Exploring Data

Chapter 1 Notes

This chapter introduces you to the concept of an Exploratory Data Analysis. You will learn how to use a variety of graphical techniques to display data as well as how to describe distributions numerically. Emphasis will be placed on interpreting information from these summaries in the context of the data.

Objectives:

* Use a variety of graphical techniques to display a distribution. These should include bar graphs, pie charts, stemplots, histograms, ogives, time plots, and boxplots.
* Interpret graphical displays in terms of the shape, center, and spread of the distribution as well as gaps and outliers.
* Use a variety of numerical techniques to describe a distribution. These should include mean, median, quartiles, five-number summary, interquartile range, standard deviation, range, and variance.
* Interpret numerical measures in the context of the situation in which they occur.
* Learn to identify outliers.
* Explore the effects of a linear transformation of a data set.

Case Study: Nielsen Ratings

What does it mean to say that a TV show was ranked number 1? The Nielsen Media Research company randomly samples about 5100 households and 13,000 individuals each week. The TV viewing habits of this sample are captured by metering equipment, and data are sent automatically in the middle of the night to Nielsen. Broadcasters and companies that want to air commercials on TV use the data on who is watching TV and what they are watching. The results of this data gathering appear as ratings on a weekly basis. For more information on the Nielsen TV ratings, go to [**www.nielsenmedia.com**](http://www.nielsenmedia.com)**,** and click on “Inside TV Ratings.” Then under “Related,” select “What Are TV Ratings?”

Here are the top prime-time shows for viewers aged 18 to 49 during the week of August 20, 2012.

.



Which network is winning the ratings battle? At the end of this chapter, you will be asked to use what you have learned to answer this question.

Exploratory Data Analysis: Statistical practice of analyzing distributions of data through graphical displays and numerical summaries.

**Distribution:** Description of the values a variable takes on and how often the variable takes on those values.

An EDA allows us to identify patterns and departures from patterns in distributions.

**Displaying Distributions with Graphs:**

What was the difference between a categorical and a quantitative variable?

Let’s start with categorical variables and displays of their distributions.

**Example 1:**

The radio audience rating service Arbitron places the country's 13,838 radio stations into categories that describe the kind of programs they broadcast. Here is the distribution of station formats:**1**



It's a good idea to check data for consistency. Do the counts equal 13, 838? Do the percents add up to 100%

**Round-off Error:** The result of rounding off values; it doesn’t mean our work is wrong. It just displays the effects of rounding off results.

What is the problem with viewing data arranged in this format?

****

One way to display categorical variables is to create a pie graph. This is excellent manner of showing the relationship between the whole and the various categories. A pie chart must be made with ALL categories.

To find the angle for the pie chart: take the % of the category times 360 and then round to the nearest degree.

****

Another way to exhibit categorical data is using bar graphs. These are easier to make and often easier to read. They are also more flexible i.e. categories can be left off, side-by-side displays for comparisons etc.

**Example 2:** Can a pie chart be used to display the set of data below? Explain why or why not.

Portable MP3 music players, such as the Apple iPod, are popular—but not equally popular with people of all ages. Here are the percents of people in various age groups who own a portable MP3 player:**2**



**Example 3:** Do the data tell us what we want to know?

Let's say that you plan to buy radio time to advertise your Web site for downloading MP3 music files. How helpful are the data from example 1? Not very. You are interested, not in counting *stations,* but in counting *listeners.* For example, 14.6% of all stations are religious, but they have only a 5.5% share of the radio audience. In fact, you aren't even interested in the entire radio audience, because MP3 users are mostly young people. You really want to know what kinds of radio stations reach the largest numbers of young people. **Always think about whether the data you have help answer your questions.** Would the data from example 2 help you answer your question?

**Displaying Quantitative Variables**:

Stemplots:

* Are also called stem and leaf plots
* Give a quick picture of the shape of a distribution while including the actual numerical values in the graph
* Works best for small data sets with positive numbers



**Remember to include a key and all appropriate labels for a stemplot!**

**Example 4:** The stemplot below represents the percent of literate citizens from various Islamic nations. What do we notice about the data?

