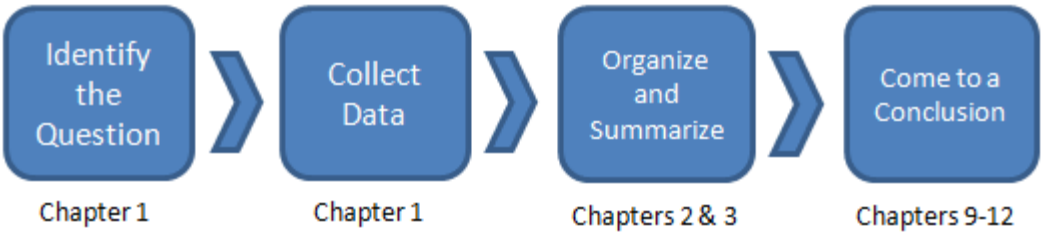
Chapter 3: Numerically Summarizing Data

Let's again review the process of statistics we introduced in Section 1.1:



In Chapter 1, we focused on how to collect data. In Chapter 2, we talked about how to organize and summarize data using tables in graphs. In this chapter, we'll introduce various ways to summarize data numerically, along with one new graphical representation - the box plot. In general, we have three ways to summarize the distribution of a random variable - shape, center, and spread. Shape was discussed back in Section 2.2, but the center and spread will be introduced here in Section 3.1 and Section 3.2, respectively. Section 3.3 talks about estimating measures of center and spread from grouped data.

In Section 3.4, we talk about summarizing information about an individual observation in relation to the rest of the sample/population. (We call these measures of *position*.)

And finally, in Section 3.5, we introduce a new graphical representation of data called the box plot. We'll be using this plot frequently throughout the semester.

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

Section 3.1: Measures of Central Tendency

Objectives

By the end of this lesson, you will be able to...

1. determine the arithmetic mean of a variable from raw data
2. determine the median of a variable from raw data
3. explain what it means for a statistics to be resistant
4. determine the mode of a variable from raw data
5. use the mean and median to help identify the shape of a distribution

It's often very helpful to get a sense of what a "typical" individual might be in a population. This is what we mean when we say we're looking at measures of "center" or "central tendency".

Before we get into specifics, we need to clarify whether we're talking about typical individual from a population or from a sample.

A **parameter** is a descriptive measure of a population.
A **statistic** is a descriptive measure of a sample.

Arithmetic Mean

You already know the arithmetic mean, though maybe not by name. It's more commonly referred to as the "average". It's calculated just by finding the some of the values and dividing by the number of observations. As mentioned above, we'll have two different means - one for the population and one if we're talking about a sample.

The **population arithmetic mean**, μ (pronounced "mew"), is computed using all the individuals in the population. The **sample arithmetic mean**, $\bar{x}$ (pronounced "x-bar"), is computed using sample data.

$$\text{population arithmetic mean: } \mu = \frac{x_1 + x + \ldots + x_N}{N} = \frac{\sum x_i}{N}$$

$$\text{sample arithmetic mean: } \bar{x} = \frac{x_1 + x + \ldots + x_n}{n} = \frac{\sum x_i}{n}$$

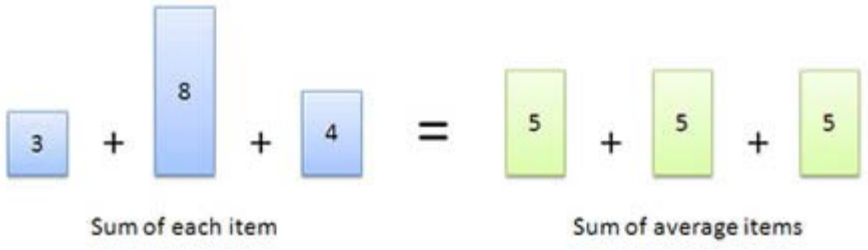
This is pretty formulaic, but the concept should be relatively familiar.

To give another explanation, I'm going to reference one of my favorite web sites, BetterExplained. The author, Kalid Azad, presents topics in a non-traditional way, and I feel it's much more accessible and easier to understand than traditional texts. Here's what Kalid writes about the arithmetic mean:

The Arithmetic Mean

The arithmetic mean is the most common type of average:

$$average = \frac{sum}{number}$$



Let's say you weigh 150 lbs, and are in an elevator with a 100lb kid and 350lb walrus. What's the average weight?

The real question is "If you replaced this merry group with 3 identical people and want the same load in the elevator, what should each clone weigh?"

In this case, we'd swap in three people weighing 200 lbs each [(150 + 100 + 350)/3], and nobody would be the wiser.

Pros:

- It works well for lists that are simply combined (added) together.
- Easy to calculate: just add and divide.
- It's intuitive — it's the number "in the middle", pulled up by large values and brought down by smaller ones.

Cons:

- The average can be skewed by outliers — it doesn't deal well with wildly varying samples. The average of 100, 200 and -300 is 0, which is misleading.

The arithmetic mean works great 80% of the time; many quantities are added together. Unfortunately, there's always those 20% of situations where the average doesn't quite fit.

Source: BetterExplained, Kalid Azad
Article: How to Analyze Data Using the Average
Used with permission.

So, let's try an example.

Example 1

Suppose we record the exam scores from a sample of six students from a class of 30 (see table below).

student	exam score
Joseph	62
Alicia	83

Kendra	77
Cheryl	92
Adrian	89
Brian	75

Find the sample mean, along with its appropriate symbol.

[reveal answer]

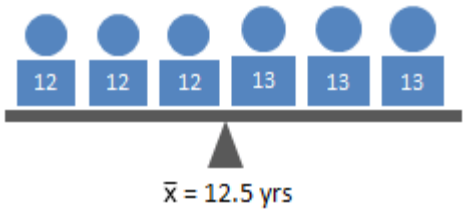
When necessary, **round the mean to one more digit than the original data**. i.e. If the data are whole numbers, you should round the mean to the tenths place (as in the previous example). If the data are already to the tenths place, you should round to the hundredths place.

You might also consider watching this video regarding rounding (in Quicktime or iPod format).

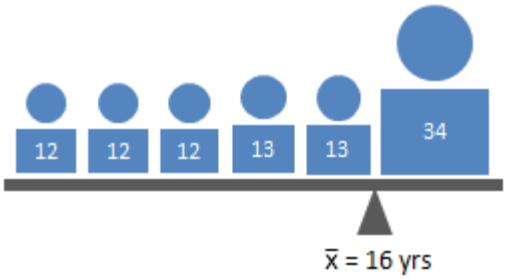
Example 2

One point that should be emphasized again is the effect of outliers on the arithmetic mean. Because it adds all the values together, the arithmetic mean can be skewed by extremely large or extremely small values.

