

Chapter 3: Numerical Descriptions of Data

Chapter 1 discussed what a population, sample, parameter, and statistic are, and how to take different types of samples. Chapter 2 discussed ways to graphically display data. There was also a discussion of important characteristics: center, variations, distribution, outliers, and changing characteristics of the data over time. Distributions and outliers can be answered using graphical means. Finding the center and variation can be done using numerical methods that will be discussed in this chapter. Both graphical and numerical methods are part of a branch of statistics known as **descriptive statistics**. Later descriptive statistics will be used to make decisions and/or estimate population parameters using methods that are part of the branch called **inferential statistics**.

Section 3.1: Measures of Center

This section focuses on measures of central tendency. Many times you are asking what to expect on average. Such as when you pick a major, you would probably ask how much you expect to earn in that field. If you are thinking of relocating to a new town, you might ask how much you can expect to pay for housing. If you are planting vegetables in the spring, you might want to know how long it will be until you can harvest. These questions, and many more, can be answered by knowing the center of the data set. There are three measures of the “center” of the data. They are the mode, median, and mean. Any of the values can be referred to as the “average.”

The **mode** is the data value that occurs the most frequently in the data. To find it, you count how often each data value occurs, and then determine which data value occurs most often.

The **median** is the data value in the middle of a sorted list of data. To find it, you put the data in order, and then determine which data value is in the middle of the data set.

The **mean** is the arithmetic average of the numbers. This is the center that most people call the average, though all three – mean, median, and mode – really are averages.

There are no symbols for the mode and the median, but the mean is used a great deal, and statisticians gave it a symbol. There are actually two symbols, one for the population parameter and one for the sample statistic. In most cases you cannot find the population parameter, so you use the sample statistic to estimate the population parameter.

Population Mean:

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

N is the size of the population.

x represents a data value.

$\sum x$ means to add up all of the data values.

Sample Mean:

$$\bar{x} = \frac{\sum x}{n}, \text{ pronounced } x \text{ bar.}$$

n is the size of the sample.

x represents a data value.

$\sum x$ means to add up all of the data values.

The value for $\bar{x}$ is used to estimate μ since μ can't be calculated in most situations.

Example #3.1.1: Finding the Mean, Median, and Mode

Suppose a vet wants to find the average weight of cats. The weights (in pounds) of five cats are in table #3.1.1.

Table #3.1.1: Weights of cats in pounds

6.8	8.2	7.5	9.4	8.2
-----	-----	-----	-----	-----

Find the mean, median, and mode of the weight of a cat.

Solution:

Before starting any mathematics problem, it is always a good idea to define the unknown in the problem. In this case, you want to define the variable. The symbol for the variable is x .

The variable is x = weight of a cat

Mean:

$$\bar{x} = \frac{6.8 + 8.2 + 7.5 + 9.4 + 8.2}{5} = \frac{40.1}{5} = 8.02 \text{ pounds}$$

Median:

You need to sort the list for both the median and mode. The sorted list is in table #3.1.2.

Table #3.1.2: Sorted List of Cats' Weights

6.8	7.5	8.2	8.2	9.4
-----	-----	-----	-----	-----

There are 5 data points so the middle of the list would be the 3rd number. (Just put a finger at each end of the list and move them toward the center one number at a time. Where your fingers meet is the median.)

Table #3.1.3: Sorted List of Cats' Weights with Median Marked

6.8	7.5	8.2	8.2	9.4
-----	-----	-----	-----	-----

The median is therefore 8.2 pounds.

Mode:

This is easiest to do from the sorted list that is in table #3.1.2. Which value appears the most number of times? The number 8.2 appears twice, while all other numbers appear once.

Mode = 8.2 pounds.

A data set can have more than one mode. If there is a tie between two values for the most number of times then both values are the mode and the data is called bimodal (two modes). If every data point occurs the same number of times, there is no mode. If there are more than two numbers that appear the most times, then usually there is no mode.

In example #3.1.1, there were an odd number of data points. In that case, the median was just the middle number. What happens if there is an even number of data points? What would you do?

Example #3.1.2: Finding the Median with an Even Number of Data Points

Suppose a vet wants to find the median weight of cats. The weights (in pounds) of six cats are in table #3.1.4. Find the median

Table #3.1.4: Weights of Six Cats

6.8	8.2	7.5	9.4	8.2	6.3
-----	-----	-----	-----	-----	-----

Solution:

Variable: x = weight of a cat

First sort the list if it is not already sorted.

There are 6 numbers in the list so the number in the middle is between the 3rd and 4th number. Use your fingers starting at each end of the list in table #3.1.5 and move toward the center until they meet. There are two numbers there.

Table #3.1.5: Sorted List of Weights of Six Cats

6.3	6.8	7.5	8.2	8.2	9.4
-----	-----	-----	-----	-----	-----

To find the median, just average the two numbers.

$$\text{median} = \frac{7.5 + 8.2}{2} = 7.85 \text{ pounds}$$

The median is 7.85 pounds.

Example #3.1.3: Finding Mean and Median using Technology

Suppose a vet wants to find the median weight of cats. The weights (in pounds) of six cats are in table #3.1.4. Find the median

Solution:

Variable: x = weight of a cat

You can do the calculations for the mean and median using the technology.

The procedure for calculating the sample mean ($\bar{x}$) and the sample median (Med) on the TI-83/84 is in figures 3.1.1 through 3.1.4. First you need to go into the STAT menu, and then Edit. This will allow you to type in your data (see figure #3.1.1).

Figure #3.1.1: TI-83/84 Calculator Edit Setup

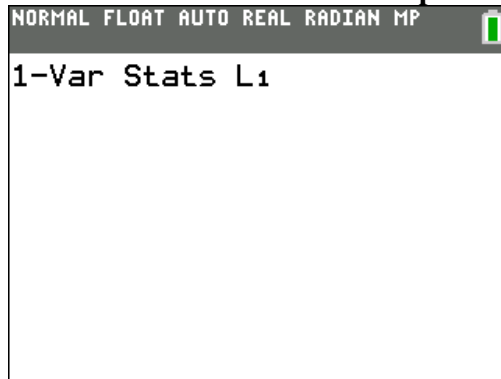
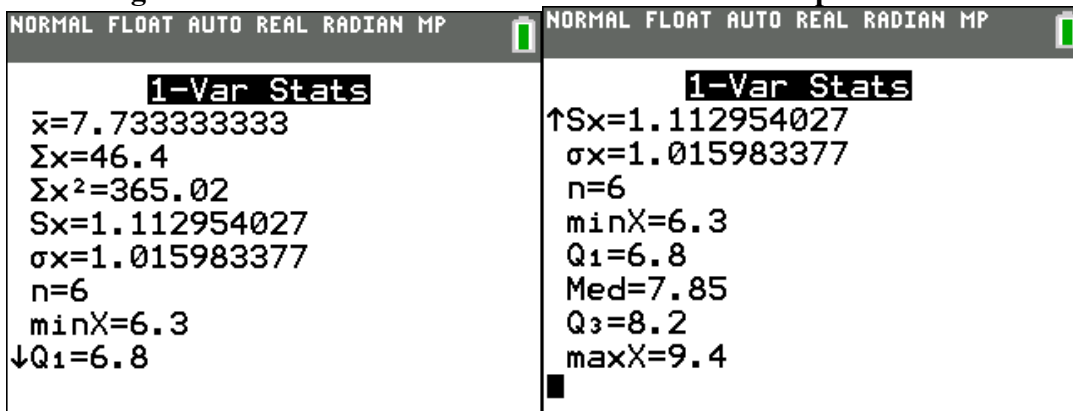
NORMAL FLOAT AUTO REAL RADIAN MP					
L1	L2	L3	L4	L5	1
6.8	-----	-----	-----	-----	
8.2					
7.5					
9.4					
8.2					
6.3					

L1(?)=					

Once you have the data into the calculator, you then go back to the STAT menu, move over to CALC, and then choose 1-Var Stats (see figure #3.1.2). The calculator will now put 1-Var Stats on the main screen. Now type in L1 (2nd button and 1) and then press ENTER. (Note if you have the newer operating system on the TI-84, then the procedure is slightly different.) If you press the down arrow, you will see the rest of the output from the calculator. The results from the calculator are in figure #3.1.3.

Figure #3.1.2: TI-83/84 Calculator CALC Menu

EDIT	TESTS
1: 1-Var Stats	
2: 2-Var Stats	
3: Med-Med	
4: LinReg(ax+b)	
5: QuadReg	
6: CubicReg	
7: QuartReg	

Figure #3.1.3: TI-83/84 Calculator Input for Example #3.1.3 Variable**Figure #3.1.4: TI-83/84 Calculator Results for Example #3.1.3 Variable**

The commands for finding the mean and median using R are as follows:

```
variable<-c(type in your data with commas in between)
```

```
To find the mean, use mean(variable)
```

```
To find the median, use median(variable)
```

So for this example, the commands would be

```
weights<-c(6.8, 8.2, 7.5, 9.4, 8.2, 6.3)
```

```
mean(weights)
```

```
[1] 7.733333
```

```
median(weights)
```

```
[1] 7.85
```

Example #3.1.4: Affect of Extreme Values on Mean and Median

Suppose you have the same set of cats from example 3.1.1 but one additional cat was added to the data set. Table #3.1.6 contains the six cats' weights, in pounds.

Table #3.1.6: Weights of Six Cats

6.8	7.5	8.2	8.2	9.4	22.1
-----	-----	-----	-----	-----	------

Find the mean and the median.

Solution:

Variable: x = weight of a cat

$$\text{mean} = \bar{x} = \frac{6.8 + 7.5 + 8.2 + 8.2 + 9.4 + 22.1}{6} = 10.37 \text{ pounds}$$

The data is already in order, thus the median is between 8.2 and 8.2.

$$\text{median} = \frac{8.2 + 8.2}{2} = 8.2 \text{ pounds}$$

The mean is much higher than the median. Why is this? Notice that when the value of 22.1 was added, the mean went from 8.02 to 10.37, but the median did not change at all. This is because the mean is affected by extreme values, while the median is not. The very heavy cat brought the mean weight up. In this case, the median is a much better measure of the center.

An outlier is a data value that is very different from the rest of the data. It can be really high or really low. Extreme values may be an outlier if the extreme value is far enough from the center. In example #3.1.4, the data value 22.1 pounds is an extreme value and it may be an outlier.

If there are extreme values in the data, the median is a better measure of the center than the mean. If there are no extreme values, the mean and the median will be similar so most people use the mean.

The mean is not a resistant measure because it is affected by extreme values. The median and the mode are resistant measures because they are not affected by extreme values.

