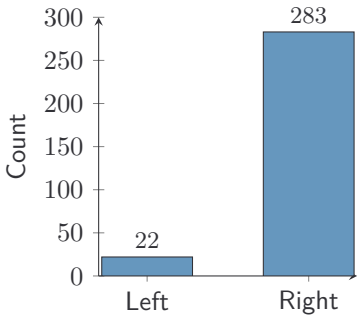
Chapter 6

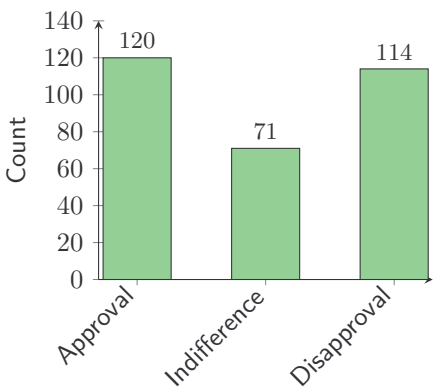
Two-Way Tables & Risk

Visualizing Categorical Data: Bar Charts

Handedness Distribution



Pineapple Preference Distribution



From Quantitative to Categorical Associations

Previously: We studied how two quantitative variables can be associated (e.g., scatterplots, correlation).

- Example: Height vs. weight, hours studied vs. exam score
- Tools: scatterplots, regression, correlation coefficient

Now: We try to answer similar questions when both variables are categorical.

- Are two categorical variables associated?
- How do we summarize and visualize their relationship?
- What tools can we use to compare groups?

Categorical Data

The following are examples of questions we might answer:

- Does a new vaccine actually **reduce the risk** of getting sick?
- Is a hiring process **fair across demographics**?
- Do left-handed people really prefer different foods than right-handed people?
- When a hospital reports better outcomes, is it truly better- or does it just treat **easier cases**?

🔑 **Key Point:** To answer these questions, we need tools for analyzing two categorical variables **simultaneously**.

The Framework: Exposure and Outcome

In this chapter, we generally view our data as having two distinct roles. For every person (or observation) in our dataset, we track:

1. Exposure (X)

- Also called the **Explanatory Variable**.
- This defines the **groups** we are comparing. It does **not** need to be the row variable of the contingency table, but it frequently is.
- *Examples: Vaccination status, Handedness, Study Method.*

2. Outcome (Y)

- Also called the **Response Variable**.
- This is the **result** we are measuring.
- *Examples: Getting the flu, Pineapple preference, Passing the exam.*

🔑 **Key Point:** Our goal is to determine if the likelihood of the **Outcome** depends on the **Exposure**.

The Core Question, Restated

Throughout this chapter, we will ask variations of this question:

Does the proportion with a given outcome differ between exposure groups?

If YES (proportions differ):
The variables are **associated**.
Exposure tells us something about outcome.
We want to quantify the difference.

If NO (proportions are the same):
The variables are **independent**.
Exposure tells us nothing about outcome.

Intended Learning Outcomes

- Construct and interpret two-way tables
- Calculate marginal distributions
- Calculate conditional distributions
- Distinguish between marginal and conditional distributions
- Calculate relative risk and odds ratio
- Interpret RR and OR in context
- Recognize Simpson's paradox
- Identify confounding variables

PART 1

Two-Way Tables & Distributions

Analyzing Relationships Between Categorical Variables

- So far, we have analyzed **one categorical variable** at a time using bar graphs and pie charts.
- Many real-world questions involve **two categorical variables**:

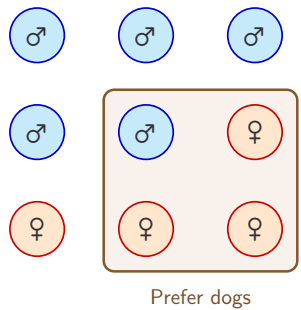
1. Does **treatment type** affect **recovery status**?
2. Is **political affiliation** related to **opinion on an issue**?
3. Does **vaccination status** relate to **disease outcome**?

🔑 **Key Point:** To study two categorical variables together, we use **two-way tables**.

Visualizing the Data

Example: Data collected on 9 individuals, tracking

- **Variables:** Gender (♂, ♀) and Pet Preference (Cat vs Dog).

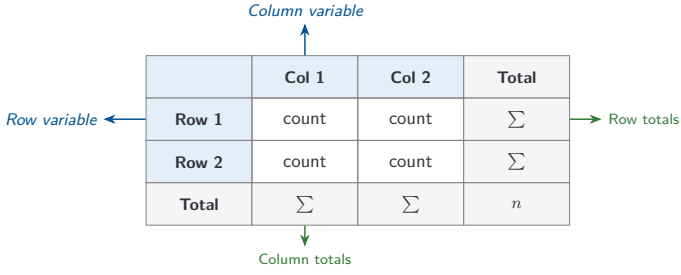


Two-Way Contingency Tables

Two-Way Table

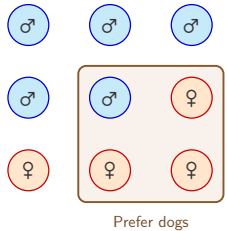
A **two-way table** (or contingency table) displays counts for each combination of values of two categorical variables.

- One variable defines the **rows**
- The other variable defines the **columns**
- Each cell contains a **count**



Example 6.0: Constructing a Table

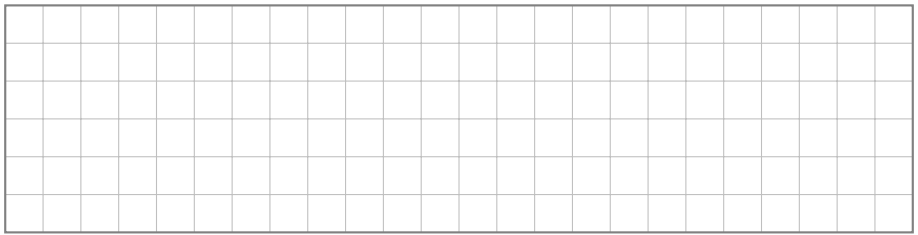
Context: Recall our toy example with 9 individuals, tracking Gender (♂, ♀) and Pet Preference (Cat vs Dog). Construct the two-way contingency table showing the relationship between Gender and Pet Preference.



