How likely is it that the next person you meet walking down a street has more than the average number of legs?

Statistics

Class Notes

Chapter 3: Measures of Center, Dispersion, and Position (Sections 3.1, 3.2, 3.4, 3.5)

Measures of Central Tendency: Mean, Median, and Mode (Section 3.1)

Let's say we have a data set, like household incomes, ages, scores on an exam, or heights of giraffes on the Serengeti. How can we summarize them so we get the bigger picture?

We will see three ways to measure the "center" of this data. They are the **mean** (usually what is called the average but *not* always), the **median** (the middle number of the data), and the **mode** (the most common value).

Definition: The **arithmetic mean** of a variable is computed by adding all the values in the data set and dividing by how many numbers you had.

Because we usually have a population and a sample to concern ourselves with, we need two different symbols for the *population* mean (μ, pronounced "mew") and the *sample* mean ($\bar{x}$, pronounced "x bar").

μ

Which is considered a parameter and which is a statistic?

μ (from pop.) $\bar{x}$ (from sample)

Let's get a bit technical with this definition. We say we add up the values and divide by how many numbers we have, right? In math speak, for μ, that looks like

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\Sigma x_i}{N}$$

The Σ (Greek letter sigma) is shorthand for "add". The subscripts of "i" simply mean the 1st, 2nd, 3rd, etc. values in the list. Recall, we say there are N obervations in the population.

When we see this for $\bar{x}$, we will use n for the sample size. That will look like

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\Sigma x_i}{n}$$

We will always round to one more decimal place than that in the raw data.

To find a mean, the data must be quantitative. Imagine trying to average the responses to the question "What is your favorite movie?" You can't.

expl 1: Let's try this out. The following data represent the travel times to work (in minutes) for *all* seven employees of a start-up web development company.

23, 36, 23, 18, 5, 26, 43

a.) Find the mean of these numbers. Should you label it μ or $\bar{x}$?

$\mu \approx 25$ minutes

b.) Let's say we sampled from this population and got the four numbers 23, 23, 5, and 43. Find the mean of these numbers. Should you label it μ or $\bar{x}$?

$\bar{x} = 23.5$ minutes

Definition: The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.

expl 2: Line up the data values from example 1 in increasing order and find the middle value. Label it as M.

~~5~~ ~~18~~ ~~23~~ 23 ~~26~~ ~~36~~ ~~43~~

$M = 23$ minutes

Think of the median as that on a highway, right down the middle.

When you are told that the **"average"** value is such-and-such, what does that mean? Sometimes this refers to the mean and sometimes this refers to the median. Often, more digging is required to see which is intended.

Like the mean, the data must be quantitative to find its median. Imagine trying to find the median of responses to the question "What is your favorite movie?"

expl 3: Let's change this example up a bit. What if an eighth employee joins this web development company? Find the median now.

23, 36, 23, 18, 5, 26, 43, 33

~~5~~ ~~18~~ ~~23~~ 23 26 ~~33~~ ~~36~~ ~~43~~

?

There is no middle number. So, what do you think you should do?

$$M = \frac{23+26}{2} = 24.5 \text{ minutes}$$

Instructions for TI Calculators:

expl 4: A company pays its employees the following salaries.

$25,000	$26,000	$26,000	$27,000	$28,000
$30,000	$30,000	$35,000	$36,000	$200,000

a.) Find both the mean and median of this data. Do this on the calculator. Here's how..

Enter the data values in column **L1** in the **STAT** editor. We do this by pressing the **STAT** button and then selecting **EDIT > 1: Edit...** from the menu. If necessary, clear out any data in **L1** by arrowing up to the column heading and pressing **CLEAR**. When you arrow back down, any data should be gone. Enter the values of the salaries in **L1,** pressing **ENTER** after each one.

Then press the **STAT** button again. But this time, arrow over to select **CALC > 1: 1-Var Stats.** That will put this expression on the home screen. Press **ENTER** and the calculator will fill with many statistics. (Some newer calculators will have an intermediate screen, where you need to select **L1** for **List** and clear out any entry in the **FreqList:** row. Arrow down and select **Calculate.**)

Look for $\bar{x}$ and record it here. (The calculator will call this $\bar{x}$ even if you know the data is from a population.) Give it a dollar sign and comma.

$\bar{x} = \$46,300$

Arrow down and you will see "Med=" which is the median. Record it here, with a dollar sign and comma.

Med = $29,000

expl 4b.) If you wanted to stress the company's great salaries to prospective employees, which "average" would you provide? Why?

Lets display the mean ($46,300) cause it's higher.

4c.) Why do you think the mean and median are so different?

($46,300 — $29,000

~~25000~~ 26000 26000 27000 28000 30000 30000 35000 36000 200000

When we find mean, the sum includes that large $200,000 and so is driven up.
But the median does not take the $200,000 value into account — It's just another number.

4d.) Give a likely explanation for the outlier salary of $200,000.

It's the advertising manager who just got a bonus (and probably owns a Porsche)

Definition: Resistant: A numerical summary of data is said to be **resistant** if extreme values (very large or small) relative to the data do *not* affect its value substantially.

Considering the data above, would you say the mean or the median is resistant? Why?

The median is resistant.

The Distribution of a Variable and Means versus Medians:

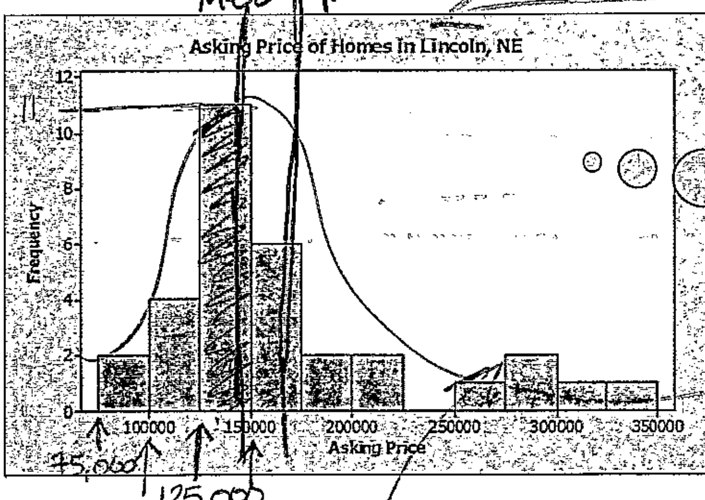
Let's see how the mean and median are affected by a variable's distribution.

expl 5: The data below represent the asking price of homes (in dollars) for sale in Lincoln, NE.

