



## Chapter 12

---

# Relationships Between Categorical Variables

---

# Thought Question 1:

Students in a statistics class were asked whether they preferred an **in-class or a take-home final exam** and were then categorized as to **whether they had received an A** on the midterm.

Of the 25 A students, 10 preferred a take-home exam, whereas of the 50 non-A students, 30 preferred a take-home exam.

How would you display these data in a table?

# Thought Question 2:

Suppose a news article claimed that **drinking coffee doubled your risk of developing a certain disease**. Assume the statistic was based on legitimate, well-conducted research.

What **additional information** would you want about the risk before deciding whether to quit drinking coffee?

*(Hint: Does this statistic provide any information on your actual risk?)*



## Thought Question 3:

A study classified pregnant women according to whether they smoked and whether they were able to get pregnant during the first cycle in which they tried to do so. What do you think is the **question of interest**? Attempt to answer it. Here are the results:

	Pregnancy Occurred After		
	First Cycle	Two or More Cycles	Total
Smoker	29	71	100
Nonsmoker	198	288	486
Total	227	359	586

## Thought Question 4:

A recent study estimated that the “**relative risk**” of a woman developing lung cancer if she smoked was 27.9.

What do you think is meant by the term *relative risk*?

# 12.1 Displaying Relationships Between Categorical Variables: Contingency Tables



- Count the number of individuals who fall into each combination of categories.
- Present counts in table = **contingency table**.
- Each row and column combination = *cell*.
- Row = *explanatory* variable.
- Column = *response* variable.

# Example 1: Aspirin and Heart Attacks



## Case Study 1.2:

**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.

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

# Conditional Percentages and Rates



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

## **Example 1: 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 and Rates



**Rate:** the number of individuals per 1000  
or per 10,000 or per 100,000.

**Percentage:** rate per 100

## Example 1: Percentage and Rate Added

	<b>Heart Attack</b>	<b>No Heart Attack</b>	<b>Total</b>	<b>Heart Attacks (%)</b>	<b>Rate per 1000</b>
Aspirin	104	10,933	11,037	0.94	9.4
Placebo	189	10,845	11,034	1.71	17.1
Total	293	21,778	22,071		

## Example 2: Young Drivers, Gender, and Driving Under the Influence of Alcohol



**Case Study 6.5:** Court case challenging law that differentiated the ages at which young men and women could buy 3.2% beer.

### Results of Roadside Survey for Young Drivers

	Drank Alcohol in Last 2 Hours?		Total	Percentage Who Drank
	Yes	No		
Males	77	404	481	16.0%
Females	16	122	138	11.6%
Total	93	526	619	15.0%

Percentage slightly higher for males, but difference in percentages is not *statistically significant*.

*Source:* Gastwirth, 1988, p. 526.

## Example 3: Ease of Pregnancy for Smokers and Nonsmokers

### Retrospective Observational Study:

Variable A = explanatory variable = smoker or nonsmoker

Variable B = response variable = pregnant in first cycle or not

### Time to Pregnancy for Smokers and Nonsmokers

	Pregnancy Occurred After			Percentage in First Cycle
	First Cycle	Two or More Cycles	Total	
Smoker	29	71	100	29%
Nonsmoker	198	288	486	41%
Total	227	359	586	

Much higher percentage of nonsmokers than smokers were able to get pregnant during first cycle, but we *cannot conclude* that smoking *caused* a delay in getting pregnant.

# 12.2 Relative Risk, Increased Risk, and Odds



A population contains 1000 individuals,  
of which 400 carry the gene for a disease.

Equivalent ways to express this proportion:

- Forty *percent* (40%) of all individuals carry the gene.
- The *proportion* who carry the gene is 0.40.
- The *probability* that someone carries the gene is .40.
- The *risk* of carrying the gene is 0.40.
- The *odds* of carrying the gene are 4 to 6 (or 2 to 3, or 2/3 to 1).

# Risk, Probability, and Odds



**Percentage with trait =**

$$(\text{number with trait}/\text{total}) \times 100\%$$

**Proportion with trait =** number with trait/total

**Probability of having trait =** number with trait/total

**Risk of having trait =** number with trait/total

**Odds of having trait =**

$$(\text{number with trait}/\text{number without trait}) \text{ to } 1$$

# Baseline Risk and Relative Risk



**Baseline Risk:** risk without treatment or behavior

- Can be difficult to find.
- If placebo included,  
baseline risk = risk for placebo group.

**Relative Risk:** of outcome for two categories of explanatory variable is ratio of risks for each category.

- *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 4: Relative Risk of Developing Breast Cancer

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  
 $= 65/4540 = 0.0143$
- Relative risk  $= 0.0190/0.0143 = 1.33$

*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

$$\begin{aligned}\text{Increased Risk} &= (\text{change in risk}/\text{baseline risk}) \times 100\% \\ &= (\text{relative risk} - 1.0) \times 100\%\end{aligned}$$

## Example 5: 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 =  $\frac{31 \times 4475}{1597 \times 65} = 1.34$

# Relative Risk and Odds Ratios in Journal Articles



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

## **Example:**

Suppose 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.

# **12.3 Misleading Statistics about Risk**



**Common ways the media misrepresent statistics about risk:**

- 1. The baseline risk is missing.**
- 2. The time period of the risk is not identified.**
- 3. The reported risk is not necessarily your 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.

# 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 woman 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.

# 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] last 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.

# 12.4 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 7: 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 7: 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.

# Case Study 12.1: Assessing Discrimination in Hiring and Firing

## Layoffs by Ethnic Group for Labor Department Employees

Ethnic Group	Laid Off?		Total	% Laid Off
	Yes	No		
African American	130	1382	1512	8.6
White	87	2813	2900	3.0
Total	217	4195	4412	

- Selection ratio of those laid off =  $3.0/8.6 = 0.35$
- Selection ratio of those retained =  $91.4/97 = 0.94$
- Discrepancy handled using Odds Ratio =  $\frac{130 \times 2813}{1382 \times 87} = 3.04$

*Odds of being laid off compared with being retained are **three times higher** for African Americans than for whites.*

Source: Gastwirth and Greenhouse, 1995.

# For Those Who Like Formulas



To represent the *observed numbers* in a  $2 \times 2$  contingency table, we use the notation:

Variable 1	Variable 2		Total
	Yes	No	
Yes	$a$	$b$	$a + b$
No	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$n$

## Relative Risk and Odds Ratio

Using the notation for the observed numbers, if variable 1 is the explanatory variable and variable 2 is the response variable, then we can compute

$$\text{relative risk} = \frac{a(c + d)}{c(a + b)}$$

$$\text{odds ratio} = \frac{ad}{bc}$$