Why is a key important?

**Clusters: a bunch of data grouped together.**

  Key: 9|9 = 99

**Back-to-back stemplot:**

**Example 5:** Here is the literacy rate of males vs. females for the Islamic nations. Describe what the data are telling us.

key 9|2 = 92

**Stemplots do not work well for large data sets where each stem must hold a large number of leaves.** Fortunately, there are two modifications of the basic stemplot that are helpful when plotting the distribution of a moderate number of observations. You can double the number of stems in a plot by ***splitting stems*** into two: one with leaves 0 to 4 and the other with leaves 5 through 9. When the observed values have many digits, ***trimming*** the numbers by removing the last digit or digits before making a stemplot is often best. You must use your judgment in deciding whether to split stems and whether to trim, though statistical software will often make these choices for you. Remember that the purpose of a stemplot is to display the shape of a distribution.

**Example 6:**  The stemplot on the left was the original. The right one shows the data with a split stem. Why is the second picture a much better representation of the data?



**Histograms:**

Stemplots display the actual values of the observations. This feature makes stemplots awkward for large data sets. Moreover, the picture presented by a stemplot divides the observations into groups (stems) determined by the number system rather than by judgment. Histograms do not have these limitations.

**Histograms:**

A ***histogram*** breaks the range of values of a variable into *classes* and displays only the count or percent of the observations that fall into each class. You can choose any convenient number of classes, but you should *always choose classes of equal width.*

Histograms are slower to construct by hand than stemplots and do not display the actual values observed. For these reasons we prefer stemplots for small data sets. The construction of a histogram is best shown by example. Any statistical software package will of course make a histogram for you, as will your calculator.

**To construct a frequency histogram, we often construct a frequency distribution first to organize the data.**

**Example:** The data set represents the number of minutes 50 internet subscribers spent on the internet during their most recent session. Construct a frequency distribution that has **7 classes.**

 50 40 41 17 11 7 22 44 28 21 19 23 37 51 54 42 88 41 78 56 72 56 17 7 69 30 80 56 29 33 46 31 39 20 18 29 34 59 73 77 36 39 30 62 54 67 39 31 53 44

|  |  |
| --- | --- |
| Class | Frequency, f |
| 7-18 | 6 |
| 19-30 | 10 |
| 31-42 | 13 |
| 43-54 | 8 |
| 55-66 | 5 |
| 67-78 | 6 |
| 79-90 | 2 |
| Totals | 50 |

 **Lower class limit** is the smallest number in a class.

**Upper class limit:** is the largest number in a class.

 **Classes can NOT overlap or have gaps between them!**

**Class Width**: calculate the width of each class using the formula

 Usually, the number of classes to be used will be given to you. If not, use between 5-20 classes depending on the number of data pieces and the spread.

Additional Features can be included in a frequency distribution to provide a better understanding of the data

* **Midpoint (class mark):**We will use the class mark in our calculations to represent a class when we do not have the individual data values.

* **Relative Frequency:** the percentage of the data that falls in that class

 Round to 2 decimal places; the sum of the relative frequencies should be approximately 1!

* **Cumulative Frequency:** the sum of the frequency for that class and all previous classes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Frequency, f | Midpoint | Relative Frequency | Cumulative Frequency |
| 7-18 | 6 | 12.5 |  0.12 | 6 |
| 19-30 | 10 | 24.5 | 0.20 | 16 |
| 31-42 | 13 | 36.5 | 0.26 | 29 |
| 43-54 | 8 | 48.5 | 0.16 | 37 |
| 55-66 | 5 | 60.5 | 0.10 | 42 |
| 67-78 | 6 | 72.5 | 0.12 | 48 |
| 79-90 | 2 | 84.5 | 0.04 | 50 |
| Totals | 50 |  | 1.00 |  |

 **Creating histograms on the calculator:**

1. Enter the data into a list.

Press STAT ENTER. This is the list editor. Clear the list of any entries before you use it. To clear: arrow up and highlight the L1 and press then down arrow.

CLEAR



Enter the entries in the list for the list you using.

