

Chapter 3: Inference for Contingency Tables-I

Dipankar Bandyopadhyay

Department of Biostatistics,
Virginia Commonwealth University

BIOS 625: Categorical Data & GLM

[Acknowledgements to Tim Hanson and Haitao Chu]

3.1.1 Odds Ratios

- The sample odds ratio $\hat{\theta} = n_{11}n_{22}/n_{12}n_{21}$ can be zero, undefined, or ∞ if one or more of $\{n_{11}, n_{22}, n_{12}, n_{21}\}$ are zero.
- An alternative is to add $1/2$ observation to each cell $\tilde{\theta} = (n_{11} + 0.5)(n_{22} + 0.5)/(n_{12} + 0.5)(n_{21} + 0.5)$. This also corresponds to a particular Bayesian estimate.
- Both $\hat{\theta}$ and $\tilde{\theta}$ have skewed sampling distributions with small $n = n_{++}$. The sampling distribution of $\log \hat{\theta}$ is relatively symmetric and therefore more amenable to a Gaussian approximation.
- An approximate $(1 - \alpha) \times 100\%$ CI for $\log \theta$ is given by

$$\log \hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}.$$

A CI for θ is obtained by exponentiating the interval endpoints.

- When $\hat{\theta} = 0$ this doesn't work ($\log 0 = -\infty$).
- Can use $n_{ij} + 0.5$ in place of n_{ij} in MLE estimate and standard error yielding

$$\log \tilde{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11} + 0.5} + \frac{1}{n_{12} + 0.5} + \frac{1}{n_{21} + 0.5} + \frac{1}{n_{22} + 0.5}}.$$

- Perhaps better approach would involve inverting score or LRT tests for $H_0 : \theta = \theta_0$.
- Exact approach involves testing $H_0 : \theta = t$ for various values of t subject to rows and columns fixed, and simulating a p-value. Those values of t that gives p-values greater than 0.05 defined the 95% CI. This is related to Fisher's exact test, sketched in Sections 3.5 and 16.6.4

3.1.2 Aspirin and Heart attacks

- $n = 1360$ stroke patients randomly assigned to aspirin or placebo (product multinomial sampling) & followed about 3 years.

	Heart attack	No heart attack	Total
Placebo	28	656	684 (fixed)
Aspirin	18	658	676 (fixed)

- 95% CI for $\log \theta$ using $\hat{\theta}$ is $(-0.157, 1.047)$ and so the CI for θ is $(e^{-0.157}, e^{1.047}) = (0.85, 2.85)$.
- We cannot reject that $H_0 : \theta = 1$ (at significance level $\alpha = 0.05$). We conclude that there is not enough evidence to support that heart attacks are related to aspirin intake (**Note: Absence of evidence is not evidence of absence**).
- Now, read the example in the book [Page 71].

3.1.3 Difference in proportions

- Assume (1) multinomial sampling or (2) product binomial sampling where n_{i+} are fixed (fixed row totals as in heart attack data). Let $\pi_1 = P(Y = 1|X = 1)$ and $\pi_2 = P(Y = 1|X = 2)$.
- The sample proportion for each level of X is the MLE $\hat{\pi}_1 = n_{11}/n_{1+}$, $\hat{\pi}_2 = n_{21}/n_{2+}$. Using either large sample results or the CLT we have

$$\hat{\pi}_1 \overset{\bullet}{\sim} N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_{1+}}\right) \perp \hat{\pi}_2 \overset{\bullet}{\sim} N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_{2+}}\right).$$

- Since the difference of two independent normals is also normal, we have

$$\hat{\pi}_1 - \hat{\pi}_2 \overset{\bullet}{\sim} N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_{1+}} + \frac{\pi_2(1-\pi_2)}{n_{2+}}\right).$$

- Plugging in MLEs for unknowns, we estimate the standard deviation of the difference in sample proportions by the standard error

$$\hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2) = \sqrt{\frac{\hat{\pi}_1(1 - \hat{\pi}_1)}{n_{1+}} + \frac{\hat{\pi}_2(1 - \hat{\pi}_2)}{n_{2+}}}.$$

- A Wald CI for the unknown difference has endpoints

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{\pi}_1 - \hat{\pi}_2).$$

- For the aspirin data, this yields $0.0143 \pm 1.96(0.00978)$ for the 95% CI $(-0.005, 0.033)$. How?
- $\hat{\pi}_1 - \hat{\pi}_2 = 28/684 - 18/676 = 0.0143$, and so on ...

