Chapter 4
Descriptive Statistical Measures
Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers
Populations and Samples

- **Population** - all items of interest for a particular decision or investigation
  - *all* married drivers over 25 years old
  - *all* subscribers to Netflix

- **Sample** - a subset of the population
  - a list of individuals who rented a comedy from Netflix in the past year

The purpose of sampling is to **obtain sufficient information to draw a valid conclusion** about a population.

Is the Netflix sample above a good sample? Why?
Other ways to select a sample?
We typically label the elements of a data set using subscripted variables, $x_1, x_2, \ldots$, and so on, where $x_i$ represents the $i^{th}$ observation. Upper-case letters like $X$ represent often random variables.

It is common practice in statistics to use Greek letters, such as $\mu$ (mu; mean), $\sigma$ (sigma; std. deviation), and $\pi$ (pi; proportion), to represent population measures and italic letters such as by $\bar{x}$ (called x-bar), $s$, and $p$ to represent sample statistics.

$N$ represents the number of items in a population and $n$ represents the number of observations in a sample.
Content

- Notation
- Measures of Location
  - Mean
  - Median
  - Mode
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers
Measures of Location: Arithmetic Mean

- Population mean:
  \[ \mu = \frac{\sum_{i=1}^{N} x_i}{N} \]  

- Sample mean:
  \[ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \]  

- Excel function: =AVERAGE(data range)

- Property of the mean:
  \[ \sum_{i} (x_i - \bar{x}) = 0 \]  

- Outliers can affect the value of the mean.

- Mean valid for interval/ratio variables and often questionable for ordinal variables.
Example 4.1: Computing Mean Cost per Order

*Purchase Orders* database

- Using formula:

\[
\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

\[=\text{SUM(B2:B95)/COUNT(B2:B95)}\]

Mean = \$2,471,760/94

= \$26,295.32

- Using Excel AVERAGE Function

\[=\text{AVERAGE(B2:B95)}\]
### Outlier

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
<th>Person</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>3</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>999</td>
<td>5</td>
<td>999</td>
</tr>
<tr>
<td>6</td>
<td>22</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>11</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>8</td>
<td>25</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td><strong>141.00</strong></td>
<td><strong>Mean</strong></td>
<td><strong>18.43</strong></td>
</tr>
</tbody>
</table>

Wikipedia: In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to **variability in the measurement** or it may indicate **experimental error**; the latter are sometimes excluded from the data set.
The median specifies the middle value when the data are arranged from least to greatest.

- Half the data are below the median, and half the data are above it.
- For an odd number of observations, the median is the middle of the sorted numbers.
- For an even number of observations, the median is the mean of the two middle numbers.

We could use the Sort option in Excel to rank-order the data and then determine the median. The Excel function \(\text{=MEDIAN(data range)}\) could also be used.

- The median is meaningful for ratio, interval, and ordinal data.
- Not affected by outliers.
Example 4.2: Finding the Median Cost per Order

- Sort the data from smallest to largest. Since we have 90 observations, the median is the average of the 47th and 48th observation.

\[
\text{Median} = \frac{($15,562.50 + $15,750.00)}{2} = $15,656.25
\]

= \text{MEDIAN(B2:B94)}
# Median

<table>
<thead>
<tr>
<th>Person</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17.00</td>
</tr>
<tr>
<td>2</td>
<td>21.00</td>
</tr>
<tr>
<td>3</td>
<td>15.00</td>
</tr>
<tr>
<td>4</td>
<td>18.00</td>
</tr>
<tr>
<td>5</td>
<td>999.00</td>
</tr>
<tr>
<td>6</td>
<td>22.00</td>
</tr>
<tr>
<td>7</td>
<td>11.00</td>
</tr>
<tr>
<td>8</td>
<td>25.00</td>
</tr>
<tr>
<td>Mean</td>
<td>141.00</td>
</tr>
<tr>
<td>Median</td>
<td>19.50</td>
</tr>
</tbody>
</table>

Median is insensitive to outliers!
The **mode** is the observation that occurs most frequently.

