Categorical Data Analysis

Chapter 3

Deyuan Li
School of Management
Fudan University

Fall 2015

Outline

1. 3.1 Confidence Intervals for Association Parameters
2. 3.2 Testing Independence in Two-way Contingency Tables
3. 3.3 Following-up Chi-squared Tests
4. 3.4 Two-way Tables with Ordered Classifications
5. 3.5 Small-Sample Tests of Independence

3.1 Confidence Intervals for Association Parameters

3.1.1 Interval estimation of odds ratios

For a 2 $\times$ 2 table,

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

- the sample odds ratio

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

- If any $n_{ij} = 0 \;\Rightarrow\; \hat{\theta} = 0$ or ∞.
- If both entries in a row or column are zero $\Rightarrow\; \hat{\theta}$ is undefined.

3.1.1 Interval estimation of odds ratios

The Wald $100(1-\alpha)\%$ CI due to Woolf (1955).

(1) calculate the estimated SE for $\log\hat{\theta}$ (see Section 3.1.7):

(2) construct the CI for $\log\theta$ as $\log\hat{\theta} \pm z_{\alpha/2}\,\hat{\sigma}(\log\hat{\theta})$;

(3) obtain the CI for θ: $[\hat{\theta}e^{-z_{\alpha/2}\hat{\sigma}(\log\hat{\theta})}, \hat{\theta}e^{z_{\alpha/2}\hat{\sigma}(\log\hat{\theta})}]$

Note: the CI does not exist when $\hat{\theta} = 0$ or ∞.
$\Rightarrow$ Other approaches (Section 3.1.8).

3.1.2 Aspirin and myocardial infarction example

Table 3.1 Swedish Study on Aspirin Use and Myocardial Infarction（心机梗塞）.

	Myocardial Infarction		
	Yes	No	Total
Placebo	28	656	684
Aspirin	18	658	676

A randomized clinical trial, placebo vs. aspirin.

Outcome = death due to myocardial infarction within 3 years

3.1.2 Aspirin and myocardial infarction example

- $\hat{\theta} = 1.56$, $\log\hat{\theta} = 0.445$ and $\hat{\sigma}(\log\hat{\theta}) = 0.307$.
- A 95% CI for $\log\theta$ is $0.445 \pm 1.96 \times 0.307 = (-0.157, 1.047)$.
- The 95% CI for θ is $[\exp(-0.157), \exp(1.047)] = (0.85, 2.85)$.
- Conclusion:
  1. The width of the CI is quite large. ⇒ The estimate of the true odds ratio is rather imprecise.
  2. The CI for θ contains 1.0. ⇒ It is possible that the true odds of death are equal for aspirin and placebo.

3.1.3 Interval estimation of difference of proportions

Consider two independent binomial samples.

For sample i with sample size n_i, the probability of success is π_i and the observed number of successes is y_i.

We are interested in $\pi_1 - \pi_2$.

3.1.3 Interval estimation of difference of proportions

The Wald $100(1-\alpha)\%$ CI for $\pi_1 - \pi_2$.

(1) calculate sample proportions $\hat{\pi}_i = y_i/n_i$;
$E[\hat{\pi}_i] = \pi_i$ and $Var(\hat{\pi}_i) = \pi_i(1-\pi_i)/n_i$;

(2) calculate the estimated SE for $\hat{\pi}_1 - \hat{\pi}_2$:

(3) construct the CI for $\pi_1 - \pi_2$ as $(\hat{\pi}_1 - \hat{\pi}_2) \pm z_{\alpha/2}\,\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2)$.

3.1.3 Interval estimation of difference of proportions

Aspirin and myocardial infarction example (Table 3.1)

(1) Placebo: $\hat{\pi}_1 = 28/684 = 0.0409$; Aspirin: $\hat{\pi}_2 = 18/676 = 0.0266$. $\hat{\pi}_1 - \hat{\pi}_2 = 0.0409 - 0.0266 = 0.014$.

(2) standard error:

$$\begin{aligned}\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2) &= [.0409(1 - .0409)/684 + .0266(1 - .0266)/676]^{1/2} \\ &= .0098.\end{aligned}$$

(3) The 95% CI for $\pi_1 - \pi_2$ is

$$0.014 \pm 1.96 \times 0.0098 = (-0.005, 0.033).$$

Conclusion: the death rate for those taking placebo was between -0.005 and 0.033, larger than that for those taking aspirin.

3.1.4 Interval estimation of relative risk

The Wald $100(1-\alpha)\%$ CI for r.

(1) calculate sample relative risk $\hat{r} = \hat{\pi}_1/\hat{\pi}_2$ and $\log\hat{r}$;

(2) calculate the estimated SE for $\log\hat{r}$:

(3) construct the CI for $\log r$ as $\log\hat{r} \pm z_{\alpha/2}\,\hat{\sigma}(\log\hat{r})$;

(4) obtain the CI for r: $[\hat{r}e^{-z_{\alpha/2}\hat{\sigma}(\log\hat{r})}, \hat{r}e^{z_{\alpha/2}\hat{\sigma}(\log\hat{r})}]$

3.1.4 Interval estimation of relative risk

Aspirin and myocardial infarction example (Table 3.1)

(1) $\hat{r} = 0.0409/0.0266 = 1.54$ and

$$\hat{\sigma}(\log \hat{r}) = \left[\frac{1-.0409}{.0409 \times 684} + \frac{1-.0266}{.0266 \times 676}\right]^{1/2} = 0.297.$$

(2) The 95% CI for $\log r$ is
$\log(1.54) \pm 1.96 \times 0.297 = (-0.150, 1.014)$.

(3) The 95% CI for r is
$[\exp(-0.150), \exp(1.014)] = (0.86, 2.75)$.

Conclusion: the death rate for those taking placebo was between 0.86 and 2.75 times that for those taking aspirin.

Aspirin and myocardial infarction example

Overall conclusion:

According to either difference of proportions or relative risk, takin aspirin could give some benefit, but no effect or a slight negative effect are also possible.

