Categorical Data Analysis Note
Edgard Maboudou

Chapter 2: Two-way Contingency Tables

I Introduction

- Bivariate Analysis: Suppose X and Y are 2 categorical variables – two way table.
  X has I categories or levels i.e. X takes on values $1, 2, \ldots, I$;
  Y has J categories or levels i.e. Y takes on values $1, 2, \ldots, J$;
- There are IJ cells in a cross-classification of X and Y.
- X is the row variable, which is indexed by i.
- Y is the column variable, which indexed by j.
- Display the IJ possible combinations of outcomes in a rectangular table having I rows for the categories of X and J columns for the categories of Y.
- A table of this form in which the cells contain frequency counts of outcomes is called a contingency table.
- A contingency table that cross classifies two variables is called a two-way table.
- A table which cross classifies three variables is called a three-way table.
- A “2-way contingency table” is a cross-classification of observations by the levels of 2 discrete variables.
- The cells of the table contain frequency counts.
- The number of variables is often referred to as the “dimension of the table”.
- The “size” of the table often refers to the number of cells.
- A two-way table having I rows and J columns is called an $I \times J$ table.
- The size of a two-way table is $I \times J$.
- Focus for now on a two-way table

- In some situations, Y is a response variable and X is an explanatory variable.
- In other situations, both are response variables.

1. Joint, Marginal, and Conditional Distributions

- Notation: Joint probability $\pi_{ij} = P(X = i, Y = j)$. This is the probability that (X, Y) falls in the cell in row i and column (j).

- The probabilities $\{\pi_{ij}\}$ form the joint distribution of X and Y. Note that ,
$$\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij} = 1.$$
- The marginal distribution of X is π_{i+}, which is obtained by the row sums or the sum of cell probabilities across the rows, that is,
$$\pi_{i+} = P(X = i) = \sum_{j=1}^{J}\pi_{ij}$$
- The marginal distribution of Y is π_{+j}, which is obtained by the column sums or the sum of cell probabilities across the columns, that is,
$$\pi_{+j} = P(Y = j) = \sum_{i=1}^{I}\pi_{ij}$$
- Cell counts are denoted by $\{n_{ij}\}$, with
$$n = \sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij}.$$
- Cell proportions are
$$p_{ij} = \frac{n_{ij}}{n}$$
- This is the proportion of observations in the $(i, j)^{th}$ cell.
- The marginal frequencies are row totals $\{n_{i+}\}$ and column totals $\{n_{+j}\}$

- Let Y be a binary response variable and X be an explanatory variable, it is informative to construct separate probability distributions for Y at each level of X,
- i.e. we would be interested in the conditional probability of Y given X: $\pi_{j|i} = P(Y = j|X = i) = \pi_i$ and is called a conditional distribution.
- If Y is the response variable and X is the explanatory variable, we would be interested in the conditional probability of Y given X: $\pi_{j|i} = P(Y = j|X = i) = \pi_i$
- Corresponding sample proportions are denoted using p. Example: p_{ij} for π_{ij}, p_{i+} for π_{i+}, $p_{j|i} = p_i$ for $\pi_{j|i} = \pi_i$
- Corresponding cell counts on frequencies are n_{ij}, n_{i+}. For instance, for 2×2 table, we would have:

$X \setminus Y$	1	2	$\rightarrow \sum$	Prop.
1	n_{11}	n_{12}	n_{1+}	$p_{11} = \frac{n_{11}}{n}$
2	n_{21}	n_{22}	n_{2+}	$p_{1+} = \frac{n_{1+}}{n}$
$\sum \downarrow$	n_{+1}	n_{+2}	n	

- Divide any cell by n to get corresponding proportion.

2. Example

2013 workers were classified according to whether or not they have a stressful job and whether or not they develop coronary heart disease (CHD).

Stress\CHD	Y	N	
Y	97	307	404
N	200	1409	1609
	297	1716	2013

Solution: First, divide by n to get the sample proportion

$X \setminus Y$	Y	N	
Y			
N			

- Here, "CHD" would be the response variable and "Stress" the explanatory variable.

- So, we would be interested in the conditional distribution of CHD given stress
- Estimate of $P(CHD = 1|Stress = 1) = \pi_1$

- Estimate of $P(CHD = 1|Stress = 2) = \pi_2$

- The difference in these proportions may suggest that $\pi_1 \neq \pi_2$.
- This would mean that CHD and Stress are dependent.
- Equivalently, we can compare their joint probability to the product of the marginal probabilities

$$\pi_{ij} = \pi_{i+}\pi_{+j} \ \forall i, j \quad \Leftrightarrow \quad \text{independent}$$

3. Independence

- Two variables are statistically independent if all joint probabilities equal the product of their marginal probabilities

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

- or Conditional distributions of Y are identical at each levels of X,

$$\pi_{j|i} = \pi_{+j} \ \forall i, j$$

.

