

## **Due Date Deadline: Tuesday August 29<sup>th</sup>**

Dear Prospective AP and Honors Statistics Student,

First of all, congratulations on making a fantastic decision in choosing to take AP/Honors Statistics, the study of collecting, organizing, analyzing, and drawing conclusions with data, a course that can and should change the way you see the world and how you make decisions.

The below readings and videos will serve as your main source of notes for Unit 1 (reference the digital version online to see the colors in the graphs more easily). This assignment has been created for you to give evidence of your ability- in writing- to express your ability to learn from reading. It will help prepare you for the reading, writing, and reasoning that is required in the upcoming course by teaching you the course's first unit. Embedded in the reading are two problem sets that you should complete to apply your learning.

Additionally, a video is linked to this document to talk you through some key concepts and to verbally support your learning. The last few pages of this packet are a guided notes sheet that you can use while you follow along with the videos.

Video link: *For some reason, tech issues split the video into two clips. Links are below.*

Part 1:

[https://drive.google.com/file/d/14rUzBru2rt\\_vsv3mBV22FpldrtNZ9cSZ/view?usp=sharing](https://drive.google.com/file/d/14rUzBru2rt_vsv3mBV22FpldrtNZ9cSZ/view?usp=sharing)

Part 2:

<https://drive.google.com/file/d/1dXJY90q20YRUXpzLDuSZp3j--ePI7YqQ/view?usp=sharing>

**Note-taking tip!** Annotate your notes! Don't just copy down what I write, but instead, add annotations with reminders and key details so that they jog your memory when you look back at the notes.

In addition to watching the videos, your job is to prove that you can learn the given content through several readings of the following text. I should reemphasize that the learning of the material should be your goal. [Feel free to supplement your learning using YouTube and other internet sites to look up information you are uncertain about after reading as necessary, or email me with questions.] Produce thorough responses to the provided problems to exhibit your learning, while referring back to the reading and videos as necessary. **Completed summer assignments will serve as your materials for an "open-note" assessment during the first week of class.** For examples of model answers, see those given in the reading.

You will need a graphing calculator for this course, but do not need one to complete the summer assignment. If you have been using a TI-83, TI-84 or TI-89, that will work. If you will also be taking Calculus in the fall and will be using a TI-Nspire, I recommend using that for Statistics as well. I will be using/teaching from the TI-Nspire and will provide one for you in the fall.

Success in the first unit of the class is contingent upon this assignment's successful completion by the due date. Your first exam will be on this material within two weeks of the school year's start.

I can be contacted regarding the course or for assistance with the packet at the email address below as you progress. If you are unsure of what is expected with this packet, contact me.

Please join the remind group for the class for an easy way to contact me, and for reminders throughout the school year. Text @wyoapstat to 81010.

Looking forward to meeting and working with you soon,

Miss Tierney  
[mtierney@wyoarea.org](mailto:mtierney@wyoarea.org)

## SECTION ONE: **Displaying Quantitative Data**

**E**nron Corporation was once one of the world's biggest corporations. From its humble beginnings as an interstate natural gas supply company in 1985, it grew steadily throughout the 1990s, diversifying into nearly every form of energy transaction and eventually dominating the energy trading business.

Its stock price followed this spectacular growth. In 1985 Enron stock sold for about \$5 a share, but at the end of 2000, Enron stock closed at a 52-week high of \$89.75 and the company's stock was worth more than \$6 billion. Less than a year later it hit a low of \$0.25 a share, having lost more than 99% of its value. Many employees who had taken advantage of generous stock plans lost retirement packages worth hundreds of thousands of dollars. Just how volatile was Enron stock? And were there hints of trouble that might have been seen?

Rather than look at the stock prices themselves, let's look at how much they changed from month to month. For example, on February 3 (the first trading day of the month) of 1997, Enron stock sold for \$0.75 less than it had on January 2 (the first trading day of that month). Table 4.1 gives the monthly changes in stock price (in dollars) for the 5 years leading up to the company's failure.



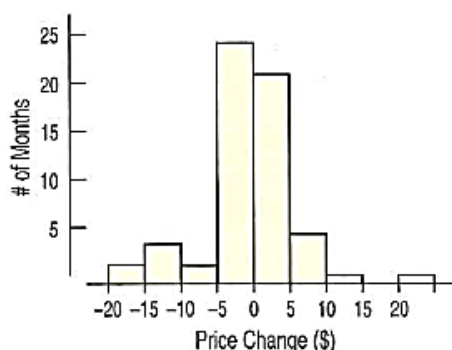
	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1997	-\$1.44	-0.75	-0.69	-0.88	0.12	0.75	0.81	-1.75	0.69	-0.22	-0.16	0.34
1998	0.78	0.62	2.44	-0.28	2.22	-0.50	2.06	-0.88	-4.50	4.12	1.16	-0.50
1999	3.28	3.34	-1.22	0.47	5.62	-1.59	4.31	1.47	-0.72	-0.38	-3.25	0.03
2000	5.72	21.06	4.50	4.56	-1.25	-1.19	-3.12	8.00	9.31	1.12	-3.19	-17.75
2001	14.38	-1.08	-10.11	-12.11	5.84	-9.37	-4.74	-2.69	-10.61	-5.85	-17.16	-11.59

Monthly stock price change in dollars of Enron stock for the period January 1997 to December 2001. Negative amounts indicate that the stock lost value. **Table 4.1**

It's hard to tell very much from the data. Don't try too hard. Tables with lots of numbers are hard to understand by just reading them. You might get a rough idea of how much the stock changed from month to month, but even that would be approximate. "Looks like it's usually less than 10 dollars in either direction," you might say.

Instead, let's follow the first rule of data analysis and make a picture. What kind of picture should we make? It can't be a bar chart or a pie chart. Those are for categorical variables, and the values here are in dollars. We'll want to treat price change as a *quantitative* variable, not a categorical one.

## Histograms: Displaying the Distribution of Price Changes



**Monthly price changes of Enron stock**  
This histogram displays the distribution of price changes by showing how many months had price changes in each of the intervals (bins). **Figure 4.1**

How can we display a quantitative variable? With a categorical variable, life was easy. When there were only a few categories, we could make one pile for each.

With quantitative variables, life is different. Usually we can't list all the individual values; there are just too many of them. So instead, we slice up the entire span of values covered by the quantitative variable into equal-width piles called **bins**. Then we count the number of values that fall into each bin. The bins and the counts in each bin give the **distribution** of the quantitative variable.

We can display the bin counts in a display called a histogram. Like a bar chart, a **histogram** plots the bin counts as the heights of bars. For the Enron data, the cases are months, so the height of each bar shows the number of months that have price changes falling into that bin. Here, the *Who* are months and the *What* are price changes. The histogram displays the distribution of price changes by showing the number of *months* that have price changes in each of the bins.

Does the distribution of *Price Change* look as you expected? It's often a good idea to *imagine* what the distribution might look like before you ask a computer or calculator to make a picture for you. That way you'll be ready

if the picture shows something unexpected, and less likely to let bad data or wrong data fool you.

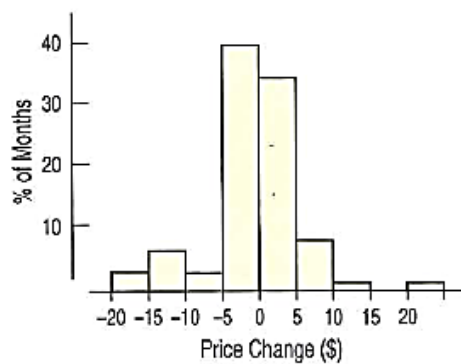
The first bar, which counts losses between \$15 and \$20, has only two months in it. We can see that although they vary, most of the monthly price changes are smaller than \$5 in either direction. Only in a very few months were the changes larger than \$10 in either direction. There appear to be about as many positive as negative price changes.

### Determining bins by hand

The first column of the Enron data shows monthly price changes of -\$1.44, \$0.78, \$3.28, \$5.72, and \$14.38. Each observation becomes a count in one of the bins of the histogram. In Figure 4.1 we've made each bin \$5 wide. So, the first data point (-\$1.44) goes into the bin from -\$5 to \$0. The next one, \$0.78, goes into the bin from \$0 to \$5 as does the third value. \$5.72 is bigger than 5 and less than 10, so it goes into the \$5 to \$10 bin, and \$14.38 goes into the \$10 to \$15 bin. Continuing this for the entire data set, we find four data values between -\$15 and -\$10. The histogram shows this as a bar of height 4 for that bin. We also see that there are 5 times as many price changes between \$0 and \$5, something that we couldn't easily see from the table.

*Where do values on the borders of the bins go? The choice is up to you, but you have to be consistent. The standard rule is to put values at the edge of the bins into the next higher bin, placing \$10.00 in the \$10 to \$15 bin rather than in the \$5 to \$10 bin.*

Even if you use technology to make the histogram, you may have a choice of how many bins to use. Unless the data set is very large, you'll probably want to have between 5 and 20 bins, so you'll have to choose the width of the bin with that in mind. You might also want to choose the bins so that the border numbers come out "nicely." Always be sure that you use the *same* bin width for all the bins so that the display preserves the area principle.



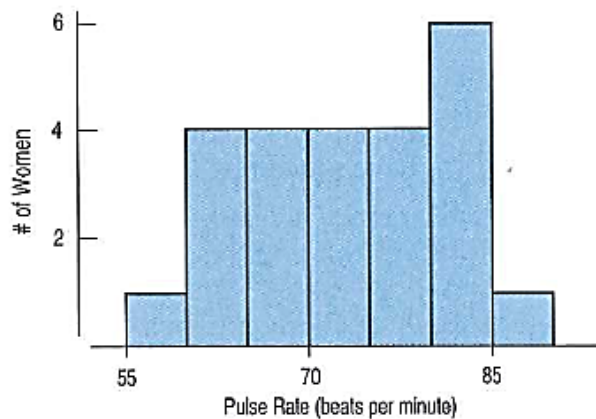
A bar chart has spaces between the bars because the categories could appear in any order. But in a histogram, there are no spaces because the bins slice up *all the values* of a quantitative variable. (A gap, as seen in Figure 4.2 between 15 and 20, indicates that no values fell in that bin.) Both kinds of display satisfy the area principle because the area covered by each bar corresponds to the count of the cases falling in the range covered by that bar. Sometimes it is useful to make a **relative frequency histogram**, replacing the counts on the vertical axis with the *percentage* of the total number of cases falling in each bin. Of course, the shape of the histogram is exactly the same; only the labels are different.