3.1.5 Deriving standard errors with the delta method

For large n, suppose T_n is normally distributed about a parameter θ with SE $\sigma/\sqrt{n}$. That is,

$$\sqrt{n}\,(T_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

Suppose g is twice differentiable at θ. Then $g(T_n)$ is an estimator of $g(\theta)$. Using Taylor expansion for $g(t)$ around θ, we obtain

$$\sqrt{n}\,[g(T_n) - g(\theta)] \approx \sqrt{n}\,(T_n - \theta)\, g'(\theta)$$

for large n (see Section 14.1.2). Thus,

this is called the *delta method*.

3.1.5 Deriving standard errors with the delta method

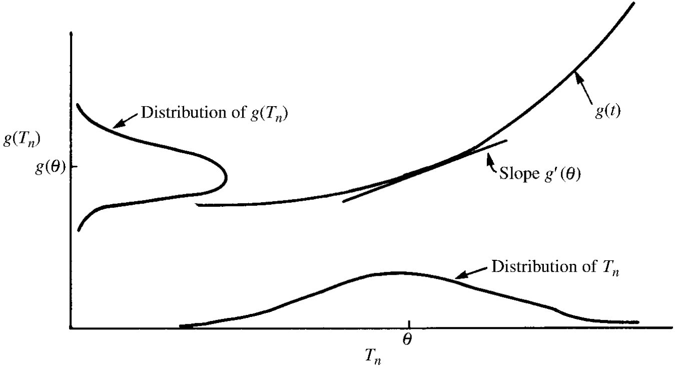

FIGURE 3.1 Depiction of delta method.

3.1.5 Deriving standard errors with the delta method

Since $g'(\theta)$ and $\sigma^2 = \sigma^2(\theta)$ usually depend on the unknown parameter θ, the asymptotic variance $[g'(\theta)]^2\sigma^2$ is unknown.

Substitute T_n for θ in $g'(\theta)$ and $\sigma^2(\theta)$ to obtain $\hat{g}'(T_n)$ and $\hat{\sigma}^2(T_n)$, then

$$\sqrt{n}\,[g(T_n) - g(\theta)]/|\hat{g}'(T_n)|\,\hat{\sigma}(T_n) \xrightarrow{d} N(0, 1).$$

Hence, a large-sample Wald 95% CI for $g(\theta)$ is

$$g(T_n) \pm 1.96\,|\hat{g}'(T_n)|\,\hat{\sigma}(T_n)/\sqrt{n}.$$

3.1.6 Delta method applied to sample logit

$\hat{\pi} = y/n$, the ML estimate of the binomial parameter π, with y successes in n trials. We will derive the asymptotic normality of $\log(\hat{\pi}/(1-\hat{\pi}))$.

- Let $T_n = \hat{\pi}$, $\theta = \pi$ and $g(t) = \log(t/(1-t))$. Recall $E(T_n) = E(\hat{\pi}) = \pi$, $\text{var}(T_n) = \sigma^2/n = \text{var}(\hat{\pi}) = \pi(1-\pi)/n$ and hence $\sigma^2 = \pi(1-\pi)$.
- By CLT, $\hat{\pi}$ has a large-sample normal distribution.
- Then,

$$g'(\theta) = g'(\pi) = \frac{d\log\pi}{d\pi} - \frac{d\log(1-\pi)}{d\pi} = \frac{1}{\pi} + \frac{1}{1-\pi} = \frac{1}{\pi(1-\pi)},$$

$$[g'(\theta)]^2\,\sigma^2 = \frac{1}{\pi^2(1-\pi)^2} \times \pi(1-\pi) = \frac{1}{\pi(1-\pi)}.$$

3.1.6 Delta method applied to sample logit

- By the delta method,

$$\sqrt{n}\Big(\log\frac{\hat{\pi}}{1-\hat{\pi}}-\log\frac{\pi}{1-\pi}\Big)\overset{d}{\to}N\Big(0,\ \frac{1}{\pi(1-\pi)}\Big)$$

  or

$$[g(\hat{\pi})-g(\pi)]\overset{d}{\to}N(0,\ 1/[n\pi(1-\pi)]).$$

Remark

(1) For $\pi=0$ or 1, the true variance does not exist.

(2) Since $\hat{\pi}=0$ or 1 with positive probability, the logit can equal $-\infty$ or ∞ with positive probability.

(3) For unknown π, we use

$$[g(\hat{\pi})-g(\pi)]\to N(0,\ 1/[n\hat{\pi}(1-\hat{\pi})]).$$

3.1.7 Delta method for log odds ratio

A multi-parameter version of the delta method

Suppose that $\{n_i,\ i = 1, \ldots, c\}$ have a multinomial $(n,\ \{\pi_i\})$ distribution.

The sample proportion $\hat{\pi}_i = n_i/n$ has

$$E(\hat{\pi}_i) = \pi_i, \quad \text{var}(\hat{\pi}_i) = \pi_i(1 - \pi_i)/n$$

and (see Section 14.1.4)

$$\text{cov}(\hat{\pi}_i, \hat{\pi}_j) = -\pi_i \pi_j / n, \quad \text{for } i \neq j.$$

The sample proportions $(\hat{\pi}_1,\ \hat{\pi}_2, \ldots, \hat{\pi}_{c-1})$ have a large-sample multivariate normal distribution.

3.1.7 Delta method for log odds ratio

Let $g(\boldsymbol{\pi})$ denote a differentiable function of $\{\pi_i\}$, with sample value $g(\hat{\boldsymbol{\pi}})$. Let $\phi_i = \partial g(\boldsymbol{\pi})/\partial \pi_i$, for $i = 1, \ldots, c$.

Then,

$$\sqrt{n}\,[g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})]/\sigma \to N(0,\, 1)$$

or

$$\Big(g(\hat{\boldsymbol{\pi}}) - g(\boldsymbol{\pi})\Big) \to N(0,\, \sigma^2/n),$$

where

$$\sigma^2 = \sum \pi_i\, \phi_i^2 - (\sum \pi_i\, \phi_i)^2.$$

3.1.7 Delta method for log odds ratio

Because

In practice, replacing $\{\pi_i\}$ and $\{\phi_i\}$ in σ^2 by their sample values yields an ML estimate $\hat{\sigma}^2$ of σ^2. Then $\widehat{\text{SE}}(g(\hat{\boldsymbol{\pi}})) = \hat{\sigma}/\sqrt{n}$, and a large-sample Wald CI for $g(\boldsymbol{\pi})$ is $g(\hat{\boldsymbol{\pi}}) \pm z_{\alpha/2}\,\hat{\sigma}/\sqrt{n}$.

