Introduction to Probability and Statistics
Twelfth Edition

Chapter 2
Describing Data
with Numerical Measures

Some graphic screen captures from Seeing Statistics ©
Some images © 2001-(current year) www.arttoday.com
Describing Data with Numerical Measures

• Graphical methods may not always be sufficient for describing data.

• **Numerical measures** can be created for both **populations** and **samples**.
  
  – A **parameter** is a numerical descriptive measure calculated for a **population**.
  
  – A **statistic** is a numerical descriptive measure calculated for a **sample**.
Measures of Center

- A measure along the horizontal axis of the data distribution that locates the center of the distribution.
Arithmetic Mean or Average

• The mean of a set of measurements is the sum of the measurements divided by the total number of measurements.

\[
\bar{x} = \frac{\sum x_i}{n}
\]

where \( n \) = number of measurements

\( \sum x_i \) = sum of all the measurements
Example

• The set: 2, 9, 1, 5, 6

\[
\bar{x} = \frac{\sum x_i}{n} = \frac{2 + 9 + 11 + 5 + 6}{5} = \frac{33}{5} = 6.6
\]

If we were able to enumerate the whole population, the population mean would be called \( \mu \) (the Greek letter “mu”).
Median

- The **median** of a set of measurements is the middle measurement when the measurements are ranked from smallest to largest.

- The **position of the median** is 

\[ \frac{.5(n + 1)} \]

once the measurements have been ordered.
Example

• The set: 2, 4, 9, 8, 6, 5, 3 \( n = 7 \)
• Sort: 2, 3, 4, 5, 6, 8, 9
• Position: \( .5(n + 1) = .5(7 + 1) = 4^{\text{th}} \)
  
  Median = 4^{\text{th}} \text{ largest measurement}

• The set: 2, 4, 9, 8, 6, 5 \( n = 6 \)
• Sort: 2, 4, 5, 6, 8, 9
• Position: \( .5(n + 1) = .5(6 + 1) = 3.5^{\text{th}} \)
  
  Median = (5 + 6)/2 = 5.5 — average of the 3^{rd} and 4^{th} measurements
Mode

• The **mode** is the measurement which occurs most frequently.

• The set: 2, 4, 9, 8, 8, 5, 3
  – The mode is **8**, which occurs twice

• The set: 2, 2, 9, 8, 8, 5, 3
  – There are two modes—**8** and **2** (bimodal)

• The set: 2, 4, 9, 8, 5, 3
  – There is **no mode** (each value is unique).
Example

The number of quarts of milk purchased by 25 households:

0 0 1 1 1 1 1 2 2 2 2 2 2 2 2 2 3 3 3 3 3 4 4 4 4 5

- **Mean?**
  \[ \bar{x} = \frac{\sum x_i}{n} = \frac{55}{25} = 2.2 \]

- **Median?**
  \[ m = 2 \]

- **Mode? (Highest peak)**
  \[ \text{mode} = 2 \]
Extreme Values

- The mean is more easily affected by extremely large or small values than the median.

- The median is often used as a measure of center when the distribution is skewed.
Extreme Values

Symmetric: Mean = Median

Skewed right: Mean > Median

Skewed left: Mean < Median
Measures of Variability

- A measure along the horizontal axis of the data distribution that describes the spread of the distribution from the center.
The Range

• The range, \( R \), of a set of \( n \) measurements is the difference between the largest and smallest measurements.

• Example: A botanist records the number of petals on 5 flowers:

\[ 5, 12, 6, 8, 14 \]

• The range is \( R = 14 - 5 = 9 \).

• Quick and easy, but only uses 2 of the 5 measurements.
The Variance

• The variance is a measure of variability that uses all the measurements. It measures the average deviation of the measurements about their mean.

• Flower petals: 5, 12, 6, 8, 14

\[ \bar{x} = \frac{45}{5} = 9 \]
The Variance

• The variance of a population of \( N \) measurements is the average of the squared deviations of the measurements about their mean \( \mu \).

\[
\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}
\]

• The variance of a sample of \( n \) measurements is the sum of the squared deviations of the measurements about their mean, divided by \((n - 1)\).

\[
s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}
\]
In calculating the variance, we squared all of the deviations, and in doing so changed the scale of the measurements.

To return this measure of variability to the original units of measure, we calculate the standard deviation, the positive square root of the variance.