Asking Prices ~~of Homes in Lincoln,~~ Nebraska (dollars)			
79,995	128,950	149,900	189,900
99,899	130,950	151,350	203,950
105,200	131,800	154,900	217,500
111,000	132,300	159,900	260,000
120,000	134,950	163,300	284,900
121,700	135,500	165,000	299,900
125,950	138,500	174,850	309,900
126,900	147,500	180,000	349,900

The mean is $168,320.
The median is $148,700.

Med | Mean



Here is the histogram of this data. It uses a class width of $25,000. Which classes hold most of the data?

$125,000 – $150,000

How would you describe the distribution? In other words, is it symmetric, skewed right, or skewed left?

tail on right.

Since the mean is less resistant than the median, it is pulled up by the large amounts seen on the right of the histogram. Draw vertical lines on the histogram to mark the mean and median.

The mean is the point where the histogram is perfectly balanced.

0, 0, 0.5, 1, 1, 1.5, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2

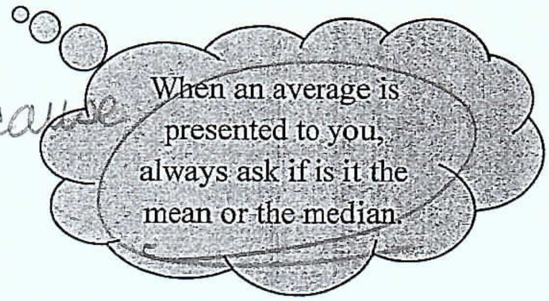
median

expl 6: Let's return to a riddle asked at the very beginning of these Notes. How likely is it that the next person you meet walking down a street has more than the average number of legs? It is common to imagine the "average" to be two legs. Would that number be the median or the mean? In that case, it would be *very unlikely* to come upon a person with more than this "average" of two legs.

median = 2 very unlikely

But if we use the mean for "average", what would you guess its value to be. Make an educated guess and write an approximate number. What, now, is the likelihood of coming upon someone with more legs than average? (almost certain)

The mean would be slightly less than 2 because the data is not all 2s. (but most are 2s).



Definition: The **mode** of a variable is the most frequent observation that occurs in the data set.

A set of data can have no mode, one mode, two modes (**bimodal**) or more than two modes (**multimodal**). *If no observation occurs more than once*, we say the data have **no mode.**

Modes, unlike means and medians, can be found for qualitative data. You could find the mode of responses to the question "What is your favorite movie?"

expl 7: Mr. Kramer gets a yearly evaluation from his students. Using a scale of [strongly agree, agree, neutral, disagree, strongly disagree] students were asked which most fits their level of agreement to the statement "The teacher is fair." The replies are listed below. Make a frequency chart and determine the mode.

strongly disagree	disagree	neutral	strongly agree
agree	strongly disagree	disagree	disagree
agree	strongly disagree	strongly agree	strongly agree
agree	disagree	strongly agree	strongly agree

Outcomes	Count (Frequency)
Strongly disagree	3
disagree	4
neutral	1
agree	3
strongly agree	5

The mode is "Strongly agree"

Instructions for STATCRUNCH:

Within MSL problems, you will see a little icon that looks like overlapping rectangles next to the data. Click on it and select "Open in StatCrunch". This will open StatCrunch and import the data. Alternatively, if you have your own data to enter, open StatCrunch from the left-hand MSL menu and make your way to the spreadsheet. Enter the data in column 1 and label it if you want.

Select **Stats > Summary Stats > Columns**. You will need to tell it where the data is ("Select column(s)" at top). By default, it will calculate lots of stuff including stuff we have not covered yet. You can select more to display under "Statistics". If you just need mean, median, and mode, select "Mean" and scroll down to Control-click "Median" and "Mode". You will see those selected items appear to the right of the selection list. Press the "Compute!" button and it will output a little window with the results.

Measures of Dispersion: Range, Standard Deviation, and Variance (Section 3.2)

Once we know the mean of a set of data, we might be interested in knowing how close the actual values are to that mean. Are the values spread out or close together and all gathered around the mean? We have a few ways to describe what we will call **dispersion.**

All of these measures require that the data be **quantitative.**

The simplest (and quickest) is the range.

Definition: The **range, *R*,** of a variable is the *difference* between the largest data value and the smallest data value. That is,

Range = R = Largest Data Value − Smallest Data Value

expl 8: Let's try this out. The following data represent the travel times to work (in minutes) for all seven employees of a start-up web development company. Find the range of these numbers.

23, 36, 23, 18, 5, 26, 43

R = range = 43 − 5 = 38 minutes

Another very useful measure is the standard deviation. Its definition below is a little daunting but the standard deviation can be thought of as the average distance each value is from the mean.

Definition: The **population standard deviation** of a variable is the square root of the sum of squared deviations about the population mean divided by the number of observations in the population, N. That is, it is the square root of the mean of the squared deviations about the population mean.

σ

The population standard deviation is symbolically represented by σ (lowercase Greek sigma).

Wow, that's a mouthful. Here is the formula.

$$\sigma = \sqrt{\frac{(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2}{N}}$$

$$= \sqrt{\frac{\Sigma(x_i-\mu)^2}{N}}$$

We take each value and find its distance to the mean. We then square those distances, add them up, and divide by N.

where $x_1, x_2, \ldots, x_N$ are the N observations in the population and μ is the population mean.

σ^2 = variance

That gives us **variance** which we will see later. Square root it and we get the **standard deviation**.

That is well and good, but you may also see an equivalent (computational) formula for the **population standard deviation** that is sometimes used. It follows.

$$\sigma = \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{N}}{N}}$$

We do not use this unless we do the calculation by hand (sometimes). The calculator will output the bits for you.

