Lecture 9: Categorical Data Analysis in Complex Surveys

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Multinomial Sampling

Each couple in an SRS of 500 married couples from a large population is asked whether (1) the household owns at least one personal computer; (2) household subscribes to cable television. The following contingency table presents the outcomes:

Observed Count		Computer? Yes	Computer? No	
Cable?	Yes	119	188	307
	No	88	105	193
		207	293	500

Are households with a computer more likely to subscribe to cable?

Expected Count		Computer? Yes	Computer? No	
Cable?	Yes	127.1	179.9	307
	No	79.9	113.1	193
		207	293	500

Pearson's chi-square test statistic is

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}} = 2.281.$$

The **likelihood ratio chi-square test statistic** is

$$G^2 = 2 \sum_{\text{all cells}} (\text{observed count}) \ln\left(\frac{\text{observed count}}{\text{expected count}}\right) = 2.275.$$

If owning a computer and subscribing to cable are independent events, the odds that a cable subscriber will own a computer should equal the odds that a non-cable-subscriber will own a computer. We estimate the odds of owning a computer if the household subscribes to cable as 119/188 and estimate the odds of owning a computer if the household does not subscribe to cable as 88/105. The **odds ratio** is therefore estimated as

$$\frac{\frac{119}{188}}{\frac{88}{105}} = 0.755.$$

If the null hypothesis of independence is true, we expect the odds ratio to be close to one. Equivalently, we expect the logarithm of the odds ratio to be close to zero. The log odds is -0.28 with asymptotic standard error

$$\sqrt{\frac{1}{119}+\frac{1}{88}+\frac{1}{188}+\frac{1}{105}} = 0.186;$$

an approximate 95% confidence interval (CI) for the log odds is $-0.28 \pm 1.96(0.186) = [-0.646, 0.084]$. This CI includes 0, and confirms the result of the hypothesis test that there is no evidence against independence. ■

Testing Independence of Factors

Each of n independent observations is cross-classified by two factors: row factor R with r levels and column factor C with c levels. Each observation has probability p_{ij} of falling into row category i and column category j, giving the following table of true probabilities. Here, $p_{i+} = \sum_{j=1}^{c} p_{ij}$ is the probability that a randomly selected unit will fall in row category i, and $p_{+j} = \sum_{i=1}^{r} p_{ij}$ is the probability that a randomly selected unit will fall in column category j.

		C				
		1	2	$\cdots$	c	
R	1	p_{11}	p_{12}	$\cdots$	p_{1c}	p_{1+}
	2	p_{21}	p_{22}	$\cdots$	p_{2c}	p_{2+}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	r	p_{r1}	p_{r2}	$\cdots$	p_{rc}	p_{r+}
		p_{+1}	p_{+2}	$\cdots$	p_{+c}	1

The null hypothesis of independence is

$$H_0 : p_{ij} = p_{i+}p_{+j} \quad \text{for } i = 1, \ldots, r \quad \text{and} \quad j = 1, \ldots, c.$$

Let $m_{ij} = np_{ij}$ represent the expected counts. If H_0 is true, $m_{ij} = np_{i+}p_{+j}$, and m_{ij} can be estimated by

$$\hat{m}_{ij} = n\hat{p}_{i+}\hat{p}_{+j} = n\frac{x_{i+}}{n}\frac{x_{+j}}{n},$$

where $\hat{p}_{ij} = x_{ij}/n$, $\hat{p}_{+j} = \sum_{i=1}^{r} \hat{p}_{ij}$, and $\hat{p}_{i+} = \sum_{j=1}^{c} \hat{p}_{ij}$. Pearson's chi-square test statistic is

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(x_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} = n \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}}.$$

The likelihood ratio test statistic is

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} \ln\left(\frac{x_{ij}}{\hat{m}_{ij}}\right) = 2n \sum_{i=1}^{r}\sum_{j=1}^{c} \hat{p}_{ij} \ln\left(\frac{\hat{p}_{ij}}{\hat{p}_{i+}\hat{p}_{+j}}\right).$$

The p-values will be approximately correct if (a) the expected count in each cell is greater than 1 and (b) $n \geq 5 \times$ (number of cells).

