Data Displays

Statistics is the science of data. *Data* ("data" is a plural; "datum" is the singular) is a collection of related observations.

Types of observations
Numerical variables (also referred to as *quantitative variables*) are observations that are expressed as a number. Examples of numerical variables are the daily high temperature in Seattle, the number of calories in each meal at a restaurant, or the populations of different cities.

Categorical variables (also referred to as *qualitative variables*) are observations that place things within a set of known categories. Examples of categorical variables are the colors of cars on the highway or people's favorite desserts.
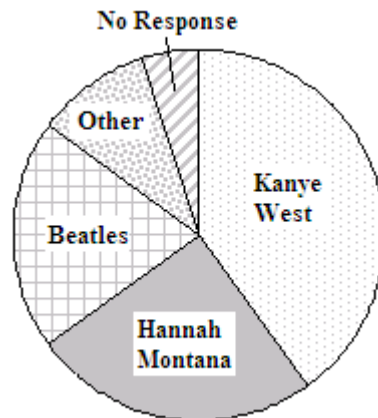
Data displays
Creating a data display is often the most effective way to discover trends or communicate conclusions about the observations. There are many types of data displays at many levels of complexity. In this section we will focus on pie charts and histograms, two common and important types of data displays.

Pie charts are often encountered in everyday life. In a pie chart, the size of each wedge represents the percentage of the total that corresponds to a particular observation.

Example: A Pie-Chart for the musical preferences of Freshmen.

| Artist	# of students	Percentages	Wedge sizes
Kanye West	8	$8/20 = .4$ (40%)	$(.40)(360^o) = 144^o$ for Kanye West
Hanna Montana	5	$5/20 = .25$ (25%)	$(.25)(360^o) = 90^o$ for Hannah Montana
Beatles	4	$4/20 = .2$ (20%)	$(.20)(360^o) = 72^o$ for the Beatles
Other	2	$2/20 = .1$ (10%)	$(.10)(360^o) = 36^o$ for "other"
No response	1	$1/20 = .05$ (5%)	$(.05)(360^o) = \underline{18^o}$ for "no response"
Total	**20**		$360^o \leftarrow$ total must be 360°

Musical Preferences for freshmen

Histograms are another common type of data display. They are often used in the media to report quantitative information and are also widely used by practicing scientists and engineers to reason about their observations. The height of each bar of a histogram shows the number of observations that fall within a given range of values.

Example: A histogram for annual rainfall in the United States. The table below gives data for the average number of inches of rain that fall per year in each state.[1]

Alabama	56.9	Hawaii	23.5	Massachusetts	43.8	New Mexico	8.9	South Dakota	17.5
Alaska	53.2	Idaho	11.7	Michigan	32.2	New York	39.3	Tennessee	48.5
Arizona	7.1	Illinois	33.3	Minnesota	26.4	North Carolina	42.5	Texas	34.7
Arkansas	49.2	Indiana	39.1	Mississippi	52.8	North Dakota	15.4	Utah	15.3
California	17.3	Iowa	34.7	Missouri	33.9	Ohio	37.8	Vermont	33.7
Colorado	15.3	Kansas	28.6	Montana	11.4	Oklahoma	30.9	Virginia	45.2
Connecticut	44.4	Kentucky	43.6	Nebraska	30.3	Oregon	37.4	Washington	27.7
Delaware	41.4	Louisiana	59.7	Nevada	7.9	Pennsylvania	40.3	West Virginia	40.7
Florida	49.9	Maine	43.5	New Hampshire	36.5	Rhode Island	41.9	Wisconsin	30.9
Georgia	48.6	Maryland	41.8	New Jersey	41.9	South Carolina	51.6	Wyoming	13.3

To make a histogram you first have to decide on the ranges of values to use for each bar. This range is called the *bin width*. For this data set, 10 inches is a reasonable bin width. Next you need to count how many states are in the range 0-10 inches, 10-20 inches, and so on. You then make the height of each bar equal the number of states that fall in its range. The 30-40 inch bar, for example, is 14 units tall because 14 states have an average annual rainfall between 30 and 40 inches.

Inches of rain per year

---

[1] http://www.betweenwaters.com/etc/usrain.html
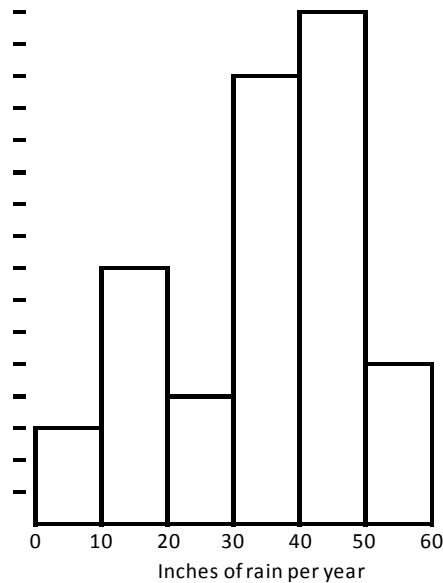
Exercises

1. A listing posted by Arby's restaurant gives, for each of the sandwiches it sells, the type of meat in the sandwich, the serving size in ounces, and the number of calories.

    The categorical variables are _____.

    The numerical variables are _____.

2. In a group of Lakeside students who gave blood, the distribution of blood types was:

   O: 32   B: 21   A: 24   AB: 12

    Draw an accurate a pie chart for this situation. Round percents to the nearest 0.1%. For the wedges, round your angles to the nearest whole degree. Show your work.

3. Use the histogram to answer the following questions about the heights of freshman boys. One unit on the vertical scale represents one boy.

    a. How many boys were from 70 to 74 inches tall?

    b. How many total boys were involved in the study?

    Heights of freshman boys

    c. How many boys were 71 inches tall?
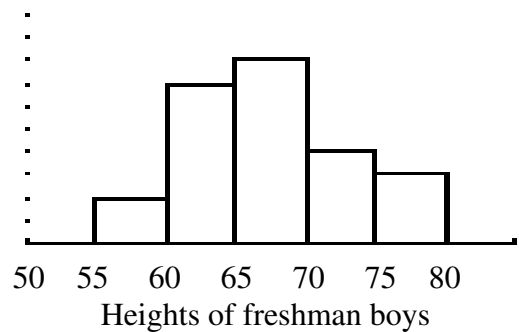
4. According to *The World Almanac,* record high temperatures (in degrees Fahrenheit) for each of the 50 states are as follows:

112	110	107	116	120	100	118	112	108	113
127	117	114	110	120	120	116	115	121	117
134	118	118	113	105	118	122	117	120	110
105	114	118	119	118	110	114	122	111	112
109	105	106	104	114	112	109	110	111	114

Graph the data in a histogram. Use a bin width of 5 degrees, starting at 100. Be sure to label your graph. Observations on the border between two bins should be included only in the higher bin. For example, a temperature of 105 degrees should be included in the 105-110 bin and not the 100-105 bin.

5. Miguel was grouping the number of athletes in each grade. He said that he didn't have any categorical data, as "grades" are numerical (9 through 12) and the number of athletes is also a numerical variable. Is he correct? Explain.

6. A *stem and leaf* plot is a variation on the histogram where the observations themselves create the bars. The *stem* of the graph is the set of vertical numbers. The *leaves* are the horizontal numbers.

The graph below shows the amount of money (rounded to the nearest dollar) found in the pockets of randomly chosen Pike Place Market shoppers. Each number on the "stem" represents a multiple of 10, and each "leaf" is a number from the units place. Repeated numbers mean that observation occurred more than once. For example, the 4 4 4 in the 3rd row means three people had $24 in their pockets.

