MATH 109 Lecture Notes

2 Descriptive Statistics

2.1 Frequency Distributions

When the values of a quantitative variable are divided into classes or intervals, a frequency distribution displays the count (frequency) of subjects that fall into each class.

The range of a quantitative variable is the maximum (largest) data value minus the minimum (smallest) data value. When the variable is divided into classes, the minimum value in a class is called the lower class limit and the maximum value in a class is called the upper class limit. The class width is the difference between the lower class limit of the next class and the lower class limit of that class.

The midpoint of a class is the value exactly in the middle of the class limits (the average of the class limits).

The relative frequency of a class is the number of subjects in that class (frequency) divided by the total number of subjects (the size of the sample or the population).

The cumulative frequency of a class is the frequency of that class, added to the frequencies of all lower classes.

Example 2.1: Below is a list of the score of each winning team in the first 49 Super Bowls. Construct a frequency distribution with class width 10 for this data.

35	33	16	23	16	24	14	24	16	21
32	27	35	31	27	26	27	38	38	46
39	42	20	55	20	37	52	30	49	27
35	31	34	23	34	20	48	32	24	21
29	17	27	31	31	21	34	43	28	

Definition 1: A frequency histogram is a visual representation of a quantitative variable for a data set. It is built by compartmentalizing the variable values into classes and graphing the frequencies as bars. There should be no gaps or spaces between adjacent bars. Empty space should represent an actual gap in the data values. A relative frequency histogram graphs the relative frequencies rather than the number of subjects.

Note: To create a histogram, the data MUST be quantitative. When the bars of a histogram are drawn vertically, the horizontal axis measures the data values and the vertical axis measures the frequencies (or relative frequencies).

Example 2.2: Create a frequency histogram for the data in Example 1.

Example 2.3: Create a relative frequency histogram for the data below. What is the class width?

Number of Points	Frequency
0-4	1
5-9	8
10-14	12
15-19	14
20-24	9
25-29	3
30-34	2

Definition 2: A line graph which displays frequencies or relative frequencies is called a (relative) frequency polygon. A line graph which displays cumulative frequencies is called an ogive.

Example 2.4: Below is a list of the number of goals scored by each NHL team in the 2014-2015 season. Create a frequency polygon and an ogive for the data.

248	228	239	214	259	220
236	226	245	227	237	237
231	232	217	223	227	209
257	218	224	209	198	212
176	183	206	165	193	153

2.2 More Graphs and Displays

Definition 3: A stem-and-leaf display, or stem-and-leaf plot, is a visual representation of a quantitative variable that uses classes, but that also shows all individual data values.

Example 2.5: Below is a stem-and-leaf plot for the number of points scored by the Super Bowl winner for each of the first 48 Super Bowls. (Note: 5|5 means 55 points scored.)

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

What is the class width for the given stem-and-leaf plot?

Note: Notice that in the description of the stem-and-leaf plot, a key is given for the scale. This is a necessity for these types of displays.

Example 2.6: Below is the list of number of points scored by the Super Bowl loser for each of the first 49 Super Bowls. Build a stem-and-leaf plot with each of the following bin sizes: 10, 5.