The mode is most useful for data sets that contain a relatively small number of unique values.

You can easily identify the mode from a frequency distribution by identifying the value or group having the largest frequency or from a histogram by identifying the highest bar.

Excel function: `=MODE.SNGL(data range)`.

For multiple modes: `=MODE.MULT(data range)`.
Example 4.3: Finding the Mode

- *Purchase Orders* database: A/P Terms
  - Mode = 30 months

- Cost per order
  - Mode is the group between $0 and $13,000
Measures of Location: Midrange

- The **midrange** is the average of the greatest and least values in the data set.

- Caution must be exercised when using the midrange because extreme values easily distort the result. This is because the midrange uses only two pieces of data, whereas the mean uses all the data; thus, it is usually a much rougher estimate than the mean and is often used for only small sample sizes.

- Best not used!
The Excel file *Computer Repair Times* includes 250 repair times for customers.

- What repair time would be reasonable to quote to a new customer?
- Median repair time is 2 weeks; mean and mode are about 15 days.
- Examine the histogram.
Example 4.5 (continued)

90% are completed within 3 weeks

Distribution is important!
Content

- Notation
- Measures of Location
- Measures of Dispersion
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
  - Empirical Rules
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers
Measures of Dispersion

- **Dispersion** refers to the degree of variation in the data; that is, the numerical spread (or compactness) of the data.

- **Key measures:**
  - Range
  - Interquartile range
  - Variance
  - Standard deviation
The range is the simplest and is the difference between the maximum value and the minimum value in the data set.

In Excel, compute as =MAX(data range) - MIN(data range).

The range is affected by outliers, and is often used only for very small data sets.
Example 4.6: Computing the Range

- *Purchase Orders* data

- For the cost per order data:
  - Maximum = $127,500
  - Minimum = $68.78

- Range = $127,500 - $68.78 = $127,431.22
The interquartile range (IQR), or the midspread is the difference between the first and third quartiles, Q3 – Q1.

This includes only the middle 50% of the data and, therefore, is not influenced by extreme values.
Example 4.7: Computing the Interquartile Range

- *Purchase Orders* data
- For the Cost per order data:
  - Third Quartile = \( Q_3 = $27,593.75 \)
  - First Quartile = \( Q_1 = $6,757.81 \)
- Interquartile Range = \( $27,593.75 - $6,757.81 = $20,835.94 \)
The variance is the “average” of the squared deviations from the mean.

For a population:

\[ \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \]  

- In Excel: `=VAR.P(data range)`

For a sample:

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

- In Excel: `=VAR.S(data range)`

Note the difference in denominators!
Example 4.8 Computing the Variance

- **Purchase Orders Cost per order data**

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Observation</td>
<td>Cost per order</td>
<td>(xi - mean)</td>
<td>(xi - mean)^2</td>
</tr>
<tr>
<td>2</td>
<td>x1</td>
<td>$2,700.00</td>
<td>-$23,595.32</td>
<td>$556,739,085.74</td>
</tr>
<tr>
<td>3</td>
<td>x2</td>
<td>$19,250.00</td>
<td>-$7,045.32</td>
<td>$49,636,521.91</td>
</tr>
<tr>
<td>4</td>
<td>x3</td>
<td>$15,937.50</td>
<td>-$10,357.82</td>
<td>$107,284,417.52</td>
</tr>
<tr>
<td>5</td>
<td>x4</td>
<td>$18,150.00</td>
<td>-$8,145.32</td>
<td>$66,346,224.04</td>
</tr>
<tr>
<td>93</td>
<td>x92</td>
<td>$74,375.00</td>
<td>$48,079.68</td>
<td>$2,311,655,710.74</td>
</tr>
<tr>
<td>94</td>
<td>x93</td>
<td>$72,250.00</td>
<td>$45,954.68</td>
<td>$2,111,832,692.12</td>
</tr>
<tr>
<td>95</td>
<td>x94</td>
<td>$6,562.50</td>
<td>-$19,732.82</td>
<td>$389,384,151.56</td>
</tr>
<tr>
<td>96</td>
<td>Sum of cost/order</td>
<td>$2,471,760.00</td>
<td>Sum of squared deviations</td>
<td>$82,825,295,365.68</td>
</tr>
<tr>
<td>97</td>
<td>Number of observations</td>
<td>94</td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>Mean cost/order</td>
<td>$26,295.32</td>
<td>Variance</td>
<td>890,594,573.82</td>
</tr>
<tr>
<td>100</td>
<td>Excel VAR.S function</td>
<td></td>
<td></td>
<td>890,594,573.82</td>
</tr>
</tbody>
</table>