Now, the above standard deviation concerned data gotten from an entire population. However, often we have sample data. Here, we see *similar but slightly different formulas* for the sample standard deviation.

Definition: The **sample standard deviation, *s*,** of a variable is the square root of the sum of squared deviations about the sample mean divided by $n-1$, where n is the sample size.

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

$$= \sqrt{\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n-1}}$$

We do almost the same as if it was population data, *but we divide by one less than the sample size.*

where $x_1, x_2, \ldots, x_n$ are the n observations in the sample and $\bar{x}$ is the sample mean.

Notice how we use s to denote the standard deviation for a *sample* and σ for that of a *population.*

s^2 = variance

Ch 3

The computational formula follows.

$$s = \sqrt{\frac{\Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n}}{n-1}}$$

Is it resistant?:

Since the range is found by subtracting the minimum value from the maximum value, it is affected by extreme values. So, we say the range is *not* resistant.

The standard deviation uses all of the values in its calculation. Therefore, it is also affected by extreme values, so it is *not* considered resistant either.

expl 9: Complete the table to find the standard deviation of this sample data set.

Data Set	$(x_i - \bar{x})^2$
12	$(12-19.3333)^2 \approx 53.78$
16	$(16-19.3333)^2 \approx 11.11$
18	$(18-19.3333)^2 \approx 1.78$
20	$(20-19.3333)^2 \approx 0.44$
24	$(24-19.3333)^2 \approx 21.78$
26	$(26-19.3333)^2 \approx 44.44$
What is $n-1$?	total $= \sum(x_i - \bar{x})^2 \approx 133.33$
$n-1 = 6-1$ $= 5$	$s^2 = \frac{\sum(x_i-\bar{x})^2}{n-1} \approx 26.67$
	$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \approx \sqrt{26.67}$

$n=6$

The mean is 19.3333.

Notice how this squared difference gets larger as the value gets farther from the mean.

Variance

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

≈ 5.2 units

☆ On average, each data value is 5.2 units from the mean (of 19.3333).

Same as mean median

Doing this on the calculator is the same as the process we used for finding the mean and median.

Enter the data values in column **L1** in the **STAT** editor. We do this by pressing the **STAT** button and then selecting **EDIT > 1: Edit…** from the menu. If necessary, clear out any data in **L1** by arrowing up to the column heading and pressing **CLEAR**. When you arrow back down, any data should be gone. Enter the values of the salaries in **L1,** pressing **ENTER** after each one.

Then press the **STAT** button again. But this time, arrow over to select **CALC > 1: 1-Var Stats.** That will put this expression on the home screen. Press **ENTER** and the calculator will fill with many statistics. (Some newer calculators will have an intermediate screen, where you need to select **L1** for **List** and clear out any entry in the **FreqList:** row. Arrow down and select **Calculate**.)

$Sx \approx 5.16$

Look for Sx and record it here. (You will also see σx which is the population standard deviation. Again, the calculator does *not* know if the data is from a population or sample. You must decide which standard deviation to record.)

Instructions for STATCRUNCH:

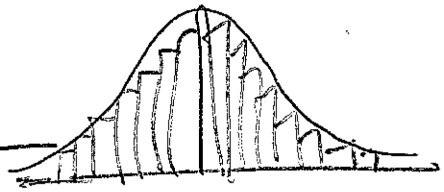
Within MSL problems, you will see a little icon that looks like overlapping rectangles next to the data. Click on it and select "Open in StatCrunch". This will open StatCrunch and import the data. Select **Stats** > **Summary Stats** > **Columns**. You will need to tell it where the data is ("Select column(s)" at top). By default, it will calculate lots of stuff including *sample* standard deviation and variance, mean and median, and range. You can select more to display under "Statistics". If you know the data is from a *population*, "unadjusted" (abbreviated "unadj.") variance and standard deviation (abbreviated "std. dev.") is what you need.

Definition: The **variance** of a variable is the square of the standard deviation. The **population variance** is σ^2 and the **sample variance** is s^2.

Notice the table in the last example has us calculate the variance on the way to the standard deviation. The units of the variance are squared units (for instance, if the variable is in feet, the variance would be in square feet). That makes interpreting the variance a little more challenging. We will see the variance later in inferential statistics.

3.2 **Worksheet: Comparison of two data sets with the same mean:**

We will investigate two data sets with the same mean. One data set is more spread out than the other. How do you think their standard deviations will be related?



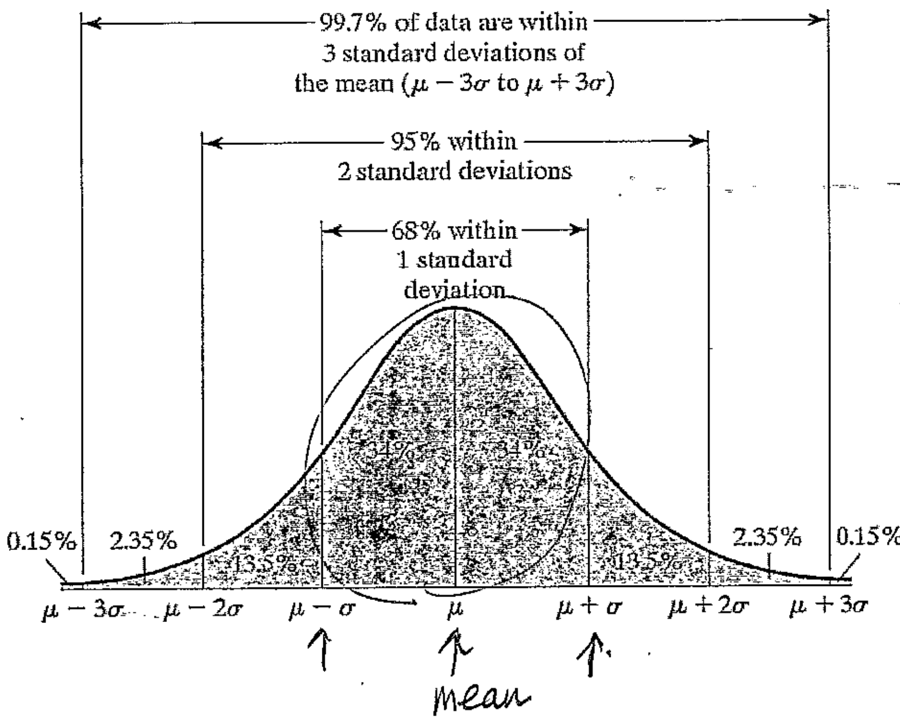
The Empirical Rule:

If a distribution is roughly bell shaped, then

- Approximately 68% of the data will lie within 1 standard deviation of the mean. That is, approximately 68% of the data lie between $\mu - 1\sigma$ and $\mu + 1\sigma$.
- Approximately 95% of the data will lie within 2 standard deviations of the mean. That is, approximately 95% of the data lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean. That is, approximately 99.7% of the data lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

Note: We also use the Empirical Rule based on sample data with $\bar{x}$ and s used in place of μ and σ.

Here is a picture that illustrates the Empirical Rule.



Notice how μ is positioned in the exact middle, showing the center of the data. Let's see this in action to get a better understanding.

population data

expl 10: The following data represent the serum HDL cholesterol of the 54 female patients of a family doctor.

Circled 45.7 to 69.1 (part g)

41	48	43	38	35	37	44	44	44
62	75	77	58	82	39	85	55	54
67	69	69	70	65	72	74	74	74
60	60	60	61	62	63	64	64	64
54	54	55	56	56	56	57	58	59
45	47	47	48	48	50	52	52	53

a) Compute the *population* mean and standard deviation. Use a calculator.

b) Draw a histogram to verify the data is bell-shaped.

c.) Draw a quick sketch of a bell-shaped curve, labeling the mean in the middle. Then mark the various standard deviations (plus or minus 1, 2, and 3) from the mean using a reasonable scale. You must calculate these values and label them on the horizontal axis.

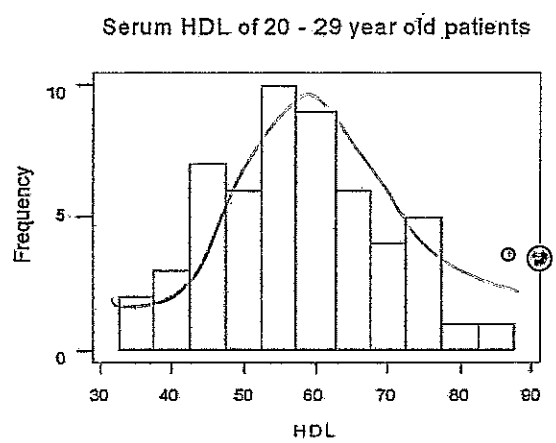
We complete example 10 here.

a) Compute the *population* mean and standard deviation. Use a calculator.

If you want, you can verify this information. The calculator tells us that $\mu = 57.4$ and $\sigma = 11.7$.

Parts *a* and *b* are done for us.

b) Draw a histogram to verify the data is bell-shaped.



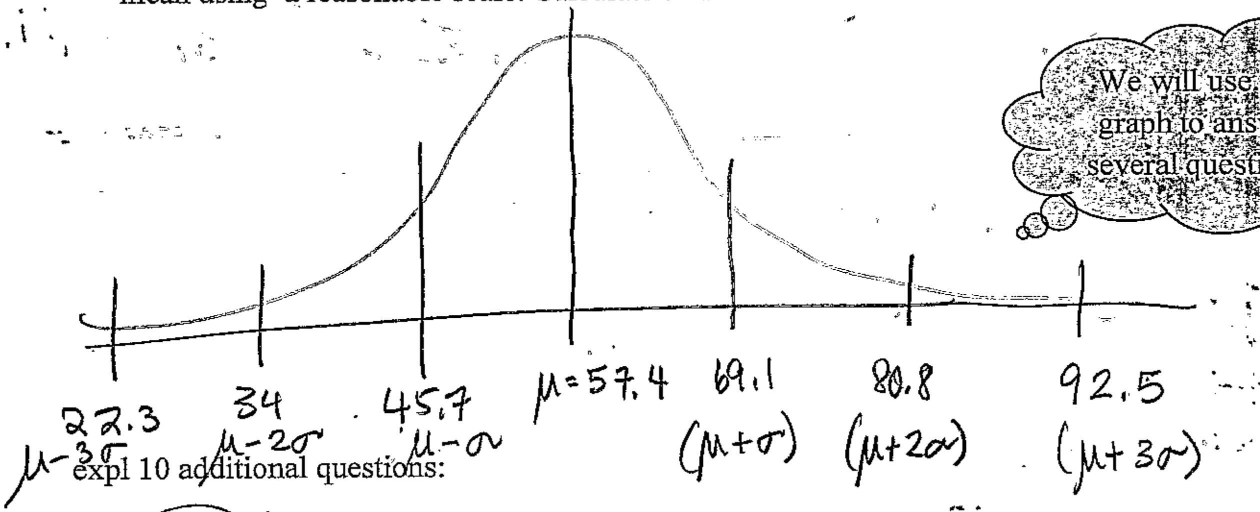
Does the histogram look roughly bell-shaped?

yes

expl 10: (continued)

c.) Draw a quick sketch of a bell-shaped curve for this data, labeling the mean ($\mu = 57.4$) in the middle. Then mark the various standard deviations (plus or minus 1, 2, and 3) from the mean using a reasonable scale. Calculate these values and label them on the horizontal axis.

We will use this graph to answer several questions.



expl 10 additional questions:

d.) What is the percentage of all patients that have serum HDL within 1, 2, and 3 standard deviations of the mean according to the Empirical Rule?

within 1 standard deviation from mean: 68%

within 2 standard deviations from mean: 95%

within 3 standard deviations from mean: 99.7%

(from pg 11)

e.) Determine the percentage of all patients that have serum HDL between 22.3 and 92.5 according to the Empirical Rule. Refer to the gaph you produced on the previous page and your answer to part *d*.

