CHAPTER 2: Describing Distributions with Numbers

Basic Practice of Statistics
7th Edition
In Chapter 2 we cover …

- Measuring center: the mean
- Measuring center: the median
- Comparing the mean and the median
- Measuring variability: the quartiles
- The five-number summary and boxplots
- Spotting suspected outliers and the modified boxplot
- Measuring variability: the standard deviation
- Choosing measures of center and variability
- Examples of technology
- Organizing a statistical problem
The most common measure of center is the arithmetic average, or mean.

- To find the mean, $\bar{x}$ (pronounced “x-bar”), of a set of observations, add their values and divide by the number of observations. If the $n$ observations are $x_1, x_2, x_3, \ldots, x_n$, their mean is:
  
  $$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- or, in more compact notation: $\bar{x} = \frac{1}{n} \sum x_i$

- Wait a second what is this ??? What is this notation ?
Example

- Here are the travel times in minutes of 10 randomly chosen Salt Lake City workers: 10, 30, 5, 25, 40, 20, 10, 15, 30, 20
- What is the mean?
Problem

Five men in a room have a mean height of 70 inches. A tall man, 80 inches, enters the room. Now the mean height is:

a) $\frac{500}{6}$ inches.
b) $\frac{350}{6}$ inches.
c) $\frac{430}{6}$ inches.
d) $\frac{430}{5}$ inches.
Median ($M$)

- The **median**, $M$, is the midpoint of a distribution, the number such that half of the observations are smaller and the other half are larger.
- At least half of the *ordered* values are less than or equal to the median value.
- At least half of the *ordered* values are greater than or equal to the median value.
- Location of the median: $L(M) = (n+1)/2$, where $n =$ sample size.
- **Example:** If 25 data values are recorded, the Median would be the $(25+1)/2 = 13^{th}$ ordered value.
Median

- Example 1 data: 2 4 6
  Median \((M) = 4\)

- Example 2 data: 2 4 6 8
  Median = 5 (ave. of 4 and 6)

- Example 3 data: 6 2 4
  Median \(\not= 2\)
  (order the values: 2 4 6, so Median = 4)
Example

Here are the travel times in minutes of 10 randomly chosen Salt Lake City workers: 10, 30, 5, 25, 40, 20, 10, 15, 30, 20

What is the median?
Problem

Find the median of the following nine numbers.

43  54  55  63  67  68  69  77  85

a) 65
b) 64
c) 67
d) 64.6
Problem

Consider the following data.

43 54 55 63 67 68 69 77 85

Suppose that the last value is actually 115 instead of 85. What effect would this new maximum have on the median of the data?

a) increase the value of the median
b) decrease the value of the median
c) no effect
# Measuring center

Use the data below to calculate the mean and median of the commuting times (in minutes) of 15 randomly selected North Carolina workers.

<table>
<thead>
<tr>
<th>30</th>
<th>20</th>
<th>10</th>
<th>40</th>
<th>25</th>
<th>10</th>
<th>20</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>40</td>
<td>5</td>
<td>30</td>
<td>12</td>
<td>10</td>
<td>10</td>
<td>60</td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{30 + 20 + 10 + 40 + \ldots + 10 + 10}{15} = 22.5 \text{ minutes}
\]

Key: 1|5 represents a North Carolina worker who reported a 15-minute travel time to work.

\[
M = 20 \text{ minutes}
\]
Comparing the mean and the median

The mean and the median measure center in different ways, and both are useful.

- The mean and the median of a roughly symmetric distribution are close together.
- If the distribution is exactly symmetric, the mean and the median are exactly the same.
- In a skewed distribution, the mean is usually farther out in the long tail than is the median.
Comparing the mean and the median

Symmetric Distribution

Right-Skewed Distribution

Left-Skewed Distribution

Mean = Median

Median Mean

Mean Median
Problem

Based on the distribution of the data, approximately which values represent the mean and median cost of a haircut?

a) mean = $19, median = $24
b) mean = $24, median = $19
Question

A recent newspaper article in California said that the **median** price of single-family homes sold in the past year in the local area was $136,000 and the **mean** price was $149,160. Which do you think is more useful to someone considering the purchase of a home, the median or the mean?
Answer

Both! Average is affected by outliers while median is not. For example, if one house is extremely expensive, then the average will rise. The median would ignore that outlier.
Spread, or Variability

- If all values are the same, then they all equal to the mean. There is no variability.
- Variability exists when some values are different from (above or below) the mean.
- We will discuss the following measures of spread: range, quartiles, variance, and standard deviation.
Range

- One way to measure spread is to give the smallest (minimum) and largest (maximum) values in the data set;
  \[ \text{Range} = \max - \min \]
- The range is strongly affected by outliers
  (e.g. one house is extremely expensive and the rest all have the same price. The range is large while there is little variability!)
Measuring variability: quartiles

- A measure of center alone can be misleading.
- A useful numerical description of a distribution requires both a measure of center and a *measure of spread*. We could look at the largest and smallest values (and we will!), but like the mean, they are (obviously) affected by extreme values—so we will examine other percentiles.