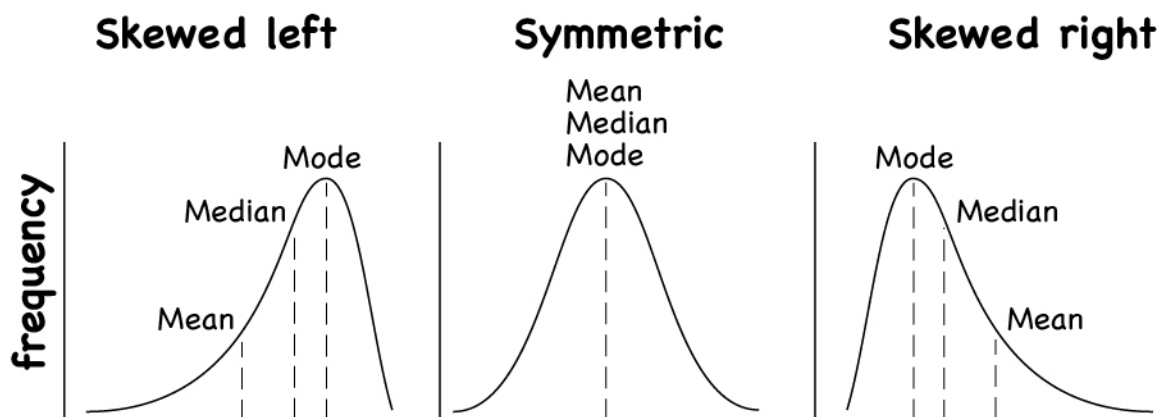
As a consumer you need to be aware that people choose the measure of center that best supports their claim. When you read an article in the newspaper and it talks about the “average” it usually means the mean but sometimes it refers to the median. Some articles will use the word “median” instead of “average” to be more specific. If you need to make an important decision and the information says “average”, it would be wise to ask if the “average” is the mean or the median before you decide.

As an example, suppose that a company wants to use the mean salary as the average salary for the company. This is because the high salaries of the administration will pull the mean higher. The company can say that the employees are paid well because the average is high. However, the employees want to use the median since it discounts the extreme values of the administration and will give a lower value of the average. This will make the salaries seem lower and that a raise is in order.

Why use the mean instead of the median? The reason is because when multiple samples are taken from the same population, the sample means tend to be more consistent than other measures of the center. The sample mean is the more reliable measure of center.

To understand how the different measures of center related to skewed or symmetric distributions, see figure #3.1.5. As you can see sometimes the mean is smaller than the median and mode, sometimes the mean is larger than the median and mode, and sometimes they are the same values.

Figure #3.1.5: Mean, Median, Mode as Related to a Distribution



One last type of average is a weighted average. Weighted averages are used quite often in real life. Some teachers use them in calculating your grade in the course, or your grade on a project. Some employers use them in employee evaluations. The idea is that some activities are more important than others. As an example, a fulltime teacher at a community college may be evaluated on their service to the college, their service to the community, whether their paperwork is turned in on time, and their teaching. However, teaching is much more important than whether their paperwork is turned in on time. When the evaluation is completed, more weight needs to be given to the teaching and less to the paperwork. This is a weighted average.

Weighted Average

$$\frac{\sum xw}{\sum w} \text{ where } w \text{ is the weight of the data value, } x.$$

Example #3.1.5: Weighted Average

In your biology class, your final grade is based on several things: a lab score, scores on two major tests, and your score on the final exam. There are 100 points available for each score. The lab score is worth 15% of the course, the two exams are worth 25% of the course each, and the final exam is worth 35% of the course. Suppose you earned scores of 95 on the labs, 83 and 76 on the two exams, and 84 on the final exam. Compute your weighted average for the course.

Solution:

Variable: x = score

The weighted average is $\frac{\sum xw}{\sum w} = \frac{\text{sum of the scores times their weights}}{\text{sum of all the weights}}$

$$\text{weighted average} = \frac{95(0.15) + 83(0.25) + 76(0.25) + 84(0.35)}{0.15 + 0.25 + 0.25 + 0.35} = \frac{83.4}{1.00} = 83.4\%$$

A weighted average can be found using technology.

The procedure for calculating the weighted average on the TI-83/84 is in figures 3.1.6 through 3.1.9. First you need to go into the STAT menu, and then Edit. This will allow you to type in the scores into L1 and the weights into L2 (see figure #3.1.6).

Figure #3.1.6: TI-83/84 Calculator Edit Setup

[illegible]

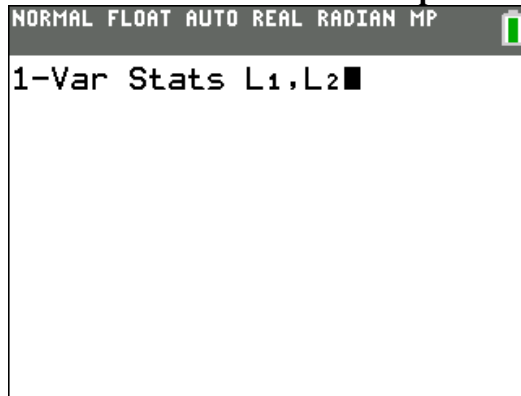
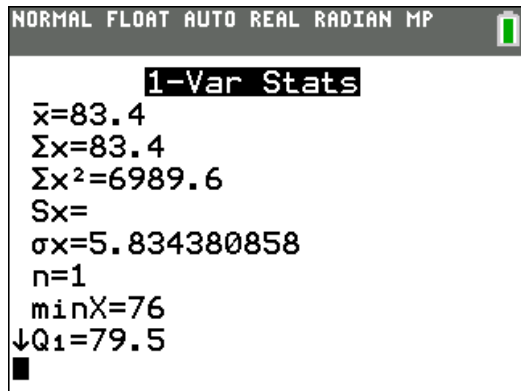
Once you have the data into the calculator, you then go back to the STAT menu, move over to CALC, and then choose 1-Var Stats (see figure #3.1.7). The calculator will now put 1-Var Stats on the main screen. Now type in L1 (2nd button and 1), then a comma (button above the 7 button), and then L2 (2nd button and 2) and then press ENTER. (Note if you have the newer operating system on the TI-84, then the procedure is slightly different.) The results from the calculator are in figure #3.1.9. The $\bar{x}$ is the weighted average.

Figure #3.1.7: TI-83/84 Calculator CALC Menu

```

EDIT 0000 TESTS
1:1-Var Stats
2:2-Var Stats
3:Med-Med
4:LinReg(ax+b)
5:QuadReg
6:CubicReg
7:QuartReg

```


Figure #3.1.8: TI-83/84 Calculator Input for Weighted Average**Figure #3.1.9: TI-83/84 Calculator Results for Weighted Average**

The commands for finding the mean and median using R are as follows:

```
x<-c(type in your data with commas in between)
w<-c(type in your weights with commas in between)
weighted.mean(x,w)
```

So for this example, the commands would be

```
x<-c(95, 83, 76, 84)
w<-c(.15, .25, .25, .35)
weighted.mean(x,w)
[1] 83.4
```

Example #3.1.6: Weighted Average

The faculty evaluation process at John Jingle University rates a faculty member on the following activities: teaching, publishing, committee service, community service, and submitting paperwork in a timely manner. The process involves reviewing student evaluations, peer evaluations, and supervisor evaluation for each teacher and awarding him/her a score on a scale from 1 to 10 (with 10 being the best). The weights for each activity are 20 for teaching, 18 for publishing, 6 for committee service, 4 for community service, and 2 for paperwork.

- a) One faculty member had the following ratings: 8 for teaching, 9 for publishing, 2 for committee work, 1 for community service, and 8 for paperwork. Compute the weighted average of the evaluation.

Solution:

Variable: x = rating

The weighted average is $\frac{\sum xw}{\sum w} = \frac{\text{sum of the scores times their weights}}{\text{sum of all the weights}}$.

$$\text{evaluation} = \frac{8(20) + 9(18) + 2(6) + 1(4) + 8(2)}{20 + 18 + 6 + 4 + 2} = \frac{354}{50} = 7.08$$

- b) Another faculty member had ratings of 6 for teaching, 8 for publishing, 9 for committee work, 10 for community service, and 10 for paperwork. Compute the weighted average of the evaluation.

Solution:

$$\text{evaluation} = \frac{6(20) + 8(18) + 9(6) + 10(4) + 10(2)}{20 + 18 + 6 + 4 + 2} = \frac{378}{50} = 7.56$$

- c) Which faculty member had the higher average evaluation?

Solution:

The second faculty member has a higher average evaluation.

You can find a weighted average using technology. On the
The last thing to mention is which average is used on which type of data.

Mode can be found on nominal, ordinal, interval, and ratio data, since the mode is just the data value that occurs most often. You are just counting the data values.
Median can be found on ordinal, interval, and ratio data, since you need to put the data in order. As long as there is order to the data you can find the median.
Mean can be found on interval and ratio data, since you must have numbers to add together.

Section 3.1: Homework

- 1.) Cholesterol levels were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985) and are in table #3.1.7. Find the mean, median, and mode.

Table #3.1.7: Cholesterol Levels

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

- 2.) The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Pacific Ocean are listed in table #3.1.8 (Lee, 1994). Find the mean, median, and mode.

Table #3.1.8: Lengths of Rivers (km) Flowing to Pacific Ocean

River	Length (km)	River	Length (km)
Clarence	209	Clutha	322
Conway	48	Taieri	288
Waiau	169	Shag	72
Hurunui	138	Kakanui	64
Waipara	64	Rangitata	121
Ashley	97	Ophi	80
Waimakariri	161	Pareora	56
Selwyn	95	Waihao	64
Rakaia	145	Waitaki	209
Ashburton	90		

- 3.) The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed in table #3.1.9 (Lee, 1994). Find the mean, median, and mode.

Table #3.1.9: Lengths of Rivers (km) Flowing to Tasman Sea

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

- 4.) Eyeglassmatic manufactures eyeglasses for their retailers. They research to see how many defective lenses they made during the time period of January 1 to March 31. Table #3.1.10 contains the defect and the number of defects. Find the mean, median, and mode.

Table #3.1.10: Number of Defective Lenses

Defect type	Number of defects
Scratch	5865
Right shaped – small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape – big	1105
Lost in lab	976
Spots/bubble – intern	976

- 5.) Print-O-Matic printing company's employees have salaries that are contained in table #3.1.1.