99.7 %

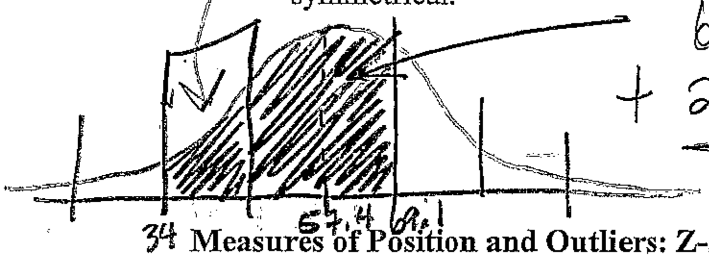
f.) Determine the percentage of all patients that have serum HDL between 45.7 and 69.1 according to the Empirical Rule. Refer to the gaph you produced on the previous page and your answer to part *d*.

68 %

g.) Determine the actual percentage of patients that have serum HDL between 45.7 and 69.1. Do this by looking at the *actual data*. Compare this to the answer in part *f.*

actual data $= \frac{35}{54} \approx 64.8\%$ — The Empirical Rule got us pretty darn close.

h.) Determine the percentage of all patients that have serum HDL between 34 and 69.1 according to the Empirical Rule. Refer to the gaph you produced on the previous page and your answer to part *d.* You will also use the fact that this bell-shaped graph is symmetrical.



68% for $\mu-\sigma$ to $\mu+\sigma$

$+ \; 27/2$ for this piece

$68 + 13.5 = \boxed{81.5\%}$

Measures of Position and Outliers: Z-scores, Percentiles, Quartiles, and Interquartile Range (Section 3.4)

We have talked about the center and the spread of a set of data values. Now, we look at how the values are positioned about each other.

Using z-scores to compare data:

Consider the Los Angeles Angels, a baseball team in the American League. During the 2014 season, the Angels scored 773 runs. During that same season, the Colorado Rockies, who play in the National League, scored 755 runs.

It seems as though the Angels outperformed the Rockies. But did they really?

The National League and American League have an important difference. The National League makes their pitchers hit (and they are rather notorious for doing it poorly). The American League, on the other hand, uses designated hitters to replace the pitchers when it is their time at bat.

The National League averages 640 runs with a standard deviation of 55.9 runs. The American League averages 677.4 runs with a standard deviation of 51.7 runs.

The idea of **z-scores** will allow us to compare each team, not directly with each other, but with their respective leagues, and then against each other.

Definition: The z-**score** represents the distance that a data value is from the mean in terms of the number of standard deviations. We find it by subtracting the mean from the data value and dividing this result by the standard deviation. There is both a population z-score and a sample z-score.

$$z = \frac{x-\mu}{\sigma} \quad \text{or} \quad z = \frac{x-\bar{x}}{s}$$

The z-score has *no* units (like feet or seconds). They have a mean of 0 and a standard deviation of 1.

expl 11: Use the population information given for each league to find the z-scores for the Angels and the Rockies. Compare them.

Angels (2014): 773 runs
The American League averages 677.4 runs with a standard deviation of 51.7 runs.

Rockies (2014): 755 runs
The National League averages 640 runs with a standard deviation of 55.9 runs.

$$z = \frac{x-\mu}{\sigma}$$

$$z = \frac{x-\mu}{\sigma} = \frac{773-677.4}{51.7}$$

$z \approx 1.85$ (They did 1.85 "standard devs" better than their league's mean.)

$$z = \frac{x-\mu}{\sigma} = \frac{755-640}{55.9}$$

$z \approx 2.06$

(They did 2.06 "stand. devs" better than their league's mean.)

So, who did better? Explain.

Rockies

They performed better in comparison to their league.

expl 12: What would cause a z-score to be negative versus being positive?

If they had less runs than the average amt of runs.

(generally, if $x < \mu$)

man woman

expl 13: Bob and Maggie ran a marathon. The mean time to complete the marathon for men was 242 minutes (with a standard deviation of 57 minutes). The mean time for women was 273 minutes (with a standard deviation of 52 minutes). Bob's z-score is -0.51 and Maggie's z-score is -0.62. Who did better?

Consider what a better score means?

Maggie did better cause her z-score puts her time more under the women's mean than Bob was under the men's mean. (Less time is better)

Definition: Percentiles: The ***k*th percentile**, denoted, P_k, of a data set is a value such that k percent of the observations are *less than or equal to* the value.

You may have gotten SAT or ACT results back and learned that you scored in the 74th percentile. What does that mean, with respect to the other test-takers?

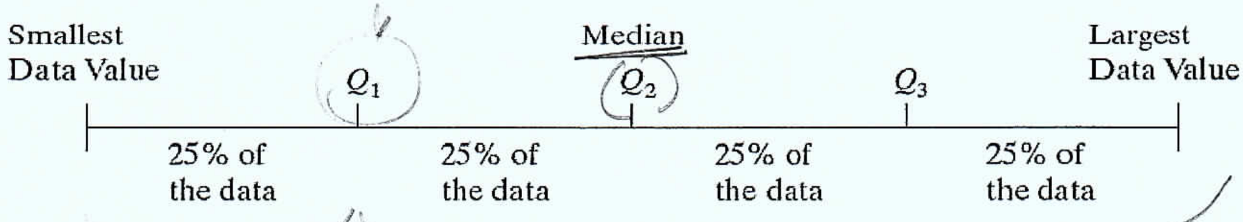
Your score is such that 74 % of all other test takers are less than or equal to your score.

Percentiles break the data up into 100 parts, essentially. We could divide the data up into just four parts. This is called **quartiles**.

Definition: Quartiles divide data sets into fourths, or four equal parts.

- The 1st quartile, denoted Q_1, divides the bottom 25% of the data from the top 75%. Therefore, the 1st quartile is equivalent to the 25th percentile.
- The 2nd quartile, denoted Q_2, divides the bottom 50% of the data from the top 50% of the data. The 2nd quartile is equivalent to the 50th percentile, which is, in fact, the **median**.
- The 3rd quartile, denoted Q_3, divides the bottom 75% of the data from the top 25% of the data. The 3rd quartile is equivalent to the 75th percentile.

