# Introductory Statistics with R

## Contents



# 1. Sampling and Data

## 1.1. Key Terms in Statistics

### Objectives

- Distinguish the difference between a population and a sample.
- Distinguish between a parameter and a statistic.
- Identify variables and data.

### The Definition of Statistics

The science of **statistics** deals with the collection, analysis, interpretation, and presentation of **data**. We see and use data in our everyday lives.

In this course, you will learn how to organize and summarize data. Organizing and summarizing data is called **descriptive statistics**. Two ways to summarize data are by graphing and by using numbers (for example, by finding an average). After you have studied probability and probability distributions, you will use formal methods for drawing conclusions from “good” data. The formal methods are called **inferential statistics**. Statistical inference uses probability to determine how confident we can be that our conclusions are correct.

Effective interpretation of data (inference) is based on good procedures for producing data and thoughtful examination of the data. You will encounter what will seem to be too many mathematical formulas for interpreting data. The goal of statistics is not to perform numerous calculations using the formulas, but to gain an understanding of your data. The calculations can be done using a

calculator or a computer. The understanding must come from you. If you can thoroughly grasp the basics of statistics, you can be more confident in the decisions you make in life.

---

### Example 1.1.1

We asked $14$ people chosen randomly from a statistics class how many hours they sleep per night. We obtained the following data:

5, 5.5, 6, 6, 6, 6.5, 6.5, 6.5, 6.5, 7, 7, 8, 8, 9

We can summarize this data in a **histogram**, as shown in Figure 1.1.1.

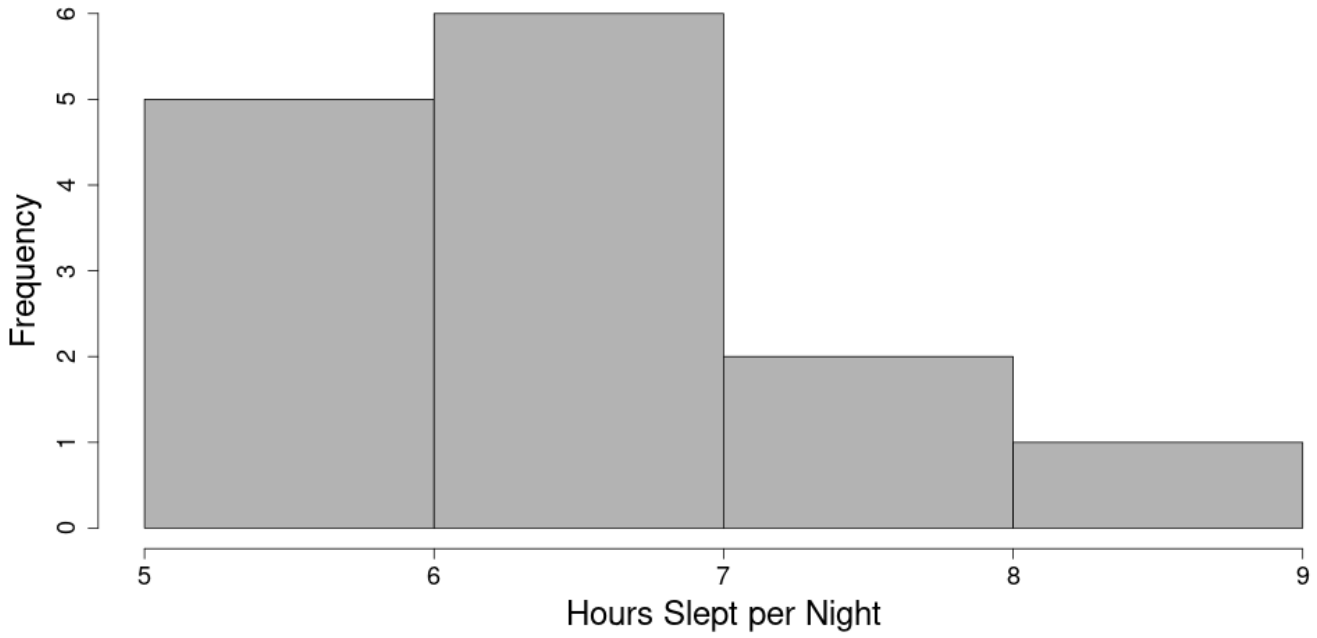



*Fig. 1.1.1* A histogram of the data above. The height of each bar in the histogram tells us how many data values we have in the range covered by the histogram. For example, we can see that 5 of the individuals surveyed sleep between 5 and 6 hours per night.

By looking at the histogram, we can see that most people we surveyed got no more than $7$ hours of sleep per night.

We can summarize our data by calculating the average number of hours slept: the $14$ statistics students surveyed slept an average of $6.679$ hours per night. Since this was the average for our sample of $14$ students, we might infer that the average amount of sleep of *all* statistics students is close to $6.679$ hours per night.

Note that this example deals in a very simple manner with the fundamental components of statistics mentioned above. We are provided *data* and told how the data was *collected*. We *present* this data graphically using a histogram, then provide a simple *analysis* of the data, concluding that most people surveyed get no more than $7$ hours of sleep per night. We calculated the average number of hours those surveyed slept each night; this *descriptive statistic* summarized our data. Using the average we calculated, we *inferred* that it was likely that if we averaged the amount of sleep of all statistics students, it would be close to the average amount of sleep of the $14$ students we surveyed.

In this course, we will be studying each of these concepts in much greater detail.

Skip to main content

## Key Terms

In statistics, we generally want to study a **population**. You can think of a population as a collection of persons, things, or objects under study. To study the population, we select a **sample**. The idea of **sampling** is to select a portion (or subset) of the larger population and study that smaller portion (the sample) to infer information about the population.

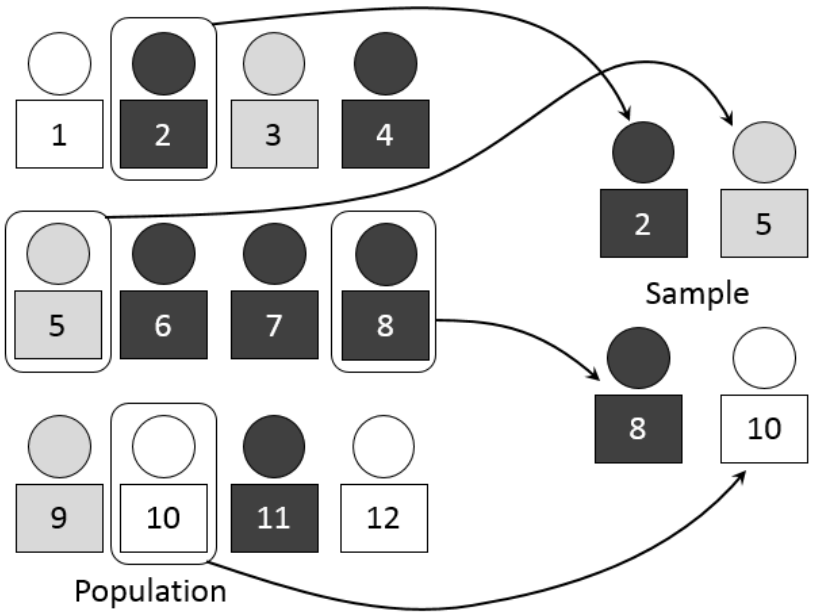



*Fig. 1.1.2* A sample of four individuals is chosen from a population of twelve individuals.[1]

Because it usually takes a lot of time and money to examine an entire population, sampling is a very practical technique. For example:

- During a national election, it is impossible for a pollster to survey all the millions of voters in the country. Instead, the pollster will sample only a few hundred voters from the population. The opinions expressed by those sampled in the survey (hopefully) represent the views of voters in the entire country.
- An ice cream manufacturer wouldn't want to taste-test every carton of ice cream produced to ensure its quality, because the testing process would ruin the ice cream. Instead, the manufacturer could taste-test a sample of ice cream cartons. If any quality control issues are present in the ice cream cartons sampled, it likely indicates there are problems in the larger population.

---

### Example 1.1.2

An advertising company has created two ads for an American restaurant chain, one featuring a burger and one featuring a salad. They want to know which of the two ads Americans would find more appealing. To find out, the advertising company surveys $50$ people and asks them which of the two ads they prefer. What is the population and what is the sample in this example?

#### Solution

- Population: All Americans
- Sample: The 50 people that were surveyed

Skip to main content

We sample the population so we can estimate a population **parameter**. A parameter is a number that summarizes the characteristics of a population. Because it can be difficult to study an entire population, we often cannot find the exact value of a parameter. Instead, we estimate the population parameter by calculating a sample **statistic**. A statistic is a number that summarizes the characteristics of a sample. A statistic is easier to calculate than a parameter because we only need to study a sample—not the entire population—to calculate a statistic. We can use a sample statistic to estimate a population parameter.

The **mean** of a population (commonly referred to as the **average** of a population) is one example of a parameter. Suppose we want to find the mean age of people living in California. Since it is difficult to find the age of every person in California, we could sample just a few Californians and calculate the mean age of the sample. We can use this statistic—the mean age of the Californians sampled—to estimate the parameter—the mean age of all Californians. For instance, if the average age of Californians in our sample is 35.2 years old, we would estimate that the average age of all Californians is around 35.2 years old.

Another common parameter is the **proportion** of a population. The proportion of the population is a fraction of the population that has some characteristic. We can express proportions as fractions, percents, or decimals. Suppose on your way to school you counted 50 cars, 8 of which were red. So in your sample of 50 cars, the proportion of red cars is $\frac{8}{50}$. (As a percent, this proportion is 16%. As a decimal, it is 0.16.) We can use this statistic from our sample of 50 cars to estimate the proportion of the population of all cars that are red; that is, we estimate that around 16% of all cars are red.

One of the main concerns in the field of statistics is how accurately a statistic estimates a parameter. The accuracy really depends on how well the sample represents the population. The sample must contain the characteristics of the population in order to be a representative sample. We are interested in both the sample statistic and the population parameter in inferential statistics. In a later chapter, we will use the sample statistic to test the validity of the established population parameter.

**Tip**

The first letter of *parameter* and the first letter of *statistic* will help you remember in which setting each word is used. A **p**arameter summarizes characteristics of a **p**opulation. A **s**tatistic summarizes characteristics of a **s**ample.

---

## Example 1.1.4

The Bureau of Labor Statistics (BLS) wants to know the average annual income of statisticians. To find out, they ask 150 statisticians across the country what their annual income is, then average their results. What is the parameter and the statistic of interest in this example?

### Solution

- Population Parameter: The average annual income of all statisticians
- Sample Statistic: The average annual income of the 150 statisticians that were surveyed

---

To calculate a sample statistic, we must gather information from each member sampled. A **variable** is a characteristic or measurement that can be determined for each member of a population. Variables are usually denoted by campital letters such as $X$ or $Y$. To calculate a sample statistic, we must gather **data** corresponding to the variable we are studying from each member sampled. The data are the actual values of the variable. They may be numbers or they may be words. A **datum** is a single data value. For example:

- Let $X$ equal the number of points earned by a math student on an exam. Examples of data for $X$ might be 91, 76, or 83.

Skip to main content

- Let $Y$ be a person's political party affiliation. Some examples of data for $Y$ include "Republican", "Democrat", and "Independent".

**Note**

In practice, the word *data* is sometimes treated as a plural noun, in which case plural verbs and pronouns are used, and at other times *data* is treated as a singular collective noun, in which case singular verbs and pronouns are used. In most cases, either form is acceptable: "The data *are* accurate" or "The data *is* accurate" have the same meaning.[2] This is similar to how *people* is a plural noun ("The people *are* happy") but *population* is a singular collective noun ("The population *is* happy), even though both words indicate a group of several individuals.

In formal writing, the plural form of *data* is often preferred, and several writing style guides, such as MLA and APA, only allow the plural form.

---

## Example 1.1.5

A medical doctor wishes to know the average time a laproscopic gallbladder removal surgery takes at her hospital. To find out, she samples 14 gallbladder removal surgeries to find out and notes how long each surgery takes. Three of the surgeries sampled respectively took 1.5 hours, 2.25 hours, and 1.25 hours. What is the variable and what is the data in this example?

### Solution

- Variable: $X$ = The time it takes to perform one gallbladder surgery.
- Data: 1.5 hours, 2.25 hours, 1.25 hours

---

## Example 1.1.6

We want to know the average (mean) amount of money first year college students spend at ABC College on school supplies that do not include books. We randomly surveyed 100 first year students at the college. Three of those students spent $150, $200, and $225, respectively.

Identify the population, sample, parameter, statistic, variable, and data in this example.

### Solution

- Population: First year college students at ABC College
- Sample: The 100 first year students surveyed
- Population Parameter: The average (mean) amount of money spent (excluding books) by first year college students at ABC College
- Sample Statistic: The average (mean) amount of money spent (excluding books) by first year college students in the sample
- Variable: $X$ = The amount of money spent (excluding books) by one first year student attending ABC College
- Data: $150, $200, and $225

---

An insurance company would like to determine the proportion of all medical doctors who have been involved in one or more malpractice lawsuits. The company selects 500 doctors at random from a professional directory and determines the number in the sample who have been involved in a malpractice lawsuit.

Identify the population, sample, parameter, statistic, variable, and data in this example.

Solution

- Population: All medical doctors
- Sample: The 500 doctors selected at random from the professional directory
- Population Parameter: The proportion of medical doctors who have been involved in one or more malpractice suits
- Sample Statistic: The proportion of medical doctors who have been involved in one or more malpractice suits in the sample
- Variable: $X$ = Whether or not one doctor has been involved in one or more malpractice suits
- Data: Either “Yes, was involved in one or more malpractice suits,” or “No, was not involved in one or more malpractice suits”

---

[1] Figure 1.1.2 was created by Dan Kernler and is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

[2] See Merriam-Webster’s entry for *data* for more information.

# 1.2. Qualitative Data and Quantitative Data

## Objectives

- Distinguish between qualitative and quantitative data.
- For quantitative data, distinguish between discrete and continuous data.

## Qualitative Data and Quantitative Data

Data may come from a population or from a sample. Lowercase letters like x or y generally are used to represent data values. Most data can be put into the following categories:

- Qualitative
- Quantitative

**Qualitative data** are the result of categorizing or describing attributes of a population. Qualitative data are also often called **categorical data**. Hair color, blood type, ethnic group, the car a person drives, and the street a person lives on are examples of qualitative data. Qualitative data are generally described by words or letters. For instance, hair color might be black, dark brown, light brown, blonde, gray, or red. Blood type might be AB+, O-, or B+.

**Quantitative data** are always numbers, and are often also called **numerical data**. Quantitative data are the result of **counting** or **measuring** attributes of a population. Weight, amount of money, pulse rate, number of people living in your town, and number of students who take statistics are examples of quantitative data. Quantitative data may be either **discrete** or **continuous**.

Researchers often prefer to use quantitative data over qualitative data because it lends itself more easily to mathematical analysis. For example, it does not make sense to find an average of qualitative data like hair color or blood type.

Skip to main content

All data that are the result of counting are called **quantitative discrete data**. These data take on only certain numerical values. If you count the number of phone calls you receive for each day of the week, you might get values such as $0$, $1$, $2$, or $3$. It doesn't make sense to receive $1.25$ phone calls in a day.

Data that are not only made up of counting numbers, but that may include fractions, decimals, or irrational numbers, are called **quantitative continuous data**. Continuous data are often the results of measurements like lengths, weights, or times. A list of the duration in minutes for all the phone calls that you make in a week, with numbers like $2.4$, $7.5$, or $11.0$, would be quantitative continuous data.

> **Tip**
>
> One way to remember the difference between qualitative data and quantitative data is that **qual**itative data measures the **qual**ities of the members of a population, whereas **quant**itative data deals with attributes that are measured using **quant**ities or numbers.

---

## Example 1.2.1

Students each carry a number of books in their backpacks. You sample five students. Two students carry 3 books, one student carries 4 books, one student carries 2 books, and one student carries 1 book.

1. What are the data in this example?
2. Are the data qualitative or quantitative?
3. If the data are quantitative, are they continuous or discrete?

### Solution

1. The data are the numbers of books the students carry: 3, 3, 4, 2, 1.
2. Since this data are counts of the quantities of books each student caries, they are quantitative data.
3. The data are discrete. A student can have 2 books in their backpack or 3 books in their backpack, but not 2.68 books. Only whole numbers are allowed.

---

## Example 1.2.2

Students each have a backpack, and each backpack has a certain weight. You sample five students. The weights (in pounds) of their backpacks are 6.2, 7, 6.8, 9.1, 4.3.

1. What are the data in this example?
2. Are the data qualitative or quantitative?
3. If the data are quantitative, are they continuous or discrete?

### Solution

1. The data are the weights of the backpacks: 6.2, 7, 6.8, 9.1, 4.3.
2. Since the data are measures of how much the backpacks weigh, they are quantitative data.

Skip to main content

3. The data are continuous. A student can have a backpack that weighs 6.2 pounds or a backpack that weighs 6.8 pounds, or a backpack that weighs any amount in between.

---

### Example 1.2.3

Students each have a backpack of a certain color. Again, you sample five students. One student has a red backpack, two students have black backpacks, and two students have green backpacks.

1. What are the data in this example?
2. Are the data qualitative or quantitative?
3. If the data are quantitative, are they continuous or discrete?

#### Solution

1. The data are the colors of the backpacks the students carry: red, black, black, green, green.
2. Since this data describes a quality that each backpack possesses—its color—these are qualitative data.
3. Since our data are qualitative, not quantitative, this question does not apply.

---

### Example 1.2.4

Determine the correct data type (quantitative or qualitative). Indicate whether quantitative data are continuous or discrete. Hint: Data that are discrete often start with the words “the number of.”

1. the number of pairs of shoes you own
2. the type of car you drive
3. the distance it is from your home to the nearest grocery store
4. the number of classes you take per school year.
5. the type of calculator you use
6. weights of sumo wrestlers
7. number of correct answers on a quiz

#### Solution

1. Quantitative discrete
2. Qualitative
3. Quantitative continuous
4. Quantitative discrete
5. Qualitative
6. Quantitative continuous
7. Quantitative discrete

## 1.3. Sampling Techniques

Skip to main content

## Objectives

- Identify the standard methods of obtaining data and the advantages and disadvantages of each.
- Construct a random sample from a population using simple random sampling, systematic sampling, stratified sampling, and cluster sampling.

## Sampling and Bias

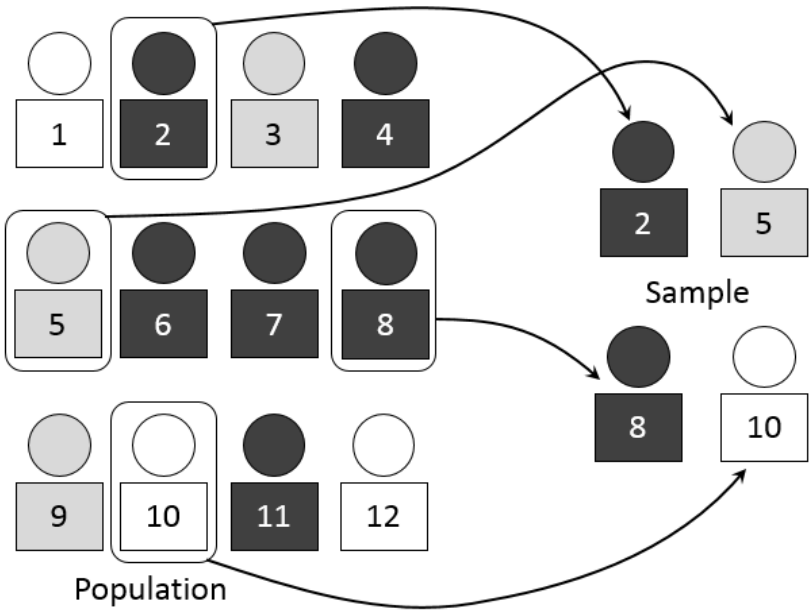
Gathering information about an entire population often costs too much or is virtually impossible. Instead, we gather information about a sample of the population. **A sample should have the same characteristics as the population it is representing.** If the sample does not have the same characteristics as the population, the statistic may not be a good estimator for the parameter. When this happens, we say the sample statistic is **biased**.

For example, suppose a pollster wants to find out if the United States population believes a college education is important for a successful career. If the pollster only samples college students, then we might expect the proportion of the sample that believes college is important to be higher than the proportion of the population that believes college is important. Because the sample (which includes only college students) does not have the same characteristics as the population (which includes college students, college graduates, those who have never been to college, etc.), the sample statistic probably isn't a good estimate of the population parameter.

To try to make sure that the sample has the same characteristics as the population, most statisticians use various methods of **random sampling**. There are several different methods of random sampling. In each form of random sampling, each member of a population initially has an equal chance of being selected for the sample. Each method has pros and cons. We describe some of the most common methods below.

## Simple Random Sampling

The easiest method to describe is called a **simple random sample**. In simple random sampling, each member of the population has an equal chance of being chosen for the sample and is chosen independent of any other member in the population.




Skip to main content

For example, suppose Professor Baldwin wants to form a six-person study group from the students in his precalculus class, which has 30 students. To choose a simple random sample of size six from the students of his class, Professor Baldwin could put all 30 names in a hat, shake the hat, close his eyes, and pick out six names. A more technological way is for Professor Baldwin to first list the last names of the members of his class together with a number, as in Table 1.3.1.

*Table 1.3.1* Professor Baldwin's Class Roster.

| Number | Name | Number | Name | Number | Name |
|---|---|---|---|---|---|
| 1 | Anselmo | 11 | Khan | 21 | Roquero |
| 2 | Bautista | 12 | Legeny | 22 | Roth |
| 3 | Bayani | 13 | Lundquist | 23 | Rowell |
| 4 | Cheng | 14 | Macierz | 24 | Salangsang |
| 5 | Cuarismo | 15 | Motogawa | 25 | Slade |
| 6 | Cuningham | 16 | Okimoto | 26 | Stratcher |
| 7 | Fontecha | 17 | Patel | 27 | Tallai |
| 8 | Hong | 18 | Price | 28 | Tran |
| 9 | Hoobler | 19 | Quizon | 29 | Wai |
| 10 | Jiao | 20 | Reyes | 30 | Wood |

Professor Baldwin could now use a computer to generate six random numbers and choose the individuals on the table matching the numbers. We can do this in R with the `sample` function:

```
sample(x, size)
```

Here, `x` is a list of the members of the population, and `size` is the size of the sample we desire. The `sample` function will randomly choose a simple random sample from population `x` of the given `size`.

In this example, we would want `x` to be a list of numbers from 1 to 30 corresponding to the numbers on our table. R makes it easy to create a list of consecutive numbers. Simply type `start:end`, where `start` is the first number in the list and `end` is the last number in the list. For a list from 1 to 30, we would type `1:30`.

And since we want to choose six students from the class to be in the study group, `size` will be 6.

```
# Randomly choose 6 students from the class
sample(1:30, size = 6)
```

4 · 25 · 19 · 7 · 17 · 18

In this case, the `sample` function randomly chose students 4, 25, 19, 7, 17, and 18 (Cheng, Slade, Quizon, Fontecha, Patel, and Price) to be in the study group. (See Table 1.3.2.)

**Simple Ra**

Skip to main content

For large samples, simple random sampling tends to do a very good job at representing the characteristics of a population. However, the smaller the sample is, the more likely it is that some characteristics or groups of a population won't be accurately represented in the sample. Also, to make sure that each member of the population has an equal chance of being chosen, simple random sampling requires a list of the full population, which may be difficult or impossible to obtain for large populations. Because of this, simple random sampling can sometimes be tedious, time consuming, and expensive to implement.

**Warning**

Humans don't have a good intuition for randomness. To demonstrate this point, the author asked 113 of his math students to pick a random number from 1 to 10. Figure 1.3.2 shows a bar chart on the left of how many times each number was chosen by the students, as well as a bar chart on the right of how many times we would expect each number to be chosen by a truly uniformly random process.

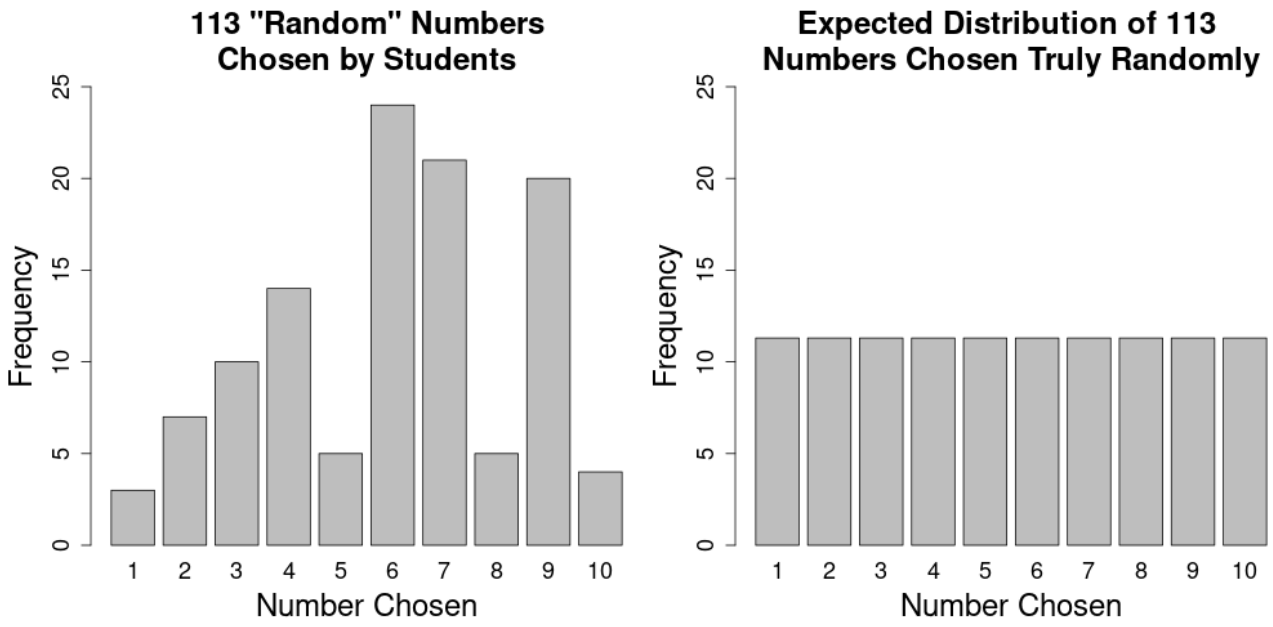



*Fig. 1.3.2* The bar chart on the left shows how many times each number from 1 to 10 was chosen by 113 students. The bar chart on the right shows how many times we would expect each number to be chosen if the numbers were actually chosen randomly. The dramatic differences in the bar charts suggest that the students did not truly choose their numbers randomly.

With a truly random process, we would expect 1 to be chosen about as often as 6 is chosen. However, only three of the 113 students chose 1, while twenty-four students chose 6.

When you need random values, **do not trust yourself to be random.** Instead, use a random process like rolling a die, picking values out of a hat, or (as we will most often do in this class) using a computer to generate random values.

## Example 1.3.1

Table 1.3.3 contains real data of the revenues per capita from each city in Riverside County, California in 2019. The data in the table is displayed how data in a spreadsheet might be displayed. Sample five cities using simple random sampling.

*Table 1.3.3* Real data from the cities in Riverside County, California in 2019. The data is organized how data might be organized in a spreadsheet with each row labeled with a number and each column labeled with a letter. [2]

| | A | B | C | D | E |
|---|---|---|---|---|---|
| **1** | **City Name** | **Fiscal Year** | **Total Revenues** | **Estimated Population** | **Revenues Per Capita** |
| **2** | Banning | 2019 | 80260834 | 31044 | 2585 |
| **3** | Beaumont | 2019 | 74407907 | 48401 | 1537 |
| **4** | Blythe | 2019 | 18024910 | 19428 | 928 |
| **5** | Calimesa | 2019 | 13892401 | 9159 | 1517 |
| **6** | Canyon Lake | 2019 | 6168512 | 11285 | 547 |
| **7** | Cathedral City | 2019 | 90568810 | 54907 | 1649 |
| **8** | Coachella | 2019 | 49670997 | 46351 | 1072 |
| **9** | Corona | 2019 | 365886198 | 168101 | 2177 |
| **10** | Desert Hot Springs | 2019 | 28443202 | 29251 | 972 |
| **11** | Eastvale | 2019 | 38063775 | 66078 | 576 |
| **12** | Hemet | 2019 | 97943137 | 84754 | 1156 |
| **13** | Indian Wells | 2019 | 50539127 | 5445 | 9282 |
| **14** | Indio | 2019 | 159055776 | 89406 | 1779 |
| **15** | Jurupa Valley | 2019 | 53992209 | 106318 | 508 |
| **16** | Lake Elsinore | 2019 | 86746399 | 62949 | 1378 |
| **17** | La Quinta | 2019 | 83793774 | 42098 | 1990 |
| **18** | Menifee | 2019 | 88732376 | 93452 | 949 |
| **19** | Moreno Valley | 2019 | 211345835 | 208297 | 1015 |
| **20** | Murrieta | 2019 | 110754720 | 118125 | 938 |
| **21** | Norco | 2019 | 53558820 | 26386 | 2030 |
| **22** | Palm Desert | 2019 | 111757506 | 53625 | 2084 |
| **23** | Palm Springs | 2019 | 271251311 | 48733 | 5566 |
| **24** | Perris | 2019 | 111143322 | 76971 | 1444 |

Skip to main content

| | A | B | C | D | E |
|---|---|---|---|---|---|
| **1** | **City Name** | **Fiscal Year** | **Total Revenues** | **Estimated Population** | **Revenues Per Capita** |
| **25** | Rancho Mirage | 2019 | 68949584 | 18489 | 3729 |
| **26** | Riverside | 2019 | 942831959 | 328101 | 2874 |
| **27** | San Jacinto | 2019 | 52582546 | 48878 | 1076 |
| **28** | Temecula | 2019 | 140192073 | 113826 | 1232 |
| **29** | Wildomar | 2019 | 23506765 | 36066 | 652 |

Solution

Rather than assign each city a number, we can use the row numbers already on the spreadsheet. Note, though, that row $1$ contains the column headers. The actual data is contained in rows $2$ to $29$. To perform simple random sampling, we use the `sample` function to select five random row numbers between $2$ and $29$.

```
sample(2:29, size = 5)
```

2 · 10 · 13 · 23 · 17

Our sample includes the city in row 2 (Banning), row 10 (Desert Hot Springs), row 13 (Indian Wells), row 23 (Palm Springs), and row 17 (La Quinta).

---

## Systematic Sampling

In a **systematic sample**, the population is in some order, and every $k$th member of the population is sampled for some number $k$ called the **sampling interval**. (For example, if $k = 7$, then every $7$th member of the population is included.) To make sure we sample members throughout the whole population, we calculate the sampling interval $k$ using the formula

$$k = \frac{N}{n},$$

where $N$ is the size of the population, and $n$ is the size of the sample we want. To make the sampling random, one of the first $k$ members of the population is randomly chosen as the first member of the sample and the starting point of the sampling process.

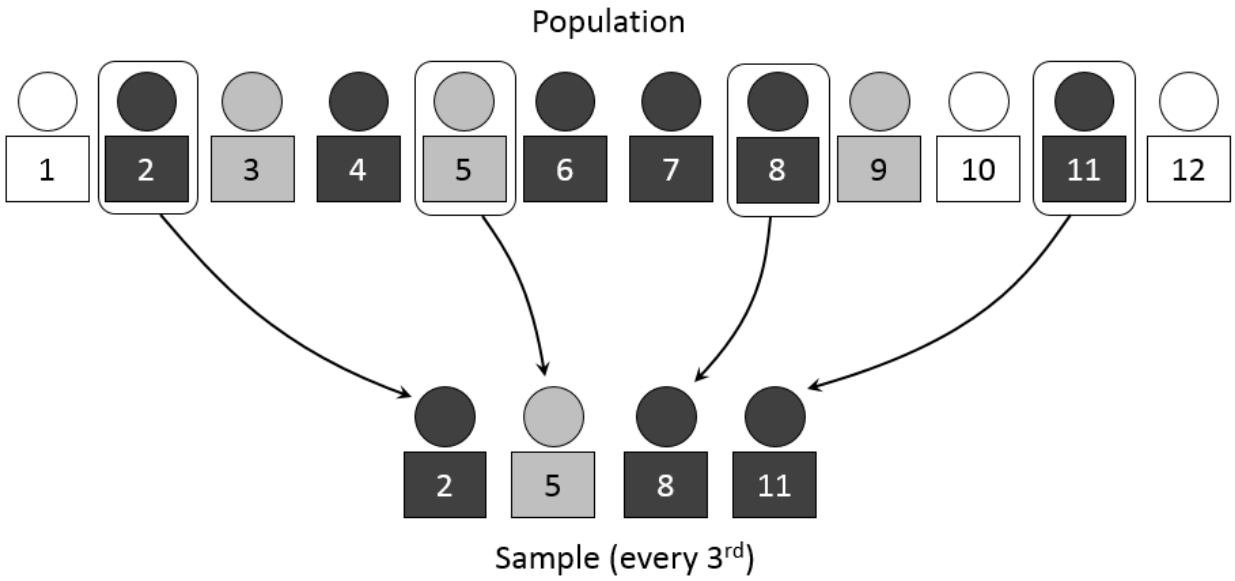



*Fig. 1.3.3* Four individuals are chosen from a population of twelve individuals using systematic sampling. The second individual is randomly chosen as a starting point. Since there are $N = 12$ individuals in the population and the size of the sample is $n = 4$, the interval between each individual sampled is $k = 12/4 = 3$. So every 3rd individual is chosen from the starting point.[3]

For example, suppose Professor Baldwin wants to choose six students from his precalculus class for a study group using systematic sampling. The students are already in an ordered list in Table 1.3.1. To begin the systematic sampling, Professor Baldwin first calculates the sampling interval $k$. Since there are $N = 30$ total students in the class, and since Professor Baldwin wants to sample $n = 6$ of the students for the study group, the sampling interval is

$$k = \frac{30}{6} = 5.$$

So Professor Baldwin will sample every $5$th student.

To determine which student is the first student chosen to be sampled, Professor Baldwin uses R to randomly select one of the first 5 students.

must randomly choose a starting point. To make sure the choice is truly random, we can use the `sample` function. We will choose $1$ student from the list of $30$ students as a starting point. Systematic sampling is frequently chosen because it is a simple method.

```
# Randomly choose one student as a starting point
sample(1:5, size = 1)
```

4

So the $4$th student in Table 1.3.1, Cheng, is the first person in the study group and the starting point of the systematic sample.

Since the sampling interval is $k = 5$, Professor Baldwin next samples every $5$th student in the population after Cheng. After Cheng, the next student in the study group is student $4 + 5 = 9$, Hoobler. The next student after Hoobler is student number $9 + 5 = 14$, Macierz. The next student is number $14 + 5 = 19$, Quizon, then student number $19 + 5 = 24$, Salangsang. The final student is student number $24 + 5 = 29$, Wai. (See Table 1.3.4.)



Skip to main content

Systematic sampling is a usually effective sampling technique that is easy to understand and implement. However, if members of the population with a certain characteristic regularly repeat in the order, systematic sampling may over-sample or under-sample these members. For example, suppose that a microchip manufacturer performs quality testing on every $9$th microchip in the assembly line. If there is a fault on the assembly line that causes every $3$rd microchip manufactured to malfunction, then the quality testing will either sample none of the faulty microchips (if the systematic sampling is started on a working microchip), or the quality testing will sample only faulty microchips (if the systematic sampling is started on a faulty microchip). In either case, the sample would be biased.

---

## Example 1.3.2

The spreadsheet below contains real data of the revenues per capita of each city in Riverside County, California in 2019. Sample five cities using systematic sampling.

*Table 1.3.5* Real data from the cities in Riverside County, California in 2019. The data is organized how data might be organized in a spreadsheet with each row labeled with a number and each column labeled with a letter. [2]

| | A | B | C | D | E |
|---|---|---|---|---|---|
| **1** | **City Name** | **Fiscal Year** | **Total Revenues** | **Estimated Population** | **Revenues Per Capita** |
| **2** | Banning | 2019 | 80260834 | 31044 | 2585 |
| **3** | Beaumont | 2019 | 74407907 | 48401 | 1537 |
| **4** | Blythe | 2019 | 18024910 | 19428 | 928 |
| **5** | Calimesa | 2019 | 13892401 | 9159 | 1517 |
| **6** | Canyon Lake | 2019 | 6168512 | 11285 | 547 |
| **7** | Cathedral City | 2019 | 90568810 | 54907 | 1649 |
| **8** | Coachella | 2019 | 49670997 | 46351 | 1072 |
| **9** | Corona | 2019 | 365886198 | 168101 | 2177 |
| **10** | Desert Hot Springs | 2019 | 28443202 | 29251 | 972 |
| **11** | Eastvale | 2019 | 38063775 | 66078 | 576 |
| **12** | Hemet | 2019 | 97943137 | 84754 | 1156 |
| **13** | Indian Wells | 2019 | 50539127 | 5445 | 9282 |
| **14** | Indio | 2019 | 159055776 | 89406 | 1779 |
| **15** | Jurupa Valley | 2019 | 53992209 | 106318 | 508 |
| **16** | Lake Elsinore | 2019 | 86746399 | 62949 | 1378 |
| **17** | La Quinta | 2019 | 83793774 | 42098 | 1990 |
| **18** | Menifee | 2019 | 88732376 | 93452 | 949 |
| **19** | Moreno Valley | 2019 | 211345835 | 208297 | 1015 |
| **20** | Murrieta | 2019 | 110754720 | 118125 | 938 |
| **21** | Norco | 2019 | 53558820 | 26386 | 2030 |
| **22** | Palm Desert | 2019 | 111757506 | 53625 | 2084 |
| **23** | Palm Springs | 2019 | 271251311 | 48733 | 5566 |
| **24** | Perris | 2019 | 111143322 | 76971 | 1444 |

Skip to main content

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | City Name | Fiscal Year | Total Revenues | Estimated Population | Revenues Per Capita |
| 25 | Rancho Mirage | 2019 | 68949584 | 18489 | 3729 |
| 26 | Riverside | 2019 | 942831959 | 328101 | 2874 |
| 27 | San Jacinto | 2019 | 52582546 | 48878 | 1076 |
| 28 | Temecula | 2019 | 140192073 | 113826 | 1232 |
| 29 | Wildomar | 2019 | 23506765 | 36066 | 652 |

Solution

We must first determine the sampling interval $k$ between the cities in the sample. Since there are $N = 28$ cities in the population, and we want a sample size of $n = 5$, we calculate

$$k = \frac{28}{5} = 5.6.$$

But the size of the sampling interval must be a whole number, so we round $5.6$ to a sampling interval of $k = 6$.

Next, we will randomly choose one city with which to begin the systematic sampling. We use the `sample` function to choose one city from the from the first $k = 6$ cities in the population. (Note that the first $6$ cities are on rows $2$ to $7$.)

```
sample(2:7, size = 1)
```

3

We will begin our systematic sampling with the city on row $3$ (Eastvale). Since the sampling interval is $k = 6$, we sample every $6$th city after Eastvale. The next city to be sampled is Corona on row $3 + 6 = 9$. After Corona is the city on row $9 + 6 = 15$, Jurupa Valley. Next is Norco on row $15 + 6 = 21$. The last city to be included in the sample is San Jacinto on row $21 + 6 = 27$.

---

## Stratified Sampling

To choose a **stratified sample**, divide the population into groups called strata, then randomly choose a **proportionate** number of members from each stratum. For example, if a certain stratum consists of 27% of the population, the sample should have 27% of its members taken from that strata. Usually, strata are chosen to guarantee that certain important characteristics of the population are present in the sample. For instance, imagine a pollster wants to make sure their sample accurately represents each ethnicity of the population. The pollster can do this using stratified sampling where the strata are the different ethnicities of the population. If the population is 13.4% African American and 5.9% Asian, then the pollster's stratified sample will be 13.4% African American and 5.9% Asian.

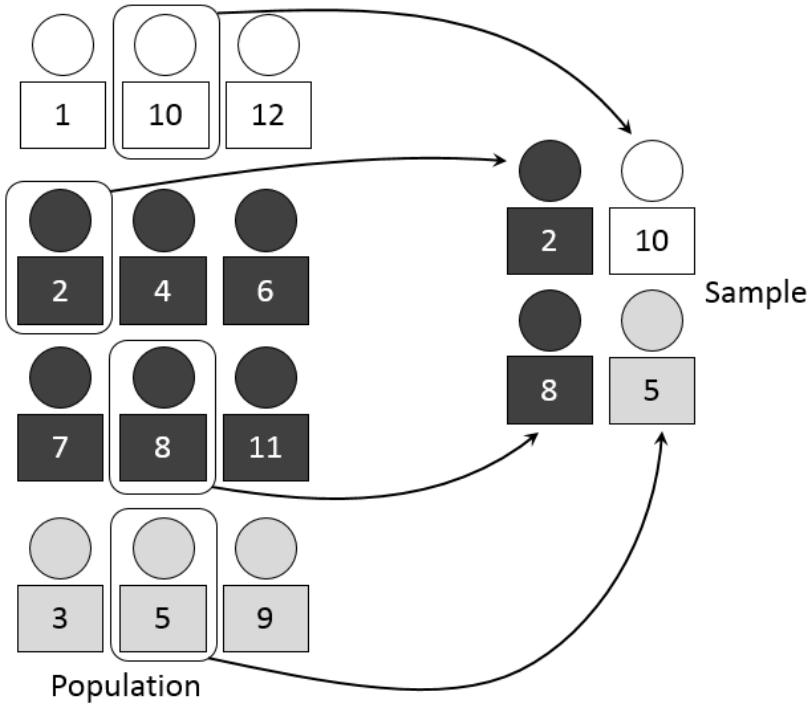



*Fig. 1.3.4* Four individuals are chosen from a population of twelve individuals by dividing the population into strata (groups) based on color, then sampling a proportionate number of individuals from each strata. In this example, one individual is sampled from the white stratum, two individuals from the black stratum, and one individual from the gray stratum. Twice as many individuals were sampled from the black stratum than from the other strata because the black stratum is twice as large as the other strata.[4]

To better illustrate stratified sampling, suppose Professor Baldwin was to choose six students from his precalculus class for a study group, but wants students with an A, students with a B, and students with a C or lower to be proportionately represented in the study group. He divides the 30 students in the class into three different strata based on their grade, as shown in Table 1.3.6. (Note that we have assigned each student a new number in the stratum they are in. This will make it easier to randomly sample students in each stratum later.)

*Table 1.3.6* Students are divided into strata based on their grade in the course.

| Grade A Students | | Grade B Students | | Grade C or Lower Students | |
|---|---|---|---|---|---|
| **Number in Stratum** | **Name** | **Number in Stratum** | **Name** | **Number in Stratum** | **Name** |
| 1 | Bautista | 1 | Anselmo | 1 | Cuningham |
| 2 | Cheng | 2 | Bayani | 2 | Macierz |
| 3 | Hong | 3 | Cuarismo | 3 | Motogawa |
| 4 | Hoobler | 4 | Fontecha | 4 | Rowell |
| 5 | Jiao | 5 | Khan | 5 | Wood |
| 6 | Lundquist | 6 | Legeny | | |
| 7 | Patel | 7 | Okimoto | | |
| 8 | Salangsang | 8 | Price | | |
| 9 | Stratcher | 9 | Quizon | | |
| 10 | Tran | 10 | Reyes | | |
| | | 11 | Roquero | | |
| | | 12 | Roth | | |
| | | 13 | Slade | | |
| | | 14 | Tallai | | |
| | | 15 | Wai | | |

There are 10 students with an A out of the 30 total students in the class, so the proportion of students in the class in the “Grade A” stratum is $\frac{10}{30}$. Since the size of the study group is 6, there should be $\frac{10}{30} \times 6 = 2$ students from the “Grade A” stratum in the study group. Similarly, the “Grade B” stratum has 15 students out of the 30 total students in the class, so there should be $\frac{15}{30} \times 6 = 3$ students from the “Grade B” stratum in the study group. And there should be $\frac{5}{30} \times 6 = 1$ student from the “Grade C” or Lower stratum in the study group since 5 of the 30 students in the class have a grade of C or lower.

We can use simple random sampling using the `sample` function to choose the students for the study group in each stratum. We want to choose 2 students from the 10 students in the “Grade A” stratum.

```
# Randomly choose 2 students from the Grade A stratum
sample(1:10, size = 2)
```

1 · 9

So student 1 (Bautista) and student 9 (Stratcher) from the “Grade A” stratum will be in the study group.

Similarly, we choose 3 of the 15 students in the “Grade B” stratum for the study group.

```
# Randomly choose 3 students from the Grade B stratum
sample(1:15, size = 3)
```

2 · 5 · 15

We include student 2 (Bayani), student 5 (Khan), and student 15 (Wai) from the “Grade B” stratum in the study group.

Finally, we want to select 1 of the 5 students in the “Grade C or Lower” stratum for the study group.

```
# Randomly choose 1 student from the Grade C or Lower stratum
sample(1:5, size = 1)
```

2

The last student in the study group is student 2 (Macierz) from the “Grade C or Lower” stratum. (See Table 1.3.7.)

Unlike other sampling methods, stratified sampling guarantees that important characteristics of the population (the characteristics that determine the strata) are properly represented in the sample. However, it can be difficult to implement, since the size of each stratum and which stratum each member of the population belong to must be known beforehand.

**Stratified**

*Table 1.3.*

| Stratum |
| --- |
| Grade A |
| Grade A |
| Grade B |
| Grade B |
| Grade B |
| Grade C Lower |

## Example 1.3.3

An economist wants to survey 400 renters regarding their income and budgets. She wants to make sure her sample accurately reflects the different prices of rental units in the United States, so she will use stratified sampling to collect her sample, where the strata are the different price levels of rental units according to the nationwide data obtained by the Census Bureau shown in Table 1.3.8.

*Table 1.3.8* Occupied Units Paying Rent in the United States in 2019.[5]

| Monthly Rent | Percentage |
| --- | --- |
| Less that $500 | 9.2% |
| $500 to $999 | 34.1% |
| $1,000 to $1,499 | 29.9% |
| $1,500 to $1,999 | 15.2% |
| $2,000 to $2,499 | 6.2% |
| $2,500 to $2,999 | 2.7% |
| $3,000 or More | 2.7% |

How many renters with rents between $1,500 and $1,999 should the economist include in her sample?

Skip to main content

To find the number of renters that should be included in the sample with rents between $1,500 and $1,999, we need to multiply the proportion of renters with rents between $1,500 and $1,999 in the population with the size of the sample. From the table, we can see that the proportion of renters in this strata is 15.2% = 0.152. (Note, when performing mathematics with percentages, we must always first convert percentages to decimals.) Since the economist wants a sample of size 400, we simply multiply 0.152 by 400 to find the number of individuals we should sample from the strata.

```
0.152 * 400
```

60.8

We can't sample exactly 60.8 people, so we round up to 61. Thus, of the 400 renters in the sample, 61 of those renters should pay monthly rents of between $1,500 and $1,999.

---

## Cluster Sampling

To choose a **cluster sample**, divide the population into clusters (groups) and then randomly select some of the clusters. *All* the members from the selected clusters are in the sample.

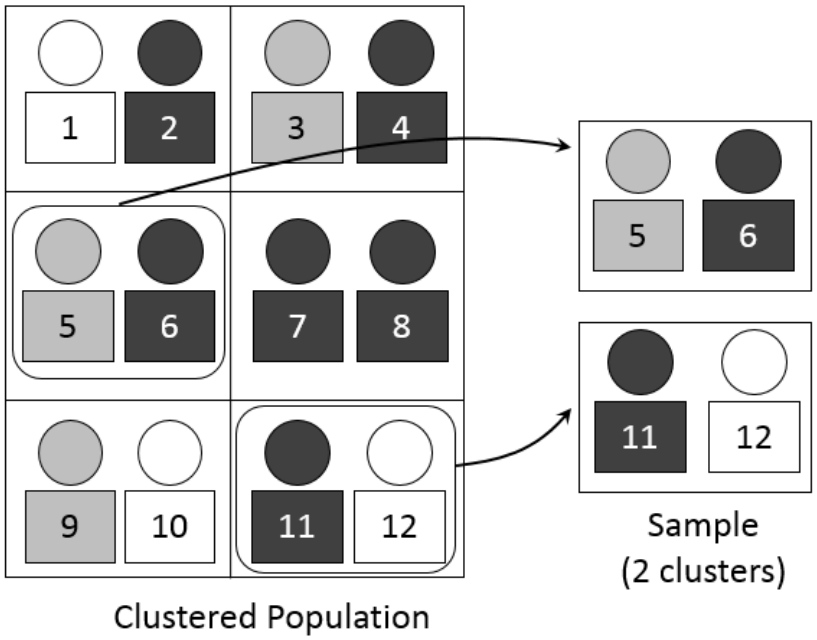



*Fig. 1.3.5* Four individuals are chosen from a population of twelve individuals by dividing the population into six clusters, then randomly selecting two of the clusters. All individuals in the two chosen clusters are sampled. [6]

For example, suppose Professor Baldwin's precalculus students are seated in class at tables in groups of three. (See Figure 1.3.6.) He wants to use cluster sampling to select a six-student study group, where the clusters are the groups of three students at each table.

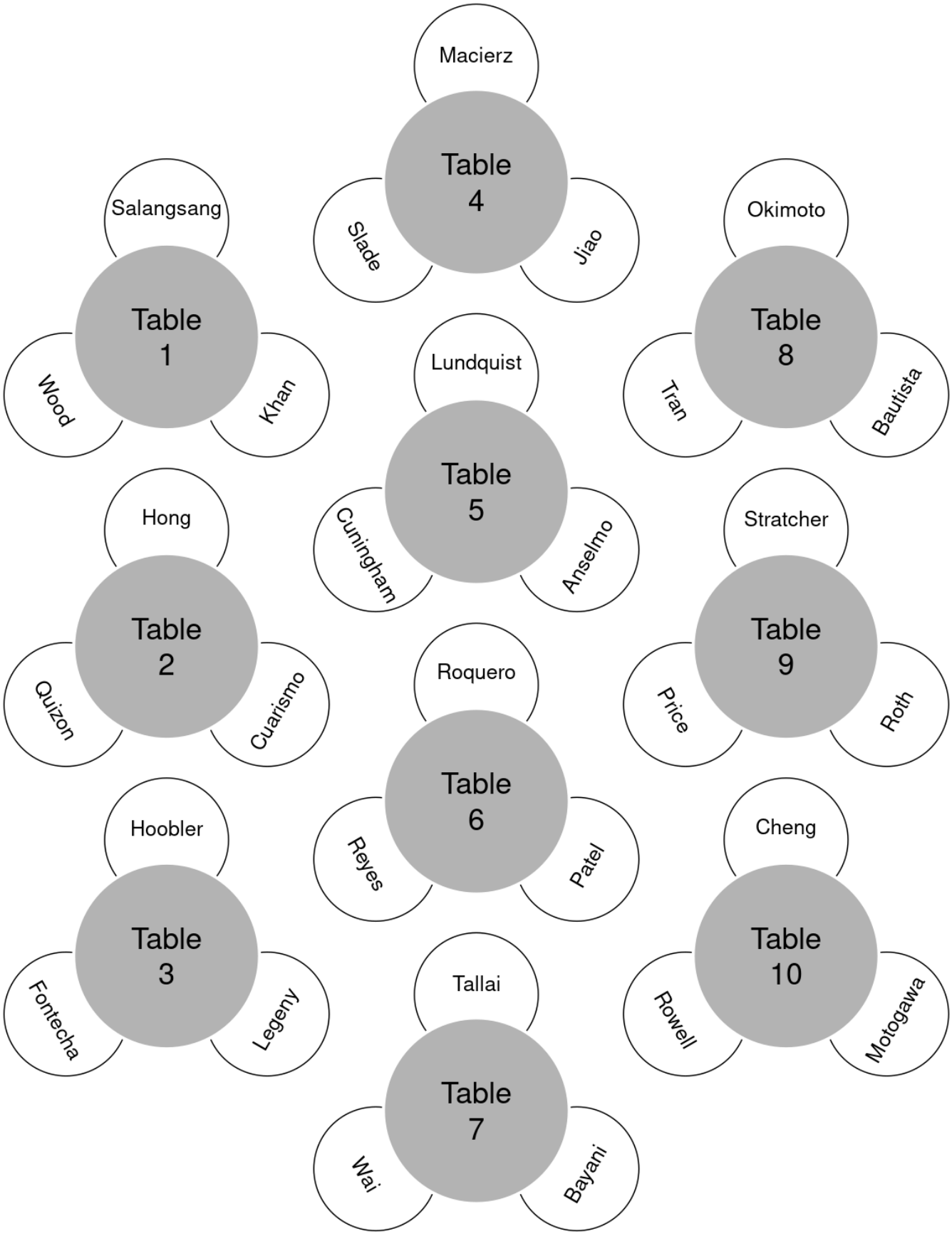



*Fig. 1.3.6* Three students sit at each of the ten tables in Professor Baldwin's precalculus classroom. The students at each table form a cluster.

Skip to main content

Since each cluster has three students, Professor Baldwin will need to randomly select two of the ten tables to create the six-person study group. As with the other sampling methods described above, we can make a random selection with the `sample` function.

```
# Randomly select two of the ten tables
sample(1:10, size = 2)
```

2 · 8

The students in cluster 2 and cluster 8 have been randomly selected to be included in the study group. So *all* the students at table 2 (Hong, Quizon, and Cuarismo) and *all* the students at table 8 (Okimoto, Tran, and Bautista) have been selected for the study group. None of the students at any of the other tables are included in the study group. (See Table 1.3.9.)

Cluster Sa

Cluster sampling can be more convenient to implement than other sampling methods because it often involves sampling members of a population that are close together. Because of this, cluster sampling can be an economical sampling method. However, the clusters must be carefully chosen to reflect the population, or the sample may be biased.

---

### Example 1.3.4

About 130 planes fly out of Ontario International Airport in California each day. Airport management wants to survey customer satisfaction using cluster sampling by sampling the passengers on six different flights on a given day. Use R to randomly choose the flights that should be sampled.

#### Solution

The clusters in this sample are the flights. We will use the `sample` function to randomly choose six flights from the 130 total.

```
sample(1:130, size = 6)
```

36 · 67 · 77 · 79 · 26 · 106

We will sample *all* the passengers on the 36th flight to leave for the day, the 67th flight to leave for the day, the 77th flight to leave for the day, the 79th flight to leave for the day, the 26th flight to leave for the day, and the 106th flight to leave for the day.

---

## Convenience Sampling: A Method to Avoid

A type of sampling that is non-random is **convenience sampling**. Convenience sampling involves using results that are convenient and readily available. For example, a computer software store conducts a marketing study by interviewing potential customers who happen to be in the store browsing through the available software. The results of convenience sampling may be very good in some cases and highly biased (favor certain outcomes) in others. Because the outcome can be biased, convenience sampling should generally be avoided.

For example, suppose Professor Baldwin chooses a six-student study group by conveniently choosing the six students sitting nearest to the front of the classroom. Note that Professor Baldwin's choice is *not* random: students who tend to sit at the back of the classroom did not have an equal chance of being chosen for the study group. The study group is also very likely to be biased: it is well established that students who sit at the front of the classroom tend to do better in the class, which means the study group is likely to have a higher proportion of students with high grades than the class as a whole.

Skip to main content

## Sample Size and Variability

When collecting a sample, the **sample size** is important. The larger the size of the sample, the more likely it is that the sample statistic will accurately approximate the population parameter. A random sample of 10 members of a population is less likely to give a good approximation of the population parameter than a random sample of 100 members. The examples you have seen in this book so far have been small. Samples of only a few hundred observations, or even smaller, are sufficient for many purposes.

Additionally, two or more samples from the same population, taken randomly, and having close to the same characteristics of the population will likely be different from each other. This **sample variability** is completely natural. Suppose Doreen and Jung both decide to study the average amount of time students at their college sleep each night. Doreen and Jung each use simple random sampling to sample 500 students. Doreen's sample will be different from Jung's sample. Neither sample is wrong, but purely by chance, Doreen will sample students that Jung doesn't sample, and Jung will sample students that Doreen doesn't sample. So we shouldn't be surprised if the students in Doreen's sample sleep an average of 7.38 hours per night, while the students in Jung's sample sleep an average of 7.41 hours per night. While the results are different, both statistics are likely a good approximation of the population average.

If Doreen and Jung took larger samples (i.e. the number of data values is increased), their sample statistics (the average amount of time a student sleeps) might be closer to the actual population average. But still, their samples would be, in all likelihood, different from each other.

---

**[1]** Figure 1.3.1 was created by Dan Kernler and is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

**[2](1,2)** Riverside County data on 2019 city revenues per capita obtained from https://data.ca.gov/dataset/city-revenues-per-capita

**[3]** Figure 1.3.3 was created by Dan Kernler and is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

**[4]** Figure 1.3.4 was created by Dan Kernler and is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

**[5]** The data in Table 1.3.8 is from the United States Census Bureau's 2019 American Community Survey 1-Year Estimates.

**[6]** Figure 1.3.5 was created by Dan Kernler and is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

# 1.4. Tables and Charts

## Objectives

- Create and interpret frequency tables.
- Create and interpret bar charts, histograms, and pie charts.

## Frequency

Once you have a set of data, you will need to organize it so that you can analyze how frequently each datum occurs in the set. The **frequency** of a value is the number of times the value occurs in the data.

For example, imagine twenty students in an organic chemistry class obtain the following grades:

By counting the number of times each value occurs, we can construct a **frequency table** for the data, shown in Table 1.4.1.

*Table 1.4.1* The frequency with which each grade is obtained in an organic chemistry class. The frequency of a value is the number of times the value occurs in a data set.

| Grade | Frequency |
|---|---|
| A | 5 |
| B | 7 |
| C | 5 |
| D | 1 |
| F | 2 |
| | **Total = 20** |

By inspecting the table, we see, for example, that 7 students received a B in Organic Chemistry.

It is sometimes more useful to know how many times a particular value occurs in comparison to the total number of values in the data. (A surgeon who has performed 99 surgeries out of 100 successfully may be a good surgeon. A surgeon who has performed 99 surgeries out of 1,000 successfully may not be a good surgeon.) A **relative frequency** is the ratio (fraction or proportion) of the number of times a value of the data occurs to the total number of outcomes. To find the relative frequencies, divide each frequency by the total number values. Relative frequencies can be written as fractions, percents, or decimals.

For the example of the twenty-student organic chemistry class above, the **relative frequency table** is provided in Table 1.4.2.

*Table 1.4.2* The frequency and relative frequency with which each grade is obtained in an organic chemistry class. The relative frequency for each value is obtained by dividing the frequency of the value by the total size of the data set.

| Grade | Frequency | Relative Frequency |
|---|---|---|
| A | 5 | $\frac{5}{20} = 0.25$ |
| B | 7 | $\frac{7}{20} = 0.35$ |
| C | 5 | $\frac{5}{20} = 0.25$ |
| D | 1 | $\frac{1}{20} = 0.05$ |
| F | 2 | $\frac{2}{20} = 0.10$ |
| | **Total = 20** | **Total = 1.00** |

Note that the sum of the relative frequencies should always add up to 1.00 or 100% since the table accounts for all (that is, 100%) of the data values.

We could also **group** data and calculate the frequency and relative frequency for each group. For example, Table 1.4.3 represents the heights, in inches, of a sample of 150 male semiprofessional soccer players.

*Table 1.4.3* The frequency and relative frequency of soccer player heights.

| Height (Inches) | Frequency | Relative Frequency |
|---|---|---|
| 60-62 | 8 | $\frac{8}{150} = 0.0533$ |
| 62-64 | 4 | $\frac{4}{150} = 0.0267$ |
| 64-66 | 23 | $\frac{23}{150} = 0.1533$ |
| 66-68 | 60 | $\frac{60}{150} = 0.4000$ |
| 68-70 | 25 | $\frac{25}{150} = 0.1667$ |
| 70-72 | 18 | $\frac{18}{150} = 0.1200$ |
| 72-74 | 10 | $\frac{10}{150} = 0.0667$ |
| 74-76 | 2 | $\frac{2}{150} = 0.0133$ |
| | **Total = 150** | **Total = 1.00** |

So, for example, a soccer player in the sample that is 64.5 inches tall will be counted in the group of athletes with heights between 64-66 inches.

---

## Example 1.4.1

Adrian asked 75 instructors whether they preferred teaching face-to-face, online, or hybrid. His results can be found in the incomplete frequency table below. Complete the table.

*Table 1.4.4* An incomplete frequency table showing what types of classes 75 instructors prefer to teach.

| Type of Class | Frequency | Relative Frequency |
|---|---|---|
| Face-to-Face | 42 | |
| Online | | |
| Hybrid | 7 | |

### Solution

Let's first find the missing frequency. We know there are 75 total instructors in the sample, and each instructor is represented in exactly one of the three categories. So all the frequencies added together, including the missing frequency (let's call it $x$), should

Skip to main content

$$42 + x + 7 = 75.$$

Solving for $x$, we get

$$x = 75 - 42 - 7 = 26.$$

So there are 26 instructors who prefer teaching online.

Now that we have all the frequencies, we can find the relative frequency of each category. To do so, we divide the frequency of each category by the total number of instructors in the sample. We can use R to do the calculations.

```
# Relative Frequency for Face-to-Face Modality
42/75

# Relative Frequency for Online Modality
26/75

# Relative Frequency for Hybrid Modality
7/75
```

0.56
0.346666666666667
0.0933333333333333

Table 1.4.5 is the completed table. The values we added to the table are in bold, and we rounded the relative frequencies to four decimal places.

*Table 1.4.5* A completed frequency table showing what types of classes 75 instructors prefer to teach.

| Type of Class | Frequency | Relative Frequency |
|---|---|---|
| Face-to-Face | 42 | **0.56** |
| Online | **26** | **0.3467** |
| Hybrid | 7 | **0.0933** |

## Bar Charts

Once you have organized your data, it can be useful to present the data graphically. For qualitative or categorical data, **Bar charts** or **bar graphs** can be used to visualize the frequency or relative frequency of each category. In a bar chart, each category has a rectangular bar associated with it, where the height of the bar (or the width of the bar, for a horizontal bar chart) represents the frequency or relative frequency of the category. The bars in a bar chart are separated by a small space or gap.

### Example 1.4.2

An elementary school teacher records for one week how many students were absent each day. Figure 1.4.1 below summarizes the data she obtained. Use the bar chart to answer the following questions.

Skip to main content

1. On which day were the most students absent?
2. How many students were absent on Wednesday?
3. What percentage of absences occurred on Tuesday?

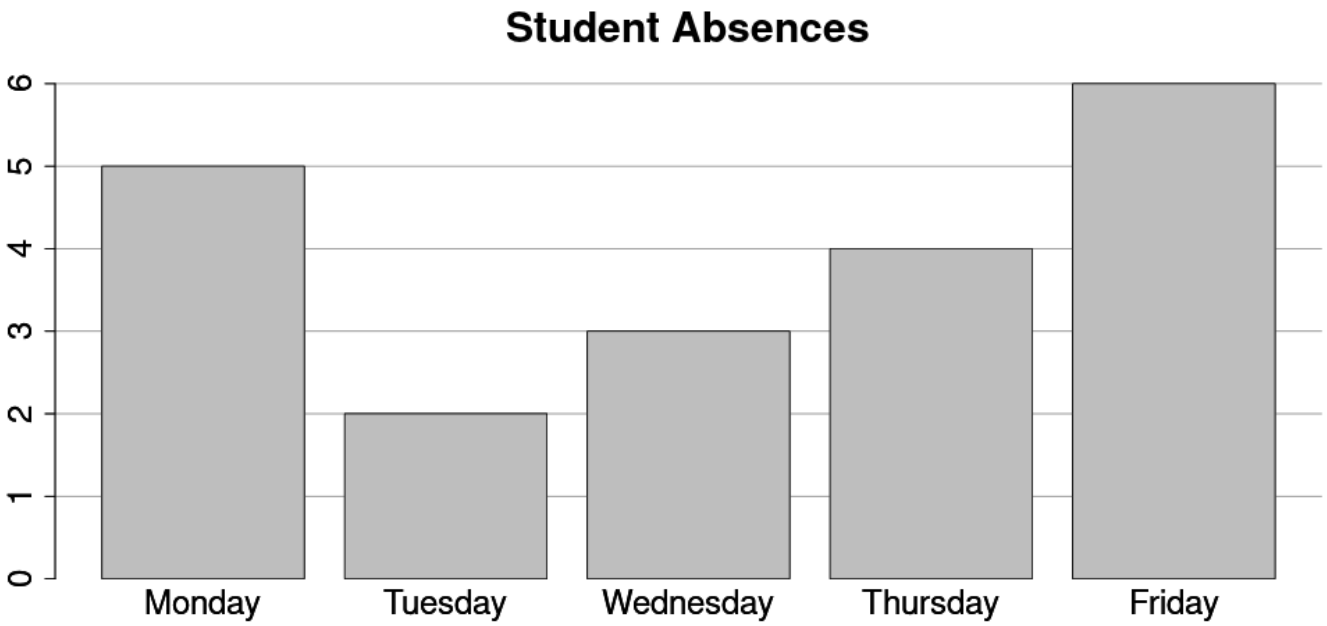



*Fig. 1.4.1* A bar chart illustrating the number of absences in an elementary school class each day for one week.

Solution

Part 1

Friday has the highest bar on the bar chart, so the most students were absent on Friday.

Part 2

The bar for Wednesday has a height of 3, meaning 3 students were absent on Wednesday.

Part 3

To find what *percentage* of absences occurred on Tuesday, we must first find what *fraction* of absences occurrred on Tuesday. Looking at the bar chart, we see that 2 absences occurred on Tuesday. We need to divide this by the total number of absences that occurred during the week. There were 5 absences on Monday, 2 on Tuesday, 3 on Wednesday, 4 on Thursday, and 6 on Friday, for a total of

$$5 + 2 + 3 + 4 + 6 = 20$$

absences that occurred during the week. So the *fraction* of absences that occurred on Tuesday is

Now, to calculate the *percentage* of absences, we multiply this fraction by 100%.

```
2/20 * 100
```

10

So 10% of the week's absences occurred on Tuesday.

---

We can use R to construct a bar chart for us by specifying both the categories of the data and the frequencies (or relative frequencies) of those categories. To do so, we use the `barplot` function:

```
barplot(height, names)
```

Here, `height` is a list of the heights (the frequencies or relative frequencies) of the bars, and `names` is the corresponding list of categories that will be used to label the bars. The order of the `height` list must correspond to the order of the `names` list so that the bar with the first height is labeled with the first name, the bar with the second height is labeled with the second name, etc.

---

## Example 1.4.3

The population in Park City is made up of children, working-age adults, and retirees. Table 1.4.6 shows the three age groups, the number of people in the city in each age group, and the proportion of people in each age group. Construct a bar graph showing the proportions.

*Table 1.4.6* Population in Park City by age group.

| Age Groups | Number of People | Proportion of Population |
|---|---|---|
| Children | 67,059 | 19.11% |
| Working-age Adults | 152,198 | 43.37% |
| Retirees | 131,662 | 37.52% |

**Note**

Table 1.4.6 is a relative frequency table even though the table uses different labels for the frequency and relative frequency columns than we saw in the other examples. The frequency column is labeled "Number of People". The relative frequency column is labeled "Proportion of Population", and relative frequencies are given in percents instead of fractions or decimals.

### Solution

In this example, we have three categories of age groups ("Children", "Working-age adults", "Retirees"). We are given the frequency of each age group in the "Number of people" column and the relative frequency in the "Proportion of the population" column. We will use R to construct a bar graph of the relative frequencies of each category.

```r
proportions = c(19.11, 43.37, 37.52)
age_groups = c("Children", "Working-age Adults", "Retirees")

barplot(height = proportions, names = age_groups)
```

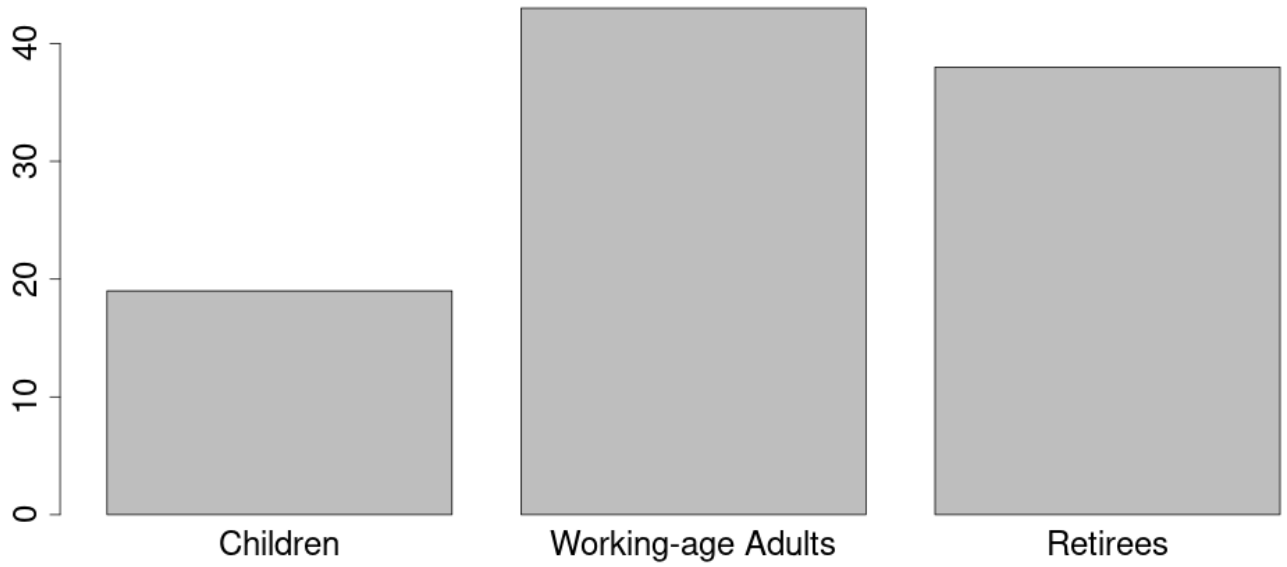



*Fig. 1.4.2* A bar chart illustrating the proportion of each age group relative to the population of Park City.

**Warning**

Note that the labels “Children”, “Working-age Adults”, and “Retirees” in the code block in this example are each in double quotes. Double quotes surrounding text tells R that the text is just a label, not instructions the computer should follow. If you don’t include the double quotes, R will return an error. When creating a list of `names` for a bar plot:

- Make sure *each* label in the list is contained within double quotes.
- Make sure *only* the label is in the double quotes. In particular, the comma separating the labels should be outside the double quotes.

For the example above, here are ways to define the labels for the bar plot correctly and incorrectly in R.

**Incorrect**

```
age_groups = c("Children, Working-age Adults, Retirees")
```

**Why?**

All labels are between the same set of double quotes. Each label should have its own set of double quotes.

**Incorrect**

```
age_groups = c("Children," "Working-age Adults," "Retirees")
```

**Why?**

The commas separating the labels are inside the double quotes, which tells R that the commas are part of the labels. But we actually want the commas to separate the labels. The commas should be outside the double quotes.

**Correct**

```
age_groups = c("Children", "Working-age Adults", "Retirees")
```

**Why?**

Each label has its own set of double quotes, and only the labels are inside the double quotes. The commas are outside the double quotes.

## Example 1.4.4

The columns in the table below contain: the race or ethnicity of students in U.S. Public Schools for the class of 2011, percentages for the Advanced Placement examine population for that class, and percentages for the overall student population. Create a bar

graph with the student race or ethnicity (qualitative data) on the x-axis, and the Advanced Placement examinee population percentages on the y-axis.

*Table 1.4.7* Race/Ethnicity of AP Examinee Population vs. Overall Student Population. Note a key is provided which assigns a letter to each race/ethnicity. This makes it easier to reference the long race/ethnicity names.

| Race/Ethnicity | AP Examinee Population | Overall Student Population |
|---|---|---|
| A = Asian, Asian American, or Pacific Islander | 10.3% | 5.7% |
| B = Black or African American | 9.0% | 14.7% |
| C = Hispanic or Latino | 17.0% | 17.6% |
| D = American Indian or Alaskan Native | 0.6% | 1.1% |
| E = White | 57.1% | 59.2% |
| F = Not Reported/Other | 6.0% | 1.7% |

**Note**

Table 1.4.7 does not include a frequency column. Instead, it provides the relative frequencies (as percents) of two different populations: the AP examinee population and the overall student population.

### Solution

This table provides us with the relative frequecy of each ethnicity in two different populations: the population of all AP examinees and the overall student population. In this case, we will only use the data in the “AP Examinee Population” column to construct our bar chart; we will not use the “Overall Student Population” data in this example.

```
APproportion = c(10.3, 9.0, 17.0, 0.6, 57.1, 6.0)
race = c("A", "B", "C", "D", "E", "F")

barplot(height = APproportion, names = race)
```

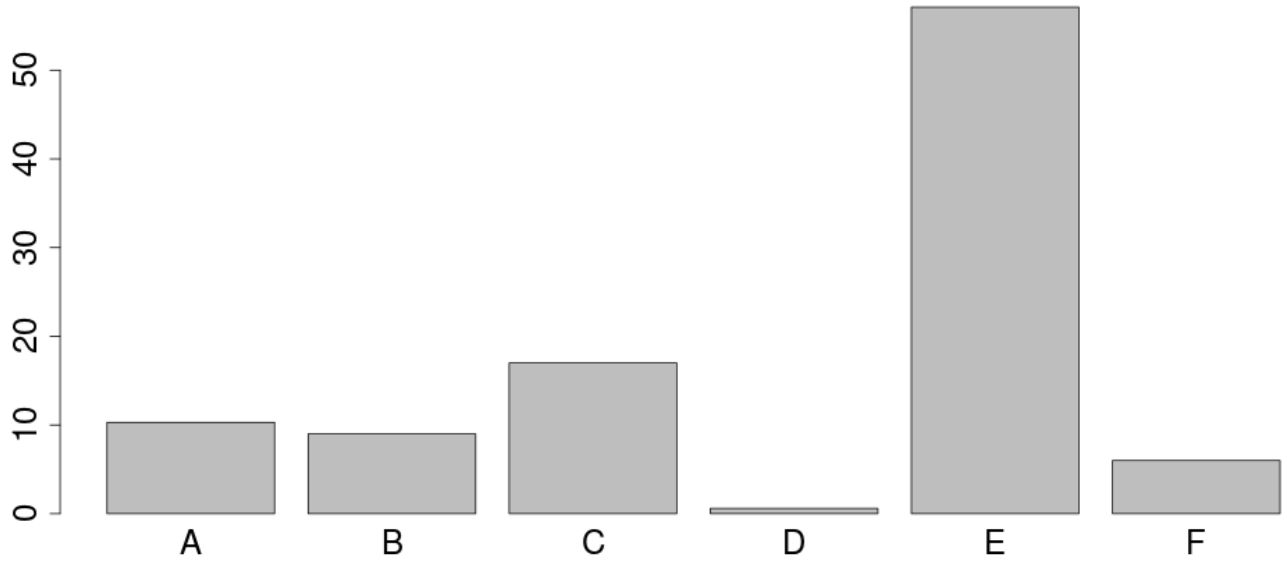



*Fig. 1.4.3* A bar chart illustrating the proportion of each race/ethnicity in the AP examinee population.