Table #3.1.11: Salaries of Print-O-Matic Printing Company Employees

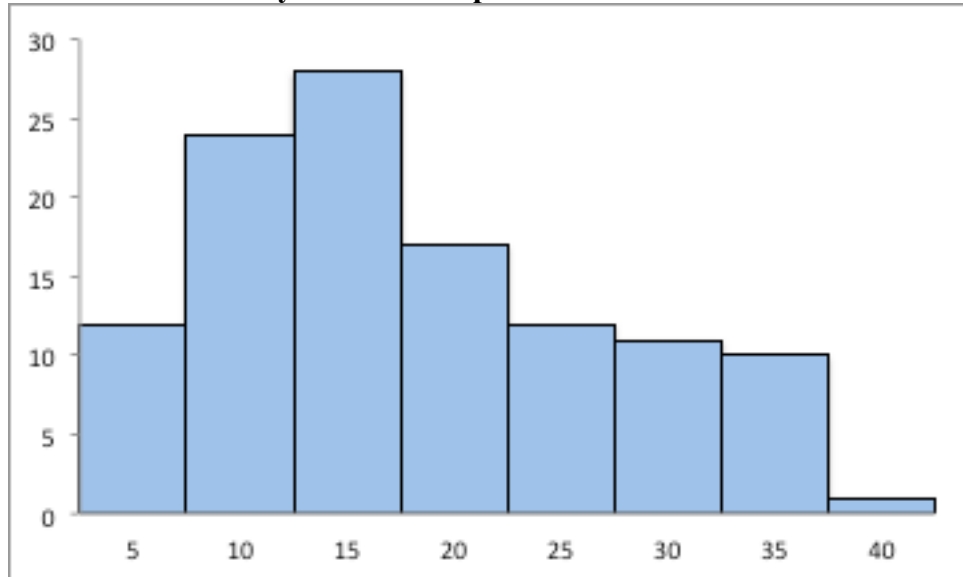
Employee	Salary (\$)
CEO	272,500
Driver	58,456
CD74	100,702
CD65	57,380
Embellisher	73,877
Folder	65,270
GTO	74,235
Handwork	52,718
Horizon	76,029
ITEK	64,553
Mgmt	108,448
Platens	69,573
Polar	75,526
Pre Press Manager	108,448
Pre Press Manager/ IT	98,837
Pre Press/ Graphic Artist	75,311
Designer	90,090
Sales	109,739
Administration	66,346

- a.) Find the mean and median.
b.) Find the mean and median with the CEO's salary removed.

- c.) What happened to the mean and median when the CEO's salary was removed? Why?
- d.) If you were the CEO, who is answering concerns from the union that employees are underpaid, which average of the complete data set would you prefer? Why?
- e.) If you were a platen worker, who believes that the employees need a raise, which average would you prefer? Why?
- 6.) Print-O-Matic printing company spends specific amounts on fixed costs every month. The costs of those fixed costs are in table #3.1.12.
- Table #3.1.12: Fixed Costs for Print-O-Matic Printing Company**
- | Monthly charges | Monthly cost (\$) |
|--------------------|-------------------|
| Bank charges | 482 |
| Cleaning | 2208 |
| Computer expensive | 2471 |
| Lease payments | 2656 |
| Postage | 2117 |
| Uniforms | 2600 |
- a.) Find the mean and median.
- b.) Find the mean and median with the bank charges removed.
- c.) What happened to the mean and median when the bank charges was removed? Why?
- d.) If it is your job to oversee the fixed costs, which average using the complete data set would you prefer to use when submitting a report to administration to show that costs are low? Why?
- e.) If it is your job to find places in the budget to reduce costs, which average using the complete data set would you prefer to use when submitting a report to administration to show that fixed costs need to be reduced? Why?
- 7.) State which type of measurement scale each represents, and then which center measures can be use for the variable?
- a.) You collect data on people's likelihood (very likely, likely, neutral, unlikely, very unlikely) to vote for a candidate.
- b.) You collect data on the diameter at breast height of trees in the Coconino National Forest.
- c.) You collect data on the year wineries were started.
- d.) You collect the drink types that people in Sydney, Australia drink.
- 8.) State which type of measurement scale each represents, and then which center measures can be use for the variable?
- a.) You collect data on the height of plants using a new fertilizer.
- b.) You collect data on the cars that people drive in Campbelltown, Australia.
- c.) You collect data on the temperature at different locations in Antarctica.
- d.) You collect data on the first, second, and third winner in a beer competition.

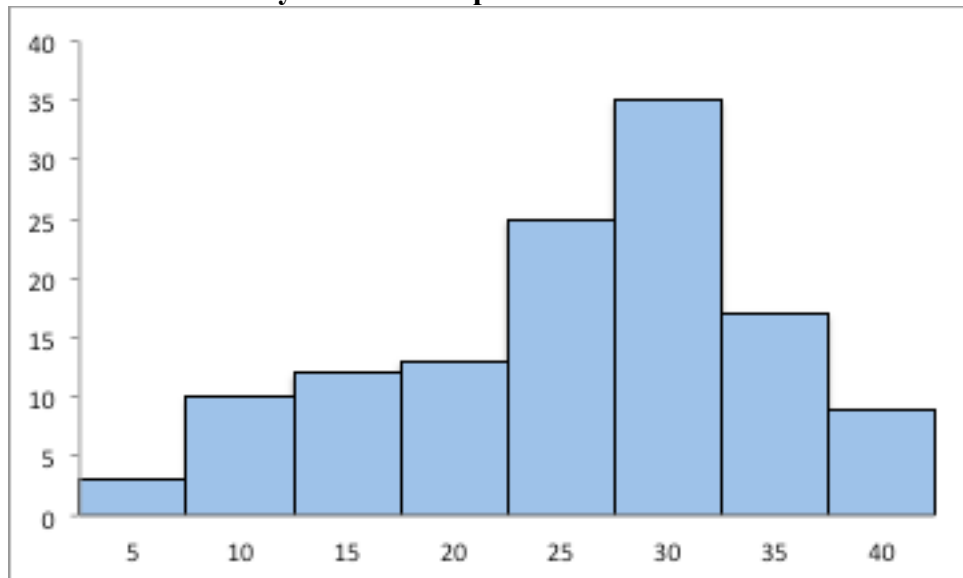
- 9.) Looking at graph #3.1.1, state if the graph is skewed left, skewed right, or symmetric and then state which is larger, the mean or the median?

Graph #3.1.1: Skewed or Symmetric Graph



- 10.) Looking at graph #3.1.2, state if the graph is skewed left, skewed right, or symmetric and then state which is larger, the mean or the median?

Graph #3.1.2: Skewed or Symmetric Graph



- 11.) An employee at Coconino Community College (CCC) is evaluated based on goal setting and accomplishments toward the goals, job effectiveness, competencies, and CCC core values. Suppose for a specific employee, goal 1 has a weight of 30%, goal 2 has a weight of 20%, job effectiveness has a weight of 25%, competency 1 has a goal of 4%, competency 2 has a goal has a weight of 3%, competency 3 has a weight of 3%, competency 4 has a weight of 3%, competency 5 has a weight of 2%, and core values has a weight of 10%. Suppose the employee has scores of 3.0 for goal 1, 3.0 for goal 2, 2.0 for job effectiveness, 3.0 for competency 1, 2.0 for competency 2, 2.0 for competency 3, 3.0 for competency 4, 4.0 for competency 5, and 3.0 for core values. Find the weighted average score for this employee. If an employee has a score less than 2.5, they must have a Performance Enhancement Plan written. Does this employee need a plan?
- 12.) An employee at Coconino Community College (CCC) is evaluated based on goal setting and accomplishments toward goals, job effectiveness, competencies, CCC core values. Suppose for a specific employee, goal 1 has a weight of 20%, goal 2 has a weight of 20%, goal 3 has a weight of 10%, job effectiveness has a weight of 25%, competency 1 has a goal of 4%, competency 2 has a goal has a weight of 3%, competency 3 has a weight of 3%, competency 4 has a weight of 5%, and core values has a weight of 10%. Suppose the employee has scores of 2.0 for goal 1, 2.0 for goal 2, 4.0 for goal 3, 3.0 for job effectiveness, 2.0 for competency 1, 3.0 for competency 2, 2.0 for competency 3, 3.0 for competency 4, and 4.0 for core values. Find the weighted average score for this employee. If an employee that has a score less than 2.5, they must have a Performance Enhancement Plan written. Does this employee need a plan?
- 13.) A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives an 85 on test 1, a 76 on test 2, an 83 on test 3, a 74 on the homework, a 65 on the project, and a 79 on the final, what grade did the student earn in the course?
- 14.) A statistics class has the following activities and weights for determining a grade in the course: test 1 worth 15% of the grade, test 2 worth 15% of the grade, test 3 worth 15% of the grade, homework worth 10% of the grade, semester project worth 20% of the grade, and the final exam worth 25% of the grade. If a student receives a 92 on test 1, an 85 on test 2, a 95 on test 3, a 92 on the homework, a 55 on the project, and an 83 on the final, what grade did the student earn in the course?

Section 3.2: Measures of Spread

Variability is an important idea in statistics. If you were to measure the height of everyone in your classroom, every observation gives you a different value. That means not every student has the same height. Thus there is variability in people's heights. If you were to take a sample of the income level of people in a town, every sample gives you different information. There is variability between samples too. Variability describes how the data are spread out. If the data are very close to each other, then there is low variability. If the data are very spread out, then there is high variability. How do you measure variability? It would be good to have a number that measures it. This section will describe some of the different measures of variability, also known as variation.

In example #3.1.1, the average weight of a cat was calculated to be 8.02 pounds. How much does this tell you about the weight of all cats? Can you tell if most of the weights were close to 8.02 or were the weights really spread out? What are the highest weight and the lowest weight? All you know is that the center of the weights is 8.02 pounds. You need more information.

The **range** of a set of data is the difference between the highest and the lowest data values (or maximum and minimum values).

$$\text{Range} = \text{highest value} - \text{lowest value} = \text{maximum value} - \text{minimum value}$$

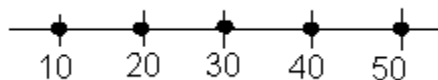
Example #3.2.1: Finding the Range

Look at the following three sets of data. Find the range of each of these.

a) 10, 20, 30, 40, 50

Solution:

Graph #3.2.1: Dot Plot for Example #3.2.1a

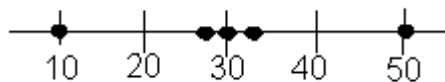


$$\text{mean} = 30, \quad \text{median} = 30, \quad \text{range} = 50 - 10 = 40$$

b) 10, 29, 30, 31, 50

Solution:

Graph #3.2.2: Dot Plot for Example #3.2.1b



$$\text{mean} = 30, \quad \text{median} = 30, \quad \text{range} = 50 - 10 = 40$$

c) 28, 29, 30, 31, 32

Solution:

Graph #3.2.3: Dot Plot for Example #3.2.1



$$\text{mean} = 30, \quad \text{median} = 30, \quad \text{range} = 32 - 28 = 4$$

Based on the mean, median, and range in example #3.2.1, the first two distributions are the same, but you can see from the graphs that they are different. In example #3.2.1a the data are spread out equally. In example #3.2.1b the data has a clump in the middle and a single value at each end. The mean and median are the same for example #3.2.1c but the range is very different. All the data is clumped together in the middle.

The range doesn't really provide a very accurate picture of the variability. A better way to describe how the data is spread out is needed. Instead of looking at the distance the highest value is from the lowest how about looking at the distance each value is from the mean. This distance is called the **deviation**.

Example #3.2.2: Finding the Deviations

Suppose a vet wants to analyze the weights of cats. The weights (in pounds) of five cats are 6.8, 8.2, 7.5, 9.4, and 8.2. Find the deviation for each of the data values.

Solution:

Variable: x = weight of a cat

The mean for this data set is $\bar{x} = 8.02$ pounds.