Here is a nice picture of how these quartiles break up the data.



bottom 25%

top 75%

middle 50% of values

IQR = Q3 − Q1

expl 14: Find the median (Q_2) and then Q_1 and Q_3 for the following data.

This is a stem-and-leaf plot for 17 states detailing the percentage of people who are aged 65 or older. In this graphic, the entry 10 | 6 means 10.6 %. (Source: Statistical Abstract of US, 1995)

4	6
5	
6	
7	
8	
9	
10	1 1 6
11	6
12	1 6 7 7 8 9
13	4 9
14	2 8
15	4
16	
17	
18	4

First, convert the plot to a listing of the values in order.

Next, find the median (Q_2) of the data.

Once you have divided the data into two halves, find the median of each half. These will be Q_1 and Q_3.

#17 outlier = 4.6%

M = 12.7 %

4.6 | 10.1 10.1 10.6 11.6 12.1 12.6 12.7 12.7 12.8 12.9 13.4 13.9 14.2 14.8 15.4 18.4

lower fence (#17) 6.675

$$Q_1 = \frac{10.6 + 11.6}{2}$$

$$Q_1 = 11.1\ \%$$

$$Q_3 = \frac{13.9 + 14.2}{2}$$

$$Q_3 = 14.05\ \%$$

upper fence (#17) 18.475

These quartiles are part of the calculator output when you perform **1:1-Var Stats** in the **STAT > CALC** menu.

Definition: The **interquartile range, IQR**, is the range of the middle 50% of the observations in a data set. That is, the IQR is the difference between the third and first quartiles and is found using the formula

$$IQR = Q_3 - Q_1$$

Return to page ~~3~~ 16 and label this on the graphic.

Quartiles and IQR are resistant. They are *not* much affected by outliers.

expl 15: Find the interquartile range of the data in the previous example.

$$IQR = Q_3 - Q_1$$
$$= 14.05 - 11.1 = 2.95\ \%$$

Use the same units as the data.

expl 16: Let's use the results from the last two examples to answer some questions.

a.) What percentage of the data has a value that is less than or equal to 11.1? Write your answer in a sentence that explains the full meaning of the data.

$Q_1 = 11.1\ \%$

So, 25% of the states in our sample have a senior population (65 and older) that is less than or equal to 11.1%.

b.) What percentage of the data has a value that is greater than 12.7? Write your answer in a sentence that explains the full meaning of the data.

So, 50% of the states in our sample have a senior population that is greater than 12.7%.

c.) Between which two values do the middle 50% of the data lie? Write your answer in a sentence that explains the full meaning of the data.

The IQR is a good measure of the dispersion of the data.

Concerning the states' senior populations, the middle 50% of the data lies between 11.1% and 14.05% (of the population are senior citizens).

For the remainder of the semester, when asked to find the distribution of a data set, we will describe its shape (symmetric, skewed left, or skewed right), its center (mean or median), and its spread (standard deviation or interquartile range). Use medians and interquartile ranges if you have skewed data.

Checking Data for Outliers:

To this point, we have described outliers to be values that appear way larger or smaller than the other values. However, there is a more defined statistical method we see now.

Step 1 Determine the first and third quartiles of the data.

Step 2 Compute the interquartile range.

Step 3 Determine the fences. **Fences** serve as cutoff points for determining outliers.

Lower Fence = $Q_1 - 1.5(\text{IQR})$

Upper Fence = $Q_3 + 1.5(\text{IQR})$

Step 4 If a data value is less than the lower fence or greater than the upper fence, it is considered an **outlier**.

expl 17: Refer back to the data in the example concerning the percentage of states with populations aged 65 or older.

a.) Rewrite the first and third quartiles as well as the interquartile range.

$Q_1 = 11.1$ $\quad$ $IQR = Q_3 - Q_1 = 2.95$

$Q_3 = 14.05$

b.) Follow step 3 above to find the lower and upper fences. Follow step 4 above to determine if the data set has any outliers. What are the outlier(s)?

lower fence $= Q_1 - 1.5IQR = 11.1 - 1.5(2.95) = 6.675$

upper fence $= Q_3 + 1.5IQR = 14.05 + 1.5(2.95) = 18.475$

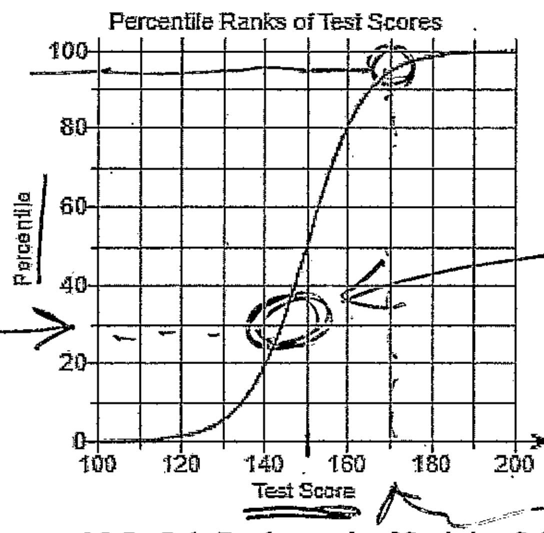
The outlier is Alaska at 4.6% (data came from Data Displays)

Go back to pg 17

Ogive Graphs:

Ogive (pronounced oh-jive) graphs show off percentiles in an interesting way. The vertical axis in an ogive is the **cumulative relative frequency** and can also be interpreted as a **percentile**. Let's practice interpreting this one which displays the percentiles of a fictional standardized test's scores.

(a) ? 95



(b) Test score = ?

(source: MyStatLab, Fundamentals of Statistics, Sullivan)

expl 18: Use the above ogive to answer the following questions.

a.) Estimate and interpret (complete sentence below) the percentile rank of the test score 170.

A test score of 170 equates to the 95 th percentile. That means that 95 % of the scores would be less than or equal to 170.

b.) What test score falls at the 30th percentile? Do your best to estimate it from the graph. Complete the sentence below.

So, 30% of the scores are less than or equal to the (estimated) score of 145.

The Five-Number Summary and Boxplots (Section 3.5)

We saw quartiles in the previous section. Five-number summaries use those to define a useful way to explore the data. A boxplot is a graphic that shows these numbers off.

