BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part II

Contingency Tables

Objectives

Probability structure for contingency tables

Proportions, odds ratios, and other risk measures

Test for independence

- Alternative: not independent
- Alternative: linear trend (ordinal data)

Association in Three-way Tables

Agresti (2002), Section 2.1 - 2.5, 2.7

Self-read: Exact inference for small sample

Agresti (2002), Section 2.6

Contingency Table

A two-way contingency table is a cross-classification of observations by the levels of two discrete variables. The cells of the table contain frequency count.

A two-way contingency table with I rows and J columns is called an $I \times J$ table. Denote

- X: row variable with I categories $i = 1, 2, \ldots, I$
- Y: column variable with J categories $j = 1, 2, \ldots, J$

The simplest one is a 2×2 table.

[Example] Belief in Afterlife by Gender

	Have belief	No belief
Female	509	116
Male	398	104

Notations - Counts

We observe (X, Y) for a sample of n subjects. Let n_{ij} be the number of subjects having $(X = i, Y = j)$.

Grand total: n

Cell count: $\{n_{ij}\}_{i,j} \quad i = 1, \ldots, I; j = 1, \ldots, J$

Row total: $n_{i+} = n_{i1} + n_{i2} + \cdots + n_{iJ} = \sum_{j=1}^{J} n_{ij}, \quad i = 1, \ldots, I$

Column total: $n_{+j} = n_{1j} + n_{2j} + \cdots + n_{Ij} = \sum_{i=1}^{I} n_{ij}, \quad j = 1, \ldots, J$

Row totals and column totals are marginal totals.

$$\sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij} = \sum_{i} n_{i+} = \sum_{j} n_{+j} = n$$

Notations - Probabilities

Joint probabilities

$\pi_{ij} = \Pr(X = i, Y = j), \quad i = 1, \ldots, I; j = 1, \ldots, J$

Marginal probabilities

Row margins

$\pi_{i+} = \pi_{i1} + \pi_{i2} + \cdots + \pi_{iJ} = \sum_{j=1}^{J} \pi_{ij}, \quad i = 1, \ldots, I$

Column margins

$\pi_{+j} = \pi_{1j} + \pi_{2j} + \cdots + \pi_{Ij} = \sum_{i=1}^{I} \pi_{ij}, \quad j = 1, \ldots, J$

$$\sum_{i,j} \pi_{ij} = \sum_{i} \pi_{i+} = \sum_{j} \pi_{+j} = 1$$

Conditional probabilities: probability of a level of one variable given the level of the other variable.

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}, \quad \text{and} \quad \pi_{i|j} = \frac{\pi_{ij}}{\pi_{+j}}$$

Notations - Sample Proportions

The sample cell proportion is $p_{ij} = \frac{n_{ij}}{n}$

The marginal row proportion is $p_{i+} = \frac{n_{i+}}{n}$

The marginal column proportion is $p_{+j} = \frac{n_{+j}}{n}$

[Example] Belief in Afterlife by Gender - with notations

	Have belief	No belief	Total	Proportion
Females	$n_{11} = 509$ $p_{11} = \frac{509}{1127} = 0.452$	$n_{12} = 116$ $p_{12} = \frac{116}{1127} = 0.103$	$n_{1+} = 625$	$p_{1+} = \frac{625}{1127} = 0.555$
Males	$n_{21} = 398$ $p_{21} = \frac{398}{1127} = 0.353$	$n_{22} = 104$ $p_{22} = \frac{104}{1127} = 0.092$	$n_{2+} = 502$	$p_{2+} = \frac{502}{1127} = 0.445$
Total	$n_{+1} = 907$	$n_{+2} = 220$	$n = 1127$	
Proportion	$p_{+1} = 0.805$	$p_{+2} = 0.195$		

Sampling Schemes and Study Design

What are some ways of generating the tables of counts?

- Unrestricted sampling (Poisson)
- Sampling with fixed total sample size (Multinomial)
- Sampling with fixed certain marginal totals (Product-Multinomial, Hypergeometric)

Multinomial Sampling

Draw a sample of n subjects from a population and record (D, E) for each subject. For example,

D: 1 - no disease; 2 - disease

E: 1 - low exposure level; 2 - high exposure level

Then the joint distribution of n_{ij} is multinomial with index n and parameter $\pi = \{\pi_{ij}\}$, where the grand total n is known.

Sometimes we express the parameter in terms of the cell means $\mu_{ij} = E(n_{ij}) = n\pi_{ij}$.

In this case, you can estimate joint probability, marginal probability, and conditional probability.

Poisson Sampling

Collect data until the sunset, or until the experimenter runs out of the supplies, or ... etc.