II Sampling Designs (2×2 Table)

These are extensions of the Poisson, Binomial, and multinomial models that we have discussed for 1 variable, in chapter 1, to 2 variables.

1. **Poisson Sampling**
   - No margins of a table are fixed by design. Each cell is considered an independent Poisson random variable.
   - Each cell contains a frequency over a period of time.
   - the n_{ij}'s are independent Poisson random variables.

2. **Independent Binomial Sampling**
   - Independent samples from each level of X
   - One margin is fixed by design while the other is free to vary. Classified according to level of Y. Thus, marginal totals are fixed, i.e. n_{1+} and n_{2+} are fixed.
   - Conditional distributions of Y at each level of X are binomial.
   - note that we can estimate the conditional distribution of Y given X, but not the joint distribution of X and Y.

3. **Multinomial Sampling**
   - the total number of observations, n, is fixed by design but not the row or column totals and they are classified according to the 2 variables.
   - The margins are free to vary

4. **Pseudo-Independent Binomial Sampling**
   - When one variable is considered the response and the other variable is considered the explanatory variable, but only the total n is fixed by design,
   - we may want to treat the data as if it were independent binomial samples.

5. **Analysis**
   - Most analysis do not depend on which sampling scheme was used.
   - When one variable is considered the response and the other variable is considered the explanatory variable, but only the total n is fixed by design, we may want to treat the data as if it were independent binomial samples.

- Different sampling models usually lead to the same inferential methods.
- Importance of Considering Sampling Design: sampling and design do make a difference regarding conclusions that can be made.

III. Measuring Association in 2×2 Tables

- Ways to study and analyze the relationship between two variables.
- Multiple ways to do measure association:
  1. Differences of Proportions
  2. Relative risk
  3. Odds Ratios

1. Differences of proportions – independent binomial sampling – compare conditional probabilities

- Assume that the row totals are fixed and hence we have a binomial model.
- Suppose the two categories of Y are success and failure.
- Let $\pi_{1|1} = \pi_1 =$ Probability of "success" given row 1
- $\pi_{2|1} = 1 - \pi_{1|1} = 1 - \pi_1 =$ probability of "failure" given row 1
- and $\pi_{1|2} = \pi_2 =$ Probability of "success" given row 2.
- $\pi_{2|2} = 1 - \pi_{1|2} = 1 - \pi_2 =$ probability of "failure" given row 2
- These are conditional probabilities.
- The difference in probabilities $\pi_1 - \pi_2$ compares the success probabilities in the two rows.
- In this setting, we want to compare the conditional probabilities

$X \setminus Y$	1	2
1	π_1	$1 - \pi_1$
2	π_2	$1 - \pi_2$

- If X and Y are independent, then $\pi_1 = \pi_2$ and $\pi_1 - \pi_2 = 0$. We compare π_1 and π_2 (test or CI).

- Standard Inference for 2 Populations: $H_0 : \pi_1 = \pi_2$ v.s. $H_0 : \pi_1 \neq \pi_2$
- Let p_1 and p_2 be sample proportions of success for the two rows.
- The sample difference $p_1 - p_2$ estimates $\pi_1 - \pi_2$.
- Let's denote n_{1+} and n_{2+} by n_1 and n_2 respectively.
- If the counts in two rows are independent samples, the estimated standard error of $\pi_1 - \pi_2$ is

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

- For example, a large sample $(1-\alpha) \times 100\%$ CI for $\pi_1 - \pi_2$ is

$$p_1 - p_2 \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

  where $z_{\alpha/2}$ denotes the standard normal percentile having a right tail probability equals to $\alpha/2$.

Example: A survey was conducted to examine the attitude of males and females about abortion. Of 500 females, 309 supported legalized abortion. Of 600 males, 319 supported legalized abortion. Let $Y = 1$ be "supported legalized abortion".

$X \setminus Y$	1	2	
F	309	191	500
M	319	281	600
	628	472	1100

Solution

Solution continued

Method 2: Ratio of Proportion – Relative Risk (R.R.)

- In 2×2 tables, the relative risk of a "success"is the ratio of the success probabilities for the two groups

$$R.R. = \frac{\pi_1}{\pi_2}.$$

- Why it might be a good idea to use $R.R.$ rather than Z-test?

