Chapter 2 Contingency Tables

Outline

1. 2.1 Probability Structure For Contingency Tables
2. 2.2 Comparing Proportions in Two-by-Two Tables
3. 2.3 The Odds Ratio
4. 2.4 Chi-Square Tests of Independence
5. 2.5 Testing Independence for Ordinal Data
6. 2.6 Exact Inference for Small Samples
7. 2.7 Association in Three-Way Tables
8. Homework 2

Table 2.1 cross classifies a sample of Americans according to their gender and their opinion about an afterlife. For the females in the sample, for example, 509 said they believed in an afterlife and 116 said they did not or were undecided. **Does an association exist between gender and belief in an afterlife? Is one gender more likely than the other to believe in an afterlife, or is belief in an afterlife independent of gender?**

Table 2.1. Cross Classification of Belief in Afterlife by Gender

	Belief in Afterlife	
Gender	Yes	No or Undecided
Females	509	116
Males	398	104

Source: Data from 1998 General Social Survey.

Outline

1. 2.1 Probability Structure For Contingency Tables
2. 2.2 Comparing Proportions in Two-by-Two Tables
3. 2.3 The Odds Ratio
4. 2.4 Chi-Square Tests of Independence
5. 2.5 Testing Independence for Ordinal Data
6. 2.6 Exact Inference for Small Samples
7. 2.7 Association in Three-Way Tables
8. Homework 2

- Single categorical variable: summarize the data by counting the number of observations in each category. The sample proportions in the categories estimate the category probabilities.
- Two categorical variables X, Y:

$$I - \#\{\text{categories of } X\}, J - \#\{\text{categories of } Y\}.$$

A rectangular table having I rows for the categories of X and J columns for the categories of Y has cells that display the IJ possible combinations of outcomes.

- A table of this form that displays counts of outcomes in the cells is called a *contingency table*. A table that cross classifies two variables is called a *two-way contingency table*; one that cross classified three variables is called a *three-way contingency table*, and so forth. A two-way table with I rows and J columns is called an $I \times J$ table.

2.1.1 Joint, Marginal, and Conditional Probabilities

- *The Joint distribution of X and Y*: $\pi_{ij} = P(X = i, Y = j)$, $\sum_{i,j} \pi_{ij} = 1$.
- *The Marginal Distribution*: row and column totals of the joint probabilities. $\{\pi_{i+}\}$ for the row variable, $\{\pi_{+j}\}$ for the column variable.
- Similarly, $\{p_{ij}\}$ are cell proportions in a sample joint distribution. $\{n_{ij}\}$ - cell counts, $\{n_{i+}\}$ - marginal frequencies are the row total, $\{n_{+j}\}$ - column totals, $n = \sum_{i,j} n_{ij}$ - sample size. $p_{ij} = n_{ij}/n$.
- *Conditional Distribution*: the probability distribution for Y at each level of X.

2.1.2 Example: Belief in Afterlife

Table 2.1 cross classified $n = 1127$ respondents to a General Social Survey by their gender and by their belief in an afterlife. Table 2.2 illustrates the cell count notation for these data.

Table 2.2, Notation for Table 2.1

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Females	$n_{11} = 509$	$n_{12} = 116$	$n_{1+} = 625$
Males	$n_{21} = 398$	$n_{22} = 104$	$n_{2+} = 502$
Total	$n_{+1} = 907$	$n_{+2} = 220$	$n = 1127$

2.1.2 Example: Belief in Afterlife

Remark

In Table 2.1, belief in the afterlife is a response variable and gender is an explanatory variable. We therefore study the conditional distributions of belief in the afterlife, given gender. For females, the proportion of “yes” responses was $509/625 = 0.81$ and the proportion of “no” responses was $116/625 = 0.19$. The proportions $(0.81, 0.19)$ form the sample conditional distribution of belief in the afterlife. For males, the sample conditional distribution is $(0.79, 0.21)$.

2.1.3 Sensitivity and Specificity in Diagnostic Tests

Positive & Negative

The result of a diagnostic test is said to be *positive* if it states that the disease is present and *negative* if it sates that the disease is absent.

The accuracy of diagnostic tests

Given that a subject has the disease, the probability the diagnostic test is positive is called the *sensitivity*. Given that the subject does not have the disease, the probability the test is negative is called the *specificity*.

2.1.3 Sensitivity and Specificity in Diagnostic Tests

The accuracy of diagnostic tests

The higher the sensitivity and specificity, the better the diagnostic test.

$$X - \text{true state of a person} \begin{cases} 1 & \text{diseased,} \\ 2 & \text{not diseased.} \end{cases}$$

$$Y - \text{outcome of diagnostic test} \begin{cases} 1 & \text{positive,} \\ 2 & \text{negative.} \end{cases}$$

$$\text{sensitivity} = P(Y = 1|X = 1), \ \text{specificity} = P(Y = 2|X = 2).$$

2.1.3 Example

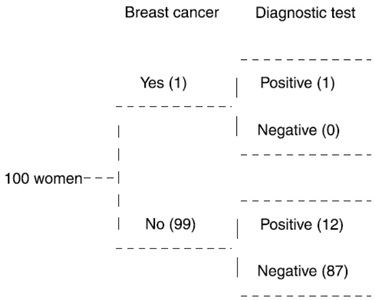


Figure 2.1. Tree diagram showing results of 100 mammograms, when sensitivity $= 0.86$ and specificity $= 0.88$.

2.1.3

Example

For a women with breast cancer, there is a 0.86 probability of detecting it. Figure 2.1 shows that of the 13 women with a positive test result, the proportion $1/13 = 0.08$ actually have breast cancer. The small proportion of errors for the large majority of women who do not have breast cancer swamps the large proportion of correct diagnoses for the few women who have it.

Monty Hall Problem

Monty Hall Problem

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Monty Hall Problem: Standard Assumptions

- The host must always open a door that was not picked by the contestant (Mueser and Granberg 1999).
- The host must always open a door to reveal a goat and never the car.
- The host must always offer the chance to switch between the originally chosen door and the remaining closed door.