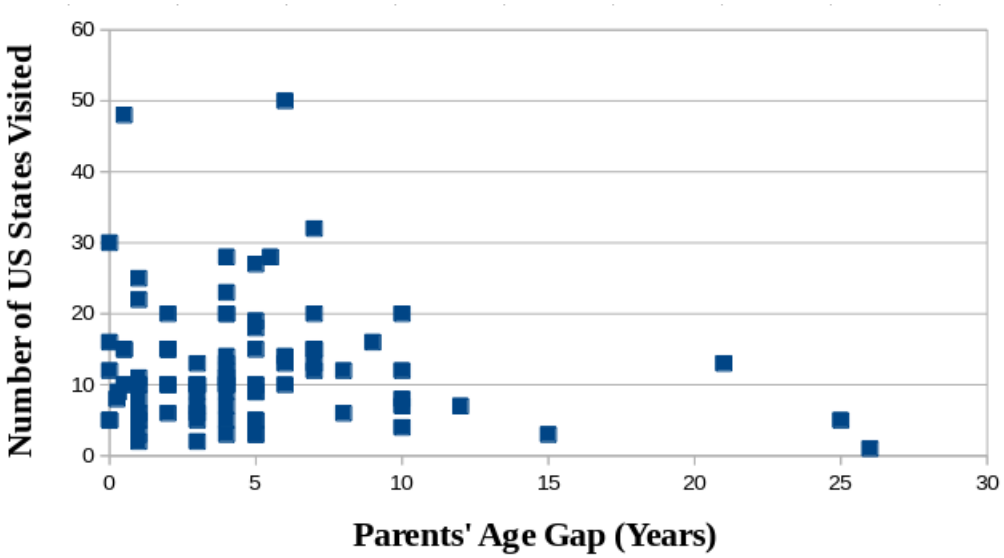
10, 14, 7, 7, 13, 3, 7, 7, 6, 17, 14, 10, 31, 19, 10, 21, 17, 9, 16, 10, 20, 10, 16, 10, 19, 24, 17, 13, 26, 17, 21, 24, 19, 16, 7, 17, 21, 29, 21, 10, 17, 14, 23, 17, 25, 17, 31, 8, 24

Definition 4: A dotplot is like a histogram, except that each case is represented by a dot (stacked on top of each other) rather than a solid bar.

Example 2.7: Construct a dotplot for the data from the previous example.

Definition 5: A scatter diagram, or scatterplot, is a visual representation of data comparing two quantitative variables. The two variables must make up a paired data set, meaning that each value of one variable corresponds to a value of the other, and both of these values "come from" the same individual subject. If you call one of the variables x and the other y, you plot each individual in the sample as a point (x, y) corresponding to its values for the two variables of interest.

Example 2.8: Below is a scatterplot displaying data from our first day of class survey. It shows the relationship between parents' age gap and number of states visited. Does there appear to be a pattern in the data? As the gap between parents' ages increases, what tends to happen to the number of states visited?

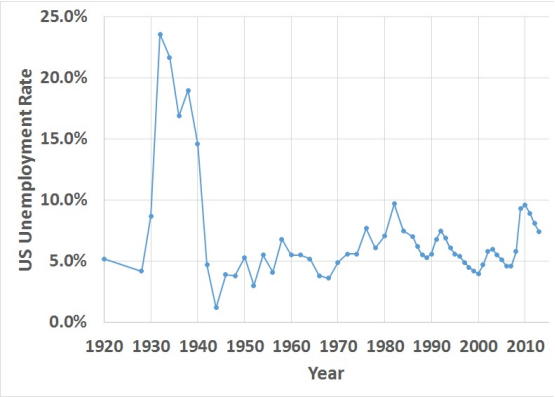


Example 2.9: Below is a table displaying data from the first day of class survey (for 2 different classes, including ours). Create a scatterplot using this data. Does there appear to be a pattern?

Height (in)	Hours sleep per night	Height (in)	Hours sleep per night
75	7	66	7
75	6	66	7
67	8	66	6
64	7	71	8
62	7	74	7.5
74	9	69	7
71	6	64	6
67	8	73	6
73	8	73	6
66	8	63	11
68	7	60	7
73	7	68	7
67	6	63	9
62.5	6	63	8
68	7	72	7
70	6	63	8

Definition 6: When we have data measured over time, we can look for patterns by plotting the data in time order. Displaying values against time is called a time-series chart.