3.1.7 Delta method for log odds ratio

Application to the log odds ratio

- Take $g(\boldsymbol{\pi}) = \log\theta = \log\pi_{11} + \log\pi_{22} - \log\pi_{12} - \log\pi_{21}$. Since $\phi_{11} = \partial\log\theta/\partial\pi_{11} = 1/\pi_{11}$, $\phi_{12} = -1/\pi_{12}$, $\phi_{21} = -1/\pi_{21}$, $\phi_{22} = 1/\pi_{22}$, we obtain

$$\sum_i\sum_j \pi_{ij}\,\phi_{ij} = \pi_{11}/\pi_{11} - \pi_{12}/\pi_{12} - \pi_{21}/\pi_{21} + \pi_{22}/\pi_{22} = 0$$

$$\sigma^2 = \sum_i\sum_j \pi_{ij}\,\phi_{ij}^2 = \frac{\pi_{11}}{(\pi_{11})^2} + \frac{\pi_{12}}{(\pi_{12})^2} + \frac{\pi_{21}}{(\pi_{21})^2} + \frac{\pi_{22}}{(\pi_{22})^2} = \sum_i\sum_j \frac{1}{\pi_{ij}}.$$

- The estimated asymptotic SE of $\log\hat{\theta}$ is

$$\widehat{\text{SE}}(\log\hat{\theta}) = \hat{\sigma}/\sqrt{n} = \sqrt{\sum_i\sum_j \frac{1}{\hat{\pi}_{ij}}}\,/\,\sqrt{n} = \Bigg[\sum_i\sum_j \frac{1}{n\,\hat{\pi}_{ij}}\Bigg]^{1/2}.$$

Application to the log relative risk. (Similarly)

3.1.8 Score and profile likelihood confidence intervals

CI's based on inverting Wald tests sometimes work poorly for small to moderate *n*. Alternative CI's based on inverting likelihood-ratio or score test often perform better.

Score method for the difference of proportions

For $H_0 : \pi_1 - \pi_2 = \Delta$, the score statistic ($Z_S = u(\beta)/\iota(\beta)$) is

$$z(\Delta) = \frac{(\hat{\pi}_1 - \hat{\pi}_2) - \Delta}{\sqrt{\hat{\pi}_1(\Delta)\left[1 - \hat{\pi}_1(\Delta)\right]/n_1 + \hat{\pi}_2(\Delta)\left[1 - \hat{\pi}_2(\Delta)\right]/n_2}},$$

where $\hat{\pi}_i(\Delta)$ denotes the ML estimate of π_i subject to the constraint $\pi_1 - \pi_2 = \Delta$.

No closed-form expressions for $\hat{\pi}_i(\Delta)$. The score CI is the set of Δ such that $|z(\Delta)| < z_{\alpha/2}$.

3.1.8 Score and profile likelihood confidence intervals

Likelihood-ratio method for the odds ratio

- The multinomial likelihood for a 2×2 table is a function of $\{\pi_{11}, \pi_{12}, \pi_{21}\}$, or equivalently of $\{\theta, \pi_{1+}, \pi_{+1}\}$.
- For testing $H_0 : \theta = \theta_0$, the π_{1+} and π_{+1} are *nuisance parameters*（讨厌参数）.
- The null ML estimates $\hat{\pi}_{1+}(\theta_0)$ and $\hat{\pi}_{+1}(\theta_0)$ maximize the likelihood under the H_0.
- The *profile log-likelihood function*（剖面对数似然函数）is $L(\theta_0, \hat{\pi}_{1+}(\theta_0), \hat{\pi}_{+1}(\theta_0))$, viewed as a function of θ_0.

3.1.8 Score and profile likelihood confidence intervals

- Evaluated at $\theta_0 = \hat{\theta}$, this is the maximized log likelihood $L(\hat{\theta}, \hat{\pi}_{1+}, \hat{\pi}_{+1})$, which occurs at the sample proportions $\hat{\pi}_{1+} = n_{1+}/n$ and $\hat{\pi}_{+1} = n_{+1}/n$.
- The profile CI for θ is the set of θ_0 for which

$$-2[L(\theta_0, \hat{\pi}_{1+}(\theta_0), \hat{\pi}_{+1}(\theta_0)) - L(\hat{\theta}, \hat{\pi}_{1+}, \hat{\pi}_{+1})] < \chi_1^2(\alpha).$$

- The profile likelihood approach is available in SAS (the CLODDS=PL option in MODEL statement of PROC LOGISTIC).

Outline

1. 3.1 Confidence Intervals for Association Parameters
2. 3.2 Testing Independence in Two-way Contingency Tables
3. 3.3 Following-up Chi-squared Tests
4. 3.4 Two-way Tables with Ordered Classifications
5. 3.5 Small-Sample Tests of Independence

3.2 Testing Independence in Two-way Contingency Tables

For multinomial sampling with probabilities $\{\pi_{ij}\}$ in an $I \times J$ contingency table, the null hypothesis of statistical independence is

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

3.2.1 Pearson and likelihood-ratio chi-squared tests

Pearson chi-squared test

- In Section 1.5.2 the Pearson $X^2 = \sum_i (n_i - \mu_i)^2/\mu_i$.
- Here, a test of H_0: independence uses

$$X^2 = \sum_i \sum_j (n_{ij} - \mu_{ij})^2/\mu_{ij},$$

  where $\mu_{ij} = n\,\pi_{i+}\,\pi_{+j} = E(n_{ij})$ under H_0.