A relative frequency histogram is faithful to the area principle by displaying the percentage of cases in each bin instead of the counts. **Figure 4.2**

## Stem-and-Leaf Displays

[A.K.A. STEMPLOTS]

Histograms provide an easy-to-understand summary of the distribution of a quantitative variable, but they don't show the data values themselves. Here's a histogram of the pulse rates of 24 women, taken by a researcher at a health clinic:



The pulse rates of 24 women at a health clinic. **Figure 4.3**

The story seems pretty clear. We can see the entire span of the data and can easily see what a typical pulse rate might be. But is that all there is to these data?

A stem-and-leaf display is like a histogram, but it shows the individual values. It's also easier to make by hand. Here's a stem-and-leaf display of the same data:

```

8 | 8
8 | 000044
7 | 6666
7 | 2222
6 | 8888
6 | 0444
5 | 6
Pulse Rate
(8|8 means 88 beats/min)

```

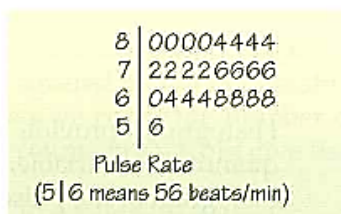
Turn the stem-and-leaf on its side (or turn your head to the right) and squint at it. It should look roughly like the histogram of the same data. Does it? Well, it's



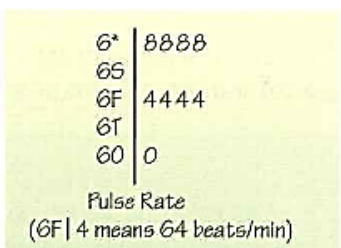
backwards because now the higher values are on the left, but other than that it has the same shape.<sup>1</sup>

What does the line that says 8|8 at the top of the display mean? It stands for a pulse of 88 beats per minute. We've taken part of the number (the "tens" place) and made that the "stem." Then we've sliced off the trailing digit (the "ones" place) and turned it into a "leaf." The next line down is 8|000044. That shows that there were four pulse rates of 80 and two pulse rates of 84 bpm. As you can see, we always label our plot and include a key explaining how to interpret the numbers.

To make a stem-and-leaf display, we cut each data value into leading digits (which become the "stem") and trailing digits (the "leaves"). Then we use the stems to label the *bins*. For the pulse rate data, we chose the first digit as the stem and displayed each stem on two lines (one with digits 0-4 and the other with 5-9). We could have used only one line for each stem. It would have used only 4 bins, but it looks a bit crowded:



By splitting the stems in our original plot, we got a better look at the distribution. Sometimes, we may want to go further. If we decide that splitting the stem in two lines is not enough, we could try splitting it into 5, putting the digits 8 and 9 on the top line, 6 and 7 on the next, etc. This is too many bins for these data, but we'd have something like this (for the pulses from 60 to 69):



Here we've used letters to show which digits will appear on each line as leaves: O for 0 and One, T for Two and Three, F for Four and Five, S for Six and Seven and \* (oops, we were doing so well . . .) for 8 and 9.

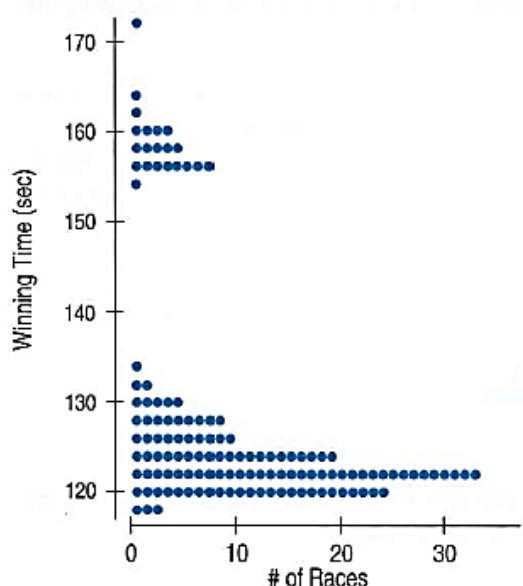
If we had had more digits reported (if the pulses had been reported to the nearest tenth bpm, like 56.2, 57.8, etc.) we would still use only *one* digit for the leaf, rounding the data values to one decimal place after the stem before splitting them into stem and leaves.<sup>2</sup> Or, we might have decided to use each value 56, 57, 58 as the stem and use the number after the decimal place as the leaf, but we would have needed a lot of values for that to make sense.

Does a stem-and-leaf display satisfy the area principle? It does as long as each digit takes up the same amount of space. When you make a stem-and-leaf display by hand, be careful to write thin numerals like “1” and fat ones like “3” to take up the *same amount* of horizontal space. That way, each bar’s length will be proportional to the number of observations that fall into its bin.

**Stem-and-leaf displays** contain all the information found in a histogram and, when carefully drawn, satisfy the area principle and show the distribution. In addition, stem-and-leaf displays preserve the individual data values. Few data displays can do all this as effectively. Unlike a histogram, stem-and-leaf displays also show the digits in the bins, so they can reveal unexpected patterns in the data. Did you notice anything about the pulse rates? How many seconds do you think the nurse waited while counting beats? It is clear at a glance that all the values are even. But look a bit closer. Every pulse rate is divisible by 4, something we couldn’t possibly tell from the histogram. The nurse probably waited only 15 seconds and multiplied by 4 rather than counting for a whole minute.

Unlike most other displays discussed in this book, stem-and-leaf displays are great pencil-and-paper constructions. Consider using them whenever you have a quantitative variable. They are well suited to moderate amounts of data—say, between ten and a few hundred values. For larger data sets, histograms do a better job.

## Dotplots



A dotplot of Kentucky Derby winning times plots each race as its own dot, showing the bimodal distribution. **Figure 4.4**

A **dotplot** is a simple display. It just places a dot along an axis for each case in the data. It’s like a stem-and-leaf display, but with dots instead of digits for all the leaves. Dotplots are a great way to display a small data set (especially if you forget how to write the digits from 0 to 9). Here’s a dotplot of the time (in seconds) that the winning horse took to win the Kentucky Derby in each race between the first Derby in 1875 and the 2004 Derby.

Dotplots show basic facts about the distribution. We can find the slowest and quickest races by finding times for the topmost and bottommost dots. It’s also clear that there are two clusters of points, one just below 160 seconds and the other at about 122 seconds. Something strange happened to the Derby times. Once we know to look for it, we can find out that in 1896 the distance of the Derby race was changed from 1.5 miles to the current 1.25 miles. That explains the two clusters of winning times.

Some dotplots stretch out horizontally, with the counts on the vertical axis, like a histogram. Others, such as the one shown here, run vertically, like a stem-and-leaf display. Some dotplots place points next to each other when they would otherwise overlap. Others just place them on top of one another. Newspapers sometimes offer dotplots with the dots made up of little pictures.



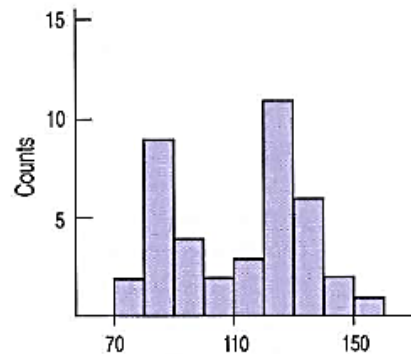
# Shape, Center, and Spread

The **mode** is sometimes defined as the single value that appears most often. That definition is fine for categorical variables because all we need to do is count the number of cases for each category. For quantitative variables, the mode is more ambiguous. What's the mode of the Enron data? No price change occurred more than twice, but two months had drops of \$0.50. Should that be the mode? Probably not. It makes more sense to use the word "mode" in the more general sense of peak in a histogram, rather than as a single summary value.

Step back from a histogram or stem-and-leaf display. What can you say about the distribution? When you describe a distribution, you should always tell about three things: its **shape**, **center**, and **spread**.

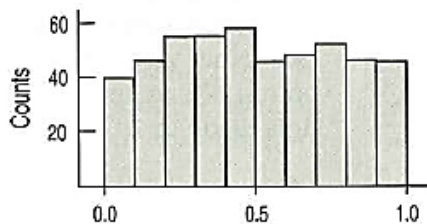
## What Is the Shape of the Distribution?

1. Does the histogram have a single, central hump or several separated bumps? These humps are called **modes**.<sup>3</sup> The Enron stock price changes have a single mode at just about \$0. A histogram with one main peak, such as the price changes, is dubbed **unimodal**; histograms with two peaks are **bimodal**, and those with three or more are called **multimodal**.<sup>4</sup> For example, here's a bimodal histogram.



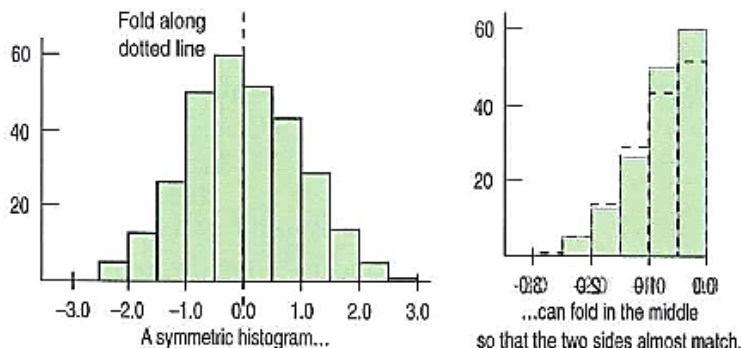
A bimodal histogram has two apparent peaks. **Figure 4.5**

A histogram that doesn't appear to have any mode and in which all the bars are approximately the same height is called **uniform**.