Example 2.10:



1. Estimate the unemployment rate in 1970.

2. Estimate the unemployment rate in the year you were born.

3. In what year was the unemployment rate its lowest since 1920?

4. In what year was the unemployment rate its highest since 1920?

Definition 7: A bar chart displays the distribution of a qualitative variable. You may label the x-axis with the variable values and the y-axis with the counts (frequencies) and draw vertical bars representing the distribution, OR you may label the y-axis with the variable values and the x-axis with the counts and draw horizontal bars representing the distribution. A relative frequency bar chart is constructed in the same way as a bar chart, except that it displays the proportions of the values (relative frequencies) instead of the counts. A Pareto chart is a bar chart where the bars are drawn in decreasing order of height.

Example 2.11: Below are some frequency distributions for the data collected on the first day of class for some stats classes (including this one). Create a bar chart or Pareto chart for each.

Gender	Count
Male	52
Female	68
Total	120

Area Code	Count
315	2
319	1
330	1
412	4
440	2
571	1
585	4
610	1
714	1
716	2
724	4
814	11
Total	34

Eye Color	Count
Black	4
Blue	28
Brown	55
Green	14
Hazel	17
Total	118

Definition 8: A pie chart displays the distribution of values of a qualitative variable, as a proportion of the total number of cases. In order to use a pie chart, the sum of the distributions must be 100% of the total cases. This means there cannot be any overlap among the variable values, and each case must be identified with exactly one value.

Example 2.12: For which of the variables from our class survey would a bar chart be appropriate? For which would a pie chart be appropriate?

Variable	Bar Chart?	Pie Chart?
Gender		
Handedness		
Favorite Genre of Music		
Area code		
Number of US States visited		
Birthplace		
Favorite sport		
Number of Semesters at Mercyhurst		
Major		

2.3 Measures of Central Tendency

Once we have a visual representation of quantitative data, we'd like to summarize what we see to be able to draw some conclusions / gain some insights about it. There are three main attributes of a distribution that we will report: shape, center, and variation.

Definition 9: The center of a distribution describes the "typical" value. A statistic that is meant to describe the center of a distribution is sometimes called a measure of central tendency.

Definition 10: The mean of a distribution is given by

$$\overline{y} = \frac{Total}{n} = \frac{\sum y}{n}$$

where y denotes the data values, n is the number of values, and $\overline{y}$ is notation for "the mean of the y values." (This is the same as you've always calculated averages before.) The mean of a population is denoted by μ, the Greek lowercase letter mu.

Definition 11: The median of a distribution is the "middle value." It is the value for which exactly half of the data lies above it and exactly half of the data lies below it.

1. Assume your distribution has n data values (we always will assume this in this class).
2. Rewrite your values in ascending (or descending) order.
3. If n is odd, then the median of the distribution is the $\frac{n+1}{2}^{\text{th}}$ value. (Example: If you have 7 values, then the median is the $\frac{7+1}{2} = 4^{\text{th}}$ value. This is the one in the middle!)
4. If n is even, then the median of the distribution is the average of the values in the $\frac{n}{2}$ and $\frac{n}{2}+1$ positions. (Example: If you have 8 values, then the median is the average of the 4th and 5th values.)

Definition 12: The mode of a data set is the value that occurs with the greatest frequency. If no entry is repeated (the frequency of every value is 1), the data set has no mode. It is possible for a data set to have more than one mode.

Example 2.13: Find the mean, median, and mode of each of the following data sets:

1. The values: 8, 17, -3, 8, 1, 23, 4, 2, -5

2. The number of points scored by Super Bowl winners (through 2014 season).

3. The number of points scored by each of the last 14 Super Bowl losers:
   17, 21, 29, 21, 10, 17, 14, 23, 17, 25, 17, 31, 8, 24

To compute the mean of data in a frequency distribution, we use the midpoint of each class as the value, and we must count each class according to its frequency:

$$\frac{\sum(m \cdot f_m)}{n}$$

Note that in a frequency distribution, the number of observation is the sum of the frequencies:

$$n = \sum f$$

Example 2.14: Find the median and approximate the mean of the birth weight for babies born in the United States in 2004 using the frequency distribution given below.