Monty Hall Problem: Direct Solution

Consider the events C_1, C_2 and C_3 indicating the car is behind respectively door 1, 2 or 3. All these 3 events have probability $\frac{1}{3}$. The player initially choosing door 1 is described by the event X_1. As the first choice of the player is independent of the position of the car, also the conditional probabilities are $P(C_i|X_1) = \frac{1}{3}$. The host opening door 3 is described by H_3. For this event it holds:

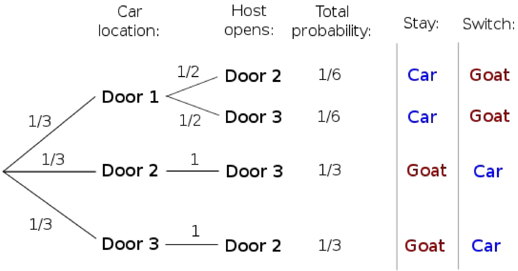
$$P(H_3|C_1, X_1) = \frac{1}{2},\ P(H_3|C_2, X_1) = 1,\ P(H_3|C_3, X_1) = 0$$

Monty Hall Problem: Direct Solution

Then, if the player initially selects door 1, and the host opens door 3, the conditional probability of winning by switching is

$$
\begin{aligned}
P(C_2|H_3, X_1) &= \frac{P(H_3|C_2, X_1)P(C_2 \cap X_1)}{P(H_3 \cap X_1)} \\
&= \frac{P(H_3|C_2, X_1)P(C_2 \cap X_1)}{P(H_3|C_1, X_1)P(C_1 \cap X_1) + P(H_3|C_2, X_1)P(C_2 \cap X_1) + P(H_3|C_3, X_1)P(C_3 \cap X_1)} \\
&= \frac{P(H_3|C_2, X_1)}{P(H_3|C_1, X_1) + P(H_3|C_2, X_1) + P(H_3|C_3, X_1)} = \frac{1}{\frac{1}{2} + 1 + 0} = \frac{2}{3}
\end{aligned}
$$

Monty Hall Problem: Solution 2



Car location:	Host opens:	Total probability:	Stay:	Switch:
Door 1	Door 2	1/6	Car	Goat
	Door 3	1/6	Car	Goat
Door 2	Door 3	1/3	Goat	Car
Door 3	Door 2	1/3	Goat	Car

Three Prisoners Problem

Three prisoners, A, B and C, are in separate cells and sentenced to death. The governor has selected one of them at random to be pardoned. The warden knows which one is pardoned, but is not allowed to tell. Prisoner A begs the warden to let him know the identity of one of the others who is going to be executed. “If B is to be pardoned, give me C’s name. If C is to be pardoned, give me B’s name. And if I’m to be pardoned, flip a coin to decide whether to name B or C.” The warden tells A that B is to be executed. Prisoner A is pleased because he believes that his probability of surviving has gone up from 1/3 to 1/2, as it is now between him and C. Prisoner A secretly tells C the news, who is also pleased, because he reasons that A still has a chance of 1/3 to be the pardoned one, but his chance has gone up to 2/3. What is the correct answer?

Three Prisoners Problem: solution

Call A, B and C the events that the corresponding prisoner will be pardoned, and b the event that the warden mentions prisoner B as the one not being pardoned, then, using Bayes' formula, the posterior probability of A being pardoned, is:

$$\begin{aligned} P(A|b) = & \frac{P(b|A)P(A)}{P(b|A)P(A) + P(b|B)P(B) + P(b|C)P(C)} \\ & = \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2} \times \frac{1}{3} + 0 \times \frac{1}{3} + 1 \times \frac{1}{3}} = \frac{1}{3}. \end{aligned}$$

Three Prisoners Problem

Why the Paradox?

The tendency of people to provide the answer 1/2 neglects to take into account that the warden may have tossed a coin before he gave his answer. The warden may have answered B because A is to be released and he tossed a coin. Or, C is to be released. The probabilities of the two events are not equal.

2.1.4 Independence

Statistically Independent

Two variables are said to be *statistically independent* if the population conditional distributions of Y are identical at each level of X. When two variables are independent, the probability of any particular column outcome j is the same in each row.

2.1.4 Independence

When both variables are response variables, we can describe their relationship using their joint distribution, or the conditional distribution of Y given X, or the conditional distribution of X given Y.

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \quad \text{for } i = 1, ..., I \text{ and } j = 1, ..., J.$$

2.1.5 Binomial and Multinomial Sampling

- It is often sensible to assume that cell counts in contingency tables have one of these distributions.
- When the columns are a response variable and the rows are an explanatory variable, it is sensible to divide the cell counts by the row totals to form conditional distributions on the response. In doing so, we inherently treat the row totals as fixed and analyze the data the same way as if the two rows formed separate samples.

2.1.5 Binomial and Multinomial Sampling

- When the total sample size n is fixed and we cross classify the sample on two categorical response variables, the multinomial distribution is the actual joint distribution over the cells. The cells of the contingency table are the possible outcomes, and the cell probabilities are the multinomial parameters.

Outline

1. 2.1 Probability Structure For Contingency Tables
2. 2.2 Comparing Proportions in Two-by-Two Tables
3. 2.3 The Odds Ratio
4. 2.4 Chi-Square Tests of Independence
5. 2.5 Testing Independence for Ordinal Data
6. 2.6 Exact Inference for Small Samples
7. 2.7 Association in Three-Way Tables
8. Homework 2

Introduction

Response variables having two categories are called binary variables. For instance, belief in afterlife is binary when measured with categories (yes, no). Many studies compare two groups on a binary response, Y . The data can be displayed in a 2×2 contingency table, in which the rows are the two groups and the columns are the response levels of Y . This section presents measures for comparing groups on binary responses.

2.2.1 Difference of Proportions

- *The difference of proportions* $\pi_1 - \pi_2$ compares the success probabilities in the two rows. Let p_1 and p_2 denote the *s*ample proportions of successes. The sample difference $p_1 - p_2$ estimates $\pi_1 - \pi_2$.
- When the counts in the two rows are independent binomial samples, the estimated standard error of $p_1 - p_2$ is

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

where the sample sizes for the two groups(n_{1+}, n_{2+}) is denoted by n_1 and n_2.

2.2.1 Difference of Proportions

- A large-sample $100(1-\alpha)\%$(Wald) confidence interval for $\pi_1 - \pi_2$ is

$$(p_1 - p_2) \pm z_{\alpha/2}(SE)$$

2.2.2 Example: Aspirin and Heart Attacks

Example

Table 2.3 Cross Classification of Aspirin Use and Mocardial Infarction

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037

*S*ource: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262-264, 1988.

2.2.2 Example: Aspirin and Heart Attacks

Table 2.3 is from a report on the relationship between aspirin use and myocardial infarction (heart attacks) by the Physicians Health Study Research Group at Harvard Medical School. The Physicians' Health Study was a five-year randomized study testing whether regular intake of aspirin reduces mortality from cardiovascular disease. Every other day, the male physicians participating in the study took either one aspirin tablet or a placebo. The study was "blind"- the physicians in the study did not know which type of pill they were taking.