To calculate the **quartiles**:

- Arrange the observations in increasing order and locate the median \( M \).
- The **first quartile**, \( Q_1 \), is the median of the observations located to the left of the median in the ordered list.
- The **third quartile**, \( Q_3 \), is the median of the observations located to the right of the median in the ordered list.
Quartiles

- Three numbers which divide the ordered data into four equal sized groups.
- $Q_1$ has 25% of the data below it.
- $Q_2$ has 50% of the data below it. (Median)
- $Q_3$ has 75% of the data below it.
The minimum and maximum values alone tell us little about the distribution as a whole. Likewise, the median and quartiles tell us little about the tails of a distribution.

To get a quick summary of both center and spread, combine all five numbers.

The five-number summary of a distribution consists of the smallest observation, the first quartile, the median, the third quartile, and the largest observation—written in order from smallest to largest.

Minimum  $Q_1$  $M$  $Q_3$  Maximum
### Weight Data: Sorted

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>124</td>
<td>148</td>
<td>170</td>
<td>185</td>
<td>215</td>
<td></td>
</tr>
<tr>
<td>101</td>
<td>125</td>
<td>150</td>
<td>170</td>
<td>185</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>127</td>
<td>150</td>
<td>172</td>
<td>186</td>
<td>260</td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>128</td>
<td>152</td>
<td>175</td>
<td>187</td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>130</td>
<td>155</td>
<td>175</td>
<td>192</td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>130</td>
<td>157</td>
<td>180</td>
<td>194</td>
<td></td>
<td></td>
</tr>
<tr>
<td>119</td>
<td>133</td>
<td>165</td>
<td>180</td>
<td>195</td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>135</td>
<td>165</td>
<td>180</td>
<td>203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>120</td>
<td>139</td>
<td>165</td>
<td>180</td>
<td>210</td>
<td></td>
<td></td>
</tr>
<tr>
<td>123</td>
<td>140</td>
<td>170</td>
<td>185</td>
<td>212</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[ L(M) = \frac{(53+1)}{2} = 27 \]
\[ L(Q1) = \frac{(26+1)}{2} = 13.5 \]
Weight Data: Quartiles

- $Q_1 = 127.5$
- $Q_2 = 165$ (Median)
- $Q_3 = 185$
Five-Number Summary

- minimum = 100
- $Q_1 = 127.5$
- $M = 165$
- $Q_3 = 185$
- maximum = 260

\[ \text{Interquartile Range (IQR)} = Q_3 - Q_1 = 57.5 \]

$IQR$ gives spread of middle 50% of the data
Boxplots

The five-number summary divides the distribution roughly into quarters. This leads to a new way to display quantitative data, the boxplot.

How to Make a Boxplot

- A central box spans the quartiles $Q_1$ and $Q_3$.
- A line in the box marks the median $M$.
- Lines extend from the box out to the smallest and largest observations.
Weight Data: Boxplot

(min)  Q₁  M  Q₃  (max)

100  125  150  175  200  225  250  275

Weight
Five-number summary and boxplots

Consider a second travel times data set, these from New York. Find the five-number summary and construct a boxplot.

<table>
<thead>
<tr>
<th>TravelTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>30</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>85</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>65</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>60</td>
</tr>
<tr>
<td>40</td>
</tr>
<tr>
<td>45</td>
</tr>
</tbody>
</table>

- **Min** = 5
- **Q₁** = 15
- **M** = 22.5
- **Q₃** = 42.5
- **Max** = 85
Spotting suspected outliers and modified boxplots

- Having observed that the extremes (minimum and maximum) don’t describe the spread of the majority of the data, we turn to the difference of the quartiles:

  The **interquartile range**, or **IQR**, is the distance between the first and third quartiles

  \[
  IQR = Q_3 - Q_1
  \]

- In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers.

**The 1.5 × IQR Rule for Outliers**

Call an observation a suspected outlier if it falls more than \(1.5 \times IQR\) above the third quartile or below the first quartile.
Spotting suspected outliers: example

- In the New York travel time data, $Q_1 = 15$ minutes, $Q_3 = 42.5$ minutes, and $IQR = 27.5$ minutes.
- For these data, $1.5 \times IQR = 1.5(27.5) = 41.25$
- $Q_1 - 1.5 \times IQR = 15 - 41.25 = -26.25$
- $Q_3 + 1.5 \times IQR = 42.5 + 41.25 = 83.75$
Spotting suspected outliers: example modified boxplot

- Any travel time shorter than –26.25 minutes or longer than 83.75 minutes is considered an outlier.
- So the maximum observation, 85 minutes, would be a suspected outlier.