Pocket Cash

0	0 1 2 2 3 7
1	0 3 5 6 6 8
2	1 2 4 4 4 8
3	0 2 5
4	6 9

Key: 2|1 means $21

a. How many shoppers had between $10 and $19 (inclusive)?

b. How many shoppers had exactly $16?

Exactly $38?

c. How many shoppers were involved in the study?

d. What is the range of amounts, from smallest to largest? (Note: it is helpful to include the smallest and largest values, but the *range* is actually just one number.)

e. Starting from the smallest number and counting up, what is the middle number?

7. Put the following math test performance data on a stem-and-leaf graph (we also call these *stemplots*). Generally, we first put the data in the appropriate row, and then make another plot with the data points in order.

Math Test Scores (out of 100 points):
71, 93, 95, 82, 83, 77, 71, 88, 84, 73, 91, 92, 85, 82, 76, 63, 50, 92, 87, 75, 71, 80, 68, 93

Initial Stemplot (unordered leaves)   Revised Stemplot (ordered leaves)

a. How many students had scores below 70?

b. If scores in the 60s are Ds, in the 70s Cs, 80s Bs, and 90s As, what was the most common grade? What was the middle grade?

c. Does a stemplot have any advantages or disadvantages over a histogram?

Measures of Central Tendency

A primary purpose of statistics is to *summarize* data. The simplest summary of a data set is to describe it with a single number that represents the "typical" observation. Depending on the situation, any of three different numbers are commonly used as "typical". These *measures of central tendency* are the *mean*, the *median*, and the *mode*.

The mean is the result of adding up the values of all the observations in a data set then dividing by the number of observations. It is usually (though not always!) what is meant by the term *average*.

The median is the "middle number". More precisely, it is the number that half of the observations in the data set are less than and half are greater than. If there are an odd number of observations, the median is *the* middle number. If there are an even number of observations, the median is the mean of the *two* middle numbers.

The mode of a set of observations is the number (or category) that occurs most often. It is possible for a set of data to have more than one mode. In cases when a set of data has more than two modes, the mode is generally not a useful measure.

Example: The mean, median and mode of heights of Seattle Storm basketball players

$$\left(\begin{array}{c} \text{mean} \\ \text{height} \end{array}\right) = \frac{69+73+74+77+74+76+71+71+77+72+71}{11}$$

$$\left(\begin{array}{c} \text{mean} \\ \text{height} \end{array}\right) \cong 73.18 \text{ inches}$$

Name	Position	Height
Sue Bird	G	5' 9" = 69 in.
Swin Cash	F	6' 1" = 73 in.
Ify Ibekwe	F	6' 2" = 74 in.
Lauren Jackson	F-C	6' 5" = 77 in.
Camille Little	F	6' 2" = 74 in.
Ashley Robinson	C	6' 4" = 76 in.
Katie Smith	G	5' 11" = 71 in.
Belinda Snell	G	5' 11" = 71 in.
Krystal Thomas	F	6' 5" = 77 in.
Le'coe Willingham	F	6' 0" = 72 in.
Tanisha Wright	G	5' 11" = 71 in.

To find the median it is usually best to first order the data:
$$\text{heights} = \{69, 71, 71, 71, 72, \underline{73}, 74, 74, 76, 77, 77\}$$
There are five players shorter than 73 inches and five taller than 73 inches, so 73 inches is the median height.

The height that occurs most often is 71 inches, so it is the mode of this data set.

Example: Median of a data set with an even number of observations
Imagine there was a 12th player added to the Storm roster who was 78 inches tall. The ordered data would now look like:
$$\text{heights} = \{69, 71, 71, 71, 72, \underline{73}, \underline{74}, 74, 76, 77, 77, 78\}.$$

There are five numbers less than 73 inches, which is the 6th largest, and five numbers greater than (or equal to) 74 inches, which is the 7th largest. The median is thus 73.5 inches, the mean of these *two* middle observations.

Exercises

1. Which measure of central tendency would the following represent?

    Jojo was happy that his GPA had gone up to 3.37.   mean / median / mode

    The typical student in the class wore glasses.   mean / median / mode

    In her family of five, Bertha is the third tallest.   mean / median / mode

2. The table below shows Kobe Bryant's scoring for the first 10 games of the 2007-2008 NBA season.

    a. Find the mean, median, and mode number of his points scored.

Game	Pts
@ UTA	28
vs. DEN	24
vs. SEA	35
vs. NJN	31
@ BOS	28
@ MIL	27
@ IND	32
vs. CHI	18
vs. DET	19
@ HOU	30

    Mean:

    Median:

    Mode:

    b. NBA.com listed his PPG (points per game) as 27.2. What measure of central tendency were they using?

    c. Assuming he maintained these scoring trends for the rest of the season, what will be his most likely point total in the 30th game?

    d. Again assuming he maintained these scoring trends, how many total points did he score over the entire 82-game regular season?

3. A six-question quiz was given, with each question worth 5 points. The scores were as follows:

Score	Frequency
10	2
15	5
20	8
25	15
30	10

Frequency means how many of each score occurred. For example, 5 students scored 15. Find the mean, median, and modal score.

4. The histogram shows sales data for the past month at an auto dealership. The horizontal scale shows the number of cars sold by individual salespeople. The vertical scale gives the number of people who sold that number of cars.

**# of people
(scale = 1)**
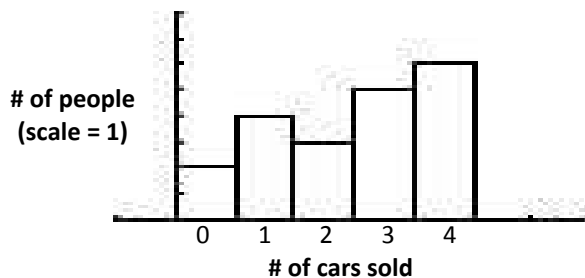
# of cars sold

a. How many salespeople sold only one car this past month?

b. How many salespeople work at this dealership?

c. How many cars were sold at this dealership last month?

d. What was the mean number of cars sold by a salesperson?

e. What was the median number of cars sold by a salesperson?

f. What was the modal number of cars sold by a salesperson?

5. If possible, find the mean, median, and mode from the following histogram. If you can't find the exact number, can you estimate any of the measures of central tendency? (Note: the first bin is from $80-$89.99)
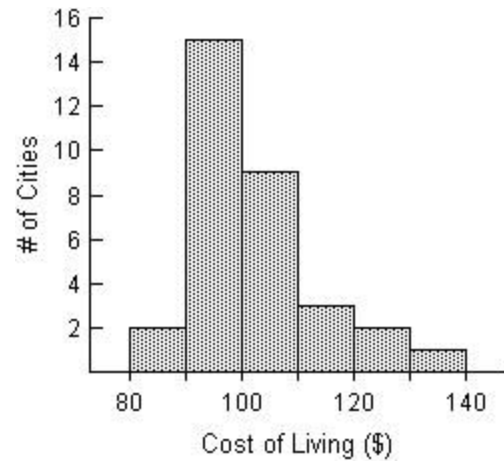
6. Harold scored 87, 82, 95, and 89 on his first four tests.
  a. What score will he need to get on his next test to have a mean score of 90?

  b. Could Harold receive a score on his next test that would give him a median of 90? If so, what score would he need? If not, why not?

