Introductory Statistics

WORKBOOK FOR MA 206 AT FCC

by Josiah Hartley

Assistant Professor of Mathematics
Frederick Community College

Student Edition

Frederick Community College

2018

Attributions This workbook was developed for use with the OpenStax College Statistics Textbook. Some of the structure and many of the examples come from that text.

OpenStax College, *Introductory Statistics*. OpenStax College. 19 September 2013. <http://cnx.org/content/col11562/latest/>

Thanks This work was supported by a summer grant from FCC in 2016, and I would like to thank the college administration for their support.

Contents

Chapter 1

Sampling and Data

What is Statistics? Broadly speaking, the study of statistics is the study of how to make sense of data.

Example: US Census Every 10 years, the U.S. Census Bureau undertakes the enormous task of gathering all kinds of data on people residing in the country. In between the huge national surveys, the Census Bureau collects data with smaller surveys like the American Community Survey.

What would you do? If your job was to collect a national census, and your results looked like the table below, what would you want to do with this?

	Age	Sex	Primary Language	Working	Earnings	Owns a Car	⋯
Household 1							
Person 1	54	M	English	Yes	\$89,500	Yes	⋯
Person 2	51	F	English	No	N/A	Yes	⋯
Person 3	19	F	English	Yes	\$23,000	Yes	⋯
Household 1							
Person 1	78	M	Spanish	No	\$32,800	Yes	⋯
Person 2	82	F	Spanish	No	\$28,350	No	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

SECTION 1.1 Definitions and Key Terms

Two Goals:

1. To organize the data in such a way that it makes sense.
2. To set it up so that someone with a question could come along later and find an answer to their question.

In other words, you'd want to clearly display the data so that you can explain it to someone who has never seen it before, but you also want to have a way that someone studying a particular topic (like car ownership in the U.S.) could "ask" the data a question and get an answer.

These two goals are related to the two sides of statistics: **descriptive statistics** and **inferential statistics**.

Descriptive vs. Inferential Statistics

Descriptive Statistics: Organizing and summarizing data.

There are many ways to do this, including the use of graphs and charts, but the goal is always the same: to give the reader a clear, concise idea of what the data looks like without having to show them something like the whole table above.

Inferential Statistics: Drawing conclusions from the data.

For instance, we might take a poll to compare two political candidates, and we need to know whether the results we get are valid, or whether they were a fluke.

Definitions

ex: entire US

Population: The group that we are interested in.

ex: randomly pick 100 households

Sample: The group that we can actually feasibly study. The census mentioned above is a rare example in which the entire population is studied (this is incredibly expensive and time-consuming). More often, a reasonably-sized sample is selected and studied.

If we get a **representative sample**, we assume that the population looks similar enough to the sample that by studying the sample, we can get a good idea of what the population looks like (if you want to know how a pot of soup tastes, you only have to take one sip).

ex: average salary of every US worker

Parameter: A number that describes something about the **population**.

Statistic: A number that describes something about the **sample**.

ex: average salary of every worker in our sample

Every parameter has a corresponding statistic; since we're assuming that we can't study the entire population, we get the statistic from the sample, and we assume that the statistic is a good estimate for the parameter.

In general, when we're dealing with the sample, we're doing descriptive statistics (**describing** the sample) and whenever we're using the sample to draw conclusions (another word for **inferences**) about the population, we're doing inferential statistics.

Variables

Variable: Something that we record about our sample. After we record it (collecting data like in the table at the beginning of the chapter) we can start to describe it by taking the average or drawing a graph or something.

ex: salary

Numerical (or quantitative) Variables: Variables that we find by measuring or counting.

ex: number of cars in a household or age of household members

Discrete Quantitative Variables: Numerical variables that come from counting. They are limited to specific values.

For instance, "number of children" is a discrete variable, because one cannot have 3.14 children; the answer will always be 0, 1, 2, etc.

Continuous Quantitative Variables: Numerical variables that come from measuring. They can be any number in a valid range.

For instance, "height" is a continuous variable, because one's height can be any value (within a reasonable range), provided that we can measure as precisely as we want.

Categorical (or qualitative) Variables: Variables that divide people or things into categories.

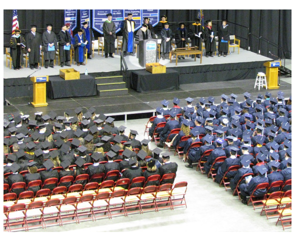
ex: sex or political affiliation

Note that categorical variables can also by numerical; think of your student ID number. Your ID number categorizes you; it doesn't measure or count something about you. You wouldn't think about taking the average ID number of students, because that would be a meaningless result.

Summary of Key Terms

1. **Population:** The group that we are interested in.
2. **Sample:** The subset of the population that we can feasibly study.
3. **Parameter:** A number that describes something about a population variable.
4. **Statistic:** A number that describes something about a sample variable.
5. **Variable:** Something that we record about our sample.
6. **Quantitative variable:** A numerical variable that we find by counting (discrete) or measuring (continuous).
7. **Qualitative variable:** A variable that divides people or things into categories.

EXAMPLE 1 USING THESE KEY TERMS



A study was conducted at a local college to analyze the average cumulative GPA's of students who graduated last year. Fill in the letter of the phrase that best describes each of the items below.

1. ____ Population
2. ____ Statistic
3. ____ Parameter
4. ____ Sample
5. ____ Variable

(a) all students who attended the college last year

(b) the cumulative GPA of one student who graduated from the college last year

(c) a group of students who graduated from the college last year, randomly selected

(d) the average cumulative GPA of students who graduated from the college last year

(e) all students who graduated from the college last year

(f) the average cumulative GPA of students in the study who graduated from the college last year

Solution 1. e; 2. f; 3. d; 4. c; 5. b

SECTION 1.2 Sampling Methods

Good sampling is one of the most important parts of a statistical study. Remember, the key is that we want the sample to be **representative** of the population.

Random Sampling

To get a representative sample, we select our sample randomly.

Random Sampling:

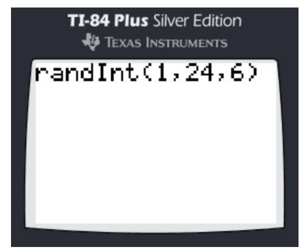
1. **Simple Random Sampling:**

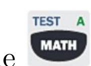
2. **Stratified Sampling:**

3. **Cluster Sampling:**

4. **Systematic Sampling:**

5. **Convenience Sampling:**



On your graphing calculator, press the 

button, scroll over to the PRB menu, and select `5:randInt(` to access the random number generator. If you enter three numbers, separated by commas, as shown, the calculator will return 6 numbers between 1 and 24. If you just enter two numbers, the calculator will return one number between those bounds.

EXAMPLE 1 QUIZ SCORE SAMPLES

Use the random number generator on your calculator to generate different types of samples from the data below. Find the average score for each sample and compare your results with your classmates.

This table displays six sets of quiz scores (out of 10 points) for an elementary statistics class.

# 1	# 2	# 3	# 4	# 5	# 6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

1. Create a stratified sample of 12 quiz scores, using the columns as the strata.

   Data: Average Score:

2. Create a cluster sample by picking two of the rows.

   Data: Average Score:

3. Create a simple random sample of 12 quiz scores.

   Data: Average Score:

4. Create a systematic sample of 12 quiz scores.

   Data: Average Score:

5. Create a convenience sample of 12 quiz scores.

   Data: Average Score:

SAMPLING METHODS EXAMPLE 2

Determine the type of sampling used in each of the following scenarios.

1. A soccer coach selects six players from a group of boys aged eight to ten, seven players from a group of boys aged 11 to 12, and three players from a group of boys aged 13 to 14 to form a recreational soccer team.

   Sampling Method:

2. A pollster interviews all human resource personnel in five different high tech companies.

   Sampling Method:

3. A high school educational counselor interviews 50 female teachers and 50 male teachers.

   Sampling Method:

4. A medical researcher interviews every third cancer patient from a list of cancer patients at a local hospital.

   Sampling Method:

5. A high school counselor uses a computer to generate 50 random numbers and then picks students whose names correspond to the numbers.

   Sampling Method:

6. A student interviews classmates in his algebra class to determine how many pairs of jeans a student at his school owns, on the average.

   Sampling Method:

REPRESENTATIVE SAMPLES EXAMPLE 3

Decide whether each of the following sampling methods is likely to produce a representative sample.

1. To find the average annual income of all adults in the United States, sample representatives in the US Congress.

2. To find out the most popular cereal among children under the age of 10, stand outside a large supermarket one day and poll every twentieth child under the age of 10 who enters the supermarket.

Sample Size

If two people take samples from the same group, their samples will almost certainly be different. This doesn't mean that one is right and one is wrong, though. There is simply some natural variability in samples.

Bigger Samples are *Often* Better One way to reduce this natural variability is to take larger samples, where the variation will get drowned out.

- For national polls, somewhere between 1000 and 2000 people is usually considered a big enough sample.

Be Careful: Just having a big sample doesn't guarantee good results.

In general, self-selected samples (or volunteer samples) are not representative of the population. For this reason, surveys with voluntary responses are not reliable. People who volunteer their opinion for online reviews, for instance, tend to be strongly positive or negative; the voluntary sample misses everyone in the middle who doesn't have a strong opinion.

The most famous example of this comes from the 1936 presidential election, where the incumbent Democrat, Franklin D. Roosevelt, was challenged by the Republican governor of Kansas, Alf Landon. The *Literary Digest*, a weekly magazine, boasted that it had correctly predicted the results of the last 4 elections by sending out questionnaires to its huge sample of readers. In 1936, the *Digest* sent out 10 million questionnaires and received over 2 million responses, predicting that Landon would unseat Roosevelt with a handy victory. When Election Day came, though, Roosevelt received over 60% of the popular vote, carrying every state except for Maine and Vermont (including Landon's home state). It was one of the most lopsided victories in U.S. history. The reason for the failure of this poll was largely based on the voluntary response nature–those who responded were more likely to be those who were unhappy with the current administration; people who were happy with Roosevelt's programs had no incentive to fill out the questionnaire and send it in.

Largely due to this failure and embarrassment, the *Literary Digest* folded within a few years. In contrast, a young pollster named George Gallup (whose name is borne by the Gallup polls today) made his name in the 1936 election by correctly predicting the winner with a much smaller, carefully chosen sample.

Other Considerations

Besides a small or biased sample, there are other conditions that can throw off a statistical study, such as

-
-
-
-
-
-

SECTION 1.4 Experimental Design and Ethics

Observational Study:

Experimental Study:

Experimental Study Terms

- **Experimental unit:**
- **Response:**
- **Explanatory variable:**
- **Treatment:**
- **Lurking variable:**
- **Control group:**
- **Placebo:**
- **Blinding:**

EXPERIMENTAL DESIGN EXAMPLE 1

Researchers want to investigate whether taking aspirin regularly reduces the risk of heart attack. Four hundred men between the ages of 50 and 84 are recruited as participants. The men are divided randomly into two groups: one group will take aspirin, and the other group will take a placebo. Each man takes one pill each day for three years. At the end of the study, researchers count the number of men in each group who have had heart attacks.

Identify each of the following in this study:

- **Population:**
- **Sample:**
- **Experimental units:**
- **Control group:**
- **Explanatory variable:**
- **Treatments:**
- **Response variable:**

Homework 1 Name:

For problems 1–2, identify each of the given components of the study.

1. A fitness center is interested in the mean amount of time a client spends exercises in the center each week.

(a) Population:

(b) Sample:

(c) Parameter:

(d) Statistic:

(e) Variable:

2. A marriage counselor is interested in the proportion of clients she counsels who stay married.

(a) Population:

(b) Sample:

(c) Parameter:

(d) Statistic:

(e) Variable:

For problems 3–8, determine the type of sampling used.

3. A high school principal polls 50 freshmen, 50 sophomores, 50 juniors, and 50 seniors regarding policy changes for after school activities.

Sampling Method:

4. To check their accuracy, the Census Bureau draws a sample of several city blocks and recounts everyone in those blocks.

Sampling Method:

5. A pollster walks around a busy shopping mall and asks people passing by how often they shop at the mall.

Sampling Method:

6. A restaurant samples 100 sales from the past week by numbering all their receipts, generating 100 random numbers, and picking the receipts that correspond to those numbers.

Sampling Method:

7. Police at a DUI checkpoint stop every tenth car to check whether the driver is sober.

Sampling Method:

8. The provost at a university wants to know how a particular policy is affecting faculty, so she randomly selects 3 members of each department to survey.

Sampling Method:

For problems 9–10, determine whether an observational study or experimental study is more appropriate. Explain your answer.

9. A scientist wants to determine whether people who live at higher altitudes are more or less likely to develop respiratory problems than people who live at lower altitudes.

10. A scientist wants to determine whether a certain nasal spray is effective in reducing allergic reactions to pollen.

11. A recent study compared the heart rates of 19 infants born to nonsmoking mothers with those of 17 infants born to mothers who smoked an average of 15 cigarettes a day while pregnant and after giving birth. The heart rates of the infants at one year of age were 20% slower on the average for the smoking mothers.

(a) What were the experimental units?

(b) What was the response?

(c) What were the treatments?

(d) Give one suggestion that you would make to improve this study, as it is stated.

SECTION 1.3 Frequency Tables

Frequency tables simply count how many times each data value occurs and list this count.

Qualitative Data

STUDENT GRADES EXAMPLE 1

Suppose a group of students earned the following final grades:

B, C, A, B, B, D, C, C, C, F, A, C, B, B, B, C, B, D

The frequency table for this data would look like the following.

Grade	Frequency
A	
B	
C	
D	
F	

Relative Frequency: The proportion of the whole group that is in each category.

Grade	Frequency	Relative Frequency
A		
B		
C		
D		
F		

Quantitative Data

EXAMPLE 2 IPADS SOLD

A store tracked how many iPads were sold each day for fifty days, and their data is below.

4	2	3	2	5	5	1	3	3	2
3	2	2	3	2	2	2	3	0	1
3	1	1	5	4	1	2	4	3	5
2	0	0	3	2	3	3	3	2	2
0	4	2	4	3	1	1	4	0	1

The frequency distribution looks like the following.

Value	Frequency	Relative Frequency
0	5	
1	8	
2	14	
3	13	
4	6	
5	4	

EXAMPLE 3 WORKING STUDENTS

Twenty students were asked how many hours they worked per day. Their responses are as follows:

5, 6, 3, 3, 2, 4, 7, 5, 2, 3, 5, 6, 5, 4, 4, 3, 5, 2, 5, 3

Construct a frequency table (including a relative frequency column) to organize this data.

Hours Worked	Frequency	Relative Frequency

Grouped Frequency Tables

Sometimes it could be inconvenient to have a separate row for each data value. For instance, consider the following wait times on a customer service line (measured in minutes):

0.6	1.2	1.3	2.5	2.8	3.2	3.2	3.5	3.8	3.9
3.9	4.4	4.4	4.6	4.6	4.6	4.8	4.9	5.1	5.2
5.4	5.5	5.8	6.1	6.4	6.9	7.0	8.0	8.1	8.1
8.3	8.7	9.0	9.3	9.3	9.5	9.5	9.7	9.8	9.9
10.2	10.5	10.9	12.2	12.5	13.1	13.3	13.6	14.4	17.4

If we had a separate category for each time, there would be a bunch of categories with a frequency of 1 or 2. Instead, we'll construct a **grouped frequency table**:

-

-

-

-

-

EXAMPLE 4 CUSTOMER SERVICE WAIT TIMES

Construct a grouped frequency table for the data above, using a class width of 3 minutes.

Number of Minutes	Frequency	Relative Frequency

Chapter

2 Descriptive Statistics

We've already started doing descriptive statistics, with frequency tables in the last section. The point of descriptive statistics is to organize and present data in a way that is easy to read and interpret.

In the first part of this chapter, we will cover visual summaries of data, including histograms, bar graphs, stem-and-leaf plots, and line and time series graphs. Then, in the next part, we'll see numerical summaries of data.

There, we'll measure four things:

-
-
-
-

SECTION 2.1 Bar Graphs and Histograms

Bar graphs and histograms are pretty similar; the differences are fairly subtle, but in short, a histogram is a specific type of bar graph.

Bar Graphs:

-
-

Histograms:

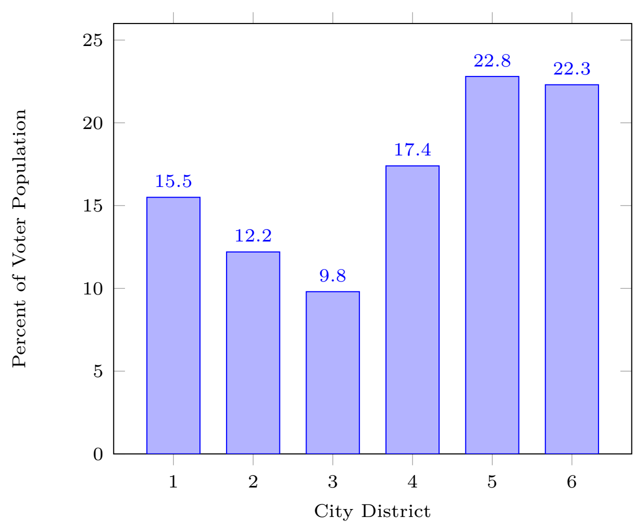
-
-
-
-
-

BAR GRAPH: REGISTERED VOTERS EXAMPLE 1

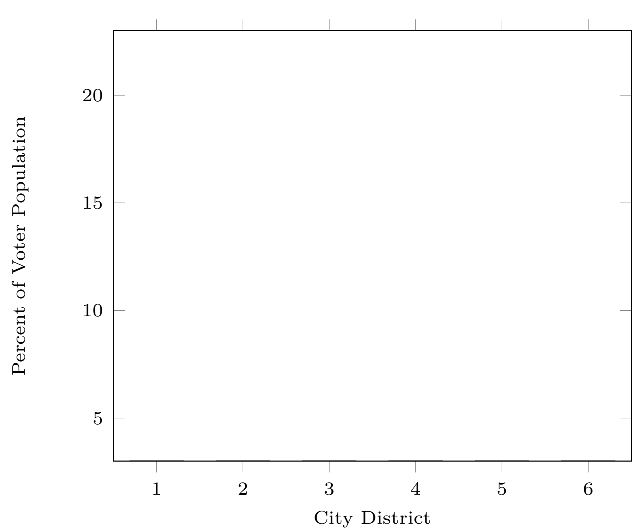
A city is broken down into six districts, and the following table shows the percentage of the total registered voter population that lives in each district, as well as the percentage total of the entire population.

District	Registered Voter Population	Overall City Population
1	15.5%	19.4%
2	12.2%	15.6%
3	9.8%	9.0%
4	17.4%	18.5%
5	22.8%	20.7%
6	22.3%	16.8%

The bar graph below shows the registered voter population by district.

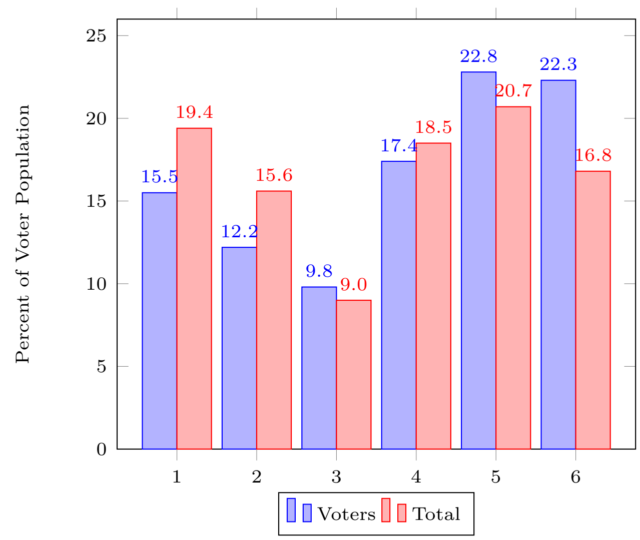


Construct a bar graph that shows the percentage of total population of the city by district.

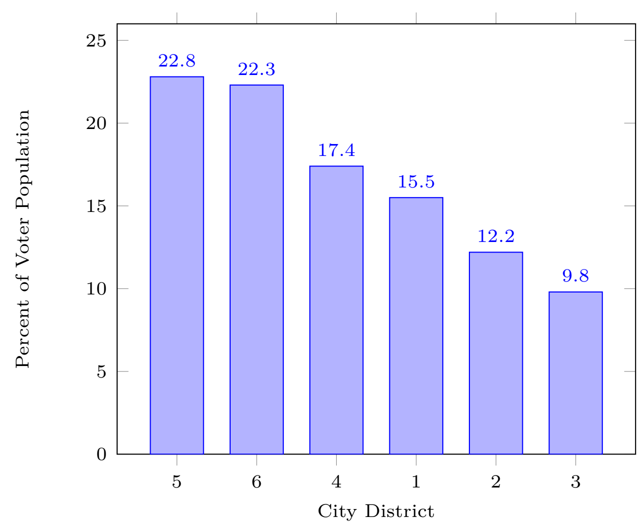


There are a couple of other things that we can do with bar graphs like those in the last example:

1. We could make a combined bar graph.



2. We could order the bars in descending order; this is called a **Pareto chart** (technically, a Pareto chart is a bit more involved, but that's the main feature). This is meant to highlight the largest category.



HISTOGRAM: IPADS SOLD — EXAMPLE 2

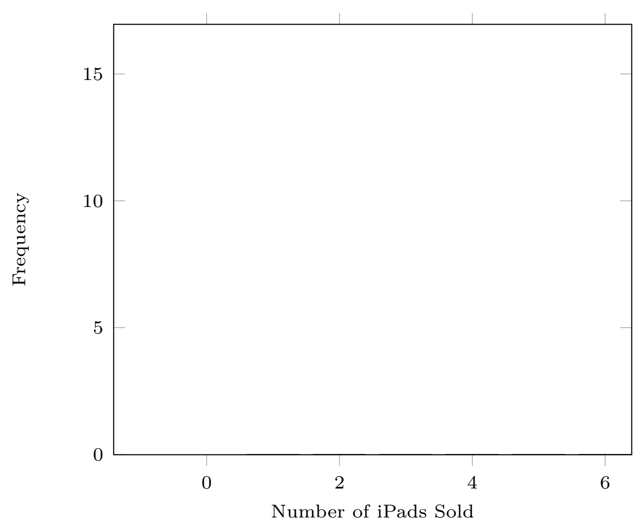
Recall the following data on the number of iPads sold in a store over the last 50 days.

4	2	3	2	5	5	1	3	3	2
3	2	2	3	2	2	2	3	0	1
3	1	1	5	4	1	2	4	3	5
2	0	0	3	2	3	3	3	2	2
0	4	2	4	3	1	1	4	0	1

Remember, the frequency table looked like the following.

Value	Frequency	Relative Frequency
0	5	0.10
1	8	0.16
2	14	0.28
3	13	0.26
4	6	0.12
5	4	0.08

The histogram for this data looks like this:



EXAMPLE 3 HISTOGRAM: CUSTOMER SERVICE TIMES

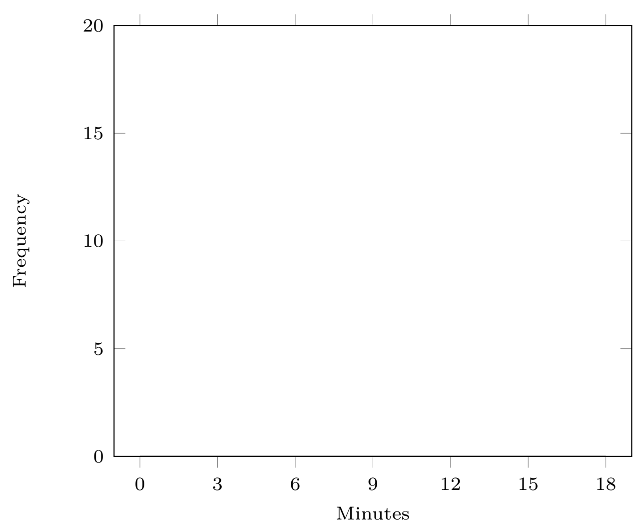
Construct a histogram to display the following customer service data (wait time in minutes).

0.6	1.2	1.3	2.5	2.8	3.2	3.2	3.5	3.8	3.9
3.9	4.4	4.4	4.6	4.6	4.6	4.8	4.9	5.1	5.2
5.4	5.5	5.8	6.1	6.4	6.9	7.0	8.0	8.1	8.1
8.3	8.7	9.0	9.3	9.3	9.5	9.5	9.7	9.8	9.9
10.2	10.5	10.9	12.2	12.5	13.1	13.3	13.6	14.4	17.4

We already found the frequency table:

Number of Minutes	Frequency	Relative Frequency
$0 \leq t < 3$	5	0.10
$3 \leq t < 6$	18	0.36
$6 \leq t < 9$	9	0.18
$9 \leq t < 12$	11	0.22
$12 \leq t < 15$	6	0.12
$15 \leq t < 18$	1	0.02

Now draw the histogram:



Drawing Histograms on TI Graphing Calculators

1. Enter data into `L1` by pressing 

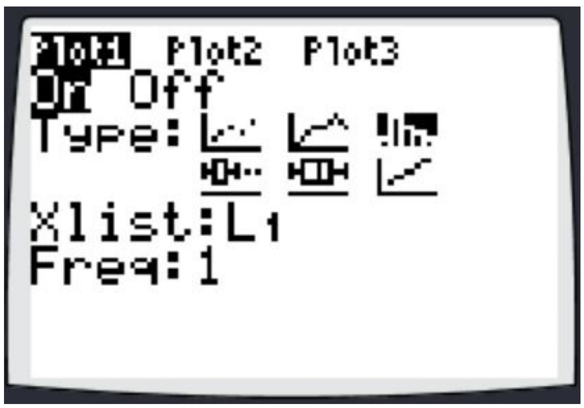
 1:Edit...

2. Press 

 to access the `STAT PLOT` menu. Press 

 to open the first `STAT PLOT`. You'll see the following menu.



3. Select the icon for a histogram and press 

4. To see the graph, you'll probably need to adjust the window. Press 

 to access the following menu.

Adjust `Xmin`, `Xmax`, `Ymin`, and `Ymax` to get the histogram in the picture. To adjust the width of the bars on the histogram (the class width), change `Xscl`.

SECTION 2.2 Other Graphs

We'll look at three other graphical summaries of data:

1. Stem-and-leaf plots
2. Scatter plots
3. Time series plots

Stem-and-Leaf Plots

-
-
-

EXAMPLE 1 STEM-AND-LEAF PLOT: TEST SCORES

The test scores for a class look like the following.

81	86	78	80	81	82	92	90	79	83	84	95
84	79	80	83	79	87	84	80	85	88	80	78

The stems for these data points are the first digits, and the leaves are the second digits.

Stems	Leaves
7	
8	
9	

EXAMPLE 2 STEM-AND-LEAF PLOT: DISTANCE TO SUPERMARKET

The distances (in km) from a particular home to the closest supermarkets are shown below.

1.1	1.5	2.3	2.5	2.7	3.2	3.3
3.3	3.5	3.8	4.0	4.2	4.5	4.5
4.7	4.8	5.5	5.6	6.5	6.7	12.3

Construct a stem-and-leaf plot for this data, noting that the leaves are the digits to the right of the decimal.

Stems	Leaves

Scatter Plots

-
-

SCATTER PLOT: TV PRICE EXAMPLE 3

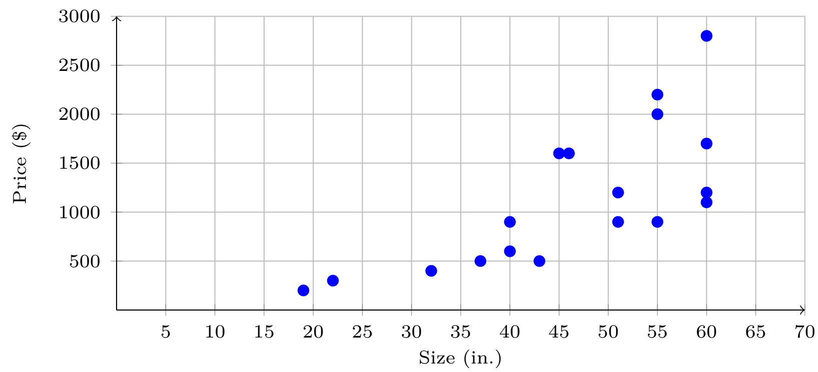
The following table shows, for a sample of Samsung LCD TVs, their size and their price.

Size (in.)	Price ($)	Size (in.)	Price ($)
43	500	60	1200
55	900	45	1600
51	900	19	200
32	400	55	2200
51	1200	60	1700
37	500	55	2000
60	2800	22	300
60	1100	40	600
46	1600	40	900

Note: We could pick either variable to be x, but we typically let x be the explanatory–or predictor–variable; in that example, it makes more sense to say that the size of a TV predicts its price than to say that the price of a TV predicts its size.

It turns out, though, that if we switch x and y, nothing crucial really changes.

To construct a scatter plot, put the size along the x axis, put price along the y axis, and plot a point for each TV.

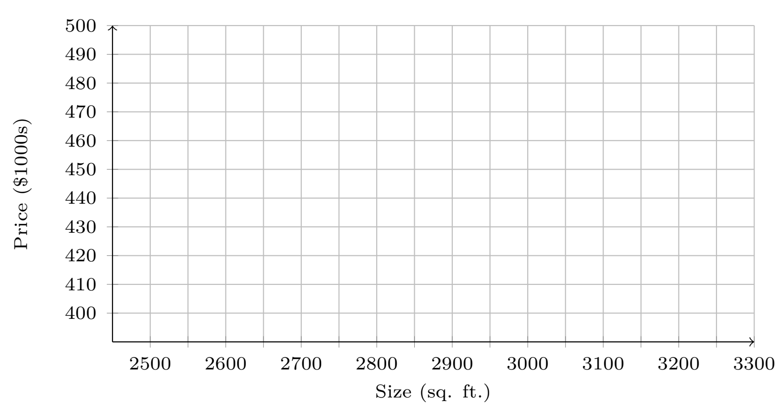


EXAMPLE 4 SCATTER PLOT: HOME PRICES

The following table shows a sample of homes on the market, recording their size in square feet and their price in thousands of dollars (so for instance, the first home is selling for $400,000).

Size (sq. ft.)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

Construct a scatter plot for this data.



Time Series Plots

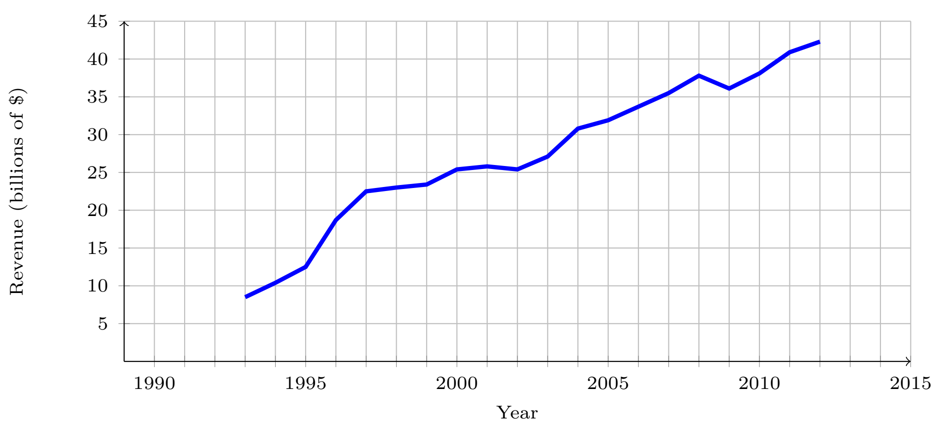
-
-
-

TIME SERIES: DISNEY REVENUE EXAMPLE 5

The following table records the annual revenue for Disney from 1993 to 2012.