Table #3.2.1: Deviations of Weights of Cats

x	$x - \bar{x}$
6.8	$6.8 - 8.02 = -1.22$
8.2	$8.2 - 8.02 = 0.18$
7.5	$7.5 - 8.02 = -0.52$
9.4	$9.4 - 8.02 = 1.38$
8.2	$8.2 - 8.02 = 0.18$

Now you might want to average the deviation, so you need to add the deviations together.

Table #3.2.2: Sum of Deviations of Weights of Cats

x	$x - \bar{x}$
6.8	$6.8 - 8.02 = -1.22$
8.2	$8.2 - 8.02 = .018$
7.5	$7.5 - 8.02 = -0.52$
9.4	$9.4 - 8.02 = 1.38$
8.2	$8.2 - 8.02 = 0.18$
Total	0

This can't be right. The average distance from the mean cannot be 0. The reason it adds to 0 is because there are some positive and negative values. You need to get rid of the negative signs. How can you do that? You could square each deviation.

Table #3.2.3: Squared Deviations of Weights of Cats

x	$x - \bar{x}$	$(x - \bar{x})^2$
6.8	$6.8 - 8.02 = -1.22$	1.4884
8.2	$8.2 - 8.02 = .018$	0.0324
7.5	$7.5 - 8.02 = -0.52$	0.2704
9.4	$9.4 - 8.02 = 1.38$	1.9044
8.2	$8.2 - 8.02 = 0.18$	0.0324
Total	0	3.728

Now average the total of the squared deviations. The only thing is that in statistics there is a strange average here. Instead of dividing by the number of data values you divide by the number of data values minus 1. In this case you would have

$$s^2 = \frac{3.728}{5-1} = \frac{3.728}{4} = 0.932 \text{ pounds}^2$$

Notice that this is denoted as s^2 . This is called the variance and it is a measure of the average squared distance from the mean. If you now take the square root, you will get the average distance from the mean. This is called the standard deviation, and is denoted with the letter s .

$$s = \sqrt{.932} \approx 0.965 \text{ pounds}$$

The standard deviation is the average (mean) distance from a data point to the mean. It can be thought of as how much a typical data point differs from the mean.

The **sample variance** formula:

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

where $\bar{x}$ is the sample mean, n is the sample size, and Σ means to find the sum

The **sample standard deviation** formula:

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

The $n - 1$ on the bottom has to do with a concept called degrees of freedom. Basically, it makes the sample standard deviation a better approximation of the population standard deviation.

The **population variance** formula:

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

where σ is the Greek letter sigma and σ^2 represents the population variance, μ is the population mean, and N is the size of the population.

The **population standard deviation** formula:

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

Note: the sum of the deviations should always be 0. If it isn't, then it is because you rounded, you used the median instead of the mean, or you made an error. Try not to round too much in the calculations for standard deviation since each rounding causes a slight error.

Example #3.2.3: Finding the Standard Deviation

Suppose that a manager wants to test two new training programs. He randomly selects 5 people for each training type and measures the time it takes to complete a task after the training. The times for both trainings are in table #3.2.4. Which training method is better?

Table #3.2.4: Time to Finish Task in Minutes

Training 1	56	75	48	63	59
Training 2	60	58	66	59	58

Solution:

It is important that you define what each variable is since there are two of them.

Variable 1: X_1 = productivity from training 1

Variable 2: X_2 = productivity from training 2

To answer which training method better, first you need some descriptive statistics. Start with the mean for each sample.

$$\bar{x}_1 = \frac{56 + 75 + 48 + 63 + 59}{5} = 60.2 \text{ minutes}$$

$$\bar{x}_2 = \frac{60 + 58 + 66 + 59 + 58}{5} = 60.2 \text{ minutes}$$

Since both means are the same values, you cannot answer the question about which is better. Now calculate the standard deviation for each sample.

Table #3.2.5: Squared Deviations for Training 1

x_1	$x_1 - \bar{x}_1$	$(x_1 - \bar{x}_1)^2$
56	-4.2	17.64
75	14.8	219.04
48	-12.2	148.84
63	2.8	7.84
59	-1.2	1.44
Total	0	394.8

Table #3.2.6: Squared Deviations for Training 2

x_2	$x_2 - \bar{x}_2$	$(x_2 - \bar{x}_2)^2$
60	-0.2	0.04
58	-2.2	4.84
66	5.8	33.64
59	-1.2	1.44
58	-2.2	4.84
Total	0	44.8

The variance for each sample is:

$$s_1^2 = \frac{394.8}{5-1} = 98.7 \text{ minutes}^2$$

$$s_2^2 = \frac{44.8}{5-1} = 11.2 \text{ minutes}^2$$

The standard deviations are:

$$s_1 = \sqrt{98.7} \approx 9.93 \text{ minutes}$$

$$s_2 = \sqrt{11.2} \approx 3.35 \text{ minutes}$$

From the standard deviations, the second training seemed to be the better training since the data is less spread out. This means it is more consistent. It would be better for the managers in this case to have a training program that produces more

consistent results so they know what to expect for the time it takes to complete the task.

You can do the calculations for the descriptive statistics using the technology. The procedure for calculating the sample mean ($\bar{x}$) and the sample standard deviation (s_x) for X_2 in example #3.2.3 on the TI-83/84 is in figures 3.2.1 through 3.2.4 (the procedure is the same for X_1). Note the calculator gives you the population standard deviation (σ_x) because it doesn't know whether the data you input is a population or a sample. You need to decide which value you need to use, based on whether you have a population or sample. In almost all cases you have a sample and will be using s_x . Also, the calculator uses the notation of s_x instead of just s . It is just a way for it to denote the information. First you need to go into the STAT menu, and then Edit. This will allow you to type in your data (see figure #3.2.1).

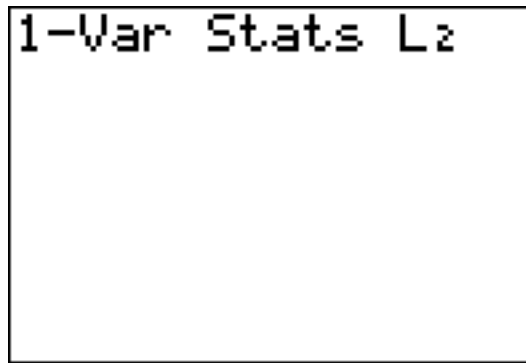
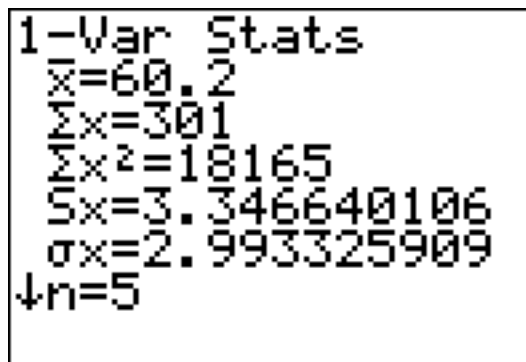
Figure #3.2.1: TI-83/84 Calculator Edit Setup

L1	L2	L3	2
56	60	-----	
75	58		
48	66		
63	59		
59	58		
-----	-----		
L2(6) =			

Once you have the data into the calculator, you then go back to the STAT menu, move over to CALC, and then choose 1-Var Stats (see figure #3.2.2). The calculator will now put 1-Var Stats on the main screen. Now type in L2 (2nd button and 2) and then press ENTER. (Note if you have the newer operating system on the TI-84, then the procedure is slightly different.) The results from the calculator are in figure #3.2.4.

Figure #3.2.2: TI-83/84 Calculator CALC Menu

EDIT	TESTS
1: 1-Var Stats	
2: 2-Var Stats	
3: Med-Med	
4: LinReg(ax+b)	
5: QuadReg	
6: CubicReg	
7: QuartReg	

Figure #3.2.3: TI-83/84 Calculator Input for Example #3.2.3 Variable X_2 **Figure #3.2.4: TI-83/84 Calculator Results for Example #3.2.3 Variable X_2** 

The processes for finding the mean, median, range, standard deviation, and variance on R are as follows:

variable<-c(type in your data)

To find the mean, use mean(variable)

To find the median, use median(variable)

To find the range, use range(variable). Then find maximum – minimum.

To find the standard deviation, use sd(variable)

To find the variance, use var(variable)

For the second data set in example #3.2.3, the commands and results would be

```
productivity_2<-c(60, 58, 66, 59, 58)
```

```
mean(productivity_2)
```

```
[1] 60.2
```

```
median(productivity_2)
```

```
[1] 59
```

```
range(productivity_2)
```

```
[1] 58 66
```

```
sd(productivity_2)
```

```
[1] 3.34664
```

```
var(productivity_2)
```

```
[1] 11.2
```

In general a “small” standard deviation means the data is close together (more consistent) and a “large” standard deviation means the data is spread out (less consistent).

Sometimes you want consistent data and sometimes you don’t. As an example if you are making bolts, you want to lengths to be very consistent so you want a small standard deviation. If you are administering a test to see who can be a pilot, you want a large standard deviation so you can tell who are the good pilots and who are the bad ones.

What do “small” and “large” mean? To a bicyclist whose average speed is 20 mph, $s = 20$ mph is huge. To an airplane whose average speed is 500 mph, $s = 20$ mph is nothing. The “size” of the variation depends on the size of the numbers in the problem and the mean. Another situation where you can determine whether a standard deviation is small or large is when you are comparing two different samples such as in example #3.2.3. A sample with a smaller standard deviation is more consistent than a sample with a larger standard deviation.

Many other books and authors stress that there is a computational formula for calculating the standard deviation. However, this formula doesn’t give you an idea of what standard deviation is and what you are doing. It is only good for doing the calculations quickly. It goes back to the days when standard deviations were calculated by hand, and the person needed a quick way to calculate the standard deviation. It is an archaic formula that this author is trying to eradicate it. It is not necessary anymore, since most calculators and computers will do the calculations for you with as much meaning as this formula gives. It is suggested that you never use it. If you want to understand what the standard deviation is doing, then you should use the definition formula. If you want an answer quickly, use a computer or calculator.

Use of Standard Deviation

One of the uses of the standard deviation is to describe how a population is distributed by using Chebyshev’s Theorem. This theorem works for any distribution, whether it is skewed, symmetric, bimodal, or any other shape. It gives you an idea of how much data is a certain distance on either side of the mean.

Chebyshev’s Theorem

For any set of data:

At least 75% of the data fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.

At least 88.9% of the data fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.

At least 93.8% of the data fall in the interval from $\mu - 4\sigma$ to $\mu + 4\sigma$.

Example #3.2.4: Using Chebyshev's Theorem

The U.S. Weather Bureau has provided the information in table #3.2.7 about the total annual number of reported strong to violent (F3+) tornados in the United States for the years 1954 to 2012. ("U.S. tornado climatology," 17)

Table #3.2.7: Annual Number of Violent Tornados in the U.S.

46	47	31	41	24	56	56	23	31	59
39	70	73	85	33	38	45	39	35	22
51	39	51	131	37	24	57	42	28	45
98	35	54	45	30	15	35	64	21	84
40	51	44	62	65	27	34	23	32	28
41	98	82	47	62	21	31	29	32	

- a.) Use Chebyshev's theorem to find an interval centered about the mean annual number of strong to violent (F3+) tornados in which you would expect at least 75% of the years to fall.

