

UNIT 4: Does the Designated Hitter Increase Offense in Major League Baseball?

Exploring Numerical Data

In 1973 the American League adopted a rule which allows a designated hitter (DH) to bat in place of the pitcher. This was intended to increase offense, which would increase interest in the games, which would in turn increase revenue for AL teams. In this unit we will use data from the 2008 season to investigate if teams using a DH score more runs than teams with no DH.

Here are the run totals for the 14 American League teams and the 16 National League teams for 2008:

Team	League	Runs
Baltimore Orioles	AL	782
Boston Red Sox	AL	845
Chicago White Sox	AL	811
Cleveland Indians	AL	805
Detroit Tigers	AL	821
Kansas City Royals	AL	691
Los Angeles Angels	AL	765
Minnesota Twins	AL	829
New York Yankees	AL	789
Oakland Athletics	AL	646
Seattle Mariners	AL	671
Tampa Bay Rays	AL	774
Texas Rangers	AL	901
Toronto Blue Jays	AL	714

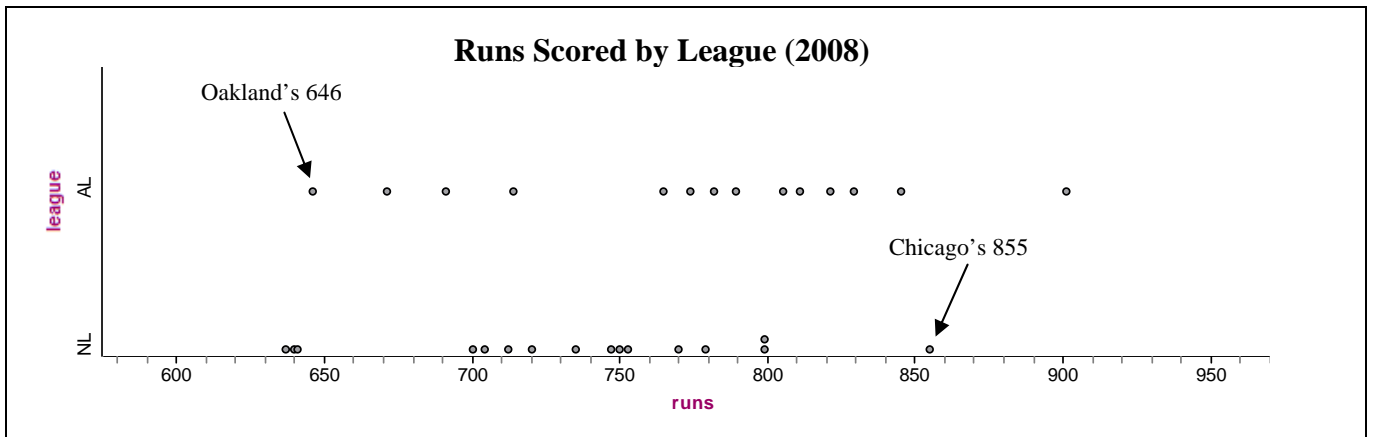
Team	League	Runs
Arizona Diamondbacks	NL	720
Atlanta Braves	NL	753
Chicago Cubs	NL	855
Cincinnati Reds	NL	704
Colorado Rockies	NL	747
Florida Marlins	NL	770
Houston Astros	NL	712
Los Angeles Dodgers	NL	700
Milwaukee Brewers	NL	750
New York Mets	NL	799
Philadelphia Phillies	NL	799
Pittsburgh Pirates	NL	735
San Diego Padres	NL	637
San Francisco Giants	NL	640
St. Louis Cardinals	NL	779
Washington Nationals	NL	641

In the next few units we will be working with numerical variables as opposed to categorical variables. If you recall from the first three units, categorical variables are variables where the possible responses fall into categories such as win/lose, make/miss, or success/failure. Possible responses to numerical variables are (surprise!) numbers instead of categories. Some examples of numerical variables are: points per game in basketball, lap speed in NASCAR, time of possession in football, and number of runs in baseball.

Graphing Numerical Variables: Dotplots and Histograms

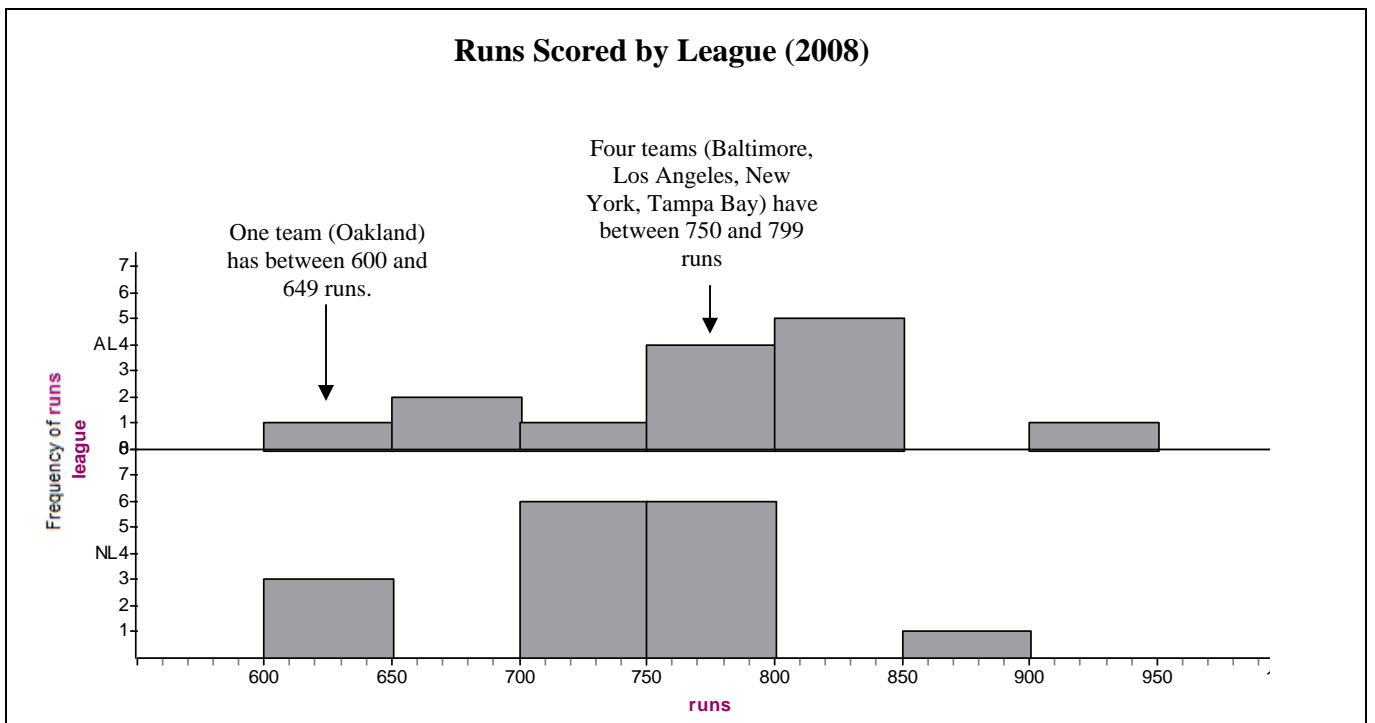
Even though we are working with a different type of variable, many of the lessons we learned in dealing with categorical variables still apply. One of the most important lessons was the need to graph the data.

We have already been introduced to one of easiest and best ways to graph numerical variables: the dotplot. This type of plot displays a number line with appropriately placed dots representing each observation. Here are dotplots comparing the runs scored for AL and NL teams in 2008.



Notice that the axes are clearly labeled and both leagues are plotted on the same scale. This is essential for making comparisons!