- A difference between two proportions of a certain fixed size may have greater importance when both proportions are near 0 or 1 than when they are near the middle of the range.

- e.g. the difference between 0.010 and 0.001 is the same as the difference between 0.410 and 0.401, namely 0.009 but the former one may be more important than the later one.

- Examples of such cases include a comparison of drugs on the proportion of subjects who have adverse reactions when using the drug.
- $R.R.$ is helpful with small probabilities.
- When $\frac{\pi_1}{\pi_2} = 1$, the response is independent of the groups. Conditional probability equals marginal probability.
- The sample relative risk is $\widehat{R.R.} = \frac{p_1}{p_2}$.
- Its distribution (of $\frac{p_1}{p_2}$) is heavily skewed and cannot be well approximated by the normal distribution, unless the sample sizes are quite large.
- The log of the relative risk has a sampling distribution that is approximately normal with variance

$$\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$$

- This permits the construction of a confidence interval (CI) which is symmetric around $\log(RR)$.
- A $(1-\alpha) \times 100\%$ CI of $\log\left(\frac{\pi_1}{\pi_2}\right) = \ln\left(\frac{\pi_1}{\pi_2}\right)$ is

$$\log\left(\frac{p_1}{p_2}\right) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}$$

- Hence, a large sample $(1-\alpha) \times 100\%$ confidence interval of $\frac{\pi_1}{\pi_2}$ is given by

$$\exp\left\{\log\left(\frac{p_1}{p_2}\right) \pm z_{\alpha/2}\sqrt{\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}}\right\}$$

Example

In a study, 140 individuals were given a placebo while 139 were given a daily dose of ascorbic acid (Vitamin C). For each individual, it was determined whether or not they developed a cold sometime during winter season.

$X \setminus Y$	Cold	No Cold	
Placebo	31	109	140
Vitamin C	17	122	139
	48	231	279

Solution

Method 3: Odds Ratio

1- Odd of Success – Odds Ratio

- Assume a binary variable, within row 1, the odds of success for population 1 is:

$$\Omega_1 = \frac{\pi_1}{1-\pi_1} = \frac{P(Y=1|X=1)}{P(Y=2|X=1)} = \frac{P(S)}{P(F)}$$

- Similarly, within row 2, the odds of success for population 2 is:

$$\Omega_2 = \frac{\pi_2}{1-\pi_2} = \frac{P(Y=1|X=2)}{P(Y=2|X=2)}$$

- Note: If we know Ω_i, we can compute π_i since

- Odds are non-negative and values greater than 1 indicates a success is more likely than a failure.
- $\Omega = 1 \Longleftrightarrow$ success and failure equally likely
- $\Omega > 1 \Longleftrightarrow$ success more likely than failure
- $\Omega < 1 \Longleftrightarrow$ failure more likely than success
- A common measure of association is the odds ratio

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

- In a 2×2 table,

2- Properties of θ

(i) Odds ratios are non-negative i.e. $\theta \in [0, +\infty)$

(ii) When X and Y are independent, conditional distributions of Rows 1 and 2 are same, that is, $\pi_1 = \pi_2$ and this implies, $\theta = 1$.

(iii) If $1 < \theta < +\infty$, the odds of success are higher in row 1 than in row 2.

(iv) If $0 < \theta < 1$, the odds of success are less likely in row 1 than in row 2.

- Values of θ farther from 1 (too small or too large) in a given direction indicates stronger level of association.
- If the order of the rows or the order of the columns is reversed (but not both), the new value of θ is the inverse of the original value.
- This ordering is usually arbitrary, so whether we get $\theta = 4.0$ or 0.25 is simply a matter of how we label the rows and columns.

3- Interpretation of $\theta = 2$

4- More on the odds ratio

- Recall that $\pi_1 = P(Y = 1|X = 1)$ and $\pi_2 = P(Y = 1|X = 2)$.

$$\pi_1 = P(Y = 1|X = 1) = \frac{P(Y = 1, X = 1)}{P(X = 1)} = \frac{\pi_{11}}{\pi_{1+}}$$

$$1 - \pi_1 = \frac{\pi_{1+} - \pi_{11}}{\pi_{1+}} = \frac{\pi_{12}}{\pi_{1+}} = P(Y = 2|X = 1).$$

  Similarly, $1 - \pi_2 = \frac{\pi_{2+} - \pi_{21}}{\pi_{2+}} = \frac{\pi_{22}}{\pi_{2+}} = P(Y = 2|X = 2)$.