3.1.4 Estimating relative risk

- Like the odds ratio, the relative risk $\pi_1/\pi_2 \in (0, \infty)$ and tends to have a skewed sampling distribution in small samples. Let $r = \hat{\pi}_1/\hat{\pi}_2$ be the sample relative risk. Large sample normality implies

$$\log r = \log \hat{\pi}_1/\hat{\pi}_2 \overset{\bullet}{\sim} N(\log \pi_1/\pi_2, \sigma(\log r)).$$

where

$$\sigma(\log r) = \sqrt{\frac{1 - \pi_1}{\pi_1 n_{1+}} + \frac{1 - \pi_2}{\pi_2 n_{2+}}}.$$

- Plugging in $\hat{\pi}_i$ for π_i gives the standard error and CIs are obtained as usual for $\log \pi_1/\pi_2$, then exponentiated to get the CI for π_1/π_2 .
- Applying this to the heart attack data we obtain a 95% CI for π_1/π_2 as (0.86, 2.75). The probability of a heart attack on placebo is between 0.86 and 2.75 times greater than on aspirin.

Seat Belts and Traffic Deaths Example: Page 70-71

- Read the book.
- SAS code follows
- `norow` and `nocol` remove row and column percentages from the table (not shown); these are conditional probabilities
- `measures` give estimates and CIs for odds ratio and relative risk
- `riskdiff` gives estimate and CI for $\pi_1 - \pi_2$
- `exact` plus `or` or `riskdiff` gives exact p-values for hypothesis tests of no difference and/or CIs

SAS code

```
data table;
input use$ outcome$ count @@;
datalines;
no fatal 54 no nonfatal 10325
yes fatal 25 yes nonfatal 51790
;
proc freq data=table order=data; weight count;
tables use*outcome / measures riskdiff norow nocol;
* exact or riskdiff; * exact test for H0:  $\pi_1 = \pi_2$  takes forever...;
run;
```

Inference for $\pi_1 - \pi_2$, π_1/π_2 and θ

Statistics for Table of use by outcome

Column 1 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.0052	0.0007	0.0038	0.0066	0.0039	0.0068
Row 2	0.0005	0.0001	0.0003	0.0007	0.0003	0.0007
Total	0.0013	0.0001	0.0010	0.0016	0.0010	0.0016
Difference	0.0047	0.0007	0.0033	0.0061		

Difference is (Row 1 - Row 2)

Column 2 Risk Estimates

	Risk	ASE	(Asymptotic) 95% Confidence Limits		(Exact) 95% Confidence Limits	
Row 1	0.9948	0.0007	0.9934	0.9962	0.9932	0.9961
Row 2	0.9995	0.0001	0.9993	0.9997	0.9993	0.9997
Total	0.9987	0.0001	0.9984	0.9990	0.9984	0.9990
Difference	-0.0047	0.0007	-0.0061	-0.0033		

Difference is (Row 1 - Row 2)

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	10.8345	6.7405	17.4150
Cohort (Col1 Risk)	10.7834	6.7150	17.3165
Cohort (Col2 Risk)	0.9953	0.9939	0.9967

Three CIs give three equivalent tests...

- Note that $(54/10379)/(25/51815) = 10.78$ and $(10325/10379)/(51790/51815) = 0.995$
- Col1 risk is relative risk of *dying* and Col2 risk is relative risk of *living*
- We can test for (a) $H_0 : \theta = 1$, $H_0 : \pi_1/\pi_2 = 1$, and $H_0 : \pi_1 - \pi_2 = 0$. All are equivalent, i.e., living is independent of wearing a seat belt.