Year	Revenue (billions of $)	Year	Revenue (billions of $)
1993	8.5	2003	27.1
1994	10.4	2004	30.8
1995	12.5	2005	31.9
1996	18.7	2006	33.7
1997	22.5	2007	35.5
1998	23.0	2008	37.8
1999	23.4	2009	36.1
2000	25.4	2010	38.1
2001	25.8	2011	40.9
2002	25.4	2012	42.3

The time series plot for this data looks like the following.



Homework 2 Name:

1. Nineteen immigrants to the U.S. were asked how many years, to the nearest year, they have lived in the U.S. The data are as follows:

$$2,\ 5,\ 7,\ 2,\ 2,\ 10,\ 20,\ 15,\ 0,$$
$$7,\ 0,\ 20,\ 5,\ 12,\ 15,\ 12,\ 4,\ 5,\ 10.$$

Build a frequency table (in the space to the right) for this data set, including a relative frequency column.

---

2. The following data are the manufacturer's suggested retail prices (in thousands of dollars) for models and styles of 2008 BMW cars. Construct a (grouped) frequency table, using a class width of 10, with 20 as the lower class limit for the first class.

28.8	52.0	34.5	49.5	53.8	46.2	50.4
34.9	58.8	36.4	76.6	56.5	54.8	52.4
33.1	54.4	35.6	83.7	65.0	52.5	39.1
32.7	37.4	76.8	83.9	63.0	44.6	34.6
41.2	79.9	100.3	36.7	46.8	39.3	43.0
124.1	106.1	42.7	49.8	41.2	43.5	115.0

---

3. The students in Ms. Ramirez's math class have birthdays in each of the four seasons. The table below shows the four seasons and the number of students who have birthdays in each season. Construct a bar graph to represent the data.

Season	Number of Students
Spring	8
Summer	9
Fall	11
Winter	6

4. Fifty part-time students were asked how many courses they are taking this term. The incomplete results are shown below. Fill in the blanks in the table.

# of Courses	Frequency	Relative Frequency
1	30	0.6
2	15	
3		

5. Several children were asked how many TV shows they watch each day.

Number of Shows	Frequency
0	12
1	18
2	36
3	7
4	2

What percentage of children watch two or more shows each day?

6. Construct a histogram for the following.

Pulse Rates for Women	Frequency
60–69	12
70–79	14
80–89	11
90–99	1
100–109	1
110–119	0
120–129	1

7. Construct a histogram for the following.

Speed in a 30 MPH Zone	Frequency
42–45	25
46–49	14
50–53	7
54–57	3
58–61	1

8. The mpg ratings for 30 cars are shown below. Construct a stem-and-leaf plot for this data set.

19, 19, 19, 20, 21, 21, 25, 25, 25, 26,
26, 26, 28, 29, 31, 31, 32, 32, 33, 34,
35, 36, 37, 37, 38, 38, 38, 38, 41, 43

9. Construct a scatter plot for the data given below.

Study Hours	Exam Score
5	84
7	92
4.5	82
7	80
8	90
6.5	78
5.5	74
4	75

---

10. The following table presents the amount spent on national defense by the U.S. government every other year from 1995 through 2009 (adjusted for inflation). Construct a time-series plot for this data set.

Year	Spending
1995	366.9
1997	351.8
1999	348.5
2001	370.1
2003	472.8
2005	546.4
2007	574.6
2009	675.1

SECTION 2.3 Measures of the Location of Data

Suppose you're applying for law school, so you take the LSAT (the Law School Admission Test), and you score 155. If that's all you know, that number is pretty meaningless. What you **really** want to know is how well you did *relative to everyone else who took the test.* If I told you that you scored better than 63% of people who took the LSAT, that would be a much better indication of your success.

Therefore, we often want to know where a particular data point (your LSAT score, your baby's weight, etc.) falls in the data set. To do this, we have several **measures of position**, two of which we'll look at in this section:

1.

2.

Percentiles

Percentiles are exactly what was described above: if you scored 155 on the LSAT, you did better than 63% of test-takers, so we would say that you were in the 63rd percentile for the test.

Percentiles

Definition:

PERCENTILES: SLEEP TIME — EXAMPLE 1

Fifty college students were asked how much sleep they get per school night (rounded to the nearest hour). The following frequency table records their results.

Hours of Sleep	Frequency
4	2
5	5
6	7
7	12
8	14
9	7
10	3

1. Find the 28th percentile.

2. Find the 80th percentile (for them).

3. Find the 40th percentile (for them).

Quartiles

Quartiles are nothing more than specific percentiles that split the data into quarters:

EXAMPLE 2 PHYSICS EXAM SCORES

A physics class earned the following scores on an exam:

47	48	53	56	57	58	60	61	61	62
63	64	71	72	74	75	76	82	89	95

(note that the scores are already ordered; if they weren't, we would have to start by ordering them)

1. Find the first quartile.

2. Find the median (for them).

3. Find the third quartile (for them).

Five Number Summary

The five number summary summarizes a data set by giving the following five statistics:

$$\text{Min}, Q_1, \text{Median}, Q_3, \text{Max}$$

We'll use this in the next section to draw **Box Plots**.

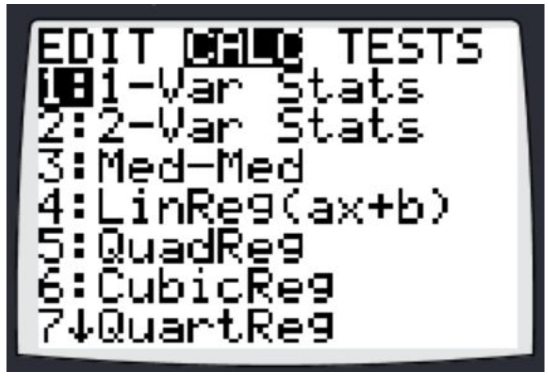
PHYSICS EXAM SCORES EXAMPLE 3

The five number summary for the test score data set in the previous example is

Using Your Calculator

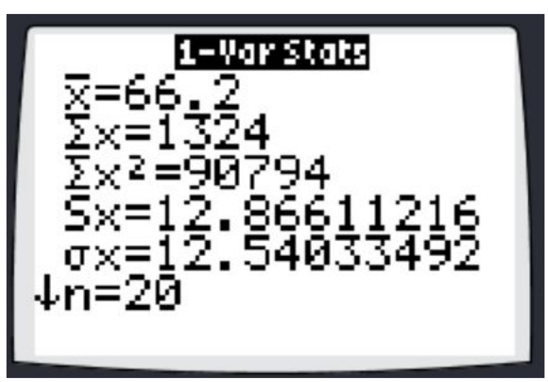
To find the five number summary on your graphing calculator, start by entering the data into L1 (press  1:Edit... to access the data).

Then press  again and scroll over to the CALC menu along the top.



Select the first option: 1-Var-Stats and you'll see the following:

Leave everything as is (if you entered your data into a different list than L1, you could select that here; if your data as a frequency table, you could select which list to use as the frequency list). Select Calculate and you'll get something like



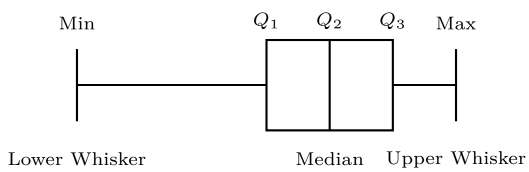
There's a lot of information here, but if you scroll down, you'll find minX, Q1, Med, Q3, and maxX, the five number summary.

SECTION 2.4 Box Plots

A box plot is nothing more than a graph that shows the five number summary for a data set. Its value is in showing where the data is clustered and where it is spread out.

Basic Box Plots

The following is a simple box plot:



(of course, in an example, we'll include a labeled axis for scale)

Remember, each segment of this box plot contains a quarter of the data.

BOX PLOT EXAMPLE 1

The following data are the number of pages in 40 books on a shelf.

136	140	178	190	205	215	217	218	232	234
240	255	270	275	290	301	303	315	317	318
326	333	343	349	360	369	377	388	391	392
398	400	402	405	408	422	429	450	475	512

Construct a box plot.

EXAMPLE 2 COMPARING TWO GROUPS

Test scores for a statistics class held during the day are

99	56	78	55	32	90	80	81	56	59
45	77	85	84	70	72	68	32	79	90

Test scores for a statistics class held during the evening are

98	78	68	83	81	89	88	76	65	45
98	90	80	84	85	79	78	98	90	25

Create side by side box plots to compare these two classes.

Day Class	Evening Class
Min = 32	Min = 25
$Q_1 = 56$	$Q_1 = 77$
Med = 74.5	Med = 82
$Q_3 = 82.5$	$Q_3 = 89.5$
Max = 99	Max = 98

Box Plots with Outliers

Outliers are unusual data points. Sometimes outliers are the result of mistakes in recording the data; sometimes there really are out-of-the-ordinary observations.

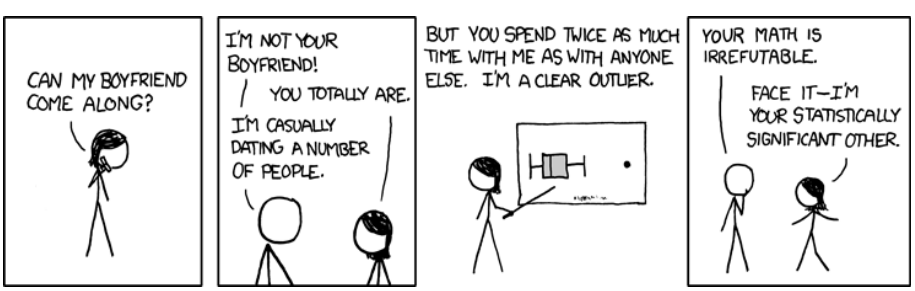
There is a rule of thumb using box plots that can identify outliers.

Outliers

Interquartile Range:

Outliers:

Sometimes box plots are drawn like in the previous examples, with no mention of outliers (the fences extend all the way to the minimum and maximum). Other times, box plots show the outliers.



xkcd.com

EXAMPLE 3 BOX PLOT WITH OUTLIERS

Test scores for a statistics class held during the evening are

98	78	68	83	81	89	88	76	65	45
98	90	80	84	85	79	78	98	90	25

$$\begin{aligned} &\text{Min} = 25 \\ &Q_1 = 77 \\ &\text{Med} = 82 \\ &Q_3 = 89.5 \\ &\text{Max} = 98 \end{aligned}$$

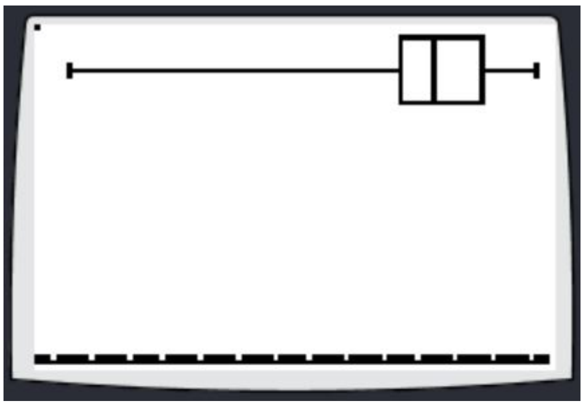
Drawing Box Plots with TI Calculator

After entering the data, press 

to access the `STAT PLOT` menu. Turn the first plot `ON`, then select one of the two types of box plots (one with outliers and one without.

Then press 

and adjust the window to the appropriate maximum and minimum x values.



Homework 3 Name:

1. Listed below are 32 ages for Academy Award winning best actors in order from smallest to largest.

18	18	21	22	25	26	27	29
30	31	31	33	36	37	37	41
42	47	52	55	57	58	62	64
67	69	71	72	73	74	76	77

(a) Find the 37th percentile. ________

(b) Find the 72nd percentile. ________

(c) Find the median. ________

2. Consider the following questions.

(a) When considering finish times in a race, is it more desirable to have a time with a high or low percentile? ________

(b) The 20th percentile of run times in a particular race is 5.2 minutes. Write a sentence interpreting the 20th percentile in this context.

(c) A cyclist in the 90th percentile of a race finished in 1 hour and 12 minutes. Is he among the fastest or slowest racers? ________

3. On an exam, would it be more desirable to earn a grade with a low or a high percentile? ________

4. Jesse was ranked 37th in his graduating class of 180 students. At what percentile is Jesse's ranking? ________

5. Suppose that you are buying a house. You and your realtor have determined that the most you can pay for a house is $240,000. In the market in which you are buying, this price is in the 34th percentile. Can you afford 34% or 66% of the houses in this market? ________

6. The University of California has two criteria used to set admission standards for freshmen to be admitted to a college in the UC system:

(a) The UC system seeks to admit the top 12% of high school students in the state. What percentile does this correspond to? ________

(b) Students whose GPAs are at or above the 96th percentile within their high school are eligible even if they are not in the top 12% of the students in the entire state. What percentage of students from each high school fit this criteria? ________

7. Sixty-five randomly selected car salespersons were asked the number of cars they generally sell in one week. Fourteen of them answered that they sell three cars, nineteen sell four cars, twelve sell five cars, nine sell six cars, and eleven sell seven cars.

(a) Find and list the five-number summary.

(b) Construct a box plot for this data.

(c) Are there any outliers? If so, list them. Show your work.

SECTION 2.5 Measures of Center

So far in this chapter, we've summarized data visually. Now we come to summarizing data numerically. We'll start by summarizing where the data is centered, using three measures:

1.

2.

3.

Mean, Median, and Mode

Mean or Average:

For a sample of size n,

For a population of size N,

Median:

The median is at position

Mode:

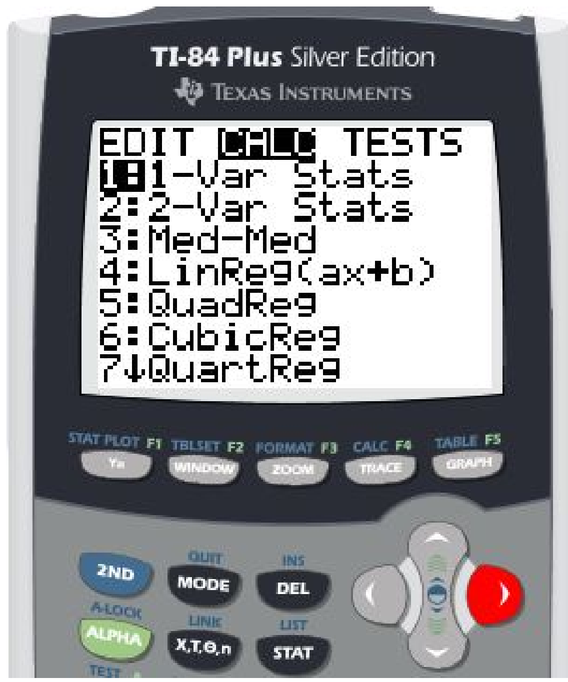
The easiest way to find the mean and median is by using your calculator:

1. Enter the data by pressing 

`1:Edit...`

2. Press 

again and scroll over to the `CALC` menu.

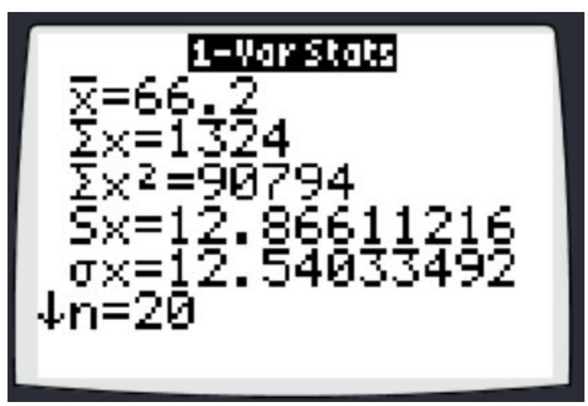


3. Press 

to select `1-Var Stats`

4. Scroll down to `Calculate` on the menu and press 

5. Scroll through the resulting list to find $\overline{x}$ and `Med`.



From a Frequency Table

MEAN AND MEDIAN EXAMPLE 1

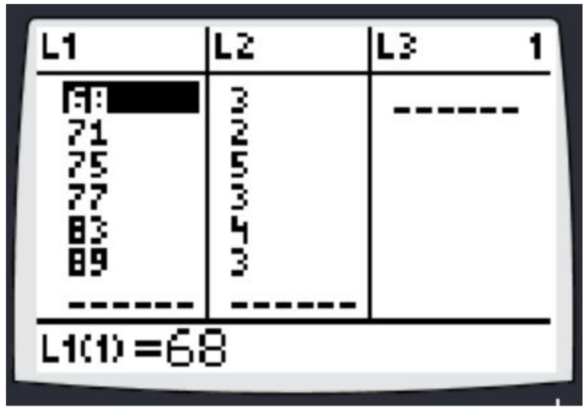
Calculate the mean and median of the data set summarized by the following frequency table.

Score	Frequency
68	3
71	2
75	5
77	3
83	4
89	3

Mean

Median

This can be done using calculator as well. Enter the data as a frequency table:



On the TI-83, when you hit enter on `1-Var Stats`, you'll be taken back to the home screen with that entered. Type in `L1, L2` by pressing `2nd` $\longrightarrow$ `1` $\longrightarrow$ `,` $\longrightarrow$ `2nd` $\longrightarrow$ `2`

Then when you select `1-Var Stats`, enter `L2` as the `FreqList` (using `2nd` $\longrightarrow$ `2`).

EXAMPLE 2 MEAN AND MEDIAN

Find the mean, median, and mode for the data set summarized by the following frequency table, the frequency of airline no-shows.

Number of No-Shows	Frequency
0	37
1	31
2	20
3	16
4	12
5	4

Mean:

Median:

Mode:

SECTION 2.6 Skewness and the Mean, Median, and Mode

Why do we need more than one measure of center? Can't we just find the mean and call it a day?

COMPARING MEAN AND MEDIAN — EXAMPLE 1

At a small startup, there is one CEO who takes home \$1,000,000 a year and 9 interns who make \$40,000 a year each.

Mean: The average salary at this company is

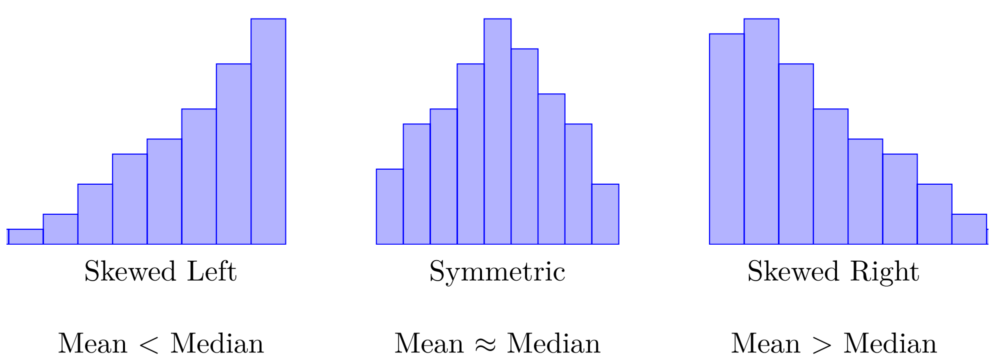
Median: The median salary at this company is

Which of these is a better measure of what a "typical" member of this corporation makes?

This is an example of a **skewed** data set, with one outlier.

Sensitivity to Outliers

The data set in that example is said to be **skewed to the right**, because the outlier is on the right, or upper side of the data set.



If the mean is larger than the median, there are outliers on the upper side pulling the mean up, so the data is skewed to the right (vice versa if the mean is smaller than the median)

EXAMPLE 2 SKEWED OR SYMMETRIC?

Is the following data set approximately symmetric or skewed?

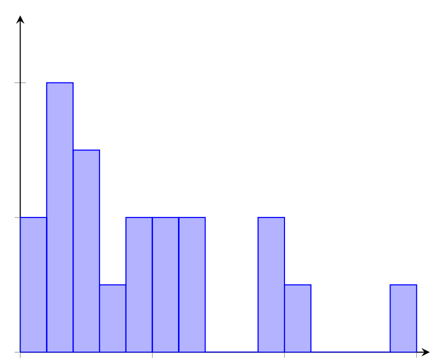
2010 Winter Olympics Gold Medal Wins by Country

0	0	1	1	1	1	2	2	2	3
4	4	5	5	6	6	9	9	10	14

Mean:

Median:

Conclusion:



Because the mean is more sensitive to outliers than the median is, the median is often a better measure of the center than the mean. Thus, when you go shopping for a house, or looking at typical salaries in a particular field, try searching for the *median* house price or salary.

SECTION 2.7 Measures of Spread

Knowing where a data set is centered is good, but we also would like to be able to measure how spread out a data set is.

The simplest measure of spread is the range of a data set:

There's a better measure of spread, though: the **standard deviation**.

Standard Deviation

The standard deviation is essentially the average distance of the data points from the mean, the center.

For a sample of size n,

For a population of size N,

Why is it so complicated? Why can't we just find the distances from the mean ($x - \overline{x}$, called the **deviations**) and average them? The problem is, if we add up all the deviations, the sum will always equal 0, so taking the average won't work.

That's why we square the deviations, then average[1] them, and then take the square root again.

[1]Notice the $n-1$ in the denominator. This isn't quite an average. The reason for the $n-1$ is somewhat complicated, but basically, it's there so that the sample standard deviation is an *unbiased estimator* of the population standard deviation.

EXAMPLE 1 STANDARD DEVIATION

The ages of ten fifth-grade students are given below.

11	10	9.5	11	11.5
10.5	10	11	10	9.5

Find the standard deviation of this data set.

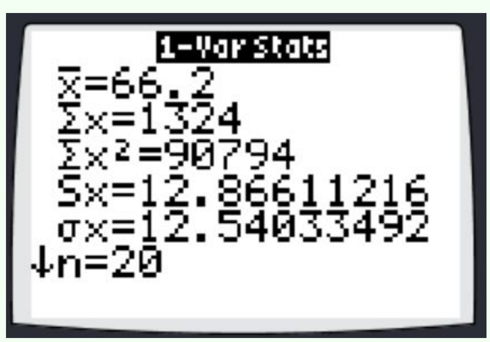
Data	**Deviations**	**Sq. Deviations**
x	$x - \overline{x}$	$(x - \overline{x})^2$
11	0.475	0.225625
10	-0.525	0.275625
9.5	-1.025	1.050625
11	0.475	0.225625
11.5	0.975	0.950625
10.5	-0.025	0.000625
10	-0.525	0.275625
11	0.475	0.225625
10	-0.525	0.275625
9.5	-1.025	1.050625

The sum of the squared deviations is 4.55625; divide this by 9 and take the square root:

Of course, we don't do this process in practice; we just use `1-Var Stats` on the calculator.

S_x or σ_x?

There are two standard deviations listed in `1-Var Stats`: S_x and σ_x.



The difference between them is that S_x is the sample standard deviation, and σ_x is the population standard deviation.
Which one to use depends on whether the data set in question is the entire population of interest, or a sample from that population.

z-scores

USING THE STANDARD DEVIATION EXAMPLE 2

On a baseball team, the ages of each of the players are as follows:

21	21	22	23	24
24	25	25	28	29
29	31	32	33	33
34	35	36	36	36
36	38	38	38	40

1. Find the mean and standard deviation.

   Mean:

   Standard Deviation:

2. Find the value that is one standard deviation below the mean.

In this example, the standard deviation gives us a measure of position that is more powerful than it seems at the moment: the z-score.

The z-score is the number of standard deviations that a particular data point falls above or below the mean.

EXAMPLE 3 Z-SCORES

In the baseball team age data set, find the z-scores that correspond to the following ages:

(a) 26

(b) 32

z-score

The z-score is the number of standard deviations that a particular data point falls above or below the mean.

To find the z-score for a particular data point, subtract the mean and divide the answer by the standard deviation:

What good are z-scores? The first application is in comparing data points in different data sets.

COMPARING TEST SCORES — EXAMPLE 4

Scores on the SAT and ACT are normally distributed:

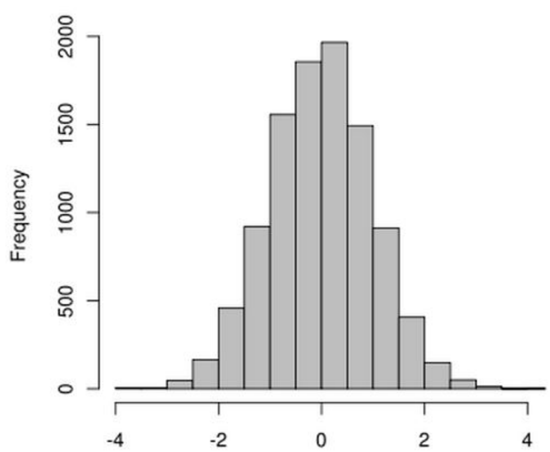
Test	Mean	Std. Deviation
SAT	500	100
ACT	18	6

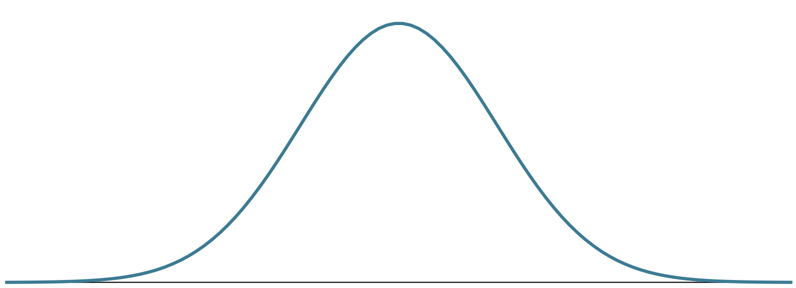
You score 550 on the SAT and 24 on the ACT. On which test did you have a better score, relative to everyone else who took the test?

The z-scores for each test score are

Empirical Rule

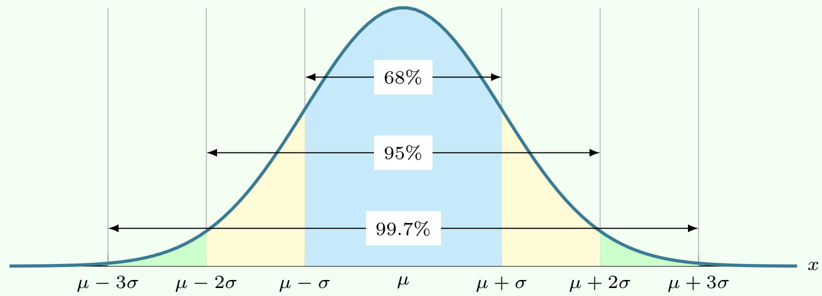
Another application is the Empirical Rule. This rule applies to data sets for which the histogram is **symmetric** and **bell-shaped**:





The Empirical Rule

- Approximately 68% of the data is within **one** standard deviation of the mean.
- Approximately 95% of the data is within **two** standard deviations of the mean.
- Approximately 99.7% of the data is within **three** standard deviations of the mean.



Note that this diagram uses μ for the population mean (as opposed to $\overline{x}$ for the sample mean) and σ for the population standard deviation (as opposed to s for the sample standard deviation).

This will come back later when we study the **Normal Distribution**.

Chebyshev's Rule

This rule applies to any data set, regardless of whether or not it is symmetric and bell-shaped.

Chebyshev's Rule

- At least 75% of the data is within **two** standard deviations of the mean.
- At least 89% of the data is within **three** standard deviations of the mean.

Homework 4 Name:

1. Listed below are the weights (in pounds) of a sample of players on the San Francisco 49ers football team.

177	205	210	210	232	205
185	185	178	210	245	259
278	270	280	295	272	286
215	241	302	265	290	228

(a) Find the mean of this sample. ________

(b) Find the median of this sample. ________

(c) Is this sample symmetric or somewhat skewed?

____

(d) Find the mode of this sample. ________

(e) Find the standard deviation.

____

(f) When Steve Young played for the 49ers, he weighed 205 pounds. How many standard deviations above or below the mean was he?

____

(g) The same year, the mean weight for the Dallas Cowboys was 240.08 pounds with a standard deviation of 44.38 pounds, and Emmit Smith weighed in at 209 pounds. With respect to his team, who was lighter, Smith or Young?

____

2. Listed below are the daily high temperatures in Phoenix for May of a given year.

94	93	92	96	93	87	85	84
88	93	96	99	103	103	101	101
94	94	95	97	99	103	100	101
98	96	94	98	98	100	101	

(a) Find the mean of this sample. ________

(b) Find the median of this sample. ________

(c) Is this sample symmetric or somewhat skewed?

____

(d) Find the mode of this sample. ________

(e) Find the standard deviation.

____

(f) If the temperature one day was 101 degrees, how many standard deviations above or below the mean would that be?

____

(g) The same month, the mean temperature in New York City was 83 degrees, with a standard deviation of 3.6 degrees. Which day would be relatively hotter that month, a 102 degree day in Phoenix or a 87 degree day in New York City?

____

3. Suppose you are studying home prices, and you take a sample with a mean price of \$300,000 and a standard deviation of \$50,000. Suppose this data set is normally distributed.

(a) Find the z-score of a home in this sample that sells for \$420,000. ________

(b) Find the z-score of a home in this sample that sells for \$230,000. ________

(c) What percentage of houses should sell at between \$200,000 and \$400,000? ________

(d) According to Chebyshev's Rule, what is the minimum percentage of houses that should sell at between \$150,000 and \$450,000?

____

4. Twenty-five students were asked the number of movies they watched the previous week. The results are shown below.

Number of Movies	Frequency
0	5
1	9
2	6
3	4
4	1

(a) Find the mean of this sample. __________

(b) Find the median of this sample. __________

(c) Find the mode of this sample. __________

5. Forty students were asked the number of pairs of sneakers they owned. The results are shown below.

Number of Pairs	Frequency
1	2
2	5
3	8
4	12
5	12
6	0
7	1

(a) Find the mean of this sample. __________

(b) Find the median of this sample. __________

(c) Find the mode of this sample. __________

Chapter

3 Probability

Much of the study of statistics needs a grounding in the basics of probability, so in this chapter we'll start with the basics; you most likely have some intuitive understanding of probability, but our goal is to formalize much of this.

When a weather forecaster gives a prediction, an actuary estimates insurance payouts, or a basketball commentator describes how likely it is that a player will make the next free throw, they are using (to varying extents) some of the principles outlined in this chapter. You may not realize how much probability gets used around you.

SECTION 3.1 Basic Concepts

You probably have some idea of what we mean when we say "probability," but here's a definition to clarify:

What is Probability?

Probability:

Probability of something occurring:

Vocabulary

- **Outcome:**

- **Sample Space:**

(a)

(b)

(c)

- **Event:**

The probability of an event A is written $P(A)$, or we could write P(rolling a 4) or $P(4)$, if that is clear enough in context.

TWO SIBLINGS — EXAMPLE 1

Consider randomly selecting a family with 2 children where the order in which different gender siblings are born is significant. That is, a family with a younger girl and an older boy is different from a family with an older girl and a younger boy. What would the sample space look like?

Solution

If we let G denote a girl, B denote a boy, then we have the following:

$$S = \{GG, BB, GB, BG\}$$

This notation represents families with 2 girls, 2 boys, an older girl and a younger boy, an older boy and a younger girl.

THREE SIBLINGS — EXAMPLE 2

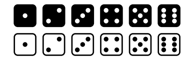
What would the sample space S look like if we considered a family with 3 children? Remember, the order of children born is significant.