Another useful type of graph is a histogram. Instead of displaying each individual observation like in a dotplot, a histogram divides the possible responses into categories and counts how many observations are in each category. The number in each category (frequency) is then represented by the height of the bar representing that category. Here are histograms for the number of runs scored in the AL and NL during 2008:



In these histograms the variable runs is divided into categories of 600-649, 650-699, 700-749, etc. and the number of teams in each category is counted. The number of teams in each category is represented by the heights of the bars. If you choose to make a histogram by hand, you can

choose the width of the categories but it should be the same for all categories. You can also choose where to start the first category to make the boundaries nice. If you use technology to make a histogram, it will often make these choices for you but allow you the opportunity to adjust them if you want.

When data sets are small, we will typically use dotplots so we can see each individual observation. However, when data sets are large, histograms become more useful. Imagine a dotplot showing the incomes of all Americans—there would be over 300 million dots!

Comparing Distributions of Numerical Variables:

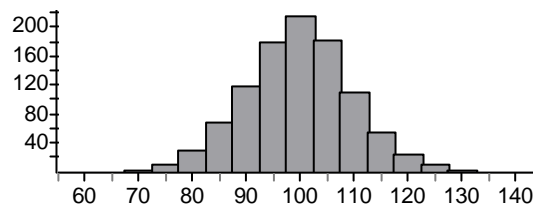
When comparing distributions of numerical variables, it is important to consider four key characteristics of the distributions:

1. Shape
2. Center
3. Spread
4. Unusual Values

Shape:

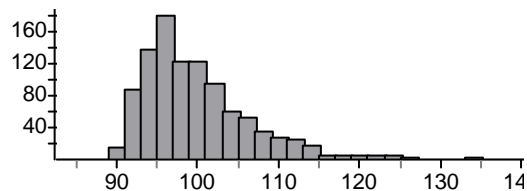
There are several phrases we can use to describe the shape of a distribution:

Symmetric, Single Peaked:



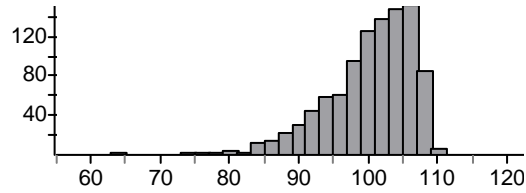
In reality, almost no distributions of real data are exactly symmetric. So, when a statistician uses the word symmetric, she probably means “roughly symmetric.”

Skewed Right, Single Peaked:



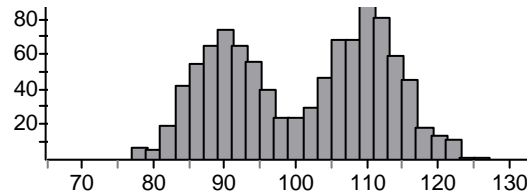
In a distribution that is skewed to the right, the data seem stretched out to the right. The direction of the long tail tells you which way the data is skewed.

Skewed Left, Single Peaked:



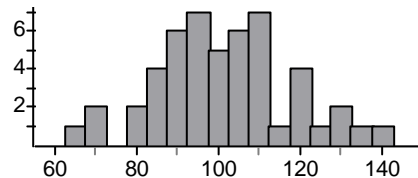
In a distribution that is skewed to the left, the data seemed stretched out to the left.

Double Peaked:



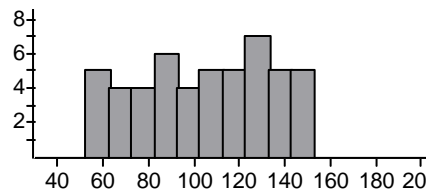
Double peaked distributions often arise when two different populations are mixed together in one graph. For example, if you graphed the heights of a sample of horse jockeys and NBA players, you would see a double peaked distribution. In this case you should describe the data around each peak separately.

Single or Double Peaked?



Even though the heights of the bars go up and down quite a bit, this distribution is still single peaked. When looking at real data, especially when the amount of data is small, you will often see some small deviations from the overall shape.

Uniform:

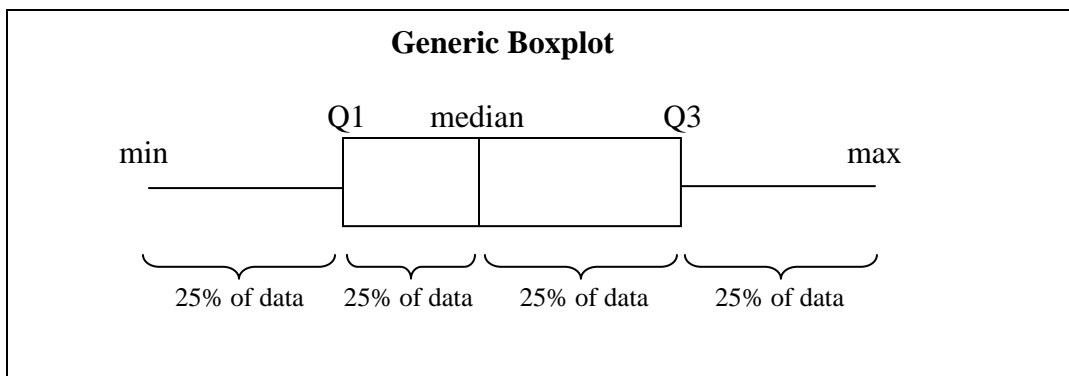


When players are in uniform, they all look the same. Likewise, in a uniform distribution, the heights of the bars are all about the same. Just like symmetric distributions, rarely will real data be exactly uniform. Using the phrase “approximately uniform” is probably a better description.

Looking back at the graphs of the AL and NL data, it appears that the AL data is slightly skewed to the left while the NL data is roughly symmetric. Both are probably single peaked.

Graphing Numerical Variables: Boxplots

A third very useful type of graph is called a boxplot (or box-and-whisker plot). A boxplot shows five key positions in a distribution, often called the 5-number summary (the minimum, the first quartile, the median, the third quartile, and the maximum). These five numbers divide the data into four approximately equally sized groups representing the lowest 25% of the data, the second 25%, the third 25% and the highest 25% of the data. This is represented in the generic boxplot below, where Q1 represents the first quartile and Q3 represents the third quartile:



In the generic boxplot above, the left whisker shows where the lowest 25% of the data are located and the right whisker shows where the highest 25% of the data are located. The box shows where the middle 50% of the data is located and the median shows the boundary between the lower and upper half of this middle 50%.

Finding the 5 number summary:

1. Put the data in numerical order
2. Find the median (the middle number). If there are two middle numbers, use the average of those two numbers for the median.
3. To find Q1, find the median of the lower half of the data.
4. To find Q3, find the median of the upper half of the data.

Note: On steps 3 and 4, if the median is one of the values in the data set, do not include it in either the lower half or the upper half of the data. Some books (and some computer software, including Fathom) choose a different approach by including the median in both halves when it is one of the observations. This will lead to slightly different values for the quartiles, which is OK.

Here is the AL data in order:

646 671 691 714 765 774 782 789 805 811 821 829 845 901

Since there are 2 middle numbers (782 and 789) the median is: $\frac{782+789}{2} = 785.5$.