A helpful way to illustrate this is to think of the mean as the center of gravity - like the balance point. Suppose we consider the ages of the six Jackson cousins, Hudson, Abella, Amelia, Jillian, Katelyn, and Jessica. The figure below represents their ages and the corresponding sample mean. (*Sample*, in this case, because this isn't *all* of the Jackson cousins.)



If we replace Jessica with her father, who is 34 years old, we get something like this:



You can see very clearly here the effect of including the dad. 16 years old does not really represent the "middle" value.

Technology

Here's a quick overview of the formulas for finding arithmetic mean in StatCrunch.

1. Select **Stat** > **Summary Stat** > **Columns**.
2. Select the variable you want to summarize (e.g., "Heights")--leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated.
5. Click "Calculate" and another window with these numbers calculated will pop up.

 You can also visit the video page for links to see videos in either Quicktime or iPod format.

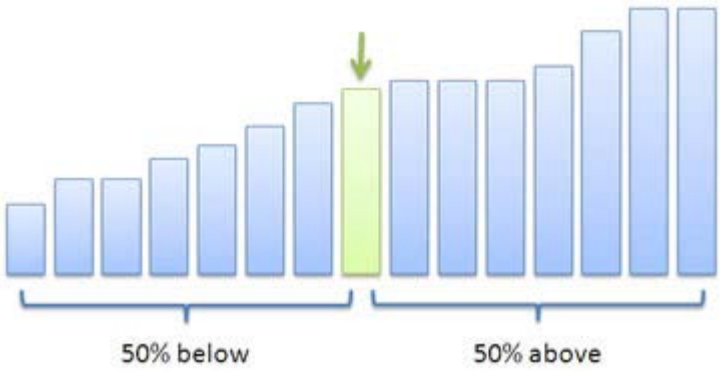
Median

As we mentioned at the end of the previous page, we need another measure of center when the data include outliers. The most common choice is called the **median**.

> The **median** of a variable is the value that lies in the middle of the data when arranged in ascending order. That is, half the data are below the median and half the data are above the median. We use M to represent the median.

Like the previous topic, I really appreciate how Kalid Azad explained the median on his web site, BetterExplained. Here's what he wrote:

> ### Median
>
> The median is “the item in the middle”. But doesn’t the average (arithmetic mean) imply the same thing? What gives?
>
> Humor me for a second: what’s the “middle” of these numbers?
>
> 1, 2, 3, 4, 100
>
> Well, 3 is the middle of the list. And although the average (22) is somewhere in the “middle”, 22 doesn’t really represent the distribution. We’re more likely to get a number closer to 3 than to 22. The average has been pulled up by 100, an outlier.
>
> The median solves this problem by taking the **number in the middle of a sorted list**. If there’s two middle numbers (even number of items), just take their average. Outliers like 100 only tug the median along one item in the sorted list, instead of making a drastic change: the median of 1 2 3 4 is 2.5.



Pros:

- Handles outliers well — often the most accurate representation of a group
- Splits data into two groups, each with the same number of items

Cons:

- Can be harder to calculate: you need to sort the list first
- Not as well-known; when you say “median”, people may think you mean “average”

Some jokes run along the lines of “Half of all drivers are below average. Scary, isn’t it?”. But really, in your head, you know they should be saying “half of all drivers are below *median*“.

Figures like housing prices and incomes are often given in terms of the median, since we want an idea of **the middle of the pack**. Bill Gates earning a few billion extra one year might bump up the average income, but it isn’t relevant to how a regular person’s wage changed. We aren’t interested in “adding” incomes or house prices together — we just want to find the middle one.

Again, the type of average to use depends on how the data is used.

Source: BetterExplained, Kalid Azad
Article: How to Analyze Data Using the Average
Used with permission.

Example 3

Let's again consider the exam scores from a sample of six students from a class of 30 (see table below).

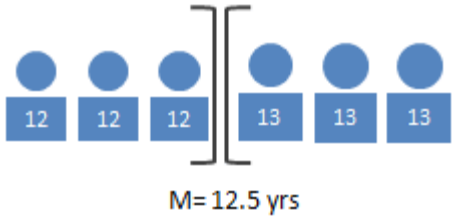
student	exam score
Joseph	62
Alicia	83
Kendra	77
Cheryl	92
Adrian	89
Brian	75

Find the sample median.

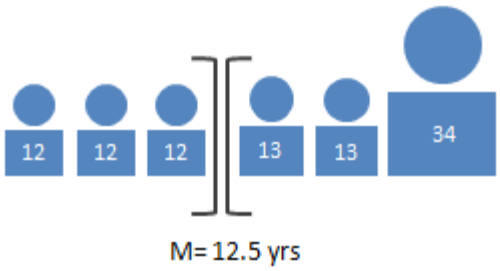
[reveal answer]

Example 4

To illustrate how the median deals with outliers, let's again consider the ages of the six Jackson cousins. The figure below represents their ages and the corresponding sample median.



If we replace Jessica with her father, who is 34 years old, we get something like this:



You can immediately see the benefit of using the median - it is not affected by the age of Jessica's father.

Technology

Here's a quick overview of the formulas for finding median in StatCrunch.

1. Select **Stat** > **Summary Stat** > **Columns**.
2. Select the variable you want to summarize (e.g., "Heights")--leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated.
5. Click "Calculate" and another window with these numbers calculated will pop up.


You can also visit the video page for links to see videos in either Quicktime or iPod format.

Mode

Often, we just want to know what "most" people think on an issue. We don't call it that, but we're really looking

at is called the **mode**.

> The **mode** of a variable is the most frequent observation of the variable.

Look at any poll from the Pew Research Center. Any time an article discusses the "most common" or "most popular" choice, it's talking about the mode.

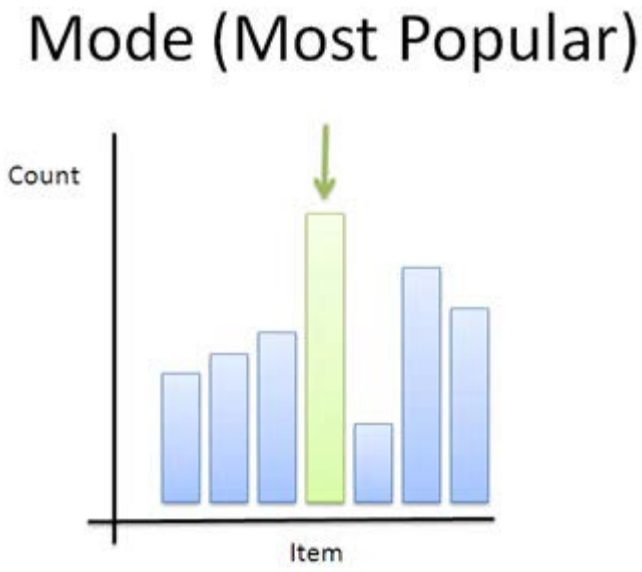
As with the previous two measures of central tendency, I like Kalid Azad's explanation of the mode on his web site, BetterExplained. Here's what he wrote:

Mode

The mode sounds strange, but it just means **take a vote**. And sometimes a vote, not a calculation, is the best way to **get a representative sample** of what people want.