Delta Method

It's probably worth reading or at least skimming 3.1.5, 3.1.6, 3.1.7, 3.1.8 (pp. 72-75). Idea of the Delta method is straightforward (see page 72 Fig. 3.1) and wildly useful.

- Let T_n be a statistic that is asymptotically normally distributed, i.e., $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$.
- Let g be a function that is at least twice differentiable at θ . Then using the Taylor series expansion for $g(t)$, we have $\sqrt{n}[g(T_n) - g(\theta)] \approx \sqrt{n}(T_n - \theta)g'(\theta)$.
- Thus, $\sqrt{n}[g(T_n) - g(\theta)] \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2)$.

Pearson and likelihood-ratio chi-squared tests

- Assume one $\text{mult}(n, \pi)$ distribution for the whole table. Let $\pi_{ij} = P(X = i, Y = j)$; we must have $\pi_{++} = 1$.
- If the table is 2×2 , we can just look at $H_0 : \theta = 1$.
- In general, independence holds if $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, or equivalently, $\mu_{ij} = n\pi_{i+}\pi_{+j}$.
- That is, independence implies a constraint; the parameters $\pi_{1+}, \dots, \pi_{I+}$ and $\pi_{+1}, \dots, \pi_{+J}$ define all probabilities in the $I \times J$ table under the constraint.

- Pearson's statistic is

$$X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}},$$

where $\hat{\mu}_{ij} = n(n_{i+}/n)(n_{+j}/n)$, the MLE under H_0 .

- There are $I - 1$ free $\{\pi_{i+}\}$ and $J - 1$ free $\{\pi_{+j}\}$. Then $IJ - 1 - [(I - 1) + (J - 1)] = (I - 1)(J - 1)$.

When H_0 is true, $X^2 \overset{\bullet}{\sim} \chi^2_{(I-1)(J-1)}$.

The LRT statistic boils down to

$$G^2 = 2 \sum_{i=1}^I \sum_{j=1}^J n_{ij} \log(n_{ij}/\hat{\mu}_{ij}),$$

and is also $G^2 \overset{\bullet}{\sim} \chi^2_{(I-1)(J-1)}$ when H_0 is true.

- $X^2 - G^2 \xrightarrow{P} 0$.
- The approximation is better for X^2 than G^2 in smaller samples.
- The approximation can be okay when some $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ are as small as 1, but most are at least 5.
- When in doubt, use small sample methods.
- Everything holds for product multinomial sampling too (fixed marginals for one variable)!

3.3.1 Pearson and standardized residuals

- Rejecting $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ does not tell us about the nature of the association.
- The Pearson residual is

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}},$$

where, as before, $\hat{\mu}_{ij} = n_{i+}n_{+j}/n$ is the estimate under $H_0 : X \perp Y$.

- When $H_0 : X \perp Y$ is true, under multinomial sampling $e_{ij} \stackrel{\bullet}{\sim} N(0, v)$, where $v < 1$, in large samples.
- Note that $\sum_{i=1}^I \sum_{j=1}^J e_{ij}^2 = X^2$.

- Standardized Pearson residuals are Pearson residuals divided by their standard error under multinomial sampling.

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

where $p_{i+} = n_{i+}/n$ and $p_{+j} = n_{+j}/n$ are MLEs under the null model.

- Values of $|r_{ij}| > 3$ happen very rarely when $H_0 : X \perp Y$ is true and $|r_{ij}| > 2$ happen only roughly 5% of the time.
- Pearson residuals and their standardized version tell us which cell counts are much larger or smaller than what we would expect under $H_0 : X \perp Y$.
- Example:** we analyze Table 3.2 in Agresti 2002 with the following SAS code, modified from Alan Agresti's website...