The median of the lower 7 numbers (Q1) is 714:

646 671 691 714 765 774 782

The median of the upper 7 numbers (Q3) is 821:

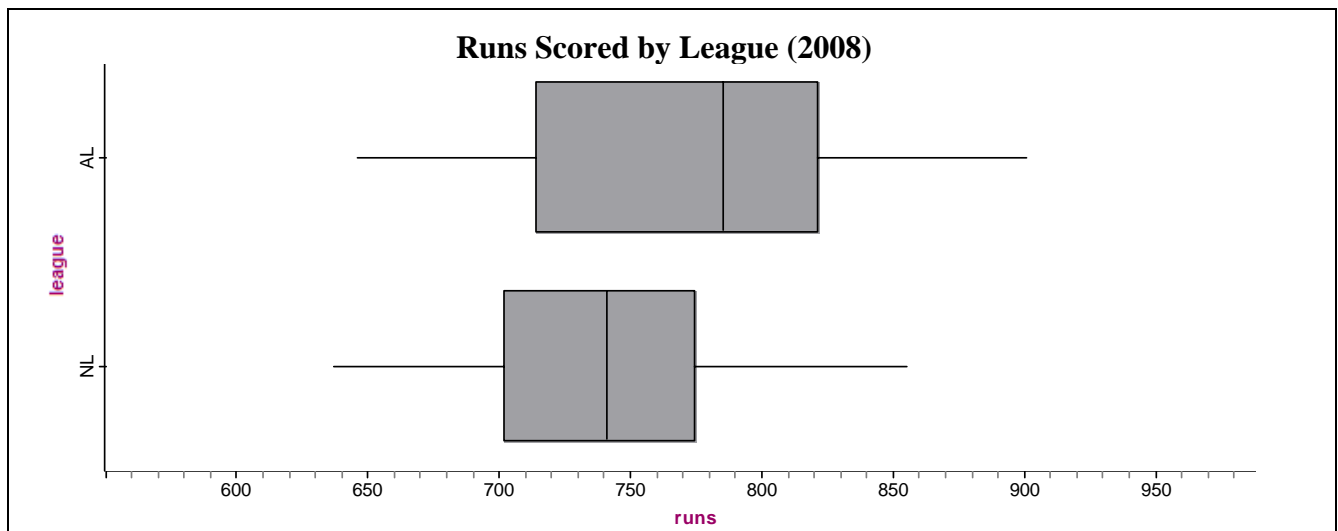
789 805 811 821 829 845 901

So, the 5-number summary for the AL data is:

- Minimum = 646
- Q1 = 714
- Median = 785.5
- Q3 = 821
- Maximum = 901

Likewise, the 5-number summary for the NL data is: 637, 702, 741, 774.5, 855.

We can now make boxplots for the AL and NL data using the values in the 5-number summary:



Comparing Distributions Again

Recall from earlier that when comparing distributions you should always address shape, center, spread and any unusual values. Using the boxplots we can now do a more thorough comparison of the AL and NL data.

Shape: The AL distribution is skewed left since the distance from the minimum to the median is longer than the distance from the median to the maximum. However, the NL distribution is approximately symmetric. We cannot tell if there are any peaks in the distributions, however, since boxplots do not reveal this feature of the distribution.

Center: When comparing centers using boxplots, it is easiest to compare the medians of each distribution since they are prominently displayed on the boxplots. In this case, the median of the AL distribution is about 40 more than the median of the NL distribution.

Another common way to find the center of a distribution is to calculate the mean value (average value). To find the mean simply add up all the observations and divide by the number of observations.

For example:

$$\text{The mean of the AL data is: } \frac{646 + 671 + \dots}{14} = \frac{10844}{14} = 774.6 \text{ runs}$$

$$\text{The mean of the NL data is: } \frac{637 + 671 + \dots}{16} = \frac{11741}{16} = 733.8 \text{ runs}$$

Using either the mean or the median, we can tell that the center of the AL distribution is higher than the center of the NL distribution.

Spread: To describe the spread (or variability) of the data we can use the range or the interquartile range (IQR). We will also learn a third way to measure variability (the standard deviation) in a future unit.

The range of a data set is the distance between the minimum value and the maximum value. For example:

$$\text{The range of the AL data is: } 901 - 646 = 255 \text{ runs}$$

$$\text{The range of the NL data is: } 855 - 637 = 218 \text{ runs}$$

According to this measure of spread, there was more variability in the AL data than in the NL data. On a boxplot, the range is represented by the length of the plot from whisker tip to whisker tip.

Another way to measure spread is with the interquartile range (IQR). This is simply the range of the middle 50% of the data (or the distance between Q1 and Q3).

The IQR of the AL data is: $821 - 714 = 107$ runs

The IQR of the NL data is: $774.5 - 702 = 72.5$ runs

According to this measure of spread, there was also more variability in the AL data than in the NL data. On a boxplot, the IQR is represented by the length of the box. Just as a reminder, whether or not you include the median when you calculate the quartiles will have an effect on the value of the IQR.

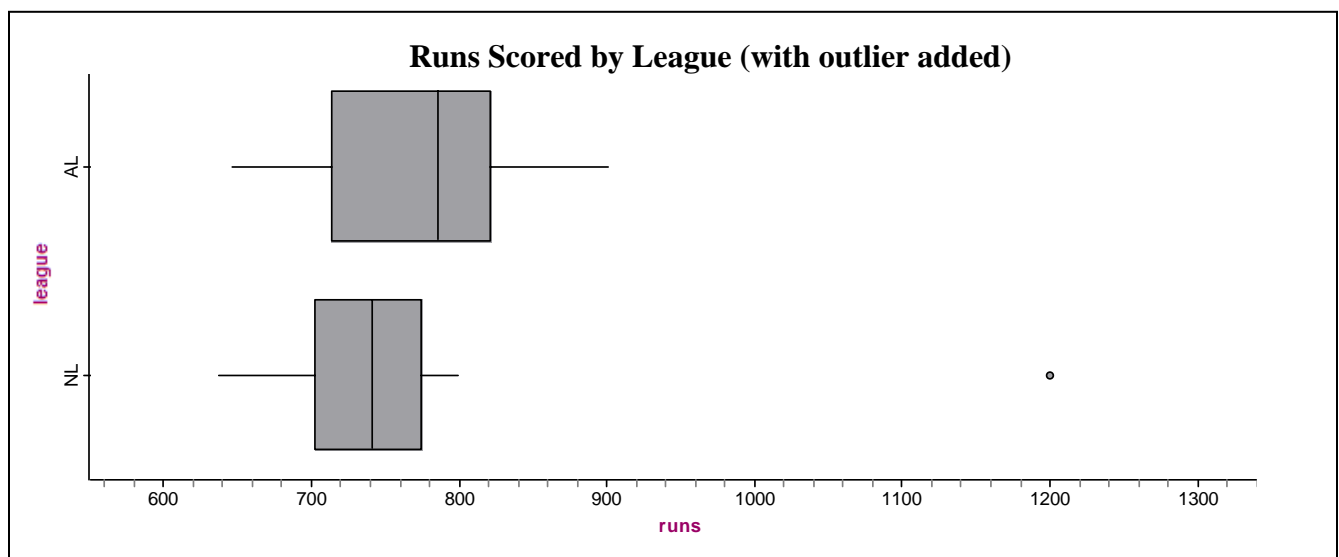
Unusual Values: If there are values that do not fall in the pattern of the rest of the data, these are called outliers. Typically, outliers are unusually high or unusually low values in a distribution. These are important to notice for several reasons:

1. They could indicate a mistake in data entry (for example, a team scoring 8444 runs in a season would be quite remarkable!).
2. They could indicate a special phenomenon (for example, in a graph of career earnings, Tiger Woods would probably be an outlier).
3. They can heavily influence the values of summary statistics such as the mean and the range. For example, consider what would happen if the top scoring team in the NL had 1200 runs instead of 855. How would this affect the mean? median? range? IQR? Clearly, the mean and range would increase, but the median and IQR would stay the same. This points out an important property of the median and IQR: they are not influenced by unusual values very much compared to the mean and range.

Outliers are typically marked separately on a boxplot, such as the example below. One common way to identify outliers is to see if there are any values more than 1.5 IQRs away from each quartile.

$$\text{Outliers} < Q1 - 1.5(\text{IQR}) \quad \text{and} \quad \text{Outliers} > Q3 + 1.5(\text{IQR})$$

Here is an example with the largest NL value changed from 855 to 1200. This observation certainly stands out as unusual! It would also greatly increase the mean and range of the NL data, but the median and IQR would be unaffected.