- The odds ratio

$$\theta = \frac{\pi_1(1 - \pi_2)}{\pi_2(1 - \pi_1)} = \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}.$$

- That is θ can be computed directly from the joint distribution.

  1. As the odds ratio treats the variables symmetrically, it is unnecessary to identify one classification as a response variable to calculate it.
  2. It does not depend on the choice of a response and explanatory variables. If you switch X and Y, θ is the same.
  3. When both variables are responses, the odds ratio can be defined using the joint probability as

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

     and called cross - product ratio.
  4. It can be computed from the conditional probability $X|Y$.

5. Remark. Since $p_{ij} = \frac{n_{ij}}{n}$, the sample odds ratio reduces to

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

5- Inference for Odds Ratio

- For small to moderate sample size, the distribution of sample odds ratio $\widehat{\theta}$ is highly skewed.
- So, consider the log odds ratio, $\log\theta$
- X and Y are independent implies $\log\theta = 0$.
- Log odds ratio is symmetric about zero in the sense that reversal of rows or reversal of columns changes its sign only.
- The sample log odds ratio, $\log\widehat{\theta}$ has a less skewed distribution and can be approximated by the normal distribution well.
- The asymptotic standard error of $\log\widehat{\theta}$ is given by

$$ASE(\log\widehat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

- We can get a large sample CI for $\log\theta$ using the following result

$$\log\widehat{\theta} \pm z_{\alpha/2} ASE(\log\widehat{\theta})$$

- A large sample $(1-\alpha)\times 100\%$ confidence interval for θ is:

$$\exp\left\{\log\widehat{\theta} \pm z_{\alpha/2} ASE(\log\widehat{\theta})\right\}$$

- Note: The notation “log” means “natural logarithm”.

Example

Back to Vitamin C example:

$X \setminus Y$	Cold	No Cold	
Placebo	31	109	140
Vitamin C	17	122	139
	48	231	279

- $\widehat{\theta} =$

- **Interpretation**:

- A 90% CI for θ is

- **Conclusion**:

Some Observations

- Recall the formula for sample odds ratio

$$\widehat{\theta} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$

- The sample odds ratio is 0 or 1 if any $n_{ij} = 0$ and it is undefined if both entries in a row or column are zero.
- Consider the slightly modified formula

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

- In the ASE formula also, n_{ij}'s are replaced by $n_{ij}+0.5$.
- A sample odds ratio equals to 1.832 does not mean that p_1 is 1.832 times p_2.
- A simple relation:

$$\theta = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)} = R.R. \times \frac{1-\pi_2}{1-\pi_1}$$

- If p_1 and p_2 are close to 0, the odds ratio and relative risk take similar values, i.e.

$$\theta = R.R. \times \frac{1-\pi_2}{1-\pi_1} \approx R.R.$$

- This relationship between odds ratio and relative risk is useful.

IV. Types of Studies

Y: response and X: explanatory variable.

1. **Cross-sectional design**.
   — Take a sample from the population of interest and record which group a person falls into and the outcome of interest.
   — Fix n and the observations are classified according to both variables.
   — Can estimate joint probability and consequently conditional probability $Y|X$.

2. **Prospective design** or "look into the future".
   — Take a sample, wait some period of time, then count the number of outcomes/events/attributes of interest.
   — There are 2 kinds of prospective studies: Clinical trials and Cohort Studies
   — Clinical trials (experiments): Subjects are randomly assigned to groups.

— Cohort study: Subjects make their own choice as to which group they belong or "come as they are".
— Fixed row sums, n_{1+} and n_{2+}, that is sampling from the 2 levels of Y.
— Can estimate conditional distribution of $Y|X$, but not the joint distribution.

3. **Retrospective design** or "look into the past".
— Fixed column sums, n_{+1} and n_{+2}, that is sampling from the 2 levels of X.
— Sample those with and those without attribute of interest.
— Used to ensure that you have enough cases for events that are relatively rare in the population.
— Can estimate conditional distribution of $X|Y$.

- Odds ratio can be estimated for all 3 types of design.

- R.R. is computed from the conditional distribution of $Y|X$.

- In general, we cannot get the conditional distribution from a retrospective study.

- However, $\theta =$

Example

49 women aged 50-59 at diagnosis of cervical cancer are compared to 310 controls.

		disease status		
		Cancer	Control	
age at 1st	≤ 25	42	203	245
pregnancy	> 25	7	107	114
		49	310	359

- This is a retrospective study.

