Numerical Summaries

- Center of the data
  - mean
  - median
- Variation
  - range
  - quartiles (interquartile range)
  - variance
  - standard deviation
Mean or Average

- Traditional measure of center
- Sum the values and divide by the number of values
Median ($M$)

- A *resistant measure* of the data’s center
- 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
- If $n$ is *odd*, the median is the middle ordered value
- If $n$ is *even*, the median is the average of the two middle ordered values
Median ($M$)

Location of the median: $L(M) = (n+1)/2$, where $n = \text{sample size}$.

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 \( \neq 2 \)
  \( \text{(order the values: 2 4 6, so Median = 4)} \)
Comparing the Mean & Median

- The mean and median of data from a symmetric distribution should be close together. The actual (true) mean and median of a symmetric distribution are exactly the same.
- In a skewed distribution, the mean is farther out in the long tail than is the median [the mean is ‘pulled’ in the direction of the possible outlier(s)].
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.
Case Study

Airline fares

appeared in the New York Times on November 5, 1995

“...about 60% of airline passengers ‘pay less than the average fare’ for their specific flight.”

- How can this be?

13% of passengers pay more than 1.5 times the average fare for their flight
Spread, or Variability

- If all values are the same, then they all equal 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;

\[
\text{Range} = \text{max} - \text{min}
\]

- 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!)
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.
Obtaining the Quartiles

- **Order** the data.
- For $Q_2$, just find the median.
- For $Q_1$, look at the lower half of the data values, those to the left of the median location; find the median of this lower half.
- For $Q_3$, look at the upper half of the data values, those to the right of the median location; find the median of this upper half.
### 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>

\[
L(M) = (53 + 1)/2 = 27 \\
L(Q1) = (26 + 1)/2 = 13.5
\]
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

$\text{Interquartile Range (IQR) } = Q_3 - Q_1$

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

- Central box spans $Q_1$ and $Q_3$.
- A line in the box marks the median $M$.
- Lines extend from the box out to the minimum and maximum.
Weight Data: Boxplot

- min: 100
- Q1: 125
- M: 150
- Q3: 175
- max: 275

Weight Data Range:

- min: 100
- Q1: 125
- M: 150
- Q3: 175
- max: 275

Weight

Chapter 2
Example from Text: Boxplots

```
\begin{figure}
\centering
\includegraphics[width=\textwidth]{boxplots.png}
\caption{Comparison of earnings for high school and bachelor's degree levels.}
\end{figure}
```
Identifying Outliers

- The central box of a boxplot spans $Q_1$ and $Q_3$; recall that this distance is the Interquartile Range ($IQR$).

- We call an observation a suspected outlier if it falls more than $1.5 \times IQR$ above the third quartile or below the first quartile.
Variance and Standard Deviation

- Recall that variability exists when some values are different from (above or below) the mean.
- Each data value has an associated deviation from the mean:
Deviations

- what is a *typical* deviation from the mean? *(standard deviation)*

- small values of this typical deviation indicate small variability in the data

- large values of this typical deviation indicate large variability in the data
Variance

- Find the mean
- Find the deviation of each value from the mean
- Square the deviations
- Sum the squared deviations
- Divide the sum by \( n-1 \)

*(gives typical squared deviation from mean)*
Variance Formula
Standard Deviation Formula

*typical deviation from the mean*

\[
s = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

[standard deviation = square root of the variance]
Variance and Standard Deviation
Example from Text

Metabolic rates of 7 men (cal./24hr.) :
1792  1666  1362  1614  1460  1867  1439
Variance and Standard Deviation

Example from Text

<table>
<thead>
<tr>
<th>Observations $x_i$</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>
</tbody>
</table>

sum = 0                                sum = 214,870
Variance and Standard Deviation

Example from Text
Choosing a Summary

- Outliers affect the values of the mean and standard deviation.
- The five-number summary should be used to describe center and spread for skewed distributions, or when outliers are present.
- Use the mean and standard deviation for reasonably symmetric distributions that are free of outliers.
- Best to use both!
### 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>30</td>
<td>99</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>10</td>
<td>12</td>
<td>13</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>13</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>3</td>
<td>5</td>
<td>14</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>15</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>4</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5.5 + (5.5 - 1) x 1.5 = 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