Weight (grams)	Number of babies born (in thousands)
0-999	30
1000-1999	97
2000-2999	935
3000-3999	2698
4000-4999	344
5000-5999	5

Example 2.15: Find the median and approximate the mean of the data set below, which displays the frequency distribution for the running time for movies released in the United States in 2014.

Running time (minutes)	Number of movies
30-39	1
40-49	0
50-59	0
60-69	1
70-79	9
80-89	45
90-99	98
100-109	87
110-119	60
120-129	29
130-139	19
140-149	6
150-159	4
160-169	4
170-179	0
180-189	1
190-199	0
200-209	0
210-219	0
220-229	0
230-239	0
240-249	0
250-259	1

Describing Shapes

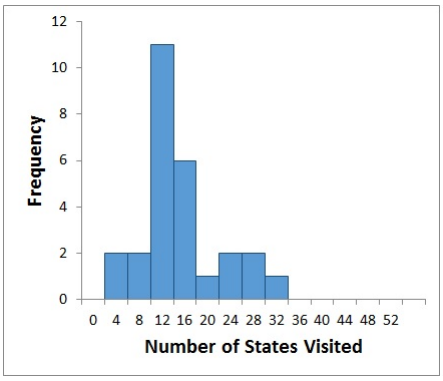
Does the distribution have any hills/mounds? These correspond to modes. Distributions with one mode are called unimodal. Distributions with two modes are called bimodal. Distributions with more than two modes are called multimodal. Distributions with no modes (roughly the same height for all values) are called uniform.

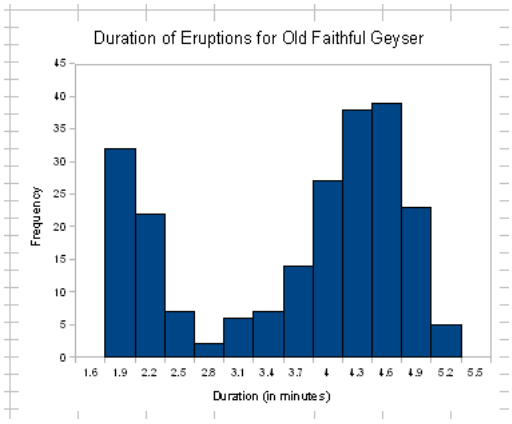
Is the distribution symmetric? A distribution is symmetric if the middle of the distribution acts like a mirror; it looks the same on the left of the middle as it does on the right.

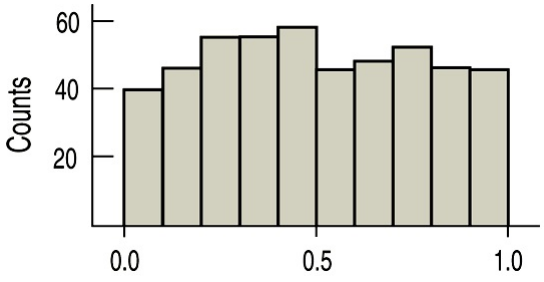
Does the distribution get thinner on one end (or the other, or both)? The thin end(s) of a distribution are called the tail(s). If the left tail is longer, we say the distribution is skewed to the left. If the right tail is longer, we say the distribution is skewed to the right.

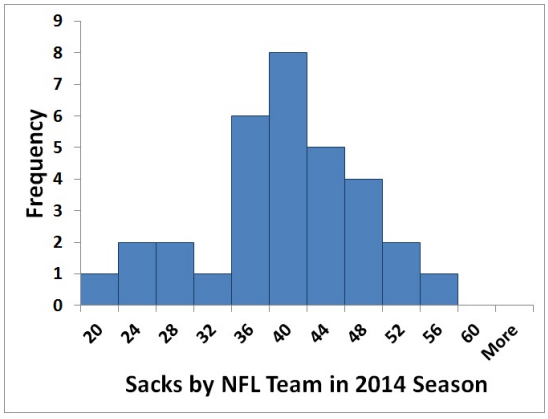
Are there any unusual values? Values that are especially unusual (when compared with the rest of the values in the distribution) are called outliers. An outlier may be a special case with a clear explanation, an error or typo, or a true mystery. For right now, it is enough to simply point out when a value looks like it doesn't belong. As we progress through the course we will learn different ways of handling outliers, and specific ways of identifying them (besides just looking and guessing).