- Usually $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown, we use their ML estimates, $\hat{\pi}_{i+} = n_{i+}/n$ and $\hat{\pi}_{+j} = n_{+j}/n$. (Show the proof!)
- Hence $\hat{\mu}_{ij} = n\,\hat{\pi}_{i+}\,\hat{\pi}_{+j} = n_{i+}\,n_{+j}/n$ and $X^2 = \sum_i \sum_j (n_{ij} - \hat{\mu}_{ij})^2/\hat{\mu}_{ij}$.

3.2.1 Pearson and likelihood-ratio chi-squared tests

Under H_0,

3.2.1 Pearson and likelihood-ratio chi-squared tests

- $X^2 \sim \chi^2_{df}$:
  1. if $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are known, then $df = IJ - 1$;
  2. if $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown, $\{\hat{\mu}_{ij}\}$ require estimating $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, then df$= (IJ - 1) - (I - 1) - (J - 1) = (I - 1)(J - 1)$.
- The larger the value of X^2, the more evidence against independence.
- We can show that the score test produces the X^2 statistic.

3.2.1 Pearson and likelihood-ratio chi-squared tests

Likelihood-ratio chi-squared test

- For multinomial sampling, the kernel of the likelihood is

$$\prod_i \prod_j \pi_{ij}^{n_{ij}}, \quad \text{where all } \pi_{ij} \geq 0 \quad \text{and} \quad \sum_i \sum_j \pi_{ij} = 1.$$

- Under H_0: independence, $\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n^2$. In the general case, $\hat{\pi}_{ij} = n_{ij}/n$. The likelihood-ratio chi-squared statistic is, with $\{\hat{\mu}_{ij} = n_{i+}\, n_{+j}/n\}$,

3.2.1 Pearson and likelihood-ratio chi-squared tests

- Under H_0, $\{\pi_{ij}\}$ are determined by $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, so the dimension is $(I-1)+(J-1)$.

  In the general case, $\{\pi_{ij}\}$ are subject to $\sum_i \sum_j \pi_{ij} = 1$, so the dimension is $IJ-1$.

  The difference in these dimensions equals $(I-1)(J-1)$.
- For large sample, the G^2 has a chi-squared distribution with $df = (I-1)(J-1)$.

3.2.1 Pearson and likelihood-ratio chi-squared tests

Comparison:

- X^2 and G^2 are asymptotically equivalent, i.e., $X^2 - G^2$ converges in probability to zero.
- The convergence to chi-squared is quicker for X^2 than G^2.
- The approximation is usually poor for G^2 when $n/(IJ) < 5$.
- The approximation can be decent for X^2 when I or J is large, some $\hat{\mu}_{ij}$ are as small as 1 but most exceed 5.

3.2.2 Education and religious belief

Table 3.2 Education and Religious Beliefs

	Religious Beliefs			
Highest Degree	Fundamentalist	Moderate	Liberal	Total
Less than high school	178 (137.8)[1] (4.5)[2]	138 (161.5) (-2.6)	108 (124.7) (-1.9)	424
High school or junior college	570 (539.5) (2.6)	648 (632.1) (1.3)	442 (488.4) (-4.0)	1660
Bachelor or graduate	138 (208.7) (-6.8)	252 (244.5) (0.7)	252 (188.9) (6.3)	642
Total	886	1038	802	2726

[1] Estimated expected frequencies for testing independent.

[2] Standard Pearson residuals.

3.2.2 Education and religious belief

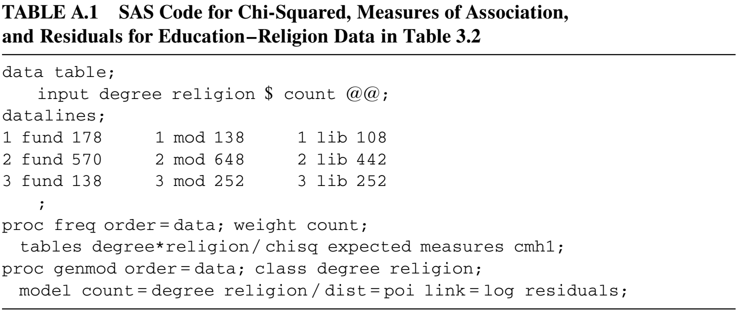

TABLE A.1 SAS Code for Chi-Squared, Measures of Association, and Residuals for Education–Religion Data in Table 3.2

```
data table;
   input degree religion $ count @@;
datalines;
1 fund 178     1 mod 138     1 lib 108
2 fund 570     2 mod 648     2 lib 442
3 fund 138     3 mod 252     3 lib 252
   ;
proc freq order=data; weight count;
  tables degree*religion/chisq expected measures cmh1;
proc genmod order=data; class degree religion;
  model count=degree religion/dist=poi link=log residuals;
```

$X^2 = 69.2$ and $G^2 = 69.8$, with df$= (3-1)(3-1) = 4$. P-values are < 0.0001.

$\Rightarrow$ There is strong evidence of an association between religious beliefs and degree of education.

Outline

1. 3.1 Confidence Intervals for Association Parameters
2. 3.2 Testing Independence in Two-way Contingency Tables
3. 3.3 Following-up Chi-squared Tests
4. 3.4 Two-way Tables with Ordered Classifications
5. 3.5 Small-Sample Tests of Independence

3.3 Following-up Chi-squared Tests

The *P*-value of a significance test of independence provides little information about the nature or strength of the association.

Some methods following up the test help to learn more about the association.

3.3.1 Pearson and standardized residuals

n_{ij}: observations in cell (i, j);
$\hat{\mu}_{ij}$: estimated expected frequency for cell (i, j).

Pearson residual: $e_{ij} = (n_{ij} - \hat{\mu}_{ij})/\sqrt{\hat{\mu}_{ij}}$.
Note $\sum_i \sum_j e_{ij}^2 = X^2$, i.e., Pearson statistic.