Example 6.1: Pineapple on Pizza

Context: Last term, I asked DS1000A students about their handedness and pineapple-on-pizza preference. The results are summarized below:

Pineapple?	**Left-handed**	**Right-handed**	**Total**
Approval	11	109	120
Indifference	5	66	71
Disapproval	6	108	114
Total	22	283	305



Uncertainty: Why We Need Proportions

Key insight: When we select an individual from a population, we **don't know in advance** what their exposure or outcome will be.

- Pick a random student from this class: are they left-handed or right-handed?
- Pick a random patient from a hospital: did they receive Treatment A or B?
- Pick a random website visitor: did they click the "Buy" button?

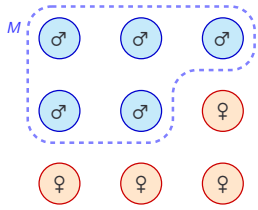
🔑 **Key Point:** Because we face **uncertainty**, we describe patterns using **proportions**: "What fraction of people in Group A have Outcome Y?"

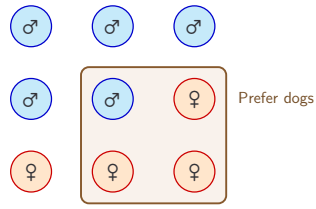
These proportions tell us **how likely** different outcomes are within different groups. Comparing them tells us whether exposure and outcome are **associated**.

Marginal Distributions

Marginal Distribution

The **marginal distribution** of a variable describes the distribution of that variable alone, .



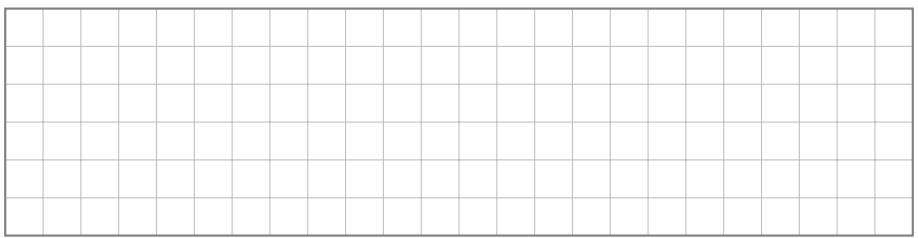


Example 6.1: Marginal Distribution of Handedness

Context: Using the DS1000A pineapple-on-pizza data:

Pineapple?	**Left-handed**	**Right-handed**	**Total**
Approval	11	109	120
Indifference	5	66	71
Disapproval	6	108	114
Total	22	283	305

Find the marginal distribution of handedness.

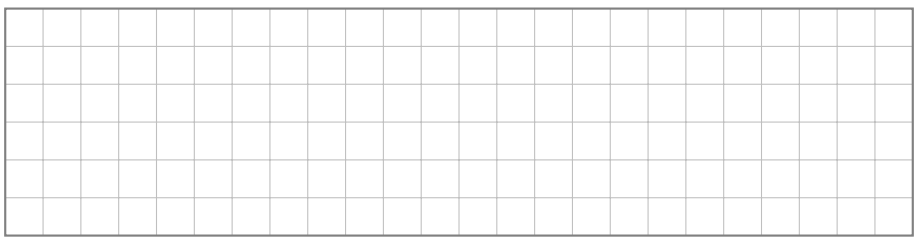


Example 6.1: Marginal Distribution of Pineapple Preference

Context: Using the DS1000A pineapple-on-pizza data:

Pineapple?	**Left-handed**	**Right-handed**	**Total**
Approval	11	109	120
Indifference	5	66	71
Disapproval	6	108	114
Total	22	283	305

Find the marginal distribution of pineapple preference.



Notation: Conditional Probability

Before we calculate conditional distributions, let's introduce a standard piece of notation that makes it easier to write down what we mean.

The "Given" Bar (|)

In probability and statistics, the vertical bar | is read as "**given**."

$$P(A \mid B)$$

This translates to: "The probability of outcome A **given** that we already know condition B is true."

- **Marginal:** $P(\text{Approval})$ means "What % of **everyone** approves?"
- **Conditional:** $P(\text{Approval} \mid \text{Left-Handed})$ means "If we look **only** at Left-Handed people, what % of **them** approve?"

Conditional Distributions

Conditional Distribution

The **conditional distribution** of a variable describes the distribution of that variable **within a specific group** defined by the other variable. There is a separate conditional distribution for each value of the conditioning variable.

Pineapple?	Left-handed	Right-handed	Total
Approval	11	109	120
Indifference	5	66	71
Disapproval	6	108	114
Total	22	283	305

Find the conditional distribution of pineapple preference *given* handedness.

Left handed people $P(\text{Outcome}|\text{Left})$:

- Approval:
- Indifference:
- Disapproval:

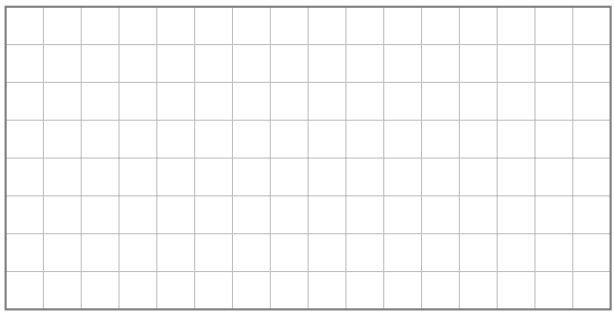
Right handed people $P(\text{Outcome}|\text{Right})$:

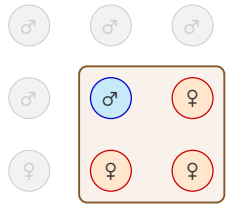
- Approval:
- Indifference:
- Disapproval:

Visualizing Conditional Probability

Conditional means we **zoom in** on a specific group (e.g., dog enthusiasts) and ignore everyone else.

Example: Find the conditional distribution of gender among dog enthusiasts.





The "faded" nodes no longer count.

Reading and Computing Conditional Proportions

Key Point: $P(A \mid B) =$ "probability of A given B" $= \dfrac{\#\text{ with both } A \text{ and } B}{\#\text{ with } B}$

The key: The denominator is **restricted** to only those with condition B.