No margins of a table are fixed by design. Each cell is considered to be an independent Poisson variable. That is, the cell counts will be independent with Poisson distribution with $n_{ij} \sim Poisson(\mu_{ij})$, for $i = 1, \ldots, I$ and $j = 1, \ldots, J$.

From a likelihood standpoint, we get the same inferences about $\pi = \{\pi_{ij}\}$, whether n is regarded as fixed or random. Therefore, Poisson data may be analyzed as if they were multinomial if n is not of interest.

Product-multinomial Sampling

Collect data on a predetermined number of individuals for each category of one variable and classify them according to the other, i.e., one margin is fixed by design while the other(s) is free to vary.

- Cohort Studies
- Case-Control Studies

Viewing data as product-multinomial is appropriate when

- Row totals are truly fixed by design
- Row totals are not fixed, but we are only interested in $P(Y|X)$. That is, when Y is the outcome of interest, and X is a covariate that we do not wish to model.

From the likelihood standpoint, if the data are collected as multinomial distribution but the parameters of interest are functions of $\pi_{j|i}$, then the correct likelihood-based inference may be obtained by treating the data as if they were product-multinomial.

Cohort Studies

In a *cohort* study, sampling is carried out separately at different exposure levels, leading to distinct cohorts.

1. Identify two subgroups based on the levels of exposure E. That is, $E = 1$ with low exposure level and $E = 2$ with high exposure level.
2. Take a random sample from each of these two subgroups separately, of sizes $n_{E=1}$ and $n_{E=2}$, respectively.
3. Measure subsequently the absence $(D = 1)$ and presence $(D = 2)$ of disease for individuals in both samples.

In this case, the joint probability and marginal probability are not meaningful. For conditional probability, you can estimate $P(D = 1|E = 1)$, $P(D = 1|E = 1)$,$P(D = 1|E = 2)$, and $P(D = 2|E = 2)$, but not the other way around.

Case-control Studies

In a *case-control* study, separate samples are selected from cases ($D = 2$) and controls ($D = 1$).

1. Identify two subgroups of the population on the basis of the presence or absence of D.
2. Take a simple random sample from each of these two subgroups separately, of size $n_{D=1}$ and $n_{D=2}$, respectively
3. Measure subsequently the exposure level E (1 or 2) for individuals in both random samples

In this case, the joint probability and marginal probability are not meaningful. For conditional probability, you can estimate $P(E = 1|D = 1)$, $P(E = 2|D = 1)$,$P(E = 1|D = 2)$, and $P(E = 2|D = 2)$, but not the other way around.

Hypergeometric Sampling

In these rare examples, we may encounter data where both the row totals and column totals are fixed by design.

Even when both sets of marginal totals are not fixed by design, some statisticians like to condition on them and perform "exact" conditional inference when the sample size is small and asymptotic approximations are unlikely to work well.

Two-by-Two Tables

Many studies compare two groups on a binary response Y, e.g., Y is Outcome that takes two values, success and failure. The data can be displayed in a 2×2 table.

		Outcome	
		Success	Failure
Group	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

In this table, π_1 and π_2 are the probabilities of success for Group 1 for Group 2, respectively. They are both conditional probabilities:
$\pi_1 = \pi_{Y=1|X=1} = \pi_{Success|Group=1}$, $\pi_2 = \pi_{Y=1|X=2} = \pi_{Success|Group=2}$.

Risk measures:

- Difference of proportions: $\pi_1 - \pi_2$
- Relative risk: π_1/π_2
- Odds ratio: $\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$

Difference of Proportions

	Y		
X	Success	Failure	Sample Size
1	p_1	$1-p_1$	n_1
2	p_2	$1-p_2$	n_2

The difference of proportions $\pi_1 - \pi_2$ is estimated by its sample counterpart $p_1 - p_2$. The estimated standard error of $p_1 - p_2$ is

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

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

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

- To test $H_0 : \pi_1 = \pi_2$, use the z test statistic (Score test)

$$z = \frac{p_1 - p_2}{SE_{\text{pooled}}} = \frac{p_1 - p_2}{\sqrt{p_{\text{pooled}}(1-p_{\text{pooled}})(n_1^{-1} + n_2^{-1})}}$$

where $p_{\text{pooled}} = (n_1 p_1 + n_2 p_2)/(n_1 + n_2)$.

[Example] Aspirin and Heart Attacks among Male Physicians

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	$n_1 = 11,034$
Aspirin	104	10,933	$n_2 = 11,037$

We have
$p_1 = 189/11034 = 0.0171$
$p_2 = 104/11037 = 0.0094$

The sample difference of proportions is $p_1 - p_2 = 0.0077$. Its SE is

$$SE = \sqrt{\frac{(0.0171)(0.9829)}{11034} + \frac{(0.0094)(0.9906)}{11037}} = 0.0015$$

So a 95% confidence interval for the true difference $\pi_1 - \pi_2$ is

$$0.0077 \pm 1.96(0.0015) = (0.005, 0.011)$$

Since this interval contains only positive values, we conclude that $\pi_1 > \pi_2$. For males, taking aspirin appears to result in a diminished risk of heart attack.