Population standard deviation : $\sigma = \sqrt{\sigma^2}$

Sample standard deviation : $s = \sqrt{s^2}$
Two Ways to Calculate the Sample Variance

Use the Definition Formula:

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

\[
s = \sqrt{s^2} = \sqrt{15} = 3.87
\]

<table>
<thead>
<tr>
<th>(x_i)</th>
<th>(x_i - \bar{x})</th>
<th>((x_i - \bar{x})^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>-4</td>
<td>16</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>8</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td>25</td>
</tr>
</tbody>
</table>
| Sum   | 45              | 0               | 60

\[
\sum (x_i - \bar{x})^2 = 60
\]

Copyright ©2006 Brooks/Cole
A division of Thomson Learning, Inc.
Two Ways to Calculate the Sample Variance

Use the Calculational Formula:

\[ s^2 = \frac{\sum x_i^2 - (\sum x_i)^2}{n} \]

\[ s^2 = \frac{465 - \frac{45^2}{5}}{4} = \frac{15}{4} = 15 \]

\[ s = \sqrt{s^2} = \sqrt{15} = 3.87 \]

<table>
<thead>
<tr>
<th>( x_i )</th>
<th>( x_i^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>144</td>
</tr>
<tr>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>8</td>
<td>64</td>
</tr>
<tr>
<td>14</td>
<td>196</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td><strong>45</strong></td>
</tr>
</tbody>
</table>
• The value of $s$ is **ALWAYS** positive.
• The larger the value of $s^2$ or $s$, the larger the variability of the data set.
• **Why divide by $n - 1$?**
  – The sample standard deviation $s$ is often used to estimate the population standard deviation $\sigma$. Dividing by $n - 1$ gives us a better estimate of $\sigma$. 
Measures of Relative Standing

- Where does one particular measurement stand in relation to the other measurements in the data set?
- How many standard deviations away from the mean does the measurement lie? This is measured by the $z$-score.

$$z = \frac{x - \bar{x}}{s}$$

Suppose $s = 2$. $s$

$x = 9$ lies $z = 2$ std dev from the mean.
Measures of Relative Standing

• How many measurements lie below the measurement of interest? This is measured by the $p^{th}$ percentile.

$p$-th percentile
Examples

• 90% of all men (16 and older) earn more than $319 per week.

$319 is the 10\textsuperscript{th} percentile.

50\textsuperscript{th} Percentile $\equiv$ Median

25\textsuperscript{th} Percentile $\equiv$ Lower Quartile ($Q_{1}$)

75\textsuperscript{th} Percentile $\equiv$ Upper Quartile ($Q_{3}$)
The value of the $p$th percentile can be calculated as follows:

**Position of $p$th percentile**

$$
0.\,p\,(n + 1)
$$

where $p$ denotes the number of the percentile and $n$ represents the sample size.
Example

The prices ($) of 18 brands of walking shoes:
40 60 65 65 65 68 68 70 70
70 70 70 70 74 75 75 90 95

Position of 42nd percentile = 0.42(18+1) 
= 7.98

42^{nd} percentile = 68 + .98(70 - 68) = 69.96

Interpretation?
80^{th} percentile?
Quartiles and the IQR

- The **lower quartile** \((Q_1)\) is the value of \(x\) which is larger than 25% and less than 75% of the ordered measurements.
- The **upper quartile** \((Q_3)\) is the value of \(x\) which is larger than 75% and less than 25% of the ordered measurements.
- The range of the “middle 50%” of the measurements is the **interquartile range**, 
  \[ IQR = Q_3 - Q_1 \]
Calculating Sample Quartiles

- The lower and upper quartiles ($Q_1$ and $Q_3$), can be calculated as follows:
- The position of $Q_1$ is $\frac{.25(n + 1)}{}$.
- The position of $Q_3$ is $\frac{.75(n + 1)}{}$.

Once the measurements have been ordered. If the positions are not integers, find the quartiles by interpolation.
Example

The prices ($) of 18 brands of walking shoes:
40 60 65 65 65 68 68 70 70
70 70 70 70 74 75 75 90 95

Position of $Q_1 = .25(18 + 1) = 4.75$

Position of $Q_3 = .75(18 + 1) = 14.25$

$Q_1$ is 3/4 of the way between the 4th and 5th ordered measurements, or

$Q_1 = 65 + .75(65 - 65) = 65.$
Example

The prices ($)) of 18 brands of walking shoes:

40  60  65  65  65  68  68  70  70
70  70  70  70  74  75  75  90  95

Position of $Q_1 = .25(18 + 1) = 4.75$

Position of $Q_3 = .75(18 + 1) = 14.25$

✓ $Q_3$ is 1/4 of the way between the 14th and 15th ordered measurements, or

$Q_3 = 74 + .25(75 - 74) = 74.25$

✓ and

$IQR = Q_3 - Q_1 = 74.25 - 65 = 9.25$
Using Measures of Center and Spread: The Box Plot

The Five-Number Summary:

<table>
<thead>
<tr>
<th>Min</th>
<th>Q₁</th>
<th>Median</th>
<th>Q₃</th>
<th>Max</th>
</tr>
</thead>
</table>

- Divides the data into 4 sets containing an equal number of measurements.
- A quick summary of the data distribution.
- Use to form a box plot to describe the shape of the distribution and to detect outliers.
Constructing a Box Plot

- Calculate $Q_1$, the median, $Q_3$ and IQR.
- Draw a horizontal line to represent the scale of measurement.
- Draw a box using $Q_1$, the median, $Q_3$. 
Constructing a Box Plot

- Isolate outliers by calculating
  - Lower fence: $Q_1 - 1.5 \text{ IQR}$
  - Upper fence: $Q_3 + 1.5 \text{ IQR}$

- Measurements beyond the upper or lower fence are outliers and are marked (*).
Constructing a Box Plot

✓ Draw “whiskers” connecting the largest and smallest measurements that are NOT outliers to the box.
Example

Amt of sodium in 8 brands of cheese:

260  290  300  320  330  340  340  520

\[ Q_1 = 292.5 \quad m = 325 \quad Q_3 = 340 \]
Example

IQR = 340 - 292.5 = 47.5
Lower fence = 292.5 - 1.5(47.5) = 221.25
Upper fence = 340 + 1.5(47.5) = 411.25
Outlier: \( x = 520 \)

\[ Q_1 \]
\[ Q_3 \]
Interpreting Box Plots

- Median line in center of box and whiskers of equal length—symmetric distribution
- Median line left of center and long right whisker—skewed right
- Median line right of center and long left whisker—skewed left
Key Concepts

I. Measures of Center

1. Arithmetic mean (mean) or average
   a. Population: $\mu$
   b. Sample of size $n$: $\bar{x} = \frac{\sum x_i}{n}$

2. Median: position of the median $= .5(n +1)$

3. Mode

4. The median may be preferred to the mean if the data are highly skewed.

II. Measures of Variability

1. Range: $R = \text{largest} - \text{smallest}$
Key Concepts

2. Variance
   a. Population of \( N \) measurements:
   \[ \sigma^2 = \frac{\sum(x_i - \mu)^2}{N} \]
   b. Sample of \( n \) measurements:
   \[ s^2 = \frac{\sum(x_i - \bar{x})^2}{n - 1} = \frac{\sum x_i^2 - (\sum x_i)^2}{n} \]

3. Standard deviation
   Population standard deviation : \( \sigma = \sqrt{\sigma^2} \)
   Sample standard deviation : \( s = \sqrt{s^2} \)

4. A rough approximation for \( s \) can be calculated as \( s \approx R/4 \).
   The divisor can be adjusted depending on the sample size.
Key Concepts

IV. Measures of Relative Standing

1. Sample $z$-score:
2. $p$th percentile; $p\%$ of the measurements are smaller, and 
   $(100 - p)\%$ are larger.
3. Lower quartile, $Q_1$; position of $Q_1 = .25(n + 1)$
4. Upper quartile, $Q_3$; position of $Q_3 = .75(n + 1)$
5. Interquartile range: $IQR = Q_3 - Q_1$

V. Box Plots

1. Box plots are used for detecting outliers and shapes of 
   distributions.
2. $Q_1$ and $Q_3$ form the ends of the box. The median line is in 
   the interior of the box.
Key Concepts

3. Upper and lower fences are used to find outliers.
   a. Lower fence: \( Q_1 - 1.5 \times (\text{IQR}) \)
   b. Outer fences: \( Q_3 + 1.5 \times (\text{IQR}) \)

4. **Whiskers** are connected to the smallest and largest measurements that are not outliers.