Definition: The **five-number summary** of a set of data consists of the smallest data value (or minimum), Q_1, the median, Q_3, and the largest data value (or maximum). We organize the five-number summary in increasing order.

Minimum $\quad Q_1 \quad$ Median $\quad Q_3 \quad$ Maximum

Recall the interquartile range is resistant. Along with the minimum and maximum values, we get a good picture of the data.

Worksheet: Effect of an Outlier:
This worksheet will give us practice in finding means, medians, standard deviations, quartiles, and five-number summaries. We will see how an outlier affects some of these measures more than others. Recall, this is the idea of resistance.

expl 19: Every six months, the United States Federal Reserve Board conducts a survey of credit card plans in the U.S. The following data are the interest rates charged by ten credit card issuers randomly selected for the July 2005 survey. Determine the five-number summary of the data.

Institution	Rate	
Pulaski Bank and Trust Company	6.5%	← Min = 6.5%
Bank of Louisiana	9.9%	
Rainier Pacific Savings Bank	12.0%	← Q_1 = 12.0%
Infibank	13.0%	
United Bank, Inc.	13.3%	
First National Bank of The Mid-Cities	13.9%	← Median = $\frac{13.3+13.9}{2}$ = 13.6% Med
Lafayette Ambassador Bank	14.3%	
Wells Fargo Bank NA	14.4%	← Q_3 = 14.4%
Firstbank of Colorado	14.4%	
Bar Harbor Bank and Trust Company	14.5%	← Max = 14.5%

Definition: Boxplot: A **boxplot** (or **box-and-whiskers plot**) is a graphic that shows the five-number summary, with any outliers clearly marked. You will see a scale along the bottom to give meaning to the numbers.

The steps of creating a boxplot are below.

Step 1 Determine the lower and upper fences (where $IQR = Q_3 - Q_1$):

Lower Fence $= Q_1 - 1.5(IQR)$

Upper Fence $= Q_3 + 1.5(IQR)$

Step 2 Draw a number line long enough to include the maximum and minimum values. Above the number line, draw vertical lines at Q_1, M, and Q_3. Use these vertical lines to draw a rectangular box.

Step 3 Temporarily label the lower and upper fences.

Step 4 Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. (Basically, do not draw all the way to outliers. They are dealt with next.) These lines are called **whiskers**.

Step 5 Any data values less than the lower fence or greater than the upper fence are outliers and are marked with an asterisk (*). Erase the upper and lower fences (from step 3).

expl 20: Draw a boxplot for the interest rates data in the last example. Follow the steps below.

a.) (Step 1) Determine the lower and upper fences (where $IQR = Q_3 - Q_1$): $IQR = Q_3 - Q_1 = 14.4 - 12.0 = 2.4$

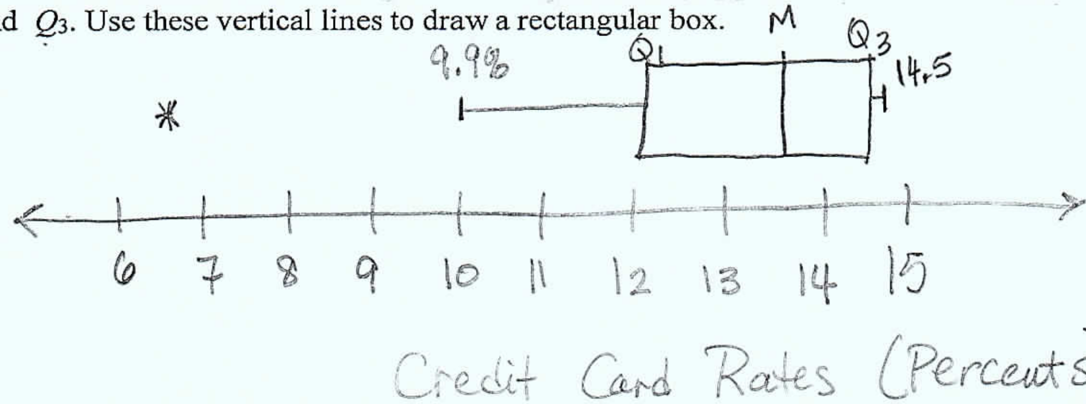
Lower Fence $= Q_1 - 1.5(IQR)$ $= 12.0 - 1.5(2.4) = 8.4\%$

Upper Fence $= Q_3 + 1.5(IQR)$ $= 14.4 + 1.5(2.4) = 18\%$

Give yourself a scale from 6 to 18.

b.) (Step 2) Draw a number line long enough to include the maximum and minimum values, leaving enough space above it to draw the plot. Above the number line, draw vertical lines at Q_1, M, and Q_3. Use these vertical lines to draw a rectangular box.

Ted's Survey of Credit Cards



expl 20: (continued) (Step 3) Temporarily label the lower and upper fences.

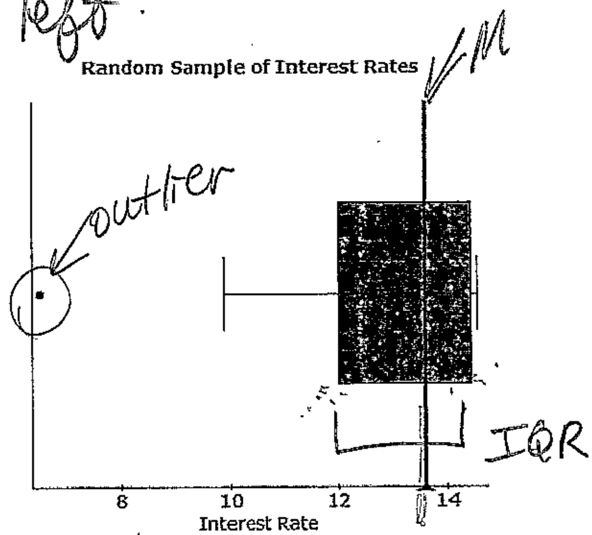
(Step 4) Draw a line from Q_1 to the smallest data value that is larger than the lower fence. Draw a line from Q_3 to the largest data value that is smaller than the upper fence. (Basically, do not draw all the way to outliers. They are dealt with next.) These lines are called **whiskers.**

(Step 5) Any data values less than the lower fence or greater than the upper fence are outliers and are marked with an asterisk (*). You can now erase the upper and lower fences.