Making Graphs on the TI-84:

1. Enter the AL data in L1 and the NL data in L2.

L1	L2	L3	Z
646	799		
671	799		
774	735		
901	637		
714	640		
-----	779		
	641		

L2(16) = 641

2. Press StatPlot (2nd: Y=) and choose Plot1

```

STAT PLOTS
1: Plot1...On
  □ L1 1 □
2: Plot2...Off
  □ L2 1 +
3: Plot3...Off
  □ L3 1 □
4↓ PlotsOff
  
```

3. Turn Plot1 On, choose the Boxplot icon (with outliers), and enter L1 for Xlist

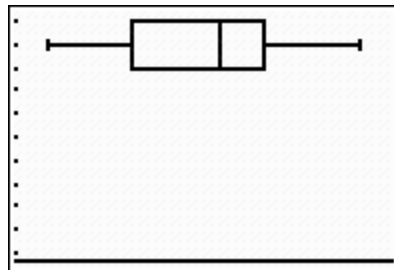
```

Plot1 Plot2 Plot3
Off Off
Type: [L1] [L2] [L3]
Xlist: L1
Freq: 1
Mark: [ ] + .
  
```

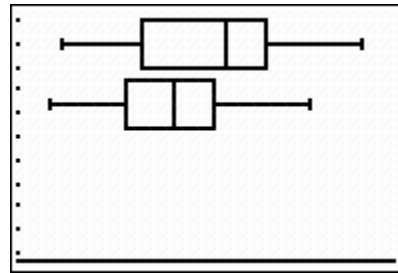
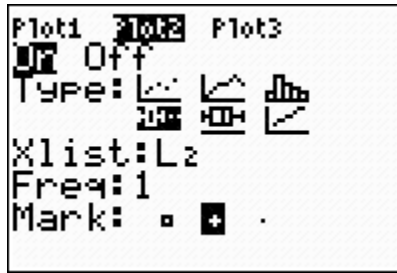
4. Press Zoom : 9ZoomStat to have the calculator choose a nice window for the graph:

```

MEMORY
4↑ ZDecimal
5: ZSquare
6: ZStandard
7: ZTrig
8: ZInteger
9↓ ZoomStat
0: ZoomFit
  
```



5. To see the boxplot for the NL distribution at the same time, repeat the same process using Plot2 and L2:



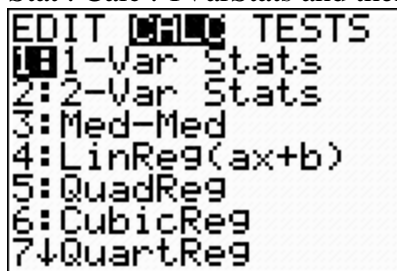
Note: You can also make histograms on the TI-84, but you can only view one at a time. Follow the same steps as above, except for choosing the histogram icon.

Calculating Summary Statistics on the TI-84:

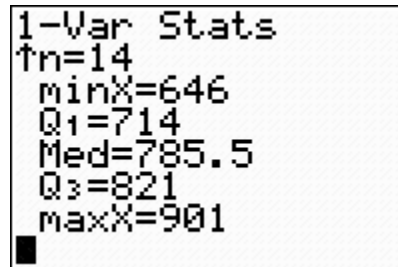
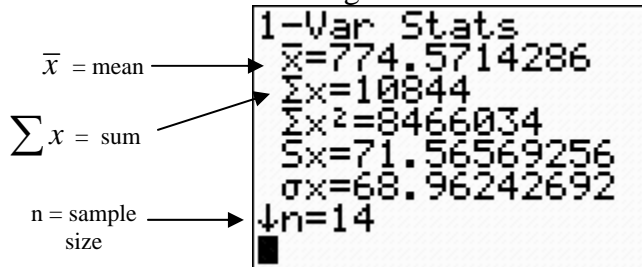
1. Enter the AL data in L1 and the NL data in L2

L1	L2	L3	2
782	720	-----	
845	753		
811	855		
805	704		
821	747		
691	770		
765	712		
L2(1)=720			

2. Press Stat : Calc : 1VarStats and then specify L1



3. The following statistics are calculated:



} 5 number summary

Here is the output for the NL data (1VarStats L2):

```
1-Var Stats
x̄=733.8125
Σx=11741
Σx²=8672521
Sx=61.55129974
σx=59.59678971
↓n=16
█
```

```
1-Var Stats
n=16
minX=637
Q1=702
Med=741
Q3=774.5
maxX=855
```

Putting it all together: Comparing Distributions

When you are asked to compare two distributions, you should make an appropriate graph and compare the shapes, centers, and spreads of the distributions and discuss any unusual values.

For example, when asked to compare the distributions of runs for AL and NL teams in 2008 (using the original data), you should make a graph and say something like: “The distribution of runs for the AL is slightly skewed left while the distribution of runs for the NL is approximately symmetric. The AL distribution has a higher median than the NL and is also slightly more spread out than the NL distribution (both the range and IQR are higher for the AL).”

Taking the Next Step: *Performance vs. Ability*

Based on our comparison of the distribution of runs for the AL and the NL, it is clear that the average offensive *Performance* of the AL teams is better than the average offensive *Performance* of the NL teams in 2008. However, as we learned in previous units, *Performance* is part *Ability* and part *Random Chance*. What we are really interested in is if the teams with the DH in the AL have a better *Ability* to score runs.

So, we want to test the following hypotheses:

H_0 : Teams with a DH have the same average *Ability* to score runs as teams without a DH

H_a : Teams with a DH have a greater average *Ability* to score runs than teams without a DH

The measure we will use to test these hypotheses is the difference in their means, although we can also do a test for the difference in their medians.

Recall from earlier that the mean of the AL distribution was 774.6 runs and the mean of the NL distribution was 733.8 runs, giving a difference of 40.8 runs in favor of the league with the DH. Of course, there are two plausible explanations for this difference: 1. The AL really does have a greater *Ability* to score runs or 2. The AL doesn't have a greater *Ability* to score runs and the difference in *Performance* was due to *Random Chance*.

To investigate how likely it is to get a difference in average *Performances* of 40.8 or larger simply due to *Random Chance*, we start by assuming that there is no difference in the *Ability* to score runs for the two leagues. To simulate this, write the 30 run totals on 30 index cards. Then, shuffle them up and deal them into two stacks. One stack should have 14 cards to represent the AL and the other stack should have 16 cards to represent the NL. Finally, calculate the mean of each stack and find the difference in the means (AL – NL).

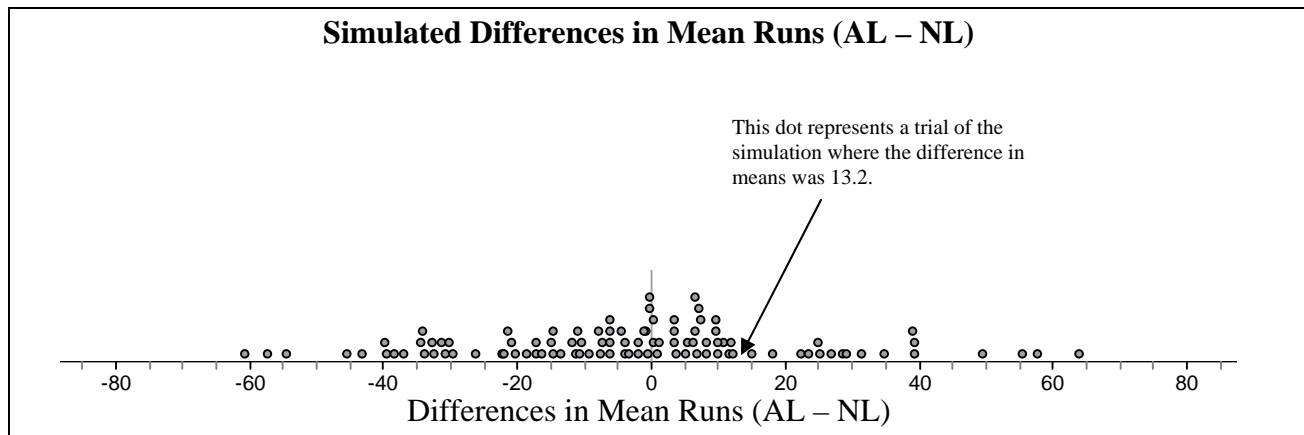
By randomly assigning each run total to a league, we are assuming that there is no difference in abilities between the leagues so that a team that scored 855 runs would score 855 runs no matter which league it was in.