\[
 s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n - 1} \quad (4.5)
\]
Measures of Dispersion: Standard Deviation

- The **standard deviation** is the square root of the variance.
  - Note that the dimension of the variance is the square of the dimension of the observations, whereas the dimension of the standard deviation is the same as the data. This makes the standard deviation more practical to use in applications.

- For a population:
  \[ \sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \]  
  - In Excel: `=STDEV.P(data range)`

- For a sample:
  \[ s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}} \]  
  - In Excel: `=STDEV.S(data range)`
Example 4.9 Computing the Standard Deviation

- *Purchase Orders* Cost per order data

- Using the results of Example 4.8, take the square root of the variance:

\[ \sqrt{890,594,573.82} = 29,842.8312. \]

- Alternatively, use the STDEV.S function for the data range.
Standard Deviation as a Measure of Risk

Excel file: Closing Stock Prices
Intel (INTC):
  Mean = $18.81
  Standard deviation = $0.50

General Electric (GE):
  Mean = $16.19
  Standard deviation = $0.35

INTC is a higher risk investment than GE.
Empirical Rules

- For many data sets encountered in practice:
  - Approximately 68% of the observations fall within one standard deviation of the mean \( \bar{x} - s \) and \( \bar{x} + s \)
  - Approximately 95% fall within two standard deviations of the mean \( \bar{x} \pm 2s \)
  - Approximately 99.7% fall within three standard deviations of the mean \( \bar{x} \pm 3s \)

- These rules are commonly used to characterize the natural variation in manufacturing processes and other business phenomena.
Empirical Rules

- The empirical Rule comes from the normal distribution.

Most data does not follow a normal distribution!
Chebyshev’s Theorem

For any data set (any distribution), the proportion of values that lie within +/- \( k \) \((k > 1)\) standard deviations of the mean is at least \(1 - \frac{1}{k^2}\)

Examples:
- For \( k = 2 \): at least \(\frac{3}{4}\) or 75% of the data lie within two standard deviations of the mean
- For \( k = 3 \): at least \(\frac{8}{9}\) or 89% of the data lie within three standard deviations of the mean
Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers
A standardized value, commonly called a z-score, provides a relative measure of the distance an observation is from the mean, which is independent of the units of measurement.

The z-score for the i\(^{th}\) observation in a data set is calculated as follows:

\[ z_i = \frac{x_i - \bar{x}}{s} \]  

Excel function: =STANDARDIZE(x, mean, standard_dev).

Standardized data is needed by many predictive methods since it makes variables comparable.
Example 4.12 Computing z-Scores