Table 3.2 Frequency of Education and Religious Beliefs

Highest degree	Religious beliefs		
	Fundamentalist	Moderate	Liberal
Less than high school	178	138	108
High school or junior college	570	648	442
Bachelor or graduate	138	252	252

```

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 format;
value $dc
  '1' = '< HS'
  '2' = 'HS or JC'
  '3' = '>= BA/BS';
value $rc
  'fund' = 'Fundamentalist'
  'mod' = 'Moderate'
  'lib' = 'Liberal';
proc freq order=data; weight count;
  format Religion $rc. Degree $dc.;
  tables Degree*Religion / chisq expected measures cmh1;
proc genmod order=data; class Degree Religion;
  format Religion $rc. Degree $dc.;
  model count = Degree Religion / dist=poi link=log residuals;
run;

```

Annotated output from proc freq:

Degree	Religion			
Frequency				
Expected				
Percent				
Row Pct				
Col Pct	Fundamen	Moderate	Liberal	Total
	talist			
< HS	178	138	108	424
	137.81	161.45	124.74	
	6.53	5.06	3.96	15.55
	41.98	32.55	25.47	
	20.09	13.29	13.47	
HS or JC	570	648	442	1660
	539.53	632.09	488.38	
	20.91	23.77	16.21	60.90
	34.34	39.04	26.63	
	64.33	62.43	55.11	
>= BA/BS	138	252	252	642
	208.66	244.46	188.88	
	5.06	9.24	9.24	23.55
	21.50	39.25	39.25	
	15.58	24.28	31.42	
Total	886	1038	802	2726
	32.50	38.08	29.42	100.00

More...

Statistics for Table of Degree by Religion

Statistic	DF	Value	Prob
Chi-Square	4	69.1568	<.0001
Likelihood Ratio Chi-Square	4	69.8116	<.0001

Annotated output from proc genmod:

The GENMOD Procedure

Observation Statistics

	Raw	Pearson	Deviance	Std	Pearson	Likelihood
Observation	Residual	Residual	Residual	Deviance Residual	Residual	Residual
1	40.192213	3.4237736	3.2748138	4.3376139	4.5349167	4.4235336
2	-23.44974	-1.845523	-1.893142	-2.618003	-2.552151	-2.586795
3	-16.74248	-1.499038	-1.534598	-1.987766	-1.941705	-1.969288
4	30.469522	1.3117699	1.2997048	2.5297817	2.5532655	2.5470879
5	15.909037	0.632782	0.630155	1.280585	1.2859234	1.2846328
6	-46.37857	-2.098646	-2.133249	-4.060564	-3.994696	-4.012984
7	-70.66184	-4.891741	-5.216165	-7.261384	-6.809756	-7.046419
8	7.5406572	0.4822874	0.4798392	0.6974074	0.7009655	0.6992834
9	63.121006	4.5928481	4.3672118	5.9453678	6.2525411	6.0887236

- The Std Pearson Residual column has the r_{ij} . Values larger than 3 in magnitude indicate severe departures from independence.
- Observations 1 and 9, corresponding to “less than high school, fundamentalist” and “at least BS/BA, liberal” are over-represented relative to independence.
- Observations 6 and 7, corresponding to “HS or JC, liberal” and “at least BS/BA, fundamentalist” are under-represented.
- That is, we tend to see concentrations along the diagonal, so increased education is associated with increasingly liberal religious views.
- These data are ordinal; part of `proc freq` output is the γ statistic:

Statistics for Table of Degree by Religion		
Statistic	Value	ASE
Gamma	0.2178	0.0281

We see a moderate, positive association.

Education and Religious Beliefs

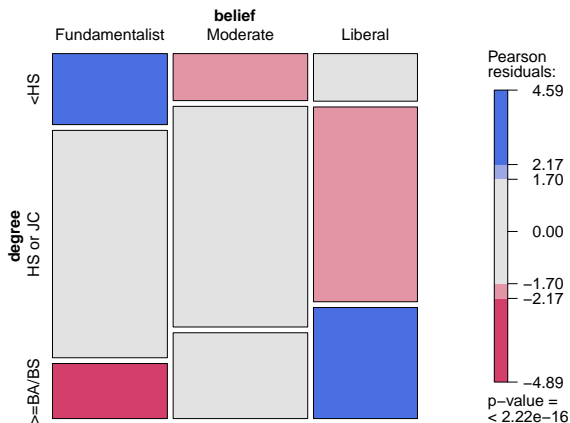


Figure : Mosaic Plot: Education by Religion

3.3.3 Partitioning Chi-squared

- Recall from ANOVA the partitioning of SS Treatments via orthogonal contrasts. We can do something similar with contingency tables.
- A χ^2_ν random variable X^2 can be written

$$X^2 = Z_1^2 + Z_2^2 + \cdots + Z_\nu^2,$$

where Z_1, \dots, Z_ν are *iid* $N(0, 1)$ & so Z_1^2, \dots, Z_ν^2 are *iid* χ_1^2 .