- Let X = Age (explanatory variable) and Y = Disease status (response variable). We are interested in comparing $P(\text{cancer } | X \leq 25)$ to $P(\text{cancer } | X > 25)$.
- We have a retrospective design $\Longleftrightarrow$ these conditional probabilities cannot be estimated.
- However, it is known that cervical cancer is a fairly rare disease ($\theta \approx R.R.$).

V. Goodness-of-Fit Test ($I \times J$ Table)

- Consider a null hypothesis, H_0, regarding the probability structure of this table.
- Let μ_{ij} be the expected cell frequency for the ij-th cell when H_0 is true ($\mu_{ij} = n\pi_{ij}$).
- The Pearson Chi-Square statistic for testing H_0 is

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}.$$

- If n is large and H_0 is true, $X^2 \sim \chi^2_{IJ-1-t}$ where t = # of underlying parameters that are needed to be estimated in getting estimates of μ_{ij}.

Example 1

A random sample of 100 observations is classified according to 2 variables X and Y. Suppose we wish to test $H_0 : \pi_{11} = .1, \pi_{12} = .15, \pi_{21} = .25, \pi_{22} = .5$.

$X \setminus Y$	1	2	
1	12	16	
2	29	43	
			100

Solution

Example 2

A random sample of 100 observations is classified according to 2 variables X and Y. Suppose we wish to test $H_0 : \pi_{11} = 2\pi_{21}, \pi_{12} = 2\pi_{22}$.

$X \setminus Y$	1	2	
1	25	42	
2	15	18	
	40	60	100

Solution

Solution continued

VI. Test of Independence ($I \times J$ Table)

- Are X and Y related?
  $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j (Joint = product of marginal).
- $\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$ and μ_{ij} can be estimated by $\widehat{\mu}_{ij} = np_{i+}p_{+j} =$
- So, the Pearson Chi-Squared statistic is

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - \widehat{\mu}_{ij})^2}{\widehat{\mu}_{ij}}.$$

- For "large" samples, X^2 has an approximate chi-squared distribution.
- A good rule: "Large" means $\mu_{ij} \geq 5$ for all (i, j).
- For the null hypothesis, need to estimate $I - 1$ π_{i+}'s and $J - 1$ π_{+j}'s, so $(I - 1) + (J - 1)$ parameters
- For the alternative hypothesis, need to estimate $IJ - 1$ parameters
- Hence, $t = (I - 1) + (J - 1)$, and then $df = IJ - 1 - t = (I - 1)(J - 1)$.
- An alternative test statistic is the likelihood ratio statistic, defined as

$$G^2 := 2 \sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{\widehat{\mu}_{ij}}\right),$$

- Like X^2, $G^2 \sim \chi^2_{(I-1)(J-1)}$ for large n.

Example

Rats were injected with a drug that cause breast cancer, then each rat was fed a controlled diet for 15 weeks. At the end of the feeding period, each rat was checked for cancer. Is the development of cancer related to diet?

Cancer \ Diet	HF wo. Fiber	HF w. Fiber	LF wo. Fiber	LF w. Fiber	
Y	27	20	19	14	80
N	3	10	11	16	40
	30	30	30	30	120

Are cancer and diet dependent, then test for independence.
$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j. Test for independence.

Method 1. Use Pearson's Chi-Squared Statistic

$X^2 =$

Method 2. Likelihood Ratio Statistic

$G^2 =$

VII. Understanding Dependence – Residuals for Cells in a Contingency Table

- Reject test of independence ⇒ X and Y are related.
- Can better understand this relationship by looking at residuals, and partitioning our Chi-Squared statistic into pieces.

1. Residuals

- The residuals are $n_{ij} - \widehat{\mu}_{ij}$
- Problem: These tend to be large when $\widehat{\mu}_{ij}$ is large.
- Pearson Residuals or often called "standardized residual,"

$$\frac{n_{ij} - \widehat{\mu}_{ij}}{\sqrt{\widehat{\mu}_{ij}}}$$

- Problem with Pearson Residuals: The variance (standard deviation) of Pearson residuals is a bit too small.
- The standardized adjusted residuals for the (i, j)th cell is

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

- Approximately standard normal $N(0, 1)$ for large n if the null hypothesis is true.
- Standardized adjusted residuals far from zero (say 2 or 3 units) correspond to cells that exhibit lack of independence.

