n_1, n_2, \dots, n_k$ .
2. Rank the combined  $n = n_1 + n_2 + \dots + n_k$  observations.
3. Let  $r_i$  be the sum of the ranks associated with the observations from sample  $i$ .
4. If all the populations are identical, the  $r_i$ 's should be approximately the same.

### The Kruskal–Wallis Test

Suppose  $k > 2$  independent random samples of sizes  $n_1, n_2, \dots, n_k$  are obtained from continuous distributions. A hypothesis test concerning the general populations with significance level  $\alpha$  has the form:

$H_0$ : The  $k$  samples are from identical populations.

$H_a$ : At least two of the populations are different.

Combine all observations, and rank these values from smallest (1) to largest ( $n$ ). Equal values are assigned the mean rank for their positions. Let  $R_i$  be the sum of the ranks associated with the  $i$ th sample.

$$\text{TS: } H = \left[ \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} \right] - 3(n+1)$$

Critical values for the Kruskal–Wallis test statistic are available. However, if  $H_0$  is true and either

1.  $k = 3, \quad n_i \geq 6, \quad (i = 1, 2, 3)$  or
2.  $k > 3, \quad n_i > 5, \quad (i = 1, 2, \dots, k)$

then  $H$  has an approximate chi-square distribution with  $k - 1$  degrees of freedom.

$$\text{RR: } H \geq \chi_{\alpha, k-1}^2$$

**Remarks**

1. Reject  $H_0$  only for *large* values of the test statistic  $H$ .
2. If we reject the null hypothesis, there is evidence to suggest that at least two populations are different. Further analysis is needed to determine which pairs of populations are different, and how they differ; the means, medians, variances, shapes of the distributions, or other characteristics could be dissimilar.

**Example 14.4.1** A consumer group conducted a study to compare the power of rototillers manufactured for home gardeners. Random samples of 6.5 hp rototillers from three companies were obtained, and the maximum torque was measured (in ft-lbs) for each. The data are given in the following table.

Troy-Bilt (1)	14.1	13.6	12.5	14.8	14.4	13.6	12.4	14.3
Toro (2)	14.1	12.5	14.1	13.0	14.7	15.3	13.4	14.5
Stihl (3)	15.0	15.0	14.4	12.7	14.8	17.8	14.4	14.6

The underlying populations are assumed to be continuous. Use the Kruskal–Wallis test to determine whether there is any evidence that the maximum torque populations are different. Use a significance level of  $\alpha = 0.05$ .



**Example 14.4.2** Reflective vests are worn by runners, cyclists, road crews, firemen, EMTs, and others out at night. The brightness of a vest is measured by the coefficient of retroreflection with units candela per lux per meter squared ( $\text{cd}/\text{lux}/\text{m}^2$ ). Random samples of vests from four companies were obtained, and the coefficient of retroreflection at 50 feet was measured for each. The data are given in the following table.

Nathan (1)	251	233	241	234	240	252	232		
Plastex (2)	267	289	278	275	267	263	237	276	
Occulux (3)	295	293	258	227	255	222	193	226	
Wearguard (4)	204	263	260	295	298	265	205	269	307

The underlying populations are assumed to be continuous. Use the Kruskal–Wallis test to determine whether there is any evidence that the coefficient of retroreflection populations are different. Use  $\alpha = 0.05$  and find bounds on the  $p$  value associated with this test.



## 14.5 The Runs Test

1. Consider the *order* in which observations are drawn from a population.
2. Determine only whether there is evidence that the *sequence* of observations is not random.
3. Test is based on the number of runs: smaller subsequences in which the observations are the same.

### Definition

A **run** is a series, or subsequence, of one or more identical observations.

**Example 14.5.1** Suppose a convenience store sells two kinds of sugar-free gums: Trident and Orbit. A sample of customers who purchased sugar-free gum was obtained, and the sequence of responses is given in the following table.

T	O	T	T	T	T	O	O	T	O	O	T	T	T	O	O	T	O	T	O
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Find the number of runs in this sample.

### Remarks

1. Runs test: appropriate for testing whether a sequence of observations is not random.

It can be used if the data can be divided into two mutually exclusive categories.

May also be used for quantitative data that can be classified into one of two categories.

2. The smallest possible number of runs in any sample is 1.

The largest possible number of runs in a sample depends on the number of observations in each category.

3. The statistical test is based on the total number of runs.

If the order of observations is random, then the total number of runs should not be very large or very small.

Table of critical values that uses the exact distribution for the number of runs.

A normal approximation that can be used if number of observations in each category is large.

### The Runs Test

Suppose a sample is obtained in which each observation is classified into one of two mutually exclusive categories. Assume there are  $m$  observations in one category and  $n$  observations in the other.

$H_0$ : The sequence of observations is random.

$H_a$ : The sequence of observations is not random.

TS:  $V$  = the total number of runs.

RR:  $V \geq v_1$  or  $V \leq v_2$

The critical values  $v_1$  and  $v_2$  are obtained from Table 11 such that  $P(V \geq v_1) \approx \alpha/2$  and  $P(V \leq v_2) \approx \alpha/2$ .