Now there are 8 possibilities:

TOSSING A COIN AND ROLLING A DIE — EXAMPLE 3

Suppose we toss a fair coin and then roll a six-sided die once. Describe the sample space S.

Solution



Let T denote Tails, and H denote Heads. Then

Note: in the definition of probability, we said that probability is a **proportion**.

Probability

The basic rules of probability are:

1.

2.

3.

4.

Often we use percentages to represent probabilities. For example, a weather forecast might say that there is 85% chance of rain in Frederick tomorrow. Or there is 67% chance that the Baltimore Orioles will win their next series. Or a particular poker player has a 35% chance of winning the game with his current hand. As you might have already guessed, 100% chance corresponds to 1, and 0% corresponds to 0.

Theoretical Probability

There are two types of probability: **theoretical** and **empirical**. Theoretical probability is used when the set of all equally-likely outcomes is known. To compute the theoretical probability of an event A, denoted $P(A)$, we use the formula below:

Theoretical probability

This makes sense with the definition of probability, namely that it is the proportion of times we would expect E to occur if we repeated the experiment many times. This proportion comes from dividing the number of possibilities that correspond to E by the total number of possibilities there are.

In the example below, one could probably find the probability by intuition, but it's good to know how to apply the formula, even in what seems to be a simple experiment.

ROLLING A DIE — EXAMPLE 4

Assume you are rolling a fair six-sided die. What is the probability of rolling an odd number?

THREE SIBLINGS — EXAMPLE 5

Consider the earlier example about a family with three children. Remember, the sample space looked like

Let's find the probability of a few combinations of kids:

(a) $P(\text{three girls}) =$

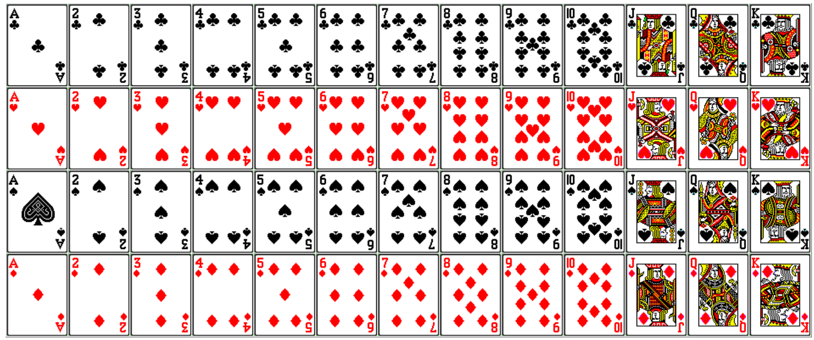
(b) $P(\text{at least two boys}) =$

(c) $P(\text{exactly one girl}) =$

(d) $P(\text{youngest is a boy}) =$

(e) $P(\text{oldest and youngest same}) =$

In the next example, it is not necessary to list all possible outcomes of an experiment. However, if you are not familiar with a standard deck of 52 cards, the diagram below should be helpful.



EXAMPLE 6 DRAWING A CARD

Suppose you draw one card from a standard 52-card deck. What is the probability of drawing an Ace?

EXAMPLE 7 DRAWING ANOTHER CARD

(a) When drawing a card from a standard 52-card deck, what is the probability of drawing a face card? *Face cards include Jacks, Queens, and Kings.*

(b) What is the probability of drawing the King of Hearts?

COOKIE JAR **EXAMPLE 8**

Lisa's cookie jar contains the following: 5 peanut butter, 10 oatmeal raisin, 12 chocolate chip, and 8 sugar cookies. If Lisa selects one cookie, what is the probability she gets a peanut butter cookie?

Empirical Probability

As long as we can list–or at least count–the sample space and the number of outcomes that correspond to our event, we can calculate basic probabilities by dividing, as we have done so far. But there are many situations where this isn't feasible.

For instance, take the example of a batter coming to the plate in a baseball game. There's no way to even begin to list all the possible outcomes that could occur, much less count how many of them correspond to the batter getting a hit. We'd still like to be able to estimate the likelihood of the batter getting a hit during this at-bat, though. Just as sports fan do, then, we turn to this batter's previous performance; if he's gotten a hit in 200 of his last 1000 at-bats, we assume that the probability of a hit this time is $\frac{200}{1000} = 0.200$.

Empirical probability is used when we observe the number of occurrences of an event. It is used to calculate probabilities based on the *real data* that we observed and collected. To compute the empirical probability of an event E, denoted $P(E)$, we use the formula below:

Empirical probability

This can also be used to answer questions about sampling randomly from a population if we know the breakdown of the group.

EXAMPLE 9 FCC STUDENTS

Consider the following information about FCC students' enrollment:

Gender	Enrollment
Female	3653
Male	2580

If one person is randomly selected from all students at FCC, what is the probability of selecting a male?

The next example contains a two-way table, often referred to as *contingency* table, which breaks down information about a group based on two criteria. For example, the table below breaks down a group of 130 FCC students based on gender and which hand is their dominant hand:

Gender	Right-handed	Left-handed
Female	58	13
Male	47	12

In order to use this to calculate probabilities if we randomly select someone from the group, we need to calculate totals for each category: the number of males, the number of females, the number of left-handed people, and the number of right-handed people. This is done by simply summing each row and column; if we do that, we obtain the completed table below.

Gender	Right-handed	Left-handed	**Total**
Female	58	13	71
Male	47	12	59
Total	105	25	130

FCC STUDENTS EXAMPLE 10

Consider the following information about a group of 130 FCC students:

Gender	Right-handed	Left-handed	**Total**
Female	58	13	71
Male	47	12	59
Total	105	25	130

(a) If one person is randomly selected from the group, what is the probability this student is left-handed?

(b) Find the probability of selecting a female student.

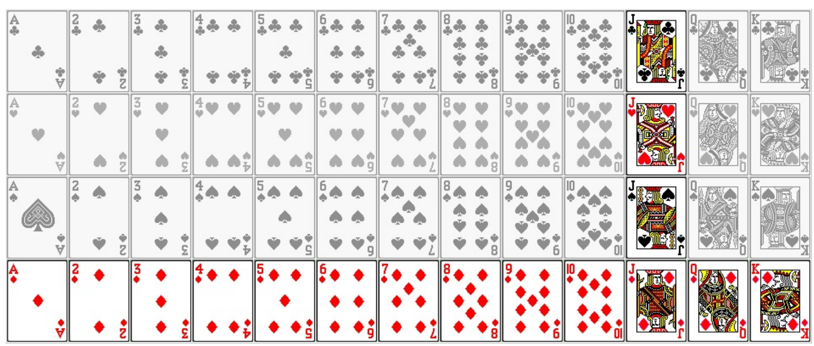
SECTION 3.2 The Addition Rule and Complements

In this section, we will focus on computing probabilities of events involving "or" as well as learning the concept of mutually exclusive events. We will also discuss complementary events and their probabilities.

To start, recall the experiment of drawing one card from a standard deck of cards. Let J denote drawing a Jack, and Q denote drawing a Queen. What is the probability of drawing a Jack? It is, of course, 4/52, and the same goes for the probability of drawing a Queen. Now, what is the probability of drawing a Jack *OR* Queen? By looking back at the deck of cards, we can see that there are 8 cards that are either Jacks or Queens, so

which happens to be the sum of their individual probabilities.

What about, though, if we wanted to find the probability of drawing a Jack or a diamond? Could we just add their individual probabilities (4/52 and 13/52, respectively)? Let's check by looking back at the cards and see which correspond to Jacks or diamonds.



Notice that there are 16 cards that match that description, so the probability is 16/32, which ISN'T the sum of the individual probabilities. What went wrong?

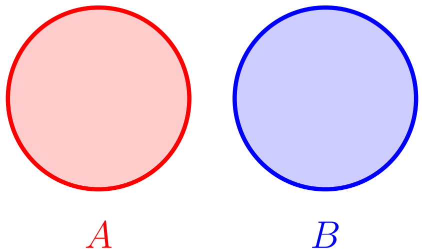
Mutually Exclusive Events

The answer can be found by looking at the diagram above. Notice that if we add up the number of Jacks and the number of diamonds (for a total of 19), we *double count* the Jack of diamonds. This brings us to an important definition that determines how we find the probability of one event OR another occurring: we need to find whether the events are **mutually exclusive** or **disjoint**. That is, can these two events happen at the same time?

Disjoint (mutually exclusive) outcomes

Can we draw a card that is both Jack and Queen? Clearly, there is no such card, therefore these events are disjoint. Another familiar example of disjoint events would be getting an even or odd number when rolling a die. Each number is either even or odd, thus these two events are also mutually exclusive. Above, though, we showed that drawing a Jack and drawing a diamond are NOT mutually exclusive, since you can draw the Jack of diamonds.

Notice that the terms **disjoint** and **mutually exclusive** are equivalent and interchangeable. The Venn diagram below illustrates the concept of mutually exclusive events: two events A and B do not overlap; they are disjoint.



Before we formally define a formula for computing probabilities of disjoint events, let us solve some problems by using the rules we already know.

ROLLING A DIE **EXAMPLE 1**

Suppose you roll a fair six-sided die once. What is the probability of rolling a 6 or an odd number?

Addition rule for mutually exclusive events

EXAMPLE 2 DRAWING A CARD

Suppose you draw one card from a standard 52-card deck.

(a) What is the probability that you get an Ace or a face card?

(b) What is the probability of getting a number or a red Jack?

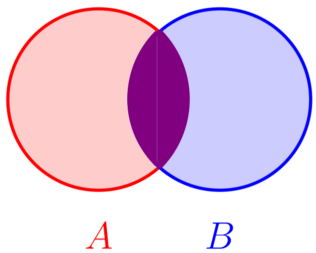
(c) What is the probability of selecting a red card or a black card?

EXAMPLE 3 MARBLES

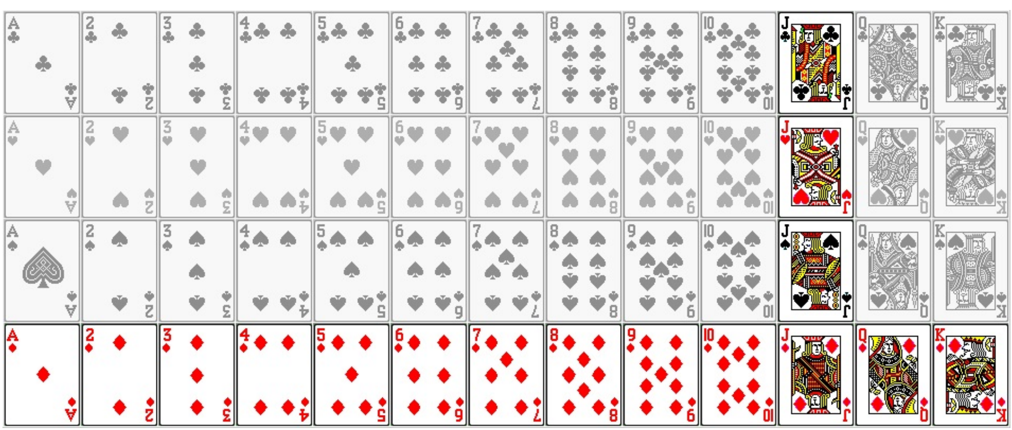
A large bag contains 28 marbles: 7 are blue, 8 are yellow, 3 are white, and 10 are red. If one marble is randomly selected, what is the probability that it's either red or yellow?

Overlapping Events

What if the events of interest are not mutually exclusive? How do we compute probabilities of events that are not disjoint? Pictorially, we can visualize this situation with the following diagram, where the red intersection of two circles represents all outcomes when two events both happen. For example, if we consider FCC students, selecting a female and selecting a full-time students would not be mutually exclusive events, since there are certainly female students who go to school full time.



Let's go back to the deck of cards to see how to calculate probabilities in situations like this. We'll again use the example of drawing a Jack or a diamond.



As we noted already, these are not mutually exclusive events. Because of that, adding the probability of drawing a Jack (4/52) and the probability of drawing a diamond (13/52) gave an incorrect answer of 17/52, where the correct probability–as we noted earlier–is 16/52. Again, this is because we *double counted* the Jack of diamonds, once when we calculated the probability of drawing a Jack and once when we calculated the probability of a diamond.

The way to correct for this double counting is to subtract off the overlap; thus, we'll add up the probability of drawing a Jack and the probability of drawing a diamond and then subtract the probability of drawing both together (i.e. of drawing the Jack of diamonds):

In general, to calculate probabilities of compound events that are not mutually exclusive, we will use the General Addition rule:

General Addition rule

Notice that this is a more general form of the addition rule we stated earlier, with mutually exclusive events. If two events are mutually exclusive, the probability of them occurring together is 0, so the general addition rule simplifies down in that case to the simpler addition rule.

EXAMPLE 4 DRAWING A CARD

Suppose you draw one card from a standard 52-card deck.

(a) What is the probability that you get a King or a spade?

(b) What is the probability that you get a Queen or a face card?

FCC STUDENTS EXAMPLE 5

Consider the following information about a group of 130 FCC students:

Gender	Right-handed	Left-handed	**Total**
Female	58	13	71
Male	47	12	59
Total	105	25	130

(a) If one person is randomly selected from the group, what is the probability this student is female or left-handed?

(b) Compute the probability of selecting a male or a right-handed student.

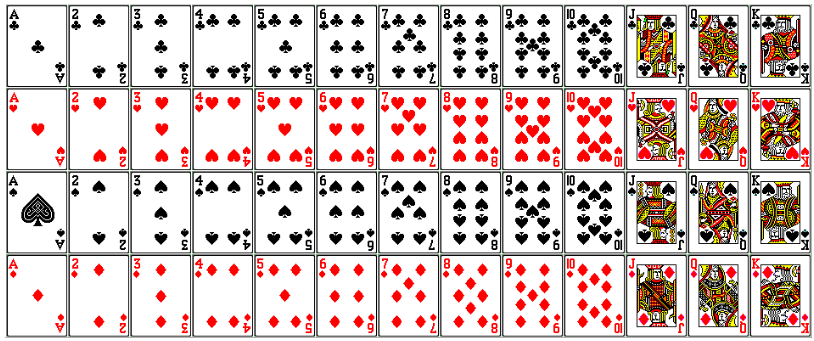
SPEEDING TICKETS AND CAR COLOR EXAMPLE 6

The table below shows the number of survey subjects who have received and not received a speeding ticket in the last year, and the color of their car. Find the probability that a randomly chosen person has a red car *or* got a speeding ticket

	Speeding ticket	No speeding ticket	**Total**
Red car	15	135	150
Not red car	45	470	515
Total	60	605	665

Complements

The probability of an event not occurring can be just as useful as computing the probability of that event happening. The best way to introduce this concept is to consider an example. Let's revisit the standard 52-card deck, where we randomly select one card:



What is the probability of not drawing an Ace? Well, you know that there are 4 Aces in the deck, so $52 - 4 = 48$ cards that are not Aces. We compute:

$$P(\text{not}\quad \text{Ace}) = \frac{48}{52} \approx 0.923$$

Now, notice that

$$\frac{48}{52} = 1 - \frac{4}{52},\ \text{where}\ P(Ace) = \frac{4}{52}$$

This is not a coincidence. If you recall the basic rules of probability, the sum of probabilities of all outcomes must be 1. In this case, the card you draw is either an Ace or it's not, so it makes sense that the probabilities of these two events add up to 1.

Complement of an event

NOT HEARTS! EXAMPLE 7

If you pull a random card, what is the probability it is not a heart?

MULTIPLE CHOICE QUESTION EXAMPLE 8

A multiple choice question has 5 answers, and exactly one of them is correct. If you were to guess, what is the probability of not getting the correct answer?

FCC STUDENTS' DEMOGRAPHICS EXAMPLE 9

According to the FCC website, female students make up 57% of the Fall 2014 student body. If one student is randomly selected, what is the probability the student is not female?

SECTION 3.3 The Multiplication Rule

We began by calculating the probabilities of single events occurring, and then we learned how to combine events using *OR*. Now we ask a different question: suppose we know how to calculate the probability of A and the probability of B on their own; how can we calculate the probability that A *AND* B both occur? To set this up, we'll look at two situations: flipping a coin twice and drawing two cards *without replacement* (this will be important).

Independent

Flipping a coin twice If we flip a coin twice in succession, the sample space is

Now suppose we ask the following questions:

1. What is the probability that the first flip results in a head?

2. What is the probability that the second flip results in a tail?

3. What is the probability that the first flip results in a head *AND* the second flip results in a tail?

Notice that the probability of both happening together is the probability of one times the probability of the other:

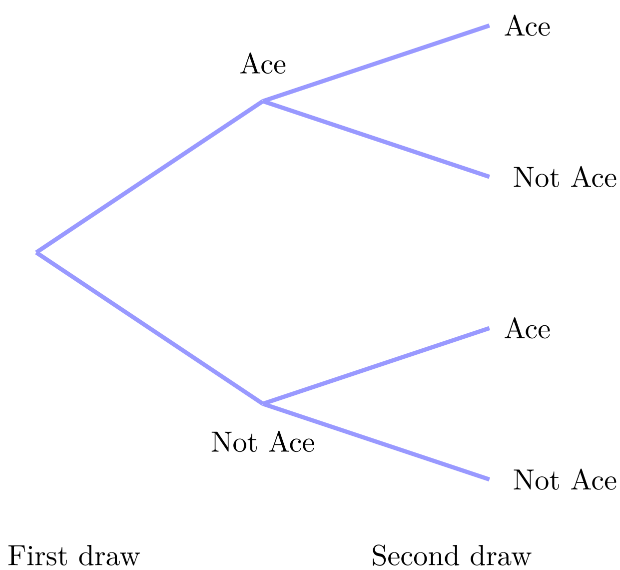
Seeing this, and noting the title of the section, we may be tempted to jump to the conclusion that the probability of A *AND* B is simply the probability of A times the probability of B. However, the next scenario illustrates that we need to be a bit more careful.

Just as we found with the addition rule, there is a simple version that works if a certain condition is met, and if not, there is a more general version of the multiplication rule.

Drawing two cards without replacement Suppose we draw one card, and then *without* placing it back and re-shuffling the deck, we draw a second card. What is the probability that we draw two Aces?

Not independent

This situation is different from the previous one, because now what happens on the first draw affects the probabilities for the second draw. In other words, the probability of drawing an Ace the first time is 4/52. If we draw an Ace the first time, there are only 3 Aces left and 51 total cards left, so the probability of drawing an Ace the second time is 3/51. However, if we do not draw an Ace the first time, there are still 4 Aces in the deck, so the probability of drawing an Ace the second time is 4/51. We can illustrate this with a branching tree diagram.



Now the probability of drawing an Ace both times is the probability of drawing an Ace the first time multiplied by the probability of drawing an Ace the second time **given that we drew an Ace the first time**. Notice on the tree diagram that this corresponds to following the upward branch both times.

This is because *only* if we draw an Ace the first time do we have any chance of fulfilling the scenario; if we fail to draw an Ace the first time, it doesn't matter what we do the second time–we've already failed.

Thus, the probability of drawing an Ace both times is

This is what we call *conditional probability*, and it's what we have to consider for the general multiplication rule.

Independence

Independence

Note that saying that two events are *independent* is different than saying that two events are *mutually exclusive*.

-

-

EXAMPLE 1 INDEPENDENT EVENTS

Determine whether these events are independent:

1. A fair coin is tossed two times. The two events are $A =$ first toss is Heads and $B =$ second toss is Heads.

2. The two events $A =$ *It will rain tomorrow in Frederick MD* and B $=$ *It will rain tomorrow in Thurmont MD*

3. You draw a red card from a deck, then draw a second card without replacing the first.

4. You draw a face card from the deck, then replace it and re-shuffle the deck before drawing a second card.

Now we are ready to formally state the rule that we used in the first scenario at the beginning of the section.

The Multiplication Rule for Independent Events

Probabilities of independent events

If A and B are independent, then the probability of both A and B occurring is

We can generalize this to finitely many independent events $A_1, A_2, \ldots, A_k$

COINS AND DICE — EXAMPLE 2

Suppose you flip a coin and roll a six-sided die once. What is the probability you get Tails and an even number?

DRAWING CARDS WITH REPLACEMENT — EXAMPLE 3

Assume you have a 52 card deck, and you select two cards at random. Also assume that you replace and reshuffle after each selection. Find the probability of drawing a king first and then a black card.

EXAMPLE 4 LEFT-HANDED POPULATION

About 9% of people are left-handed. Suppose 2 people are selected at random from the U.S. population. Because the sample size of 2 is very small relative to the population, it is reasonable to assume these two people are independent. What is the probability that both are left-handed?

EXAMPLE 5 BOYS AND GIRLS

Assuming that probability of having a boy is 0.5, find the probability of a family having 3 boys.

Multiplication Rule for Dependent Events

Conditional probability:

Notation:

Example: Drawing two Aces:

Multiplication formula for dependent events

If events A and B are not independent, then

Note that this, like with the addition rule, is the general multiplication rule; if A and B are independent, $P(B|A) = P(B)$ (because the probability of B is the same regardless of whether A has occurred or not) and the general multiplication formula becomes the simpler form for independent events that we have already seen.

DRAWING CARDS WITHOUT REPLACEMENT — EXAMPLE 6

If you pull 2 cards out of a deck, what is the probability that both are spades?

SOCK COLORS — EXAMPLE 7

In your drawer you have 10 pairs of socks, 6 of which are white. If you reach in and randomly grab two pairs of socks, what is the probability that both are white?

EXAMPLE 8 M&M'S

A bag of M&M's contains the following breakdown of colors:

Red	Yellow	Brown	Blue	Orange	Green
12	18	24	22	13	17

Suppose you pull two M&M's out of the bag (without replacing candy after each pull). Find the following probabilities:

1. The probability of drawing two red candies

2. The probability of drawing a blue candy and then a brown candy

3. The probability of not drawing 2 green candies

We'll conclude this section with an example of calculating conditional probability from a contingency table.

CONDITIONAL PROBABILITY AND CONTINGENCY TABLES

EXAMPLE 9

Again using the data regarding 130 FCC students, broken down by gender and dominant hand:

Gender	Right-handed	Left-handed	**Total**
Female	58	13	71
Male	47	12	59
Total	105	25	130

1. What is the probability that a randomly chosen student is female, given that the student is left-handed?

2. What is the probability that a randomly chosen student is right-handed, given that the student is male?

MEDICAL TEST

EXAMPLE 10

A certain disease infects 100 out of every 100,000 people. The test for this disease is correct 99% of the time. If you get a positive result, what is the probability that you have the disease?

Homework 5 Name:

1. A fair die is rolled. Find the probability of getting at least 5.

____

2. A ball is drawn randomly from a jar that contains 6 red balls, 2 white balls, and 5 yellow balls. Find the probability of drawing a white ball.

____

3. Suppose you write each letter of the alphabet on a different slip of paper and put the slips into a hat. What is the probability of drawing one slip of paper from the hat at random and getting a consonant?

____

4. In a survey, 205 people indicated they prefer cats, 160 indicated they prefer dogs, and 40 indicated they don't enjoy either pet. Find the probability that if a person is chosen at random, they prefer cats.

____

5. Lisa has a large bag of coins. After counting the coins, she recorded the counts in the table below. She then decided to draw some coins at random, replacing each coin before the next draw.

Quarters	Nickels	Dimes	Pennies
27	18	34	21

(a) What is the probability that Lisa obtains a quarter on the first draw?

____

(b) What is the probability that Lisa obtains a penny or a dime on the second draw?

____

(c) What is the probability that Lisa obtains at most 10 cents worth of money on the third draw?

____

(d) What is the probability that Lisa does not get a nickel on the fourth draw?

____

(e) What is the probability that Lisa obtains at least 10 cents worth of money on the fifth draw?

____

6. I asked my Facebook friends to complete a two-question survey. They answered the following questions: Which beverage do you prefer in the morning: coffee or tea? What is your gender? I summarized the results in following table:

	Coffee	Tea	**Total**
Female	37	24	**61**
Male	22	31	**53**
Total	**59**	**55**	**114**

(a) What is the probability that I select a friend who prefers coffee?

____

(b) What is the probability that I select a friend who is female?

____

(c) What is the probability that I select a friend who is male and prefers tea?

____

7. A jar contains 6 red marbles numbered 1 to 6 and 8 blue marbles numbered 1 to 8. A marble is drawn at random from the jar. Find the probability the marble is red or odd-numbered.

____

8. You draw one card from a standard 52-card deck.

1. What is the probability of selecting a King or a Queen?

   ____

2. What is the probability of selecting a face card or a 10?

   ____

3. What is the probability of selecting a spade or a heart?

   ____

4. What is the probability of selecting a red card or a black card?

   ____

9. You are dealt a single card from a standard 52-card deck.

1. Find the probability that you are not dealt a diamond.

   ____

2. Find the probability that you are not dealt a face card.

   ____

3. Find the probability that you are not dealt an Ace.

   ____

4. Find the probability that you are not dealt a jack or a king.

   ____

10. You have a box of chocolates that contains 50 pieces, of which 30 are solid chocolate, 15 are filled with cashews and 5 are filled with cherries. All the candies look exactly alike. You select a piece, eat it, select a second piece, eat it, and finally eat one last piece. Find the probability of selecting a solid chocolate followed by two cherry-filled chocolates.

____

11. The table below shows the number of survey subjects who have received and not received a speeding ticket in the last year, and the color of their car. Find the probability that a randomly chosen person has a red car *or* got a speeding ticket.

	Speeding ticket	No speeding ticket	**Total**
Red car	15	135	150
Not red car	45	470	515
Total	60	605	665

____

12. You roll a fair six-sided die twice. Find the probability of rolling a 6 the first time and a number greater than 2 the second time.

____

13. A math class consists of 25 students, 14 female and 11 male. Three students are selected at random, one at a time, to participate in a probability experiment. Compute the probability that:

(a) A male is selected, then two females.

____

(b) A female is selected, then two males.

____

(c) Two females are selected, then one male.

____

(d) Three males are selected.

____

14. A large cooler contains the following drinks: 6 lemonade, 8 Sprite, 15 Coke, and 7 root beer. You randomly pick two cans, one at a time (without replacement). Compute the following probabilities:

(a) What is the probability that you get 2 cans of Sprite?

____

(b) What is the probability that you do not get 2 cans of Coke?

____

(c) What is the probability that you get either 2 root beers or 2 lemonades?

____

(d) What is the probability that you get one can of Coke and one can of Sprite?

____

15. Suppose a math class contains 25 students, 14 females (three of whom speak French) and 11 males (two of whom speak French). Compute the probability that a randomly selected student speaks French, given that the student is female.

____

16. Suppose a math class contains 30 students, 18 females (four of whom speak French) and 12 males (three of whom speak French). Compute the probability that a randomly selected student is male, given that the student speaks French.

____

17. My top drawer contains different colored socks: 14 are white, 10 are black, 6 are pink, and 4 are blue. All socks in the drawer are loose. Every morning I randomly select 2 socks, one at a time. Calculate the following probabilities, giving both fraction and decimal answers, rounding to 4 decimal places:

1. What is the probability that I get a blue pair of socks? ________
2. What is the probability that I do not get a blue pair of socks? ________
3. What is the probability that I either get a white pair or a blue pair of socks? ________
4. What is the probability that I get one black sock and one white sock? ________

18. A poll was taken of 14,056 working adults aged 40-70 to determine their level of education. The participants were classified by sex and by level of education. The results were as follows.

Education Level	Male	Female	Total
High School or Less	3141	2434	5575
Bachelor's Degree	3619	3761	7380
Master's Degree	534	472	1006
Ph.D.	52	43	95
Total	7346	6710	14,056

A person is selected at random. Compute the following probabilities:

(a) The probability that the selected person does not have a Ph.D. ________

(b) The probability that the selected person is female and has a Master's degree ________

(c) The probability that the selected person is male or has a Ph.D ________

(d) The probability that the selected person is male, given he has a Master's degree ________

(e) The probability that the selected person does not have a Master's degree, given it is a male ________

(f) The probability that the selected person is female, given that she has a Bachelor's degree ________

(g) The probability that the selected person has a Ph.D, given it is a female ________

Chapter

4 Discrete Random Variables

If we roll five dice, what is the probability that all five of them come up odd? Four of them?

If we just wanted to know the probability that a single die would come up odd, that would be straightforward, but this question is harder.

In this chapter, we'll start working with **random variables**, which can be used to answer questions like this one.

Random Variable:

Discrete Random Variable:

SECTION 4.1 Probability Distribution Functions

- Random variables are denoted with capital letters:

- Values that the random variable can take are denoted with lowercase letters:

- The probability that x occurs is written

Probability Distribution Function

EXAMPLE 1 ROLLING TWO DICE

Roll two dice and let X be the sum of the values. Build the probability distribution function for this experiment.

Sample space:

$$\begin{aligned} S = \{&(1,1),\ (1,2),\ (1,3),\ (1,4),\ (1,5),\ (1,6) \\ &(2,1),\ (2,2),\ (2,3),\ (2,4),\ (2,5),\ (2,6) \\ &(3,1),\ (3,2),\ (3,3),\ (3,4),\ (3,5),\ (3,6) \\ &(4,1),\ (4,2),\ (4,3),\ (4,4),\ (4,5),\ (4,6) \\ &(5,1),\ (5,2),\ (5,3),\ (5,4),\ (5,5),\ (5,6) \\ &(6,1),\ (6,2),\ (6,3),\ (6,4),\ (6,5),\ (6,6)\} \end{aligned}$$

The sum can be anything from 2 to 12:

FLIP A COIN TWICE — EXAMPLE 2

Flip a coin twice; let X be the number of heads. Build the probability distribution function for this experiment.

Sample space:

Probability distribution function:

COLLEGE CLASSES — EXAMPLE 3

There are 5000 undergrads at a college. The following frequency table describes how many of them are taking a given number of courses.

Number of Courses	Frequency
1	478
2	645
3	568
4	1864
5	1357
6	88

If X is the number of courses that a randomly chosen student is taking, find the probability distribution for X.

Notice that the probability column is exactly what you would get if you put a relative frequency column on the frequency table.

Two Rules

1.

2.

EXAMPLE 4 SUPERMARKET CUSTOMERS

The number of customers in line at a supermarket express checkout counter is a random variable with the following probability distribution.

x	0	1	2	3	4	5
$P(x)$	0.10	0.25	0.30	0.20	0.10	0.05

1. Find $P(2)$.

2. Find $P(\text{no more than } 1)$.

3. Find the probability that no one is in line.

4. Find the probability that at least three people are in line.

SECTION 4.2 Expected Value

An investor is considering a \$10,000 investment in a start-up company. She estimates that she has a probability 0.25 of a \$20,000 loss, probability 0.20 of a \$10,000 profit, probability 0.15 of a \$50,000 profit, and probability 0.40 of breaking even (a profit of \$0). Would you advise her to make the investment?

To answer a question like this, we need to find the **expected value** of this random variable.

Expected Value

In the example with the investor, we have the following probability distribution, where X represents the earnings on this investment:

x	$P(x)$
\$0	0.40
\$10,000	0.20
\$50,000	0.15
-\$20,000	0.25

The expected value is

Since the expected value is positive, that means she can expect to make a profit.

Note: It's impossible for her to actually make \$4500, so this expected value isn't what we expect her to make. Instead, this means that if a bunch of investors made this investment, the majority of them would make money, and their average profits would be \$4500.

