4.1 Displaying Quantitative Variables

**Histograms**

A histogram is a graph for a quantitative variable. Since there are no categories, we usually slice up all the possible values into bins and then count the number of cases that fall in each bin.

**Relative Frequency Histograms**

This graph reports the percentage of cases in each bin. The two graphs are the same, except for the labeling of the vertical axis.

**Stem-and-Leaf Displays**

\[ \text{Stem-and-leaf displays are like histograms, but they also give the individual values. Below is the AIG stock data.} \]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>77.26</td>
<td>72.95</td>
<td>73.72</td>
<td>71.57</td>
<td>68.42</td>
<td>65.99</td>
<td>61.22</td>
<td>64.10</td>
<td>58.04</td>
<td>68.26</td>
<td>65.03</td>
</tr>
<tr>
<td>2003</td>
<td>59.74</td>
<td>49.57</td>
<td>49.41</td>
<td>54.58</td>
<td>56.52</td>
<td>57.88</td>
<td>58.80</td>
<td>51.51</td>
<td>59.39</td>
<td>68.83</td>
<td>58.73</td>
</tr>
<tr>
<td>2004</td>
<td>68.62</td>
<td>79.70</td>
<td>72.86</td>
<td>74.23</td>
<td>79.93</td>
<td>72.61</td>
<td>48.80</td>
<td>68.58</td>
<td>70.67</td>
<td>62.19</td>
<td>62.37</td>
</tr>
<tr>
<td>2005</td>
<td>66.74</td>
<td>68.96</td>
<td>61.55</td>
<td>51.77</td>
<td>55.81</td>
<td>55.66</td>
<td>68.27</td>
<td>68.86</td>
<td>60.54</td>
<td>62.04</td>
<td>67.06</td>
</tr>
<tr>
<td>2006</td>
<td>68.33</td>
<td>67.02</td>
<td>67.35</td>
<td>64.28</td>
<td>63.14</td>
<td>58.74</td>
<td>58.40</td>
<td>62.00</td>
<td>65.25</td>
<td>67.02</td>
<td>69.86</td>
</tr>
<tr>
<td>2007</td>
<td>70.45</td>
<td>68.99</td>
<td>68.34</td>
<td>68.25</td>
<td>71.78</td>
<td>71.75</td>
<td>68.64</td>
<td>65.21</td>
<td>66.02</td>
<td>66.12</td>
<td>56.86</td>
</tr>
</tbody>
</table>

**How do stem-and-leaf displays work?**

We break each number in our dataset into two parts:

1) Use the first digit of a number (called the stem) to name the bins. The stem is to the left of the solid line.
2) Use the next digit of the number (called the leaf) to make the “bars”. The leaf is to the right of the solid line. For example, for the number 21, we would write 2 | 1 with 2 serving as the stem, 1 as the leaf, and a solid line in between.
4.1 Displaying Quantitative Variables

Stem-and-Leaf Displays

Example: Show how to display the data 21, 22, 24, 33, 33, 36, 38, 41 in a stem-and-leaf display.

2 | 124
3 | 3368
4 | 1

Note: If you turn your head sideways to look at the display, it resembles the histogram for the same data.

4.2 Shape

Modes

Peaks or humps seen in a histogram are called the modes of a distribution.

A distribution whose histogram has one main peak is called unimodal, two peaks – bimodal (see figure), three or more – multimodal.

Modes

A distribution whose histogram doesn’t appear to have any mode and in which all the bars are approximately the same height is called uniform.

Symmetry

A distribution is symmetric if the halves on either side of the center look, at least approximately, like mirror images.

When describing a distribution, attention should be paid to

• its shape,
• its center, and
• its spread.

We describe the shape of a distribution in terms of its modes, its symmetry, and whether it has any gaps or outlying values.
4.2 Shape

Symmetry

The thinner ends of a distribution are called the tails. If one tail stretches out farther than the other, the distribution is said to be skewed to the side of the longer tail. The distribution below is skewed to the right.