- **Purchase Orders** Cost per order data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observation</td>
<td>Cost per order</td>
<td>z-score</td>
</tr>
<tr>
<td>2</td>
<td>x1</td>
<td>$2,700.00</td>
</tr>
<tr>
<td>3</td>
<td>x2</td>
<td>$19,250.00</td>
</tr>
<tr>
<td>4</td>
<td>x3</td>
<td>$15,937.50</td>
</tr>
<tr>
<td>5</td>
<td>x4</td>
<td>$18,150.00</td>
</tr>
<tr>
<td>6</td>
<td>x5</td>
<td>$23,400.00</td>
</tr>
<tr>
<td>91</td>
<td>x90</td>
<td>$6,750.00</td>
</tr>
<tr>
<td>92</td>
<td>x91</td>
<td>$16,625.00</td>
</tr>
<tr>
<td>93</td>
<td>x92</td>
<td>$74,375.00</td>
</tr>
<tr>
<td>94</td>
<td>x93</td>
<td>$72,250.00</td>
</tr>
<tr>
<td>95</td>
<td>x94</td>
<td>$6,562.50</td>
</tr>
<tr>
<td>96</td>
<td>Mean</td>
<td>$26,295.32</td>
</tr>
<tr>
<td>97</td>
<td>Standard Deviation</td>
<td>$29,842.83</td>
</tr>
</tbody>
</table>

\[
\text{Cost per order data} = \frac{(B2 - \text{Mean})}{\text{Standard Deviation}}, \text{ or } \text{STANDARDIZE}(B2, \text{Mean}, \text{Standard Deviation}).
\]
Excel Descriptive Statistics Tool

This tool provides a summary of numerical statistical measures for sample data.

Data >
Data Analysis >
Descriptive Statistics

- Enter Input Range
- Labels (optional)
- Check Summary Statistics box

- The data must be in a single row or column. If the data are in multiple columns, the tool treats each row or column as a separate data set

You may need to activate the Data Analysis Tool Pack in File>Options>Add-Ins
Example 4.15: Using the Descriptive Statistics Tool

- Purchase Orders database

Note: Results of the Analysis Toolpak do not change when changes are made to the data.
Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers
Descriptive Statistics for Categorical Data: The Proportion

- The proportion, denoted by $p$, is the fraction of data that have a certain characteristic.

- Proportions are key descriptive statistics for categorical data, such as defects or errors in quality control applications or consumer preferences in market research.

- Example: Proportion of female students is 60%.
**Example 4.18: Computing a Proportion**

- Proportion of orders placed by Spacetime Technologies
  
  $= \text{COUNTIF(A4:A97, “Spacetime Technologies”)}/94$

  $= 12/94 = 0.128$

---

<table>
<thead>
<tr>
<th>Supplier</th>
<th>Order No.</th>
<th>Item No.</th>
<th>Item Description</th>
<th>Item Cost</th>
<th>Quantity</th>
<th>Cost per order</th>
<th>A/P Terms (Months)</th>
<th>Order Date</th>
<th>Arrival Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spacetime Technologies</td>
<td>A0111</td>
<td>6489</td>
<td>O-Ring</td>
<td>$3.00</td>
<td>900</td>
<td>$2,700.00</td>
<td>25</td>
<td>10/10/11</td>
<td>10/18/11</td>
</tr>
<tr>
<td>Steelpin Inc.</td>
<td>A0115</td>
<td>5319</td>
<td>Shielded Cable/ft.</td>
<td>$1.10</td>
<td>17,500</td>
<td>$19,250.00</td>
<td>30</td>
<td>08/20/11</td>
<td>08/31/11</td>
</tr>
<tr>
<td>Steelpin Inc.</td>
<td>A0123</td>
<td>4312</td>
<td>Bolt-nut package</td>
<td>$3.75</td>
<td>4,250</td>
<td>$15,937.50</td>
<td>30</td>
<td>08/25/11</td>
<td>09/01/11</td>
</tr>
<tr>
<td>Steelpin Inc.</td>
<td>A0204</td>
<td>5319</td>
<td>Shielded Cable/ft.</td>
<td>$1.10</td>
<td>16,500</td>
<td>$18,150.00</td>
<td>30</td>
<td>09/15/11</td>
<td>10/05/11</td>
</tr>
<tr>
<td>Steelpin Inc.</td>
<td>A0205</td>
<td>5677</td>
<td>Side Panel</td>
<td>$195.00</td>
<td>120</td>
<td>$23,400.00</td>
<td>30</td>
<td>11/02/11</td>
<td>11/13/11</td>
</tr>
<tr>
<td>Steelpin Inc.</td>
<td>A0207</td>
<td>4312</td>
<td>Bolt-nut package</td>
<td>$3.75</td>
<td>4,200</td>
<td>$15,750.00</td>
<td>30</td>
<td>09/01/11</td>
<td>09/10/11</td>
</tr>
<tr>
<td>Alum Sheeting</td>
<td>A0223</td>
<td>4224</td>
<td>Bolt-nut package</td>
<td>$3.95</td>
<td>4,500</td>
<td>$17,775.00</td>
<td>30</td>
<td>10/15/11</td>
<td>10/20/11</td>
</tr>
</tbody>
</table>
Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
  - Correlation
- Outliers
Two variables have a strong statistical relationship with one another if they appear to “move” together.