Relative Risk

Difference in proportions may be misleading when the proportions are close to 0 or 1. For example, the difference between 0.01 and 0.001 is the same as the difference between 0.41 and 0.401 but the ratios are very different.

Relative Risk $= \pi_1/\pi_2$.

- Relative risk is equal to 1 if and only if $\pi_1 = \pi_2$
- Relative risk is estimated by the sample relative risk $= p_1/p_2$
- A large-sample confidence interval for $\log(\pi_1/\pi_2)$ is

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

- A large-sample $(1-\alpha)\%$ confidence interval for π_1/π_2 is obtained by exponentiating the above result:

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

[Example] Aspirin and Heart Attacks among Male Physicians

The sample relative risk is

$$\frac{p_1}{p_2} = \frac{0.0171}{0.0094} = 1.82$$

The sample proportion of MI cases was 82% higher for the group taking placebo.

A large-sample 95% CI for is $\log(\pi_1/\pi_2)$ is

$$\log(1.82) \pm 1.96\sqrt{\frac{1-0.0171}{0.0171 \times 11034} + \frac{1-0.0094}{0.0094 \times 11037}} = (0.361, 0.837)$$

A large-sample 95% CI for is π_1/π_2 is

$$(e^{0.361}, e^{0.837}) = (1.434, 2.309)$$

The proportion of MI is at least 43.4% higher for the placebo group.

Odds and Odds Ratio

For a probability of success π, the *odds* of success is defined as

$$\text{odds} = \frac{\pi}{1-\pi}$$

It is the ratio of success probability over failure probability.

By algebra, we have

$$\pi = \frac{\text{odds}}{\text{odds}+1}$$

In our 2×2 table, we have the odds of success in row 1 (Group =1) as $\pi_1/(1-\pi_1)$ and the odds of success in row 2 (Group =2) as $\pi_2/(1-\pi_2)$. The ratio of two odds is defined as the odds ratio and denoted by θ.

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

Properties of the Odds Ratio

- $\theta = 1$ if and only if $\pi_1 = \pi_2$
- $\theta > 1$ implies row 1 has higher odds and the probability of success
- $\theta < 1$ implies row 1 has lower odds and the probability of success
- The farther away θ is from 1, the stronger the association between row and column variables.
- If the order of the two rows or the two columns are switched, the new odds ratio is the inverse of the old one

  If the row 1 odds is one fourth of the row 2 odds, the odds ratio is 0.25. If the rows are switched, the odds ratio is 4, which is equal to 1/0.25
- If row variable and column variable are switched, the new odds ratio equals the old one.
- When both π_1 and π_2 are close to 0, the odds ratio $\approx$ relative risk.

When both variables are response variable, the odds ratio can be defined using joint probabilities:

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

Regardless the multinomial distribution over four cells or independent binomial for the two rows, the odds of row 1 can be estimated by n_{11}/n_{12} and the odds of row 2 can be estimated by n_{21}/n_{22}. Therefore, the odds ratio can be estimated by

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

- When any cell has 0 counts, the sample odds ratio will be 0 or ∞. In this case, or even when some cells have small counts, one can add 0.5 to each cell and estimate θ by

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

- When computing SE, replace n_{ij} by $n_{ij}+0.5$.

[Example] Aspirin and Heart Attacks among Male Physicians

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	$n_1 = 11,034$
Aspirin	104	10,933	$n_2 = 11,037$

For physicians taking placebo, the estimated odds of MI is

$$n_{11}/n_{12} = 189/10845 = 0.0174$$

For those taking aspirin, the estimated odds of MI is

$$n_{21}/n_{22} = 104/10933 = 0.0095$$

So the sample odds ratio is

$$\hat{\theta} = \frac{0.0174}{0.0095} = 1.832.$$

- 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.

The above calculation is equivalent to

$$\hat{\theta} = \frac{189 \times 10933}{10845 \times 104} = 1.832.$$

Adding 0.5 to each of the four cells, we have another estimate of the odds ratio

$$\tilde{\theta} = \frac{189.5 \times 10933.5}{10845.5 \times 104.5} = 1.828.$$

This estimate is close to what we obtained before $\hat{\theta} = 1.832$, since no cell count is small.