- Note that North Carolina has no suspected outliers.
Problem

Below is a boxplot for the tar content of 25 different cigarettes. What is a plausible set of values for the five-number summary?

- a) Min = 13, $Q_1 = 10$, Median = 12.6, $Q_3 = 14$, Max = 15
- b) Min = 1, $Q_1 = 8.5$, Median = 12.6, $Q_3 = 15$, Max = 17
- c) Min = 1, $Q_1 = 8.5$, Median = 11.5, $Q_3 = 13$, Max = 15
- d) Min = 8.5, $Q_1 = 10$, Median = 11.5, $Q_3 = 15$, Max = 17
Problem

Which group has the largest spread?

a) married females
b) single females
c) married males
d) single males
Measuring variability: standard deviation

The most common measure of spread looks at how far each observation is from the mean. This measure is called the standard deviation.

The variance, $s^2$, of a set of observations is an average of the squares of the deviations of the observations from their mean. In symbols, the variance of the $n$ observations $x_1, x_2, x_3, \ldots, x_n$, is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

Again, more briefly:

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

The standard deviation, $s$, is the square root of the variance, $s^2$. 

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n - 1} \sum (x_n - \bar{x})^2}$$
Calculating the Standard Deviation (1 of 2)

EXAMPLE: Consider the following data on the SAT critical reading scores for 5 Georgia Southern University freshman in 2010.

1) Calculate the mean.
2) Calculate each deviation.
   
   \[ \text{deviation} = \text{observation} - \text{mean} \]

   \[ \begin{align*}
   490 - 548 &= -58 \\
   \text{etc.} \\
   650 - 548 &= 102
   \end{align*} \]
Calculating the standard deviation (2 of 2)

3) Square each deviation.
4) Find the “average” squared deviation. Calculate the sum of the squared deviations divided by (n-1)…this is the variance.
5) Calculate the square root of the variance…this is the standard deviation.

<table>
<thead>
<tr>
<th>(x_i)</th>
<th>((x_i\text{-mean}))</th>
<th>((x_i\text{-mean})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>650</td>
<td>650 – 548 = 102</td>
<td>(102)^2 = 10404</td>
</tr>
<tr>
<td>490</td>
<td>490 – 548 = -58</td>
<td>(-58)^2 = 3364</td>
</tr>
<tr>
<td>580</td>
<td>580 – 548 = 32</td>
<td>(32)^2 = 1024</td>
</tr>
<tr>
<td>450</td>
<td>450 – 548 = -98</td>
<td>(-98)^2 = 9604</td>
</tr>
<tr>
<td>570</td>
<td>570 – 548 = 22</td>
<td>(22)^2 = 484</td>
</tr>
</tbody>
</table>

Sum = ?  Sum = ?

“Average” squared deviation = 24,880/(5 – 1) = 6220. This is the variance.

Standard deviation = square root of variance = \(\sqrt{6220} = 78.87\)
Example 2: Calculating the Standard Deviation

Metabolic rates of 7 men (cal./24hr.):

\[
\bar{x} = \frac{1792 + 1666 + 1362 + 1614 + 1460 + 1867 + 1439}{7}
\]

\[
= \frac{11,200}{7}
\]

\[
= 1600
\]
Example 2: Calculating the Standard Deviation

<table>
<thead>
<tr>
<th>Observations</th>
<th>Deviations $x_i - \bar{x}$</th>
<th>Squared deviations $(x_i - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1792</td>
<td>1792–1600 = 192</td>
<td>$(192)^2 = 36,864$</td>
</tr>
<tr>
<td>1666</td>
<td>1666 –1600 = 66</td>
<td>$(66)^2 = 4,356$</td>
</tr>
<tr>
<td>1362</td>
<td>1362 –1600 = -238</td>
<td>$(-238)^2 = 56,644$</td>
</tr>
<tr>
<td>1614</td>
<td>1614 –1600 = 14</td>
<td>$(14)^2 = 196$</td>
</tr>
<tr>
<td>1460</td>
<td>1460 –1600 = -140</td>
<td>$(-140)^2 = 19,600$</td>
</tr>
<tr>
<td>1867</td>
<td>1867 –1600 = 267</td>
<td>$(267)^2 = 71,289$</td>
</tr>
<tr>
<td>1439</td>
<td>1439 –1600 = -161</td>
<td>$(-161)^2 = 25,921$</td>
</tr>
<tr>
<td>sum = 0</td>
<td>sum = 214,870</td>
<td></td>
</tr>
</tbody>
</table>
Example 2: Calculating the Standard Deviation

\[ s^2 = \frac{214,870}{7 - 1} = 35,811.67 \]

\[ s = \sqrt{35,811.67} = 189.24 \text{ calories} \]
Example 3: Number of Books Read for Pleasure: Sorted

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td></td>
<td>30</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td></td>
<td>99</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>13</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>14</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td>20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.5+(5.5-1)x1.5=12.25
Five-Number Summary: Boxplot

Median = 3
interquartile range (iqr) = 5.5 - 1.0 = 4.5
range = 99 - 0 = 99

Mean = 7.06  s.d. = 14.43
Problem
Complete the calculations for the standard deviation of Mark McGwire’s yearly home runs.