You should also label the horizontal axis and title your plot.

c.) Do you think this distribution is symmetric, skewed left, or skewed right? Explain.

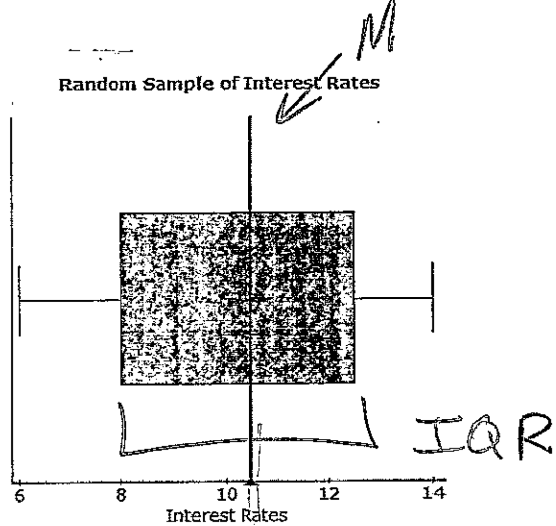
Bigger whisker to left



expl 21: Here we see the boxplot as generated by StatCrunch for the interest rate data.

Let's suppose another sample was taken of ten banks and whose interest rates are summarized in the second boxplot. The medians are marked with a red, vertical line.

Let's explore the comparisons we can make. Notice the scales are roughly lined up.



a.) How would you compare the spread of the two data sets as measured by the interquartile range?

2nd graph has more Spread (width of box)

b.) How would you compare the centers of the two data sets?

2nd set has lower median

c.) How would you compare the overall ranges (maximum minus minimum values) of the two data sets?

about the same

d.) Overall, which data set would you say has higher interest rates?

First one

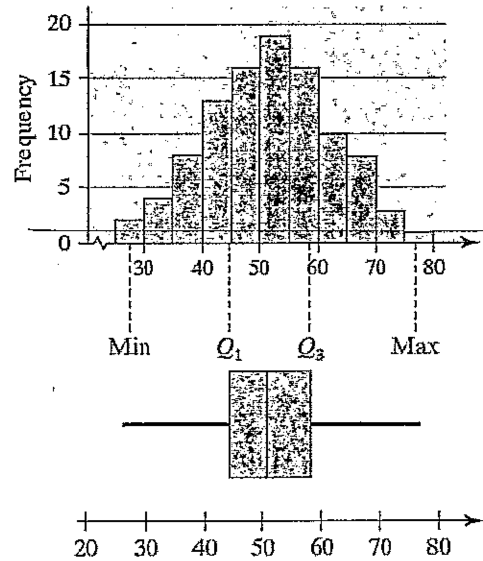
Distributions and the Boxplot:

A boxplot shows the distribution of a data set nicely. You can see similarities between a data set's histogram and its boxplot. Consider the pictures below.

Here, we see a symmetric distribution. The boxplot is nice and centered.

The median is centered in the box.

The whiskers are of equal lengths.

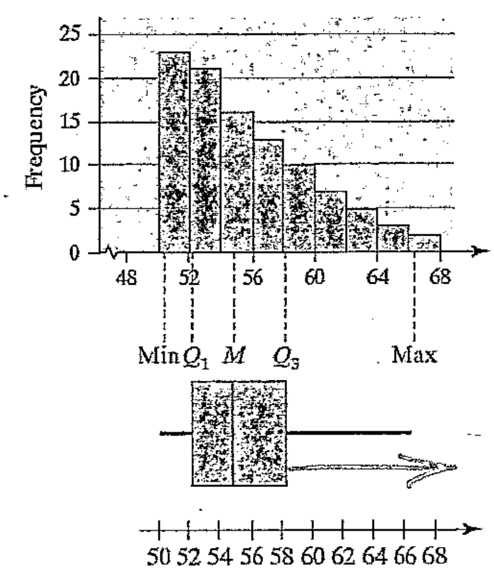


(b) Symmetric

Here, we see a distribution that is skewed right.

Notice how the median is off-center, just left of the box's center.

The right whisker mimics the tail we see on the histogram. It is longer than the left whisker.

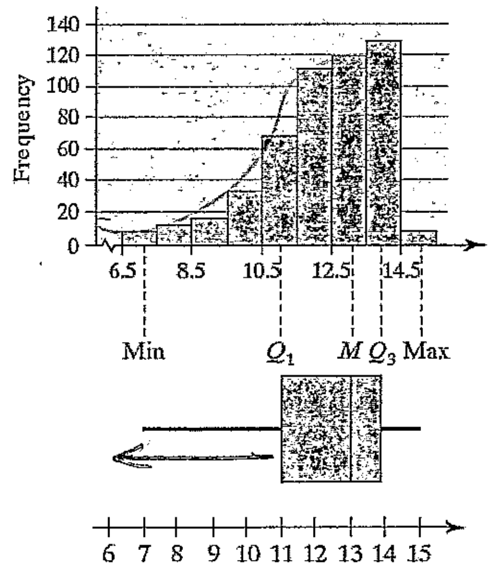


(a) Skewed right

Here, we see the opposite of the last picture. We see a distribution that is skewed left.

Notice how the median is off-center, just right of the box's center.

The left whisker mimics the tail we see on the histogram. It is longer than the right whisker.



(c) Skewed left

Making a Boxplot with Technology:
In Stat Crunch in MyStatLab (left-hand menu in MSL), follow these steps.

1. Enter the raw data if needed. Name the column variable.

2. Select **Graph** and highlight **Boxplot**.

3. Click on the variable whose boxplot you want to draw under "Select Column(s)". Check the boxes "Use fences to identify outliers" and "Draw boxes horizontally". Enter a label for the x-axis. (You'll have to scroll down.) Enter a title and click **Compute!**

Making boxplots with other technology such as your calculator is described in the book.