Under H_0, $\{e_{ij}\}$ are asymptotically normal with mean 0, but their variances are less than 1.0, averaging $[(I-1)(J-1)]/$(number of cells), i.e.,

$$Var(e_{ij}) \approx (I-1)(J-1)/(IJ) < 1.$$

3.3.1 Pearson and standardized residuals

Standardized Pearson residual:

- this residual is asymptotically standard normal under H_0.

$$\frac{n_{ij} - \hat{\mu}_{ij}}{[\hat{\mu}_{ij}\,(1 - p_{i+})\,(1 - p_{+j})]^{1/2}},$$

  where

$$p_{i+} = \frac{n_{i+}}{n} \quad \text{and} \quad p_{+j} = \frac{n_{+j}}{n}.$$

- A standardized Pearson residual that exceeds about 2 or 3 in absolute value indicates lack of fit of H_0 in that cell.

3.3.2 Education and religious fundamentalism revisited

Table 3.2 also shows standardized Pearson residuals.

- Cells (1,1) and (3,3) have large **positive** standardized residuals.

  ⇒ Significantly **more** subjects were at these combinations than H_0 predicts.

- Cells (2,3) and (3,1) have large **negative** standardized residuals.

  ⇒ Significantly **less** subjects were at these combinations than H_0 predicts.

3.3.2 Education and religious fundamentalism revisited

Odds ratios describe this trend.

- The 2×2 table constructed from the 1st and 3rd rows and the 1st and 3rd columns of Table 3.2 has a sample odds ratio of $(178 \times 252)/(108 \times 138) = 3.0$.
- For those with a bachelor's or graduate degree (3rd row), the estimated odds of selecting liberal (3rd column) instead of fundamentalist (1st column) were 3.0 times the estimated odds for those with less than a high school education (1st row).

3.3.3 Partitioning chi-squared

A chi-squared statistic having $df = \nu$ has partitions into independent chi-squared components, e.g., into ν components each having $df = 1$.

Conversely, if X_i^2, $i = 1, \ldots, c$, are independent, and $X_i^2 \sim \chi^2_{\nu_i}$, then $X^2 = \sum_{i=1}^c X_i^2 \sim \chi^2_{df}$ with $df = \sum_{i=1}^c \nu_i$.

A partitioning of a chi-squared test may show that an association reflects primarily differences between certain categories or groupings of categories.

3.3.3 Partitioning chi-squared

For $2 \times J$ tables (2 rows, J columns)

- partition G^2, which has $df = J - 1$, into $J - 1$ *independent* components, each has $df = 1$.
- The jth component is G^2_j for a 2×2 table where the 1st column combines columns 1 through j of the full table and the 2nd column is column $j + 1$, for $j = 1, \ldots, J - 1$.

	$1 \quad \cdots \quad j$	$j+1$	$\cdots \quad J$
1	$n^j_{11} = \sum_{b \leq j} n_{1b}$	$n^j_{12} = n_{1(j+1)}$	
2	$n^j_{21} = \sum_{b \leq j} n_{2b}$	$n^j_{22} = n_{2(j+1)}$	

- $G^2 = G^2_1 + \cdots + G^2_{J-1}$, where G^2_j compares the first j columns to the last column.

3.3.3 Partitioning chi-squared

For $I \times J$ tables (I rows, J columns)

(1) $J-1$ independent $I \times 2$ components with df=$I-1$, for $j = 1, \ldots, J-1$.

	$1 \quad \cdots \quad j$	$j+1$	$\cdots \quad J$
1	$n^j_{11} = \sum_{b \leq j} n_{1b}$	$n^j_{12} = n_{1(j+1)}$	
2	$n^j_{21} = \sum_{b \leq j} n_{2b}$	$n^j_{22} = n_{2(j+1)}$	
$\cdots$	$\cdots$	$\cdots$	
I	$n^j_{I1} = \sum_{b \leq j} n_{Ib}$	$n^j_{I2} = n_{I(j+1)}$	

(2) $(I-1)(J-1)$ independent 2×2 components with df=1, for $i = 2, \ldots, I;\ j = 2, \ldots, J$.

	$1 \quad \cdots \quad j-1$	j	$\cdots \quad J$
1 $\cdots$ $i-1$	$n^{ij}_{11} = \sum_{a<i} \sum_{b<j} n_{ab}$	$n^{ij}_{12} = \sum_{a<i} n_{aj}$	
i	$n^{ij}_{21} = \sum_{b<j} n_{ib}$	$n^{ij}_{22} = n_{ij}$	
$\cdots$			

3.3.4 Origin of schizophrenia （精神分裂症） example

Table 3.3: Most Influential School of Psychiatric Thought and Ascribed Origin of Schizophrenia

School of	Origin of Schizophrenia		
Psychiatric Thought	Biogenic	Environmental	Combination
Eclectic	90	12	78
Medical	13	1	6
Psychoanalytic	19	13	50

注：将精神病医生按照其大学学习时的理论学派（折衷派、医学派、精神分析派）和其个人对精神分裂症起因的观点（生物遗传、环境、综合）分类。

Table 3.3: 3 × 3, $G^2 = 23.04$ with df$= 4$.

3.3.4 Origin of schizophrenia example

Table 3.4: Subtables Used in Partitioning Chi-Squared for Table 3.3

				Bio+						Bio+	
	Bio	Env		Env	Com		Bio	Env		Env	Com
Ecl	90	12	Ecl	102	78	E+M	103	13	E+M	116	84
Med	13	1	Med	14	6	Psy	19	13	Psy	32	50

E+M: Ecl+Med

From Table 3.4, we have

1. Subtables 1 and 2 have small G^2 values.
   $\Rightarrow$ There is little evidence of a difference between the eclectic (Ecl) and medical (Med) schools.

3.3.4 Origin of schizophrenia example

Table 3.4: Subtables Used in Partitioning Chi-Squared for Table 3.3

				Bio+						Bio+	
	Bio	Env		Env	Com		Bio	Env		Env	Com
Ecl	90	12	Ecl	102	78	E+M	103	13	E+M	116	84
Med	13	1	Med	14	6	Psy	19	13	Psy	32	50

E+M: Ecl+Med

2. Subtables 3 and 4 have large G^2 values.

   ⇒ Subtable 4: Members of the psychoanalytic school (Psy) seem more likely than the other schools (Ecl + Med) to choose the combination (Com) origin.