Ex. Previous example. We look at the standardized adjusted residual for $(1, 1)$th cell

$$\frac{27 - 20}{\sqrt{20(1 - \frac{80}{120})(1 - \frac{30}{120})}} \approx 3.14$$

Cancer \ Diet	HF/NF	HF/F	LF/NF	LF/F	
Y					0
N					0
	0	0	0	0	

Some Comments

Reduced Table

Cancer \ Diet	HF/NF	LF/F	
Y			
N			
	30	30	

2. Partitioning Chi-Squares

- Another way to investigate the nature of association
- The sum of independent chi-squared statistics are themselves chi-squared statistics with degrees of freedom equal to the sum of the degrees of freedom for the individual statistics.
- $\chi^2_d = \chi^2_{d_1} + \chi^2_{d_2} + \cdots + \chi^2_{d_r}$ where $d = d_1 + d_2 + \cdots + d_r$ and $\chi^2_{d_i}$ are independent.
- "Partitioning chi-squared" uses this fact, but in reverse:

- We start with a chi-squared statistic with $df > 1$ and break it into component parts, each with $df = 1$

- This works with G^2 exactly but only approximately with X^2.

- Why partition?

- Partitioning chi-squared statistics helps to show that an association which was significant for the overall table primarily reflects differences between some categories and/or some groups of categories.

Ex. Partition G^2 into 3 parts $G^2 = G_1^2 + G_2^2 + G_3^2$.

– **How to partition?**

For G_1^2, we use only the first 2 columns.

$X \setminus Y$	HF/NF	HF/F	
Y			47
N			13
	30	30	

– $G_1^2 =$

– df =

– For G_2^2, we combine the first two columns and compare with the 3rd column (HF/NF + HF/F = HF).

$X \setminus Y$	HF	LF/NF	
Y			66
N			24
	60	30	90

– For G_3^2, we combine the first three columns and compare with the 4th column (No LF/F).

$X \setminus Y$	Not LF/F	LF/F	
Y			80
N			40
	90	30	120

– $df =$

When n is small, we use another type of test.

VIII. Exact Test for Independence (for small n) – Fisher Exact Test

- When samples are small, the distributions of X^2 and G^2 are not well approximated by the chi-squared distribution
- Solution: Perform "exact tests" (or "estimates of exact tests").
- Fisher's test conditions on the margins of the observed 2 × 2 table i.e test is based on conditioning on the marginals.
- Consider the set of all tables with the exact same margins as the observed table.

- In this set of tables, once you know the value in 1 cell, you can fill in the rest of the cells.

- Therefore, to find the probability of observing a table, we only need to find the probability of 1 cell in the table (rather than the probabilities of 4 cells).

- Typically, we use the (1, 1) cell, and compute the probabilities that $n_{11} = y$.

- That is for a 2×2 table, $n_{1+}, n_{2+}, n_{+1}, n_{+2}$ are fixed, which means that there is one free variable, say n_{11}.

- Computing Probabilities of Tables assuming $H_0 : \theta = 1$

- When $\theta = 1$, n_{11} has a hypergeometric distribution with probability function:
$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}.$$

- Sometimes both marginal are fixed by the experiment (Tea taster in textbook).

- More often, both are not fixed, but we test as if fixed.

- The p-value equals

- p-value $= \sum$ (hypergeometric probabilities of tables that favor H_a, including the probability for the observed table).

- To compute the p-value, we need the alternative H_a.

- $H_0 : \theta = 1$ versus $H_a : \theta < 1$

  - Find the odds ratio of the observed table,
  $$\theta = n_{11}n_{22}/n_{12}n_{21}$$

  - Compute the hypergeometric probabilities for the tables where the odds ratios are less than odds ratio from the observed table, including the probability for the observed table.

- $H_0 : \theta = 1$ versus $H_a : \theta > 1$

- Compute the hypergeometric probabilities for tables where $\hat{\theta} >$ the odds ratio from the observed table, including the probability for the observed table.

- $H_0 : \theta = 1$ versus $H_a : \theta \neq 1$

- For this case, we use a different criterion.

- p-value = sum of hypergeometric probabilities of tables that are no more likely than the observed table.

Ex. A new treatment for a disease is to be compared with the current method. The current method is used on 6 patients and the new method is used on 9 patient with the following results:

	Success	Failure	
current	2	4	6
new	8	1	9
	10	5	15

Is there evidence to say that the new method is better?
Test: $H_0 : \theta = 1$ v.s. $H_a : \theta < 1$.

IX. Three Way Tables $2 \times 2 \times 2$

1. Introduction

- Common situation: what effect does the explanatory variable X have on response Y implies bivariate analysis.
- What if the relationship between X and Y depends on the values of some other variable?
- Ex: Consider the outcome (Success or Failure) of 2 medical treatments classified by sex of the patients.