Because the names of the races/ethnicities are so long, there isn't enough room on the bar chart to use the actual names of the races/ethnicities as the bar labels. Instead, we use a key for the labels ("A" for "Asian, Asian American, or Pacific Islander", "B" for "Black or African American", "C" for "Hispanic or Latino", etc.).

---

## Pie Charts

Qualitative data can also be visualized using a pie chart. In a **pie chart**, a circle is sliced (like a pie) so that the number of categories in the data is equal to the number of slices, and the size of each slice is proportional to the corresponding category's quantity.

We can create a pie chart in R using the `pie` function:

```
pie(x, labels)
```

Here, `x` is a list with a frequency (or relative frequency) for each category, and `labels` is a list with a label for each category. The two lists must be in the same order.

The `x` and `labels` variables in the `pie` function are equivalent, respectively, to the `height` and `names` variables in the `barplot` function.

Let's revisit the population of Park City in Example 1.4.5, but this time construct a pie chart of the data.

---

### Example 1.4.5

The population in Park City is made up of children, working-age adults, and retirees. Table 1.4.8 shows the three age groups, the

Skip to main content

proportions.

*Table 1.4.8* Population in Park City by age group.

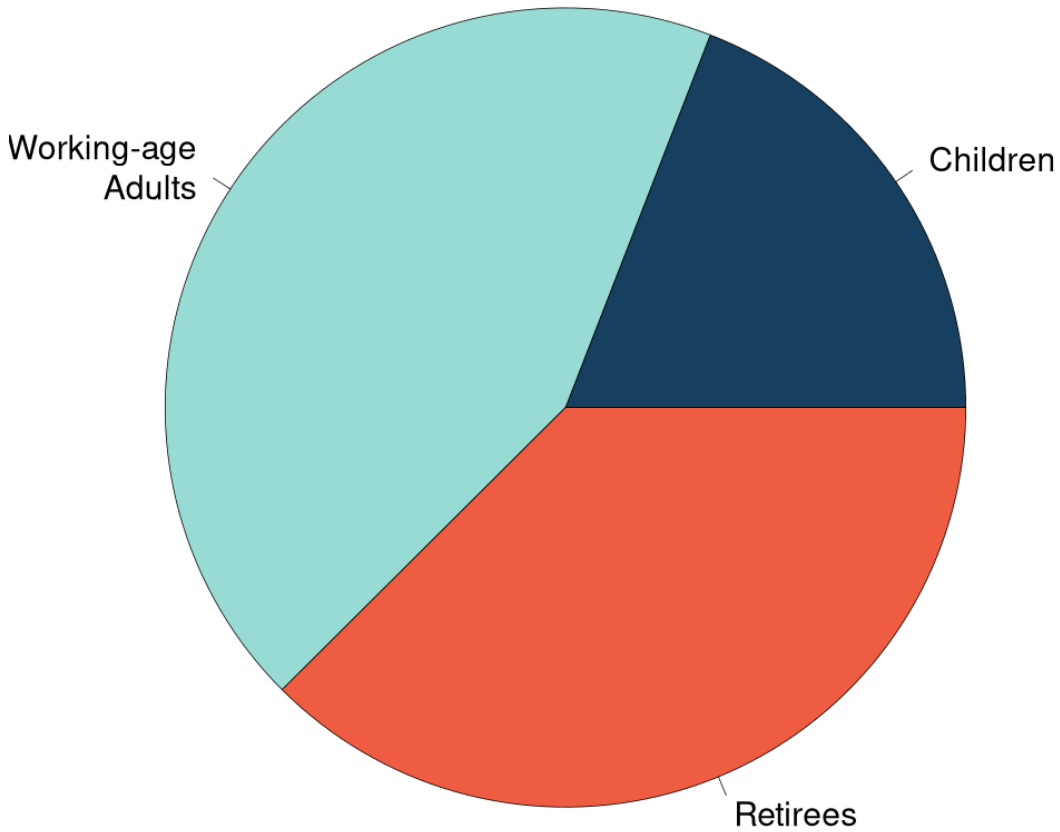
| Age Groups | Number of People | Proportion of Population |
|---|---|---|
| Children | 67,059 | 19.11% |
| Working-age Adults | 152,198 | 43.37% |
| Retirees | 131,662 | 37.52% |

Solution

We use R to construct a pie chart.

```
proportions = c(19.11, 43.37, 37.52)
age_groups = c("Children", "Working-age Adults", "Retirees")

pie(proportions, labels = age_groups)
```

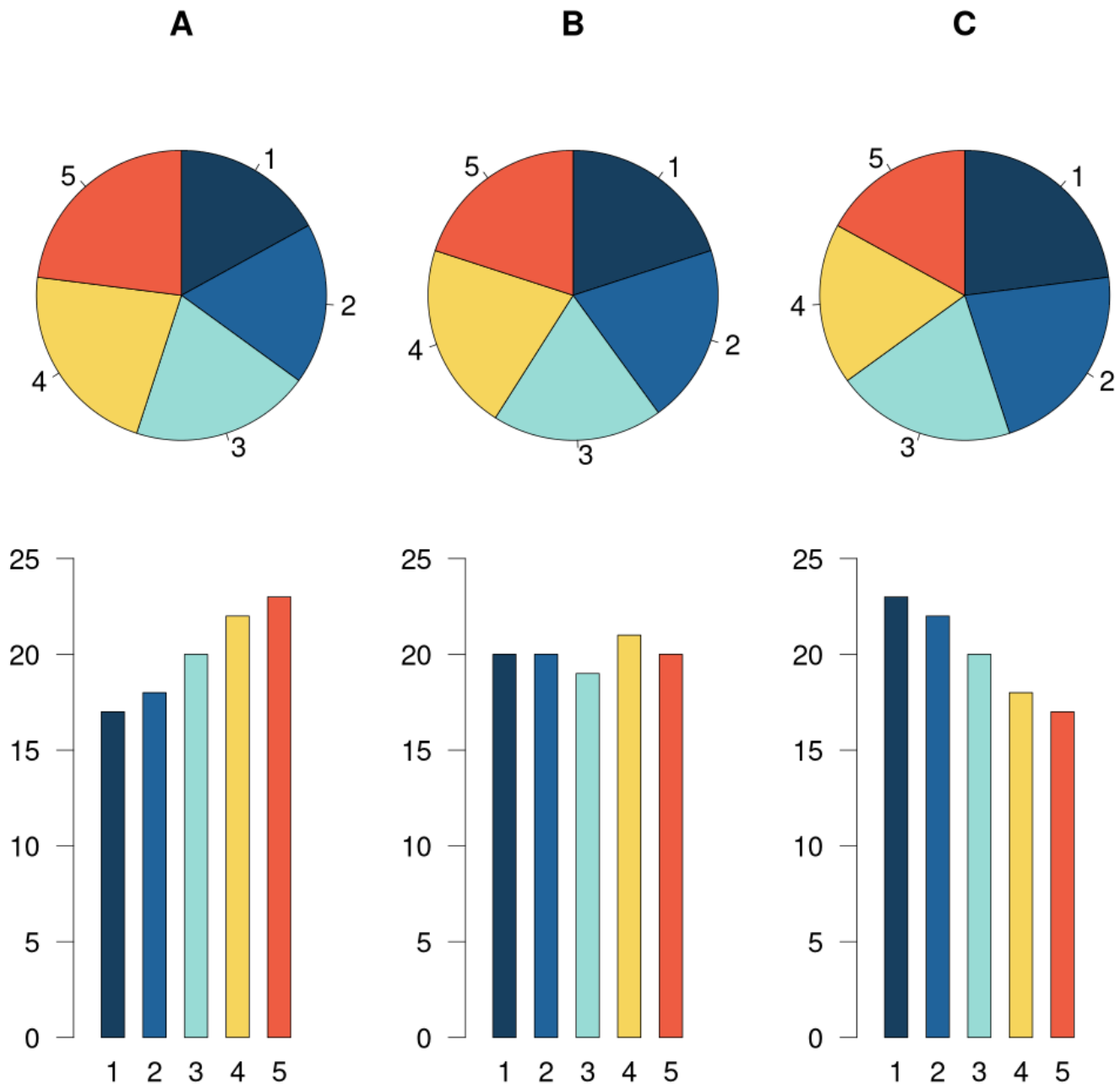



Skip to main content

*Fig. 1.4.4* A pie chart illustrating the proportion of each age group relative to the population of Park City.

Compare the bar chart in Example 1.4.5 with the pie chart in Example 1.4.7. Which do you think is easier to read? Notice that it is a little difficult to decide if the “working-age adults” category or the “retirees” category is larger with the pie chart, but it is immediately obvious which category is larger using the bar chart. This isn’t just a quirk with this data: bar charts tend to be easier to read than pie charts.

As another example, consider the pie charts and bar charts below. Notice how hard it is to compare the sizes of the categories in the pie charts; notice how much easier it is to compare the sizes of the categories in the corresponding bar charts.




Skip to main content

*Fig. 1.4.5* Three examples comparing pie charts and bar charts illustrating the same data. In each case, the qualitative and quantitative features of the data are easier to interpret using the bar chart.[1]

While pie charts are still commonly used in media and in business, most visualization experts generally recommend against using pie charts in favor of other visualization tools like bar charts.

## Histograms

We apporach graphs for quantitative data similarly, though with a few important distinctions.

First, with quantitative data, it is often up to the statistician how to group or categorize the data. For example, if we are gathering data on eye color (qualitative data), we naturally have a category of blue eyes and a category of brown eyes; there is no opportunity to adjust the categories. But if we are gathering data on the heights of adult males (quantitative data), the statistician could partition the data so they get a group of individuals 50-59 inches tall, or they could partition the data so they get a group of individuals 50-54 inches tall and another group of individuals 55-59 inches tall.

Second, we can measure how far apart quantitative data is. It doesn't make sense to ask how far apart blue eyes are from brown eyes. But it does make sense to ask how far apart 50 inches is from 59 inches. (They are 59 - 50 = 9 inches apart.)

A common way to visualize quantitative data is to use a histogram. A **histogram** is similar to a bar chart, but it is better at displaying quantitative data. Unlike a bar chart, a histogram consists of contiguous (adjoining) boxes so there is no gap between the bars. This helps emphasize the "distance" between quantitative data. A histogram has both a horizontal axis and a vertical axis. The horizontal axis is labeled with what the data represents (for instance, distance from your home to school). The vertical axis is labeled either frequency or relative frequency (or percent frequency or probability). The histogram can suggest shape of the data and can suggest where the center of the data is and how spread out the data is.

---

### Example 1.4.6

An economist surveys gas stations in the Bay Area. Figure 1.4.6 below summarizes the gas prices at the gas stations he surveyed. Use the histogram to answer the following questions.

1. About how many gas stations had gas prices above $4.70?
2. About what percentage of gas stations surveyed have gas prices between $4.40 and $4.60?

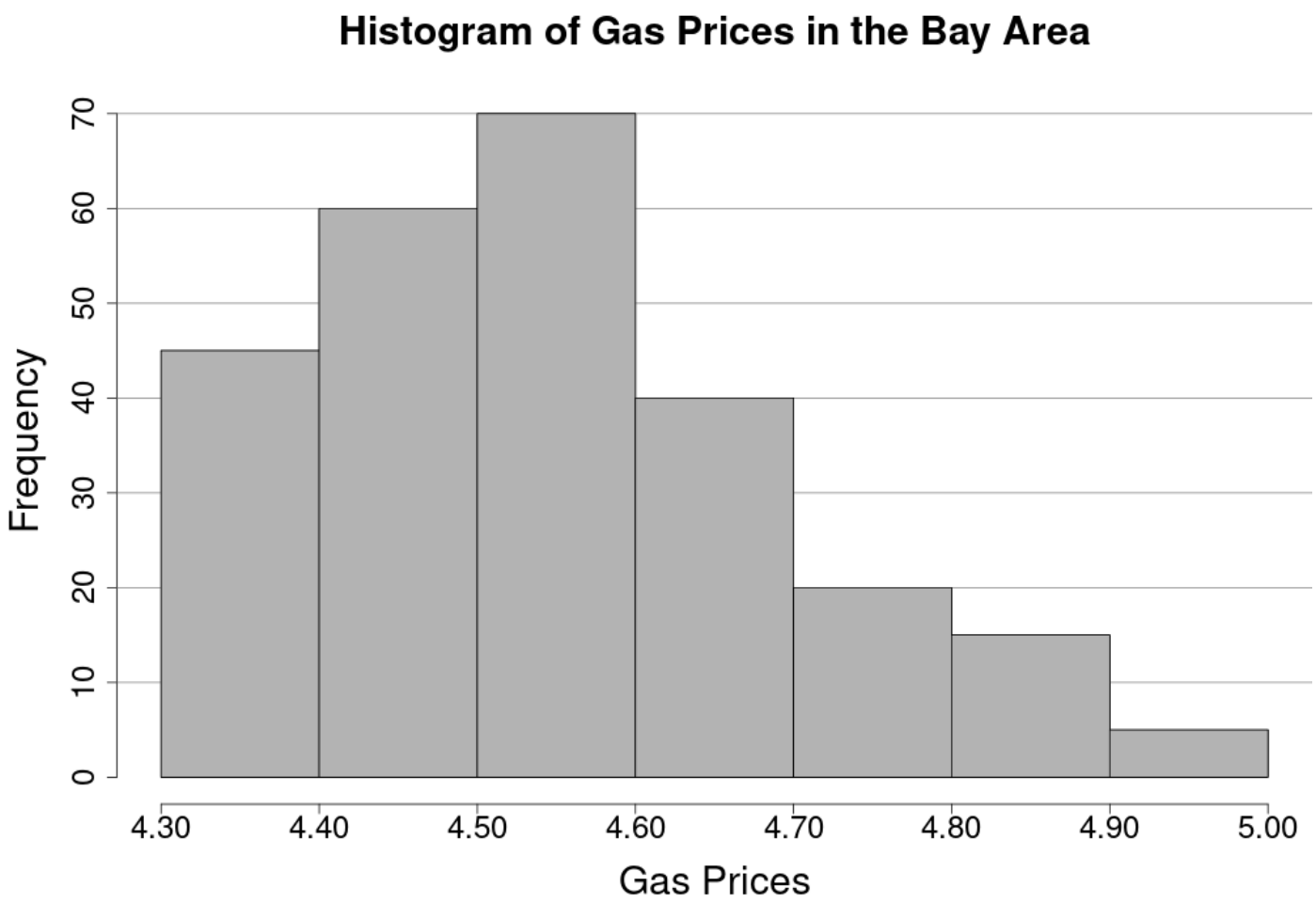



*Fig. 1.4.6* A histogram of gas prices in the Bay Area. Note there is no gap in between the bars, and each bar represents a range of values instead of a single category or number.

Solution

Part 1

Looking at the histogram, we can see that there are 20 gas stations with prices between $4.70 and $4.80, there are 15 gas stations with prices between $4.80 and $4.90, and there are 5 gas stations prices between $4.90 and $5.00. So there are about 20 + 15 + 5 = 40 gas stations with prices above $4.70.

Part 2

To find the *percentage* of gas stations with prices between $4.40 and $4.60, we must first find the *fraction* of gas stations with prices between $4.40 and $4.60.

We start by finding the *number* of gas stations with prices between $4.40 and $4.60: there are 60 gas stations with prices between $4.40 and $4.50 and 70 gas stations with prices between $4.50 and $4.60, meaning there are a total of

$$60 + 70 = 130$$

Skip to main content

gas stations with prices between $4.40 and $4.60.

To find the *fraction* of gast stations with prices between $4.40 and $4.60, we need to divide the number of gas stations with prices between $4.40 and $4.60 (which we just calculated is 130) by the total number of gas stations surveyed. We can calculate the total number of gas stations surveyed by adding up the number of gas stations in each range of prices on the histogram. There are:

- 45 gas stations with prices between $4.30 and $4.40
- 60 gas stations with prices between $4.40 and $4.50
- 70 gas stations with prices between $4.50 and $4.60
- 40 gas stations with prices between $4.60 and $4.70
- 20 gas stations with prices between $4.70 and $4.80
- 15 gas stations with prices between $4.80 and $4.90
- 5 gas stations with prices between $4.90 and $5.00

So there were

$$45 + 60 + 70 + 40 + 20 + 15 + 5 = 255$$

gas stations surveyed in total.

Then the *fraction* of gas stations surveyed with prices between $4.40 and $4.60 is

$$130/255.$$

To find the *percentage* of gas stations with prices between $4.40 and $4.60, we multiply the fraction by 100%.

```
130/255 * 100
```

50.9803921568627

So about 50.9804% of gas stations surveyed had prices between $4.40 and $4.60.

---

We can use R to construct a histogram from a data set. To construct a histogram in R, use the `hist` function:

```
hist(x)
```

Here, `x` is the list of data we want a histogram of.

---

## Example 1.4.7

The following data are the heights (in inches to the nearest half inch) of 100 male semiprofessional soccer players. The heights are continuous data, since height is measured.

60, 60.5, 61, 61, 61.5, 63.5, 63.5, 63.5, 64, 64, 64, 64, 64, 64, 64, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 64.5, 66, 66, 66, 66, 66, 66, 66, 66, 66, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 66.5, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67, 67.5, 67.5, 67.5, 67.5, 67.5, 67.5, 68, 68, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69, 69.5, 69.5, 69.5, 69.5, 69.5, 70, 70, 70, 70, 70, 70,

Construct a histogram of the data. Where do the data seem to be concentrated?

Solution

```
data = c(60, 60.5, 61, 61, 61.5, 63.5, 63.5, 63.5, 64, 64, 64, 64, 64, 64, 64, 64.5, 64.5, 64.5, 64.5, 64.5,
hist(data)
```

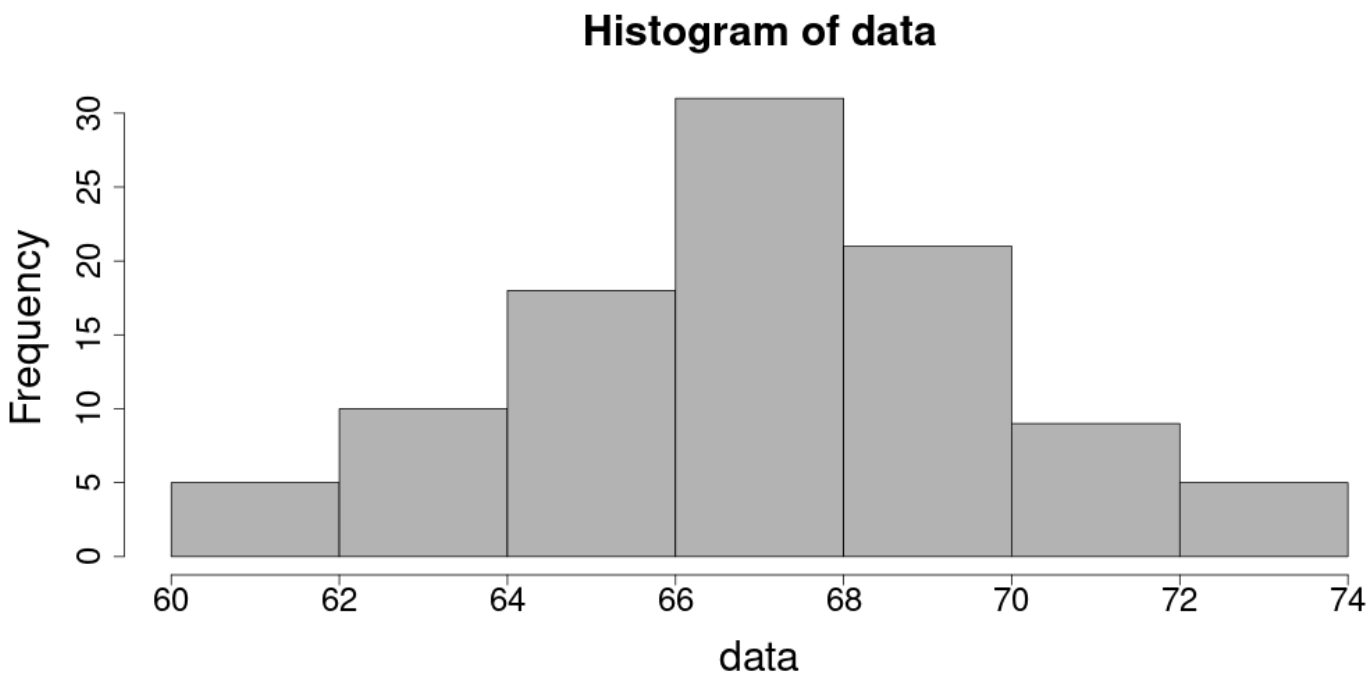



*Fig. 1.4.7* A histogram of the heights of 100 semiprofessional soccer players.

From the histogram, we can see that most of the players' heights are between 64 inches and 70 inches.

**Note**

For histograms in R, each bar represents the frequency of values that lie between the endpoints to left and right of the bar, but *not including* the endpoint to the left and *including* the value to the right. For instance, the bar between 70 and 72 in this example counts only the data 70.5, 70.5, 70.5, 71, 71, 71, 72, 72, 72. So it includes the data points of value 72 (the right endpoint) but not the data points of value 70 (the left endpoint). The one exception to this rule is the bar on the far left, which includes values at both endpoints.

## Example 1.4.8

Create a histogram for the following data: the number of books bought by 50 part-time college students at ABC College. The number of books is discrete data, since books are counted.

0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6

Skip to main content

```
books = c(0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,
hist(books)
```


*Fig. 1.4.8* A histogram of number of books bought by 50 part-time students at ABC college.

**Note**

A histogram may not be the best way to visualize the data in this example. We can see that the bar ranging from 1 to 2 has a height of 10, meaning there are 10 students at ABC college that purchased between 1 and 2 books. But what does that mean? A student can purchase 1 book, or they can purchase 2 books, but they can't purchase 2.7 books. Because of this, a bar chart, where the bars are labeled "0 books", "1 book", "2 books", etc., may be a better way to visualize the data.

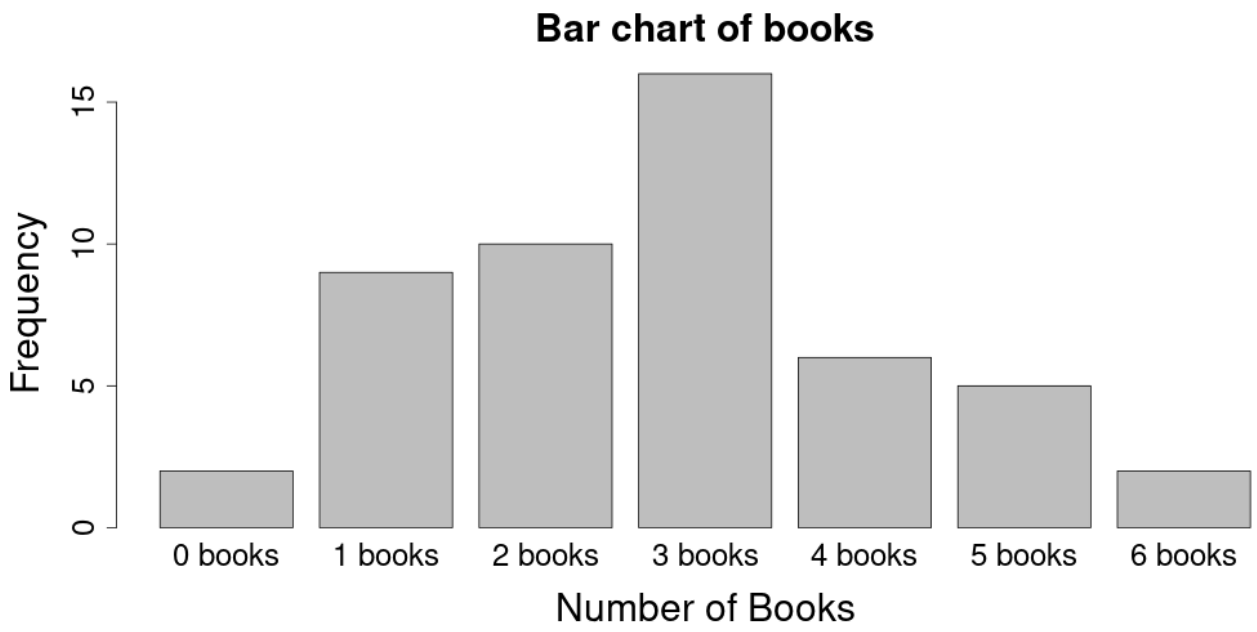



*Fig. 1.4.9* A bar chart of the number of books bought by part-time students at ABC college. Note that *in this case*, a bar chart communicates the data better than a histogram.

This does *not* mean that a bar chart is always better than a histogram. In some situations (like this one), a bar chart may do the best job at communicating the data. In other situations (like for continuous data), a histogram may do the best job at communicating the data. A wise statistician will choose the best tool for the job.

[1] Figure 1.4.5 was adapted from an image created by Shutz and is licensed under the Creative Commons Attribution 1.0 Generic license.

# 2. Descriptive Statistics

## 2.1. Measures of Position

### Objectives

- Calculate percentiles and quartiles and interpret the results.
- Construct and interpret box plots.

Skip to main content

## Percentiles

A common measure of position is **percentiles**. Percentiles are values that divide ordered data into 100 equally sized groups. To score in the 90th percentile of an exam does not mean, necessarily, that you received 90% on a test. It means that 90% of test scores are the same or less than your score and 10% of the test scores are the same or greater than your test score.

Percentiles are useful for comparing values. For this reason, universities and colleges use percentiles extensively. One instance in which colleges and universities use percentiles is when SAT results are used to determine a minimum testing score that will be used as an acceptance factor. For example, suppose Duke accepts SAT scores at or above the 75th percentile. This means Duke will only accept students with an SAT scores higher than at least 75% of all SAT scores.

Percentiles are mostly used with very large populations. Therefore, if you were to say that 90% of the test scores are less (and not the same or less) than your score, it would be acceptable because removing one particular data value is not significant in a very large population.

To calculate percentiles by hand, the data must be ordered from smallest to largest. But we can use R to quickly find percentiles without ordering the data using the `quantile` function. **Quantiles** are values that divide ordered data into any number of equally-sized groups, so a percentile is a type of **quantile**.

The syntax for the `quantile` function is:

```
quantile(x, probs)
```

In this case, `x` is the list of data that we want the percentiles of and `probs` is the percentage or list of percentages, converted to decimal form, corresponding to our percentiles. The name `probs` is an abbreviation for 'probabilities.' We will learn more about why that abbreviation is appropriate later.

The following example illustrates how to use the `quantile` function to find a percentile.

---

### Example 2.1.1

Listed are 29 ages for Academy Award winning best actors. Find the 73rd percentile. Explain, in words, what this value represents.

18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 59, 62, 65, 67, 71, 72, 73, 74, 76, 77

#### Solution

We will use the `quantile` function in R with `probs = 0.73`, since 73% = 0.73. Note that it is important to use the decimal form of the percentage; if we use `probs = 73`, we will get an error.

```
ages = c(18, 21, 22, 25, 26, 27, 29, 30, 31, 33, 36, 37, 41, 42, 47, 52, 55, 57, 58, 59, 62, 65, 67, 71, 72,

quantile(ages, probs = 0.73)
```

**73%:** 63.32

So the 73rd percentile is 63.32. That means that of the 29 Academy Award winning best actors in our sample, 73% of them were younger than 63.32 years old.

Skip to main content

# Quartiles

**Quartiles** are values that divide ordered data into 4 equally sized groups, called **quarters**. A quartile is another kind of quantile.

- The first quartile, $Q_1$, is the same as the 25th percentile.
- The second quartile, $Q_2$, is the same as the 50th percentile.
- The third quartile, $Q_3$, is the same as the 75th percentile.

The four quarters that we get from the quartiles each contain about the same amount of data values.

- The first quarter consists of the data values less than or equal to $Q_1$ (the 25th percentile). This quarter contains 25% of the data values.
- The second quarter consists of the data values between $Q_1$ (the 25th percentile) and $Q_2$ (the 50th percentile). This quarter contains 25% of the data values.
- The third quarter consists of the data values between $Q_2$ (the 50th percentile) and $Q_3$ (the 75th percentile). This quarter contains 25% of the data values.
- The fourth quarter consists of the data values greater than or equal to$Q_3$ (the 75th percentile). This quarter contains 25% of the data values.

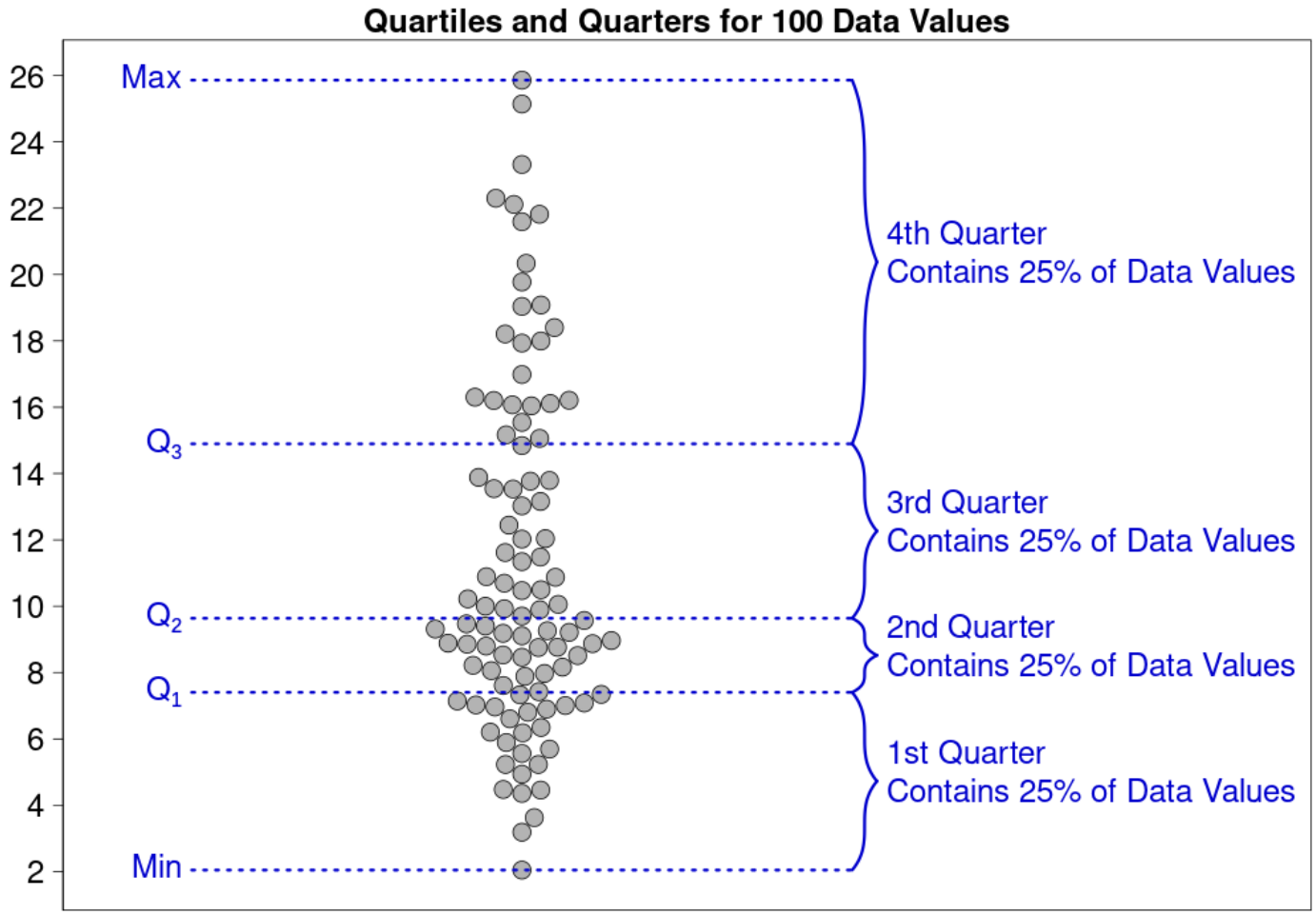



*Fig. 2.1.1* Quartiles and quarters for 100 random data values. The quartiles ($Q_1$, $Q_2$, and $Q_3$) divide the data into quarters. Note that each quarter of the data contains 25% of data values. The data is more spread out in larger quarters and more squished together in smaller quarters, but both large quarters and small quarters contain the same number of data values.

Skip to main content

The second quartile, $Q_2$, is more often called the **median** and represented by the variable $M$. It is a number that measures the "center" of the data. You can think of the median as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. Half the values are the same value or smaller than the median, and half the values are the same value or larger than the median. For example, consider the following data.

1, 11.5, 6, 7.2, 4, 8, 9, 10, 6.8, 8.3, 2, 2, 10, 1

Ordered from smallest to largest:

1, 1, 2, 2, 4, 6, 6.8, 7.2, 8, 8.3, 9, 10, 10, 11.5

Since there are 14 observations, the median is between the seventh value, 6.8, and the eighth value, 7.2. To find the median, average the two values together by adding them and dividing by 2.

$$\frac{6.8 + 7.2}{2} = 7$$

The median is 7. Half of the values are smaller than 7 and half of the values are larger than 7.

Let's confirm the median is 7 using R. Since the median is the 50th percentile, we will use `probs = 0.50`, since 50% = 0.50.

```
data = c(1, 11.5, 6, 7.2, 4, 8, 9, 10, 6.8, 8.3, 2, 2, 10, 1)

quantile(data, probs = 0.50)
```

**50%:** 7

From our R calculation, we see that the median, or the 50th percentile, is 7. This matches the result we obtained from our calculation by hand above. But recall that when we calculated the median by hand, we had to first order the data from smallest to largest. By contrast, we were able to leave the data we plugged into R unordered and let the computer do the hard work.

We can use the `quantile` function allows us to quickly and easily find any quartile, not just the median.

**Note**

Be careful not to confuse the term *quartile* with the term *quantile*. **Quart**iles split the data into four **quart**ers, while **quant**iles split the data using more general **quant**ities.

---

## Example 2.1.2

For the following 13 real estate prices, calculate the quartiles.

$389,950; $230,500; $158,000; $479,000; $639,000; $114,950; $5,500,000; $387,000; $659,000; $529,000; $575,000; $488,800; $1,095,000

### Solution

We will use the `quantile` function to calculate our quartiles.

- To calculate our first quartile $Q_1$, we will use `probs = 0.25`.

Skip to main content

- To calculate our third quartile $Q_3$, we will use `probs = 0.75`.

But notice that our list of real estate prices above isn't ready to be input into R: it won't understand the dollar sign, and we must use a comma to separate values.

389950, 230500, 158000, 479000, 639000, 114950, 5500000, 387000, 659000, 529000, 575000, 488800, 1095000

> **Warning**
>
> R will not allow the use of a comma to mark the thousands place of a number; commas are only used in R to separate values or items in a list. For example, if you type `389,950` into R, it will interpret it as a list of two numbers: 389 and 950. Instead, we omit the comma and type simply `389950`.

```
prices = c(389950, 230500, 158000, 479000, 639000, 114950, 5500000, 387000, 659000, 529000, 575000, 488800,

#First Quartile
quantile(prices, probs = 0.25)

#Median/Second Quartile
quantile(prices, probs = 0.50)

#Third Quartile
quantile(prices, probs = 0.75)
```

**25%:** 387000
**50%:** 488800
**75%:** 639000

So $Q_1 = \$387,000$, $Q_2 = \$488,800$, and $Q_3 = \$639,000$.

So we deduce that:

- About 25% of the real estate prices in our sample are less than $Q_1 = \$387,000$ (in the first quarter of the data).
- About 25% of the real estate prices in our sample are between $Q_1 = \$387,000$ and $Q_2 = \$488,800$ (in the second quarter of the data).
- About 25% of the real estate prices in our sample are between $Q_2 = \$488,800$ and $Q_3 = \$639,000$ (in the third quarter of the data).
- About 25% of the real estate prices in our sample are greater than $Q_3 = \$639,000$ (in the fourth quarter of the data).

The quartiles quartered the data into groups each containing 25% of the data points.

---

## Box Plots

**Box plots** (also called **box-and-whisker plots** or **box-whisker plots**) give a good graphical image of the concentration of the data. They also show how far the extreme values are from most of the data. A box plot is constructed from five values: the minimum value, the first quartile, the second quartile or median, the third quartile, and the maximum value. The box plot illustrates the four quarters of the data and how squished together or spread out the data in each quarter is.

To construct a box plot, use a horizontal or vertical number line and a rectangular box. The smallest and largest data values label the endpoints of the axis. The first quartile marks one end of the box and the third quartile marks the other end of the box.

Skip to main content

**Approximately the middle 50 percent of the data fall inside the box.** The “whiskers” extend from the ends of the box to the smallest and largest data values. The box plot gives a good, quick picture of how the data is spread out.

To illustrate a box plot, consider this small dataset.

1, 2, 2, 3, 5, 7, 9, 9, 11, 12, 15

We can easily find the quartiles for this data using R:

```
data = c(1, 2, 2, 3, 5, 7, 9, 9, 11, 12, 15)

quantile(data, probs = 0.25)
quantile(data, probs = 0.50)
quantile(data, probs = 0.75)
```

**25%:** 2.5
**50%:** 7
**75%:** 10

So $Q_1 = 2.5$, $Q_2 = 7$, and $Q_3 = 10$. We also note that 1 is the minimum value and 15 is the maximum value of the data. Observe that these quartiles and maximum/minimum values are plotted using horizontal lines in the box plot for the data in Figure 2.1.2.

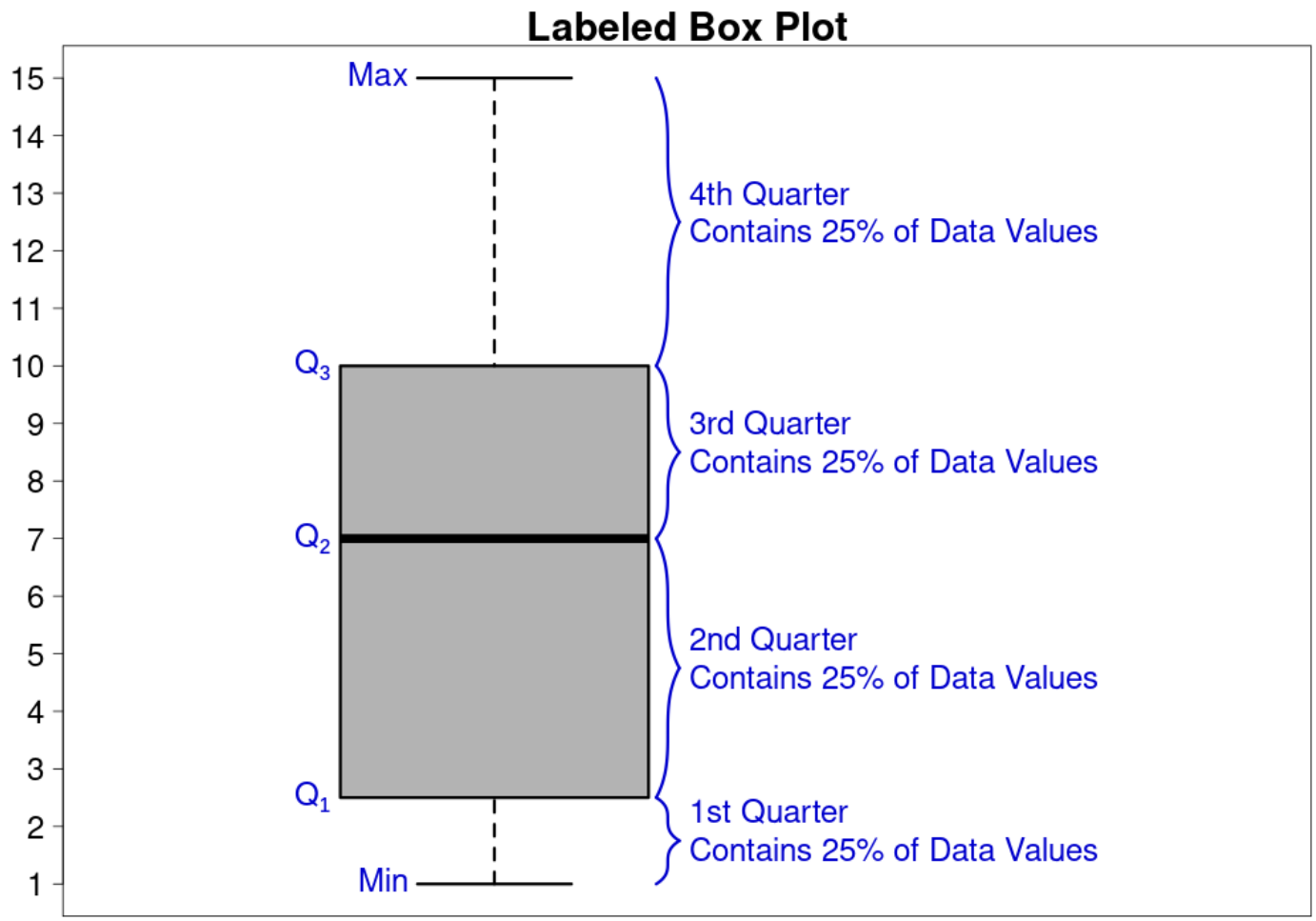



*Fig. 2.1.2* The locations of the quartiles ($Q_1$, $Q_2$, and $Q_3$) and the minimum and maximum values of the data are indicated with horizontal lines. While the data itself is not plotted, we can see how spread out the data is in each quarter. Note that each

quarter of the data contains 25% of data values. The data is more spread out in larger quarters and more squished together in smaller quarters, but both large quarters and small quarters contain the same number of data values.

Note that box plots are different than other graphs like bar graphs. In a bar graph, the height of a bar tells us how many data values are in a particular category. The higher the bar is, the more data values are in that category. But each quarter of a box plot contains the same number of data values. The size of a quarter tells us how spread out the data values in that quarter are. The larger the quarter is, the more spread out the data values are in that quarter.

**Note**

You may encounter box plots that have dots marking especially extreme values. In these cases, the extreme values are left out of the box plot so that the data doesn't look more spread out than it really is.

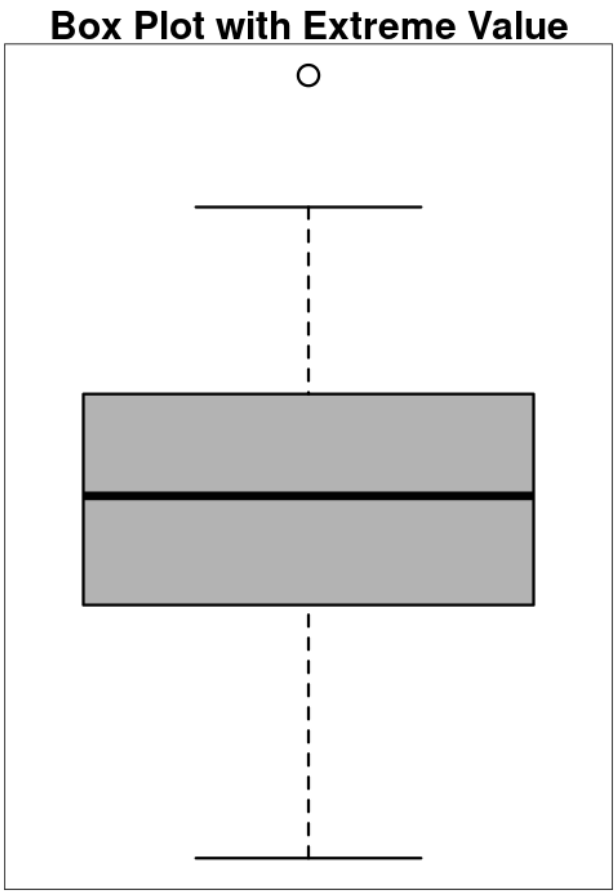



*Fig. 2.1.3* A box plot with one extreme value excluded from the main box plot and instead plotted as a separate point.

We can construct a boxplot in R using the boxplot function:

```
boxplot(x)
```

Here, `x` is the list of data we want to create a boxplot for.

**Note**

Sometimes there will be a slight difference between the quartile values we obtain using the `quantile` function and the partition boundaries on the box plot we obtain using the `boxplot` function. This is due to small differences in the way the `quantile` function and the `boxplot` function calculate $Q_1$ and $Q_3$. In practice, this small difference usually doesn't matter as both still partition the data into regions containing *about* 25% of the data.

---

## Example 2.1.3

The following data are the heights of $40$ students in a statistics class. Construct a box plot using R. In which quarter is the data spread out the least? In which quarter is the data spread out the most?

60, 60, 61, 62, 62, 63, 63, 64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 67, 67, 68, 68, 69, 70, 70, 70, 70, 70, 71, 71, 72, 72, 73, 74, 74, 75, 77

### Solution

Simply construct a box plot of the data.

```
heights = c(60, 60, 61, 62, 62, 63, 63, 64, 64, 64, 65, 65, 65, 65, 65, 65, 65, 65, 65, 66, 66, 67, 67, 68,

boxplot(heights)
```

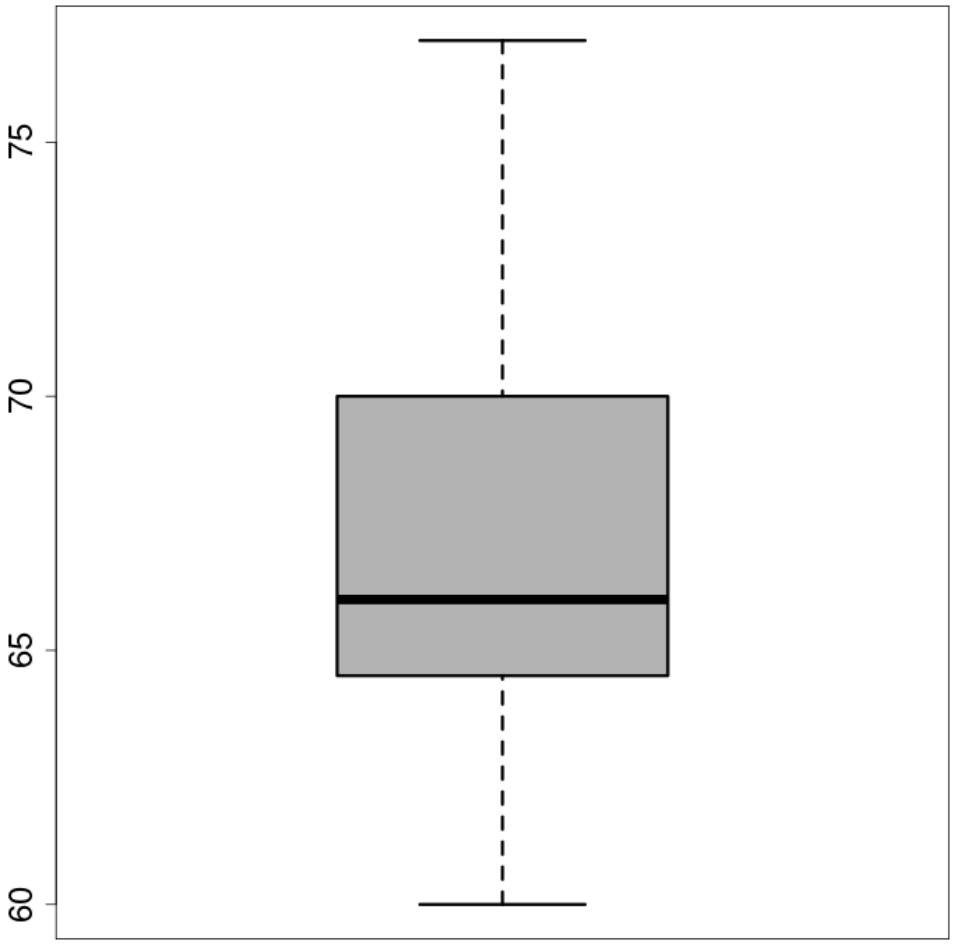



*Fig. 2.1.4* A box plot of the example data.

From the box plot, we can see that the region between the first quartile $Q_1$ and the median $M = Q_2$ has the smallest spread of data, and the region between the third quartile $Q_3$ and the maximum value has the largest spread of data.

---

By providing the `boxplot` function with multiple sets of data, we can generate multiple box plots side-by-side. This is useful for comparing how different data sets are distributed.

---

## Example 2.1.4

Use R to construct side-by-side boxplots for the two sets of data below. Which data set has the larger spread of data?

Data set 1: 64, 73, 101.3, 71.1, 94, 85.5, 99.9, 42, 92.4, 103, 54.7, 72.5, 101.9, 86.4, 47.2, 71.5, 71

Data set 2: 56.8, 71, 49, 69.4, 67.6, 71, 68.5, 61.1, 50.9, 71, 71, 71, 72, 72, 71, 64, 67.2, 52.5, 71, 69, 72, 71.3, 58.3, 71, 71

### Solution

We will create two box plots side-by-side, one for each set of data.

Skip to main content

```
data1 = c(64, 73, 101.3, 71.1, 94, 85.5, 99.9, 42, 92.4, 103, 54.7, 72.5, 101.9, 86.4, 47.2, 71.5, 71)
data2 = c(56.8, 71, 49, 69.4, 67.6, 71, 68.5, 61.1, 50.9, 71, 71, 71, 72, 72, 71, 64, 67.2, 52.5, 71, 69, 72

boxplot(data1, data2)
```

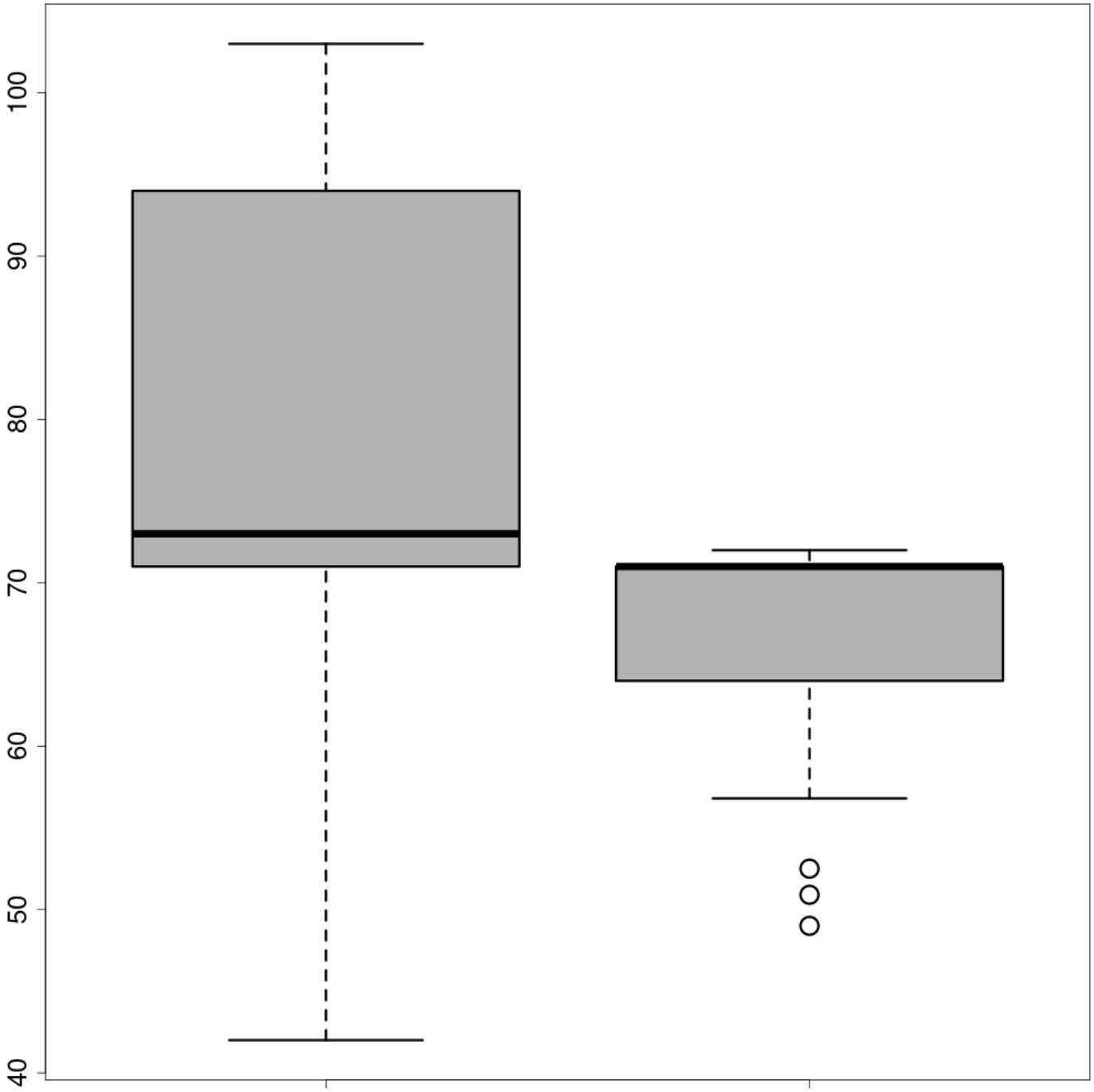



*Fig. 2.1.5* Two box plots side-by-side makes it easy to compare how different data sets are distributed.

By examining the box plots, we quickly see that the data in data set 1 is much more spread out than the data in data set 2.

## 2.2. Measures of Central Tendency

Skip to main content

## Objectives

- Calculate the mean and median for a given data set.

## Mean and Median

The “center” of a data set is also a way of describing location. The two most widely used measures of the “center” of the data are the **mean** (average) and the **median**. To calculate the mean weight of 50 people, add the 50 weights together and divide by 50. In the last section, we learned that the median is the second quartile $Q_2$ or the 50th percentile. To find the median weight of the 50 people, order the data and find the number that splits the data into two equal parts. The median is generally a better measure of the center when there are extreme values or outliers because it is not affected by the precise numerical values of the outliers. The mean is the most common measure of the center.

**Note**

The words *mean* and *average* are often used interchangeably. The substitution of one word for the other is common practice. The technical term for the type of mean we a discussing in this section is *arithmetic mean*, and *average* is technically a general term for any center location. For example, the *median* of a data set is also technically an average of the data set because it is one measure of the center location. However, in practice among non-statisticians, the *average* of a data set is commonly understood as the *arithmetic mean*.

We use different symbols to differentiate between the mean of a sample and the mean of a population. The letter used to represent the sample mean is an x with a bar over it (pronounced “x bar”): $\bar{x}$. The Greek letter $\mu$ (pronounced “mew” and spelled in English “mu”) represents the population mean. Both the sample mean and population mean are calculated the same way: add together all the data values in the sample or population, then divide by the number of data values in the sample of population. In practice, we usually calculate the sample mean $\bar{x}$. If we know enough about the population to calculate the population mean $\mu$, then we don’t need to collect a smaller sample to estimate features of the population.

When each value in the data set is not unique, the mean can be calculated by multiplying each distinct value by its frequency and then dividing the sum by the total number of data values. To see that both ways of calculating the mean are the same, consider this sample of 11 data values:

1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 4

$$\bar{x} = \frac{1+1+1+2+2+3+4+4+4+4+4}{11} = 2.727$$

$$\bar{x} = \frac{3(1)+2(2)+1(3)+5(4)}{11} = 2.727$$

In the second calculation, the frequencies are 3, 2, 1, and 5 since the data contains 3 ones, 2 twos, 1 three, and 5 fours.

In R, these calculations look like:

```
xbar = (1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4 + 4)/11
```

Skip to main content

2.72727272727273

```
xbar = (3*1 + 2*2 + 1*3 + 5*4)/11
xbar
```

2.72727272727273

> **Note**
>
> In each case, we first calculate the mean and store the value in the variable `xbar`. To have the computer display the value that we stored in `xbar`, we simply type `xbar` by itself on the next line. If we don't type that extra `xbar`, the computer will store the value of the mean in the variable, but it won't tell use what the mean is.

Observe that all we do to calculate the mean is add all the data values together, then divide by the number of data values. This concept can be more succinctly expressed using the formula

$$\bar{x} = \frac{\sum x}{n}.$$

We use $\sum$ (the capital Greek letter sigma) when we want to add up or find the sum of values. In this case, the formula is telling us to add up all the $x$'s, where we use $x$ as a placeholder for the data values in the sample. Then we divide the sum of $x$'s by $n$, where $n$ is the number of data values in the sample. Note, though we've used the sample mean $\bar{x}$ in the formula, the formula is essentially the same for the population mean $\mu$:

$$\mu = \frac{\sum x}{N},$$

where the $x$'s are the the data values in the population, and $N$ is the number of data values in the population.

Calculating a mean is easy using R. We can use the `sum` function to add up the values in a list, which gives us $\sum x$. And we can use the `length` function to find out how many values are in a list, which gives us the sample size $n$ (or population size $N$ if working with an entire population). Both the `sum` function and the `length` function have just one argument:

```
sum(x)
```

```
length(x)
```

In both cases, `x` is a list of data values.

So for the above sample data, we can calculate the sample mean using R as follows:

```
x = c(1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 4)
n = length(x)

xbar = sum(x)/n
xbar
```

2.72727272727273

Skip to main content

**Note**

We will usually store a list of data in variable `x` to make the computer code as similar as possible to the formula. For example, `xbar = sum(x)/n` has an obviously similar structure to $\bar{x} = \sum x/n$. However, any variable name would work. For example, the following code calculates the mean just as well as the code above:

```
values = c(1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 4)
n = length(values)

xbar = sum(values)/n
xbar
```

We've already discussed the median: the median $M$ is the same as $Q_2$, the second quartile, or the 50th percentile. The median is the 'middle value' of the data: exactly half the data are greater than the median, and exactly half the data are less than the median. We've seen that we can find the median in R using `quantile(x, probs=0.50)`, where `x` is the list of data values we want the median of. For example, to find the median of the data above, we would type:

```
x = c(1, 1, 1, 2, 2, 3, 4, 4, 4, 4, 4)

quantile(x, probs = 0.50)
```

**50%:** 3

The median is $M = 3$. Exactly half the values in the data are less than the median, and exactly half the values is the data are greater than the median.

---

## Example 2.2.1

The following data show the number of months patients wait on a transplant list before getting surgery. The data are ordered from smallest to largest. Calculate the mean and median.

3, 4, 5, 7, 7, 7, 7, 8, 8, 9, 9, 10, 10, 10, 10, 10, 11, 12, 12, 13, 14, 14, 15, 15, 17, 17, 18, 19, 19, 19, 21, 21, 22, 22, 23, 24, 24, 24, 24

### Solution

```
x = c(3, 4, 5, 7, 7, 7, 7, 8, 8, 9, 9, 10, 10, 10, 10, 10, 11, 12, 12, 13, 14, 14, 15, 15, 17, 17, 18, 19, 1

# Find the Mean
n = length(x)

xbar = sum(x)/n
xbar

# Find the Median
M = quantile(x, probs = 0.50)
M
```

13.9487179487179

**50%:** 13

Skip to main content

(We use $\bar{x}$ instead of $\mu$ because we are dealing with only a *sample* of patients on the transplant list, not the entire *population* of transplant patients.)

---

## Example 2.2.2

All exam scores from a Calculus class are shown below. Find the mean score and the median score.

42, 24, 35, 31.5, 32.5, 32, 48.5, 35.5, 38, 40, 20.5, 37, 34, 41.5, 43, 49, 48, 28, 35, 48.5, 22, 35.5, 44.5, 39, 21.5, 34.5, 40

```
x = c(42, 24, 35, 31.5, 32.5, 32, 48.5, 35.5, 38, 40, 20.5, 37, 34, 41.5, 43, 49, 48, 28, 35, 48.5, 22, 35.5

# Find the Mean
N = length(x)

mu = sum(x)/N
mu

# Find the Median
M = quantile(x, probs = 0.50)
M
```

36.3148148148148
**50%:** 35.5

So the mean score is about $\mu = 36.3148$ points and the median is $M = 35.5$ points.

(We use $\mu$ instead of $\bar{x}$ because we are dealing with the whole *population* of exam scores, not a *sample* of the class exam scores.

---

## Example 2.2.3

Suppose that we sample 50 people in one city. One person earns $5,000,000 per year and the other 49 each earn $30,000. Find the mean and the median of the data. Which is the better measure of the “center,” the mean or the median?

### Solution

The data in this example includes lots of large numbers. We could calculate the mean and median of the data the same way we have in the previous examples, but it would take a long time to list out all 50 numbers. Instead, we can take advantage of the fact that many of the values are repeated to perform the calcualtions more quickly.

Since there is one person who earns $5,000,000 per year and 49 people who earn $30,000 per year, we calculate the mean as follows:

```
xbar = (1*5000000 + 49*30000)/50
xbar
```

129400

So the mean annual income is $\bar{x} = \$129,400$ per year.

To calculate the median, remember that the median is the middle value. Imagine lining up the data from smallest to largest. The $5,000,000 value lies on the edge of our list of values. All the other values are $30,000, so the median, the middle value, must be

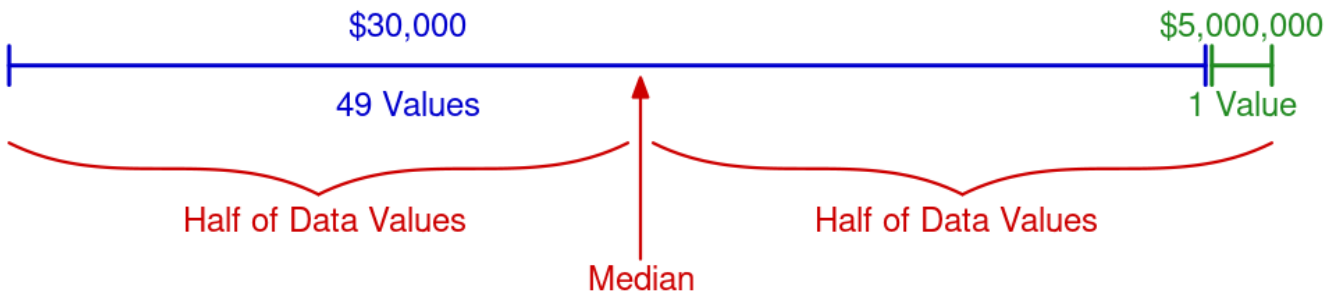



*Fig. 2.2.1* If we imagine lining up the data values, we see that the median must be $M = \$30,000$.

The median is a better measure of the "center" than the mean in this case because 49 of the values are $30,000 and one is $5,000,000. The $5,000,000 value is an outlier and significantly skews the mean. The median of $M = \$30,000$ gives us a better sense of the income of an ordinary person in the city than the mean of $\bar{x} = \$129,400$.

# 2.3. Measures of Dispersion

## Objectives

- Calculate the standard deviation of a data set.
- Distinguish between the population standard deviation and sample standard deviation.
- Identify the outliers of a data set.

## The Standard Deviation

An important characteristic of any set of data is the variation in the data. In some data sets, the data values are concentrated closely near the mean; in other data sets, the data values are more widely spread out from the mean. The most common measure of variation is the standard deviation. The **standard deviation** is a number that measures the average distance of data values from their mean.

Because distance is never a negative value, the standard deviation is always positive or zero. The standard deviation is small when the data are all concentrated close to the mean, exhibiting little variation. The standard deviation is larger when the data values are more spread out from the mean, exhibiting more variation.

Suppose that we are studying the amount of time customers wait in line at the checkout at supermarket A and supermarket B. the average wait time at both supermarkets is five minutes. At supermarket A, the standard deviation for the wait time is two minutes; at supermarket B the standard deviation for the wait time is four minutes.

Because supermarket B has a higher standard deviation, we know that there is more variation in the wait times at supermarket B. Overall, wait times at supermarket B are more spread out from the mean; wait times at supermarket A are more concentrated near the mean.

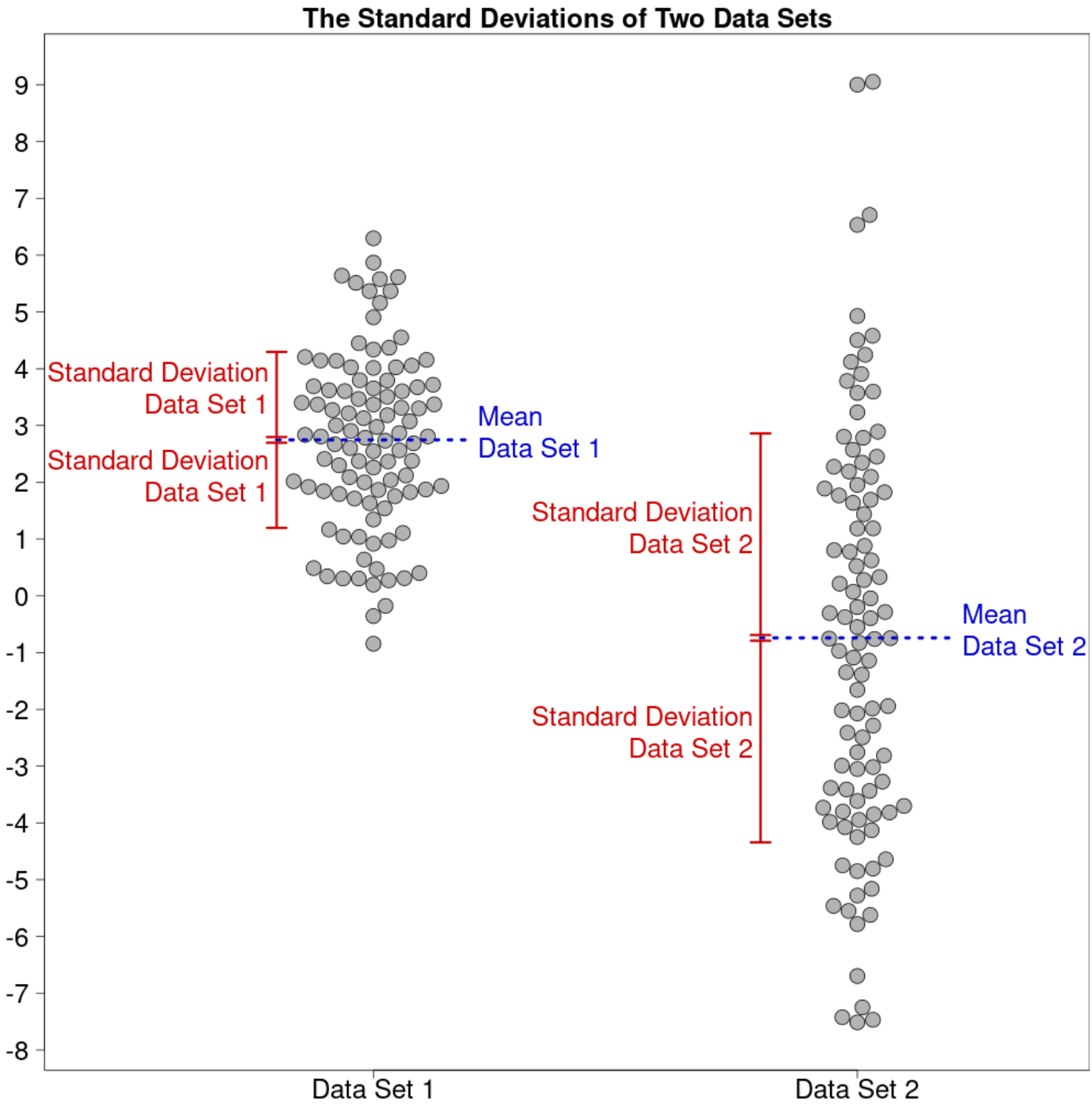



*Fig. 2.3.1* The standard deviation is a measure of the average distance of data values from their mean. If we compare the two data sets in the figure, the values of data set 1 are squashed closer to the mean, so data set 1 has a smaller standard deviation. The values of data set 2 are more spread out from the mean, so data set 2 has a larger standard deviation.

## Calculating the Population Standard Deviation

We will first look at calculating the standard deviation for a population. If $x$ is a data value, then the difference "$x - \text{mean}$" is called its **deviation**; it's a measure of the distance from the value $x$ to the mean of the data. There is a deviation for each value in a data set. The deviations are used to calculate the standard deviation. For a population, the deviation for a data value $x$ is $x - \mu$.

Skip to main content

For example, suppose a population has mean $\mu = 5$, and consider two data values of the population: $x = 2$ and $x = 8$. The deviation of the value $x = 2$ from the mean is

$$x - \mu = 2 - 5 = -3.$$

The deviation of the value $x = 8$ from the mean is

$$x - \mu = 8 - 5 = 3.$$

Both $x = 2$ and $x = 8$ are a distance of $3$ units away from $\mu = 5$. But the signs (positives or negatives) of the deviations also tell us direction: since the deviation of $x = 2$ is $-3$, $x = 2$ is $3$ units *below* the mean; since the deviation of $x = 8$ is $+3$, $x = 8$ is $3$ units *above* the mean.

When calculating the standard deviation, we want to only focus on the average distance from the mean, not the direction. To remove the information on direction (the sign) from each deviation, we square each deviation. For $x = 2$, the squared deviation is

$$(x - \mu)^2 = (2 - 5)^2 = (-3)^2 = 9.$$

For $x = 8$, the squared deviation is

$$(x - \mu)^2 = (8 - 5)^2 = 3^2 = 9.$$

Since squaring a deviation always results in a positive number, it removes the information on direction. Unfortunately, squared deviations tell us the *square* of the distance of each value from the mean, not simply the distance. We will correct this problem later.

Next, we find the mean of all the squared deviations in the usual way: by adding up all the squared deviations in the population and dividing by the population size $N$:

$$\frac{\sum (x - \mu)^2}{N}.$$

This is the average *squared* distance of the population data from the mean. We need to "unsquare" the average by taking the square root. This gives us the formula for the population standard deviation:

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}.$$

We denote the population standard deviation by $\sigma$ (the lower-case Greek letter sigma).

Note that to calculate the population standard deviation $\sigma$, we first must calculate the population mean $\mu$ since we use $\mu$ in the formula for $\sigma$.

Also, to use the formula for population standard deviation in R, we will need to know how to take a square root in R. We can take the square root of a value in R using the `sqrt` function:

```
sqrt(x)
```

Skip to main content

```
sqrt(4)
```

2

---

## Example 2.3.1

The Cook Political Report scores the partisan lean of each state in the U.S. based on the results of recent elections. The Partisan Voting Index (PVI) for each of the 50 states (as of April 2021) is below, where more negative values means the state leans more toward the Democratic party, and more positive values means the state leans more toward the Republican party.[1]

15, 9, 3, 16, -14, -3, -7, -6, 3, 3, -15, 19, -7, 11, 6, 11, 16, 12, -1, -14, -14, 1, -1, 10, 11, 11, 13, 0, 0, -6, -3, -10, 3, 20, 6, 20, -6, 2, -8, 8, 16, 14, 5, 13, -15, -2, -8, 23, 2, 26

1. Calculate the mean state PVI.
2. Calculate the standard deviation of the state PVIs.

### Solution

Note that this list includes the PVIs for the entire *population* of 50 states, so we will be calculating the *population* mean and the *population* standard deviation.

### Part 1

```
x = c(15, 9, 3, 16, -14, -3, -7, -6, 3, 3, -15, 19, -7, 11, 6, 11, 16, 12, -1, -14, -14, 1, -1, 10, 11, 11,

# Calculate the Mean
N = length(x)

mu = sum(x)/N
mu
```

3.76

The mean PVI is $\mu = 3.76$. This means that states on average lean Republican.

### Part 2

```
# Calculate the Standard Deviation
sigma = sqrt( sum( (x - mu)^2 )/N )
sigma
```

10.6743805440878

The standard deviation of the state PVIs is $\sigma = 10.6744$. On average, state PVIs differ from the mean by $10.6744$ points.

Skip to main content

**Warning**

Be careful that you don't interpret the mean state PVI that we calculated above to be the same as the PVI for the whole United States. Many Democratic leaning states have much larger populations than Republican leaning states, but we didn't account for population differences in our analysis. For example, California (with a PVI of -14) has a population that is 68.5 times as large as the population of Wyoming (with a PVI of 26), but both states were treated equally in our calculations above. The lesson: we must be very careful not to misinterpret or over-interpret statistical results.

---

## Calculating the Sample Standard Deviation

In many real-life situations, it is impractical to collect the data needed to calculate the standard deviation for an entire population. Instead, statisticians estimate the standard deviation of a population by calculating the standard deviation of a sample. The formula for the sample standard deviation is almost the same as the formula for the population standard deviation.

**Formula for the Sample Standard Deviation:**

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Here, $x$ is a data value in our sample, $\bar{x}$ is the sample mean, and $n$ is the sample size. We denote the sample standard deviation by $s$.

For comparison, let's also review the formula for calculating the population standard deviation.

**Formula for the Population Standard Deviation:**

$$\sigma = \sqrt{\frac{\sum(x - \mu)^2}{N}}$$

Here, $x$ is a data value in our population, $\mu$ is the population mean, and $N$ is the population size. We denote the population standard deviation by $\sigma$.