7. Make up a set of six numbers that has a mean of 100 and a median of 95.

8. In a group of 6 teenage students (ages 13 to 19), the mean age is 16, the median is 16.5, and the mode is 17. If a 21 year-old joins the group, what would be the new mean, median, and mode?

9. A volleyball coach asks his 7-player team "Should we scare our next opponents by telling them our mean height, or should we give them false hope by telling them our median height?"

   Give an example of a team that might reflect this situation.

10. A travel magazine determines the daily cost of food for travelers visiting many different cities. To summarize the results, the magazine publishes this graph:
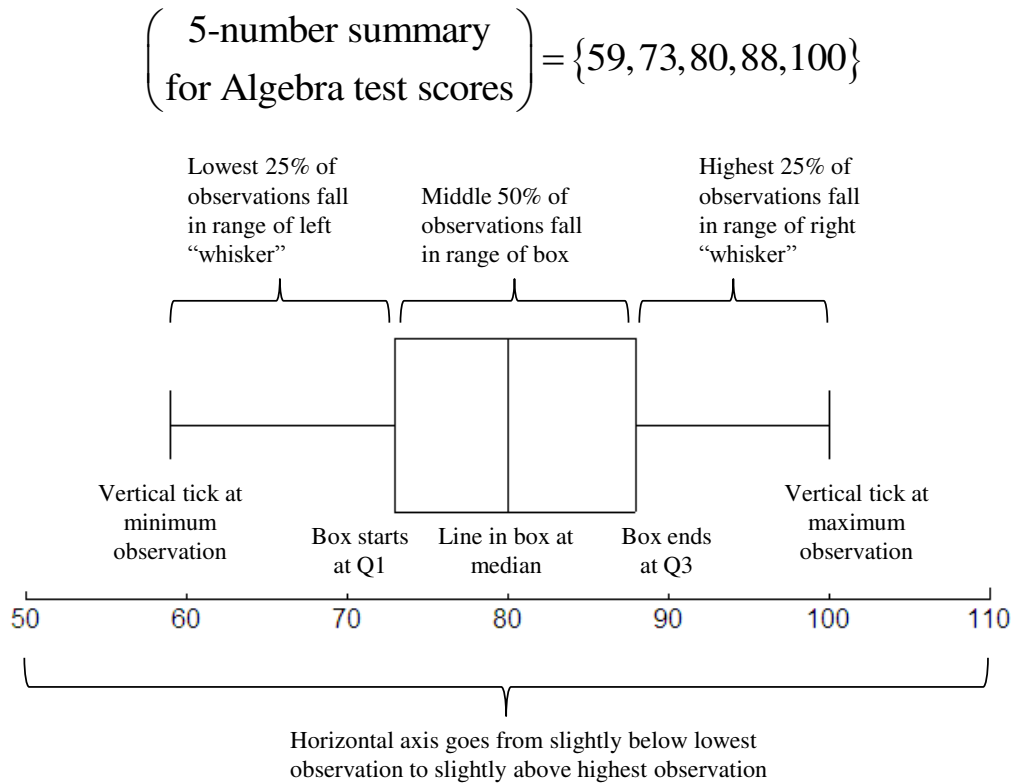

DAILY FOOD COST IN DOLLARS
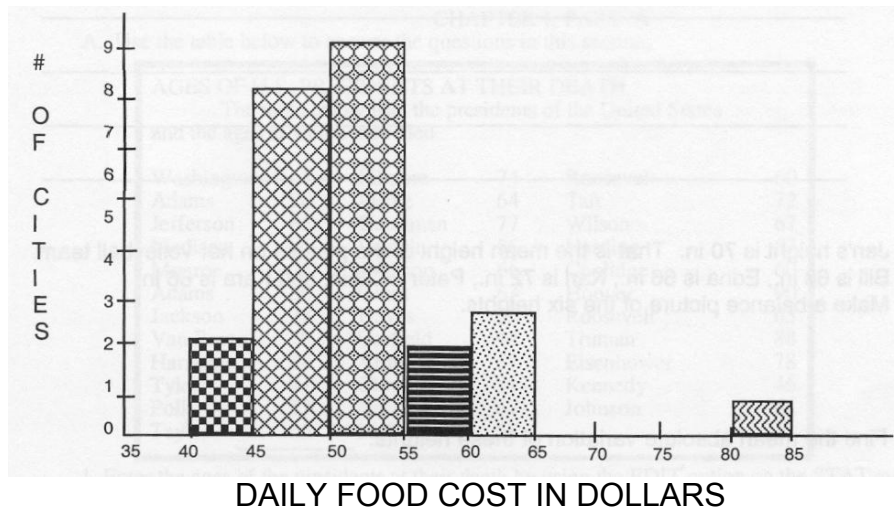
   a. How many cities are represented in the above graph?

   b. In which interval (40-to-45, 45-to-50, etc.) does the median daily food cost fall? How do you know this?

Distribution of Data: The 5-number Summary and Boxplots

In the previous section we learned about three different ways to summarize a set of observations with a single number. While this is concise and useful, we often want a more detailed summary to give us a better sense of trends in the set of observations.

The *five-number summary* is commonly used to give a more detailed picture of a data set than can be found by looking only at a measure of central tendency.

The five numbers of the five number summary

First number	*minimum*	The lowest value
Second number	*first quartile-Q_1* (25th *percentile*)	The number where 25% of the observations are lower and 75% are higher. Q_1 is also the median of the lower half of the data.
Third number	*median* (*second quartile-Q_2* or 50th *percentile*)	The number where 50% of the observations are lower and 50% are higher
Fourth number	*third quartile-Q_3* (75th *percentile*)	The number where 75% of the observations are lower and 25% are higher. Q_3 is also the median of the higher half of the data.
Fifth number	*maximum*	The highest value

Example: Here are student scores on an Algebra test.

59, 60, 64, 67, **72, 74**, 77, 77, 78, 79, **80**, 80, 81, 83, 84, **88, 88**, 91, 93, 94, 100

Middle of Lower Half
Q1 = 73

Middle of Upper Half
Q3 = 88

Low   **Median**   **High**

$$\left(\begin{array}{c}\text{5-number summary} \\ \text{for Algebra test scores}\end{array}\right) = \{59, 73, 80, 88, 100\}$$

A *boxplot* (or *box and whisker plot*) is a type of data display that allows you to visualize a data set through its 5-number summary. The graphic below uses the test score data set from the previous example to explain how a boxplot works.

$$\begin{pmatrix} \text{5-number summary} \\ \text{for Algebra test scores} \end{pmatrix} = \{59, 73, 80, 88, 100\}$$

Lowest 25% of observations fall in range of left "whisker"

Middle 50% of observations fall in range of box

Highest 25% of observations fall in range of right "whisker"

Vertical tick at minimum observation

Box starts at Q1

Line in box at median

Box ends at Q3

Vertical tick at maximum observation

50   60   70   80   90   100   110

Horizontal axis goes from slightly below lowest observation to slightly above highest observation

Recall that the *range* of an entire data set is the difference between the smallest and largest values. The range of this test score data is $100 - 59 = 41$ points.

The *interquartile range* (*IQR*) of a data set is the difference between Q3 and Q1. You can also think of it as the range of the middle 50% of the data or as the width of the box part of the boxplot. The interquartile range of this test score data is $88 - 73 = 15$ points.

Outliers are observations that are either particularly low or particularly high. One common way of identifying outliers involves the IQR.