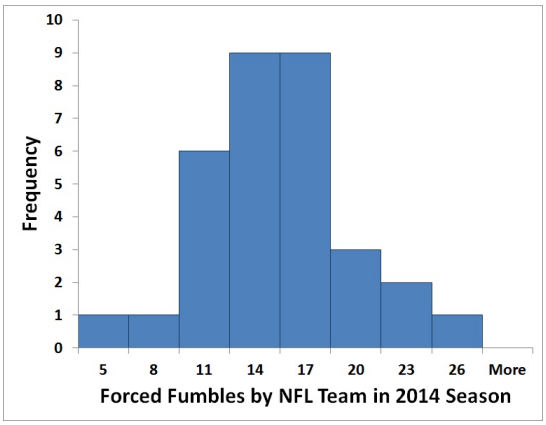
Example 2.16: Describe the shape of each distribution that follows.

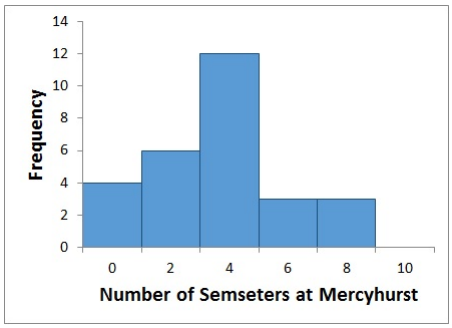












2.4 Measures of Variation

Definition 13: The range of a quantitative variable is the difference between the maximum value and the minimum value.

Definition 14: The standard deviation describes how far the values in a data set are from the mean of that data set, on average. The population standard deviation σ compares the entire population (of size N) with the population mean μ. Sample standard deviation s compares the sample (of size n) to the sample mean $\overline{x}$. The variance is the standard deviation squared (either σ^2 or s^2, depending on if you are analyzing a population or a sample).

Below are the formulas needed to compute population/sample variance/standard deviation.

	Variance	Standard Deviation
Population	$\sigma^2 = \frac{\sum\left[(x-\mu)^2\right]}{N}$	$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum\left[(x-\mu)^2\right]}{N}}$
Sample	$s^2 = \frac{\sum\left[(x-\overline{x})^2\right]}{n-1}$	$s = \sqrt{s^2} = \sqrt{\frac{\sum\left[(x-\overline{x})^2\right]}{n-1}}$

Example 2.17: Find the range and standard deviation of each of the following data sets:

1. The values: 1, 23, 4, 2, -5

2. The number of points scored by Super Bowl winners (through 2014 season).

3. The number of points scored by each of the last 14 Super Bowl losers:
   17, 21, 29, 21, 10, 17, 14, 23, 17, 25, 17, 31, 8, 24

To compute the standard deviation of data in a frequency distribution, we use the midpoint of each class as the value, and we must count each class according to its frequency:

$$\frac{\sum \left[(m-\overline{x})^2 \cdot f_m\right]}{n}$$

Recall that in a frequency distribution, the number of observation is the sum of the frequencies:

$$n = \sum f$$

Example 2.18: Approximate the standard deviation of the birth weight for babies born in the United States in 2004 using the frequency distribution given below.

Weight (grams)	Number of babies born (in thousands)
0-999	30
1000-1999	97
2000-2999	935
3000-3999	2698
4000-4999	344
5000-5999	5

Example 2.19: Approximate the standard deviation of the data set below, which displays the frequency distribution for the running time for movies released in the United States in 2014.