   ⇒ Subtable 3: Among those who chose either the biogenetic (Bio) or environmental (Env) origin, members of the psychoanalytic school (Psy) seem more likely than the other schools (Ecl + Med) to choose the environmental (Env) origin.

3.3.6 Limitations of chi-squared tests

1. The chi-squared tests of independence merely indicate the degree of evidence of association, not the nature and strength of the association.
   $\Rightarrow$ Study residuals, decompose chi-squared into components, estimate parameters such as odds ratios.
2. The chi-squared tests (using X^2 or G^2) require large sample.
3. The chi-squared tests are invariant to reordering of rows and columns.
   $\Rightarrow$ They treat both classifications as nominal.
   $\Rightarrow$ They are not suitable for ordinal variables.

3.3.7 Why consider independence?

One reason to consider independence is the benefits of model parsimony（模型精简）, i.e., the simplest model explains most of the data.

If the independence model approximate the true probabilities well, then unless n is vary large, $\{\hat{\pi}_{ij} = \hat{\pi}_{i+}\hat{\pi}_{+j} = n_{i+}n_{+j}/n^2\}$ tends to be better than the sample proportions $\{p_{ij} = n_{ij}/n\}$.

$\{\hat{\pi}_{ij}\}$ smooth the sample counts (not only on a single cell count).

- Although $\{\hat{\pi}_{ij}\}$ may be biased, they have smaller variance because they are based on estimating fewer parameters, i.e., $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$, instead of $\{\pi_{ij}\}$.
- $\{\hat{\pi}_{ij}\}$ can have smaller MSE, unless n is so large that the bias term dominates the variance.

3.3.7 Why consider independence?

Note: If *W* is an estimator for a parameter θ, then

As indicated in Section 3.1.7 (consider the $\hat{\pi}_i$ there as the p_{ij} here), $E(p_{ij}) = \pi_{ij}$, i.e., unbiased, so

The independence estimators are biased, so the MSE expression cannot be simplified:

$$\text{MSE}(\{\hat{\pi}_{ij}\}) = \sum_i \sum_j E(\hat{\pi}_{ij} - \pi_{ij})^2.$$

3.3.7 Why consider independence?

Example

Table 3.5 Cell Probabilities for Comparison of Estimators

$(1+\delta)/9$	$1/9$	$(1-\delta)/9$
$1/9$	$1/9$	$1/9$
$(1-\delta)/9$	$1/9$	$(1+\delta)/9$

in Table 3.5,

$$\pi_{ij} = \pi_{i+}\,\pi_{+j}\,[1+\delta(i-2)(j-2)]$$

with $-1<\delta<1$ and $\pi_{i+}=\pi_{+j}=1/3$ (for all i, j and δ).

3.3.7 Why consider independence?

Since

$$\begin{aligned}\sum_i \sum_j \pi_{ij}^2 &= 2 \times \left(\frac{1+\delta}{9}\right)^2 + 5 \times \left(\frac{1}{9}\right)^2 + 2 \times \left(\frac{1-\delta}{9}\right)^2 \\ &= \frac{9+4\delta^2}{81} = \frac{1}{9} + \frac{4\delta^2}{81},\end{aligned}$$

we have

$$\mathrm{MSE}(\{p_{ij}\}) = \frac{1}{n}\left\{1 - \left(\frac{1}{9} + \frac{4\delta^2}{81}\right)\right\} = \frac{1}{n}\left\{\frac{8}{9} - \frac{4\delta^2}{81}\right\}.$$

Rather tedious calculations yield

$$\mathrm{MSE}(\{\hat{\pi}_{ij}\}) = \frac{1}{n}\left\{\frac{4}{9} - \frac{4}{9n}\right\} + \frac{4\delta^2}{81}\left\{1 - \frac{2}{n} + \frac{2}{n^2} - \frac{2}{n^3}\right\}.$$

3.3.7 Why consider independence?

Table 3.6: Comparison of Total MSE(*10,000) for Sample Proposition and Independence Estimators

	$\delta = 0$		$\delta = 0.1$		$\delta = 0.2$		$\delta = 0.6$		$\delta = 1.0$	
n	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$	p	$\hat{\pi}$
10	889	489	888	493	887	505	871	634	840	893
50	178	91	178	95	177	110	174	261	168	565
100	89	45	89	50	89	65	87	220	84	529
500	18	9	18	14	18	28	17	186	17	500
∞	0	0	0	5	0	20	0	178	0	494

$\delta = 0$ means independence model.

3.3.7 Why consider independence?

Table 3.6 lists the total MSE values for various δ and n:

- When $\delta = 0$, $\text{MSE}(\{p_{ij}\}) = 8/(9n)$, whereas $\text{MSE}(\{\hat{\pi}_{ij}\}) \approx 4/(9n)$ for large n.

  $\Rightarrow$ The independence estimator $\{\hat{\pi}_{ij}\}$ is much better than the sample proportions $\{p_{ij}\}$.
- When $\delta \approx 0$ and n is not large, $\text{MSE}(\{\hat{\pi}_{ij}\}) \approx \text{MSE}(\{p_{ij}\})/2$.
- When $\delta \neq 0$, $\text{MSE}(\{\hat{\pi}_{ij}\}) \to 4\delta^2/81$, whereas $\text{MSE}(\{p_{ij}\}) \to 0$, as $n \to \infty$.

Outline

1. 3.1 Confidence Intervals for Association Parameters
2. 3.2 Testing Independence in Two-way Contingency Tables
3. 3.3 Following-up Chi-squared Tests
4. 3.4 Two-way Tables with Ordered Classifications
5. 3.5 Small-Sample Tests of Independence

3.4.1 Linear trend alternative to independence

When the row variable X and the column variable Y are ordinal, a positive or negative trend in the association is common.

A test for linear trend ($H_0 : X \perp Y$ v.s. $H_a : X \approx kY$)

1) Assign scores（赋分）to the categories.

   The scores have the same ordering as the categories. They assign distances between categories and actually treat the measurement scale (score) as interval.

   Let $u_1 \leq u_2 \leq \cdots \leq u_I$ denote scores for the rows,
   and $v_1 \leq v_2 \leq \cdots \leq v_J$ for the columns.