---

4.2 Shape

Symmetry

The distribution below is skewed to the left.

---

4.2 Shape

Outliers

Always be careful to point out the outliers in a distribution: those values that stand off away from the body of the distribution. Outliers …

- can affect every statistical method we will study.
- can be the most informative part of your data.
- may be an error in the data.
- should be discussed in any conclusions drawn about the data.

---

4.3 Center

To find the mean of the variable $y$ (we could call it $x$) add all the values of the variable and divide that sum by the number of data values, $n$.

The mean is a natural summary for unimodal, symmetric distributions.

We will use the Greek letter sigma to represent sum, so the equation for finding the mean can be written as shown.

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

The mean is considered to be the balancing point of the distribution.
4.3 Center

If a distribution is skewed, contains gaps, or contains outliers, then it is better to use the median – the center value that splits the histogram into two equal areas.

The median is found by counting in from the ends of the data until we reach the middle value.

The median is said to be resistant because it isn’t affected by unusual observations or by the shape of the distribution.

4.4 Spread of the Distribution

The range is a single value and it is not resistant to unusual observations. Concentrating on the middle of the data avoids this problem.

The quartiles are the values that frame the middle 50% of the data. One quarter of the data lies below the lower quartile, Q1, and one quarter lies above the upper quartile, Q3.

The interquartile range (IQR) is defined to be the difference between the two quartile values.

\[ IQR = Q3 - Q1 \]

4.4 Spread of the Distribution

If a distribution is roughly symmetric, we’d expect the mean and median to be close. The histogram below depicts monthly trading volume of AIG shares (in millions of shares) for the period 2002 to 2007. The mean is 170.1 million shares and the median is 135.9 million shares.

4.4 Spread of the Distribution

The IQR is a reasonable summary of spread, but because it only uses the two quartiles of data, it ignores much of the information about how individual values vary.

By contrast, the standard deviation takes into account how far each value is from the mean.

Like the mean, the standard deviation is appropriate only for symmetric distributions and can be influenced by outlying observations.

4.4 Spread of the Distribution

As the name implies, the standard deviation uses the deviations of each data value from the mean.

The average of the squared deviations of the values of the variable \( y \) from the mean is called the variance and is denoted by \( s^2 \):

\[ s^2 = \frac{\sum (y - \bar{y})^2}{n - 1} \]
4.4 Spread of the Distribution

The variance plays an important role in measuring spread, but the units are the square of the original units of the data. Taking the square root of the variance corrects this issue and gives us the **standard deviation**.

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

4.5 Shape, Center, and Spread – A Summary

Which measures of center and spread should be used for a distribution?

- If the shape is skewed, the median and IQR should be reported.
- If the shape is unimodal and symmetric, the mean and standard deviation and possibly the median and IQR should be reported.

4.5 Shape, Center, and Spread – A Summary

- If there are multiple modes, try to determine if the data can be split into separate groups.
- If there are unusual observations point them out and report the mean and standard deviation with and without the values.
- Always pair the median with the IQR and the mean with the standard deviation.

4.6 Standardizing Variables

To compare different variables, the values are standardized by measuring how far they are from the mean in terms of the standard deviation. We measure the distance from the mean and divide by the standard deviation, and the result is the standardized value. The standardized value tells how many standard deviations each value is above or below the overall mean. Usually we call it a **z-score**.

\[ z = \frac{x - \bar{x}}{s} \]

For example, a z-score of 2 indicates that a data value is two standard deviations above the mean.

4.6 Standardizing Variables

**For Example:** Real Estate

**Question:** A real estate analyst finds from data on 350 recent sales, that the average price was $175,000 with a standard deviation of $55,000. The size of the houses (in square feet) averaged 2100 sq. ft. with a standard deviation of 650 sq. ft.