A *Lower Outlier* is any observation that is less than $Q1 - 1.5 \cdot IQR$. For the test score data any lower outliers would be lower than $73 - 1.5 \cdot 15 = 50.5$ points, so there are no lower outliers.

An *Upper Outlier* is any observation that is greater than $Q3 + 1.5 \cdot IQR$. For the test score data any upper outliers would be above $88 + 1.5 \cdot 15 = 110.5$, so there are no upper outliers.

Example: Temperature data

Seattle High Temperature (F)

	By date		High to low		
	2010	**2011**	**2010**	**2011**	
1-May	56	62	72	70	← MAX
2-May	53	53	71	68	
3-May	54	58	71	66	
4-May	53	62	71	66	
5-May	48	57	69	65	
6-May	61	54	68	64	
7-May	65	56	67	64	
8-May	65	58	65	63	← Q3
9-May	67	57	65	62	
10-May	55	59	65	62	
11-May	62	51	63	62	
12-May	62	60	62	61	
13-May	69	64	62	61	
14-May	72	66	62	61	
15-May	71	51	61	60	
16-May	71	56	61	60	← MEDIAN
17-May	71	64	61	59	
18-May	63	66	60	59	
19-May	68	68	59	58	
20-May	56	70	57	58	
21-May	55	54	57	58	
22-May	61	60	56	57	
23-May	57	63	56	57	
24-May	62	65	55	56	← Q1
25-May	61	56	55	56	
26-May	57	61	54	56	
27-May	60	58	54	54	
28-May	53	62	53	54	
29-May	54	59	53	53	
30-May	59	61	53	51	
31-May	65	61	48	51	← MIN

<u>5-number summaries</u>

$$(\text{Year } 2010) = \{48, 55, 61, 65, 72\}$$

$$(\text{Year } 2011) = \{51, 56, 60, 63, 70\}$$

<u>Boxplots</u>
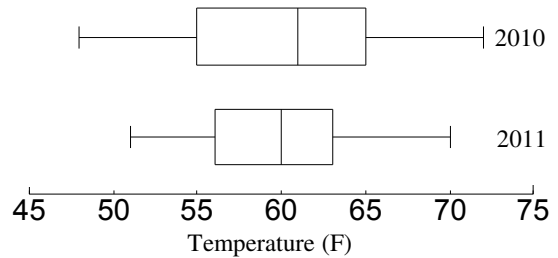
In 2010 the range of the May high temperatures was $72 - 48 = 24°$ and the interquartile range was $65 - 55 = 10°$. No highs were above $65 + 1.5 \cdot 10 = 80°$ nor below $55 - 1.5 \cdot 10 = 40°$, so there were no upper or lower outliers.

In 2011 the range of the May high temperatures was $70 - 51 = 19°$ and the interquartile range was $63 - 56 = 7°$. No highs were above $63 + 1.5 \cdot 7 = 73.5°$ nor below $56 - 1.5 \cdot 7 = 45.5°$, so there were again no upper or lower outliers.

Exercises

1. Three different auto dealerships were competing to see who had the best sales staff. Each dealership had four salespeople, and at the end of the month they had the total sales as shown below. For each dealership, find the mean and median number of sales and the range.
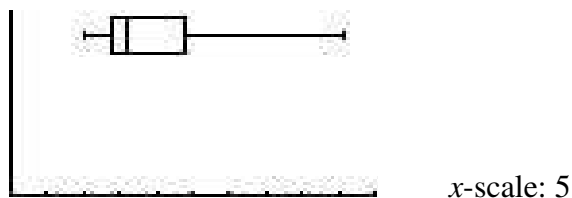
Dealer	Sales	Mean	Median	Range
A	5, 5, 5, 5			
B	2, 4, 6, 8			
C	0, 0, 10, 10			

2. Twelve babies spoke for the first time at the following ages (in months):

   15, 26, 10, 9, 15, 20, 18, 11, 8, 20, 12, 12

   Find the 5-number summary: Min_____, Q1_____, Med_____, Q2_____, Max_____

3. A data set has a 5-number summary of {10, 14, 16, 24, 46}

   a. Find the Range_____   and the *Interquartile Range*_____.

   b. The boxplot for this data set would look like this:

   x-scale: 5

   When interpreting a box-plot, note that the graph is divided into four main parts, each representing 25% of the data. **The width of each part does not indicate more data; it only indicates that the data is spread out more in that 25%.**

   Which 25% has the smallest spread?

   The top 25% of the data is spread between what two values?

4. A set of data has a five-number summary of {40, 67.5, 200, 285, 420}
   a. What is the interquartile range?
   b. Draw a box-and-whisker plot of the data. Put an appropriate scale on the segment, mark the extremes, median, and quartiles, then draw the box and the whiskers.

5. The annual numbers of deaths from tornadoes in the United States from 1990 through 2000 (according to NOAA) are: 53, 39, 39, 33, 69, 30, 25, 67, 130, 94, 40

Find (by hand) the 5-number summary, and make a boxplot of the data.

6. The stemplot shows rainfall data for Pas Christian, Mississippi from 1989 to 2005. Find the 5-number summary.

```
7 | 2
6 | 4
5 | 1 5
4 | 3 4
3 | 2 3 3 5 6 8
2 | 7 8
1 |
0 | 4 6 6
```

7. A group of 6th grade boys and a group of 6th grade girls were tested on how many sit-ups they could do in 30 seconds. The data is as follows:

Boys: 17, 18, 22, 22, 29, 22, 24, 23, 21, 17, 23

Girls: 20, 25, 22, 27, 12, 19, 28, 24, 22, 21, 25, 26, 25, 16

Using your calculator, put the data for the boys into L1, and the girls into L2. Find the 5-number summary for both sets, and graph side-by-side box and whisker plots. Copy the plots and your 5-numbers onto this page.

Using the graphs as a guide, which set (boys or girls) would you say are more sit-up fit? Explain. (Use the interquartile range in your answer)

8. Three classes took the same science test. Here is a boxplot showing the scores in each class.
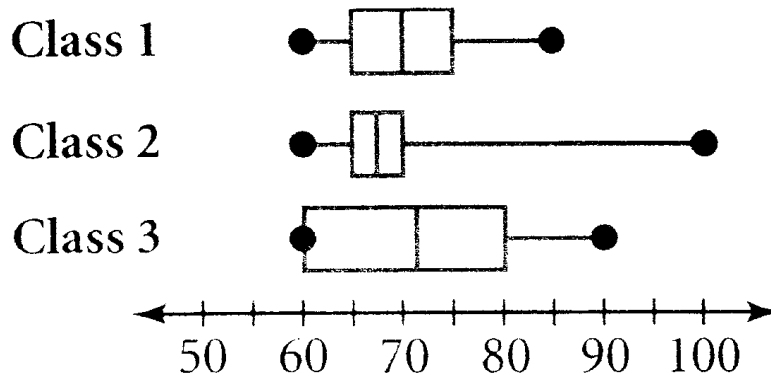
a. Which class had the highest-scoring student, and what was this student's score?

b. Rank the three classes based on their median test scores.

      best: middle: worst:

c. If a student had to score at least 65 to pass the test, what percent of the students in Class 1 passed the test?

d. The teacher of Class 3 allows all students who scored less than 80 to retake the test. What percent of Class 3 is allowed to retake the test?

e. To get an A on the test, students need to be in the top 25% of their class. What score would a student in Class 1 need to get in order to get an A?