In a uniform histogram, the bars are all about the same height. The histogram doesn't appear to have a mode. **Figure 4.6**

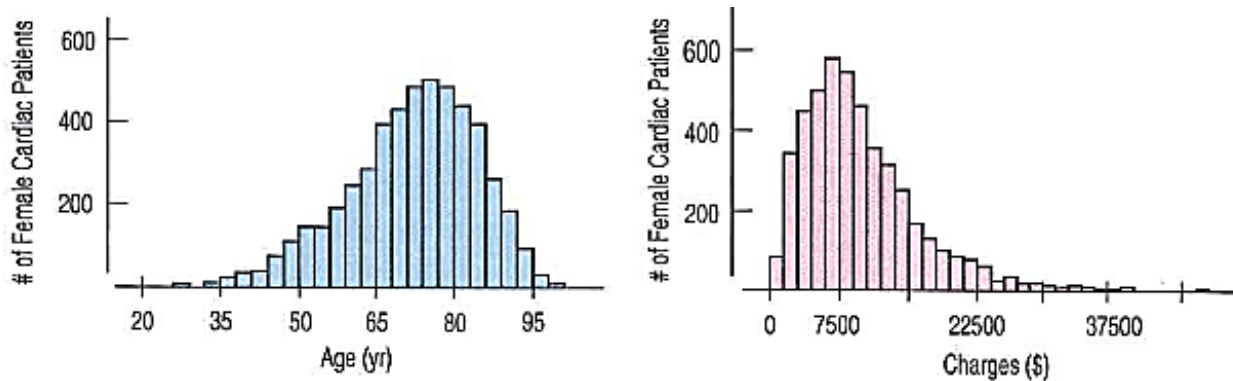
2. Is the histogram **symmetric**? Can you fold it along a vertical line through the middle and have the edges match pretty closely, or are more of the values on one side?



**Figure 4.7**

The (usually) thinner ends of a distribution are called the **tails**. If one tail stretches out farther than the other, the histogram is said to be **skewed** to the side of the longer tail.



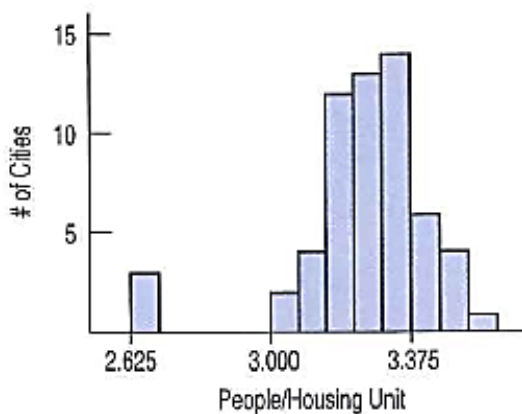


Two skewed histograms showing data on two variables for all female heart attack patients in New York state in one year. The blue one (age in years) is skewed to the left. The purple one (charges in \$) is skewed to the right. **Figure 4.8**

3. *Do any unusual features stick out?* Often such features tell us something interesting or exciting about the data. You should always mention any stragglers, or **outliers**, that stand off away from the body of the distribution. If you're collecting data on nose lengths and Pinocchio is in the group, you'd probably notice him, and you'd certainly want to mention it.

Outliers can affect almost every method we discuss in this course. So we'll always be on the lookout for them. An outlier can be the most informative part of your data. Or it might just be an error. But don't throw it away without comment. Treat it specially and discuss it when you tell about your data. Or find the error and fix it, if you can. Be sure to look for outliers. Always.

(In the next chapter you'll learn a handy rule of thumb for deciding when a point might be considered an outlier.)



A histogram with outliers. There are three cities in the leftmost bar.

**Figure 4.9**

Are there any **gaps** in the distribution? The Kentucky Derby data that we saw in the dotplot on page 49 has a large gap between two groups of times, one near 120 seconds and one near 160. Gaps help us see multiple modes, and encourage us to notice when the data may come from different sources or contain more than one group.

● **Does some of this seem vague? It is!** How you characterize a distribution is often a judgment call, and it's always relative. Is that point way out on the right really an outlier, or is it just an indication of a long tail to the distribution? Generally, we start by looking at the main body of the data. If a point is separated from it by only a small gap, then it's not that unusual and probably not an outlier. If the main body seems roughly symmetric, then more distant stragglers are best regarded as outliers. If the main body of the data is skewed, then the long tail that continues that skewness is part of the overall pattern, so points would need to be farther away to be called outliers.

It may surprise you (especially if you thought this was going to be a math class) that many of the most important concepts in Statistics aren't as precise as in mathematics. When we summarize data, we want to emphasize overall features. We don't want to focus on all the *details* of the data set we're looking at, so some of the statistical concepts are deliberately left vague.

Whether a histogram is symmetric or skewed, whether it has one or more modes, whether a point is far enough from the body of the plot to be an outlier—these are all somewhat vague concepts. You may be uncomfortable with this at first, because you're used to finding a correct, precise answer. In Statistics there may be more than one right answer. That means you're entitled to your own (informed) opinion, provided you can justify it.

Keep an eye out for vague concepts in Statistics. We'll point out when concepts are vague and help you to navigate through them so you can understand the story told by the data. In fact, here come two more. . . . ●

## Where Is the Center of the Distribution?

If you had to pick a single number to describe all the data, what would you pick? The center is an easy description of a typical value and a concise summary of the whole batch of numbers. When a histogram is unimodal and symmetric, it's easy to find its center. It's right in the middle. (Where else would you look?) The center of the Enron price changes is about \$0. That tells us that over the period we've examined, the stock went down about as often as it went up.

For distributions with other shapes, the situation isn't as clear. If the histogram is skewed, defining the center is more of a challenge. And if the histogram has more than one mode, the center might not even be a useful concept. You might be looking at two different groups thrown together, so it's probably a good idea to find out why you don't have a single mode.

The next chapter discusses some ways to locate centers numerically. For now we'll just eyeball a picture of the distribution and give a rough idea of where the center seems to be.

Why do banks favor a single line that feeds several teller windows rather than separate lines for each teller? The average waiting time is the same. But the time you can expect to wait is less variable when there is a single line, and people prefer consistency.

## How Spread Out Is the Distribution?

The center gives a typical value, but not everyone is typical. Variation matters. Statistics is about variation, but how can we describe it? We can look to see whether all the values of the distribution are tightly clustered around the center or spread out. Because distributions that vary a lot around the center are harder to predict or model, we often prefer distributions with less variability. Would you rather invest in a stock whose price gyrates wildly or one that grows steadily?<sup>5</sup>

You're not finished describing a distribution until you discuss its spread. Don't worry. We'll have a lot more to say about spread in the next two chapters.

## Displaying Quantitative Data Step-By-Step

With the current state of the economy,<sup>6</sup> more attention has been paid to the compensation of chief executive officers (CEOs) of major companies. Let's look at the CEO salaries of the 800 largest corporations in 1994.

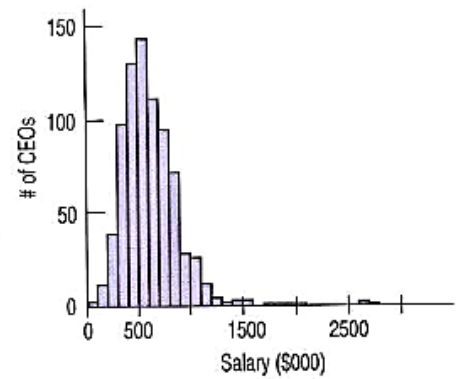


## Show

**Mechanics** We almost always make histograms with a computer or graphing calculator.

## REALITY CHECK

It is always a good idea to think about what we expected to see and to check whether the histogram is close to what we expected. Could CEOs really earn this much? Yes, six-to seven-figure salaries are about what we might expect.



## Tell

**Conclusion** Describe the shape, center, and spread of the distribution. Be sure to report on the symmetry, number of modes, and any gaps or outliers. Remember to interpret all results in context.

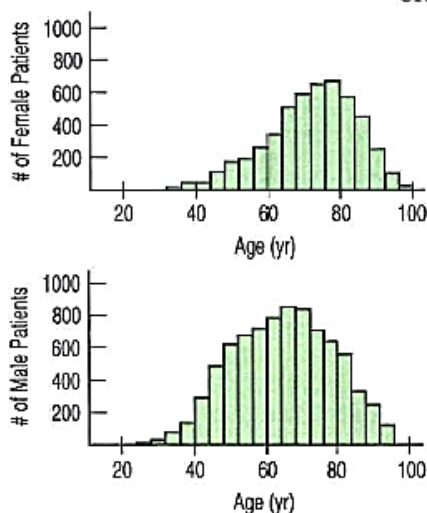
The main body of the distribution is unimodal and nearly symmetric,<sup>7</sup> centered around \$500,000, with slightly more than half of CEOs earning salaries higher than that. But there are some high outliers. Some CEOs' salaries are higher than what is typical for most CEOs of large corporations. Even though the vast majority of CEOs have salaries below \$1,000,000 a year, there are a few with salaries between \$2,500,000 and \$3,000,000 a year.

## Comparing Distributions

Up to now, we've looked at one distribution at a time. But this route can take us only so far. While it may be interesting to know the distribution of CEO salaries, it might be more interesting to know how salaries in high-tech industries compare with those in manufacturing. The fact is that most interesting results about data involve making comparisons or modeling relationships.

For example, many common diseases show different patterns in women and men. In the past couple of decades, researchers have been more careful to collect data on women. What kinds of questions might we ask about these data? Here's one: Do men and women tend to get heart attacks at different ages? That's the

kind of question that's well suited to investigating with a graph. Here are two histograms of the ages of every heart attack patient hospitalized during 1993 in New York State, one for women and one for men.