1. Set up the histogram in the plots editor (2nd Y=). Choose Plot1. Turn it on and choose histogram 3rd picture in the top row). Your XList should be the list you stored your data in.



1. Create the histogram. Be sure there is nothing in the Y = menu. Choose ZOOM 9:ZOOMSTAT

This will force the calculator to create the histogram based on the data we are using.

 

The calc. set up the histogram.

This is set up to our class width and boundaries.

1. Adjust the histogram: The TI-84 chooses the class sizes on its own. To force it to use the class width we calculated, press WINDOW. Set xscl to 12 to our class width, adjust xmin to 7 and xmax to 90.
2. Determine the heights of the bars. Use TRACE to show the height of the bar. The left and right arrow keys moves you from class to class.



1. Use the information for the height of the bars to commit your histogram to paper. You must label both the horizontal and vertical axis accurately on your paper.

**\*Note on class width: Make your width too big and you get too few bins. Make your width too small and you get too many bars.**

  

 **It is usually best to let the calculator do most of the work; then use your calculation for class width as the xscl. Adjusting the xmin or xmax may occasionally need to be done too.**

**Tips for Histograms:**

Here are some important properties of histograms to keep in mind when you are constructing a histogram.

* Our eyes respond to the area of the bars in a histogram, so *be sure to choose classes that are all the same width.* Then area is determined by height and all classes are fairly represented.
* There is no one right choice of the classes in a histogram. Too few classes will give a “skyscraper” graph, with all values in a few classes with tall bars. Too many will produce a “pancake” graph, with most classes having one or no observations. Neither choice will give a good picture of the shape of the distribution. Five classes is a good minimum. Bottom line: *Use your judgment in choosing classes to display the shape.*
* Statistical software and graphing calculators will choose the classes for you. The default choice is often a good one, but you can change it if you want. *Beware of letting the device choose the classes.*
* *Use histograms of percents for comparing several distributions with different numbers of observations.* Large sets of data are often reported in the form of frequency tables when it is not practical to publish the individual observations. In addition to the frequency (count) for each class, we may be interested in the fraction or percent of the observations that fall in each class. A histogram of percents looks just like a frequency histogram. Simply re-label the vertical scale to read in percents.

**Histograms versus Bar Graphs**

Although histograms resemble bar graphs, their details and uses are distinct. A histogram shows the distribution of counts or percents among the values of a single quantitative variable. A bar graph displays the distribution of a categorical variable. The horizontal axis of a bar graph identifies the values of the categorical variable. Draw bar graphs with blank space between the bars to separate the items being compared. Draw histograms with no space, to indicate that all values of the variable are covered.

**Describing Distributions:**

There are several factors we need to consider when we examine distributions. These are the things that should be mentioned when you are asked to describe a distribution.

 **Shape of the distribution:**



 **Symmetric:** the graph has a vertical line of symmetry; you will not find a perfectly symmetrical graph!

 Real world examples:

 **Uniform:** the heights of the bars are approximately the same across the distribution

 Real world examples:

 **Skewed left:** the distribution forms a tail to the left with the majority of the data clustered to the right

 Real world examples:

 **Skewed right:** the distribution forms a tail to the right with the majority of the data clustered to the left

 Real world examples:

 **Bimodal:** the distribution has two peaks and may or may not be symmetric also

 Real world examples:

\* Note: Distributions have a WIDE variety of shapes. In describing a distribution, we want to use the BEST word(s) to describe what we are seeing. Imagine trying to paint a picture of the distribution for a blind person when you are describing the distribution. Also be sure to mention, any strange data values you see in the picture.

When describing distributions: we also talk about the following items:

* **Center:** a value that separates the histogram roughly in half; it is an estimate
* **Spread (Variation):** the scope (range) of the values; in other words, give the max and the min
* **Clusters:** groups of data “clumped” together
* **Gaps:** holes where no data values fall
* **Outliers:** Extreme values in the distribution; sometimes happen due to errors in measuring or recording the data; can also be true values

When describing distributions, remember SOCS

SHAPE, OUTLIERS, CENTER, SPREAD

**Dealing with outliers:**

With small data sets, you can spot outliers by looking for observations that stand apart (either high or low) from the overall pattern of a histogram or stemplot. **Identifying outliers is a matter of judgment. Look for points that are clearly apart from the body of the data, not just the most extreme observations in a distribution.** You should search for an explanation for any outlier. Sometimes outliers point to errors made in recording the data. In other cases, the outlying observation may be caused by equipment failure or other unusual circumstances. In the next section we'll learn a rule of thumb that makes identifying outliers more precise.