2.2.2 Example: Aspirin and Heart Attacks

We treat the two rows in Table 2.3 as independent binomial samples. Of the $n_1 = 11,034$ physicians taking placebo, 189 suffered myocardial infarction (MI) during the study, a proportion of $p_1 = 189/11,034 = 0.0171$. Of the $n_2 = 11,037$ physicians taking aspirin, 104 suffered MI, a proportion of $p_2 = 0.0094$. The sample difference of proportions is $0.0171 - 0.0094 = 0.0077$. From equation of SE, this difference has an estimated standard error of

$$SE = \sqrt{\frac{(0.0171)(0.9829)}{11,034} + \frac{(0.0094)(0.9906)}{11,037}}$$

What's the 95% confidence interval for the true difference $\pi_1 - \pi_2$?

2.2.3 Relative Risk

Question

Is the difference between 0.010 and 0.001 the same as the difference between 0.410 and 0.401?

The ratio of proportions is a more relevant descriptive measure.

Relative Risk

For 2 × 2 tables, the *relative risk* is the ratio

$$\text{relative risk} = \frac{\pi_1}{\pi_2}$$

2.2.3 Relative Risk

- Two groups with *sample* proportions p_1 and p_2 have a sample relative risk of p_1/p_2.
- Table 2.3, the sample relative risk is $p_1/p_2 = 0.0171/0.0094 = 1.82$. The sample proportion of MI cases was 82% higher for the group taking placebo.
- Using the difference of proportions alone to compare two groups can be misleading when the proportions are both close to zero.

2.2.3 Relative Risk

- The sampling distribution of the sample relative risk is highly skewed unless the sample sizes are quite large. A 95% confidence interval for the true relative risk is $(1.43, 2.30)$. This indicates that the risk of MI is at least 43% higher for the placebo group.
- The ratio of failure probabilities $(1-\pi_1)/(1-\pi_2)$.

Outline

1. 2.1 Probability Structure For Contingency Tables
2. 2.2 Comparing Proportions in Two-by-Two Tables
3. 2.3 The Odds Ratio
4. 2.4 Chi-Square Tests of Independence
5. 2.5 Testing Independence for Ordinal Data
6. 2.6 Exact Inference for Small Samples
7. 2.7 Association in Three-Way Tables
8. Homework 2

Definition

For a probability of success π, the *odds* of success are defined to be

$$\text{odds} = \pi/(1-\pi)$$

- The odds are nonnegative, $\text{odds} > 1$, a success is more likely than a failure.
- The success probability itself is the function of the odds,

$$\pi = \text{odds}/(\text{odds}+1)$$

Definition

In 2×2 tables, within row 1, the *odds* of success are $\text{odds}_1 = \pi_1/(1-\pi_1)$, and within row 2 the odds of success equal $\text{odds}_2 = \pi_2/(1-\pi_2)$. The ratio of the odds from the two rows,

$$\theta = \frac{\text{odds}_1}{\text{odds}_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

is the *odds ratio*. Whereas the relative risk is a ratio of two probabilities, the odds ratio θ is a ratio of two odds.

2.3.1 Properties of the Odds Ratio

- The odds ratio is nonnegative. When X and Y are independent, $\pi_1 = \pi_2$, so $\theta = 1$. This is a baseline for comparison.
- $\theta > 1$, the odds of success are higher in row 1 than in row 2. Thus, subjects in row 1 are more likely to have successes than are subjects in row 2, that is $\pi_1 > \pi_2$.
- $\theta < 1$, a success is less likely in row 1 than in row 2, that is, $\pi_1 < \pi_2$.

2.3.1 Properties of the Odds Ratio(Continue)

- Two values for θ represent the same strength of association, but in opposite direction, when one value is the inverse of the other.
- The odds ratio does not change value when the table orientation reverses so that the rows become the columns and the columns become the rows.

2.3.1 Properties of the Odds Ratio(Continue)

- When both variables are response variables, the odds ratio can be defined using joint probabilities as

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

  The odds ratio is also called the *cross-product ratio*.

- The sample odds ratio is

$$\hat{\theta} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

2.3.2 Example: Odds Ratio for Aspirin Use and Heart Attacks

Example

Let us revisit Table 2.3 from Section 2.2.2 on aspirin use and myocardial infarction. For the physicians taking placebo, the estimated odds of MI is $n_{11}/n_{12} = 189/10,845 = 0.0174$. The estimated odds is $104/10,933 = 0.0095$ for those taking aspirin. The sample odds ratio is $\hat{\theta} = 0.0174/0.0095 = 1.832 = (189 \times 10,933)/(10,845 \times 104)$. The estimated odds of MI for male physicians taking placebo equal 1.83 times the estimated odds for male physicians taking aspirin. The estimated odds were 83% higher for the placebo group.

2.3.3 Inference for Odds Ratios and Log Odds Ratios

- Unless the sample size is extremely large, the sampling distribution of the odds ratio is highly skewed.
- Because of the skewness, statistical inference for the odds ratio uses an alternative but equivalent measure-its natural logarithm, $\log(\theta)$. Independence corresponds to $\log(\theta) = 0$.
- The log odds ratio is symmetric about zero, in the sense that reversing rows or reversing columns changes its sign.

2.3.3 Inference for Odds Ratios and Log Odds Ratios

- $\log(\hat{\theta})$ has an approximating normal distribution with mean of $\log(\theta)$ and standard error of

$$SE = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \tag{1}$$

- A large-sample confidence interval for $\log(\theta)$ is

$$\log \hat{\theta} \pm z_{\alpha/2}(SE)$$

2.3.3 Inference for Odds Ratios and Log Odds Ratios(Continue)

- The sample odds ratio $\hat{\theta}$ equals 0 or ∞ if any $n_{ij} = 0$, and it is undefined if both entries in a row or column are zero. The slightly amended estimator

$$\tilde{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)}$$

corresponding to adding 1/2 to each cell count, is preferred when any cell counts are very small. In that case, the SE formula (1) replaces $\{n_{ij}\}$ by $\{n_{ij} + 0.5\}$.

2.3.4 Relationship Between Odds Ratio and Relative Risk

Remark

A sample odds ratio of 1.83 does not mean that p_1 is 1.83 times p_2.

$$\text{Odds ratio} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \text{Relative risk} \times (\frac{1-p_2}{1-p_1})$$

This relationship between the odds ratio and the relative risk is useful. For some data sets direct estimation of the relative risk is not possible, yet one can estimate the odds ratio and use it to approximate the relative risk, as the next example illustrates.

2.3.5 The Odds Ratio Applies in Case-Control Studies

Table 2.4. Cross Classification of Smoking Status and Myocardial Infarction

Ever Smoker	MI Cases	Control
Yes	172	173
No	90	346

*S*ource: A.Gramenzi et al., *J. Epidemiol. Community Health*, **43**:214-217, 1989. Reprinted with permission by BMJ Publishing Group.