Which is more unusual, a house in this town that costs $340,000, or a 5000 sq. ft. house?
4.6 Standardizing Variables

**Answer:** Convert both values to $z$-scores and compare which is more extreme.

For the $340,000 house:

$$z = \frac{x - \overline{x}}{s} = \frac{340,000 - 175,000}{55,000} = 3.0$$

For the 5000 sq. ft. house:

$$z = \frac{x - \overline{x}}{s} = \frac{5000 - 2100}{650} = 4.46$$

4.7 Five-Number Summary and Boxplots

**Answer:** Because the $z$-score of 4.46 is a greater number of standard deviations away from its mean, it is the more extreme value.

Therefore, the 5000 square foot house is more unusual.

The **five-number summary** of a distribution reports its median, quartiles, and extremes (maximum and minimum).

Below is the five-number summary of monthly trading volume of AIG shares (in millions of shares) for the period 2002 to 2007.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>515.62</td>
</tr>
<tr>
<td>Q3</td>
<td>182.32</td>
</tr>
<tr>
<td>Median</td>
<td>135.87</td>
</tr>
<tr>
<td>Q1</td>
<td>121.04</td>
</tr>
<tr>
<td>Min</td>
<td>83.91</td>
</tr>
</tbody>
</table>

4.7 Five-Number Summary and Boxplots

**For Example:** Real Estate

The central box shows the middle 50% of the data, between the quartiles – the height of the box equals 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.

The outliers are displayed individually to keep them out of the way in judging skewness and to display them for special attention.

To make a boxplot:

1) Locate the median and quartiles on an axis and draw a three short lines. For AIG data, approximate values are $Q_1= 121$, median = 136, and $Q_3 = 82$.

2) Then connect the quartile lines to form a box.
4.7 Five-Number Summary and Boxplots

3) Determine the "fences" by computing the upper fence 1.5 IQRs above the upper quartile and the lower fence 1.5 IQRs below the lower quartile.

4) Draw lines (whiskers) from each end of the box up and down to the most extreme data values found within the fences.

5) Display any outliers with special symbols like dots or asterisks.

Example: Gretzky
Wayne Gretzky scored 50% more points than anyone else who played professional hockey. Here are the number of games Gretzky played during each of his 20 seasons. Create a boxplot.

80, 80, 80, 80, 80, 80, 81, 82, 82, 79, 79, 78, 78, 74, 74, 73, 70, 64, 48, 45

Example (continued): Gretzky
Wayne Gretzky scored 50% more points than anyone else who played professional hockey. Here are the number of games Gretzky played during each of his 20 seasons. Create a boxplot.

Example (continued): Gretzky
Wayne Gretzky scored 50% more points than anyone else who played professional hockey. Here are the number of games Gretzky played during each of his 20 seasons. Describe the distribution. What unusual features do you see?

80, 80, 80, 80, 80, 80, 81, 82, 82, 79, 79, 78, 78, 74, 74, 73, 70, 64, 48, 45

Key: 1 represents 80 games
4.7 Five-Number Summary and Boxplots

**Example (continued): Gretzky**
Wayne Gretzky scored 50% more points than anyone else who played professional hockey. Here are the number of games Gretzky played during each of his 20 seasons. Describe the distribution. What unusual features do you see?

The distribution of the number of games played per season by Wayne Gretzky is skewed to the left with 2 outliers. He may have been injured during these seasons. The season with 64 games is also separated by a gap. The median is 79 games, the range is 37 games, and the IQR is 6.5 games.

4.8 Comparing Groups

What happened in 2008? We can further partition 2008 into months and compare those boxplots as well:

---

4.8 Comparing Groups

Example: Downloads
The chief financial officer of a music download site just secured the rights to offer downloads of a new music album. She’d like to see how well it’s selling, so she collected the number of downloads per hour for the past 24 hours.

Below the AIG data is displayed in yearly boxplots.

---

4.8 Comparing Groups

Example: Downloads
Compare the a.m. downloads to the p.m. downloads by displaying the two distributions side-by-side. Summarize.
4.8 Comparing Groups

Example: Downloads
There are generally more downloads in the afternoon than in the morning. The median number of afternoon downloads is around 22 as compared with 14 for the morning hours. The p.m. downloads are also much more consistent. The entire range of the p.m. hours, 15, is about the size of the IQR for a.m. hours. Both distributions appear to be fairly symmetric, although the a.m. hour distribution has some high points which seem to give some asymmetry.