EXAMPLE 1 EXPECTED VALUE

Find the mean of the random variable with the following probability distribution.

x	$P(x)$
0	0.03125
1	0.15625
2	0.31250
3	0.31250
4	0.15625
5	0.03125

EXAMPLE 2 CIRCUIT BOARD DEFECTS

The following table presents the probability distribution of the number of defects X in a randomly chosen printed circuit board.

x	0	1	2	3
$P(x)$	0.5	0.3	0.1	0.1

Compute the mean μ_X.

LOTTERY EXAMPLE 3

In the NY State Numbers Lottery, you pay \$1 and pick a number from 000 to 999. If your number comes up, you win \$500, which is a profit of \$499. If you lose, you lose \$1. What is your expected value?

The probability distribution below describes your possible winnings.

x	$P(x)$
\$499	0.001
-\$1	0.999

Therefore, the expected value is

When casinos design games, you better believe the player's expected value is negative. This is why "the house always wins" in the long run, even if a few players win money now and again.

CRAPS EXAMPLE 4

In the game of craps, two dice are rolled, and people bet on the outcome. For example, you can bet \$1 that the dice will total 7, and if you win, your profit is \$4. What is your expected value?
The probability distribution below describes your possible winnings.

x	$P(x)$
\$4	$\frac{6}{36}$
-\$1	$\frac{30}{36}$

Therefore, the expected value is

EXAMPLE 5 CARNIVAL GAME

Consider the following game: you flip a fair coin. If you get heads on the flip, you win $200 and the game is over. If you get tails on the flip, you get to flip the coin a second time; if you get heads on the second flip, you win $40 and the game is over. If you get tails on the second flip, you win nothing and the game is over.

(a) Fill in the following probability distribution with the possible winnings and their associated probabilities.

x	
$P(x)$	

(b) What is the expected value for this game?

(c) What is the probability that you win at most $40 when you play this game once?

EXAMPLE 6 DRAW FOUR CARDS

You are playing a game in which you draw four cards from a standard deck of 52 cards, and the cards are replaced in the deck after each draw. You guess the suit of each card before it is drawn; you pay $1 to play, and if you guess the correct suit each time, you get your money back plus $275. Will you make money playing this game in the long run?

Since the card is replaced each time, the trials are independent, so the probability of guessing the correct suit four times in a row is

Therefore, the probability distribution looks like

The expected value of this game is

BIASED COIN EXAMPLE 7

Suppose you play a game with a biased coin, where the probability of heads is 2/3 and the probability of tails is 1/3. You toss the coin once; if your toss is heads, you pay \$6, and if your toss is tails, you win \$10. If you play this game many times, will you come out ahead?

The probability distribution looks like

Therefore, the expected value of this game is

Homework 6 Name:

1. Does the following table represent a probability distribution?

x	55	65	75	85
$P(x)$	-0.3	0.6	0.4	0.2

2. You conduct an experiment where you draw a single card and see if it is a face card or not. In other words, let X represent the number of face cards that you get when you draw that one card. Construct the probability distribution for X.

3. You flip a coin three times and let X be the number of tails. Construct the probability distribution for X.

4. The General Social Survey asked 1676 people how many hours per day they were able to relax. The results are shown below.

Number of Hours	0	1	2	3	4	5	6	7	8
Frequency	114	186	336	251	316	231	149	33	60

Let X be the number of hours of relaxation for a person sampled at random from this population.

(a) Construct the probability distribution of X.

(b) Find the probability that a person relaxes more than 4 hours per day.

(c) Find the probability that a person relaxes no more than 2 hours per day.

(d) Find the probability that a person doesn't relax at all.

(e) Compute the mean of this probability distribution.

5. Find the mean for the following probability distribution.

x	4	6	7	9
$P(x)$	0.33	0.11	0.21	0.35

6. Find the mean for the following probability distribution.

x	15	17	19	22	26
$P(x)$	0.15	0.23	0.25	0.27	0.10

7. A theater group holds a fund-raiser, selling 100 raffle tickets for \$5 each. The prize for the winning ticket is \$150. If you purchase 4 tickets, what is your expected value?

8. You buy a lottery ticket that costs \$10. There are only 100 tickets to be sold in this lottery. There is one \$500 prize, two \$100 prizes, and four \$25 prizes. Find your expected gain.

9. A game involves selecting a card from a standard 52-card deck and tossing a coin. If the card is a face card, and the coin lands on heads, you win \$6. If the card is a face card, and the coin lands on tails, you win \$2. If the card is not a face card, you lose \$2, no matter what the coin shows. Find the expected value of this game.

10. You are offered the following deal: you roll a die, and if you roll a six, you win \$10. If you roll a four or a five, you win \$5. If you roll anything else, you pay \$6. What are your expected winnings per game?

SECTION 4.3 Binomial Distribution

Remember, a random variable describes the results of an experiment. There is a certain kind of experiment that often arises, like the following example:

Ex: Suppose you run a manufacturing plant making light bulbs. Through extensive testing, you've found that the probability that a particular light bulb is defective is 0.02%. In a batch of 1000 light bulbs, what is the probability that 1 light bulb is defective? What about 2, or 3? 10? Fewer than 10, or more than 10?

All of these questions can be answered when we recognize that this problem is an example of a **binomial random variable**.

Binomial Experiment

A binomial random variable comes from a binomial experiment, which has the following characteristics:

1.

2.

3.

4.

5.

EXAMPLE 1 ARE THESE BINOMIAL?

Decide whether each of the following is a binomial random variable.

(a) A coin is tossed ten times. Let X be the number of heads.

(b) Five basketball players attempt a free throw. Let X be the number of free throws made.

(c) A random sample of 250 voters is chosen from a list of 10,000 registered voters. Let X be the number of who support the incumbent mayor for reelection.

Notation

The question with a binomial random variable is always: "what is the probability of x successes?" or "what is the probability of more/less than x successes?"

For example, the probability of 5 successes would be written $P(X = 5)$, and the probability of less than or equal to 5 successes would be written $P(X \leq 5)$.

Both kinds of questions can be answered if we have the probability distribution. For example, if there are three trials, we'd want to fill out the following table:

x	$P(x)$
0	
1	
2	
3	

For n trials, there can be anywhere from 0 to n successes. The probability of any number of successes will, of course, depend on the probability p of each individual success.

Okay, so how do we actually calculate the probability of a certain number of successes?

Binomial Distribution Formula

The probability of x successes in a binomial experiment is

When we actually do problems, though, we don't typically use the formula.

Using the Table

Rather than using the formula, we can use a table that was built with the formula (looking up a value in the table is quicker than evaluating the formula).

The table is essentially the probability distribution function for different numbers of trials and different probabilities of success. Below is a segment of this table (the full table is in the appendix).

Binomial Probabilities

n	k	$\binom{n}{k}$	0.01	0.05	0.10	0.15	0.20	0.2
2	0	1	0.9801	0.9025	0.8100	0.7225	0.6400	0.562
	1	2	0.0198	0.0950	0.1800	0.2550	0.3200	0.375
	2	1	0.0001	0.0025	0.0100	0.0225	0.0400	0.062
3	0	1	0.9703	0.8574	0.7290	0.6141	0.5120	0.421
	1	3	0.0294	0.1354	0.2430	0.3251	0.3840	0.421
	2	3	0.0003	0.0071	0.0270	0.0574	0.0960	0.140
	3	1		0.0001	0.0010	0.0034	0.0080	0.015
4	0	1	0.9606	0.8145	0.6561	0.5220	0.4096	0.316
	1	4	0.0388	0.1715	0.2916	0.3685	0.4096	0.421
	2	6	0.0006	0.0135	0.0486	0.0975	0.1536	0.210
	3	4		0.0005	0.0036	0.0115	0.0256	0.046
	4	1			0.0001	0.0005	0.0016	0.003
5	0	1	0.9510	0.7738	0.5905	0.4437	0.3277	0.237
	1	5	0.0480	0.2036	0.3281	0.3915	0.4096	0.395
	2	10	0.0010	0.0214	0.0729	0.1382	0.2048	0.263
	3	10		0.0011	0.0081	0.0244	0.0512	0.087
	4	5			0.0005	0.0022	0.0064	0.014
	5	1				0.0001	0.0003	0.001
6	0	1	0.9415	0.7351	0.5314	0.3771	0.2621	0.178

For instance, for three trials where $p = 0.15$, the probability distribution function is

x	$P(x)$
0	0.6141
1	0.3251
2	0.0574
3	0.0034

(notice that the probabilities all add up to 1, so this is a valid probability distribution function).

EXAMPLE 2 BINOMIAL PROBABILITIES

If $n = 3$ and $p = 0.15$, find the following probabilities:

(a) $P(X = 2)$

(b) $P(X \leq 1)$

(c) $P(X > 2)$

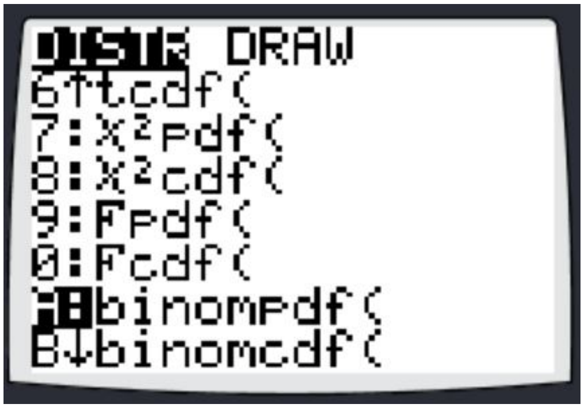
EXAMPLE 3 MULTIPLE-CHOICE QUIZ

A student takes a quiz with four multiple-choice questions, each with five possible answers. What is the probability that the student gets at least three correct answers if she guesses on each question?

Using Your Calculator

There's an even easier way to calculate binomial probabilities: using the built in function on your calculator. If you press 

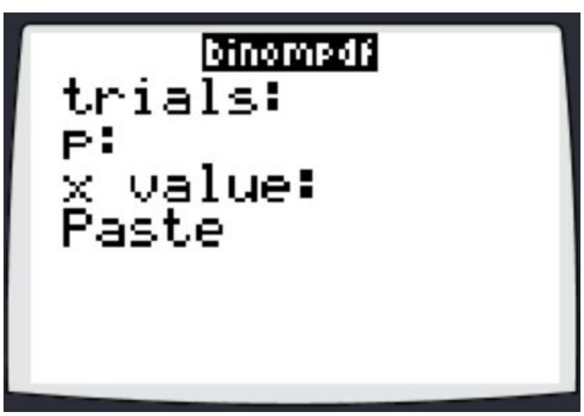
and scroll down, you'll find two relevant options: `binompdf` and `binomcdf`.



`binompdf:`

`binomcdf:`

If, for instance, you select `binompdf`, you might see a menu like the following:



After you enter the number of trials, the probability of success, and the number of successes in question, click `Paste`, and you'll see something like the following:

Note: On a TI-83, you should instead type in the three pieces, separated by commas. The syntax is

`binompdf(n,p,x)` or `binomcdf(n,p,x)`

EXAMPLE 4 AIRLINE FLIGHTS

At a particular airport, 81% of the flights arrived on time last year. If 15 flights are randomly selected, find the probability that

(a) exactly 10 of the flights are on time.

(b) exactly 12 of the flights are on time.

(c) 11 or fewer flights are on time.

(d) fewer than 10 flights are on time.

(e) more than 9 flights are on time.

(f) 11 or more flights are on time.

GOOGLE SEARCHES EXAMPLE 5

According to a Nielsen report, 65% of Internet searches in May 2010 used Google. If a sample of 25 searches are randomly selected, find the probability that

(a) exactly 20 of them used Google.

(b) 15 or fewer used Google.

(c) more than 22 used Google.

(d) fewer than 12 used Google.

(e) 17 or more used Google.

Mean: The expected value of a binomial random variable is

EXAMPLE 6 COLLEGE ENROLLMENT

The *Statistical Abstract of the United States* reported that 67% of students who graduated from high school in 2007 enrolled in college. Thirty high school graduates are sampled; find the probability that

(a) exactly 18 of them enroll in college.

(b) more than 15 of them enroll in college.

(c) fewer than 20 of them enroll in college.

How many of these students would you expect to enroll in college?

EXAMPLE 7 DRIVER'S EXAM

Sixty-five percent of people pass the state driver's exam on the first try. A group of 50 individuals who have taken the driver's exam is randomly selected. Find the probability that 30–35 of them passed on the first try.

There are two ways to tackle this problem:

One Way:

Another Way:

Summary

- To find $P(X =$ number), use `binompdf(n,p,`number`)`.
- To find $P(X \leq$ number), use `binomcdf(n,p,`number`)`.
- To find $P(X <$ number), use `binomcdf(n,p,`number-1`)`.
- To find $P(X \geq$ number), use $1-$`binomcdf(n,p,`number-1`)`.
- To find $P(X >$ number), use $1-$`binomcdf(n,p,`number`)`.
- To find $P(a \leq X \leq b)$, use `binomcdf(n,p,b)` $-$ `binomcdf(n,p,a-1)`.
  (pay attention to the inequalities)

Homework 7 Name:

1. A fair coin is flipped six times, and the number of heads is counted.

(a) Find the probability that the coin will land on heads exactly 3 times.

____

(b) Find the probability that the coin will land on heads more than 4 times.

____

(c) Find the probability that the coin will land on heads at least 2 times.

____

(d) Find the probability that the coin will land on heads fewer than 3 times.

____

2. At a particular college, 55% of the students population is female.

(a) Find the probability that a class of 10 students will have at least 4 females.

____

(b) Find the probability that a class of 25 students will have fewer than 10 females.

____

(c) Find the probability that a class of 8 students will have exactly 5 females.

____

(d) Find the probability that a class of 12 students will have more than 8 females.

____

3. A basketball player tends to make 78% of his free throw attempts. Suppose he attempts 12 free throws in a particular game.

(a) Find the probability that he makes exactly 10 of his free throws.

____

(b) Find the probability that he makes at most 11 of his free throws.

____

(c) Find the probability that he makes between 8 and 10 free throws.

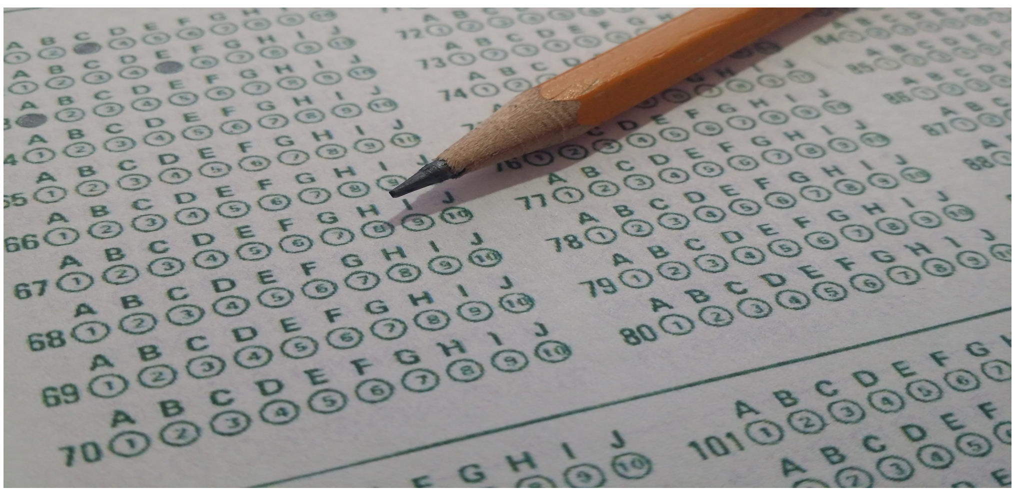
____

(d) Find the probability that he makes 9 or more of his free throws.

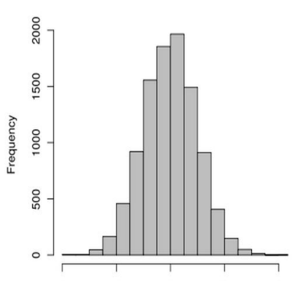
____

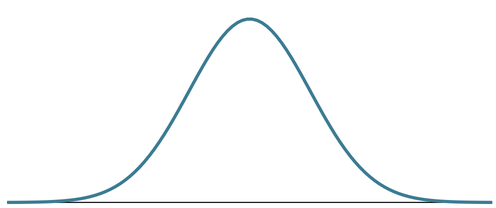
Chapter

6 The Normal Distribution



Standardized test scores tend to have a symmetric, bell-shaped distribution. What does that mean? That means that if we counted how many people got each score, and built a histogram (especially a relative frequency histogram), we'd get something that looked like the picture on the left. If those boxes got thinner and thinner (as we measured scores more finely), that histogram would start to look like the smooth curve on the right.





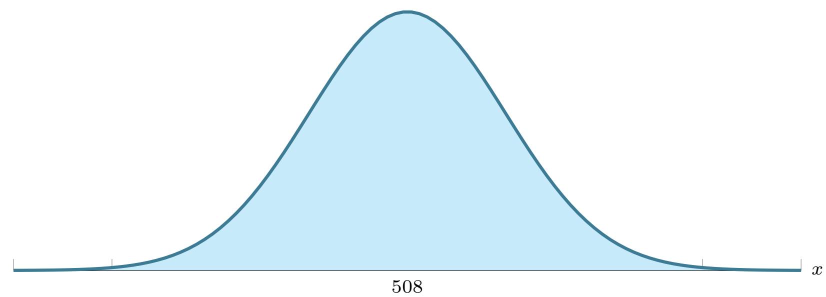
There are some quantities like these test scores that naturally have a distribution like this, but the normal distribution is more important for reasons that we'll see later.

SECTION 6.1 The Normal Distribution

This bell-shaped curve is similar to a probability distribution function (it's called a **probability density function**).

Just like a probability distribution, the density curve tells us how likely a given outcome is, based on the height of the density curve at that point.

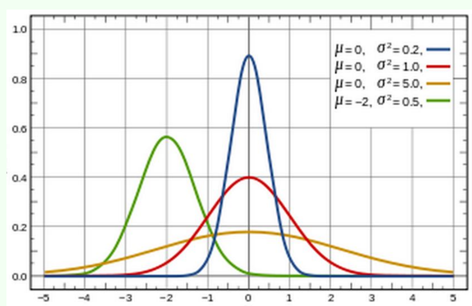
For instance, the average SAT verbal score is 508, and the distribution of scores looks like the following:



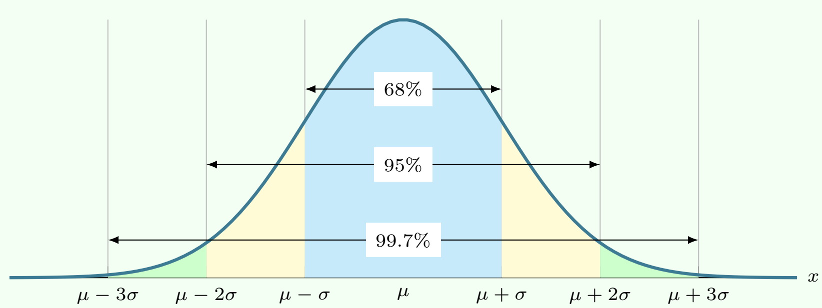
This means that the most common score is 508 (and thus that's the most likely result for a randomly chosen test taker). Not only that, but this distribution gives us a precise description of how scores are clustered around this center.

Normal Distribution

Two parameters define a particular normal distribution: the mean (center) and standard deviation (spread).



Recall the Empirical Rule:



THE INTELLIGENCE QUOTIENT — EXAMPLE 1

IQ is normally distributed with a mean of 100 and a standard deviation of 16. Use the Empirical Rule to the find the data that is within one, two, and three standard deviations of the mean.

CAR SALES — EXAMPLE 2

Suppose you know that the prices paid for cars are normally distributed with a mean of \$17,000 and a standard deviation of \$500. Use the 68–95–99.7 Rule to find the percentage of buyers who paid less than \$16,000.

z Table

What if we want to know about points that don't happen to be exactly one, two, or three standard deviations away from the mean?

Recall:

Now:

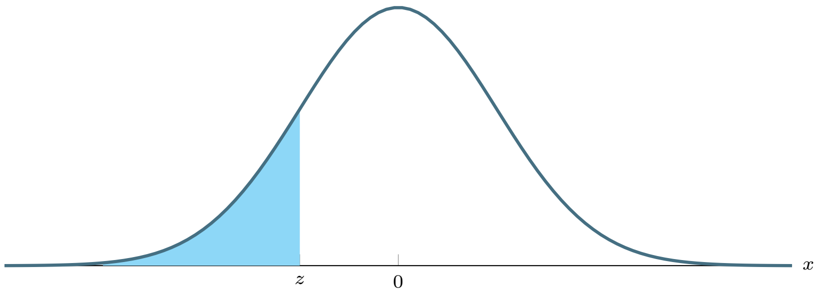
Area under normal curve:

Standard normal distribution:

We can use a table like the one below to find the area under a curve in given ranges.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0620	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379

More specifically, the table gives the proportion of values **below** any given z score:



Reading the z table: How do we read this?

For instance, to find the proportion to the left of $z = -1.73$, go down to the -1.7 row and over to the 0.03 column and read the proportion:

$$0.0418$$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0620	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170

EXAMPLE 3 USING Z TABLE

Find the area under the standard normal curve that is

(a) to the left of $z = 0.47$.

(b) to the right of $z = -1.24$.

(c) between $z = 0.86$ and $z = 1.15$.

(d) outside the interval between $z = -0.44$ and $z = 2.10$.

EXAMPLE 4 USING Z TABLE

A normal distribution has mean $\mu = 20$ and standard deviation $\sigma = 4$.

(a) What proportion of the population is less than 18?

(b) What is the probability that a randomly chosen value will be greater than 25?

Note: the normal distribution describes a **continuous** random variable, where we never talk about the probability that X *equals* a given value. This is because this probability is technically zero. This doesn't affect our problems much, except that we can talk interchangeably about

$$P(X \leq x) \qquad \text{or} \qquad P(X < x).$$

Using Your Calculator

Here again, there's an easier way, using your calculator. There's a built-in function called `normalcdf` that can calculate the proportion of the data in any given range for a normal distribution with any mean and standard deviation.

1. Press 

, then scroll to the second option: `2: normalcdf(`
2. You might see the following menu:

3. Enter values for the lower and upper bounds that you're interested in, as well as the mean and standard deviation of the given data set, and press `Paste`:

If you don't get the menu from the previous step, just enter the information the way it is shown here, as

```
normalcdf(lower,upper,mean,stdev)
```

EXAMPLE 5 PREGNANCY LENGTHS

The average length of a pregnancy is 272 days and the standard deviation is 9 days. Find the probability that

(a) a randomly chosen pregnancy will last less than 252 days.

(b) a randomly chosen pregnancy will last more than 252 days.

(c) a randomly chosen pregnancy will last between 252 and 298 days.

EXAMPLE 6 BLOOD PRESSURE

The Centers for Disease Control and Prevention reported that diastolic blood pressures of adult women in the US are approximately normally distributed with mean 80.5 and standard deviation 9.9.

(a) What proportion of women have blood pressures lower than 70?

(b) What is the probability that a randomly chosen woman would have blood pressure between 75 and 90?

(c) A diastolic blood pressure greater than 90 is classified as hypertension (high blood pressure). What proportion of women have hypertension?

Working Backwards: Percentiles

What if we turn the question around? Instead of asking what percentage of people fall into a certain range, we could ask what range corresponds to a given percentage. Of course, we've already done this, and we called it finding percentiles.

With a normal distribution, we can do this either with the z table or with the calculator. We'll do an example of each, but after this, we'll stick with the calculator method.

IQ SCORES EXAMPLE 7

IQ scores have a mean of 100 and a standard deviation of 15.

(a) Find the 90th percentile using the z table.

This means to find the point with 90% of the data below it. To use the table, locate a proportion as close as possible to 0.9000. The closest we can get is 0.8997, but that's good enough.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

(b) Find the value with 20% of the data above it, using the z table.

If 20% of the data is above a certain point, 80% must be below it.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621

EXAMPLE 8 CHERRY TREES

Cherry trees in a certain orchard have heights that are normally distributed with mean $\mu = 112$ inches and standard deviation $\sigma = 14$ inches.

(a) What proportion of trees are more than 120 inches tall?

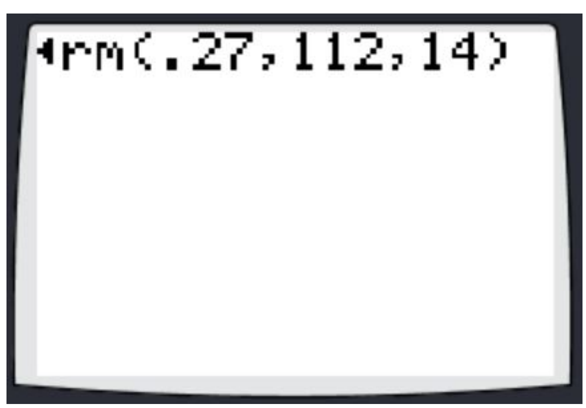
(b) What is the probability that a randomly chosen tree is either less than 100 inches tall or more than 125 inches tall?

(c) Find the 27th percentile of the tree heights.

Press 

and select `3: invNorm(` to pull up the following menu:

Enter the desired proportion (0.27 in this case for 27%), as well as the mean and standard deviation of the data set and press paste:



If the menu didn't show up for you, type it in as shown:

```
invNorm(proportion,mean,stdev)
```

In this case, the answer is

PREGNANCY PERCENTILES EXAMPLE 9

Recall that the average length of a pregnancy is 272 days and the standard deviation is 9 days. Find the 65th percentile of pregnancy lengths.

Homework 8 Name:

1. Suppose that the speeds at which cars pass through a checkpoint are normally distributed with a mean of $\mu = 61$ miles per hour and a standard deviation of $\sigma = 4$. Use the Empirical Rule to determine the probability that the next car that passes through the checkpoint will be traveling slower than 65 miles per hour.

____

2. Find the area under the standard normal curve that is

(a) to the left of $z = -1.22$. __________

(b) to the right of $z = 0.96$. __________

(c) between $z = -1.11$ and $z = -0.27$. __________

(d) outside the interval between $z = 0.1$ and $z = 1.94$. __________

3. A normal distribution has mean $\mu = 85$ and standard deviation $\sigma = 17$.

(a) What proportion of the population is below 67? __________

(b) What is the probability that a randomly chosen value from this population will be between 80 and 120? __________

4. The selling prices for homes in a certain community are normally distributed with mean $\mu = \$321,000$ and standard deviation $\sigma = \$38,000$.

(a) Find the probability that a randomly chosen house in this community will sell for less than $350,000. __________

(b) Find the probability that a randomly chosen house in this community will sell for more than $300,000. __________

(c) Find the probability that a randomly chosen house in this community will sell for between $290,000 and $370,000. __________

(d) Find the 45th percentile for home prices in this community. __________

5. The customers in a retail store have ages that are normally distributed with mean $\mu = 28.9$ years and standard deviation $\sigma = 6.3$ years.

(a) Find the probability that a randomly chosen customer will be younger than 31. __________

(b) Find the probability that a randomly chosen customer will be between 20 and 30 years old. __________

(c) Find the probability that a randomly chosen customer will be older than 16. __________

(d) Find the 80th percentile of customer ages. __________

Chapter

7 The Central Limit Theorem

The normal distribution can be used to describe some quantities that naturally fit it, but it is more valuable because of what we'll use it for throughout the rest of the course: the normal distribution lies behind much of what we'll do, and the Central Limit Theorem is what makes the connection.

For instance, when pollsters try to predict the outcome of an election, how do they know how good their predictions are going to be? Based on the theory that we'll see in this chapter and the next, they have a margin of error for their polls that gives an estimate of how reliable they are.

SECTION 7.1 The Central Limit Theorem

The Central Limit Theorem is one of the most profound and useful results in all of statistics and probability, and yet it isn't all that hard to understand.

The idea is this: take some quantity that isn't necessarily normally distributed. For instance, consider annual salaries in the U.S.

1.

2.

3.

4.

Central Limit Theorem

The point of the Central Limit Theorem is this:

Illustrating the Central Limit Theorem

The following table records the number of days that a cookie recipe lasted at a diner.

Recipe #	X	Recipe #	X	Recipe #	X	Recipe #	X
1	1	16	2	31	3	46	2
2	5	17	2	32	4	47	2
3	2	18	4	33	5	48	11
4	5	19	6	34	6	49	5
5	6	20	1	35	6	50	5
6	1	21	6	36	1	51	4
7	2	22	5	37	1	52	6
8	6	23	2	38	2	53	5
9	5	24	5	39	1	54	1
10	2	25	1	40	6	55	1
11	5	26	6	41	1	56	2
12	1	27	4	42	6	57	4
13	1	28	1	43	2	58	3
14	3	29	6	44	6	59	6
15	2	30	2	45	2	60	5

1. Calculate the population mean and standard deviation:

$$\mu_X =$$
$$\sigma_X =$$

2. Use a random number generator to select five samples of size $n = 5$ each. Record the mean of each sample. Then copy the means from students around you until you have at least 30 sample means.

3. Calculate the mean and standard deviation of this sampling distribution:

$$\mu_{\overline{X}} =$$
$$\sigma_{\overline{X}} =$$

4. Repeat this process, taking five samples of size $n = 10$ and recording the sample means for your samples and those of your classmates.

5. Calculate the mean and standard deviation of this sampling distribution:

$$\mu_{\overline{X}} =$$
$$\sigma_{\overline{X}} =$$

6. Draw a histogram for the original population, with a class width of one.

7. Draw a histogram for the first sampling distribution (where $n = 5$), with a class width of one half.

8. Draw a histogram for the first sampling distribution (where $n = 10$), with a class width of one half.

9. See what observations you can make.

Using the Central Limit Theorem

COLLEGE AGE EXAMPLE 1

The mean age of college students in 2008 was $\mu = 25$ years, with a standard deviation of $\sigma = 6.8$ years. A simple random sample of 64 students is drawn. What is the probability that the average age of the students in the sample is greater than 26 years?

EXAMPLE 2 BULL WEIGHT

If the mean weight of a bull is 1135 pounds, with a standard deviation of 97 pounds, would it be unusual for the mean weight of 100 head of cattle to be less than 1100 pounds?

This example sets up the kind of thing we'll do later: it would be unusual to get a sample like this if the claim is true about the mean weight of a bull, so if we DID get a sample like this, we might doubt that claim.

EXAMPLE 3 GAS MILEAGE

The EPA rates the mean highway gas mileage of the 2011 Ford Edge to be 27 miles per gallon. Assume the standard deviation is 3 miles per gallon. A rental car company buys 60 of these cars.

(a) What is the probability that the average mileage of the fleet is greater than 26.5 miles per gallon?

(b) What is the probability that the average mileage of the fleet is between 26 and 26.8 miles per gallon?

(c) Would it be unusual if the average mileage of the fleet were less than 26 miles per gallon?

NYC RENT EXAMPLE 4

The Real Estate Group NY reports that the mean monthly rent for a one-bedroom apartment without a doorman in Manhattan is \$2631. Assume the standard deviation is \$500. A real estate firm samples 100 apartments.

(a) What is the probability that the sample mean rent is greater than \$2700?

(b) What is the probability that the sample mean rent is between \$2500 and \$2600?

(c) Find the 60th percentile of the sample mean.

(d) Would it be unusual if the sample mean were greater than \$2800?

(e) Do you think it would be unusual for an individual apartment to have a rent greater than \$2800?

EXAMPLE 5 BATTERY LIFE

A battery manufacturer claims that the lifetime of a certain type of battery has a population mean of $\mu = 40$ hours and a standard deviation of $\sigma = 5$ hours. Let $\overline{x}$ represent the mean lifetime of the batteries in a simple random sample of size 100.