5. Skewed distributions usually have a long whisker in the direction of the skewness, and the median line is drawn away from the direction of the skewness.
Grouped Data
Sample Mean

• Ungrouped data:

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + \ldots + x_n}{n}
\]

- \( n \) = number of observation

• Grouped data:

\[
\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i^*}{n} = \frac{1}{n} \left( \sum_{i=1}^{k} f_i x_i^* \right)
\]

- \( n \) = sum of the frequencies
- \( k \) = number class
- \( f_i \) = frequency of each class
- \( x_i^* \) = midpoint of each class

Copyright ©2006 Brooks/Cole
A division of Thomson Learning, Inc.
Example: - *ungrouped data*

- Resistance of 5 coils:
  3.35, 3.37, 3.28, 3.34, 3.30 ohm.

- The average:

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{3.35 + 3.37 + 3.28 + 3.34 + 3.30}{5} = 3.33 \ \Omega
\]
## Example: - grouped data

Frequency Distributions of the life of 320 tires in 1000 km

<table>
<thead>
<tr>
<th>Boundaries</th>
<th>Midpoint $x_i^*$</th>
<th>Frequency, $f_i$</th>
<th>$f_i x_i^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>23.5-26.5</td>
<td>25.0</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>26.5-29.5</td>
<td>28.0</td>
<td>36</td>
<td>1008</td>
</tr>
<tr>
<td>29.5-32.5</td>
<td>31.0</td>
<td>51</td>
<td>1581</td>
</tr>
<tr>
<td>32.5-35.5</td>
<td>34.0</td>
<td>63</td>
<td>2142</td>
</tr>
<tr>
<td>35.5-38.5</td>
<td>37.0</td>
<td>58</td>
<td>2146</td>
</tr>
<tr>
<td>38.5-41.5</td>
<td>40.0</td>
<td>52</td>
<td>2080</td>
</tr>
<tr>
<td>41.5-44.5</td>
<td>43.0</td>
<td>34</td>
<td>1462</td>
</tr>
<tr>
<td>44.5-47.5</td>
<td>46.0</td>
<td>16</td>
<td>736</td>
</tr>
<tr>
<td>47.5-50.5</td>
<td>49.0</td>
<td>6</td>
<td>294</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>n = 320</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i^*}{n} = \frac{11,549}{320} = 36.1
\]

Copyright ©2006 Brooks/Cole
A division of Thomson Learning, Inc.
Sample Variance

- **Ungrouped data:**

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

- **Grouped data:**

\[
s^2 = \frac{\sum_{i=1}^{k} (f_i x_i^*)^2 - \left(\sum_{i=1}^{k} f_i x_i^*\right)^2}{n(n-1)}
\]
Example: - *ungrouped data*

- Sample of six moisture content (%) of kraft paper are:
  
  \[
  6.7, 6.0, 6.4, 6.4, 5.9, \text{ and } 5.8
  \]

  \[
  s = \sqrt{\frac{(231.26) - (37.2)^2 / 6}{(6 - 1)}} = 0.35
  \]

- Sample standard deviation, \( s = 0.35 \% \)
### Example: - *grouped data*

**Standard deviation for a grouped sample:**

**Table: Car speeds in km/h**

<table>
<thead>
<tr>
<th>Boundaries</th>
<th>$x_i^*$</th>
<th>$f_i$</th>
<th>$f_i x_i^*$</th>
<th>$f_i x_i^{*2}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>72.5-81.5</td>
<td>77.0</td>
<td>5</td>
<td>385</td>
<td>29645</td>
</tr>
<tr>
<td>81.5-90.5</td>
<td>86.0</td>
<td>19</td>
<td>1634</td>
<td>140524</td>
</tr>
<tr>
<td>90.5-99.5</td>
<td>95</td>
<td>31</td>
<td>2945</td>
<td>279775</td>
</tr>
<tr>
<td>99.5-108.5</td>
<td>104.0</td>
<td>27</td>
<td>2808</td>
<td>292032</td>
</tr>
<tr>
<td>108.5-117.5</td>
<td>113</td>
<td>14</td>
<td>1582</td>
<td>178766</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>96</td>
<td></td>
<td>9354</td>
<td>920742</td>
</tr>
</tbody>
</table>

$$s = \sqrt{\frac{\sum_{i=1}^{k} (f_i x_i^{*2}) - \left( \sum_{i=1}^{k} f_i x_i^* \right)^2}{n}}$$

$$s = \sqrt{\frac{(920,742) - (9354)^2/96}{(96-1)}} = 9.9$$
Mode – Grouped Data

Mode
• Mode is the value that has the highest frequency in a data set.
• For grouped data, class mode (or, modal class) is the class with the highest frequency.
• To find mode for grouped data, use the following formula:

\[
\text{Mode} = L + \frac{d_1}{d_1 + d_2} \cdot l
\]