The distributions of ages for female and male heart attack patients differ in interesting ways. **Figure 4.10**

<b>WHO</b>	Heart attack patients
<b>WHAT</b>	Age (years)
<b>WHEN</b>	1993
<b>WHERE</b>	NY state

What can we tell from these histograms? We'll start, as usual, by looking at shape, center, and spread.

Notice first that the *shapes* are different. The men's distribution is nearly symmetric, while the women's is clearly skewed to the left. Young women are much less likely to have heart attacks than young men. The *center* of the men's distribution appears to be in the low 60s, but for women, it's closer to 70. So, men tend to have heart attacks at an earlier age. The women's ages are highly clustered between 60 and 85 years, unlike for the men.

In general, the age at which a man might have a heart attack is less predictable than for a woman because the men's ages are more *spread* out than the women's.



## Comparing Infant Death Rates Step-By-Step

In 2001 the infant death rate in the United States was 6.8 deaths per 1000 live births. How does the rate differ from region to region? The Kaiser Family Foundation collected data from all 50 states and the District of Columbia, allowing us to compare infant death rates in the Northeast and Midwest to those in the South and West.

Since there are only 51 data values, a back-to-back stem-and-leaf plot is an effective display of these distributions. We simply run the stems down the middle of the plot, then place leaves for states in the Northeast and Midwest on the right and those for states in the South and West on the left. Look at the plot in the step-by-step comparison below. You'll see that the highest rate among southern and western states was 10.7 infant deaths per 1000 live births, compared to a high of 8.8 deaths per thousand for one of the northeastern or midwestern states.

### Show

**Mechanics** Be sure to label the display clearly. Be careful when you read the values on the left: 8|4| means 4.8 deaths per 1000.

### Tell

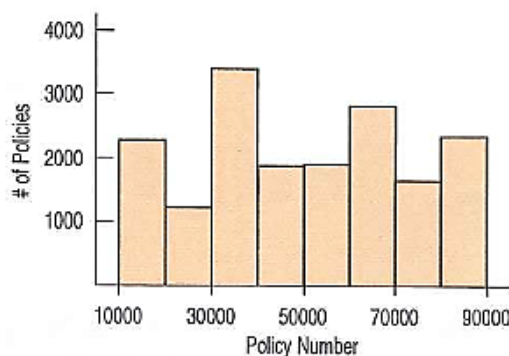
**Conclusion** Compare the shapes, centers, and spreads of the two distributions. Also mention any unusual features, and be sure to interpret everything in context. If you had the original table of these data showing the states' names, you should say in your report that the unusually low rate was for New Hampshire.

Infant Death Rates (by state), 2001	
South and West	Northeast and Midwest
5 6 7	10
4 8	9
1 9 7 3 5 1 6	8 8 0
3 6 2 3	7 2 7 5 4 1 4 4
7 9 2 2 4	6 8 5 1 1 8
8 8 9 9 4 4 9 7	5 8 5 0 6 3
8	4
	3 8
	2
	1

(3|8| means 3.8 deaths per 1000 live births)

In general, infant death rates appear to have been somewhat higher for states in the South and West than in the Northeast and Midwest. The distribution is roughly symmetric for the northeastern and midwestern states, but may be slightly skewed to the right for the South and West. Nationally, most states had infant death rates between 5 and 9 deaths per 1000 live births, but the rates varied more widely in the West and South, where 5 states had rates above 9. Infant death rates were more consistent in the Northeast and Midwest; no states were above 9, but one state had an unusually low 3.8 infant deaths per 1000 live births.

## What Can Go Wrong?



It's not appropriate to display these data with a histogram. **Figure 4.14**

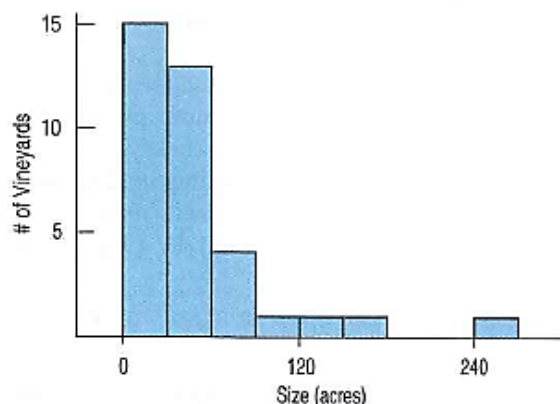
- **Don't make a histogram of a categorical variable.** Just because the variable contains numbers doesn't mean it's quantitative. Here's a histogram of the insurance policy numbers of some workers. It's not very informative because the policy numbers are categorical. A histogram or stem-and-leaf display of a categorical variable makes no sense. A bar chart or pie chart may do better.
- **Don't look for shape, center, and spread of a bar chart.** A bar chart showing the sizes of the piles displays the distribution of a categorical variable, but the bars could be arranged in any order left to right. Concepts like symmetry, center, and spread make sense only for quantitative variables.

# TERMS

<b>Distribution</b>	The distribution of a variable gives <ul style="list-style-type: none"><li>• the possible values of the variable.</li><li>• the frequency or relative frequency of each value.</li></ul>
<b>Histogram (relative frequency histogram)</b>	A histogram uses adjacent bars to show the distribution of values in a quantitative variable. Each bar represents the frequency (or relative frequency) of values falling in an interval of values.
<b>Stem-and-leaf display</b>	A stem-and-leaf display shows quantitative data values in a way that sketches the distribution of the data. It's best described in detail by example.
<b>Dotplot</b>	A dotplot graphs a dot for each case against a single axis.
<b>Shape</b>	To describe the shape of a distribution, look for <ul style="list-style-type: none"><li>• single vs. multiple modes.</li><li>• symmetry vs. skewness.</li></ul>
<b>Center</b>	A value that attempts the impossible by summarizing the entire distribution with a single number, a "typical" value.
<b>Spread</b>	A numerical summary of how tightly the values are clustered around the "center."
<b>Mode</b>	A hump or local high point in the shape of the distribution of a variable is called a "mode." The apparent location of modes can change as the scale of a histogram is changed.
<b>Unimodal</b>	Having one mode. This is a useful term for describing the shape of a histogram when it's generally mound-shaped. Distributions with two modes are called <b>bimodal</b> . Those with more than two are <b>multimodal</b> .
<b>Uniform</b>	A distribution that's roughly flat is said to be uniform.
<b>Symmetric</b>	A distribution is symmetric if the two halves on either side of the center look approximately like mirror images of each other.
<b>Tails</b>	The tails of a distribution are the parts that typically trail off on either side. Distributions can be characterized as having long tails (if they straggle off for some distance) or short tails (if they don't).
<b>Skewed</b>	A distribution is skewed if it's not symmetric and one tail stretches out farther than the other. Distributions are said to be <b>skewed left</b> when the longer tail stretches to the left, and <b>skewed right</b> when it goes to the right.
<b>Outliers</b>	Outliers are extreme values that don't appear to belong with the rest of the data. They may be unusual values that deserve further investigation, or just mistakes; there's no obvious way to tell. Don't delete outliers automatically—you have to think about them. Outliers can affect many statistical analyses, so you should always be alert for them.
<b>Timeplot</b>	A timeplot displays data that change over time. Often, successive values are connected with lines to show trends more clearly.

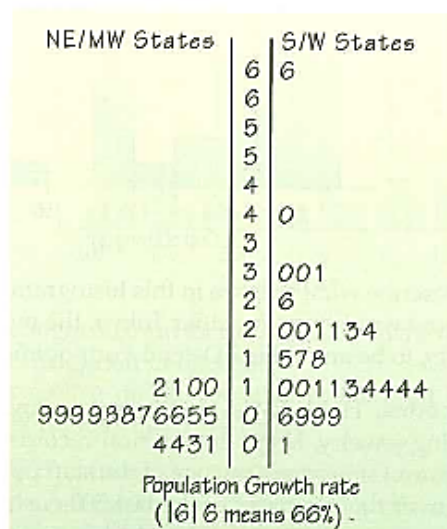
## Problem Set #1

- 1) **Vineyards.** The histogram shows the sizes (in acres) of 36 vineyards in the Finger Lakes region of New York.



- Approximately what percentage of these vineyards are under 60 acres?
- Write a brief description of this distribution (shape, center, spread, unusual features).

- 3) **Population growth.** Here is a “back-to-back” stem-and-leaf display that shows two data sets at once—one going to the left, one to the right. It compares the percent change in population for two regions of the United States (based on census figures for 1990 and 2000). The fastest growing states were Nevada at 66% and Arizona at 40%. Write a few sentences describing the difference in growth rates for the two regions of the United States. To show the distributions better, this display breaks each stem into two lines, putting leaves 0–4 on one stem and leaves 5–9 on the other.



2)

**Gasoline.** In June 2004, 16 gas stations in Ithaca, NY, posted these prices for a gallon of regular gasoline.

2.029	2.119	2.259	2.049
2.079	2.089	2.079	2.039
2.069	2.269	2.099	2.129
2.169	2.189	2.039	2.079

- Make a stem-and-leaf display of these gas prices. Use split stems; for example, use two 2.1 stems, one for prices between \$2.10 and \$2.149, the other for prices \$2.15 to \$2.199.
- Describe the shape, center, and spread of this distribution.
- What unusual feature do you see?

4)

**Hurricanes.** The data below give the number of hurricanes that happened each year from 1944 through 2000 as reported by *Science* magazine.

3, 2, 1, 2, 4, 3, 7, 2, 3, 3, 2, 5, 2, 2, 4, 2, 2, 6, 0, 2, 5, 1, 3, 1, 0,  
3, 2, 1, 0, 1, 2, 3, 2, 1, 2, 2, 2, 3, 1, 1, 1, 3, 0, 1, 3, 2, 1, 2, 1, 1,  
0, 5, 6, 1, 3, 5, 3

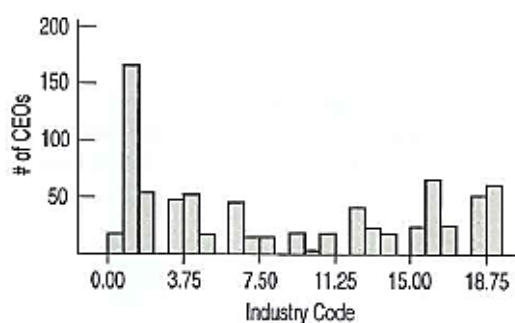
- Create a dotplot of these data.
- Describe the distribution.