Here is an example of one trial of the simulation:

AL: 714 691 799 829 720 765 641 821 901 770 789 714 704 782
 NL: 700 640 845 671 779 805 855 774 646 735 637 811 799 747 750 753

The means are AL = 759.9 and NL = 746.7 for a difference of 13.2 runs.

Here are the results of 100 trials of the simulation:



A difference of 40.8 or larger only occurred in 4 of the 100 trials (p -value = .04). This indicates that it is fairly unlikely to get a difference in *Performances* this large by *Random Chance* alone when the *Abilities* of the two leagues are the same. Thus, we can conclude that the data does provide fairly convincing evidence that the presence of the DH increases the *Ability* to score runs.

Caution about Causation

Even though we may be convinced that AL teams have a higher ability to score runs, this doesn't prove that the presence of the DH is the reason for the increase. It is possible that American League teams place a higher value on offense and National League teams place a higher value on

pitching. Or, it is possible that American League teams play in smaller ballparks which make it easier to hit home runs.

As always, we cannot automatically conclude that two variables have a cause-and-effect relationship just because they have an association. It is possible that changes in one variable cause changes in the other, but there are usually other possible causes as well.

Another Example: Home Court Advantage in the WNBA?

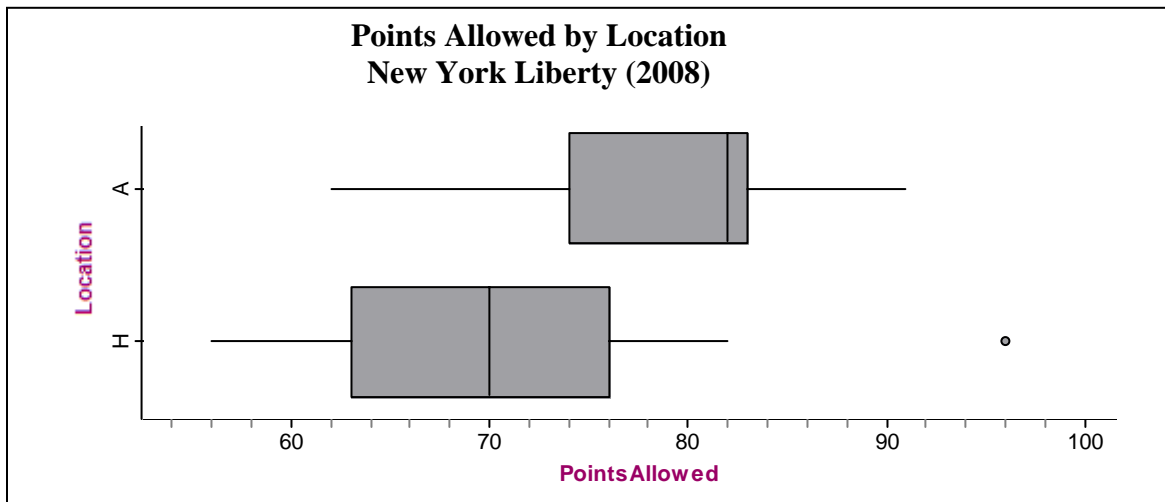
In unit 2 we investigated home field advantage in the NFL by considering the outcome of the game (a categorical variable with two outcomes: win or lose). In this unit we will continue this investigation, but focus on a numerical variable: points allowed. In basketball, the home crowd is usually very vocal and supportive of the home team and not so supportive of the visiting team (watch the fans as the other team tries to shoot a free throw!) So, it shouldn't be a surprise if a team allowed fewer points to be scored at home than on the road. Can't you just hear the fans chanting? De-fense! De-fense!

Here are the points allowed during the 2008 regular season for the WNBA's New York Liberty:

Home: 77 60 63 73 70 76 82 96 64 56 71 68 69 71 61 82 61

Away: 72 89 77 83 91 82 78 84 83 83 71 73 76 83 87 74 62

To initially compare these distributions, consider the boxplots below:



Overall, it seems like the Liberty allow fewer points at home since the center for the home games is much lower than the center for the away games. Also, the variability of the home distribution is much greater (both the IQR and range are larger), especially considering the outlier at 96 points (a game they won 102-96). The shape of the away distribution is skewed to the left, while the home distribution is roughly symmetric except for the outlier.

To measure their difference in *Performance* at home and on the road, we can compute the difference in their home and away means: On the road they give up an average of 79.3 points

and at home they give up an average of 70.6 points. Since the mean of the home distribution is so much lower, it appears that the Liberty's *Ability* to prevent scoring is greater at home than on the road. Of course, it is always possible that their *Ability* is the same in both locations and the difference in *Performance* is due to *Random Chance*.

So, we will test the following hypotheses:

H_0 : The Liberty have the same *Ability* to prevent scoring at home and on the road

H_a : The Liberty have a greater *Ability* to prevent scoring at home than on the road

The measure we will use to test these hypotheses is the difference in their means:

$$\begin{aligned} & \text{Mean Points Allowed in Away Games} - \text{Mean Points Allowed in Home Games} \\ &= 79.3 - 70.6 \\ &= 8.7 \end{aligned}$$

To see how likely it is to get a difference of 8.7 by *Random Chance*, we will perform a simulation on the TI-84.

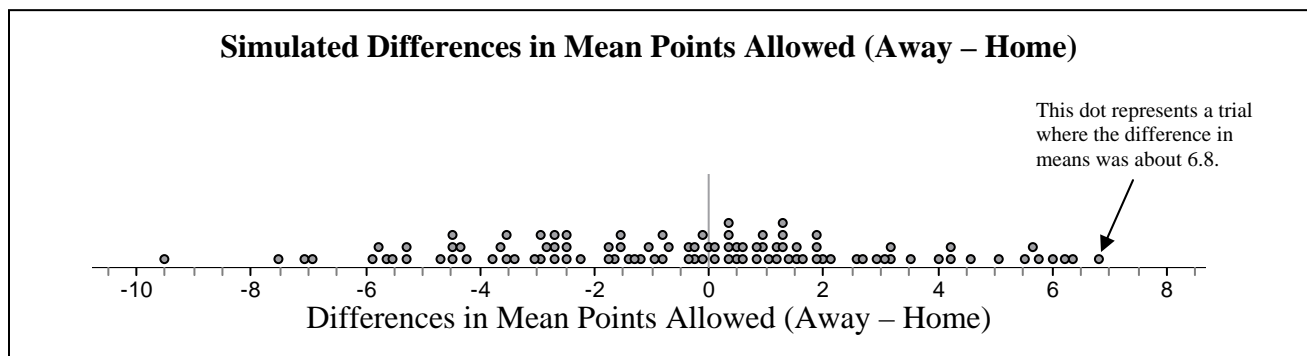
Simulation using the TI-84:

1. Enter the 34 observations in List 1
2. In the heading of List 2, enter the command "Rand(34)". This will place 34 random numbers into List 2.
3. In the Stat>Edit menu, choose SortA and enter "SortA(L2,L1)". This will put the random numbers in order and shuffle the original 34 observations.
4. Find the mean of the first 17 observations in L1. This will be the mean points allowed in the home games. You will have to do this by hand or by transferring these 17 observations to a new list.
5. Likewise, find the mean of the last 17 observations in L1. This will be the mean points allowed in the away games.
6. Find the difference in the medians: Away – Home

Here are the results of 1 trial of this simulation:

$$\begin{aligned} \text{Home: } & 82 \ 64 \ 71 \ 68 \ 69 \ 61 \ 82 \ 72 \ 77 \ 83 \ 91 \ 84 \ 83 \ 73 \ 76 \ 74 \ 62 \ (\text{mean} = 74.8) \\ \text{Away: } & 77 \ 60 \ 63 \ 73 \ 70 \ 76 \ 96 \ 56 \ 71 \ 61 \ 89 \ 82 \ 78 \ 83 \ 71 \ 83 \ 87 \ (\text{mean} = 75.1) \\ \text{Difference in means (Away} - \text{Home)} &= 75.1 - 74.8 = 0.3 \end{aligned}$$

Here are the results of 100 trials of this simulation:



As you can see from the dotplot above, a difference of 8.7 or greater didn't occur a single time (p -value = 0). Since it would be very unusual to get a difference in *Performance* of 8.7 or more if the Liberty's *Ability* to prevent scoring was the same at home as on the road, we can conclude that there is convincing evidence that the Liberty's *Ability* to prevent scoring is better at home.