Sometimes, we can remove an outlier as “bad data” but there must be a clear reason that is explained! Bad data could be proven measurement errors, etc. Be careful when deciding if something can be removed and don’t delete a data point without good cause or you could be changing what the data is saying!

**Ogives (cumulative frequency histograms):**

 A histogram does a good job of displaying the distribution of values of a quantitative variable. But it tells us little about the relative standing of an individual observation. If we want this type of information, we should construct a relative cumulative frequency graph, often called an ogive (pronounced O-JIVE).

**Example 7:** Let’s use the data from homework problem 6 regarding the age of the presidents at their inauguration. We are interested in the relative position of President Clinton.

Here is the frequency table with the relative frequency, cumulative frequency, and relative cumulative frequency columns added.



To draw the ogive, use the horizontal axis as the age at inauguration and the vertical axis and the relative cumulative frequency.

What about Bill Clinton who was 46 at inauguration? What is his relative standing?

What age represents the 50th percentile?



**Time plots:**

Whenever data are collected over time, it is a good idea to plot the observations in time order. **Displays of the distribution of a variable that ignore time order, such as stemplots and histograms, can be misleading when there is systematic change over time.**

**Time Plot:** A time plot of a variable plots the time on the x-axis vs. the value of the variable on the vertical axis. Connect the dots to help emphasize the change over time.



Average Monthly gas price per gallon:

What seems to be the trend in gas prices over time?

How could this graph be improved?

**Describing Data with Numbers:**

 A good description of a distribution should include its shape and numbers describing its center and its spread. Information about the shape should come from a graphical display. In this section, we will learn how to calculate numbers that describe the center and the spread. You can calculate these numerical measures for any quantitative variable. But to interpret measures of center and spread, and to choose among the several measures you will learn, you must think about the shape of the distribution and the meaning of the data. The numbers, like graphs, are aids to understanding, not “the answer” in themselves.

**Measuring the center:**

 Numerical description of a distribution begins with a measure of its center or average. The two common measures of center are the ***mean*** and the ***median.***

**Mean:** The mean is the average of the data set.

* To find the mean, add the data values and divide by n, the number of data points.
* Formula for mean: 
	+ If we are working with a sample: represents the mean and n = the number of data values
	+ If we are working with a population, then μ represents the mean and N = the number of data values.
* **Median:** The median M of a distribution is the middle value where half of the data falls below and half above.
	+ BE SURE TO PUT ALL DATA IN ORDER FROM LEAST TO GREATEST!!!!
* Special Note: The mode of a data set is the value that occurs the most frequently.

**Example 8:**

 The table represents the gas mileage of several popular two-seater cars. Find the mean and median highway mileage.



Make a histogram of the data. What do you notice?

Is it possible to drop this point from the data? Explain.

Recalculate the mean and median now.

This example illustrates the idea of **resistance.**

The mean is not resistant to extreme values. An outlier or a skewed distribution will affect the mean by pulling the mean towards the extreme value or the tail. Explain why.

This is a weakness of the mean as a measure of center. If the data set is skewed or contains an outlier, the mean is not the best measure of center to report as a “typical” value of the data.

What about the median? Is it a resistant measure? Explain.

**Mean vs. Median**:

**Example 9:** Sketch an arrow to show the approximate location of the mean and median in each distribution.



**Don't confuse the “average” value of a variable (the mean) with its “typical” value, which we might describe by the median.**

**Measuring Spread**

 A measure of center alone can be misleading. Two nations with the same median family income are very different if one has extremes of wealth and poverty and the other has little variation among families. A drug with the correct mean concentration of active ingredient is dangerous if some batches are much too high and others much too low. We are interested in the *spread* or *variability* of incomes and drug potencies as well as their centers. ***The simplest useful numerical description of a distribution consists of both a measure of center and a measure of spread.*** The spread describes the variability among the data values.When reporting about a data set, both a measure of center and a measure of spread need to be reported to give a full picture of the data.