Let's say you're throwing a party and need to pick a day (1 is Monday and 7 is Sunday). The "best" day would be the option that satisfies the most people: an average may not make sense. ("Bob likes Friday and Alice likes Sunday? Saturday it is!").

Similarly, colors, movie preferences and much more can be measured with numbers. But again, the ideal choice may be the mode, not the average: the "average" color or "average" movie could be... unsatisfactory (Rambo meets Pride and Prejudice).



Pros:

- Works well for exclusive voting situations (this choice or that one; no compromise)
- Gives a choice that the most people wanted (whereas the average can give a choice that nobody wanted).
- Simple to understand

Cons:

- Requires more effort to compute (have to tally up the votes)
- "Winner takes all" — there's no middle path

The term "mode" isn't that common, but now you know what button to look for when playing around with your favorite statistics program.

Source: BetterExplained, Kalid Azad

Article: How to Analyze Data Using the Average
Used with permission.

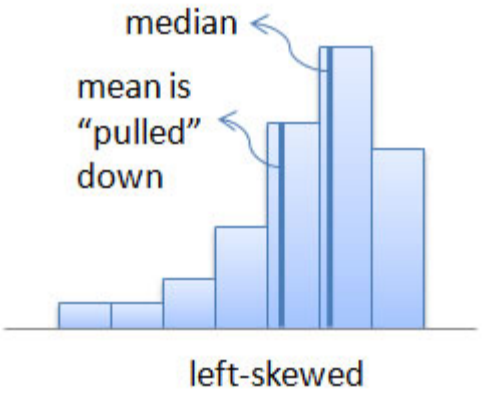
Technology

To see how to find the mode using technology, open the appropriate video from the list below. These videos include all measures of center included in this section, plus other descriptive statistics.

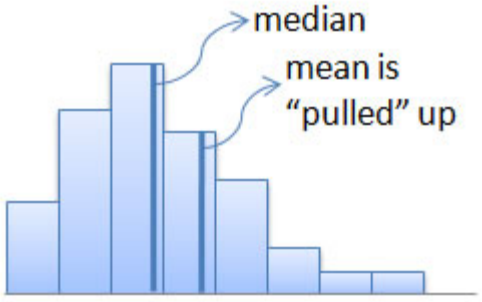
Visit the video page for links to see videos in either Quicktime or iPod format.

Using the Mean and Median to Identify the Distribution Shape

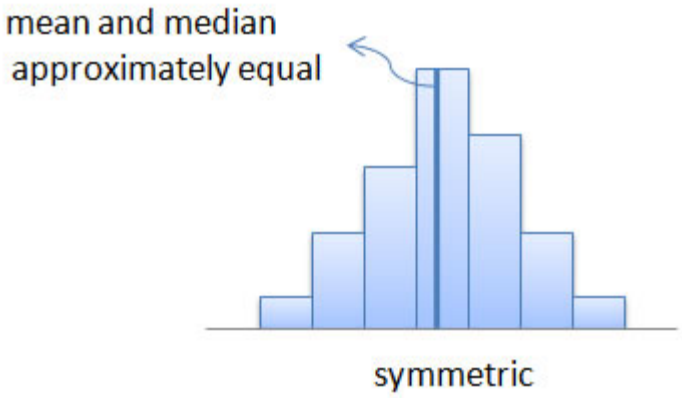
In Section 2.2, we talked about different ways to describe the distribution shape. With these new measures of center, we can now use the mean and median to get an idea of the distribution shape as well.



left-skewed



right-skewed



<< previous section | next section >>

1 2 3 4 5 6 7 8 9 10 11 12 13

Section 3.2: Measures of Dispersion

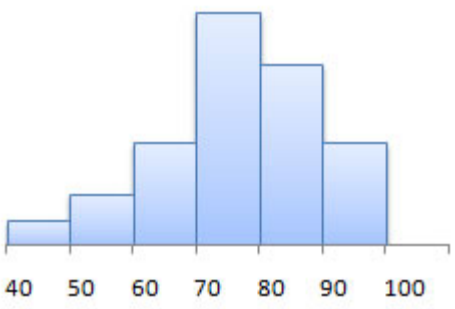
Objectives

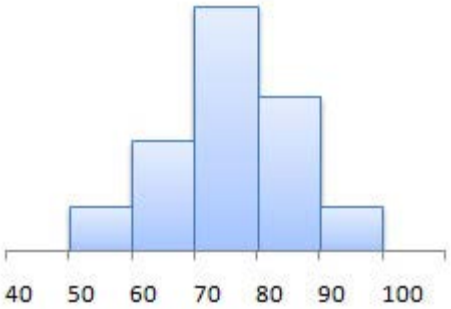
By the end of this lesson, you will be able to...

1. compute the range of a variable from raw data
2. compute the variance of a variable from raw data
3. compute the standard deviation of a variable from raw data
4. use the empirical Rule to describe data that are bell-shaped

Consider the following two sets of exam scores:

48 57 58 65 68 69 71 73 73	58 58 61 62 63 64 66 71 71
74 75 77 78 78 78 79 80 85	72 73 75 76 76 77 77 78 78
87 88 89 89 89 95 96 97 99	79 81 82 82 84 85 88 90 91





Often, we want to compare two data sets and look for differences. In this case, however, the sample mean for the first set is 78.3, with a median of 78, while the second set has a sample mean of 78.7, also with a median of 78.

We can see that the measures of center are not enough to distinguish between the two sets, so we'll need to somehow compare their "dispersion", or spread. The first statistic we'll learn about to help do that is called the **range**.

Range

The **range**, **R**, of a variable is the difference between the largest and smallest data values.

Example 1

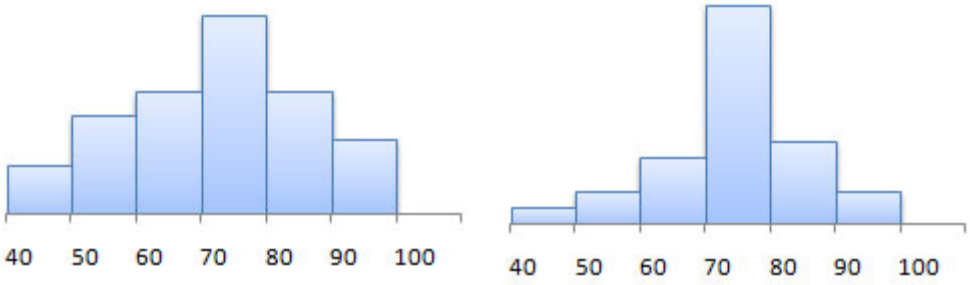
Looking at our data from above, we see that the range for the first set of exam scores is 99-48 = 51, while the range for the second set is 91-58 = 33. As we can see from the histograms, the second set of exam scores

has *less dispersion*.