Most of the differences between the sample standard deviation $s$ and the population standard deviation $\sigma$ are trivial. Because we are dealing with a sample instead of a population, we use $\bar{x}$ instead of $\mu$ to notate the mean, and we use $n$ instead of $N$ to notate the number of data values. But the most significant difference is that the when calculating the sample standard deviation, we divide by *one less* than the number of data values: $n - 1$. In contrast, when calculating the population standard deviation, we simply divide by the number of data values: $N$.

**Why?**

Why do we divide by $n - 1$ instead of $n$ in the sample standard deviation formula? The sample standard deviation uses the distances from each value to the *sample* mean in its calculation: $(x - \bar{x})$. We want the sample standard deviation to approximate the population standard deviation, but the population standard deviation averages the distances from each value to the *population* mean: $(x - \mu)$. The sample mean $\bar{x}$ is usually close to $\mu$, but they are rarely the same. Because of this small difference in what is being calculated, we need a small correction to the formula. Using techniques that are more advanced than what we will cover in this text, statisticians can show that dividing by $n - 1$ instead of $n$ corrects this

### Example 2.3.2

The data below represents students' scores on the first exam from one of Susan Dean's precalculus classes.

33, 42, 49, 49, 53, 55, 55, 61, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 90, 92, 94, 94, 94, 94, 96, 100

1. Calculate the sample mean.
2. Calculate the sample standard deviation.

Solution

Part 1

```
x = c(33, 42, 49, 49, 53, 55, 55, 61, 63, 67, 68, 68, 69, 69, 72, 73, 74, 78, 80, 83, 88, 88, 88, 90, 92, 94

# Calculate the Sample Mean
n = length(x)

xbar = sum(x)/n
xbar
```

73.5161290322581

The mean score is $\bar{x} = 73.5161$.

Part 2

```
# Calculate the Sample Standard Deviation
s = sqrt( sum( (x - xbar)^2 ) / (n - 1) )
s
```

17.923673298633

So the standard deviation of the scores is $s = 17.9237$.

## Outliers

**Outliers** are data values that deviate from the mean much greater than the standard amount for the data. Statisticians are interested in identifying outliers because they may indicate an error in the data. For example, suppose we collect the following data for the price of a cup of coffee at several local coffee shops:

2.69, 2.79, 2.99, 3.00, 2.69, 3.15, 2.89, 2.69, 2.90, **259**, 2.79, 2.80, 3.10, 2.59, 2.75

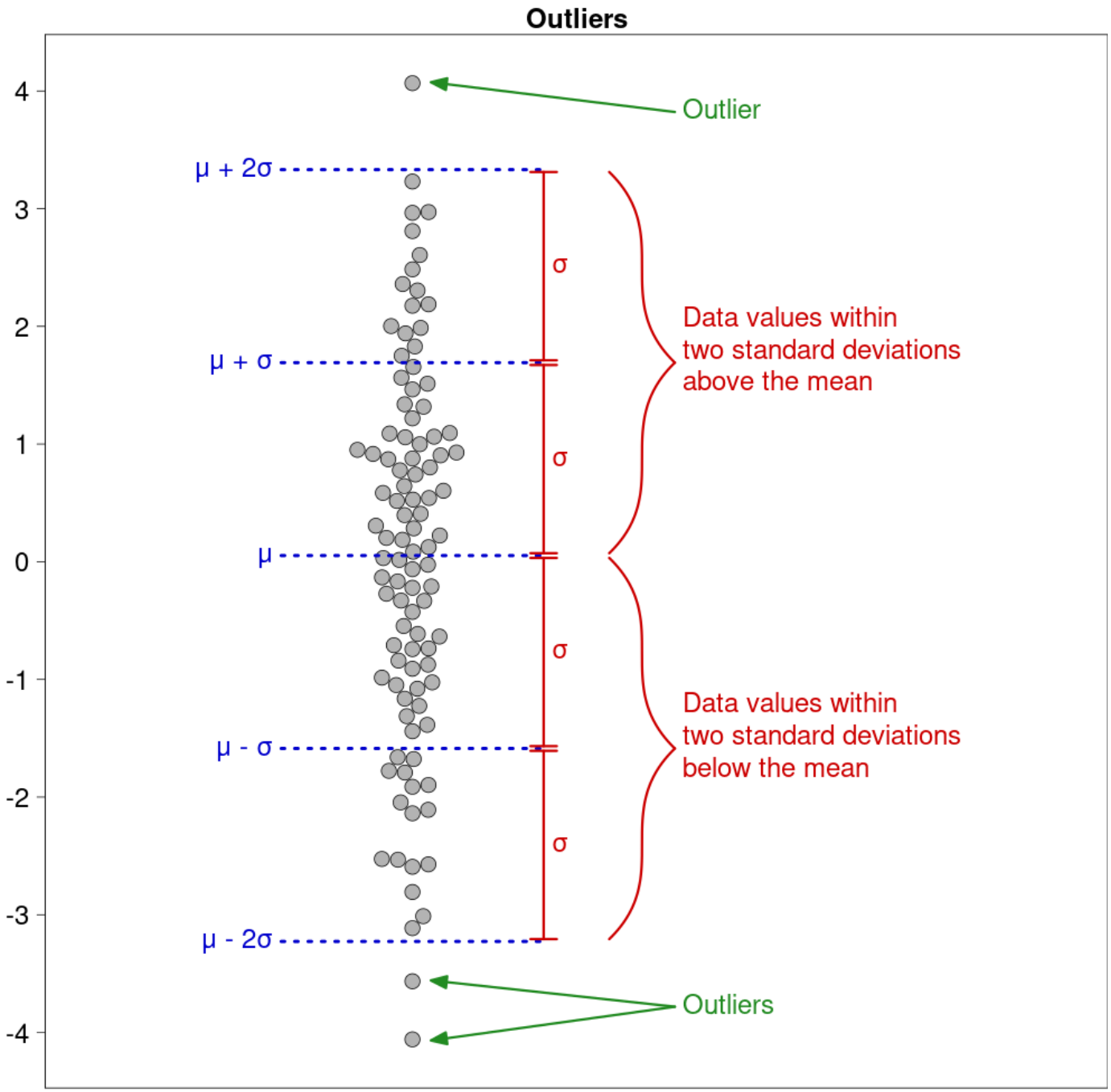
The value in bold is a clear outlier. It seems extremely unlikely that this cup of coffee actually cost $259. Instead, the data value may have been recorded incorrectly, so that the true cost for this cup of coffee was $2.59, not $259.

There are many different methods statisticians use to calculate outliers depending on the distribution of the data they are working

standard deviations away from the mean. In mathematical notation, this means a data value is an outlier in a population if it is more than $\mu + 2\sigma$ or less than $\mu - 2\sigma$. In a sample, a data value is an outlier if it is more than $\bar{x} + 2s$ or less than $\bar{x} - 2s$. This is a reasonable definition because the average distance between data values and their mean is one standard deviation, so most data values lie within a distance of two standard deviations from their mean. Only relatively extreme data values are more than two standard deviations away from the mean.

This definition is more a “rule of thumb” than a rigid law. Professional statisticians may use different rules to identify outliers depending on the circumstances.

Also, just because a value is an outlier does not mean there is anything wrong with it. While some outliers may be the result of errors, other outliers may be perfectly valid data values that are just unusually large or small. You should determine whether outliers are valid on a case-by-case basis.




Skip to main content

*Fig. 2.3.2* Given a set of data, most data values are within two standard deviations from their mean, either above or below the mean. In this text, we consider outliers to be data values that are more than two standard deviations away from the mean.

---

## Example 2.3.3

A dating website surveyed 50 of their recently-married customers to see how many long-term relationships they had before they were married. The results are below.

1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 9, 10, 11, 11, 13, 13, 14, 15

1. Find the mean number of relationships of the individuals surveyed.
2. Find the standard deviation of the data.
3. Does the data contain any outliers? If so, what are they?

### Solution

Note that the 50 customers surveyed are only a small *sample* of all the dating website's customers, so we will calculate the *sample* mean and *sample* standard deviation.

### Part 1

```
x = c(1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6,

# Calculate the Mean
n = length(x)

xbar = sum(x)/n
xbar
```

5.42

On average, those surveyed had $\bar{x} = 5.42$ relationships before they were married.

### Part 2

```
s = sqrt( sum( (x - xbar)^2 )/(n - 1) )
s
```

3.55763048866483

The sample standard deviation is $s = 3.558$ relationships.

### Part 3

To find any outliers, we will calculate two standard deviations above the mean ($\bar{x} + 2s$) and two standard deviations below the

```
xbar + 2*s
xbar - 2*s
```

12.5352609773297

-1.69526097732966

Returning to the original list of data values, we can see that $13$, $13$, $14$, and $15$ are greater than $\bar{x} + 2s = 12.5353$. Any values smaller than $\bar{x} - 2s = -1.6953$ would also be outliers, but there aren't any values this small in the data.

So our data contains a total of four outliers: $13$, $13$, $14$, and $15$.

---

[1] Wasserman, David; Flinn, Ally (April 15, 2021). "Introducing the 2021 Cook Political Report Partisan Voter Index". The Cook Political Report. Retrieved April 15, 2021.

# 3. Probability

## 3.1. Introduction to Probability

### Objectives

- Understand fundamental terminology used in probability and be able to apply it appropriately.
- Express an event of a probability experiment as a list of outcomes in set notation.
- Understand the law of large numbers.
- Calculate basic probability involving sample spaces with equally likely outcomes.

### Experiments, Outcomes, and Events

Before we consider probability, we first must define a few key terms. To begin, it will be useful to focus on activities that can be repeated and where all the possible outcomes are known. These kinds of activities are called **experiments**. For example, flipping a fair coin is an experiment because it can be repeated (we can flip the coin again), and all the possible outcomes are known (the coin will land on either heads or tails). Another example of an experiment is drawing two cards from a standard 52 card deck because it can be repeated (we can return the two cards to the deck, shuffle, then take two new cards from the deck) and we know all the possible outcomes (since we know which cards are in the deck, we know all the cards that might be drawn).

To be precise, we define an **outcome** of an experiment as a possible result of the experiment. Before an experiment takes place, there may be several possible outcomes that have a chance of happening. But after the experiment has concluded, exactly one outcome has happened. For example, if you toss a coin, the coin may come up heads or tails—two possible outcomes. But once the coin has landed, the coin has either come up heads or come up tails—exactly one outcome has actually occurred.

The **sample space** of an experiment is the **set** or collection of all possible outcomes. The uppercase letter $S$ is used to denote sample space. We can express the set of all possible outcomes using **set notation**: we list all the outcomes and enclose the list in curly brackets. For example, if you flip one fair coin, the sample space is $S = \{H, T\}$ where $H$ = Heads and $T$ = Tails are the outcomes. As another example, if you roll a six-sided die, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$, where each number represents the side of the die that is face up.

Skip to main content

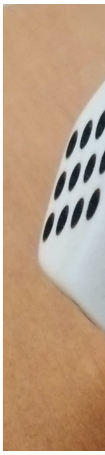
An **event** is any combination of outcomes. We usually use upper case letters like $A$ and $B$ to represent events, and we use set notation to represent the outcomes in the event. For example, if the experiment is to roll a six-sided die, then let event $A$ be rolling an even number, and let event $B$ be rolling a number greater than $2$. In set notation, $A = \{2, 4, 6\}$ and $B = \{3, 4, 5, 6\}$. If the experiment results in *any* outcome in the event, then the event has happened. And note that while an experiment can have only one outcome, it can result in multiple events happening. For example, if we roll a $4$ with a six-sided die, then event $A$ has happened (since $4$ is an even number), *and* event $B$ has happened (since $4$ is a number greater than $2$).

It is often useful to know how many outcomes are in an event. For any event $E$, we let $n(E)$ represent the number of outcomes in the event. Again, consider rolling a six-sided die, and let $D = \{2, 3, 5, 6\}$. There are four outcomes in event $D$, so $n(D) = 4$. Similarly, $n(S) = 6$ because $S$ is the entire sample space, and there are six possible outcomes total when rolling a six-sided die.

---

## Example 3.1.1

Imagine you roll a twelve-sided die.



1. Write the sample space $S$ in set notation. What is $n(S)$?
2. Let event $A =$ rolling an odd number. What outcomes are in $A$? Express your answer in set notation. What is $n(A)$?

### Solution

#### Part 1

The sample space is the set of all possible outcomes. Since we are rolling a twelve-sided die, the die land on any number from $1$ to $12$. In set notation, the sample space is

$$S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}.$$

Clearly, there are twelve possible outcomes, so $n(S) = 12$.

#### Part 2

Since event $A$ is rolling an odd number, in set notation,

$$A = \{1, 3, 5, 7, 9, 11\}.$$

There are six outcomes in $A$, so $n(A) = 6$.

---

## Example 3.1.2

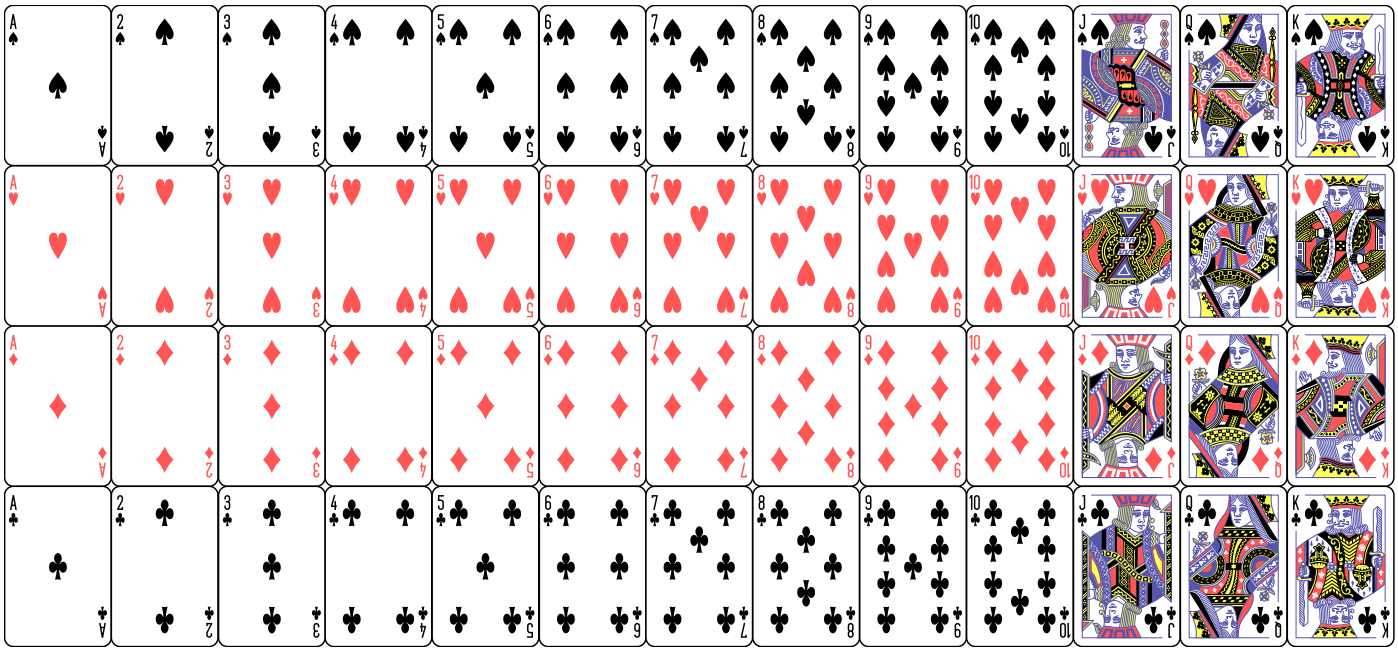


*Fig. 3.1.2* A standard deck of $52$ playing cards.[2]

A deck is divided into 4 suits: Spades (♠), Hearts (♥), Diamonds (♦), and Clubs (♣).

Spades (♠) and Clubs (♣) are black. Hearts (♥) and Diamonds (♦) are red.

There are 13 ranks in each suit: Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, and King.

The Jacks, Queens, and Kings in a deck are called face cards.

Prudence draws one card from a standard deck of $52$ cards.

1. Find $n(S)$, where $S$ is the sample space.
2. Let event $F =$ drawing a black face card. What outcomes are in $F$? Express your answer in set notation. What is $n(F)$?
3. Let event ♠ $=$ drawing a spade. What is $n(\spadesuit)$?

Solution

Part 1

The sample space $S$ is the set of all possible outcomes. Each card that Prudence can draw represents one possible outcome. Since Prudence draws from a deck containing 52 different cards, $n(S) = 52$.

Part 2

In set notation, $F = \{J\spadesuit, Q\spadesuit, K\spadesuit, J\clubsuit, Q\clubsuit, K\clubsuit\}$. Since there are six black face cards, $n(F) = 6$.

Skip to main content

**Note**

In this example, we chose to represent each card with a letter (J, Q, or K) and an image of the suit (♠ or ♣). But there are other ways we could have represented each outcome. For example, we could have written down the full name of each card:

$$F = \{\text{Jack of Spades, Queen of Spades, King of Spades,}\\ \text{Jack of Clubs, Queen of Clubs, King of Clubs}\}.$$

It's not important how we represent these outcomes as long as the meaning is clear.

Part 3

Looking at the deck pictured above, we can see that thirteen of the cards are spades. So $n(\spadesuit) = 13$.

---

## Basic Probability

**Probability** is a measure of how likely it is that some event occurs when an experiment is performed. For example, if we flip a fair coin, there is a $50\%$ probability that the coin comes up heads and a $50\%$ probability that the coin comes up tails. This means that the coin is equally as likely to come up heads as it is to come up tails. Probabilities are always between $0\%$ and $100\%$ because an event cannot have less than a $0\%$ chance of happening or more than a $100\%$ chance of happening.

While it is common to use percentages when speaking of probabilities, it is more convenient mathematically to express probabilities in decimal form. In decimal form, the probability that a flipped coin comes up heads is $0.5$ (since $0.5 = 50\%$). Similarly, in decimal form, probabilities are always between $0$ and $1$. If a probability is in percentage form, it must first be converted to decimal form before it can be used in calculations.

We denote the probability of an event $E$ as $P(E)$. For example, if we toss a fair coin and $H =$ the coin comes up heads, then $P(H) = 0.5$.

The probability of an event tells us how often we would expect the event to occur if we repeated an experiment many times. If $P(A) = 0.25$, then we expect event $A$ to happen in $25\%$ of the experiments if we repeated the experiment many times. If $P(A) = 0.672$, then we expect event $A$ to happen in $67.2\%$ of the experiments if we repeated the experiment many times. If $P(A) = 0$, then event $A$ will happen in $0\%$ of the experiments; that is, event $A$ will never be the result of an experiment. If $P(A) = 1$, then event $A$ will happen in $100\%$ of the experiments; that is, event $A$ will happen in every experiment.

The more times you conduct an experiment, the more likely it is that the proportion of times an event occurs will be close to the probability of the event. For example, if we tossed a coin just two times, we wouldn't be too surprised if $100\%$ of the tosses came up heads. If we tossed the coin five times, we'd be a little more surprised if $100\%$ of the tosses came up heads. But if we toss a coin $1,000$ times, we would be *very* surprised if $100\%$ of the tosses come up heads. With $1,000$ tosses, we expect the proportion of tosses that come up heads to be much closer to the probability of $0.5 = 50\%$. (See Figure 3.1.3.)

This important characteristic of probability experiments is known as the **law of large numbers** which states that as the number of repetitions of an experiment is increased, the proportion obtained in the experiments tends to become closer and closer to the theoretical probability. Even though the outcomes of the experiments do not happen according to any set pattern or order, overall, the long-term observed proportion will approach the theoretical probability.

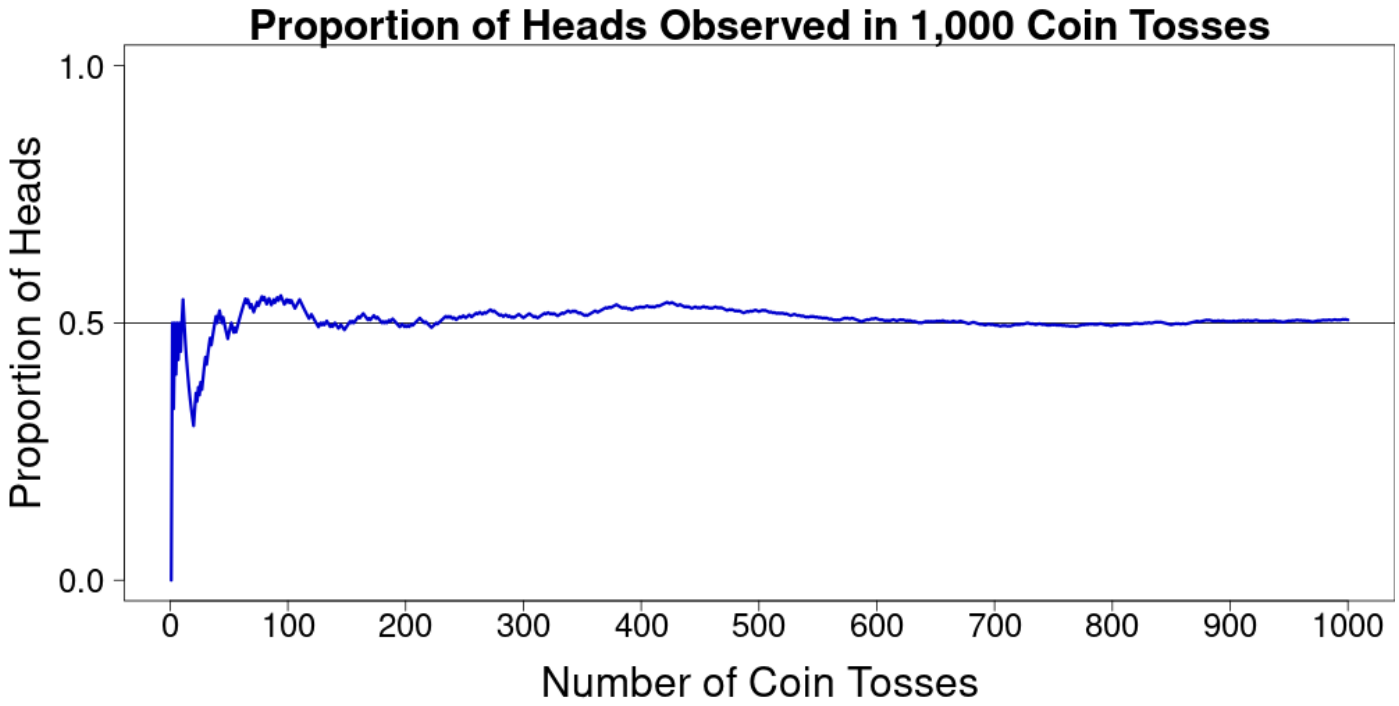



*Fig. 3.1.3* A computer was used to simulate 1,000 coin tosses. The graph relates the proportion of heads that have come up to the number of times the coin has been tossed. Notice that when the coin has been tossed only a few times, the graph looks wild, and the proportion of heads can be quite different than the probability of tossing a head, $0.5$. As the number of coin tosses increases, the graph grows less erratic, and the proportion of heads tends to get closer to $0.5$.

When all the outcomes of an experiment are **equally likely**—that is, when each outcome in the sample space has the same probability of occuring—the probability of an event $E$ can be calculated using the formula

$$P(E) = \frac{n(E)}{n(S)}.$$

That is, to find $P(E)$, we divide the number of possible outcomes in event $E$ by the total number of possible outcomes in the sample space $S$.

For example, imagine we toss a fair coin. Since it is just as likely for a coin to come up heads as it is to come up tails, all the outcomes are equally likely. Let $H$ be the event where the coin comes up heads. Since there are only two possible outcomes in the sample space (heads or tails), $n(S) = 2$. Since there is only one outcome in $H$ (heads), $n(H) = 1$. So the probability of tossing a coin and having it come up heads is

$$P(H) = \frac{n(H)}{n(S)} = \frac{1}{2} = 0.5.$$

As another example, suppose we roll a six-sided die. Each value of the die has an equal probability of being rolled, so all the outcomes are equally likely. Let $A$ be the event of rolling a value that's not a $2$. Then $A = \{1, 3, 4, 5, 6\}$, and $n(A) = 5$. Since there are $6$ possible outcomes in total,

$$P(A) = \frac{n(A)}{n(S)} = \frac{5}{6} = 0.8333.$$

Skip to main content

The probability of rolling a number that isn't a $2$ is $0.8333$. In other words, if we roll a six-sided die, there is an $83.33\%$ chance that we do not roll a $2$.

It is important to realize that in many situations the outcomes are not equally likely. For example, imagine a weather forecast predicts that there is a $70\%$ chance of rain tomorrow. If we view tomorrow's weather as an experiment, then there are two possible outcomes: rain, or no rain. But the probability that it rains tomorrow is $70\% = 0.7$, meaning the probability that it doesn't rain is only $30\% = 0.3$. Since the two possible outcomes have different probabilities, the outcomes are not equally likely.

When the outcomes of an experiment are not equally likely, we cannot use the above formula to calculate an event's probability. We will learn techniques for working with outcomes that are not equally likely later.

---

## Example 3.1.3

Imagine you draw a card from a well-shuffled standard $52$-card deck. (See Figure 3.1.2.)

1. Let event $N =$ drawing a $7$. Find $P(N)$.
2. Let event $\clubsuit =$ drawing a club. Find $P(\clubsuit)$.
3. Let event $F =$ drawing a face card. Find $P(F)$.
4. Let $S$ be the sample space. Find $P(S)$.

### Solution

First, note that all outcomes are equally likely because each card has the same chance of being drawn. Also, since there are $52$ cards in the deck that could possibly be drawn, $n(S) = 52$.

#### Part 1

A deck of cards contains four $7$s, so $n(N) = 4$. Thus, the probability of drawing a $7$ is

$$P(N) = \frac{n(N)}{n(S)} = \frac{4}{52} = 0.0769.$$

There is a $7.69\%$ chance that the card you draw is a $7$.

#### Part 2

There are thirteen clubs in a deck, so $n(\clubsuit) = 13$. The probability of drawing a club is

$$P(\clubsuit) = \frac{n(\clubsuit)}{n(S)} = \frac{13}{52} = 0.25.$$

There is a $25\%$ chance that the card you draw is a club.

#### Part 3

Each of the four suits has three face cards (Jack, Queen, and King), so there are twelve face cards altogether. That means

Skip to main content

$$P(F) = \frac{n(F)}{n(S)} = \frac{12}{52} = 0.2308.$$

There is a $23.07\%$ chance that the card you draw is a face card.

### Part 4

We already know $n(S) = 52$, so the probability of drawing a card in the sample space is

$$P(S) = \frac{n(S)}{n(S)} = \frac{52}{52} = 1.$$

There is a $100\%$ chance that the card you draw is in the sample space. In retrospect, this is should be obvious since the sample space contains all the possible outcomes of an experiment, and when an experiment is performed, one of those possible outcomes is guaranteed to occur.

---

## Example 3.1.4

Table 3.1.1 describes the distribution of a random sample of $466$ individuals from the U.S., organized by political party and views on the circumstances under which abortion should be legal. One individual from the sample is selected at random.

*Table 3.1.1* The political party identification and views on the circumstances under which abortion should be legal of a sample of $466$ individuals.

| | Legal under any circumstances | Legal under certain circumstances | Illegal in all circumstances |
|---|---|---|---|
| **Republican** | 18 | 66 | 38 |
| **Independent** | 64 | 96 | 34 |
| **Democrat** | 76 | 62 | 12 |

1. Let event $D =$ the probability the individual is a Democrat. Find $P(D)$.
2. Let event $C =$ the probability that the individual believes abortion should be legal only under certain circumstances. Find $P(C)$.

### Solution

Note that since the individual is selected randomly, each individual has an equal probability of being chosen. So the outcomes are equally likely. And since there are $466$ individuals in the sample that could be selected, $n(S) = 466$.

### Part 1

To find the probability that the randomly chosen individual is a Democrat, we must first determine how many Democrats are in the sample. Looking at Table 3.1.1, we see that $76$ Democrats believe abortion should be legal under any circumstances, $62$ Democrats believe abortion should be legal only under certain circumstances, and $12$ Democrats believe abortion should be illegal under all circumstances. Adding these together, we see that the total number of Democrats in the sample is

$$n(D) = 76 + 62 + 12 = 150.$$

So the probability that the randomly chosen individual is a Democrat is

$$P(D) = \frac{n(D)}{n(S)} = \frac{150}{466} = 0.3219.$$

There is a $32.19\%$ chance that the individual we randomly choose from the sample is a Democrat.

Part 2

To find the probability that the randomly chosen individual believes abortion should be legal only under certain circumstances, we must first determine how many individuals with this belief are in the sample. Looking at Table 3.1.1, we see $66$ Republicans have this belief, $96$ Independents have this belief, and $62$ Democrats have this belief. Then the total number of individuals in the sample who believe abortion should be legal only under certain circumstances is

$$n(C) = 66 + 96 + 62 = 224.$$

Then the probability that the randomly chosen individual believes abortion should be legal only under certain circumstances is

$$P(C) = \frac{n(C)}{n(S)} = \frac{224}{466} = 0.4807.$$

There is a $48.07\%$ chance that the individual we randomly choose will believe that abortion should be legal in certain circumstances.

---

[1] Figure 3.1.1 was created by Saharasav and is licensed under the Creative Commons Attribution-Share Alike 4.0 International license.

[2] Figure 3.1.2 was adapted from an image created by Dmitry Fomin made available under the Creative Commons CC0 1.0 Universal Public Domain Dedication.

# 3.2. Conditional Probability

## Objectives

- Construct composite events using $\text{AND}$ and $\text{OR}$ and calculate their probability.
- Calculate conditional probability.

## $\text{AND}$ and $\text{OR}$ Events

Let $A$ and $B$ be any two events. Then $A$ $\text{AND}$ $B$ is an event with all the outcomes that are both in $A$ and in $B$. For example, suppose $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$. Then event $A$ $\text{AND}$ $B = \{3, 4\}$ since those two outcomes are the only outcomes in both event $A$ and event $B$.

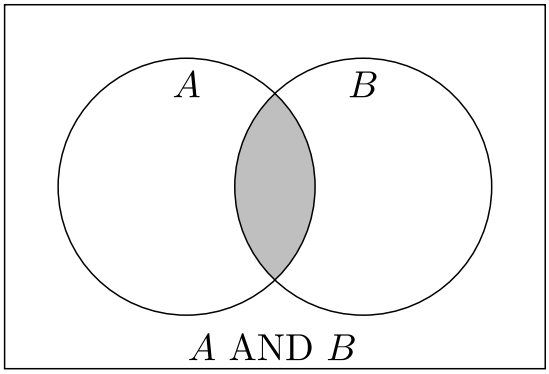



*Fig. 3.2.1* A venn diagram of two events A and B. The shaded portion of the diagram represents event $A$ AND $B$. It is the portion of the diagram that is both in $A$ and in $B$.

Similarly, $A$ OR $B$ is an event with all outcomes that are either in $A$ or in $B$ (or both). If $A = \{1, 2, 3, 4\}$ and $B = \{3, 4, 5, 6\}$, then event $A$ OR $B = \{1, 2, 3, 4, 5, 6\}$ since each of these outcomes are in $A$ or in $B$. Note that when an outcome is in both $A$ and $B$, it is still only listed once in $A$ OR $B$.

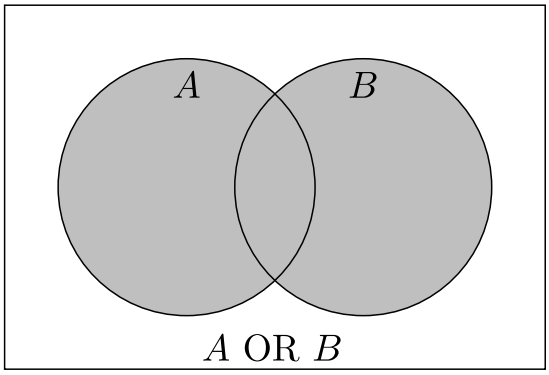



*Fig. 3.2.2* A venn diagram of two events A and B. The shaded portion of the diagram represents event $A$ OR $B$. It is the portion of the diagram that is in any part of $A$ or in any part of $B$.

---

Example 3.2.1

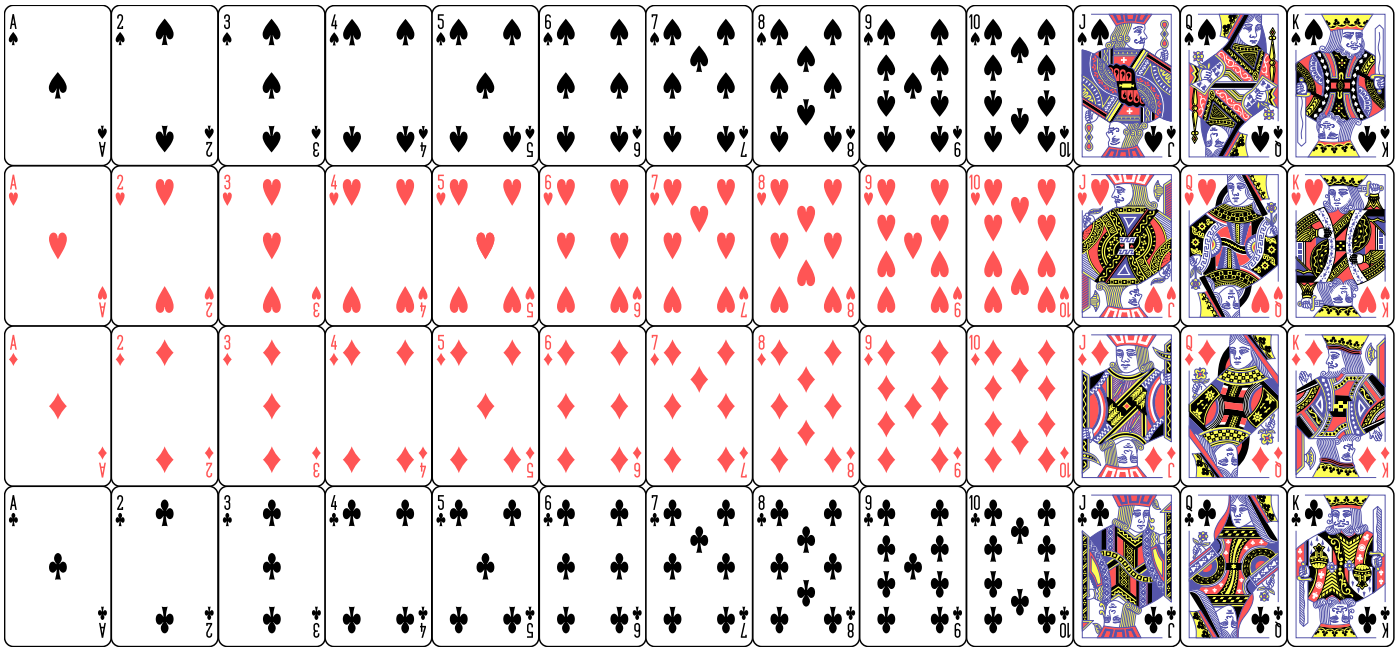


*Fig. 3.2.3* A standard deck of $52$ playing cards.[1]
A deck is divided into 4 suits: Spades (♠), Hearts (♥), Diamonds (♦), and Clubs (♣).
Spades (♠) and Clubs (♣) are black. Hearts (♥) and Diamonds (♦) are red.
There are 13 ranks in each suit: Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack, Queen, and King.
The Jacks, Queens, and Kings in a deck are called face cards.

Gloriana drew one card from a standard deck of $52$ cards.

1. What is the probability that Gloriana's card is both a Queen and a Heart?
2. What is the probability that Gloriana's card is a Queen or a Heart?

### Solution

Let $Q$ = the event where Gloriana draws a Queen and ♥ = the event where Gloriana draws a Heart. Note that since each card has an equal probability of being chosen, all outcomes are equally likely. And since there are $52$ cards in the deck, $n(S) = 52$.

### Part 1

We can express the probability that Gloriana's card is both a Queen and a Heart mathematically as $P(Q \text{ AND } ♥)$. Since all the outcomes are equally likely, we can calculate the probability using the formula

$$P(Q \text{ AND } ♥) = \frac{n(Q \text{ AND } ♥)}{n(S)}.$$

Looking at Figure 3.2.3, we can see that the Queen of Hearts is the only card that is both a Queen and a Heart, meaning $n(Q \text{ AND } ♥) = 1$. So the probability that Gloriana's card is both a Queen and a Heart is

$$P(Q \text{ AND } ♥) = \frac{n(Q \text{ AND } ♥)}{n(S)} = 1/52 = 0.0192.$$

Skip to main content

Part 2

The probability that Gloriana's card is a Queen or a Heart can be found using the formula

$$P(Q \text{ OR } \heartsuit) = \frac{n(Q \text{ OR } \heartsuit)}{n(S)}.$$

Looking at Figure 3.2.3, we can see the cards that are either a Queen or a Heart are

$$Q \text{ OR } \heartsuit = \{Q\spadesuit, Q\heartsuit, Q\diamondsuit, Q\clubsuit, A\heartsuit, 2\heartsuit, 3\heartsuit, 4\heartsuit, 5\heartsuit, 6\heartsuit, 7\heartsuit, 8\heartsuit, 9\heartsuit, 10\heartsuit, J\heartsuit, K\heartsuit\}.$$

(Note that we counted the Queen of Hearts only once.) This means $n(Q \text{ OR } \heartsuit) = 16$, so

$$P(Q \text{ OR } \heartsuit) = \frac{n(Q \text{ OR } \heartsuit)}{n(S)} = \frac{16}{52} = 0.3077.$$

---

## Conditional Probability When Outcomes are Equally Likely

The **conditional probability** of $A$ given $B$ is written $P(A|B)$. $P(A|B)$ is the probability that event $A$ is true given that we already know event $B$ is true.

For example, imagine you roll a six-sided die, and it rolls where you can't see it. You want to calculate the probability that the die came up a $4$, $5$, or $6$. Let $D = \{4, 5, 6\}$. Since you can't see the die, the probability of that you rolled a $4$, $5$, or $6$ is

$$P(D) = \frac{n(D)}{n(S)} = \frac{3}{6} = 0.5.$$

Now suppose your friend checks the die and tells you that you have rolled an even number. This extra information changes how you calculate the probability because now you know some outcomes are not possible. For example, it is no longer possible that you have rolled a $5$. Since you know the result is even, there are only **two** outcomes in $D = \{4, 5, 6\}$ that are still possible: $4$ and $6$. Similarly, of the outcomes in sample space $S = \{1, 2, 3, 4, 5, 6\}$, there are only **three** outcomes that are still possible: $2$, $4$, and $6$. If event $E$ is rolling an even number, then the probability that you rolled a $4$, $5$, or $6$ *given* that the number you rolled is even is

$$P(D|E) = \frac{2}{3} = 0.6666.$$

As this example demonstrates, when outcomes are equally likely, the conditional probability $P(A|B)$ is calculated almost like the usual probability $P(A)$: in both cases, we divide the possible number of outcomes in $A$ by the total number of possible outcomes. The difference is that the extra condition in $P(A|B)$ changes what outcomes are possible. If we know event $B$ is true, the only outcomes in $A$ that are possible are those outcomes that are also in $B$; that is, the outcomes in $A$ AND $B$ are the only outcomes in $A$ that are possible. And if we know event $B$ is true, the set of all possible outcomes is reduced from the entire sample space $S$ to just the outcomes in $B$. Putting these ideas together gives us the formula for $P(A|B)$ when outcomes are equally likely:

$$P(A|B) = \frac{n(A \text{ AND } B)}{n(B)}.$$

Skip to main content

**Warning**

The order we write events in matters when working with conditional probability: generally, $P(A|B) \neq P(B|A)$. Be careful that you use the right expression and formula for your application.

---

## Example 3.2.2

Table 3.2.1 describes the distribution of a random sample of $466$ individuals from the U.S., organized by political party and views on the circumstances under which abortion should be legal. One individual from the sample is selected at random.

*Table 3.2.1* The political party identification and views on the circumstances under which abortion should be legal of a sample of $466$ individuals.

| | Legal under any circumstances | Legal under certain circumstances | Illegal in all circumstances |
|---|---|---|---|
| **Republican** | 18 | 66 | 38 |
| **Independent** | 64 | 96 | 34 |
| **Democrat** | 76 | 62 | 12 |

1. Find the probability that the individual believes abortion should be illegal in all circumstances given that they are Republican.
2. Find the probability that the individual is a Republican given that they believe abortion should be illegal in all circumstances.

### Solution

Note that since the individual is selected randomly, each individual has an equal probability of being chosen. So the outcomes are equally likely. Let $I$ be the event where the individual believes abortion should be illegal in all circumstances, and let $R$ be the event where the individual is a Republican.

#### Part 1

We want to find $P(I|R)$. The formula for this conditional probability is

$$P(I|R) = \frac{n(I \text{ AND } R)}{n(R)}.$$

Looking at the table, there are $38$ individuals who believe abortion should be illegal in all circumstances and are Republican, so $n(I \text{ AND } R) = 38$. We also calculate that the total number of Republicans is $n(R) = 18 + 66 + 38 = 122$. So

$$P(I|R) = \frac{n(I \text{ AND } R)}{n(R)} = \frac{38}{122} = 0.3115.$$

#### Part 2

Skip to main content

$$P(R|I) = \frac{n(R \text{ AND } I)}{n(I)}.$$

Since $R$ AND $I$ is the same as $I$ AND $R$ (because the order of an AND event doesn't matter), our work in Part 1 means $n(R \text{ AND } I) = n(I \text{ AND } R) = 38$. And the total number of individuals who believe abortion should be illegal in all circumstances is $n(I) = 38 + 34 + 12 = 84$. So

$$P(R|I) = \frac{n(R \text{ AND } I)}{n(I)} = \frac{38}{84} = 0.4524.$$

---

## Conditional Probability When Outcomes are Not Equally Likely

When calculating conditional probability when outcomes are not all equally likely, we need to adapt the formula we learned above. The generalized formula for calculating conditional probability is

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

The formula we learned in the previous section for equally likely events is actually a special case of this more general formula. While the formula in the previous section only works when the outcomes are all equally likely, this formula works *always* in *all* circumstances.

---

### Example 3.2.3

Kristina plays basketball. Sometimes she must take two free throw shots. She makes the first shot $59\%$ of the time, and she makes both free throw shots $32\%$ of the time. What is the probability that she makes the second shot given that she made the first shot?

#### Solution

Let $F_1$ be the event where Kristina makes the first free throw shot, and let $F_2$ be the event where Kristina makes the second free throw shot. According to the problem statement, the probability that Kristina makes the first free throw shot is $P(F_1) = 0.59$, and the probability that she makes both free throw shots is $P(F_1 \text{ AND } F_2) = 0.32$. Then the probability that Kristina makes the second shot given that she made the first shot is

$$P(F_2|F_1) = \frac{P(F_2 \text{ AND } F_1)}{P(F_1)} = \frac{0.32}{0.59} = 0.5424.$$

---

**[1]** Figure 3.2.3 was adapted from an image created by Dmitry Fomin, which is made available under the Creative Commons CC0 1.0 Universal Public Domain Dedication.

## Objectives

- Use the multiplication rule to find the probability of $\text{AND}$ events.
- Identify when events are independent. When they are, use the simplified multiplication rule to find the probability of $\text{AND}$ events.
- Find the probability of samples chosen with replacement and chosen without replacement.

## The Multiplication Rule

Recall the general formulas for conditional probability:

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}, \qquad P(B|A) = \frac{P(A \text{ AND } B)}{P(A)}.$$

By multiplying $P(B)$ to the other side of the first equation and $P(A)$ to the other side of the second equation, we obtain the **multiplication rules**

$$P(A \text{ AND } B) = P(B) \cdot P(A|B), \qquad P(A \text{ AND } B) = P(A) \cdot P(B|A).$$

We can use either rule to calculate $P(A \text{ AND } B)$. We usually decide which rule to use based on what we already know. For example, if we only know $P(A)$, $P(B)$, and $P(B|A)$, it would be difficult to use the first rule to calculate $P(A \text{ AND } B)$ since we do not know $P(A|B)$. But we can easily use the second rule to calculate because we know $P(B|A)$ and $P(A)$.

---

### Example 3.3.1

Carlos plays college soccer. He makes a goal $65\%$ of the time he shoots. Carlos is going to attempt two goals in a row in the next game, and he tends to shoot in streaks. The probability that he makes the second goal *given* that he made the first goal is $0.90$. What is the probability that Carlos makes both goals?

#### Solution

Let's first translate this problem into math. Let $G_1 =$ the event Carlos is successful on his first goal attempt, and let $G_2 =$ the event Carlos is successful on his second goal attempt. Because Carlos makes a goal $65\%$ of the time he shoots,

$$P(G_1) = 0.65, P(G_2) = 0.65.$$

We are told that the probability that Carlos makes the second goal *given* that he made his first goal is $0.90$. Mathematically, that means

$$P(G_2|G_1) = 0.90.$$

We want to find the probability that Carlos makes both goals; that is, we want to find

$$P(G_1 \text{ AND } G_2).$$

Skip to main content

$$P(G_1 \text{ AND } G_2) = P(G_1) \cdot P(G_2|G_1) = 0.65 \cdot 0.90 = 0.585.$$

We calculate that Carlos has a $58.5\%$ chance of making both his first goal attempt and his second goal attempt.

---

## Independent Events

We say that events $A$ and $B$ are **independent** if:

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

In words, this means that whether or not event $B$ occurs has no impact on the probability of $A$; the likelihood that event $A$ happens is independent of $B$. If two events are *not* independent, then we say that they are **dependent**.

The day of the week and whether it rains are independent events: it's just as likely to rain on a Tuesday as it is to rain on a Friday. The outcome of the first event (which day of the week it is) does not change the probability of the second event (whether or not it rains).

Drawing one card, then drawing another card without putting the first card back is an example of two events that are dependent. If you draw an Ace of Spades for your first card, you can't draw an Ace of Spades for your second draw. The outcome of the first event (what card you draw first) changes the probability of the second event (what card you draw second).

Note, you need only show that **one** of the above equations is true to establish that two events are independent; if either one of the equations is true, then the other equation is guaranteed to be true as well.

Since $P(B|A) = P(B)$ when $A$ and $B$ are independent, our multiplication rule $P(A \text{ AND } B) = P(A) \cdot P(B|A)$ above simplifies to

$$P(A \text{ AND } B) = P(A) \cdot P(B)$$

in the case where $A$ and $B$ are independent. Note that this simplified multiplication rule works *only* when $A$ and $B$ are independent.

If it is not known whether A and B are independent or dependent, assume they are dependent until you can show otherwise.

---

### Example 3.3.2

A school fundraiser is selling chocolate bars and kettle corn. The probability that an adult will buy a chocolate bar when asked is $0.4$. The probability that an adult will buy a bag of kettle corn when asked is $0.6$. The probability than an adult will buy a chocolate bar *given* that they bought a bag of kettle corn is $0.4$.

For a randomly chosen adult, let $C =$ the adult buys a chocolate bar, and let $K =$ the adult buys a bag of kettle corn.

1. Are $C$ and $K$ independent events?
2. What is the probability that the adult buys both a chocolate bar and a bag of kettle corn?

Start by translating the problem into math. According to the problem statement:

$$\begin{aligned} P(C) &= 0.4, \\ P(K) &= 0.6, \\ P(C|K) &= 0.4. \end{aligned}$$

Part 1

Events $C$ and $K$ are independent if either $P(C|K) = P(C)$ or $P(K|C) = P(K)$. Since we know only $P(C|K)$, we will use the first equation to test if the events are independent. In this case,

$$P(C|K) = 0.4 = P(C).$$

Since $P(C|K) = P(C)$, events $C$ and $K$ *are* independent. The adult is no more or less likely to buy a chocolate bar if they buy kettle corn.

Part 2

Since $C$ and $K$ are independent, we can use the simpler multiplication rule for independent events to find $P(C \text{ AND } K)$. We calculate

$$P(C \text{ AND } K) = P(C) \cdot P(K) = 0.4 \cdot 0.6 = 0.24.$$

There is a $24\%$ chance that the adult will buy a chocolate bar and a bag of kettle corn.

**Note**

Because $C$ and $K$ are independent events, we used the simpler multiplication rule which works only for independent events,

$$P(C \text{ AND } K) = P(C) \cdot P(K).$$

However, we *could* have used the more complex general multiplication rule we learned at the beginning of the section,

$$P(C \text{ AND } K) = P(K) \cdot P(C|K),$$

which *always* works, including for independent events. If we use the general multiplication rule for this problem, we get

$$P(C \text{ AND} K) = P(K) \cdot P(C|K) = 0.6 \cdot 0.4 = 0.24,$$

which is the same answer we get using the simpler multiplication rule. Since both rules give the same result, we prefer to use the simpler rule when working with independent events.

## Example 3.3.3

Felicity attends Modesto Junior College in Modesto, CA. The probability that Felicity enrolls in a math class is $0.2$ and the probability that she enrolls in a speech class is $0.65$. The probability that she enrolls in a math class *given* that she enrolls in speech class is $0.25$.

Let $M =$ Felicity enrolls in math class, and let $S =$ Felicity enrolls in speech class.

1. Are $M$ and $S$ independent events?
2. What is the probability that Felicity enrolls in both math class and speech class?

### Solution

Start by translating the problem into math. Based on the problem statement:

$$\begin{aligned} P(M) &= 0.2, \\ P(S) &= 0.65, \\ P(M|S) &= 0.25. \end{aligned}$$

#### Part 1

Events $M$ and $S$ are independent if either $P(M|S) = P(M)$ or $P(S|M) = P(S)$. Since we know only $P(M|S)$, we will use the first equation to test if the events are independent. In this case,

$$P(M|S) = 0.25 \neq 0.2 = P(M).$$

Since $P(M|S) \neq P(M)$, events $M$ and $S$ are *not* independent. In this case, if Felicity enrolls in speech class, she is becomes more likely to enroll in math class.

#### Part 2

Since $M$ and $S$ are not independent events, we must use the general multiplication rule to calculate $P(M \text{ AND } S)$, not the simpler multiplication rule for independent events. Thus,

$$P(M \text{ AND } S) = P(S) \cdot P(M|S) = 0.65 \cdot 0.25 = 0.1625.$$

There is a $16.25\%$ chance that Felicity enrolls in both math class and speech class.

---

## Example 3.3.4

Esmeralda plays a game where she flips a coin then rolls a six-sided die. She wins if the coin comes up heads and the die comes up $4$. What is the probability that Esmeralda wins the game?

### Solution

Let $H =$ the event that the coin comes up heads and $F =$ the event that the die comes up $4$. We want to find $P(H \text{ AND } F)$. Note, though, that these two events are obviously independent: the outcome of the coin flip will not change the probability of the die

$$P(H \text{ AND } F) = P(H) \cdot P(F).$$

We calculate that $P(H) = \frac{1}{2}$ and $P(F) = \frac{1}{6}$, so that

$$P(H \text{ AND } F) = P(H) \cdot P(F) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12} = 0.0833.$$

So there is an $8.33\%$ chance that Esmeralda wins the game.

---

## Sampling With Replacement and Without Replacement

Sampling may be done **with replacement** or **without replacement**.

When sampling is done **with replacement**:

1. A member of the population is randomly selected to be in the sample.
2. The randomly selected member is replaced back in the population. The member may be sampled again, and the population is identical to how it was before the member was sampled.
3. The process is repeated until the desired sample size is obtained.

For example, suppose you want to draw two cards with replacement from a standard deck of $52$ cards. You draw your first card, take a look at it, then replace the card back in the deck (so the deck still has all $52$ cards in it). After reshuffling, you draw your second card. Note that because the card you drew first is still in the deck, it is possible that you will draw it again.

Because the population stays the same after each pick when sampling with replacement, the different picks are independent of each other. When calculating the probability of a sample taken with replacement, we can calculate the probability of each pick without considering the outcomes of the other picks.

When sampling is done **without replacement**:

1. A member of the population is randomly selected to be in the sample.
2. The randomly selected member is *not* replaced back in the population. The member may not be sampled again, and the population is different than it was before because it no longer includes the member that was sampled.
3. The process is repeated until the desired sample size is obtained.

For example, suppose you want to draw two cards without replacement from a standard deck of $52$ cards. You draw your first card but do *not* replace it back in the deck (so the deck now has only $51$ cards in it). When you draw your second card, the deck is a little different than it was when you drew your first card because now it has one less card in it. Note that because the card you drew first is no longer in the deck, it is not possible to draw the card again.

Because the population changes after each pick when sampling without replacement, the different picks are generally *not* independent of each other. When calculating the probability of a sample taken without replacement, to calculate the probability of each pick, we must be careful to take into account the outcomes of the picks that came before.

---

### Example 3.3.5

You have a well-shuffled deck of 52 cards. (See Figure 3.1.2.) You draw two cards from the deck. Find the probability that the two

1. with replacement.
2. without replacement.

Solution

Let $Q_1 =$ the event where the first card you draw is a Queen and $Q_2 =$ the event where the second card you draw is a Queen.

Part 1

You are drawing two cards *with replacement*, meaning you draw the first card from the deck, then put the card back in the deck and re-shuffle before drawing the second card. These two events are independent. Since there are $4$ Queens in the $52$-card deck, we calculate

$$P(Q_1 \text{ AND } Q_2) = P(Q_1) \cdot P(Q_2) = \frac{4}{52} \cdot \frac{4}{52} = \frac{16}{2704} = 0.0059.$$

There is a $0.59\%$ chance that the two cards you draw are both Queens if you draw the cards with replacement.

Part 2

You are drawing two cards *without replacement*, so we must assume the two events are dependent. First, note that $P(Q_1) = \frac{4}{52}$ since there are $4$ Queens and $52$ total cards in the deck. But since you don't replace the first drawn Queen back in the deck, there are only $3$ Queens and $51$ cards in the deck when you draw the second card, meaning $P(Q_2|Q_1) = \frac{3}{51}$. Therefore,

$$P(Q_1 \text{ AND } Q_2) = P(Q_1) \cdot P(Q_2|Q_1) = \frac{4}{52} \cdot \frac{3}{51} = \frac{12}{2652} = 0.0045.$$

So there is a $0.45\%$ chance that the two cards you draw are both Queens if you draw the cards without replacement.

# 3.4. Addition Rules and Mutually Exclusive Events

## Objectives

- Use the addition rule to find the probability of $\text{OR}$ events.
- Identify when events are mutually exclusive. When they are, use the simplified addition rule to find the probability of $\text{OR}$ events.
- Find complementary events and their probabilities.

## The Addition Rule

Given two events $A$ and $B$, we would like to know how to calculate $P(A \text{ OR } B)$, the probability that the outcome is in event $A$ or in event $B$. We know the probability that the outcome is in event $A$ is $P(A)$, and the probability that the outcome is in event $B$ is $P(B)$, so it may be reasonable to conclude that we can calculate $P(A \text{ OR } B)$ by simply adding $P(A)$ and $P(B)$. But this isn't quite right.

Skip to main content

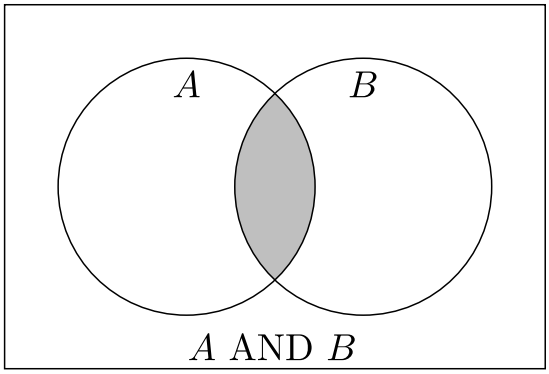



*Fig. 3.4.1* A venn diagram of event $A$ and event $B$. The shaded portion of the diagram represents event $A \text{ AND } B$. Outcomes that are both in $A$ and in $B$ would be in this intersection.

Observe from the Venn Diagram that generally event $A$ and event $B$ share some outcomes. The set of shared outcomes is event $A \text{ AND } B$. We count the set of shared outcomes $A \text{ AND } B$ once when we calculate $P(A)$, then *again* when we calculate $P(B)$, meaning we count the probability of the outcomes in $A \text{ AND } B$ *twice*. We only want to count each outcome *once*. To correct this, we need to subtract the excess: since we counted $P(A \text{ AND } B)$ twice instead of just once, we need to subtract $P(A \text{ AND } B)$ once to remove the excess. This gives us our **addition rule**:

$$P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B).$$

This equation can be expressed visually as

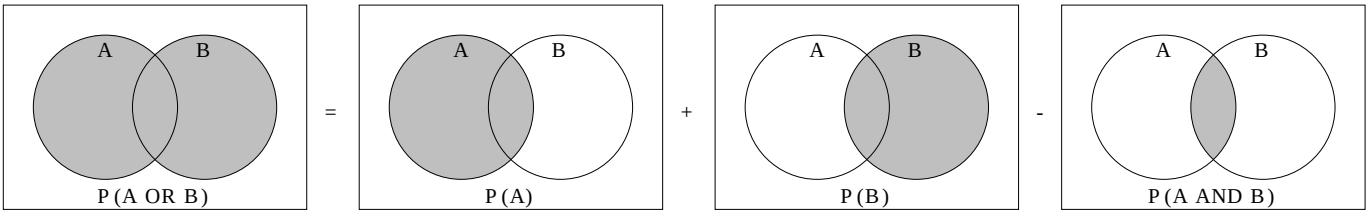



*Fig. 3.4.2* A diagram of the addition rule $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$. Notice event $A \text{ AND } B$, represented by the intersection of the two circles, is counted once when we calculate $P(A)$ and again when we calculate $P(B)$. Because we double count this event, we have to subtract $P(A \text{ AND } B)$ once to correct for the overcounting.

---

## Example 3.4.1

Carlos plays college soccer. He makes a goal $65\%$ of the time he shoots. Carlos is going to attempt two goals in a row in the next game, and he tends to shoot in streaks. The probability that he makes the second goal *given* that he made the first goal is $0.90$.

1. What is the probability that Carlos makes both goals?
2. What is the probability that Carlos makes the first goal or the second goal?

### Solution

Let's first translate this problem into math. Let $G_1 =$ the event Carlos is successful on his first goal attempt, and let $G_2 =$ the

Skip to main content

$$P(G_1) = 0.65, P(G_2) = 0.65.$$

We are told that the probability that Carlos makes the second goal *given* that he made his first goal is $0.90$. Mathematically, that means

$$P(G_2|G_1) = 0.90.$$

Part 1

We want to find the probability that Carlos makes both goals; that is, we want to find

$$P(G_1 \text{ AND } G_2).$$

Since we know $P(G_2|G_1)$ (and we don't know $P(G_1|G_2)$), we use the multiplication rule

$$P(G_1 \text{ AND } G_2) = P(G_1) \cdot P(G_2|G_1) = 0.65 \cdot 0.90 = 0.585.$$

We calculate that Carlos has a $58.5\%$ chance of making both his first goal attempt and his second goal attempt.

Part 2

We want to find $P(G_1 \text{ OR } G_2)$. Using the addition rule and the value of $P(G_1 \text{ AND } G_2)$ that we found in Part 1, we calculate

$$\begin{aligned} P(G_1 \text{ OR } G_2) &= P(G_1) + P(G_2) - P(G_1 \text{ AND } G_2) \\ &= 0.65 + 0.65 - 0.585 \\ &= 0.715. \end{aligned}$$

So there is a $71.5\%$ chance that Carlos makes his first goal or his second goal.

---

## Mutually Exclusive Events

$A$ and $B$ are **mutually exclusive** events if they cannot occur at the same time. This means that $A$ and $B$ do not share any outcomes. Thus, two events $A$ and $B$ are mutually exclusive if and only if $P(A \text{ AND } B) = 0$.

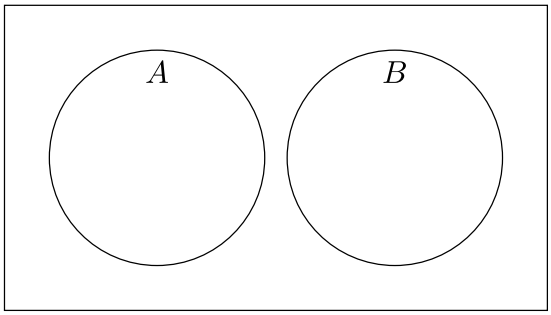



*Fig. 3.4.3* A venn diagram of two mutually exclusive events A and B. The circles representing the two events do not intersect, so $A$ and $B$ have no outcomes in common. This means $P(A \text{ AND } B) = 0$ since it is impossible for events $A$ and $B$ to occur at the same time.

Skip to main content

For example, imagine you roll a die. Let event $E$ be rolling an even number, and let event $D$ be rolling an odd number. Notice it is impossible for event $E$ and event $D$ to both occur at the same time since there is no outcome where the die comes up both an even number and an odd number at the same time. This means $P(E \text{ AND } D) = 0$, so events $E$ and $D$ are mutually exclusive.

As another example, suppose the sample space of some experiment is $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ where each outcome is equally likely. Let $I = \{1, 2, 3, 4, 5\}$, $J = \{4, 5, 6, 7, 8\}$, and $K = \{7, 9\}$.

- $I \text{ AND } J = \{4, 5\}$, so $P(I \text{ AND } J) = \frac{2}{10}$. Since $P(I \text{ AND } J) \neq 0$, events $I$ and $J$ are *not* mutually exclusive.
- $J \text{ AND } K = \{7\}$, so $P(J \text{ AND } K) = \frac{1}{10}$. Since $P(J \text{ AND } K) \neq 0$, events $J$ and $K$ are *not* mutually exclusive.
- $I \text{ AND } K = \{\}$, meaning events $I$ and $K$ have no outcomes in common and $I \text{ AND } K$ is empty. So $P(I \text{ AND } K) = \frac{0}{10} = 0$. Therefore, events $I$ and $K$ *are* mutually exclusive.

If two events $A$ and $B$ are mutually exclusive, then because $P(A \text{ AND } B) = 0$, the addition rule $P(A \text{ OR } B) = P(A) + P(B) - P(A \text{ AND } B)$ simplifies to

$$P(A \text{ OR } B) = P(A) + P(B).$$

Note that this simplified addition rule works *only* when $A$ and $B$ are mutually exclusive.

If it is not known whether two events $A$ and $B$ are mutually exclusive, assume they are not until you can show otherwise.

---

## Example 3.4.2

Keyarah draws one card from a standard deck of $52$ cards. (See Figure 3.1.2.) Let ♠ be event where the card is a spade, and let ♣ be the event where the card is a club.

1. What is the probability that Keyarah draws a spade? What is the probability that Keyarah draws a club?
2. Given that Keyarah draws a spade, what is the possibility that the card is also a club?
3. What is the probability that the card Keyarah draws is a spade and a club?
4. Are ♠ and ♣ mutually exclusive?
5. What is the probability that the card Keyarah draws is a spade or a club?

### Solution

#### Part 1

The outcomes of this experiment are all equally likely since each card in the deck has the same chance of being drawn. Since there are $13$ spades in a deck and $52$ cards total,

$$P(\spadesuit) = \frac{n(\spadesuit)}{n(S)} = \frac{13}{52} = 0.25.$$

There is a $25\%$ chance that Keyarah draws a spade.

Similarly, there are $13$ clubs in a deck, so

$$P(\clubsuit) = \frac{n(\clubsuit)}{n(S)} = \frac{13}{52} = 0.25.$$

There is a $25\%$ chance that Keyarah draws a club.

Part 2

Since the outcomes are all equally likely, the conditional probability can be found using the formula

$$P(\clubsuit|\spadesuit) = \frac{n(\clubsuit \text{ AND } \spadesuit)}{n(\spadesuit)}.$$

We know there are $13$ spades in the deck, so $n(\spadesuit) = 13$. But there aren't any cards that are both a club and a spade, so $n(\clubsuit \text{ AND } \spadesuit) = 0$. Thus

$$P(\clubsuit|\spadesuit) = \frac{n(\clubsuit \text{ AND } \spadesuit)}{n(\spadesuit)} = \frac{0}{13} = 0.$$

Part 3

Since the outcomes are all equally likely, we can calculate

$$P(\spadesuit \text{ AND } \clubsuit) = \frac{n(\spadesuit \text{ AND } \clubsuit)}{n(S)} = \frac{0}{52} = 0.$$

Alternatively, we could use our results from Part 1 and Part 2 to calculate this quantity using the multiplication rule:

$$P(\spadesuit \text{ AND } \clubsuit) = P(\spadesuit) \cdot P(\clubsuit|\spadesuit) = 0.25 \cdot 0 = 0.$$

Both methods are valid in this case, and both give us the same result: there is a $0\%$ chance that Keyarah draws a card that is both a spade and a club.

Part 4

Since $P(\spadesuit \text{ AND } \clubsuit) = 0$, events $\spadesuit$ and $\clubsuit$ are mutually exclusive.

Part 5

Because $\spadesuit$ and $\clubsuit$ are mutually exclusive, we can use the simplified addition rule:

$$P(\spadesuit \text{ OR } \clubsuit) = P(\spadesuit) + P(\clubsuit) = 0.25 + 0.25 = 0.5.$$

There is a $50\%$ chance that Keyarah draws a spade or a club.

---

Studies show that about one woman in seven (approximately $14.3\%$) who live to be $90$ will develop breast cancer. Suppose that of those women who develop breast cancer, a test is negative $2\%$ of the time; that is, the test gives a false negative result. Also suppose that in the general population of women, the test for breast cancer is negative about $85\%$ of the time. Let $B =$ woman develops breast cancer and let $N =$ tests negative. Suppose one woman is selected at random.

1. What is the probability that the woman develops breast cancer? What is the probability that woman tests negative?
2. Given that the woman has breast cancer, what is the probability that she tests negative?
3. What is the probability that the woman has breast cancer *and* tests negative?
4. Are $B$ and $N$ mutually exclusive?
5. What is the probability that the woman has breast cancer or tests negative?

Solution

Part 1

The probability that the randomly chosen woman develops breast cancer is $P(B) = 0.143$. The probability that the woman tests negative for breast cancer is $P(N) = 0.85$.

Part 2

Given that the woman has breast cancer, the probability that she tests negative is $P(N|B) = 0.02$.

Part 3

To find the probability that the woman has breast cancer *and* tests negative, we use the multiplication rule:

$$P(B \text{ AND } N) = P(B) \cdot P(N|B) = 0.143 \cdot 0.02 = 0.0029.$$

There is a $0.29\%$ chance that the woman has breast cancer and tests negative.

Part 4

Events $B$ and $N$ are *not* mutually exclusive since $P(B \text{ AND } N) = 0.0029 \neq 0$.

Part 5

Since $B$ and $N$ are not mutually exclusive, we need to use the more general addition rule:

$$P(B \text{ OR } N) = P(B) + P(N) - P(B \text{ AND } N) = 0.143 + 0.85 - 0.0029 = 0.9901.$$

---

## Example 3.4.4

An urn has 5 red marbles, 3 blue marbles, and 6 green marbles, and 6 yellow marbles. Seamus pulls one marble at random out of the urn. What is the probability that the marble is red or blue?

Solution

Let $R$ = the event of drawing a red marble and $B$ = the event of drawing a blue marble. First, note that $R$ and $B$ are mutually exclusive: the marble Seamus pulls out can't be both red and blue at the same time, so $P(R \text{ AND } B) = 0$. Thus,

$$P(R \text{ OR } B) = P(R) + P(B) = \frac{5}{20} + \frac{3}{20} = \frac{8}{20} = 0.4.$$

---

## Complementary Events

The **complement** of an event $A$ is the event where $A$ does not happen. The complement of $A$ is often written $A^c$. For example, if $R$ is the event that it rains tomorrow, then $R^c$ is the event that it does not rain tomorrow. If $W$ is the event that you win in a game of tic-tac-toe, then $W^c$ is the event that you do not win the game of tic-tac-toe (meaning either you lose or the game is a draw).

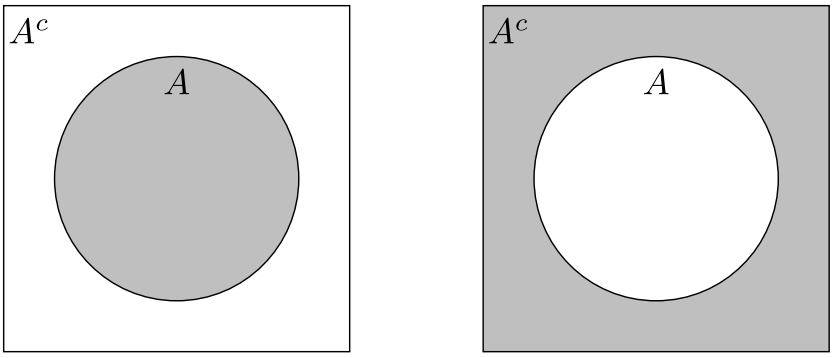



*Fig. 3.4.4* The shaded portion in the diagram on the left represents event $A$. The shaded portion in the diagram on the right represents event $A^c$. Notice that $A^c$ consists of the entire sample space that is not in $A$, so any outcome not in $A$ is in $A^c$ (and vice versa). Together, $A$ and $A^c$ fill the sample space.

Mathematically, $A^c$ is the event with all of the outcomes that are not in $A$. For example, let the sample space of an experiment be $S = \{1, 2, 3, 4, 5, 6, 7\}$. Let $E = \{2, 5, 6\}$. Then the complement of $E$ has every outcome not in $E$:

$$E^c = \{1, 3, 4, 7\}.$$

For two complementary events $A$ and $A^c$:

- $A$ and $A^c$ are mutually exclusive.
- Every possible outcome is in either $A$ or in $A^c$.

Because every possible outcome is either in $A$ or $A^c$, there is a $100\%$ chance that either $A$ is true or $A^c$ is true. That is,

$$P(A \text{ OR } A^c) = P(A) + P(A^c) = 1.$$

This observation is useful for calculating probabilities of complementary events.

For example, suppose you play a game where it is only possible to win or lose (so it is not possible for the game to end in a draw). Let $W$ be the event where you win, and let $L$ be the event where you lose. So in every outcome of the game where you do not win, you lose (and vice-versa). And since you cannot win and lose at the same time, events $W$ and $L$ are complementary. This

Now imagine there is a $32\%$ chance you lose the game, so $P(L) = 0.32$. Since $W$ is the complement of $L$, we can use the probability that you lose to find the probability that you win. By rearranging the equation $P(W) + P(L) = 1$, we get $P(W) = 1 - P(L)$. Using this formula, we calculate

$$P(W) = 1 - P(L) = 1 - 0.32 = 0.68.$$

So if there is a $32\%$ chance you lose the game, then there is a $68\%$ chance you win the game.

---

## Example 3.4.5

Letitia rolls a six-sided die. Let $D$ be the event where she rolls a $2$ or a $5$.

1. Find $D^c$.
2. Find $P(D)$.
3. Find $P(D^c)$.

### Solution

#### Part 1

Since the die has six sides, the sample space is $S = \{1, 2, 3, 4, 5, 6\}$. We are told that $D = \{2, 5\}$. The outcomes in $D^c$ are all the outcomes that are not in $D$:

$$D^c = \{1, 3, 4, 6\}.$$

#### Part 2

Since each number on the die is equally likely to come up, we calculate

$$P(D) = \frac{n(D)}{n(S)} = \frac{2}{6} = 0.3333.$$

There is a $33.33\%$ chance that Letitia rolls a $2$ or a $5$.

#### Part 3

We can calculate $P(D^c)$ in two ways.

In this case, it is easy to calculate the probability of $D^c$ directly. Since there are four outcomes in $D^c$,

$$P(D^c) = \frac{n(D^c)}{n(S)} = \frac{4}{6} = 0.6667.$$

Alternatively, since $D$ and $D^c$ are complementary events, we can use $P(D)$, which we already calculated, to find $P(D^c)$:

Skip to main content

$$P(D^c) = 1 - P(D) = 1 - 0.3333 = 0.6667.$$

Using either method, we find that the probability that Letitia does not roll a $2$ or a $5$ is $66.67\%$.

---

## Example 3.4.6

Muhammad plays a game where he draws three cards from a standard deck of $52$ cards without replacement. If all three cards are red, Muhammad wins. Otherwise, he loses.

1. Find the probability that Muhammad wins.
2. Find the probability that Muhammad loses.

### Solution

#### Part 1

Let $W$ be the event where Muhammad wins the game. Because the cards are drawn without replacement, the draws are not independent. We must be careful to take into consideration the results of previous draws when calculating the probability for each draw. Muhammad wins when the three cards he draws are all red.

For the first draw, the deck has all $52$ cards, including $26$ red cards. The probability that the first card Muhammad draws is red is $\frac{26}{52}$.

For the second draw, we assume Muhammad has already drawn one red card from the deck. Then the deck has only $25$ red cards left and $51$ total cards left, so the probability that the second card drawn is a King is $\frac{25}{51}$.

Finally, for the third card drawn, we assume Muhammad has already drawn two red cards from the deck. Since Muhammad has already drawn two red cards, there are $24$ red cards and $50$ total cards left in the deck, so the probability that the third card drawn is a red card is $\frac{24}{50}$.

So the probability that Muhammad wins by drawing three red cards from the deck is

$$P(W) = \frac{26}{52} \cdot \frac{25}{51} \cdot \frac{24}{50} = 0.1176.$$

There is an $11.76\%$ chance that Muhammad wins the game.

#### Part 2

Let $L$ be the event where Muhammad loses the game. We *could* try to calculate $P(L)$ directly, but it would take a lot of work because there is a lot of ways for Muhammad to lose. Muhammad loses if any of the following happens:

- He draws a black card, then a black card, then a black card.
- He draws a red card, then a black card, then a black card.
- He draws a black card, then a red card, then a black card.
- He draws a black card, then a black card, then a red card.
- He draws a red card, then a red card, then a black card.

Skip to main content

- He draws a black card, then a red card, then a red card.

To calculate $P(L)$ directly, we would need to calculate the probability for all of these possible ways for Muhammad to lose, then add them together. That would take a lot of work.