* **RANGE**: The range describes the spread between the maximum and the minimum value of a data set.
* Range = Max – min
* Is the range resistant? Explain.

Percentile: the pth percentile breaks the data set such that p% of the data falls at or below it .

 Example: A test score at the 30th percentile means that 30% of the test scores are lower and 70% are higher.

**The Quartiles:**

 The quartiles break a data set into quarters.

To calculate the quartiles:

1. Find the median, Q2
2. The first Quartile, Q1, is the median of the bottom half of the data.
3. The third Quartile, Q3, is the median of the bottom half of the data.

**Example 10:**

Find the quartiles for the highway mileage of the two-seater cars listed below.

 13 15 16 16 17 19 20 22 23 23 23 24 25 25 26 28 28 28 29 32

**Five-Number Summary and Boxplots:**

Five-Number Summary: This summary is used to help describe the spread or variation found in the data set.

* This summary is used to help describe the spread or variation found in the data set.

Consists of:

 the minimum, Q1 , Median (Q2), Q3 , the maximim

**Example 11:** What is the five-number summary for the two-seater cars?

* **Box Plots**: A box plot is used to show the variation or spread of the data. Also called a bow and whisker plot.
* A box plot is a graph of the five-number summary. It is drawn above or below a number line.
* The Quartiles are used to draw a box. Mark the median with a line in the box.
* Draw a line to connect the box with the max and min.

**Example 12:** Draw the box plot for the two-seater cars.

Because boxplots show less detail than histograms or stemplots, they are best used for side-by-side comparison of more than one distribution. When you look at a boxplot, first locate the median, which marks the center of the distribution. Then look at the spread. The quartiles show the spread of the middle half of the data, and the extremes (the smallest and largest observations) show the spread of the entire data set. We see at once that city mileages are lower than highway mileages. The minicompact cars have slightly higher median gas mileages than the two-seaters, and their mileages are markedly less variable. In particular, the low gas mileages of the Ferraris and Lamborghinis in the two-seater group pull the group minimum down.



**Interpreting a box plot:**

1. 50% of the data falls in the box.
2. Each whisker contains 25% of the data.
3. It is possible for 2 of the numbers in the 5-number summary to be equal.

 

Q1 and the minimum are the same.

The median is equal to either Q1 or Q3

1. We can see outliers fairly easily; an extremely long whisker indicates that an outlier is likely on that end of the data set.



1. We can also get an idea of the shape of the distribution.

\* The box plot in part 4 is most likely skewed right.

 \* This box plot is fairly symmetrical.



**The 1.5 IQR Outlier Rule for suspected outliers:**

* **Interquartile Range (IQR):** the difference between the third and first quartiles.
	+ IQR = Q3-Q1

  The IQR is resistant to outliers. Why?

* **No single numerical measure of spread, such as *IQR*, is very useful for describing skewed distributions.**

**No single numerical measure of spread, such as *IQR*, is very useful for describing skewed distributions.**

**We can detect skewness by examining and comparing the lengths of the first and fourth quarters of the data set.**

* The 1.5IQR Rule for Outliers: Call an observation a suspected outlier if it falls more than 1.5IQR above the third quartile or below the first quartile.

Note: Modified box plots on the calculator are using this test!

**Example 13:** Apply the 1.5IQR Rule to the two-seater data. Do we have any suspected outliers?

**Standard Deviation:**

The five-number summary is not the most common numerical description of a distribution. That distinction belongs to the combination of the mean to measure center and the ***standard deviation*** to measure spread. The standard deviation measures spread by looking at how far the observations are from their mean.