2.3.5 The Odds Ratio Applies in Case-Control Studies

Example

Table 2.4 refers to a study that investigated the relationship between smoking and myocardial infarction. The first column refers to 262 young and middle-aged women (age < 69) admitted to 30 coronary care units in northern Italy with acute MI during a 5-year period. Each case was matched with two control patients admitted to the same hospitals with other acute disorders. The controls fall in the second column of the table. All subjects were classified according to whether they had ever been smokers. The “yes group consists of women who were current smokers or ex-smokers, whereas the “no group consists of women who never were smokers.We refer to this variable as smoking status.

2.3.5 The Odds Ratio Applies in Case-Control Studies

We would normally regard MI as a response variable and smoking status as an explanatory variable. In this study, however, the marginal distribution of MI is fixed by the sampling design, there being two controls for each case. The outcome measured for each subject is whether she ever was a smoker. The study, which uses a *retrospective* design to look into the past, is called a *case control study*. Such studies are common in health-related applications, for instance to ensure a sufficiently large sample of subjects having the disease studied.

2.3.6 Types of Observational Studies

Prospective Design

By contrast to the study summarized by Table 2.4, imagine a study that follows a sample of women for the next 20 years, observing the rates of MI for smokers and nonsmokers. Such a sampling design is *prospective*.

2.3.6 Types of Observational Studies

Types of prospective studies

- In *cohort studies*, the subjects make their own choice about which group to join (e.g., whether to be a smoker), and we simply observe in future time who suffers MI.
- In *clinical trials*, we randomly allocate subjects to the two groups of interest, such as in the aspirin study described in Section 2.2.2, again observing in future time who suffers MI.

2.3.6 Types of Observational Studies

Cross-sectional design

A *cross-sectional design*, samples women and classifies them simultaneously on the group classification and their current response. As in a case-control study, we can then gather the data at once, rather than waiting for future events

Case-control, cohort, and cross-sectional studies are *observational studies*. We observe who chooses each group and who has the outcome of interest. By contrast, a clinical trial is an *experimental study*, the investigator having control over which subjects enter each group, for instance, which subjects take aspirin and which take placebo.

Outline

1. 2.1 Probability Structure For Contingency Tables
2. 2.2 Comparing Proportions in Two-by-Two Tables
3. 2.3 The Odds Ratio
4. 2.4 Chi-Square Tests of Independence
5. 2.5 Testing Independence for Ordinal Data
6. 2.6 Exact Inference for Small Samples
7. 2.7 Association in Three-Way Tables
8. Homework 2

Expected frequencies

Consider the null hypothesis (H_0) that cell probabilities equal certain fixed values $\{\pi_{ij}\}$. For a sample of size n with cell counts $\{n_{ij}\}$, the values $\{\mu_{ij} = n\pi_{ij}\}$ are *expected frequencies*. They represent the values of the expectations $\{E(n_{ij})\}$ when H_0 is true.

To judge whether the data contradict H_0, we compare $\{n_{ij}\}$ to $\{\mu_{ij}\}$. If H_0 is true, n_{ij} should be close to μ_{ij} in each cell. The larger the differences $\{n_{ij} - \mu_{ij}\}$, the stronger the evidence against H_0. The test statistics used to make such comparisons have large-sample chi-squared distributions.

2.4.1 Pearson Statistic and the Chi-Squared Distribution

Pearson chi-squared statistic

The *Pearson chi-squared statistic* for testing H_0 is

$$X^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}.$$

2.4.1 Pearson Statistic and the Chi-Squared Distribution

Pearson chi-squared statistic

It was proposed in 1900 by Karl Pearson, the British statistician known also for the Pearson product-moment correlation estimate, among many contributions. This statistic takes its minimum value of zero when all $n_{ij} = \mu_{ij}$. For a fixed sample size, greater differences $\{n_{ij} - \mu_{ij}\}$ produce larger X^2 values and stronger evidence against H_0.

2.4.1 Pearson Statistic and the Chi-Squared Distribution

Larger X^2 values are more contradictory to H_0, the P-value is the null probability that X^2 is at least as large as the observed value. The X^2 statistic has approximately a chi-squared distribution, for large n. The P-value is the chi-squared right-tail probability above the observed X^2 value. The chi-squared approximation improves as $\{\mu_{ij}\}$ increase, and $\{\mu_{ij} \geq 5\}$ **is usually sufficient for a decent approximation**.

2.4.1 Pearson Statistic and the Chi-Squared Distribution

The chi-squared has mean equal to its degrees of freedom (df), and standard deviation equals $\sqrt{(2df)}$. As df increases, the distribution concentrates around larger values and is more spread out. The distribution is skewed to the right, but it becomes more bell-shaped (normal) as df increases. Figure 2.2 displays chi-squared densities having $df = 1, 5, 10, 20$.

2.4.1 Pearson Statistic and the Chi-Squared Distribution

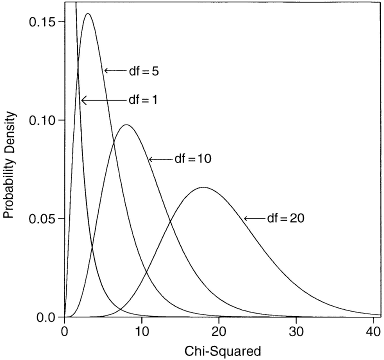


Figure 2.2

2.4.2 Likelihood-Ratio Statistic

The likelihood-ratio test determines the parameter values that maximize the likelihood function (a) under the assumption that H_0 is true, (b) under the more general condition that H_0 may or may not be true. The test statistic uses the ratio of the maximized likelihoods, through

$$-2\log\left(\frac{\text{maximum likelihood when parameters satisfy } H_0}{\text{maximum likelihood when parameters are unrestricted}}\right).$$

The test statistic value is nonnegative. When H_0 is false, the ratio of maximized likelihoods tends to be far below 1; then, -2 times the log ratio tends to be a large positive number, more so as the sample size increases.

2.4.2 Likelihood-Ratio Statistic

The likelihood-ratio chi-squared statistic

For two-way contingency tables with likelihood function based on the multinomial distribution, the likelihood-ratio statistic is

$$G^2 = 2\sum n_{ij}\log(\frac{n_{ij}}{\mu_{ij}}).$$

This statistic is called the *likelihood-ratio chi-squared statistic*. Like the Pearson statistic, G^2 takes its minimum value of 0 when all $n_{ij} = \mu_{ij}$, and larger values provide stronger evidence against H_0.

2.4.2 Likelihood-Ratio Statistic

X^2 and G^2 provide separate test statistics, but they share many properties and usually provide the same conclusions. When H_0 is true and the expected frequencies are large, the two statistics have the same chi-squared distribution, and their numerical values are similar.