<table>
<thead>
<tr>
<th>Year</th>
<th>$x_i$</th>
<th>$x_i - \bar{x}$</th>
<th>$(x_i - \bar{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>49</td>
<td>11.167</td>
<td>124.70</td>
</tr>
<tr>
<td>1988</td>
<td>32</td>
<td>-5.833</td>
<td>34.02</td>
</tr>
<tr>
<td>1989</td>
<td>33</td>
<td>-4.833</td>
<td>23.36</td>
</tr>
<tr>
<td>1990</td>
<td>39</td>
<td>1.167</td>
<td>1.36</td>
</tr>
<tr>
<td>1991</td>
<td>22</td>
<td>-15.833</td>
<td>250.68</td>
</tr>
<tr>
<td>1992</td>
<td>42</td>
<td>4.167</td>
<td>17.36</td>
</tr>
<tr>
<td>1993</td>
<td>9</td>
<td>-28.833</td>
<td>831.34</td>
</tr>
<tr>
<td>1994</td>
<td>9</td>
<td>-28.833</td>
<td>831.34</td>
</tr>
<tr>
<td>1995</td>
<td>39</td>
<td>1.167</td>
<td>1.36</td>
</tr>
<tr>
<td>1996</td>
<td>52</td>
<td>14.167</td>
<td>200.70</td>
</tr>
<tr>
<td>1997</td>
<td>58</td>
<td>20.167</td>
<td>406.71</td>
</tr>
<tr>
<td>1998</td>
<td>70</td>
<td>32.167</td>
<td>1034.72</td>
</tr>
<tr>
<td>Sum</td>
<td>454</td>
<td>0</td>
<td>3757.7</td>
</tr>
</tbody>
</table>

**a)** $\frac{3757.7}{12}$

**b)** $\sqrt{\frac{3757.7}{12}}$

**c)** $\frac{3757.7}{\sqrt{12}}$

**d)** $\sqrt{\frac{3757.7}{11}}$
Properties of $s$

- $n - 1$ is called the degrees of freedom.
- $s$ measures variability about the mean and should be used only when the mean is chosen as the measure of center.
- $s$ is always zero or greater than zero. $s = 0$ only when there is no variability. This happens only when all observations have the same value. Otherwise, $s > 0$.
- As the observations become more variable about their mean, $s$ gets larger.
- $s$ has the same units of measurement as the original observations. For example, if you measure weight in kilograms, both the mean $\bar{x}$ and the standard deviation $s$ are also in kilograms. This is one reason to prefer $s$ to the variance $s^2$, which would be in squared kilograms.
- Like the mean $\bar{x}$, $s$ is not resistant. A few outliers can make $s$ very large.
Choosing measures of center and variability

We now have a choice between two descriptions for center and variability:

- mean and standard deviation
- median and interquartile range

Choosing a Summary

The five-number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with strong outliers. Use $\bar{x}$ and $s$ only for reasonably symmetric distributions that are free of outliers.
Examples of technology

- The displays below come from a Texas Instruments graphing calculator, JMP statistical software, and the Microsoft Excel spreadsheet program.
- Once you know what to look for, you can read output from any technological tool.
Organizing a statistical problem

As you learn more about statistics, you will be asked to solve more complex problems.

Here is a four-step process you can follow.

Organizing a Statistical Problem: A Four-Step Process

- **State:** What is the practical question, in the context of the real-world setting?
- **Plan:** What specific statistical operations does this problem call for?
- **Solve:** Make graphs and carry out calculations needed for the problem.
- **Conclude:** Give your practical conclusion in the setting of the real-world problem.
Problem
An individual requires a warm, stable climate. Which city is better?

(a) Atlanta, because the mean is lower.
(b) San Diego, because the mean is higher.
(c) Atlanta, because the standard deviation is higher.
(d) San Diego, because the standard deviation is lower.
Problem

You have $100,000 to invest, and you don’t like to take risks. Which mutual fund should you choose?

a) SWBDX, because the minimum is higher.
b) SWLBX, because the maximum is higher.
c) SWBDX, because the standard deviation is lower.
d) SWLBX, because the standard deviation is higher.