9. These boxplots show the fuel efficiency (miles per gallon) for cars from three different manufacturers (A, B, and C).
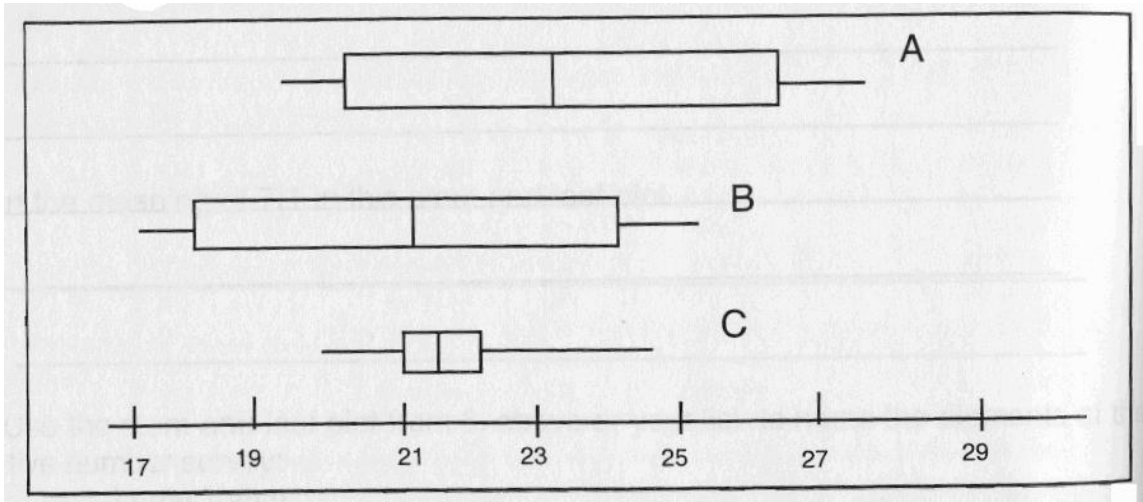
a. Rank the three manufacturers based on the median fuel efficiency.

　　　　best:　　　　　　　　middle:　　　　　　worst:

b. If someone wanted to buy the most fuel efficient car available, which manufacturer would it come from?

c. If someone wanted to send a protest letter to the manufacturer of the least fuel efficient automobile, who should get it?

d. What percent of the cars from manufacturer C have a fuel efficiency of at least 21 mpg?

e. The *range* of a set of data is the difference between the maximum and minimum values. Which manufacturer has the smallest range of fuel efficiencies? What is that range?

The Standard Deviation and Standardized Scores

When analyzing data we use the mean, median and/or mode to tell us about a "typical" observation from a data set. We are also often interested in how much the observations vary from one another. In the last section we learned about the interquartile range, which tells us the spread in the middle 50% of the observations. The *standard deviation* is another very important measure of the variation in a set of observations. It tells you the average difference between a typical observation and the mean.

There are actually two different standard deviations studied in advanced statistics. For the purposes of this class, we will use the *population standard deviation*, which is designated by σ (lower-case sigma).

The formula for the standard deviation for *n* a data set with *n* observations is

$$\sigma = \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+(x_3-\bar{x})^2+...+(x_n-\bar{x})^2}{n}} \quad \text{or,} \quad \sigma = \sqrt{\frac{\sum(x_i-\bar{x})^2}{n}}$$

The symbol $\sum$ (which is a capital sigma) means to add up all of the terms. There are a lot of strange symbols in this formula. Let's break it apart step-by-step:

To find the standard deviation, we just need to perform some simple arithmetic:
- Find the mean (called $\bar{x}$, pronounced *x*-bar)
- Find the amount each value varies from the mean. $\quad (x_i - \bar{x})$
- Square each of the $(x_i - \bar{x})$ values. $\quad (x_i - \bar{x})^2$
- Add them up. $\quad \sum(x_i - \bar{x})^2$
- Divide by the number of values, *n*. $\quad \dfrac{\sum(x_i - \bar{x})^2}{n}$
- Finally take the square root of the whole thing. $\quad \sqrt{\dfrac{\sum(x_i - \bar{x})^2}{n}}$

Whew! That is a lot of steps, but though tedious, it is not too hard to actually compute. This process is actually easier to perform than describe.

Example: Say our data set consists of {2,3,6,7,9}. The mean ($\bar{x}$) is 5.4, and the sum of the squared differences from the mean is then

$$(2-5.4)^2+(3-5.4)^2+(6-5.4)^2+(7-5.4)^2+(9-5.4)^2=33.2. \text{ Thus, } \sigma = \sqrt{\frac{33.2}{5}} = 2.58$$

We would describe our set as having a mean of 5.4 and a standard deviation of 2.58. If we were to select an observation at random from this data set, it would be, *on average*, 2.58 units away from the mean of 5.4.

One important use of the standard deviation is to compare observations from different data sets.

Example: Randy took two different tests. He scored a 77 in math and an 86 in physics. Relative to the other students, in which class did he do better?

To answer this question we need to look at the data for the physics and math tests, which is shown in the table below. Specifically, we will calculate the mean and standard deviation of each data set and then consider Randy's scores in the context of these numbers.

Math	Physics
68	76
71	79
71	81
77	84
75	85
70	84
77	82
70	76
70	77
69	74
74	86

In his math class the mean test score was 72 and in physics it was 80. Calculate the standard deviations:

Math

$$(68-72)^2 +(71-72)^2 +(71-72)^2 +(77-72)^2 +(75-72)^2 +(70-72)^2 +\ldots$$
$$(77-72)^2 +(70-72)^2 +(70-72)^2 +(69-72)^2 +(74-72)^2 =102$$
$$\frac{102}{11} \cong 9.27 \quad \rightarrow \quad \boxed{\sigma = \sqrt{9.27} = 3.05}$$

Physics

$$(76-72)^2 +(79-72)^2 +(81-72)^2 +(84-72)^2 +(85-72)^2 +(84-72)^2 +\ldots$$
$$(82-72)^2 +(76-72)^2 +(77-72)^2 +(74-72)^2 +(86-72)^2 =178$$
$$\frac{178}{11} =16 \quad \rightarrow \quad \boxed{\sigma = \sqrt{16} = 4}$$

In math his score was $77-72=5$ points above the mean. Since the standard deviation was 3.05, he was $5/3.05 =1.64$ standard deviations above the mean in math. In physics his score was $86-80=6$ points above the mean. The standard deviation on this test was 4, so he was $6/4=1.5$ standard deviations above the mean in physics. So, relative to the other students in his classes, he scored better in math!

We can make comparisons between observations from different data sets by finding their *standard scores* (often called a *z-scores*). The z-score is the number of standard deviations away from the mean for a given observation. We find the z-score of an observation by subtracting the mean from it and then dividing by the standard deviation.

The formula is easy: $z = \dfrac{x - \bar{x}}{\sigma}$, where $\bar{x}$ is the mean, x is the value of the observation of interest and σ is the standard deviation of the data set.

For the two examples above, Randy's z-score for physics was $z = (86 - 80)/4 = 1.5$ and in math it was $z = (77 - 72)/3.05 = 1.64$.

Exercises

1. Lakeside's girls basketball team has a mean height of 165 cm with a standard deviation of 12 cm. The Ingraham girls team has a mean height of 165 cm. with a standard deviation of 6 cm. What does this tell you about the heights of players on the two teams?