2.4.3 Test of Independence

The null hypothesis

In two-way contingency tables with joint probabilities $\{\pi_{ij}\}$ for two response variables, the null hypothesis of statistical independence is

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \text{for} \quad \text{all } i \quad \text{and } j.$$

To test H_0, we identify $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ as the expected frequency. Here, μ_{ij} is the expected value of n_{ij} assuming independence. Usually, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown, as is this expected value.

2.4.3 Test of Independence

Estimated expected frequencies

$$\hat{\mu}_{ij} = np_{i+}p_{+j} = n(\frac{n_{i+}}{n})(\frac{n_{+j}}{n}) = \frac{n_{i+}n_{+j}}{n}.$$

Test statistics

For testing independence in $I \times J$ contingency tables, the Pearson and likelihood ratio statistics equal

$$X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \qquad G^2 = 2\sum n_{ij}\log(\frac{n_{ij}}{\hat{\mu}_{ij}}).$$

Their large-sample chi-squared distributions have $df = (I-1)(J-1)$.

2.4.3 Test of Independence

Degree of freedom (df)

The *df* value means the following: under H_0, $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ determine the cell probabilities. There are $I-1$ nonredundant row probabilities. Because they sum to 1, the first $I-1$ determine the last one through $\pi_{I+} = 1-(\pi_{1+}+\cdots+\pi_{I-1,+})$. Similarly, there are $J-1$ nonredundant column probabilities. So, under H_0, there are $(I-1)+(J-1)$ parameters.

2.4.3 Test of Independence

Degree of freedom (df)

The alternative hypothesis H_a merely states that there is not independence. It does not specify a pattern for the IJ cell probabilities. The probabilities are then solely constrained to sum to 1, so there are $IJ - 1$ nonredundant parameters. The value for df is the difference between the number of parameters under H_a and H_0, or

$$df = (IJ - 1) - [(I - 1) + (J - 1)] = IJ - I - J + 1 = (I - 1)(J - 1).$$

2.4.4 Example: Gender Gap in Political Affiliation

Table 2.5 Cross Classification of Party Identification by Gender

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	762	327	468	1557
	(703.7)	(319.6)	(533.7)	
Males	484	239	477	1200
	(542.3)	(246.4)	(411.3)	
Total	1246	566	945	2757

Note: Estimated expected frequencies for hypothesis of independence in parentheses. Data from 2000 General Social Survey.

2.4.4 Example: Gender Gap in Political Affiliation

Example

Table 2.5, from the 2000 General Social Survey, cross classifies gender and political party identification. Subjects indicated whether they identified more strongly with the Democratic or Republican party or as Independents. Table 2.5 also contains estimated expected frequencies for H_0: independence. For instance, the first cell has $\mu_{11} = n_{1+}n_{+1}/n = (1557 \times 1246)/2757 = 703.7$.

How to conduct this test? Can we reject H_0?

2.4.5 Residuals for Cells in a Contingency Table

The standardized residual

Larger differences between n_{ij} and $\hat{\mu}_{ij}$ tend to occur for cells that have larger expected frequencies, so the raw difference $n_{ij} - \hat{\mu}_{ij}$ is insufficient. For the test of independence, a useful cell residual(standardized residual) is

$$\frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}}.$$

2.4.5 Residuals for Cells in a Contingency Table

When H_0 is true, each standardized residual has a large-sample standard normal distribution. A standardized residual having absolute value that exceeds about 2 when there are few cells or about 3 when there are many cells indicates lack of fit of H_0 in that cell. (Under H_0, we expect about 5% of the standardized residuals to be farther from 0 than ± 2 by chance alone.)

2.4.5 Residuals for Cells in a Contingency Table

Table 2.6 Standardized Residuals for Table 2.5

	Party Identification		
Gender	Democrat	Independent	Republican
Females	762 (4.50)	327 (0.70)	468 (-5.32)
Males	484 (-4.50)	239 (-0.70)	477 (5.32)

How to calculate these standardized residuals (in Parentheses)?

2.4.5 Residuals for Cells in a Contingency Table

Information from Table 2.6

- Large positive residuals for female Democrats and male Republicans.
- Large negative residuals for female Republicans and male Democrats.
- An odds ratio describes this evidence of a gender gap.
- For each political party, Table 2.6 shows that the residual for females is the negative of the one for males.

2.4.6 Partitioning Chi-Squared

- Chi-squared statistics sum and break up into other chi-squared statistics. $df = df_1 + df_2$.
- Chi-squared statistics having $df > 1$ can be broken into components with fewer degrees of freedom. A partitioning may show that an association primarily reflects differences between certain categories or groupings of categories.
- For testing independence in $2 \times J$ tables, $df = (J - 1)$ and a chi-squared statistic can partition into $J - 1$ components.
- The G^2 statistic has exact partitionings. The Pearson X^2 does not equal the sum of X^2 values for the separate tables in a partition.

2.4.7 Comments About Chi-Squared Tests

- Chi-squared tests of independence, like any significance test, have limitations. They merely indicate the degree of evidence for an association. They are rarely adequate for answering all questions we have about a data set.
- The X^2 and G^2 chi-squared tests also have limitations in the types of data sets for which they are applicable. To play safe, you can instead use a small-sample procedure whenever at least one expected frequency is less than 5.
- The $\{\hat{\mu}_{ij} = n_{i+}n_{+j}/n\}$ used in X^2 and G^2 depend on the row and column marginal totals, but not on the order in which the rows and columns are listed.

Outline

1. 2.1 Probability Structure For Contingency Tables
2. 2.2 Comparing Proportions in Two-by-Two Tables
3. 2.3 The Odds Ratio
4. 2.4 Chi-Square Tests of Independence
5. **2.5 Testing Independence for Ordinal Data**
6. 2.6 Exact Inference for Small Samples
7. 2.7 Association in Three-Way Tables
8. Homework 2

When the rows and/or the columns are ordinal, the chi-squared test of independence using test statistic X^2 or G^2 ignores the ordering information. Test statistics that use the ordinality by treating ordinal variables as quantitative rather than qualitative (nominal scale) are usually more appropriate and provide greater power.

When the variables are ordinal, a trend association is common. As the level of X increases, responses on Y tend to increase toward higher levels, or responses on Y tend to decrease toward lower levels.

2.5.1 Linear Trend Alternative to Independence

To detect a trend association, a simple analysis assigns scores to categories and measures the degree of *linear trend*. The test statistic, which is sensitive to positive or negative linear trends, utilizes correlation information in the data. Let $u_1 \leq u_2 \leq \cdots \leq u_I$ denote scores for the rows, and let $v_1 \leq v_2 \leq \cdots \leq v_J$ denote scores for the columns. The scores have the same ordering as the category levels.You should choose the scores to reflect distances between categories, with greater distances between categories regarded as farther apart.