When two variables appear to be related, you might suspect a cause-and-effect relationship.

**Caution: Correlation does not prove causation!** Statistical relationships may exist even though a change in one variable is not caused by a change in the other.
**Measures of Association: Covariance**

- **Covariance** is a measure of the linear association between two variables, $X$ and $Y$. Like the variance, different formulas are used for populations and samples.

- **Population covariance:**
  \[
  \text{cov}(X, Y) = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}
  \]  
  (4.17)

  - Excel function: `=COVARIANCE.P(array1,array2)`

- **Sample covariance:**
  \[
  \text{cov}(X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}
  \]  
  (4.18)

  - Excel function: `=COVARIANCE.S(array1,array2)`

- The covariance between $X$ and $Y$ is the average of the product of the deviations of each pair of observations from their respective means.
Example 4.20: Computing the Covariance

Colleges and Universities data

![Graduation % vs Median SAT](chart)

<table>
<thead>
<tr>
<th></th>
<th>Graduation % (X)</th>
<th>Median SAT (Y)</th>
<th>X - Mean(X)</th>
<th>Y - Mean(Y)</th>
<th>(X - Mean(X))(Y - Mean(Y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>93</td>
<td>1315</td>
<td>9.755</td>
<td>51.898</td>
<td>506.2698875</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>1220</td>
<td>-3.245</td>
<td>-43.102</td>
<td>139.8617243</td>
</tr>
<tr>
<td>4</td>
<td>88</td>
<td>1240</td>
<td>4.755</td>
<td>-23.102</td>
<td>-109.8525614</td>
</tr>
<tr>
<td>47</td>
<td>86</td>
<td>1250</td>
<td>2.755</td>
<td>-13.102</td>
<td>-36.09745939</td>
</tr>
<tr>
<td>48</td>
<td>91</td>
<td>1290</td>
<td>7.755</td>
<td>26.898</td>
<td>208.5964182</td>
</tr>
<tr>
<td>49</td>
<td>93</td>
<td>1336</td>
<td>9.755</td>
<td>72.898</td>
<td>711.1270304</td>
</tr>
<tr>
<td>50</td>
<td>93</td>
<td>1350</td>
<td>9.755</td>
<td>86.898</td>
<td>847.698459</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Sum</th>
<th>Count</th>
<th>Covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>51</td>
<td>83.245</td>
<td>12641.77551</td>
<td>49</td>
<td>263.3703231</td>
</tr>
<tr>
<td>55</td>
<td>COVARIANCE.S</td>
<td>263.3703231</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measures of Association: Correlation

- **Correlation** is a measure of the linear relationship between two variables, $X$ and $Y$, which does not depend on the units of measurement.
- Correlation is measured by the correlation coefficient, also known as the **Pearson product moment correlation coefficient**.
- Correlation coefficient for a population:
  \[ \rho_{xy} = \frac{\text{cov}(X, Y)}{\sigma_x \sigma_y} \]  
  (4.19)
- Correlation coefficient for a sample:
  \[ r_{xy} = \frac{\text{cov}(X,Y)}{s_x s_y} \]  
  (4.20)
- The correlation coefficient is scaled between -1 and 1.
- Excel function: =CORREL(array1,array2)
Examples of Correlation

(a) Positive Correlation

(b) Negative Correlation

(c) No Correlation

(d) A Nonlinear Relationship with No Linear Correlation

Why is correlation important?
### Example 4.21 Computing the Correlation Coefficient

- **Colleges and Universities data**
- Is a school's graduation rate related to the SAT score of incoming students?