Instead, note that for this game, Muhammad can only win or lose. The game cannot end in a draw or a tie. That means $W$ and $L$ are complementary events: Muhammad wins if he doesn't lose ($W = L^c$), and he loses if he doesn't win ($L = W^c$). So we can use $P(W)$, which we already calculated, to quickly and easily find $P(L)$:

$$P(L) = 1 - P(W) = 1 - 0.1176 = 0.8824.$$

There is an $88.24\%$ chance that Muhammad loses the game.

# 3.5. Discrete Probability Distributions

## Objectives

- Identify and Construct discrete probability distributions.
- Find the expected value of discrete probability distributions.
- Find the standard deviation of discrete probability distributions.

## Discrete Probability Distributions

A **probability distribution** is a table, formula, or rule that gives the probability for each outcome of an experiment. When the possible outcomes of an experiment only take on discrete values (that is, the outcomes can take on only certain values and not the values in between), we say the experiment has a **discrete probability distribution**. Almost all of the experiments we have studied in this text so far have **discrete probability distributions**.

For example, the probability distribution of rolling a six-sided die is described by Table 3.5.1. Note that the outcomes of rolling a die can only take on certain discrete value (specifically, the numbers $1$, $2$, $3$, $4$, $5$, and $6$), but not the values in between (like $2.457$).

*Table 3.5.1* The discrete probability distribution for rolling a fair six-sided die.

| $X$ | $P(X)$ |
|---|---|
| $1$ | $\frac{1}{6} = 0.1667$ |
| $2$ | $\frac{1}{6} = 0.1667$ |
| $3$ | $\frac{1}{6} = 0.1667$ |
| $4$ | $\frac{1}{6} = 0.1667$ |
| $5$ | $\frac{1}{6} = 0.1667$ |
| $6$ | $\frac{1}{6} = 0.1667$ |

In contrast, if the outcome of an experiment can take on take on the value of any fraction, decimal, or irrational number within the range of allowed values, we say the experiment has a **continuous probability distribution**. For example, if we conduct an experiment where we measure the height of a randomly chosen person, the outcome could be that the individual is $64.41$ inches tall, or maybe $72.683$ inches tall, or it could be any decimal value in between. We will learn more about continuous probability distributions later.

A **discrete probability distribution** has two characteristics:

1. Since each outcome can't have less than a $0\%$ chance of happening or more than a $100\%$ chance of happening, the probability of each outcome must be between $0$ and $1$.
2. Since one of the outcomes is guaranteed to happen, the sum of the probabilities of all the outcomes must be $1$.

---

## Example 3.5.1

A child psychologist is interested in the number of times a newborn baby's crying wakes its mother after midnight. For a random sample of $50$ mothers, the information in Table 3.5.2 was obtained. Does Table 3.5.2 represent a discrete probability distribution?

*Table 3.5.2* Relative frequency table for the number of times a newborn baby's cry wakes a sample of $50$ mothers after midnight.

| Number of times cry wakes mother after midnight | Relative Frequency |
|---|---|
| $0$ | $\frac{2}{50} = 0.04$ |
| $1$ | $\frac{11}{50} = 0.22$ |
| $2$ | $\frac{23}{50} = 0.46$ |
| $3$ | $\frac{9}{50} = 0.18$ |
| $4$ | $\frac{4}{50} = 0.08$ |
| $5$ | $\frac{1}{50} = 0.02$ |

Solution

Table 3.5.2 represents a discete probability distribution because:

1. Each outcome has a probability between $0$ and $1$.
2. The sum of the probabilities of all the outcomes is $1$:

$$\frac{2}{50} + \frac{11}{50} + \frac{23}{50} + \frac{9}{50} + \frac{4}{50} + \frac{1}{50} = 1$$

This means that if we select one of the $50$ mothers at random, the table tells us the probability of each possible outcome. For example, the table tells us that the probability that the selected mother was woken up twice after midnight by her crying baby is $0.46$.

---

## Example 3.5.2

Suppose Nancy has classes **three days** a week. She attends classes **three days a week 80%** of the time, **two days 15%** of the time, **one day 4%** of the time, and **no days 1%** of the time. Suppose one week is randomly selected.

1. What is $X$?
2. $X$ can take on what values?
3. Construct a probability distribution table like the one in example 4.1. The table should have two columns labeled $X$ and $P(X)$. What does the $P(X)$ column sum up to?

Solution

Part 1

$X =$ the number of days Nancy attends class.

Part 2

$X$ can take on the values $0$, $1$, $2$, or $3$.

Part 3

*Table 3.5.3* A discrete probability distribution for the number of days of the week that Nancy attends class.

| $X$ | $P(X)$ |
|---|---|
| $0$ | $0.01$ |
| $1$ | $0.04$ |
| $2$ | $0.15$ |
| $3$ | $0.80$ |

The $P(X)$ column sums up to:

$$0.01 + 0.04 + 0.15 + 0.80 = 1.00.$$

This is expected. If the probabilities did not add up to one, Table 3.5.3 wouldn't be a discrete probability distribution.

## Expected Value

The **expected value** is often referred to as the **"long-term" average** or **mean**. This means if you repeat an experiment many times, over the long term you would **expect** this average.

Recall that the **Law of Large Numbers** states that as the number of trials in a probability experiment increases, the theoretical probability of an event and the proportion of experiments where that event occurs gets closer and closer.

For example, when you roll a six-sided die, probabiliy says you should expect each face of the die to come up in $\frac{1}{6}$ of the rolls. But if you only roll the die $6$ times, we wouldn't be shocked if you rolled $2$ fives (meaning you rolled a five in $\frac{2}{6}$ of the rolls). However, if you rolled the die $6{,}000{,}000$ times, we would expect you to roll a five in very nearly $\frac{1}{6}$ of the rolls. As the number of die rolls increases, the proportion of fives that you roll tends to get closer and closer to the probability of rolling a five.

Skip to main content

When evaluating the long-term results of statistical experiments, we often want to know the "average" outcome. This "long-term average" is known as the **mean** or **expected value** of the experiment and is denoted by the Greek letter $\mu$. In other words, after conducting many trials of an experiment, you would expect this average value.

The expected value of a discrete probability distribution function can be found using the formula

$$\mu = \sum x \cdot P(x).$$

In words, we multiply each value $x$ by the probability $P(x)$ that the value will occur. Then we add up the products $x \cdot P(x)$.

To illustrate this idea, again return to the idea of rolling a six-sided die. The possible outcomes of rolling a die are $x = 1, 2, 3, 4, 5, 6$, and the probability of each outcome is $\frac{1}{6}$. Then the expected value is

$$\begin{aligned}\mu &= \sum x \cdot P(x) \\ &= 1 \cdot P(1) + 2 \cdot P(2) + 3 \cdot P(3) + 4 \cdot P(4) + 5 \cdot P(5) + 6 \cdot P(6) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \\ &= 3.5,\end{aligned}$$

meaning if we rolled the die many times, we would expect the average of all the rolls to be close to $3.5$.

We can use R to do this same calculation more quickly.

```
x = c(1, 2, 3, 4, 5, 6)
Px = c(1/6, 1/6, 1/6, 1/6, 1/6, 1/6)

mu = sum(x * Px)
mu
```

3.5

---

## Example 3.5.3

A men's soccer team plays soccer zero, one, or two days a week. The probability that they play zero days is $0.2$, the probability that they play one day is $0.5$, and the probability that they play two days is $0.3$. On average, how many games per week would we expect the team to play next year?

### Solution

We need to calculate the expected value, which estimates the average number of games the team plays per week this next year. The discrete probability distribution is given in Table 3.5.4, where $X$ is the number of games played per week.

*Table 3.5.4* A discrete probability distribution for the number of days the soccer team plays in a week.

| $X$ | $P(X)$ |
|---|---|
| $0$ | $0.2$ |
| $1$ | $0.5$ |
| $2$ | $0.3$ |

We can use R to find the expected value.

```
x = c(0, 1, 2)
Px = c(0.2, 0.5, 0.3)

mu = sum(x*Px)
mu
```

1.1

The mean or expected value of the distribution is $\mu = 1.1$ games per week. That is, we expect the soccer team to play an average of $1.1$ games per week next year.

---

## Example 3.5.4

Suppose you play a game of chance in which five numbers are randomly chosen from $0$ to $9$ by a computer. If you match all five numbers in order, you win $\$100,000$. If you lose, you pay $\$2$. Over the long term, what is your expected profit per game?

### Solution

We need to calculate the expected value because it will tell us how much we expect to profit per game on average.

Let's first construct a table of the discrete probability distribution. Let $X$ be the amount of money you profit from playing the game. The values that $X$ can take on are $x = 100{,}000$ and $x = -2$. (Note, since we are ultimately interested in the profit, $X$ does not take on the values of the randomly chosen numbers.)

Now let's look at the probability of winning the game. Let

- $N_1 =$ the event you pick the right 1st number
- $N_2 =$ the event you pick the right 2nd number

Skip to main content

- $N_4 =$ the event you pick the right 4th number
- $N_5 =$ the event you pick the right 5th number

The probability that you win is

$$P(x = 100,000) = P(N_1 \text{ AND } N_2 \text{ AND } N_3 \text{ AND } N_4 \text{ AND } N_5)$$

Note that since the numbers are chosen *with replacement*, $N_1$, $N_2$, $N_3$, $N_4$, and $N_5$ are all independent events. Thus

$$\begin{aligned} P(N_1 \text{ AND } N_2 \text{ AND } N_3 \text{ AND } N_4 \text{ AND } N_5) &= P(N_1) \cdot P(N_2) \cdot P(N_3) \cdot P(N_4) \cdot P(N_5) \\ &= \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \\ &= \frac{1}{10^5} \\ &= 0.00001 \end{aligned}$$

So the probability that you win is $P(100{,}000) = 0.00001$. Since you can only win or lose at this game (the game can't end in a draw), the probability that you lose is

$$P(-2) = 1 - P(100{,}000) = 1 - 0.00001 = 0.99999.$$

Table 3.5.5 describes the probability distribution for this game.

*Table 3.5.5* A discrete probability distribution for the amount of money you win or lose playing the game.

| $X$ | $P(X)$ |
|---|---|
| $100{,}000$ | $0.00001$ |
| $-2$ | $0.99999$ |

Now use R to calculate the expected value.

```
x = c(100000, -2)
Px = c(0.00001, 0.99999)

mu = sum(x*Px)
mu
```

-0.99998

The expected value is $\mu = -\$0.99998$. Since the expected value is negative, this means that if you played this game many times, you would lose an average of almost $1 per game.

The standard deviation $\sigma$ of a discrete probability distribution is

$$\sigma = \sqrt{\sum [(x-\mu)^2 P(x)]},$$

where $\mu$ is the expected value of the distribution. The standard deviation tells us the standard (or average) amount the outcome of an experiment deviates from the expected value.

---

## Example 3.5.5

Patients who have had an appendectomy must stay in the hospital for $1$, $2$, or $3$ days. Patients stay in the hospital for $1$ day $40\%$ of the time, for $2$ days $55\%$ of the time, and for $3$ days $5\%$ of the time.

1. How long should a patient expect to stay in the hospital?
2. What is the standard deviation of the distribution?

### Solution

Let's first construct the probability distribution. Let $X$ be the number of days the patient stays in the hosplital. The possible values of $X$ are $x = 1, 2, 3$. Table 3.5.6 is the probability distribution for the number of days appendectomy patients stay in the hospital after surgery.

*Table 3.5.6* A discrete probability distribution for the number of days appendectomy patients stay in the hospital after surgery.

| $X$ | $P(X)$ |
|---|---|
| $1$ | $0.40$ |
| $2$ | $0.55$ |
| $3$ | $0.05$ |

### Part 1

The expected value gives the best estimate for the number of days a patient should expect to stay in the hospital.

```
x = c(1, 2, 3)
Px = c(0.40, 0.55, 0.05)

mu = sum(x*Px)
mu
```

1.65

The expected value is $\mu = 1.65$. A patient should expect to spend, on average, $1.65$ days in the hospital after an appendectomy.

### Part 2

We can use R to calculate the standard deviation.

```
sigma = sqrt( sum( (x - mu)^2 * Px ) )
sigma
```

0.57227615711298

So the standard deviation is $\sigma = 0.572$. This means it is not uncommon for a patient to spend $0.572$ days less or $0.572$ days more than the expected $1.65$ days in the hospital after surgery

---

## Example 3.5.6

A hospital researcher is interested in the number of times the average post-op patient will ring the nurse during a $12$-hour shift. For a random sample of $50$ patients, Table 3.5.7 was obtained.

*Table 3.5.7* A relative frequency table for the number of times $50$ patients rings the nurse in a $12$-hour shift.

| Number of Times Patient Rings Nurse | Relative Frequency |
|---|---|
| $0$ | $\frac{4}{50}$ |
| $1$ | $\frac{8}{50}$ |
| $2$ | $\frac{16}{50}$ |
| $3$ | $\frac{14}{50}$ |
| $4$ | $\frac{6}{50}$ |
| $5$ | $\frac{2}{50}$ |

1. Find the mean.
2. Find the standard deviation.

### Solution

First, note that Table 3.5.7 is a discrete probability distribution, where $X$ is the number of times the patient rings the nurse, and

Part 1

Remember, for a probability distribution, the mean and the expected value are the same thing.

```
x = c(0, 1, 2, 3, 4, 5)
Px = c(4/50, 8/50, 16/50, 14/50, 6/50, 2/50)

mu = sum(x*Px)
mu
```

2.32

The mean or expected value is $\mu = 2.32$. This means, on average, we expect a patient to ring the nurse $2.32$ times during a $12$-hour shift.

Part 2

```
sigma = sqrt( sum( (x - mu)^2 * Px ) )
sigma
```

1.22376468326227

The standard deviation is $\sigma = 1.2238$. This means that it is common for a patient to ring the nurse $1.2238$ times less or $1.2238$ times more than the expected $2.32$ rings.

# 3.6. The Binomial Distribution

## Objectives

- Identify experiments that fit a binomial probability distribution.
- Calculate the expected value and standard deviation for a binomial distribution.
- Calculate probabilities using a binomial distribution.

## The Binomial Distribution

There are three characteristics of a **binomial experiment**.

- There are a fixed number of trials. Think of trials as repetitions of an experiment. The letter $n$ denotes the number of trials.

  For example, if we conduct a binomial experiment by rolling a six-sided die ten times, then the binomial experiment has $n = 10$ trials.

- There are only two possible outcomes, called “success” and “failure,” for each trial. The letter $p$ denotes the probability of a success for one trial. Since “success” and “failure” are complementary outcomes, the probability of failure is $(1 - p)$.

  Imagine we roll a six-sided die ten times, and we want to know many twos we roll. Then each of the $n = 10$ trials is counted a “success” when we roll a two and is a “failure” when we roll anything that isn’t a two. In this case, the probability a trial is a

Skip to main content

success is $p = \frac{1}{6}$, and the probability a trial is a failure is $(1 - p) = (1 - \frac{1}{6}) = \frac{5}{6}$.

- The $n$ trials of a binomial experiment are independent and are repeated using identical conditions. Because the $n$ trials are independent and identical, the probability of success $p$ and probability of failure $(1 - p)$ are the same for each trial.

  In our binomial experiment of $n = 10$ die rolls, we have a probabiliy of $p = \frac{1}{6}$ of rolling a two on the second trial no matter what happened during the first trial. The probability of success and the probability of failure is the same for every roll of the die.

The outcomes of a binomial experiment fit a **binomial probability distribution**, which is a special kind of discrete probability distribution. A binomial distribution is a formula that gives the probability that there were exactly $x$ successes in the $n$ trials, where we can choose any whole number $x$ between $0$ and $n$.

The expected value of a binomial distribution is

$$\mu = n \cdot p.$$

The standard deviation of a binomial distribution is

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)}.$$

**Note**

Be aware that the success of a trial in a binomial experiment isn't necessarily good, and the failure of a trial isn't necessarily bad. Rather, we call the outcome we are trying to measure a success regardless of whether it is good or bad.

For example, we may consider a test for cancer a success if cancer is detected since that is what we are trying to measure, even though a cancer diagnosis is definitely bad.

---

## Example 3.6.1

At ABC College, the withdrawal rate from an elementary physics course is $30\%$ for any given term. This implies that, for any given term, $70\%$ of the students stay in the class for the entire term. We track $20$ students in an elementary physics course to see if they withdraw from the course. Assume the probability that each student withdraws is independent from the other students.

1. Is this a binomial experiment? Why or why not?
2. How many students are expected to withdraw from the course?
3. What is the standard deviation?

### Solution

#### Part 1

There are three conditions an experiment must satisfy to be a binomial experiment.

1. Are there are a fixed number of trials? Yes. In this case, we are looking to see if each of the $20$ students withdraws from the

2. Does each trial only have two possible outcomes? Yes. Each student will either withdraw or not withdraw from the course.
3. Are the trials independent and identical? Yes. We are told in the prompt to assume that the probability that each student withdraws is independent from the other students, so the probability that a student will withdraw is identical for each student.

Since these three conditions are met, this is a binomial experiment.

### Part 2

To find how many students are expected to withdraw from the course, we need to find the expected value. Since there are $20$ students, the experiment consists of $n = 20$ trials. Since we are interested in how many students will withdraw from the course, we will call a trial a success if the student withdraws from the course. So the probability of success is $p = 0.30$. So the expected value is

$$\mu = n \cdot p = 20 \cdot 0.30 = 6.$$

We expect $6$ of the $20$ students to drop the course.

### Part 3

Since the probability of success of any one trial is $p = 0.30$, the probability of failure is $(1 - p) = (1 - 0.30) = 0.70$. Then the standard deviation of the binomial distribution is

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)} = \sqrt{20 \cdot 0.30 \cdot 0.70} = 2.0494.$$

---

## Example 3.6.2

Suppose you play a game where you draw a single card from a well-shuffled $52$-card deck. You win the game if you draw a Spade. If you do not draw a spade, you lose. You play the game $50$ times.

1. Is this a binomial experiment? Why or why not?
2. How many times do you expect to win the game?
3. What is the standard deviation?

### Solution

### Step 1

There are three conditions an experiment must satisfy to be a binomial experiment.

1. Are there are a fixed number of trials? Yes. In this case, you play the game $50$ times, so the experiment has $n = 50$ trials.
2. Does each trial only have two possible outcomes? Yes. Each time you play the game, you can either win or lose.
3. Are the trials independent and identical? Yes. Each time you play the game, you are drawing from a full deck of $52$ cards. Since you play with the full deck each game, the games are independent and identical.

Step 2

We want to find the expected value. There are $n = 50$ trials since you play the game $50$ times. Since we are interested in how many times you win the game, a trail is a success if you win. Since there are $13$ spades in a deck of $52$ cards, the probability of success is $p = \frac{13}{52} = 0.25$. Then the expected value is

$$\mu = n \cdot p = 50 \cdot 0.25 = 12.5.$$

We expect you to win $12.5$ of the $50$ games.

Step 3

Since the probability of success is $p = 0.25$, the probability of failure is $(1 - p) = (1 - 0.25) = 0.75$. Then the standard deviation is

$$\sigma = \sqrt{n \cdot p \cdot (1 - p)} = \sqrt{50 \cdot 0.25 \cdot 0.75} = 3.0619.$$

---

## Notation and Calculation

When we want to say that $X$ is a random variable with a binomial distribution, we write

$$X \sim B(n, p)$$

where $n$ is the number of trials and $p$ is the probability of success of any one trial.

We can calculate the the probability of a binomial distribution using the R function

```
dbinom(x, size, prob)
```

where `x` is the number of successes we want to calculate the probability of obtaining, `size` is the number of trials $n$, and `prob` is the probability of success $p$ of one trial. Alternatively, `x` could be a list of values, in which case the function will return a list of probabilities, one for each value in `x`.

---

### Example 3.6.3

It has been stated that about $41\%$ of adult workers have a high school diploma but do not pursue any further education. Suppose $20$ adult workers are randomly selected.

1. Find the probability that exactly 7 of the 20 workers have a high school diploma but do not pursue further education.
2. Find the probability that at most 4 workers have a high school diploma but do not pursue further education.
3. Find the probability that more than 15 workers have a high school diploma but do not pursue further education.
4. Find the probability that more than 8 but less than 18 workers have a high school diploma but do not pursue further education.

For this problem, $X =$ the number of workers, of the $20$ selected, that have a high school diploma but do not pursue further education. The number of trials is $n = 20$, and the probability of "success" is $p = 0.41$. So

$$X \sim B(20, 0.41)$$

Part 1

We want to find $P(X = 7)$. We will use the `dbinom` function.

```
dbinom(x = 7, size = 20, prob = 0.41)
```

0.158480894003918

The probability that exactly $7$ of the $20$ adults have a high school diploma but do not pursue further education is $P(X = 7) = 0.1585$, or $15.85\%$.

Part 2

We want to find $P(X \leq 4)$. In other words, we want to know the probability that $x$ is $0$ or $1$ or $2$ or $3$ or $4$. We will again use the `dbinom` function, but instead of passing a single value, we will pass a list of values. Let's first see how this works.

```
values = c(0, 1, 2, 3, 4)
dbinom(x = values, size = 20, prob = 0.41)
```

2.6124033550459e-05 · 0.000363079788328412 · 0.00239694199243926 · 0.00999402932440777 · 0.0295162645725093

This gives us a list of $5$ probabilities, one for each value we passed. For example, we can see from the list that $P(X = 4) = 0.0295$.

But these values are all mutually exclusive (because, for example, we can't have at the same time $2$ and $3$ adults with a high school diploma but who do not pursue higher education; it can be either $2$, or $3$, not both). So

$$P(X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4);$$

that is, we just need to add up these probabilities. We can do this quickly using the `sum` function.

```
values = c(0, 1, 2, 3, 4)
probs = dbinom(x = values, size = 20, prob = 0.41)
sum(probs)
```

0.0422964397112352

So $P(X \leq 4) = 0.0423$. That is, there is a $4.23\%$ chance that no more than $4$ adults of the $20$ selected have a high school diploma but do not pursue higher education.

Part 3

We want $P(X > 15)$. "More than $15$" means that $16$ or $17$ or $18$ or $19$ or $20$ of the adults have a high school diploma but do not pursue further education. (Note, we stop at $20$ because that's the total number of adults. We can't have more successes than we have trials.)

```
values = c(16, 17, 18, 19, 20)
probs = dbinom(x = values, size = 20, prob = 0.41)
sum(probs)
```

0.000443150625878507

So $P(x > 15) = 0.0004$. There is only a $0.04\%$ chance that more than $15$ of the $20$ adults selected have a high school diploma but do not pursue further education.

### Part 4

We want $P(8 < x < 18)$.

```
values = c(9, 10, 11, 12, 13, 14, 15, 16, 17)
probs = dbinom(x = values, size = 20, prob = 0.41)
sum(probs)
```

0.440587402134619

So $P(8 < x < 18) = 0.4406$. There is a $44.06\%$ chance that between $8$ and $18$ adults of the $20$ selected have a high school diploma but do not pursue futher education.

---

## Example 3.6.4

The lifetime risk of developing pancreatic cancer is about $1.28\%$. Suppose we randomly sample $200$ people. Let $X =$ the number of people who will develop pancreatic cancer.

1. Is it more likely that five people or six people of the $200$ people in the sample will develop pancreatic cancer? Justify your answer numerically.
2. Find the probability that less than eight people develop pancreatic cancer.

### Solution

For this problem, $X \sim B(200, 0.0128)$.

### Part 1

We need to find $P(X = 5)$ and $P(X = 6)$ and see which probability is greater.

```
values = c(5, 6)
dbinom(x = values, size = 200, prob = 0.0128)
```

0.0706549781513899 · 0.0297735726407802

Skip to main content

So $P(x = 5) = 0.0707$ and $P(x = 6) = 0.0298$. Since the probability is greater when $x = 5$ than when $x = 6$, it is more likely that $5$ of the $200$ people develop pancreatic cancer than it is for $6$ of the $200$ people to develop pancreatic cancer.

Part 2

We want to find $P(X < 8)$.

```
values = c(0, 1, 2, 3, 4, 5, 6, 7)
probs = dbinom(x = values, size = 200, prob = 0.0128)
sum(probs)
```

0.995434178943854

So $P(x < 8) = 0.9954$. There is a 99.54% chance that fewer than $8$ of the $200$ people sampled develop pancreatic cancer.

# 3.7. Continuous Probability Distributions

## Objectives

- Identify continuous probability density functions.
- Calculate probabilities by finding the areas under probability density functions.

## Continuous Probability Distributions

If the outcome of an experiment can take on take on the value of any fraction, decimal, or irrational number within the range of allowed values, we say the experiment has a **continuous probability distribution**. For example, if we conduct an experiment where we measure the height of a randomly chosen person, the outcome could be that the individual is $64.41$ inches tall, or maybe $72.683$ inches tall, or it could be any decimal value in between.

A continuous probability distribution can be described using a **probability density function**. The graph of a probability density function can be used to tell us how likely any range of outcomes is likely to be. Specifically, the area under the curve within a range of values is the probability that the outcome will be within that range of values. In short, when dealing with probability density functions,

$$\text{Probability} = \text{Area}.$$

Because you can't have less than $0\%$ probability, the graph of a probability density function can only take on non-negative values above the $x$-axis. And since the total combined probability of all outcomes in an experiment is $100\%$, the total area under a probability density function is always $1$.

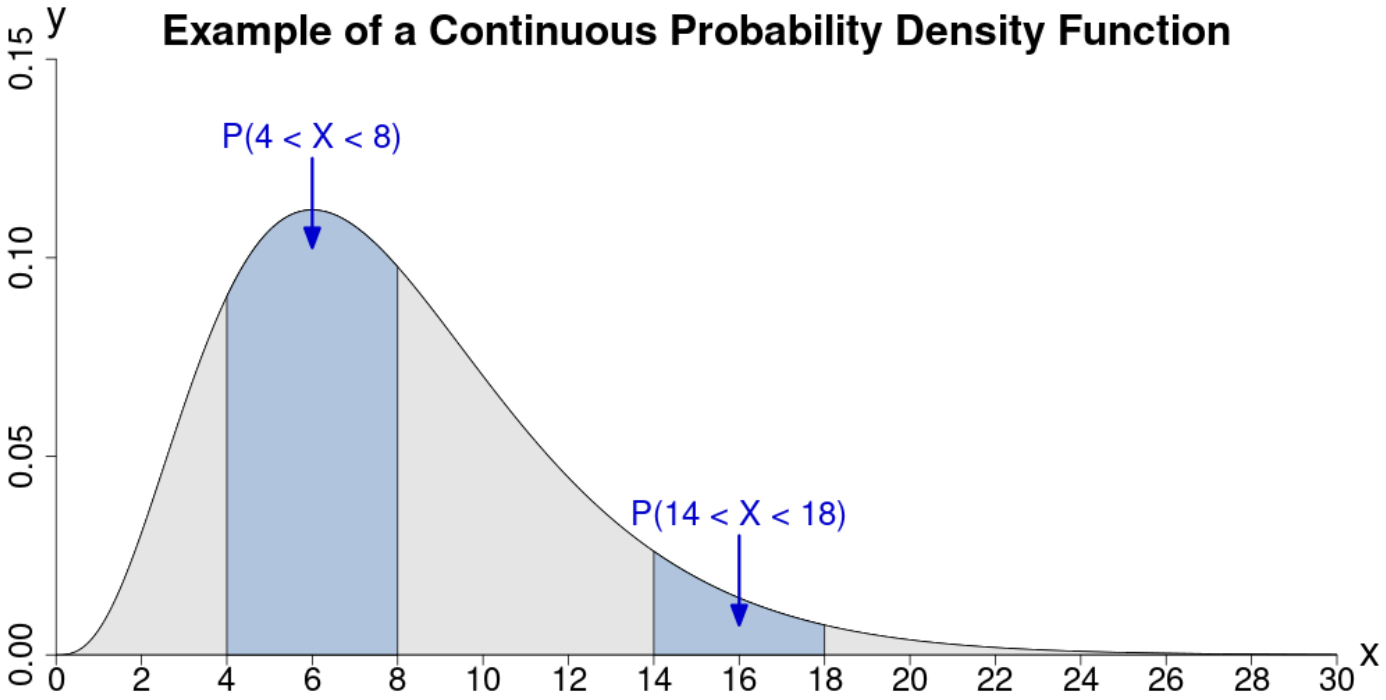



*Fig. 3.7.1* An example of a continuous probability density function. The area under the curve between $x = 4$ and $x = 8$ (the shaded region on the left) is $P(4 < X < 8)$. The area under the curve between $x = 14$ and $x = 18$ (the shaded region on the right) is $P(14 < X < 18)$. Since the area of the left region is clearly larger than the area of the right region, an outcome between $x = 4$ and $x = 8$ is more likely than an outcome between $x = 14$ and $x = 18$.

In this section, we will deal only with calculating the area for regions with shapes we are familiar with from geometry, like rectangles, circles, and triangles. We will learn how to calculate the areas of some regions with more complex shapes later.

**Note**

Since, for continuous probability distributions, we find probability by calculating the area under a probability density function, we can only find probabilities for *ranges* of values, not for *individual* values. It makes sense to find $P(14 < X < 18)$ because the length of the base of the area over that range from $x = 14$ to $x = 18$ is $18 - 14 = 4$. It doesn't make sense to find $P(X = 16)$ because the region over $x = 16$ has no base of any length, so it doesn't make sense to talk about the area of the region in this context.

For the same reason, a probability like $P(X < 8)$ would have the same value as $P(X \leq 8)$ when dealing with. The length of the base up to $x = 7.99999\ldots$ is effectively the same as the length of the base up to $x = 8$.

## Example 3.7.1

Does the function pictured in Figure 3.7.2 represent a probability density function? Why or why not?

Skip to main content

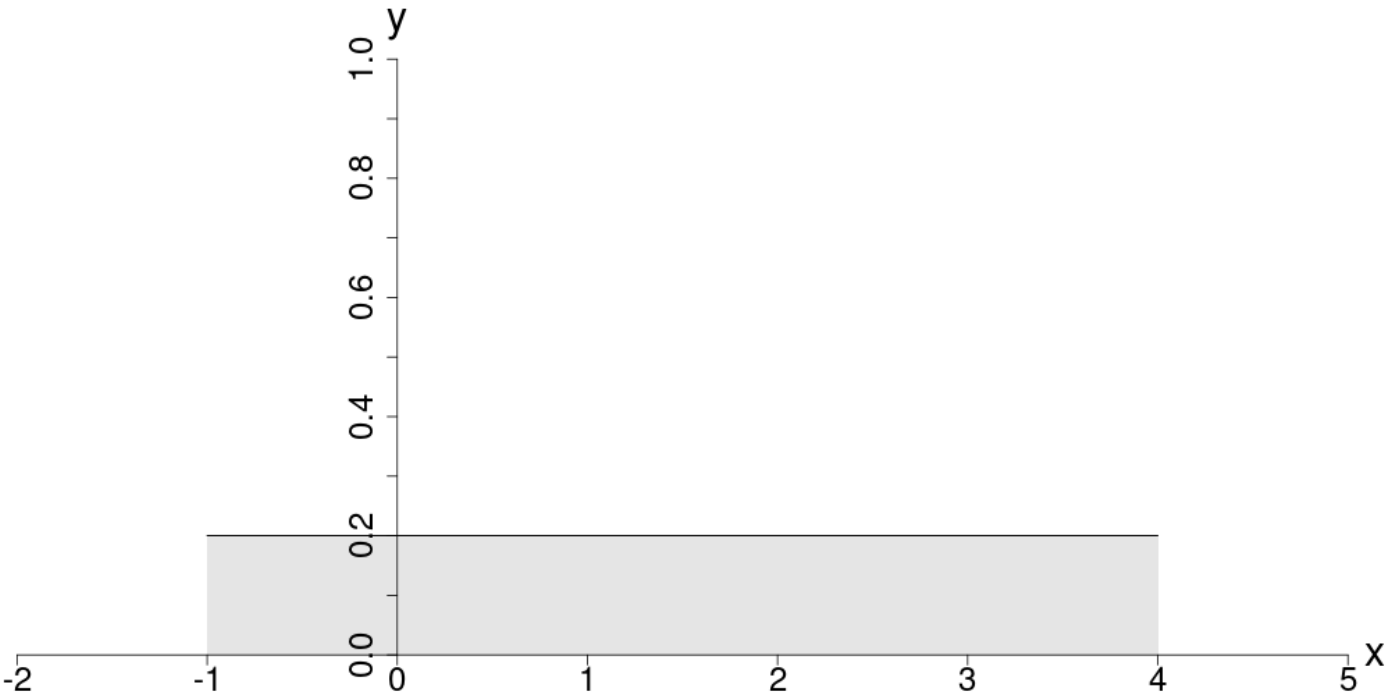



*Fig. 3.7.2* A horizontal line segment of height $y = 0.2$ from $x = -1$ to $x = 4$.

Solution

For a graph or function to be a probability density function, it needs to satisfy two criteria.

First, is the graph entirely above the $x$-axis so no probabilities are less than zero? Yes, we can see that the graph is entirely above the $x$-axis.

Second, is the total area under the graph equal to one? The region under the graph is a rectangle. We can calculate the area of a rectangle by multiplying the length of the base with the height. We can see from the graph that the height of the rectangle is $y = 0.2$. Since the rectangle ranges from $x = -1$ to $x = 4$, the length of the base is $4 - (-1) = 5$. So the total area under the graph is

$$A = \text{Base} \cdot \text{Height} = 5 \cdot 0.2 = 1.$$

Since the total area under the graph is $1$, the function is a probability density function.

---

## Example 3.7.2

Does the function pictured in Figure 3.7.3 represent a probability density function? Why or why not?

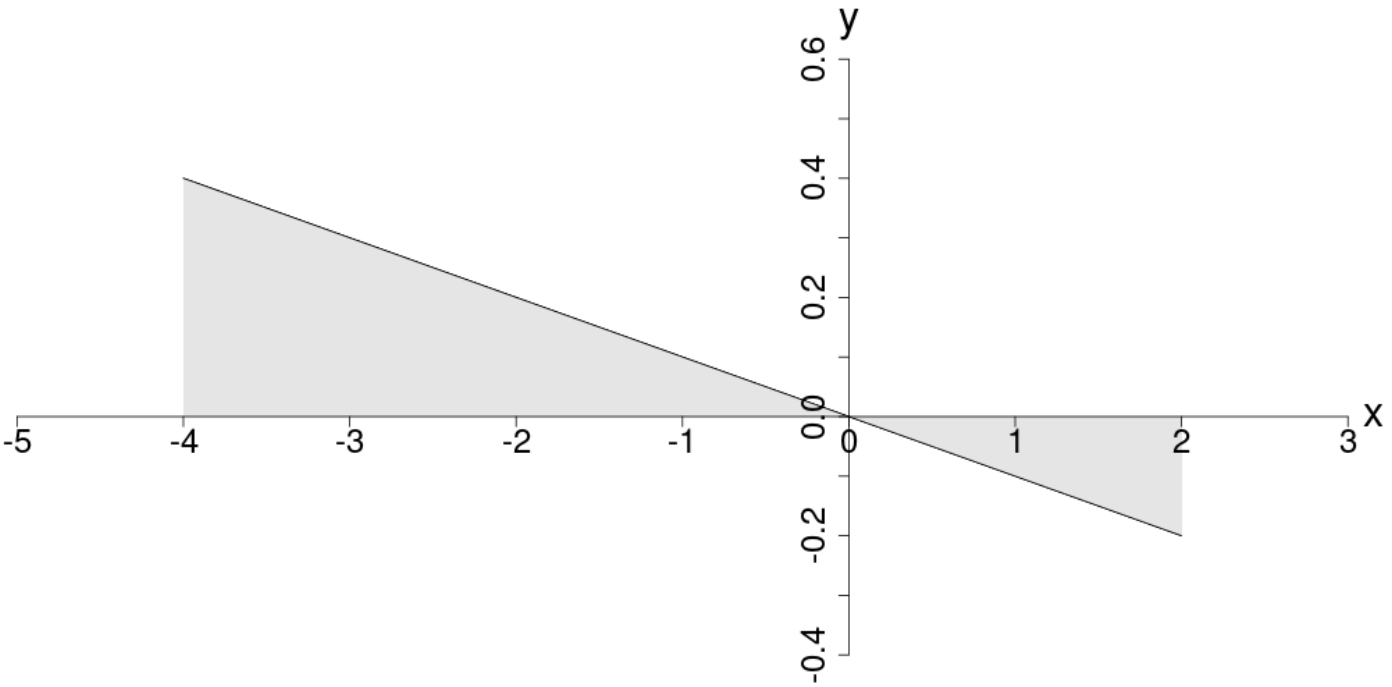



*Fig. 3.7.3* A line segment from $(x, y) = (-4, 0.4)$ to $(x, y) = (2, -0.2)$.

Solution

For a graph or function to be a probability density function, it needs to satisfy two criteria.

First, is the graph entirely above the $x$-axis so no probabilities are less than zero? No, we can see on that the graph is below the $x$-axis between $x = 0$ and $x = 2$. Since the function fails this first criterion, we know it is not a probability density function. We do not need to check the second criterion.

---

## Example 3.7.3

Does the function pictured in Figure 3.7.4 represent a probability density function? Why or why not?

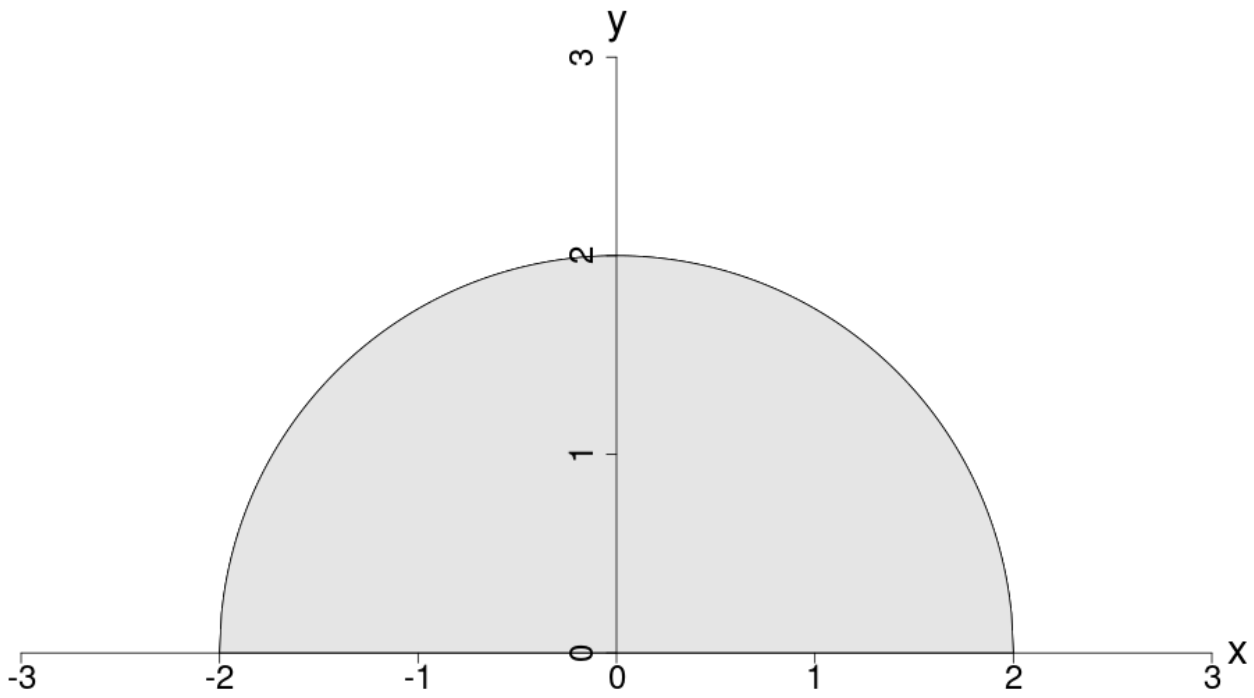



*Fig. 3.7.4* The upper half of a circle of radius $2$.

Solution

For a graph or function to be a probability density function, it needs to satisfy two criteria.

First, is the graph entirely above the $x$-axis so no probabilities are less than zero? Yes, we can see that the graph is entirely above the $x$-axis.

Second, is the total area under the graph equal to one? We can calculate the area of the entire region under the graph using the formula for the area of a full circle:

$$A = \pi \cdot r^2,$$

where $\pi \approx 3.14$ and $r$ is the radius of the circle. But in our case, we only have half a circle, so we will multiply the formula by $\frac{1}{2}$. Since this half-circle has a radius of $r = 2$, the area of the half-circle is

$$A = \frac{1}{2} \cdot \pi \cdot r^2 \approx \frac{1}{2} \cdot 3.14 \cdot 2^2 = 6.28.$$

So the total area under the half-circle is about $6.28$. Because $6.28 \neq 1$, this graph isn't a probability density function.

---

## Example 3.7.4

Consider the probability density function $y = 0.5$ from $x = 1$ to $x = 3$.

1. Find $P(X > 1.5)$.
2. Find $P(2 < X < 2.5)$

Skip to main content

Solution

First, let's graph the probability density function. In this case, it is a horizontal line of height $y = 0.5$ that stretches from $x = 1$ to $x = 3$, as shown in Figure 3.7.5. Note that it really is a probability density function since the area under the graph is

$$A = \text{Base} \cdot \text{Height} = 2 \cdot 0.5 = 1.$$

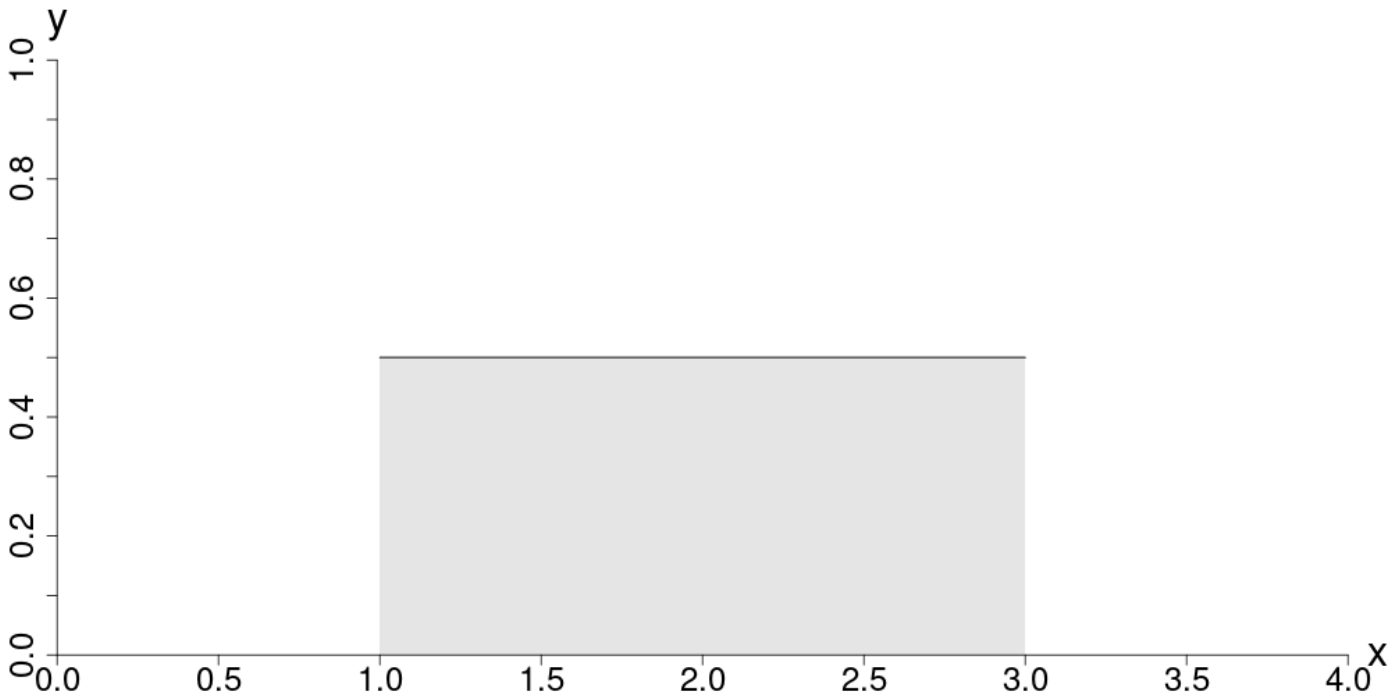



*Fig. 3.7.5* A horizontal line segment of height $y = 0.5$ from $x = 1$ to $x = 3$.

Part 1

To find $P(X > 1.5)$, we need to calculate the area under the graph over the range where $x > 1.5$. (See Figure 3.7.6.)

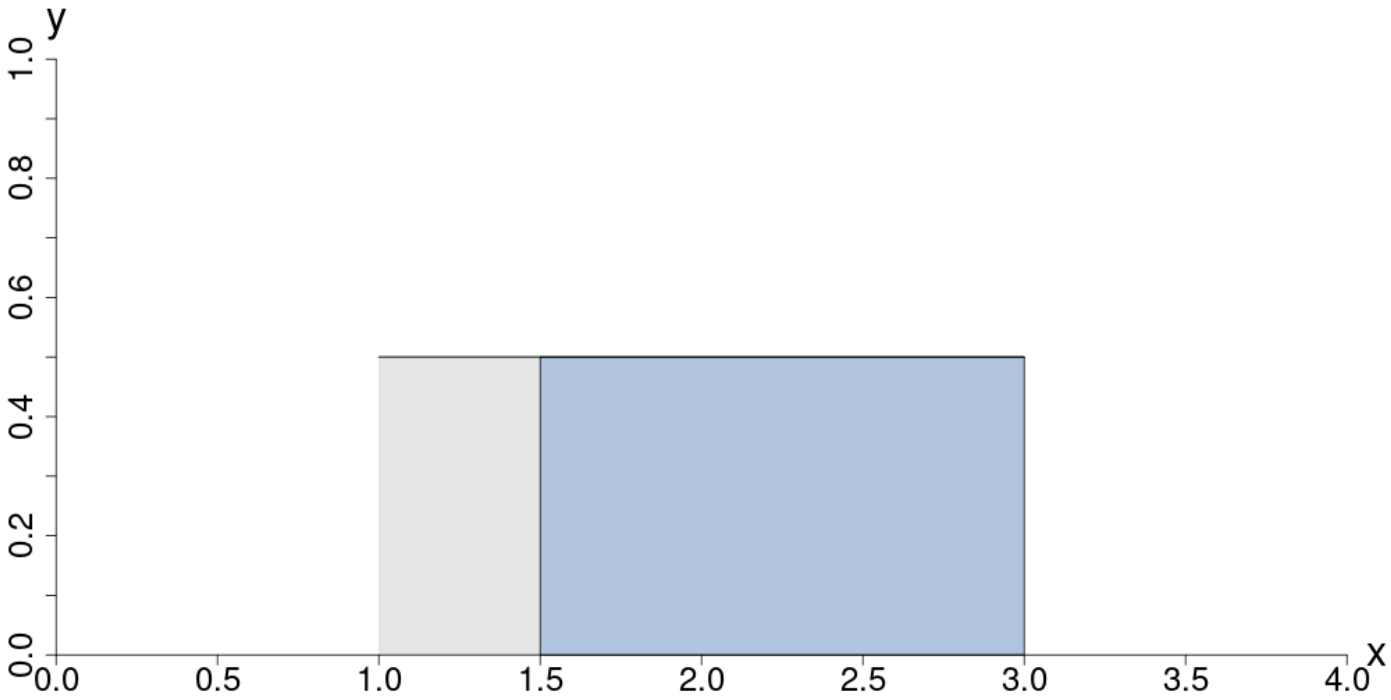



*Fig. 3.7.6* A horizontal line segment of height $y = 0.5$ from $x = 1$ to $x = 3$. The area of the region shaded blue is $P(X > 1.5)$.

This region has a rectangular shape, so we can find the area by multiplying the base of the region times its height. The height is $y = 0.5$. The length of the base from $x = 1.5$ to $x = 3.0$ is $3.0 - 1.5 = 1.5$. So the area of the region is

$$A = \text{Base} \cdot \text{Height} = 1.5 \cdot 0.5 = 0.75.$$

So $P(X > 1.5) = 0.75$.

Part 2

To find $P(2 < X < 2.5)$, we need to calculate the area under the graph over the range between $x = 2$ and $x = 2.5$. (See Figure 3.7.7.)

Skip to main content

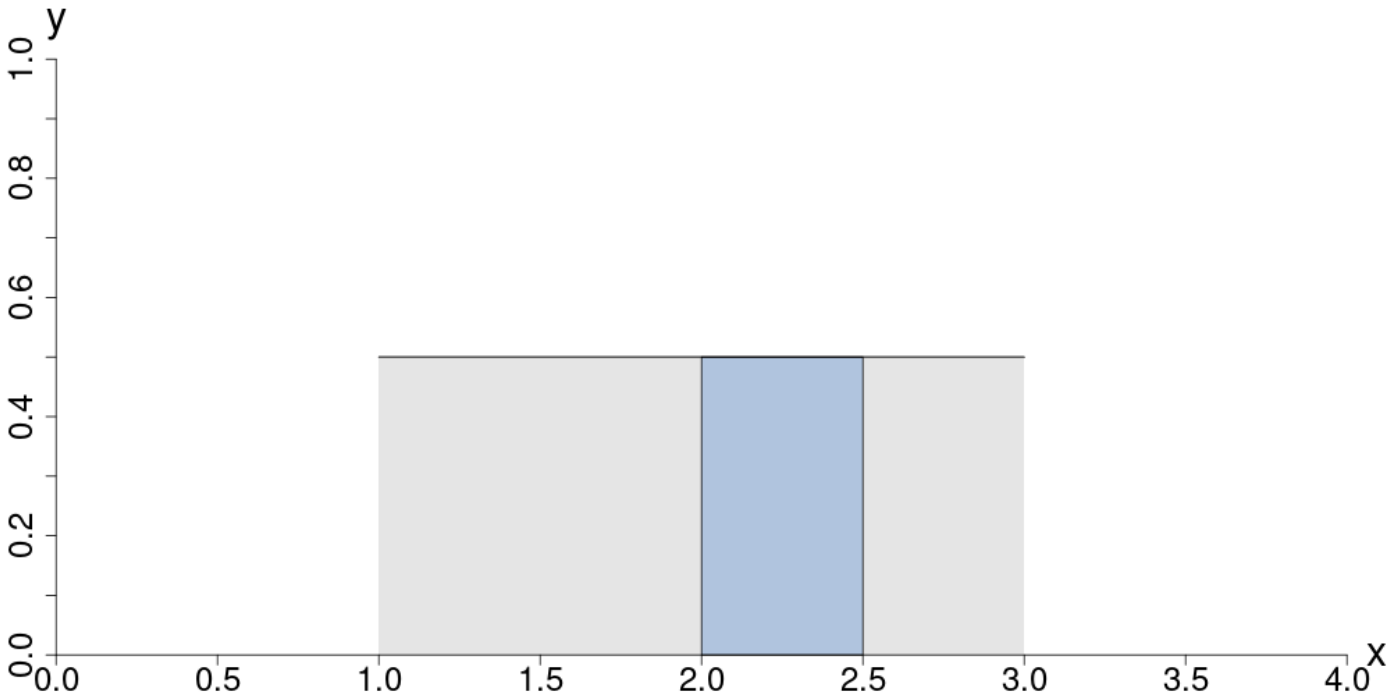



*Fig. 3.7.7* A horizontal line segment of height $y = 0.5$ from $x = 1$ to $x = 3$. The area of the region shaded blue is $P(2 < X < 2.5)$.

This region has a rectangular shape, so we can find the area by multiplying the base of the region times its height. The height is $y = 0.5$. The length of the base from $x = 2$ to $x = 2.5$ is $2.5 - 2 = 0.5$. So the area of the region is

$$A = \text{Base} \cdot \text{Height} = 0.5 \cdot 0.5 = 0.25.$$

So $P(2 < X < 2.5) = 0.25$.

---

## Example 3.7.5

Consider the probability density function $y = \frac{1}{8}x + \frac{3}{8}$ from $x = -3$ to $x = 1$.

1. Find $P(X < -1.5)$.
2. Find $P(X > -1.5)$.

### Solution

Begin by plotting the probability density function. If you remember algebra, you may recall that $y = \frac{1}{8}x + \frac{3}{8}$ is the equation of a line. We know this line only goes from $x = -3$ to $x = 1$. If we plug $x = -3$ into the equation, we get

$$y = \frac{1}{8}(-3) + \frac{3}{8} = 0,$$

so the line starts at the point $(x, y) = (-3, 0)$. Similarly, if we plug $x = 1$ into the equation, we get

Skip to main content

$$y = \frac{1}{8}(1) + \frac{3}{8} = 0.5,$$

so the line ends at the point $(x, y) = (1, 0.5)$. (See Figure 3.7.8)

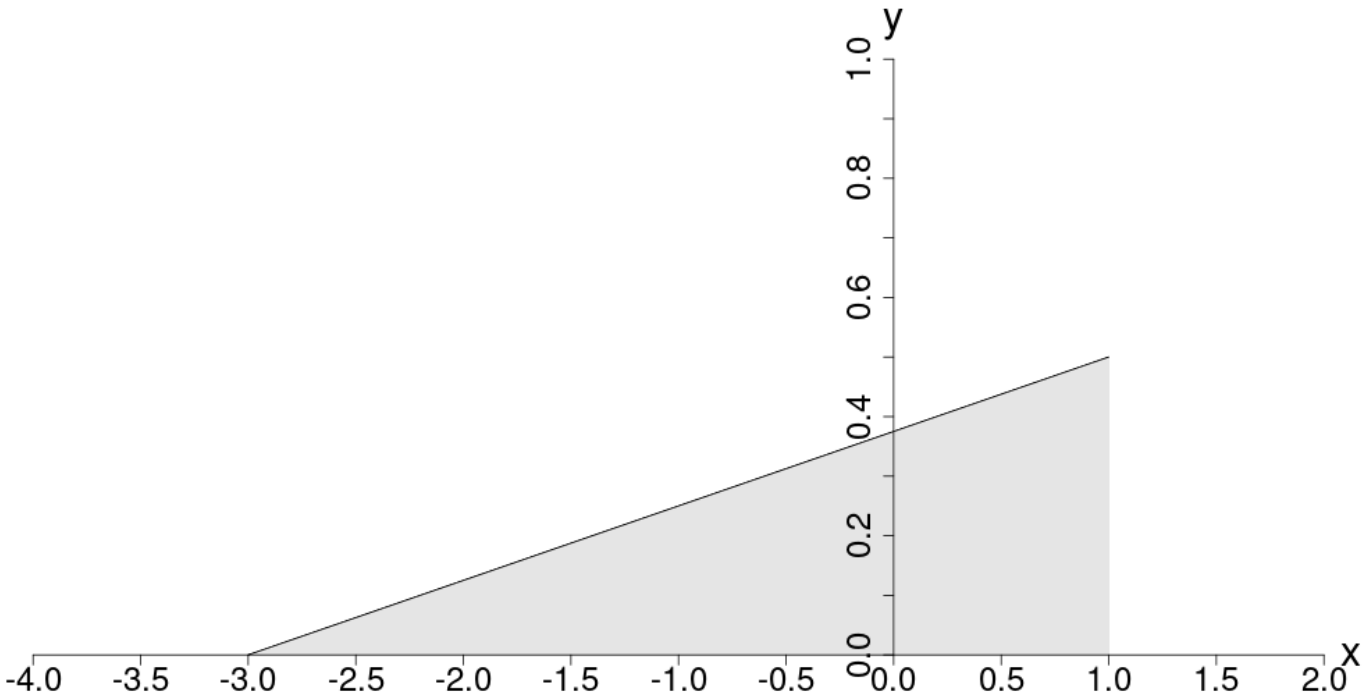



*Fig. 3.7.8* A line segment of from $(x, y) = (-3, 0)$ to $(x, y) = (1, 0.5)$.

Note the total area under the function has a triangular shape. Since the length of the base of the triangle is $1 - (-3) = 4$, and the height is $0.5$, the area of the triangle is

$$A = \frac{1}{2} \cdot \text{Base} \cdot \text{Height} = \frac{1}{2} \cdot 4 \cdot 0.5 = 1,$$

which means the function really is a probability density function, as the problem statement claims.

### Part 1

To find $P(X < -1.5)$, we need to calculate the area under the graph over the range where $x < 1.5$. (See Figure 3.7.9.)

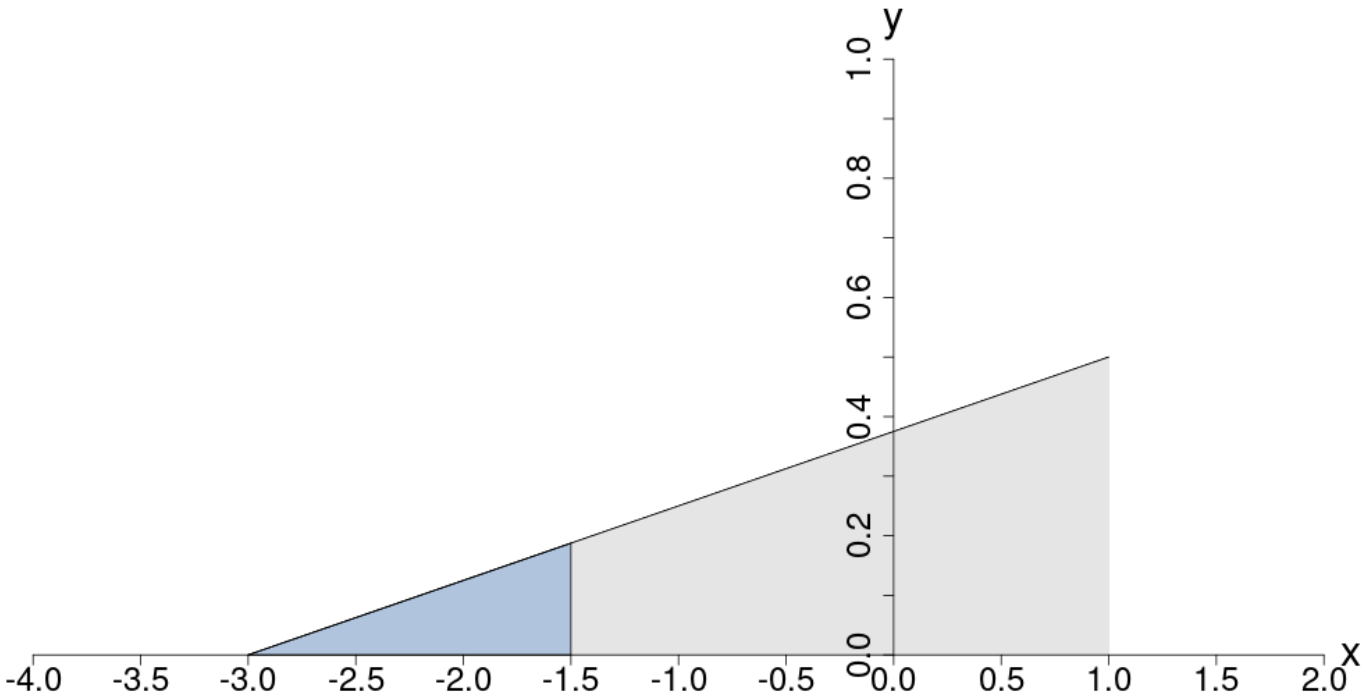



*Fig. 3.7.9* A line segment of from $(x, y) = (-3, 0)$ to $(x, y) = (1, 0.5)$. The area of the shaded blue region is $P(X < -1.5)$.

This region has a triangular shape, so we can find the area using the formula for the area of a triangle. The length of the base of the triangle is $(-1.5) - (-3.0) = 1.5$. To find the height of the triangle, we can plug the $x$-value at which the triangle is highest, $x = -1.5$, into the probability density function:

$$y = \frac{1}{8}(-1.5) + \frac{3}{8} = 0.1875.$$

So the height of the triangle is $0.1875$. Then the area of the triangle is

$$A = \frac{1}{2} \cdot \text{Base} \cdot \text{Height} = \frac{1}{2} \cdot 1.5 \cdot 0.1875 = 0.1406.$$

That is, $P(X < -1.5) = 0.1406$.

Part 2

To find $P(X > -1.5)$, we need to calculate the area under the graph over the range where $x > 1.5$. (See Figure 3.7.10.)

Skip to main content

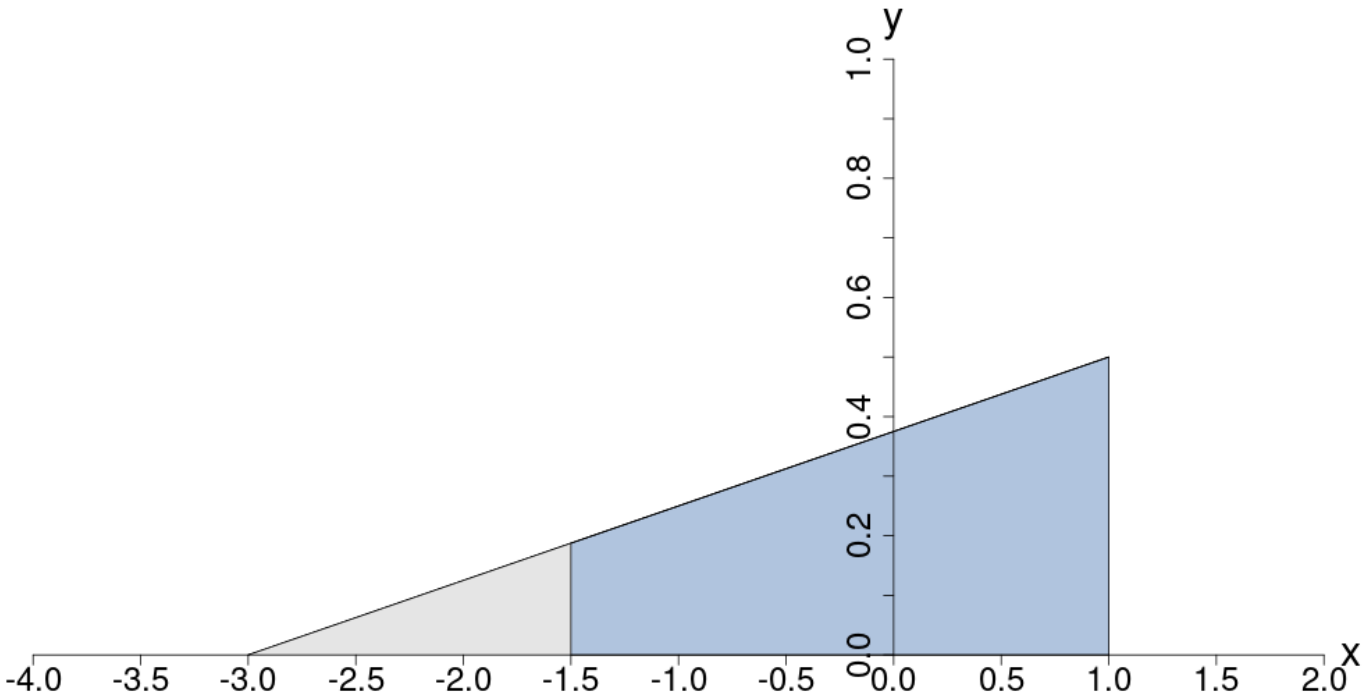



*Fig. 3.7.10* A line segment of from $(x, y) = (-3, 0)$ to $(x, y) = (1, 0.5)$. The area of the shaded blue region is $P(X > -1.5)$.

This region has the shape of a trapezoid. The formula for calculating the area of a trapezoid isn't too hard, but you probably don't remember it. In this particular case, there is an easier way to calculate the area of the region. We just need to observe that the area over $x < -1.5$ and the area over $x > -1.5$ are complementary, since together they cover the total area under the probability density function. That means

$$P(X < -1.5) + P(X > -1.5) = 1.$$

But we just found that $P(X < -1.5) = 0.1406$, meaning

$$P(X > -1.5) = 1 - P(X < -1.5) = 1 - 0.1406 = 0.8594.$$

So $P(X > -1.5) = 0.8594$.

# 4. The Normal Distribution and the Central Limit Theorem

## 4.1. Properties of The Normal Distribution

### Objectives

- Recognize the structure of the normal distribution, including the roles of the mean and standard deviation.
- Use the properties of the normal distribution, such as symmetry, to infer probability.

Skip to main content

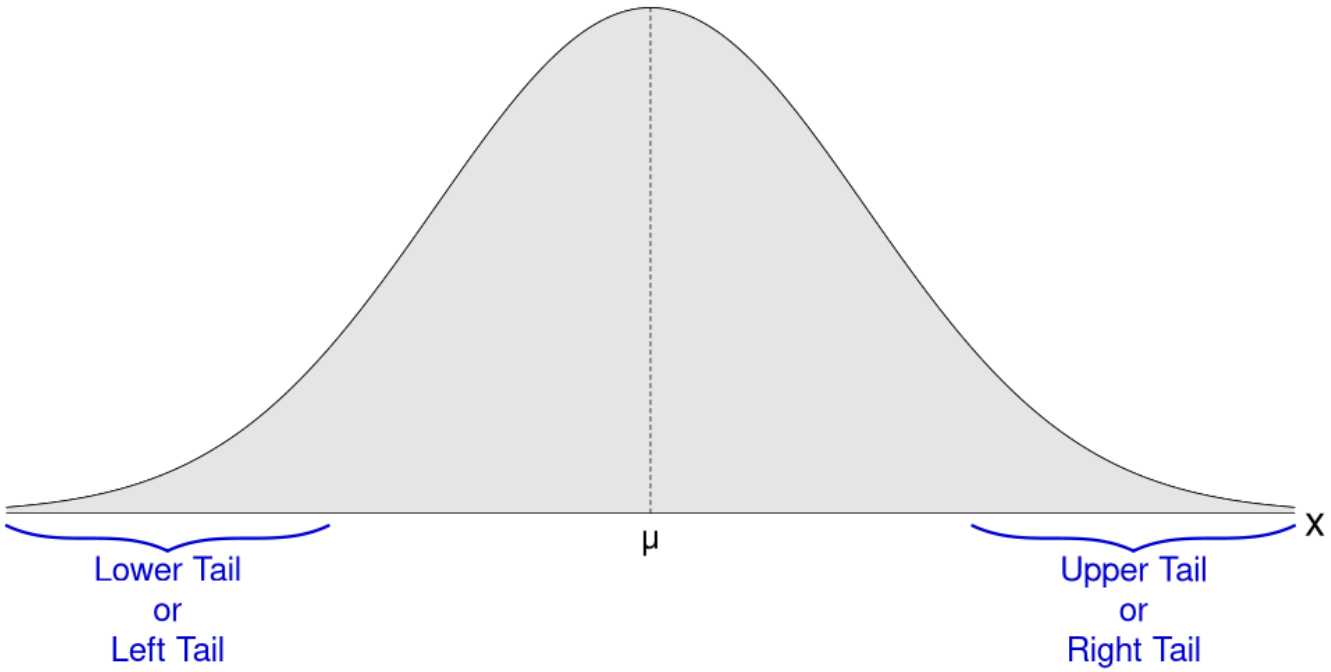



*Fig. 4.1.1* The probability density function of a normal distribution with mean $\mu$. A normal distribution has two tails: the lower tail or left tail, and the upper tail or right tail. Also, note that a normal distribution is symmetric about the mean: the vertical line through $\mu$ acts like a mirror.

The **normal distribution** is a continuous probability distribution, and it is the most important of all probability distributions. The graph of its probability density function is a classic bell-shaped curve with its mean at the center. Many real-world variables have bell-shaped distributions, such as birth weight, IQ score, annual income, and real estate prices. Because the bell-shaped curve is so prevalent, the normal distribution is used in almost all disciplines, including psychology, business, economics, the sciences, nursing, and mathematics.

The shape of a normal distribution means values tend to cluster around the mean of the distribution. The closer a value is to the mean, the more likely it is to happen. The farther a value is from the mean, the less likely it is to happen.

The two narrow ends of a normal distribution called the two **tails** of the distribution. Values much smaller than the mean are in the **lower tail** or **left tail** of the distribution. Values much larger than the mean are in the **upper tail** or **right tail** of the distribution.

A normal distribution can be described using two parameters: the distribution's mean, $\mu$, and its standard deviation, $\sigma$. If a random variable $X$ is normally distributed with mean $\mu$ and standard deviation $\sigma$, we write

$$X \sim N(\mu, \sigma).$$

The position and shape of a normal distribution depends only on its mean and standard deviation. The position of a normal distribution depends on the value of its mean, $\mu$. The peak of the distribution is always directly over its mean on the horizontal axis. (See Figure 4.1.2.) The shape of a normal distribution depends on the value of its standard deviation, $\sigma$. A normal distribution with a larger standard deviation will be more spread out with a shorter peak. A normal distribution with a smaller standard deviation will be more squished together with a taller peak. (See Figure 4.1.3.)

A normal distribution is symmetric about its mean, $\mu$, so the vertical line through the mean acts like a mirror. For example, if we know the probability in a section in the lower tail of a normal distribution, the mirror section in the upper tail of the distribution will

Skip to main content

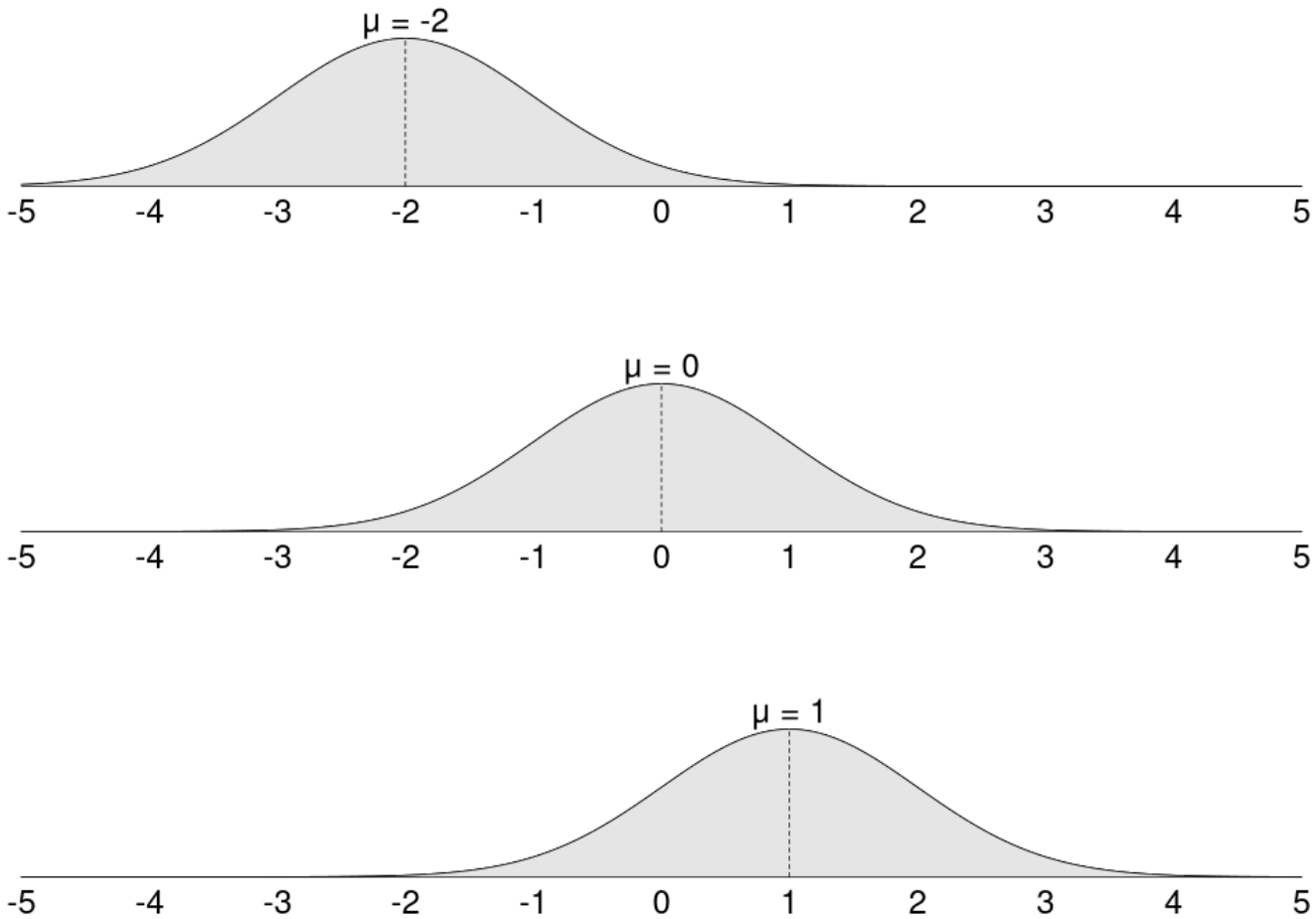



*Fig. 4.1.2* The graphs of three different normal distributions. Each of the distributions have different means, but they have the same standard deviation. The peak of a normal distribution is always positioned above the distribution mean, so normal distributions with different means will have different positions.