(a) If the claim is true, what is $P(\overline{x} \leq 39.8)$?

(b) Based on the answer to part (a), if the claim is true, is a sample mean lifetime of 39.8 hours unusually short?

(c) If the sample mean lifetime of the 100 batteries were 39.8 hours, would you find the manufacturer's claim to be plausible?

(d) If the claim is true, what is $P(\overline{x} \leq 38.5)$?

(e) Based on the answer to part (d), if the claim is true, is a sample mean lifetime of 38.5 hours unusually short?

(f) If the sample mean lifetime of the 100 batteries were 38.5 hours, would you find the manufacturer's claim to be plausible?

Homework 9 Name:

1. Assume that the systolic blood pressure of 30-year-old males is normally distributed with an average of 122 mmHg and a standard deviation of 10 mmHg. A random sample of 16 men from this age group is selected.

(a) Find the probability that the average blood pressure of the men in this sample is greater than 125 mmHg.

____

(b) Find the probability that the average blood pressure of the men in this sample is lower than 121 mmHg.

____

(c) Would it be unusual to have a sample mean greater than 125 mmHg?

____

2. Assume that the average weight of an NFL player is 245.7 pounds with a standard deviation of 34.5 pounds. A random sample of 32 NFL players is selected.

(a) Find the probability that the average weight of the players in this sample is less than 247 pounds.

____

(b) Find the probability that the average weight of the players in this sample is greater than 243 pounds.

____

(c) Would it be unusual if the average weight of the players in this sample were greater than 250 lbs?

____

3. A report claimed that the average annual consumption of milk in the U.S. was 23.4 gallons per person with a standard deviation of 7.1 gallons per person. A random sample of 60 Americans is selected.

(a) If the claim is true, find the probability that the average annual milk consumption for the people in this sample is between 23.2 and 23.5 gallons.

____

(b) If the claim is true, find the 70th percentile of the sample mean.

____

(c) If the sample mean were 23.9 gallons or greater, would you find the claim of this report to be plausible?

____

4. A category of runners claim to run the marathon in an average of 145 minutes with a standard deviation of 14 minutes. A random sample of 49 runners is selected.

(a) If the claim is true, find the probability that the sample mean is between 144 and 145 minutes.

____

(b) If the claim is true, find the 80th percentile of the sample mean.

____

(c) If the sample mean were 144 minutes or lower, would the claim be plausible?

____

Chapter

8 Confidence Intervals

Suppose your company makes iPhone cases, and you want to ensure their quality, specifically the dimensions. How can you check the average width, let's say, of all the cases you make, so that you know they'll fit properly?

Well, you could theoretically measure every single case, but in a big production facility, this isn't feasible, because the time and effort that it will add will cut into your profits. Instead, you can take a small sample, measure the average width in your sample, and use that to estimate the average width of your population.

We can do better, though. The sample mean is simply a **point estimate** of the population mean, but in this chapter we'll find how to come up with an interval that estimates the mean.

SECTION 8.1 One Population Mean, Normal

Point estimate:

Confidence interval:

In other words, instead of saying

> "I think the average width of our iPhone cases is 67 mm"

you could say

> "I am 95% confident that the average width of our iPhone cases is between 66.8 and 67.2 mm."

Notice how the second statement is much more precise (if less natural, perhaps). Also,

-
-
-

Constructing a Confidence Interval

Assumption: We know the population standard deviation σ. Also,

Recall: Central Limit Theorem

BATTERY LIFE EXAMPLE 1

A battery manufacturer claims that the lifetime of a certain type of battery has a population mean of $\mu = 40$ hours and a standard deviation of $\sigma = 5$ hours. Let $\overline{x}$ represent the mean lifetime of the batteries in a simple random sample of size 100.

(a) If the claim is true, what is $P(\overline{x} \leq 39.8)$?

`normalcdf(-1000000,39.8,40,0.5)` $= 0.3446$

(b) Based on the answer to part (a), if the claim is true, is a sample mean lifetime of 39.8 hours unusually short?

Not really.

(c) If the sample mean lifetime of the 100 batteries were 39.8 hours, would you find the manufacturer's claim to be plausible?

Yeah, I think so.

(d) If the claim is true, what is $P(\overline{x} \leq 38.5)$?

`normalcdf(-1000000,38.5,40,0.5)` $= 0.0013$

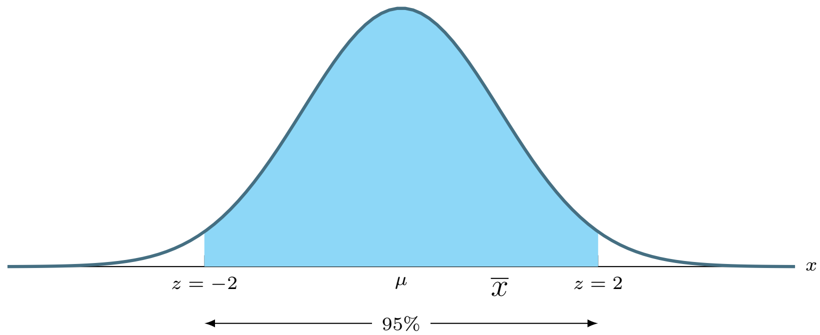
(e) Based on the answer to part (d), if the claim is true, is a sample mean lifetime of 38.5 hours unusually short?

Yes.

(f) If the sample mean lifetime of the 100 batteries were 38.5 hours, would you find the manufacturer's claim to be plausible?

No.

Finding a confidence interval essentially means finding all the values for the population mean that would not make our sample mean **unusual** (where here "unusual" depends on our confidence level).



For instance, in the sampling distribution above, we have a good idea of how likely it is that the sample mean will fall into a given range. Based on the Empirical Rule, we know that there is a 68% chance that the sample mean will be within one standard deviation of the population mean, a 95% chance that it will be within two standard deviations of the population mean, and a 99.7% chance that it will be within three standard deviations. For any other probabilities, we can consult the z table or our calculators.

Okay, let's try an example.

EXAMPLE 2 CONFIDENCE INTERVAL

If you get a sample mean of 23, and you know that the sampling distribution has standard deviation

$$\frac{\sigma}{\sqrt{n}} = 1.5,$$

find the 95% confidence interval for the population mean μ.

The population mean is unknown, but we know that whatever it is, our sample mean is 95% likely to be within two standard deviations of it (two standard deviations equals 3 in this case). The sample mean could be 3 lower or 3 higher, so our confidence interval goes from $23 - 3$ to $23 + 3$:

$$23 \pm 3 = (20, 26).$$

Either notation is acceptable for a confidence interval. Note that the point estimate is the sample mean, 23, and the margin of error is the standard deviation ($\sigma/\sqrt{n}$) times the number of standard deviations that correspond to a 95% confidence level.

Finding a confidence interval, then, consists of three pieces:

1.

2.

3.

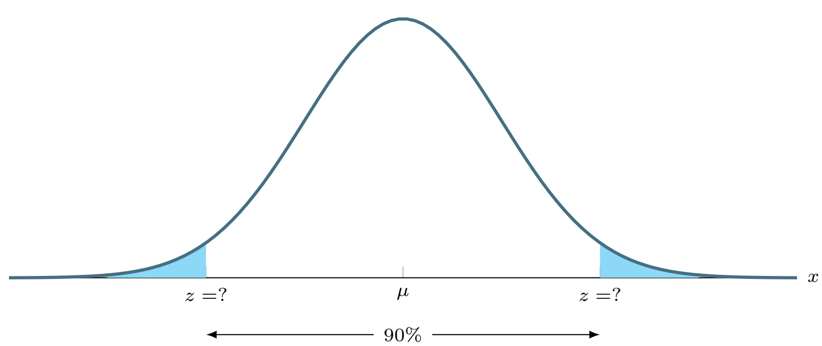
Then the confidence interval is

Finding $z_{\alpha/2}$

Okay, with a 95% confidence interval, the z value was pretty easy, because we know that 95% of the data is within two standard deviations based on the Empirical Rule. But what if we wanted a 90% confidence interval or a 99% confidence interval? The Empirical Rule has nothing to say about those, so we need to use the z table or our calculator.

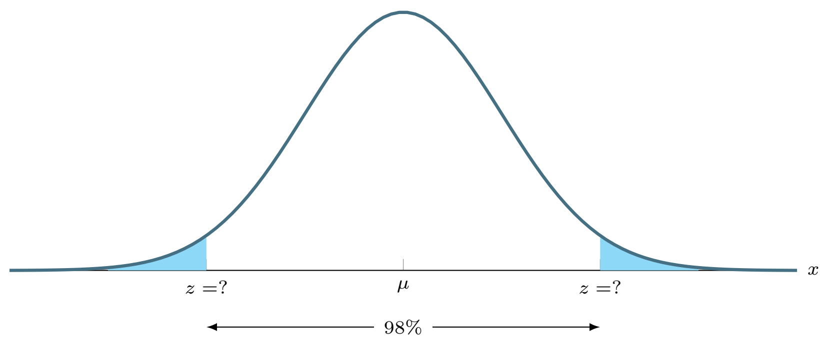
FINDING Z EXAMPLE 3

How many standard deviations do you need to go out to cover 90% of the data?



EXAMPLE 4 FINDING Z

Find $z_{\alpha/2}$ for a 98% confidence interval.



Full Examples

CEREAL BOX WEIGHT — EXAMPLE 5

A machine that fills cereal boxes is supposed to put 20 ounces of cereal in each box. A simple random sample of 6 boxes is found to contain a sample mean of 20.25 ounces of cereal. It is known from past experience that fill weights are normally distributed with a standard deviation of 0.2 ounces. Construct a 92% confidence interval for the mean fill weight.

SAT SCORES — EXAMPLE 6

A college admissions officer takes a simple random sample of 100 entering freshmen and computes their mean mathematics SAT score to be 458. Assume the population standard deviation is $\sigma = 116$. Construct a 99% confidence interval for the population mean score.

EXAMPLE 7 BABY WEIGHT

According to the National Health Statistics Reports, a sample of 360 one-year-old baby boys in the US had a mean weight of 25.5 pounds. Assume the population standard deviation is $\sigma = 5.3$ pounds. Construct a 94% confidence interval.

Changing the Confidence Level

What does changing the confidence level do?

COMPONENT LIFETIMES — EXAMPLE 8

In a simple random sample of 100 electronic components produced by a certain method, the mean lifetime was 125 hours. Assume that component lifetimes are normally distributed with population standard deviation $\sigma = 20$ hours. Construct 90%, 95%, and 98% confidence intervals.

Note: As the confidence level increases, the confidence intervals get wider (still centered at the same place).

Calculating the Sample Size

Suppose we want a given margin of error: what sample size do we need in order to make that happen?

Note: A larger sample size leads to a smaller margin of error.

EXAMPLE 9 COLLEGE STUDENT AGE

The population standard deviation for the age of students at a college is 15 years. If we want to be 95% confident that the sample mean age is within two years of the true population mean age of these students, how many randomly selected students must be surveyed?

Using Your Calculator

There is also a built-in function in your calculator that can find confidence intervals for problems like this one.

1. Press 

 and scroll over to the `TESTS` menu.

2. Select `7:ZInterval` and you'll have two options: using `Data` or `Stats`.

3. In either case, enter the population standard deviation as σ and the confidence level (as a decimal).

EXAMPLE 10 BLACKBERRY PRICES

A random sample of 11 BlackBerry Bold 9000 smartphones being sold over the Internet in 2010 had the following prices, in dollars:

230	484	379	300	239	350
300	395	230	410	460	

Assume the population standard deviation is $\sigma = 71$. Calculate a 95% confidence interval for the population mean price.

After entering the data, press 

, scroll over to `TESTS` menu, and select `7:ZInterval`. Enter $\sigma = 71$, leave `C-Level` as 0.95, and press `Calculate`. You'll see the following:

The confidence interval, then, is

$$(301.41, 385.32).$$

Homework 10 Name:

1. Find $z_{\alpha/2}$ for a 92% confidence interval.

2. Find $z_{\alpha/2}$ for a 96% confidence interval.

3. A retail store samples 40 orders and finds that the average charge is \$78.25. If the standard deviation of all their orders is known to be \$22.50, find a 90% confidence interval for the mean charge.

4. For the same retail store, find the minimum sample size needed to find a 90% confidence interval for the mean, if we want to know the average charge to within \$5.00 of the population mean.

5. A random sample of 35 teenagers found that they averaged 7.3 hours of sleep per night. Assume that the population standard deviation is 1.8 hours. Calculate a 95% confidence interval for the mean.

6. For this group of teenagers, find the minimum sample size needed to find a 95% confidence interval that will have a margin of error of 0.4 hours or less.

7. In a random sample of 15 cars of the same model, the average gas mileage is 26.7 miles per gallon. Assume that the population is normally distributed with a standard deviation of 5.2 miles per gallon. Find a 92% confidence interval for the mean.

8. The manufacturer claims that this model of car has an average gas mileage of 26 miles per gallon. Do you find this claim to be valid or not?

9. A random sample of 20 paperback novels average 425.1 pages in length. Assume that the page counts for all paperback novels are normally distributed with a standard deviation of 92.8 pages. Calculate a 99% confidence interval for the population mean.

10. Calculate the minimum sample size for this group of paperback novels that is needed to identify a 98% confidence interval with a margin of error of 52 pages or less.

SECTION 8.2 One Population Mean, Student t

What we did in the last section is rarely, if ever, done.

Note:

-

-

Instead of a CI that looks like

we'll have one know that looks like

Notice:

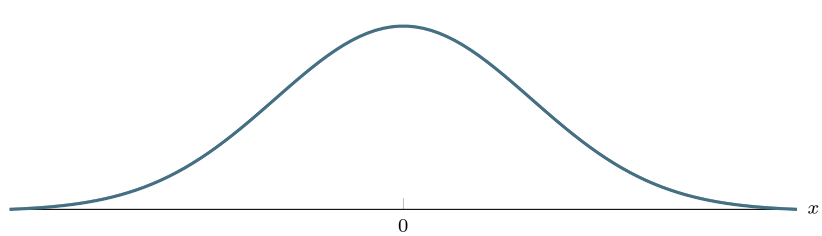
-

-

The t Distribution

The t distribution is necessary when σ is unknown and the sample size is small, but nowadays, it's used pretty much all the time (since σ is always unknown).

When the sample size is large, the t distribution is nearly indistinguishable from the normal distribution.



The t distribution:

-

-

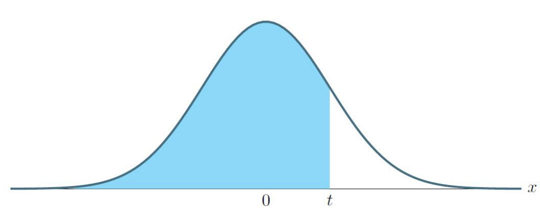
-

-

-

To find $t_{\alpha/2}$, we can use either a table or a calculator (TI-84+ and up). The table looks like this (the full table is in the appendix):

Student's t Distribution



$df = n - 1$	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501

Be careful: different books record the t table in different ways; all the results will be equal, but you may need to read the table differently.

FINDING T EXAMPLE 1

Find $t_{\alpha/2}$ to construct a 90% confidence interval based on a sample of 7 items.

FINDING T WITH A CALCULATOR EXAMPLE 2

Find the same t value using a calculator.

The TI-84+ and later models have a `invT` function located directly beneath `invNorm`.

Press 

to access the `DISTR` menu, then select 4: invT. This function requires two inputs: `area` and `df`. Enter 0.95 and 6, respectively:

The answer is the same.

Confidence Intervals with t

Now that we can find $t_{\alpha/2}$, we can find t confidence intervals:

All we have to do is

1.

2.

3.

4.

EXAMPLE 3 POTATO CHIP BAGS

A potato chip company wants to evaluate the accuracy of its potato chip bag-filling machine. Bags are labeled as containing 8 ounces of potato chips. A simple random sample of 12 bags had mean weight 8.12 ounces with a sample standard deviation of 0.1 ounce. Construct a 99% confidence interval for the population mean weight of bags of potato chips.

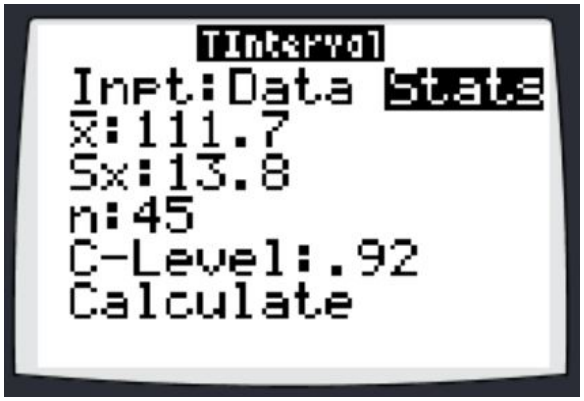
Using Your Calculator

The process for finding a t interval is identical to that for finding a z interval, except that you need to select `8: TInterval` in the `STAT TESTS` menu.

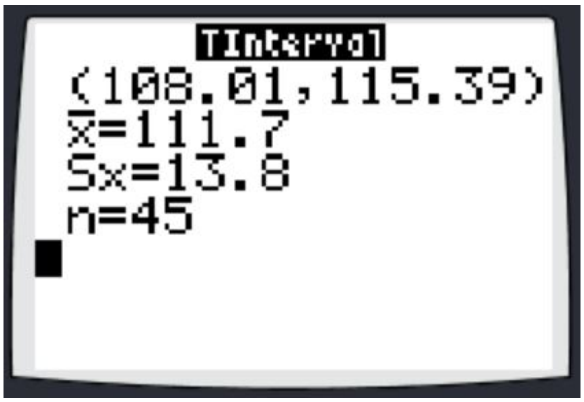
MOVIE LENGTHS EXAMPLE 4

A random sample of 45 Hollywood movies made since the year 2000 had a mean length of 111.7 minutes, with a standard deviation of 13.8 minutes. Construct a 92% confidence interval for the population mean.

On your calculator, press `STAT` and scroll over to `TESTS`. Scroll down or press the 8 key to select `8: TInterval`.



Enter the given information and click `Calculate`.



Therefore, the 92% confidence interval is

$$(108.01, 115.39).$$

EXAMPLE 5 CEREAL BOX WEIGHTS

Boxes of cereal are labeled as containing 14 ounces. Following are the weights, in ounces, of a sample of 12 boxes. It is reasonable to assume that the population is approximately normal.

14.02	13.97	14.11	14.12	14.10	14.02
14.15	13.97	14.05	14.04	14.11	14.12

Construct a 95% confidence interval. Based on this confidence interval, are the boxes labeled correctly?

EXAMPLE 6 ONLINE COURSE SATISFACTION

A sample of 263 students who were taking online courses were asked to describe their overall impression of online learning on a scale of 1–7, with 7 representing the most favorable impression. The average score was 5.53, and the standard deviation was 0.92. Construct a 99% confidence interval for the population mean score.

SECTION 8.3 One Population Proportion

This is where you may have run across the term *margin of error* before: political polls often give a percentage of voters in each category, and then state something like a margin of error of three percent. This means, of course, that the percentages could be three percent higher or lower than what is reported.

In this section, we'll deal with confidence intervals for proportions like that. In each example, there will be a sample size, and a number within that sample that respond one way. We want to know what this tells us about what proportion of the population would respond that way.

-
-
-
-
-

The confidence interval, just like before, is

The formula is more complicated than before, but the problems are actually simpler, since there's less to keep track of (all we really need is x and n).

EXAMPLE 1 ANDROID LOYALTY

The Nielsen Company surveyed 225 owners of Android phones and found that 160 of them planned to get another Android as their next phone.

(a) Construct a 95% confidence interval for the proportion of Android users who plan to get another Android.

(b) Assume that an advertisement claimed that 70% of Android users plan to get another Android. Does the confidence interval contradict this claim?

Using Your Calculator

Press the 

button and scroll over to the `TESTS` menu. Scroll down to `A: 1-PropZInt`.

When you press enter and see the menu, all you have to enter is x, n, and the confidence level.

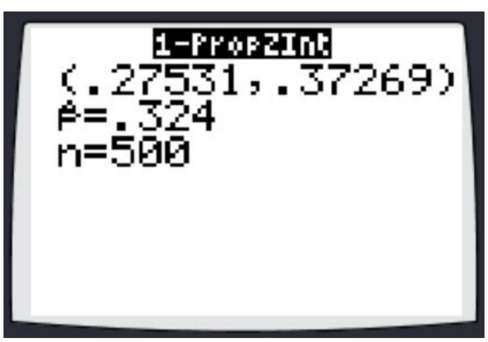
WORKING FROM HOME

EXAMPLE 2

According to the U.S Census Bureau, 43% of men who worked at home were college graduates. In a sample of 500 women who worked at home, 162 were college graduates. Construct a 98% confidence interval for the proportion of women who work at home who are college graduates. Is it reasonable to believe that this is the same as the proportion for men?

Enter the `1-PropZInt` menu:

When you press `Calculate`, you'll see



The confidence interval, then, is

$$(0.27531, 0.37269).$$

Therefore, we conclude that no, the proportion for women must be lower than the proportion for men (43%).

EXAMPLE 3 ISP QUALITY CONTROL

An Internet service provider sampled 540 customers, and found that 75 of them experienced an interruption in high-speed service during the previous month. Construct a 90% confidence interval for the proportion of all customers who experienced an interruption. The company's quality control manager claims that no more than 10% of its customers experienced an interruption during the previous month. Does the confidence interval contradict this claim?

EXAMPLE 4 HEALTH INSURANCE

In 2008, the General Social Survey asked 182 people whether they received health insurance as a benefit from their employer. A total of 60 people said they did. Construct a 95% confidence interval for the proportion of people who receive health insurance from their employer.

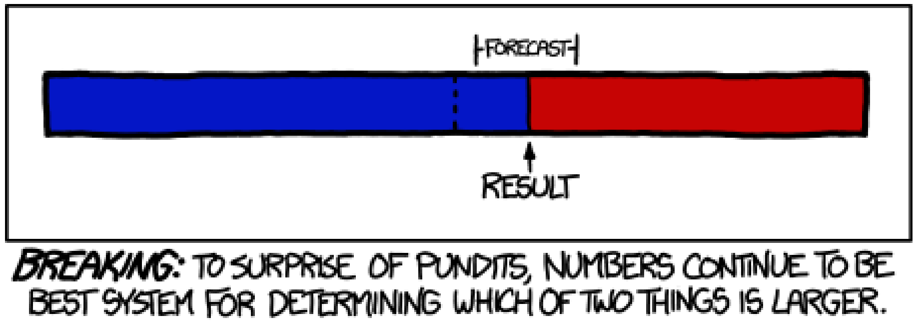
Calculating Sample Size

Note: A larger sample leads to a smaller margin of error.

What if you have a specific margin of error in mind? What sample size do you need?

FIND SAMPLE SIZE **EXAMPLE 5**

Suppose a polling company wants to ensure that their next political poll has a margin of error of three percentage points or less, using a 95% confidence level. How many voters should the company poll to make this happen?


|-FORECAST-|
RESULT
BREAKING: TO SURPRISE OF PUNDITS, NUMBERS CONTINUE TO BE BEST SYSTEM FOR DETERMINING WHICH OF TWO THINGS IS LARGER.

xkcd.com

Homework 11 Name:

1. Find $t_{\alpha/2}$ for a 95% confidence interval, based on a sample of 10 items.

2. Find $t_{\alpha/2}$ for a 98% confidence interval, based on a sample of 13 items.

3. The data set below shows the amount of trash generated by ten households (in pounds per day). Assume that the population is normally distributed.

3.9	4.6	15.6	10.5	16.0
6.7	12.0	9.2	13.8	16.8

Construct a 95% confidence interval for the mean based on the sample.

4. The table below shows the current ages of eight randomly selected aircraft passengers (in years). Assume that the population is normally distributed.

7.0	13.5	18.6	25.8
24.7	17.8	10.5	14.2

Construct a 90% confidence interval for the mean based on this sample.

5. A sample of 22 households used an average of 346.2 gallons of water per day, with a standard deviation of 50.5 gallons (note that this is a *sample* standard deviation). Assume that the population of household water usage is normally distributed. Find a 92% confidence interval for the mean based on this sample.

6. A sample of 26 mechanical engineers averaged 41.7 years of age, with a standard deviation of 6.9 years. Assume that the population age is normally distributed. Find a 98% confidence interval for the mean based on this sample.

7. A random sample of 175 registered voters revealed that 94 of them voted in the last election. Construct a 95% confidence interval to estimate the true proportion of voter turnout.

8. In this study on voting habits, how many voters must be sampled to construct a 95% confidence interval with a margin of error of 0.03 or less?

9. In a random sample of 220 men (18 or older), 139 are married. Construct a 98% confidence interval to estimate the true proportion of married men (18 or older).

10. In this study on married men, what sample size is necessary to construct a 90% confidence interval with a margin of error of 0.05 or less?

Chapter

9 Hypothesis Testing with One Sample

If a car manufacturer claims that one of their models averages more than 38 miles per gallon on the highway, how can we verify their claim? That process is called **hypothesis testing**: a claim is made (i.e. a hypothesis) and we test it.

As we'll see, hypothesis testing is closely linked to what we've already done with confidence intervals, but a hypothesis test is a way of clearly laying out the evidence that confirms or contradicts a claim like the gas mileage one.

To perform a hypothesis test, we'll describe two contradictory hypotheses (like guilty and not guilty in a criminal trial), and based on the evidence, we'll make a decision in favor of one of them.

SECTION 9.1 Null and Alternative Hypotheses

Remember this example from the section on confidence intervals?

EXAMPLE 1 CEREAL BOX WEIGHT

A machine that fills cereal boxes is supposed to put 20 ounces of cereal in each box. A simple random sample of 6 boxes is found to contain a sample mean of 20.25 ounces of cereal. It is known from past experience that fill weights are normally distributed with a standard deviation of 0.2 ounces. Construct a 92% confidence interval for the mean fill weight.

Confidence interval:

$$(20.12, 20.38)$$

At the end of the problem, we have a confidence interval, but we also have a conclusion about the claim that was made: we can conclude that the average weight of the boxes is *more than* 20 ounces, since the entire interval is above 20. If we'd gotten something like

$$(19.88, 20.56)$$

we would not have been able to conclude that the average weight is more than 20 ounces or less than 20 ounces.

Note:

If we did a hypothesis test with this example, we'd find (with 92% confidence) the same conclusion. Again, a hypothesis test is a different way to go about it, but the hypothesis test and the confidence interval will draw the same conclusions.

Kinds of Hypothesis Tests

There are many hypothesis tests that can be done, but we'll stick to ones that are similar to what we did with confidence intervals:

-
-
-
-
-

In general, a hypothesis test tests a claim about a population parameter based on a sample.

Hypotheses

At the heart of a hypothesis test are the two contradictory hypotheses.

- **Null hypothesis:**
- **Alternate hypothesis:**

Note: Terminology At the end of a hypothesis test, we'll either say

or

We'll never say

Options: Compare the parameter (mean or proportion) to some value

	H_0		H_1
$=$	Equal	$\neq$	Not equal (greater than or less than)
$\geq$	Greater than or equal to	$<$	Less than
$\leq$	Less than or equal to	$>$	Greater than

Note: the equals sign is always on the null hypothesis. Some people always use $=$ as the null hypothesis in every case.

EXAMPLE 2 RESTAURANT BILLS

Last year, the mean amount spent by customers at a restaurant was \$35. The restaurant owner believes that the mean may be higher this year.

NEWBORN WEIGHT — EXAMPLE 3

In a recent year, the mean weight of newborn boys in a certain country was 6.6 pounds. A doctor wants to know whether the mean weight of newborn girls differs from this.

GAS MILEAGE — EXAMPLE 4

A certain model of car can be ordered with either a large or small engine. The mean number of miles per gallon for cars with a small engine is 25.5. An automotive engineer thinks that the mean for cars with the larger engine will be less than this.

REGISTERED VOTERS — EXAMPLE 5

A pollster thinks that less than 30% of registered voters in the county voted.

EXAMPLE 6 MEAN GPA

We want to test whether the mean GPA of American college students differs from 2.0.

EXAMPLE 7 PLACEMENT TESTS

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. Test if the percentage of U.S. students who take advanced placement exams differs from 6.6%.

EXAMPLE 8 DRIVER'S TEST

On a state driver's test, about 40% pass on the first try. We want to test if more than 40% pass on the first try in a different state.

SECTION 9.2 Type I and Type II Errors

Since H_0 and H_1 are contradictory, one (and only one) of them must be true.

-
-
-
-

Medical test analogy: H_0 is that you don't have a disease

- **Type I Error:**
- **Type II Error:**

Trial analogy: H_0 is that you're innocent (innocent until proven guilty)

- **Type I Error:**
- **Type II Error:**

Summary:

EXAMPLE 1 ERROR TYPES

Suppose Frank tests his rock-climbing equipment, and H_0 is that his equipment is safe.

Type I Error:

Type II Error:

EXAMPLE 2 ERROR TYPES

The victim of a car accident is brought to the emergency room, and H_0 is that she is alive when she comes in.

Type I Error:

Type II Error:

SECTION 9.3 Distribution Needed for Testing

The distributions we'll use for hypothesis testing are the same ones we used for confidence intervals:

-

-

-

SECTION 9.4 Drawing a Conclusion

Example from textbook: Didi and Ali are at a birthday party of a very wealthy friend. They hurry to be first in line to grab a prize from a tall basket that they cannot see inside. There are 200 plastic bubbles in the basket and they have been told that there is only one with a $100 bill. Didi is the first person to reach into the basket and pull out a bubble. Her bubble contains a $100 bill.

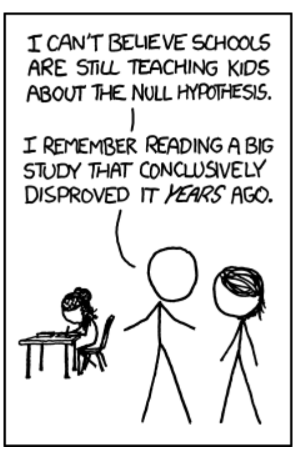
If the claim were true, the probability of this happening would be $1/200 = 0.005$, a very unlikely thing. Because a "rare event" has occurred, they begin to doubt that the information they were given was true. (In reality, they would weigh this against the probability that the person who told them this was lying, and if they trusted the person, they wouldn't doubt their word, because they'd assume that the probability of that person lying was even lower than 0.005).

This is similar to a hypothesis test: we make an **assumption** (that may or may not be true). Then we take a sample.

-

-

The p Value



xkcd.com

The way that we measure an rare event like this is by using a probability called the **p value**.

-

-

-

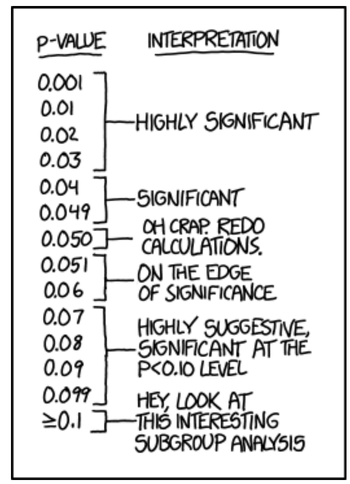
What is "low"? We consider p to be low if it is below some predetermined **significance level**, called α. This is usually 0.05 or something similarly low.