Unfortunately, the range isn't always enough to distinguish between two sets of data, which we'll see on the next page.

Let's look at another two sets of exam scores.

48 49 52 55 57 58 62 64 65	48 55 57 61 64 65 68 71 71
66 67 72 72 73 75 78 78 78	72 73 73 74 75 78 78 79 79
79 82 84 86 88 89 93 94 95	79 79 82 84 85 88 89 92 95

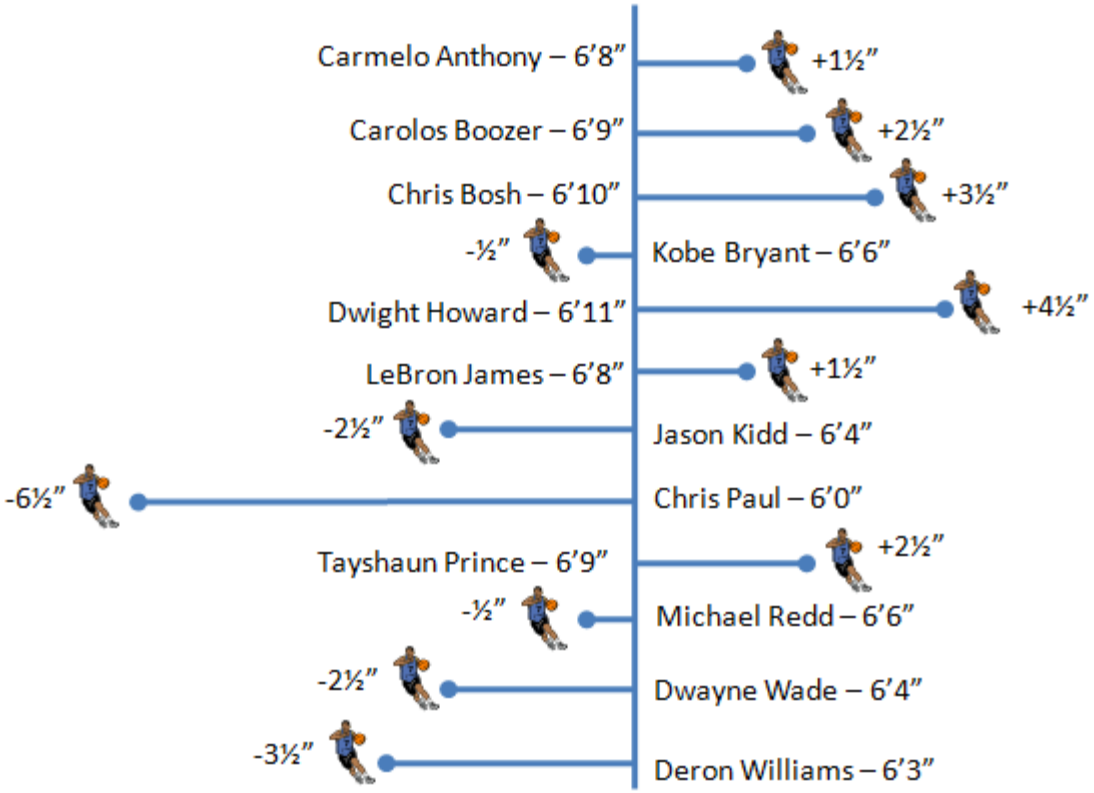


In this case, we can see that the range is R = 95-48 = 47 for both sets, but they clearly don't have the same dispersion. The second set is much more condensed, with the bulk of the scores in the C range - 70-79.

Variance

To describe the dispersion in cases like these, we'll need to somehow describe how far a "typical" observation is from the mean, rather than looking at the extreme values.

An obvious choice would be to just look at the average distance from the mean. The figure below shows the heights of the 2008 US Men's Olympic Basketball team and each player's corresponding difference from the mean. (Source: USA Basketball)



$\mu = 6'\ 6\frac{1}{2}"$

If we try to take the average difference from the mean, we have a problem - we get 0!

$$\frac{1.5 + 2.5 + 3.5 - 0.5 + 4.5 + 1.5 - 2.5 - 6.5 + 2.5 - 0.5 - 2.5 - 3.5}{12} = 0$$

Why is this? Well, it's because the mean acts as a balancing point, as we talked about earlier. In fact, this will *always* happen - the average difference from the mean will equal zero.

So what do we do? The obvious choice would be to take the average *distance* from the mean. That is, take the absolute value of each of the differences. It's a good thought, but anyone who's taken calculus will tell you that absolute values can be pretty difficult to work with, so instead, we square them.

This creates a new measure of dispersion, called the **variance**.

The **population variance**, σ^2, of a variable is the sum of the squared deviations about the population mean divided by N.

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N} = \frac{\sum(x_i - \mu)^2}{N}$$

This is pretty complex, but we'll have technology to do most of the work for us.

Like the mean, we also have a sample version of this calculation, but unlike the mean, it's actually different.

The **sample variance**, s^2, is computed by determining the sum of the squared deviations about the sample mean and dividing the result by n-1.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{n - 1} = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

The first thing most students ask when they see this (I did, too) is "Why n-1 instead of just n?" It's a good question, and a difficult to answer in plain English. The key is to look at the purpose of using the sample variance (or any sample statistics, for that matter). That purpose is to get an estimate for the true population variance.

Unless we have data for the entire population, our estimate will likely be incorrect. If we look at the average of *all possible* sample variances, though, that average should be the same as the population variance we're trying to estimate. In other words, we'll be wrong most of the time, but the average of all of our attempts will be correct.

The thing is, if we divide by N in the sample variance formula above, our estimate will, on average, be too low. (We can actually prove this mathematically, but it's pretty heady stuff. It's usually not covered until a graduate course in probability and statistics.) We call an estimate like this **biased**, since it consistently under-estimates the parameter it's trying to predict.

Interestingly enough, dividing by n-1 makes the estimate **unbiased**. (This can also be proven mathematically.) So it may seem like an odd thing to do, but there's very solid mathematical reasoning behind it.

If you'd like more information on this, you can read a more thorough analysis in your text on pages 141-142.

Example 2

Let's refer back to the heights of the players on the US Men's Olympic basketball team, and let's treat this as a sample of all the basketball players in the US.

Player	Height	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
Carmelo Anthony	6'8"	1.5	2.25
Carlos Boozer	6'9"	2.5	6.25
Chris Bosh	6'10"	3.5	12.25
Kobe Bryant	6'6"	-0.5	0.25
Dwight Howard	6'11"	4.5	20.25
LeBron James	6'8"	1.5	2.25
Jason Kidd	6'4"	-2.5	6.25
Chris Paul	6'0"	-6.5	42.25
Tayshaun Prince	6'9"	2.5	6.25
Michael Redd	6'6"	-0.5	0.25
Dwayne Wade	6'4"	-2.5	6.25
Deron Williams	6'3"	-3.5	12.25
			117

So the sample variance is then:

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{117}{12 - 1} \approx 10.6$$

You may notice that I rounded the variance to the tenths place.