Of course, it is always possible we made a Type I error: deciding they have a greater *Ability* to prevent scoring at home when in fact they do not.

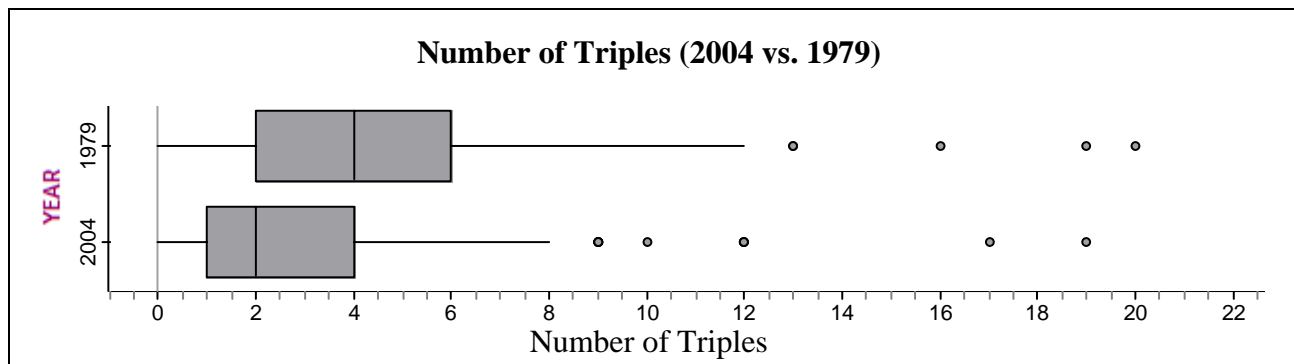
Caution: Comparing Skewed Distributions

When a distribution is skewed (or has outliers) this can greatly affect the value of the mean so that it is no longer a good indication of what is typical. However, since the median is more resistant to skewness and outliers, it will still be a good way to describe a typical value in a distribution. Thus, when we are comparing distributions that are skewed, we should consider comparing their medians rather than their means.

Example: Decline of the Triple?

In baseball, a triple is when a player hits the ball and makes it safely to third base. Typically, a triples hitter needs lots of speed to make it all the way to third base. But many people have suggested that in recent years the number of triples has been going down as teams increasingly favor power hitters over speedy ones and want to avoid the risk of turning a sure-double into an out. Is this true? Has the *Ability* of Major League Baseball players to hit triples gone down?

Here are the distributions of Number of Triples for all players with at least 500 plate appearances in 1979 (140 players) and 2004 (162 players):



Comparing these distributions we see that both are skewed to the right with several high outliers. The median for 1979 was 4 which is twice as large as the median for 2004 which was 2. The ranges for both distributions are about the same, but the IQR for 1979 was larger, showing that there was more variability in the past in addition to a higher center.

We want to test the following hypotheses:

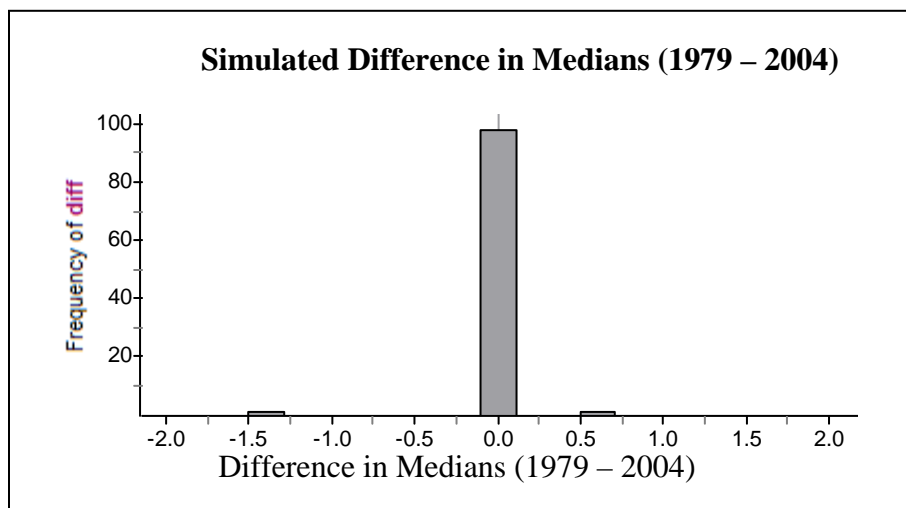
H_0 : The *Ability* of Major League Hitters to hit triples was the same in 1979 and 2004.

H_a : The *Ability* of Major League Hitters to hit triples was greater in 1979 than 2004.

We will use the Difference in Medians as our measure to test these hypotheses. The actual difference in medians was 2.

To see how likely it is to get a difference in medians of 2 or larger just by *Random Chance*, we can conduct a simulation exactly like the other simulations in this unit, except that we will calculate the difference in medians instead of the difference in means.

Here are the results of 100 trials of a simulation where all 302 observations were shuffled up and randomly assigned to a pile of 140 (representing the 1979 players) or a pile of 162 (representing the 2004 players):



As you can see, a difference of 0 was by far the most common occurrence (98 times out of 100). In one trial there was a difference of -1.5 and in another trial there was a difference of 0.5 (remember that it is possible to have a median that ends with “.5”. For example, if the two middle numbers are 3 and 4, the median would be 3.5).

Since we never got a difference of 2 or more, the p -value is approximately 0. That is, there is a roughly 0% chance of getting a difference in medians as large as 2 by *Random Chance* alone. So, we can rule out *Random Chance* as an explanation for the observed difference in the median number of triples and conclude that the *Ability* of Major League hitters to hit triples was greater in 1979 than 25 years later in 2004.

However, we need to be careful with the word *Ability* in this case. Even if we have convincing evidence that the typical number of triples has gone down, we don't necessarily know the cause of this decrease (we just know it wasn't due to *Random Chance*). For example, perhaps runners are just slower in recent years. Or, perhaps ballparks are smaller these days, not only making it more tempting to try to hit it over the fence but also giving less room for the ball to bounce around and allow runners the time to get to third. It could also be a change in strategy, balancing that the risk of being thrown out at third with playing it safe and staying on second base (we will learn more about questions like these in a future unit).

Connections: Looking Forward...Looking Back

In this unit we learned to how to graph and compare distributions of a numerical variable such as runs scored or points allowed. In many ways this was parallel to what we learned in units 1 and 2 about graphing and comparing distributions of a categorical variable such as outcomes of games and outcomes of shots. This also foreshadows what we will learn in unit 8 when we investigate relationships between numerical variables, such as driving distance and scoring average in golf.

In this unit we also continued our exploration of the home field advantage, this time focusing on a numerical variable (points allowed) rather than a categorical variable (outcome of game). We also continued to implement the same basic shuffling technique for simulating athletic *Performances* and will continue to use this technique in nearly every unit the rest of the course.