2. A paper appearing in *Animal Behavior* ("The Echolocation of Flying Insects by Bats") gave the following distances (in cm.) at which a small population of bats first detected a nearby insect:

$$62, 23, 27, 56, 52, 34, 42,\ 40, 68, 45, 83$$

   a. Compute the mean distance at which the bat first detects an insect, and find the standard deviation. Interpret your results.

   b. As with the five-number summary, you should find the mean and standard deviation by hand if asked to use the formula. If you're not required to do so, you can put the data into a list, go to STAT-CALC, then 1-Var Stats, and input your list number. Remember, the standard deviation you want is σx. Use your calculator to verify your answer to part a.

3. Without doing the calculations, which set would have the largest standard deviation? Explain.

   A: 7, 9, 11, 14, 14, 14, 17, 18   B: 7, 7, 8, 8, 14, 17, 18, 18

4. Can the standard deviation ever be a negative number? Why or why not?

5. Histograms for test scores are shown for three different classes.

1ˢᵗ period   2ⁿᵈ Period   3ʳᵈ Period
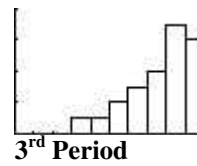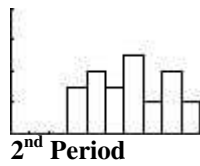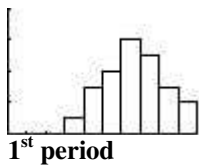
For all histograms, (*x*-scale: 10, *y*-scale: 2)

a. Which class was most likely to have had the highest mean score? Explain your answer.

b. Which class had the smallest standard deviation? Explain.

6. 10 different models of new energy efficient dishwashers had prices of $850, 900, 1400, 1200, 1050, 1000, 750, 1250, 1050, and 560. Find the mean and standard deviation *from the formulas*.

7. Repeat exercise 4 using your calculator.

8. When professional baseball players went on strike a few years back, the owners told the media that the players were overpaid, making a typical salary of $1.1 million. The players union, on the other hand, said that a typical player was making only $320,000. To answer the following questions, you need to think about how professional athletes are paid.

a. Both groups were telling the truth. Which type of central tendency was used by the owners? The players? Why would each group pick the method they used?

b. Which measures would best describe the distribution of pay for the baseball players: the mean and standard deviation, or the 5-number summary? Explain the rationale for your choice
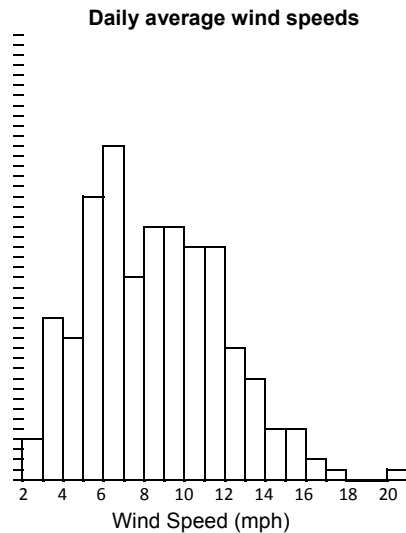
9. Calculate the z-scores of the following observations from data sets with the given means and standard deviations.

    a. 45, if the mean=36 and the standard deviation is 3.

    b. 600, if mean=500 and std. dev. =25.

    c. 92, if mean =100 and std. dev. =10.

10. What does it mean (in general) if a standard score is negative?

11. Riley was told his standard score on a test was 2. If the mean was 80, and the std. dev. was 6, what score did he have?

12. Bertha had a vertical jump-reach of 30 inches, while the mean was 36. If her standard score was −1.5, what was the standard deviation?
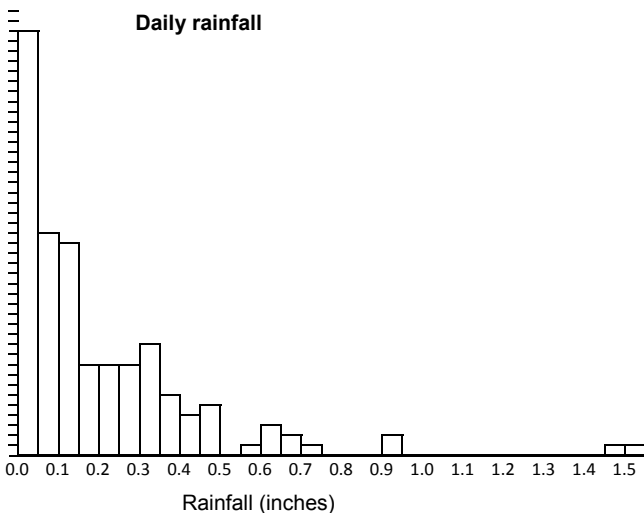
The Normal Distribution

The *distribution* of a set of data refers to the pattern with which the observations are arranged. Below you will find cumulative temperature, wind speed and rainfall data in Seattle during the month of March over the last 8 years. In the histograms one vertical unit constitutes one day. From looking at the histograms you can see that these three data sets have distinctive distributions. For example, the temperature distribution has a clear and narrow peak, the wind speed distribution has a broader peak (if it has a peak all) and the rainfall distribution is highly asymmetric. Note that there were 99 days with zero inches of rain. These days had to be omitted from the rainfall histogram to get it to fit on the page.

Daily high temperatures

Temperature (degrees F)
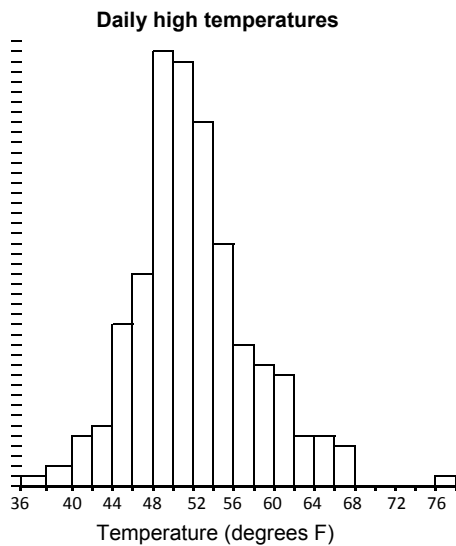
5 # summary = {37, 49, 52, 55, 77}
IQR = 6
Mean = 52.6
St. dev. = 5.8

Daily average wind speeds

Wind Speed (mph)

5 # summary = {2.2, 6.0, 8.3, 10.9, 20.8}
IQR = 4.9
Mean = 8.5
St. dev. = 3.3

Daily rainfall

Rainfall (inches)

5 # summary = {0, 0, 0.03, 0.15, 1.5}
IQR = 0.15
Mean = 0.12
St. dev. = 0.21

Often, a histogram of data will tend to be mound-shaped, meaning that there are more data points close to the middle and fewer as you move towards either the left or right ends. If the data set is large enough for the underlying pattern of the distribution to show, it will have a characteristic bell-shape. These bell-shaped curves may fit what is known as a *Normal Distribution.*

Data following a normal distribution will have the following characteristics:
1. The mean will be equal to the median, and both will occur at the peak of the histogram.
2. The histogram will be symmetric about it's peak.
3. There will only be one peak, and the tails of the graph (to the right and left) will become close to zero for values far above and below the peak.
4. 68% of the data will be within ±1 standard deviation of the mean and 95% will be within ±2 standard deviations.

The normal distribution is the most commonly observed pattern in data sets from all fields of study. Remember, though, that (like all mathematics…) the normal distribution is only a good model. Rarely, if ever, will a population be *exactly* normal. Furthermore, there are many important things that ***do not*** follow the normal distribution.