- 5) **CEO data revisited.** For each CEO, a code is listed that corresponds to the industry of the CEO's company. Here are a few of the codes and the industries to which they correspond.

Industry	Industry Code
Financial services	1
Food/drink/tobacco	2
Health	3
Insurance	4
Retailing	6
Forest products	9
Aerospace/defense	11
Energy	12
Capital goods	14
Computers/communications	16
Entertainment/information	17
Consumer nondurables	18

A recently hired investment analyst has been assigned to examine the industries and the compensations of the CEOs. To start the analysis, he produces the following histogram of industry codes.



- What might account for the gaps seen in the histogram?
- Is the histogram unimodal?
- What advice might you give the analyst about the appropriateness of this display?

6)

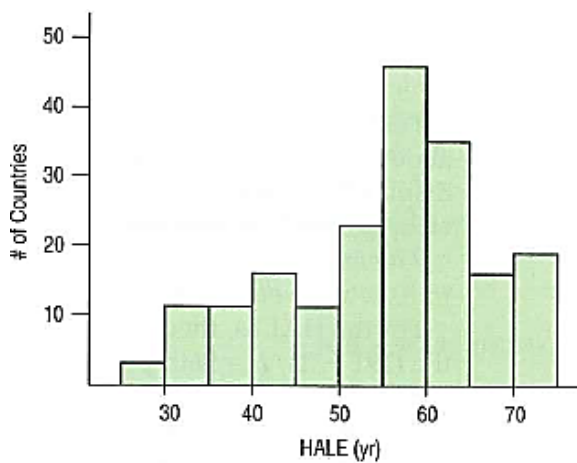
**Cholesterol.** A study examining the health risks of smoking measured the cholesterol levels of people who had smoked for at least 25 years and people of similar ages who had smoked for no more than 5 years and then stopped. Create histograms for both groups and write a brief report comparing their cholesterol levels.

Smokers				Ex-smokers		
225	211	209	284	250	134	300
258	216	196	288	249	213	310
250	200	209	280	175	174	328
225	256	243	200	160	188	321
213	246	225	237	213	257	292
232	267	232	216	200	271	227
216	243	200	155	238	163	263
216	271	230	309	192	242	249
183	280	217	305	242	267	243
287	217	246	351	217	267	218
200	280	209		217	183	228

## SECTION TWO: Describing Distributions Numerically

The World Health Organization (WHO) collects health data worldwide on every member country of the United Nations. One traditional measure of the overall health of a country has been the life expectancy of its citizens—the number of years that a newborn can expect to live. Starting in 1999, WHO scientists introduced a revised measure to take account of the fact that illness and injuries can affect the quality of life. The “healthy life expectancy” (HALE) adjusts for years of ill health to give a measure of years of healthy life.

Here is a histogram of the HALEs for all 191 United Nations countries:

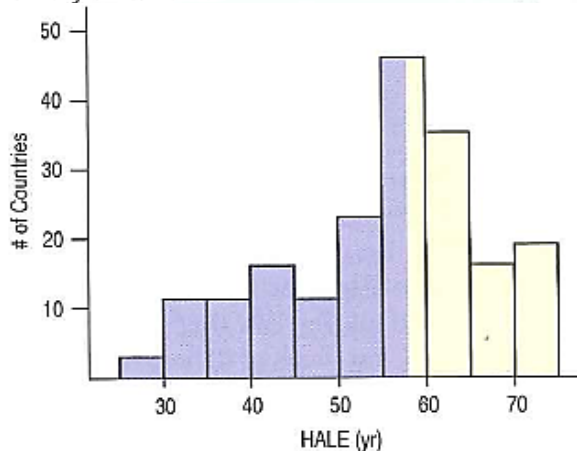


HALEs for the 191 United Nations countries. In 2001, Japan had the top HALE of 73.6 years, while Sierra Leone had the lowest at 26.5 years. **Figure 5.1**

When distributions are mixed together, the result is often a multimodal and/or skewed distribution. These data are combined from countries of different economic and social conditions, so a skewed distribution is no surprise.

## Finding the Center: The Median

What is a typical HALE? Try to put your finger on the histogram at the value you think is typical. (Read the value from the horizontal axis and remember it.) When we think of a typical value, we usually look for the center of the distribution. Where do you think the center of this distribution is? For a unimodal, symmetric distribution, it's easy. We'd all agree on the center of symmetry, where we would fold the histogram to match the two sides. But when the distribution is skewed and possibly multimodal, as this one is, it's not immediately clear what we even mean by the center.



The median splits the histogram into two halves of equal area. Notice that the halves of this histogram have very different shapes. **Figure 5.2**

Histograms follow the area principle, and each half of the data has about 95 countries, so each colored region has the same area in the display. The middle value that divides the histogram into two equal areas is called the **median**.

The median has the same units as the data. Be sure to include the units whenever you discuss the median.

For the HALEs, there are 191 countries, so the median is found at the  $(191 + 1)/2 = 96$ th place in the sorted data. This median HALE is 57.7 years.

The median is one way to find the center of the data. But there are many others. We'll look at an even more important measure later in this chapter.

Knowing the median, we could say that a typical healthy life expectancy, worldwide, was about 57.7 years. How much does that really say? How well does the median describe the data? After all, not all countries have HALEs near 57.7 years. Whenever we find the center of data, the next step is always to ask how well it actually summarizes the data.



## Spread: Home on the Range


If every country had a HALE of 57.7, knowing the median would tell us everything about the distribution of life expectancy worldwide. The more the data vary, however, the less the median alone can tell us. So we need to measure how much the data values vary around the center. In other words, how spread out are they? When we describe a distribution numerically, we always report a measure of its spread along with its center.

How should we measure the spread? We could simply look at the extent of the data. How far apart are the two extremes? The **range** of the data is defined as the difference between the maximum and minimum values:

$$\text{Range} = \text{max} - \text{min}.$$

Notice that the range is a *single number*, not an interval of values, as you might think from its use in common speech. The maximum HALE is 73.6 years and the minimum is 26.5 years, so the range is  $73.6 - 26.5 = 47.1$  years.

### Finding quartiles by hand



A simple way to find the quartiles is to split the batch into two halves at the median. (When  $n$  is odd, some statisticians include the median in both halves; others omit it.) The lower quartile is the median of the lower half, and the upper quartile is the median of the upper half.

Here are our two examples again.

The ordered values of the first batch were -17.5, 2.8, 3.2, 13.9, 14.1, 25.3, and 45.8, with a median of 13.9. Notice that 7 is odd; we'll include the median in both halves to get -17.5, 2.8, 3.2, 13.9 and 13.9, 14.1, 25.3, 45.8.

Each half has 4 values, so the median of each is the average of its 2nd and 3rd values. So, the lower quartile is  $(2.8 + 3.2)/2 = 3.0$  and the upper quartile is  $(14.1 + 25.3)/2 = 19.7$ .

The second batch of data had the ordered values -17.5, 2.8, 3.2, 13.9, 14.1, 25.3, 35.7, and 45.8.

Here  $n$  is even, so the two halves of 4 values are -17.5, 2.8, 3.2, 13.9 and 14.1, 25.3, 35.7, 45.8.

Now the lower quartile is  $(2.8 + 3.2)/2 = 3.0$  and the upper quartile is  $(25.3 + 35.7)/2 = 30.5$ .

The range (like the midrange) has the disadvantage that a single extreme value can make it very large, giving a value that doesn't really represent the data overall. For example, in the CEO compensations from Chapter 4, the range is  $\$202,991,184 - \$28,816 = \$202,962,368$ ! Most of the compensations were between \$0 and \$5,000,000, so the range doesn't give a very accurate impression of the spread.

## Spread: The Interquartile Range

A better way to describe the spread of a variable might be to ignore the extremes and concentrate on the middle of the data. We could, for example, find the range of just the middle half of the data. What do we mean by the middle half? Divide the data in half at the median. Now divide both halves in half again, cutting the data into four quarters. We call these new dividing points **quartiles**. One quarter of the data lies below the **lower quartile**, and one quarter of the data lies above the **upper quartile**, so half the data lies between them. The quartiles border the middle half of the data.

The difference between the quartiles tells us how much territory the middle half of the data covers and is called the **interquartile range**. It's commonly abbreviated IQR (and pronounced "eye-cue-are," not "ikker"):

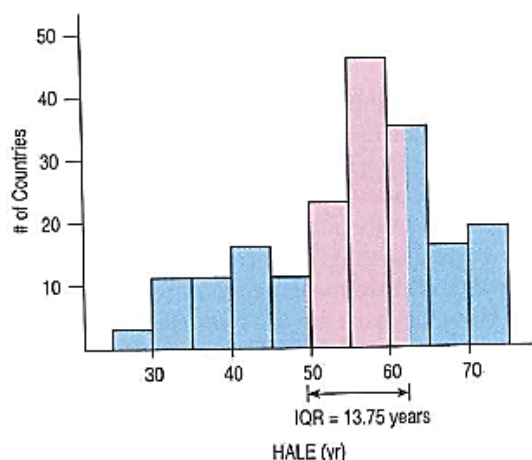
$$\text{IQR} = \text{upper quartile} - \text{lower quartile}.$$

For the HALE data, there are 95 values below and 95 values above the median. Including the median with each half of the data gives 96 values in each half. To find the median of each half, we'd average the 48th and 49th values. For the HALEs, the lower quartile is 48.9 years and the upper quartile is 62.65 years. The difference between the quartiles gives the IQR:

$$\text{IQR} = 62.65 - 48.9 \text{ years} = 13.75 \text{ years}.$$

Now we know that the middle half of the countries (in terms of HALE) extend for a (interquartile) range of 13.75 years. This seems like a reasonable summary of the spread of the distribution, as we can see from the histogram:





The IQR contains the middle 50% of the values of the distribution. It also gives a visual indication of the spread of the data. Here we see that the IQR is 13.75 years. **Figure 5.3**

The IQR is often a reasonable summary of the spread of a distribution. For the CEO compensations, the median is \$1,304,470 and the quartiles are at \$787,304 and \$2,518,628, so the IQR is  $\$2,518,628 - \$787,304 = \$1,731,324$ . The IQR gives a different impression of how spread out CEOs' salaries are than we might get from the range of \$202,962,368.