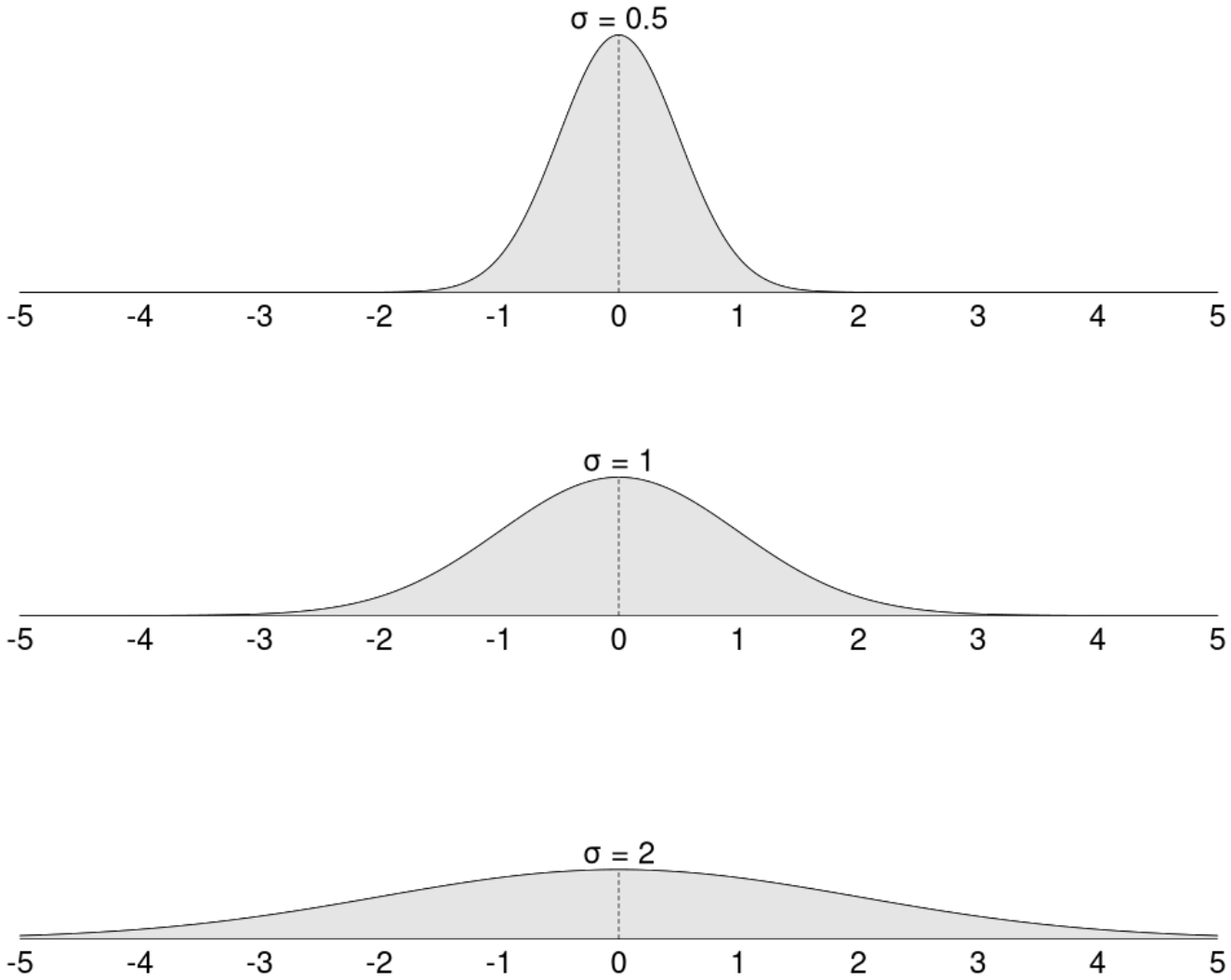



*Fig. 4.1.3* The graphs of three different normal distributions. Each of the distributions has the same mean, but they have different standard deviations. A normal distribution with a larger standard deviation is more spread out with a shorter peak. A normal distribution with a smaller standard deviation is more squished together with a taller peak.

---

## Example 4.1.1

Kelsey tells you that $P(X > 3.5) = 0.691$ for some normally distributed random variable $X$. Can you find $P(X < 3.5)$?

Solution

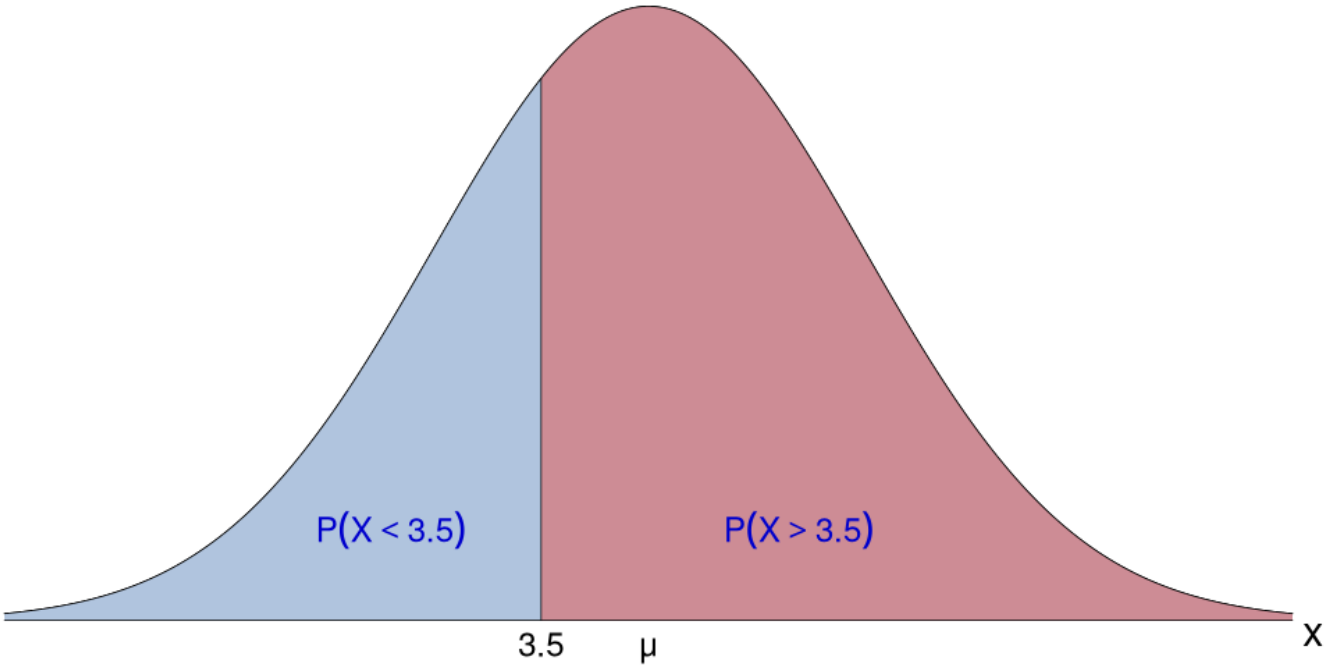



*Fig. 4.1.4* This graph of a normal distribution illustrates how the graph of the distribution of $X$ might look. The area of the region under the curve to the right of $x = 3.5$ is equal to $P(X > 3.5)$. The area of the region to the left of $x = 3.5$ is equal to $P(X < 3.5)$. In particular, note that the regions corresponding to $P(X < 3.5)$ and $P(X > 3.5)$ together account for the entire area under the curve. Since the entire area under the curve is equal to $1$, this means $P(X < 3.5) + P(X > 3.5) = 1$.

Kelsey has left out some pretty important information. Kelsey told us $X$ has a normal distribution, but she didn't say what the mean or the standard deviation of the distribution is, which means we don't know the position or exact shape of the distribution.

In this case, though, we have all the information we need to find $P(X < 3.5)$. A normal distribution is a kind of continuous probability distribution, so the total area under the curve of the distribution is $1$. This means

$$P(X < 3.5) + P(X > 3.5) = 1.$$

Since we know $P(X > 3.5) = 0.691$, we calculate

$$P(X < 3.5) = 1 - P(X > 3.5) = 1 - 0.691 = 0.309.$$

---

## Example 4.1.2

Suppose $X$ is a normally distributed random variable with mean $\mu = -2$. If $P(X < -5) = 0.274$, what is $P(X > 1)$?

Solution

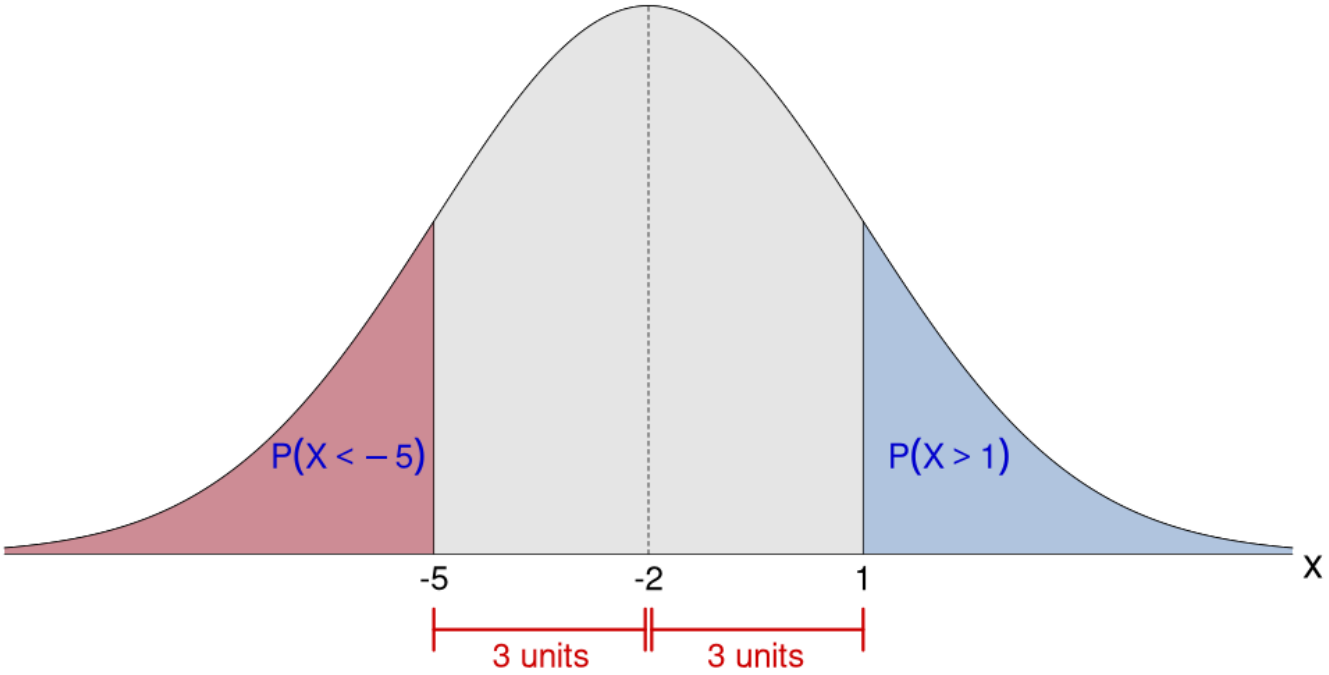



*Fig. 4.1.5* This graph of a normal distribution illustrates how the graph of the distribution of $X$ might look. The area of the shaded region to the left of $x = -5$ is equal to $P(X < -5)$. The area of the shaded region to the right of $x = 1$ is equal to $P(X > 1)$. Note that the boundaries of both shaded regions are the same distance away from the mean at $\mu = -2$. Because a normal distribution is symmetric about the mean, these two regions are mirror images of each other, so they have equal areas.

We don't have all the information we need to fully characterize this normal distribution. We know that the mean is $\mu = -2$, but we don't know the standard deviation.

But, if we're a little clever, we can still find $P(X > 1)$. First, recall how we can represent $P(X < -5)$ and $P(X > 1)$ graphically: if we consider the curve of this normal distribution, $P(X < -5)$ is equal to the area of the region under the curve and to the left of $x = -5$, and $P(X > 1)$ is equal to the area of the region under the curve and to the right of $x = 1$. (See Figure 4.1.5.) Now note that $x = -5$ and $x = 1$ are both a distance of $3$ units away from the mean $\mu = -2$, so both regions are the same distance from the mean. Since the normal distribution is symmetric about the mean, the region corresponding to $P(X > 1)$ is the mirror image to the region corresponding to $P(X < -5)$, so the two regions have equal areas. Since probability equals area, it must be that

$$P(X < -5) = P(X > 1).$$

Since $P(X < -5) = 0.274$, it follows that $P(X > -1) = 0.274$, also.

---

## Example 4.1.3

Figure 4.1.6 shows the graphs of two different normal distributions, labeled $A$ and $B$. Which distribution has the greater mean? Which distribution has the greater standard deviation?

Skip to main content

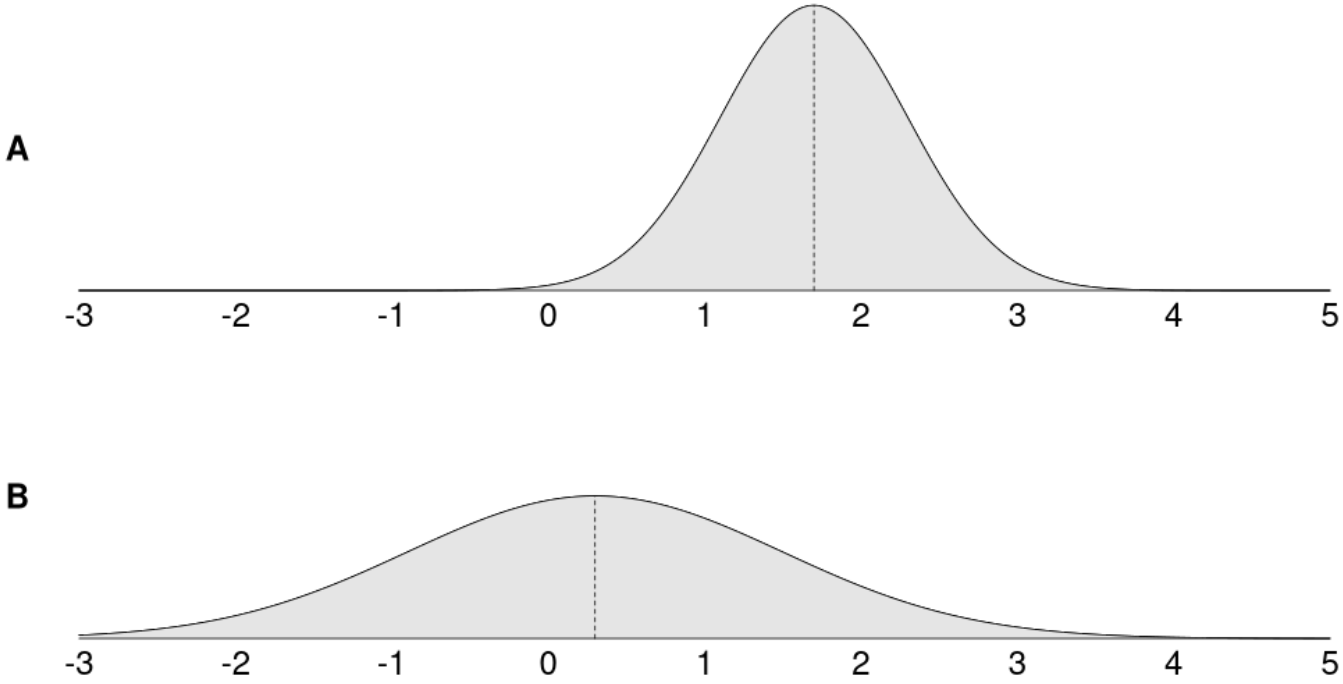



*Fig. 4.1.6* Two normal distributions, labeled A and B, with different means and standard deviations.

Solution

The mean of a normal distribution tells us the position of the distribution. The tip of the peak of a normal distribution is always directly over the mean. With this in mind, we can see that the mean of distribution $A$ is between $1$ and $2$, while the mean of distribution $B$ is between $0$ and $1$. So distribution $A$ has a greater mean than distribution $B$.

The standard deviation of a normal distribution determines the shape of the distribution. It tells us how spread out a distribution is. In this case, distribution $B$ is more spread out than distribution $A$. This means distribution $B$ has a greater standard deviation than distribution $A$.

# 4.2. The Standard Normal Distribution

## Objectives

- Calculate probabilities involving the standard normal distribution.
- Calculate quantiles of the standard normal distribution corresponding to given probabilities.

## The Standard Normal Distribution

To begin to understand how to calculate probabilities of normal distributions, we will first study how to find probabilities of special kind of normal distribution called the **standard normal distribution**. The standard normal distribution is the particular normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The standard normal distribution is so important that we reserve the random variable $Z$ exclusively for the standard normal distribution, so $Z \sim N(0, 1)$.

Skip to main content

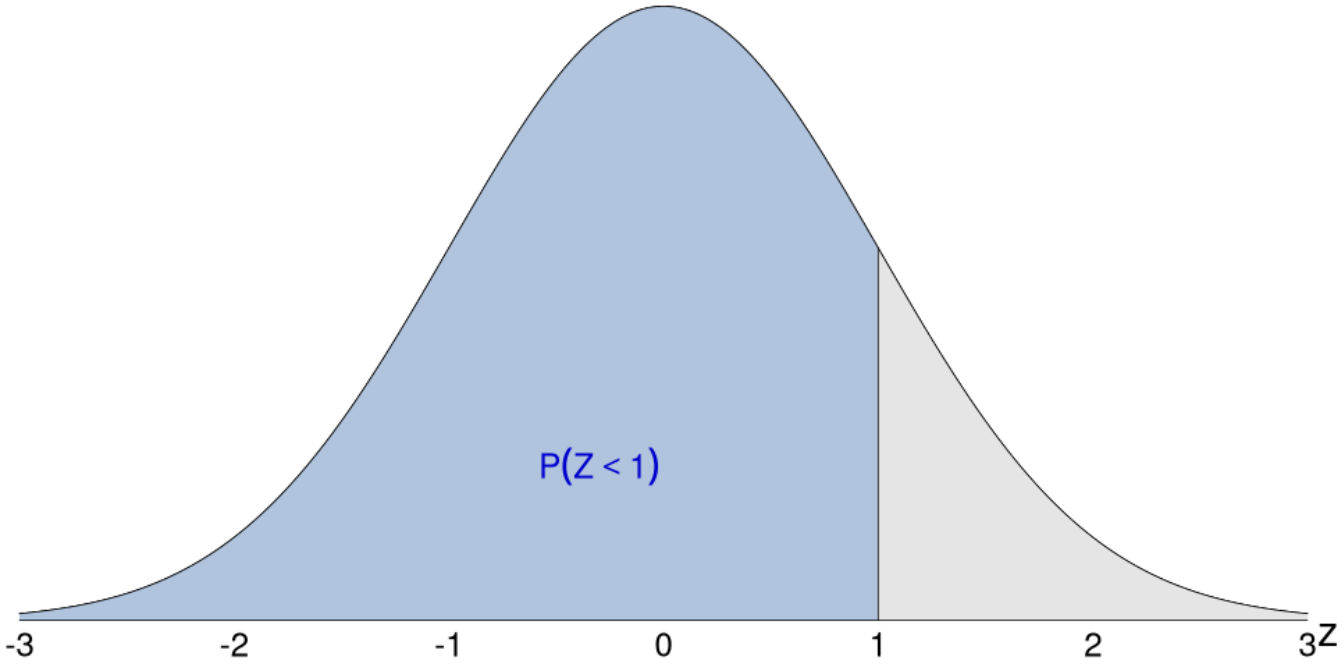



*Fig. 4.2.1* The standard normal distribution with the area to the left of $z = 1$ shaded. The shaded area is equal to $P(Z < 1)$.

Suppose we want to calculate $P(Z < 1)$. This probability is equal to the area under the curve of the standard normal distribution and to the left of $z = 1$. (See Figure 4.2.1.) This region has an unusual shape, so we can't calculate the area of this region using the usual geometric formulas for shapes like rectangles and triangles. For this reason, probabilities of normal distributions are found using computers or calculators instead of being calculated directly.

## Finding Probability Given a $z$-value

For $Z \sim N(0, 1)$, we can calculate the probability $P(Z < z)$ of randomly choosing a value less than $z$ using the R function `pnorm`:

```
pnorm(q)
```

The argument `q` is the $z$-value we want to find the probability of being less than. (Here, `q` stands for **q**uantile, since the probability that `pnorm(q)` returns is the theoretical proportion of the values that are less than argument `q`.)

For example, let's use R to find $P(Z < 1)$.

```
pnorm(q = 1)
```

0.841344746068543

So $P(Z < 1) = 0.8413$.

Note that `pnorm(q = z)` only directly computes probability of the form $P(Z < z)$, the probability that a randomly chosen value is *less* than $z$. What if we want to calculate $P(Z > z)$, the probability that a randomly chosen value is *greater* than $z$? Or, even trickier, what if we want to calculate $P(z_1 < Z < z_2)$, the probability that a randomly chosen value is *between* the values $z_1$ and

Skip to main content

## Example 4.1.4

The data in a research study has a standard normal distribution.

1. If you randomly choose a data value from the study, what is the probability that the data value will be less than $z = -0.83$?
2. If you randomly choose a data value from the study, what is the probability that the data value will be greater than $z = -0.57$ ?
3. If you randomly choose a data value from the study, what is the probability that the data value will be between $z = -0.94$ and $z = 1.25$?

### Solution

#### Part 1

Graphically, we want to find the area under the curve of the standard normal distribution and to the left of $z = -0.83$. This area is equal to $P(Z < -0.83)$.

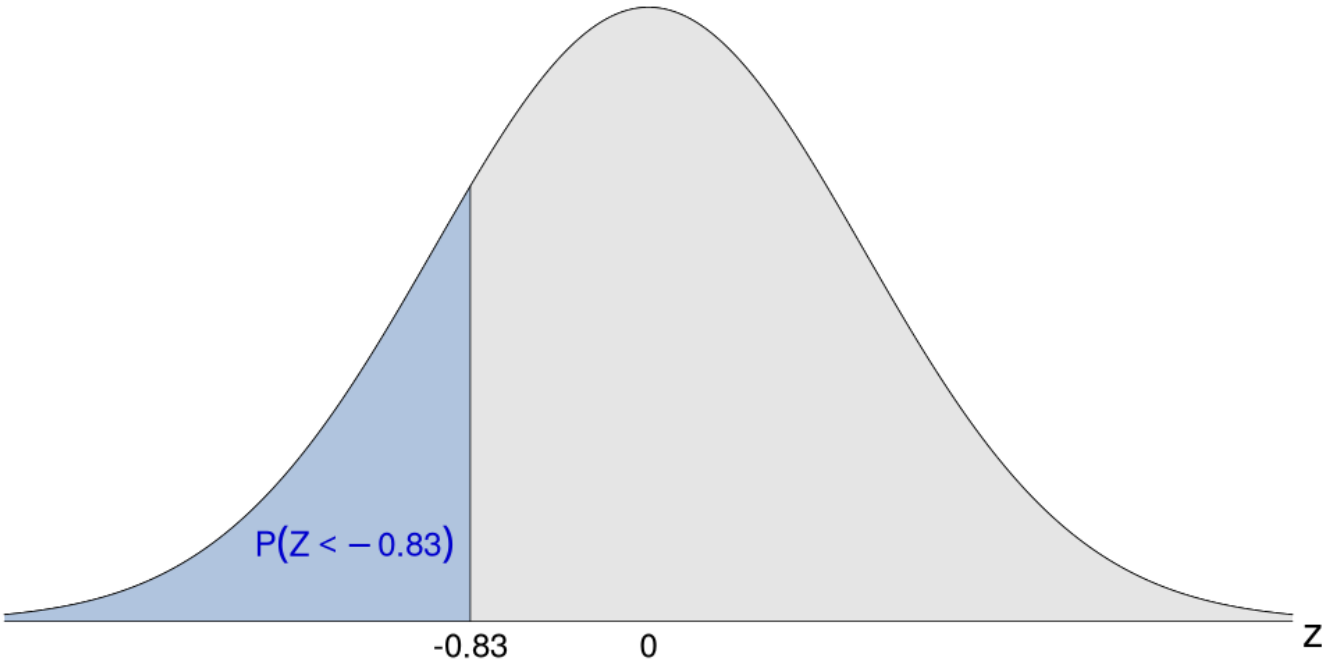



*Fig. 4.2.2* The standard normal distribution with the area equal to $P(Z < -0.83)$ shaded.

We use the `pnorm` function to calculate the probability.

```
pnorm(q = -0.83)
```

0.203269391828068

So $P(Z < -0.83) = 0.2033$. This means that if we randomly choose a data value from the research study, there is about a 20.33% chance that the value will be less than $z = -0.83$.

Skip to main content

Part 2

We want to find $P(Z > -0.57)$. This is a bit tricky since `pnorm(q = -0.57)` gives the probability that a randomly chosen value is *less* than $z = -0.57$, but we want the probability that a randomly chosen value is *greater* than $z = -0.57$. How can we use R to calculate $P(Z > -0.57)$?

Instead of calculating $P(Z > -0.57)$ directly, we start with the *total* area under the *entire* curve (which is equal to $1$), then take away the area that we *don't* want (which equals $P(Z < -0.57)$). (See Figure 4.2.3.) This idea expressed as an equation gives us

$$P(Z > -0.57) = 1 - P(Z < -0.57).$$

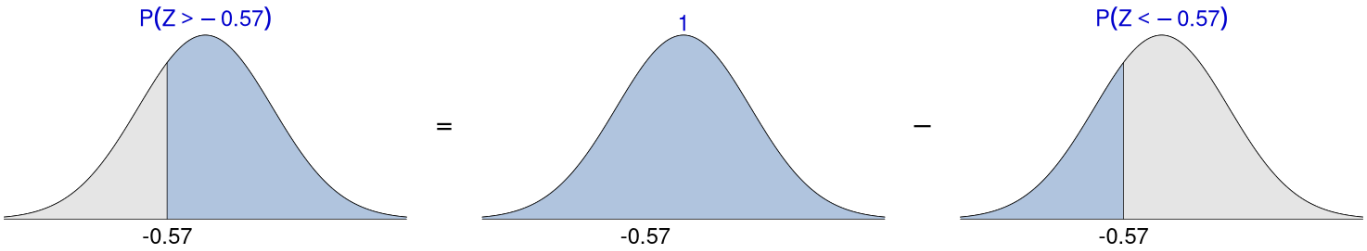



*Fig. 4.2.3* A diagram illustrating that $P(Z > -0.57) = 1 - P(Z < -0.57)$. Rather than calculating $P(Z > -0.57)$ directly, we take the total area under the standard normal distribution (which equals $1$), then take away the part that we don't want (the part equal to $P(Z < -0.57)$).

With this formula, we can compute $P(Z > -0.57)$ using R.

```
1 - pnorm(q = -0.57)
```

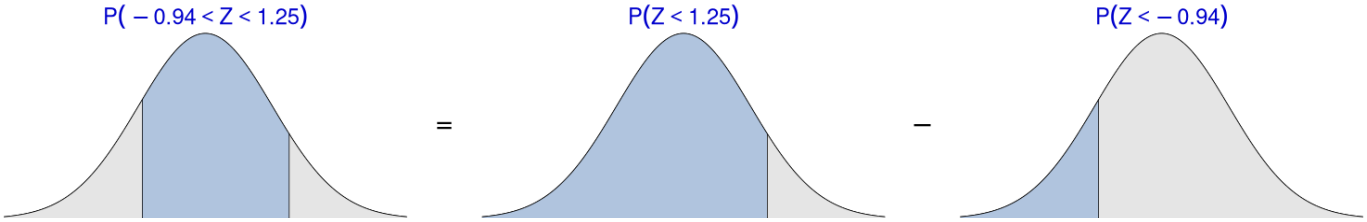
0.715661150953676

Then $P(Z > -0.57) = 0.7157$. So if we randomly choose one of the data values from the research study, there is about a $71.57\%$ chance that the data value will be greater than $z = -0.57$.

Part 3

We need to calculate $P(-0.94 < Z < 1.25)$. Once again, we can't calculate this probability directly using the `pnorm` function since, in this instance, we want the probability that a randomly chosen value is *in between* two other values, $z = -0.94$ and $z = 1.25$. Similar to what we did before, we begin by finding *all* of the area under the curve to the left of the larger $z$-value (which is equal to $P(Z < 1.25)$ in this case), then take away the extra area to the left of the smaller $z$-value (which equals $P(Z < -0.94)$). (See Figure 4.2.4.) This gives us the equation

$$P(-0.94 < Z < 1.25) = P(Z < 1.25) - P(Z < -0.94).$$




Skip to main content

*Fig. 4.2.4* A diagram illustrating that $P(-0.94 < Z < 1.25) = P(Z < 1.25) - P(Z < -0.94)$. Rather than calculating $P(Z > -0.57)$ directly, we take all the area under the curve to the left of $z = 1.25$ (which equals $P(Z < 1.25)$), then take away the part that we don't want (the part equal to $P(Z < -0.94)$).

Using this formula, we can use R to find $P(-0.94 < Z < 1.25)$.

```
pnorm(q = 1.25) - pnorm(q = -0.94)
```

0.72074144599452

So $P(-0.94 < Z < 1.25) = 0.7207$, meaning there is about a $72.07\%$ chance that a randomly chosen data value from the research study will be between $z = -0.94$ and $z = 1.25$.

---

## Example 4.1.5

Let $Z \sim N(0, 1)$.

1. Find $P(Z < 0.67)$.
2. Find $P(Z > -1.91)$.
3. Find $P(0.34 < Z < 1.62)$.

### Solution

#### Part 1

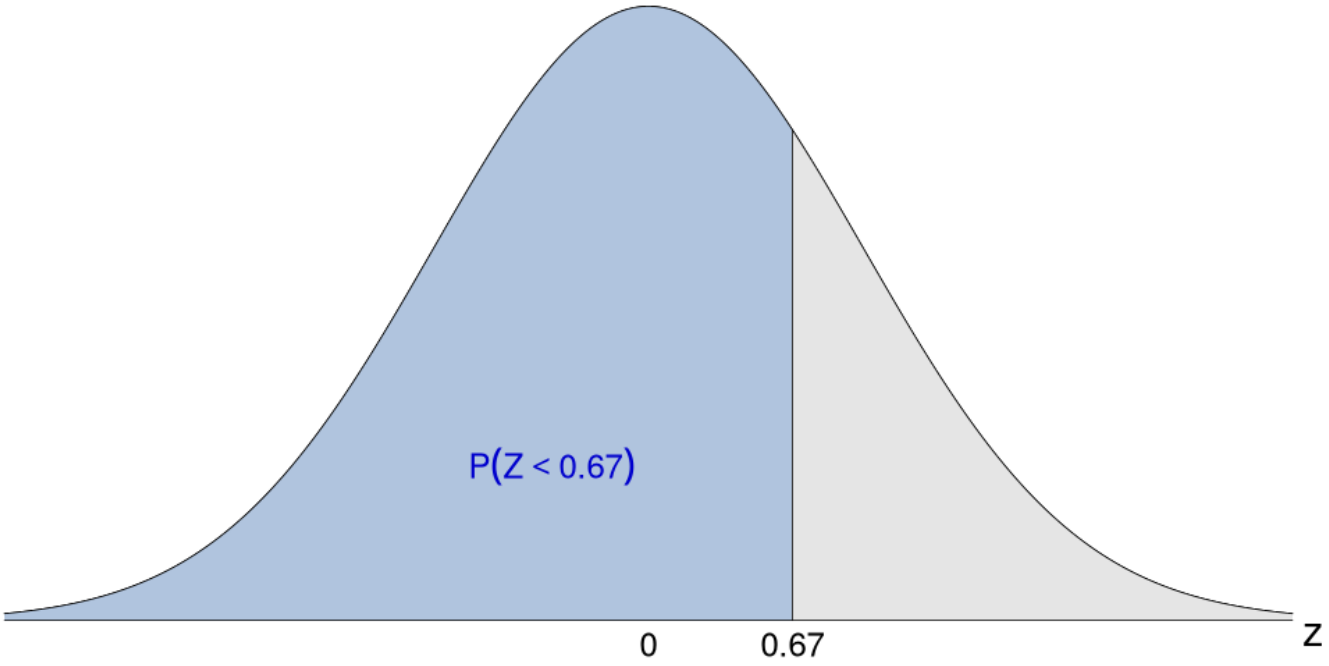



*Fig. 4.2.5* The area equal to $P(Z < 0.67)$.

Skip to main content

We want to find $P(Z < 0.67)$, the probability that a randomly chosen value from the distribution is *less* than $z = 0.67$. We don't need to do anything special to use R to calculate this probability. We just use the `pnorm` function without any changes.

```
pnorm(q = 0.67)
```

0.74857110490469

So $P(Z < 0.67) = 0.74857$.

Part 2

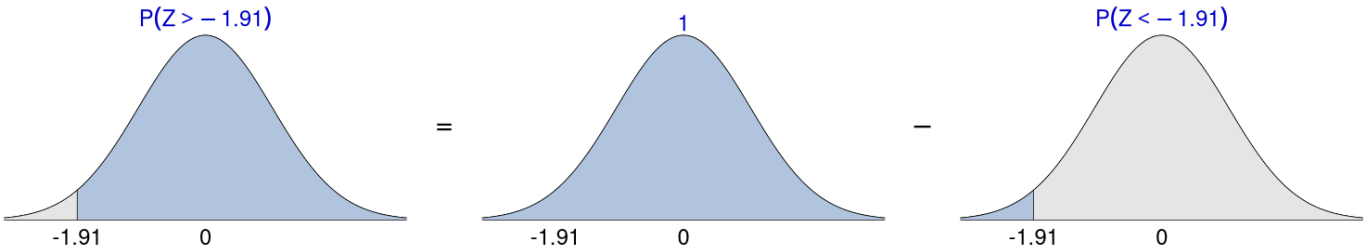



*Fig. 4.2.6* A diagram illustrating that $P(Z > -1.91) = 1 - P(Z < -1.91)$.

We want to find $P(Z > -1.91)$, the probability that a randomly chosen value from the distribution is *greater* than $z = -1.91$. To find $P(Z > -1.91)$, we take the total area under the standard normal distribution (which equals $1$), then take away the part that we don't want (the part equal to $P(Z < -1.91)$).

```
1 - pnorm(q = -1.91)
```

0.971933393340227

So $P(Z > -1.91) = 0.97193$.

Part 3

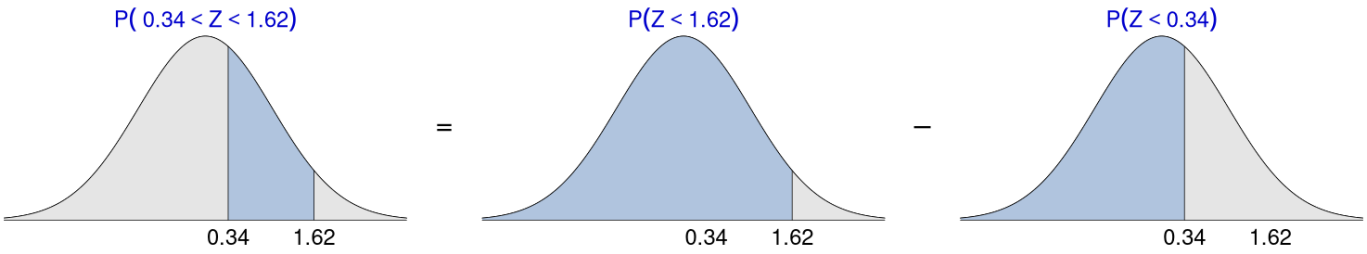



*Fig. 4.2.7* A diagram illustrating that $P(0.34 < Z < 1.62) = P(Z < 1.62) - P(Z < 0.34)$.

We want to find $P(0.34 < Z < 1.62)$, the probability that a randomly chosen value from the distribution is *between* $z = 0.34$ and $z = 1.62$. To calculate $P(0.34 < Z < 1.62)$, we take *all* the area under the curve to the left of the larger $z$ value (which equals $P(Z < 1.62)$ in this case), then take away the area that we don't want to the left of the smaller $z$ value (which equals $P(Z < 0.34)$).

```
pnorm(q = 1.62) - pnorm(q = 0.34)
```

0.31431212550972

So $P(0.34 < Z < 1.62) = 0.31431$.

---

## Finding a $z$-value Given Probability

We've seen that, given a $z$-value, the R code `pnorm(q = z)` will tell us $P(Z < z)$, the probability that a randomly chosen value from the standard normal distribution is smaller than $z$. But how do we do the inverse? That is, given a probability, how do we find a $z$-value so that $P(Z < z)$ is equal to that probability?

We can do this using the `qnorm` function:

```
qnorm(p)
```

Here, `p` is the given **p**robability. The function `qnorm(p)` tells us the $z$-value (or **q**uantile) so that $P(Z < z)$ equals `p`.

To illustrate the inverse relationship between the `pnorm` function and the `qnorm` function, let us take an arbitrary $z$-value, say $z = 0.52$. Let's use the `pnorm` function to find $P(Z < 0.52)$.

```
pnorm(q = 0.52)
```

0.698468212453034

So $P(Z < 0.52) = 0.69847$.

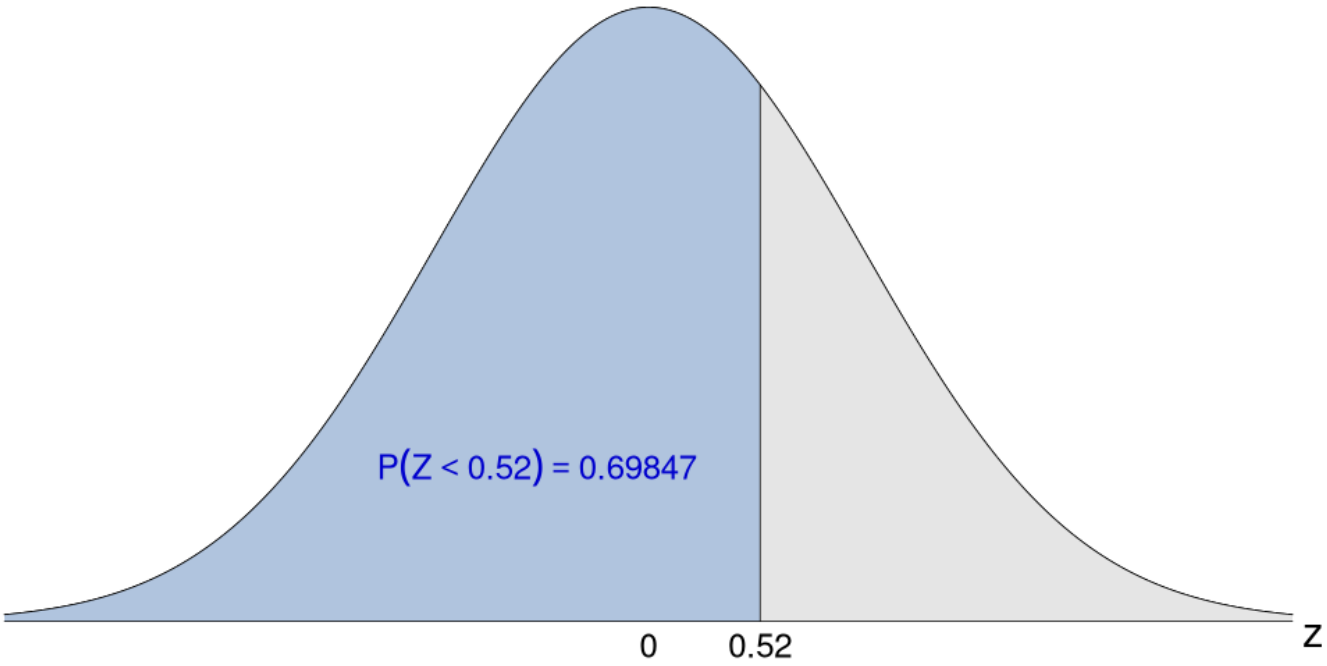



*Fig. 4.2.8 The standard normal distribution with the area equal to $P(Z < 0.52) = 0.69847$ shaded*

Skip to main content

Now let's see what happens when we plug this probability, $p = 0.69847$, into the `qnorm` function.

```
qnorm(p = 0.69847)
```

0.520005129376629

We get $z = 0.52$, which is the $z$-value we started with.

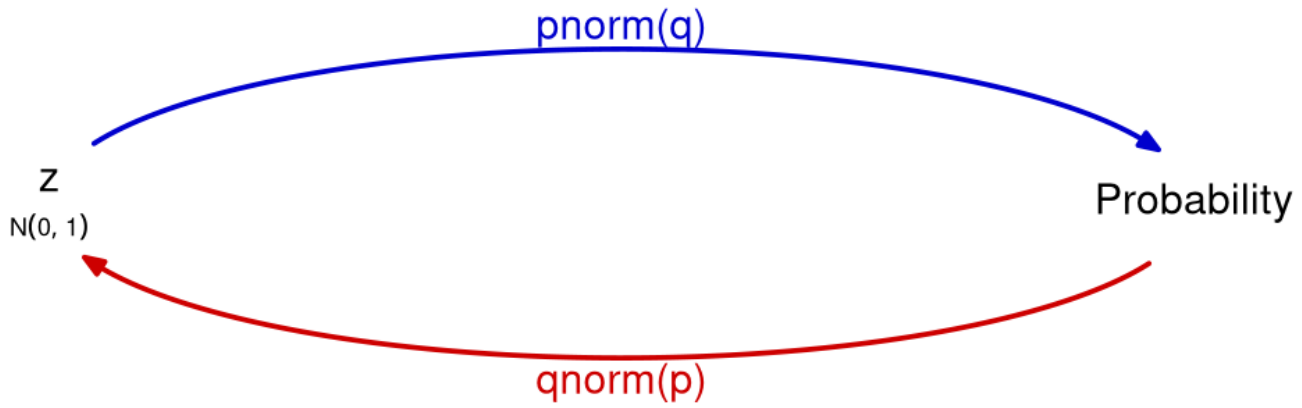



*Fig. 4.2.9* The `pnorm` and `qnorm` functions are inverse functions. For the `pnorm` function, we input the $z$-value, and the function outputs the corresponding probability $P(Z < z)$. For the `qnorm` function, we input a probability, and the function outputs the $z$-value so $P(Z < z)$ equals the probability we input.

Similar to the `pnorm` function, the `qnorm` function expects a probability with corresponding area in the lower tail of the standard normal distribution, not in the upper tail of the distribution. That is, if we input a probability into the `qnorm` function, the function will output the $z$-value so that the probability equals $P(Z < z)$, not $P(Z > z)$. As before, we can work around this limitation with a little arithmetic. The following examples illustrate the different cases.

---

## Example 4.1.6

Find $a$ so that $P(Z < a) = 0.85$.

### Solution

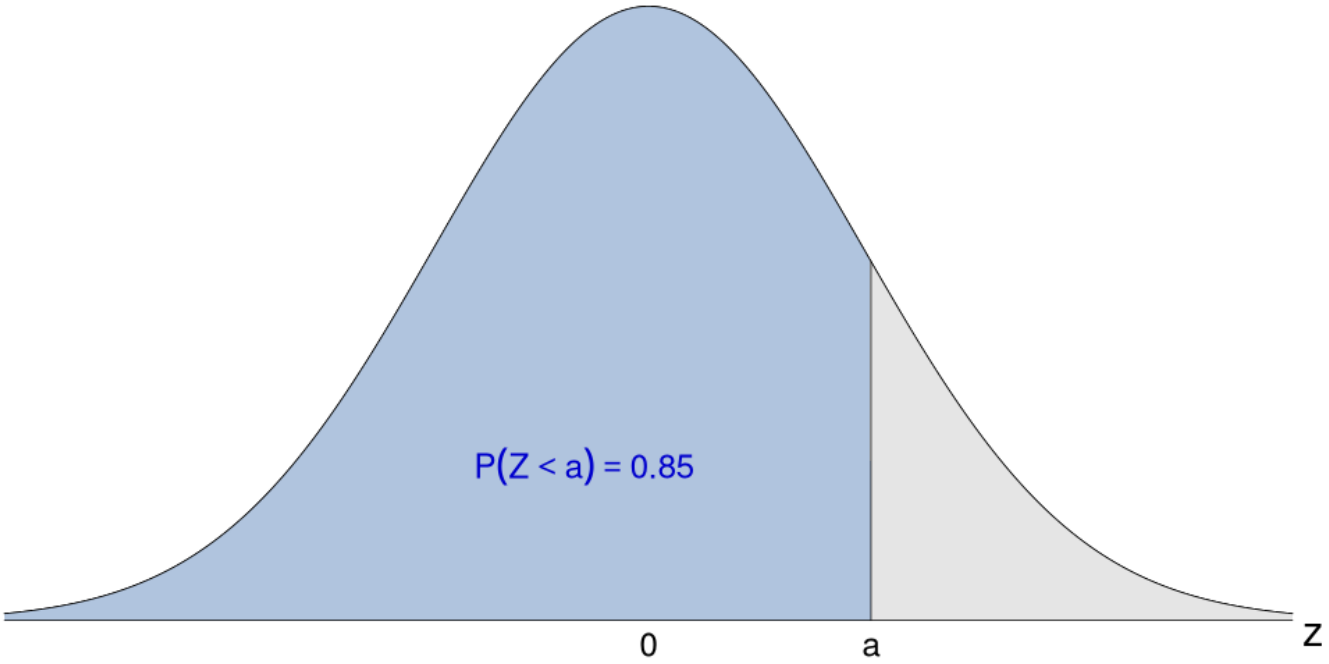



*Fig. 4.2.10* The standard normal distribution. The area under the curve to the left of some unknown value $z = a$ is shaded. We want to find $z = a$ so that $P(Z < a) = 0.85$.

We want to find some value $z = a$ so that $P(Z < a) = 0.85$. In this case, we already know the probability, and we want to find the corresponding $z$-value. That means we need to use the `qnorm` function.

```
qnorm(p = 0.85)
```

1.03643338949379

So $a = 1.03643$, and $P(Z < 1.03643) = 0.85$. In other words, we have found that $z = 1.03643$ is larger than $85\%$ of the values in the distribution.

---

## Example 4.1.7

Find $b$ so that $P(Z > b) = 0.59$.

### Solution

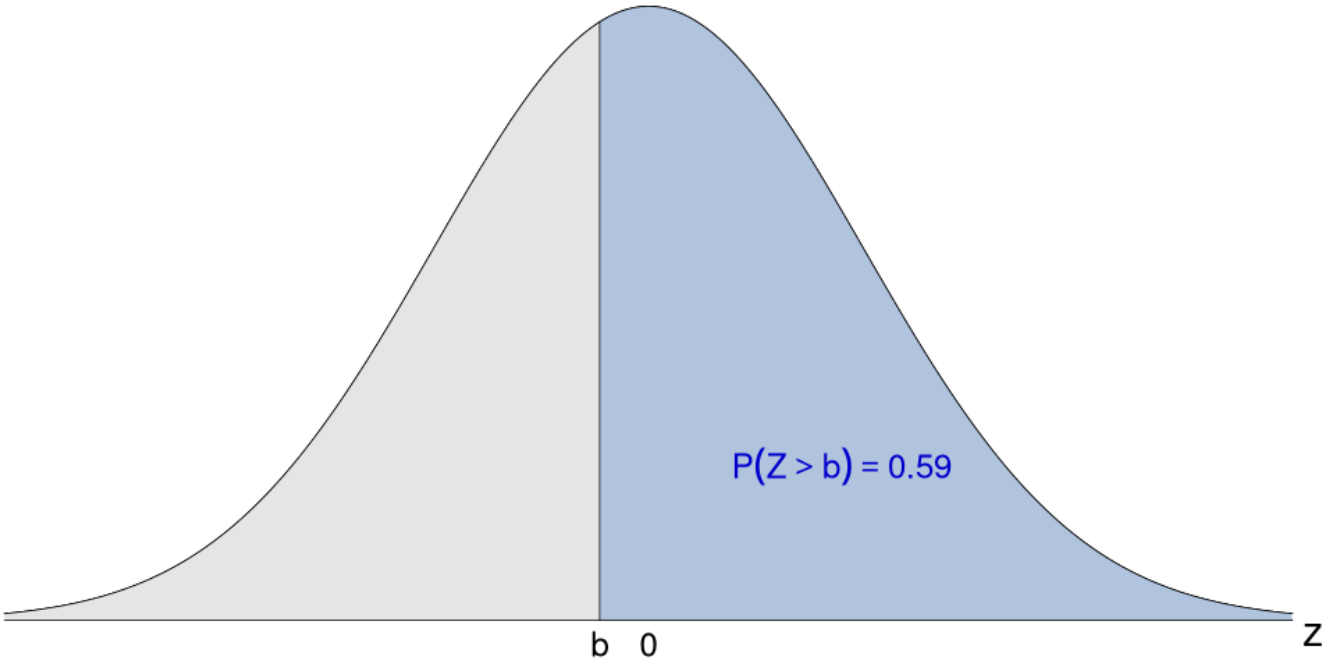



*Fig. 4.2.11* The standard normal distribution. The area under the curve to the right of some unknown value $z = b$ is shaded. We want to find $z = b$ so that $P(Z > b) = 0.85$.

We want to find some value $z = b$ so that $P(Z > b) = 0.59$. In this case, we already know the probability, and we want to find the corresponding $z$-value. That means we need to use the `qnorm` function.

However, the `qnorm` function expects a probability with corresponding area in the lower tail, but the area corresponding to our probability fills the upper tail of the distribution. To use the `qnorm` function, we first need to re-express the problem to one dealing with probability in the lower tail. To do so, note that since the total area under the curve is equal to $1$, it must be that $P(Z < b) = 1 - P(Z > b) = 1 - 0.59$. Since $P(Z < b)$ is the form of probability that `qnorm` expects as input, this formula gives us a way to find $b$ using the `qnorm` function.

```
qnorm(p = 1 - 0.59)
```

-0.227544976641149

So $b = -0.22754$, and $P(Z > -0.22754) = 0.59$. In other words, we have found that $z = -0.22754$ is less than than $59\%$ of the values in the distribution.

---

## Example 4.1.8

Find the $34$th percentile of the standard normal distribution.

### Solution

Skip to main content

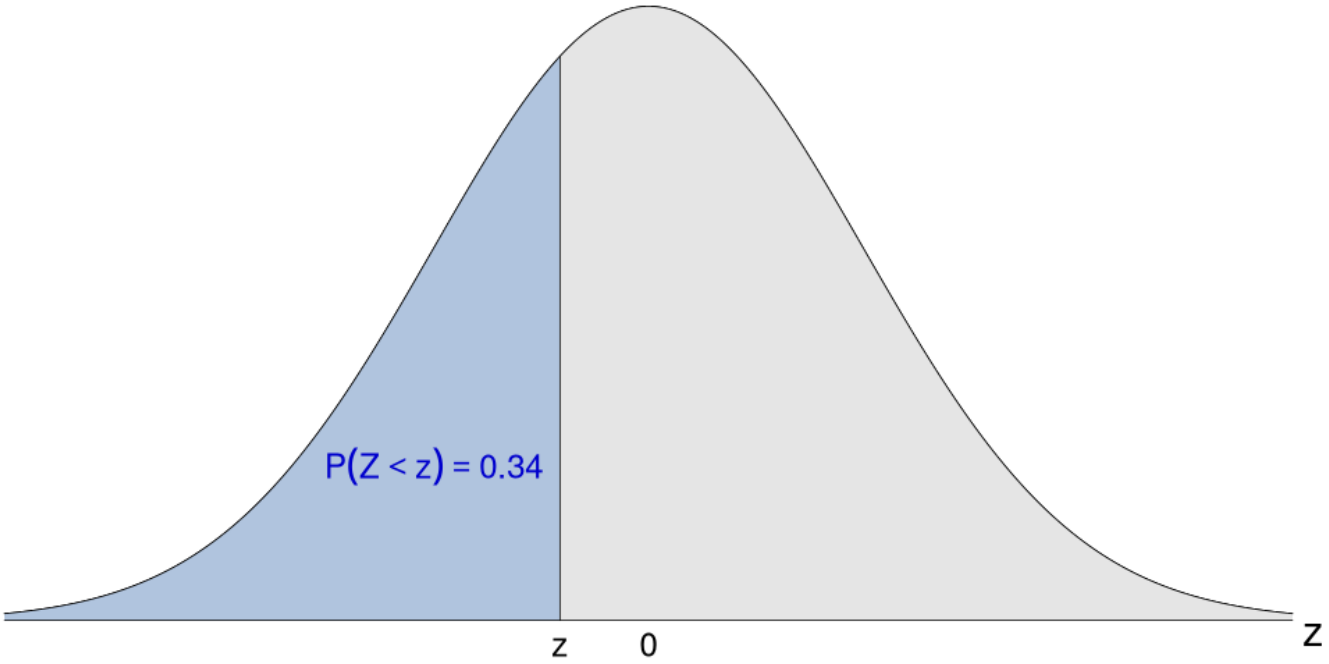



*Fig. 4.2.12* The standard normal distribution. The area under the curve to the left of some unknown value $z$ is shaded. We want to find $z$ so that $P(Z < z) = 0.34$.

Recall that the $34$th percentile of the standard normal distribution would be the $z$-value larger than $34\%$ of the possible values in the distribution. In the language of probability, we want to find $z$ so that $P(Z < z) = 0.34$. Then since we know the probability, we can use the `qnorm` function to find the corresponding $z$-value.

```
qnorm(p = 0.)
```

-0.412463129441405

So $P(Z < -0.41246) = 0.34$, meaning the $34$th percentile of the standard normal distribution is $z = -0.41246$.

---

## Example 4.1.9

Find the $z$-value less than $95\%$ of values in the standard normal distribution.

### Solution

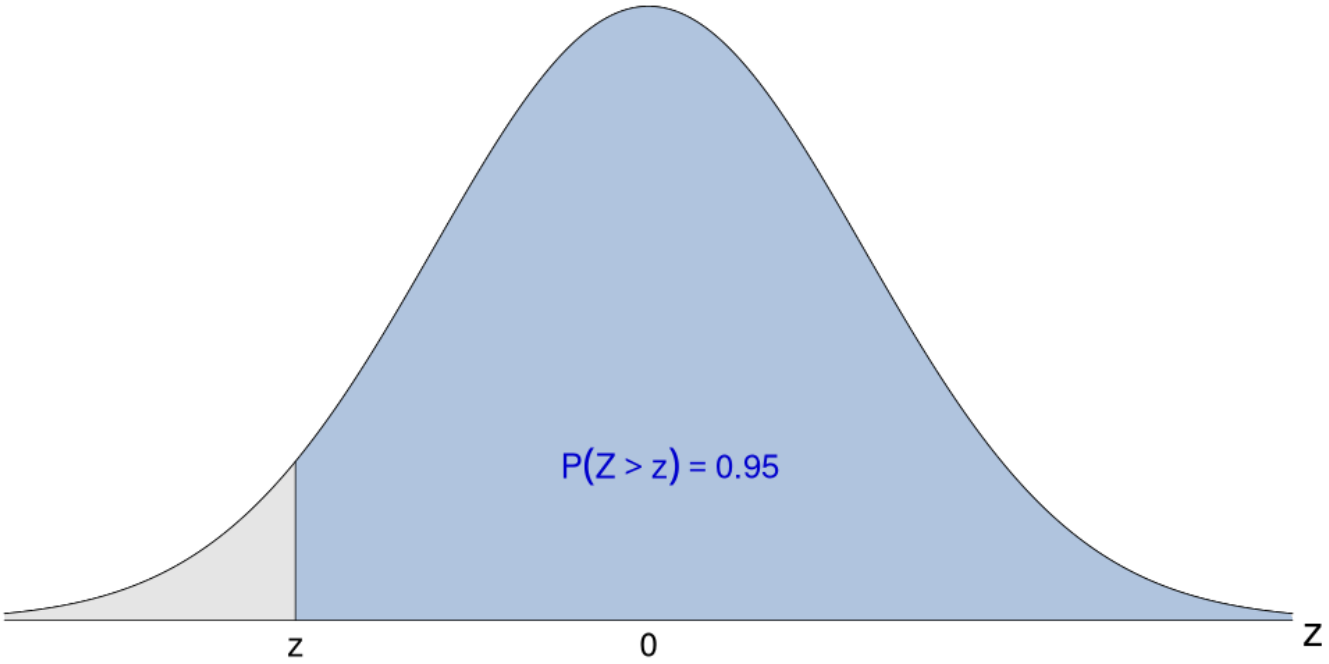



*Fig. 4.2.13* The standard normal distribution. The area under the curve to the right of some unknown value $z$ is shaded. We want to find $z$ so that $P(Z > z) = 0.95$.

We want the $z$-value smaller than $95\%$ of values in the standard normal distribution, which means we want $z$ so that $P(Z > z) = 0.95$. Since we know the probability, we want to use the `qnorm` function to find the corresponding $z$-score. However, the probability we are given corresponds to an area in the upper tail of the distribution. The `qnorm` function only works for probabilities in the lower tail of the distribution.

With a little arithmetic, we can see that if $P(Z > z) = 0.95$, then $P(Z < z) = 1 - P(Z > z) = 1 - 0.95 = 0.05$. In other words, if a $z$-value smaller than exactly $95\%$ of the values in the distribution, it is also larger than the remaining $5\%$ of the values. Now that we have the probability in the form the `qnorm` function expects, we can use the function to find $z$.

```
qnorm(p = 1 - 0.95)
```

-1.64485362695147

So $P(Z > -1.64485) = 0.95$, meaning $z = -1.64485$ is less than $95\%$ of the values in the standard normal distribution.

# 4.3. The Normal Distribution and $z$-Scores

## Objectives

- Use $z$-scores to find probabilities for non-standard normal distributions.
- Use $z$-scores to find quantiles for non-standard normal distributions.

## Finding Probability Given an $x$-value by Computing the $z$-score

Skip to main content

In the last section, we learned how to find probabilities for the standard normal distribution, $Z \sim N(0,1)$. To find probabilities generally for a non-standard normal distribution, $X \sim N(\mu, \sigma)$, we first convert the $x$-value of interest into a $z$**-score**.

The $z$-score of an $x$-value is how many standard deviations the $x$-value is away from the mean and in which direction. For example, consider $X \sim N(5, 2)$, the normal distribution with mean $\mu = 5$ and standard deviation $\sigma = 2$. Then $x = 11$ is three standard deviations above (or to the right of) the mean since

$$11 = \mu + 3\sigma = 5 + 3(2),$$

so the $z$-score of $x = 11$ is $z = 3$.

Similarly, $x = -3$ is four standard deviations below (or the the left of) the mean since

$$-3 = \mu + (-4)\sigma = 5 + (-4)(2),$$

so the $z$-score of $x = -3$ is $z = -4$.

Generally, if $X \sim N(\mu, \sigma)$, an $x$-value and its $z$-score are related by the formula $x = \mu + z\sigma$. If we use algebra to rearrange this equation so that the $z$ variable is by itself, we find that formula for finding the $z$-score of an $x$-value is

$$z = \frac{x - \mu}{\sigma}.$$

When calculating a probability for a non-standard normal distribution, we first find the $z$-score corresponding to an $x$-value because it transforms the distribution we are working with from a non-standard normal distribution $X \sim N(\mu, \sigma)$ to the standard normal distribution $Z \sim N(0, 1)$. From the previous section, we know how to find a probability of the standard normal distribution using the `pnorm` function. So once we tranform the distribution to the standard normal distribution by finding the $z$-score, we can use the `pnorm` function to find the probability.

For example, suppose we want to find the probability $P(X < x)$ to the left of an $x$-value in a non-standard normal distribution $X \sim N(\mu, \sigma)$. We first calculate the $z$-score using the formula $z = \frac{x-\mu}{\sigma}$. Then the probability we are looking for, $P(X < x)$, is equal to the probability $P(Z < z)$ to the left of the $z$-score in the standard normal distribution $Z \sim N(0, 1)$. So we can find $P(X < x)$ by first finding the $z$-score of $x$, then calculating $P(Z < z)$ using the `pnorm` function.

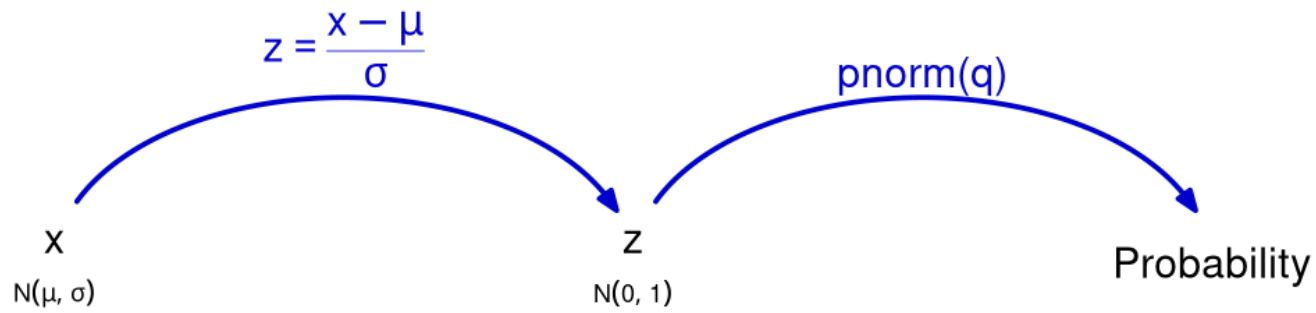



*Fig. 4.3.1* To find the probability associated with an $x$-value in a non-standard normal distribution, first tranform the problem to be over the standard normal distribution by finding the $z$-score using the formula $z = \frac{x-\mu}{\sigma}$, then use the `pnorm` function to compute the probability.

Skip to main content

## Example 4.2.1

Let $X \sim N(15, 2.4)$.

1. Find $P(X < 13.5)$.
2. Find $P(X > 16.1)$.
3. Find $P(12.4 < X < 19.5)$.

### Solution

Recall that $X \sim N(15, 2.4)$ means that $X$ is a normally distributed random variable with mean $\mu = 15$ and standard deviation $\sigma = 2.4$.

### Part 1

To find $P(X < 13.5)$, we first need to calculate the $z$-score of $x = 13.5$:

$$z = \frac{x - \mu}{\sigma} = \frac{13.5 - 15}{2.4} = -0.625.$$

So the $z$-score of $x = 13.5$ is $z = -0.625$, meaning $P(X < 13.5) = P(Z < -0.625)$.

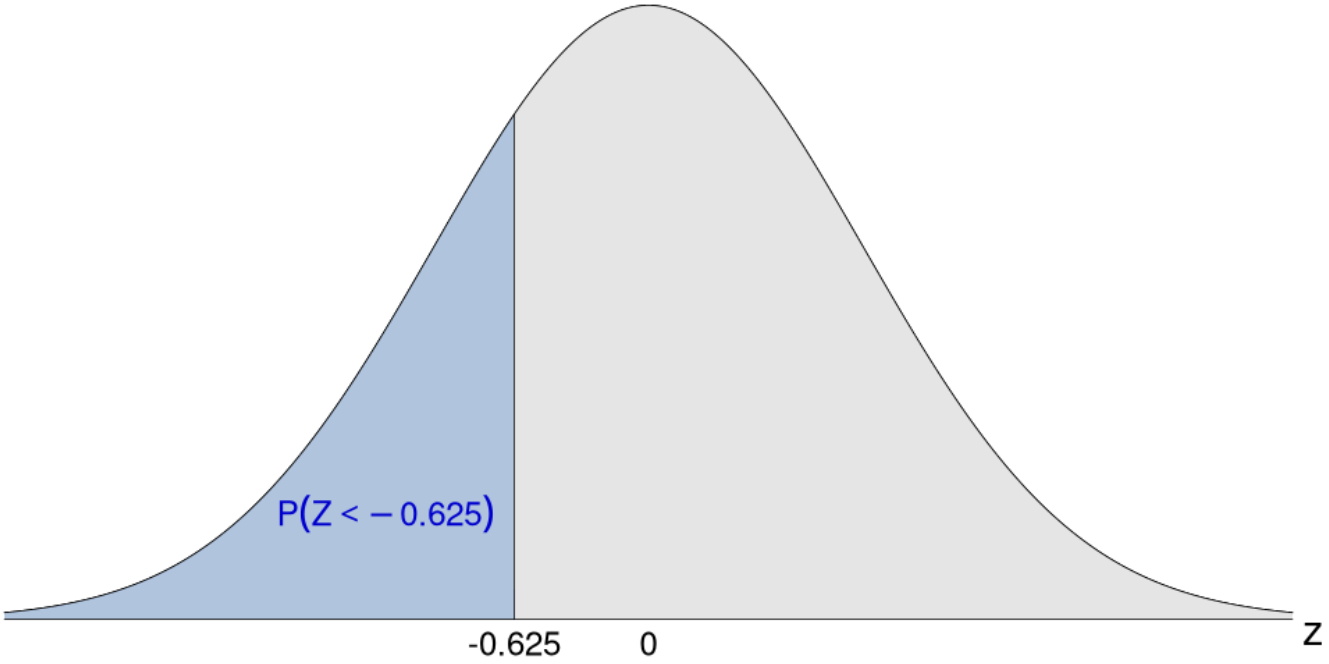



*Fig. 4.3.2* The standard normal distribution with the area equal to $P(Z < -0.625)$ shaded.

Now that we have converted the probability from a non-standard normal distribution to the standard normal distribution, we can use R to find the probability.

```
pnorm(q = -0.625)
```

Skip to main content

Thus, $P(X < 13.5) = P(Z < -0.625) = 0.26599$.

Part 2

For this part, we want to find $P(X > 16.1)$. To calculate this probability, we first find the $z$-score of $x = 16.1$:

$$z = \frac{x - \mu}{\sigma} = \frac{16.1 - 15}{2.4} = 0.45833.$$

Thus $P(X > 16.1) = P(Z > 0.45833)$.

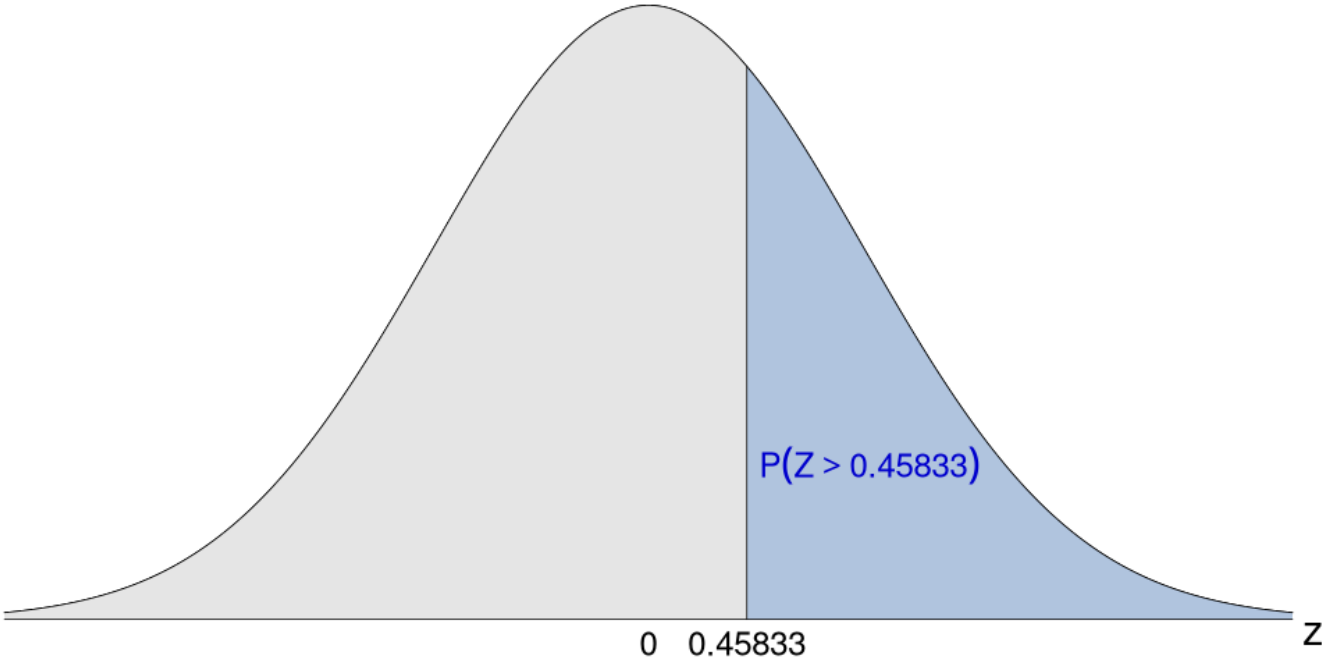



*Fig. 4.3.3* The standard normal distribution with the area equal to $P(Z > 0.45833)$ shaded.

Next, we use R to calculate the probability. Recall that the `pnorm` function only gives the area to the left of the input $z$-value. To find the area to the right of a $z$-value, we start with the total area under the standard normal distribution (which is $1 = 100\%$), then subtract away the part that we don't want, which is the area left of the $z$-value.

```
1 - pnorm(q = 0.45833)
```

0.323357687025688

We find that $P(X > 16.1) = P(Z > 0.45833) = 0.32336$.

Part 3

We want to find $P(12.4 < X < 19.5)$. We need to find two $z$-scores: one for $x = 12.4$, and one for $x = 17.5$. The $z$-score of $x = 12.4$ is

The $z$-score of $x = 19.5$ is

$$z = \frac{x - \mu}{\sigma} = \frac{19.5 - 15}{2.4} = 1.875.$$

So $P(12.4 < X < 19.5) = P(-1.08333 < Z < 1.875)$.

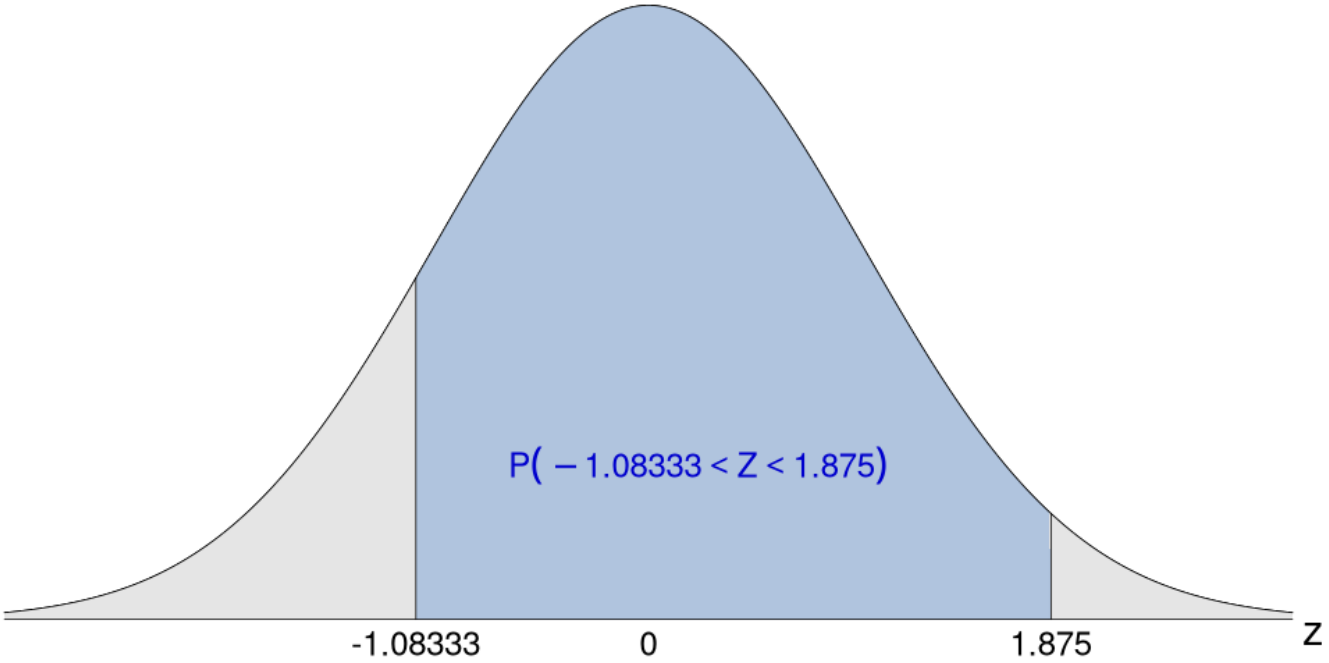



*Fig. 4.3.4* The standard normal distribution with the area equal to $P(-1.08333 < Z < 1.875)$ shaded.

Next, we'll use R to calculate the probability. Since we want the probability between two $z$-values, recall that we first calculate all the area to the left of the larger $z$-value, then subtract the excess area to the left of the smaller $z$-value.

```
pnorm(q = 1.875) - pnorm(q = -1.08333)
```

0.830272651276599

Thus, $P(12.4 < X < 19.5) = P(-1.08333 < Z < 1.875) = 0.83027$.

---

## Example 4.2.2

The heights of adult men in the United States are normally distributed with a mean of 70 inches (5 ft, 10 in) and a standard deviation of 4 inches. If you choose an adult male from the U.S. population at random, what is the probability that:

1. he will be shorter than $60$ inches ($5$ ft)?
2. he will be taller than $72$ inches ($6$ ft)?
3. he will be between $60$ inches and $72$ inches tall?

Skip to main content

Solution

The distribution of the heights of adult males is $X \sim N(70, 4)$. This is a non-standard normal distribution. In each case, to find the probabilities, we will first convert to the standard normal distribution by finding the appropriate $z$-scores.

Part 1

We want $P(X < 60)$. To find the $z$-score associated with $x = 60$, we calculate

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 70}{4} = -2.5.$$

So the $z$-score of $x = 60$ is $z = -2.5$. This means $P(X < 60) = P(Z < -2.5)$.