Example Table:

	Flu	No Flu	Total
Vaccinated	26	3874	3900
Unvaccinated	70	3830	3900

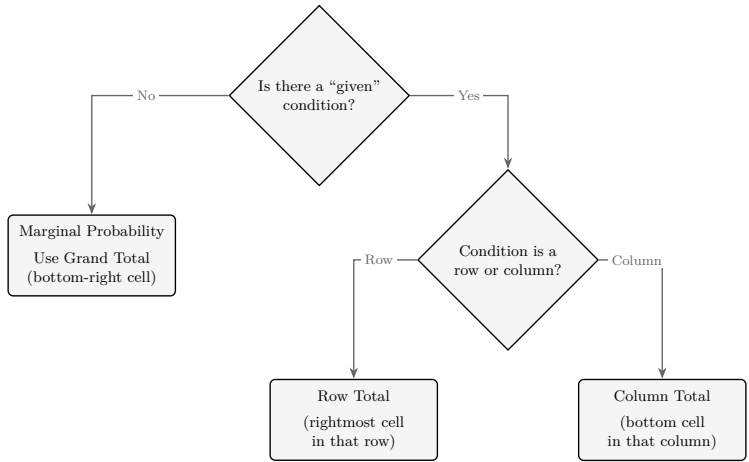
Computing:

$$P(\text{Flu} \mid \text{Vaccinated}) = \frac{26}{3900}$$

$$P(\text{Flu} \mid \text{Unvaccinated}) = \frac{70}{3900}$$

Caution: The denominator changes depending on what comes after the "|" symbol.

Which Denominator to Use?

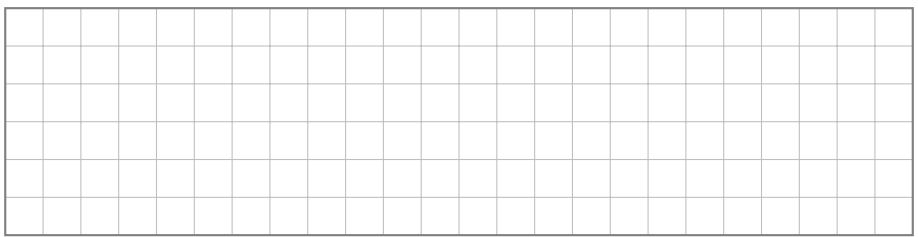


Example 6.1: Conditional Distribution of Handedness

Context: Using the DS1000A pineapple-on-pizza data:

Pineapple?	**Left-handed**	**Right-handed**	**Total**
Approval	11	109	120
Indifference	5	66	71
Disapproval	6	108	114
Total	22	283	305

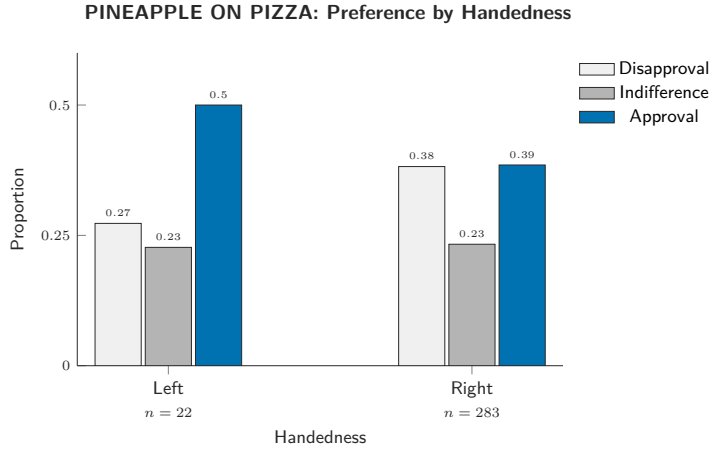
Find the conditional distribution of handedness **given** pineapple preference of **indifference**.



Visualizing Conditional Distributions: Grouped Bar Chart

Grouped Bar Chart

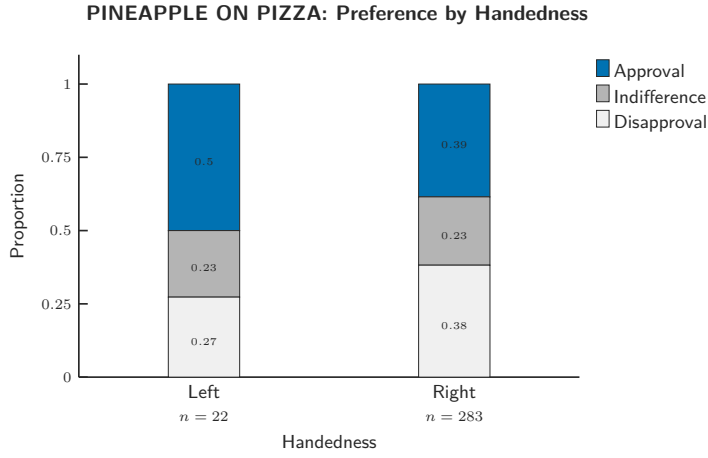
- Bars grouped by one variable
- Colors represent the other variable
- Easy to compare within groups



Visualizing Conditional Distributions: Stacked Bar Chart

Segmented (Stacked) Bar Chart

- Each bar represents 100%
- Segments show proportions
- Easy to compare proportions

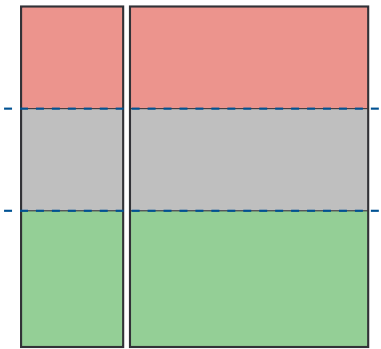


Visualizing Associations: Mosaic Plots

A **mosaic plot** shows cell counts as rectangular areas. Width = column proportion, Height = row proportion within column.

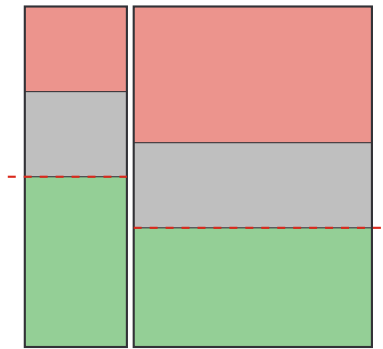
Independent Variables

(Equal row heights across columns)



Associated Variables

(Different row heights across columns)



Independence of Categorical Variables

Independence

Two categorical variables are **independent** if the conditional distribution of one variable is the _____ of the other variable.

Key Point: Checking for independence:

Compare conditional distributions across groups. If they are (approximately) equal, the variables may be independent.

What Would Independence Look Like?