Solution:

Variable: x = number of strong or violent (F3+) tornadoes

Chebyshev's theorem says that at least 75% of the data will fall in the interval from $\mu - 2\sigma$ to $\mu + 2\sigma$.

You do not have the population, so you need to estimate the population mean and standard deviation using the sample mean and standard deviation. You can find the sample mean and standard deviation using technology:

$$\bar{x} \approx 46.24, s \approx 22.18$$

So,

$$\mu \approx 46.24, \sigma \approx 22.18$$

$$\mu - 2\sigma \text{ to } \mu + 2\sigma$$

$$46.24 - 2(22.18) \text{ to } 46.24 + 2(22.18)$$

$$46.24 - 44.36 \text{ to } 46.24 + 44.36$$

$$1.88 \text{ to } 90.60$$

Since you can't have fractional number of tornados, round to the nearest whole number.

At least 75% of the years have between 2 and 91 strong to violent (F3+) tornados. (Actually, all but three years' values fall in this interval, that means that

$$\frac{56}{59} \approx 94.9\% \text{ actually fall in the interval.)}$$

- b.) Use Chebyshev's theorem to find an interval centered about the mean annual number of strong to violent (F3+) tornados in which you would expect at least 88.9% of the years to fall.

Solution:

Variable: x = number of strong or violent (F3+) tornadoes

Chebyshev's theorem says that at least 88.9% of the data will fall in the interval from $\mu - 3\sigma$ to $\mu + 3\sigma$.

$$\begin{aligned} &\mu - 3\sigma \text{ to } \mu + 3\sigma \\ &46.24 - 3(22.18) \text{ to } 46.24 + 3(22.18) \\ &46.24 - 66.54 \text{ to } 46.24 + 66.54 \\ &-20.30 \text{ to } 112.78 \end{aligned}$$

Since you can't have negative number of tornados, the lower limit is actually 0. Since you can't have fractional number of tornados, round to the nearest whole number.

At least 88.9% of the years have between 0 and 113 strong to violent (F3+) tornados.

(Actually, all but one year falls in this interval, that means that $\frac{58}{59} \approx 98.3\%$ actually fall in the interval.)

Chebyshev's Theorem says that at least 75% of the data is within two standard deviations of the mean. That percentage is fairly high. There isn't much data outside two standard deviations. A rule that can be followed is that if a data value is within two standard deviations, then that value is a common data value. If the data value is outside two standard deviations of the mean, either above or below, then the number is uncommon. It could even be called unusual. An easy calculation that you can do to figure it out is to find the difference between the data point and the mean, and then divide that answer by the standard deviation. As a formula this would be

$$\frac{x - \mu}{\sigma}.$$

If you don't know the population mean, μ , and the population standard deviation, σ , then use the sample mean, $\bar{x}$, and the sample standard deviation, s , to estimate the population parameter values. However, realize that using the sample standard deviation may not actually be very accurate.

Example #3.2.5: Determining If a Value Is Unusual

- a.) In 1974, there were 131 strong or violent (F3+) tornados in the United States. Is this value unusual? Why or why not?

Solution:

Variable: x = number of strong or violent (F3+) tornadoes

To answer this question, first find how many standard deviations 131 is from the mean. From example #3.2.4, we know $\mu \approx 46.24$ and $\sigma \approx 22.18$. For $x = 131$,

$$\frac{x - \mu}{\sigma} = \frac{131 - 46.24}{22.18} \approx 3.82$$

Since this value is more than 2, then it is unusual to have 131 strong or violent (F3+) tornados in a year.

- b.) In 1987, there were 15 strong or violent (F3+) tornados in the United States. Is this value unusual? Why or why not?

Solution:

Variable: x = number of strong or violent (F3+) tornadoes

For this question the $x = 15$,

$$\frac{x - \mu}{\sigma} = \frac{15 - 46.24}{22.18} \approx -1.41$$

Since this value is between -2 and 2 , then it is not unusual to have only 15 strong or violent (F3+) tornados in a year.

Section 3.2: Homework

- 1.) Cholesterol levels were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985) and are in table #3.2.8.

Table #3.2.8: Cholesterol Levels

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

Find the mean, median, range, variance, and standard deviation using technology.

- 2.) The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Pacific Ocean are listed in table #3.2.9 (Lee, 1994).

Table #3.2.9: Lengths of Rivers (km) Flowing to Pacific Ocean

River	Length (km)	River	Length (km)
Clarence	209	Clutha	322
Conway	48	Taieri	288
Waiau	169	Shag	72
Hurunui	138	Kakanui	64
Waipara	64	Waitaki	209
Ashley	97	Waihao	64
Waimakariri	161	Pareora	56
Selwyn	95	Rangitata	121
Rakaia	145	Ophi	80
Ashburton	90		

- Find the mean and median.
 - Find the range.
 - Find the variance and standard deviation.
- 3.) The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed in table #3.2.10 (Lee, 1994).

Table #3.2.10: Lengths of Rivers (km) Flowing to Tasman Sea

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

- Find the mean and median.
- Find the range.
- Find the variance and standard deviation.

- 4.) Eyeglassmatic manufactures eyeglasses for their retailers. They test to see how many defective lenses they made the time period of January 1 to March 31. Table #3.2.11 gives the defect and the number of defects.

Table #3.2.11: Number of Defective Lenses

Defect type	Number of defects
Scratch	5865
Right shaped – small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape – big	1105
Lost in lab	976
Spots/bubble – intern	976

- Find the mean and median.
- Find the range.
- Find the variance and standard deviation.

- 5.) Print-O-Matic printing company's employees have salaries that are contained in table #3.2.12.

Table #3.2.12: Salaries of Print-O-Matic Printing Company Employees

Employee	Salary (\$)	Employee	Salary (\$)
CEO	272,500	Administration	66,346
Driver	58,456	Sales	109,739
CD74	100,702	Designer	90,090
CD65	57,380	Platens	69,573
Embellisher	73,877	Polar	75,526
Folder	65,270	ITEK	64,553
GTO	74,235	Mgmt	108,448
Pre Press Manager	108,448	Handwork	52,718
Pre Press Manager/ IT	98,837	Horizon	76,029
Pre Press/ Graphic Artist	75,311		

Find the mean, median, range, variance, and standard deviation using technology.

- 6.) Print-O-Matic printing company spends specific amounts on fixed costs every month. The costs of those fixed costs are in table #3.2.13.

Table #3.2.13: Fixed Costs for Print-O-Matic Printing Company

Monthly charges	Monthly cost (\$)
Bank charges	482
Cleaning	2208
Computer expensive	2471
Lease payments	2656
Postage	2117
Uniforms	2600

- a.) Find the mean and median.
 b.) Find the range.
 c.) Find the variance and standard deviation.
- 7.) Compare the two data sets in problems 2 and 3 using the mean and standard deviation. Discuss which mean is higher and which has a larger spread of the data.
- 8.) Table #3.2.14 contains pulse rates collected from males, who are non-smokers but do drink alcohol ("Pulse rates before," 2013). The before pulse rate is before they exercised, and the after pulse rate was taken after the subject ran in place for one minute.

Table #3.2.14: Pulse Rates of Males Before and After Exercise

Pulse before	Pulse after	Pulse before	Pulse after
76	88	59	92
56	110	60	104
64	126	65	82
50	90	76	150
49	83	145	155
68	136	84	140
68	125	78	141
88	150	85	131
80	146	78	132
78	168		

Compare the two data sets using the mean and standard deviation. Discuss which mean is higher and which has a larger spread of the data.

- 9.) Table #3.2.15 contains pulse rates collected from females, who are non-smokers but do drink alcohol ("Pulse rates before," 2013). The before pulse rate is before they exercised, and the after pulse rate was taken after the subject ran in place for one minute.

Table #3.2.15: Pulse Rates of Females Before and After Exercise

Pulse before	Pulse after	Pulse before	Pulse after
96	176	92	120
82	150	70	96
86	150	75	130
72	115	70	119
78	129	70	95
90	160	68	84
88	120	47	136
71	125	64	120
66	89	70	98
76	132	74	168
70	120	85	130

Compare the two data sets using the mean and standard deviation. Discuss which mean is higher and which has a larger spread of the data.

- 10.) To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) immediately before and after the Reiki treatment (Olson & Hanson, 1997) and the data is in table #3.2.16.

Table #3.2.16: Pain Measurements Before and After Reiki Treatment

VAS before	VAS after	VAS before	VAS after
6	3	5	1
2	1	1	0
2	0	6	4
9	1	6	1
3	0	4	4
3	2	4	1
4	1	7	6
5	2	2	1
2	2	4	3
3	0	8	8

Compare the two data sets using the mean and standard deviation. Discuss which mean is higher and which has a larger spread of the data.

- 11.) Table #3.2.17 contains data collected on the time it takes in seconds of each passage of play in a game of rugby. ("Time of passages," 2013)

Table #3.2.17: Times (in seconds) of rugby plays

39.2	2.7	9.2	14.6	1.9	17.8	15.5	53.8	17.5	27.5
4.8	8.6	22.1	29.8	10.4	9.8	27.7	32.7	32	34.3
29.1	6.5	2.8	10.8	9.2	12.9	7.1	23.8	7.6	36.4
35.6	28.4	37.2	16.8	21.2	14.7	44.5	24.7	36.2	20.9
19.9	24.4	7.9	2.8	2.7	3.9	14.1	28.4	45.5	38
18.5	8.3	56.2	10.2	5.5	2.5	46.8	23.1	9.2	10.3
10.2	22	28.5	24	17.3	12.7	15.5	4	5.6	3.8
21.6	49.3	52.4	50.1	30.5	37.2	15	38.7	3.1	11
10	5	48.8	3.6	12.6	9.9	58.6	37.9	19.4	29.2
12.3	39.2	22.2	39.7	6.4	2.5	34			

- Using technology, find the mean and standard deviation.
 - Use Chebyshev's theorem to find an interval centered about the mean times of each passage of play in the game of rugby in which you would expect at least 75% of the times to fall.
 - Use Chebyshev's theorem to find an interval centered about the mean times of each passage of play in the game of rugby in which you would expect at least 88.9% of the times to fall.
- 12.) Yearly rainfall amounts (in millimeters) in Sydney, Australia, are in table #3.2.18 ("Annual maximums of," 2013).

Table #3.2.18: Yearly Rainfall Amounts in Sydney, Australia

146.8	383	90.9	178.1	267.5	95.5	156.5	180
90.9	139.7	200.2	171.7	187.2	184.9	70.1	58
84.1	55.6	133.1	271.8	135.9	71.9	99.4	110.6
47.5	97.8	122.7	58.4	154.4	173.7	118.8	88
84.6	171.5	254.3	185.9	137.2	138.9	96.2	85
45.2	74.7	264.9	113.8	133.4	68.1	156.4	

- Using technology, find the mean and standard deviation.
- Use Chebyshev's theorem to find an interval centered about the mean yearly rainfall amounts in Sydney, Australia, in which you would expect at least 75% of the amounts to fall.
- Use Chebyshev's theorem to find an interval centered about the mean yearly rainfall amounts in Sydney, Australia, in which you would expect at least 88.9% of the amounts to fall.