Consider again the three data sets from the previous page. The daily rainfall distribution is not at all bell-shaped and it does not satisfy any of the conditions for normally distributed data. It is still a perfectly valid distribution, we just cannot use properties of the normal distribution to reason about it. Both the high temperature and wind speed data sets are bell-shaped and both approximately fit the characteristics of normally distributed data. Perhaps the wind speed distribution is a bit asymmetric, but this may just be due to the data set being too small for the true underlying pattern to show through.

Here is a primary reason why the normal distribution is so helpful:

If you know (or can estimate) the mean and standard deviation of a data set, and you have reason to believe that the data set is normally distributed, then you can make detailed and accurate predictions about the data you are studying.

Specifically, if you know the mean and standard deviation of a data set, you can find the value of *any* percentile for that data. The Z-score is the key to this process. If an observation from a normally distributed has a Z-score of -2, then it is at the 2.28 percentile. This means that 2.28% of the data will be less then this observation and 100-2.28 = 97.32% will be greater then it. If an observation has a Z-score of 0.5, then it will be at 69.15 percentile. Each Z-score for normally distributed data corresponds to a specific percentile. The converse is also true. Each percentile for normally distributed data corresponds to a specific Z-score. The correspondence between Z-scores and percentiles is given in the *standard normal table* at the end of this chapter.

Example: The March high-temperature data had a mean of $\bar{x} = 52.6$ and a standard deviation of $\sigma = 5.8$. In what percentile would a 56 degree day fall? What temperature would correspond to a day in the 1st percentile?

First find the Z-score for a 56 degree day in march: $Z = \dfrac{56 - 52.6}{5.8} = 0.59$

Now look on the table. The bold numbers in the far left column give the first two digits in the Z-score. We find 0.5 is the 6th row from the top on the second page. The non-bold numbers inside the table give percentiles. Each column of the percentile numbers is for a different hundreds digit on the Z-score. The first percentile to the right of the bold 0.5, .6915, corresponds to Z = 0.50 The second percentile to the right of the bold 0.5 is for Z = 0.51, and so on. We must go to the last column of the 0.5 row to find .7224. This is the percentile corresponding to our Z-score of 0.59. We thus conclude that a 56 degree day in March is at the 72.24 percentile.

To find the temperature of a first percentile day, we have to work backwards. Start at the table and locate a *percentile* of .01. On the first page of the table, the 12th row, 3rd column is a percentile of .0102 and the 12th row 4th column is a percentile of .0099. This means that a percentile of *exactly* .0100 would have a Z-score between -2.32 and -2.33. Assume that Z = -2.33 is close enough to the 1st percentile. This means that the temperature of such a day is 2.33 standard deviations below the mean temperature, which works out to 52.6 – (2.33)(5.8) = 39.1 degrees. Thus, the Normal Distribution predicts that about 1 out of 100 days in March will be 39 degrees or colder.

Exercises

1. Say that scores on a standardized exam are normally distributed with a mean of 1000 and a standard deviation of 100. You scored 1125. You want to see where your score is in relation to other test-takers.

    a. Find your standard score (your Z-score).

    b. Use the table to find your percentile.

    c. If 5000 students took this exam, how many of them scored *better* then you?

2. For data that follows the normal distribution, what percentage of the observations are

    a. less than $z = 2.1$?

    b. greater than $z = -1.76$?

    c. less than $z = 3$?

3. What is the z-score at the mean?

4. What percent of the observations are between $z = -1.5$ and $z = .76$?

5. What z-score corresponds to a value at the 99.18^{th} percentile?

6. The Wechsler IQ test is said to be normally distributed, with a mean of 100 and a standard deviation of 15. If the results are normally distributed, what percent of people should have scores

    a. Greater than 100?

    b. Less than 70?

    c. Between 85 and 115?

8. The mean weight of Bubba's rabbits was 450 grams with a standard deviation of 15 grams. Assuming a normal distribution, find the percent of rabbits that are

    a. Less than 435 grams.

    b. Greater than 480 grams.

    c. Between 435 grams and 480 grams.

    d. If Bubba has 3000 rabbits, how many would you expect to be larger than 480 grams?

9. The average salary for a Queens College full professor is $85,900. If the salaries are normally distributed with a standard deviation of $11,000, what percent of professors

   a. Make more than $90,000?

    b. Make less than $75,000?

10. The average commute to work (according to the 2000 census) is 25.5 minutes. If commuting times across the US are normally distributed with a standard deviation of 6.1 minutes, what percent of commutes take 30 minutes or more?

11. The average price of a personal computer is $949. If computer prices are normally distributed, and the standard deviation is $100, the least expensive 10% of computers will cost less than what amount?

12. The length of human pregnancies from conception to birth is normally distributed with an average length of 266 days, and a standard deviation of 16 days.
  a. What percent of pregnancies last less than 250 days? (a little more than 8 months)

  b. What percent last between 240 and 270 days?

  c. How long do the longest 20% of pregnancies last?

Additional Exercises

1.1 Data Displays

1. In a survey of 123 Lakeside freshmen, 45 took Spanish, 15 took Latin, 27 took Chinese, and the rest took French. Assuming all freshmen took a language, how many degrees are needed in each wedge of a pie chart for this data?

2. An individual has managed to save $9,000. The pie chart shows how they divided up their savings. If the stocks wedge measures 132°, the bonds wedge measures 92°, and the CD wedge measures 84°, how much money does this person have in each of their four categories?

■ stocks

■ bonds

■ CD

■ savings account
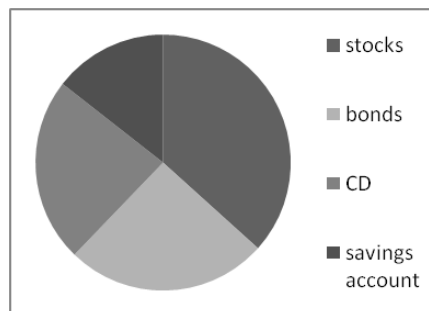
3. Make a histogram for the following data set, which gives the number of members of the US House of Representatives that come from each of the 50 states. Use a bin-width of three and have your first bin include states with 1, 2, and 3 representatives. Also, you should exclude California and its 53 representatives from your histogram.

1	1	2	3	5	6	8	9	13	19
1	1	2	3	5	7	8	9	13	25
1	2	3	4	5	7	8	10	15	29
1	2	3	4	5	7	9	11	18	32
1	2	3	4	6	8	9	13	19	53

4. The histogram displays data for the heights of the tallest buildings in the world. Each vertical unit represents one building.
   a. How many buildings are included in this histogram?
   b. How many buildings are taller than 1,400 feet?
   c. What is the bin width of this histogram?
   d. Between what two heights is the world's tallest building?

1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8

Height in thousands of feet
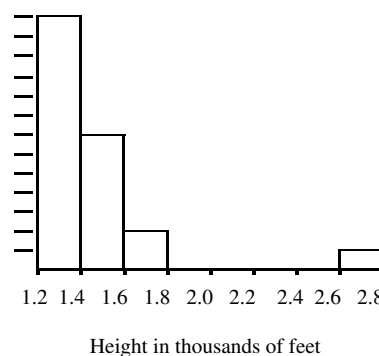
1.2 Measures of Central Tendency

1. Find the mean, median and mode of the following sets of numbers
   a. {-5, 3, 2, -1, 0, 0}   b. {12, 12, 13, 15, 12}   c. {123, 111, 115, 300}

2. Find the mean, median and modal score for the quiz result data in the table below. There were 6 points possible on this quiz.

score	0	1	2	3	4	5	6
# students	2	4	3	3	8	5	1

3. Six friends are taking the Lakeside physical fitness test. The first five did 15, 8, 14, 10, and 16 pull-ups. How many does the last friend need to do so that they can have a combined mean of 14 pull-ups? How many must he do for them to have a combined median of 14?