Where:

- \(L\) = Lower boundary of the modal class
- \(d_1\) = frequency of modal class – frequency of the class before
- \(d_2\) = frequency of modal class – frequency of the class after
- \(l\) = width of the modal class
### Example of Grouped Data (Mode)

Based on the grouped data below, find the mode

<table>
<thead>
<tr>
<th>Class Limit</th>
<th>Boundaries</th>
<th>f</th>
<th>Cum Freq</th>
<th>Cum Rel Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>99 - 108</td>
<td>98.5 - 108.5</td>
<td>6</td>
<td>6</td>
<td>0.150</td>
</tr>
<tr>
<td>109 - 118</td>
<td>108.5 - 118.5</td>
<td>7</td>
<td>13</td>
<td>0.325</td>
</tr>
<tr>
<td>119 - 128</td>
<td>118.5 - 128.5</td>
<td>13</td>
<td>26</td>
<td>0.650</td>
</tr>
<tr>
<td>129 - 138</td>
<td>128.5 - 138.5</td>
<td>8</td>
<td>34</td>
<td>0.850</td>
</tr>
<tr>
<td>139 - 148</td>
<td>138.5 - 148.5</td>
<td>6</td>
<td>40</td>
<td>0.975</td>
</tr>
</tbody>
</table>

**Modal Class** = 119 – 128 (third)

- $L = 118.5$
- $l = 10$
- $d_1 = 13 - 7 = 6$
- $d_2 = 13 - 8 = 5$

**Mode** = $L + \frac{d_1}{d_1 + d_2} l$

= $118.5 + \frac{6}{10} \cdot 10$

= 123.95
Percentile: - grouped data

\[ Sp = LBC_p + \frac{k - CFB}{f_p} \cdot l \]

\[ k = \frac{n \times p}{100} \]

- \( n \) = number of observations
- \( LBC_p \) = Lower Boundary of the percentile class
- \( CFB \) = Cumulative frequency of the class before \( C_p \)
- \( l \) = Class width
- \( f_p \) = Frequency of percentile class
Example - Percentile for grouped data

- Using the previous table of freq, find the median and 80\textsuperscript{th} percentile?

<table>
<thead>
<tr>
<th>Class Limit</th>
<th>Boundaries</th>
<th>f</th>
<th>Cum Freq</th>
<th>Cum Rel Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>99 - 108</td>
<td>98.5 - 108.5</td>
<td>6</td>
<td>6</td>
<td>0.150</td>
</tr>
<tr>
<td>109 - 118</td>
<td>108.5 - 118.5</td>
<td>7</td>
<td>13</td>
<td>0.325</td>
</tr>
<tr>
<td>119 - 128</td>
<td>118.5 - 128.5</td>
<td>13</td>
<td>26</td>
<td>0.650</td>
</tr>
<tr>
<td>129 - 138</td>
<td>128.5 - 138.5</td>
<td>8</td>
<td>34</td>
<td>0.850</td>
</tr>
<tr>
<td>139 - 148</td>
<td>138.5 - 148.5</td>
<td>6</td>
<td>40</td>
<td>0.975</td>
</tr>
</tbody>
</table>
Median, $S_{50}$

\[ k = \frac{40 \times 50}{100} = 20 \]

\[ C_{50} = (119 - 128) \text{ (3rd class)} \]

\[ LBC_p = 118.5 \]

\[ CFB = 13 \]

\[ l = 10 \]

\[ f_p = 13 \]

\[ S_{50} = 118.5 + \left( \frac{20 - 13}{13} \right) \times 10 \]

\[ = 128.885 \]

80th percentile, $S_{80}$

\[ k = \frac{40 \times 80}{100} = 32 \]

\[ C_{80} = (129 - 138) \text{ (4th class)} \]

\[ LBC_p = 128.5 \]

\[ CFB = 26 \]

\[ l = 10 \]

\[ f_p = 8 \]

\[ S_{80} = 128.5 + \left( \frac{32 - 26}{8} \right) \times 10 \]

\[ = 136 \]
Coefficient of variation

The coefficient of variation (CV) of a data set describes the standard deviation as a percent of the mean.

Population: \[ CV = \frac{\sigma}{\mu} \cdot 100\% \]

Sample: \[ CV = \frac{s}{\bar{x}} \cdot 100\% \]