2.5.1 Linear Trend Alternative to Independence

Sample Covariance

Let $\bar{u} = \sum_i u_i p_{i+}$ denote the sample mean of the row scores, and let $\bar{v} = \sum_j v_j p_{+j}$ denote the sample mean of the column scores. The sum $\sum_{ij}(u_i - \bar{u})(v_j - \bar{v})p_{ij}$ weights cross-products of deviation scores by their relative frequency. This is the *sample covariance* of X and Y. The correlation r between X and Y equals the covariance divided by the product of the sample standard deviations of X and Y. That is,

$$r = \frac{\sum_{ij}(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{[\sum_i (u_i - \bar{u})^2 p_{i+}][\sum_j (v_j - \bar{v})^2 p_{+j}]}}.$$

2.5.1 Linear Trend Alternative to Independence

Testing Independence

For testing H_0: independence against the two-sided $H_a : \rho \neq 0$, a test statistic is

$$M^2 = (n-1)r^2.$$

2.5.1 Linear Trend Alternative to Independence

Testing Independence

This test statistic increases as r increases in magnitude and as the sample size n grows. For large n, M^2 has approximately a chi-squared distribution with $df = 1$. Large values contradict independence, so, as with X^2 and G^2, the P-value is the right-tail probability above the observed value. The square root, $M = \sqrt{(n-1)}r$, has approximately a standard normal null distribution. It applies to one-sided alternative hypotheses, such as $H_a : \rho > 0$.

2.5.2 Example: Alcohol Use and Infant Malformation

Table 2.7 Infant Malformation and Mother's Alcohol Consumption

Alcohol	Malformation			Percentage	Standardized
Consumption	Absent	Present	Total	Present	residual
0	17066	48	17114	0.28	-0.18
< 1	14464	38	14502	0.26	-0.71
1-2	788	5	793	0.63	1.84
3-5	126	1	127	0.79	1.06
≥ 6	37	1	38	2.63	2.71

Source: B. I. Graubard and E. L. Korn, *Biometrics*, 43: 471-476, 1987. Reprinted with permission from the Biometric Society.

2.5.2 Example: Alcohol Use and Infant Malformation

Example

Table 2.7 refers to a prospective study of maternal drinking and congenital malformations. After the first 3 months of pregnancy, the women in the sample completed a questionnaire about alcohol consumption. Following childbirth, observations were recorded on the presence or absence of congenital sex organ malformations. Alcohol consumption, measured as average number of drinks per day, is an explanatory variable with ordered categories. Malformation, the response variable, is nominal.

2.5.2 Example: Alcohol Use and Infant Malformation

When a variable is nominal but has only two categories, statistics (such as M^2) that treat the variable as ordinal are still valid. For instance, we could artificially regard malformation as ordinal, treating “absent” as “low” and “present” as “high”. Any choice of two scores, such as 0 for “absent” and 1 for “present”, yields the same value of M^2.

2.5.2 Example: Alcohol Use and Infant Malformation

Information from Table 2.7

From Table 2.7, the percentage of malformation cases has roughly an increasing trend across the levels of alcohol consumption. The first two are similar and the next two are also similar. The sample percentages and the standardized residuals both suggest a possible tendency for malformations to be more likely at higher levels of alcohol consumption

2.5.2 Example: Alcohol Use and Infant Malformation

Test Using M^2

To use the ordinal test statistic M^2, we assign scores to alcohol consumption that are midpoints of the categories; that is, $v_1 = 0$, $v_2 = 0.5$, $v_3 = 1.5$, $v_4 = 4.0$, $v_5 = 7.0$, the last score being somewhat arbitrary. From PROC FREQ in SAS, the sample correlation between alcohol consumption and malformation is $r = 0.0142$. The test statistic $M^2 = (32,573)(0.0142)^2 = 6.6$ has P-value $= 0.01$, suggesting strong evidence of a nonzero correlation. The standard normal statistic $M = 2.56$ has P $= 0.005$ for $H_a : \rho > 0$.

2.5.3 Extra Power with Ordinal Tests

- For testing H_0: independence, X^2 and G^2 refer to the most general H_a possible, whereby cell probabilities exhibit *any* type of statistical dependence.
- When the row and column variables are ordinal, one can attempt to describe the association using a single extra parameter. For instance, M^2.
- When the association truly has a positive or negative trend, the ordinal test using M^2 has a power advantage over the tests based on X^2 or G^2.
- Another advantage of chi-squared tests having small *df* values relates to the accuracy of chi-squared approximations.

2.5.4 Choice of Scores

- For most data sets, the choice of scores has little effect on the results. Different choices of ordered scores usually give similar results. This may not happen, however, when the data are very unbalanced, such as when some categories have many more observations than other categories.
- An alternative approach assigns ranks to the subjects and uses them as the category scores. For all subjects in a category, one assigns the average of the ranks that would apply for a complete ranking of the sample from 1 to n. These are called *midranks*.

2.5.4 Choice of Scores

- The M^2 statistic using midrank scores for each variable is sensitive to detecting nonzero values of a rank correlation called **Spearman's rho**. Alternative ordinal tests for $I \times J$ tables utilize versions of other ordinal association measures. E.g. gamma and Kendall's tau-b.

2.5.5 Trend Tests for $I \times 2$ and $2 \times J$ Tables

Wilcoxon Test

When X is binary, the table has size $2 \times J$. Such tables occur in comparisons of two groups, such as when the rows represent two treatments. The M^2 statistic then detects differences between the two row means of the scores $\{v_j\}$ on Y . Small P-values suggest that the true difference in row means is nonzero. With midrank scores for Y, the test is sensitive to differences in mean ranks for the two rows. This test is called the *Wilcoxon* or *Mann - Whitney test.*

2.5.5 Trend Tests for $I \times 2$ and $2 \times J$ Tables

Cochran-Armitage Trend Test

How the proportion of "successes" varies across the levels of X. For the chosen row scores, M^2 detects a linear trend in this proportion and relates to models presented in Section 3.2.1. Small P-values suggest that the population slope for this linear trend is nonzero.

2.5.6 Nominal- Ordinal Tables

The M^2 test statistic treats both classifications as ordinal. When one variable (say X) is nominal but has only two categories, we can still use it. When X is nominal with more than two categories, it is inappropriate. One possible test statistic finds the mean response on the ordinal variable (for the chosen scores) in each row and summarizes the variation among the row means. The statistic, which has a large-sample chi-squared distribution with $df = (I - 1)$, is rather complex computationally.We defer discussion of this case to Section 6.4.3. When $I = 2$, it is identical to M^2.