Hypothetical: Suppose pineapple preference was truly independent of handedness. Then the conditional distributions would be (approximately) equal:

	Left-handed	Right-handed	Marginal
Approval	39.3%	39.3%	39.3%
Indifference	23.3%	23.3%	23.3%
Disapproval	37.4%	37.4%	37.4%

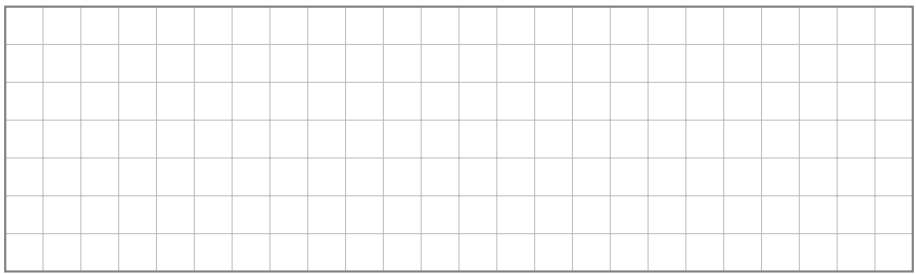
Key Point: Under independence, every conditional distribution equals (approximately) the **marginal distribution**.

Caution: Compare this with the **actual data**: left-handed approval was 50.0% vs. right-handed 38.5%. These are quite far apart, suggesting the variables are **not** independent.

Example 6.3: Instrument and Music Genre

Context: A survey of 120 students asked about their primary instrument and favorite music genre. Are instrument and favorite genre independent?

Instrument	Classical	Pop	Total
Piano	25	15	
Guitar	10	30	
Violin	20	20	
Total			

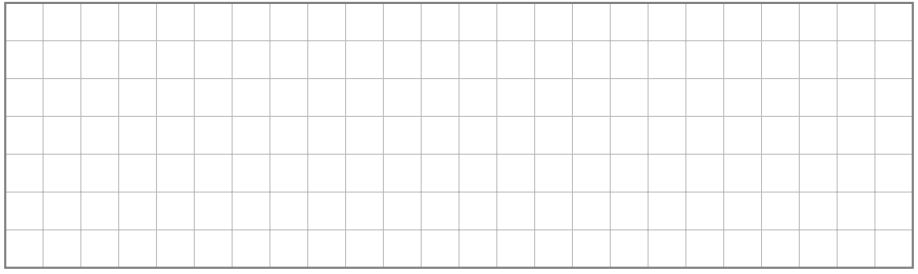


Example 6.4: Movie Genre and Snack Preference

Context: A movie theater surveyed 500 customers about their movie genre choice and snack preference.

Genre	Popcorn	Candy	Total
Action	84	56	140
Comedy	126	84	210
Drama	90	60	150
Total	300	200	500

Are genre and snack preference independent?



PART 2

Quantifying Associations

Beyond Independence

In Part 1, we asked a binary question: **Are these variables independent?**

- If $P(\text{Outcome}|\text{Group A}) \neq P(\text{Outcome}|\text{Group B})$, we say they are **associated**.
- But simply knowing an association *exists* is rarely enough for decision-making.

The Next Logical Questions:

1. **Direction:** Does the exposure increase or decrease the likelihood of the outcome?
2. **Strength:** How *large* is that effect? Is it negligible or substantial?

Quantifying the Association

To make decisions, we need to measure the **effect size** of the relationship.

Scenario A:
Treatment A: 1.0% recovery
Treatment B: 1.1% recovery

Associated? Yes.
Meaningful? Perhaps not.

Scenario B:
Treatment A: 1.0% recovery
Treatment B: 50.0% recovery

Associated? Yes.
Meaningful? Absolutely.

Risk Measures

Risk Measures (like Relative Risk and Odds Ratios) are statistics that quantify the **strength** of the association between two categorical variables.

Why “Risk”?

The terminology comes from **epidemiology** and **biostatistics**, where these tools were developed to study disease.

- In those fields, the “Success” outcome is often a negative event (death, infection, side effect).
- Therefore, the probability of the outcome is called the **Risk**.

Probability, Risk, and Odds

Key Terms

For a categorical variable with category c:

- **Percentage:** $\dfrac{\#\text{ in category } c}{\text{total } \#} \times 100\%$
- **Proportion/Probability/Risk:** $\dfrac{\#\text{ in category } c}{\text{total } \#}$
- **Odds:** $\dfrac{\#\text{ in category } c}{\#\text{ not in category } c} = \dfrac{\text{Probability}}{1 - \text{Probability}}$

Converting Between Probability and Odds

Key Point: You can convert back and forth between probability and odds:

Probability $\rightarrow$ Odds

$$\text{Odds} = \frac{p}{1-p}$$

Example: If $p = 0.75$,
Odds =

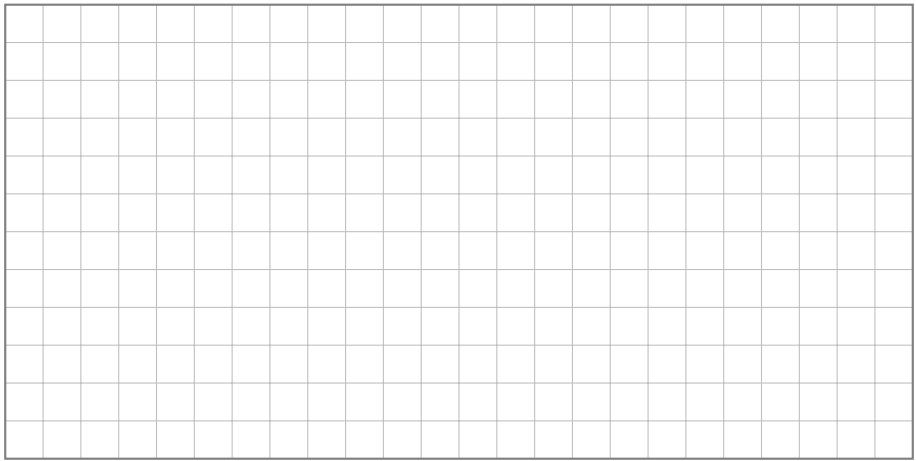
Odds $\rightarrow$ Probability

$$p = \frac{\text{Odds}}{1+\text{Odds}}$$

Example: If Odds $= 4$,
$p =$

Example 6.5: Rare Outcomes

Context: In a population of 10,000 individuals, 6 have heterochromia (two different colored eyes). Find the percentage, proportion, probability, risk, and odds of heterochromia.



Example 6.6: Conditional Probability

Context: A neighborhood survey on rollerblading and having children.

