

**ANNOUNCEMENTS**
Quiz begins at 1pm today and ends at noon Friday

**TODAY**
Sections 6.1 to 6.3. Read whatever we don’t have time to finish in those sections (if anything).

**HOMEWORK** (Due Fri, Oct 8)
Chapter 6: #23, 36
Don’t forget to put *Section number* on homework
S1 = 180 ICS @ 4pm    S2 = 174 ICS @ 4pm
S3 = 5pm    S4 = 6pm

---

**Chapter 6**

**Relationships Between Categorical Variables**

---

**6.1 Displaying Relationships Between Categorical Variables:**

Contingency Tables

- You did this in Discussion #1!
- Count the number of individuals who fall into each *combination* of categories.
- Present counts in table, called a **contingency table** or **two-way table**.
- Each row and column combination = **cell**.
- Row = *explanatory* variable.
- Column = *response* variable.

**Example: Aspirin and Heart Attacks**

**Case Study 1.6:**
Variable A = explanatory variable = aspirin or placebo
Variable B = response variable = heart attack or no heart attack

**Contingency Table** with explanatory as row variable, response as column variable, four cells. (Don’t count “Total” row and column.)

<table>
<thead>
<tr>
<th></th>
<th>Heart Attack</th>
<th>No Heart Attack</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aspirin</td>
<td>104</td>
<td>10,933</td>
<td>11,037</td>
</tr>
<tr>
<td>Placebo</td>
<td>189</td>
<td>10,845</td>
<td>11,034</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>293</strong></td>
<td><strong>21,778</strong></td>
<td><strong>22,071</strong></td>
</tr>
</tbody>
</table>

---

**Conditional Percentages (Rows)**

**Question of Interest:** Do the percentages in each category of the response variable change when the explanatory variable changes?

**Example: Find the Conditional (Row) Percentages**

**Aspirin Group:**
Percentage who had heart attacks = \( \frac{104}{11,037} = 0.0094 \) or 0.94%

**Placebo Group:**
Percentage who had heart attacks = \( \frac{189}{11,034} = 0.0171 \) or 1.71%

---

**Conditional Percentages (Columns)**

**Not usually of interest**

**Example: Find the Column Percentages**

**Heart Attack Group:**
Percentage who took aspirin = \( \frac{104}{293} = .355 \) or 35.5%

**No Heart Attack Group:**
Percentage who took aspirin = \( \frac{10,933}{21,778} = .50 \) or 50%
Review of Visual Displays for Contingency Tables: Bar Graphs (see p. 24)

Hopefully you learned this in Discussion 1!

- If there is a logical explanatory variable, create separate group of bars for each category of the explanatory variable.
- Within each group, draw bars for each category of the response variable.
- Use row percents, so heights of bars sum to 100% within each group of bars.
- Sometimes makes more sense to use actual counts.

Example 6.1 Smoking and Divorce

Data on smoking habits and divorce history for the 1669 respondents who had ever been married.

<table>
<thead>
<tr>
<th>Ever Divorced?</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smoker?</td>
<td>228</td>
<td>737</td>
<td>965</td>
</tr>
<tr>
<td>No</td>
<td>514</td>
<td>514</td>
<td>1028</td>
</tr>
<tr>
<td>Total</td>
<td>742</td>
<td>1251</td>
<td>1993</td>
</tr>
</tbody>
</table>

Among nonsmokers, only 32% have been divorced, 68% have not. Use counts rather than percents for the bar graph, to see numbers. Among smokers, 49% have been divorced, 51% have not.

Example 6.2 Tattoos and Ear Pierces

Responses from \( n = 565 \) men to two questions:
1. Do you have a tattoo?
2. How many total ear pieces do you have?

<table>
<thead>
<tr>
<th>Ear Pieces</th>
<th>No Tattoo</th>
<th>Tattoo</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>261</td>
<td>42</td>
<td>263</td>
</tr>
<tr>
<td>1 or more</td>
<td>54</td>
<td>16</td>
<td>70</td>
</tr>
<tr>
<td>Total</td>
<td>315</td>
<td>58</td>
<td>373</td>
</tr>
</tbody>
</table>

- No clear explanatory and response variable, so could do the table and bar graph in either direction. Note table is one version, graph is the other.
- Notice that the bars represent the percent with different numbers of pierces within each tattoo category, so the heights sum to 100% within each of the two sets of bars.