- 13.) The number of deaths attributed to UV radiation in African countries in the year 2002 is given in table #3.2.19 ("UV radiation: Burden," 2013).

Table #3.2.19: Number of Deaths from UV Radiation

50	84	31	338	6	504	40	7	58
204	15	27	39	1	45	174	98	94
199	9	27	58	356	5	45	5	94
26	171	13	57	138	39	3	171	41
1177	102	123	433	35	40	456	125	

- a.) Using technology, find the mean and standard deviation.
- b.) Use Chebyshev's theorem to find an interval centered about the mean number of deaths from UV radiation in which you would expect at least 75% of the numbers to fall.
- c.) Use Chebyshev's theorem to find an interval centered about the mean number of deaths from UV radiation in which you would expect at least 88.9% of the numbers to fall.
- 14.) The time (in 1/50 seconds) between successive pulses along a nerve fiber ("Time between nerve," 2013) are given in table #3.2.20.

Table 3.2.20: Time (in 1/50 seconds) Between Successive Pulses

10.5	1.5	2.5	5.5	29.5	3	9	27.5	18.5	4.5
7	9.5	1	7	4.5	2.5	7.5	11.5	7.5	4
12	8	3	5.5	7.5	4.5	1.5	10.5	1	7
12	14.5	8	3.5	3.5	2	1	7.5	6	13
7.5	16.5	3	25.5	5.5	14	18	7	27.5	14

- a.) Using technology, find the mean and standard deviation.
- b.) Use Chebyshev's theorem to find an interval centered about the mean time between successive pulses along a nerve fiber in which you would expect at least 75% of the times to fall.
- c.) Use Chebyshev's theorem to find an interval centered about the mean time between successive pulses along a nerve fiber in which you would expect at least 88.9% of the times to fall.
- 15.) Suppose a passage of play in a rugby game takes 75.1 seconds. Would it be unusual for this to happen? Use the mean and standard deviation that you calculated in problem 11.
- 16.) Suppose Sydney, Australia received 300 mm of rainfall in a year. Would this be unusual? Use the mean and standard deviation that you calculated in problem 12.
- 17.) Suppose in a given year there were 2257 deaths attributed to UV radiation in an African country. Is this value unusual? Use the mean and standard deviation that you calculated in problem 13.
- 18.) Suppose it only takes 2 (1/50 seconds) for successive pulses along a nerve fiber. Is this value unusual? Use the mean and standard deviation that you calculated in problem 14.

Section 3.3: Ranking

Along with the center and the variability, another useful numerical measure is the ranking of a number. A **percentile** is a measure of ranking. It represents a location measurement of a data value to the rest of the values. Many standardized tests give the results as a percentile. Doctors also use percentiles to track a child's growth.

The ***k*th percentile** is the data value that has *k*% of the data at or below that value.

Example #3.3.1: Interpreting Percentile

a.) What does a score of the 90th percentile mean?

Solution:

This means that 90% of the scores were at or below this score. (A person did the same as or better than 90% of the test takers.)

b.) What does a score of the 70th percentile mean?

Solution:

This means that 70% of the scores were at or below this score.

Example #3.3.2: Percentile Versus Score

If the test was out of 100 points and you scored at the 80th percentile, what was your score on the test?

Solution:

You don't know! All you know is that you scored the same as or better than 80% of the people who took the test. If all the scores were really low, you could have still failed the test. On the other hand, if many of the scores were high you could have gotten a 95% or so.

There are special percentiles called quartiles. Quartiles are numbers that divide the data into fourths. One fourth (or a quarter) of the data falls between consecutive quartiles.

To find the quartiles:

- 1) Sort the data in increasing order.
 - 2) Find the median, this divides the data list into 2 halves.
 - 3) Find the median of the data below the median. This value is *Q1*.
 - 4) Find the median of the data above the median. This value is *Q3*.
- Ignore the median in both calculations for *Q1* and *Q3*

If you record the quartiles together with the maximum and minimum you have five numbers. This is known as the five-number summary. The five-number summary consists of the minimum, the first quartile (*Q1*), the median, the third quartile (*Q3*), and the maximum (in that order).

The interquartile range, *IQR*, is the difference between the first and third quartiles, *Q1* and *Q3*. Half of the data (50%) falls in the interquartile range. If the *IQR* is “large” the data is spread out and if the *IQR* is “small” the data is closer together.

Interquartile Range (*IQR*)

$$IQR = Q3 - Q1$$

Determining probable outliers from *IQR*: fences

A value that is less than $Q1 - 1.5 * IQR$ (this value is often referred to as a *low fence*) is considered an outlier.

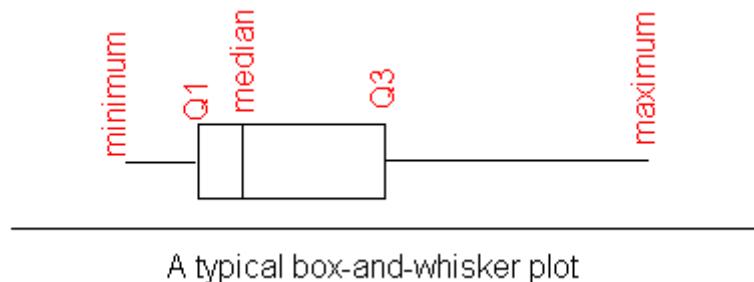
Similarly, a value that is more than $Q3 + 1.5 * IQR$ (the *high fence*) is considered an outlier.

A box plot (or box-and-whisker plot) is a graphical display of the five-number summary. It can be drawn vertically or horizontally. The basic format is a box from *Q1* to *Q3*, a vertical line across the box for the median and horizontal lines as whiskers extending out each end to the minimum and maximum. The minimum and maximum can be represented with dots. Don’t forget to label the tick marks on the number line and give the graph a title.

An alternate form of a Box-and-Whiskers Plot, known as a modified box plot, only extends the left line to the smallest value greater than the *low fence*, and extends the left line to the largest value less than the *high fence*, and displays markers (dots, circles or asterisks) for each outlier.

If the data are *symmetrical*, then the box plot will be visibly symmetrical. If the data distribution has a left skew or a right skew, the line on that side of the box plot will be visibly long. If the plot is symmetrical, and the four quartiles are all about the same length, then the data are likely a near *uniform* distribution. If a box plot is symmetrical, and both outside lines are noticeably longer than the *Q1* to median and median to *Q3* distance, the distribution is then probably *bell-shaped*.

Figure #3.3.1: Typical Box Plot



Example #3.3.3: Five-number Summary for an Even Number of Data Points

The total assets in billions of Australian dollars (AUD) of Australian banks for the year 2012 are given in table #3.3.1 ("Reserve bank of," 2013). Find the five-number summary and the interquartile range (*IQR*), and draw a box-and-whiskers plot.

Table #3.3.1: Total Assets (in billions of AUD) of Australian Banks

2855	2862	2861	2884	3014	2965
2971	3002	3032	2950	2967	2964

Solution:

Variable: x = total assets of Australian banks

First sort the data.

Table #3.3.2: Sorted Data for Total Assets

2855	2861	2862	2884	2950	2964	2965	2967	2971	3002	3014	3032
------	------	------	------	------	------	------	------	------	------	------	------

The minimum is 2855 billion AUD and the maximum is 3032 billion AUD.

There are 12 data points so the median is the average of the 6th and 7th numbers.

Table #3.3.3: Sorted Data for Total Assets with Median

2855	2861	2862	2884	2950	2964	2965	2967	2971	3002	3014	3032
------	------	------	------	------	------	------	------	------	------	------	------

$$\frac{2964 + 2965}{2} = 2964.5 \text{ billion AUD}$$

To find $Q1$, find the median of the first half of the list.

Table #3.3.4: Finding $Q1$

2855	2861	2862	2884	2950	2964
------	------	------	------	------	------

Q1

$$Q1 = \frac{2862 + 2884}{2} = 2873 \text{ billion AUD}$$

To find $Q3$, find the median of the second half of the list.

Table #3.3.5: Finding Q_3

2965	2967	2971	3002	3014	3032
------	------	------	------	------	------

Q_3

$$Q_3 = \frac{2971 + 3002}{2} = 2986.5 \text{ billion AUD}$$

The five-number summary is (all numbers in billion AUD)

Minimum: 2855

Q_1 : 2873

Median: 2964.5

Q_3 : 2986.5

Maximum: 3032

To find the interquartile range, IQR , find $Q_3 - Q_1$.

$$IQR = 2986.5 - 2873 = 113.5 \text{ billion AUD}$$

This tells you the middle 50% of assets were within 113.5 billion AUD of each other.

You can use the five-number summary to draw the box-and-whiskers plot:

Graph #3.3.1: Box Plot of Total Assets of Australian Banks

Total Assets of Australian Banks in 2012
(Billions of Australian Dollars (AUD))



The distribution is skewed right because the right tail is longer.

Example #3.3.4: Five-number Summary for an Odd Number of Data Points

The life expectancy for a person living in one of 11 countries in the region of South East Asia in 2012 is given below ("Life expectancy in," 2013). Find the five-number summary for the data and the IQR , then draw a box-and-whiskers plot.

Table #3.3.6: Life Expectancy of a Person Living in South-East Asia

70	67	69	65	69	77
65	68	75	74	64	

Solution:

Variable: x = life expectancy of a person

Sort the data first.

Table #3.3.7: Sorted Life Expectancies

64	65	65	67	68	69	69	70	74	75	77
----	----	----	----	----	----	----	----	----	----	----

The minimum is 64 years and the maximum is 77 years.

There are 11 data points so the median is the 6th number in the list.

Table #3.3.8: Finding the Median of Life Expectancies

64	65	65	67	68	69	69	70	74	75	77
----	----	----	----	----	----	----	----	----	----	----

Median = 69 years

Finding the $Q1$ and $Q3$ you need to find the median of the numbers below the median and above the median. The median is not included in either calculation.

Table #3.3.9: Finding $Q1$

64	65	65	67	68
----	----	----	----	----

$Q1$

Table #3.3.10: Finding $Q3$

69	70	74	75	77
----	----	----	----	----

$Q3$

$Q1$ = 65 years and $Q3$ = 74 years.

The five-number summary is (in years)

Minimum: 64

$Q1$: 65

Median: 69

$Q3$: 74

Maximum: 77

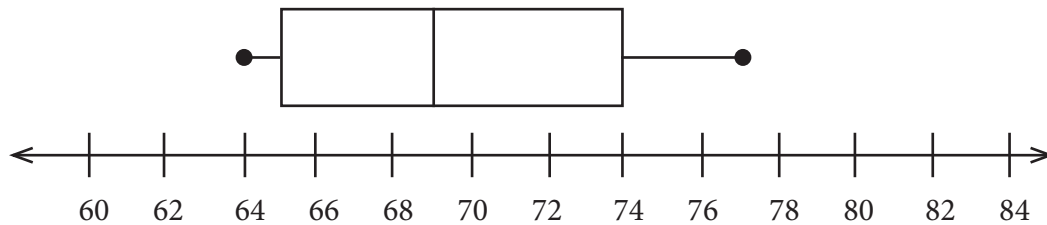
To find the interquartile range (*IQR*)

$$IQR = Q3 - Q1 = 74 - 65 = 9 \text{ years}$$

The middle 50% of life expectancies are within 9 years.