Inference for Odds Ratio

Similar to relative risk, the sampling distribution of odds ratio is highly skewed to the right (long right tail). A log-transformation of θ makes its sampling distribution more symmetric.

$$\begin{array}{lll} \theta \in (0,1), & \text{or } \log(\theta) \in (-\infty, 0): & \text{negative association} \\ \theta = 1, & \text{or } \log(\theta) = 0: & \text{no association} \\ \theta \in (1,\infty), & \text{or } \log(\theta) \in (0,\infty): & \text{positive association} \end{array}$$

The asymptotic distribution of $\log(\hat{\theta})$ is $N(\log(\theta), (SE)^2)$, where

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

A large-sample $(1-\alpha)\%$ CI for $\log\theta$ is

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

A large-sample $(1-\alpha)\%$ CI for θ is

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

[Example] Aspirin and Heart Attacks among Male Physicians

Since $\hat{\theta} = 1.832$, $\log(\hat{\theta}) = 0.605$. The sampling SE for $\log(\hat{\theta})$ is

$$SE = \sqrt{\frac{1}{189} + \frac{1}{10933} + \frac{1}{104} + \frac{1}{10845}} = 0.123$$

The asymptotic distribution of $\log(\hat{\theta})$ is $N(0.605, 0.123^2)$.

A 95% CI for $\log \theta$ is

$$0.605 \pm 1.96 \times 0.123 = (0.365, 0.846)$$

A 95% CI for θ is

$$(e^{0.365}, e^{0.846}) = (1.44, 2.33)$$

This interval does not contain 1, we know the null hypothesis $H_0 : \theta = 1$ (i.e., there is no association) will be rejected at significance level 0.05. The true odds seem to be different for the two groups.

Independence

Definition: $X(i = 1, 2, \ldots, I)$ and $Y(j = 1, 2, \ldots, J)$ are *statistically independent* if the conditional distribution of Y is the same at each level of X, i.e.,

$$\pi_{j|1} = \pi_{j|2} = \ldots = \pi_{j|I} = \pi_{+j} \quad \text{for all } j.$$

In the 2×2 table introduced earlier,

		Outcome	
		Success	Failure
Group	1	π_1	$1 - \pi_1$
	2	π_2	$1 - \pi_2$

- $\pi_1 = \pi_2$
- Difference of proportions $\pi_1 - \pi_2 = 0$
- Relative risk $\pi_1/\pi_2 = 1$
- Odds ratio $\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 1$

We have learned all the three tests this far.

Alternative Definition for Independence

Definition: $X(i = 1, 2, \ldots, I)$ and $Y(j = 1, 2, \ldots, J)$ are *statistically independent* if and only if

$$P(X = i, Y = j) = P(X = i)P(Y = j) \quad \text{for all } i \text{ and } j.$$

That is,

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

The Chi-squared test of independence is developed based on this definition.

Expected Frequency and Its Estimator

H_0: X and Y are independent versus H_1: X and Y are dependent

Recall that H_0 implies that, for all (i, j)

$$\Pr(X = i, Y = j) = \Pr(X = i)\Pr(Y = j), \quad \text{i.e.} \quad \pi_{ij} = \pi_{i+}\pi_{+j}.$$

Denote μ_{ij} as the expected frequency for cell (i, j). By definition, $\mu_{ij} = E(n_{ij}) = n\pi_{ij}$. Under the H_0, $\mu_{ij} = n\pi_{i+}\pi_{+j}$.

Denote $\hat{\mu}_{ij}$ as the estimated expected frequencies under the H_0. By MLE,

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}.$$

Note that n_{ij} is the MLE of μ_{ij} under H_1.

Tests of Independence

The Pearson Chi-squared statistic (Score test) is

$$X^2 = \sum_{\text{all cell}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

The Likelihood Ratio Statistic is

$$G^2 = -2 \log \left(\frac{\text{Maximum likelihood under } H_0}{\text{Maximum likelihood under } H_1} \right) = 2 \sum_{i,j} n_{ij} \log \left(\frac{n_{ij}}{\hat{\mu}_{ij}} \right)$$

Facts About the Tests

- X^2 and G^2 are ≥ 0 and 0 is achieved when all $n_{ij} = \hat{\mu}_{ij}$.
- If H_0 is true, X^2 and G^2 are approximately a Chi-squared distribution for large n with degrees of freedom equal to

$$\begin{aligned} df &= \#\text{ of parameters under } H_1 - \#\text{ of parameters under } H_0 \\ &= (IJ - 1) - [(I - 1) + (J - 1)] \\ &= (I - 1)(J - 1) \end{aligned}$$

- As n increases, X^2 converges to Chi-square faster than does G^2
- The p-value is determined by the upper-tail of the distribution

$$p\text{-value} = \Pr(\chi^2_{df} \geq X^2)$$

$$p\text{-value} = \Pr(\chi^2_{df} \geq G^2)$$

- The chi-squared approximation is usually decent when $\{\hat{\mu}_{ij} \geq 5\}$