6.2 Risk, Relative Risk, Odds Ratio, and Increased Risk

Relative Risk = \( \frac{\text{Risk in category 1}}{\text{Risk in category 2}} \)

- Risk in denominator often the baseline risk.

Example:
- For those who drive under the influence of alcohol, the relative risk of an accident is 15.
- The risk of an accident for those who drive under the influence is 15 times the risk for those who don’t drive under the influence.
- In this example, numerator is risk under the influence, and denominator is risk when sober.

Baseline Risk and Relative Risk

Baseline Risk: risk without treatment, behavior, trait, etc, of interest. (Placebo instead of aspirin, don’t smoke, drive sober, don’t have gene for disease, etc.)

- Can be difficult to find.
- In many medical studies with placebo included, baseline risk = risk for placebo group.

Interpreting relative risk:
- Relative risk of 3: Risk of developing disease for one group is 3 times what it is for another group.
- Relative risk of 1: Risk is same for both categories of the explanatory variable (or both groups).
Example from *New York Times*

**January 13, 2009**

- “Drivers talking on cell phones are four times as likely to have an accident as drivers who are not.”
- In statistical terms, the “4” is called the **relative risk**.
- It’s the **risk** of having an accident on cell phone, compared to the **baseline risk** of an accident, under ordinary (no cell phone) conditions.

How did they find the relative risk of 4?

- Based on driving simulators and accident data combination, so don’t have actual data
- So, here is hypothetical data based on 10,000 trips, that would give relative risk of 4:

<table>
<thead>
<tr>
<th>Cell Phone?</th>
<th>Accident</th>
<th>No Accident</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>16</td>
<td>984</td>
<td>1000</td>
</tr>
<tr>
<td>No</td>
<td>36</td>
<td>8964</td>
<td>9000</td>
</tr>
<tr>
<td>Total</td>
<td>52</td>
<td>9948</td>
<td>10,000</td>
</tr>
</tbody>
</table>

Computations for relative risk:

- Risk of accident using cell phone = 16/1000 = .016
- Baseline risk (not using cell phone) = 36/9000 = 4/1000 = .004
- Relative risk = .016/.004 = 4
- Drivers on cell phone are 4 times as likely to have an accident.

Example: Cell phones and accidents

Recall risk is 16/1000 compared to 4/1000

Relative risk of accident on cell phone is 4.

Percent increase in risk of accident on cell phone

\[
(4 - 1) \times 100\% = 300\%
\]

or

\[
\frac{\text{Difference in risks}}{\text{Baseline risk}} \times 100\% = \frac{(16 - 4)}{4} \times 100\% = 300\%
\]

Drivers talking on cell phones have a 300% increase in the risk of an accident. Same as saying they are 4 times as likely to have an accident.

**Example 6.1: Smoking and Divorce Risk**

<table>
<thead>
<tr>
<th>Table 6.1 Smoking and Divorce, GSS Surveys 1991-1993</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ever Divorced?</strong></td>
</tr>
<tr>
<td>Smoker?  Yes No Total</td>
</tr>
<tr>
<td>Yes      233  247  480</td>
</tr>
<tr>
<td>No       224  810 1034</td>
</tr>
<tr>
<td>Total    457 1067 1521</td>
</tr>
</tbody>
</table>

For smokers:
Risk of divorce = 238/485 = 0.491 or 49.1%.

For nonsmokers:
Risk of divorce = 374/1184 = 0.316 or 31.6%.

Relative Risk of divorce = \( \frac{0.491}{0.316} = 1.53 \)

In this sample, the risk of divorce for smokers is 1.53 times the risk of divorce for nonsmokers.
Smoking and Divorce Risk—“Increased risk” is more meaningful with moderate rel. risk:

Relative Risk of divorce for smokers = 1.53

Percent increase in risk of divorce for smokers = (1.53 – 1) x 100% = 53%

Difference in risks = (49 – 32) / 32 x 100% = 53%

The risk of divorce is 53% higher for smokers than it is for nonsmokers.