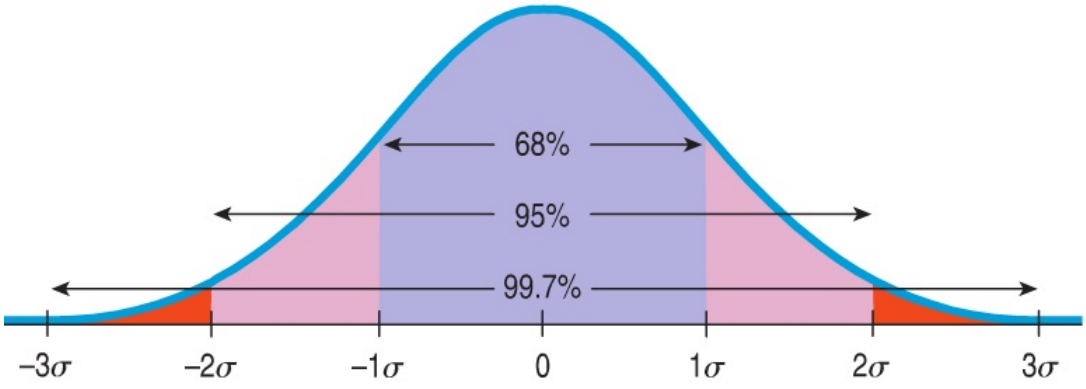
Running time (minutes)	Number of movies
30-39	1
40-49	0
50-59	0
60-69	1
70-79	9
80-89	45
90-99	98
100-109	87
110-119	60
120-129	29
130-139	19
140-149	6
150-159	4
160-169	4
170-179	0
180-189	1
190-199	0
200-209	0
210-219	0
220-229	0
230-239	0
240-249	0
250-259	1

Interpreting Standard Deviation

If the standard deviation of a data set is small, then the values of the variable are close together. If the standard deviation of a data set is large, then the values of the variable spread farther apart.

The Empirical Rule. In a *bell-shaped distribution*, the standard deviation provides even more information about the spread of the data.

- Approximately 68% of the values are between $\mu-\sigma$ and $\mu+\sigma$ (ie, within one standard deviation from the mean).
- Approximately 95% of the values are between $\mu - 2\sigma$ and $\mu + 2\sigma$ (ie, within two standard deviations from the mean).
- Approximately 99.7% of the values are between $\mu - 3\sigma$ and $\mu + 3\sigma$ (ie, within three standard deviations from the mean).
- Only about 0.3% of the values will be more than three standard deviations from the mean, making these values very unusual/different from the rest of the data.



Example 2.20: Below is a table that shows the length of eruption in seconds for a random sample of eruptions of the Old Faithful geyser.

108	102	103	110	104	113	100
108	99	109	102	100	116	101
99	106	109	105	103	95	107
105	90	111	110	102	105	110
103	104	101	106	120	103	92
103	110	101	104	90	101	108
94	110					

1. Find the mean and standard deviation.

2. Draw a histogram and determine if the Empirical Rule is appropriate.

3. Use the Empirical Rule to estimate the percentage of eruptions that last between 92 and 116 seconds. Then determine the actual percentage of the sample between 92 and 116 seconds.

4. Use the Empirical Rule to estimate the percentage of eruptions that last less than 98 seconds. Then determine the actual percentage of the sample less than 98 seconds.

2.5 Measures of Position

Definition 15: The k^{th} percentile, denoted P_k, of a data set for a quantitative variable is a value such that k percent of the observations are less than or equal to the value.

Definition 16: Fractiles of a data set for a quantitative variable are values that partition the data into equal parts (fractions). Ex: Quartiles break the data into equal fourths, quintiles break the data into equal fifths, etc.

Definition 17: The 25^{th} percentile is called the first quartile or lower quartile and is denoted by Q1. The 75^{th} percentile is called the third quartile or upper quartile and is denoted by Q3. The 0^{th} percentile is the minimum value of the data set. The 100^{th} percentile is the maximum value of the data set. The 50^{th} percentile is the median.

Definition 18: The Five Number Summary of a data set is comprised of the minimum value, Q1, Median, Q3, and maximum value.