The lower and upper quartiles are also known as the 25th and 75th **percentiles** of the data, respectively, since the lower quartile falls above 25% of the data and the upper quartile falls above 75% of the data. If we count this way, the median is the 50th percentile. We could, of course, define and calculate any percentile that we want. For example, the 10th percentile would be the number that falls above the lowest 10% of the data values.

## 5-Number Summary

### NOTATION ALERT:

We always use Q1 to label the lower (25%) quartile and Q3 to label the upper (75%) quartile. We skip the number 2 because the median would, by this system, naturally be labeled Q2—but we don't usually call it that.

The **5-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum). The 5-number summary for the HALE data looks like this.

<b>Max</b>	73.6 years
<b>Q3</b>	62.65
<b>Median</b>	57.7
<b>Q1</b>	48.9
<b>Min</b>	26.5

It's a good idea to report the number of data values and the identity of the cases (the *Who*). Here there are 191 countries.

## Rock Concert Deaths: Making Boxplots

In their "Rock and Roll Wall of Shame," Crowd Management Strategies lists the names of people who have died at rock concerts from lax safety controls. The Crowdsafe® Database lists the ages, names, causes, and locations of these unfortunate concertgoers. During the period 1999–2000 there were 66 people who died from "crowd crush." How old were these victims? Here's a 5-number summary of their ages:

<b>Max</b>	47 years
<b>Q3</b>	22
<b>Median</b>	19
<b>Q1</b>	17
<b>Min</b>	13

Whenever we have a 5-number summary of a (quantitative) variable, we can display the information in a **boxplot**. To make a boxplot of the ages of rock concert victims, follow these steps:

1. Draw a single vertical axis spanning the extent of the data.<sup>1</sup> Draw short horizontal lines at the lower and upper quartiles and at the median. Then connect them with vertical lines to form a box. The box can have any width that looks OK.<sup>2</sup>
2. To help us construct the boxplot, we erect “fences” around the main part of the data. We place the upper fence 1.5 IQRs above the upper quartile and the lower fence 1.5 IQRs below the lower quartile. For the rock concert data, we compute

$$\text{Upper fence} = Q3 + 1.5 \text{ IQRs} = 22 + 1.5 \times 5 = 29.5 \text{ years.}$$

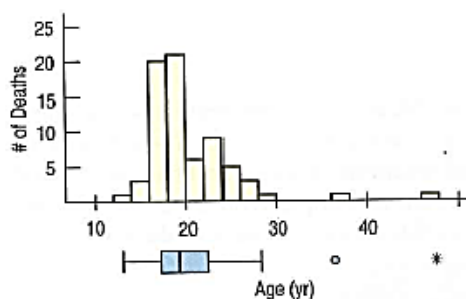
and

$$\text{Lower fence} = Q1 - 1.5 \text{ IQRs} = 17 - 1.5 \times 5 = 9.5 \text{ years.}$$

The fences are just for construction and are not part of the display. We show them here with dotted lines for illustration. You should never include them in your boxplot.

3. We use the fences to grow “whiskers.” Draw lines from the ends of the box up and down to the *most extreme data values found within the fences*. If a data value falls outside one of the fences, we do *not* connect it with a whisker.
4. Finally, we add the outliers by displaying any data values beyond the fences with special symbols. (We often use a different symbol for “far outliers”—data values farther than 3 IQRs from the quartiles.)

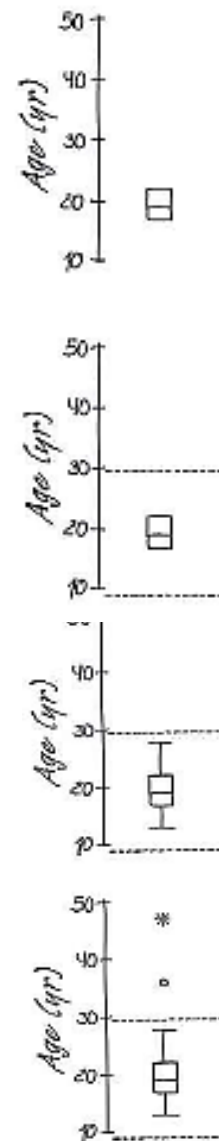
Now that we’ve drawn the boxplot, let’s summarize what it shows. The center of a boxplot is (remarkably enough) a box that shows the middle half of the data, between the quartiles. The height of the box is equal to the IQR. If the median is roughly centered between the quartiles, then the middle half of the data is roughly symmetric. If it is not centered, the distribution is skewed. The whiskers show skewness as well if they are not roughly the same length. Any outlier candidates are displayed individually, both to keep them out of the way for judging skewness and to encourage you to give them special attention. They may be mistakes, or they may be the most interesting cases in your data.



By turning a boxplot and putting it on the same scale, we can compare the boxplot and histogram of the rock concert deaths and see how each represents the distribution. **Figure 5.4**

From the boxplot we see that half of the victims were between 17 and 22 years old. The boxplot makes it look like the distribution of ages is roughly symmetric, with most of the victims between about 13 and 28 years old, but the histogram shows that the distribution is slightly skewed to the right. It is both an advantage and a disadvantage of boxplots that they simplify our view of the distribution like this. Boxplots are particularly good at pointing out outliers. Here, there are two victims who were substantially older. In a careful analysis of these data, we’d want to learn more about them.

Boxplots complement histograms by providing more specific information about the center, the quartiles, and outliers. When looking at one variable, it’s a good idea to look at the boxplot and histogram together. The main use for boxplots is to compare groups. That’s when they really start to shine.



## Comparing Groups with Boxplots

Histograms show a lot about the shape of the distribution, but get a little unwieldy when we want to look at more than one group at a time. Boxplots work well for comparing groups because they let the fundamental story show through. When we place them side-by-side, we can easily see which group has the higher median, which has the greater IQR, where the central 50% of the data are located, and which has the greater overall range. And we can get a general idea of symmetry from whether the medians are centered within their boxes and whether the whiskers extend roughly the same distance on either side of the boxes. Equally important, we can see past any outliers in making these comparisons because they’ve been separated from the rest of the data.



## Comparing Groups Step-By-Step

A student designed an experiment to test the efficiency of various coffee containers by placing hot (180 °F) liquid in each of 4 different container types 8 different times. After 30 minutes she measured the temperature again and recorded the difference in temperature. Because these are temperature *differences*, smaller differences mean that the liquid stayed hot—probably what we want in a coffee mug.

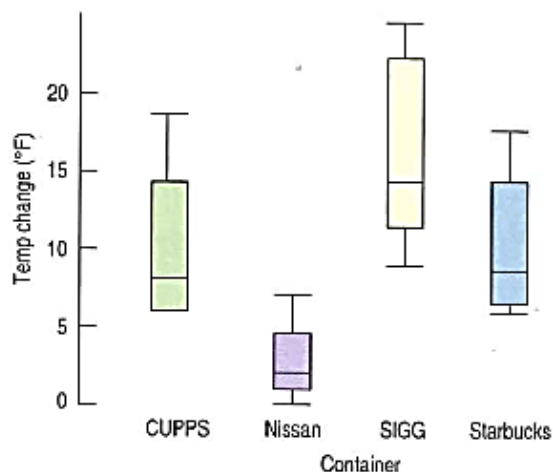
What can we say about the effectiveness of these four mugs? Let's see what story the data tell us.

### Show

**Mechanics** Report the 5-number summaries of the four groups. Including the IQR is a good idea as well.

Make a picture. Because we want to compare the distributions for four groups, boxplots are an obvious choice.

	Min	Q1	Median	Q3	Max	IQR
CUPPS	6 °F	6	8.25	14.25	18.50	8.25
Nissan	0	1	2	4.50	7	3.50
SIGG	9	11.50	14.25	21.75	24.50	10.25
Starbucks	6	6.50	8.50	14.25	17.50	7.75



### Tell

**Conclusion** Interpret what the boxplots and summaries say about the ability of these mugs to maintain heat. Compare the shapes, centers, and spreads, and note any outliers.

The individual distributions are all slightly skewed to the high end. The Nissan cup does the best job of keeping liquids hot, with a median loss of only 2 °F, and the SIGG cup does the worst, typically losing 14 °F. The difference is large enough to be important; a coffee drinker would be likely to notice a 14° drop in temperature. And the mugs are clearly different; 75% of the Nissan tests showed less heat loss than any of the other mugs in the study. The IQR of results for the Nissan cup is also the smallest of these test cups, indicating that it is a consistent performer.

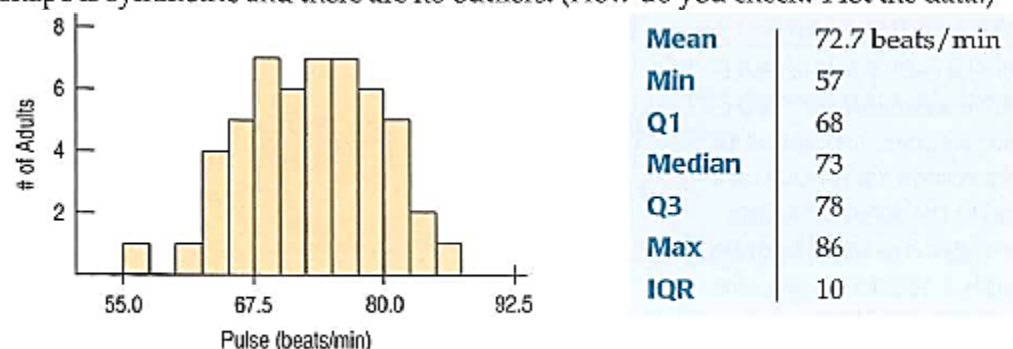
## Summarizing Symmetric Distributions

Medians do a good job of locating the center of a distribution even when the shape is skewed. But when we have symmetric data, there's another alternative. You probably already know how to average values. In fact, to find the median when  $n$  is even, we said you should average the two middle values, and you didn't even flinch. In general, to average values, add them up and divide by  $n$ , the number of values.