4.9 Identifying Outliers

Example: Real Estate
A boxplot shows an extreme outlier, at $339.4 million. A check on the Internet shows this is clearly a mistake, warranting removal.

4.9 Identifying Outliers

What should be done with outliers?

They should be understood in the context of the data. An outlier for a year of data may not be an outlier for the month in which it occurred and vice versa.

They should be investigated to determine if they are in error. The values may have simply been entered incorrectly. If a value can be corrected, it should be.

They should be investigated to determine why they are so different from the rest of the data. For example, were extra sales or fewer sales seen because of a special event like a holiday.

4.10 Time Series Plots

A display of values against time is sometimes called a time series plot. Below we have a time series plot of the AIG daily closing prices in 2007.

Example: Real Estate
A real estate report lists the following prices for sales of single family homes in a small town in Virginia (rounded to the nearest thousand). Write a couple of sentences describing house prices in this town.

<table>
<thead>
<tr>
<th>Price (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>155,000</td>
</tr>
<tr>
<td>139,000</td>
</tr>
<tr>
<td>150,000</td>
</tr>
<tr>
<td>140,000</td>
</tr>
<tr>
<td>329,000</td>
</tr>
<tr>
<td>178,000</td>
</tr>
<tr>
<td>194,000</td>
</tr>
<tr>
<td>160,000</td>
</tr>
<tr>
<td>172,000</td>
</tr>
<tr>
<td>330,000</td>
</tr>
<tr>
<td>270,000</td>
</tr>
<tr>
<td>231,000</td>
</tr>
<tr>
<td>122,000</td>
</tr>
<tr>
<td>339,435,000</td>
</tr>
<tr>
<td>136,000</td>
</tr>
<tr>
<td>167,000</td>
</tr>
<tr>
<td>136,000</td>
</tr>
<tr>
<td>330,000</td>
</tr>
<tr>
<td>159,000</td>
</tr>
<tr>
<td>128,000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Price (in thousands)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100000</td>
</tr>
<tr>
<td>200000</td>
</tr>
<tr>
<td>300000</td>
</tr>
<tr>
<td>350000</td>
</tr>
<tr>
<td>400000</td>
</tr>
<tr>
<td>450000</td>
</tr>
<tr>
<td>500000</td>
</tr>
<tr>
<td>550000</td>
</tr>
<tr>
<td>600000</td>
</tr>
<tr>
<td>650000</td>
</tr>
<tr>
<td>700000</td>
</tr>
<tr>
<td>750000</td>
</tr>
<tr>
<td>800000</td>
</tr>
<tr>
<td>850000</td>
</tr>
<tr>
<td>900000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Date</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>01/01/2007</td>
<td>70.00</td>
</tr>
<tr>
<td>04/01/2007</td>
<td>50.00</td>
</tr>
<tr>
<td>07/01/2007</td>
<td>70.00</td>
</tr>
<tr>
<td>10/01/2007</td>
<td>70.00</td>
</tr>
<tr>
<td>01/01/2008</td>
<td>70.00</td>
</tr>
</tbody>
</table>
4.10 Time Series Plots

Time series plots often show a great deal of point-to-point variation, but general patterns do emerge from the plot. Time series plots may be drawn with the points connected. Below the AIG data from before is displayed this way.

When a time series is stationary (without a strong trend or change in variability), then a histogram can provide a useful summary.

However, when the time series is not stationary like the AIG prices after 2007, a histogram is unlikely to display much of interest; a time series plot would be more informative.

4.10 Time Series Plots

To better understanding the trend of times series data, plot a smooth trace. A trace is typically created using a statistics software package and will be discussed in a later section.

The AIG data has been plotted with a smooth trace below.

Unless there is strong evidence for doing otherwise, we should resist the temptation to think that any trend we see will continue indefinitely.