Graph #3.3.2: Box Plot of Life Expectancy

Life Expectancy of Southeast Asian Countries in 2011

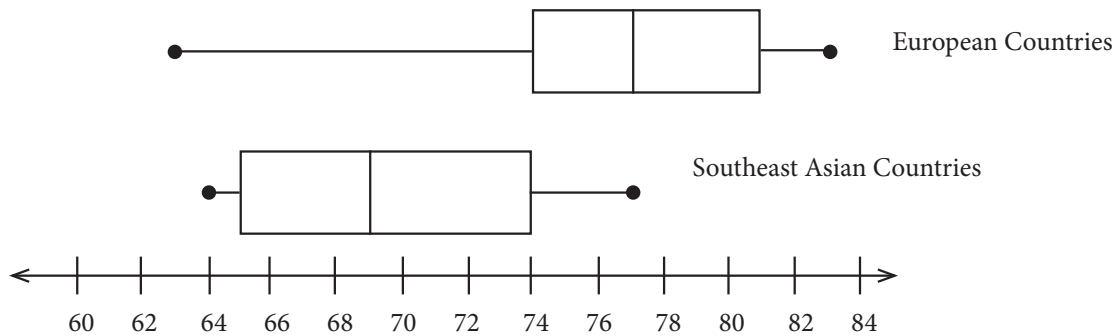


This distribution looks somewhat skewed right, since the whisker is longer on the right. However, it could be considered almost symmetric too since the box looks somewhat symmetric.

You can draw 2 box plots side by side (or one above the other) to compare 2 samples. Since you want to compare the two data sets, make sure the box plots are on the same axes. As an example, suppose you look at the box-and-whiskers plot for life expectancy for European countries and Southeast Asian countries.

Graph #3.3.3: Box Plot of Life Expectancy of Two Regions

Life Expectancy of European Countries
and Southeast Asian Countries in 2011



Looking at the box-and-whiskers plot, you will notice that the three quartiles for life expectancy are all higher for the European countries, yet the minimum life expectancy for the European countries is less than that for the Southeast Asian countries. The life expectancy for the European countries appears to be skewed left, while the life expectancies for the Southeast Asian countries appear to be more symmetric. There are of course more qualities that can be compared between the two graphs.

To find the five-number summary using R, the command is:

```
variable<-c(type in data with commas)
summary(variable)
```

This command will give you the five number summary and the mean.

For example #3.3.4, the commands would be

```
expectancy<-c(70, 67, 69, 65, 69, 77, 65, 68, 75, 74, 64)
summary(expectancy)
```

The output would be:

```
Min. 1st Qu. Median Mean 3rd Qu. Max.
64.00 66.00 69.00 69.36 72.00 77.00
```

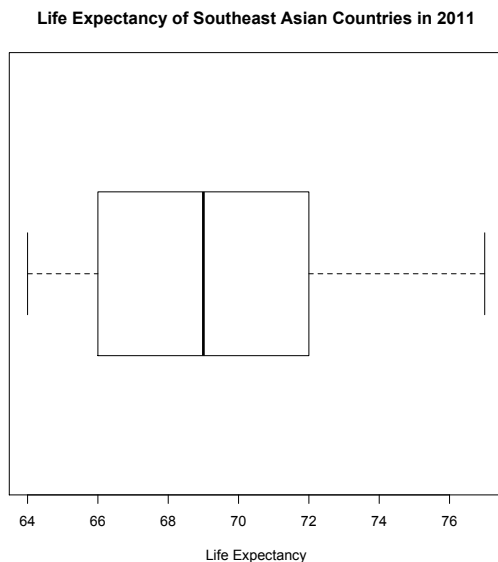
To draw the box plot the command is `boxplot(variable, main="title you want", xlab="label you want", horizontal = TRUE)`. The `horizontal = TRUE` orients the box plot to be horizontal. If you leave that part off, the box plot will be vertical by default.

For example #3.3.4, the command is

```
boxplot(expectancy, main="Life Expectancy of Southeast Asian Countries in
2011",horizontal=TRUE, xlab="Life Expectancy")
```

You should get the box plot in graph #3.3.4.

Graph #3.3.4: Box plot for Life Expectance in Southeast Asian Countries



This is known as a modified box plot. Instead of plotting the maximum and minimum, the box plot has as a lower line $Q1 - 1.5 * IQR$, and as an upper line, $Q3 + 1.5 * IQR$. Any values below the lower line or above the upper line are considered outliers. Outliers are plotted as dots on the modified box plot. This data set does not have any outliers.

Example #3.3.5: Putting it all together

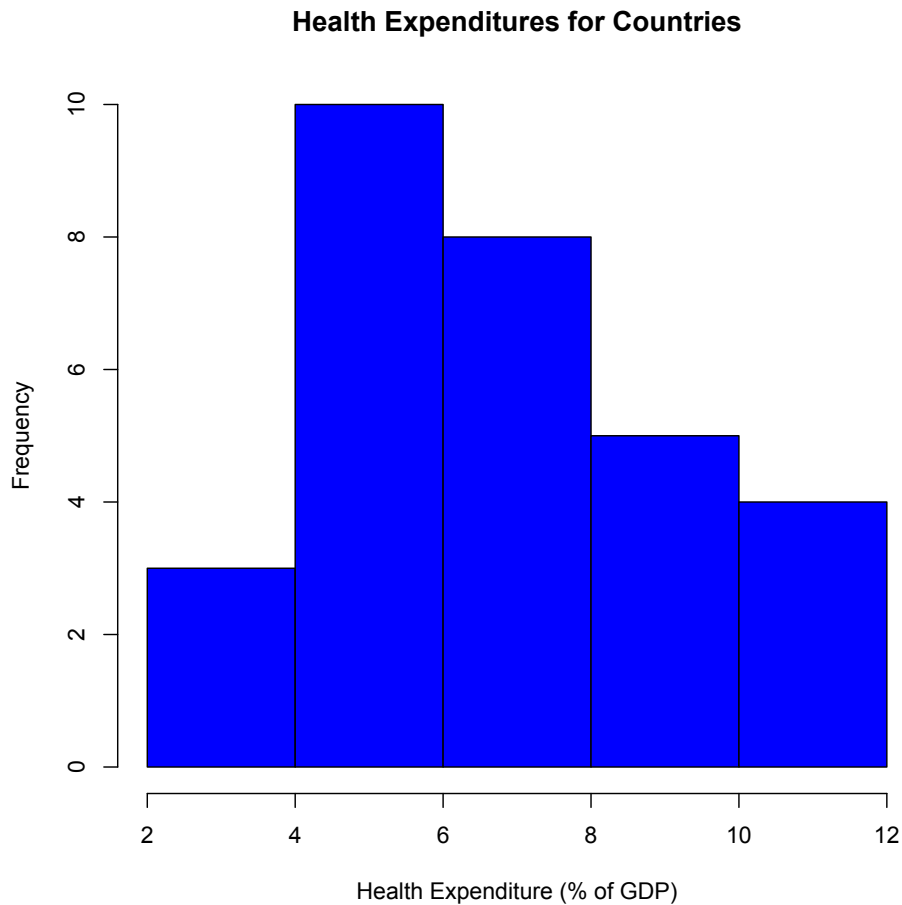
A random sample was collected on the health expenditures (as a % of GDP) of countries around the world. The data is in Table #3.3.11. Using graphical and numerical descriptive statistics, analyze the data and use it to predict the health expenditures of all countries in the world.

Table #3.3.11: Health Expenditures as a Percentage of GDP

3.35	5.94	10.64	5.24	3.79	5.65	7.66	7.38	5.87	11.15
5.96	4.78	7.75	2.72	9.50	7.69	10.05	11.96	8.18	6.74
5.89	6.20	5.98	8.83	6.78	6.66	9.45	5.41	5.16	8.55

Solution:

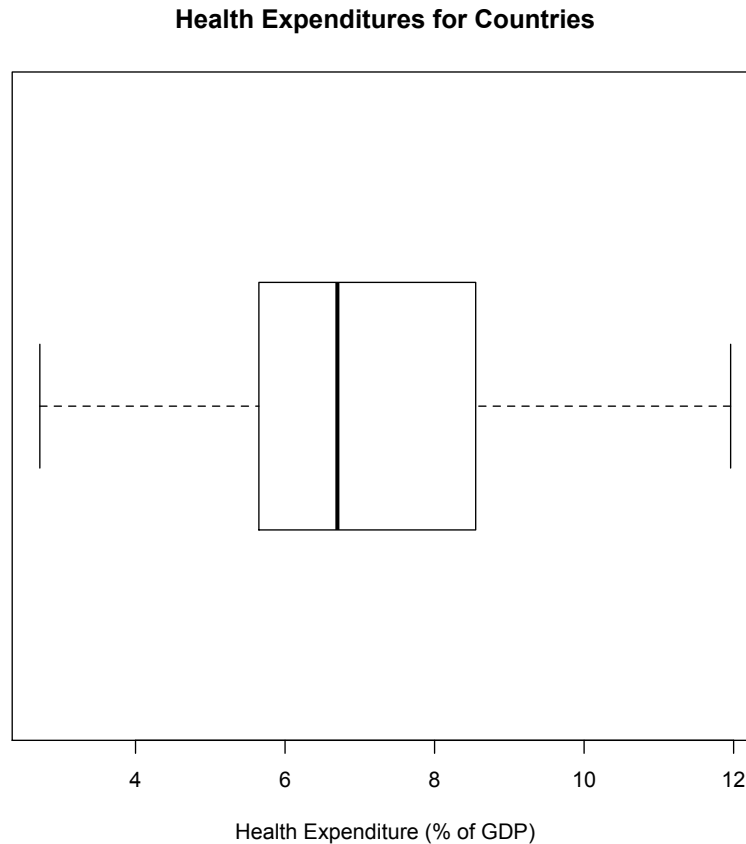
First, it might be useful to look at a visualization of the data, so create a histogram.

Graph #3.3.5: Histogram of Health Expenditure

From the graph, the data appears to be somewhat skewed right. So there are some countries that spend more on health based on a percentage of GDP than other countries, but the majority of countries appear to spend around 4 to 8% of their GDP on health.

Numerical descriptions might also be useful. Using technology, the mean is 7.03%, the standard deviation is 2.27%, and the five-number summary is minimum = 2.72%, Q1 = 5.71%, median = 6.70%, Q3 = 8.46%, and maximum = 11.96%. To visualize the five-number summary, create a box plot.