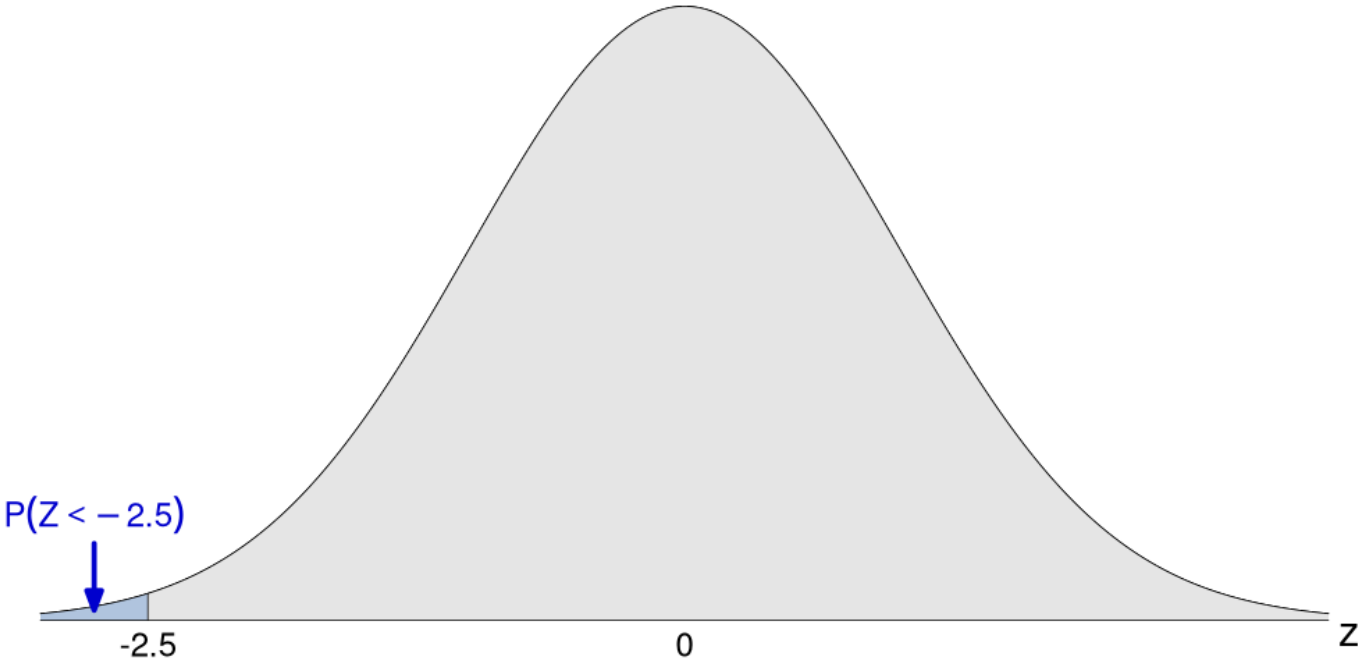



*Fig. 4.3.5* The standard normal distribution with the area equal to $P(Z < -2.5)$ shaded.

Now that we have converted the probability from a non-standard normal distribution to the standard normal distribution, we can use R to find the probability.

```
pnorm(q = -2.5)
```

0.00620966532577613

Then $P(X < 60) = P(Z < -2.5) = 0.00621$. In other words, there is only a $0.621\%$ chance that the adult male you randomly choose from the population is less than $5$ feet tall.

Part 2

We want $P(X > 72)$. To find the $z$-score associated with $x = 72$, we calculate

$$z = \frac{x - \mu}{\sigma} = \frac{72 - 70}{4} = 0.5.$$

So the $z$-score of $x = 72$ is $z = 0.5$. This means $P(X > 72) = P(Z > 0.5)$.

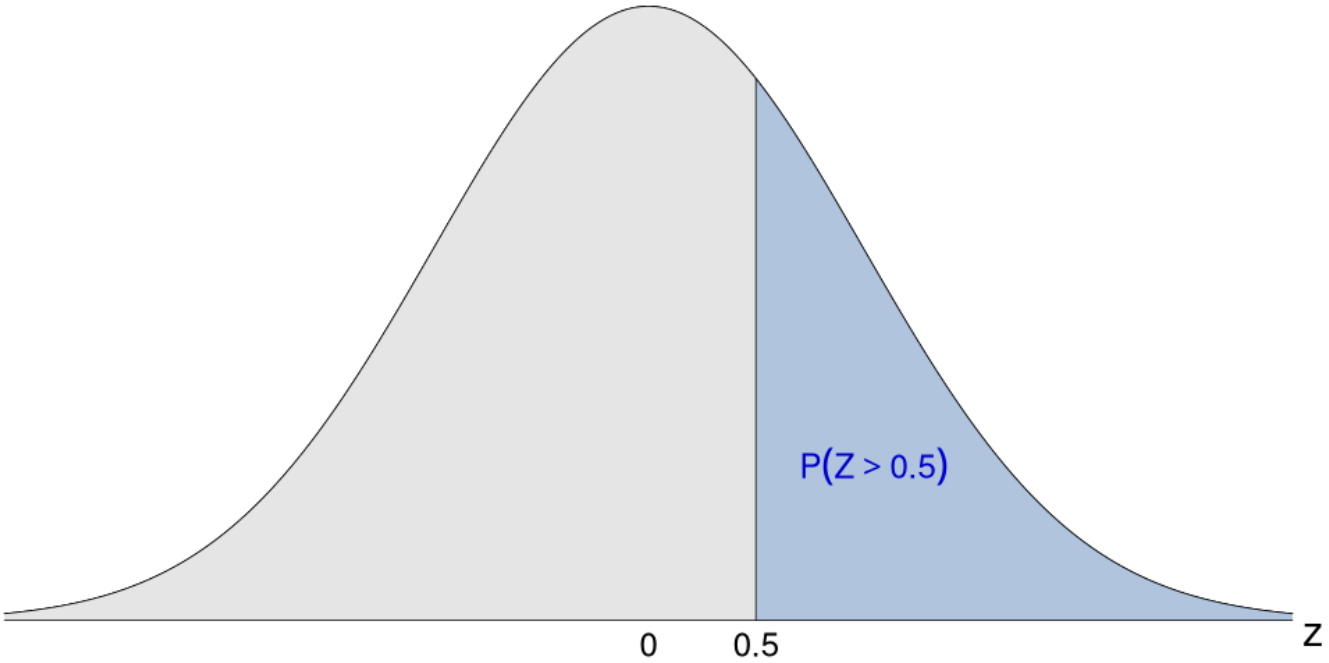



*Fig. 4.3.6* The standard normal distribution with the area equal to $P(Z > 0.5)$ shaded.

Now that we have converted the probability from a non-standard normal distribution to the standard normal distribution, we can use R to find the probability.

```
1 - pnorm(q = 0.5)
```

0.308537538725987

So $P(x > 72) = P(z > 0.5) = 0.30854$. In other words, there is a $30.854\%$ chance that the adult male you randomly choose from the population is more than $6$ feet tall.

### Part 3

We want $P(60 < X < 72)$. We need a $z$-score for both $x = 60$ and $x = 72$. Fortunately, we've already done the hard work. We already found the $z$-score of $x = 60$ is $z = -2.5$ in part 1, and we found the $z$-score of $x = 72$ is $z = 0.5$ in part 2. This means that

$$P(60 < X < 72) = P(-2.5 < Z < 0.5).$$

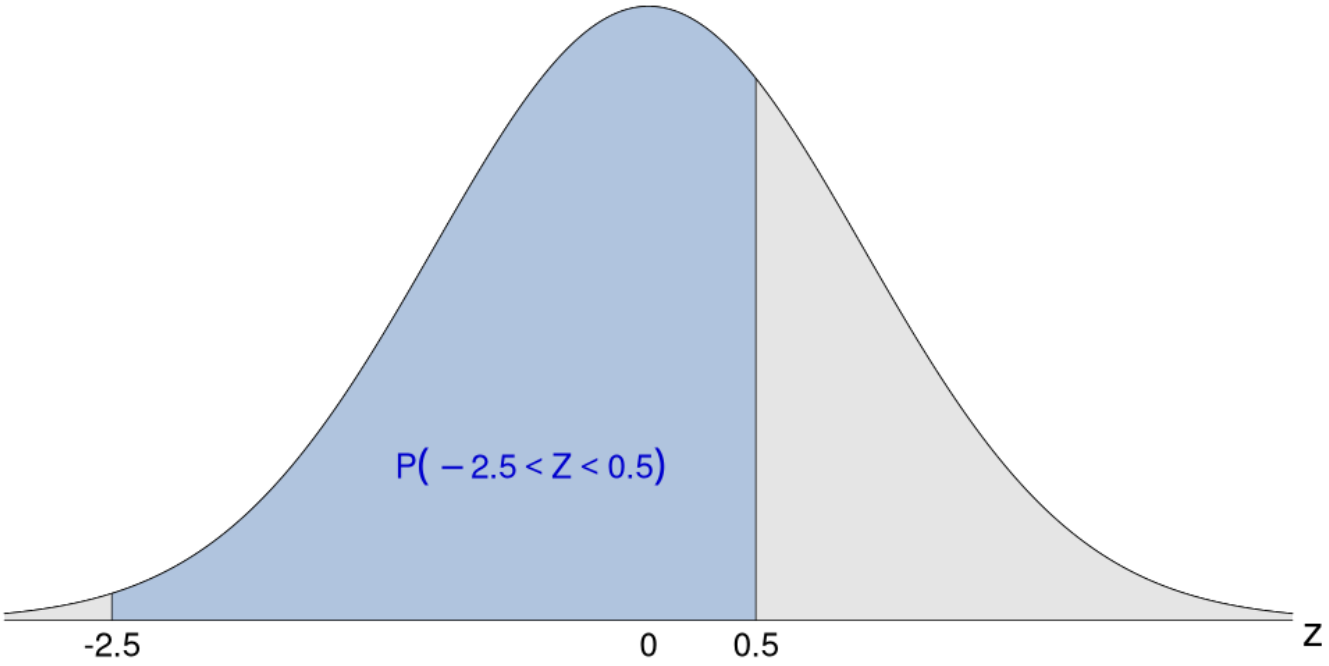



*Fig. 4.3.7* The standard normal distribution with the area equal to $P(-2.5 < Z < 0.5)$ shaded.

Now that we have converted the probability from a non-standard normal distribution to the standard normal distribution, we can use R to find the probability.

```
pnorm(q = 0.5) - pnorm(q = -2.5)
```

0.685252795948237

So $P(60 < X < 72) = P(-2.5 < Z < 0.5) = 0.68525$. That is, there is a $68.525\%$ chance that the adult male you choose is between $5$ feet and $6$ feet tall.

---

## Finding an $x$-value Given Probability by Computing the $z$-score

In the above examples, we saw how to find a probability involving a non-standard normal distribution. Next, we will learn how to do the inverse: given a probability, we want to find a corresponding $x$-value in a non-standard normal distribution. For example, we might want to find $x$ so that $P(X < x)$ is equal to a certain probability.

To do so, we first use the `qnorm` function (which is the inverse of the `pnorm` function) to find the $z$-score associated with our initial probability. Once we have the $z$-score, we use the formula

$$x = \mu + z\sigma$$

to calculate the $x$-value we are looking for.

Skip to main content

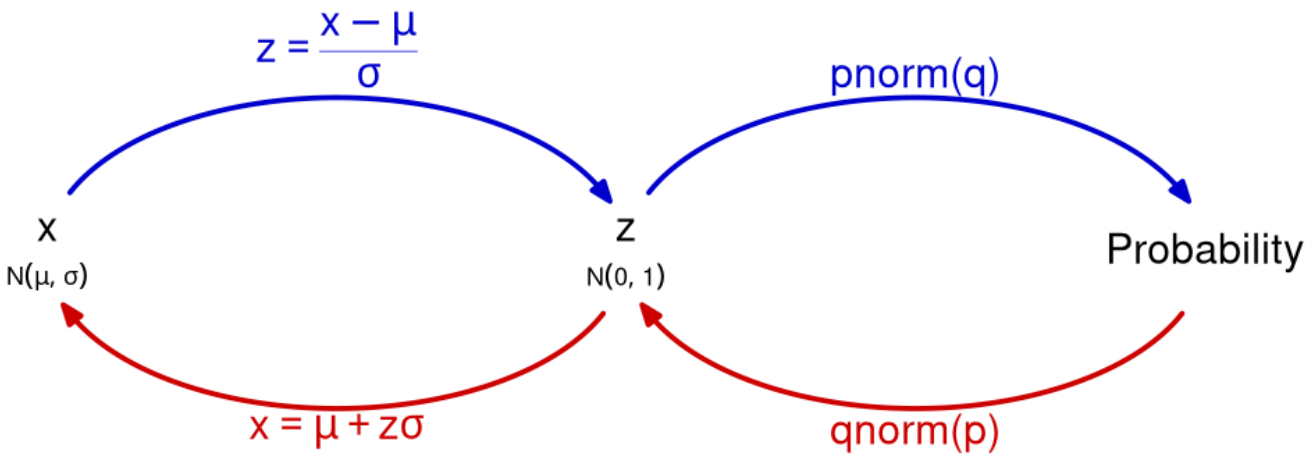



*Fig. 4.3.8* The approach we take to solve a problem involving a non-standard normal distribution depends on whether we know an $x$-value and want to find an associated probability, or we know a probability and want to find an associated $x$-value. But in either case, we must find a $z$-score as an intermediary step.

---

### Example 4.2.3

Let $X \sim N(5, 0.6)$.

1. Find $a$ so that $P(X < a) = 0.75$.
2. Find $b$ so that $P(X > b) = 0.9$.

#### Solution

First, observe that the random variable $X$ is normally distributed with mean $\mu = 5$ and standard deviation $\sigma = 0.6$.

#### Part 1

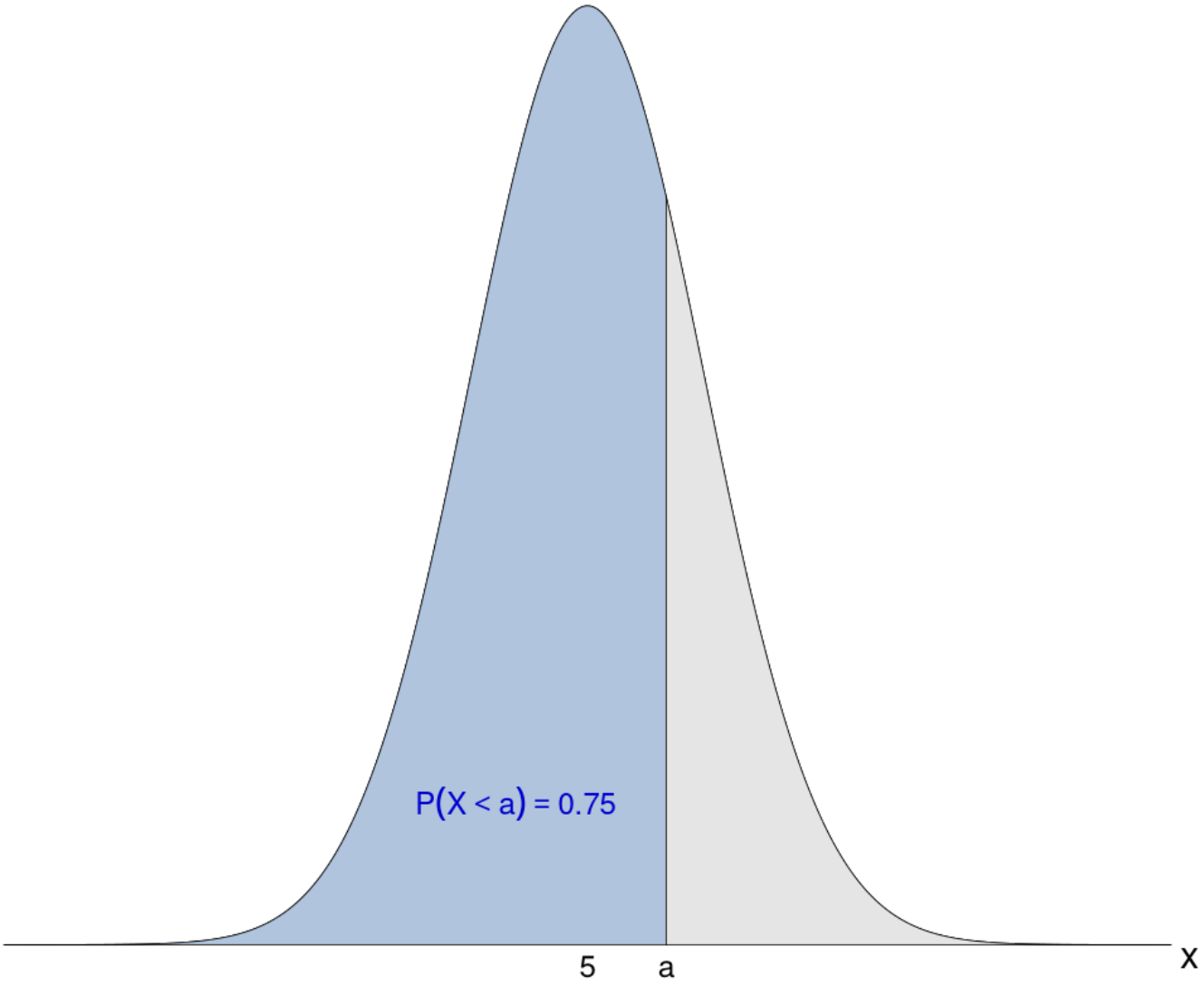



*Fig. 4.3.9* The distribution $N(5, 0.6)$ is shown. We want to find the value $a$ so that $P(X < a) = 0.75$.

We want to find $a$ so that $P(X < a) = 0.75$. Note that we are given a probability ($0.75$), and we want to find a particular $x$-value in the distribution ($x = a$). To do this, we first need to use the `qnorm` function to find the $z$-score associated with the probability:

```
qnorm(p = 0.75)
```

0.674489750196082

So the $z$-score is $z = 0.67449$, meaning $P(Z < 0.67449) = 0.7$.

Now that we have the $z$-score, we can calculate the value of $x = a$:

$$a = \mu + z\sigma = 5 + 0.67449(0.6) = 5.40469.$$

Thus, $P(X < 5.40469) = 0.75$.

Skip to main content

Part 2

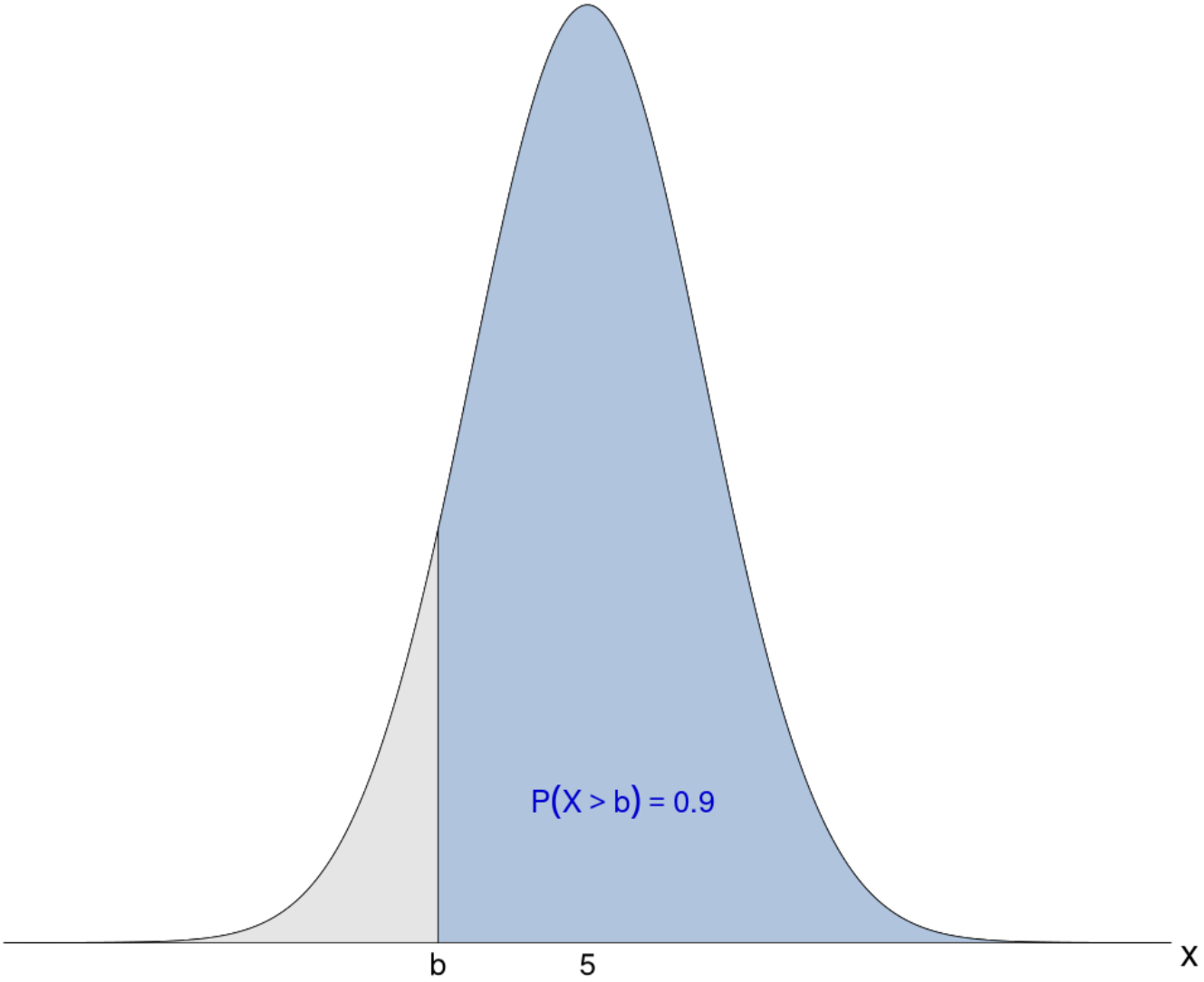



*Fig. 4.3.10* The distribution $N(5, 0.6)$ is shown. We want to find the value of $b$ so that $P(X > b) = 0.9$.

We want to find the value of $b$ so that $P(X > b) = 0.9$. Again, we are given a probability ($0.9$), and we need to find an $x$-value associated with that probability ($x = b$).

We will begin by using the `qnorm` function to find the $z$-value associated with the probability $0.9$. However, recall that the `qnorm` function assumes that the input represents a *left*-tailed probability. The probability of $0.9$ in this example is a *right*-tailed probability. To use the `qnorm` function, we first observe that since the probability to the right of $b$ is $P(X > b) = 0.9$, the probability to the left of $b$ is $P(X < b) = 1 - P(X > b) = 1 - 0.9$. With this observation, we are able to calculate the $z$-score:

```
qnorm(p = 1 - 0.9)
```

-1.2815515655446

We get $z = -1.28155$.

Skip to main content

$$b = \mu + z\sigma = 5 + -1.28155(0.6) = 4.23107.$$

Thus, $P(X > 4.23107) = 0.9$.

---

## Example 4.2.4

Scores on a calculus test are normally distributed with a mean score of $73$ and a standard deviation of $9$.

1. Find the $80$th percentile.
2. The highest $32\%$ of scores are higher than which score?

### Solution

Part 1

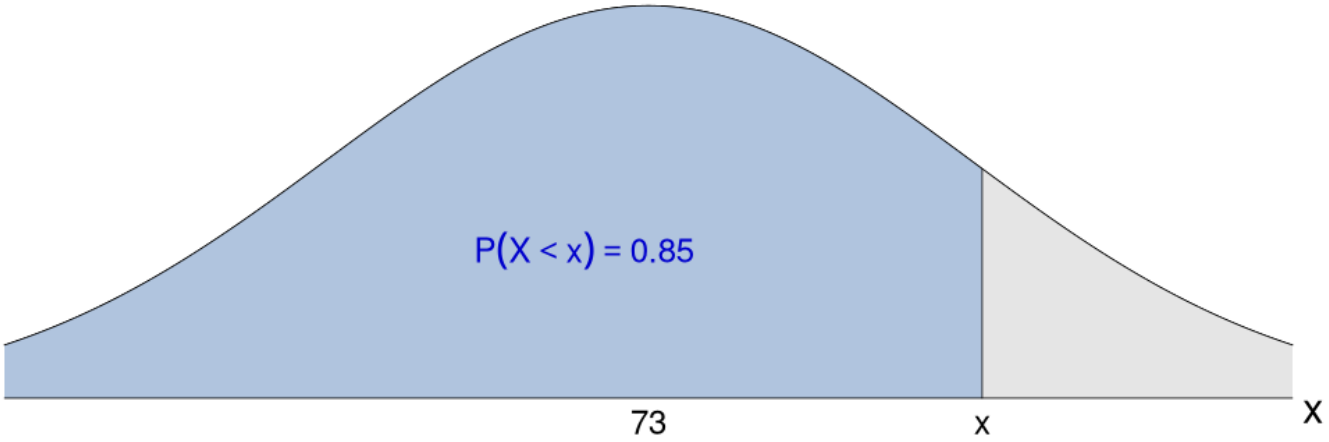



*Fig. 4.3.11* The distribution $N(73, 9)$ is shown. We want to find the $x$-value that is the $85$th percentile.

In this example, the $85$th percentile is the test score that is larger than $85\%$ of all test scores. Note that we are given a probability ( $85\% = 0.85$), and we need to find a particular $x$-value (the test score that is the $85$th percentile). We will first use the `qnorm` function to find the $z$-score associated with the given probability.

```
qnorm(p = 0.85)
```

1.03643338949379

We get $z = 1.03643$. This $z$-score is the $85$th percentile of the *standard* normal distribution. To find the $x$-value that is the $85$th percentile of test scores, we calculate

$$x = \mu + z\sigma = 73 + 1.03643(9) = 82.32790.$$

So a test score of $x = 82.32790$ is greater than $85\%$ of scores on the calculus test.

Skip to main content

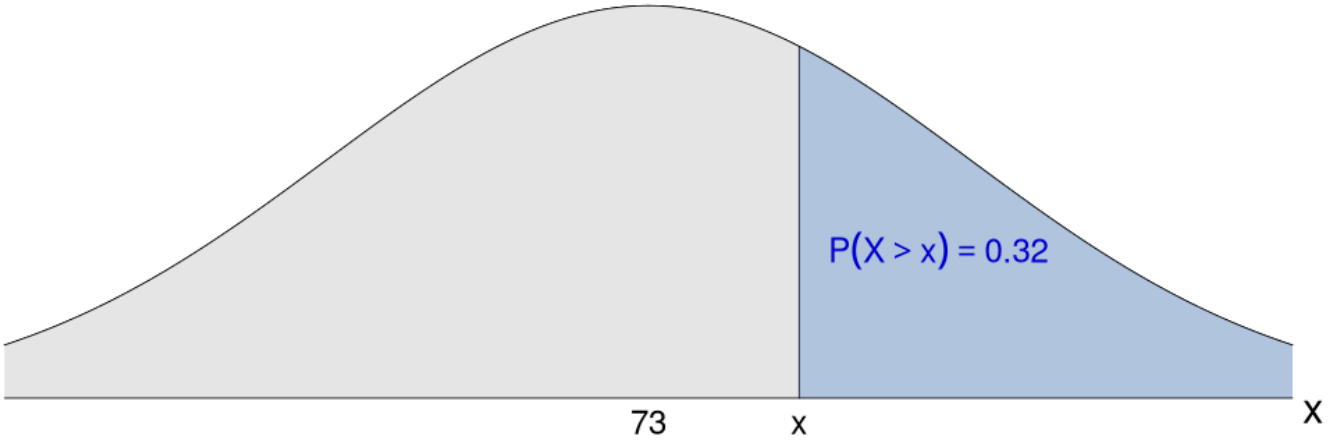



*Fig. 4.3.12* The distribution $N(73, 9)$ is shown. We want to find the $x$-value with $36\%$ of test scores higher than it.

Again, we are given a probability ($32\% = 0.32$), and we want to find an $x$-value (the test score with $32\%$ of scores higher than it). We can use the `qnorm` function to find the $z$-score associated with the probability. However, the probability we are given is located in the *upper* tail of the distribution. The `qnorm` function expects probability in the *lower* tail of the distribution. But this is easy to calculate. Since the probability in the upper tail of the distribution is $0.32$, the area in the lower tail of the distribution is $1 - 0.32$.

```
qnorm(p = 1 - 0.32)
```

0.467698799114508

We get $z = 0.46770$. This means that the $32\%$ of possible $z$-values in the *standard* normal distribution are larger than $z = 0.46770$. To find the $x$-value which is the test score with $32\%$ of test scores greater than it, we calculate

$$x = \mu + z\sigma = 73 + 0.46770(9) = 77.20929.$$

So $32\%$ of test scores were higher than a score of $77.20929$.

# 4.4. The Central Limit Theorem for Means

## Objectives

- Apply the central limit theorem for means to identify sampling distributions of sample means.
- Compute probabilities involving sampling distributions of means.

## The Central Limit Theorem

Suppose we want to find the average hourly wage of workers in Brazil. There are millions of workers in Brazil, so we can't practically survey all Brazilian workers to find the true population mean of their hourly wages. Instead, we can approximate the population mean by surveying a random sample of, say, $500$ Brazilian workers and calculating the mean of this smaller sample. This sample mean would probably not be *exactly* equal to the population mean, but the sample mean would likely be a good *approximation* of the population mean.

Skip to main content

But our random sample of $500$ Brazilian workers is only one particular sample of many random samples that are possible. Because there are millions of Brazilian workers, there are many possible ways a sample of $500$ Brazilian workers might be randomly chosen. For instance, if we were to collect a second random sample of $500$ Brazilian workers, this second sample would almost certainly include $500$ different workers than those workers which were selected for our first sample. The sample mean from this new second sample would probably be a little different than the sample mean from our first sample, though both sample means would probably give a good approximation of the population mean.

So far, we've imagined taking just two different samples of $500$ Brazilian workers from the population. Now imagine *all* the possible ways we can sample $500$ Brazilian workers from the population. How would the different sample means from all those samples be distributed? Well, like with the two sample means we considered above, we would expect most of the possible sample means to be close to the population mean. In fact, the further we get from the population mean, the fewer sample means we will to find.

For example, it is *possible* that in random sample of $500$ Brazilian workers, just by chance, all $500$ workers are extremely well-paid. The sample mean of this sample would be considerably higher than the population mean since we only sampled well-paid workers. However, though it is *possible* to randomly select a sample like this, it is *extremely unlikely*. The vast majority of possible samples would include workers at a variety of income levels. Only a very small proportion of samples would have a sample mean as far from the population mean as in this example.

We call the distribution of sample means taken from the possible random samples of a population the **sampling distribution** of the sample mean. From the above discussion, we expect the sampling distribution to be distributed like a bell curve, with most sample means clustered near the population mean, and few samples means far from the population mean.

If we formalize this idea mathematically, we get the **central limit theorem** for sample means:

**Theorem 4.4.1 (The Central Limit Theorem for Sample Means)**

Let $X$ be a random variable for a population with mean $\mu$ and standard deviation $\sigma$. Let $\bar{X}$ be the random variable of the sampling distribution of the sample mean, where the samples drawn from the population are of size $n$. If the sample size $n$ is big enough (so that $n \geq 30$), then the sampling distribution is roughly a normal distribution with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$. That is,

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

There are several important observations about this theorem that we should note:

- The central limit theorem does not make any requirements on how the population is distributed. This is why the central limit theorem is so powerful. The population can have any distribution. We do not even need to know how the population is distributed. No matter how the population is distributed, the sampling distribution of sample means (with $n \geq 30$) will approximate a normal distribution.
- The theorem states that the distribution of sample means $\bar{X}$ only *approximates* a normal distribution. But in practice, as long as the sample size $n \geq 30$, the distribution of sample means approximates the normal distribution so closely that we can treat the distribution as if it is normally distributed. (In the special case that the population is also normally distributed, we can drop the requirement that $n \geq 30$. If the population is normally distributed, then the sampling distribution will also be normally distributed for any sample size $n$.)
- The standard deviation of the population, $\sigma$, is different from the standard deviation of the sampling distribution, $\frac{\sigma}{\sqrt{n}}$. To

- Note that the larger the sample size $n$ is, the smaller the standard error $\frac{\sigma}{\sqrt{n}}$ becomes. That is, sample means from larger samples will tend to better approximate the population mean than sample means from smaller samples.
- The central limit theorem only applies for samples that are randomly selected. If a sample is not selected randomly, the central limit theorem may not apply.

In the previous section, we learned that if we have a normal distribution with mean $\mu$ and standard deviation $\sigma$, the formulas

$$z = \frac{x - \mu}{\sigma}, \qquad x = \mu + z\sigma,$$

allow us to calculate a $z$-score from an $x$-value or and $x$-value from a $z$-score. We use these same formulas in this section, but because we are focusing on the sampling distribution for sample means instead of just the population distribution, some of the variables in the formula look a little different. In particular, we use $\bar{x}$ instead of $x$ (since we are dealing with the distribution of sample means, not particular values of the population), and we use the standard error $\frac{\sigma}{\sqrt{n}}$ instead of the population standard deviation $\sigma$ (since the standard error is the standard deviation of the sampling distribution). Making these substitutions, the formulas above become

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}, \qquad \bar{x} = \mu + z\frac{\sigma}{\sqrt{n}}.$$

---

## Example 4.4.1

The expected value of a fair six-sided die roll is $\mu = 3.5$. The standard deviation is $\sigma = 1.7078$.

1. If we roll the die $50$ times, what is the likelihood that the sample mean is smaller than $3.0$?
2. If we roll the die $100$ times, what is the likelihood that the sample mean is smaller than $3.0$?
3. Is it more likely that our sample mean is smaller than $3.0$ if we roll the die $50$ times or $100$ times? Why?

### Solution

#### Part 1

By the central limit theorem, we know that the distribution of sample means of samples of size $n = 50$ are normally distributed with population mean

$$\mu = 3.5$$

and with standard error

$$\frac{\sigma}{\sqrt{n}} = \frac{1.7078}{\sqrt{50}} = 0.2415$$

We want to find $P(\bar{X} < 3.0)$. Since $\bar{X} \sim N(3.5, 0.2415)$, we first find the $z$-score for $\bar{x} = 3.0$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.0 - 3.5}{0.2415} = -2.0704.$$

So $P(\bar{X} < 3.0) = P(Z < -2.0704)$. We use R to find the probability.

```
pnorm(q = -2.0704)
```

0.0192074509255941

So $P(\bar{X} < 3.0) = P(Z < -2.0704) = 0.0192$. That is, there is only a $1.92\%$ chance that our $50$ die rolls have an average of less than $3.0$.

### Part 2

By the central limit theorem, we know that the distribution of sample means of samples of size $n = 100$ are normally distributed with population mean

$$\mu = 3.5$$

and with standard error

$$\frac{\sigma}{\sqrt{n}} = \frac{1.7078}{\sqrt{100}} = 0.1708.$$

We want to find $P(\bar{X} < 3.0)$. Since $\bar{X} \sim N(3.5, 0.1708)$, we first need to find the $z$-score associated with $\bar{x} = 3.0$:

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{3.0 - 3.5}{0.1708} = -2.9274.$$

So $P(\bar{X} < 3.0) = P(Z < -2.9274)$. We use R to find the probability.

```
pnorm(q = -2.9274)
```

0.00170904480232814

So $P(\bar{X} < 3.0) = P(Z < -2.9274) = 0.0017$. That is, there is only a $0.17\%$ chance that our $100$ die rolls have an average of less than $3.0$.

### Part 3

The probability of getting a sample mean of less than $3.0$ after $50$ die rolls is $1.92\%$. But the probability of getting a sample mean of less than $3.0$ after $100$ die rolls is only $0.17\%$, more than $10$ times smaller. While both are unlikely, we are much more likely to have a sample mean of less than $3.0$ after only $50$ die rolls than we are after $100$ die rolls.

This is because the larger the sample size is, the smaller the standard error is, and, therefore, the more likely the sample mean from a randomly selected sample will be close to the population mean or expected value.

## Example 4.4.2

The length of time taken on the SAT for a group of students is normally distributed with a mean of $2.5$ hours and a standard deviation of $0.85$ hours. A sample size of $n = 30$ is drawn randomly from the population. Find the probability that the sample mean is between $2.3$ hours and $2.7$ hours.

### Solution

By the central limit theorem, sample means of samples of size $n = 30$ are normally distributed with

$$\mu = 2.5$$

and a standard error of

$$\frac{\sigma}{\sqrt{n}} = \frac{0.85}{\sqrt{30}} = 0.1552;$$

that is, $\bar{X} \sim N(2.5, 0.1552)$.

We want to find $P(2.3 < \bar{X} < 2.7)$. To do so, we first must find the $z$-scores associated with $\bar{x} = 2.3$ and $\bar{x} = 2.7$. We calculate

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.3 - 2.5}{0.1552} = -1.2887,$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{2.7 - 2.5}{0.1552} = 1.2887.$$

Then $P(2.3 < \bar{X} < 2.7) = P(-1.2887 < Z < 1.2887)$. We will use R to find the probability. We will first find *all* the area to the left of $z = 1.2887$, then subtract off the excess area to the left of $z = -1.2887$.

```
pnorm(q = 1.2887) - pnorm(q = -1.2887)
```

0.802497597503648

So $P(2.3 < \bar{X} < 2.7) = P(-1.2887 < Z < 1.2887) = 0.8025$. There is an $80.25\%$ chance that the sample of $30$ individuals has a sample mean test time of between $2.3$ and $2.7$ hours.

---

## Example 4.4.3

A statistics instructor at Mt. San Jacinto College assigns each student in his class to randomly sample $40$ other students at the college and find out how far they travel to get to campus, then calculate the sample mean. If the population mean is $26$ miles with a standard deviation of $32$ miles, then the top $10\%$ of sample means submitted by the instrutor's students should be greater than what value?

### Solution

We will need to approach this problem a little differently than the way we approached the last two examples, but we start off the same way: by using the central limit theorem to find the properties of the sampling distribution of sample means. Since we are looking at samples of size $n = 40$, the central limit theorem tells us that the sampling distribution is normally distributed with mean

$$\mu = 26$$

and a standard error of

$$\frac{\sigma}{\sqrt{n}} = \frac{32}{\sqrt{40}} = 5.05964.$$

So $\bar{X} \sim N(26, 5.05964)$.

In this problem, we are given a probability ($10\% = 0.10$), and we need to find an $\bar{x}$-value associated with it. We start by using R to find the $z$-score associated with the probability:

```
qnorm(p = 1 - 0.10)
```

1.2815515655446

So $P(Z > 1.28155) = 0.10$, meaing $10\%$ of all $z$-values are greater than $z = 1.28155$.

Then we calculate

$$\bar{x} = \mu + z\frac{\sigma}{\sqrt{n}} = 26 + 1.28155(5.05964) = 32.48420.$$

Then the top $10$ of sample means will be greater than $\bar{x} = 32.48420$ miles.

# 4.5. The Central Limit Theorem for Proportions

## Objectives

- Apply the central limit theorem for proportions to identify sampling distributions of sample proportions.
- Computer probabilities involving sampling distributions of proportions.

## Population Proportions

A **population proportion** is the fraction, ratio, or percentage of the population that possesses a certain characteristic. For example, in 2006, the **proportion** of adult Americans that were married was $55.7\%$. Note a few important properties about proportions:

- For each member of the population, there are only two options: either the member possesses the characteristic or the member doesn't possess the characteristic. In the above example, since $55.7\%$ of U.S. adults were married in 2006, we can conclude that $100\% - 55.7\% = 44.3\%$ of U.S. adults were not married in 2006.
- The value of a proportion is always between $0$ and $1$. In the above example, the proportion $55.7\% = 0.557$ lies between $0$

Skip to main content

- A proportion is the probability that, if we randomly select a member of the population, that the selected member will possess the characteristic. In the example above, if we were to randomly select an adult in the year 2006, there is a $55.7\%$ chance that the selected adult would be married.

A population proportion $p$ is calculated using the formula

$$p = \frac{X}{N},$$

where $X$ is the number of individuals in the population with the desired characteristic and $N$ is the size of the population.

For example, if we know that Menifee High School has a population of $4{,}582$ students, of whom $1{,}231$ are freshman, the proportion of Menifee High School students that are freshmen is

$$p = \frac{X}{N} = \frac{1{,}231}{4{,}582} = 0.2687.$$

So $26.87\%$ of all Menifee High School students are freshmen.

## The Central Limit Theorem for Proportions

We can also find the proportion of a sample, denoted by $\hat{p}$ (which is read "*p*-hat"). The formula for a sample proportion is

$$\hat{p} = \frac{x}{n},$$

where $x$ is the number of individuals in the *sample* with the desired characteristic, and $n$ is the size of the *sample*.

Recall from the central limit theorem for sample means says that, if the size of a sample is large enough, we can generally expect the mean of a sample to approximate the mean of the population. We might naturally ask the same question of proportions: if we take a large enough sample from a population, can we expect the sample proportion to be close to the population proportion?

The answer is "Yes". This is the idea of the **central limit theorem** for sample proportions.

**Theorem 4.5.1 (The Central Limit Theorem for Sample Proportions)**

Consider a population, and let $p$ be the proportion of the population that possesses some characteristic. Let $\hat{P}$ be the random variable of the sampling distribution of the sample proportions for this population, where the samples are of size $n$. If the sample size $n$ is big enough (so that $np \geq 5$ and $n(1-p) \geq 5$), then the sampling distribution is roughly a normal distribution with mean $p$ and standard deviation $\sqrt{\frac{p(1-p)}{n}}$. That is,

$$\hat{P} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

We reiterate the observations about the central limit theorem that we made in the previous section:

Skip to main content

distributed. No matter how the population is distributed, the sampling distribution of sample proportions (with $np \geq 5$ and $n(1-p) \geq 5$) will approximate a normal distribution.

- The theorem states that the distribution of sample proportions $\hat{P}$ only *approximates* a normal distribution. But in practice, as long as the sample size $n$ is large enough so that $np \geq 5$ and $n(1-p) \geq 5$, the distribution of sample proportions approximates the normal distribution so closely that we can treat the distribution as if it is normally distributed.
- We call the standard deviation of the sampling distribution, $\sqrt{\frac{p(1-p)}{n}}$, the **standard error** to distinguish it from the standard deviation of a population.
- Note that the larger the sample size $n$ is, the smaller the standard error $\sqrt{\frac{p(1-p)}{n}}$ becomes. That is, sample proportions from larger samples will tend to better approximate the population proportion than sample proportions from smaller samples.
- The central limit theorem only applies for samples that are randomly selected. If a sample is not selected randomly, the central limit theorem may not apply.

As in the previous section, we need to adjust our formulas

$$z = \frac{x - \mu}{\sigma}, \qquad x = \mu + z\sigma,$$

which allow us to calculate a $z$-score from an $x$-value or and $x$-value from a $z$-score, to use the appropriate symbols for the sampling distribution of sample proportions. We use $\hat{p}$ instead of $x$ (since we are dealing with the distribution of sample proportions, not particular values of the population); we use the mean of the sampling distribution, which is the population proportion $p$, instead of $\mu$; and we use the standard error $\sqrt{\frac{p(1-p)}{n}}$ instead of the population standard deviation $\sigma$ (since the standard error is the standard deviation of the sampling distribution). Making these substitutions, the formulas above become

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}, \qquad \hat{p} = p + z\sqrt{\frac{p(1-p)}{n}}.$$

---

## Example 4.4.1

A poll is conducted of $1{,}000$ Americans to determine whether or not they approve of the President of the United States. Suppose that, in reality, $54\%$ of all Americans approve of the President.

1. What is the probability that fewer than $50\%$ of those polled approve of the President?
2. What is the probability that more than $56\%$ of those polled approve of the President?
3. What is the probability that between $50\%$ and $56\%$ of those polled approve of the President?

### Solution

First, note that the sample size is $n = 1{,}000$ and the population proportion is $p = 0.54$. Then by the central limit theorem, we know that the sample proportions are normally distributed with mean

$$p = 0.54$$

and standard error

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.54(1-0.54)}{1000}} = 0.0158.$$

So $\hat{P} \sim N(0.54, 0.0158)$.

Part 1

We want to find $P(\hat{p} < 0.50)$. First find the $z$-score of $\hat{p} = 0.50$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.50 - 0.54}{0.0158} = -2.5316.$$

So $P(\hat{p} < 0.50) = P(z < -2.5316)$. Let's use R to calculate the probability.

```
pnorm(q = -2.5316)
```

0.0056771717907339

Then $P(\hat{p} < 0.50) = P(z < -2.5316) = 0.0057$. There is a $0.57\%$ chance that a sample of $1{,}000$ Americans will yield a sample proportion less than $50\%$.

Part 2

We want to find $P(\hat{p} > 0.56)$. First find the $z$-score of $\hat{p} = 0.56$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.56 - 0.54}{0.0158} = 1.2658.$$

So $P(\hat{p} > 0.56) = P(z > 1.2658)$. Let's use R to calculate the probability.

```
1 - pnorm(q = 1.2658)
```

0.102792347555097

Then $P(\hat{p} > 0.56) = P(z > 1.2658) = 0.1028$. There is a $10.28\%$ chance that a sample of $1{,}000$ Americans will yield a sample proportion more than $56\%$.

Part 3

We want $P(0.50 < \hat{p} < 0.56)$. We know from parts 1 and 2 that the $z$-score of $\hat{p} = 0.50$ is $z = -2.5316$ and the $z$-score of $\hat{p} = 0.56$ is $z = 1.2658$. So $P(0.50 < p' < 0.56) = P(-2.5316 < z < 1.2658)$.

The probability is the entire area under the normal density function between $z = -2.5316$ and $z = 1.2658$. To find this, we will use R to first find all the area to the left of the larger $z$-score, $z = 1.2658$, then subtract off the excess area to the left of the smaller $z$-score, $z = -2.5316$.

```
pnorm(q = 1.2658) - pnorm(q = -2.5316)
```

0.891530480654169

So $P(0.50 < \hat{p} < 0.56) = P(-2.5316 < z < 1.2658) = 0.8915$. There is an $89.15\%$ chance that a sample of $1{,}000$ Americans will have a sample proportion of between $50\%$ and $56\%$.

---

### Example 4.4.2

Apparently having nothing better to do, Yan flips $20$ coins, calculates the proportion of heads that come up, then repeats the process. We would expect $25\%$ of the time the proportion of heads Yan gets will be smaller than what value?

#### Solution

Let's start by calculating the properties of the sampling distribution according to the central limit theorem. We know the mean of the sampling distribution is the population proportion. We are not directly told what the population proportion is, but we know coins come up heads $50\%$ of the time, so the proportion of heads in the population is

$$p = 0.50.$$

We can use $p$ to calculate the standard error:

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(1-0.5)}{20}} = 0.11180.$$

So $\hat{P} \sim N(0.50, 0.11180)$.

Now we need to find the $\hat{p}$-value so that $25\%$ of possible sample proportions will be smaller than that $\hat{p}$-value.

We start by using R to find the $z$-score associated with the probability we are given.

```
qnorm(p = 0.25)
```

-0.674489750196082

This means $P(z < -0.67449) = 0.25$.

Now we use this $z$-score to calculate

$$\hat{p} = p + z\sqrt{\frac{p(1-p)}{n}} = 0.50 + 0.67449(0.11180) = 0.42459.$$

So $P(\hat{p} < 0.42459) = 0.25$. That is, when Yan tosses his $20$ coins, we expect the proportion of heads among the $20$ coins to be smaller than $\hat{p} = 0.42459 = 42.459\%$ only $25\%$ of the time.

## 5 Confidence Intervals

# 5.1. Estimating Population Means ($\sigma$ Known)

## Objectives

- Construct confidence intervals for population means in populations where the standard deviation is known.
- Identify and interpret the margin of error for a confidence interval.

## Introduction

Suppose you were trying to determine the mean rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a **point estimate** of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a **point estimate** for the true proportion.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. The sample data help us to make an estimate of a population parameter. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we can construct a range of values, called a confidence interval, where we have a high confidence the population parameter falls.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. It provides a range of reasonable values in which we expect the population parameter to fall. There is no guarantee that a given confidence interval does capture the parameter, but there is a predictable probability of success.

## Calculating the Confidence Interval

To construct a confidence interval for a single unknown population mean $\mu$, **where the population standard deviation $\sigma$ is known**, we need to know $\bar{x}$ and we need to know the **margin of error**, $E$. The sample mean $\bar{x}$ is a **point estimate** of the unknown population mean $\mu$. It is a single value or point that estimates the population parameter. The margin of error is a measure of how far $\bar{x}$ might be from $\mu$.

A confidence interval has the form

$$(\text{point estimate} - \text{margin of error},\ \text{point estimate} + \text{margin of error}),$$

or, in symbols,

$$(\bar{x}–E, \bar{x} + E).$$

We say $\bar{x} - E$ is the **lower bound** of the confidence interval, and $\bar{x} + E$ is the **upper bound** of the confidence interval.

The margin of error, $E$, depends on the **confidence level** (abbreviated $\mathrm{CL}$). The confidence level is often considered the probability that the calculated confidence interval estimate will contain the true population parameter. However, it is more accurate to state that the confidence level is the percent of confidence intervals that contain the true population parameter when repeated samples are taken. Most often, the person constructing the confidence interval chooses a confidence level of $90\%$ or higher because that person wants to be reasonably certain of their conclusions.

Skip to main content

Another probability, $\alpha$ (the Greek letter *alpha*) is the probability that the interval does not contain the unknown population parameter. Mathematically, $\alpha + \text{CL} = 1$.

---

## Example 5.1.1

Suppose we have sample data collected from a population. Imagine we've already found that the sample mean is $\bar{x} = 7$ and that the margin of error is $E = 2.5$ for a confidence level of $\text{CL} = 95\%$. The population mean is unkown. Construct a confidence interval for the population mean.

### Solution

Since we are told what $\bar{x}$ and $E$ are already, calculating the confidence interval is straightforward. The confidence interval is

$$(\bar{x} - E, \bar{x} + E) = (7 - 2.5, 7 + 2.5) = (4.5, 9.5).$$

So we are $95\%$ confident that the population mean $\mu$ is between $4.5$ and $9.5$.

> **Note**
>
> In the media, margins of error and confidence intervals are often expressed a little differently than how we have expressed them here. The media might say the margin of error for this example is $\pm 2.5$ and that the confidence interval is $7 \pm 2.5$. This is just a different way to express the same idea.

---

A confidence interval for a population mean $\mu$ with a known standard deviation $\sigma$ is based on the fact that the sample means follow an approximately normal distribution. Recall that by the central limit theorem, this sampling distribution is distributed as

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

But in real-world scenarios, we generally do not know the population mean $\mu$, which is also the mean of the sampling distribution. If we did know $\mu$, we would have no need to estimate it by calculating a confidence interval. So we approximate the population mean $\mu$ with the point estimate $\bar{x}$. By making this substitution, the distribution becomes

$$N\left(\bar{x}, \frac{\sigma}{\sqrt{n}}\right).$$

We will use this approximate sampling distribution to calculate our confidence interval.

**Note**

We replace $\mu$ in the sampling distribution with its point estimate $\bar{x}$, but why do we leave in the population standard deviation $\sigma$? If we don't know $\mu$, why should we expect to know $\sigma$? After all, we generally need to know $\mu$ to calculate $\sigma$.

This is a good question. Generally, if we do not know the mean of a population, we would *not* expect to know the standard deviation of the population. However, it turns out approximating $\sigma$ with the sample standard deviation $s$ in the sampling distribution formulas is more delicate and more complicated than approximating $\mu$ with $\bar{x}$. In a future section, we will learn how to use $s$ instead of $\sigma$ in the calculation. For now, since we are just beginning to learn about confidence intervals, we make the somewhat unrealistic assumption that we know the population standard deviation $\sigma$ but not the population mean $\mu$ in order to simplify the calculation.

Suppose we want to estimate the mean of a population, but we know the standard deviation is $\sigma = 3$. We sample $n = 35$ members of the population and calculate a sample mean of $\bar{x} = 52$. We want to construct a $90\%$ confidence interval for the population mean. To get a $90\%$ confidence interval, we must include the central $90\%$ of the probability of the normal distribution. If we include the central $90\%$, we leave out a total of $\alpha = 10\%$ in both tails of the normal distribution, or $\frac{\alpha}{2} = 5\%$ in each tail.

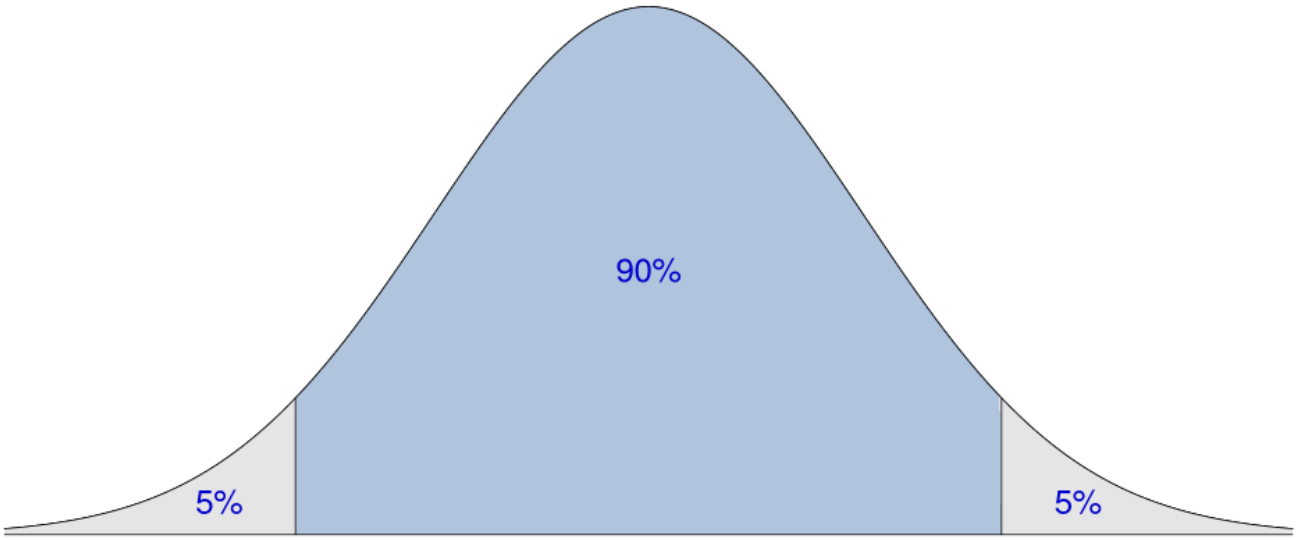



*Fig. 5.1.1* If we want to construct a confidence interval with $\text{CL} = 90\%$, that means $\alpha = 1 - \text{CL} = 10\%$. This is split between the two tails, so each tail has $\frac{\alpha}{2} = 5\%$.

We want to find the margin of error $E$ of our confidence interval. We know that the upper bound to our confidence interval is $\bar{x} + E$. But we also know that we can express the upper bound of our confidence interval using a $z$-score as $\bar{x} + z\frac{\sigma}{\sqrt{n}}$. Since $\bar{x} + E = \bar{x} + z\frac{\sigma}{\sqrt{n}}$, subtracting $\bar{x}$ off from both sides tells us that the margin of error is

$$E = z\frac{\sigma}{\sqrt{n}}.$$

For this particular problem, we want the $z$-score with an area of $0.05$ to the right (since we want $\frac{\alpha}{2} = 5$ probability to the right of

Skip to main content

```
qnorm(p = 1 - 0.05)
```

1.64485362695147

So the $z$-score we will use is $z = 1.645$. To make it clear what this $z$-score represents, we often write $z_{0.05} = 1.645$, meaning $1.645$ is the $z$-score that has an area of $0.05$ to its right.

Now we can calculate the margin of error:

$$E = z_{0.05}\frac{\sigma}{\sqrt{n}} = 1.645\left(\frac{3}{\sqrt{35}}\right) = 0.83409.$$

So the margin of error is $E = 0.83409$. Then the confidence interval is

$$(\bar{x} - E, \bar{x} + E) = (52 - 0.83409, 52 + 0.83409) = (51.16591, 52.83409).$$

We are $90\%$ confident that the population mean $\mu$ is between $51.16591$ and $52.83409$.

Based on this example, we can see that the steps to finding a confidence interval for a population mean are:

1. Find the sample mean $\bar{x}$.
2. Find $z_{\alpha/2}$, the $z$-score with an area of $\alpha/2$ to its right.
3. Calculate the margin of error using the formula $E = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
4. Construct the confidence interval $(\bar{x} - E, \bar{x} + E)$.

**Note**

Note that we only found the $z$-score for the right tail in this example. Why didn't we find the $z$-score for the left tail? Because the standard normal distribution is symmetric about its mean $\mu = 0$, we don't need to perform a separate calculation to find the $z$-score for the left tail: it is just the negative of the $z$-score for the right tail. In this case, since the $z$-score for the right tail is $z = 0.83409$, we know by symmetry that the $z$-score for the left tail is $z = -0.83409$. This negative is already reflected in the formula for the confidence interval: the negative in the lower bound $\bar{x} - E$ is there because of the negative from the left-tailed $z$-score. So we only need to find the $z$-score for the right tail; the formula takes care of the rest.

---

## Example 5.1.2

The standard deviation of the weights of elephants is known to be approximately $50$ pounds. Forty-five newborn elephants are sampled and found to have the following weights, in pounds:

333, 248, 303, 248, 153, 168, 280, 256, 195, 234, 366, 250, 325, 266, 164, 253, 262, 343, 244, 425, 345, 343, 277, 215, 226, 254, 289, 296, 268, 195, 268, 202, 249, 256, 284, 257, 205, 215, 251, 257, 144, 323, 238, 257, 218

Construct a $95\%$ confidence interval for the mean weight of a newborn elephant.

First, let's note what we are given. We are told that

$$\begin{aligned} \sigma &= 50, \\ n &= 45, \\ \text{CL} &= 0.95. \end{aligned}$$

Step 1: Find the sample mean $\bar{x}$.

Recall that $\bar{x} = \frac{\sum x}{n}$. We will use R to calculate this.

```
x = c(333, 248, 303, 248, 153, 168, 280, 256, 195, 234, 366, 250, 325, 266, 164, 253, 262, 343, 244, 425, 34
n = length(x)

xbar = sum(x)/n
xbar
```

258.844444444444

So the sample mean is $\bar{x} = 258.8444$.

Step 2: Find $z_{\alpha/2}$.

Since we want a $95\%$ confidence interval, the area under the normal distribution outside our confidence interval is

$$\alpha = 1 - \text{CL} = 1 - 0.95 = 0.05$$

So we want an area of $\alpha/2 = 0.05/2 = 0.025$ in each tail of the normal distribution. We can calculate $z_{\alpha/2} = z_{0.025}$, the $z$-score with an area of $0.025$ to its right, using R.

```
qnorm(p = 1 - 0.025)
```

1.95996398454005

So $z_{\alpha/2} = z_{0.025} = 1.960$.

Step 3: Calculate the Margin of Error.

Using the formula, the margin of error is

$$E = z_{0.025}\frac{\sigma}{\sqrt{n}} = 1.960\left(\frac{50}{\sqrt{45}}\right) = 14.6090.$$

Step 4: Construct the Confidence Interval.

Our confidence interval is

Skip to main content

We are $95\%$ confident that the population mean weight of newborn elephants is between $244.2354$ pounds and $273.4534$ pounds.

---

## Example 5.1.3

A Menifee High School math teacher, Mr. DeLeon, wants to know the average GPA of students at the high school. He randomly selects $30$ students and asks what their GPA is. He obtains the following data:

3.55, 3.51, 3.27, 4.30, 3.17, 3.61, 3.24, 3.74, 3.40, 3.91, 3.00, 1.88, 2.54, 3.15, 4.35, 2.62, 4.01, 3.69, 3.82, 3.18, 2.60, 3.49, 3.05, 2.91, 3.28, 2.97, 3.09, 3.49, 3.49, 3.05

Mr. DeLeon assumes the population standard deviation is $\sigma = 0.5$. Construct a $98\%$ confidence interval for the mean GPA.

### Solution

First, note that

$$\begin{aligned} n &= 30 \\ \sigma &= 0.5 \\ \text{CL} &= 0.98 \end{aligned}$$

### Step 1: Find the Sample Mean $\bar{x}$.

```
x = c(3.55, 3.51, 3.27, 4.30, 3.17, 3.61, 3.24, 3.74, 3.40, 3.91, 3.00, 1.88, 2.54, 3.15, 4.35, 2.62, 4.01,
n = length(x)

xbar = sum(x)/n
xbar
```

3.312

The sample mean GPA is $\bar{x} = 3.312$.

### Step 2: Find $z_{\alpha/2}$.

Note that

$$\alpha = 1 - \text{CL} = 1 - 0.98 = 0.02.$$

Then we want area outside the confidence interval in each tail to be $\alpha/2 = 0.02/2 = 0.01$.

We use R to find $z_{\alpha/2} = z_{0.01}$, the $z$-score with an area of 0.01 to its right.

```
qnorm(p = 1 - 0.01)
```

2.32634787404084

Skip to main content

So $z_{0.01} = 2.3263$.

Step 3: Calculate the Margin of Error.

The margin of error is

$$E = z_{0.01}\left(\frac{\sigma}{\sqrt{n}}\right) = 2.3263\left(\frac{0.5}{\sqrt{30}}\right) = 0.2124.$$

Step 4: Construct the Confidence Interval.

The confidence interval is

$$(\bar{x} - E, \bar{x} + E) = (3.312 - 0.2124, 3.312 + 0.2124) = (3.0996, 3.5244).$$

Mr. DeLeon can be $98\%$ confident that the average GPA at Menifee High School is between $3.0996$ and $3.5244$.

# 5.2. Estimating Population Proportions

## Objectives

- Construct confidence intervals for population proportions.
- Identify and interpret the margin of error for a confidence interval.

## Introduction

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has $40\%$ of the vote with a margin of error of three percentage points (if the sample is large enough). Often, election polls are calculated with $95\%$ confidence, so the pollster in this example would be $95\%$ confident that the true proportion of voters who favored the candidate would fall in the interval $(0.40 - 0.03, 0.40 + 0.03) = (0.37, 0.43)$. That is, the pollster would be $95\%$ confident that the true proportion falls between $37\%$ and $43\%$.

Investors in the stock market might be interested in the true proportion of stocks that go up and down each week. Heart surgeons might be interested in the success rate for a certain procedure. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for proportion of heart procedures that are successful. Estimating population proportions using confidence intervals is useful in many disciplines.

## Calculating the Confidence Interval

The procedure to find the confidence interval, the sample size, the **margin of error**, and the **confidence level** for a proportion is similar to that for the population mean, but the formulas are different.

We know from the Central Limit Theorem that the distributions of sample proportions is

$$\hat{P} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right),$$

where $p$ is the population proportion and $n$ is the sample size. Of course, as in the case with means, if we already knew the value of the population proportion $p$, we wouldn't be estimating it using a confidence interval. We approximate the population proportion $p$ with the point estimate $\hat{p}$. Then to construct a confidence interval, we will use the distribution

$$N\left(\hat{p}, \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

We want to construct the confidence interval

$$(\hat{p} - E, \hat{p} + E),$$

where $E$ is the margin of error. As in the case with means, we know that the upper bound has a $z$-score associated with it. This implies that $\hat{p} + E = \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ for some value of $z$. Subtracting $\hat{p}$ from both sides gives us the formula for the margin of error:

$$E = z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

The value of $z$ is calculated exactly the same way when estimating proportions as it was when estimating means.

Based on these observations, we can see that the steps to finding a confidence interval for a population proportion are:

1. Find the sample proportion $\hat{p}$.
2. Find $z_{\alpha/2}$, the $z$-score with an area of $\alpha/2$ to its right.
3. Calculate the margin of error using the formula $E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
4. Construct the confidence interval $(\hat{p} - E, \hat{p} + E)$.

---

## Example 5.2.1

A student polls his school to see if students in the school district are for or against new legislation regarding school uniforms. She surveys $600$ students and finds that $480$ are against the new legislation. Compute a $99\%$ confidence interval for the true percent of students who are against the new legislation.

### Solution

First, note the given information:

$$\begin{aligned} n &= 600 \\ x &= 480 \\ \text{CL} &= 0.99. \end{aligned}$$

Skip to main content

Here, $x$ represents the number of students who are against the new legislation.

### Part 1: Find the sample proportion $\hat{p}$.

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{480}{600} = 0.8.$$

So $80\%$ of the students surveyed are against the new legislation.

### Part 2: Find $z_{\alpha/2}$.

Since $\text{CL} = 0.99$, the area under the normal density function not in the confidence interval is

$$\alpha = 1 - \text{CL} = 1 - 0.99 = 0.01.$$

This means that each tail should have an area of $\alpha/2 = 0.01/2 = 0.005$ outside the confidence interval. We can use R to find $z_{\alpha/2} = z_{0.005}$, the $z$-score with area $0.005$ to its right.

```
qnorm(p = 1 - 0.005)
```

2.5758293035489

So $z_{0.005} = 2.5758$.

### Step 3: Calculate the Margin of Error.

The margin of error is

$$E = z_{0.005}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 2.5758\sqrt{\frac{0.8(1-0.8)}{600}} = 0.0421.$$

### Step 4: Construct the Confidence Interval.

The confidence interval is

$$(\hat{p} - E, \hat{p} + E) = (0.8 - 0.0421, 0.8 + 0.0421) = (0.7579, 0.8421).$$

So we are $99\%$ confident that between $75.79\%$ and $84.21\%$ of all students in the district are against the new legislation.

---

## Example 5.2.2

The Berkman Center for Internet & Society at Harvard recently conducted a study analyzing the privacy management habits of

interval for the true proportion of teens who would report having more than $500$ Facebook friends.

### Solution

We are told that

$$\begin{aligned} n &= 50 \\ x &= 13 \\ \text{CL} &= 0.92 \end{aligned}$$

### Step 1: Find the sample proportion $\hat{p}$.

The sample proportion is

$$\hat{p} = \frac{x}{n} = \frac{13}{50} = 0.26.$$

So $26\%$ of the $50$ teens sampled have more than 500 friends on facebook.

### Step 2: Find $z_{\alpha/2}$.

The area under the normal density function inside the confidence interval is $\text{CL} = 0.92$. Thus, the area under the normal density function outside the confidence interval is

$$\alpha = 1 - \text{CL} = 1 - 0.92 = 0.08.$$

This means that the area outside the confidence interval in each tail of the normal density function is $\alpha/2 = 0.08/2 = 0.04$.

We use R to find $z_{\alpha/2} = z_{0.04}$, the $z$-score with an area of $0.04$ to its right.

```
qnorm(p = 1 - 0.04)
```

1.75068607125217

So $z_{0.04} = 1.7507$.

### Step 3: Calculate the Margin of Error.

The margin of error is

$$E = z_{0.04}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.7507\sqrt{\frac{0.26(1-0.26)}{50}} = 0.1086.$$

### Step 4: Construct the Confidence Interval.

$$(\hat{p} - E, \hat{p} + E) = (0.26 - 0.1086, 0.26 + 0.1086) = (0.1514, 0.3686).$$

We are $92\%$ confident that between $15.14\%$ and $36.86\%$ of teens have more than $500$ friends on Facebook.

# 5.3. Student's $t$-Distribution

## Objectives

- Compute probabilities using Student's $t$-distribution.

## Why Student's $t$-Distribution?

In practice, when constructing a confidence interval for a population mean, we rarely know the population standard deviation. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation $s$ as an estimate for $\sigma$ and proceeded as before to calculate a confidence interval with close enough results. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing the population standard deviation $\sigma$ with the sample standard deviation $s$ did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the Student's $t$-distribution. The name comes from the fact that Gosset wrote under the pen name "Student" when he published his results.

Up until the mid-1970s, some statisticians used the normal distribution approximation for large sample sizes and used Student's $t$-distribution only for sample sizes of at most $30$. With graphing calculators and computers, the practice now is to use Student's $t$-distribution whenever $s$ is used as an estimate for $\sigma$.

## Student's $t$-Distribution

If you draw a simple random sample of size $n$ from a population that has an approximately normal distribution with mean $\mu$ and unknown population standard deviation $\sigma$ and calculate the $t$-score

$$t = \frac{\bar{x} - \mu}{\left(\frac{s}{\sqrt{n}}\right)},$$

then the $t$-scores follow a Student's $t$-distribution **with $n$–$1$ degrees of freedom**. The $t$-score has the same interpretation as the $z$-score. It measures how far $\bar{x}$ is from its mean $\mu$. For each sample size $n$, there is a different Student's $t$-distribution.

The degrees of freedom, $n$–$1$, come from the calculation of the sample standard deviation $s$. Recall that if we have a sample of size $n$, we used $n$ deviations (that is, the $n$ values of $(x - \bar{x})$) to calculate $s$. Because the sum of the deviations is zero, we can find the last deviation once we know the other $n$–$1$ deviations. This means that the other $n$–$1$ deviations can vary freely, but once $n - 1$ deviations are known, there is only one number we can choose for the final deviation to get the sample standard deviation $s$. We call the number $n$–$1$ the degrees of freedom ($df$).

Skip to main content

- The graph for the Student's $t$-distribution is similar to the standard normal curve.
- The mean for the Student's $t$-distribution is zero and the distribution is symmetric about zero.
- The Student's $t$-distribution has more probability in its tails than the standard normal distribution because the spread of the $t$-distribution is greater than the spread of the standard normal. So the graph of the Student's $t$-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's $t$-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's $t$-distribution becomes more like the graph of the standard normal distribution.

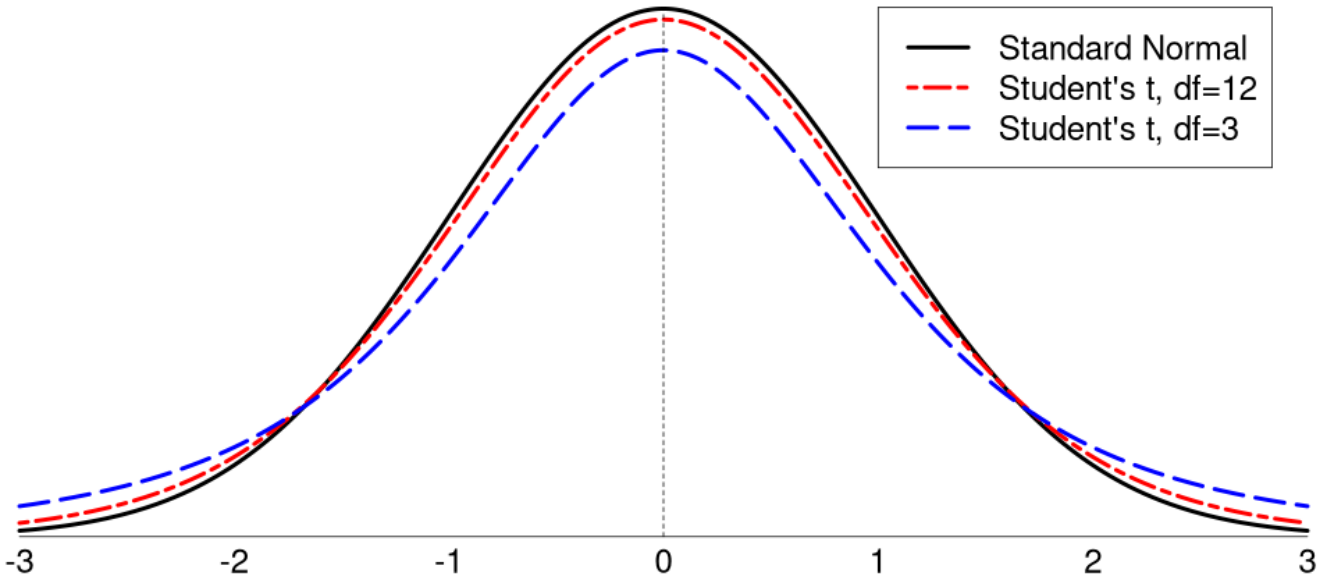



*Fig. 5.3.1* The standard normal distribution, a Student's $t$-distribution with $df = 12$, and a Student's $t$-distribution with $df = 3$.

## Student's $t$-Distribution Using R

We can use R to run calculations involving Student's $t$-distribution in almost the same way we did calculations involving the standard normal distribution. Recall that to find the area under the standard normal function to the left of a $z$-value, we use the function `pnorm(q)`. Similarly, to find the area under Student's $t$-distribution to the left of a $t$-value, we use the function

```
pt(q, df)
```

where `q` is the $t$-value, and `df` is the degrees of freedom.

---

### Example 5.3.1

Consider a $t$-distribution with $11$ degrees of freedom.

1. Find $P(T < -0.5)$.
2. Find $P(T > 0.3)$.

Skip to main content

Solution

Part 1

```
pt(q = -0.5, df = 11)
```

0.313463109591923

So $P(T < -0.5) = 0.3135$.

Part 2

Just like with the `pnorm` function, the `pt` function only gives the area or probability to the *left* of a value. But in this case, we want to find $P(t > 0.3)$, which is the probability to the *right* of $t = 0.3$. Since the total probability is always equal to $1 = 100\%$, we can find this using the formula

$$P(T > 0.3) = 1 - P(T < 0.3),$$

just like we do with the standard normal distribution. We use R to perform the calculation.

```
1 - pt(q = 0.3, df = 11)
```

0.384885178944686

So $P(T > 0.3) = 0.3849$.

Part 3

As we do with the normal distribution, to find $P(-1.0 < t < 1.2)$, we will first find all the area to the left of the larger value $t = 1.2$, then subtract the excess area to the left of the smaller value $t = -1.0$.

```
pt(q = 1.2, df = 11) - pt(q = -1.0, df = 11)
```

0.702926510623682

So $P(-1.0 < T < 1.2) = 0.7029$.

---

Given a value, we can find the corresponding probability using `pnorm` (for a standard normal distribution) or `pt` (for a $t$-distribution). We can also do the reverse: given a probability, we can find the value. For the standard normal distribution, we've seen that the function for doing this is `qnorm(p)`. Similarly, the function for finding a value given a probability for a $t$-distribution is

```
qt(p, df)
```

where `p` is the probability to the left of the $t$-value, and `df` is the degrees of freedom.

---

Consider a $t$-distribution with $22$ degrees of freedom.

1. Find $t$ so that the area to the left of $t$ is 0.12.
2. Find $t$ so that the area to the right of $t$ is 0.34.
3. Find $t_{0.05}$.

Solution

Part 1

```
qt(p = 0.12, df = 22)
```

-1.20766344702836

So the area to the left of $t = -1.0277$ is $0.12$.

Part 2

Just like with the `qnorm` function, the `qt` function expects a probability to the *left* of the $t$-value we want. But for this problem, we are asked to find a $t$-value so that the area or probability to the right of $t$ is $0.34$. To find the $t$-value, we first need to find the probability to the left of $t$. Since the total probability is equal to $1 = 100\%$, if the probability to the right of $t$ is $0.34$, that means the probability to the left of $t$ is $1 - 0.34$. We use R to run the calculation.

```
qt(p = 1 - 0.34, df = 22)
```

0.417998499316742

So the area to the right of $t = 0.4180$ is $0.34$.

Part 3

Just as we saw with $z$-scores, the notation $t_{0.05}$ is the $t$-value with area $0.05$ to its right.

```
qt(p = 1 - 0.05, df = 22)
```

1.71714437438024

So $t_{0.05} = 1.7171$.