	Kids	**No Kids**	**Total**
Rollerblader	41	9	50
Not a rollerblader	3	9	12
Total	44	18	62

Suppose that David, Anastasia and Miles were survey participants and that:

- David is focusing on his career and does not have children.
- Anastasia has never rollerbladed.
- Miles is very tall.

Example 6.6: Neighborhood Survey (cont.)

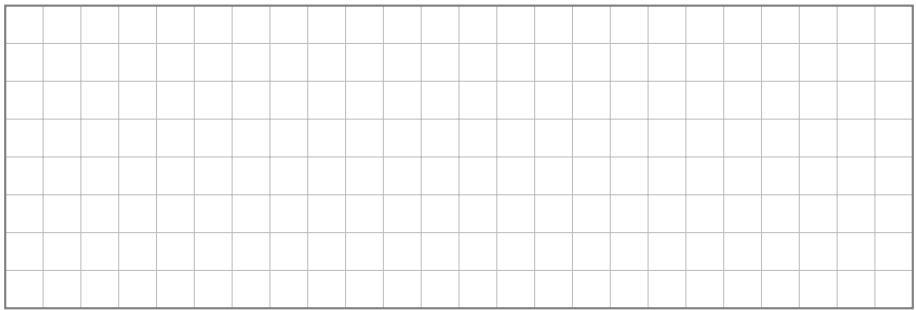
Context: A neighborhood survey on rollerblading and having children.

David is focusing on his career and does not have children. What is the probability that David rollerblades?

Example 6.6: Neighborhood Survey (cont.)

Context: A neighborhood survey on rollerblading and having children.

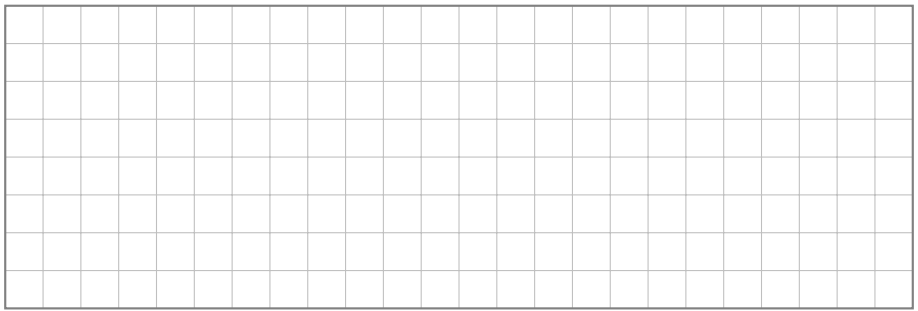
Anastasia has never rollerbladed. What are the odds that Anastasia has children?



Example 6.6: Neighborhood Survey (cont.)

Context: A neighborhood survey on rollerblading and having children.

Miles is very tall. What is the risk of Miles rollerblading?



PART 3

Comparing Groups: RR & OR

Relative Risk

Relative Risk (RR)

The **relative risk** compares the risk (probability) in one group to the risk in another group (baseline):

$$\text{Relative Risk} = \frac{P(\text{Outcome} \mid \text{Exposed})}{P(\text{Outcome} \mid \text{Unexposed})}$$

Interpreting RR:

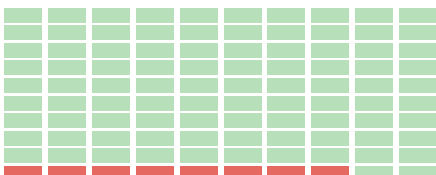
- RR $= 1$: Both groups have equal risk (no association)
- RR > 1: Exposed group has **higher** risk than baseline
- RR < 1: Exposed group has **lower** risk than baseline

Note: The baseline group is typically the **control** or **unexposed** group.

Visualizing Relative Risk: Two Towns

Example: Disease outbreak: 8% in Town A vs 2% in Town B $\implies$ RR $= 4.0$

Town A (100 residents)



■ 8 sick (8%)

Town B (100 residents)

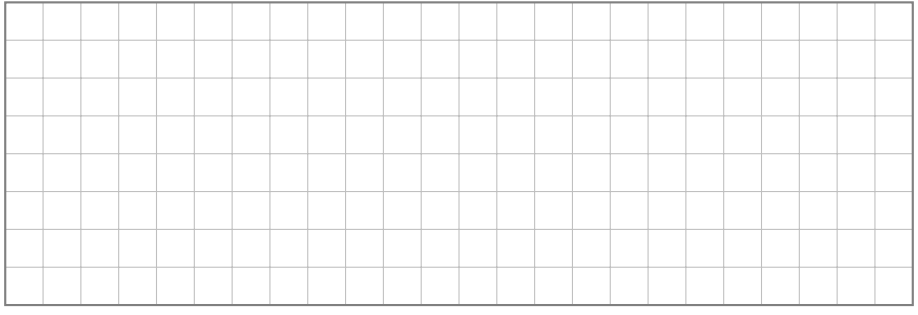
■ 2 sick (2%)

Example 6.7: Flu Vaccine Study

Context: A study comparing flu rates between vaccinated and control groups.

Group	Developed Flu	No Flu	Total
Vaccinated	26	3874	3900
Control	70	3830	3900
Total	96	7704	7800

Find the relative risk of developing flu for the vaccinated group compared to the control group.

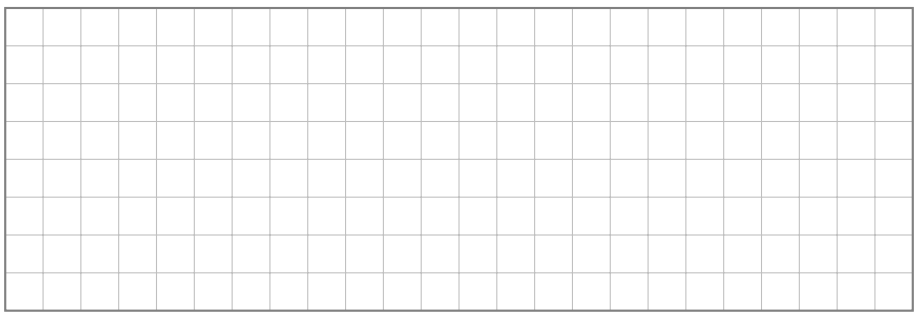


Example 6.2: Pet Ownership (Relative Risk)

Context: Dog ownership by sex.