•

YEARS OF EDUCATION

EXAMPLE 1

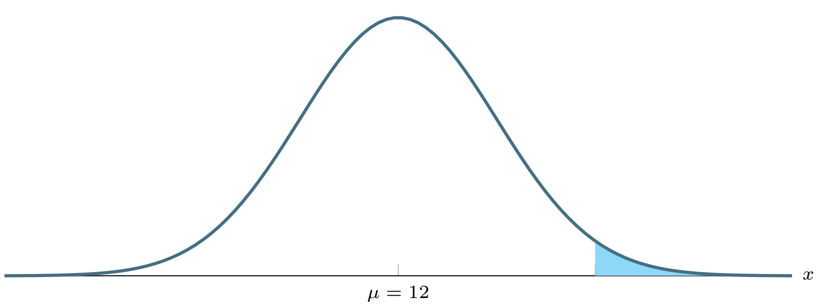
A social scientist suspects that the mean number of years of education for adults in a certain large city is greater than 12 years. She surveys 100 adults and finds that the sample mean number of years is 12.98. Assume that the population standard deviation is 3 years. Test this claim.



xkcd.com

Step 1: State the hypotheses.

Step 2: Calculate the test statistic. This is the z score of our sample, which gives an idea of how unusual our sample is, assuming that the true population mean is 12 or less (our null hypothesis).



Note that we shaded the area to the right of the sample mean, because the claim is that the mean is **greater**.

Step 3: Calculate the p value that corresponds to this area. Use the table or calculator.

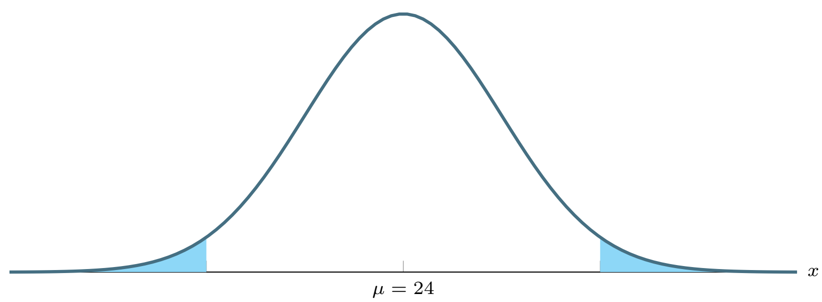
Step 4: Draw a conclusion.

EXAMPLE 2 TEST SCORES

A pre-test and post-test were given to workshop attendees. The pretest score average was 24, and the researchers want to know whether the post-test score is significantly different from the pre-test score. They sampled 50 tests and found that the sample mean was 24.8. Assume that the population standard deviation is 1.2. Use a significance level of 0.01.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.



Note that we shaded the area outside the sample mean and on the opposite side, since the claim is that it is different from 24 (greater or smaller). This is called a **two-tailed test**.

Step 3: Calculate the p value that corresponds to this area. Use the table or calculator. Remember to multiply by 2.

Step 4: Draw a conclusion.

SECTION 9.5 Full Examples

We'll do these four steps in every example in this section:

Step 1:

Step 2:

Step 3:

Step 4:

We'll show three types of examples:

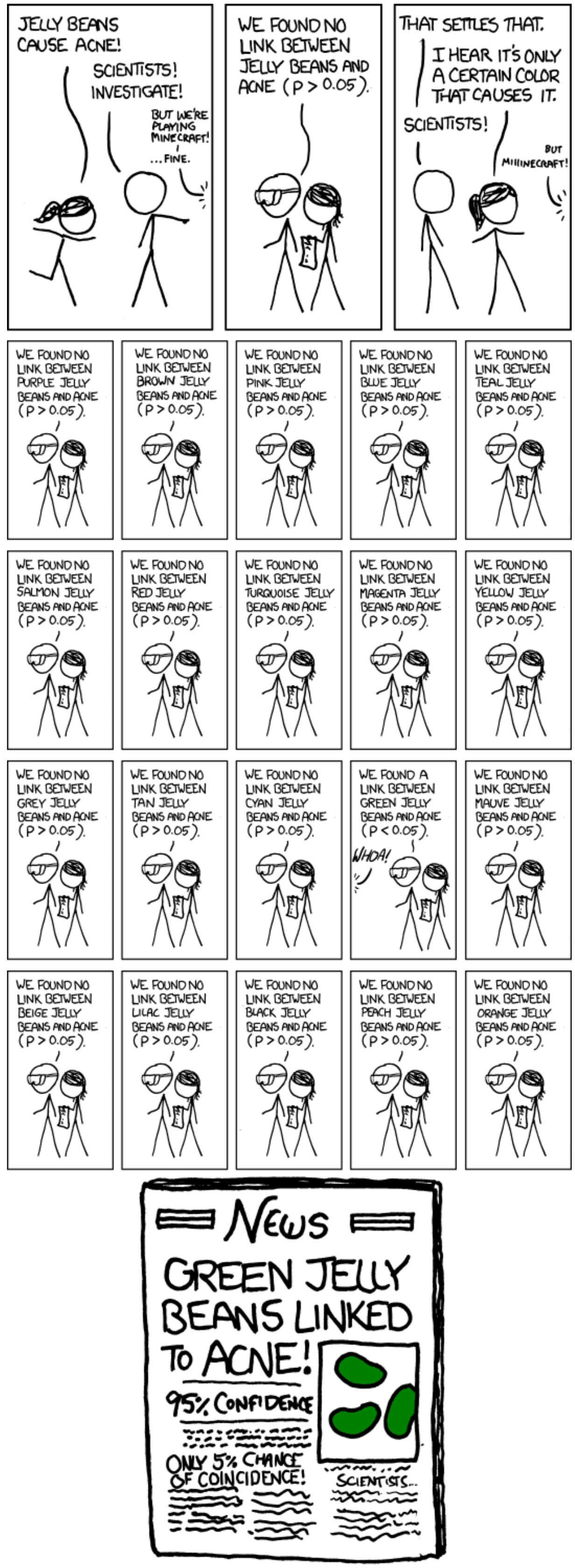
1.

2.

3.

Then we'll mix up a bunch of examples in order to get practice with deciding what kind of problem we're up against.

Note:


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

xkcd.com

Mean, Population Standard Deviation Known

FACEBOOK TIME — EXAMPLE 1

A study by the Web metrics firm Hitwise showed that in August 2008, the mean time spent per visit to Facebook was 19.5 minutes. Assume the standard deviation of the population is 8 minutes. Suppose that a simple random sample of 100 visits in August 2009 has a sample mean of 21.5 minutes. A social scientist is interested in knowing whether the mean time of Facebook visits has increased. Conduct a hypothesis test to determine this. Use a significance level of 0.05.

EXAMPLE 2 AVERAGE MALE HEIGHT

According to the National Health Statistics Reports, the mean height for U.S. men is 69.4 inches, and the population standard deviation is 2.84. In a sample of 300 men between the ages of 60 and 69, the mean height was 69.0 inches. Public health officials want to determine whether the mean height for older men is less than the mean height of all men. Conduct a hypothesis test to answer this question.

CHILD WEIGHT EXAMPLE 3

Are children heavier now than they were in the past? The National Health and Nutrition Examination Survey taken between 1999 and 2002 reported that the mean weight of six-year-old girls in the U.S. was 49.3 pounds. Another NHANES survey, published in 2008, reported that a sample of 193 six-year-old girls weighed between 2003 and 2006 had an average weight of 51.5 pounds. Assume that the population standard deviation is 15 pounds. Can you conclude that the mean weight of six-year-old girls is higher in 2006 than in 2002? Use a significance level of 0.01.

Mean, Population Standard Deviation Unknown

EXAMPLE 4 FAMILY PRACTITIONER SALARY

The Bureau of Labor Statistics reported that in May 2009, the mean annual earnings of all family practitioners in the United States was \$168,550. A random sample of 55 family practitioners in Missouri that month had mean earnings of \$154,590 with a standard deviation of \$42,750. Do the data provide sufficient evidence to conclude that the mean salary for family practitioners in Missouri is less than the national average? Use the $\alpha = 0.05$ level of significance.

BABY BOY WEIGHT EXAMPLE 5

The National Health Statistics Reports described a study in which a sample of 360 one-year-old baby boys were weighed. Their mean weight was 25.5 pounds with standard deviation 5.3 pounds. A pediatrician claims that the mean weight of one-year-old boys is greater than 25 pounds. Do the data provide convincing evidence that the pediatrician's claim is true? Use the $\alpha = 0.01$ level of significance.

EXAMPLE 6 COMMUTE TIME

A 2007 Gallup poll sampled 1019 people, and asked them how long it took them to commute to work each day. The sample mean one-way commute time was 22.8 minutes with a standard deviation of 17.9 minutes. A transportation engineer claims that the mean commute time is greater than 20 minutes. Do the data provide convincing evidence that the engineer's claim is true? Use the $\alpha = 0.05$ level of significance.

Proportion

JOB SATISFACTION EXAMPLE 7

A nationwide survey of working adults indicates that only 50% of them are satisfied with their jobs. The president of a large company believes that more than 50% of employees at his company are satisfied with their jobs. To test his belief, he surveys a random sample of 100 employees, and 54 of them report that they are satisfied with their jobs. Can he conclude that more than 50% of employees at the company are satisfied with their jobs? Use the $\alpha = 0.05$ level of significance.

EXAMPLE 8 SPAM

According to MessageLabs Ltd., 89% of all email sent in July 2010 was spam. A system manager at a large corporation believes that the percentage at his company may be 80%. He examines a random sample of 500 emails received at an email server and finds that 382 of the messages are spam. Using a significance level of $\alpha = 0.05$, can you conclude that the percentage of emails that are spam differs from 80%?

CHILDREN WITH CELL PHONES EXAMPLE 9

A marketing manager for a cell phone company claims that more than 35% of children aged 10–11 have cell phones. In a 2009 survey of 5000 children aged 10–11 by Mediamark Research and Intelligence, 1805 of them had cell phones. Can you conclude that the manager's claim is true? Use the $\alpha = 0.01$ level of significance.

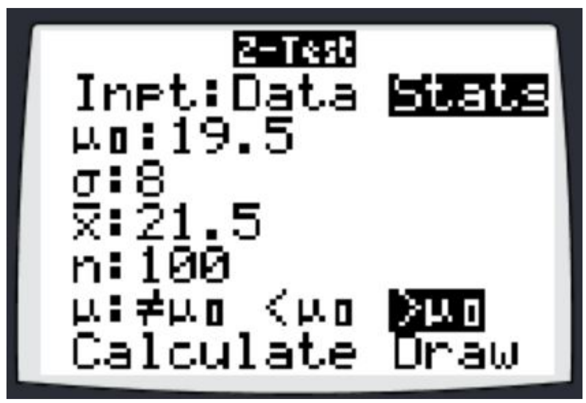
Using Your Calculator

To access the hypothesis tests on the TI calculator, press 

and scroll over to the `TESTS` menu:

1. **Mean, Population Standard Deviation Known**

Use `1: Z-Test`.



(enter given stats and select the appropriate alternate hypothesis)

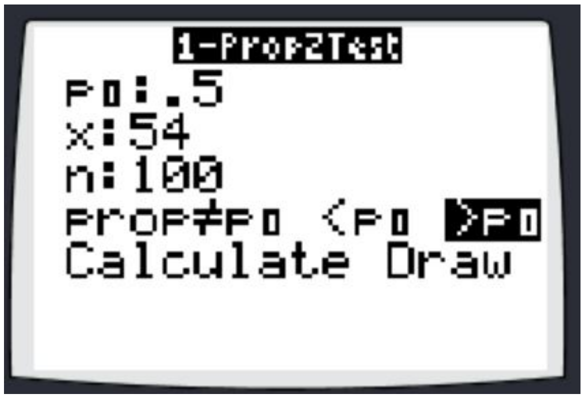
(enter the data into `L1` and select the appropriate alternate hypothesis)

2. **Mean, Population Standard Deviation Unknown**

   Use `2: T-Test`, and do it the same way as the `Z-Test`.

3. **Proportion**

   Use `5: 1-PropZTest`.



Assorted Examples

EXAMPLE 10 TIME WATCHING TV

In 2008, the General Social Survey asked a sample of 1324 people how much time they spent watching TV each day. The mean number of hours was 2.98 with a standard deviation of 2.66. A sociologist claims that people watch a mean of 3 hours of TV per day. Do the data provide sufficient evidence to disprove the claim? Use the $\alpha = 0.01$ level of significance.

SCALE CALIBRATION

EXAMPLE 11

NIST is the National Institute of Standards and Technology. Suppose that NIST technicians are testing a scale by using a weight known to weigh exactly 1000 grams. They weigh this weight on the scale 50 times and read the result each time, finding a sample mean of 1000.6 grams. If the standard deviation is known to be 2 grams, perform a hypothesis test to determine whether the scale is out of calibration. Use a significance level of 0.05.

EXAMPLE 12 ENVIRONMENTAL INTEREST

In 2008, the General Social Survey asked 1493 U.S. adults to rate their level of interest in environmental issues. Of these, 751 said that they were "very interested." Does the survey provide convincing evidence that more than half of U.S. adults are very interested in environmental issues? Use the $\alpha = 0.05$ level of significance.

TIRE LIFETIMES

EXAMPLE 13

A particular brand of tires claims that its deluxe tire averages more than 50,000 miles before it needs to be replaced. From past studies of this tire, the standard deviation is known to be 8,000. A survey of owners of that tire design is conducted. From the 28 tires surveyed, the mean lifespan was 46,500 miles. Using $\alpha = 0.05$, is the data consistent with the claim?

EXAMPLE 14 AGE OF SMOKERS

From generation to generation, the mean age when smokers first start to smoke varies. However, the standard deviation of that age remains constant at 2.1 years. A survey of 40 smokers of this generation was done to see if the mean starting age is greater than 19, and the sample mean was 18.1. Do the data support the claim at the 5% level?

GASTROPLASTY EXAMPLE 15

Vertical banded gastroplasty is a surgical procedure that reduces the volume of the stomach in order to produce weight loss. In a recent study, 82 patients with Type 2 diabetes underwent this procedure, and 59 of them experienced a recovery from diabetes. Does this study provide convincing evidence that more than 60% of those with Type 2 diabetes who undergo this surgery will recover from diabetes? Use the $\alpha = 0.05$ level of significance.

EXAMPLE 16 SICK DAYS

The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees, and the number of sick days they took for the past year are as follows:

$$12, 4, 15, 3, 11, 8, 6, 8.$$

Should the personnel team believe that the mean number is ten?

CABLE TV CHANNEL

EXAMPLE 17

A telecom company provided its cable TV subscribers with free access to a new sports channel for a period of one month. It then chose a sample of 400 television viewers and asked them whether they would be willing to pay an extra \$10 per month to continue to access the channel. A total of 25 of the 400 replied that they would be willing to pay. The marketing director of the company claims that more than 5% of all its subscribers would pay for the channel. Can you conclude that the director's claim is true? Use the $\alpha = 0.01$ level of significance.

EXAMPLE 18 TROUT IQ

A Nissan ad read, "The average man's IQ is 107. The average brown trout's IQ is 4. So why can't a man catch a brown trout?" Suppose you believe that the brown trout's mean IQ is greater than four. You catch 12 brown trout, and a fish psychologist determines that their IQs are

$$5, 4, 7, 3, 6, 4, 5, 3, 6, 3, 8, 5.$$

Conduct a hypothesis test of your belief.

Homework 12 Name:

1. A realtor claims that houses in a particular community average less than 90 days on the market. A random sample of 9 homes averaged 77.4 days on the market with a sample standard deviation of 29.6 days. Assume that the population is normally distributed, and test this realtor's claim using a significance level of 0.05.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

2. A PC manufacturer claims that its laptop batteries average more than 3.5 hours of use per charge. A sample of 45 batteries last an average of 3.72 hours. Assume that the population standard deviation is 0.7 hours, and test this claim at the $\alpha = 0.10$ significance level.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

3. A company claims that the average time a customer waits on hold is less than 5 minutes. A sample of 35 customers has an average wait time of 4.78 minutes. Assume that the population standard deviation is 1.8 minutes, and test this claim at the $\alpha = 0.05$ level of significance.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

4. A government bureau claims that more than 50% of US tax returns were filed electronically last year. A random sample of 150 tax returns for last year contained 86 that were filed electronically. Test this claim at the $\alpha = 0.10$ level of significance.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

5. A researcher is testing the claim that the average adult consumes 1.7 cups of coffee per day. A sample of 30 adults averaged 1.85 cups of coffee per day. Assume that the population standard deviation is 0.5 cups per day, and test this claim using a significance level of 0.05.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

6. A nationwide poll claims that the president's approval rating is below 64%. In a random sample of 125 people, 74 people gave the president a positive rating. Test this claim using a significance level of 0.02.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

7. A writer claims that the average cost for a family of four to attend an MLB game is not equal to $172. A random sample of 22 families reported an average cost of $189.34, with a sample standard deviation of $33.65. Assume that the population is normally distributed, and test this claim using a significance level of 0.10.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

8. Test a claim that the proportion of US households that watched the Super Bowl on TV is not 40%. In a random sample, 72 of 140 households watched the most recent Super Bowl. Use a significance level of 0.05 to test this claim.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

9. An auditor claims that the average annual salary of a project manager at a construction company exceeds $82,000. A random sample of 20 project managers had an average salary of $89,600, with a sample standard deviation of $12,700. Assume the salaries are normally distributed, and test this claim using a 0.01 significance level.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

10. An insurance company claims that the average automobile on the road today is less than 6 years old. A random sample of 15 cars had an average age of 5.4 years with a sample standard deviation of 1.1 years. Assume that the population is normally distributed, and test this claim using a significance level of 0.05.

Step 1: State the hypotheses.

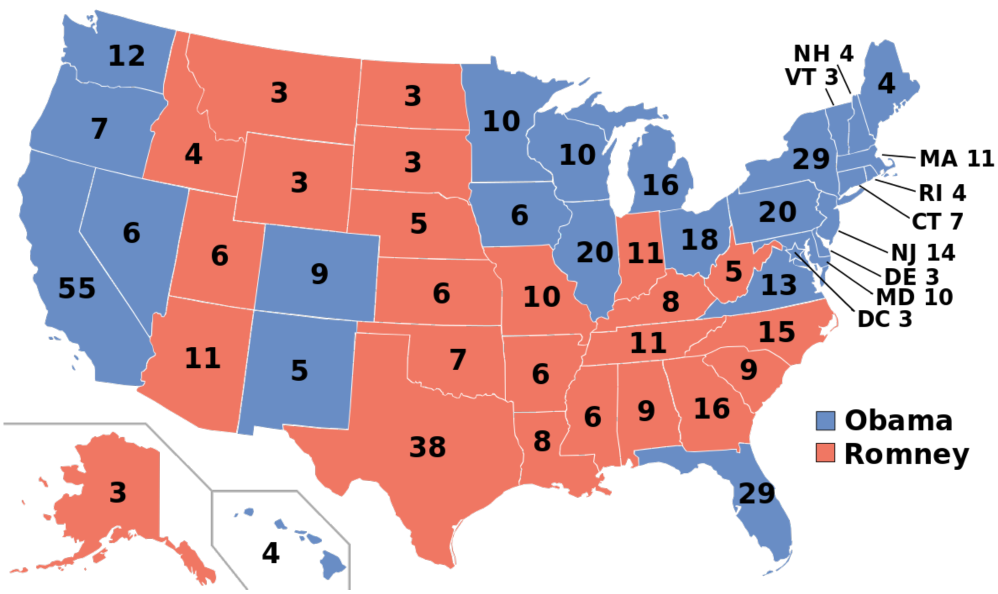
Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

Chapter

10 Hypothesis Testing with Two Samples



So far, all the hypothesis tests we've done have been to determine something about the mean or proportion in a single population; in this chapter, we briefly discuss how to compare two populations by comparing their means or proportions. For instance, we may want to compare the proportion of voters that voted Democrat in two different states. Of course, we could simply compare the sample proportions for a sample from each state, but the hypothesis tests here will give us a way to tell if there is a significant difference between them.

The formulas in this chapter are more complicated, so we'll pretty much stick to the calculator; we'll use more of the tests in this menu:

SECTION 10.1 Two Means, Sigmas Unknown

We'll use the same four steps as every hypothesis test:

Step 1: State the hypotheses.

H_0	H_1
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$

Step 2: Calculate the test statistic.

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

$$df = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\left(\dfrac{1}{n_1 - 1}\right)\left(\dfrac{s_1^2}{n_1}\right)^2 + \left(\dfrac{1}{n_2 - 1}\right)\left(\dfrac{s_2^2}{n_2}\right)^2}$$

Step 3: Calculate the p value.

`tcdf(-1000000,t,df)` or similar

Step 4: Draw a conclusion.

- If $p < \alpha$, reject H_0.
- If $p > \alpha$, fail to reject H_0.

Using Your Calculator

Since the population standard deviation is *unknown*, use the `2-SampTTest` in the `TESTS` menu. You can either enter the raw data or the summary statistics.

In either case, make sure to keep the two populations separate, enter the appropriate alternate hypothesis, and leave the `Pooled` option as `No`.

Press `Calculate` to see the t and p values, or press `Draw` to see a sketch of the distribution, with the appropriate area shaded.

COMPARING DIETS EXAMPLE 1

Are low-fat diets or low-carb diets more effective for weight loss? A sample of 77 subjects went on a low-carbohydrate diet for six months. At the end of that time, the sample mean weight loss was 4.7 kilograms with a sample standard deviation of 7.16 kilograms. A second sample of 79 subjects went on a low-fat diet. Their sample mean weight loss was 2.6 kilograms with a standard deviation of 5.90 kilograms. Can you conclude that the mean weight loss differs between the two diets? Use the $\alpha = 0.01$ level.

EXAMPLE 2 BIRTH ORDER AND IQ

In a study of birth order and intelligence, IQ tests were given to 18- and 19-year-old men to estimate the size of the difference, if any, between the mean IQs of firstborn sons and secondborn sons. The following data for 10 firstborn sons and 10 secondborn sons are consistent with the means and standard deviations reported in the article. It is reasonable to assume that the samples come from populations that are approximately normal.

Firstborn				
104	82	102	96	129
89	114	107	89	103

Secondborn				
103	103	91	113	102
103	92	90	114	113

Can you conclude that there is a difference in mean IQ between firstborn and secondborn sons? Use the $\alpha = 0.01$ level.

POSTSURGICAL TREATMENT — EXAMPLE 3

A new postsurgical treatment was compared with a standard treatment. Seven subjects received the new treatment, while seven others (the controls) received the standard treatment. The recovery times, in days, are given below.

Treatment:	12	13	15	19	20	21	24
Control:	18	23	24	30	32	35	39

Can you conclude that the mean recovery time for those receiving the new treatment is less than the mean for those receiving the standard treatment? Use the $\alpha = 0.05$ level.

EXAMPLE 4 KING TUT'S CURSE

King Tut was an ancient Egyptian ruler whose tomb was discovered and opened in 1923. Legend has it that the archaeologists who opened the tomb were subject to a "mummy's curse," which would shorten their life spans. A team of scientists conducted an investigation of the mummy's curse. They reported that the 25 people exposed to the curse had a mean life span of 70.0 years with a standard deviation of 12.4 years, while a sample of 11 Westerners in Egypt at the time who were not exposed to the curse had a mean life span of 75.0 years with a standard deviation of 13.6 years. Assume that the populations are approximately normal. Can you conclude that the mean life span of those exposed to the mummy's curse is less than the mean of those not exposed? Use the $\alpha = 0.05$ level.

SECTION 10.2 Two Means, Sigmas Known

Step 1: State the hypotheses.

H_0	H_1
$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$
$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$
$\mu_1 \geq \mu_2$	$\mu_1 < \mu_2$

Step 2: Calculate the test statistic.

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

Step 3: Calculate the p value.

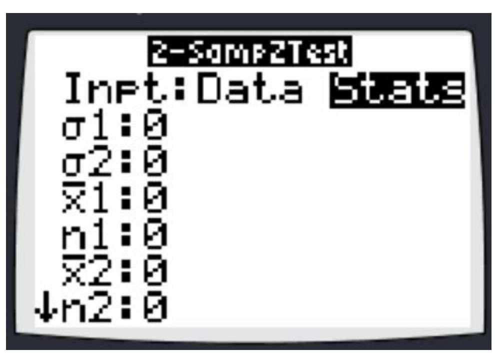
`normalcdf(-1000000,z,0,1)` or similar

Step 4: Draw a conclusion.

- If $p < \alpha$, reject H_0.
- If $p > \alpha$, fail to reject H_0.

Using Your Calculator

Since the population standard deviation is *known*, use the `2-SampZTest` in the `TESTS` menu. You can either enter the raw data or the summary statistics.



Again, make sure to keep the two populations separate, enter the appropriate alternate hypothesis, and selecte either `Calculate` or `Draw`.

EXAMPLE 1 COMPARING TWO ENGINES

The mean RPMs of two competing engines are to be compared. Ten engines of each type are randomly assigned to be tested. Both populations have normal distributions, and the following table summarizes the details.

Engine	Sample Mean RPM	Population Standard Deviation
1	1500	80
2	1600	90

Do the data indicate that Engine 2 has a higher mean RPM than Engine 1? Test at a 5% level of significance.

SECTION 10.3 Two Proportions

First of all, to conduct the test for proportions, we use what's called a **pooled proportion**:

$$p_{pooled} = \frac{x_1 + x_2}{n_1 + n_2}$$

Step 1: State the hypotheses.

H_0	H_1
$p_1 = p_2$	$p_1 \neq p_2$
$p_1 \leq p_2$	$p_1 > p_2$
$p_1 \geq p_2$	$p_1 < p_2$

Step 2: Calculate the test statistic.

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{p_{pooled}(1 - p_{pooled})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Step 3: Calculate the p value.

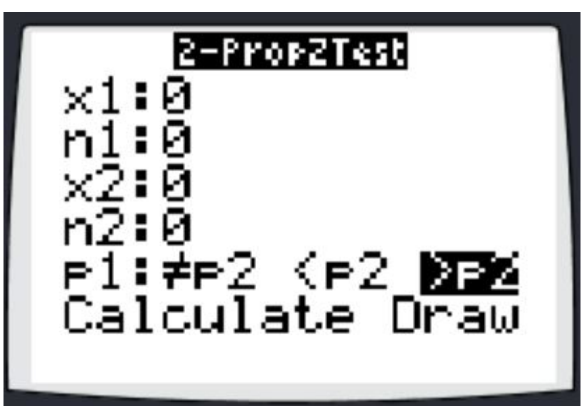
`normalcdf(-1000000,z,0,1)` or similar

Step 4: Draw a conclusion.

- If $p < \alpha$, reject H_0.
- If $p > \alpha$, fail to reject H_0.

Using Your Calculator

Use the `2-PropZTest` in the `TESTS` menu. All you have to enter is `x` and `n` for each group; make sure to keep the two groups straight. Then enter the appropriate alternate hypothesis and select either `Calculate` or `Draw`.



EXAMPLE 1 CHILDHOOD OBESITY

The National Health and Nutrition Examination Survey (NHANES) weighed a sample of 546 boys aged 6–11 and found that 87 of them were overweight. They weighed a sample of 508 girls aged 6–11 and found that 74 of them were overweight. Can you conclude that the proportion of boys who are overweight differs from the proportion of girls who are overweight?

POLLUTION AND ALTITUDE — EXAMPLE 2

In a random sample of 340 cars driven at low altitudes, 46 of them exceeded a standard of 10 grams of particulate pollution per gallon of fuel consumed. In an independent random sample of 85 cars driven at high altitudes, 21 of them exceeded the standard. Can you conclude that the proportion of high-altitude vehicles exceeding the standard is greater than the proportion of low-altitude vehicles exceeding the standard? Use the $\alpha = 0.01$ level of significance.

EXAMPLE 3 PREVENTING HEART ATTACKS

Medical researchers performed a comparison of two drugs, clopidogrel and ticagrelor, which are designed to reduce the risk of heart attack or stroke in coronary patients. A total of 6676 patients were given clopidogrel, and 6732 were given ticagrelor. Of the clopidogrel patients, 668 suffered a heart attack or stroke within one year, and of the ticagrelor patients, 569 suffered a heart attack or stroke. Can you conclude that the proportion of patients suffering a heart attack or stroke is less for ticagrelor? Use the $\alpha = 0.01$ level.

Homework 13 Name:

1. In a study on wages at day-care centers, two samples are taken, one of 52 workers from the Northeast and the other of 38 workers from the Southeast. The average hourly wages in the Northeast were \$9.60, and \$8.40 in the Southeast. Assume that the Northeast population standard deviation is \$1.25, and the Southeast population standard deviation is \$1.30. Test the hypothesis that the average hourly wage in the Northeast is higher than the average hourly wage in the Southeast, using a significance level of 0.02.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

2. The table below records data on customers' bills at a restaurant when different types of music were played. Assume that the populations are normally distributed.

	Fast Music	Slow Music
Sample Mean	\$39.65	\$42.60
Sample Size	18	23
Population Standard Deviation	\$4.21	\$5.67

Test the hypothesis that the average bill of customers exposed to fast music differs from the average bill of customers exposed to slow music, using a significance level of 0.10.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

3. In a sample of 75 flights on Airline A, 17 arrived late. In a sample of 85 flights on Airline B, 30 arrived late. Test the claim that the proportion of late flights is higher for Airline B than for Airline A.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

4. The table below shows the average costs of seven-day cruises to Alaska and the Caribbean based on a random sample of various cruise lines. Assume that the populations are normally distributed.

	Alaska	Caribbean
Sample Mean	\$884	\$702
Sample Size	8	7
Sample Standard Deviation	\$135	\$120

Test the hypothesis that the average seven-day cruise to Alaska is more expensive than the average seven-day cruise to the Caribbean, using a significance level of 0.01.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

5. The table below shows the results of a taste test between competing soda brands Cola A and Cola B. Assume that the populations are normally distributed.

	Cola A	Cola B
Sample Mean	7.92	7.22
Sample Size	38	45
Sample Standard Deviation	2.7	1.4

Test the hypothesis that Cola A is preferred over Cola B, using a significance level of 0.05.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

6. A sample of 400 Florida residents contained 272 home owners, and a sample of 600 Maryland residents contained 390 home owners. Test the claim that the proportion of home ownership in Florida exceeds that in Maryland, using a significance level of 0.02.

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

Chapter

11 Chi-Square Distribution: Goodness-of-Fit

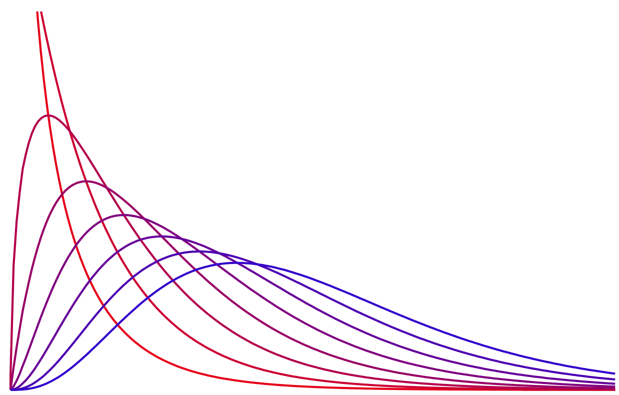
If you opened a bag of M&Ms, you may expect that the number of candies of each color is about equal. But how can you check this assumption? Naturally, you could separate them by color, then count the ones of each color, and see if they're equal. The issue with that, of course, is that you likely wouldn't find exactly the same number of each, so how much variation would you be willing to accept before you conclude that your assumption was wrong?

Like so many questions in statistics, we understand that there is inherent variability, and we need a way to distinguish between small variations and significant deviations: in this case, we'll use a **goodness-of-fit test**. First, though, we'll have to learn a bit about a new distribution: the **chi-square distribution**.