*The Normal Approximation:* As  $m$  and  $n$  increase, the statistic  $V$  approaches a normal distribution with

$$\mu_V = \frac{2mn}{m+n} + 1 \quad \text{and} \quad \sigma_V^2 = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}.$$

Therefore, the random variable  $Z = \frac{V - \mu_V}{\sigma_V}$  has approximately a standard normal distribution. The normal approximation is good when both  $m$  and  $n$  are greater than 10. In this case,  $Z$  is the test statistic and the rejection region is

RR:  $|Z| \geq z_{\alpha/2}$

**Example 14.5.2** Counterfeit cell phone batteries are easy to make and sell for a fraction of the cost of legitimate batteries. However, they can overheat and cause skin burns and fires. A random sample of consumer cell phones was obtained, and each battery was classified as legitimate (L) or fake (F). The sequence of observations is given in the following table.

---

L	L	F	F	F	L	L	L	F	L	L	F	L	F	F	F	F	L
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

---

Use the runs test with  $\alpha = 0.05$  to determine whether there is any evidence that the order in which the sample was selected was not random.



**Example 14.5.3** When a person sits on a bike, the recommended sag in a shock absorber on a high quality frame is 9.5 mm. A sample of bicycles with high quality frames was obtained, and the sag was measured on each. The data are given in the following table, in order from left to right.

---

8.7	9.7	9.3	9.6	9.2	9.7	9.4	10.1	9.2	9.6	9.8	9.5	9.5	9.0	9.9
9.3	9.6	9.8	9.5	9.8	9.6	9.3	9.7	9.4	9.2	10.1	9.4	9.4	9.0	9.2

---

Is there any evidence to suggest that the order of the observations is not random with respect to the recommended sag? Use  $\alpha = 0.01$  and find the  $p$  value associated with this test.

## 14.6 Spearman's Rank Correlation

1. Sample correlation coefficient,  $r$ : a measure of the strength of the linear relationship between two continuous variables.
2. *Spearman's rank correlation coefficient*: a nonparametric alternative, computed using ranks.
3. No assumptions about the underlying populations.

Each observation is converted to a rank.

Compute the sample correlation coefficient using the ranks in place of the actual observations.

### Spearman's Rank Correlation Coefficient

Suppose there are  $n$  pairs of observations  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ . Rank the observations in each sample separately, from smallest to largest. Let  $u_i$  be the rank of the  $i$ th observation in the first sample and let  $v_i$  be the rank of the  $i$ th observation in the second sample. **Spearman's rank correlation coefficient**,  $r_S$ , is the sample correlation coefficient between the ranks and is computed using the equation

$$r_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

where  $d_i = u_i - v_i$ .

### Remarks

1. Equal values within each sample are assigned the mean rank of their positions in the ordered list.

The equation above is not exact when there are tied observations within either sample.

In this case, compute  $r_S$  by finding the sample correlation coefficient between the ranks.

2.  $r_S$  is always between  $-1$  and  $+1$ .

Values near  $-1$  indicate a strong negative linear relationship between the ranks.

Values near  $+1$  suggest a strong positive linear relationship between the ranks.

3. Correlation does not imply causation;  $r_S$  is a measure of the linear association between the ranks.

A strong linear relationship between the ranks does not imply that the relationship between the original variables is also linear.

**Example 14.6.1** A study was conducted to determine whether productivity is affected by the air quality in offices. A random sample of adult office workers was obtained, and each was asked to perform a set of simulated tasks (typing, filing, proof-reading, etc.). The air quality was measured by considering the outdoor supply rate (in L/s) for the building and the time to complete these tasks (in hours) was used as a measure of performance. The data are given in the following table.

Adult	1	2	3	4	5	6	7	8	9	10
Air quality	18	17	21	23	7	27	18	33	20	18
Performance	6.35	6.17	5.73	5.38	5.84	5.04	6.03	4.53	6.05	6.19

Compute Spearman's rank correlation coefficient and interpret this value.

Adult $i$	Air quality $x_i$	Rank $u_i$	Performance $y_i$	Rank $v_i$	Difference $d_i$
1	18		6.35		
2	17		6.17		
3	21		5.73		
4	23		5.38		
5	7		5.84		
6	27		5.04		
7	18		6.03		
8	33		4.53		
9	20		6.05		
10	18		6.19		

Example (continued)

**Example 14.6.2** Wet milling of corn produces a variety of byproducts including sweeteners, gluten, and starch. A study was conducted to determine if the temperature of the starch slurry is associated with the final starch yield. A random sample of fifty-six pound bushels of corn was obtained, and each was wet milled. The temperature of the starch slurry (in °C) and the starch yield (in pounds) are given in the following table.

Temperature	107.7	106.7	107.0	108.0	101.9	101.3	107.7	103.6
Yield	35.1	34.8	35.2	33.9	34.8	32.5	33.7	36.2

Temperature	107.2	102.6	103.8	106.2	103.1	107.7	104.4	105.4
Yield	35.5	34.8	36.3	36.6	34.7	35.0	35.7	36.0

- Compute Spearman's rank correlation coefficient using the formula given in this section and by finding the correlation between the ranks. Compare these two numbers and describe the relationship between temperature and yield.
- Construct a scatter plot of this data. Does the scatter plot suggest there is a relationship between temperature and yield? If so, describe the relationship.

Example (continued)

Temperature	Rank	Yield	Rank	Difference
107.7		35.1		
106.7		34.8		
107.0		35.2		
108.0		33.9		
101.9		34.8		
101.3		32.5		
107.7		33.7		
103.6		36.2		
107.2		35.5		
102.6		34.8		
103.8		36.3		
106.2		36.6		
103.1		34.7		
107.7		35.0		
104.4		35.7		
105.4		36.0		