	Sex			
	M		F	
Outcome (Y)	S	F	S	F
$X = 1$	60	20	40	80
$X = 2$	100	50	10	30

Does the choice of treatment (X) affect outcome (Y)?
Case 1: No mention of (Z):

	Sex	
Outcome (Y)	S	F
$X = 1$	100	100
$X = 2$	110	80

Case 2: Just look at male:

	Sex	
	M	
Outcome (Y)	S	F
$X = 1$	60	20
$X = 2$	100	50

- This contradicts conclusion for marginal table.
- This contradiction is known as Simpson's Paradox
- How can this happen?
- Probabilities from marginal table are weighted averages of those from males and females.
- Using the law of total probability

  $$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B}).$$

  That is $P_1 = .75\left(\frac{80}{200}\right) + .33\left(\frac{120}{200}\right) = .5$ and $P_2 = .67\left(\frac{150}{190}\right) + .25\left(\frac{40}{190}\right) = .58$
- P_2 is larger because males have much higher success probability than females and the majority of men had treatment 2 while the majority of females had treatment 1.
- In this example, it is important to control or adjust for gender when looking at the relationship between choice of treatment and outcome.

Remarks:

1. Table for males and table for females are called partial tables.
2. They are treatment × outcome table conditioning on gender.
3. Odds computed from these tables are called *conditional or partial odds.*
4. Marginal table is 2 tables with combined gender.

Moral:

(a) Don't collapse tables, that is, don't use marginal tables unless appropriate.

- Appropriate if relationship between X and Y is the same in the marginal table as it is in the partial tables.
- estimated probabilities need to be about the same in all 3 tables.

(b) In designing experiments, record all potentially important variables. "Control variable(s)", that might possibly influence the relationship between X and Y.

2. Conditional and Marginal Odds ratio

- We have seen that the odds ratio for X, Y is

$$\theta = \theta_{XY} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}.$$

- Conditional Odds Ratios are odds ratios between two variables for fixed levels of the third variable.

- In an $X - Y - Z$ table, the π's and μ's are obtained by summing over Z, so we can also write

$$\theta = \theta_{XY} = \frac{\pi_{11+}\pi_{22+}}{\pi_{12+}\pi_{21+}} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}.$$

- Conditional or partial odds ratios are computed from partial tables, i.e.

$$\theta = \theta_{XY(k)} = \frac{\mu_{11(k)}\mu_{22(k)}}{\mu_{12(k)}\mu_{21(k)}}.$$

describes the XY association when $Z = k$.

- Conditional odds ratios are sometimes referred to as measures of "partial association".
- Marginal Odds Ratios are the odds ratios between two variables in the marginal table.
- The marginal odds ratios need not equal the partial (conditional) odds ratios.
- Marginal association can be very different from conditional association.
- Marginal association is meaningful only when it is identical to the conditional association.

3. Marginal vs. Conditional Independence

- No relationship between marginal and conditional independence.

a- Marginal Independence of X and Y is

$$\pi_{ij} = \pi_{i+}\pi_{+j} \Leftrightarrow \theta_{XY} = 1.$$

b- Conditional independence of X and Y given Z is

$$P(X = i, Y = j|Z = k) = P(X = i|Z = k)P(X = i|Z = k) \Leftrightarrow \theta_{XY(k)} = 1$$

Note:

- $b \not\Rightarrow a$. Conditional independence does not imply marginal independence. See table 2.11 on page 53.
- $a \not\Rightarrow b$. Marginal independence does not imply conditional independence.

Ex. π_{ijk} given as follows:

		Y	
	X	1	2
$Z = 1$	1	.1	.2
	2	.1	.05
$Z = 2$	1	.2	.1
	2	.1	.15

– Verify that (a) is true
– First, find the marginal table – collapse over Z

		Y	
		1	2
X	1	.3	.3
	2	.2	.2

– Next, look at the 2 conditional tables, partial odds ratios:

– X and Y are conditionally dependent given Z.

– When $Z = 1$, $Y = 1$ is less likely for $X = 1$ than $X = 2$. When $Z = 2$, $Y = 1$ is more likely for $X = 1$ than $X = 2$.

4. Homogeneous XY Association

- How are X and Y related? Look at conditional odds ratios.
- Generally, we have a different answer for each $Z = k$ value.
- If we get the same relationship for all cases, i.e.

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(k)}, \quad (X, Y binary)$$

then we say that we have Homogeneous XY association.