# 5.4. Estimating Population Means ($\sigma$ Unknown)

## Objectives

- Construct confidence intervals for population means in populations where the standard deviation is not known.
- Identify and interpret the margin of error for a confidence interval

Skip to main content

# Confidence Intervals using Student's $t$-Distribution

As mentioned before, we rarely actually know the population standard deviation $\sigma$ when constructing a confidence interval for the population mean $\mu$. To approximate the population standard deviation $\sigma$ with the sample standard deviation $s$, we must use a $t$-distribution with $n-1$ degrees of freedom (where $n$ is the sample size) instead of a normal distribution.

With the exception of this change, the process of constructing a confidence interval for the population mean is largely the same:

1. Find the sample mean $\bar{x}$ and the sample standard deviation $s$.
2. Find $t_{\alpha/2}$, the $t$-score with area $\alpha/2$ to its right and $n-1$ degrees of freedom.
3. Calculate the margin of error using the formula $E = t_{\alpha/2}\dfrac{s}{\sqrt{n}}$.
4. Construct the confidence interval $(\bar{x}-E, \bar{x}+E)$.

---

## Example 5.4.1

Forty-five newborn elephants are sampled and found to have the following weights, in pounds:

333, 248, 303, 248, 153, 168, 280, 256, 195, 234, 366, 250, 325, 266, 164, 253, 262, 343, 244, 425, 345, 343, 277, 215, 226, 254, 289, 296, 268, 195, 268, 202, 249, 256, 284, 257, 205, 215, 251, 257, 144, 323, 238, 257, 218

Construct a $95\%$ confidence interval for the mean weight of a newborn elephant.

### Solution

Note that we are *not* told what the population standard deviation $\sigma$ is. That means we will need to approximate it using the sample standard deviation $s$, and we'll need to use a $t$-distribution.

We are given that

$$n = 25$$
$$CL = 0.95$$

### Step 1: Find the sample mean $\bar{x}$ and the sample standard deviation $s$.

```
x = c(333, 248, 303, 248, 153, 168, 280, 256, 195, 234, 366, 250, 325, 266, 164, 253, 262, 343, 244, 425, 34
n = length(x)

xbar = sum(x)/n
xbar
```

258.844444444444

Then the sample mean is $\bar{x} = 258.84444$.

To find the sample standard deviation, recall that we use the formula

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}.$$

Let's translate this formula to R.

```
s = sqrt(sum( (x - xbar)^2 )/(n-1))
s
```

57.3842532865291

The sample standard deviation is $s = 57.38425$.

Step 2: Find $t_{\alpha/2}$.

First, note the degrees of freedom for the $t$-distribution is

$$df = n - 1 = 45 - 1 = 44.$$

Next, since $\text{CL} = 0.95$, the area outside of the confidence interval is

$$\alpha = 1 - \text{CL} = 1 - 0.95 = 0.05.$$

So $\alpha/2 = 0.05/2 = 0.025$. We want to find $t_{\alpha/2} = t_{0.025}$, the $t$-score with an area of $0.025$ to its right.

```
qt(p = 1 - 0.025, df = 44)*sd(x)/sqrt(45)
```

17.2401382726373

So $t_{0.025} = 2.01537$.

Step 3: Calculate the Margin of Error.

The margin of error is

$$E = t_{0.025} \frac{s}{\sqrt{n}} = 2.01537 \left( \frac{57.38425}{\sqrt{45}} \right) = 17.24014.$$

Step 4: Construct the Confidence Interval.

The confidence interval is

$$(\bar{x} - E, \bar{x} + E) = (258.84444 - 17.24014, 258.84444 + 17.24014) = (241.60431, 276.08458).$$

We are $95\%$ confident that the average weight of a newborn elephant is between $241.60431$ pounds and $276.08458$ pounds.

---

Skip to main content

A Menifee High School math teacher, Mr. DeLeon, wants to know the average GPA of students at the high school. He randomly asks $30$ students what their GPA is, and obtains the following data:

3.55, 3.51, 3.27, 4.30, 3.17, 3.61, 3.24, 3.74, 3.40, 3.91, 3.00, 1.88, 2.54, 3.15, 4.35, 2.62, 4.01, 3.69, 3.82, 3.18, 2.60, 3.49, 3.05, 2.91, 3.28, 2.97, 3.09, 3.49, 3.49, 3.05

Construct a $98\%$ confidence interval for the mean GPA.

### Solution

We are not told the population standard deviation $\sigma$, so we will need to approximate it using the sample standard deviation $s$ and use a $t$-distribution to find the margin of error.

We are told that

$$\begin{aligned} n &= 30 \\ CL &= 0.98 \end{aligned}$$

### Step 1: Find the Sample Mean $\bar{x}$ and the Sample Standard Deviation $s$.

```
x = c(3.55, 3.51, 3.27, 4.30, 3.17, 3.61, 3.24, 3.74, 3.40, 3.91, 3.00, 1.88, 2.54, 3.15, 4.35, 2.62, 4.01,
n = length(x)

xbar = sum(x)/n
xbar
```

3.312

So the sample mean is $\bar{x} = 3.312$.

```
s = sqrt(sum( (x - xbar)^2 )/(n-1))
s
```

0.526428434406543

The sample standard deviation is $s = 0.5264$.

### Step 2: Find $t_{\alpha/2}$.

First, note that the degrees of freedom for our $t$-distribution is

$$df = n - 1 = 30 - 1 = 29.$$

Next, since the area inside the confidence interval is $CL = 0.98$, the area outside the confidence interval is

$$\alpha = 1 - CL = 1 - 0.98 = 0.02.$$

So the area remaining in each tail of the $t$-distribution is $\alpha/2 = 0.02/2 = 0.01$. We want to find $t_{\alpha/2} = t_{0.01}$, the $t$-value with a

```
qt(p = 1 - 0.01, df = 29)
```

2.46202136015041

Then $t_{0.01} = 2.4620$.

Step 3: Calculate the Margin of Error.

The margin of error is

$$E = t_{0.01}\frac{s}{\sqrt{n}} = 2.4620\left(\frac{0.5264}{\sqrt{30}}\right) = 0.2366.$$

Step 4: Construct the Confidence Interval.

The confidence interval is

$$(\bar{x} - E, \bar{x} + E) = (3.312 - 0.2366, 3.312 + 0.2366) = (3.0754, 3.5486).$$

We are $98\%$ confident that the average GPA of students at Menifee High School is between $3.0754$ and $3.5486$.

# 6. An Introduction to Hypothesis Testing

## 6.1. An Introduction to Hypothesis Testing

### Objectives

- Understand the fundamentals of hypothesis testing.

### Introduction to Hypothesis Testing

Sometimes we want to test whether a claim made about a population parameter is true or not. For instance, a car dealer advertises that its new small truck averages $35$ miles per gallon. A tutoring service claims that $90\%$ of its students get an A or a B. A company says that women managers in their company earn an average of $\$53{,}000$ per year.

It's often impractical to survey every member of a population to confirm a claim or **hypothesis**. Instead, we can sample a few members of the population and calculate a point estimate of the parameter we are interested in. If the hypothesis is true, we would expect the point estimate to be close to the hypothesized value of the parameter. So if the point estimate is actually far away from the hypothesized value of the parameter, we might conclude that the actual value of the parameter is different from the hypothesis.

For example, suppose Julian claims his average time for the $100$-meter dash is $11.72$ seconds. We can test the hypothesis that this parameter, the population mean, is $\mu = 11.72$ seconds. Imagine we sample several of his runs and calculate a sample mean of $\bar{x} = 11.77$ seconds. Since the point estimate $\bar{x} = 11.77$ seconds is so close to Julian's claimed mean time of $\mu = 11.72$

Skip to main content

seconds, we would probably feel that Julian's claim is about right. But if, instead, we obtained a sample mean of $\bar{x} = 13.45$ seconds, this sample mean would be so far from Julian's claim of $\mu = 11.72$ seconds that we might seriously doubt his claim.

But how do we tell exactly when a point estimate is far enough from a hypothesized parameter that we should reject that hypothesis? To do this, we first calculate the sampling distribution of the point estimate *assuming the hypothesis is true*. Using the sampling distribution, we determine the theoretical probability of randomly selecting a sample with a point estimate at least as far from the parameter as the point estimate obtained from the actual sample. If it is exceptionally unlikely that we would have randomly selected a sample with a point estimate so far from the hypothesized parameter, we can conclude with some confidence that the hypothesis is probably not true.

In summary, the fundamental steps to a **hypothesis test** are as follows:

0. **Gather a random sample.**
   Consider this "step zero": before we conduct a hypothesis test, we need a point estimate from sample data to compare to the hypothesized parameter. In the real world, statisticians must carefully gather a random sample from the population to be able to perform a hypothesis test. In this text, we will generally be given the sample data in the problem statement, so we won't need to gather a sample ourselves.
1. **State the hypothesis.**
   Identify a claim about a population parameter that we want to test, and state the hypothesis mathematically.
2. **Assuming the hypothesis is true, identify the sampling distribution.**
   In this chapter, we will focus on hypotheses about the population mean or population proportion. We use the central limit theorem to identify the sampling distributions of the point estimates of these parameters.
3. **Find the probability.**
   Using the sampling distribution, we calculate the probability that the point estimate from a random sample is at least as far from the hypothesized parameter as the point estimate from the actual sample is.
4. **Draw a conclusion.**
   If the probability of obtaining our point estimate is very small, we might conclude that our assumption, that the hypothesis is true, is actually not correct. On the other hand, if the probability is not small, we would not have a good argument to reject the hypothesis.

We will formalize these ideas in the coming chapter and return to them throughout the rest of the course. For now, let's go over a few examples focusing on the big idea and the fundamental components.

---

## Example 6.1.1

A company says that women managers in their company earn an average of $\$53{,}000$ per year with a standard deviation of $\$5{,}000$. Janice wants to test the claim. She surveys $25$ women managers and obtains the following data on income per year:

48230, 53491, 52926, 54832, 48080, 47101, 50126, 47433, 57753, 55107, 57205, 46297, 61249, 66936, 5463, 44429, 52696, 51243, 49008, 51522, 46658, 51788, 60244, 46164, 51103

Assume the population is normally distributed. What do you think Janice should conclude about the claim?

### Solution

### Step 1: State the Hypothesis.

The claim or hypothesis is that the average annual income of women managers at the company is $\$53,000$. Written mathematically, the hypothesis is that

$$\mu = 53{,}000.$$

### Step 2: Assuming the hypothesis is true, identify the sampling distribution.

We are testing the mean $\mu$, and we are told what the population standard deviation $\sigma$ is. Then by the central limit theorem, sample means follow a normal distribution with mean

$$\mu = 53{,}000$$

and standard error

$$\frac{\sigma}{\sqrt{n}} = \frac{5000}{\sqrt{25}} = 1000.$$

That is, $\bar{X} \sim N(53000, 1000)$.

### Step 3: Find the probability.

First, we calculate the sample mean $\bar{x}$, which is the point estimate of the population mean $\mu$.

```
x = c(48230, 53491, 52926, 54832, 48080, 47101, 50126, 47433, 57753, 55107, 57205, 46297, 61249, 66936, 5463
n = length(x)

xbar = sum(x)/n
xbar
```

50283.36

The sample mean is $\bar{x} = 50283.36$.

The $z$-score associated with the sample mean $\bar{x}$ is

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{50283.36 - 53000}{1000} = -2.7166.$$

To test the claim, we want to find the probability that a randomly selected sample will have a sample mean at least as far away from the hypothesized population mean as the sample mean we actually got. This means we want to find $P(\bar{X} \leq 50283.36) = P(Z \leq -2.7166)$\$. Let's use R to find this.

```
pnorm(q = -2.7166)
```

0.00329781334352045

So $P(\bar{X} \leq 50283.36) = P(Z \leq -2.7166) = 0.00330$. That is, if the claim were true, there would only be a $0.33\%$ chance

Skip to main content

Step 4: Draw a conclusion.

In this case, it is very unlikely that we would obtain the point estimate we did if the claim were true. In fact, it is so unlikely (less than a 1% chance) that we feel confident in concluding that the claim is not true.

We conclude that the average salary of women managers at the company is not $\$53,000$.

---

## Example 6.1.2

A tutoring service claims that $90\%$ of its students get an A or a B. Mike wants to test the claim. He surveys $45$ students who utilized the tutoring service and found that $37$ students got an A or a B. Do you think Mike should reject the claim?

Solution

Step 1: State the Hypothesis.

The claim we are testing is that the proportion of students of the tutoring service who get an A or a B is $90\%$. Mathematically, the hypothesis is

$$p = 0.90.$$

Step 2: Assuming the hypothesis is true, identify the sampling distribution.

Assuming the hypothesis is true, we know by the central limit theorem that the sampling distribution of sample proportions is normally distributed with mean

$$p = 0.90$$

and standard error

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.90(1-0.90)}{45}} = 0.0447.$$

So $\hat{P} \sim N(0.90, 0.0447)$.

Step 3: Find the probability.

First note that the sample proportion, the point estimate of the population proportion, is

$$\hat{p} = \frac{x}{n} = \frac{37}{45} = 0.8222.$$

The $z$-score associated with the sample proportion $\hat{p}$ is

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.8222 - 0.90}{0.0447} = -1.7405.$$

We want to find the probability We want to find $P(\hat{p} \leq 0.8222) = P(z \leq -1.7405)$. We calculate this using R.

To test the claim, we want to find the probability that a randomly selected sample will have a sample proportion at least as far away from the hypothesized population proportion as the sample proportion we actually got. This means we want to find $P(\hat{P} \leq 0.8222) = P(Z \leq -1.7405)$$. Let's use R to find this.

```
pnorm(q = -1.7405)
```

0.0408856300354386

Then $P(\hat{P} \leq 0.8222) = P(Z \leq -1.7405) = 0.0409$. So if the population proportion is actually $p = 0.90$, there is a $4.09\%$ chance that a random sample gives a sample proportion of $\hat{p} = 0.8222$ or less.

### Step 4: Draw a conclusion.

There is about a $4\%$ chance of obtaining a sample proportion of $\hat{p} \leq 0.8222$ if the claim were true. This is somewhat unlikely, but it wouldn't be outlandish.

Ultimately, whether we reject a hypothesis or not is a choice. Would you reject the hypothesis if there is only a $1\%$ chance that we get a point estimate as extreme as the point estimate we actually get from our sample? What if it was a $4\%$ chance? An $8\%$ chance? The choice ultimately must be made by the statistician where to draw the line. In a later section, we will cover some guidelines on making this choice.

So what do you think? In this particular case, do you think the evidence is sufficient to reject the claim that $90\%$ of students of the tutoring service get an A or a B? Or do you think the evidence isn't strong enough to make that conclusion?

---

## Example 6.1.3

A statistics instructor claims that the class average on an exam is $65$ points. Richard thinks the class average is higher. He collects the following scores from 9 students in the class:

84, 67, 60, 55, 51, 64, 66, 70, 83

Richard assumes the scores in the class are normally distributed. Should Richard reject his instructor's claim?

### Solution

### Step 1: State the Hypothesis.

The hypothesis is that the class average is $65$ points. Mathematically, we write

$$\mu = 65.$$

Skip to main content

Step 2: Assuming the hypothesis is true, identify the sampling distribution.

We want to test the population mean $\mu$, but we are *not* given the population standard deviation $\sigma$. We will need to use the sample standard deviation $s$ and a $t$-distribution with $df = n - 1 = 9 - 1 = 8$ degrees of freedom.

To find the $t$-score, we need the mean of the sampling distribution,

$$\mu = 65.$$

The standard error is given by the formula $\frac{s}{\sqrt{n}}$, so we will need to find the sample standard deviation $s$ first.

To find the sample standard deviation $s$, we first find the sample mean $\bar{x}$.

```
x = c(84, 67, 60, 55, 51, 64, 66, 70, 83)
n = length(x)

xbar = sum(x)/n
xbar
```

66.6666666666667

The sample mean is $\bar{x} = 66.6667$.

Next, we find the sample standard deviation.

```
s = sqrt(sum( (x - xbar)^2 )/(n - 1))
s
```

11.247221879202

So the sample standard deviation is $s = 11.2472$. We can now calculate the standard deviation of the distribution of sample means:

$$\frac{s}{\sqrt{n}} = \frac{11.2472}{\sqrt{9}} = 3.7491.$$

Step 3: Find the probability.

We calculated the point estimate in step 2: the sample mean is $\bar{x} = 66.6667$. The $t$-score associated with $\bar{x} = 66.6667$ is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{66.6667 - 65}{3.7491} = 0.4446.$$

To test the claim, we want to find the probability that a randomly selected sample will have a sample mean at least as far away from the hypothesized population mean as the sample mean we actually got. This means we want to find $P(\bar{X} \geq 66.6667) = P(T \geq -0.4446)$. We calculate this using R.

```
1 - pt(q = 0.4446, df = 8)
```

Skip to main content

So $P(\bar{X} \geq 66.6667) = P(T \geq 0.4446) = 0.3342$. That is, assuming the claim is true, there is a $33.42\%$ chance that our sample would have a sample mean of $66.6667$ or greater.

Step 4: Draw a conclusion.

If the claim is true, there is a $33.42\%$ chance that a random sample of the population has a sample mean as extreme as $\bar{x} = 66.6667$. This isn't a small probability; there is a fairly large chance that we obtain a sample with the sample mean we do given that the claim is true. We usually don't reject a claim unless we obtain at a probability of at least less than $10\%$. So we do not have enough evidence to conclude that the mean exam score is greater than $65$ points.

# 6.2. Hypotheses and Errors

## Objectives

- Identify the null and alternative hypotheses for a hypothesis test.
- Identify the possible type I and type II errors in a hypothesis test.

## The Null and Alternative Hypothesis

In the previous section we considered the general strategy of how we can statistically test a hypothesis. In this section, we begin to formalize the ideas needed for hypothesis testing.

A hypothesis test begins by considering two **hypotheses**. They are called the **null hypothesis** and the **alternative hypothesis**. These hypotheses represent opposing viewpoints.

$H_a$: **The alternative hypothesis**. This is a statement about a population parameter such as the mean $\mu$ or proportion $p$. When conducting a hypothesis test, we generally seek to gather evidence in favor of this alternative hypothesis.

$H_0$: **The null hypothesis**. This hypothesis is the mathematical opposite of the alternative hypothesis. When conducting a hypothesis test, we assume the null hypothesis is true.

The null and alternative hypotheses are complementary mathematical statements. This means that either the null hypothesis or the alternative hypothesis is true, but not both. One of the hypotheses must be true. When conducting a hypothesis test, we assume the null hypothesis is true, then examine the evidence from sample data to decide if we have enough evidence to reject the null hypothesis and accept the alternative hypothesis or not.

After we have determined which hypothesis the sample data supports, we make a decision. There are two options for a decision:

- The evidence is sufficient to reject $H_0$ and accept $H_a$.
- The evidence is not sufficient reject $H_0$ and accept $H_a$.

Note that we never definitively conclude to accept the null hypothesis and reject the alternative hypothesis. The purpose of the hypothesis test is only to determine if we have sufficient evidence to accept the alternative hypothesis.

*Table 6.2.1* The mathematical symbols used to state null and alternative hypotheses. The symbols in each row are mathematical opposites.

| **Null Hypothesis** $(H_0)$ | **Alternative Hypothesis** $(H_a)$ |
|---|---|
| equal $(=)$ | not equal $(\neq)$ |
| greater than or equal to $(\geq)$ | less than $(<)$ |
| less than or equal to $(\leq)$ | greater than $(>)$ |

The symbols we use to state the null hypothesis are different than the symbols we use to state the alternative hypothesis. Table 6.2.1 shows the symbols we can use when stating each hypothesis. Also, note that the two hypotheses are complementary and will use symbols that are mathematical opposites so that exactly one of the hypotheses must be true. For example, if null hypothesis of a hypothesis test is $\mu \geq 7$, then the alternative hypothesis is $\mu < 7$ because the less-than symbol is the mathematical opposite of the greater-than-or-equal-to symbol.

**Note**

The null hypothesis $H_0$ always has a symbol with an equal in it. The alternative hypothesis $H_a$ never has a symbol with an equal in it. The choice of symbol depends on the wording of the hypothesis test.

Also, while we will only state null and alternative hypotheses using mathematically opposite symbols in this textbook, be aware that some researchers sometimes use $=$ when stating a null hypothesis and $>$ or $<$ when stating the corresponding alternative hypothesis even though the symbols are not mathematical opposites. This practice does not ultimately make any real difference to the hypothesis test.

---

## Example 6.2.1

We want to test whether the mean GPA of students in American colleges is different from $2.0$ (out of $4.0$). State the null and alternative hypotheses.

### Solution

The statement, "the mean GPA of students in American colleges is different from $2.0$" can be written mathematically as $\mu \neq 2.0$. Because we are using a not-equal-to symbol, we can see from Table 6.2.1 that this statement must be our alternative hypothesis $H_a$. Since the null hypothesis $H_0$ and the alternative hypothesis $H_a$ are mathematical opposites, the null hypothesis is $\mu = 2.0$. In summary

$$H_0 : \mu = 2.0$$
$$H_a : \mu \neq 2.0$$

---

## Example 6.2.2

Skip to main content

We want to test if, on average, college students take less than five years to graduate from college. Find the null and alternative hypotheses.

### Solution

The statement, “college students, on average, take less than five years to graduate from college”, can be written mathematically as $\mu < 5$. Since the less-than symbol does not have an equal in it, this statement is the alternative hypothesis $H_a$. Then the null hypothesis $H_0$ is $\mu \geq 5$, which is the opposite of the alternative hypothesis $H_a$. So

$$H_0 : \mu \geq 5$$
$$H_a : \mu < 5$$

---

### Example 6.2.3

In an issue of *U.S. News and World Report*, an article on school standards stated that about half of all students in France, Germany, and Israel take advanced placement exams, and a third pass. The same article stated that $6.6\%$ of U.S. students take advanced placement exams, and $4.4\%$ pass. We want to test if the percentage of U.S. students who take advanced placement exams is more than $6.6\%$. State the null and alternative hypotheses of this test.

### Solution

The statement, “the percentage of U.S. students who take advanced placement exams is more than $6.6\%$” can be written mathematically as $p > 0.066$. Since the greater-than symbol does not have an equal in it, this is the alternative hypothesis. The null hypothesis $H_0$ is $p \leq 0.066$, which is the mathematical opposite of the alternative hypothesis $H_a$. That is,

$$H_0 : p \leq 0.066$$
$$H_a : p > 0.066$$

---

## Type I and Type II Errors

When you perform a hypothesis test, there are four possible outcomes depending on the actual truth of the hypotheses and what conclusion we draw from the sample data. The possible outcomes are summarized in Table 6.2.2.

*Table 6.2.2* The different possible outcomes of a hypothesis test. The outcome depends on which hypothesis is actually true and what we conclude based on the sample data.

| | **$H_0$ is Actually True and $H_a$ is Actually False** | **$H_0$ is Actually False and $H_a$ is Actually True** |
|---|---|---|
| **Do Not Reject $H_0$ and Do Not Accept $H_a$** | Correct Outcome | Type II Error |
| **Reject $H_0$ and Accept $H_a$** | Type I Error | Correct Outcome |

The four possible outcomes in Table 6.2.2 are:

1. $H_0$ *is true and* $H_a$ *is false*, and we decide *not to reject* $H_0$ *and not to accept* $H_a$. (Correct outcome.)
2. $H_0$ *is true and* $H_a$ *is false*, but we decide to *reject* $H_0$ *and accept* $H_a$. (Incorrect Outcome known as a **type I error** or a **false positive**.)
3. $H_0$ *is false and* $H_a$ *is true*, but we decide *not to reject* $H_0$ *and not to accept* $H_a$. (Incorrect outcome known as a **type II error** or a **false negative**.)
4. $H_0$ *is false and* $H_a$ *is true*, and we decide to *reject* $H_0$ *and accept* $H_a$. (Correct outcome.)

Statisticians are trained to minimize the probability that a type I or type II error would occur in a hypothesis test. We will focus on simply identifying the type I and type II errors in different situations.

**Important**

Note that type I or type II errors are not the result of a mathematical mistake. Rather, these errors happen because of the probabilistic uncertainty inherent in random sampling.

For example, suppose a new education law was passed in California, and we want to conduct a hypothesis test to see if the new law has improved student outcomes. We randomly sample a number of students, but because the sampling is random, there is a small chance that our sample contains an unusually high proportion of high performing students. Based on this skewed sample data, we would conclude that student outcomes had improved after implementation of the new law even if, in reality, student outcomes did not actually improve. (This would be a type I error or false positive.) In this hypothetical example, we did all the math right. This error is not the result of a mathematical mistake, but because, just by chance, we collected a sample with an unusually high sample mean.

So there is a small chance that we can come to the wrong conclusion when performing a hypothesis test, resulting in a type I or type II error, even if we do all the math right.

---

## Example 6.2.4

Truc tests blood cultures to see if there are traces of pathogen X. The null hypothesis is that the blood cultures contain no traces of pathogen X. The alternative hypothesis is that the blood cultures do contain traces of pathogen X. State the type I and type II errors

Solution

- Type I error: Truc concludes that the blood cultures *do* contain traces of pathogen X, but the blood cultures actually *do not* contain traces of pathogen X.
- Type II error: Truc concludes that the blood cultures *do not* contain traces of pathogen X, but the blood cultures actually *do* contain traces of pathogen X.

In this situation, which do you think would be the more serious error: a Type I error or a Type II error?

---

### Example 6.2.5

A certain experimental drug claims a cure rate of at least $75\%$ for males with prostate cancer. State the type I and type II errors in context. Which error is the more serious?

Solution

Let's first identify our hypotheses. The statement "the drug claims a cure rate of at least $75\%$" can be written mathematically as $p \geq 0.75$. Since the greater-than-or-equal-to symbol has an "equal" in it, this inequality represents the null hypothesis. This means that the alternative hypothesis is the mathematical opposite: $p < 0.75$.

$$H_0 : p \geq 0.75$$
$$H_a : p < 0.75$$

Then:

- Type I error: We conclude that the cure rate for the drug is less than $75\%$, but the cure rate for the drug is actually at least $75\%$.
- Type II error: We conclude that the cure rate for the drug is at least $75\%$, but the cure rate for the drug is actually less than $75\%$.

In this case, the Type II error would be the more serious error. If a patient believes the drug works at least 75% of the time, this most likely will influence the patient's (and doctor's) choice about whether to use the drug as a treatment option.

# 6.3. Drawing a Conclusion

## Objectives

- Understand the meaning of the $p$-value in a hypothesis test.
- Compare the $p$-value of a hypothesis test with the predetermined level of significance $\alpha$ to draw a conclusion.

## Drawing a Conclusion

After identifying the hypotheses for a hypothesis test, we calculate the point estimate from the sample data. Next, we calculate the probability that, if the null hypothesis is true, the point estimate from a randomly selected sample would be at least as extreme as the point estimate we obtained from the given sample. This probability is called the **$p$-value** of the hypothesis test.

A small $p$-value means that, if the null hypothesis is true, there is only a small probability of getting the point estimate we do. The smaller the $p$-value, the more unlikely the outcome we get. If this probability is unlikely enough, we may conclude that it is more reasonably to reject the null hypothesis and accept the alternative hypothesis instead.

If the $p$-value is not small, we may conclude that we do not have enough evidence to reject the null hypothesis and accept the alternative hypothesis.

How do we know if the $p$-value is small enough to reject the null hypothesis? A systematic way to make a decision of whether or not to reject the null hypothesis is to compare the $p$-value to a predetermined **level of significance**, denoted by $\alpha$ (the Greek letter "alpha"). The level of significance $\alpha$ is probability that the hypothesis test results in a type I error. In this textbook, the value of $\alpha$ will always be given at the beginning of each problem. Real statisticians must themselves choose a value for $\alpha$ appropriate to each particular problem. A choice of $\alpha = 5\% = 0.05$ is a common choice for real-life hypothesis tests.

To determine whether or not to reject the null hypothesis $H_0$, do the following:

- If the $p$-value $< \alpha$ then reject $H_0$. The results from the sample data are significant. There is sufficient evidence to reject the null hypothesis $H_0$ and conclude that the alternative hypothesis $H_a$ is correct.
- If the $p$-value $\geq \alpha$ then do not reject $H_0$. The results from the sample data are not significant. There is insufficient evidence to reject the null hypothesis $H_0$ and conclude that the alternative hypothesis $H_a$ is correct.

*Table 6.3.1* Possible results of a hypothesis test and the corresponding conclusions that should be drawn from the results.

| Result | Conclusion |
|---|---|
| $p$-value $< \alpha$ | Reject $H_0$ and Accept $H_a$ |
| $p$-value $\geq \alpha$ | Do Not Reject $H_0$ and Do Not Accept $H_a$ |

## Example 6.3.1

Boy Genetics Labs claim that their procedures improve the chances of a pregnancy resulting in a boy.

1. What are the null and alternative hypotheses?
2. Sample data is gathered, and a hypothesis test is conducted at a $1\%$ level of significance. If the $p$-value is $0.025$, what conclusion should be made?

### Solution

#### Part 1

A normal pregnancy has a $50\%$ chance in being a boy and a $50\%$ chance of being a girl. So the claim that the procedures improve

hypothesis. Then the null hypothesis is $p \leq 0.50$.

$$H_0 : p \leq 0.50$$
$$H_a : p > 0.50$$

Part 2

The level of significance is $\alpha = 0.01$. So we are only comfortable with rejecting the null hypothesis if the $p$-value is smaller than $\alpha = 0.01$. Since

$$p\text{-value} = 0.025 \geq 0.01 = \alpha,$$

we do not have sufficient evidence to reject the null hypothesis.

We cannot conclude from this sample data that that Boy Genetics Labs procedures improve the chances of a pregnancy resulting in a boy.

---

## Example 6.3.2

Your friend claims that his mean golf score is $63$. You want to test to see if his mean score is actually higher.

1. What are the null and alternative hypotheses?
2. A hypothesis test is conducted at the $5\%$ level of significance. If the $p$-value is $0.045$, what conclusion should be made?

### Solution

Part 1

You want to test if your friend's mean golf score is higher than $63$. Written mathematically, you want to test if $\mu > 63$. The symbol used in this inequality has no "equal" in it, so the inequality represents the alternative hypothesis of the hypothesis test. The null hypothesis is the mathematical opposite: $\mu \leq 63$.

$$H_0 : \mu \leq 63$$
$$H_a : \mu > 63$$

Part 2

Since

$$p\text{-value} = 0.045 < 0.05 = \alpha,$$

the evidence is strong enough to reject the null hypothesis.

We conclude that your friend's mean golf score is higher than $\mu = 63$.

Skip to main content

# 6.4. Hypothesis Tests of One Population

## Objectives

- Perform full hypothesis tests on the means and the proportions of populations.
- Identify the null and alternative hypotheses of a hypothesis test.
- Assuming the null hypothesis is true, use the central limit theorem to identify the sampling distribution that should be used to calculate the $p$-value.
- Calculate the $p$-value.
- Draw a conclusion based on the $p$-value.

## Calculating the $p$-value

When calculating the $p$-value in a hypothesis test, we assume the null hypothesis $H_0$ is true, then we calculate the probability that the point estimate of a random sample supports the alternative hypothesis $H_a$ at least as strongly as the point estimate we actually obtained. This means the alternative hypothesis determines whether we perform:

- a **left-tailed test**, where the $p$-value is the probability to the left of the point estimate;
- a **right-tailed test**, where the $p$-value is the probability to the right of the point estimate; or
- a **two-tailed test**, where the $p$-value is the sum of equal probabilities in both tails of the sampling distribution.

The mathematical symbol used in the statement of the alternative hypothesis indicates which test we should use to calculate the $p$-value as illustrated in Table 6.4.1.

*Table 6.4.1* The inequality symbol in the alternative hypothesis tells us which test we should perform.

| Symbol in $H_a$ | Test |
|---|---|
| $<$ | Left-Tailed Test |
| $>$ | Right-Tailed Test |
| $\neq$ | Two-Tailed Test |

The following examples illustrate when to use a left-, right-, or two-tailed test.

---

### Example 6.4.1

A hypothesis test has the hypotheses given below. Should a left-tailed test, a right-tailed test, or a two-tailed test be performed to find the $p$-value?

$$H_0: \ \mu \geq 5$$
$$H_a: \ \mu < 5$$

Skip to main content

Solution

Because the alternative hypothesis uses a “less-than” symbol, we would perform a left-tailed test to find the $p$-value.

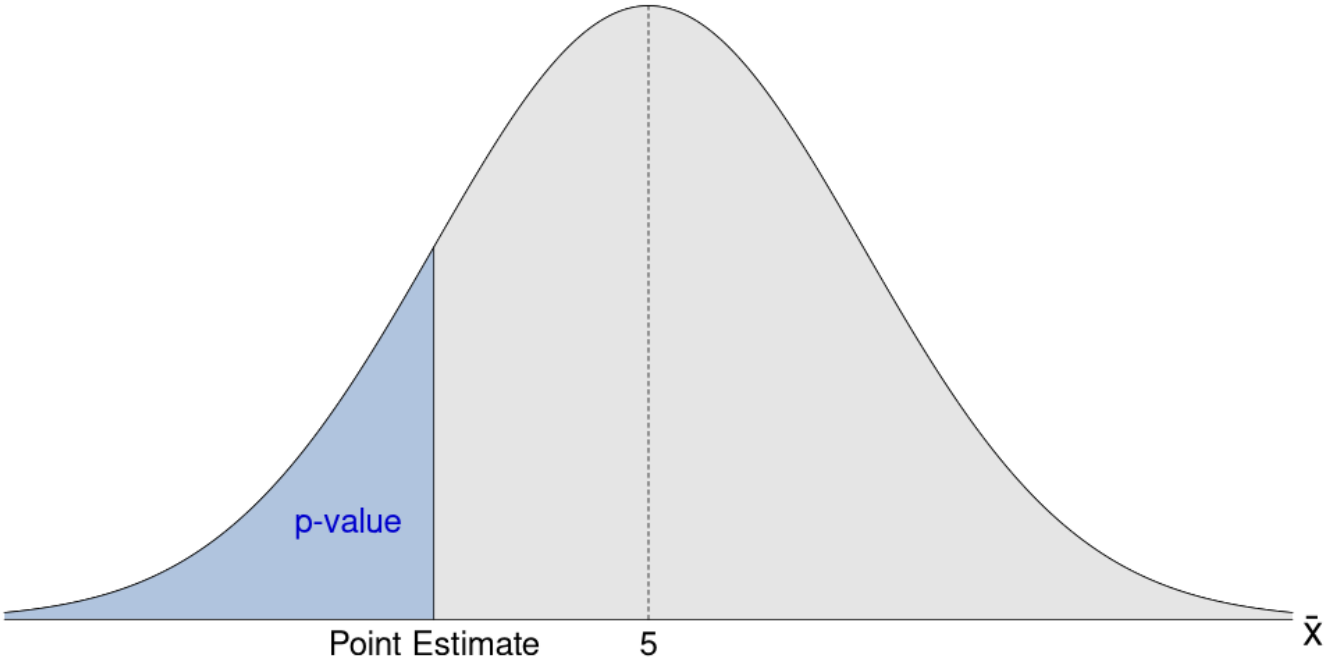



*Fig. 6.4.1* The sampling distribution of the population in this example is shown illustrates a left-tailed test. In a left-tailed test, the $p$-value is the probability to the left of the point estimate.

---

## Example 6.4.2

A hypothesis test has the hypotheses given below. Should a left-tailed test, a right-tailed test, or a two-tailed test be performed to find the $p$-value?

$$H_0 : p \leq 0.2$$
$$H_a : p > 0.2$$

Solution

Because the alternative hypothesis uses the “greater-than” symbol, we would use a right-tailed test to find the $p$-value.

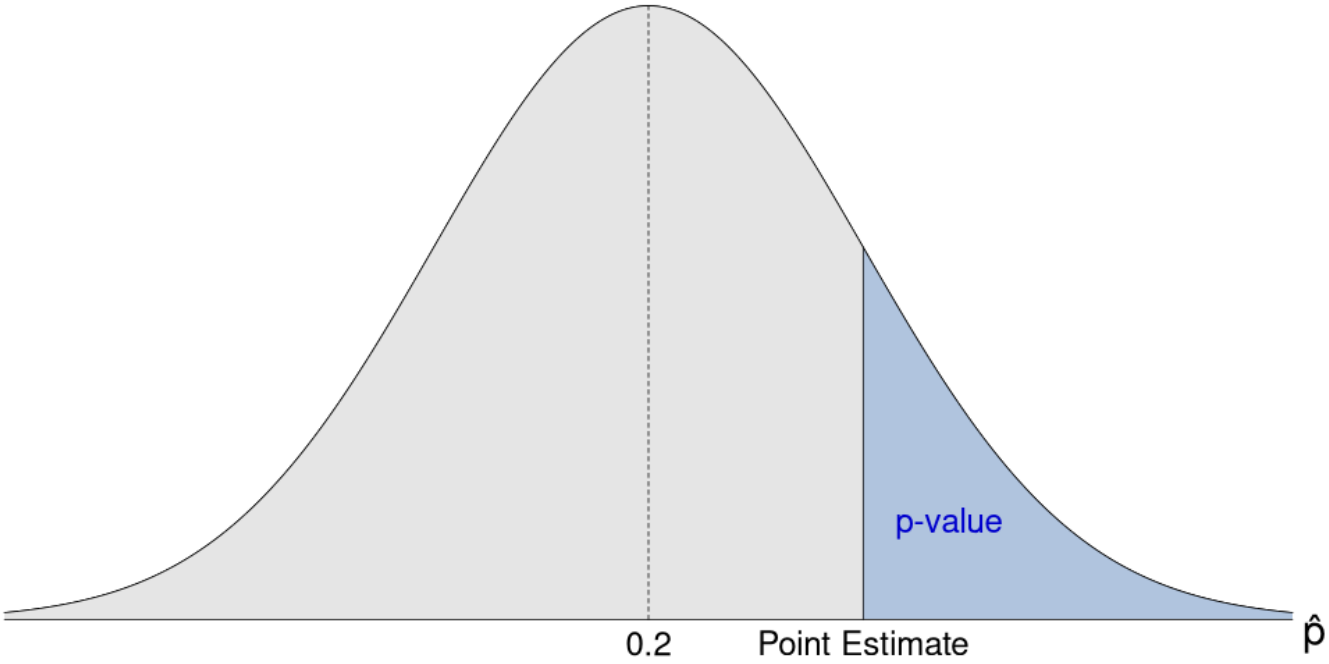



*Fig. 6.4.2* The sampling distribution of the population in this example is shown and illustrates a right-tailed test. In a right-tailed test, the $p$-value is the probability to the right of the point estimate.

---

## Example 6.4.3

A hypothesis test has the hypotheses given below. Should a left-tailed test, a right-tailed test, or a two-tailed test be performed to find the $p$-value?

$$H_0 : p = 0.7$$
$$H_a : p \neq 0.7$$

## Solution

The alternative hypothesis is this test uses a “not-equal” symbol. Therefore, we would perform a two-tailed test to find the $p$-value.

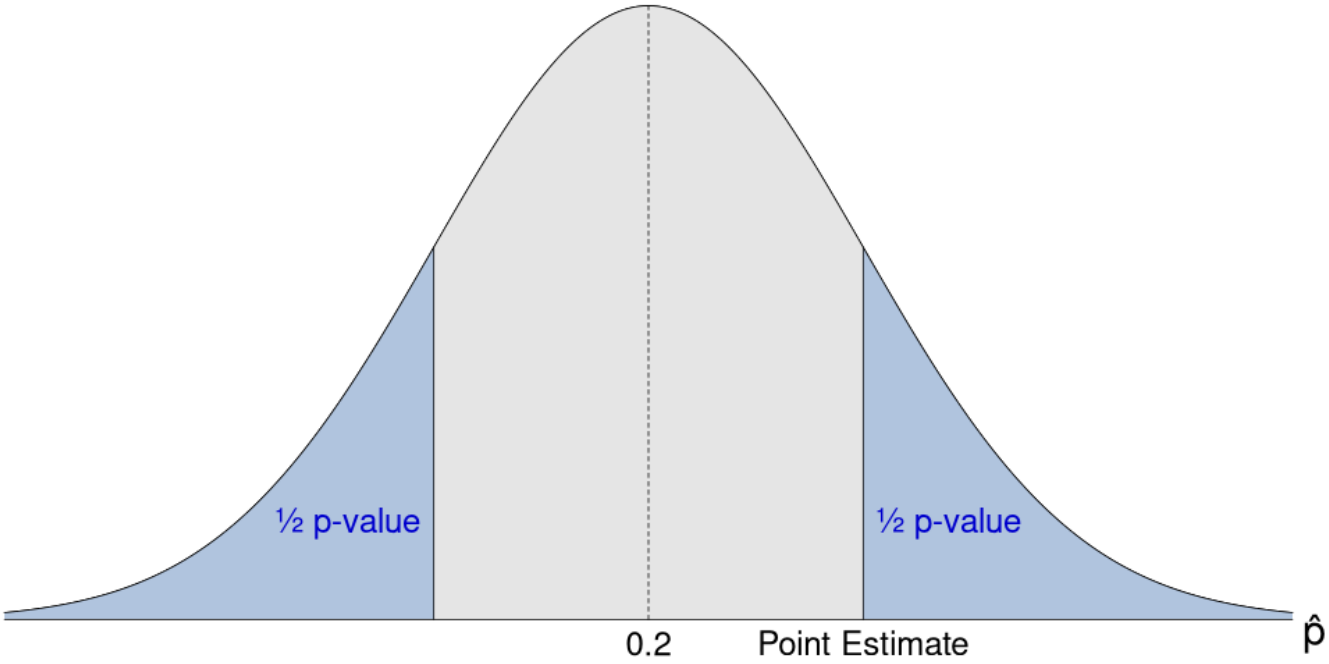



*Fig. 6.4.3* The sampling distribution of the population in this example is shown and illustrates a two-tailed test. In a two-tailed test, we first find just half of the $p$-value by calculating the probability to the left or right of the point estimate depending on which tail the point estimate is closer to. (In this figure, the point estimate is closer to the right tail, so we would calculate the probability to the right of the point estimate.) Then we double the probability to get the full $p$-value. Doubling the probability we initially calculate includes the mirror probability in the opposite tail in the $p$-value.

## Hypothesis Tests

We are finally ready to combine all the things we've learned in this chapter to perform full hypothesis tests. Here are the steps we will follow in the following examples.

1. **State the null and alternative hypotheses.**
   The null and alternative hypotheses are complementary statements and use mathematically opposite symbols. The symbol used in the null hypothesis always has an "equal" in it, and the symbol used in the alternative hypothesis never has an "equal" in it.
2. **Assuming the null hypothesis is true, identify the sampling distribution.**
   Use the central limit theorem to identify important features of the distribution like the mean and the standard error.
3. **Find the $p$-value.**
   First, calculate the point estimate using the sample data. Determine whether to perform a left-tailed, right-tailed, or two-tailed test based on the alternative hypothesis, then use the sampling distribution to find the $p$-value.
4. **Draw a conclusion.**
   Compare the $p$-value to the level of significance $\alpha$. If the $p$-value is less than $\alpha$, reject the null hypothesis and accept the alternative hypothesis. Otherwise, do not reject the null hypothesis and do not accept the alternative hypothesis. State the conclusion in context using plain words.

Jeffrey, an eight-year old, established a mean time of $16.43$ seconds for swimming the $25$-yard freestyle with a standard deviation of $0.8$ seconds. His dad, Frank, thought that Jeffrey could swim the $25$-yard freestyle faster using goggles. Frank bought Jeffrey a new pair of expensive goggles and timed Jeffrey as he swam the $25$-yard freestyle $15$ times. He obtained the following swim times (in seconds):

14.96, 15.51, 15.54, 16.14, 15.55, 16.73, 16.4, 16.59, 14.76, 17.6, 17.68, 16.71, 14.87, 15.73, 16.42

Frank assumes Jeffrey's swim times are normally distributed. Conduct a hypothesis test at the $5\%$ significance level to conclude whether or not the goggles helped Jeffrey swim faster.

Solution

Step 1: State the null and alternative hypotheses.

Frank thinks Jeffrey's mean swim time would be faster with goggles; that is, $\mu < 16.43$. Since the symbol has no equal in it, this is the alternative hypothesis. Then we can write the null and alternative hypotheses as

$$H_0 : \mu \geq 16.43$$
$$H_a : \mu < 16.43$$

Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

We are testing the population mean, and we are told the population standard deviation. Then by the central limit theorem and assuming the null hypothesis is true, sample means are normally distributed with mean

$$\mu = 16.43$$

and standard error

$$\frac{\sigma}{\sqrt{n}} = \frac{0.8}{\sqrt{15}} = 0.2066.$$

Step 3: Find the $p$-value.

The point estimate of the population mean is the sample mean $\bar{x}$.

```
x = c(14.96, 15.51, 15.54, 16.14, 15.55, 16.73, 16.4, 16.59, 14.76, 17.6, 17.68, 16.71, 14.87, 15.73, 16.42)
n = length(x)

xbar = sum(x)/n
xbar
```

16.0793333333333

The sample mean is $\bar{x} = 16.0793$.

Since the alternative hypothesis $H_a$ uses a less than symbol, we will perform a left tailed test. That is, the $p$-value is the probability

Skip to main content

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{16.0793 - 16.43}{0.2066} = -1.6975.$$

So $P(\bar{x} \leq 16.0793) = P(z \leq -1.6975)$. We use R to calculate this.

```
pnorm(q = -1.6975)
```

0.0448010855505775

Then the $p$-value is $P(\bar{x} \leq 16.0793) = P(z \leq -1.6975) = 0.0448$. That is, assuming the null hypothesis is true, that using goggles didn't improve Jeffrey's mean swim time, there is a $4.48\%$ chance that a random sample of $15$ of Jeffrey's swims with goggles would yield a sample mean of $16.0793$ seconds or less.

### Step 4: Draw a conclusion

The hypothesis test is at the $5\%$ significance level, so $\alpha = 0.05$. Since

$$p\text{-value} = 0.0448 < 0.05 = \alpha,$$

we reject the null hypothesis and accept the alternative hypothesis. The chance of obtaining the sample mean we did if the null hypothesis were true is so unlikely that we think it is more likely that the null hypothesis is not true.

We conclude that Jeffrey *does* improve his swim time using goggles.

---

## Example 6.4.5

A college football coach records the mean weight that his players can bench press as $275$ pounds. Three of his players thought that the mean weight was more than that amount. They asked $30$ of their teammates for their estimated maximum lift on the bench press exercise, obtaining the following data (in pounds):

205, 205, 205, 215, 215, 215, 225, 241, 241, 252, 252, 265, 265, 275, 275, 313, 313, 316, 316, 316, 316, 316, 338, 338, 341, 345, 345, 368, 368, 385

Conduct a hypothesis test using a $2.5\%$ level of significance to determine if the bench press mean is more than $275$ pounds.

### Solution

### Step 1: State the null and alternative hypotheses.

The three players think the mean bench press weight of the team is more than $275$ pounds. Mathematically, we write this as $\mu > 275$. Since the greater-than symbol has no equal in it, this is our alternative hypothesis. Then the hypotheses are

$$H_0 : \mu \leq 275$$
$$H_a : \mu > 275$$

Skip to main content

Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

We are testing the population mean, but we are *not* given the population standard deviation. We will need to approximate the population standard deviation using the sample standard deviation and a $t$-distribution with $df = n - 1 = 30 - 1 = 29$ degrees of freedom.

Assuming the null hypothesis is true, the mean of the distribution is

$$\mu = 275.$$

To find the standard error $\frac{s}{\sqrt{n}}$ of the distribution, we first need to find the standard deviation $s$ of the sample. To do so, first find the sample mean.

```
x = c(205, 205, 205, 215, 215, 215, 225, 241, 241, 252, 252, 265, 265, 275, 275, 313, 313, 316, 316, 316, 31
n = length(x)

xbar = sum(x)/n
xbar
```

286.166666666667

The sample mean is $\bar{x} = 286.1667$. Using this, we calculate the sample standard deviation.

```
s = sqrt(sum( (x - xbar)^2 )/(n-1))
s
```

55.8983580866344

The sample standard deviation is $s = 55.8984$. Then the standard error of the sampling distribution is

$$\frac{s}{\sqrt{n}} = \frac{55.8984}{\sqrt{30}} = 10.2056.$$

Step 3: Find the $p$-value.

We already found in step 2 that the sample mean is $\bar{x} = 286.1667$. This is the point estimate of the population mean.

Since $H_a$ uses a greater-than symbol, we will perform a right-tailed test. So the $p$-value is $P(\bar{x} \geq 286.1667)$. To calculate this, we will need the $t$-score of $\bar{x}$:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{286.1667 - 275}{10.2056} = 0.9962.$$

So $P(\bar{x} \geq 286.1667) = P(t \geq 0.9962)$. Let's use R to find this probability.

```
1 - pt(q = 0.9962, df = 29)
```

0.163696572454292

Skip to main content

Then the $p$-value is $P(\bar{x} \geq 286.1667) = P(t \geq 0.9962) = 0.1637$. That is, assuming the null hypothesis is true, that the team's mean lift weight is $275$ pounds, there is a $16.37\%$ chance that if we randomly sample $30$ team members, their mean lift weight would be at least $286.1667$ pounds.

Step 4: Draw a conclusion.

The level of significance is $2.5\%$, so $\alpha = 0.025$. Since

$$p\text{-value} = 0.1637 \geq 0.025 = \alpha,$$

we do not reject the null hypothesis and do not accept the alternative hypothesis.

The evidence is not strong enough to conclude that the team mean lift weight is greater than $275$ pounds.

---

## Example 6.4.6

Joon believes that $50\%$ of first-time brides in the United States are younger than their grooms. She performs a hypothesis test to determine if the percentage is the same or different from $50\%$. Joon samples $95$ first-time brides, and $51$ reply that they are younger than their grooms. For the hypothesis test, she uses a $1\%$ level of significance.

Solution

Step 1: State the null and alternative hypotheses.

Joon wants to know if the percent of first-time brides that are younger than their grooms is $50\%$ (that is, if $p = 0.50$) or different from $50\%$ (that is, if $p \neq 0.50$). Then the hypotheses are

$$H_0 : p = 0.50$$
$$H_a : p \neq 0.50$$

Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

We are testing the population proportion. By the central limit theorem, sample proportions are normally distributed with mean

$$p = 0.50$$

and standard error

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.50(1-0.50)}{95}} = 0.0513.$$

Step 3: Find the $p$-value.

$$\hat{p} = \frac{x}{n} = \frac{51}{95} = 0.5368.$$

Since $H_a$ uses a not-equal-to symbol, we will perform a two-tailed test. That means half of the $p$-value is in each tail. Since we are assuming the mean of the sampling distribution is $p = 0.50$, the point estimate $\hat{p} = 0.5368$ is closer to the right tail of the distribution. So we will first calculate the half of the $p$-value in the right tail, as represented by $P(\hat{p} \geq 0.5368)$. To do so, we first need to find the $z$-score for $\hat{p}$:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.5368 - 0.50}{0.0513} = 0.7173.$$

So $P(\hat{p} \geq 0.5368) = P(z \geq 0.7173)$. We use R to calculate the probability.

```
1 - pnorm(q = 0.7173)
```

0.236594503619779

Then the half of the $p$-value in the right right tail is $P(\hat{p} \geq 0.5368) = P(z \geq 0.7173) = 0.2366$. To find the full $p$-value, we need to also include the mirror probability in the left tail, so we double the probability we just calculated:

$$p\text{-value} = 2(0.2366) = 0.4732.$$

So, assuming the null hypothesis that the proportion of first-time brides younger than their grooms is $50\%$, there is a $47.32\%$ chance that a random survey would yield a sample proportion at least as extreme as $\hat{p} = 0.5368 = 53.68\%$.

Step 4: Draw a conclusion.

The level of significance of the hypothesis test is $1\%$, so $\alpha = 0.01$. Since

$$p\text{-value} = 0.4732 \geq 0.01 = \alpha,$$

we cannot reject the null hypothesis.

The evidence is insufficient to conclude that the proportion of first-time brides that are younger than their grooms is different than $50\%$.

# 7. The $\chi^2$-Distribution

## 7.1. The $\chi^2$-Distribution

### Objectives

- Calculate probabilities using the $\chi$^2-distribution.

Skip to main content

# The $\chi^2$-Distribution

In this chapter, we learn about how to conduct hypothesis tests using the $\chi^2$-distribution. The $\chi^2$-distribution ($\chi$ is the Greek letter "chi", pronounced "kai") is different than the normal distribution or the $t$-distribution. Like the $t$-distribution, the $\chi^2$-distribution depends on the degrees of freedom $df$, and there are different $\chi^2$-distributions for different values of the degrees of freedom. The mean of the $\chi^2$-distribution is $\mu = df$, where $df$ is the degrees of freedom. The standard deviation is $\sigma = \sqrt{2(df)}$.

The random variable for a $\chi^2$-distribution with $k$ degrees of freedom is the sum of $k$ independent squared standard normal variables:

$$\chi^2 = Z_1^2 + Z_2^2 + \cdots + Z_k^2.$$

Unlike the normal distribution and the $t$-distribution, the curve of the $\chi^2$-distribution is asymmetrical.

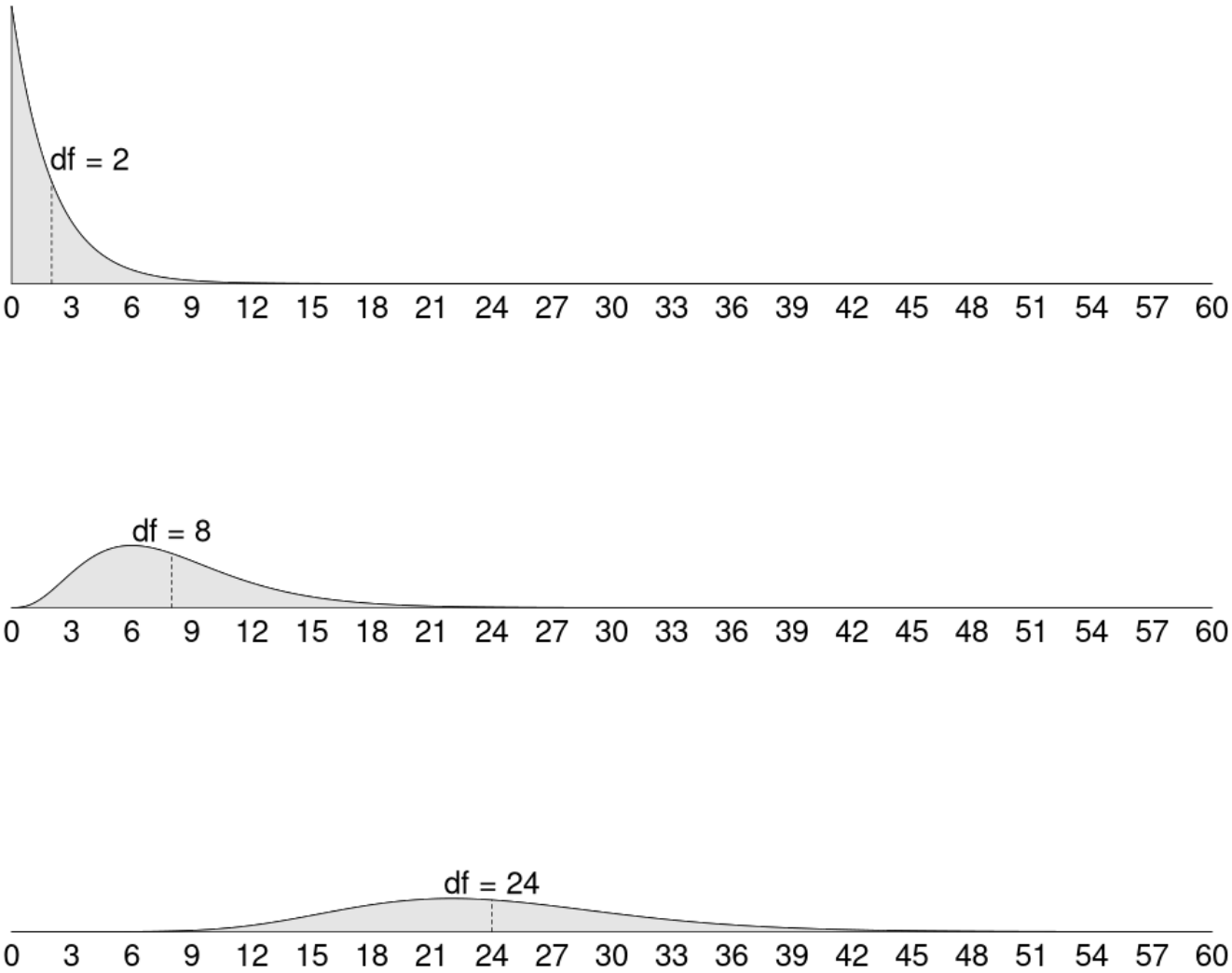



*Fig. 7.1.1* Three different $\chi^2$-distributions with different degrees of freedom. The mean $\mu$ of a $\chi^2$-distribution is equal to its degrees of freedom $df$, so $\mu = df$.

To find probability left of some value when using the $\chi^2$-distribution, we will use the R function

```
pchisq(q, df)
```

Skip to main content

where `q` is the $\chi^2$-value, and `df` is the degrees of freedom.

---

## Example 7.1.1

Consider a $\chi^2$-distribution with $4$ degrees of freedom.

1. Find $P(\chi^2 < 2)$.
2. Find $P(\chi^2 > 6)$.
3. Find $P(3 < \chi^2 < 5)$.

### Solution

#### Part 1

We can calculate $P(\chi^2 < 2)$ simply using R:

```
pchisq(q = 2, df = 4)
```

0.264241117657115

So $P(\chi^2 < 2) = 0.2642$.

#### Part 2

We want to find $P(\chi^2 > 6)$, which means we want to find the area to the *right* of $\chi^2 = 6$. The `pchisq` function only returns the probability to the *left* of the given value, sowe will use the formula $P(\chi^2 > 6) = 1 - P(\chi^2 < 6)$ to calculate what we need. In R, this formula becomes:

```
1 - pchisq(q = 6, df = 4)
```

0.199148273471456

So $P(\chi^2 > 6) = 0.1991$.

#### Part 3

To find $P(3 < \chi^2 < 5)$, we will first calculate *all* the area to the left of the larger value $\chi^2 = 5$, then subtract the excess area to the left of the smaller value $\chi^2 = 3$.

```
pchisq(q = 5, df = 4) - pchisq(q = 3, df = 4)
```

0.270527905187429

So $P(3 < \chi^2 < 5) = 0.2705$.

## Objectives

- Conduct a hypothesis test on the variance of a population.

## Variance and the Test Statistic

In the previous chapter, we learned how we test a claim about the mean or the proportion of a population. The process for testing a claim about the variance of a population is very similar.

Recall that the **variance** of a population is simply the square of the population standard deviation; that is, if $\sigma$ is the population standard deviation, then $\sigma^2$ is the population variance. While the variance and the standard deviation are not quite the same, the two parameters both measure how far data is from the mean. Testing the population variance is the same as testing the population standard deviation.

Unlike hypothesis tests for the population mean or the population proportion, we do not directly use a point estimate when testing the population variance. Instead, we use a **test statistic**. A test statistic is a quantity that summarizes the data in a sample and can be used in hypothesis testing. The $z$-scores and $t$-scores we calculated when testing population means and population proportions in the previous chapter are test statistics.

We assume the underlying population is normally distributed. Then the test statistic used when performing a hypothesis test on the population variance is

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2},$$

where $n$ is the sample size, $s$ is the sample standard deviation, and $\sigma$ is the population standard deviation hypothesized by the null hypothesis.

In this case, the sampling distribution of all possible test statistics is a $\chi^2$-distribution with $df = n - 1$ degrees of freedom.

We can use this sampling distribution and test statistic to perform a hypothesis test on the population variance. The general steps for completing a hypothesis test have not changed:

1. State the null and alternative hypotheses.
2. Assuming the null hypothesis is true, identify the sampling distribution.
3. Find the $p$-value.
4. Draw a conclusion.

---

### Example 7.2.1

A beverage company wants to make sure that the amount of soda in their two-liter bottles doesn't differ significantly bottle-to-bottle. Specifically, they want the population standard deviation to be no larger than $1.5$ milliliters. The quality control division of the company carefully measures the amount of soda in $13$ bottles and obtains the following results (in milliliters):

2002.08, 1999.25, 2002, 1996.51, 1998.99, 2000.06, 1999.73, 2001.16, 2001.19, 2002.14, 2000.21, 2002.15, 2000.06

Conduct a hypothesis test at the $10\%$ significance level to determine is the beverage company's criterion is satisfied.

Skip to main content

Solution

Step 1: State the null and alternative hypotheses.

The beverage company doesn't want the population standard deviation to be larger than $1.5$ milliliters; that is, the company wants $\sigma \leq 1.5$. Since the symbol has an 'equal' in it, this is the null hypothesis. The alternative hypothesis is the opposite of the null hypothesis.

$$H_0 : \sigma \leq 1.5$$
$$H_a : \sigma > 1.5$$

Note, we have chosen to write our hypotheses in terms of the population standard deviation $\sigma$. But we could also have written the hypotheses in terms of the population variance $\sigma^2$. Since $1.5^2 = 2.25$, the same hypotheses written in terms of the population variance are

$$H_0 : \sigma^2 \leq 2.25$$
$$H_a : \sigma^2 > 2.25$$

These two ways of writing the hypotheses are equivalent and either is acceptable.

Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

Because we are testing the population standard deviation, we will use a $\chi^2$ distribution with $df = n - 1 = 13 - 1 = 12$ degrees of freedom.

Step 3: Find the $p$-value.

To calculate the test statistic, we first need to find the sample standard deviation. To begin, we calculate the sample mean:

```
x = c(2002.08, 1999.25, 2002, 1996.51, 1998.99, 2000.06, 1999.73, 2001.16, 2001.19, 2002.14, 2000.21, 2002.1
n = length(x)

xbar = sum(x)/n
xbar
```

2000.42538461538

The sample mean is about $\bar{x} = 2000.4254$. Next, we calculate the sample standard deviation:

```
s = sqrt(sum( (x - xbar)^2 )/(n - 1))
s
```

1.62851473938995

The sample standard deviation is $s = 1.6285$. Assuming the null hypothesis is true (so that $\sigma = 1.5$), we calculate the test statistic

$$\chi^2 = \frac{(n-1)s^2}{} = \frac{(13-1)1.6285^2}{} = 14.1443.$$

Skip to main content

Noting that the alternative hypothesis uses a “greater-than” symbol, we will be performing a right-tailed test. Therefore, the $p$-value is equal to $P(\chi^2 > 14.1443)$. We use R to calculate this probability.

```
1 - pchisq(q = 14.1443, df = 12)
```

0.291588646205538

So the $p$-value is $P(\chi^2 > 14.1443) = 0.2916$. That means that, assuming the null hypothesis is true, there is a $29.16\%$ chance that a random sample of $13$ bottles would have a sample standard deviation as extreme as $s = 1.6285$.

Step 4: Draw a conclusion.

The level of significance is $10\%$, so $\alpha = 0.10$. But since

$$p\text{-value} = 0.2916 \geq 0.10 = \alpha,$$

we do not reject the null hypothesis.

There is not enough evidence to conclude that the population standard deviation of the bottles of soda is greater than $1.5$ milliliters.

# 7.3. Goodness-of-Fit Test

## Objectives

- Use a goodness-of-fit hypothesis test to test if observed values fit an expected discrete distribution.

## Goodness-of-Fit Test

In this type of hypothesis test, we determine whether the data *“fit”* a particular discrete distribution or not. For example, we may suspect the unknown data fit a binomial distribution. We use a goodness-of-fit test to determine if there is a fit or not.

The test statistic for a goodness-of-fit test is:

$$\chi^2 = \sum \frac{(O - E)^2}{E},$$

where

- $O$ = observed values (sample data)
- $E$ = expected values (from theory)

The observed values are the data values from the sample. The expected values are the values we would expect to get if the null hypothesis were true. The sampling distribution of the test statistic is a $\chi^2$-distribution. The number of degrees of freedom of the distribution is $df = k - 1$, where $k$ is the number of different categories in the hypothesized discrete distribution.

Generally, the hypotheses of a goodness-of-fit test are

Skip to main content

$$H_0 : \text{The actual population fits the expected distribution.}$$
$$H_a : \text{The actual population does not fit the expected distribution.}$$

These hypotheses may be written as sentences and should be expressed in the context of the particular problem.

A goodness-of-fit test is almost always right-tailed. The further apart observed values and expected values are from each other, the further out in the right tail the test statistic will be. (A left-tailed goodness of fit test would test if the observed values fit the expected values too well, which is not generally something we are concerned with.)

For a goodness-of-fit test to be valid, the expected value for each category needs to be at least five.

As with any hypothesis test, the fundamental steps for performing a goodness-of-fit hypothesis test are:

1. State the null and alternative hypotheses.
2. Assuming the null hypothesis is true, identify the sampling distribution.
3. Find the $p$-value.
4. Draw a conclusion.

---

## Example 7.3.1

Absenteeism of college students from math classes is a major concern to math instructors because missing class appears to increase the drop rate. Suppose that a study was done to determine if the actual student absenteeism rate follows faculty perception. The faculty expected that any randomly chosen group of $100$ students would miss class according to Table 7.3.1.

*Table 7.3.1* Expected student absenteeism frequency in math classes.

| Number of Absences per Term | Expected Number of Students |
|---|---|
| 0-2 | 50 |
| 3-5 | 30 |
| 6-8 | 12 |
| 9+ | 8 |

A random survey across of $100$ students all mathematics courses was then done to determine the actual number of absences in a course. Table 7.3.2 displays the results of the survey.

*Table 7.3.2* Observed student absenteeism frequency in math classes.

| Number of Absences per Term | Expected Number of Students |
|---|---|
| 0-2 | 35 |
| 3-5 | 40 |
| 6-8 | 20 |

Skip to main content

Perform a goodness-of-fit test at a $1\%$ level of significance to determine whether or not student absenteeism fits faculty perception.

Solution

Step 1: State the null and alternative hypotheses.

The null and alternative hypotheses are

$$H_0 : \text{Student absenteeism fits faculty perception}$$
$$H_a : \text{Student absenteeism does not fit faculty perception}$$

Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

This is a goodness-of-fit test. We are testing to see if student absenteeism is distributed in the same way that faculty expect it to be distributed; that is, we are testing whether or not faculty perception of student absenteeism is a *good fit* for actual student absenteeism. Since this is a goodness-of-fit test, the sampling distribution is a $\chi^2$ distribution. The categories of the expected distribution are $0$-$2$, $3$-$5$, $6$-$8$, and $9$+, for a total of $k = 4$ categories. Therefore, the sampling distribution has $df = k - 1 = 4 - 1 = 3$ degrees of freedom.

Step 3: Find the $p$-value.

To find the $p$-value, we first must calculate the test statistic

$$\chi^2 = \sum \frac{(O - E)^2}{E}.$$

The observed values $O$ are the actual number of students observed in each category. The expected values $E$ are the numbers of students the faculty expect to be in each category. These value are found in the two tables above. We will use R to calculate the test statistic.

```
O = c(35, 40, 20, 5)
E = c(50, 30, 12, 8)

chisq = sum( (O - E)^2 / E )
chisq
```

14.2916666666667

The test statistic is $\chi^2 = 14.2917$. Since goodness-of-fit tests are almost always right-tailed tests, we will perform a right-tailed test to find the $p$-value. That means the $p$-value is $P(\chi^2 \geq 14.2917)$.

```
1 - pchisq(q = 14.2917, df = 3)
```

0.00253382515522105

So $P(\chi^2 \geq 14.2917) = 0.0025$. That is, if student absenteeism actually does fit faculty perception, then there is only a $0.25\%$ chance that a random sample of $100$ students would deviate from the expected distribution as far as our sample did.

Skip to main content

Step 4: Make a conclusion about the null hypothesis.

We are conducting this hypothesis test at the $1\%$ level of significance, so $\alpha = 0.01$. Since

$$p\text{-value} = 0.0025 < 0.01 = \alpha,$$

we reject the null hypothesis.

We conclude that the actual distribution of student absenteeism does not fit faculty perception.

---

## Example 7.3.2

Suppose you roll a six-sided die $80$ times, with the following results:

*Table 7.3.3* Observed outcomes from rolling a six-sided die $80$ times.

| Face Value | Number of Rolls |
|---|---|
| 1 | 16 |
| 2 | 21 |
| 3 | 14 |
| 4 | 9 |
| 5 | 7 |
| 6 | 13 |

Use a goodness-of-fit test with a $5\%$ level of significance to determine whether or not the die is fair.

### Solution

Part 1: State the null and alternative hypotheses.

The null hypothesis is that the true distribution of rolls of your die matches the distribution of rolls of a fair die. This could be said more succinctly as:

$$H_0 : \text{The die is fair.}$$
$$H_a : \text{The die is not fair.}$$

Part 2: Assuming the null hypothesis is true, identify the sampling distribution.

Since we want to see if the distribution of rolls of a fair die is a *good fit* for the distribution of rolls of your die, we will use a $\chi^2$-distribution to test the hypothesis. Since there are $k = 6$ categories or outcomes, the sampling distribution has

Part 3: Find the $p$-value.

To find the $p$-value, we first must calculate the test statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E}.$$

Note that the table above are the observed values. But if we roll a fair die $80$ times, we *expect* to roll each number $\frac{1}{6}80 = 13.3333$ times:

*Table 7.3.4* Expected outcomes from rolling a six-sided die $80$ times.

| Face Value | Number of Rolls |
|---|---|
| 1 | 13.3333 |
| 2 | 13.3333 |
| 3 | 13.3333 |
| 4 | 13.3333 |
| 5 | 13.3333 |
| 6 | 13.3333 |

Let's use R to calculate the test statistic $\chi^2$:

```
O = c(16, 21, 14, 9, 7, 13)
E = c(13.3333, 13.3333, 13.3333, 13.3333, 13.3333, 13.3333)

chisq = sum( (O - E)^2/E )
chisq
```

9.40002350055875

The test statistic is $\chi^2 = 9.400$.

Since a goodness-of-fit test is almost always right-tailed, the $p$-value we want to find is equal to $P(\chi^2 \geq 9.400)$.

```
1 - pchisq(q = 9.400, df = 5)
```

0.0941343840306237

So the $p$-value is $P(\chi^2 \geq 9.400) = 0.0941$. That is, assuming the die is fair, there is a $9.41\%$ chance of obtaining the distribution we observed.

Step 4: Draw a conclusion.

Since the level of significance for this hypothesis test is $5\%$, the value of $\alpha = 0.05$. Because

Skip to main content

$$p\text{-value} = 0.0941 \geq 0.05 = \alpha,$$

we do not reject the null hypothesis.

There is not enough evidence to conclude that your die is not fair.

# 8. Correlation and Linear Regression

## 8.1. Scatter Plots

### Objectives

- Create scatter plots of bivariate data.
- Identify from a scatter plot if two variables appear to be correlated and, if so, whether the variables are positively or negatively correlated.

### Introduction

Professionals often want to know how two or more variables are related. For example, is there a relationship in a math class between the score on the midterm exam and the score on the final exam? If there is a relationship, what is the relationship and how strong is it? For instance, if a student gets an $86\%$ on the midterm exam, how accurately can we predict their score on the final exam?

Data from two paired variables together is called **bivariate data**. (In reality, statisticians use multivariate data, meaning many variables.) When two variables are related so that we can predict the value of one variable if we know the value of the other variable, we say the two variables are **correlated**. In this chapter, we will learn how to measure how correlated two variables are, how to model the correlation, and how to make predictions from correlated bivariate data.

### Scatter Plots

It will be usefuly to first examine a way to graphically display the relation between two variables $x$ and $y$. The easiest and most common way to do this is to use a **scatter plot**. A scatter plot is a plot of the known $(x, y)$ data point pairs on a coordinate plane. Plotting the data graphically allows us to more easily visualize any correlation between $x$ and $y$.

We can create a scatter plot using R with the `plot` function:

```
plot(x, y)
```

Here, `x` is a list of the $x$-values, and `y` is a list of the corresponding $y$-values. Note that corresponding values in `x` and `y` must be in the same order. For example, if one of the points we want to plot is $(x, y) = (7, 3)$, and $x = 7$ is the fifth value in list `x`, then $y = 3$ must also be the fifth value in list `y`.

---

#### Example 8.1.1

Skip to main content

Amelia plays basketball for her high school. She wants to improve to play at the college level. She notices that the number of points she scores in a game goes up in response to the number of hours she practices her jump shot each week. She records the following data:

*Table 8.1.1* The number of hours Amelia practices her jump shot each week and the points she scores in the next basketball game.

| $x$—Hours practicing jump shot | $y$—Points scored in a game |
|---|---|
| 5 | 15 |
| 7 | 22 |
| 9 | 28 |
| 10 | 31 |
| 11 | 33 |
| 12 | 36 |

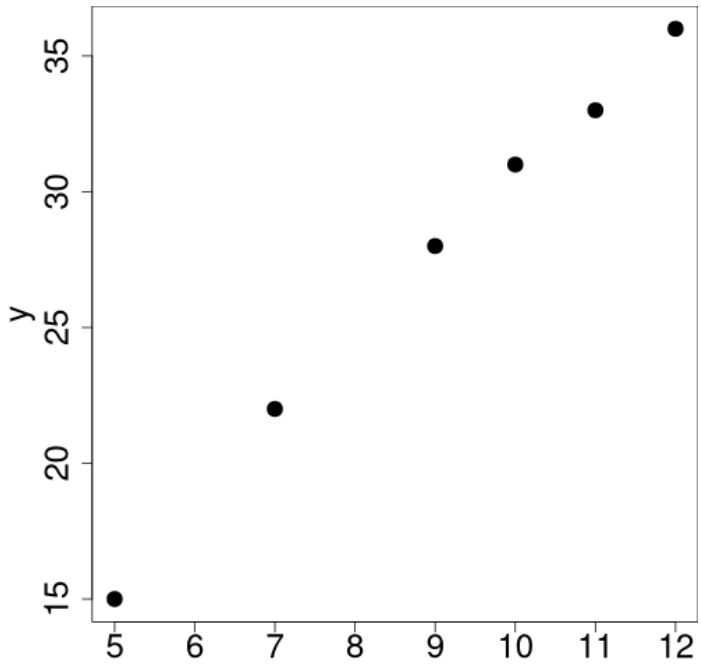
Construct a scatter plot and state if what Amelia thinks appears to be true.