<table>
<thead>
<tr>
<th></th>
<th>Graduation % (X)</th>
<th>Median SAT (Y)</th>
<th>X - Mean(X)</th>
<th>Y - Mean(Y)</th>
<th>(X - Mean(X))(Y - Mean(Y))</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>93</td>
<td>1315</td>
<td>9.755</td>
<td>51.898</td>
<td>506.2698875</td>
</tr>
<tr>
<td>3</td>
<td>80</td>
<td>1220</td>
<td>-3.245</td>
<td>-43.102</td>
<td>139.8617243</td>
</tr>
<tr>
<td>4</td>
<td>88</td>
<td>1240</td>
<td>4.755</td>
<td>-23.102</td>
<td>-109.8252814</td>
</tr>
<tr>
<td>47</td>
<td>86</td>
<td>1250</td>
<td>2.755</td>
<td>-13.102</td>
<td>-36.09745939</td>
</tr>
<tr>
<td>48</td>
<td>91</td>
<td>1290</td>
<td>7.755</td>
<td>26.898</td>
<td>208.5964182</td>
</tr>
<tr>
<td>49</td>
<td>93</td>
<td>1336</td>
<td>9.755</td>
<td>72.898</td>
<td>711.1270304</td>
</tr>
<tr>
<td>50</td>
<td>93</td>
<td>1350</td>
<td>9.755</td>
<td>86.898</td>
<td>847.698459</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Graduation % (X)</td>
<td>83.245</td>
<td>7.449</td>
</tr>
<tr>
<td>Median SAT (Y)</td>
<td>1263.102</td>
<td>62.676</td>
</tr>
</tbody>
</table>

#### Calculating Covariance

\[
\text{cov} (X, Y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n - 1}
\]

Is there a causal relationship?
Excel Correlation Tool to create a Correlation Matrix

Data > Data Analysis > Correlation

- Excel computes the correlation coefficient between all pairs of variables in the *Input Range*. *Input Range* data must be in contiguous columns.
Example 4.22: Using the Correlation Tool Results in a Correlation Matrix

- **Colleges and Universities** data

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B: Median SAT</th>
<th>C: Acceptance Rate</th>
<th>D: Expenditures/Student</th>
<th>E: Top 10% HS</th>
<th>F: Graduation %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>Median SAT</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Median SAT</td>
<td>-0.601901959</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Acceptance Rate</td>
<td>-0.284254415</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Expenditures/Student</td>
<td>0.503467995</td>
<td>-0.609720972</td>
<td>0.505782049</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Top 10% HS</td>
<td>0.564146827</td>
<td>-0.55037751</td>
<td>0.042503514</td>
<td>0.138612667</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Graduation %</td>
<td>1</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- Moderate negative correlation between acceptance rate and graduation rate, indicating that schools with lower acceptance rates have higher graduation rates.
- Acceptance rate is also negatively correlated with the median SAT and Top 10% HS, suggesting that schools with lower acceptance rates have higher student profiles.
- The correlations with Expenditures/Student suggest that schools with higher student profiles spend more money per student.
Association between categorical and continuous Variables - PivotTables

*Value Field Settings* include several statistical measures:

- Average
- Max and Min
- Product
- Standard deviation
- Variance
Example 4.19: Statistical Measures in PivotTables

- Credit Risk Data
- First, create a PivotTable.
- In the PivotTable Field List, move Job to the Row Labels field and Checking and Savings to the Values field. Then change the field settings from “Sum of Checking” and “Sum of Savings” to the averages.