Typically, we round the variance to one more digit than the data

When necessary, **round the variance to one more digit than the original data**. i.e. If the data

are whole numbers, you should round the variance to the tenths place (as in the previous example). If the data are already to the tenths place, you should round to the hundredths place.

Technology

Here's a quick overview of the formulas for finding variance in StatCrunch.

1. Select **Stat** > **Summary Stat** > **Columns**.
2. Select the variable you want to summarize (e.g., "Heights")--leave everything else as is for now.
3. Click **Next**.
4. Deselect any statistics that you do not want calculated.
5. Click **Calculate** and another window with these numbers calculated will pop up.

 You can also visit the video page for links to see videos in either Quicktime or iPod format.

One major problem with the variance is that the units don't really make sense. Take the previous example about the heights of the players on the 2008 US Men's Olympic Basketball team. If we look at the units for that variance, it's 10.64 inches *squared*. What does that have to do with the dispersion of the data? The data are in inches, not inches squared!

Standard Deviation

To remedy that, we need another measure of dispersion, called the **standard deviation**.

The **population standard deviation**, **σ**, is obtained by taking the square root of the population variance.

$$\sigma = \sqrt{\sigma^2}$$

The **sample standard deviation**, **s**, is obtained by taking the square root of the sample variance.

$$s = \sqrt{s^2}$$

So referring again to our previous example, the sample standard deviation is $\approx \sqrt{10.6} \approx 3.3$ inches. So we could then say that the "typical" player is about 3.3 inches different from the average height of the team.

Now *that* makes more sense!

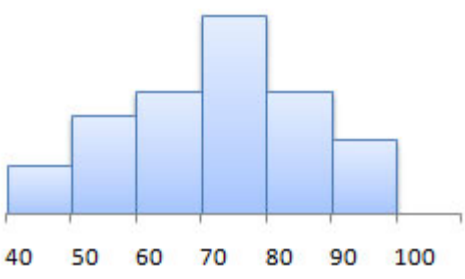
When necessary, **round the standard deviation to one more digit than the original data**. i.e. If the data are whole numbers, you should round the standard deviation to the tenths place (as in the previous example). If the data are already to the tenths place, you should round to the thousandths place.

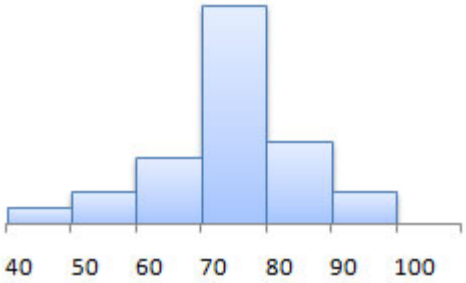
You might also consider watching this video regarding rounding (in Quicktime or iPod format).

What Does It Mean?

So what can we tell from the standard deviation? Let's go back to those two sets of exam scores. Which one do you think has a higher standard deviation?

48 49 52 55 57 58 62 64 65	48 55 57 61 64 65 68 71 71
66 67 72 72 73 75 78 78 78	72 73 73 74 75 78 78 79 79
79 82 84 86 88 89 93 94 95	79 79 82 84 85 88 89 92 95





[reveal answer]

[Why?]

Technology

Here's a quick overview of the formulas for finding standard deviation in StatCrunch.

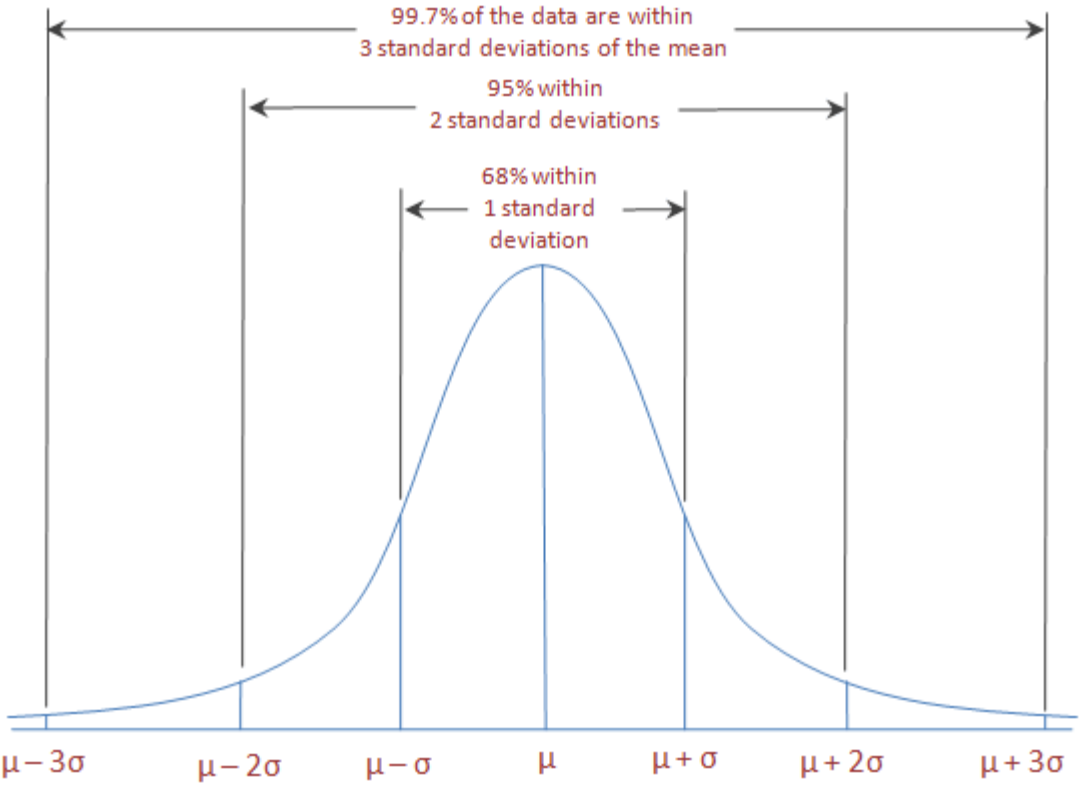
1. Select **Stat** > **Summary Stat** > **Columns**.
2. Select the variable you want to summarize (e.g., "Heights")--leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated.
5. Click "Calculate" and another window with these numbers calculated will pop up.


You can also visit the video page for links to see videos in either Quicktime or iPod format.

One nice benefit of understanding the relationship between the standard deviation and the shape of the distribution is it helps us get a sense of how much of the data should be within a certain number of standard deviations.

In particular, if the distribution is bell-shaped, we can be fairly precise about what percentage of the data should lie within 1, 2, or 3 standard deviations.

The Empirical Rule



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.
- Approximately 95% of the data will lie within 2 standard deviations of the mean.
- Approximately 99.7% of the data will lie within 3 standard deviations of the mean.