[Example:] Gender Gap in Political Affiliation

Gender	Party Identification			Total
	Democrat	Independent	Republican	
Females	$n_{11} = 762$	$n_{12} = 327$	$n_{13} = 468$	$n_{1+} = 1557$
Males	$n_{21} = 484$	$n_{22} = 239$	$n_{23} = 477$	$n_{2+} = 1200$
Total	$n_{+1} = 1246$	$n_{+2} = 566$	$n_{+3} = 945$	$n_{++} = 2757$

Given the observed counts, we can calculate the expected counts under H_0 by $\hat{\mu}_{ij} = \frac{n_{i+} n_{+j}}{n}$.

$$\hat{\mu}_{11} = \frac{1246 \times 1557}{2757} = 703.7 \qquad \hat{\mu}_{21} = \frac{1246 \times 1200}{2757} = 542.3$$

$$\hat{\mu}_{12} = \frac{566 \times 1557}{2757} = 319.6 \qquad \hat{\mu}_{22} = \frac{566 \times 1200}{2757} = 246.4$$

$$\hat{\mu}_{13} = \frac{945 \times 1557}{2757} = 533.7 \qquad \hat{\mu}_{23} = \frac{945 \times 1200}{2757} = 411.3$$

In summary,

	Party Identification			
Gender	Democrat	Independent	Republican	Total
Females	$\hat{\mu}_{11} = 703.7$	$\hat{\mu}_{12} = 319.6$	$\hat{\mu}_{13} = 533.7$	$n_{1+} = 1557$
Males	$\hat{\mu}_{21} = 542.3$	$\hat{\mu}_{22} = 246.4$	$\hat{\mu}_{23} = 411.3$	$n_{2+} = 1200$
Total	$n_{+1} = 1246$	$n_{+2} = 566$	$n_{+3} = 945$	$n_{++} = 2757$

$$X^2 = \frac{(762-703.7)^2}{703.7} + \frac{(327-319.6)^2}{319.6} + \frac{(468-533.7)^2}{533.7} + \frac{(484-542.3)^2}{542.3} + \frac{(239-246.4)^2}{246.4} + \frac{(477-411.3)^2}{411.3} = 30.1$$

$$G^2 = 2\left[762\log\left(\frac{762}{703.7}\right) + 327\log\left(\frac{327}{319.6}\right) + 468\log\left(\frac{468}{533.7}\right) + 484\log\left(\frac{484}{542.3}\right) + 239\log\left(\frac{239}{246.4}\right) + 477\log\left(\frac{477}{411.3}\right)\right] = 30.0$$

Both have $df = (2-1)(3-1) = 2$. The p-values are, respectively, $P(X > 30.1) = 2.91 \times 10^{-7}$ and $P(X > 30.0) = 3.06 \times 10^{-7}$ with $X \sim \chi_2$. Both p-values are < 0.0001, suggesting association between political party identification and gender. [R command: 1 - pchisq(30.1, df=2)]

Residuals Analysis

Residuals tell how far off are the expected and observed values for each cell. They tell us which cells drive the lack of fit.

- Raw Residual:

$$n_{ij} - \hat{\mu}_{ij}$$

- Pearson Residual:

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}} \qquad (\sum_{ij} e_{ij}^2 = X^2),$$

  where the variance of e_{ij} tends to be smaller than 1 and leads to conservative indications of cells having lack of fit.

- Standardized Pearson Residuals:

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

  which adjusts Pearson Residuals using estimated standard error of $n_{ij} - \hat{\mu}_{ij}$ under H_0. Under H_0, standardized Pearson residuals are approximately N(0,1).

[Example:] Party Identification by Gender

For the $(1, 1)$ cell,

- $n_{11} = 762$
- $\hat{\mu}_{11} = 703.7$
- $p_{1+} = 1557/2757 = 0.565$ and $p_{+1} = 1246/2757 = 0.452$

So the standardized residual for this cell is

$$\frac{762 - 703.7}{\sqrt{703.7(1 - 0.565)(1 - 0.452)}} = 4.50$$

Standardized Residuals (in parentheses) for testing independence

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

Linear Trend Alternative to Independence

When variables are ordinal, a test for trend association is common.

Usually, scores are assigned to categories and measure the degree of linear trend. These scores should reflect the ordering of categories and the distance between them. Let

- scores for rows: $u_1 \leq u_2 \leq \cdots \leq u_I$.
- scores for columns: $v_1 \leq v_2 \leq \cdots \leq v_J$.

With scores, the sample correlation r between row and column variables can be computed by

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

where $\bar{u} = \sum_i u_i p_{i+}$ is the mean of row scores and $\bar{v} = \sum_j v_j p_{+j}$ is the mean of column scores. The population version of r is ρ.