4.10 Time Series Plots

Consider the time series plot for the AIG monthly stock closing price in 2008. The histogram showed a symmetric, possibly unimodal distribution.

The time series plot shows a period of gently falling prices and then the severe decline in September, followed by very low prices.

4.11 Transforming Skewed Data

Example: Below we display the very-skewed distribution of total compensation for the CEOs of the 500 largest companies.

What is the “center” of this distribution? Are there outliers?

4.11 Transforming Skewed Data

When a distribution is skewed, it can be hard to summarize the data simply with a center and spread, and hard to decide whether the most extreme values are outliers or just part of the stretched-out tail.

One way to make a skewed distribution more symmetric is to re-express, or transform, the data by applying a simple function to all the data values.

If the distribution is skewed to the right, we often transform using logarithms or square roots; if it is skewed to the left, we may square the data values.
4.11 Transforming Skewed Data

Example: Below we display the transformed distribution of total compensation for the CEOs of the 500 largest companies.

This histogram is much more symmetric, and we see that a typical log compensation is between 6.0 and 8.0 or $1 million and $100 million in the original terms.

What Can Go Wrong?

• Watch out for multiple modes. If the data has multiple modes, consider separating the data.
• Beware of outliers.

What Have We Learned?

Make and interpret histograms to display the distribution of a variable.
• We understand distributions in terms of their shape, center, and spread.

What Can Go Wrong?

• Choose a scale appropriate to the data.
• Avoid inconsistent scales. Don’t change scales in the middle of a plot, and compare groups on the same scale.
• Label variables and axes clearly.
• Do a reality check. Make sure the calculated summaries make sense.
• Don’t compute numerical summaries of a categorical variable.

What Have We Learned?

Describe the shape of a distribution.
• A symmetric distribution has roughly the same shape reflected around the center.
• A skewed distribution extends farther on one side than on the other.
• A unimodal distribution has a single major hump or mode; a bimodal distribution has two; multimodal distributions have more.
• Outliers are values that lie far from the rest of the data.
What Have We Learned?

Compute the mean and median of a distribution, and know when it is best to use each to summarize the center.

- The **mean** is the sum of the values divided by the count. It is a suitable summary for unimodal, symmetric distributions.
- The **median** is the middle value; half the values are above and half are below the median. It is a better summary when the distribution is skewed or has outliers.

What Have We Learned?

Find a five-number summary and, using it, make a boxplot. Use the boxplot's outlier nomination rule to identify cases that may deserve special attention.

- A **five-number summary** consists of the median, the quartiles, and the extremes of the data.
- A **boxplot** shows the quartiles as the upper and lower ends of a central box, the median as a line across the box, and "whiskers" that extend to the most extreme values that are not nominated as outliers.
- Boxplots display separately any case that is more than 1.5 IQRs beyond each quartile. These cases should be considered as possible outliers.

What Have We Learned?

Compute the standard deviation and interquartile range (IQR), and know when it is best to use each to summarize the spread.

- The **standard deviation** is roughly the square root of the average squared difference between each data value and the mean. It is the summary of choice for the spread of unimodal, symmetric variables.
- The **IQR** is the difference between the quartiles. It is often a better summary of spread for skewed distributions or data with outliers.

What Have We Learned?

Use boxplots to compare distributions.

- Boxplots facilitate comparisons of several groups. It is easy to compare centers (medians) and spreads (IQRs).
- Because boxplots show possible outliers separately, any outliers don’t affect comparisons.

What Have We Learned?

Standardize values and use them for comparisons of otherwise disparate variables.

- We standardize by finding **z-scores**. To convert a data value to its z-score, subtract the mean and divide by the standard deviation.
- z-scores have no units, so they can be compared to z-scores of other variables.
- The idea of measuring the distance of a value from the mean in terms of standard deviations is a basic concept in Statistics and will return many times later in the course.

What Have We Learned?

Make and interpret time plots for time series data.

- Look for the trend and any changes in the spread of the data over time.