SECTION 11.1 The Chi-Square Distribution

It may be helpful to think back to the normal distribution and the t distribution as we meet the chi-square distribution (typically written χ^2-distribution, as χ is the Greek letter chi, pronounced 'kai').

-
-
-



-
-

Calculation

To use the calculator to find areas under the graph of the χ^2-distribution,

-

-

-

-

With that, we're ready to start testing goodness-of-fit.

SECTION 11.2 Testing Goodness of Fit

Remember, our goal now is to test how well data matches our expectation, specifically regarding how the data breaks down for a number of categories.

Let's go back to the M&M example. Suppose you opened a bag of 600 M&Ms, separated them by color, and got the following counts for each category.

Color	Number
Blue	212
Orange	147
Green	103
Red	50
Yellow	46
Brown	42

You may already see evidence that the candies are not uniformly distributed, but how can we measure this?

The Chi-Square Test Statistic

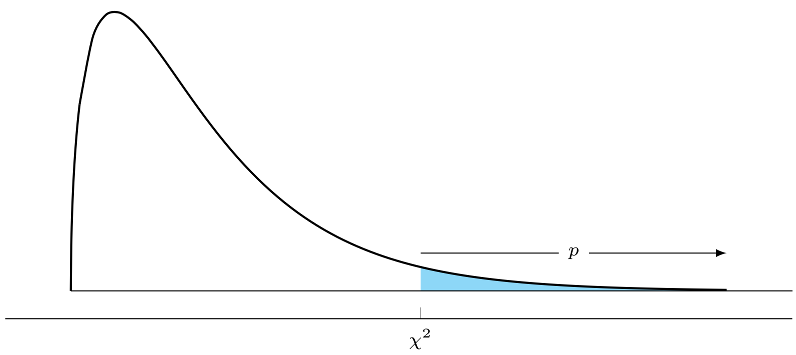
The test statistic we use to measure how closely data matches our expected distribution is found by calculating how far off each category is, then combining all those errors:

Or more concisely,

Now, the larger this value, the more difference there is between what we expected and what we observe. Therefore, the larger χ^2 is, the less likely it is that our expected distribution was correct.

The p Value

The p value for this test will be the probability that χ^2 could be at least as large as what we observe. So if we get a large value for χ^2, that probability will be relatively low (recall the shape of the χ^2 distribution):



Testing Goodness of Fit

Now that we have our test statistic (χ^2) and we can calculate a p-value, the rest of the test is just like the others that we've done.

-

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value (the area above the χ^2 statistic:

Step 4: Draw a conclusion.

-
-

EXAMPLE 1 WORK ABSENCES

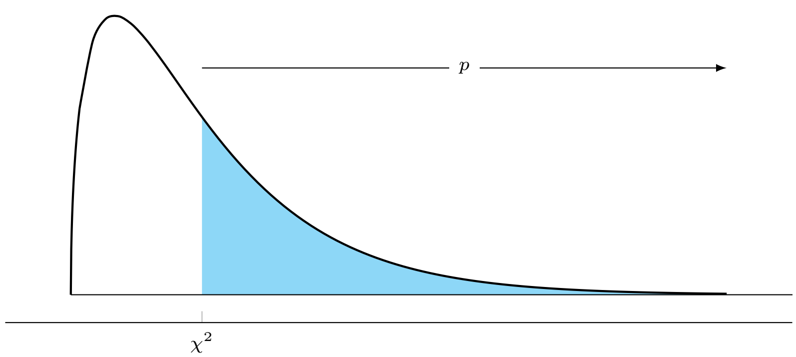
A managers wants to know which days of the week her employees are absent in a five-day work week. Most employers would like to believe that employees are absent equally during the week. She tracked the next 60 absences and recorded on which day they occurred, and the results are shown in the table below. Do the absences occur with equal frequencies during the work week? Use a significance level of 0.05.

Day	Number of Absences
Monday	15
Tuesday	12
Wednesday	9
Thursday	9
Friday	15

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Day	Observed Absences	Expected Absences	$O - E$
Monday	15		
Tuesday	12		
Wednesday	9		
Thursday	9		
Friday	15		



Step 3: Calculate the p value that corresponds to this area.

Step 4: Draw a conclusion.

EXAMPLE 2 TV OWNERSHIP

A study concluded that the number of TVs in American households is distributed as in the table below.

Number of TVs	Percent
0	10
1	16
2	55
3	11
4+	8

A new study investigates households on the west coast of the US, to see if the distribution there is similar to the entire country, or if it is unique. A random sample of 600 west coast households yielded the following data.

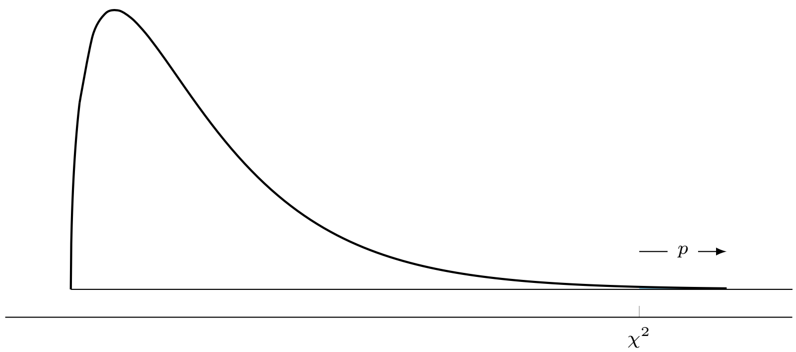
Number of TVs	Frequency
0	66
1	119
2	340
3	60
4+	15

Using a significance level of 0.01, does it seem that the distribution for the west coast is different from the country as a whole?

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Number of TVs	Observed Frequency	Expected Frequency	$O - E$
0	66		
1	119		
2	340		
3	60		
4+	15		



Step 3: Calculate the p value that corresponds to this area.

Step 4: Draw a conclusion.

EXAMPLE 3 CASINO GAME

A new casino game involves rolling 3 dice, and winnings are based on the total number of sixes rolled. A gambler played the game 100 times, with the following results.

Number of Sixes	Frequency
0	48
1	35
2	15
3	3

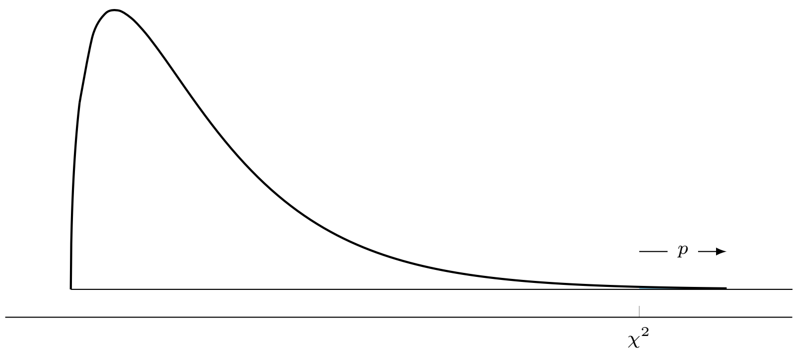
Should the casino ban this player for using rigged dice?

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Number of Sixes	Probability
0	
1	
2	
3	

Number of Sixes	Observed Frequency	Expected Frequency	$O - E$
0	48		
1	35		
2	15		
3	3		



Step 3: Calculate the p value that corresponds to this area.

Step 4: Draw a conclusion.

Homework 14 Name:

1. The table below lists the number of traffic accidents involving children within a mile of a school, listed by days of the week. Test the assumption that the frequencies are uniformly (equally) distributed, using a significance level of 0.05.

Weekday	Number of Accidents
Monday	23
Tuesday	18
Wednesday	17
Thursday	19
Friday	23

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

2. A buyer for a t-shirt shop wants to compare the proportion of shirts of each size that are sold to the proportion that were ordered. The table below shows the number that were ordered and the number that were sold for each size.

Size	Number Ordered	Number Sold
Small	23	25
Medium	45	41
Large	90	91
Extra Large	67	68

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

3. A baseball card company claims that 30% of their cards are rookies, 60% are veterans but not All-Stars, and 10% are veteran All-Stars. If a random sample of 100 cards has 50 rookies, 45 veterans, and 5 All-Stars, is this consistent with the company's claim?

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

4. A factory manager needs to understand how many products are defective based on how many are produced. The following table shows the expected numbers of defects, and the number that were actually found defective in a random sample.

Number Produced	Expected Defective	Observed Defective
0 – 100	5	5
101 – 200	6	7
201 – 300	7	8
301 – 400	8	9
401 – 500	10	11

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

5. Suppose you flip two coins 100 times. The results are 20 HH, 27 HT, 30 TH, and 23 TT. Are the coins fair?

Step 1: State the hypotheses.

Step 2: Calculate the test statistic.

Step 3: Calculate the p value.

Step 4: Draw a conclusion.

Chapter

12 Linear Regression and Correlation

Often, we want to determine whether there is a relationship between two variables, and if so, what the relationship is. For instance, when studying economics, we might study the connection between inflation and unemployment.

There are two ideas in this chapter:

- **Correlation:**

- **Regression:**

SECTION 12.1 Linear Equations

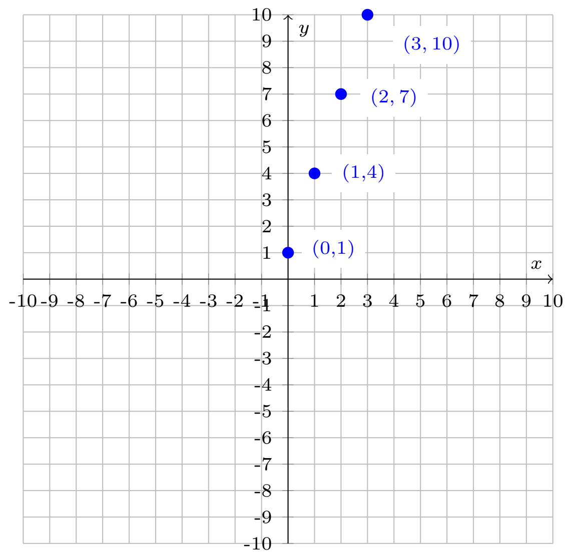
Take a look at an equation like

$$y = 3x + 1.$$

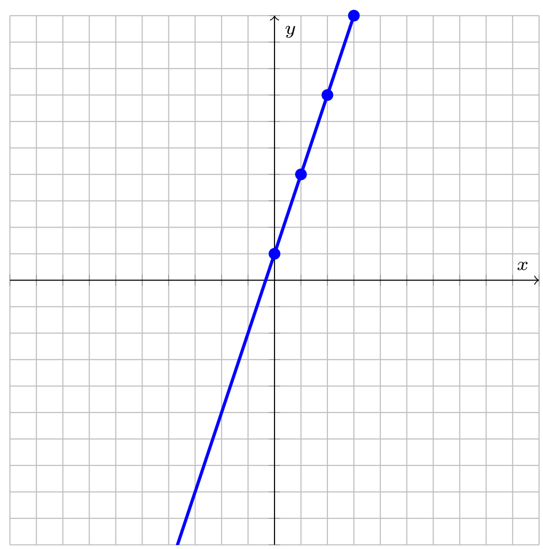
This equation gives a relationship between x and y; it simply says that whatever x is, there is a corresponding y that you get by multiplying x by 3 and adding 1. For instance,

$$\begin{aligned} &\text{if } x \text{ is } 0, \quad y \text{ is } 1\\ &\text{if } x \text{ is } 1, \quad y \text{ is } 4\\ &\text{if } x \text{ is } 2, \quad y \text{ is } 7\\ &\text{if } x \text{ is } 3, \quad y \text{ is } 10 \end{aligned}$$

We can write each of these as an ordered pair (x, y), and each of those corresponds to a point on the coordinate plane:



This is an example of a linear equation:



Slope and Intercept

Look back at that example.

x	y
0	1
1	4
2	7
3	10

-

-

Graphing with Slope and Intercept:

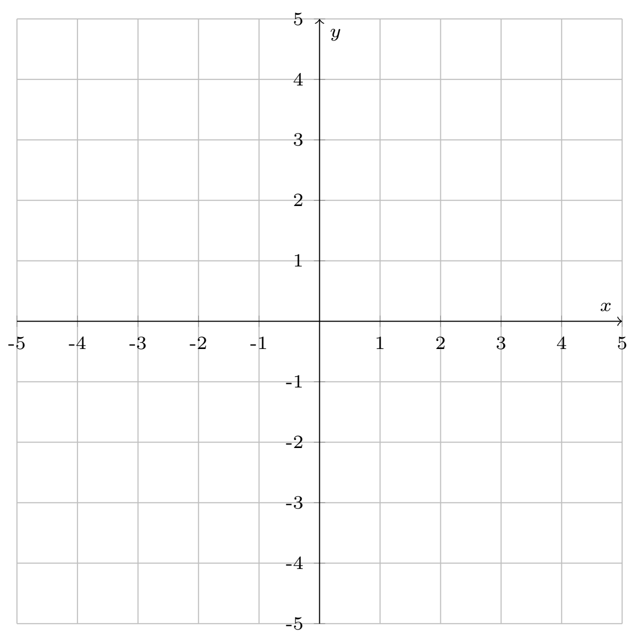
1.

2.

GRAPHING USING SLOPE AND INTERCEPT — EXAMPLE 1

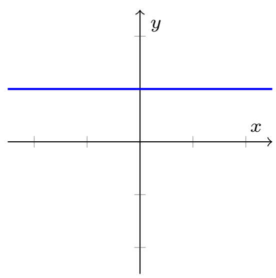
Graph the line $y = -2x + 4$.

Solution

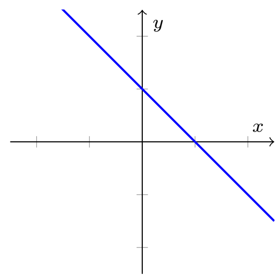


In general, we write

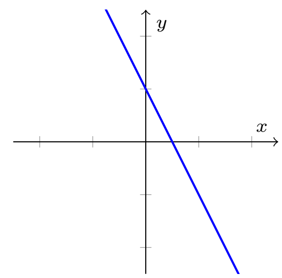
Examples of different slopes:



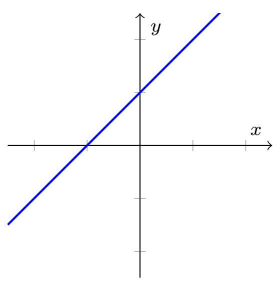
$m = 0$



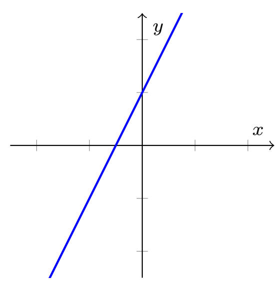
$m = -1$



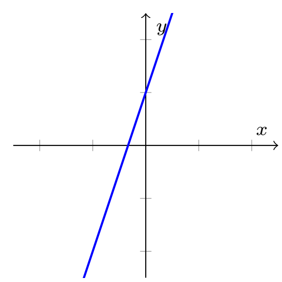
$m = -2$



$m = 1$



$m = 2$



$m = 3$

Note:

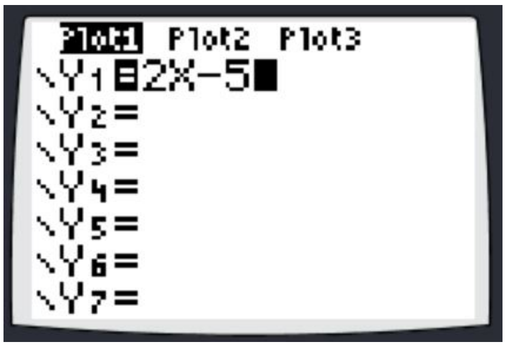
Using Your Calculator

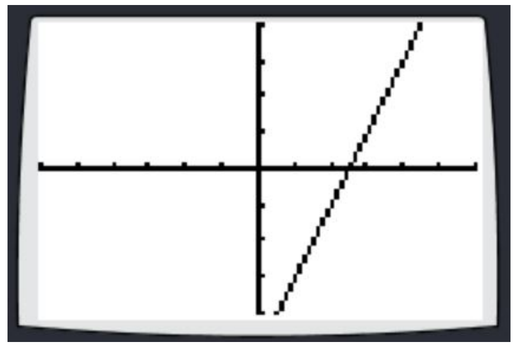
You can also use a graphing calculator to graph a linear equation for you if it is written in slope-intercept form. To do so, press the 

button in the upper lefthand corner and enter the equation, using the 

button to enter x. Then press 

to see the line.





Interpreting a Linear Equation

- x:

- y:

INTERPRETING A LINEAR EQUATION EXAMPLE 2

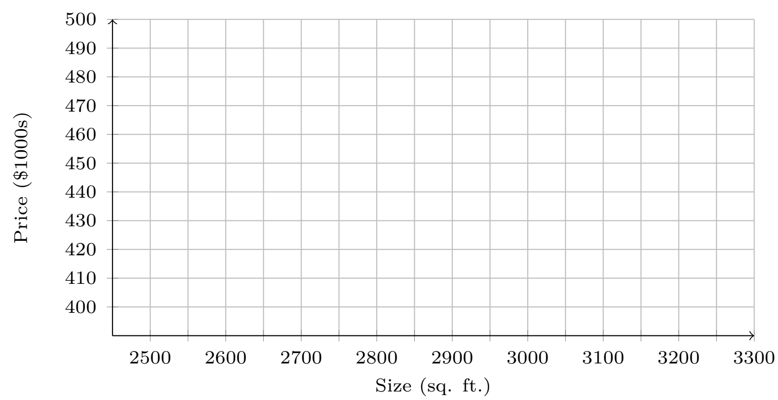
A landscaping service charges \$50 per visit, plus \$35 an hour. Write an equation that relates the cost y to the number of hours x.

SECTION 12.2 Scatter Plots and Correlation

Example: Home Prices

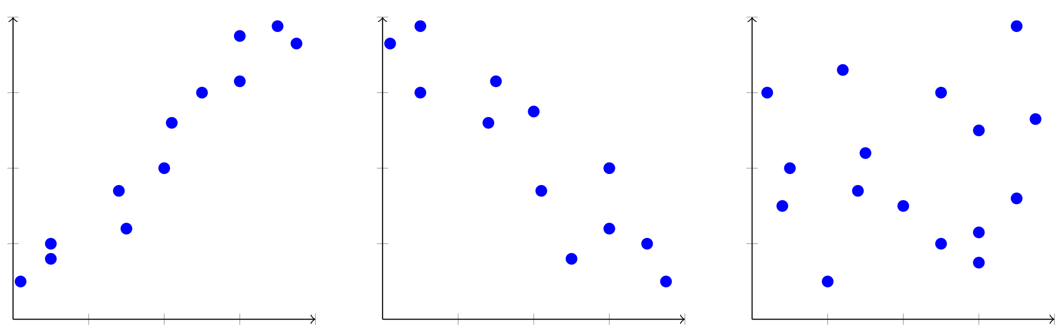
Size (sq. ft.)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

Notice:

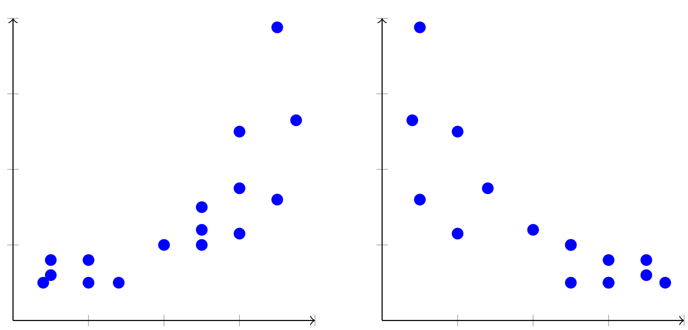


Association

-
-
-



Positive linear Negative linear No association

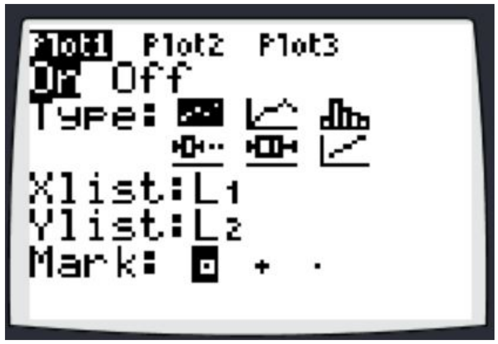


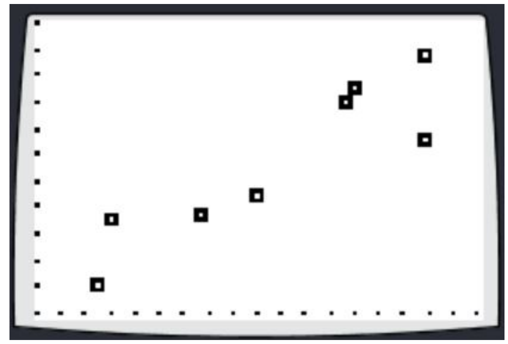
Positive nonlinear Negative nonlinear

Using Your Calculator

Make sure that there is no equation begin graphed, then enter the data. Press


to access the `STAT PLOT` menu. Select the scatter plot and select the lists where you entered the data.





Press the


button and adjust the window until the scatter plot is visible.

The Correlation Coefficient

The correlation coefficient, r, is a measure of how strong a linear relationship is between two variables.

Correlation Coefficient

FINDING R EXAMPLE 1

Find the correlation coefficient for the house price data.

Size (sq. ft.)	**Selling Price ($1000s)**
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

$$n = 8$$
$$\overline{x} = 2891.25 \qquad \overline{y} = 447$$
$$s_x = 269.49 \qquad s_y = 29.68$$

x	y	$\dfrac{x-\overline{x}}{s_x}$	$\dfrac{y-\overline{y}}{s_y}$	$\left(\dfrac{x-\overline{x}}{s_x}\right)\left(\dfrac{y-\overline{y}}{s_y}\right)$
2521	400	−1.37389	−1.58356	2.17564
2555	426	−1.24773	−0.70755	0.88283
2735	428	−0.57980	−0.64016	0.37116
2846	435	−0.16791	−0.40431	0.06789
3028	469	0.50744	0.74124	0.37613
3049	475	0.58536	0.94340	0.55223
3198	488	1.13826	1.38140	1.28783
3198	455	1.13826	0.26954	0.30681
				6.30414

$$r = \frac{6.30414}{7} = 0.90059$$

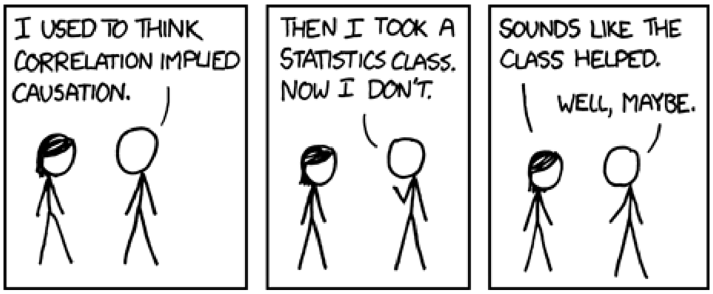
How do we interpret the correlation coefficient?

Properties of the Correlation Coefficient

-
-
-
-
-
-
-
-
-

Correlation Does Not Imply Causation

Just because two variables are highly correlated, that doesn't mean that one causes the other. In the example of the house sizes and their price, there IS a causal link, but you can't assume that in every case where there's a strong correlation.

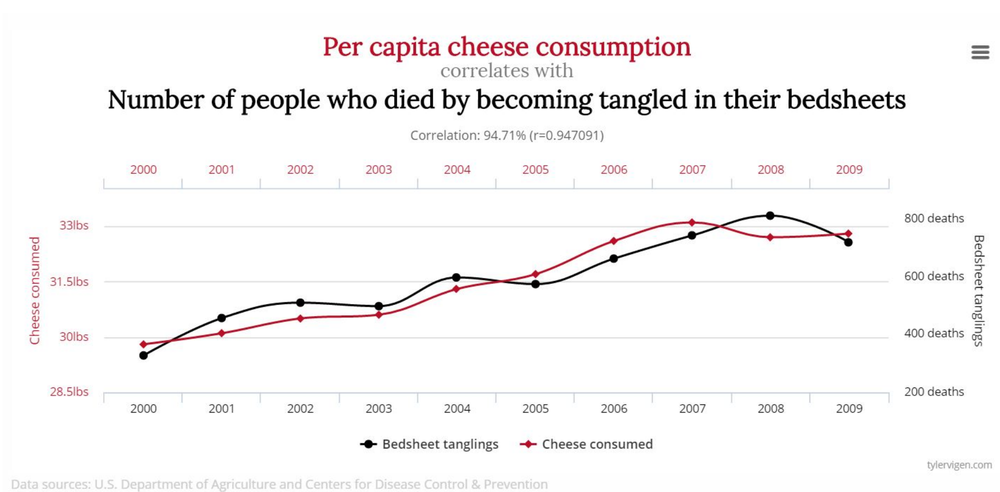


"Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there.' "
xkcd.com

For instance, the number of injuries sustained in a swimming pool is correlated with the sales of ice cream cones. Do ice cream cones cause injuries, or vice versa? Of course not; it's just that both of them are much more common in warmer weather.

In that case, we call the weather a **confounder**, a third variable that is related to the two we're interested in. If we don't consider this third variable, it can fool us into thinking that the other two cause each other.

Even if there isn't a confounder, sometimes two variables can be related by coincidence. There's a book and website by Tyler Vigen (tylervigen.com) devoted to showing such correlations. For example:



SECTION 12.3 The Regression Equation

Example: Home Prices

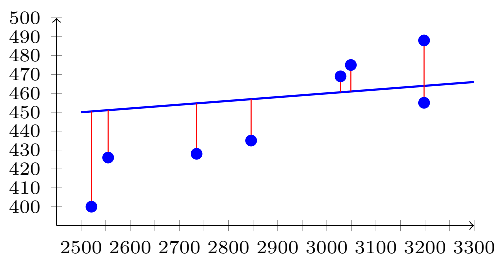
Size (sq. ft.)	Selling Price ($1000s)
2521	400
2555	426
2735	428
2846	435
3028	469
3049	475
3198	488
3198	455

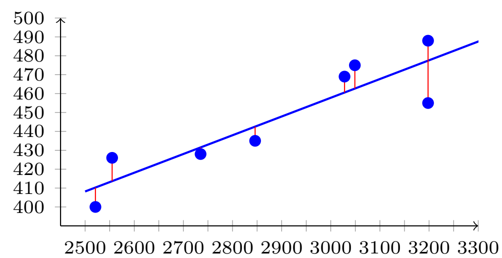
$$r = 0.9006$$

Ok, so there's a linear relationship between these variables, but what is it, actually? Finding the linear relationship

is the problem of

Note that $\hat{y}$ represents the predicted value for a given x value, and difference between the predicted value and actual value is called the





The goal of constructing the regression equation is to **minimize the squared residuals**. This line of best fit is called the **least-squares regression line**.

Equation of the Least-Squares Regression Line

Given ordered pairs (x,y) with sample means $\overline{x}$ and $\overline{y}$, sample standard deviations s_x and s_y, and correlation coefficient r, the equation of the least-squares regression line for predicting y from x is

where the slope and intercept are given by

- **Slope:**

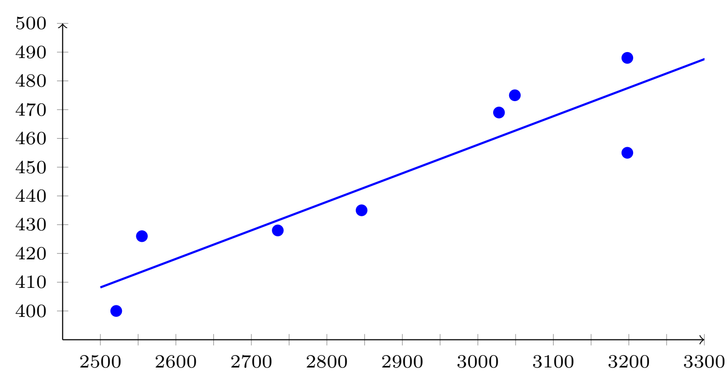
- **Intercept:**

- **Explanatory variable:**

- **Outcome or response variable:**

EXAMPLE 1 REGRESSION LINE

Compute the least-squares regression line for the house price data.



Using Your Calculator

Your calculator can also find the least-squares regression line. To do so, enter the data into `L1` and `L2`.

Then press the 

button and scroll over to the `CALC` menu. The fourth option is `4:LinReg(ax+b)`. If you select this option, you should see a menu like the following:

If you select `Calculate`, you'll see something like this:

Note: If you don't see the r value, press 

to access the catalog, then scroll down to `DiagnosticOn`. Press enter twice to turn it on, and then repeat the steps above to find the regression line.

EXAMPLE 2 REGRESSION LINE

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200.

x	y
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

Find the equation of the least-squares regression line.

SECTION 12.5 Prediction

Now that we can build the regression line, we want to know what we can do with it, and how we should interpret it.

Making Predictions

The regression line gives a predicted y value, $\hat{y}$, for each given x value (within a reasonable range). Of course, this is only a prediction, and we expect the actual value to differ slightly from the prediction.

PREDICTING **EXAMPLE 1**

The house price data led to the following regression line:

$$\hat{y} = 0.0992x + 160.194.$$

(a) Predict the price of a home with 2700 square feet.

(b) Predict the price of a home with 4500 square feet.

Interpreting the Predicted Value $\hat{y}$

Which of those answers do you think will be a better prediction?

•

•

•

More precisely: $\hat{y}$ is what we expect the *average* y value to be for all the data points with a particular x value. In the example above, we expect that the average price for homes with 2700 square feet will be \$428,034.

EXAMPLE 2 PREDICTING

The data on students' third test and final exam led to the following equation for the least-squares regression line:

$$\hat{y} = 4.827x - 175.513.$$

What final exam score would you predict for a student who scored 60 on the third test?

Note: Not every x value that you can plug into the regression equation is a meaningful one. For instance, you could try predicting the final exam score of a student who got a 90 on the third test (even though the third test scores can only go up to 80), and the equation will dutifully give you a value. Just note that that value is meaningless; you need to use common sense when making predictions.

Interpreting the Slope

-
-
-

USING SLOPE — EXAMPLE 3

Two houses differ in size by 300 square feet. How much would you expect their prices to differ?

Note: The slope doesn't mean that if x *changes* by 1, we expect y to *change* by the amount of the slope; it means that if we look at two different data points, then we can predict the difference in their y values based on the difference in their x values.

For example, if we developed a regression model to predict a person's height based on their weight, we couldn't say that if they lost weight, they'd suddenly shrink.

MAKING PREDICTIONS — EXAMPLE 4

At the final exam in a statistics class, the professor asks each student to indicate how many hours he or she studied for the exam. After grading the exam, the professor computes the least-squares regression line for predicting the final exam score from the number of hours studied. The equation of the line is $\hat{y} = 50 + 5x$.

(a) Antoine studied for 6 hours. What do you predict his exam score to be?

(b) Emma studied for 3 hours longer than Jeremy did. How much higher do you predict Emma's score to be?

Interpreting the Intercept

The slope, mathematically, is the y value of a data point whose x value is 0.

-
-

EXAMPLE 5 INTERPRETING THE INTERCEPT

For each of the following scenarios, decide whether or not the y intercept is meaningful in context.

(a) The house price example.

(b) The test score example.

(c) The least-squares regression line is $\hat{y} = 1.98 + 0.039x$, where x is the temperature in a freezer in degrees Fahrenheit, and y is the time it takes to freeze a certain amount of water into ice.