3.4.1 Linear trend alternative to independence

For the (i, j) cell, there are n_{ij} observations with scores (u_i, v_j):

$$\begin{array}{cc} & \begin{array}{cc} X & Y \end{array} \\ n_{ij} & \left\{ \begin{array}{cc} u_i & v_j \\ \vdots & \vdots \\ u_i & v_j \end{array} \right. \end{array}$$

2) Calculate correlation coefficient r between the row and column scores.

The r is based on $\sum_i \sum_j u_i \, v_j \, n_{ij}$, which relates to the covariance of X and Y.

3.4.1 Linear trend alternative to independence

3) Transform the original hypothesis to $H_0 : r = 0$ v.s. $H_a : r \neq 0$) and calculate:

$$\begin{array}{rl} \text{Two-sided} & M^2 = (n-1)\, r^2, \\ \text{One-sided} & M = \sqrt{n-1}\; r. \end{array}$$

M^2 or $|M|$ increases with $|r|$ and n. For large sample, $M^2 \sim \chi^2_1$ and $M \sim N(0, 1)$.

Large values of M^2 or $|M|$ contradict independence.

Note that H_0: independence, but the alternative may be
H_a: trend existing,
H_a: positive trend existing, or
H_a: negative trend existing.

3.4.1 Linear trend alternative to independence

Calculate *P*-value:

- for M^2, the *P*-value = Pr($M^2 \geq$ observed value);
- for $M > 0$ (positive trend), the *P*-value = Pr($M \geq$ observed value);
- for $M < 0$ (negative trend), the *P*-value = Pr($M \leq$ observed value).

A small *P*-value does NOT imply that the association is linear, merely that searching for a linear component to the association helped to build power against H_0.

The test treats the variables symmetrically, i.e., the results will be the same if we switch X with Y.

3.4.2 Job satisfaction example revisited

Table 2.8 Cross-Classification of Job Satisfaction by Income

Income (dollars)	Job Satisfaction Very Dissatisfied	Little Dissatisfied	Moderately Satisfied	Very Satisfied
<15,000	1	3	10	6
15,000-25,000	2	3	10	7
25,000-40,000	1	6	14	12
>40,000	0	1	9	11

Table 2.8 — Job satisfaction and income for 96 subjects

Ordinary chi-squared statistics for testing independence with df$=9$:
$X^2 = 6.0$ with P-value $= 0.74$ and
$G^2 = 6.8$ with P-value $= 0.66$.

$\Rightarrow$ Little evidence of association.

3.4.2 Job satisfaction example revisited

With scores (1, 2, 3, 4) for job satisfaction and scores (7.5, 20, 32.5, 60) for income (i.e., midpoints of categories in thousands of dollars), the correlation is $r = 0.200$.

$M^2 = (96 - 1)(0.200)^2 = 3.81$ with $P = 0.051$.
$\Rightarrow$ Some evidence of association.

$M = \sqrt{(96 - 1)} \times 0.2 = 1.95$ with $P = 0.026$.
$\Rightarrow$ Stronger evidence for the one-side (positive trend) alternative.

3.4.3 Monotone trend alternatives to independence

- The monotone alternative uses an ordinal measure of association, such as $\gamma = \frac{\pi_c - \pi_d}{\pi_c + \pi_d}$.
- The sample gamma $\hat{\gamma} = \frac{C-D}{C+D} \sim$ Normal for large n.
- The SE can be derived by the delta method (available in software packages, e.g., RPOC FREQ in SAS).
- The test statistic is $z = \hat{\gamma}/\text{SE}$.

3.4.3 Monotone trend alternatives to independence

For Table 2.8,

- $\hat{\gamma} = 0.221 > 0$ (Section 2.4.5). ⇒ A weak tendency for job satisfaction to be higher at higher income levels.
- The SE$=$ 0.117 from SAS. $z = 0.221/0.117 = 1.89$ with $P = 0.03$ for the one-sided alternative. ⇒ Some evidence that $\gamma > 0$.
- An approximate 95% CI for γ is $0.221 \pm 1.96 \times 0.117 = (-0.01, 0.45)$. ⇒ The true association between income and job satisfaction is at most moderately positive.

3.4.5 Choice of scores

Arbitrary choice

- Ideally, the scale is chosen by a consensus of experts, and subsequent interpretations use that same scale.
- Different scoring systems can give quite different results. However, for most data sets, different choices of monotone scores give similar results.

  Scores that are linear transforms of each other, e.g., $S_1 = (1, 2, 3, 4)$ and $S_2 = 2 \times (S_1 - 1) = (0, 2, 4, 6)$, have the same absolute correlation and hence the same M^2.

3.4.5 Choice of scores

- If the data are highly unbalanced（非平衡）, i.e., with some categories having many more observations than others, then the results may depend on the scores.
- Binary nominal variables can be treated as ordinal for statistics of ordinal analysis.

  For a binary variable, any set of distinct scores is a linear transformation of any other set.

3.4.5 Choice of scores

Table 3.7 Example for which Results Depend on Choice of Scores.

	Alcohol Consumption (average number of drinks per day)				
Malformation	0	< 1	1-2	3-5	≥ 6
Absent	17,066	14,464	788	126	37
Present	48	38	5	1	1

X=maternal drinking (5 ordinal categories derived from a continuous variable);

Y=congenital malformations (binary).

3.4.5 Choice of scores

Most of the observations are in the first two columns
$\Rightarrow$ Unbalanced.

Any score for row since it is binary.

Column scores for drinking:

Set 1: Scores = (0, 0.5, 1.5, 4.0, 7.0)
$\Rightarrow M^2 = 6.57$ with $P = 0.010$.
Set 2: Scores = (1, 2, 3, 4, 5), i.e., equally spaced,
$\Rightarrow M^2 = 1.83$ with $P = 0.18$.

3.4.5 Choice of scores

Midranks:

- assign ranks from 1 to n to all subjects;
- all subjects in a category receive the average of their ranks, i.e., *midranks*.