Outline

1 2.1 Probability Structure For Contingency Tables

2 2.2 Comparing Proportions in Two-by-Two Tables

3 2.3 The Odds Ratio

4 2.4 Chi-Square Tests of Independence

5 2.5 Testing Independence for Ordinal Data

6 2.6 Exact Inference for Small Samples

7 2.7 Association in Three-Way Tables

8 Homework 2

2.6 Exact Inference for Small Samples

The confidence intervals and tests presented so far in this chapter are large-sample methods. As the sample size n grows, "chi-squared" statistics such as X^2, G^2, and M^2 have distributions that are more nearly chi-squared. When n is small, one can perform inference using exact distributions rather than large-sample approximations.

2.6.1 Fisher's Exact Test for 2×2 Tables

For 2×2 tables, independence corresponds to an odds ratio of $\theta = 1$. Suppose the cell counts $\{n_{ij}\}$ result from two independent binomial samples or from a single multinomial sample over the four cells. A small-sample null probability distribution for the cell counts that does not depend on any unknown parameters results from considering the set of tables having the same row and column totals as the observed data. Once we condition on this restricted set of tables, the cell counts have the *hypergeometric distribution*.

2.6.1 Fisher's Exact Test for 2×2 Tables

When $\theta = 1$, the probability of a particular value n_{11} equals

$$P(n_{11}) = \frac{C_{n_{1+}}^{n_{11}} C_{n_{2+}}^{n_{+1}-n_{11}}}{C_{n}^{n_{+1}}}.$$

2.6.1 Fisher's Exact Test for 2×2 Tables

Fisher's Exact Test

To test H_0: independence, the P-value is the sum of hypergeometric probabilities for outcomes at least as favorable to H_a as the observed outcome. For $H_a : \theta > 1$. Given the marginal totals, tables having larger n_{11} values also have larger sample odds ratios $\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21})$; hence, they provide stronger evidence in favor of this alternative. The P-value equals the right-tail hypergeometric probability that n_{11} is at least as large as the observed value. This test, proposed by the eminent British statistician R. A. Fisher in 1934, is called *Fisher's exact test*.

2.6.2 Example: Fisher's Tea Taster

Four cups had milk added first, and the other four had tea added first. She was told there were four cups of each type and she should try to select the four that had milk added first. The cups were presented to her in random order.

Table 2.8 Fisher's Tea Tasting Experiment

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	

2.6.2 Example: Fisher's Tea Taster

Test

$$H_0 : \theta = 1, \quad H_a : \theta > 1.$$

For this experimental design, the column margins are identical to the row margins $(4, 4)$, because she knew that four cups had milk added first. Both marginal distributions are naturally fixed.

2.6.2 Example: Fisher's Tea Taster

The null distribution of n_{11} is the hypergeometric distribution defined for all 2×2 tables having row and column margins $(4, 4)$. The potential values for n_{11} are $(0, 1, 2, 3, 4)$. The observed table, three correct guesses of the four cups having milk added first, has probability

$$P(3) = \frac{C_4^3 C_4^1}{C_8^4} = 0.229$$

P-value$= P(3) + P(4) = 0.243.$

2.6.2 Example: Fisher's Tea Taster

Hypergeometric Distribution for Tables

n_{11}	Probability	P-value	X^2
0	0.014	1.000	8.0
1	0.229	0.986	2.0
2	0.514	0.757	0.0
3	0.229	0.243	2.0
4	0.014	0.014	8.0

Note: P-value refers to right-tail hypergeometric probability for one-sided alternative.

2.6.3 P-values and Conservatism for Actual P(Type I Error)

- The two-sided alternative $H_a : \theta \neq 1$ is the general alternative of statistical dependence, as in chi-squared tests. Its exact P-value is usually defined as the two-tailed sum of the probabilities of tables no more likely than the observed table.
- For Table 2.8, summing all probabilities that are no greater than the probability $P(3) = 0.229$ of the observed table gives $P = P(0) + P(1) + P(3) + P(4) = 0.486$. When the row or column marginal totals are equal, the hypergeometric distribution is unimodal and symmetric, and the two-sided P-value doubles the one-sided one.

2.6.3 P-values and Conservatism for Actual P(Type I Error)

Conservative

For the one-sided alternative, when H_0 is true, the probability of this outcome is 0.014. So, P(type I error) $= 0.014$, not 0.05. The test is *conservative*, because the actual error rate is smaller than the intended one.

2.6.3 P-values and Conservatism for Actual P(Type I Error)

- To diminish the conservativeness, we recommend using the mid P-value.
- It is more common that only one set is fixed, such as when rows totals are fixed with independent binomial samples. Then, alternative exact tests are *unconditional*, not conditioning on the other margin. They are less conservative than Fisher's exact test.
- Exact tests of independence for tables of size larger than 2×2 use a multivariate version of the hypergeometric distribution. Such tests are not practical to compute by hand or calculator but are feasible with software.

Outline

1. 2.1 Probability Structure For Contingency Tables
2. 2.2 Comparing Proportions in Two-by-Two Tables
3. 2.3 The Odds Ratio
4. 2.4 Chi-Square Tests of Independence
5. 2.5 Testing Independence for Ordinal Data
6. 2.6 Exact Inference for Small Samples
7. 2.7 Association in Three-Way Tables
8. Homework 2

2.7 Association in Three-Way Tables

Example

To analyze whether passive smoking is associated with lung cancer, a cross-sectional study might compare lung cancer rates between nonsmokers whose spouses smoke and nonsmokers whose spouses do not smoke. In doing so, the study should attempt to control for age, socioeconomic status, or other factors that might relate both to whether one's spouse smokes and to whether one has lung cancer. A statistical control would hold such variables constant while studying the association.

2.7 Association in Three-Way Tables

Example

Without such controls, results will have limited usefulness. Suppose that spouses of nonsmokers tend to be younger than spouses of smokers and that younger people are less likely to have lung cancer. Then, a lower proportion of lung cancer cases among nonsmoker spouses may merely reflect their lower average age and not an effect of passive smoking.

2.7.1 Partial Tables

Partial Tables

Two-way cross-sectional slices of the three-way table cross classify X and Y at separate levels of Z. These cross sections are called *partial tables*. They display the XY relationship at fixed levels of Z, hence showing the effect of X on Y while controlling for Z. The partial tables remove the effect of Z by holding its value constant.

2.7.1 Partial Tables

Marginal Table

The two-way contingency table that results from combining the partial tables is called the XY *marginal table*. Each cell count in it is a sum of counts from the same cell location in the partial tables. The marginal table contains no information about Z, so rather than controlling Z, it ignores Z.

2.7.2 Conditional Versus Marginal Associations: Death Penalty Example

Conditional Associations

The associations in partial tables are called *conditional associations*, because they refer to the effect of X on Y conditional on fixing Z at some level.