In the immediate future we will use the shuffling technique to compare distributions of a numerical variable when the values are paired (such as the number of homeruns hit by teams in two different years) and to compare the amount of spread in distributions of a numerical variable (such as comparing the consistency of two different running backs).

Stats 101: A More Traditional Approach

The entire first half of this unit (graphing and comparing distributions of a numerical variable) is very similar to what is taught in many traditional statistics courses, however the second half of the unit might look very different.

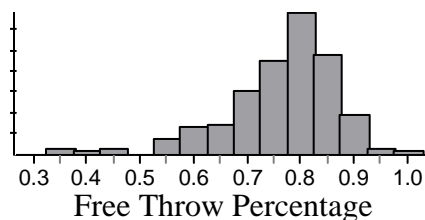
In a more traditional course there will still be plenty of problems asking students to test for a difference between two means, but instead of using a simulation to investigate the possible values of a difference in means a mathematical model (a t-distribution) is used to approximate this distribution. In theory, the simulation will give a more exact estimate of the p -value if enough trials can be conducted (ideally in the millions of trials). However, in most cases using the t-distribution to approximate the simulated distribution of differences in means will be very accurate.

Typical problems you may encounter would be comparing the mean cholesterol levels of patients receiving one of two drugs or comparing the average salaries of men and women in a certain profession.

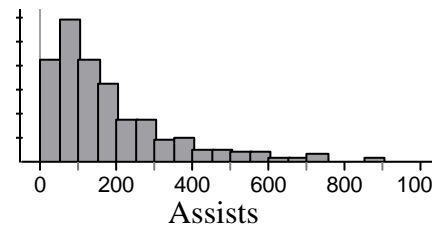
For Practice:

- Classify the following variables as categorical or numerical:
 - Outcome of the next school football game
 - Distance a golf ball travels
 - Time it takes to swim 100 meters
 - Result of next at bat in a baseball game
 - Points scored in the next basketball game
- Here are some histograms of data from individual players (minimum 60 games) the 2008-2009 NBA season. Describe the shape of each distribution.

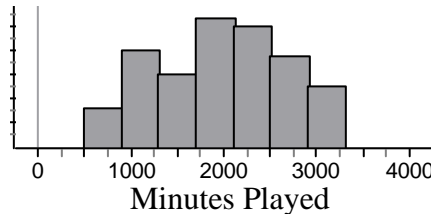
a.



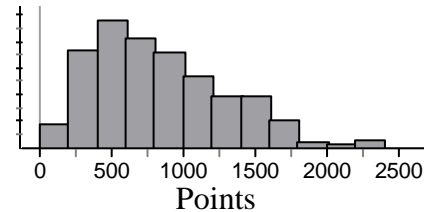
b.



c.



d.



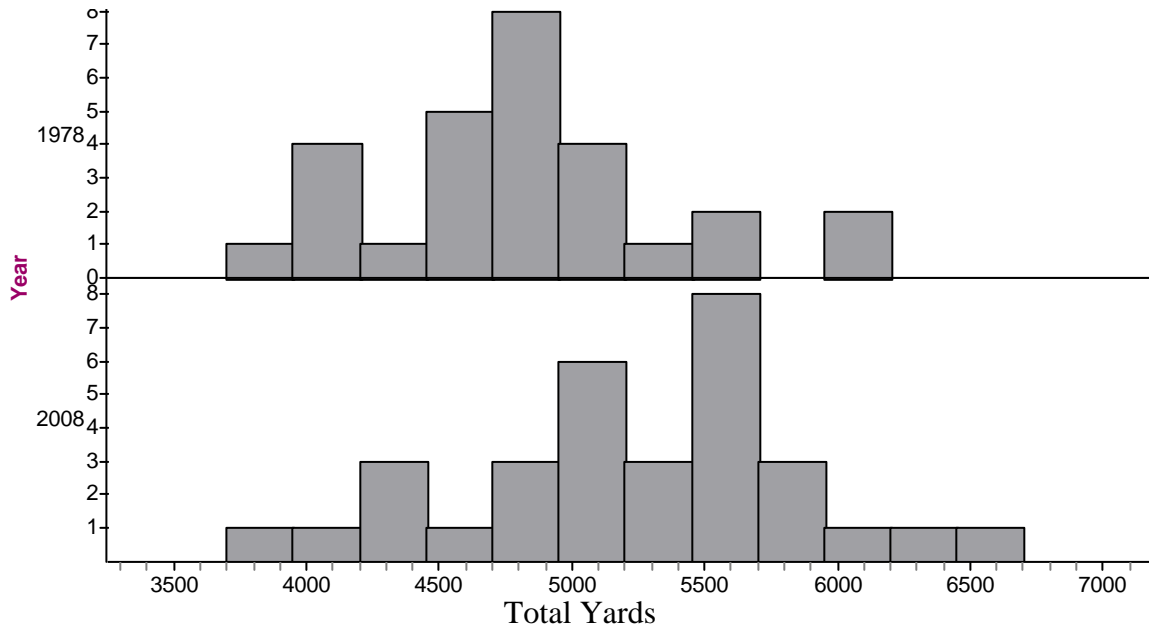
3. Here are the number of touchdowns thrown by Arizona Cardinals quarterback Kurt Warner for each game during the 2008 season. Make a dotplot of the data and describe the shape of the distribution.

1 3 2 2 2 2 2 3 1 1 3 1 1 0 4

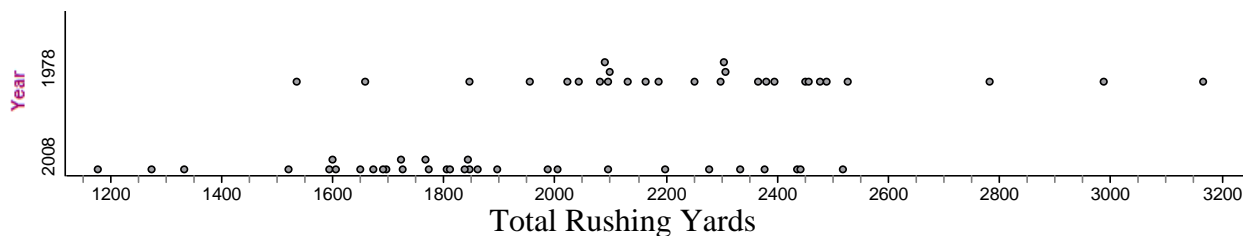
4. Here are the passing yardage totals for New York Giant quarterback Eli Manning for each game during the 2008 season. Make a histogram of the data and describe the shape of the distribution. *Hint: Use classes of size 50 starting at 100.*

216 260 289 267 196 161 199 147 191 153 240 305 123 191 181 119

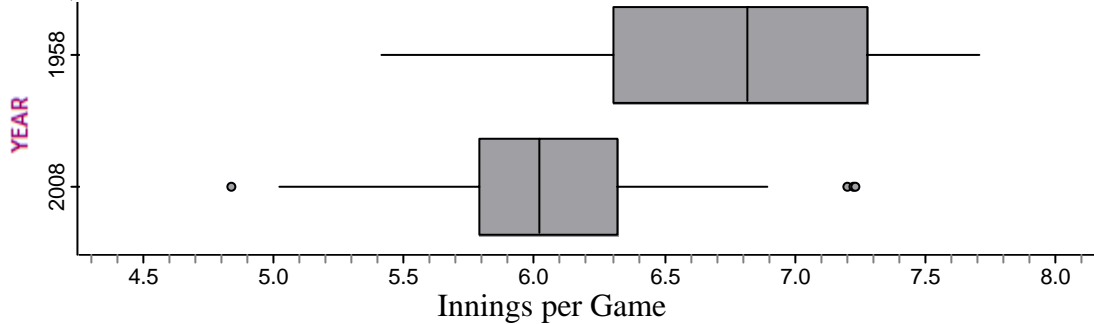
5. Make a boxplot of the data in question #4. Describe the shape, center, and spread of this distribution.
6. Has offensive *Ability* in the National Football League changed in the last 30 years? Here are histograms showing the Total Yards for each NFL team in 1978 and 2008. Compare these distributions.



