Interpretation of Standard Deviation
(See “Empirical Rule”, Chapter 2, p. 61)

If the shape of the dotplot or histogram is approximately bell-shaped, we would expect

- 68% of the data to be within 1 SD of the mean
- 95% of the data to be within 2 SD of the mean
- 99.7% of the data to be within 3 SD of the mean

Example: Student heights data (rounded to nearest 1 inch)

Mean is 66.78, SD is 3.82

66.78$\pm$3.82 is 62.96 to 70.60; 64 out of 86 students (74%) have heights within this range

Within 2SD of mean: 80/86 or 93%

Within 3SD of mean: 86/86 or 100%
**z-statistic rule for outliers**

1. Calculate the mean $\bar{x}$ and the standard deviation $s$.

2. Compute $z$-scores $z = \frac{x - \bar{x}}{s}$.

3. Any $z$-score outside $\pm 3$ is considered an outlier.

Example of CO$_2$ data:

Raw data 2.3 1.8 1.1 9.8 19.7 0.7 1.2 0.2

Mean $\bar{x} = 4.6$, standard deviation $s = 6.83$.

$z$ scores are -0.34 -0.41 -0.51 0.76 2.21 -0.57 -0.50 -0.64
Comments on this example:

1. One reason many statisticians prefer the 1.5 IQR rule is that the latter is based on resistant measures for outliers. In this example, the mean and the standard deviation are themselves inflated because of the outlier.

2. One solution might be to omit the potential outlier (in this example, USA), recompute $\bar{x}$ and $s$ based on the others, then calculate the $z$ score for USA based on those numbers. [In this case, revised $\bar{x}$ and $s$ are 2.4 and 3.3; $z$-score for the USA is 5.2, well outside $\pm 3$.]

3. However, this isn’t ideal either ... e.g. should we then omit Russia?
Chapter 3:
ASSOCIATION, CORRELATION AND REGRESSION
The **response variable** is the outcome variable on which comparisons are made.

The **explanatory variable** defines the groups to be compared with respect to values of the response variable.

**Association** means that the values of the response in some way depend on the explanatory variable. At this level of discussion, talking about association does not imply that there is an actual causal effect, because the association may be spurious (example of mortality rates in British women, grouped into smokers and non-smokers)
Contingency Tables

Used when we want to look at associations among two categorical variables.

Each entry or cell of the table contains the frequency of a particular combination of the two variables.

Note: Frequency is a count, not a proportion. We'll talk next about converting counts into proportions.
### Example Based on Political Affiliation by Year

<table>
<thead>
<tr>
<th>Party</th>
<th>2009</th>
<th>2010</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>35</td>
<td>26</td>
<td>61</td>
</tr>
<tr>
<td>Republican</td>
<td>21</td>
<td>42</td>
<td>63</td>
</tr>
<tr>
<td>Independent</td>
<td>12</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>68</td>
<td>81</td>
<td>149</td>
</tr>
</tbody>
</table>
Converting Frequencies to Proportions

The key point is that there are different ways to do this.

**Unconditional proportions**: express everything as proportion of the grand total (149).

<table>
<thead>
<tr>
<th>Party</th>
<th>2009</th>
<th>2010</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>.235</td>
<td>.174</td>
<td>.409</td>
</tr>
<tr>
<td>Republican</td>
<td>.141</td>
<td>.282</td>
<td>.423</td>
</tr>
<tr>
<td>Independent</td>
<td>.081</td>
<td>.087</td>
<td>.168</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>.456</td>
<td>.544</td>
<td>1.000</td>
</tr>
</tbody>
</table>
**Conditional proportions**: if we’re interested in comparing party affiliation by year, divide each column by the total for that column.

<table>
<thead>
<tr>
<th>Party</th>
<th>2009</th>
<th>2010</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Democrat</td>
<td>.515</td>
<td>.321</td>
<td>.409</td>
</tr>
<tr>
<td>Republican</td>
<td>.309</td>
<td>.519</td>
<td>.423</td>
</tr>
<tr>
<td>Independent</td>
<td>.176</td>
<td>.160</td>
<td>.168</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

We could also standardize by row instead of by column. Which one is more appropriate depends on the interpretation.
Associations of Categorical Variables

The question arising from all this is, when is there an association?

Two variables are associated if the conditional proportions of the response variable depend on the explanatory variable.

Note that this definition does not settle how large the samples need to be for the differences to be "significant".
Associations of Quantitative Variables

Different tools — leading role play by scatterplots.

Different uses for a scatterplot:

- Look for general associations, e.g. by plotting as trendline (option in Excel)

- A scatterplot can also be useful for detecting other features of the data, e.g. outliers.
Scatterplot of TV use against internet use
The “butterfly ballot”
Scatterplot of Buchanan vote against Bush vote in Florida 2000
Scatterplot of Buchanan vote against Gore vote in Florida 2000