An equivalent statement is that all odds ratios equal 1:

$$H_0 : \frac{p_{11}p_{ij}}{p_{1j}p_{i1}} = 1 \qquad \text{for all } i \geq 2 \text{ and } j \geq 2.$$

We may estimate any odds ratio $(p_{ij}p_{kl})/(p_{il}p_{kj})$ by substituting in estimated proportions: $(\hat{p}_{ij}\hat{p}_{kl})/(\hat{p}_{il}\hat{p}_{kj})$. If the sample is sufficiently large, the *logarithm* of the estimated odds ratio is approximately normally distributed with estimated variance

$$\hat{V}\left[\ln\left(\frac{\hat{p}_{ij}\hat{p}_{kl}}{\hat{p}_{il}\hat{p}_{kj}}\right)\right] = \frac{1}{x_{ij}} + \frac{1}{x_{kl}} + \frac{1}{x_{il}} + \frac{1}{x_{kj}}.$$

Testing Homogeneity of Proportions

Multinomial sampling is done within each population, so the sampling scheme is called product-multinomial sampling. Product-multinomial sampling is equivalent to stratified random sampling when the sampling fraction for each stratum is small or when sampling is with replacement.

The difference between product-multinomial sampling and multinomial sampling is that the row totals p_{i+} and x_{i+} are fixed quantities in product-multinomial sampling—x_{i+} is the predetermined sample size for stratum i. The null hypothesis that the proportion of observations falling in class j is the same for all strata is

$$H_0 : \frac{p_{1j}}{p_{1+}} = \frac{p_{2j}}{p_{2+}} = \cdots = \frac{p_{rj}}{p_{r+}} = p_{+j} \qquad \text{for all } j = 1, \ldots, c.$$

If the null hypothesis in (10.4) is true, again $m_{ij} = np_{i+}p_{+j}$ and the expected counts under H_0 are $\hat{m}_{ij} = np_{i+}\hat{p}_{+j}$, exactly as in the test for independence.

Test for homogeneity of proportions to test the null hypothesis that the nonresponse rate is the same for each stratum.

	Nonrespondent	Respondent	
General student	46	222	268
General tutor	41	109	150
Psychiatric student	17	40	57
Psychiatric tutor	8	26	34
	112	397	509

The two chi-square test statistics are $X^2 = 8.218$, with p-value 0.042 and $G^2 = 8.165$, with p-value 0.043. There is thus evidence of different nonresponse rates among the four groups.

Goodness of Fit testing

In the classical goodness of fit test, multinomial sampling is again assumed, with independent observations classified into k categories. The null hypothesis is

$$H_0 : p_i = p_i^{(0)} \qquad \text{for } i = 1, \ldots, k,$$

where $p_i^{(0)}$ is prespecified or is a function of parameters θ to be estimated from the data.

Example:

Number of Accidents	Number of Pilots
0	12,475
1	4,117
2	1,016
3	269
4	53
5	14
6	2
7	2

If accidents occur randomly-if no pilots are more or less “accident-prone" than others-a Poisson distribution should fit the data well.

The two chi-square test statistics are

$$X^2 = \sum_{\text{all cells}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

$$= \sum_{i=1}^{k} \frac{(n\hat{p}_i - n\hat{p}_i^{(0)})^2}{n\hat{p}_i^{(0)}}$$

$$= n \sum_{i=1}^{k} \frac{(\hat{p}_i - \hat{p}_i^{(0)})^2}{\hat{p}_i^{(0)}}$$

and

$$G^2 = 2n \sum_{i=1}^{k} \hat{p}_i \ln\left(\frac{\hat{p}_i}{\hat{p}_i^{(0)}}\right).$$

For the pilots, $X^2 = 756$ and $G^2 = 400$. If the null hypothesis is true, both statistics follow a χ^2 distribution with 4 df. Both p-values are less than 0.0001, providing evidence that a Poisson model does not fit the data.

Effects of Survey Design on Chi-Square Tests

The survey design can affect both the estimated cell probabilities and the tests of association or goodness of fit.

- Clustering: In a cluster sample with a positive intraclass correlation coefficient (ICC), the true p-value will often be much larger than the p-value reported by using the assumption of independent multinomial sampling.

  Example: Suppose that both husband and wife are asked about the household's cable and computer status for the survey, and both give the same answer. While the assumptions of multinomial sampling were met for the SRS of couples, they are not met for the cluster sample of persons-far from being independent units, the husband and wife from the same household agree completely in their answers. The ICC for the cluster sample is 1.

Example

Observed Count		Computer? Yes	No	
Cable?	Yes	238	376	614
	No	176	210	386
		414	586	1000

The estimated proportions and odds ratio are identical to those $\hat{p}_{11} = 238/1000 = 119/500$ and the odds ratio is

$$\frac{\frac{238}{376}}{\frac{176}{210}} = 0.755.$$

But $X^2 = 4.562$ and $G^2 = 4.550$.

What would be wrong?