For testing $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$, a test statistics is $M^2 = (n-1)r^2$, which is a Chi-squared distribution with $df = 1$ under a large sample.

[Example:] Infant Malformation and Mother's Alcohol Consumption. Maternal drinking = covariate, and congenital malformation = response.

Alcohol	Malformation			Percentage of Present
Consumption	Absent	Present	Total	$\frac{n_{present}}{Total} \times 100$
0	17,066	48	17,114	0.28
< 1	14,464	38	14,502	0.26
1–2	788	5	793	0.63
3–5	126	1	127	0.79
≥ 6	37	1	38	2.63

Ignoring the ordering in alcohol consumption, one can use X^2 or G^2 statistic

$$X^2 = 12.1 \text{ and } G^2 = 6.2$$

Each has $(5-1)(2-1) = 4$ degrees of freedom. The p-value is 0.02 for X^2 and 0.19 for G^2. Chi-squared distribution does not work well because of the small cell counts (i.e., the two 1s)

The percentage present indicates a possible trend: malformations are more likely at higher levels of alcohol consumption. If we assign the following scores to the row categories $v_1 = 0, v_2 = 0.5, v_3 = 1.5, v_4 = 4.0, v_5 = 7$, then $M^2 = 6.57$ and p-value $= 0.01$.

Extra Power with Ordinal Tests

- The X^2 and G^2 tests are omnibus: they detect any types of association between X and Y. Their df is $(I-1)(J-1)$.
- The M^2 test is powerful for detecting association in the direction implied by the scores. Its df is 1.
- The smaller df of M^2 test makes it more powerful if the scores are "correct".
- M^2 test suffers in power if the scores are "incorrect".
- For small to moderate sample sizes, Chi-squared distribution approximation works better for smaller df.

Choice of Scores

- For most data set, different choices of scores has little impact on M^2 statistic. This may not be the case when data are highly unbalanced: some categories have many more observations than others
- Adding or multiplying a number to scores does not change the M^2 test statistic: scores (1, 2, 3, 4, 5) yields the same M^2 test statistic as (0, 1, 2, 3, 4) or (2, 4, 6, 8, 10)
- An alternative is to assign the average rank of each category as its score. This score is called the Midrank.

	Malformation			Cum.	
Alcohol	Absent	Present	Total	Total	Midrank
0	17,066	48	17,114	17,114	$\frac{1}{2}(1 + 17114) = 8557.5$
< 1	14,464	38	14,502	31,616	$\frac{1}{2}(17115 + 31616) = 24365.5$
1–2	788	5	793	32, 409	$\frac{1}{2}(31617 + 32409 = 32013$
3–5	126	1	127	32,536	$\frac{1}{2}(32410 + 32536) = 32473$
≥ 6	37	1	38	32,574	$\frac{1}{2}(32537 + 32574) = 32555.5$

- Sensitivity analysis
- Personal judgment

Nominal-Ordinal Tables

- In the previous example, we are testing for a linear tread in proportions of "successes". The trend test M^2 under this $I \times 2$ tables is also called "Cochran-Armitage Trend Test".
- The M^2 can not be applied to the case with a nominal variable with more than 2 levels.
- Generally, if X is nominal, one can use a ANOVA-type of analysis to compare the mean scores among the categories of X. This test has $df = I - 1$. When $I = 2$, it is equivalent to M^2.

So far...

Two-way contingency table: row variable X and column variable Y.

Feature	Study design	Description	Valid estimator
Sampling scheme = Multinomial			
n is fixed	Cross-sectional	Draw a sample of n subjects from a population.	Joint/marginal/conditional prob Rel risk, diff of props, odds ratio. X^2/G^2 test for indep, Linear trend test.
Sampling scheme = Product Multinomial			
Row margins are fixed	Cohort study	Draw n_{1+} from cohort one, n_{2+} from cohort two, and follow-up to see results of Y.	Conditional probability $P(Y\|X)$ Rel risk, diff of props, odds ratio.
Col margins are fixed	Case-control	Draw n_{+1} from disease status, n_{+2} from normal persons, and recall levels of X	Conditional probability $P(X\|Y)$ Odds ratio
Sampling scheme = Poisson			
Nothing is fixed		May be analyzed as if they were multinomial if n is not of interest.	
Sampling scheme = Hypergeometric			
Both row and col margins are fixed		Used for small sample inference.	

Three-way Tables

It is possible to have three categorical variables, X, Y, and Z, then the data can be summarized by a three-dimensional table with counts $\{n_{ijk}\}$.

- $X : i = 1, 2, \ldots, I$
- $Y : j = 1, 2, \ldots, J$
- $Z : k = 1, 2, \ldots, K$

Note, in this section, we still study the association between X and Y. The Z is considered as an *extraneous variable* that are not intentionally for studying in the experiment.