Sex	Cat	Dog	Total
Female	20	30	50
Male	40	40	80

Calculate RR of dog ownership for females compared to males:



Relative Risk: Terminology

All of the following are equivalent ways to ask for **relative risk** between two groups:

- Calculate the relative risk of dog ownership for females **compared to** males.
- Find the relative risk for dog ownership for females **relative to** males.
- Determine the RR of dog ownership, females **compared with** males.
- Estimate the relative risk for dog ownership: females **vs** males.
- Compute the relative risk for dog ownership in females **versus** males.

Odds Ratio

Odds Ratio (OR)

The **odds ratio** compares the odds in one group to the odds in another group:

$$\text{Odds Ratio} = \frac{\text{Odds in Exposed Group}}{\text{Odds in Unexposed Group}}$$

Interpreting OR:

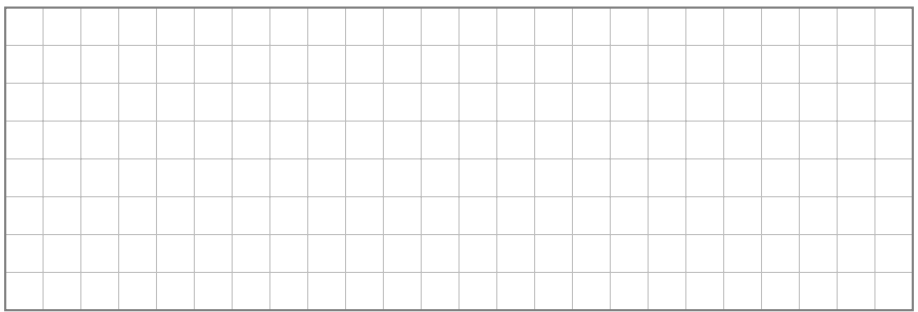
- OR $= 1$: Both groups have equal odds (no association)
- OR > 1: Exposed group has **higher** odds than baseline
- OR < 1: Exposed group has **lower** odds than baseline

Example 6.7: Flu Vaccine (Odds Ratio)

Context: Same flu vaccine study.

Group	Developed Flu	No Flu	Total
Vaccinated	26	3874	3900
Control	70	3830	3900

Calculate odds ratio of developing flu (vaccinated vs. control).

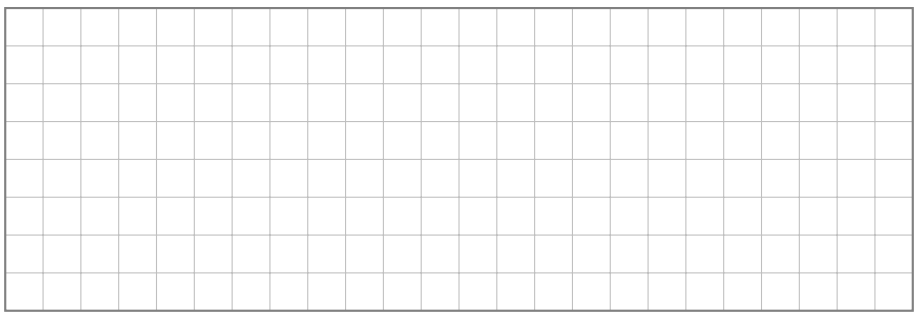


Example 6.2: Pet Ownership (Odds Ratio)

Context: Dog ownership by sex.

Sex	Cat	Dog	Total
Female	20	30	50
Male	40	40	80

Calculate OR of dog ownership for females compared to males:



Odds Ratio: Cross-Product Shortcut

For a 2×2 table, there is a shortcut to compute OR:

	Outcome Yes	**Outcome No**
Group 1 (Exposed)	a	b
Group 2 (Unexposed)	c	d

$$\text{OR} = \frac{a \times d}{b \times c}$$

Relative Risk vs. Odds Ratio

Relative Risk (RR)

- Compares **probabilities**
- Easier to interpret
- "X times more likely"
- Requires knowing group totals

Odds Ratio (OR)

- Compares **odds**
- Used in logistic regression
- "X times higher odds"
- Cross-product shortcut

PART 4

Simpson's Paradox

Example 6.8: Simpson's Paradox

Context: UC Berkeley Graduate Admissions (1973). Was there bias against women?

Status	# Men	# Women	Total
Admitted	1,198	557	1,755
Rejected	1,493	1,278	2,771
Total	2,691	1,835	4,526

Admission rates:

- Men:
- Women:

Example 6.8: By Department

When we look at the data by department:

	Number of Applicants		Admission Rate	
Dept	**Men**	**Women**	**Men**	**Women**
A	825	108	62%	82%
B	560	25	63%	68%
C	325	593	37%	34%
D	417	375	33%	35%
E	191	393	28%	24%
F	373	341	6%	7%

In 4 of 6 departments, women had **higher** admission rates than men.

Understanding Simpson's Paradox

Simpson's Paradox

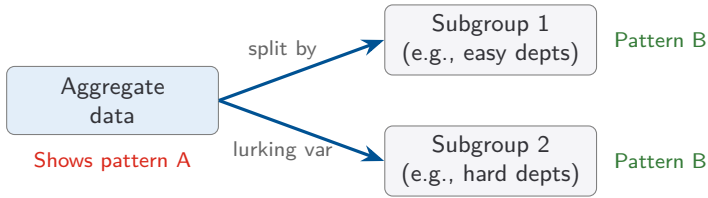
A trend that appears in aggregate data **reverses** or disappears when the data is split into subgroups.

What happened at Berkeley?

- Women applied more to competitive departments (C, D, E, F) with low admit rates
- Men applied more to less competitive departments (A, B) with high admit rates
- **Department choice** was a confounding variable!

Why Does Simpson's Paradox Happen?

The key mechanism: Groups have **unequal representation** across subgroups.



Unequal group sizes across subgroups create a misleading aggregate

🔑 **Key Point:** Simpson's Paradox occurs when a lurking variable is **unevenly distributed** across the groups being compared, and that lurking variable is also **related to the outcome**.

Example 6.10: Basketball Win Rates

Scenario: Two players, Brandon and Luca, play basketball in different settings.

Brandon's Record:

- Plays mostly against **children**
- Rarely plays against **peers**

Luca's Record:

- Plays mostly against **peers**
- Rarely plays against **children**

🔑 **Key Point:** Let's look at their overall win rates and see if we can determine who is the better player.

Overall Win Rates

Player	Wins	Total Games	Win Rate
Brandon	78	100	78%
Luca	70	100	70%

First impression: Brandon has a higher overall win rate. However, what happens when we break down their performance by opponent type?