Odds = Number in category 1 to Number in category 2
     = (Number in category 1/Number in category 2) to 1

Odds Ratio = (Odds for group 1) / (Odds for group 2)

Example:

Odds of getting a divorce to not getting a divorce for smokers are 238 to 247 or 0.96 to 1.
Odds of getting a divorce to not getting a divorce for nonsmokers are 374 to 810 or 0.46 to 1.
Odds Ratio = 0.96 / 0.46 = 2.1 => the odds of divorce for smokers are about double the odds for nonsmokers.

NOTE:

• Relative risk and Odds ratio will be similar if the values of A1 and B1 are small compared to the size of the sample. In other words, if the risk of the outcome of interest is small.
• Most studies in medical journals report the odds ratio (not the relative risk), for reasons to be explained later.

Summary table on page 201 shows formulas

<table>
<thead>
<tr>
<th>Explanatory variable</th>
<th>Response Variable</th>
<th>Category 1</th>
<th>Category 2</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Category of interest</td>
<td>A1</td>
<td>A2</td>
<td>TA</td>
<td></td>
</tr>
<tr>
<td>Baseline Category</td>
<td>B1</td>
<td>B2</td>
<td>TB</td>
<td></td>
</tr>
</tbody>
</table>

Relative risk = \( \frac{A_1}{B_1} \), Odds ratio = \( \frac{A_1}{B_1} \)

Example from Discussion 1

Commute? Parking Ticket?

<table>
<thead>
<tr>
<th>Explanatory variable</th>
<th>Ticket?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commute</td>
<td>Yes</td>
</tr>
<tr>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>Don’t commute</td>
<td>7</td>
</tr>
</tbody>
</table>

Relative risk = \( \frac{19}{25} \) = 2.95, Odds ratio = \( \frac{19}{19/18} = 4.75 \) / 18 = 0.389 = 12.2

Risk of developing breast cancer is 1.33 times greater for women who had their first child at 25 or older.
Increased Risk

**Increased Risk**

\[
\text{Increased Risk} = \left( \frac{\text{change in risk}}{\text{baseline risk}} \right) \times 100\% = \left( \frac{\text{relative risk} - 1.0}{} \right) \times 100\%
\]

**Example: Increased Risk of Breast Cancer**

- Change in risk = (0.0190 – 0.0143) = 0.0047
- Baseline risk = 0.0143
- Increased risk = (0.0047/0.0143) = 0.329 or 32.9%

There is a 33% increase in the chances of breast cancer for women who have not had a child before the age of 25.

Odds Ratio

**Odds Ratio:** ratio of the odds of getting the disease to the odds of not getting the disease.

**Example: Odds Ratio for Breast Cancer**

- Odds for women having first child at age 25 or older = 31/1597 = 0.0194
- Odds for women having first child before age 25 = 65/4475 = 0.0145
- Odds ratio = 0.0194/0.0145 = 1.34

Alternative formula: odds ratio = \(\frac{31 \times 4475}{1597 \times 65}\) = 1.34

Note that in this case, relative risk and odds ratio are similar.

Relative Risk and Odds Ratios in News and Journal Articles

Researchers often report relative risks and odds ratios **adjusted** to account for confounding variables.

**Example:**

Suppose an article reports that the relative risk for getting cancer for those with high-fat and low-fat diet is 1.3, adjusted for age and smoking status. =>

Relative risk applies (approx.) for two groups of individuals of same age and smoking status, where one group has high-fat diet and other has low-fat diet.

6.3 Misleading Statistics About Risk

**Questions to Ask:**

- What are the actual risks? What is the baseline risk?
- What is the population for which the reported risk or relative risk applies? Does it apply to you?
- What is the time period for this risk?