When X and Y are both ordinal and M^2 uses midrank scores, the correlation on which M^2 is based is called *Spearman's rho*(available in PROC CORR of SAS):

$$\rho_S = 12 \int \int (1 - F(x, y)) dF_X(x) dF_Y(y) - 3,$$

where F, F_X, F_Y are the joint cdf, marginal cdfs of X, Y, respectively.

3.4.5 Choice of scores

For Table 3.7:

- *Column 1*: The $n_{+1} = 17066 + 48 = 17114$ subjects share ranks 1 through $n_{+1} = 17114$. Each receives the midrank $= (1 + n_{+1})/2 = 8557.5$.

- *Column 2*: The $n_{+2} = 14464 + 38 = 14502$ subjects share ranks $n_{+1} + 1 = 17115$ through $n_{+1} + n_{+2} = 31616$. Each receives the midrank $= [(n_{+1} + 1) + (n_{+1} + n_{+2})]/2$ $= n_{+1} + (1 + n_{+2})/2 = 24365.5$.

- *Column 3*: The $n_{+3} = 788 + 5 = 793$ subjects share ranks $n_{+1} + n_{+2} + 1 = 31617$ through $n_{+1} + n_{+2} + n_{+3} = 32409$. Each receives the midrank $= n_{+1} + n_{+2} + (1 + n_{+3})/2 = 32013$.

3.4.5 Choice of scores

- *Column 4*: The $n_{+4} = 126 + 1 = 127$ subjects share ranks $n_{+1} + n_{+2} + n_{+3} + 1 = 32410$ through $n_{+1} + n_{+2} + n_{+3} + n_{+4} = 32536$. Each receives the midrank $= n_{+1} + n_{+2} + n_{+3} + (1 + n_{+4})/2 = 32473$.

- *Column 5*: The $n_{+5} = 37 + 1 = 38$ subjects share ranks $n_{+1} + n_{+2} + n_{+3} + n_{+4} + 1 = 32537$ through $n_{+1} + n_{+2} + n_{+3} + n_{+4} + n_{+5} = 32574$. Each receives the midrank $= n_{+1} + n_{+2} + n_{+3} + n_{+4} + (1 + n_{+5})/2 = 32555.5$.

$M^2 = 0.35$ with $P = 0.55$, i.e., an even weaker conclusion.

Outline

1. 3.1 Confidence Intervals for Association Parameters
2. 3.2 Testing Independence in Two-way Contingency Tables
3. 3.3 Following-up Chi-squared Tests
4. 3.4 Two-way Tables with Ordered Classifications
5. 3.5 Small-Sample Tests of Independence

3.5.1 Fisher's exact test for 2 × 2 tables

Under the H_0 : independence, **conditioning on both sets of marginal totals** yields the hypergeometric distribution:

$$p(t) = \Pr(n_{11} = t) = \left(\begin{array}{c} n_{1+} \\ t \end{array} \right) \left(\begin{array}{c} n_{2+} \\ n_{+1} - t \end{array} \right) / \left(\begin{array}{c} n \\ n_{+1} \end{array} \right) .$$

Given the marginal totals, n_{11} determines the other three cell counts.

The range of possible values for n_{11} is $m_- \leq n_{11} \leq m_+$, where $m_- = \max(0, n_{1+} + n_{+1} - n)$ and $m_+ = \min(n_{1+}, n_{+1})$.

3.5.1 Fisher's exact test for 2×2 tables

For 2×2 tables, independence is equivalent to $\theta = 1$.

Consider $H_0 : \theta = 1$ against $H_a : \theta > 1$.

For the given marginal totals, tables having larger n_{11} must have smaller n_{12} and n_{21}, $\Rightarrow$ larger sample odds ratios $\Rightarrow$ stronger evidence in favor of H_a.

P-value = $\Pr(n_{11} \geq t_o)$, where t_o denotes the observed value of n_{11}.

3.5.2 Fisher’s tea drinker

Fisher prepared 4 cups of tea with milk added first and 4 cups with tea added first.
$\Rightarrow$ Marginal totals of “true” variable were fixed.

Fisher’s colleague knew there were 4 cups with milk added first and 4 cups with tea added first.
$\Rightarrow$ Marginal totals of the “guess” variable were also fixed.

$\Rightarrow$ Fisher’s exact test of $H_0 : \theta = 1$ against $H_a : \theta > 1$.

3.5.2 Fisher's tea drinker

Table 3.8 Fisher's Tea Testing Experiment

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

For $t_o = 3$,

$$\Pr(n_{11} = 3) = \binom{4}{3}\binom{4}{1} \Big/ \binom{8}{4} = 0.229,$$

$$\Pr(n_{11} = 4) = \binom{4}{4}\binom{4}{0} \Big/ \binom{8}{4} = 0.014.$$

Then the P-value = $\Pr(n_{11} \geq 3) = 0.229 + 0.014 = 0.243$.
$\Rightarrow$ This result does not establish an association between the actual order of pouring and the prediction.

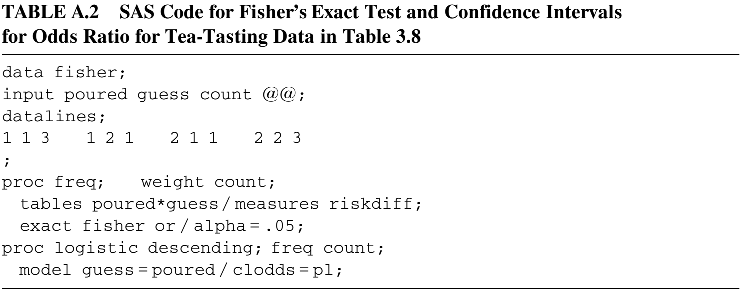

TABLE A.2 SAS Code for Fisher's Exact Test and Confidence Intervals for Odds Ratio for Tea-Tasting Data in Table 3.8

```
data fisher;
input poured guess count @@;
datalines;
1 1 3   1 2 1   2 1 1   2 2 3
;
proc freq;   weight count;
  tables poured*guess / measures riskdiff;
  exact fisher or / alpha = .05;
proc logistic descending; freq count;
  model guess = poured / clodds = pl;
```