Contingency Tables for Data from Complex Surveys

Estimate p_{ij} by

$$\hat{p}_{ij} = \frac{\sum_{k \in S} w_k y_{kij}}{\sum_{k \in S} w_k},$$

where

$$y_{kij} = \begin{cases} 1 & \text{if observation unit } k \text{ is in cell } (i,j) \\ 0 & \text{otherwise} \end{cases}$$

and w_k is the weight for observation unit k. Thus,

$$\hat{p}_{ij} = \frac{\text{sum of weights for observation units in cell } (i,j)}{\text{sum of weights for all observation units in sample}}.$$

Using the estimates $\hat{p}_{ij}$, construct the table

		C				
		1	2	$\cdots$	c	
R	1	$\hat{p}_{11}$	$\hat{p}_{12}$	$\cdots$	$\hat{p}_{1c}$	$\hat{p}_{1+}$
	2	$\hat{p}_{21}$	$\hat{p}_{22}$	$\cdots$	$\hat{p}_{2c}$	$\hat{p}_{2+}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	r	$\hat{p}_{r1}$	$\hat{p}_{r2}$	$\cdots$	$\hat{p}_{rc}$	$\hat{p}_{r+}$
		$\hat{p}_{+1}$	$\hat{p}_{+2}$	$\cdots$	$\hat{p}_{+c}$	1

to examine associations, and estimate odds ratios by $(\hat{p}_{ij}\hat{p}_{kl})/(\hat{p}_{il}\hat{p}_{kj})$. A CI for p_{ij} may be constructed by using any method of variance estimation discussed so far, or a design effect (deff) may be used to modify the SRS CI.

Effects on Hypothesis Tests and Confidence Intervals

Effects:

- Ignoring the stratification results in a conservative test, i.e., resulting in a larger p-values.
- "p-values" calculated ignoring the clustering are likely to be too small.

Ignoring clustering in chi-square tests is often more dangerous than ignoring stratification.

Example: An SRS-based chi-square test using stratified data will still indicate strong associations; it just will not uncover all weaker associations. Ignoring clustering, however, will lead to declaring associations statistically significant that really are not.

Corrections to chi-square tests

Recall that the null hypothesis of independence is

$$H_0 : p_{ij} = p_{i+}p_{+j} \qquad \text{for } i = 1, \ldots, r \quad \text{and} \quad j = 1, \ldots, c.$$

For a 2×2 table, $p_{ij} = p_{i+}p_{+j}$ for all i and j is equivalent to $p_{11}p_{22} - p_{12}p_{21} = 0$, so the null hypothesis reduces to a single equation. In general, the null hypothesis can be expressed as $(r-1)(c-1)$ distinct equations, which leads to $(r-1)(c-1)$ df for the χ^2 tests used for multinomial sampling. Let

$$\theta_{ij} = p_{ij} - p_{i+}p_{+j}.$$

Then, the null hypothesis of independence is

$$H_0 : \theta_{11} = 0, \ \theta_{12} = 0, \ \ldots, \ \theta_{r-1,c-1} = 0.$$

Wald statistic

For the 2×2 table, the null hypothesis involves one quantity,

$$\theta = \theta_{11} = p_{11} - p_{1+}p_{+1} = p_{11}p_{22} - p_{12}p_{21},$$

and θ is estimated by

$$\hat{\theta} = \hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}\hat{p}_{21}.$$

The quantity θ is a smooth function of population totals, so we estimate $V(\hat{\theta})$ using by resampling or linearizion . If the sample sizes are sufficiently large and $H_0: \theta = 0$ is true, then $\hat{\theta}/\sqrt{\hat{V}(\hat{\theta})}$ approximately follows a standard normal distribution. Equivalently, under H_0, the **Wald statistic**

$$X_W^2 = \frac{\hat{\theta}^2}{\hat{V}(\hat{\theta})}$$

approximately follows a χ^2 distribution with 1 df. In practice, we often compare X_W^2 to an F distribution with 1 and κ df, where κ is the df associated with the variance estimator. If the random group method is used to estimate the variance, then κ equals (number of groups) $-$ 1; if another method is used, κ equals (number of psus) $-$ (number of strata).

Example

A total of $n = 2588$ youths in the survey had responses for both items. The following table gives the sum of the weights for each category.