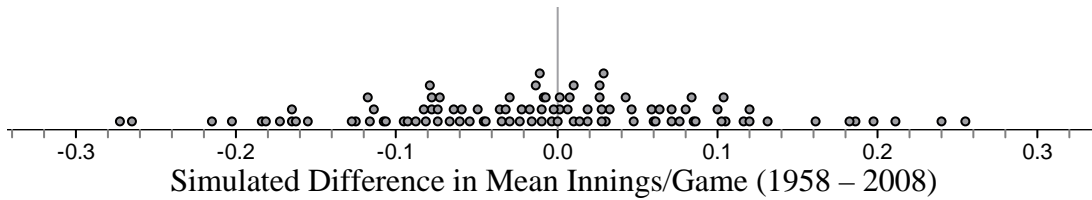
7. Here are two dotplots showing the Total Rushing Yards for each NFL team in 1978 and 2008. Compare these distributions.



8. Using the data displayed in the dotplots in question #7, construct two boxplots of the same data.
9. Many old-timers complain that modern starting pitchers in baseball don't last as long in a game as pitchers did 50 years ago. Is this true? The following boxplots show the distribution of innings per game for starting pitchers in 1958 and 2008 (minimum of 150 innings pitched).



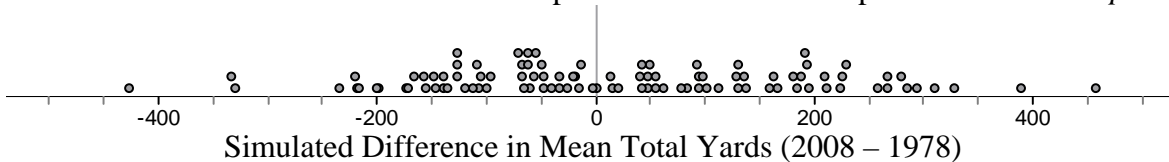
- Compare these distributions.
 - Explain how the mean is related to the median (smaller, larger, about the same) for each distribution.
 - About what percent of pitches averaged more than 6 innings in 1958? 2008?
10. The following data values are the Points Allowed in each game by the 2008 New York Jets:
14 19 48 35 14 16 24 17 3 31 13 34 24 27 13 24
- Calculate the 5-number summary for this data.
 - In one game, the Jets only allowed 3 points. Is this value an outlier? Justify.
 - In another game, the Jets allowed 48 points. Is this value an outlier? Justify.
 - Make a boxplot for this data, marking outliers separately if there are any.
11. The following data values are the Points Scored in each game by the 2008 Chicago Bears:
29 17 24 24 34 20 48 27 14 3 27 14 23 27 20 24
- Calculate the 5-number summary for this data.
 - In one game, the Bears only scored 3 points. Is this value an outlier? Justify.
 - In another game, the Bears scored 48 points. Is this value an outlier? Justify.
 - Make a boxplot for this data, marking outliers separately if there are any.
12. In question #9 you compared the distributions of innings per game for 24 pitchers in 1958 and 99 pitchers in 2008. Does the data give convincing evidence that pitchers in 1958 had the *Ability* to throw more innings per game?
- State the hypotheses we are interested in testing.
 - Describe how to simulate this situation to test the difference in mean innings/game.
 - The mean innings/game in 1958 was 6.70 and the mean innings/game in 2008 was 6.05 for a difference of 0.65. To see how likely it is to get a difference this large by *Random Chance*, a simulation was conducted assuming that pitchers in both years have the *Ability* to throw the same number of innings/game. In each of the 100 trials the difference in means was calculated and recorded on the dotplot below. Use the dotplot to estimate the *p*-value.



- d. Interpret the p -value from part (c) and make an appropriate conclusion.
- e. Which type of error, Type I or Type II is it possible you committed in part (d)? Explain.
- f. If someone concludes that pitchers in 1958 did have the *Ability* to throw more innings per game, does that mean that today's pitchers are wimpier? What other possible explanations could there be?

13. In question #6, you compared the distributions of Total Yards for 28 NFL teams in 1978 and 32 teams in 2008. The mean for 2008 was 5236 yards and the mean for 1978 was 4811 yards, giving a difference of 425 yards. Does the data give convincing evidence that NFL teams had more offensive *Ability* in 2008 than 1978?

- a. State the hypotheses we are interested in testing.
- b. Describe how to simulate this situation to test the difference in mean total yards.
- c. A simulation was conducted assuming that teams in both years have the same offensive *Ability*. In each of the 100 trials the difference in means total yards was calculated and recorded on the dotplot below. Use the dotplot to estimate the p -value.



- d. Interpret the p -value from part (c) and make an appropriate conclusion.
- e. Which type of error, Type I or Type II is it possible you committed in part (d)? Explain.
- f. If someone concludes that modern NFL teams have better offensive *Ability* than teams 30 years ago, what are some possible causes for this change?

14. Do the noisy crowds in the Metrodome help the Minnesota Vikings play better defense? Here are the points allowed at home and on the road for the 2008 Vikings:

Points Allowed at Home: 18 10 10 21 27 14 24 19

Points Allowed on Road: 24 30 27 48 19 12 16 14

- a. Graph these distributions so they can be easily compared. Write a few sentences comparing them.
- b. Find the mean Points Allowed at home and on the road. How much better does the defense play at home?
- c. State the hypotheses we are interested in testing.
- d. Describe how to simulate this situation to test the difference in mean points allowed.
- e. Conduct 10 trials of the simulation you described in part (d) and display the results in a well labeled dotplot.
- f. Estimate and interpret the p -value based on your simulation in part (e).
- g. Based on your p -value, make an appropriate conclusion.
- h. If you made an error, which type could it be, Type I or Type II? Explain.

- i. If a student concluded that the Vikings have better *Ability* to play defense at home, can we guarantee that the noisy crowd is the cause? Are there other possible causes?
15. Do the “Cameron Crazies” at Duke home games help the Blue Devils play better defense? Here are the points allowed by the Duke Men’s basketball team at home and on the road during 2008-2009 conference play.
- Points Allowed at Home: 44 56 44 54 75 101 91 81
- Points Allowed on Road: 58 56 70 74 80 67 65 79
- a. Graph these distributions so they can be easily compared. Write a few sentences comparing them.
 - b. Find the *median* Points Allowed at home and on the road. How much better does the defense play at home?
 - c. State the hypotheses we are interested in testing.
 - d. Describe how to simulate this situation to test the difference in *median* points allowed.
 - e. Conduct 10 trials of the simulation you described in part (d) and display the results in a well labeled dotplot.
 - f. Estimate and interpret the p -value based on your simulation in part (e).
 - g. Based on your p -value, make an appropriate conclusion.
 - h. If you made an error, which type could it be, Type I or Type II? Explain.
 - i. If a student concluded that Duke has a better *Ability* to play defense at home, can we guarantee that the noisy crowd is the cause? Are there other possible causes?
16. Describe the relationship between the mean and median in the 4 histograms in problem #2.

For Investigation:

1. Go online and gather data on a numerical variable in two different contexts for a player, team, or sport of your choice. The contexts could be home and away, day games and night games, from two different eras, etc. Then, write a report that analyzes the data using the methods of this unit.
2. Collect data from a team at your school or from an experiment you perform. For example, which of two types of golf balls travels further? Do baseballs go farther when hit with a metal bat than with a wood bat? Do sprinters run faster when starting in blocks or starting standing up? Then, write a report that analyzes the data using the methods of this unit.