### Solution

We get $6$ pairs of $(x, y)$ points from the data: $(5, 15)$, $(7, 22)$, $(9, 28)$, $(10, 31)$, $(11, 33)$, and $(12, 36)$. We can easily plot these by hand, but it is faster and easier to use R to obtain our scatter plot using the `plot` function.

```
x = c(5, 7, 9, 10, 11, 12)
y = c(15, 22, 28, 31, 33, 36)

plot(x, y)
```




Skip to main content

*Fig. 8.1.1* The scatterplot of the data. Each $(x, y)$ pair is represented by a point.

Notice in Figure 8.1.1 there appears to be a strong linear correlation between the number of hours Amelia spends practicing each week ($x$) and the number of points she scores in a game ($y$). That is, the points almost fall on a straight line. Also, the larger the value of $x$ is, the larger the corresponding value of $y$ tends to be. From this, we might presume that the more hours Amelia spends practicing, the more points she'll score in a game.

---

A scatter plot shows the **direction** of the correlation between the variables. A clear direction happens when either:

- Higher values of one variable tend to occur with higher values of the other variable, and lower values of one variable tend to occur with lower values of the other variable. Graphically, the points tend to go "uphill" when viewing the graph from left to right. In this case, we say the variables have a **positive** correlation.
- Higher values of one variable tend to occur with lower values of the other variable. Graphically, the points tend to go "downhill" when viewing the graph from left to right. In this case, we say the variables have a **negative** correlation.

We can determine the **strength** of the relationship by looking at the scatter plot and seeing how close the points are to a line, a power function, an exponential function, or to some other type of function.

When you look at a scatterplot, you want to notice the overall pattern and any deviations from the pattern. The following scatterplot examples illustrate these concepts.

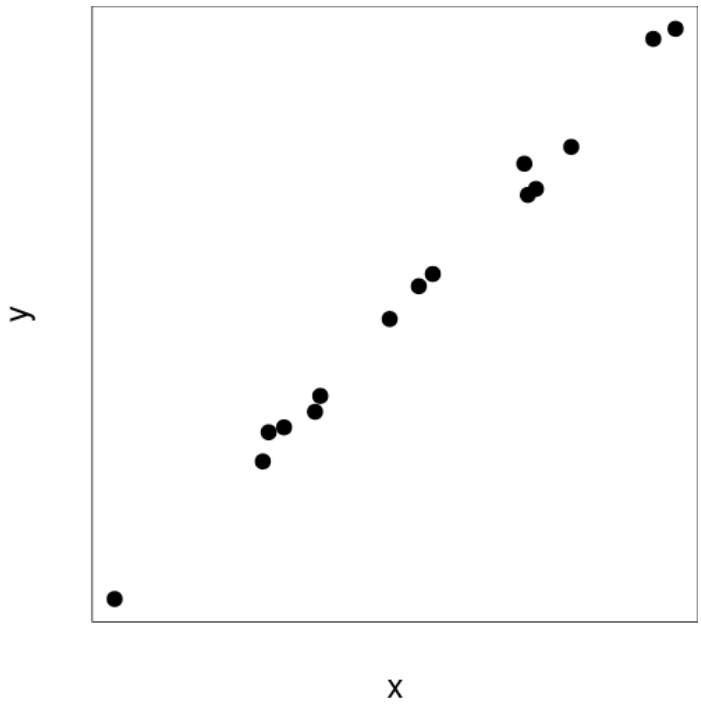



*Fig. 8.1.2* The data in this scatter plot are strongly correlated; they fall almost perfectly on a straight line. Points with higher $x$-values tend to also have higher $y$-values, so the points look like they are trending "uphill" when viewed from left to right. This means the variables are positively correlated.

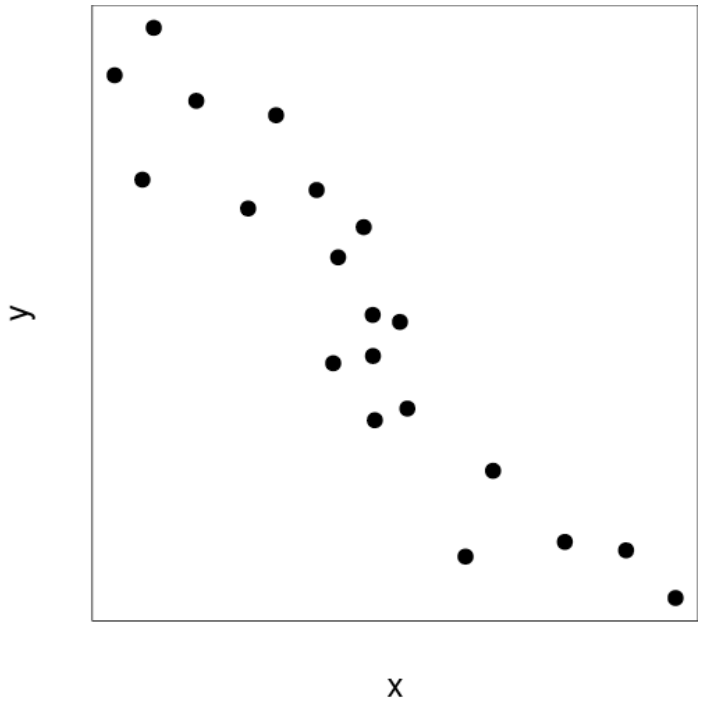



*Fig. 8.1.3* The data in this scatter plot are more weakly correlated than the data plotted in Figure 8.1.2; the data roughly fall along a line but are more spread out than in the previous figure. Points with higher $x$-values tend to also have lower $y$-values, so the points look like they are trending "downhill" when viewed from left to right. This means the variables are negatively correlated.

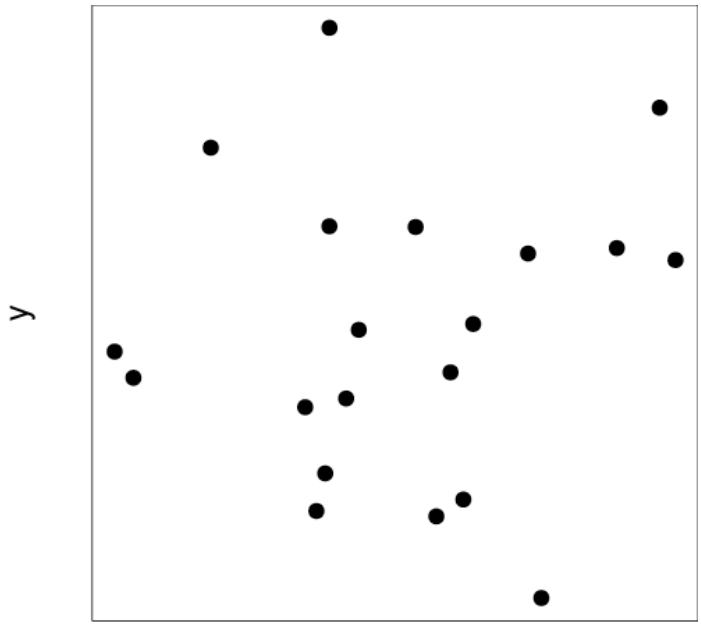



*Fig. 8.1.4* The data in this scatter plot do not appear to be correlated at all. Knowing the value of $x$ for a data point would not help us predict the value of $y$.

Skip to main content

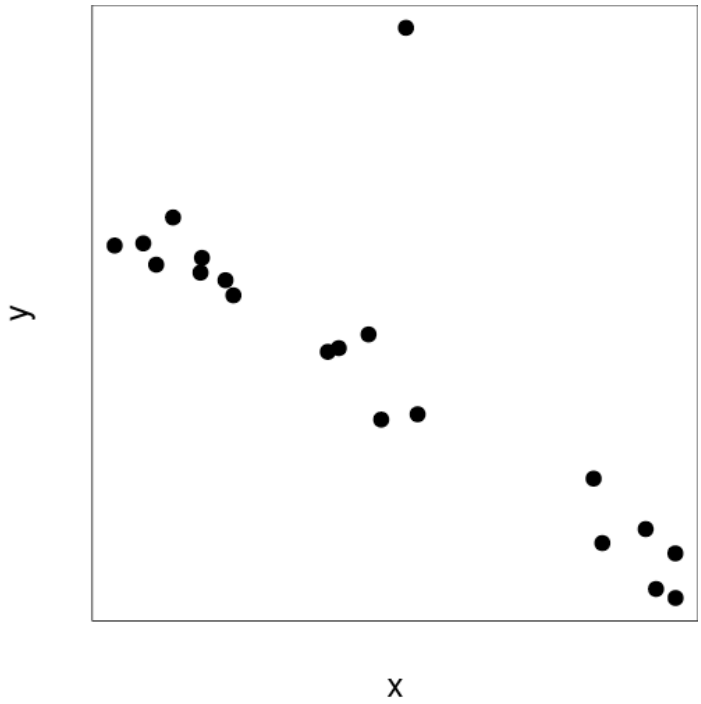



*Fig. 8.1.5* The data in this scatter plot are negatively correlated, but note that there is one point that is an outlier. When we get a point like this one that is a strong outlier, it may indicate there is a problem. For example, maybe we made a mistake when collecting the data for this point. Or maybe we made a mistake when inputting the data into the computer. It's also possible that the point is correct and it simply doesn't follow the pattern we would expect.

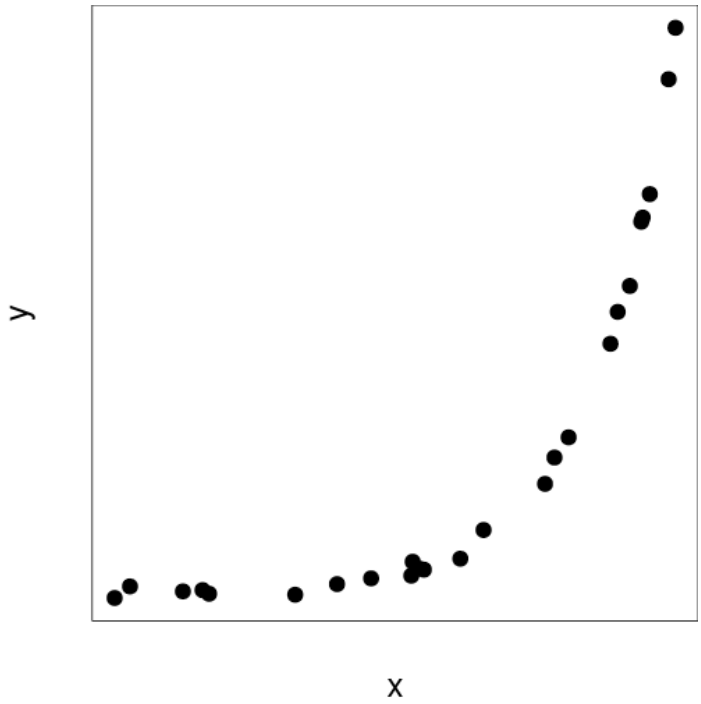



*Fig. 8.1.6* The data in this scatter plot are clearly correlated, but the data is not *linearly* correlated. Instead of following a line, the data appear to follow an exponential growth function.

In this chapter, we are interested in scatter plots that are linearly correlated. Linear correlation is quite common. The linear

Skip to main content

**linear regression** to find the line that best "fits" points that appear to show a linear relationship. We can use this line to predict a value of $y$ for a given value of $x$.

> **Warning**
>
> In this section, we have learned that data are linearly correlated when the scatter plot of the data shows a linear relationship. However, there are two exceptions that we should be aware of. Data where the the linear relationship matches a horizontal line or a vertical line are *not* correlated. This is because when a line is vertical or horizontal, one of the variables never changes, so knowing the value of one variable does not give any more predictive power over the other variable.
>
> For example, suppose we gather data each week where $x$ is the number of cars a car dealership sells that week and $y$ is the number of Tuesdays in that week. Since every week has exactly one Tuesday, the value of $y$ never changes; it is always $y = 1$. Knowing how many cars $x$ that are sold in a given week does not give us any more information about $y$; we already know $y = 1$. Similarly, knowing that $y = 1$ does not help us predict how many cars $x$ are sold in any given week.

# 8.2. Linear Equations

## Objectives

- Identify a linear equation in two variables, including the slope and $y$-intercept of the line.
- Interpret the meaning of the slope and $y$-intercept of a linear equation..
- Construct linear equations from applications.

## Linear Equations

Linear regression for two variables is based on a linear equation with one independent variable. The equation has the form:

$$y = a + bx,$$

where $a$ and $b$ are constant numbers.

The variable $x$ is called the **independent variable**, and $y$ is the **dependent variable**. While not always true, we often think of $x$ as causing or resulting in $y$; that is, $y$ depends on $x$. Typically, we choose a value to substitute for the independent variable and then solve for the dependent variable.

The graph of a linear equation of the form $y = a + bx$ is a straight line. Any line that is not vertical can be described by this equation.

## Slope and Y-Intercept of a Linear Equation

For the linear equation $y = a + bx$, the point $(0, a)$ is the $y$-**intercept** of the line, and $b$ is the **slope** of the line.

The $y$-intercept $(0, a)$ of a line is the point where the line crosses the $y$-axis. In applications, it often represents a starting value or initial value of $y = a$ when $x = 0$.

Skip to main content

The slope of a line describes the steepness and direction of the line. The larger the absolute value of slope $b$, the steeper the line. If $b$ is positive, the line has an “uphill” slope from left to right. If $b$ is negative, the line has a “downhill” slope from left to right.

In terms of the variables, a positive slope ($b > 0$) means that as the value of $x$ gets bigger, the value of $y$ gets bigger, too. Conversely, a negative slope ($b < 0$) means that as the value of $x$ gets bigger, the value of $y$ gets smaller. A line with zero slope ( $b = 0$) means $x$ can get bigger or smaller, but $y$ always stays exactly the same.

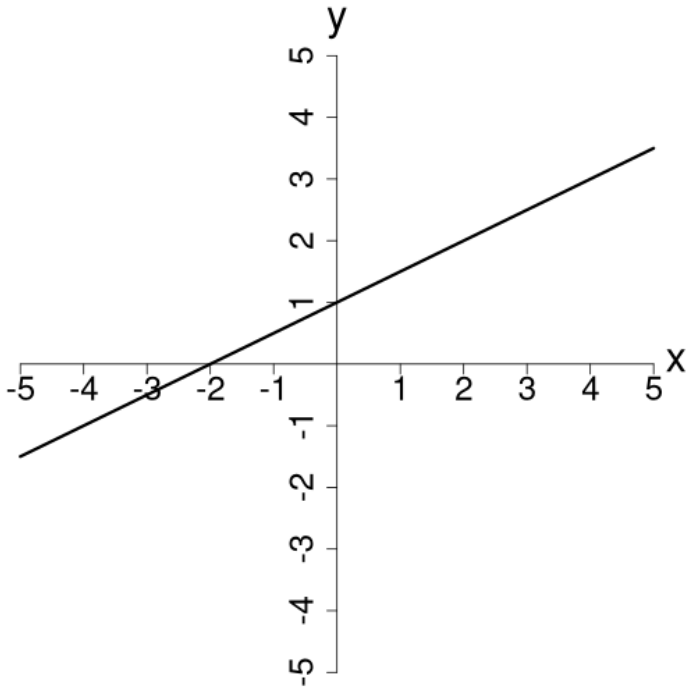
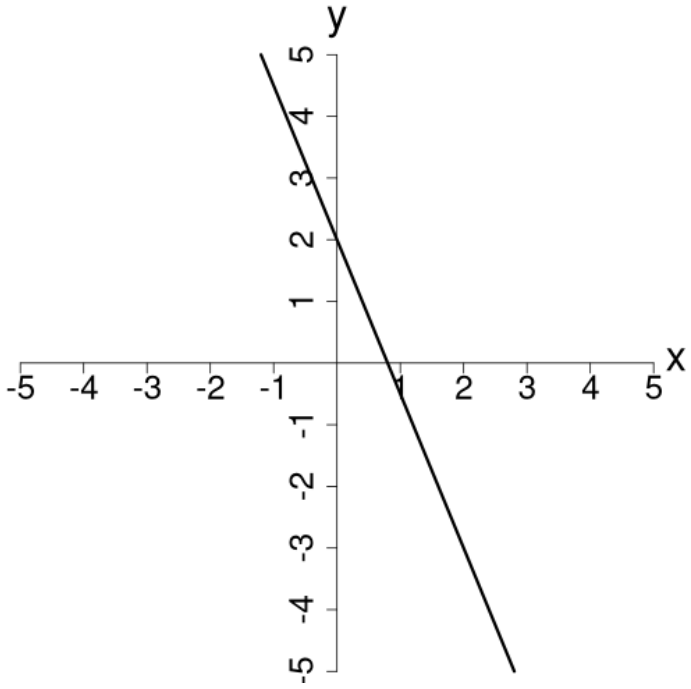



*Fig. 8.2.1* The graph of the line $y = 1 + 0.5x$. Since $a = 1$, the line crosses the $y$-axis as the $y$-intercept $(0, 1)$. Since the slope $b = 0.5$ is a positive number, the line goes “uphill” from left to right. The slope $b = 0.5$ isn’t a very big number, so the line isn’t very steep.




Skip to main content

*Fig. 8.2.2* The graph of the line $y = 2 - 2.5x$. Since $a = 2$, the line crosses the $y$-axis as the $y$-intercept $(0, 2)$. Since the slope $b = -2.5$ is a negative number, the line goes "downhill" from left to right. The absolute value of slope $b = -2.5$ is larger than the slope of the line in Figure 8.2.1, so this line is steeper.

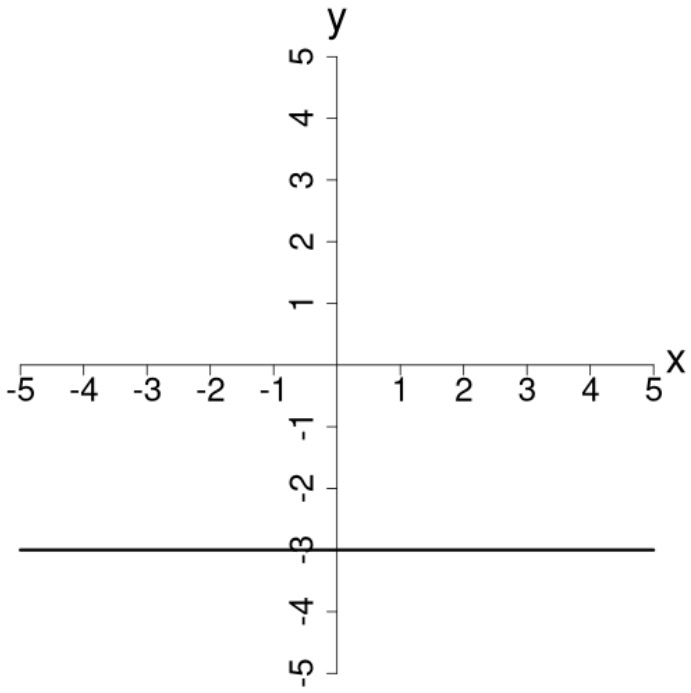



*Fig. 8.2.3* The graph of the line $y = -3$, which can also be written as $y = -3 + 0x$. Since $a = -3$, the line crosses the $y$-axis as the $y$-intercept $(0, -3)$. Since the slopeof the line is $b = 0$, the line is horizontal.

---

## Example 8.2.1

Svetlana tutors to make extra money for college. For each tutoring session, she charges a one-time fee of $\$25$ plus $\$15$ per hour of tutoring.

1. Find the equation of the line that expresses the total amount Svetlana charges in terms of the number of hours of the tutoring session.
2. What is the $y$-intercept of the line? How should we interpret this value?
3. What is the slope of the line? How should we interpret this value?
4. If Svetlana has a tutoring session that lasts $4.5$ hours, how much will she charge?

### Solution

#### Part 1

Let $x$ be the number of hours of the tutoring session.

Let $y$ be the amount Svetlana charges for the tutoring session.

Skip to main content

Svetlana charges a one-time fee of $\$25$ at the start of a tutoring session; it does not change with the length of the tutoring session. But Svetlana also charges $\$15$ per hour. We will need to multiply the hourly rate of $\$15$ by the number of hours $x$ to find the hourly charge. Combined, we have the total amount Svetlana charges for a tutoring session:

$$y = 25 + 15x.$$

### Part 2

Since $a = 25$, the $y$-intercept is $(x, y) = (0, 25)$. If someone hires Svetlana for a tutoring session that lasts $x = 0$ hours (for example, if the person hires Svetlana but doesn't show), Svetlana still charges them $y = 25$ dollars.

### Part 3

The slope is $b = 15$. Since $b$ is positive, the line has an "uphill" direction. Also, longer tutoring sessions (bigger $x$) mean Svetlana will charge more (bigger $y$).

### Part 4

If Svetlana has a tutoring session that lasts $4.5$ hours then, mathematically, $x = 4.5$. Substituting this into our linear equation will tell us $y$:

$$y = 25 + 15(4.5) = 92.5.$$

So for a $4.5$ hour tutoring session, Svetlana charges $\$92.50$.

---

## Example 8.2.2

Belisario made $36$ cookies. He eats two cookies each day.

1. Find an equation that expresses the total number of cookies that Belisario has in terms of the number of days that have passed.
2. What is the $y$-intercept? How should we interpret this value?
3. What is the slope? How should we interpret this value?
4. After $6$ days, how many cookies does Belisario have?

### Solution

### Part 1

Let $x$ be the number of days that have passed since Belisario made cookies.

Let $y$ be the number of cookies Belisario has left.

Belisario starts off with $36$ cookies, but we need to take $2$ away for each day that has passed. This gives us the equation

$$y = 36 - 2x.$$

Part 2

Since $a = 36$, the $y$-intercept is $(x, y) = (0, 36)$. That means when $x = 0$ days have passed, Belisario has $y = 36$ cookies. Or, even more simply, this is just saying that Belisario starts off with $36$ cookies.

Part 3

The slope is $b = -2$. Since $b$ is negative, the line has a “downhill” direction. Also, the more days that pass (bigger $x$), the fewer cookies Belisario will have (fewer $y$).

Part 4

To find out how many cookies Belisario has after $6$ days, we plug in $x = 6$ into the equation:

$$y = 36 - 2(6) = 24.$$

After $6$ days Belisario has $24$ cookies left.

# 8.3. The Regression Equation

## Objectives

- Construct the line of best fit for bivariate data.
- Plot a scatter plot together with its line of best fit.
- Given an $x$-value, use the line of best fit for a data set to predict the corresponding $y$-value.

## The Regression Equation

Data rarely fit a straight line exactly. Usually, we must be satisfied with rough predictions. We might have a set of data whose scatter plot appears to approximately **“fit”** a straight line. This line is called the **line of best fit** or the **least squares regression line**.

We can use R to find the coefficients of the line of best fit using the function:

```
lm(y~x)
```

Here, `x` and `y` are paired lists of values, and `y~x` means we assume `y` depends on `x`. The function `lm(y~x)` returns the coefficients of the **l**inear **m**odel that best fits the data.

For example, imagine we have the following data:

Skip to main content

*Table 8.3.1* Example data values for an independent variable $x$ and the corresponding values for a dependent variable $y$. We will find the line of best fit for this data.

| **$x$-values** | 5 | 2 | 4 | 5 | 11 | 0 | 8 | 5 |
|---|---|---|---|---|---|---|---|---|
| **$y$-values** | 6 | 4 | 2 | 7 | 13 | $-2$ | 12 | 8 |

We can use R to find the coefficients of the line of best fit.

```
x = c(5, 2, 4, 5, 11, 0, 8, 5)
y = c(6, 4, 2, 7, 13, -2, 12, 8)

lm(y~x)
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     -0.625        1.375
```

This gives us the $y$-value of the $y$-intercept, $a = -0.625$, and the slope, $b = 1.375$, of the line of best fit. So the line of best fit for the data is

$$y = -0.625 + 1.375x$$

We can graph this line *after* graphing the scatterplot using the function `abline`:

```
abline(a, b)
```

Here, `a` is the $a$ coefficient and `b` is the $b$ coefficient in the linear equation $y = a + bx$.

```
# Graph the scatter plot
plot(x, y)

# Plot the line
abline(a = -0.625, b = 1.375)
```

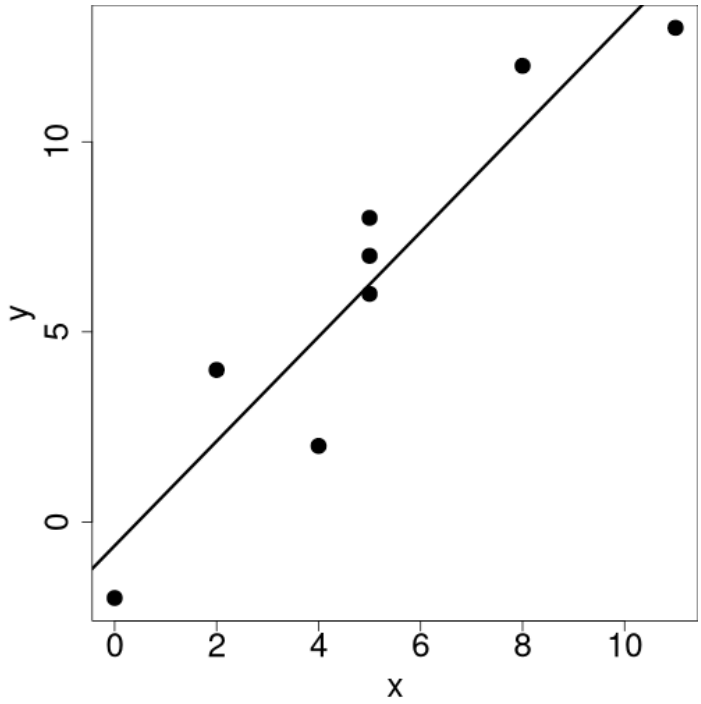



*Fig. 8.3.1* The scatter plot of the data above together with the line of best fit $y = -0.625 + 1.375x$.

The line $y = -0.625 + 1.375x$ plotted in Figure 8.3.1 is the best fit for the observed data points. We can use the equation of the best fit line to predict the $y$-value that would correlate to a given $x$-value.

For example, we might want to predict what the value of $y$ would be if $x = 6.5$. We can predict this $y$-value by plugging $x = 6.5$ into the equation of the best fit line:

$$y = -0.625 + 1.375(6.5) = 8.3125.$$

So if $x = 6.5$, we predict that $y = 8.3125$.

---

## Example 8.3.1

A random sample of $11$ statistics students produced the following data, where $x$ is the score of the third exam (with $80$ points possible), and $y$ is the score of the final exam (with $200$ points possible).

*Table 8.3.2* The scores of $11$ students on the third exam and the final exam.

| $x$ (Third Exam Score) | $y$ (Final Exam Score) |
|---|---|
| 65 | 175 |
| 67 | 133 |
| 71 | 185 |
| 71 | 163 |
| 66 | 126 |
| 75 | 198 |
| 67 | 153 |
| 70 | 163 |
| 71 | 159 |
| 69 | 151 |
| 69 | 159 |

1. Find the best fit line for the data.
2. Graph the scatter plot of the data together with the best fit line.
3. Predict a student's final exam score if their third exam score is $73$.

Solution

Part 1

We will use R to find the coefficients of the line of best fit.

```
x = c(65, 67, 71, 71, 66, 75, 67, 70, 71, 69, 69)
y = c(175, 133, 185, 163, 125, 198, 153, 163, 159, 151, 159)

lm(y~x)
```

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   -176.301        4.866
```

The coefficients of the best fit line are $a = -176.301$ and $b = 4.866$. So the best fit line is

$$y = -176.301 + 4.866x$$

Skip to main content

Part 2

```
# Graph of the scatter plot
plot(x, y)

# Plot of the best fit line
abline(a = -176.301, b = 4.866)
```

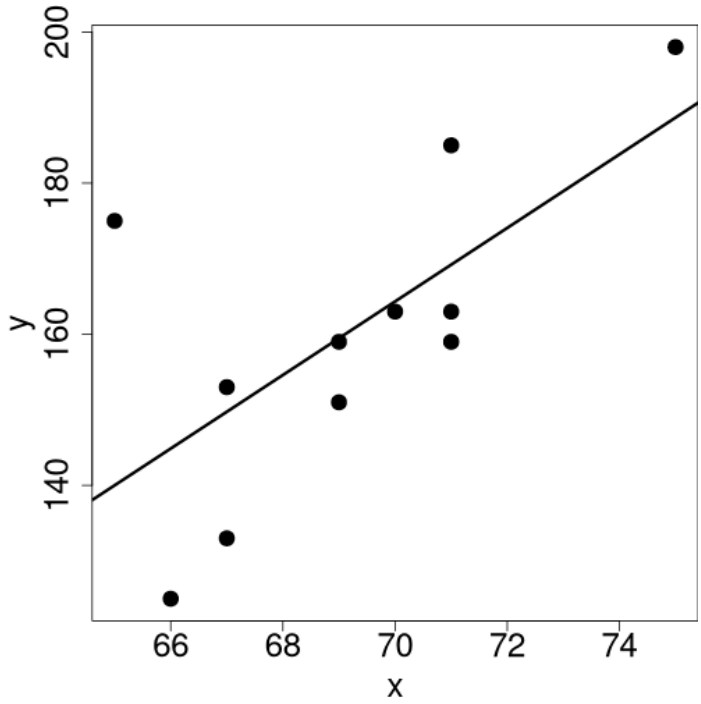



*Fig. 8.3.2* The scatter plot of the data above together with the line of best fit $y = -176.301 + 4.866x$.

Step 3

If the third exam score is $73$, we can predict the final exam score by plugging $x = 73$ into the equation for the best fit line:

$$y = -176.301 + 4.866(73) = 178.917.$$

So if a student scores $73$ out of $80$ points on the third exam, we predict that they would score about $178.917$ out of $200$ points on their final exam.

# 8.4. The Correlation Coefficient

## Objectives

- Identify and interpret the correlation coefficient for bivariate data.
- Conduct a hypothesis test to determine if the correlation coefficient for a population is significant.

## The Correlation Coefficient

Skip to main content

Besides looking at the scatter plot and seeing how closely the data points fit a line of best fit, how can you tell how well the line fits the data?

The **correlation coefficient**, $r$, developed by Karl Pearson in the early $1900$s, is a numerical value that measures the strength and direction of the linear association between an independent variable $x$ and a dependent variable $y$.

We can find the correlation coefficient in R using the `cor` function:

```
cor(x, y)
```

Here, `x` is the list of the independent $x$ values, and `y` is the corresponding list of the dependent $y$ values.

The correlation coefficient has many useful properties:

- The value of $r$ is always between -1 and +1: $-1 \leq r \leq 1$.
- The size of the correlation coefficient $r$ indicates the strength of the linear relationship between $x$ and $y$. Values of $r$ close to -1 or to +1 indicate a stronger linear relationship between $x$ and $y$.
- If $r = 0$ there is likely no linear correlation. (However, it is important to view the scatterplot, as the data may exhibit a non-linear correlation.)
- If $r = 1$, there is perfect positive correlation. If $r = -1$, there is perfect negative correlation. In both these cases, all of the original data points lie exactly on a straight line. Of course, in the real world, this will not generally happen.
- A positive value of $r$ means that the slope of the line of best fit is positive. In this case, we say the variables appear to be **positively correlated**.
- A negative value of $r$ means that the slope of the line of best fit is negative. In this case, we say the variables appear to be **negatively correlated**.

---

## Example 8.4.1

SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in the table show different depths with the maximum dive times in minutes. Find and interpret the correlation coefficient $r$ for the data.

*Table 8.4.1* The dive depth and corresponding maximum dive time at that depth for SCUBA divers.

| $x$ (Depth in Feet) | $y$ (Maximum Dive Time) |
|---|---|
| 50 | 80 |
| 60 | 55 |
| 70 | 45 |
| 80 | 35 |
| 90 | 25 |
| 100 | 22 |

Skip to main content

Solution

```
x = c(50, 60, 70, 80, 90, 100)
y = c(80, 55, 45, 35, 25, 22)

cor(x, y)
```

-0.962938502101513

So the correlation coefficient for the data is $r = -0.9629$. Since $r$ is very close to $-1$, the dive depth has a very strong linear relationship to the maximum dive time. Since $r$ is negative, the slope of the line of best fit would also be negative, meaning the dive depth and associated dive time are negatively correlated.

---

## Testing the Significance of the Correlation Coefficient

We want to perform a hypothesis test on the significance of a correlation coefficient to decide whether the linear relationship in the sample data is strong enough to use to model the relationship in the population. The correlation coefficient, $r$, tells us about the strength and direction of the linear relationship between $x$ and $y$. However, the reliability of the linear model also depends on how many observed data points are in the sample. To determine if a linear model accurately reflects the relationship in a population, we need to look at both the value of the correlation coefficient $r$ and the sample size $n$, together.

In other words, we would like to test the significance of the population correlation coefficient, denoted by $\rho$ (the Greek letter "rho"), using the sample correlation coefficient, $r$. If the evidence from the sample data is strong enough to suggest that $\rho \neq 0$, then we can conclude that the population is linearly correlated; otherwise, the evidence is not sufficient to reject the null hypothesis that $\rho = 0$, and we cannot conclude that the population is linearly correlated.

Thus, when testing the significance of the correlation coefficient, the hypotheses are always:

$$\begin{aligned} H_0 &: \ \rho = 0, \\ H_a &: \ \rho \neq 0. \end{aligned}$$

Because the alternative hypothesis uses a "not-equal-to" symbol, this kind of test is a two-tailed test.

To calculate the $p$-value for this kind of hypothesis test, we use a $t$-distribution with $df = n - 2$ degrees of freedom (where $n$ is the number of data points in our sample), and we use the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Once we have the $p$-value, we compare it to the level of significance $\alpha$ and draw our conclusion as normal:

- If $p$-value $< \alpha$, then we reject the null hypothesis. We conclude that $\rho \neq 0$; that is, we conclude that there is significant linear correlation in the population.
- If $p$-value $\geq \alpha$, then we cannot reject the null hypothesis, that $\rho = 0$. We do not have enough evidence to conclude that there is significant linear correlation in the population.

The general steps for completing a hypothesis test have not changed:

1. State the null and alternative hypotheses.

3. Find the $p$-value.
4. Draw a conclusion.

## Example 8.4.2

A random sample of $11$ statistics students produced the following data, where $x$ is the score of the third exam (with $80$ points possible), and $y$ is the score of the final exam (with $200$ points possible).

*Table 8.4.2* The scores of $11$ students on the third exam and the final exam.

| $x$ (Third Exam Score) | $y$ (Final Exam Score) |
|---|---|
| 65 | 175 |
| 67 | 133 |
| 71 | 185 |
| 71 | 163 |
| 66 | 126 |
| 75 | 198 |
| 67 | 153 |
| 70 | 163 |
| 71 | 159 |
| 69 | 151 |
| 69 | 159 |

Perform a hypothesis test with a $1\%$ level of significance to determine whether or not there is a linear correlation between a student's third exam score and their final exam score.

### Solution

#### Part 1: State the null and alternative hypotheses.

When testing the significance of the population correlation coefficient, the null and alternative hypotheses are always the same:

$$\begin{aligned} H_0 &: \rho = 0, \\ H_a &: \rho \neq 0. \end{aligned}$$

Our null hypothesis $H_0$ is that there is no linear correlation in the population between a student's third exam score and their final exam score. The alternative hypothesis $H_a$ is that there is some degree of linear correlation in the population between a student's third exam score and their final exam score.

Skip to main content

Part 2: Assuming the null hypothesis is true, identify the sampling distribution.

When testing the population correlation coefficient $\rho$, we use a $t$-distribution with $n-2$ degrees of freedom, where $n$ is the number of $(x, y)$ points in our data. Since our data has $11$ $(x, y)$ points in this case, we have $df = n - 2 = 11 - 2 = 9$ degrees of freedom.

Part 3: Find the $p$-value

We will use the test statistic

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

to find the $p$-value. To calculate this $t$-score, we first must find the sample correlation coefficient $r$:

```
x = c(65, 67, 71, 71, 66, 75, 67, 70, 71, 69, 69)
y = c(175, 133, 185, 163, 126, 198, 153, 163, 159, 151, 159)

cor(x, y)
```

0.663093590999518

The sample correlation coefficient is $r = 0.6631$. Then we calculate that the test statistic is

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.6631\sqrt{11-2}}{\sqrt{1-(0.6631)^2}} = 2.6576.$$

Since the alternative hypothesis $H_a$ uses a not-equal-to symbol, we will perform a two-tailed test. That means *half* of the $p$-value is represented by $P(t \geq 2.6576)$, which we will calculate using R.

```
1 - pt(q = 2.6576, df = 9)
```

0.0130742219589208

So half the $p$-value is $P(t \geq 2.6576) = 0.0131$. This means that the whole $p$-value is

$$p\text{-value} = 2(0.0131) = 0.0262.$$

So assuming that the null hypothesis is true—that there is no correlation in the population—there would still be a $2.62\%$ chance that a random sample of $11$ students' test scores would have a correlation coefficient at least as extreme as $r = 0.6631$.

Part 4: Draw a conclusion.

We're performing the test at the $1\%$ significance level, so $\alpha = 0.01$. Thus, since

$$p\text{-value} = 0.0262 \geq 0.01 = \alpha,$$

the evidence is not sufficient to reject the null hypothesis.

We conclude that it is possible that there is no linear correlation in the population between a student's third exam score and their final exam score.

# 9. Hypothesis Tests of Two or More Populations

## 9.1. Hypothesis Testing: Two Population Means

### Objectives

- Conduct a hypothesis test comparing the means of two populations.

### Hypothesis Test with Two Population Means

In previous sections, we learned how to test the mean of a population in relation to a particular value. But sometimes, we want to compare the means of two different populations. For example, if we have one population with mean $\mu_1$ and another population with mean $\mu_2$, we might want to test if $\mu_1$ is smaller than $\mu_2$. The null and alternative hypotheses for this hypothesis test would be

$$\begin{aligned} H_0 &: \ \mu_1 \geq \mu_2, \\ H_a &: \ \mu_1 < \mu_2. \end{aligned}$$

There is another way to write these hypotheses that will be more convenient for our purposes. If we subtract $\mu_2$ from both sides of the inequalities, our hypotheses become

$$\begin{aligned} H_0 &: \ \mu_1 - \mu_2 \geq 0, \\ H_a &: \ \mu_1 - \mu_2 < 0. \end{aligned}$$

This second way to write these hypotheses is equivalent to the first way, but it allows us to focus on the difference in the population means rather than comparing them directly. This is useful because, to find the $p$-value for a hypothesis test like this, we are going to use the distribution of the random variable $D = \bar{X}_1 - \bar{X}_2$.

Let $n_1$ be the size of the sample drawn from population $1$, and let $n_2$ be the size of the sample drawn from population $2$. Recall from the central limit theorem that $\bar{X}_1$ and $\bar{X}_2$ are normally distributed as long as $n_1, n_2 \geq 30$ or the underlying populations are normally distributed. It turns out that the random variable $D = \bar{X}_1 - \bar{X}_2$ is also normally distributed with mean $\mu_1 - \mu_2$ and standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, where $\sigma_1$ and $\sigma_2$ are the standard deviations of the respective populations. As with the central limit theorem, we still require here that either $n_1, n_2 \geq 30$ or the underlying populations are normally distributed.

Of course, we almost never know $\sigma_1$ and $\sigma_2$ in practice. Instead, we approximate the population standard deviations with the sample standard deviations from the two samples, $s_1$ and $s_2$. In this case, the mean of the sampling distribution is

$$\mu_1 - \mu_2$$

and the standard error is approximated by

$$\sqrt{s_1^2 \quad s_2^2}$$

Because we are approximating the population standard deviations with sample standard deviations, we need to use a $t$-distribution to find the $p$-value instead of the standard normal distribution. The $t$-score is given by the formula

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

and the degrees of freedom $df$ is the smaller of the values $n_1 - 1$ and $n_2 - 1$.

The fundamental steps for conducting a hypothesis test remain the same:

1. State the null and alternative hypotheses.
2. Assuming the null hypothesis is true, identify the sampling distribution.
3. Find the $p$-value.
4. Draw a conclusion.

---

## Example 9.1.1

A study is done by a community group in two neighboring colleges to determine which one graduates students with more math classes. College A samples $11$ graduates, with an average of four math classes and a standard deviation of $1.5$ math classes. College B samples nine graduates. Their average is $3.5$ math classes with a standard deviation of one math class. The community group believes that a student who graduates from college A has taken more math classes, on the average. Both populations have a normal distribution. Test at a $1\%$ significance level.

### Solution

First, gather the given information:

*Table 9.1.1* Statistics from the samples taken from college A and college B.

| | College A | College B |
|---|---|---|
| **Sample Size** | $n_A = 11$ | $n_B = 9$ |
| **Sample Mean** | $\bar{x}_A = 4$ | $\bar{x}_B = 3.5$ |
| **Sample Standard Deviation** | $s_A = 1.5$ | $s_B = 1$ |

### Step 1: State the null and alternative hypotheses.

The community group believes that a student who graduates from college A has taken more math classes, on the average, than a student who graguates from college B. Mathematically, this means the community group believes $\mu_A > \mu_B$, where $\mu_A$ is the average number of math classes a graduate from college A takes, and $\mu_B$ is the average number of math classes a graduate from college B takes. But to perform a hypothesis test, we need to test the difference in the population means, so we rewrite $\mu_A > \mu_B$ as $\mu_A - \mu_B > 0$. Thus,

$$H_0 : \mu_A - \mu_B \leq 0$$
$$H_a : \mu_A - \mu_B > 0$$

Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

If the null hypothesis is true, then the mean of the sampling distribution is

$$\mu_{\overline{D}} = \mu_A - \mu_B = 0,$$

and the standard error is

$$\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{1.5^2}{11} + \frac{1^2}{9}} = 0.5618.$$

Since $n_A - 1 > n_B - 1$, we will use a $t$-distribution with $df = n_B - 1 = 9 - 1 = 8$ degrees of freedom to find the $p$-value.

Step 3: Find the $p$-value.

The point estimate of $\mu_A - \mu_B$ is

$$\bar{x}_A - \bar{x}_B = 4 - 3.5 = 0.5.$$

The test statistic is given by the $t$-score

$$t = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{0.5 - 0}{0.5618} = 0.8900.$$

Since the alternative hypothesis $H_a$ uses a “greather-than” symbol, we will perform a right-tailed test. The $p$-value is $P(t \geq 0.8900)$.

```
1 - pt(q = 0.8900, df = 8)
```

0.199719073651568

So the $p$-value is $P(t \geq 0.8900) = 0.1997$. In other words, if the null hypothesis *is* true, there is a $19.97\%$ chance that the mean number of math classes taken by the students randomly sampled from college A is at least $0.5$ classes more than the mean number of math classes taken by the students randomly sampled from college B.

Step 4: Make a conclusion about the null hypothesis.

We are testing the hypothesis at the $1\%$ significance level, meaning that $\alpha = 0.01$. Since

$$p\text{-value} = 0.1997 \geq 0.01 = \alpha,$$

we do not reject the null hypothesis.

Skip to main content

The evidence is insufficient to conclude that the average number of math classes taken by graduates from college A is greater than the average number of math classes taken by graduates from college B.

---

## Example 9.1.2

A professor at a large community college wanted to determine whether there is a difference in the means of final exam scores between students who took his statistics course online and the students who took his face-to-face statistics class. The randomly selected final exam scores for the two courses are listed below. Test whether or not the exam scores are different at the $5\%$ significance level.

**Sample 1: Thirty final exam scores from the online class:**

67.6, 70.66, 94.1, 41.2, 38.22, 88.2, 85.3, 61.8, 64.7, 55.9, 88.2, 55.9, 82.4, 70.6, 88.2, 91.2, 58.8, 97.1, 73.5, 91.2, 85.3, 94.1, 73.5, 61.8, 64.7, 82.4, 79.4, 64.7, 35.5, 79.4

**Sample 2: Thirty final exam scores from the face-to-face class:**

77.9, 95.3, 81.2, 74.1, 98.8, 88.2, 84.9, 92.9, 87.1, 88.2, 69.4, 57.6, 69.4, 67.1, 97.6, 85.9, 88.2, 91.8, 78.8, 71.8, 98.8, 61.2, 92.9, 90.6, 97.6, 100, 95.3, 83.5, 92.9, 89.4

### Solution

#### Step 1: State the null and alternative hypotheses.

We want to test whether or not the average final exam scores for the online class are different than the average final exam scores for the face-to-face class. This means we want to know if $\mu_1 = \mu_2$ or if $\mu_1 \neq \mu_2$, where $\mu_1$ is the average final exam score for the online class and $\mu_2$ is the average final exam score for the face-to-face class. But to perform a hypothesis test, we need to express these as the difference between population means. Rewriting $\mu_1 = \mu_2$ as $\mu_1 - \mu_2 = 0$ and $\mu_1 \neq \mu_2$ as $\mu_1 - \mu_2 \neq 0$ gives our hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 \neq 0$$

#### Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

Assuming the null hypothesis is true, the mean of the distribution is

$$\mu_1 - \mu_2 = 0.$$

The standard error is given by the formula

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

which means we need to first calculate $s_1$ and $s_2$.

Skip to main content

**Finding $s_1$:**

To find $s_1$, the sample standard deviation for the online class, we first must find $\bar{x}_1$.

```
x1 = c(67.6, 70.66, 94.1, 41.2, 38.22, 88.2, 85.3, 61.8, 64.7, 55.9, 88.2, 55.9, 82.4, 70.6, 88.2, 91.2, 58.
n1 = length(x1)

xbar1 = sum(x1)/n1
xbar1
```

72.8526666666667

So the mean for the sample from the online class is $\bar{x}_1 = 72.8527$. Next, we calculate the sample standard deviation.

```
s1 = sqrt(sum( (x1 - xbar1)^2 )/(n1 - 1))
s1
```

16.9170825587289

The sample standard deviation for the online class is $s_1 = 16.9171$.

**Finding $s_2$:**

To find $s_2$, the sample standard deviation for the face-to-face class, we first must find $\bar{x}_2$.

```
x2 = c(77.9, 95.3, 81.2, 74.1, 98.8, 88.2, 84.9, 92.9, 87.1, 88.2, 69.4, 57.6, 69.4, 67.1, 97.6, 85.9, 88.2,
n2 = length(x2)

xbar2 = sum(x2)/n2
xbar2
```

84.9466666666667

So the mean for the sample from the face-to-face class is $\bar{x}_2 = 84.9467$. Next, calculate the sample standard deviation.

```
s2 = sqrt(sum( (x2 - xbar2)^2 )/(n2 - 1))
s2
```

11.7128792459363

The sample standard deviation for the face-to-face class is $s_2 = 11.7129$.

Now we can find the standard error:

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{16.9171^2}{30} + \frac{11.7129^2}{30}} = 3.7567.$$

To find the $p$-value, we will use a $t$-distribution with $df = n_1 - 1 = 30 - 1 = 29$ degrees of freedom. (Since $n_1 - 1 = n_2 - 1$, it doesn't matter which we use to calculate the degrees of freedom.)

Step 3: Find the $p$-value.

Skip to main content

The point estimate of $\mu_1 - \mu_2$ is

$$\bar{x}_1 - \bar{x}_2 = 72.8527 - 84.9467 = -12.094.$$

The test statistic is the $t$-score

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{-12.094 - 0}{3.7567} = -3.2193.$$

Since the alternative hypothesis $H_a$ uses a "not-equal-to" symbol, we will perform a two-tailed test. That means that *half* the $p$-value is $P(t \leq -3.1293)$. Let's calculate using R.

```
pt(q = -3.2193, df = 29)
```

0.00157932307641113

So $P(t \leq -3.1293) = 0.0016$, meaing the $p$-value is

$$p\text{-value} = 2(0.0016) = 0.0032.$$

Step 4: Make a conclusion about the null hypothesis.

The level of significance for this test is $5\%$, so $\alpha = 0.05$. Since

$$p\text{-value} = 0.0032 < 0.05 = \alpha,$$

we reject the null hypothesis. The chance of getting the point estimate we did if the null hypothesis were true is so small, we think it is more likely that the null hypothesis is not true.

We conclude that the average final exam score for the online statistics class does not match the average final exam score for the face-to-face statistics class.

# 9.2. Hypothesis Testing Two Population Variances

## Objectives

- Calculate probabilities using the $F$-distribution.
- Conduct hypothesis tests comparing the variances of two populations.

## The $F$-Distribution

We can compare two independent $\chi^2$-distributed random variables by considering a fraction involving the two random variables. This fraction is a new random variable, and we say that it has an $F$-distribution. The shape of an $F$-distribution is determined by the degrees of freedom of the $\chi^2$-distribution of the numerator $df_1$ and the degrees of freedom of the $\chi^2$-distribution of the

Skip to main content

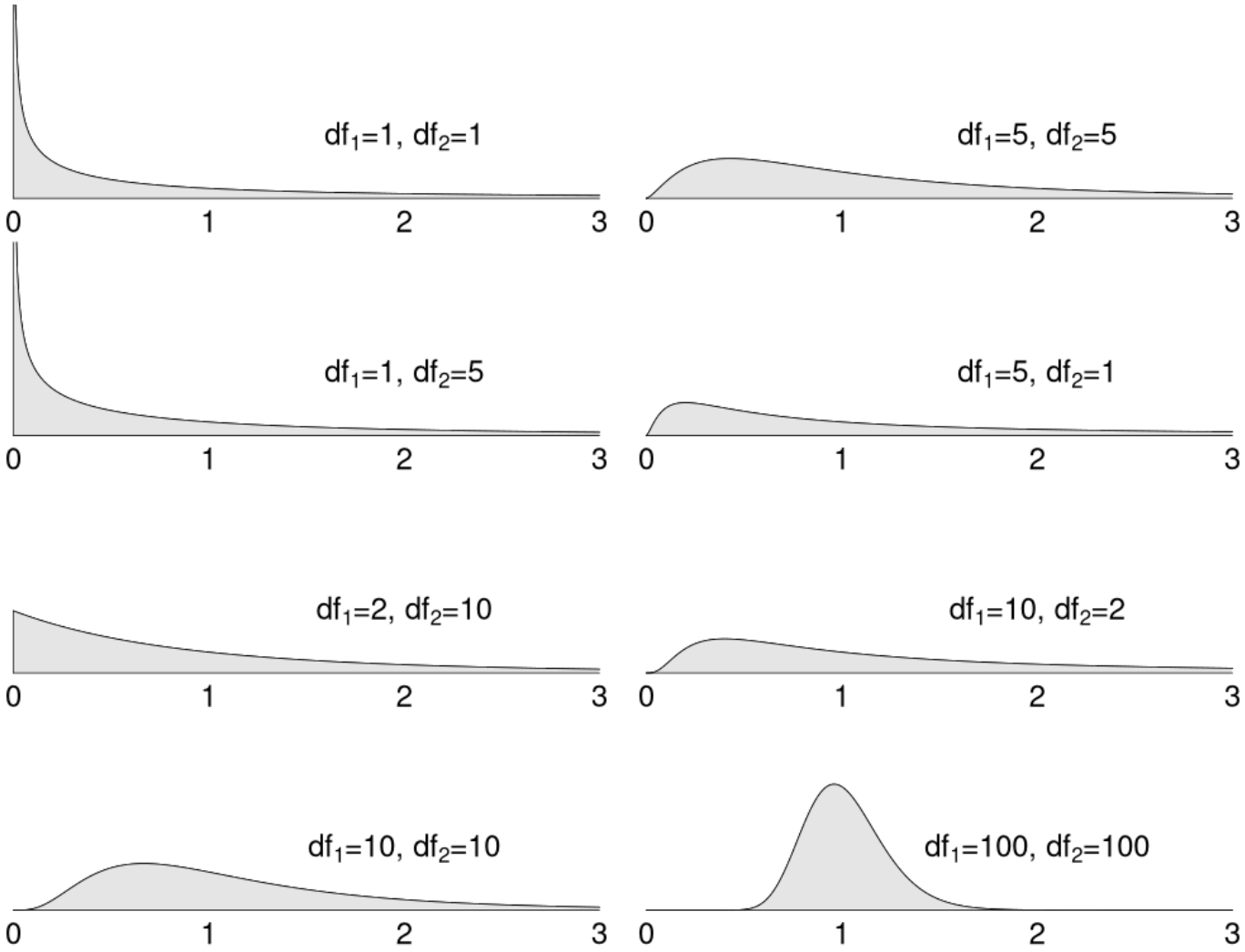



*Fig. 9.2.1* The plots of $F$-distributions with various values for $df_1$ and $df_2$.

To find the probability to the left of a value when using the $F$-distribution, we will use the R function

```
pf(q, df1, df2)
```

where we want the probability to the left of `q`, the numerator degrees of freedom is `df1`, and the denominator degrees of freedom is `df2`.

## Testing Two Population Variances

You may recall that we use a $\chi^2$-distribution when performing a hypothesis test on the variance of a population. When comparing two population variances, we have one $\chi^2$-distribution for each of the two populations, so we compare the population variances using an $F$-distribution. When the populations are independent from each other and normally distributed, the test statistic for the hypothesis test is given by

$$F = \frac{s_1^2}{s_2^2},$$

Skip to main content

where $s_1$ is the sample standard deviation of the sample taken from population $1$, and $s_2$ is the sample standard deviation of the sample taken from population $2$. If the size of sample $1$ is $n_1$ and the size of sample $2$ is $n_2$, then the numerator degrees of freedom is $df_1 = n_1 - 1$ and the denominator degrees of freedom is $df_2 = n_2 - 1$.

The general steps for completing a hypothesis test have not changed:

1. State the null and alternative hypotheses.
2. Assuming the null hypothesis is true, identify the sampling distribution.
3. Find the $p$-value.
4. Draw a conclusion.

---

## Example 9.2.1

A researcher believes that there is less variance in the heights of $10$-year-old males compared to adult males. She measures the heights of $22$ $10$-year-old males and $27$ adult males with the following results (in centimeters):

$10$ Year-Old Males (Sample A):

148.8, 141.4, 144.9, 140.2, 140.7, 139.6, 148.5, 136.9, 147.5, 148.5, 131.8, 127.9, 132.2, 138.6, 148.5, 141.1, 130.9, 139.2, 137.5, 151.9, 135, 142.6

Adult Males (Sample B):

168.8, 186.4, 172.1, 180.1, 163.7, 185.2, 173.2, 174, 165.9, 169.2, 175.7, 181.3, 172.7, 172.1, 168.3, 171.8, 195, 186.4, 179, 168.3, 171.3, 182.9, 164.4, 169, 186.8, 178.3, 191.1

Conduct a hypothesis test with a $7\%$ level of significance to test her claim.

### Solution

### Step 1: State the null and alternative hypotheses.

We want to test if the heights of 10 year-old males has a smaller variance than the heights of adult males. Written mathematically, this means that one of our hypotheses is that $\sigma_A^2 < \sigma_B^2$. We can rewrite this in a form that is consistent with the sampling distribution by dividing both sides by $\sigma_B^2$ to get

$$\frac{\sigma_A^2}{\sigma_B^2} < 1.$$

Since this inequality uses the less-than symbol, this is the alternative hypothesis. Using the opposite symbol for the null hypothesis, we obtain our hypotheses:

$$H_0 : \frac{\sigma_A^2}{\sigma_B^2} \geq 1,$$
$$H_a : \frac{\sigma_A^2}{\sigma_B^2} < 1.$$

Skip to main content

Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

Since we are testing two population variances, we will find the $p$-value using an $F$-distribution. The numerator degrees of freedom is $df_A = n_A - 1 = 22 - 1 = 21$ and the denominator degrees of freedom is $df_B = n_B - 1 = 27 - 1 = 26$.

Step 3: Find the $p$-value.

Our test statistic is $F = \frac{s_A^2}{s_B^2}$, so we need to find the standard deviation $s_A$ of sample A and the standard deviation $s_B$ of sample B. Let's start by finding $s_A$.

```
xA = c(148.8, 141.4, 144.9, 140.2, 140.7, 139.6, 148.5, 136.9, 147.5, 148.5, 131.8, 127.9, 132.2, 138.6, 148
nA = length(xA)

xbarA = sum(xA)/nA

sA = sqrt( sum( (xA - xbarA)^2 )/(nA - 1) )
sA
```

6.61266658068886

So the standard deviation of sample A is $s_A = 6.613$. The process is the same to find $s_B$.

```
xB = c(168.8, 186.4, 172.1, 180.1, 163.7, 185.2, 173.2, 174, 165.9, 169.2, 175.7, 181.3, 172.7, 172.1, 168.3
nB = length(xB)

xbarB = sum(xB)/nB

sB = sqrt( sum( (xB - xbarB)^2 )/(nB - 1) )
sB
```

8.45772965299736

The standard deviation of sample B is $s_B = 8.458$.

We can now calculate our test statistic.

```
Fscore = sA^2/sB^2
Fscore
```

0.611287694110602

Our test statistic is $$F = \frac{s_A^2}{s_B^2} = \frac{6.613^2}{8.458^2} = 0.611.$$

We are now prepared to calculate the $p$-value using the numerator degrees of freedom and denominator degrees of freedom we calculated in step 2. Since the alternative hypothesis uses a "less-than" symbol, we perform a left-tailed test.

```
pf(q = Fscore, df1 = 21, df2 = 26)
```

0.12648514626268

So the $p$-value is $P(F < 0.611) = 0.126$. In other words, assuming the null hypothesis is true, there is a $12.6\%$ chance of the variance from a sample from population A being at least as small as the variance we actually obtained relative to the variance from

Skip to main content

Step 4: Draw a conclusion.

Since $p$-value $= 0.126 \geq 0.07 = \alpha$, we do *not* reject the null hypothesis. There is not enough evidence to conclude that the variance of heights of $10$ year-old males is lower than the variance of heights of adult males.

# 9.3. ANOVA

## Objectives

- Conduct an ANOVA test to compare the means of many populations.

## Introduction to ANOVA

ANOVA stands for ANalysis Of VAriance. An ANOVA test is a hypothesis test used to compare the means among several populations. The test uses variances to help determine if the population means are equal or not. To perform a one-way ANOVA test, the following basic assumptions must be fulfilled:

- Each population is normally distributed.
- All samples are randomly selected and independent.
- The populations have equal variances (or standard deviations).

The null hypothesis is that all the population means are the same. The alternative hypothesis is that at least one pair of means is different. For example, if there are $k$ populations:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \cdots = \mu_k$$
$$H_a : \text{At least one mean isn't equal to all the other means}$$

ANOVA works by comparing the approximate variance *between* the samples (that is, how much the samples vary with each other) with the approximate variance *within* the samples (that is, how much on average each sample varies on its own). If the null hypothesis is true so that the population means are all equal, then the variance between the samples and the variance within the samples should be about equal.

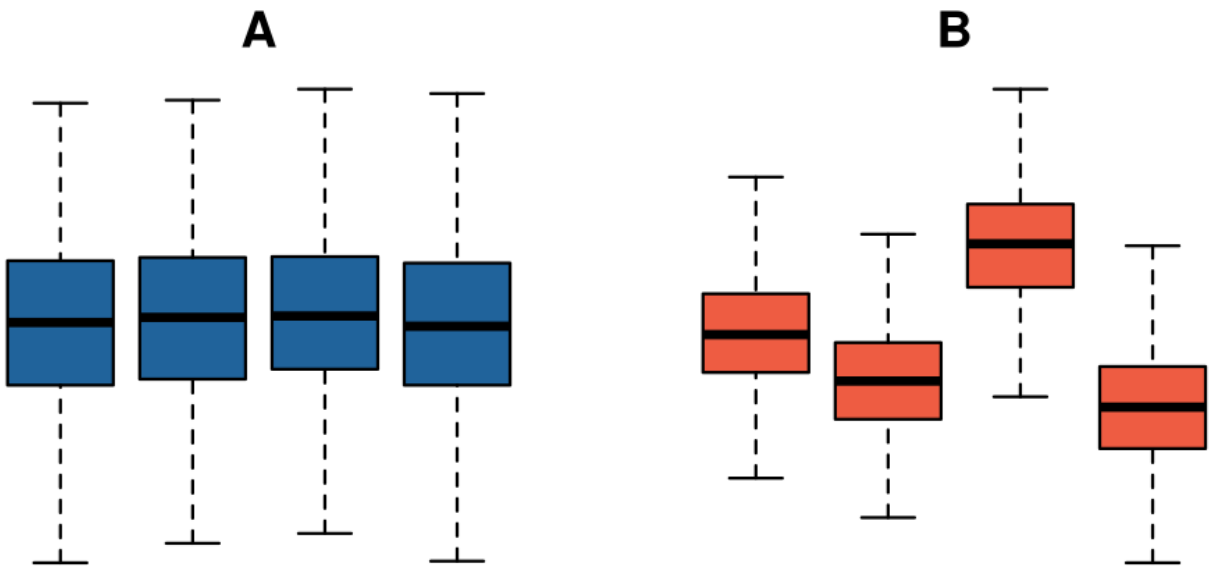



*Fig. 9.3.1* The four box plots in group A represent four populations. There is little variation between the box plots in group A. It might be reasonable to expect the populations in group A to all have the same mean. Compare this to the four populations represented by the four box plots in group B. There is significant variation between the box plots in group B. We might expect the populations in group B to have different means.

To better understand the principles behind ANOVA, consider the two groups of populations represented by the box plots in Figure 9.3.1. In group A, notice that there is little variation between the box plots of the different samples. It is reasonable to think that the underlying population means may be equal.

In comparison, there is considerable variation in group B. In fact, there is so much variation between these samples that it is very unlikely that the underlying population means between these samples are all equal.

## Notation

An ANOVA test can involve several populations and samples. To make our meaning clear, we will use the following notation in this section.

- $x_j$ are the data values in the $j$th sample.
- $n_j$ is the sample size of the $j$th sample.
- $\bar{x}_j$ is the mean of the $j$th sample. As usual, the sample mean $\bar{x}_j$ is the sum of all data values in the sample divided by sample size:

$$\bar{x}_j = \frac{\sum x_j}{n_j}.$$

- $\bar{\bar{x}}$ is the **grand mean**. It is the mean of the combined data values of all the samples. It is calculated by adding all data values from all samples together, then dividing by the total number of all the data values from all samples:

Skip to main content

$$\bar{\bar{x}} = \frac{\sum x_1 + \sum x_2 + \cdots + \sum x_k}{n_1 + n_2 + \cdots + n_k},$$

where $k$ is the total number of samples, where one sample is taken from each population.

A simple example may help to clarify the notation.

---

## Example 9.3.1

Three small samples were drawn from three populations. The sample data obtained is:

Sample $1$: 18, 21, 20, 20, 17
Sample $2$: 22, 19, 18, 21, 19, 16, 18
Sample $3$: 20, 19, 17, 21

The data values $x_1$ in sample $1$ are 18, 21, 20, 20, 17. There are $5$ data values in sample $1$, so $n_1 = 5$. We can use R to calculate the mean $\bar{x}_1$ of sample $1$.

```
x1 = c(18, 21, 20, 20, 17)
n1 = length(x1)

xbar1 = sum(x1)/n1
xbar1
```

19.2

The mean of sample $1$ is $\bar{x}_1 = 19.2$.

The data values $x_2$ in sample $2$ are 22, 19, 18, 21, 19, 16, 18. There are $7$ data values in sample $2$, so $n_2 = 7$. We can use R to calculate the mean $\bar{x}_2$ of sample $2$.

```
x2 = c(22, 19, 18, 21, 19, 16, 18)
n2 = length(x2)

xbar2 = sum(x2)/n2
xbar2
```

19

The mean of sample $2$ is $\bar{x}_2 = 19$.

The data values $x_3$ in sample $3$ are 20, 19, 17, 21. There are $4$ data values in sample $3$, so $n_3 = 4$. We can use R to calculate the mean $\bar{x}_3$ of sample $3$.

```
x3 = c(20, 19, 17, 21)
n3 = length(x3)

xbar3 = sum(x3)/n3
xbar3
```

19.25

Skip to main content

Now we can use R to calculate the grand mean $\bar{\bar{x}}$ of these three samples.

```
grandx = (sum(x1) + sum(x2) + sum(x3))/(n1 + n2 + n3)
grandx
```

19.125

The grand mean—the mean of all the data in all three samples—is $\bar{\bar{x}} = 19.125$.

---

## The Test Statistic For an ANOVA Test

To calculate the test statistic for an ANOVA test with $k$ samples (one sample from each population), we calculate two estimates of the variance.

The first estimate of the variance is $MST$, which is the **m**ean **s**quare estimate of the variance among **t**reatments. $MST$ is the variance *between* the sample means of the different samples. We calculate $MST$ using the formula

$$MST = \frac{SST}{DFT},$$

where $SST$, called the **s**um of **s**quares among **t**reatments, is the sum of squared differences between the sample means and the grand mean, and $DFT$ is the **d**egrees of **f**reedom for **t**reatments. $SST$ and $DFT$ are given by the formulas

$$\begin{aligned} SST &= n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2, \\ DFT &= k - 1. \end{aligned}$$

The second estimate of the variance is $MSE$, which is the **m**ean **s**quare estimate for **e**rrors. $MSE$ measures the variance *within* the samples. We calculate $MSE$ using the formula

$$MSE = \frac{SSE}{DFE},$$

where $SSE$, the **sum** of **s**quares for **e**rrors, is the total sum of squared differences between the data values of each sample and the sample's mean, and $DFE$ is the degrees of freedom for errors. We calculate $SSE$ and $DFE$ using the formulas

$$\begin{aligned} SSE &= \sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2 + \cdots + \sum (x_k - \bar{x}_k)^2, \\ DFE &= (n_1 - 1) + (n_2 - 1) + \cdots + (n_k - 1). \end{aligned}$$

$MST$ and $MSE$ both estimate variances, so we use an $F$-distribution to perform an ANOVA test just like we did when comparing two population variances. The $F$-statistic we use for an ANOVA test is given by

$$F = \frac{MST}{MSE}.$$

We use an $F$-distribution with numerator degrees of freedom $DFT$ and denominator degrees of freedom $DFE$ to calculate the $p$-value.

Skip to main content

If the null hypothesis is true (that is, if the means of all the populations are equal), then we would expect the variance between samples, $MST$, to be no larger than the average variance within a sample, $MSE$. In this case,

$$F = \frac{MST}{MSE} \leq 1.$$

But if the null hypothesis is not true, then we would expect the variance between samples, $MST$, to be greater than the average variance within a sample, $MSE$, in which case

$$F = \frac{MST}{MSE} > 1.$$

Because of this, an ANOVA test is always a right-tailed test.

The general steps for completing a hypothesis test have not changed:

1. State the null and alternative hypotheses.
2. Assuming the null hypothesis is true, identify the sampling distribution.
3. Find the $p$-value.
4. Draw a conclusion.

---

## Example 9.3.2

A meteorologist wishes to test if the average daily high temperature is the same for each of the past four years. She randomly selects a number of days from each year and records the daily high temperature on each selected day (in °F):

Year 1:
82, 104, 119, 56, 85, 94, 81, 106, 82, 92, 109, 71, 95, 86, 34, 89, 80, 53, 78, 99, 57, 87, 69, 98, 75, 88, 59, 104, 65, 66, 74, 73, 73, 106, 91, 89, 85, 84, 53, 81

Year 2:
80, 52, 72, 72, 43, 61, 66, 58, 118, 73, 76, 64, 81, 65, 63, 78, 72, 83, 104, 54, 69, 100, 63, 71, 38, 73, 92, 70, 86, 51, 66, 89, 74, 85

Year 3:
75, 87, 79, 96, 72, 75, 93, 84, 80, 95, 72, 56, 83, 98, 96, 54, 106, 66, 120, 77, 88, 90, 75, 98, 113, 55, 44

Year 4:
92, 68, 74, 82, 94, 44, 63, 61, 99, 71, 85, 42, 90, 94, 48, 97, 89, 65, 81, 45, 93, 81, 73, 69, 93, 75, 76, 61, 75, 102

Use an ANOVA test with a level of significance of $6\%$ to test if the average daily high temperature is the same for all four years.

### Solution

#### Step 1: State the null and alternative hypotheses.

For an ANOVA test, the null hypothesis is always that the means of all the populations are equal. The alternative hypothesis for an ANOVA test is always that at least one of the population means is not equal to the others. That is,

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a : \text{At least one mean isn't equal to the other means}$$

Step 2: Assuming the null hypothesis is true, identify the sampling distribution.

To perform an ANOVA test, we use an $F$-distribution. The numerator degrees of freedom is the degrees of freedom for treatments, $DFT$. It is one less than the number of samples. In this example, we have $k = 4$ samples, so the numerator degrees of freedom is

$$DFT = k - 1 = 4 - 1 = 3.$$

The denominator degrees of freedom is the degrees of freedom for errors, $DFE$. To calculate $DFE$, we first must know the size of each of the four samples. We will use R to count the size of each sample.

```
x1 = c(82, 104, 119, 56, 85, 94, 81, 106, 82, 92, 109, 71, 95, 86, 34, 89, 80, 53, 78, 99, 57, 87, 69, 98, 7
x2 = c(80, 52, 72, 72, 43, 61, 66, 58, 118, 73, 76, 64, 81, 65, 63, 78, 72, 83, 104, 54, 69, 100, 63, 71, 38
x3 = c(75, 87, 79, 96, 72, 75, 93, 84, 80, 95, 72, 56, 83, 98, 96, 54, 106, 66, 120, 77, 88, 90, 75, 98, 113
x4 = c(92, 68, 74, 82, 94, 44, 63, 61, 99, 71, 85, 42, 90, 94, 48, 97, 89, 65, 81, 45, 93, 81, 73, 69, 93, 7

n1 = length(x1)
n2 = length(x2)
n3 = length(x3)
n4 = length(x4)

n1
n2
n3
n4
```

40

34

27

30

So $n_1 = 40$, $n_2 = 34$, $n_3 = 27$, and $n_4 = 30$. Then the denominator degrees of freedom is

$$DFE = (n_1 - 1) + (n_2 - 1) + (n_3 - 1) + (n_4 - 1) = (40 - 1) + (34 - 1) + (27 - 1) + (30 - 1) = 127.$$

Step 3: Find the $p$-value.

To find the test statistic, we first need to calculate the sample mean of each of the four samples as well as the grand mean $\bar{\bar{x}}$.

```
xbar1 = sum(x1)/n1
xbar1

xbar2 = sum(x2)/n2
xbar2

xbar3 = sum(x3)/n3
xbar3

xbar4 = sum(x4)/n4
xbar4

grandx = (sum(x1) + sum(x2) + sum(x3) + sum(x4))/(n1 + n2 + n3 + n4)
grandx
```

81.8

72.4117647058823

82.4814814814815

76.0666666666667

78.1908396946565

We find that $\bar{x}_1 = 81.800$, $\bar{x}_2 = 72.412$, $\bar{x}_3 = 82.481$, $\bar{x}_4 = 76.067$, and the grand mean is $\bar{\bar{x}} = 78.191$.

We can now use these values to find

$$SST = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + n_3(\bar{x}_3 - \bar{\bar{x}})^2 + n_4(\bar{x}_4 - \bar{\bar{x}})^2.$$

```
SST = n1*(xbar1 - grandx)^2 + n2*(xbar2 - grandx)^2 + n3*(xbar3 - grandx)^2 + n4*(xbar4 - grandx)^2
SST
```

2288.98630610853

So $SST = 2288.986$. With $SST$ calculated and the value of $DFT$ found earlier, we are ready to find

$$MST = \frac{SST}{DFT}.$$

```
DFT = 3

MST = SST/DFT
MST
```

762.995435369511

Thus, $MST = 762.995$. This is the numerator of the test statistic.

We still need to find the denominator, $MSE$. To find $MSE$, we first calculate

$$SSE = \sum(x_1 - \bar{x}_1)^2 + \sum(x_2 - \bar{x}_2)^2 + \sum(x_3 - \bar{x}_3)^2 + \sum(x_4 - \bar{x}_4)^2.$$

```
SSE = sum( (x1 - xbar1)^2 ) + sum( (x2 - xbar2)^2 ) + sum( (x3 - xbar3)^2 ) + sum( (x4 - xbar4)^2 )
```

Skip to main content

38587.2427015251

We find that $SSE = 38587.243$. Using this value for $SSE$ together with the value for $DFE$ we found earlier, we can calculate

$$MSE = \frac{SSE}{DSE}.$$

```
DFE = 127

MSE = SSE/DFE
MSE
```

303.836556704922

So $MSE = 303.837$. This is the denominator value of the test statistic.

We now have all the pieces to calculate the test statistic,

$$F = \frac{MST}{MSE}.$$

```
F = MST/MSE
F
```

2.51120353536166

Because an ANOVA test involves so many calculations, statisticians often organize the data into a table like in Table 9.3.1.

*Table 9.3.1* A table summarizing the important calculations for the ANOVA test in this example.

| | SS | DF | MS | F |
|---|---|---|---|---|
| **Treatments** | 2288.986 | 3 | 762.995 | 2.511 |
| **Errors** | 38587.243 | 127 | 303.837 | |

Now that we have our test statistic, we are ready to find the $p$-value. Remember that an ANOVA test is always a right-tailed test.

```
1 - pf(q = F, df1 = DFT, df2 = DFE)
```

0.0616171055195667

So the $p$-value is $P(F > 2.511) = 0.062$. That is, there is about a $6.2\%$ chance that we would randomly sample data with this much variation between samples relative to the variation within samples if the population means were all actually equal.

### Step 4: Draw a conclusion.

Since the $p$-value $= 0.062 \geq 0.06 = \alpha$, we do *not* reject the null hypothesis. There is not enough evidence to conclude that the average daily high temperature for at least one year is different than average in the other years sampled.