Averaging is a common thing to do with data. Once we've averaged some data, what do you think the result is called? The *average*? No, that would be too easy. Informally, we talk about the "average person" or the "average family," but we don't actually add up families and divide by  $n$ . So, to be more precise we call this summary the mean.

Pulse rates are useful for monitoring medical conditions. Resting heart rates depend on age. For children more than 10 years old and for adults, 60 to 100 beats per minute is considered normal. At the top of the next page are a histogram and summaries for the pulse rates of 52 adults.

The histogram shows a generally symmetric distribution, and the mean and median agree quite closely. That's what we expect for symmetric data. When the shape of the distribution is symmetric, there's no numerical reason to prefer the median or the mean. As we'll see, it turns out that there's more we can do and say with the mean than with the median. The mean is appropriate, however, only when the shape is symmetric and there are no outliers. (How do you check? Plot the data!)



Pulse rates of 52 adults. Figure 5.5

## The Formula for Averaging (Say It in Greek)

You already know how to average values, but this is a good opportunity to introduce some notation that will make it easier to describe calculations later on. Here's the formula:

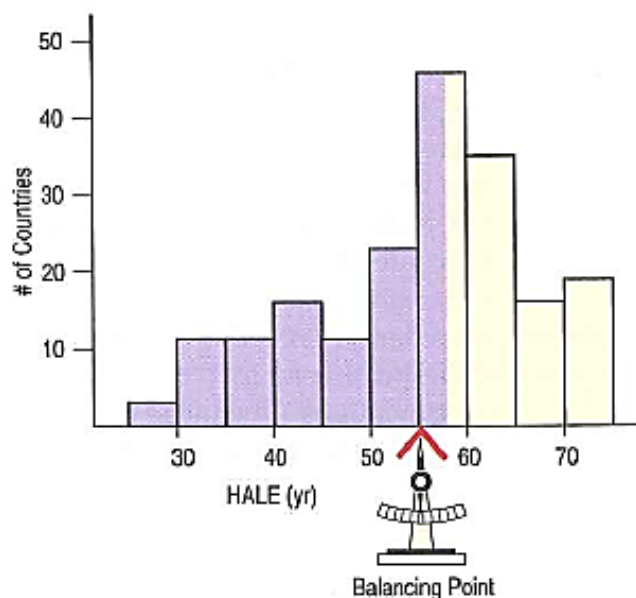
$$\bar{y} = \frac{\text{Total}}{n} = \frac{\sum y}{n}$$

The  $y$  with a line over it is pronounced "y-bar." In general, a bar over any symbol or variable name in Statistics denotes finding its mean. The symbol  $\Sigma$  is the Greek letter capital sigma—equivalent to an "S," as in "sum"—and means just that; add up all the observations. The formula says that to find the **mean**, add up all the numbers and divide by  $n$ —but you knew that.

## Mean or Median?

Does it make a difference whether we choose a mean or a median? Well, sometimes it does. The median of the HALEs is 57.7 years. The mean is only 55.26 years. Why are they different? The answer lies in the shape of the distribution. The **mean** is the point at which the histogram would balance.





The mean is located at the *balancing point* of the histogram. Since this distribution is skewed to the left, the mean is *lower* than the median. The points at the left have pulled the mean toward them, away from the median. **Figure 5.6**

Just like a child who moves away from the center of a see-saw, a bar of the histogram far from the center has more leverage, pulling the mean in its direction. If the skewness is strong or if there are straggling outliers, the mean can be pulled quite far from the median. The skewness on the left pulls the mean somewhat to the left of the median here. For the CEO compensation data from Chapter 4, the skewness is to the right. The median compensation is \$1,304,470, but the mean is \$2,818,743—more than *twice* the median. It's hard to argue that a value pulled in this way is what we meant by the center of the data. For skewed data, it's better to report the median than the mean as a measure of center.

## What About Spread? The Standard Deviation

The IQR is always a reasonable summary of spread, but because it uses only the two quartiles of the data, it ignores much of the information about how individual values vary. A more powerful approach uses the **standard deviation**, which takes into account how far *each* value is from the mean. Like the mean, the standard deviation is appropriate only for symmetric data.

One way to think about spread is to examine how far each data value is from the mean. This difference is called a *deviation*. We could just average the deviations, but the positive and negative differences always cancel each other out. So the average deviation is always zero—not very helpful.

### Finding the standard deviation by hand

To find the standard deviation, you start with the mean,  $\bar{y}$ . Then you find the *deviations* by taking  $\bar{y}$  from each value:

$(y - \bar{y})$ . Square each deviation:  $(y - \bar{y})^2$ .

Now you're nearly home. Just add these up and divide by  $n - 1$ . That gives you the variance,  $s^2$ . To find the standard deviation,  $s$ , take the square root. Here we go:

Suppose the batch of values is 4, 3, 10, 12, 8, 9, and 3.

The mean is  $\bar{y} = 7$ . So the deviations are found by subtracting 7 from each value:

Original Values	Deviations	Squared Deviations
4	$4 - 7 = -3$	$(-3)^2 = 9$
3	$3 - 7 = -4$	$(-4)^2 = 16$
10	$10 - 7 = 3$	9
12	$12 - 7 = 5$	25
8	$8 - 7 = 1$	1
9	$9 - 7 = 2$	4
3	$3 - 7 = -4$	16

Add up the squared deviations:

$9 + 16 + 9 + 25 + 1 + 4 + 16 = 80$ .

Now, divide by  $n - 1$ :  $80/6 = 13.33$ .

Finally, take the square root:

$s = \sqrt{13.33} = 3.65$

To keep them from canceling out, we *square* each deviation. Squaring always gives a positive value, so the sum won't be zero. That's great. Squaring also emphasizes larger differences—a feature that turns out to be both good and bad.

When we add up these squared deviations and find their average (almost), we call the result the **variance**:

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

Why almost? It *would* be a mean if we divided the sum by  $n$ . Instead, we divide by  $n - 1$ . Why? The simplest explanation is "to drive you crazy." But there are good technical reasons, some of which we'll see later.

The variance will play an important role later in this book, but it has a problem as a measure of spread. Whatever the units of the original data are, the variance is in *squared* units. We want measures of spread to have the same units as the data. And we probably don't want to talk about squared dollars, or *mpg*<sup>2</sup>. So, to get back to the original units, we take the square root of  $s^2$ . The result,  $s$ , is the **standard deviation**.

Putting it all together, the standard deviation of the data is found by the following formula:

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

You will almost always rely on a calculator or computer to do the calculating.

## Thinking About Variation

Statistics is about variation, so spread is an important fundamental concept in Statistics. Measures of spread help us to be precise about what we *don't* know. If many data values are scattered far from the center, the IQR and the standard deviation will be large. If the data values are close to the center, then these measures of spread will be small. If all our data values were exactly the same, we'd have no question about summarizing the center, and all measures of spread would be zero—and we wouldn't need Statistics. You might think this would be a big plus, but it would make for a boring world. Fortunately (at least for Statistics), data do vary.

Measures of spread tell how well other summaries describe the data. That's why we always (always!) report a spread along with any summary of the center.

## Shape, Center, and Spread

What should you tell about a quantitative variable? Report the shape of its distribution, and include a center and a spread. But which measure of center and which measure of spread? The rules are pretty easy:



- If the shape is skewed, report the median and IQR. You may want to include the mean and standard deviation, but you should point out why the mean and median differ. The fact that the mean and median do not agree is a sign that the distribution may be skewed. A histogram will help you make that point.
- If the shape is symmetric, report the mean and standard deviation and possibly the median and IQR as well. For symmetric data, the IQR is usually a bit larger than the standard deviation. If that's not true for your data set, look again to make sure the distribution isn't skewed and there are no outliers.
- If there are any clear outliers and you are reporting the mean and standard deviation, report them with the outliers present and with the outliers removed. The differences may be revealing. (Of course, the median and IQR are not likely to be affected by the outliers.)

We always pair the median with the IQR and the mean with the standard deviation. It's not useful to report one without the other. Reporting a center without a spread is dangerous. You may think you know more than you do about the distribution. Reporting only the spread leaves us wondering where we are.

#### How "Accurate" Should We Be?

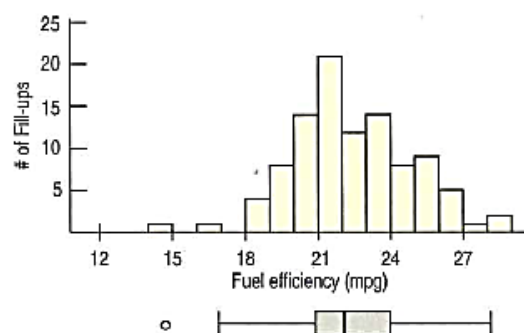
Don't think you should report means and standard deviations to a zillion decimal places; such implied accuracy is really meaningless. Although there is no ironclad rule, statisticians commonly report summary statistics to one or two decimal places more than the original data.

## Summarizing a Distribution Step-By-Step

One of the authors owned a 1989 Nissan Maxima for 8 years. Being a statistician, he recorded the car's fuel efficiency (in mpg) each time he filled the tank. He wanted to know what fuel efficiency to expect as "ordinary" for his car. (Hey, he's a statistician, what would you expect?<sup>3</sup>) Knowing this, he was able to predict when he'd need to fill the tank again, and to notice if the fuel efficiency suddenly got worse, which could be a sign of trouble. What do the data say?