2.7.2 Conditional Versus Marginal Associations: Death Penalty Example

Table 2.10 Death Penalty Verdict by Defendant's Race and Victims' Race

Victims'	Defendant's	Death Penalty		Percentage
Race	Race	Yes	No	Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

2.7.2 Conditional Versus Marginal Associations: Death Penalty Example

Example

Table 2.10 is a $2 \times 2 \times 2$ contingency table-two rows, two columns, and two layers - from an article that studied effects of racial characteristics on whether subjects convicted of homicide receive the death penalty. The 674 subjects were the defendants in indictments involving cases with multiple murders, in Florida between 1976 and 1987. The variables are Y = death penalty verdict, having categories (yes, no), and X = race of defendant and Z = race of victims, each having categories (white, black).

2.7.2 Conditional Versus Marginal Associations: Death Penalty Example

Example

We study the effect of defendant's race on the death penalty verdict, treating victims race as a control variable. Table 2.10 has a 2×2 partial table relating defendant's race and the death penalty verdict at each level of victims' race.

2.7.2 Conditional Versus Marginal Associations: Death Penalty Example

For each combination of defendant's race and victims' race, Table 2.10 lists and Figure 2.3 displays the percentage of defendants who received the death penalty.

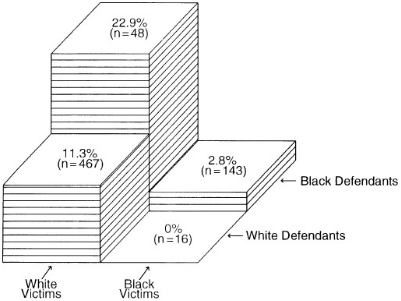


Figure 2.3. Percentage receiving death penalty, by defendant's race and victims' race.

2.7.2 Conditional Versus Marginal Associations: Death Penalty Example

- We use these to describe the conditional associations between defendant's race and the death penalty verdict, controlling for victims' race. *Controlling* for victims' race by keeping it fixed, the percentage of "yes" death penalty verdicts was higher for black defendants than for white defendants.
- The bottom portion of Table 2.10 displays the marginal table for defendant's race and the death penalty verdict. *Ignoring* victims' race, the percentage of "yes" death penalty verdicts was lower for black defendants than for white defendants. The association reverses direction compared with the partial tables.

2.7.3 Simpson's Paradox

The result that a marginal association can have different direction from the conditional associations is called *Simpson's paradox.*

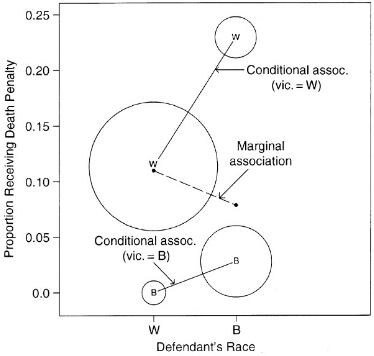


Figure 2.4. Proportion receiving death penalty by defendant's race, controlling and ignoring victims' race.

2.7.4 Conditional and Marginal Odds Ratios

Conditional Odds

Conditional associations, like marginal associations, can be described using odds ratios.We refer to odds ratios for partial tables as *conditional odds ratios*. For binary X and Y , within a fixed level k of Z, let $\theta_{XY(k)}$ denote the odds ratio between X and Y computed for the true probabilities.

From Table 2.10,$\hat{\theta}_{XY(1)} = \frac{53\times37}{414\times11} = 0.43$, $\hat{\theta}_{XY(2)} = 0$.
The conditional odds ratios can be quite different from the marginal odds ratio, for which the third variable is ignored rather than controlled.

2.7.5 Conditional Independence Versus Marginal Independence

If X and Y are independent in each partial table, then X and Y are said to be *conditionally independent, given* Z. All conditional odds ratios between X and Y then equal 1. Conditional independence of X and Y, given Z, does not imply marginal independence of X and Y.

2.7.5 Conditional Independence Versus Marginal Independence

Table 2.11 Conditional Independence Does not Imply Marginal Independence

		Response	
Clinic	Treatment	Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
Total	A	20	20
	B	20	40

2.7.5 Conditional Independence Versus Marginal Independence

- The expected frequencies in Table 2.11 show a hypothetical relationship among three variables: Y = response (success, failure), X = drug treatment (A,B), and Z = clinic(1, 2). The conditional odds ratios between X and Y at the two levels of Z are

$$\theta_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1.0, \quad \theta_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1.0$$

  Given clinic, response and treatment are conditionally independent.

- The marginal table adds together the tables for the two clinics. The odds ratio for that marginal table equals $(20 \times 40)/(20 \times 20) = 2.0$, so the variables are not marginally independent.

2.7.6 Homogeneous Association

Homogeneous Association

Let K denote the number of categories for Z. When X and Y are binary, there is *homogeneous* XY *association* when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}.$$

Conditional independence of X and Y is the special case in which each conditional odds ratio equals 1.0.

2.7.6 Homogeneous Association

- In an $I \times J \times K$ table, homogeneous XY association means that any conditional odds ratio formed using two levels of X and two levels of Y is the same at each level of Z.
- When there is homogeneous XY association, there is also homogeneous XZ association and homogeneous YZ association. Homogeneous association is a symmetric property, applying to any pair of the variables viewed across the levels of the third.
- When it occurs, there is said to be *no interaction* between two variables in their effects on the third variable.

Outline

1. 2.1 Probability Structure For Contingency Tables
2. 2.2 Comparing Proportions in Two-by-Two Tables
3. 2.3 The Odds Ratio
4. 2.4 Chi-Square Tests of Independence
5. 2.5 Testing Independence for Ordinal Data
6. 2.6 Exact Inference for Small Samples
7. 2.7 Association in Three-Way Tables
8. Homework 2

Homework

1. Analysis the GDS5037 data.

   (1) Download the data from the course web (SPOC XMU) and read the file. Suppose the samples are randomly chosen.

   (2) For all samples, there are 3 categories for patients' status: mild asthma (MMA), control, severe asthma (SA). Calculate the frequency and percentage of each category and plot a pie chart for the percentages.

   (3) Plot the frequency bar chart for the 3 categories.

   (4) Classify patients into 3 groups according to patients' status: MMA, control, SA. Calculate the sample mean and variance of IDENTIFIER "VPS39" in each group.

Homework

- (5) In the severe asthma (SA) group, calculate the proportion of male and female, respectively.

  (6) Construct a 2×2 table for gender and status (SA and control). Calculate the sample odds ratio. Test the independence between gender and status.

  (4) Construct a 2×3 table for gender and status (mild asthma-MMA, SA and control). Test the independence between gender and status.

2. Problems in textbook 2.13, 2.16, 2.27, 2.29, 2.33.

Thank you!