		Ever Violent?		
		No	Yes	Total
Family Member Incarcerated?	No	4,761	7,154	11,915
	Yes	4,838	7,946	12,784
	Total	9,599	15,100	24,699

This results in the following table of estimated proportions:

		Ever Violent?		
		No	Yes	Total
Family Member Incarcerated?	No	0.1928	0.2896	0.4824
	Yes	0.1959	0.3217	0.5176
	Total	0.3886	0.6114	1.0000

Thus,

$$\hat{\theta} = \hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}\hat{p}_{21} = \hat{p}_{11} - \hat{p}_{1+}\hat{p}_{+1} = 0.0053.$$

We can write $\theta = h(p_{11}, p_{12}, p_{21}, p_{22})$ and $\hat{\theta} = h(\hat{p}_{11}, \hat{p}_{12}, \hat{p}_{21}, \hat{p}_{12})$ for $h(a, b, c, d) = ad - bc$, so we can use linearization or a resampling method to estimate $V(\hat{\theta})$.

Using linearization, we obtain

$$X_W^2 = (0.0053)^2 / \hat{V}(\hat{\theta}) = 0.995$$

with p-value $= 0.32$.

Large tables: Bonferroni correction

The Bonferroni correction can be used for larger tables. In an $R \times C$ table the null hypothesis of independence,

$$H_0 : \theta_{11} = 0, \theta_{12} = 0, \ldots, \theta_{r-1,c-1} = 0,$$

has $m = (r-1)(c-1)$ components:

$$\begin{aligned} H_0(1) &: \theta_{11} = 0 \\ H_0(2) &: \theta_{12} = 0 \\ &\vdots \\ H_0(m) &: \theta_{(r-1)(c-1)} = 0. \end{aligned}$$

Instead of using the estimated covariance of all $\hat{\theta}_{ij}$'s as in the full multivariate Wald test, we can use the Bonferroni inequality to test each component $H_0(k)$ separately with significance level α/m.

H_0 will be rejected at level α if any of the $H_0(k)$ is rejected at level α/m—that is, if

$$\frac{\hat{\theta}_{ij}^2}{\hat{V}(\hat{\theta}_{ij})} > F_{1,\kappa,\alpha/m}$$

for $i = 1, \ldots, (r-1)$ and $j = 1, \ldots, (c-1)$. Each test statistic is compared to an $F_{1,\kappa}$ distribution, where the estimator of the variance has κ df.

Rao-Scott Tests: first order correction

When H_0 is true and compare the test statistic

$$X_F^2 = \frac{(r-1)(c-1)X^2}{E[X^2]}$$

or

$$G_F^2 = \frac{(r-1)(c-1)G^2}{E[G^2]}$$

to a $\chi^2_{(r-1)(c-1)}$ distribution.

Under H_0,

$$E[X^2] \approx E[G^2] \approx \sum_{i=1}^{r}\sum_{j=1}^{c}(1-p_{ij})d_{ij} - \sum_{i=1}^{r}(1-p_{i+})d_i^R - \sum_{j=1}^{c}(1-p_{+j})d_j^C, \quad (10.9)$$

where d_{ij} is the deff for estimating p_{ij}, d_i^R is the deff for estimating p_{i+}, and d_j^C is the deff for estimating p_{+j}. In practice, if the estimator of the cell variances has κ df, it works slightly better to compare $X_F^2/(r-1)(c-1)$ or $G_F^2/(r-1)(c-1)$ to an F distribution with $(r-1)(c-1)$ and $(r-1)(c-1)\kappa$ df.

Example

In the Survey of Youth in Custody, let's look at the relationship between age and whether the youth was sent to the institution for a violent offense (using variable *crimtype*, *currviol* was defined to be 1 if *crimtype* = 1 and 0 otherwise). Using the weights, we estimate the proportion of the population falling in each cell:

		Age Class			
		≤ 15	16 or 17	≥ 18	Total
Violent Offense?	No	0.1698	0.2616	0.1275	0.5589
	Yes	0.1107	0.1851	0.1453	0.4411
	Total	0.2805	0.4467	0.2728	1.0000

First, let's look at what happens if we ignore the clustering and pretend that the test statistic in (10.2) follows a χ^2 distribution with 2 df. With $n = 2621$ youths in the table, Pearson's X^2 statistic is

$$X^2 = n \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(\hat{p}_{ij} - \hat{p}_{i+}\hat{p}_{+j})^2}{\hat{p}_{i+}\hat{p}_{+j}} = 34.12.$$

The following design effects were estimated, using the stratification and clustering information in the survey:

Design Effects		Age Class ≤ 15	16 or 17	≥ 18	Total
Violent Offense?	No	14.9	4.0	3.5	6.8
	Yes	4.7	6.5	3.8	6.8
	Total	14.5	7.5	6.6	

We estimate $E[X^2]$ by 4.9 and use $X_F^2 = 2X^2/4.9 = 14.0$. Comparing 14.0/2 to an $F_{2,1690}$ distribution gives an approximate p-value of 0.001.