Breaking Down by Opponent Type

Games Against Children:

	Wins	Total
Brandon	75	80
Luca	10	10

- Brandon: $75/80 = 93.8\%$
- Luca: $10/10 = 100.0\%$

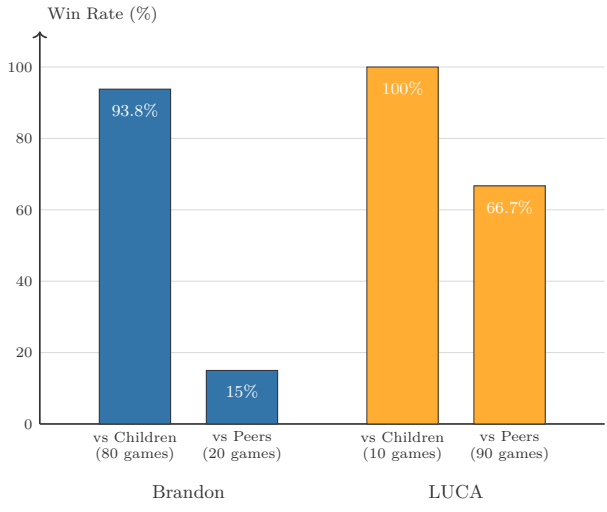
Games Against Peers:

	Wins	Total
Brandon	3	20
Luca	60	90

- Brandon: $3/20 = 15.0\%$
- Luca: $60/90 = 66.7\%$

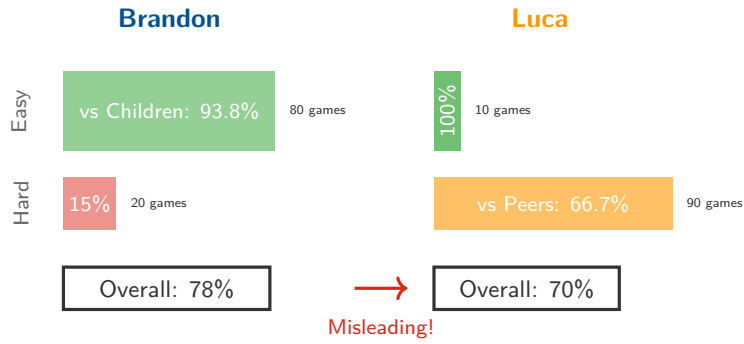
Key Point: Luca has a higher win rate in **both** opponent categories.

Visualizing the Win Rates



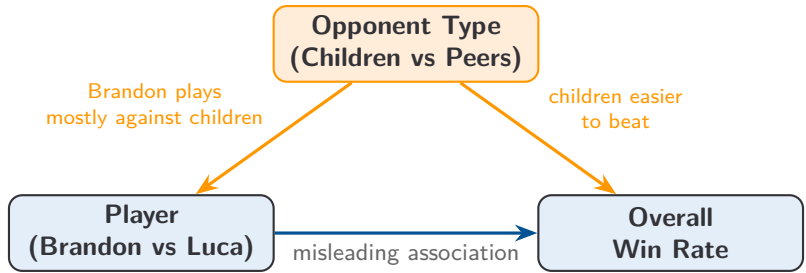
The Weighting Trap: Why Aggregates Mislead

Example: Two players with different opponent mixes.



Key Point: Bar width = sample size. Brandon's high average is inflated by playing mostly easy games.

The Hidden Variable: Opponent Type



Key Point: **Opponent type** is the confounding variable:

- Related to player (Brandon chooses children, Luca chooses peers)
- Related to outcome (children are easier opponents)
- Creates a misleading aggregate comparison

Simpson's Paradox in Basketball

Caution: The Paradox:

Brandon has a 78% overall win rate vs Luca's 70%

BUT Luca wins more often than Brandon against **both** children and peers.

Why this happens:

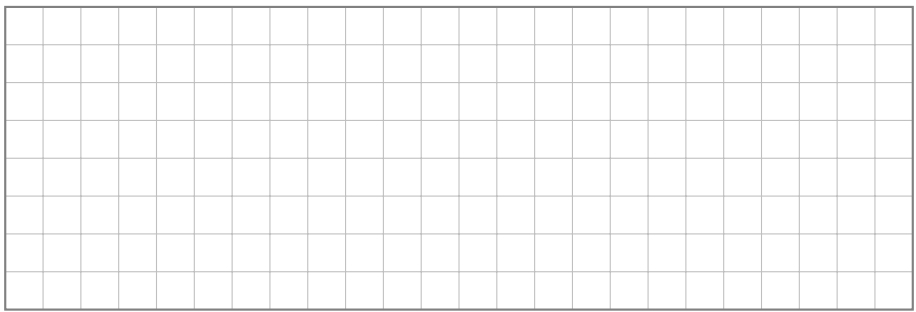
- Brandon plays **80%** of games against children (easy wins)
- Luca plays **90%** of games against peers (harder competition)
- The **mix of opponents** differs dramatically between players
- This creates a misleading aggregate picture

Example 6.9: Kidney Stone Treatments

Context: Two kidney treatments (A and B) were tested on 700 patients.

Result	Treatment A	Treatment B	Total
Success	273	289	562
Failure	77	61	138
Total	350	350	700

Identify the success rates for each treatment. Which treatment appears better overall?



Example 6.9: By Stone Size

When we segment the data by stone size, we get:

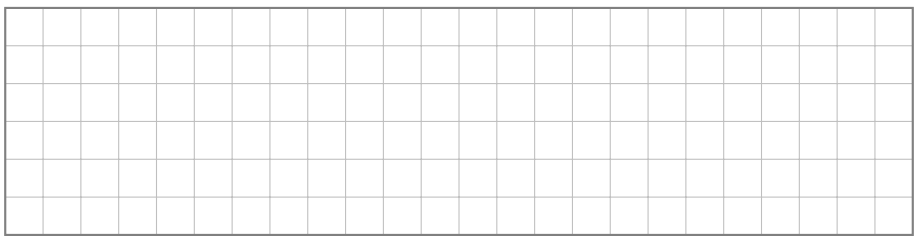
Small Stones:

	A	B
Success	81	234
Total	87	270
Rate	93%	87%

Large Stones:

	A	B
Success	192	55
Total	263	80
Rate	73%	69%

Which is the preferred treatment for each stone size?