### Show

**Mechanics** Make a histogram and boxplot. Based on the shape, choose appropriate numerical summaries.



The distribution of mileage is unimodal and symmetric with a mean of 22.4 mpg. There is a low outlier that should be investigated, but it does not influence the mean very much. The standard deviation is 2.4 mpg. The boxplot shows that half of the time, the car had a fuel efficiency between about 21 and 24 mpg.

### Tell

**Conclusion** Summarize and interpret your findings in context. Be sure to discuss the distribution's shape, center, spread, and unusual features (if any).

## Numerical Summaries on the Computer

Many statistics packages offer a pre-packaged collection of summary measures. The result might look like this:

```
Variable: Weight
N = 234
Mean = 143.3      Median = 139
St. Dev = 11.1    IQR = 14
```

Alternatively, a package might make a table for several variables and summary measures.

Variable	N	mean	median	stdev	IQR
Weight	234	143.3	139	11.1	14
Height	234	68.3	68.1	4.3	5
Score	234	86	88	9	5

It is usually easy to read the results and identify each computed summary. You should be able to read the summary statistics produced by any computer package.

Packages often provide many more summary statistics than you need. Of course, some of these may not be appropriate when the data are skewed or have outliers. It is your responsibility to check a histogram or stem-and-leaf display and decide which summary statistics to use.

It is common for packages to report summary statistics to many decimal places of "accuracy." Of course, it is rare data that have such accuracy in the original measurements. The ability to calculate to six or seven digits beyond the decimal point doesn't mean that those digits have any meaning. Generally, it's a good idea to round these values, allowing perhaps one more digit of precision than was given in the original data.

Summary statistics are easy to find in most packages.

## TERMS

**Center** We summarize the center of a distribution with the mean or the median.

**Median** The median is the middle value with half of the data above and half below it.

**Spread** We summarize the spread of a distribution with the standard deviation, interquartile range, and range.

**Range** The difference between the lowest and highest values in a data set.  $\text{Range} = \text{max} - \text{min}$ .

**Quartile** The lower quartile (Q1) is the value with a quarter of the data below it. The upper quartile (Q3) has a quarter of the data above it. The median and quartiles divide data into four equal parts.

**Interquartile range (IQR)** The IQR is the difference between the first and third quartiles.  $\text{IQR} = \text{Q3} - \text{Q1}$ .

**Percentile** The  $i$ th percentile is the number that falls above  $i\%$  of the data.



<b>5-number summary</b>	<p>A 5-number summary for a variable consists of:</p> <ul style="list-style-type: none"> <li>• The minimum and maximum</li> <li>• The quartiles: Q1 and Q3</li> <li>• The median</li> </ul>
<b>Boxplot</b>	A boxplot displays the 5-number summary as a central box with whiskers that extend to the non-outlying data values. Boxplots are particularly effective for comparing groups.
<b>Mean</b>	The mean is found by summing all the data values and dividing by the count.
<b>Variance</b>	The variance is the sum of squared deviations from the mean, divided by the count minus one.
<b>Standard deviation</b>	The standard deviation is the square root of the variance.
<b>Comparing distributions</b>	<p>When comparing the distribution of several groups, consider their</p> <ul style="list-style-type: none"> <li>• Shape</li> <li>• Center</li> <li>• Spread</li> </ul>
<b>Comparing boxplots</b>	<p>When comparing groups with boxplots,</p> <ul style="list-style-type: none"> <li>• Compare the medians; which group has the higher center?</li> <li>• Compare the IQRs; which group is more spread out?</li> <li>• Judged by the size of the IQRs, are the medians very different?</li> <li>• Check for possible outliers. Identify them if you can.</li> </ul>

## Problem Set #2

- 1) **More summaries.** Here are the annual numbers of deaths from tornadoes in the United States from 1990 through 2000. (Source: NOAA)

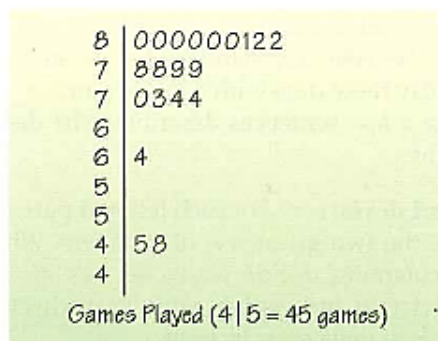
53 39 39 33 69 30 25 67 130 94 40

Find these statistics *by hand* (no calculator!):

- mean
- median and quartiles
- range and IQR

3)

**Wayne Gretzky.** In Chapter 4 (Exercise 12) you examined the number of games played by hockey great Wayne Gretzky during his 20-year career in the NHL. Here is the stem-and-leaf display:



- Would you use the median or the mean to describe the center of this distribution? Why?
- Find the median.
- Without actually finding the mean, would you expect it to be higher or lower than the median? Explain.

2)

**Standard deviation.** For each lettered part, a through c, examine the two given sets of numbers. Without doing any calculations, decide which set has the larger standard deviation and explain why. Then check by finding the standard deviations *by hand*.

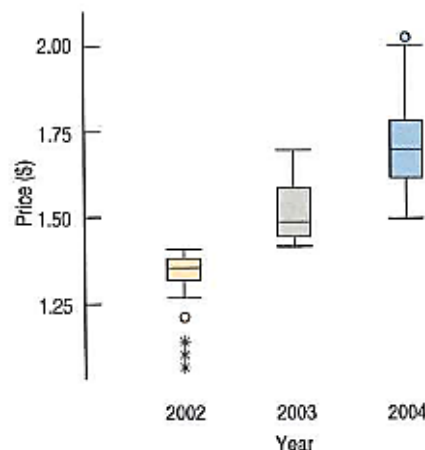
Set 1	Set 2
a) 3, 5, 6, 7, 9	2, 4, 6, 8, 10
b) 10, 14, 15, 16, 20	10, 11, 15, 19, 20

4)

**Home runs.** In 1961 Roger Maris made baseball headlines by hitting 61 home runs, breaking a famous record held by Babe Ruth. Here are Maris's home run totals for his 10 seasons in the American League. Would you consider his record-setting year to be an outlier? Explain.

8, 13, 14, 16, 23, 26, 28, 33, 39, 61

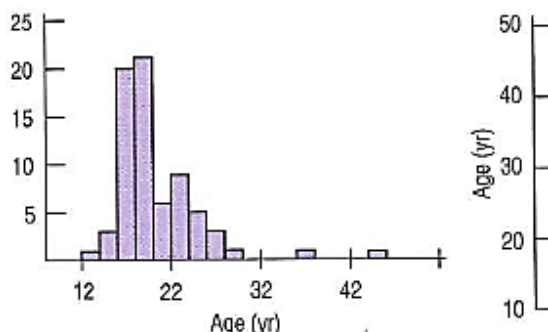
- 5) **Gas prices.** Here are boxplots of weekly gas prices at a service station in the Midwest United States (prices in \$ per gallon).



- Compare the distribution of prices over the three years.
  - In which year were the prices least stable? Explain.
- 7) **Derby speeds.** How fast do horses run? Kentucky Derby winners top 30 miles per hour, as shown in the graph on the next page. In fact, this graph shows the percentage of Derby winners that have run *slower* than a given speed. Note that few have won running less than 33 miles per hour, but about 95% of the winning horses have run less than 37 miles per hour. (A cumulative frequency graph like this is called an "ogive.")

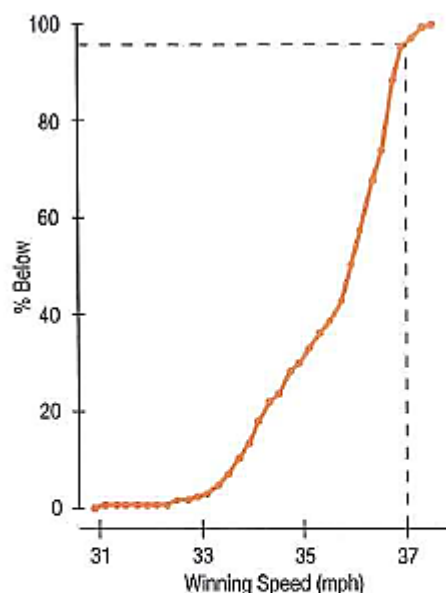
6)

**Still rockin'.** On pages 77–78, you read about the 66 deaths attributed to "crowd crush" at rock concerts during the years 1999 and 2000. Here are the histogram and boxplot of the victims' ages that we saw earlier:



- What features of the distribution can you see in both the histogram and the boxplot?
- What features of the distribution can you see in the histogram that you could not see in the boxplot?
- What summary statistic would you choose to summarize the center of this distribution? Why?
- What summary statistic would you choose to summarize the spread of this distribution? Why?

[ Be sure to use the "fence" method from the reading for your boxplot.]



- Estimate the median winning speed.
- Estimate the quartiles.
- Estimate the range and the IQR.
- Create a boxplot of these speeds.
- Write a few sentences about the speeds of the Kentucky Derby winners.



# **NOTES 1.1 Univariate Data – Graphing & Describing**

## **QUANTITATIVE (NUMERICAL) DATA –**

What's important in analyzing a distribution of quantitative data?

1. Shape –

2. Center –

3. Spread –

4. Anything unusual or noteworthy (ex. outliers, gaps)

## **Types of Graphs for Quantitative Data Distributions**

1. Dot Plots –

2. Histograms –

3. Stem and Leaf Plots (Stemplots) –

4. Box and Whisker Plots (Boxplots) –

5. Cumulative Relative Frequency Plots (Ogives) –

### **QUALITATIVE (CATEGORICAL) DATA –**

Types of Graphs for Qualitative Data Distributions

1. Pie (Circle) Charts –

2. Standard Bar Chart –

3. Segmented Bar Charts –

4. Mosaic Plots-