The idea behind the variance and the standard deviation as measures of spread is as follows. The deviations *xi* − *x* display the spread of the values *xi* about their mean *x*. Some of these deviations will be positive and some negative because some of the observations fall on each side of the mean. In fact, *the sum of the deviations of the observations from their mean will always be zero.* Squaring the deviations makes them all positive, so that observations far from the mean in either direction have large positive squared deviations. The variance is the average squared deviation. Therefore, *s*2 and *s* will be large if the observations are widely spread about their mean, and small if the observations are all close to the mean.

**Variance and Standard Deviation:**

Sample: Variance , s2 = population: variance, σ2 =

Standard deviation, s= standard deviation, σ =

**Example 14:** A person's metabolic rate is the rate at which the body consumes energy. Metabolic rate is important in studies of weight gain, dieting, and exercise. Here are the metabolic rates of 7 men who took part in a study of dieting. (The units are calories per 24 hours. These are the same calories used to describe the energy content of foods.)

  1792 1666 1362 1614 1460 1867 1439

Enter these data into your calculator or software and verify that = 1600 calories  *s* = 189.24 calories. Interpret these values in the context of the problem.

The idea of the variance is straightforward: it is the average of the squares of the deviations of the observations from their mean. The details we have just presented, however, raise some questions.

***Why do we square the deviations?*** Why not just average the distances of the observations from their mean? There are two reasons, neither of them obvious. First, the sum of the squared deviations of any set of observations from their mean is the smallest such sum possible. The sum of the unsquared distances is always zero. So, squared deviations point to the mean as center in a way that distances do not. Second, the standard deviation turns out to be the natural measure of spread for a particularly important class of symmetric unimodal distributions, the *Normal distributions*. We will meet the Normal distributions in the next chapter. We commented earlier that the usefulness of many statistical procedures is tied to distributions of particular shapes. This is distinctly true of the standard deviation.

***Why do we emphasize the standard deviation rather than the variance?*** One reason is that *s,* not *s*2, is the natural measure of spread for Normal distributions. There is also a more general reason to prefer *s* to *s*2. Because the variance involves squaring the deviations, it does not have the same unit of measurement as the original observations. The variance of the metabolic rates, for example, is measured in squared calories. Taking the square root remedies this. The standard deviation *s* measures spread about the mean in the original scale.

***Why do we “average” by dividing by n − 1 rather than n in calculating the variance?*** Because the sum of the deviations is always zero, the last deviation can be found once we know the other *n* − 1. So we are not averaging *n* unrelated numbers. Only *n* − 1 of the squared deviations can vary freely, and we average by dividing the total by *n* − 1. The number *n* − 1 is called the ***degrees of freedom*** of the variance or standard deviation. Many calculators offer a choice between dividing by *n* and dividing by *n* − 1, so be sure to use *n* − 1.

**Properties of the standard deviation:**

s measures spread about the mean and should only be used when the mean is an appropriate measure of center.

s = 0 means there is no spread!

s > 0 always since we square all the deviations!

s is not resistant

**Example 15:** Find the standard deviation for the two-seater data. Then, drop the outlier and recalculate. What do we see?

**The use of squared deviations renders *s* even more sensitive than *x* to a few extreme observations.** For example, dropping the Honda Insight from our list of two-seater cars reduces the mean highway mileage from 24.7 to 22.6 mpg. It cuts the standard deviation by more than half, from 10.8 mpg with the Insight to 5.3 mpg without it. Distributions with outliers and strongly skewed distributions have large standard deviations. The number *s* does not give much helpful information about such distributions.

**Choosing measures of center and spread:**

 How do we choose between the five-number summary and *x* and *s* to describe the center and spread of a distribution? Because the two sides of a strongly skewed distribution have different spreads, no single number such as *s* describes the spread well. The five-number summary, with its two quartiles and two extremes, does a better job.

If the distribution is fairly symmetric with no extreme values: use the mean and standard deviation.

If the distribution has an outlier or is skewed: use the median, the 5-number summary and the IQR

Remember, there is no cut and dried method! Be able to defend your choices!