- There is "no interaction between any 2 variables in their effects on the third variable".
- There is "no 3-way interaction" among the variables.
- Note: conditional independence of X and Y is a special case of homogeneous association $\theta_{XY(k)} = 1$.

  Ex. ($2 \times 2 \times 2$ table)
  X: amount of prenatal care (primary variable)
  Y: survival of infant (response variable)
  Z: clinic attended

		Infant Survival	
	Care	died	survived
Clinic	less	3	176
A	more	4	293
Clinic	less	17	197
B	more	2	23

  – Calculate two partial odds ratios for $X - Y$

  Approximately, $\theta_{XY(1)} = \theta_{XY(2)} \approx 1$ suggesting that given a clinic, it appears that survival is unrelated to prenatal care.
  Note: Homogeneous association for one pair of variable implies homogeneous association for other pairs.

- In general,

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$
$$\theta_{X(1)Z} = \theta_{X(2)Z} = \cdots = \theta_{X(J)Z}$$
$$\theta_{(1)YZ} = \theta_{(2)YZ} = \cdots = \theta_{(I)YZ}$$

- all three hold or none holds.
- Conditional independence of X and Y is a special case of homogeneous association

  Ex. XZ partial odds ratios

X. Testing for Conditional Independence

Any relationship between X and Y after adjusting for Z? That is $\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(k)} = 1$

1. Cochran-Mantel-Haenszel Test (CMH)

- From discussion of Fisher's exact test, we know that the distribution of 2×2 tables with fixed margins is hypergeometric.
- Regardless of sampling scheme, if we consider row and column totals of partial tables as fixed, we can use hypergeometric distribution to compute probabilities.
- The test for conditional association uses one cell from each partial table.
- For $2 \times 2 \times k$ table
- Under conditional independence, conditioning on marginal totals for X and Y at each level of Z, we have

H_0: X and Y are independent given Z.

$$\mu_{11k} = E[n_{11k}] = \frac{n_{1+k}n_{+1k}}{n_{++k}}$$

$$V(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{++k}^2(n_{++k}-1)}$$

T.S.

$$CMH = \frac{\left[\sum_k (n_{11k} - \mu_{11k})\right]^2}{\sum_k V(n_{11k})}$$

- If X and Y are conditionally independent (H_0 true), then approximately, $CMH \sim \chi_1^2$.

  Ex. Rabbits are given a lethal injection of streptococci and an injection of penicillin either immediately or 1.5 hours delayed. Response is "cured or died".

		Response (Y)			
	Delay (X)	cured	died	μ_{11k}	$V(n_{11k})$
1/8	none	0	6	$\frac{6\times0}{11} = 0$	0
	1.5 hrs	0	5		
1/4	none	3	3	$\frac{6\times3}{12} = 1.5$	$\frac{6\times6\times3\times9}{12^2\times11} \approx \frac{27}{44}$
	1.5 hrs	0	6		
1/2	none	6	0	$\frac{6\times8}{12} = 4$	$\frac{32}{44}$
	1.5 hrs	2	4		
1	none	5	1	$\frac{6\times11}{12} = 5.5$	$\frac{11}{44}$
	1.5 hrs	6	0		
4	none	2	0	$\frac{2\times7}{7} = 2$	0
	1.5 hrs	5	0		

$k = 5$ levels of Z

- Using chi-Square table, $\alpha = .05$, df $= 1$, $\chi^2_{.05}(1) = 3.84$.
  Conclusion: we conclude that we do not have conditional independence. The cure rate and timing of penicillin injection are dependent given the penicillin level.
- This test works best when $\theta_{XY(1)} = \cdots = \theta_{XY(k)}$
- We will see later how to test this homogeneity association, $\theta_{XY(1)} = \cdots = \theta_{XY(k)}$.
- When $\theta_{XY(1)} = \cdots = \theta_{XY(k)}$, we can consider a pooled estimator of θ.

2. Mantel-Haenszel estimator of θ:

- When $\theta_{XY(1)} = \cdots = \theta_{XY(k)}$, the "Mantel-Haenszel Estimator" of a common value of the odds ratio is

$$\widehat{\theta}_{MH} = \frac{\sum(n_{11k}n_{22k}/n_{++k})}{\sum(n_{12k}n_{21k}/n_{++k})}$$

- Note that the standard error for $\widehat{\theta}_{MH}$ is complex, so we will rely on a software to get this and therefore confidence intervals for θ_{MH}

Ex. Penicillin example.

This is the end of chapter 2, which will be in the test. You can have 1 cheat sheet of formulas, calculator, table. Always remember to always put the interpretation of your results.