(d) The least-squares regression line is $\hat{y} = -13.586 + 4.340x$, where x represents the age of an elementary school student and y represents the score on a standardized test.

LINEAR REGRESSION EXAMPLE 6

The following table lists the heights (in inches) and weights (in pounds) of 14 NFL quarterbacks in the 2009 season.

Name	Height	Weight
Peyton Manning	77	230
Tom Brady	76	225
Ben Roethlisberger	77	241
Drew Brees	72	209
Eli Manning	76	225
Carson Palmer	77	235
Phillip Rivers	77	228
Kurt Warner	74	214
Donovan McNabb	74	240
Jay Cutler	75	233
Tony Romo	74	225
Matt Ryan	76	220
Brett Favre	74	222
Kyle Orton	76	225

(a) Compute the regression line for predicting weight from height.

(b) Calculate r, the correlation coefficient.

(c) Do you think this linear regression model is going to be an accurate one?

(d) Is it possible to interpret the y-intercept?

(e) If two quarterbacks differ in height by two inches, by how much would you expect their weight to differ?

(f) Predict the weight of a quarterback who is 74.5 inches tall.

(g) Does Tom Brady weigh more or less than the weight predicted by the regression line, based on his height?

EXAMPLE 7 LINEAR REGRESSION

A blood pressure measurement consists of two numbers: the systolic pressure, which is the maximum pressure taken when the heart is contracting, and the diastolic pressure, which is the minimum pressure taken at the beginning of the heartbeat. Blood pressures were measured (in millimeters of mercury, mmHg) for a sample of 16 adults.

Systolic	134	115	113	123	119	118	130	116
Diastolic	87	83	77	77	69	88	76	70
Systolic	133	112	107	110	108	105	157	154
Diastolic	91	75	71	74	69	66	103	94

(a) Calculate r, the correlation coefficient.

(b) Do you think there is a strong linear association?

(c) Compute the regression line for predicting the diastolic pressure from the systolic pressure.

(d) Is it possible to interpret the y-intercept?

(e) If the systolic pressures of two patients differ by 10 mmHg, by how much would you predict their diastolic pressures will differ?

(f) Predict the diastolic pressure for a patient whose systolic pressure is 125 mmHg.

Homework 15 Name:

1. Consider the following data set.

x	9	5	7	13	−8	−2	6	−10
y	3	3	31	36	0	3	−2	−14

1. Calculate r, the correlation coefficient.

   ____

2. Compute the least-squares regression line for this data set.

   ____

2. Compute the least-squares regression line based on the following statistics.

$$\overline{x} = 5,\ s_x = 2,\ \overline{y} = 1350,\ s_y = 100,\ r = 0.70$$

____

3. The data set below shows the GMAT scores for five MBA students and the students' grade point averages (GPA) upon graduation.

GMAT	660	580	480	710	600
GPA	3.7	3.0	3.2	4.0	3.5

(a) Calculate r, the correlation coefficient between these two variables.

____

(b) Compute the regression line for predicting GPA from GMAT score.

____

(c) If two students' GMAT scores differ by 50 points, by how much would you expect their GPA to differ?

____

(d) Predict the GPA of a student who gets a score of 500 on the GMAT.

____

(e) Does the student with a GMAT score of 580 fall above or below the regression line?

____

4. The data set below shows the mileage and selling prices of eight used cars of the same model.

Mileage	21,000	34,000	41,000	43,000	65,000	72,000	76,000	84,000
Price	\$16,000	\$11,000	\$13,000	\$14,000	\$10,000	\$12,000	\$7,000	\$7,000

(a) Calculate r, the correlation coefficient between these two variables.

(b) Compute the regression line for predicting price from mileage.

(c) Interpret the y-intercept or state that it cannot be interpreted.

(d) If the mileages of two of these cars differ by 5,000, by how much would you expect their prices to differ?

(e) Predict the price of a car with 30,000 miles.

(f) Does the car with 43,000 miles on it fall above or below the regression line?

SECTION 12.4 Inferences with Regression

We can use confidence intervals and hypothesis tests to determine whether a linear model is a good fit. More specifically, we can tell which of many factors are good predictors.

We'll do three things in this section:

1.

2.

3.

We'll find out that the last two of these three are really the same process, so there are really just two distinct things that we'll do.

When the points that we plot in a scatter plot form a sample from a larger population, we assume that the regression line we draw is an estimate of the regression line for the population. We want to use the sample regression line to draw inferences about the population regression line.

Confidence Intervals for Slope

The population regression line is

The sample slope a is a point estimate for the population slope α.

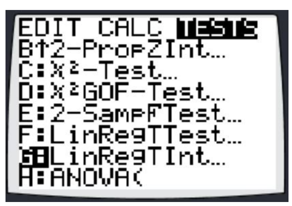
Remember, a confidence interval looks like

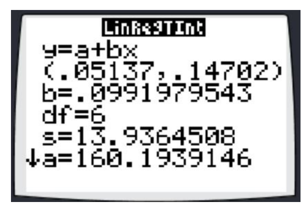
and the margin of error is a z or t value multiplied by a standard error.

Confidence Interval for Slope

Using Your Calculator Press 

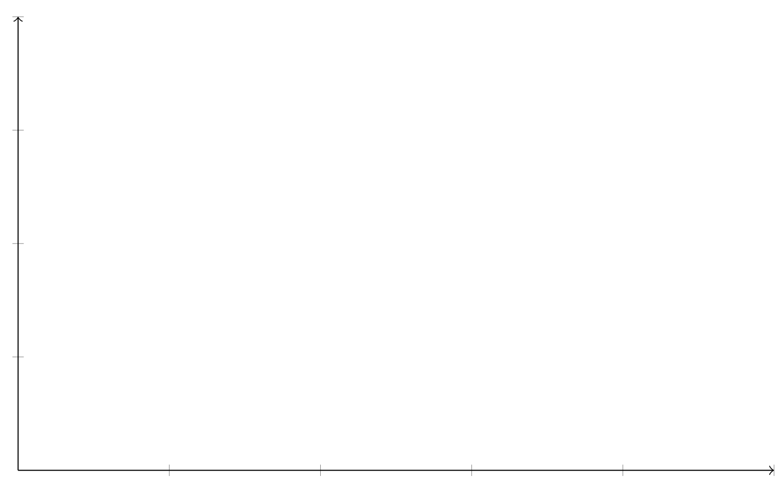
and scroll over to the `TESTS` menu. Almost at the bottom of the menu is `G:LinRegTInt`





Interpreting the Interval What does this confidence interval tell us?

-



-
-
-

Hypothesis Testing with Slope

The most common test is the same as the confidence interval: testing whether or not the population slope α is 0.

Hypothesis Testing with Slope

Step 1: State the Hypotheses

H_0	H_1
$\alpha = 0$	$\alpha \neq 0$
$\alpha \geq 0$	$\alpha < 0$
$\alpha \leq 0$	$\alpha > 0$

Step 2: Compute the Test Statistic

$$t_{\alpha/2} = \frac{a}{s_a} \qquad df = n - 2$$

Step 3: Calculate p, using the appropriate tail(s) of Student's t distribution.

Step 4: Draw a conclusion.

Using Your Calculator Use `LinRegTTest`

EXAMPLE 1 HYPOTHESIS TESTING WITH SLOPE

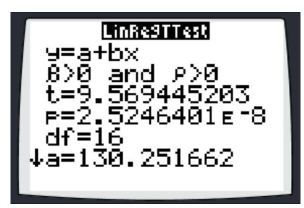
The following table displays the number of grams of fat per 100 grams of product and number of calories for a sample of 18 fast food products.

Product	Fat (x)	Calories (y)
Burger King Chicken Tenders	16.67	289
Burger King Croissan'wich	25.45	376
Domino's Cheese Pizza, Thin Crust	16.82	315
Kentucky Fried Chicken Xtra Crispy	16.55	268
Kentucky Fried Chicken Original Recipe	12.03	221
Little Caesar's Cheese Pizza Thin Crust	16.99	309
McDonald's Big and Tasty	13.68	226
McDonald's Biscuit	16.01	344
McDonald's Chocolate Triple Shake	4.51	163
McDonald's Deluxe Cinnamon Roll	16.24	367
McDonald's Hot Caramel Sundae	4.89	188
Papa John's Pepperoni Pizza Original Crust	11.86	275
Popeye's Biscuit	24.53	408
Popeye's Fried Chicken	35.39	460
Taco Bell Burrito Supreme with Beef	8.05	189
Taco Bell Nachos	22.17	366
Wendy's Chicken Nuggets	23.17	334
Wendy's Classic Double	14.20	241

Test the claim that the slope is greater than 0.

Hypotheses:

$$H_0 : \alpha \leq 0$$
$$H_1 : \alpha > 0$$



Since p is small, we reject the null hypothesis and conclude that the slope is significantly larger than 0.

The point: finding which factors are important in explaining some outcome.

Hypothesis Testing with the Correlation Coefficient

This one's a freebie:

-

-

-

-

SECTION 12.6 Outliers

Rule of Thumb for Outliers

If a point is more than two standard deviations away from the regression line, it is considered an outlier.
The standard deviation used is the standard deviation of the residuals $y - \hat{y}$.

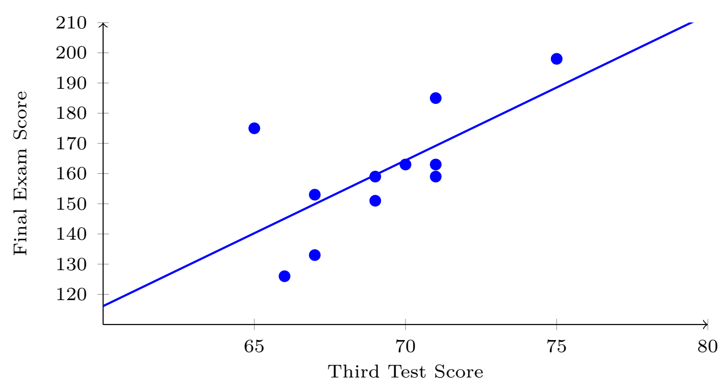
Finding s The standard deviation of the residuals is listed in the results of the `LinRegTTest`.

EXAMPLE 1 REGRESSION LINE

A random sample of 11 statistics students produced the following data, where x is the third exam score out of 80, and y is the final exam score out of 200.

x	y
65	175
67	133
71	185
71	163
66	126
75	198
67	153
70	163
71	159
69	151
69	159

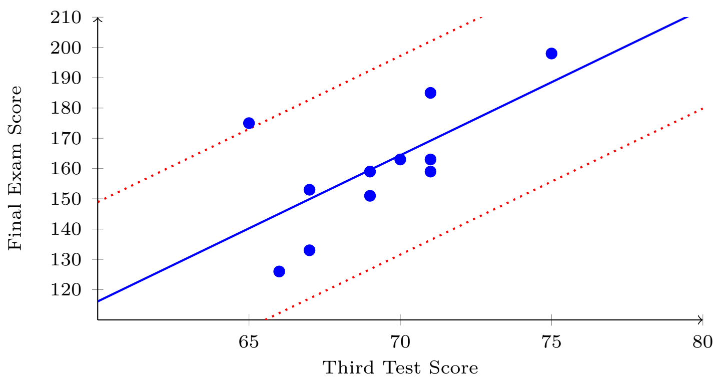
Below is the scatter plot with the regression line included.



Find any outliers.

$$s = 16.41$$

Draw bounds two standard deviations above and below the regression line.



The one outlier is the point

$$(65, 175).$$

Chapter

13

Appendix: Tables

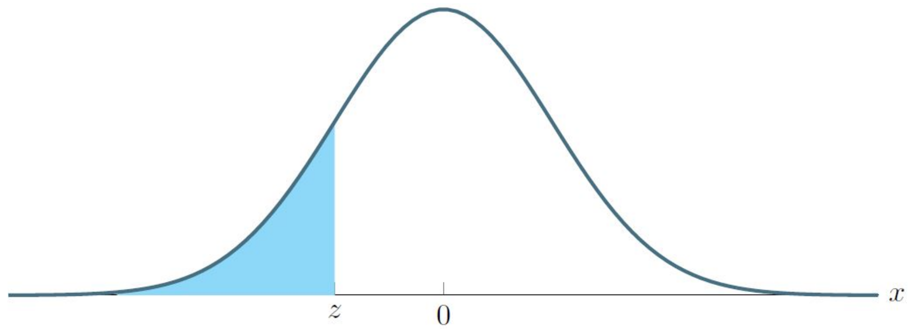
1. Binomial Probabilities
2. Cumulative Normal Distribution
3. Student's t Distribution

Table 1: Binomial Probabilities

n	k	$\binom{n}{k}$	0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.35	0.40	0.45	0.49	0.50
2	0	1	0.9801	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4444	0.4225	0.3600	0.3025	0.2601	0.2500
	1	2	0.0198	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4444	0.4550	0.4800	0.4950	0.4998	0.5000
	2	1	0.0001	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1111	0.1225	0.1600	0.2025	0.2401	0.2500
3	0	1	0.9703	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2963	0.2746	0.2160	0.1664	0.1327	0.1250
	1	3	0.0294	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4444	0.4436	0.4320	0.4084	0.3823	0.3750
	2	3	0.0003	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2222	0.2389	0.2880	0.3341	0.3674	0.3750
	3	1		0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0370	0.0429	0.0640	0.0911	0.1176	0.1250
4	0	1	0.9606	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1975	0.1785	0.1296	0.0915	0.0677	0.0625
	1	4	0.0388	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3951	0.3845	0.3456	0.2995	0.2600	0.2500
	2	6	0.0006	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.2963	0.3105	0.3456	0.3675	0.3747	0.3750
	3	4		0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.0988	0.1115	0.1536	0.2005	0.2400	0.2500
	4	1			0.0001	0.0005	0.0016	0.0039	0.0081	0.0123	0.0150	0.0256	0.0410	0.0576	0.0625
5	0	1	0.9510	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1317	0.1160	0.0778	0.0503	0.0345	0.0313
	1	5	0.0480	0.2036	0.3281	0.3915	0.4096	0.3955	0.3601	0.3292	0.3124	0.2592	0.2059	0.1657	0.1563
	2	10	0.0010	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3292	0.3364	0.3456	0.3369	0.3185	0.3125
	3	10		0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1646	0.1811	0.2304	0.2757	0.3060	0.3125
	4	5			0.0005	0.0022	0.0064	0.0146	0.0284	0.0412	0.0488	0.0768	0.1128	0.1470	0.1563
	5	1				0.0001	0.0003	0.0010	0.0024	0.0041	0.0053	0.0102	0.0185	0.0282	0.0313
6	0	1	0.9415	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0878	0.0754	0.0467	0.0277	0.0176	0.0156
	1	6	0.0571	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2634	0.2437	0.1866	0.1359	0.1014	0.0938
	2	15	0.0014	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3292	0.3280	0.3110	0.2780	0.2436	0.2344
	3	20		0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2195	0.2355	0.2765	0.3032	0.3121	0.3125
	4	15		0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0823	0.0951	0.1382	0.1861	0.2249	0.2344
	5	6			0.0001	0.0004	0.0015	0.0044	0.0102	0.0165	0.0205	0.0369	0.0609	0.0864	0.0938
	6	1					0.0001	0.0002	0.0007	0.0014	0.0018	0.0041	0.0083	0.0138	0.0156
7	0	1	0.9321	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0585	0.0490	0.0280	0.0152	0.0090	0.0078
	1	7	0.0659	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.2049	0.1848	0.1306	0.0872	0.0604	0.0547
	2	21	0.0020	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.3073	0.2985	0.2613	0.2140	0.1740	0.1641
	3	35		0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2561	0.2679	0.2903	0.2918	0.2786	0.2734
	4	35		0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1280	0.1442	0.1935	0.2388	0.2676	0.2734
	5	21			0.0002	0.0012	0.0043	0.0115	0.0250	0.0384	0.0466	0.0774	0.1172	0.1543	0.1641
	6	7				0.0001	0.0004	0.0013	0.0036	0.0064	0.0084	0.0172	0.0320	0.0494	0.0547
	7	1						0.0001	0.0002	0.0005	0.0006	0.0016	0.0037	0.0068	0.0078
8	0	1	0.9227	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0390	0.0319	0.0168	0.0084	0.0046	0.0039
	1	8	0.0746	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1561	0.1373	0.0896	0.0548	0.0352	0.0313
	2	28	0.0026	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2731	0.2587	0.2090	0.1569	0.1183	0.1094
	3	56	0.0001	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2731	0.2786	0.2787	0.2568	0.2273	0.2188
	4	70		0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1707	0.1875	0.2322	0.2627	0.2730	0.2734
	5	56			0.0004	0.0026	0.0092	0.0231	0.0467	0.0683	0.0808	0.1239	0.1719	0.2098	0.2188
	6	28				0.0002	0.0011	0.0038	0.0100	0.0171	0.0217	0.0413	0.0703	0.1008	0.1094
	7	8					0.0001	0.0004	0.0012	0.0024	0.0033	0.0079	0.0164	0.0277	0.0313
	8	1							0.0001	0.0002	0.0002	0.0007	0.0017	0.0033	0.0039
9	0	1	0.9135	0.6302	0.3874	0.2316	0.1342	0.0751	0.0404	0.0260	0.0207	0.0101	0.0046	0.0023	0.0020
	1	9	0.0830	0.2985	0.3874	0.3679	0.3020	0.2253	0.1556	0.1171	0.1004	0.0605	0.0339	0.0202	0.0176
	2	36	0.0034	0.0629	0.1722	0.2597	0.3020	0.3003	0.2668	0.2341	0.2162	0.1612	0.1110	0.0776	0.0703
	3	84	0.0001	0.0077	0.0446	0.1069	0.1762	0.2336	0.2668	0.2731	0.2716	0.2508	0.2119	0.1739	0.1641
	4	126		0.0006	0.0074	0.0283	0.0661	0.1168	0.1715	0.2048	0.2194	0.2508	0.2600	0.2506	0.2461
	5	126			0.0008	0.0050	0.0165	0.0389	0.0735	0.1024	0.1181	0.1672	0.2128	0.2408	0.2461
	6	84			0.0001	0.0006	0.0028	0.0087	0.0210	0.0341	0.0424	0.0743	0.1160	0.1542	0.1641
	7	36				0.0000	0.0003	0.0012	0.0039	0.0073	0.0098	0.0212	0.0407	0.0635	0.0703
	8	9						0.0001	0.0004	0.0009	0.0013	0.0035	0.0083	0.0153	0.0176
	9	1								0.0001	0.0001	0.0003	0.0008	0.0016	0.0020
10	0	1	0.9044	0.5987	0.3487	0.1969	0.1074	0.0563	0.0282	0.0173	0.0135	0.0060	0.0025	0.0012	0.0010
	1	10	0.0914	0.3151	0.3874	0.3474	0.2684	0.1877	0.1211	0.0867	0.0725	0.0403	0.0207	0.0114	0.0098
	2	45	0.0042	0.0746	0.1937	0.2759	0.3020	0.2816	0.2335	0.1951	0.1757	0.1209	0.0763	0.0494	0.0439
	3	120	0.0001	0.0105	0.0574	0.1298	0.2013	0.2503	0.2668	0.2601	0.2522	0.2150	0.1665	0.1267	0.1172
	4	210		0.0010	0.0112	0.0401	0.0881	0.1460	0.2001	0.2276	0.2377	0.2508	0.2384	0.2130	0.2051
	5	252		0.0001	0.0015	0.0085	0.0264	0.0584	0.1029	0.1366	0.1536	0.2007	0.2340	0.2456	0.2461
	6	210			0.0001	0.0012	0.0055	0.0162	0.0368	0.0569	0.0689	0.1115	0.1596	0.1966	0.2051
	7	120				0.0001	0.0008	0.0031	0.0090	0.0163	0.0212	0.0425	0.0746	0.1080	0.1172
	8	45					0.0001	0.0004	0.0014	0.0030	0.0043	0.0106	0.0229	0.0389	0.0439
	9	10							0.0001	0.0003	0.0005	0.0016	0.0042	0.0083	0.0098
	10	1										0.0001	0.0003	0.0008	0.0010
n	k	$\binom{n}{k}$	0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.35	0.40	0.45	0.49	0.50

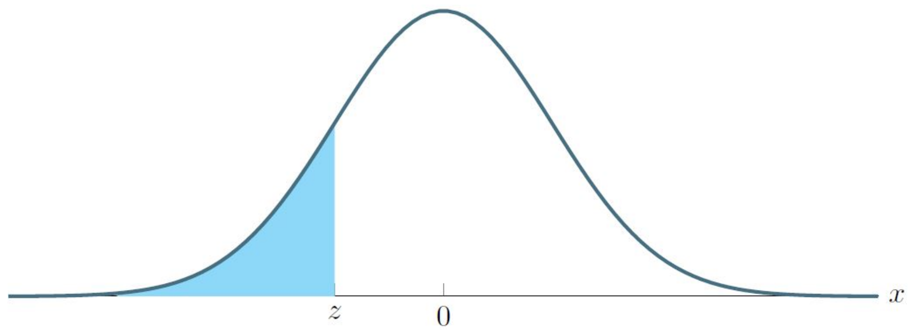
n	k	$\binom{n}{k}$	0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.35	0.40	0.45	0.49	0.50
11	0	1	0.8953	0.5688	0.3138	0.1673	0.0859	0.0422	0.0198	0.0116	0.0088	0.0036	0.0014	0.0006	0.0005
	1	11	0.0995	0.3293	0.3835	0.3248	0.2362	0.1549	0.0932	0.0636	0.0518	0.0266	0.0125	0.0064	0.0054
	2	55	0.0050	0.0867	0.2131	0.2866	0.2953	0.2581	0.1998	0.1590	0.1395	0.0887	0.0513	0.0308	0.0269
	3	165	0.0002	0.0137	0.0710	0.1517	0.2215	0.2581	0.2568	0.2385	0.2254	0.1774	0.1259	0.0888	0.0806
	4	330		0.0014	0.0158	0.0536	0.1107	0.1721	0.2201	0.2385	0.2428	0.2365	0.2060	0.1707	0.1611
	5	462		0.0001	0.0025	0.0132	0.0388	0.0803	0.1321	0.1669	0.1830	0.2207	0.2360	0.2296	0.2256
	6	462			0.0003	0.0023	0.0097	0.0268	0.0566	0.0835	0.0985	0.1471	0.1931	0.2206	0.2256
	7	330				0.0003	0.0017	0.0064	0.0173	0.0298	0.0379	0.0701	0.1128	0.1514	0.1611
	8	165					0.0002	0.0011	0.0037	0.0075	0.0102	0.0234	0.0462	0.0727	0.0806
	9	55						0.0001	0.0005	0.0012	0.0018	0.0052	0.0126	0.0233	0.0269
	10	11							0.0000	0.0001	0.0002	0.0007	0.0021	0.0045	0.0054
	11	1										0.0000	0.0002	0.0004	0.0005
12	0	1	0.8864	0.5404	0.2824	0.1422	0.0687	0.0317	0.0138	0.0077	0.0057	0.0022	0.0008	0.0003	0.0002
	1	12	0.1074	0.3413	0.3766	0.3012	0.2062	0.1267	0.0712	0.0462	0.0368	0.0174	0.0075	0.0036	0.0029
	2	66	0.0060	0.0988	0.2301	0.2924	0.2835	0.2323	0.1678	0.1272	0.1088	0.0639	0.0339	0.0189	0.0161
	3	220	0.0002	0.0173	0.0852	0.1720	0.2362	0.2581	0.2397	0.2120	0.1954	0.1419	0.0923	0.0604	0.0537
	4	495		0.0021	0.0213	0.0683	0.1329	0.1936	0.2311	0.2385	0.2367	0.2128	0.1700	0.1306	0.1208
	5	792		0.0002	0.0038	0.0193	0.0532	0.1032	0.1585	0.1908	0.2039	0.2270	0.2225	0.2008	0.1934
	6	924			0.0005	0.0040	0.0155	0.0401	0.0792	0.1113	0.1281	0.1766	0.2124	0.2250	0.2256
	7	792			0.0000	0.0006	0.0033	0.0115	0.0291	0.0477	0.0591	0.1009	0.1489	0.1853	0.1934
	8	495				0.0001	0.0005	0.0024	0.0078	0.0149	0.0199	0.0420	0.0762	0.1113	0.1208
	9	220					0.0001	0.0004	0.0015	0.0033	0.0048	0.0125	0.0277	0.0475	0.0537
	10	66							0.0002	0.0005	0.0008	0.0025	0.0068	0.0137	0.0161
	11	12								0.0000	0.0001	0.0003	0.0010	0.0024	0.0029
	12	1											0.0001	0.0002	0.0002
13	0	1	0.8775	0.5133	0.2542	0.1209	0.0550	0.0238	0.0097	0.0051	0.0037	0.0013	0.0004	0.0002	0.0001
	1	13	0.1152	0.3512	0.3672	0.2774	0.1787	0.1029	0.0540	0.0334	0.0259	0.0113	0.0045	0.0020	0.0016
	2	78	0.0070	0.1109	0.2448	0.2937	0.2680	0.2059	0.1388	0.1002	0.0836	0.0453	0.0220	0.0114	0.0095
	3	286	0.0003	0.0214	0.0997	0.1900	0.2457	0.2517	0.2181	0.1837	0.1651	0.1107	0.0660	0.0401	0.0349
	4	715		0.0028	0.0277	0.0838	0.1535	0.2097	0.2337	0.2296	0.2222	0.1845	0.1350	0.0962	0.0873
	5	1287		0.0003	0.0055	0.0266	0.0691	0.1258	0.1803	0.2067	0.2154	0.2214	0.1989	0.1664	0.1571
	6	1716			0.0008	0.0063	0.0230	0.0559	0.1030	0.1378	0.1546	0.1968	0.2169	0.2131	0.2095
	7	1716			0.0001	0.0011	0.0058	0.0186	0.0442	0.0689	0.0833	0.1312	0.1775	0.2048	0.2095
	8	1287				0.0001	0.0011	0.0047	0.0142	0.0258	0.0336	0.0656	0.1089	0.1476	0.1571
	9	715					0.0001	0.0009	0.0034	0.0072	0.0101	0.0243	0.0495	0.0788	0.0873
	10	286						0.0001	0.0006	0.0014	0.0022	0.0065	0.0162	0.0303	0.0349
	11	78							0.0001	0.0002	0.0003	0.0012	0.0036	0.0079	0.0095
	12	13										0.0001	0.0005	0.0013	0.0016
	13	1												0.0001	0.0001
14	0	1	0.8687	0.4877	0.2288	0.1028	0.0440	0.0178	0.0068	0.0034	0.0024	0.0008	0.0002	0.0001	0.0001
	1	14	0.1229	0.3593	0.3559	0.2539	0.1539	0.0832	0.0407	0.0240	0.0181	0.0073	0.0027	0.0011	0.0009
	2	91	0.0081	0.1229	0.2570	0.2912	0.2501	0.1802	0.1134	0.0779	0.0634	0.0317	0.0141	0.0068	0.0056
	3	364	0.0003	0.0259	0.1142	0.2056	0.2501	0.2402	0.1943	0.1559	0.1366	0.0845	0.0462	0.0260	0.0222
	4	1001		0.0037	0.0349	0.0998	0.1720	0.2202	0.2290	0.2143	0.2022	0.1549	0.1040	0.0687	0.0611
	5	2002		0.0004	0.0078	0.0352	0.0860	0.1468	0.1963	0.2143	0.2178	0.2066	0.1701	0.1320	0.1222
	6	3003			0.0013	0.0093	0.0322	0.0734	0.1262	0.1607	0.1759	0.2066	0.2088	0.1902	0.1833
	7	3431			0.0002	0.0019	0.0092	0.0280	0.0618	0.0918	0.1082	0.1574	0.1952	0.2088	0.2094
	8	3003				0.0003	0.0020	0.0082	0.0232	0.0402	0.0510	0.0918	0.1398	0.1756	0.1833
	9	2002					0.0003	0.0018	0.0066	0.0134	0.0183	0.0408	0.0762	0.1125	0.1222
	10	1001					0.0000	0.0003	0.0014	0.0033	0.0049	0.0136	0.0312	0.0540	0.0611
	11	364							0.0002	0.0006	0.0010	0.0033	0.0093	0.0189	0.0222
	12	91								0.0001	0.0001	0.0005	0.0019	0.0045	0.0056
	13	14										0.0001	0.0002	0.0007	0.0009
	14	1												0.0000	0.0001
15	0	1	0.8601	0.4633	0.2059	0.0874	0.0352	0.0134	0.0047	0.0023	0.0016	0.0005	0.0001	0.0000	
	1	15	0.1303	0.3658	0.3432	0.2312	0.1319	0.0668	0.0305	0.0171	0.0126	0.0047	0.0016	0.0006	0.0005
	2	105	0.0092	0.1348	0.2669	0.2856	0.2309	0.1559	0.0916	0.0599	0.0476	0.0219	0.0090	0.0040	0.0032
	3	455	0.0004	0.0307	0.1285	0.2184	0.2501	0.2252	0.1700	0.1299	0.1110	0.0634	0.0318	0.0166	0.0139
	4	1365		0.0049	0.0428	0.1156	0.1876	0.2252	0.2186	0.1948	0.1792	0.1268	0.0780	0.0478	0.0417
	5	3003		0.0006	0.0105	0.0449	0.1032	0.1651	0.2061	0.2143	0.2123	0.1859	0.1404	0.1010	0.0916
	6	5005		0.0000	0.0019	0.0132	0.0430	0.0917	0.1472	0.1786	0.1906	0.2066	0.1914	0.1617	0.1527
	7	6435			0.0003	0.0030	0.0138	0.0393	0.0811	0.1148	0.1319	0.1771	0.2013	0.1997	0.1964
	8	6435				0.0005	0.0035	0.0131	0.0348	0.0574	0.0710	0.1181	0.1647	0.1919	0.1964
	9	5005				0.0001	0.0007	0.0034	0.0116	0.0223	0.0298	0.0612	0.1048	0.1434	0.1527
	10	3003					0.0001	0.0007	0.0030	0.0067	0.0096	0.0245	0.0515	0.0827	0.0916
	11	1365						0.0001	0.0006	0.0015	0.0024	0.0074	0.0191	0.0361	0.0417
	12	455							0.0001	0.0003	0.0004	0.0016	0.0052	0.0116	0.0139
	13	105									0.0001	0.0003	0.0010	0.0026	0.0032
	14	15											0.0001	0.0004	0.0005
	15	1													
n	k	$\binom{n}{k}$	0.01	0.05	0.10	0.15	0.20	0.25	0.30	1/3	0.35	0.40	0.45	0.49	0.50

Table 2: Cumulative Normal Distribution



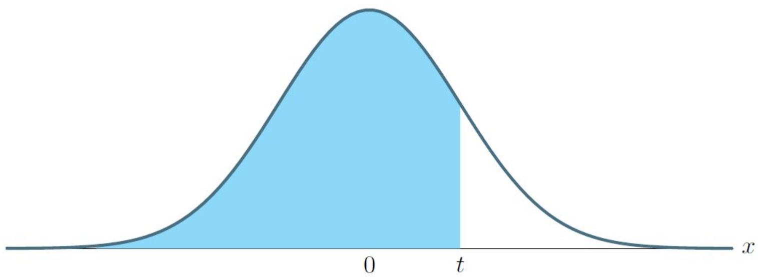
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.8	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.7	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.6	0.0002	0.0002	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
-3.5	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0620	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table 2: Cumulative Normal Distribution, continued



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

Table 3: Student's t Distribution



$df = n - 1$	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090