ANNOUNCEMENTS

- Quiz #2 begins at 4pm today and ends at 3pm Wed, Jan 23rd
- Clicker grades for Week 1 have been updated.

TODAY

Sections 4.1 to 4.3. Read whatever we don't have time to finish in those sections (probably section on "Misleading Risk," slides 27 to 30).

HOMEWORK (Due Wed, Jan 23)

Chapter 4: #14, 18, 36, and read pages 120-122 on *Misleading Statistics*

Chapter 4

Relationships Between Categorical Variables

4.1 Displaying Relationships Between Categorical Variables: Contingency Tables

- You did this in Discussion #1. (Great for some, not very well for other teams!)
- 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 (Case Study 1.6): Aspirin and Heart Attacks

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 = 104/11037 = 0.0094 or 0.94%

Placebo Group:
Percentage who had heart attacks = 189/11034 = 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 = 10933/21778 = .502 or 50.2%

4.2 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 while driving under the influence of alcohol is *15 times* the risk when not driving 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 the other 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 **four** 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 four?

- Based on driving simulators and accident data combination, so I 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 numerator risk is *smaller* than baseline (or denominator) 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 $\frac{\text{Difference in risks}}{\text{Baseline risk}} \times 100\% = \frac{(16-4)}{4} \times 100\% = \boxed{300\%}$

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.

Exercise 4.2: *Smoking and Divorce Risk*

	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/).

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

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

$= (1.53 - 1) \times 100\% = 53\%$

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

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 120 shows formulas

	Response Variable		
Explanatory variable	Category 1	Category 2	Total
Category of interest	A_1	A_2	T_A
Baseline Category	B_1	B_2	T_B

Relative risk = $\frac{A_1/T_A}{B_1/T_B}$, Odds ratio = $\frac{A_1/A_2}{B_1/B_2}$

Alternate formula for odds ratio

$$\boxed{\text{Odds ratio} = \frac{A_1 / A_2}{B_1 / B_2} = \frac{A_1 B_2}{A_2 B_1}}$$

$$\boxed{\text{Relative risk} = \frac{A_1 / T_A}{B_1 / T_B}, \qquad \text{Odds ratio} = \frac{A_1 / A_2}{B_1 / B_2}}$$

- Relative risk and Odds ratio will be similar *if* A_2 and B_2 are close to the total size of the samples (T_A and T_B). 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
Upper class? Drinker?

	Drinker?		
Explanatory variable	**Yes**	**No**	**Total**
Upper classman	20	4	24
Lower classman	7	5	12

$$\text{Relative risk} = \frac{20/24}{7/12} = \frac{.833}{.583} = 1.43, \qquad \text{Odds ratio} = \frac{20/4}{7/5} = \frac{5}{1.4} = 3.6$$

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.

Cause and Effect

- Remember, we cannot conclude that having the first child at a later age *causes* an increased risk of breast cancer. There are lots of potential confounding variables.
- Possible examples:
  - Taking birth control pills for an extended period of time.
  - Different patterns of alcohol use.

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 versus 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.

Misleading Statistics About Risk

Read next 4 slides and this section in book (pgs 120-122) on your own.

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.

4.3 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 from UC Berkeley (Data in Exercise 4.37)

Simpson's Paradox for Graduate School Admissions: Men versus Women

(Actual data not released for privacy reasons, but similar to the data shown.)

	Admit	Deny	Total	Admitted
Men	450	550	1000	450/1000 or 45%
Women	175	325	500	175/500 or 35%

Higher percent of men admitted overall. There were two Ph.D. programs involved. Which one had more serious bias in admitting a higher percentage of men? Break down data by program.

Simpson's Paradox for Grad School Admissions: Men vs Women, 2 programs

	Program A		Program B	
	Admit	**Deny**	**Admit**	**Deny**
Men	400	250	50	300
Women	50	25	125	300

Program A admitted:
400/650 = 61.5% of men
50/75 = 66.7% of women
Women fared better.

Program B admitted:
50/350 = 14.3% of men
125/425 = 29.4% of women
Women fared better.

Simpson's Paradox: Grad School Admissions

What has gone wrong?
With *combined* data it looks like women have *lower* admission rates. Yet *each* program admitted a *higher* proportion of women than men!

Explanation?
More men applied to Program A than to Program B.
More women applied to Program B than to Program A.
Program B was much harder to get into overall:
- A admitted 450/725 or 62% of applicants.
- B admitted 175/775 or 23% of applicants.
- So, lower proportion of women admitted overall.

Example 4.11 (if time, otherwise read on your own)
Blood Pressure and Oral Contraceptive Use

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

	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 slightly *higher* among *nonusers* of oral contraceptive than among users.

Blood Pressure and Oral Contraceptive Use

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.

	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.

HOMEWORK (Due Wed, Jan 23)

4.14

4.18

4.36