Missing Baseline Risk

**“Evidence of new cancer-beer connection”**

Sacramento Bee, March 8, 1984, p. A1

- Reported men who drank 500 ounces or more of beer a month (about 16 ounces a day) were **three times more likely** to develop cancer of the rectum than nondrinkers.
- Less concerned if chances go from 1 in 100,000 to 3 in 100,000 compared to 1 in 10 to 3 in 10.
- Need baseline risk (which was about 1 in 180) to help make a lifestyle decision. Often that is not known.

Reported Risk versus Your Risk

**“Older cars stolen more often than new ones”**

Davis (CA) Enterprise, 15 April 1994, p. C3

Reported among the 20 most popular auto models stolen in California the previous year, 17 were at least 10 years old.

Many factors determine which cars stolen:

- Type of neighborhood.
- Locked garages.
- Cars not locked nor have alarms.

“If I were to buy a new car, would my chances of having it stolen increase or decrease over those of the car I own now?” Article gives no information about that question.
Risk over What Time Period?

“Italian scientists report that a diet rich in animal protein and fat—cheeseburgers, french fries, and ice cream, for example—increases a woman’s risk of breast cancer threefold.”


If 1 in 9 women get breast cancer, does it mean if a women eats above diet, chances of breast cancer are 1 in 3?

Two problems:
- Don’t know how study was conducted.
- Age is critical factor. The 1 in 9 is a lifetime risk, at least to age 85. Risk increases with age.
- If study on young women, threefold increase is small.

Simpson’s Paradox: The Missing Third Variable

- Relationship appears to be in one direction if third variable is not considered and in other direction if it is.
- Can be dangerous to summarize information over groups.

Example: Simpson’s Paradox for Hospital Patients

Survival Rates for Standard and New Treatments

<table>
<thead>
<tr>
<th></th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survive</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Die</td>
<td>200</td>
<td>200</td>
</tr>
<tr>
<td>Total</td>
<td>700</td>
<td>700</td>
</tr>
</tbody>
</table>

Risk Compared for Standard and New Treatments

<table>
<thead>
<tr>
<th></th>
<th>Hospital A</th>
<th>Hospital B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk of dying with the standard treatment</td>
<td>0.50</td>
<td>0.30</td>
</tr>
<tr>
<td>Risk of dying with the new treatment</td>
<td>0.50</td>
<td>0.30</td>
</tr>
<tr>
<td>Reduction in risk</td>
<td>0.20</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Looks like new treatment is a success at both hospitals, especially at Hospital B.

Example 6.8 Blood Pressure and Oral Contraceptive Use

Hypothetical data on 2400 women. Recorded oral contraceptive use and if had high blood pressure.

Percent with high blood pressure is about the same among oral contraceptive users and nonusers.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>Number with High B.P.</th>
<th>% with High B.P.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Oral Contraceptives</td>
<td>1000</td>
<td>60</td>
<td>60 of 1000 = 6.0%</td>
</tr>
<tr>
<td>Don’t Use Oral Contraceptives</td>
<td>1400</td>
<td>100</td>
<td>100 of 1400 = 7.1%</td>
</tr>
</tbody>
</table>

Many factors affect blood pressure. If users and nonusers differ with respect to such a factor, the factor confounds the results. Blood pressure increases with age and users tend to be younger.

<table>
<thead>
<tr>
<th></th>
<th>Sample Size</th>
<th>% with High B.P.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Oral Contraceptives</td>
<td>400</td>
<td>36 (9.0%)</td>
</tr>
<tr>
<td>Don’t Use Oral Contraceptives</td>
<td>400</td>
<td>16 (4.0%)</td>
</tr>
</tbody>
</table>

In each age group, the percentage with high blood pressure is higher for users than for nonusers. => Simpson’s Paradox.
Simpson’s Paradox: Summary

- Risk of a problem is higher for Group 1 than for Group 2 in both populations.
  - Ex: Risk of high blood pressure is higher for oral contraceptive users than for non-users for both younger and older women.
- But, when populations are combined, risk of a problem is higher for Group 2 than for Group 1.
- Lesson: It can be dangerous to summarize information over groups.