Graph #3.3.6: Box Plot of Health Expenditure



So it appears that countries spend on average about 7% of their GDP on health. The spread is somewhat low, since the standard deviation is fairly small, which means that the data is fairly consistent. The five-number summary confirms that the data is slightly skewed right. The box plot shows that there are no outliers. So from all of this information, one could say that countries spend a small percentage of their GDP on health and that most countries spend around the same amount. There doesn't appear to be any country that spends much more than other countries or much less than other countries.

Section 3.3: Homework

- 1.) Suppose you take a standardized test and you are in the 10th percentile. What does this percentile mean? Can you say that you failed the test? Explain.
- 2.) Suppose your child takes a standardized test in mathematics and scores in the 96th percentile. What does this percentile mean? Can you say your child passed the test? Explain.
- 3.) Suppose your child is in the 83rd percentile in height and 24th percentile in weight. Describe what this tells you about your child's stature.
- 4.) Suppose your work evaluates the employees and places them on a percentile ranking. If your evaluation is in the 65th percentile, do you think you are working hard enough? Explain.

- 5.) Cholesterol levels were collected from patients two days after they had a heart attack (Ryan, Joiner & Ryan, Jr, 1985) and are in table #3.3.12.

Table #3.3.12: Cholesterol Levels

270	236	210	142	280	272	160
220	226	242	186	266	206	318
294	282	234	224	276	282	360
310	280	278	288	288	244	236

Find the five-number summary and interquartile range (IQR), and draw a box-and-whiskers plot

- 6.) The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Pacific Ocean are listed in table #3.3.13 (Lee, 1994).

Table #3.3.13: Lengths of Rivers (km) Flowing to Pacific Ocean

River	Length (km)	River	Length (km)
Clarence	209	Clutha	322
Conway	48	Taieri	288
Waiau	169	Shag	72
Hurunui	138	Kakanui	64
Waipara	64	Waitaki	209
Ashley	97	Waihao	64
Waimakariri	161	Pareora	56
Selwyn	95	Rangitata	121
Rakaia	145	Ophi	80
Ashburton	90		

Find the five-number summary and interquartile range (IQR), and draw a box-and-whiskers plot

- 7.) The lengths (in kilometers) of rivers on the South Island of New Zealand that flow to the Tasman Sea are listed in table #3.3.14 (Lee, 1994).

Table #3.3.14: Lengths of Rivers (km) Flowing to Tasman Sea

River	Length (km)	River	Length (km)
Hollyford	76	Waimea	48
Cascade	64	Motueka	108
Arawhata	68	Takaka	72
Haast	64	Aorere	72
Karangarua	37	Heaphy	35
Cook	32	Karamea	80
Waiho	32	Mokihinui	56
Whataroa	51	Buller	177
Wanganui	56	Grey	121
Waitaha	40	Taramakau	80
Hokitika	64	Arahura	56

Find the five-number summary and interquartile range (IQR), and draw a box-and-whiskers plot

- 8.) Eyeglassmatic manufactures eyeglasses for their retailers. They test to see how many defective lenses they made the time period of January 1 to March 31. Table #3.3.15 gives the defect and the number of defects.

Table #3.3.15: Number of Defective Lenses

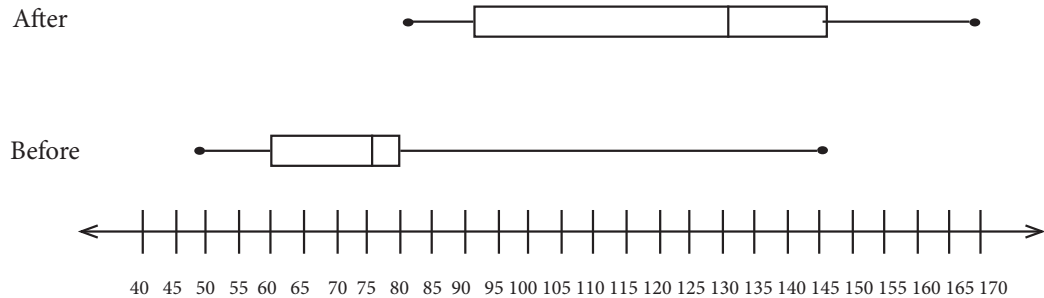
Defect type	Number of defects
Scratch	5865
Right shaped – small	4613
Flaked	1992
Wrong axis	1838
Chamfer wrong	1596
Crazing, cracks	1546
Wrong shape	1485
Wrong PD	1398
Spots and bubbles	1371
Wrong height	1130
Right shape – big	1105
Lost in lab	976
Spots/bubble – intern	976

Find the five-number summary and interquartile range (IQR), and draw a box-and-whiskers plot

- 9.) A study was conducted to see the effect of exercise on pulse rate. Male subjects were taken who do not smoke, but do drink. Their pulse rates were measured ("Pulse rates before," 2013). Then they ran in place for one minute and then measured their pulse rate again. Graph #3.3.7 is of box-and-whiskers plots that were created of the before and after pulse rates. Discuss any conclusions you can make from the graphs.

Graph #3.3.7: Box-and-Whiskers Plot of Pulse Rates for Males

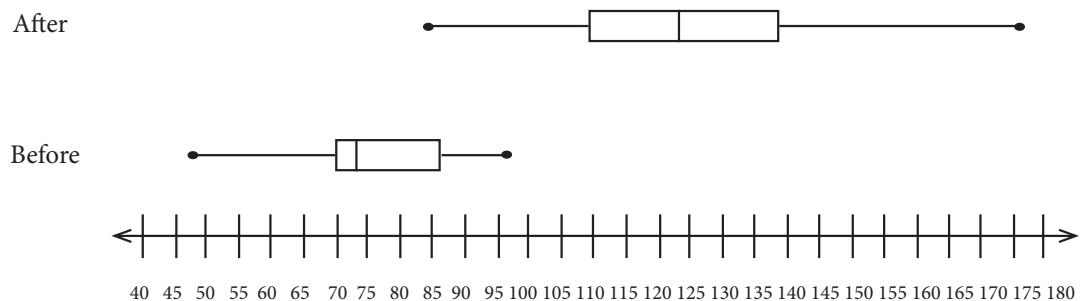
Pulse Rate of Male Subjects Before and After One Minute of Exercise



- 10.) A study was conducted to see the effect of exercise on pulse rate. Female subjects were taken who do not smoke, but do drink. Their pulse rates were measured ("Pulse rates before," 2013). Then they ran in place for one minute, and after measured their pulse rate again. Graph #3.3.8 is of box-and-whiskers plots that were created of the before and after pulse rates. Discuss any conclusions you can make from the graphs.

Graph #3.3.8: Box-and-Whiskers Plot of Pulse Rates for Females

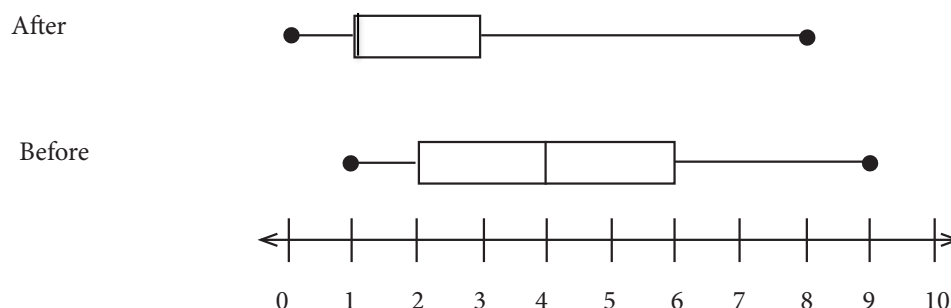
Pulse Rate of Female Subjects Before and After One Minute of Exercise



- 11.) To determine if Reiki is an effective method for treating pain, a pilot study was carried out where a certified second-degree Reiki therapist provided treatment on volunteers. Pain was measured using a visual analogue scale (VAS) immediately before and after the Reiki treatment (Olson & Hanson, 1997). Graph #3.3.9 is of box-and-whiskers plots that were created of the before and after VAS ratings. Discuss any conclusions you can make from the graphs.

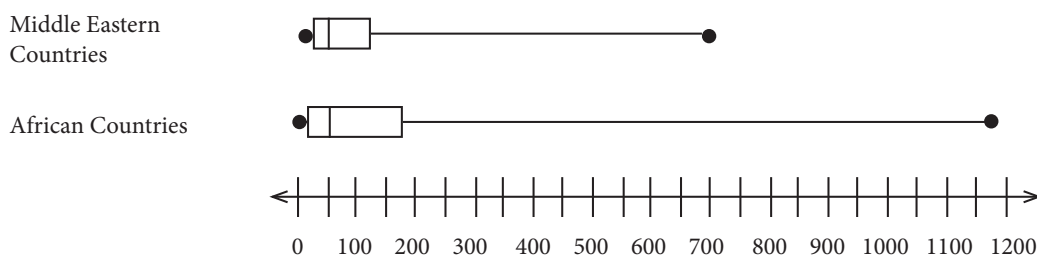
Graph #3.3.9: Box-and-Whiskers Plot of Pain Using Reiki

VAS of Subjects Before and After Reiki Treatment for Pain



- 12.) The number of deaths attributed to UV radiation in African countries and Middle Eastern countries in the year 2002 were collected by the World Health Organization ("UV radiation: Burden," 2013). Graph #3.3.10 is of box-and-whiskers plots that were created of the deaths in African countries and deaths in Middle Eastern countries. Discuss any conclusions you can make from the graphs.
- Table #3.3.10: Box-and-Whiskers Plot of UV Radiation Deaths in Different Regions**

UV Radiation Attributed Deaths in African and Middle Eastern Countries in 2002



Data Sources:

Annual maximums of daily rainfall in Sydney. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/sydrain.html>

Lee, A. (1994). *Data analysis: An introduction based on r. Auckland.* Retrieved from <http://www.statsci.org/data/oz/nzrivers.html>

Life expectancy in southeast Asia. (2013, September 23). Retrieved from <http://apps.who.int/gho/data/node.main.688>

Olson, K., & Hanson, J. (1997). Using reiki to manage pain: a preliminary report. *Cancer Prev Control*, 1(2), 108-13. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9765732>

Pulse rates before and after exercise. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/ms212.html>

Reserve bank of Australia. (2013, September 23). Retrieved from <http://data.gov.au/dataset/banks-assets>

Ryan, B. F., Joiner, B. L., & Ryan, Jr, T. A. (1985). *Cholesterol levels after heart attack.* Retrieved from <http://www.statsci.org/data/general/cholest.html>

Time between nerve pulses. (2013, September 25). Retrieved from <http://www.statsci.org/data/general/nerve.html>

Time of passages of play in rugby. (2013, September 25). Retrieved from <http://www.statsci.org/data/oz/rugby.html>

U.S. tornado climatology. (17, May 2013). Retrieved from <http://www.ncdc.noaa.gov/oa/climate/severeweather/tornadoes.html>

UV radiation: Burden of disease by country. (2013, September 4). Retrieved from <http://apps.who.int/gho/data/node.main.165?lang=en>