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.)

	Heart Attack	No Heart Attack	Total
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
Total	293	21,778	22,071

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 = 104/11,037 = 0.0094 or 0.94%

Placebo Group:
Percentage who had heart attacks = 189/11,034 = 0.0171 or 1.71%

Conditional Percentages (Columns)

Not usually of interest

	Heart Attack	No Heart Attack	Total
Aspirin	104	10,933	11,037
Placebo	189	10,845	11,034
Total	293	21,778	22,071

Example: Find the *Column* Percentages

Heart Attack Group:
Percentage who took aspirin = 104/293 = .355 or 35.5%

No Heart Attack Group:
Percentage who took aspirin = 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.

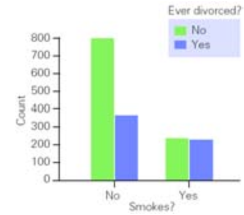
Example 6.1 *Smoking and Divorce*

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

TABLE 6.1 Smoking and Divorce, GSS Surveys 1991–1993

	Ever Divorced?		
Smoke?	*Yes*	*No*	*Total*
Yes	238	247	485
No	374	810	1184
Total	612	1057	1669

Data Source: SDA archive at UC Berkeley web site (www.csa.berkeley.edu:7502/).



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 pierces do you have?

TABLE 6.2 Ear Pierces and Tattoos for Men ($n = 565$)

Ear Pierces	No Tattoo	Tattoo	Total
0	381	43	424
1	54	16	70
2 or more	45	26	71
Total	480	85	565

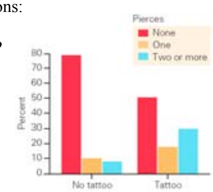


FIGURE 6.2 Column percents for the ear pierce and tattoo data

- 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

$$\textbf{Risk} = \frac{\text{Number in category}}{\text{Total number in group}}$$

Example:

Suppose in a group of 200 individuals, asthma affects 24 people. In this group the *risk* of asthma is 24/200 = 0.12 or 12%.

$$\textbf{Relative Risk} = \frac{\text{Risk in category 1}}{\text{Risk in category 2}}$$

- 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:

Cell Phone?	Accident	No Accident	Total
Yes	16	984	1000
No	36	8964	9000
Total	52	9948	10,000

Computations for relative risk:

Cell Phone?	*Accident*	No Accident	*Total*
Yes	**16**	984	**1000**
No	**36**	8964	**9000**
Total	52	9948	10,000

- *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

Percent increase in risk

$$= \frac{\text{Difference in risks}}{\text{Baseline risk}} \times 100\%$$

$$= (\text{Relative risk} - 1) \times 100\%$$

Note:

When risk is *smaller* than baseline risk, relative risk < 1 and the percent "increase" will actually be negative, so we say *percent decrease* in risk.

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

$= (4 - 1) \times 100\% = \boxed{300\%}$

or $\dfrac{\text{Difference in risks}}{\text{Baseline risk}} \times 100\% = \dfrac{(16-4)}{4} \times 100\%$

$= \boxed{300\%}$

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 6.1 Smoking and Divorce, GSS Surveys 1991–1993

Smoke?	Ever Divorced? Yes	No	Total
Yes	238	247	485
No	374	810	1184
Total	612	1057	1669

Data Source: SDA archive at UC Berkeley web site (www.csa.berkeley.edu:7502/).

- 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 = $\dfrac{49\%}{32\%} = 1.53$

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%

$$= \frac{\text{Difference in risks}}{\text{Baseline risk}} \text{ x } 100\% = \frac{(49-32)}{32} \text{ x } 100\% = 53\%$$

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.

Summary table on page 201shows formulas

	Response Variable		
Explanatory variable	Category 1	Category 2	Total
Category of interest	A_1	A_2	TA
Baseline Category	B_1	B_2	TB

$$\text{Relative risk} = \frac{A_1/TA}{B_1/TB}, \quad \text{Odds ratio} = \frac{A_1/A_2}{B_1/B_2}$$

$$\text{Relative risk} = \frac{A_1/TA}{B_1/TB}, \quad \text{Odds ratio} = \frac{A_1/A_2}{B_1/B_2}$$

NOTE:

- Relative risk and Odds ratio will be similar *if* the values of A_1 and B_1 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.*

Example from Discussion 1 Commute? Parking Ticket?

	Ticket?		
Explanatory variable	Yes	No	Total
Commute	19	4	23
Don't commute	7	18	25

$$\text{Relative risk} = \frac{19/23}{7/25} = 2.95, \quad \text{Odds ratio} = \frac{19/4}{7/18} = \frac{4.75}{0.389} = 12.2$$

New Example, compute all of these summaries: Based on observational study

First Child at Age 25 or Older?	Breast Cancer	No Breast Cancer	Total
Yes	31	1597	1628
No	65	4475	4540
Total	96	6072	6168

- Risk for women having first child at 25 or older = 31/1628 = 0.0190
- Risk for women having first child before 25 (baseline) = 65/4540 = 0.0143
- Relative risk = 0.0190/0.0143 = $\boxed{1.33}$

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

Source: Pagano and Gauvreau (1988, p. 133).

Increased Risk

Increased Risk = (change in risk/baseline risk)×100%
= (relative risk – 1.0)×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 = $\boxed{1.34}$

Alternative formula: $\text{odds ratio} = \dfrac{31 \times 4475}{1597 \times 65} = 1.34$

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,"
Prevention Magazine's Giant Book of Health Facts (1991, p. 122)

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

	Hospital A			Hospital B		
	Survive	Die	Total	Survive	Die	Total
Standard	5	95	100	500	500	1000
New	100	900	1000	95	5	100
Total	105	995	1100	595	505	1100

Risk Compared for Standard and New Treatments

	Hospital A	Hospital B
Risk of dying with the standard treatment	95/100 = 0.95	500/1000 = 0.50
Risk of dying with the new treatment	900/1000 = 0.90	5/100 = 0.05
Relative risk	0.95/0.90 = 1.06	0.50/0.05 = 10.0

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

Example: Simpson's Paradox for Hospital Patients

Estimating the Overall Reduction in Risk

	Survive	Die	Total	Risk of Death
Standard	505	595	1100	595/1100 = 0.54
New	195	905	1100	905/1100 = 0.82
Total	700	1500	2200	

What has gone wrong? With combined data it looks like the ***standard treatment is superior!*** Death rate for standard treatment is only 66% of what it is for the new treatment.
HOW?
More serious cases were treated at Hospital A (famous research hospital); more serious cases were also more likely to die, no matter what. ***And*** a higher proportion of patients at Hospital A received the new treatment.

Example 6.8 *Blood Pressure and Oral Contraceptive Use*

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

TABLE 6.5 Percent with High Blood Pressure for Users and Nonusers of Oral Contraceptives

	Sample Size	Number with High B.P.	% with High B.P.
Use Oral Contraceptives	800	64	64 of 800 = 8.0%
Don't Use Oral Contraceptives	1600	136	136 of 1600 = 8.5%

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

Example 6.8 *Blood Pressure and Oral Contraceptive Use (cont)*

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 6.6 Controlling for the Effect of Age

	Age 18–34		Age 35–49	
	Sample Size	*n and % with High B.P.*	*Sample Size*	*n and % with High B.P.*
Use Oral Contraceptives	600	36 (6%)	200	28 (14%)
Don't Use Oral Contraceptives	400	16 (4%)	1200	120 (10%)

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.