Confounding Variables

Lurking Variable

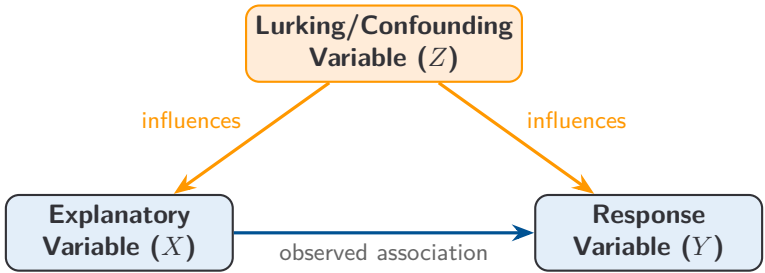
A **lurking variable** is an unobserved variable that influences both the explanatory and response variables, potentially leading to a spurious association between them.

Confounding Variable

A **confounding variable** is a variable that influences both the explanatory and response variables, and is **measured** in the study.

⚠ **Caution:** Confounding can cause associations to appear, disappear, or even reverse.

Confounding: A Causal Diagram



🔑 **Key Point:** A confounding variable Z creates a **spurious association** between X and Y because Z affects both.

Berkeley example: X = Gender, Y = Admission, Z = Department choice.

CHAPTER 6

Summary

Chapter 6 Summary

Two-Way Tables

- **Two-way tables:** counts for two categorical variables
- **Marginal:** totals (ignoring the other variable)
- **Conditional:** within each level of another variable
- **Independence:** conditional = marginal for all groups

Risk Measures

- **Relative risk:** ratio of probabilities
- **Odds ratio:** ratio of odds (cross-product: $\frac{ad}{bc}$)

Cautions

- **Simpson's paradox:** trends reverse when accounting for a third variable
- **Confounding:** affects both explanatory and response
- **Lurking:** unmeasured confounder

From Tables to Probability

Key Point: Throughout this chapter, every proportion you computed from a contingency table was a **probability**.

- A **marginal proportion** $P(Y)$ is the probability of an outcome in the whole population
- A **conditional proportion** $P(Y \mid X)$ is the probability of an outcome *within a specific group*
- **Relative risk** compares two such probabilities: $\dfrac{P(Y \mid \text{Exposed})}{P(Y \mid \text{Unexposed})}$

Looking Ahead
In Chapter 12, we develop a formal framework for probability, with precise rules for combining and reasoning about uncertain events. The intuition you built here with contingency tables carries directly into that framework.

PRACTICE

Exercises

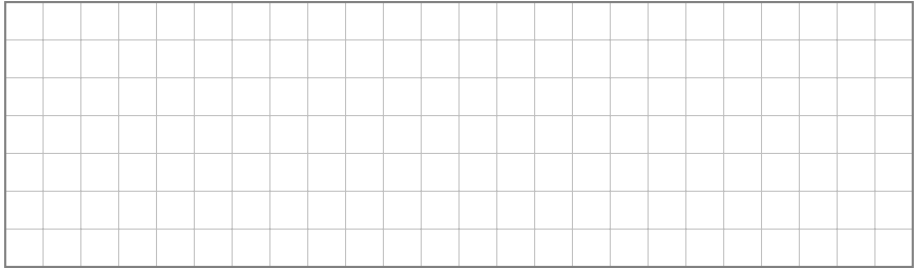
Example 6.11: Screen Time and Physical Activity

Context: A study of 400 teens classified by screen time and physical activity level.

	Active	**Inactive**	**Total**
High screen time	60	140	200
Low screen time	120	80	200
Total	180	220	400

Find:

1. Marginal distribution of physical activity level
2. Conditional distribution of activity for each screen time group
3. Are screen time and activity level independent? Explain
4. RR of being active for low vs. high screen time



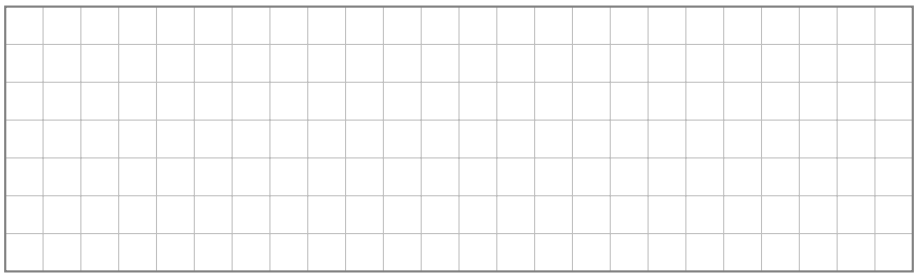
Example 6.12: Drug Side Effects

Context: A clinical trial with 500 participants testing a new drug.

	Side Effect	No SE	Total
Drug	36	214	250
Placebo	12	238	250
Total	48	452	500

Find:

1. Risk of side effect in each group
2. Relative risk (Drug vs. Placebo)
3. Odds ratio using cross-product
4. Interpret RR and OR in context



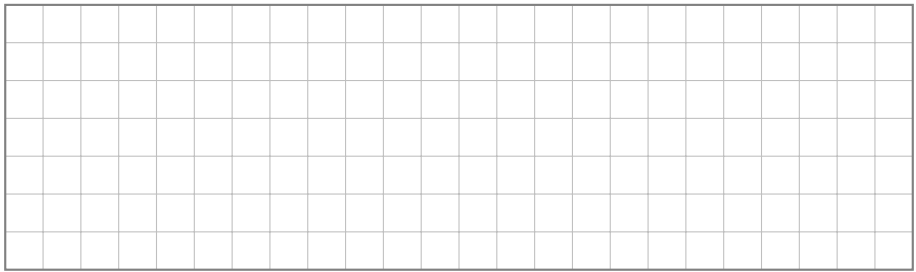
Example 6.13: Study Methods and Exam Outcomes

Context: A university tracked 300 students by study method and exam result.

	Pass	**Fail**	**Total**
Online study	90	60	150
In-person study	120	30	150
Total	210	90	300

Find:

1. Marginal distribution of exam result
2. Conditional distribution for each study method
3. RR and OR of passing (in-person vs. online)
4. Is in-person associated with better outcomes?



Example 6.14: Working Backwards

Context: In a study of 600 patients (300 per group), the relative risk of infection for the treatment group compared to the control group is RR $= 0.40$. The risk of infection in the **control group** is 0.25.

Find:

1. Risk of infection in treatment group
2. Number of infections in each group
3. Reconstruct the full two-way table
4. Odds ratio of infection (treatment vs. control)