Loglinear model (Chapter 7) can treat Z in a similar/comparable way to X and Y, and covers more type of associations.

Specifically, $2 \times 2 \times K$ table.

Partial Tables

A table of the XY counts at fixed levels of Z, showing the association of X and Y while controlling for Z.

The associations in partial tables are called *conditional associations*, because different value of Z may lead to different association.

[Conditional Odds Ratios] of X and Y given Z are odds ratio for partial tables, denoted by $\theta_{XY(k)}$. It can be estimated by

$$\hat{\theta}_{XY(k)} = \frac{n_{11k}\, n_{22k}}{n_{12k}\, n_{21k}}$$

Marginal Tables

Collapse all these partial tables that the count in each cell of this XY marginal table is the sum of the counts in the corresponding cells of all partial tables. Counts are denoted by $\{n_{ij+}\}$.

The XY marginal table ignores Z. The associations in the marginal table is called *marginal associations*.

[Marginal Odds Ratio] of X and Y can be estimated by

$$\hat{\theta}_{XY} = \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}$$

Data Format: Kidney Stone Treatment

Stone	Treatment A and Success	Treatment A and Failure	Treatment B and Success	Treatment B and Failure
Small	81	6	234	36
Large	192	71	55	25

If X =Treatment (A,B), Y = Response (Success, Failure), and Z = Stone size (Small, Large).

Partial Tables

Stone Z	Treatment X	Response Y: Success	Response Y: Failure
Small	A	81	6
	B	234	36
Large	A	192	71
	B	55	25

Marginal Tables

Treatment X	Response Y: Success	Response Y: Failure
A	273	77
B	289	61

- $n_{11+} = 81 + 192 = 273$
- $n_{12+} = 6 + 71 = 77$
- $n_{21+} = 234 + 55 = 289$
- $n_{22+} = 36 + 25 = 61$

Confounding Variable and Spurious Association

The Z may have a confounding effect on the association between X and Y.

Confounding variable: an extraneous variable that correlates with both the dependent variable and independent variable.

Spurious association: the association between two variables that is induced by the presence of a confounding variable. That is, controlling for the confounding variable, the two variables are independent.

Observation: As ice cream sales increase, drowning deaths rate increases.

Question: Association between ice cream sales and rate of drowning death?

In reality, a heat wave may have caused both. Ice cream is sold during the hot summer months at a much greater rate than during colder times, and it is during these hot summer months that people are more likely to engage in activities involving water, such as swimming. The increased drowning deaths are simply caused by more exposure to water-based activities, not ice cream.

Conditional Independence versus Marginal Independence

Conditional Independence

If X and Y are independent in each and every partial table, then X and Y are said to be *conditionally independent, given* Z.

In a $2 \times 2 \times K$ table, conditional independence means

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

- Marginal independence of X and Y does not imply conditional independence. (Spurious Association)
- Conditional independence of X and Y, given Z, does not imply marginal independence of X and Y.

[Example] Conditional independence of X and Y, given Z, does not imply marginal independence of X and Y.

Clinic Z	Treatment X	Response Y: Success	Response Y: Failure	Odds ratio
1	A	18	12	$\hat{\theta}_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1$
	B	12	8	
2	A	2	8	$\hat{\theta}_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1$
	B	8	32	
Marginal	A	20	20	$\hat{\theta}_{XY} = \frac{20 \times 40}{20 \times 20} = 2$
	B	20	40	

Here "Clinic" is a confounding factor and this marginal association is spurious association due to "Clinic".

Simpson's Paradox

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

Use the previous kidney stone treatment data as an example,

Stone Z	Treatment X	Response Y: Success	Response Y: Failure	Odds ratio
Small	A	81	6	$\hat{\theta}_{XY(1)} = \frac{81 \times 36}{6 \times 234} = 2.08$
	B	234	36	
Large	A	192	71	$\hat{\theta}_{XY(1)} = \frac{192 \times 25}{71 \times 55} = 1.23$
	B	55	25	
Total	A	273	77	$\hat{\theta}_{XY} = \frac{273 \times 61}{77 \times 289} = 0.75$
	B	289	61	

Treatment A is more effective when used on small stones, and also when used on large stones.

Treatment B is more effective when considering both sizes together.

The stone size is a confounding variable that causes Simpson's paradox.

Simpson's Paradox (cont.)

In the above example:

- Doctors tend to give the severe cases (large stones) the better treatment (A), and the milder cases (small stones) the inferior treatment (B).
- The success rate is more strongly in influenced by the severity of the case than by the choice of treatment.

In Presence of Confounding

We see that without considering confounding variables, it is possible to introduce spurious association and observe Simpson paradox.