**Remember that a graph gives the best overall picture of a distribution. Numerical measures of center and spread report specific facts about a distribution, but they do not describe its entire shape.** Numerical summaries do not disclose the presence of multiple modes or gaps, for example. **Always plot your data.**

**Changing units (linear Transformations):**

**Example 16:**  A survey was conducted asking a sample of 10 parents how many hours a week their children spent watching TV. The results were: 6, 7, 15, 24, 8, 19, 24, 13, 30, 5.Find the mean, variance and sample standard deviation for the data.

If we create the histogram for the hours of TV watched, we get the following picture:



What will happen to the graph if we cut 2 hours off each child’s time?

How does cutting 2 hours of time off each child’s viewing time affect the mean and the standard deviation? Explain.

What if we decide to cut each child’s viewing time by 10% rather than the 2 hours? How will this change the graph of the data? How will it affect the mean and the standard deviation? Explain.

In general, if we add or subtract a constant form a data set the mean will change by the same constant. The standard deviations will remain the same since we have adjusted all of the data entries by the same amount.

If we multiply or divide each data entry by a constant, then the mean and the standard deviation will both change by the same constant. Since each data entry changed by a non-constant amount, the shape and spread of the distribution has also changed. This means both the mean and standard deviation must be adjusted.

The same variable can be recorded in different units of measurement. Americans commonly record distances in miles and temperatures in degrees Fahrenheit, while the rest of the world measures distances in kilometers and temperatures in degrees Celsius. Fortunately, it is easy to convert from one unit of measurement to another. This is true because a change in the measurement unit is a ***linear transformation*** of the measurements.

*Linear transformations do not change the shape of a distribution.* If measurements on a variable *x* have a right-skewed distribution, any new variable *x*new obtained by a linear transformation *x*new = *a* + *bx* (for *b* > 0) will also have a right-skewed distribution. If the distribution of *x* is symmetric and unimodal, the distribution of *x*new remains symmetric and unimodal.

Although a linear transformation preserves the basic shape of a distribution, the center and spread may change. Because linear transformations of measurement scales are common, we must be aware of their effect on numerical descriptive measures of center and spread. Fortunately, the changes follow a simple pattern.

To see the **effect of a linear transformation** on measures of center and spread, apply these rules:

* Multiplying each observation by a positive number *b* multiplies both measures of center (mean and median) and measures of spread (interquartile range and standard deviation) by *b.*
* Adding the same number *a* (either positive, zero, or negative) to each observation adds *a* to measures of center and to quartiles but does not change measures of spread.

The measures of spread *IQR* and *s* do not change when we add the same number *a* to all of the observations because adding a constant changes the location of the distribution but leaves the spread unaltered. You can find the effect of a linear transformation *x*new = *a* + *bx* by combining these rules. For example, if *x* has mean *x*, the transformed variable *x*new has mean *a* + *bx*.

**Comparing Distributions:**

 **Parallel Box Plots:** Are used to compare 2 sets of data. They must be drawn on the same number line if we wish to compare them.

**Example 17*:*** Using the data below draw parallel box plots.

Number of homeruns hit by Babe Ruth: 54 59 35 41 46 25 47 60 54 46 49 41 34 22

 Number of homeruns hit by Roger Maris: 10 13 26 16 33 61 28 29 14 8

The top boxplot represents Babe Ruth. (22, 35, 46, 54, 60)

The bottom boxplot represents Roger Maris. (8, 13, 21, 29, 61)

 

 Who is the better homerun hitter? Justify your choice.

 Did either man have any outliers in his data?

 Who was the more consistent hitter?

**Other ways to compare data:**

1. Back to Back Stem-and-leaf plots
2. Dot plots using the same scale on the number line.

Data Analysis Toolbox

To answer a statistical question of interest involving one or more data sets, proceed as follows.

* **Data** Organize and examine the data. Answer the ***key questions*** from the Preliminary Chapter:
	1. **Who** are the individuals described by the data?
	2. **What** are the variables? In what units is each variable recorded?
	3. **Why** were the data gathered?
	4. **When, where, how, and by whom** were the data produced?
* **Graphs** Construct appropriate graphical displays.
* **Numerical summaries** Calculate relevant summary statistics.
* **Interpretation** Discuss what the data, graphs, and numerical summaries tell you in the context of the problem. Answer the question!