Definition 19: The interquartile range (denoted IQR) is the measure of spread that is "best" when using the median as our measure of center. It is the difference between the first and third quartiles. (Computed: Q3 - Q1) Based on the definition of the quartiles, the IQR gives us a range that covers 50% of the data (the "middle 50%" of the data).

Example 2.21: For each of the following data sets, find the Five Number Summary and the IQR.

1. The values: 8, 17, -3, 8, 1, 23, 4, 2, -5

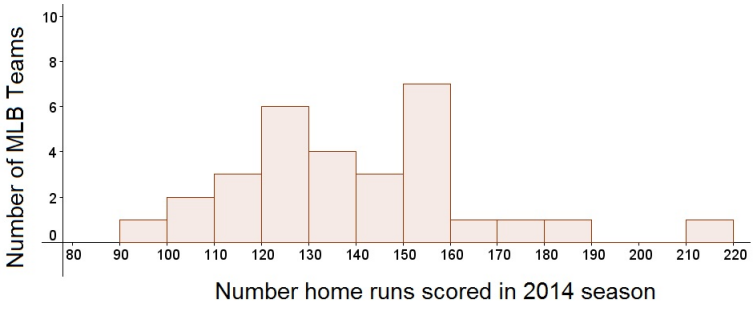
2. The number of points scored by Super Bowl winners (through 2014 season).

3. The number of points scored by each of the last 14 Super Bowl losers:
   17, 21, 29, 21, 10, 17, 14, 23, 17, 25, 17, 31, 8, 24

Boxplots

Definition 20: A boxplot is a visual representation of the quartiles of a quantitative variable.

Example 2.22: Below is a histogram and Five Number Summary for the number of home runs hit during the 2011 Regular Season by each Major League Baseball (MLB) team. We will use this information to make a boxplot for this distribution.



Min	95
Q1	123
Median	135
Q3	155
Max	211

Step 1. Draw a horizontal axis that spans from the minimum to the maximum values. Make vertical markers for the median, Q1, and Q3, then connect these with horizontal line segments to make a box (the right side of the box is Q3, the left side of the box is Q1, and there is a line somewhere in the middle of the box where the median is).

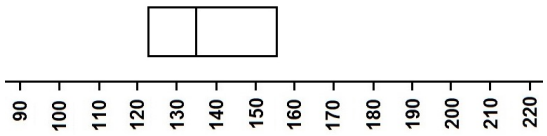


Figure 1: Step 1

Step 2. Compute the "fences" of your boxplot. These fences **should not be included** in your final diagram, but it can be helpful to pencil them in and erase them later. To find the fences, first compute the IQR (155 - 123 = 32, in this example). Then multiply the IQR by 1.5 ($32 \times 1.5 = 48$, in this example). The upper fence is Q3 plus this number (Q3 + 1.5 · IQR) and the lower fence is Q1 minus this number (Q1 - 1.5 · IQR). In this example, our fences are 203 and 75.

Step 3. Draw two more horizontal markers in the following way: If your maximum value is less than your upper fence, draw your first marker at the max value. If the maximum value is greater than your upper fence, then draw the marker at the upper fence. If your minimum value is greater than your lower fence, draw your second marker at the minimum value. If the minimum value is less than the lower fence, then draw the marker at the lower fence. Connect these two markers to the main box with single vertical lines. (Your book calls these ends the "whiskers" of the boxplot.)

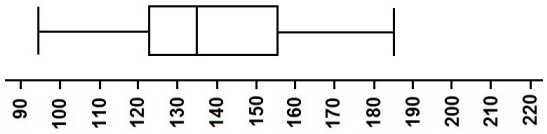


Figure 2: Step 3

Step 4. If any of your data values fall outside of the fences (and therefore were not reached by the whiskers), draw them in with a dot or star or some sort of symbol. These are the values that may be outliers.