How do we know these percentages so accurately? Well, unfortunately we can't explain it until we get to Chapter 7, but because it fits so well in here with standard deviations, you'll just have to accept it on faith at this point!

It does give us interesting information, though. Let's look at an example to illustrate.

Example 3

IQ tests are generally designed to have a mean of 100 and a standard deviation of 15. It's also known that the distribution of EQ scores tends to follow a bell-curve.

With that in mind, approximately what percentage of individuals have IQs between 85 and 115?

[reveal answer]

It's difficult to characterize a "genius" explicitly, but some put it at an IQ of about 145+. About what percent of the population are "geniuses" by this criteria?

[reveal answer]

<< previous section | next section >>

Section 3.3: Measures of Central Tendency and Dispersion from Grouped Data

Objectives

By the end of this lesson, you will be able to...

1. approximate the mean and standard deviation of a variable from grouped data*
2. compute the weighted mean

* You will not be tested on this objective.

Suppose you wanted to estimate the mean and standard deviation for an exam, but all the professor gave you was curve, maybe something like this one:

Exam 1 Scores	
90+	8
80-89	13
70-79	6
60-69	3
50-59	1

Could you do it? How?

Approximating the Mean from Grouped Data

The technique we'll use (which you may have already thought of) is to treat each individual as the midpoint of its class. So instead of 13 scores from 80-89, we'll say that there are 13 85's. (This really works best with continuous data - we should probably use a midpoint of 84 for this example. Can you see why?)

From there, we should be able to approximate both the mean and standard deviation. We just have to remember to weight each observation by the number that are in that category.

Example 1

Using the Exam 1 data from above,

Scores	Freq	Midpoint	
90+	8	95	8*95 = 760
80-89	13	85	13*85 = 1105
70-79	6	75	6*75 = 450
60-69	3	65	3*65 = 195
50-59	1	55	1*55 = 55
	31		2565

So the sample mean is then:

$$\frac{2565}{31} \approx 82.7$$

(Notice again that I'm rounding the mean to one extra decimal place.)

Let's try the sample standard deviation:

Scores	Freq	Mdpt	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	
90+	8	95	12.3	151.29	8*151.29 = 1210.32
80-89	13	85	2.3	5.29	13*5.29 = 68.77
70-79	6	75	-7.7	59.29	6*59.29 = 355.74
60-69	3	65	-17.7	313.29	3*313.29 = 939.87
50-59	1	55	-27.7	767.29	1*767.29 = 767.29
	31				3341.99

The approximate sample standard deviation is then:

$$\sqrt{\frac{3341.99}{31-1}} = \sqrt{111.3997} \approx 10.55$$

Do you ever wonder how your GPA is calculated? This is it!

Weighted Mean

A weighted mean occurs when certain values carry more weight than others. The easiest example is your GPA. An "A" in Statistics counts more than an "C" in Tennis - not because it's more important or carries a higher meaning, but because the 4 credits for Statistics outweigh the 1 credit for Tennis. That's why your GPA will be closer to an "A" than a "C" - the Statistics course counts for more.

Here's how it works:

Each letter grade is assigned a weight. At most schools, this means an A=4, B=3, etc. Some schools do have other point systems, and there are many schools that have partial points with A-, B+, etc.

When calculating your GPA, the point value for each course is weighted by the number of credits. In the quick example above, your GPA for that semester wouldn't be a "B", because the Statistics course was worth 4 credits. The real GPA would be:

$$\text{GPA} = \frac{\text{weighted points}}{\text{total credits}} = \frac{4*4 + 2*1}{4+1} = \frac{18}{5} \approx 3.6$$

Let's try one that's more interesting.

Example 2

Here's a typical course load for a 1st-year student at ECC, along with some typical grades.

Class	Credits	Grade

Statistics	4	B
Chemistry	5	A
Tennis	1	B
English	3	C
Speech	3	B

What is this student's semester GPA?

[reveal answer]

Section 3.4: Measures of Position

Objectives

By the end of this lesson, you will be able to...

1. determine and interpret z-scores
2. determine and interpret percentiles
3. determine and interpret quartiles
4. check a set of data for outliers

In Sections 3.1 and 3.2, we discussed ways to describe a "typical" individual in a population or sample. In this next section, we'll talk about ways to describe an *individual* in relation to the population.

z-Scores

Example 1

It's fairly common for upper-level statistics courses to have both undergraduate and graduate students. Given the exam scores listed below, can you determine which score is better relative to its peers, the undergraduate score of 83 or the graduate score of 88?

undergraduate student scores			
65	89	84	75
52	78	92	80
76	72	**83**	79

graduate student scores			
82	90	95	72
78	**88**	92	89

Actually, to answer this question, we need more information. In particular, we need a new way to describe *relative position*.

The **z-score** represents the number of standard deviations a data value is from the mean.

Population z-Score

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

Sample z-Score

$$z \approx \frac{x - \bar{x}}{s}$$

I can't over-emphasize the importance of the *meaning* behind the z-score. Make a note of this now - you'll be seeing this again later on in the semester - it's *very* important!

Example 1 (continued)

So let's continue with our previous example. The sample mean of the undergraduate scores is 77.1, with a standard deviation of 10.73. That gives a z-score for the undergraduate 83 of:

$$z = \frac{83 - 77.1}{10.73} \approx 0.55$$

With a sample mean of 85.75 and a standard deviation of 7.78, the graduate has a z-score of:

$$z = \frac{88 - 85.75}{7.78} \approx 0.29$$

Since the undergraduate is more than 1/2 of a standard deviation above the mean (z = 0.55), that's a better relative score.

Note: You may have noticed that I went to the hundredths place for these z-scores. That's standard practice.

Key: We use z-scores when we want to compare to individuals from different populations, *relative* to their respective populations.

Percentiles

If you've ever taken a standardized exam like the PSAT, SAT, or ACT, you've seen in the report something about your **percentile**.

The ***k*th percentile**, denoted P_k, of a set of data divides the lower k% of a data set from the upper (100-k)%.

Percentile ranks are used in a variety of fields:

- Special Education - students scoring below a certain percentile on specific tests qualify for services.
- Physicians - doctors usually track a child's weight and height and compare the growth to that of other children of the same age.

Unfortunately, there's no universally accepted way to calculate percentiles. Most software packages and calculators use a method similar to the one below (from your text), but you should be aware of the possibility of others.

Determining the *k*th percentile, denoted P_k

Step 1: Arrange the data in ascending order.
Step 2: Compute an index i using the formula

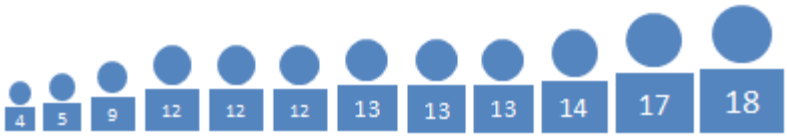
$$i = \left(\frac{k}{100}\right)(n+1)$$

Step 3:

a. If i is an integer, the kth percentile, denoted P_k, is the ith value.
b. If i is not an integer, the kth percentile is the mean of the observations on either side of i.