1.3 The 5-number summary

1. Find the 5-number summary of the following data sets, make box-and-whisker plots, find the range, find the IQR, and identify any outliers.
   a. {11, 3, 2, 13, 13, 6, 8, 5, 12}
   b. {1, 1.2, -1, 0, 0, -2, -3, 2.4, 0.5, 8, -4}
   c. {121, 87, 101, 134, 188, 167, 144, 144}

1.4 The standard deviation and standardized scores

1. Find the standard deviation of the following sets of numbers by hand then use your calculator to check.

   a. {5, 7, 3, 5}
   b. {-2, 0, 1, -3, -5, 2, 2, 4}

2. A data set has a mean of 25 and a standard deviation of 4.
   a. Find the standard score for an observation of 19 from this data set.
   b. Find the standard score for an observation of 26 from this data set.
   c. If an observation has a standard score of 2.25, what is its value?

1.5 The normal distribution

(i)

1. Use the histograms to answer the questions.
   a. Which distribution is the most asymmetric?
   b. Which distribution has the largest standard deviation?
   c. Which distribution has largest range?
   d. Which distribution has the highest mean
   e. Which distribution has the highest median?
   f. Which distribution best fits the conditions of the normal distribution?

(ii)

(iii)

2. The cows in farmer Brown's herd of cattle have a mean weight of 1,200 pounds with a standard deviation of 200 pounds. Assume that the cows weights of follow the normal distribution.
   a. What percentage of his cows weigh less than 1000 pounds?
   b. What percentage of his cows weigh between 1,100 and 1,500 pounds?
   c. 75% of his cows weigh less then _____ pounds.
   d. What is the inter-quartile range for the weights of his herd?
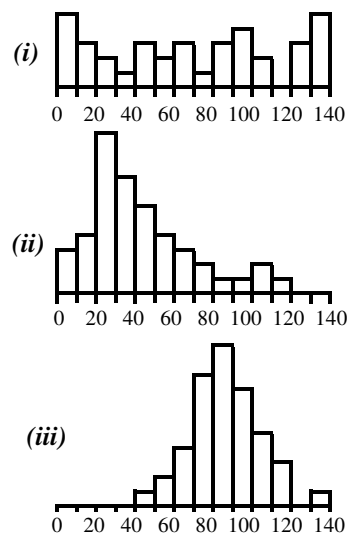
Solutions to Additional Exercises

1.1 Arithmetic Review

1. Spanish-131.7°, Latin-43.9°, Chinese-79.0°, French-105.4°

2. Stocks-$3,300, Bonds-$2,300, CD-$2,100, Savings-$1,300

4a. 23   **b.** 10   **c.** 2000 feet   **d.** 2,600 and 2,800 feet

3.

# of representatives
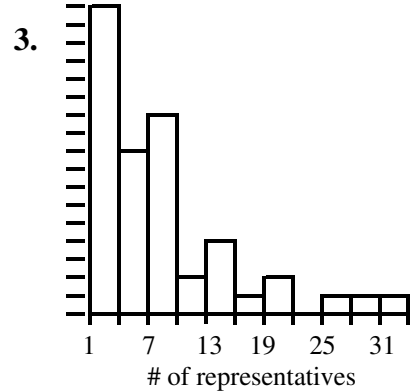
1.2 Measures of central tendency

1a. mean = -1.67, median = 0, mode = 0
1b. mean = 12.8, median = 12, mode = 12
1c. mean = 162.25, median = 119, there is no mode

2. mean = 3.15, median = 4, mode = 4

3. 21 pull-ups for mean of 14, 14 pull-ups for median of 14.

1.3 The 5-number summary

	5# summary	range	IQR	lower outliers	upper outliers
a	{2, 4, 8, 12.5, 13}	11	8.5	none	none
b	{-4, -2, 0, 1.2, 8}	12	3.2	none	8
c	{87, 111, 139, 155.5, 188}	101	44.5	none	none
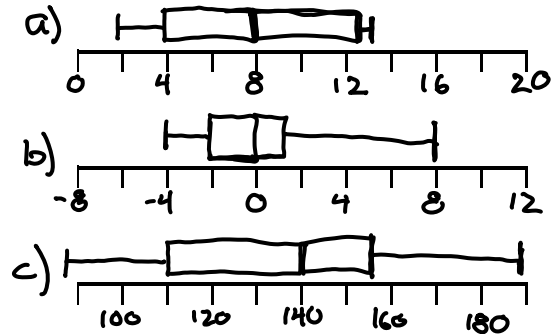
1.4 Standard deviation and standardized scores

1a. $\sigma = \sqrt{2} \approx 1.41$ **1b.** $\sigma \approx 2.80$

2a. $Z = -1.5$ **2b.** $Z = 0.25$ **2c.** observation $= 34$

1.5 The normal distribution

1a. *(ii)*   **1b.** *(i)*   **1c.** *(i)*   **1d.** *(iii)*   **1e.** *(iii)*   **1f.** *(iii)*
2a. 15.87%   **2b.** 62.47%   **2c.** About 1,334 lbs   **2d.** About 268 pounds

Standard Normal Cumulative Probability Table

Cumulative probabilities for NEGATIVE z-values are shown in the following table:
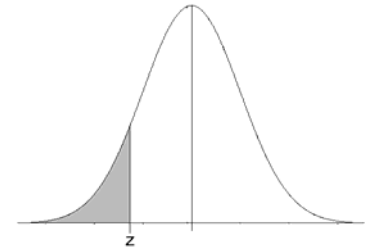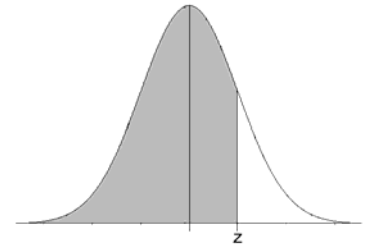
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	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	0.0630	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

Standard Normal Cumulative Probability Table

Cumulative probabilities for POSITIVE z-values are shown in the following table:

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	0.7704	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

Chapter 1 Topics Summary

Below is a brief summary of the topics covered in this chapter. Note that all of these topics are covered in greater depth within their individual sections.

1.1   Data Displays

- Categorical and numerical data
- Interpret and draw pie charts, histograms, and stem-and-leaf plots

1.2 Measures of Central Tendency

- From data, find and interpret measures of central tendency: mean, median, mode.
- Find measures of central tendency from stemplots and histograms.

1.3 Distribution of Data: The 5-Number Summary and Boxplots

- From a set of data, find the lower extreme, first quartile, median, third quartile, and upper extreme.
- Find the range and interquartile range.
- Draw and interpret box-and-whisker plots, by hand and with a calculator.

1.4 The Standard Deviation and Standardized Scores

- From a set of data, find – by hand and with calculator – the population standard deviation.
- Interpret the standard deviation in context of the distribution of a set of data.
- Compare use of the mean and standard deviation with the 5-number summary.

1.5 The Normal Distribution

- Know the general characteristics of the normal distribution.
- Understand how the standard normal table relates to the normal distribution.
- Apply the normal table in applications to normally distributed data given the mean and standard deviation.