- Stratified analysis is always a strategy for examining the association between X and Y while adjusting for effect of Z. This will lead to multiple conditional associations for each individual stratum of Z.
- If the odds ratios across different strata are relatively consistent, is there an overall measurement of the association that can be used?
  –*Yes, Mantel-Haenszel (MH) Methods*.

Mantel-Haenszel Methods: K 2×2 tables

For the kth level of Z, $k = 1, \ldots, K$, the data is

	Diseased	Non-diseased	Totals
Exposed	n_{11k}	n_{12k}	n_{1+k}
Unexposed	n_{21k}	n_{22k}	n_{2+k}
Totals	n_{+1k}	n_{+2k}	n_{++k}

The Mantel-Haenszel method

- assumes that there is a true odds ratio which is consistent across k
- provides a pooled estimate of the common odds ratio. In essence, it is a weighted average of the odds ratios from the individual strata

Note that it only makes sense to report the Mantel-Haenszel estimate if the exposure-disease relationship is consistent across the strata.

Mantel-Haenszel Estimate of Odds Ratio

The Mantel-Haenszel estimate of the odds ratio is

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} n_{11k} n_{22k} / n_{++k}}{\sum_{k=1}^{K} n_{21k} n_{12k} / n_{++k}}$$

with estimated standard error computed on the log-scale as

$$SE(\ln[\hat{\theta}_{MH}]) = \left[\frac{\sum P_k R_k}{2(\sum R_k)^2} + \frac{\sum P_k S_k + \sum Q_k R_k}{2 \sum R_k \sum S_k} + \frac{\sum Q_k S_k}{2(\sum S_k)^2} \right]^{1/2}$$

where

$$\begin{aligned} P_k &= (n_{11k} + n_{22k})/n_{++k} \\ Q_k &= (n_{12k} + n_{21k})/n_{++k} \\ R_k &= n_{11k} n_{22k}/n_{++k} \\ S_k &= n_{12k} n_{21k}/n_{++k} \end{aligned}$$

Mantel-Haenszel Estimate of Relative Risk

The Mantel-Haenszel estimate of the relative risk is

$$\hat{RR}_{MH} = \frac{\sum_{k=1}^{K} n_{11k} n_{2+k} / n_{++k}}{\sum_{k=1}^{K} n_{21k} n_{1+k} / n_{++k}}.$$

with estimated standard error computed on the log-scale as

$$SE(\ln[\hat{RR}_{MH}]) = \left\{ \frac{\sum (n_{1+k} n_{2+k} n_{+1k} - n_{11k} n_{21k} n_{++k}) / n_{++k}^2}{[\sum n_{11k} n_{2+k} / n_{++k}][\sum n_{21k} n_{1+k} / n_{++k}]} \right\}^{1/2}$$

- Mantel-Haenszel estimates can be obtained in SAS.
- It is important to keep in mind that these pooled risk estimates should only be reported if the risk is consistent (homogeneous) across the levels of the confounder.
- Apply test of homogeneity before using this combined estimate.

Homogeneous Association and Breslow-Day Test

If X and Y are binary, there is *homogeneous XY association* when

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

The alternative is, at least one pair $\theta_{XY(k)} \neq \theta_{XY(l)}$.

Test of Homogeneity: the Breslow-Day Statistic

$$X^2_{BD} = \sum_{k=1}^{K} \frac{(n_{11k} - \mu_{11k})^2}{Var(n_{11k})} \sim \chi^2_{K-1}$$

- The two-sided p-value is $p = \Pr[\chi^2_{K-1} \geq X^2_{BD}]$
- If the p-value is significant, then the null hypothesis is rejected, and it is concluded that the odds ratios are not homogeneous across strata. Specifically, it is not appropriate to report the Mantel-Haenszel pooled estimate of the odds ratio
- The test of homogeneity should be performed before deciding to report the pooled odds ratio

CMH Test for Conditional Independence

A special case for homogeneous association is when $\theta = 1$, i.e., the conditional independence. Here we introduce one test that is applicable to this case, the Cochran-Mantel-Haenszel test that takes test statistic as

$$X^2_{CMH} = \frac{\left[\sum_{k=1}^{K}(n_{11k} - \mu_{11k})\right]^2}{\sum_{k=1}^{K} Var(n_{11k})} \sim \chi^2_1,$$

where

$$\begin{aligned} \mu_{11k} &= \frac{n_{1+k}n_{+1k}}{n_{++k}}, \\ Var(n_{11k}) &= \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n^2_{++k}(n_{++k} - 1)}. \end{aligned}$$

The 2-sided p-value is $p = \Pr[\chi^2_1 \geq X^2_{CMH}]$.

If some $\theta_{XY(k)} < 1$ and other $\theta_{XY(k)} > 1$, then the CMH is NOT an appropriate test; that is, the test works well if the conditional odds ratios are in the same direction.