Example 2

Let's go back to the Jackson cousins we saw in Example 2 in Section 3.1. Suppose this time we add all the cousins, from little Zander at age 4 to Mae, who at age 18 is entering her first year at college.



Use the strategy above to find the 25th percentile by age.

[reveal answer]

Technology

Here's a quick overview of the formulas for finding percentiles in StatCrunch.

1. Select **Stat** > **Summary Stat** > **Columns**.
2. Select the variable you want to summarize (e.g., "Heights")--leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated
5. Enter the percentile you wish to calculate in the "Percentile" box.
6. Click "Calculate" and another window with these numbers calculated will pop up.

Note: Some software like Microsoft Excel interpolates instead of taking a simple average when calculating percentiles, so the results may differ slightly.

Determining the Percentile of a Data Value

The last thing we need to do with percentiles is to figure out the percentile of a particular individual. For example, if your Composite ACT score is a 28, what percentile does that leave you?

As before, there is no universally accepted way to calculate percentiles, but the following (from your text) is very common.

Finding the Percentile that Corresponds to a Data Value

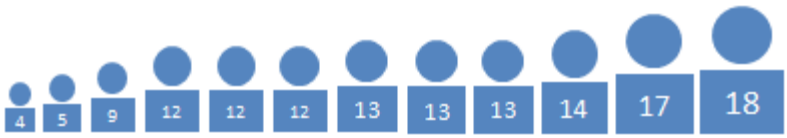
Step 1: Arrange the data in ascending order.
Step 2: Use the following formula to find the percentile of the value, x.

$$\text{percentile of } x = \frac{\text{number of data values less than } x}{n} * 100$$

Round this number to the nearest integer.

Example 3

Consider again the Jackson cousins we looked at in Example 2 above.

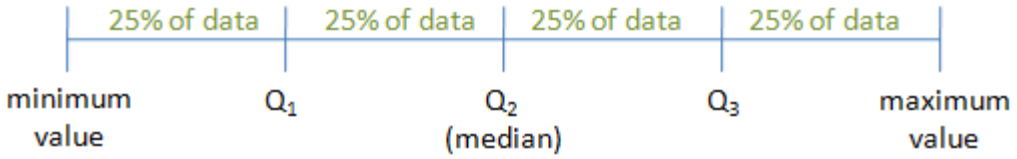


What is the percentile rank of James, the 14-year-old?

[reveal answer]

Quartiles

As the name implies, **quartiles** divide the data into four equal parts. Therefore the first quartile, Q_1, is the 25th percentile, the second quartile, Q_2 is the 50th percentile (or the median), and the third quartile, Q_3, is the 75th percentile.



Example 4

Let's consider one of the sets of hypothetical exam scores we looked at in Section 3.2.

48	57	58	65	68	69	71	73	73
74	75	77	78	78	78	79	80	85
87	88	89	89	89	95	96	97	99

Find the quartiles.

[reveal answer]

Technology

Here's a quick overview of the formulas for finding quartiles in StatCrunch.

1. Select **Stat** > **Summary Stat** > **Columns**.
2. Select the variable you want to summarize (e.g., "Heights")--leave everything else as is for now.
3. Click "Next".
4. Deselect any statistics that you do not want calculated
5. Click "Calculate" and another window with these numbers calculated will pop up.

Note: Some software like Microsoft Excel interpolates instead of taking a simple average when calculating percentiles, so the results may differ slightly.

Checking for Outliers

One good use of quartiles is they give us a sense of what values might be extreme. In Statistics, we call these values **outliers**. There are various ways to check for outliers. Most depend on the distribution and often can only characterize observations as *possible* outliers. A common technique used is the following:

Checking for Outliers by Using Quartiles

Step 1: Determine the first and third quartiles
Step 2: Compute the inter-quartile range: $\mathbf{IQR = Q_3 - Q_1}$
Step 3: Determine the fences.

Lower fence = Q_1 - 1.5(IQR)
Upper fence = Q_3 + 1.5(IQR)

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

Example 5

Let's look at those same exam scores we used in Example 4.

48 57 58 65 68 69 71 73 73
74 75 77 78 78 78 79 80 85
87 88 89 89 89 95 96 97 99

Use the above method to determine if there are any outliers.

[reveal answer]

Section 3.5: The Five-Number Summary and Boxplots

Objectives

By the end of this lesson, you will be able to...

1. compute the five-number summary
2. draw and interpret boxplots

The Five-Number Summary

The **five-number summary** of a set of data consists of the smallest data value, Q_1, the median, Q_3, and the largest value of the data.

Example 1

To illustrate, let's again look at those exam scores from Example 4 in Section 3.4.

48 57 58 65 68 69 71 73 73
74 75 77 78 78 78 79 80 85
87 88 89 89 89 95 96 97 99

Find the five-number summary.

[reveal answer]

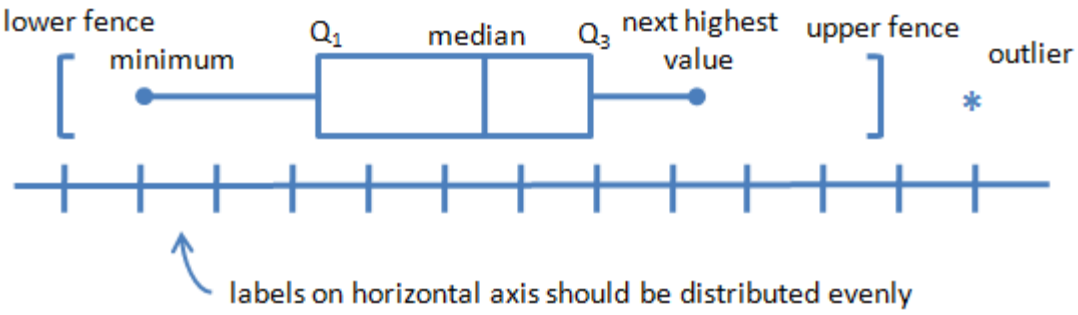
Boxplots

Using the five-number summary and the fences, we can create a new graph called a **boxplot**.

Drawing a Boxplot

Step 1: Determine the five-number summary and the lower and upper fences.
Step 2: Draw a horizontal line and label it with an appropriate scale.
Step 3: Draw vertical lines at Q_1 , M, and Q_3. Enclose these vertical lines in a box.
Step 4: Draw a line from Q1 to the smallest data value that is within the lower fence. Similarly, draw a line from Q3 to the largest value that is within the upper fence.
Step 5: Any values outside the fences are outliers and are marked with an asterisk (*).

A typical boxplot will look something like this:



Example 2

To illustrate, let's again look at those exam scores from Example 4 in Section 3.4.