<table>
<thead>
<tr>
<th>Row Labels</th>
<th>Average of Checking</th>
<th>Average of Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management</td>
<td>$606.94</td>
<td>$1,616.83</td>
</tr>
<tr>
<td>Skilled</td>
<td>$1,079.24</td>
<td>$1,836.43</td>
</tr>
<tr>
<td>Unemployed</td>
<td>$1,697.64</td>
<td>$2,760.91</td>
</tr>
<tr>
<td>Unskilled</td>
<td>$1,140.27</td>
<td>$1,741.44</td>
</tr>
<tr>
<td>Grand Total</td>
<td>$1,048.01</td>
<td>$1,812.56</td>
</tr>
</tbody>
</table>
Content

- Notation
- Measures of Location
- Measures of Dispersion
- Standardization
- Proportions for Categorical Variables
- Measures of Association
- Outliers
Identifying Outliers

- There is no standard definition of what constitutes an outlier!

- Wikipedia: “In statistics, an outlier is an observation point that is distant from other observations. […] Outliers can occur by chance in any distribution, but they often indicate either measurement error or that the population has a heavy-tailed distribution.”

- If the outlier is due to a measurement error then we often want to exclude it from the analysis.

- Some typical rules of thumb:
  - **Normal distribution**: z-scores greater than +3 or less than -3
  - **Boxplot**:
    - Extreme outliers are more than 3*IQR to the left of Q₁ or right of Q₃
    - Mild outliers are between 1.5*IQR and 3*IQR to the left of Q₁ or right of Q₃
Example 4.23: Investigating Outliers

- **Home Market Value** data

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Home Market Value</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>House Age</td>
<td>Square Feet</td>
<td>z-score</td>
<td>Market Value</td>
</tr>
<tr>
<td>4</td>
<td>33</td>
<td>1.812</td>
<td>0.5300</td>
<td>$90,000.00</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>1.914</td>
<td>0.9931</td>
<td>$104,400.00</td>
</tr>
<tr>
<td>6</td>
<td>32</td>
<td>1.842</td>
<td>0.6662</td>
<td>$93,300.00</td>
</tr>
<tr>
<td>7</td>
<td>33</td>
<td>1.812</td>
<td>0.5300</td>
<td>$91,000.00</td>
</tr>
<tr>
<td>41</td>
<td>27</td>
<td>1.484</td>
<td>-0.9562</td>
<td>$81,300.00</td>
</tr>
<tr>
<td>42</td>
<td>27</td>
<td>1.520</td>
<td>-0.7667</td>
<td>$100,700.00</td>
</tr>
<tr>
<td>43</td>
<td>28</td>
<td>1.520</td>
<td>-0.7957</td>
<td>$87,200.00</td>
</tr>
<tr>
<td>44</td>
<td>27</td>
<td>1.684</td>
<td>-0.0611</td>
<td>$96,700.00</td>
</tr>
<tr>
<td>45</td>
<td>27</td>
<td>1.581</td>
<td>-0.5188</td>
<td>$120,700.00</td>
</tr>
</tbody>
</table>

- Mean = 1,695, Standard Deviation = 220.257

None of the z-scores exceed 3. However, while individual variables might not exhibit outliers, combinations of them might.

- The last observation has a high market value ($120,700) but a relatively small house size (1,581 square feet) and may be an outlier.
Example 4.24: Applying Statistical Thinking

- Excel file *Surgery Infections*
  - Is month 12 simply random variation or some explainable phenomenon?
Example 4.24 Continued

- Three-standard deviation empirical rule:

- There is only a 0.3% (for normally distributed data) or a 11% (for any distribution) chance to see an observation outside +/- 3 std.dev.

- This suggests that month 12 is statistically different from the rest of the data.