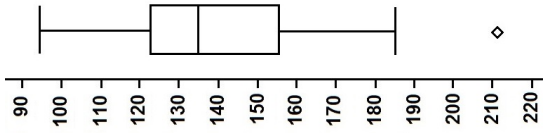


Figure 3: Step 4

Some notes about boxplots:

- The fences should never be drawn in on your final boxplot.
- The whiskers should never reach farther than the max/min values. They should also never reach farther than the fences.
- The boxplot summarizes a histogram in a smaller amount of space, but it packs a lot of information. It shows skewness and outliers. The "box" part of the boxplot represents the middle 50% of the data, and gives a visual representation of the IQR.

Example 2.23: Below is a data table showing the total number of penalty yards for each of the 32 NFL teams during the 2014 regular season.
In Class: We will find the Five Number Summary.
On your own: Use your calculator to find a histogram for this data.
In Class: We will draw the boxplot for this data.

Team	Penalty Yards	Team	Penalty Yards
Arizona Cardinals	1,899	Miami Dolphins	1,555
Atlanta Falcons	1,584	Minnesota Vikings	1,774
Baltimore Ravens	1,877	New England Patriots	1,832
Buffalo Bills	1,880	New Orleans Saints	1,559
Carolina Panthers	1,516	New York Giants	1,844
Chicago Bears	1,975	New York Jets	1,712
Cincinnati Bengals	1,688	Oakland Raiders	1,880
Cleveland Browns	1,701	Philadelphia Eagles	2,084
Dallas Cowboys	1,700	Pittsburgh Steelers	1,669
Denver Broncos	1,861	San Diego Chargers	2,034
Detroit Lions	1,838	San Francisco 49ers	1,846
Green Bay Packers	1,720	Seattle Seahawks	1,622
Houston Texans	1,620	St. Louis Rams	2,021
Indianapolis Colts	1,714	Tampa Bay Buccaneers	1,874
Jacksonville Jaguars	1,459	Tennessee Titans	1,692
Kansas City Chiefs	1,532	Washington Redskins	2,294

1459	1516	1532	1555	1559
1584	1620	1622	1669	1688
1692	1700	1701	1712	1714
1720	1774	1832	1838	1844
1846	1861	1874	1877	1880
1880	1899	1975	2021	2034
2084	2294			

Min	
Q1	
Median	
Q3	
Max	

z-scores

Sometimes we need to compare values on different scales, or even with different units. What if we want to compare runners' combined performances on races of different lengths? Below are three runners' results for 100-meter and 200-meter races. Which of the three is the "best" runner?

Runner	**100 Meter Time (seconds)**	**200 Meter Time (seconds)**
Pamela	13.2	25.7
Danielle	12.9	26.9
Samantha	12.3	27.1

We will use the mean and standard deviation of times for all runners for each race to make the values comparable. If x_{50} denotes the variable "50-meter race times" and x_{100} denotes the variable "100-meter race times," then $\overline{x_{50}}$ and $\overline{x_{100}}$ are the mean race times and s_{50} and s_{100} are their standard deviations.

Definition 21: To standardize the values, we will compute their z-scores:

$$z = \frac{x - \overline{x}}{s}$$

Example 2.24: The average time of all runners in the 100-meter race was 12.7 seconds and the mean time for the 200-meter was 26.8 seconds. The standard deviations were 0.4 seconds and 0.9 seconds, respectively. Compare the combined z-scores of the runners to determine who is the best runner.

Runner	**100 Meter z-score**	**200 Meter z-score**	**Combined z-score**
Pamela			
Danielle			
Samantha			

Note: z-scores do not have units. The mean of a distribution of z-scores is always 0 and the standard deviation is always 1. A z-score tells us how unusual a value is compared to the rest of the data. A positive z-score corresponds to a value ABOVE the mean; a negative z-score corresponds to a value BELOW the mean. The z-score value itself tells us how many standard deviations from the mean it is. (For example: a z-score of 2 is two-times-the-standard-deviation above the mean; a z-score of -1.5 is one-and-a-half-times-the-standard-deviation below the mean.)