48 57 58 65 68 69 71 73 73
74 75 77 78 78 78 79 80 85
87 88 89 89 89 95 96 97 99

Take a moment and try to sketch a boxplot of this data set, following the description above.

[reveal answer]

Technology

Here's a quick overview of how to create box plots in StatCrunch.

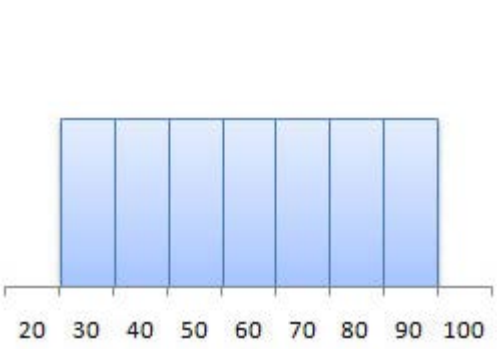
1. Enter or import the data.
2. Select **Graphics** > **Box Plot**.
3. Select the column(s) you want to create a box plot for.
4. Click **Next**.
5. Check "Use fences to identify outliers" and click **Next**.
6. Enter any modifications and click **Next**.
7. Choose a color scheme, if you wish, and click **Create Graph!**
8. You can then choose **Options** > **Copy** to copy the box plot for use elsewhere.


You can also visit the video page for links to see videos in either Quicktime or iPod format.

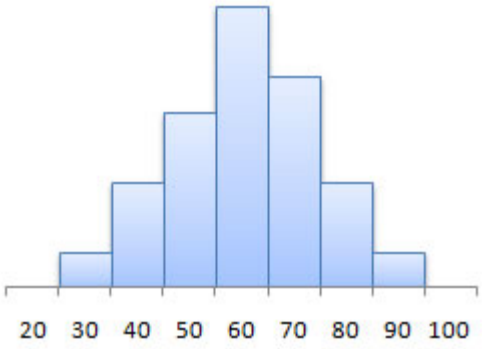
Boxplots and Distribution Shape

The last thing we want to talk about in Chapter 3 is the relationship between the shape of a boxplot and the shape of the distribution.

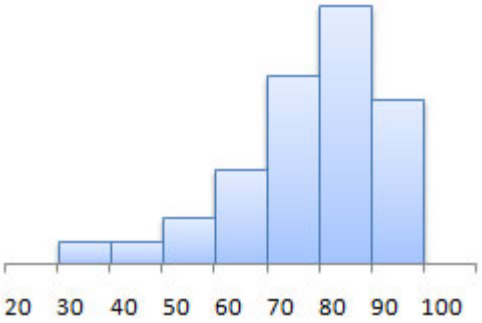
In Section 2.2, we talked about distribution shape, showing the following four standards:



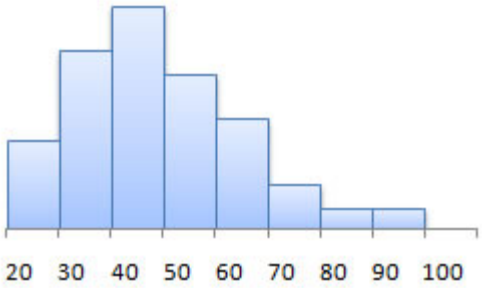
uniform



symmetric (bell-shaped)



left-skewed

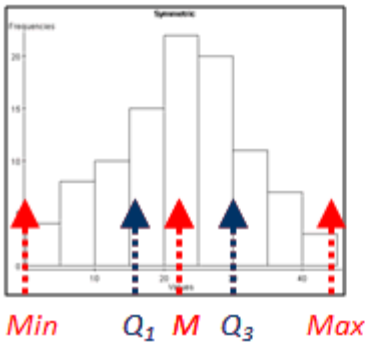


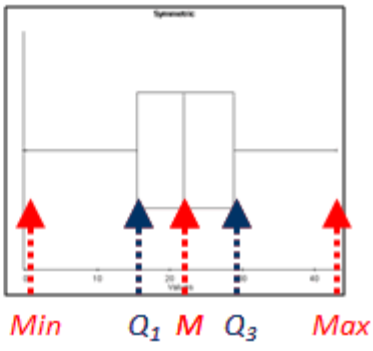
right-skewed

Let's now see how these are related to boxplots. Here's some information from your text:

Symmetric distributions

Distribution	Boxplot
Q_1 is **equally far** from the median as Q_3 is	The median line is in the **center** of the box
The minimum is **equally far** from the median as the maximum is	The left whisker is **equal** in length to the right whisker

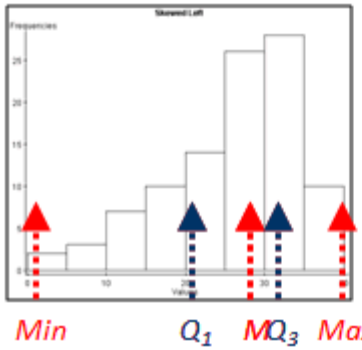


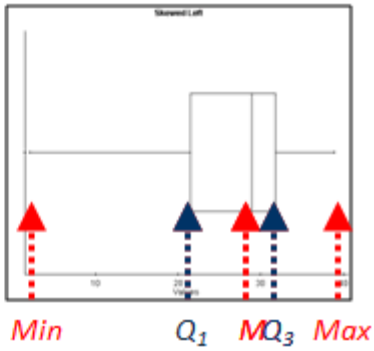


Skewed left distributions

Distribution	Boxplot

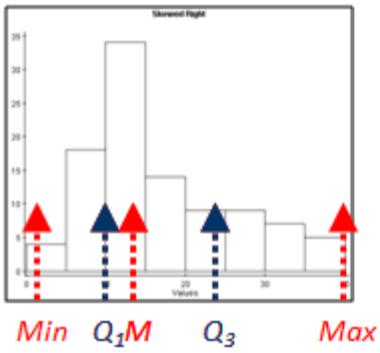
Q_1 is **further** from the median as Q_3 is	The median line is to the **right** of center in the box
The minimum is **further** from the median as the maximum is	The left whisker is **longer** than the right whisker

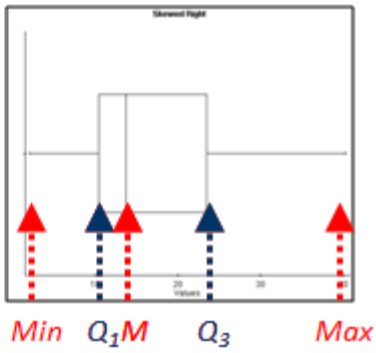




Skewed right distributions

Distribution	Boxplot
Q_1 is **closer** to the median than Q_3 is	The median line is to the **left** of center in the box
The minimum is **closer** to the median as the maximum is	The left whisker is **shorter** than the right whisker





Source: Instructor Resources; Statistics: Informed Decisions Using Data
Author: Michael Sullivan III