- Partitioning works by testing independence in a series of (collapsed) sub-tables in a particular way.
- Say t tests are performed. The i^{th} test results in G_i^2 with associated degrees of freedom $df_i = \nu_i$. Then

$$G_1^2 + G_2^2 + \cdots + G_t^2 = G^2,$$

the LRT statistic from testing independence in the overall $I \times J$ table.

- Also, $\nu_1 + \nu_2 + \cdots + \nu_t = (I - 1)(J - 1)$, the degrees of freedom for the overall test.

- One approach is to look at a series of $\nu = (I - 1)(J - 1)$ 2×2 tables (pp. 82-83) of the form:

$$\frac{\sum_{a < i} \sum_{b < j} n_{ab}}{\sum_{b < j} n_{ij}} \quad \bigg| \quad \frac{\sum_{a < i} n_{aj}}{n_{ij}}$$

for $i = 2, \dots, I$ and $j = 2, \dots, J$. Each sub-table will have $df \nu_{ij} = 1$ and $\sum_{i=2}^I \sum_{j=2}^J G_{ij}^2 = G^2$ from the overall LRT.

- Example:** Origin of schizophrenia (pp. 83-84)

Psych school	Schizophrenia origin		
	Biogenic	Environmental	Combination
Eclectic	90	12	78
Medical	13	1	6
Psychoanalytic	19	13	50

- For the full table, testing $H_0 : X \perp Y$ yields $G^2 = 23.036$ on 4 df , so $p < 0.001$.

- When we consider (Lancaster) partitioning, we get 4 tables:

Ecl Med	Bio	Env	$\hat{\theta}_{11} = 0.58$
	90	12	$G_{11}^2 = 0.294$
	13	1	$p = 0.59$
Ecl Med	Bio+Env	Com	$\hat{\theta}_{12} = 0.56$
	102	78	$G_{12}^2 = 1.359$
	14	6	$p = 0.24$
Ecl+Med Psy	Bio	Env	$\hat{\theta}_{21} = 5.4$
	103	13	$G_{21}^2 = 12.953$
	19	13	$p = 0.0003$
Ecl+Med Psy	Bio+Env	Com	$\hat{\theta}_{22} = 2.2$
	116	84	$G_{22}^2 = 8.430$
	32	50	$p = 0.004$

- Note that: $0.294 + 1.359 + 12.953 + 8.430 = 23.036$ as required.
Also: $1 + 1 + 1 + 1 = 4$.

The last two tables contribute more than 90% of the G^2 statistic.

- The first two tables suggest that eclectic and medical schools of thought tend to classify the origin of schizophrenia in roughly the same proportions.
- The last two tables suggest a difference in how the psychoanalytic school classifies the origin relative to eclectic and medical schools.
- The odds of a member of the psychoanalytical school ascribing the origin to be a combination (versus biogenic or environmental) is about 2.2 times greater than medical or eclectic. Within the last two origins, the odds of a member of the psychoanalytical school ascribing the origin to be a environmental is about 5.4 times greater than medical or eclectic.

Comments:

- Lancaster partitioning looks at a lot of tables. There might be natural, simpler groupings of X and Y levels to look at. See your text for advice and discussion on partitioning.
- Partitioning G^2 and standardized Pearson residuals are two tools to help find where association occurs in a table once $H_0 : X \perp Y$ is rejected.
- There are better methods for ordinal data, the subject of the next lecture.
- There are also exact tests of $H_0 : X \perp Y$ which we'll briefly discuss next time as well.