Sections 3.1, 3.2, 3.3

Timothy Hanson

Department of Statistics, University of South Carolina

Stat 770: Categorical Data Analysis

3.3.1 Odds ratio, SE, & CI

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.

Alternative CIs

- 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

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

- Exact approach involves testing $H_0 : \theta = t$ for various values of t subject to rows or columns fixed and simulating a p-value. Those values of t that give p-values greater than 0.05 define the 95% CI. This is related to Fisher's exact test, sketched out in Sections 3.5 and 16.6.4.

3.1.4 Aspirin and heart attacks

The following 2×2 contingency table is from a report by the Physicians' Health Study Research Group on $n = 22,071$ physicians that took either a placebo or aspirin every other day.

	Fatal attack	Nonfatal or no attack
Placebo	18	11,016
Aspirin	5	11,032

Here $\hat{\theta} = \frac{18 \times 11032}{5 \times 11016} = 3.605$ and $\log \hat{\theta} = \log 3.605 = 1.282$, and $\text{se}\{\log(\hat{\theta})\} = \sqrt{\frac{1}{18} + \frac{1}{11016} + \frac{1}{5} + \frac{1}{11032}} = 0.506$.

A 95% CI for θ is then $\exp\{1.282 \pm 1.96(0.506)\} = (e^{1.282-1.96(0.506)}, e^{1.282+1.96(0.506)}) = (1.34, 9.72)$.

3.1.3 Difference in proportions & relative risk

Assume (1) multinomial sampling or (2) product binomial sampling. The row totals n_{i+} are fixed (e.g. prospective study or clinical trial) 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

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

$se(\hat{\pi}_1 - \hat{\pi}_2)$ and CI

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

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

A Wald CI for the unknown difference has endpoints

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

For the aspirin and heart attack data,
$\hat{\pi}_1 = 18/(18 + 11016) = 0.00163$ and
$\hat{\pi}_2 = 5/(5 + 11032) = 00045.$

The estimated difference is $\hat{\pi}_1 - \hat{\pi}_2 = 0.00163 - 00045 = 0.0012$ and $se(\hat{\pi}_1 - \hat{\pi}_2) = 0.00043$ so a 95% CI for $\pi_1 - \pi_2$ is $0.0012 \pm 1.96(0.00043) = (0.0003, 0.0020)$.

Relative risk

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

$$\log r = \log \hat{\pi}_1/\hat{\pi}_2 \overset{\bullet}{\sim} N(\log \pi_1/\pi_2, \sigma^2(\log r)).$$

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.

For the aspirin and heart attack data, the estimated relative risk is $\hat{\pi}_1/\hat{\pi}_2 = 0.00163/0.00045 = 3.60$ and $se\{\log(\hat{\pi}_1/\hat{\pi}_2)\} = 0.505$, so a 95% CI for π_1/π_2 is $\exp\{\log 3.60 \pm 1.96(0.505)\} = (e^{\log 3.60 - 1.96(0.505)}, e^{\log 3.60 + 1.96(0.505)}) = (1.34, 9.70)$.

3.1.2 Seat-belts and traffic deaths

Car accident fatality records for children < 18, Florida 2008.

	Injury outcome		
Seat belt use	Fatal	Non-fatal	Total
No	54	10,325	10,379
Yes	25	51,790	51,815

- $\hat{\theta} = 54(51790)/[10325(25)] = 10.83$.
- $se(\log \hat{\theta}) = 0.242$.
- 95% CI for $\hat{\theta}$ is $(\exp\{\log(10.83) - 1.96(0.242)\}, \exp\{\log(10.83) + 1.96(0.242)\}) = (6.74, 17.42)$.
- We reject that $H_0 : \theta = 1$ (at level $\alpha = 0.05$). We reject that seatbelt use is not related to mortality.

SAS code

- `norow` and `nocol` remove row and column percentages from the table (not shown); these are conditional probabilities.
- `measures` gives 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.

```
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: pi1=pi2 takes forever...;
run;
```

SAS output: 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 all of $H_0 : \theta = 1$, $H_0 : \pi_1/\pi_2 = 1$, and $H_0 : \pi_1 - \pi_2 = 0$. *All* of these null hypotheses are equivalent to $H_0 : \pi_1 = \pi_2$, i.e. living is independent of wearing a seat belt.

A final method for testing independence is coming up in Section 3.2 that generalizes to larger $I \times J$ tables.

Delta method

It's probably worth reading or at least skimming 3.1.5, 3.1.6, 3.1.7 (pp. 72-75).

Idea is straightforward (see Fig. 3.1) & wildly useful.

Delta method is how we obtain the standard errors for $\log \hat{\theta}$ and $\log(\hat{\pi}_1/\hat{\pi}_2)$ on previous slides.

3.2 Testing independence in $I \times J$ tables

Assume one mult$(n, \boldsymbol{\pi})$ 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 $\pi_{1+}, \ldots, \pi_{I+}$ and $\pi_{+1}, \ldots, \pi_{+J}$ define all probabilities in the $I \times J$ table under the constraint.

Pearson statistic

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

This is an example of the approach in 1.5.5.

Likelihood ratio statistic

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)!

SAS code: tests for independence, seat-belt data

- `chisq` gives X^2 and G^2 tests for independence (coming up in these slides).
- `expected` gives expected cell counts under independence.
- `exact` plus `chisq` gives exact p-values for testing independence using X^2 and G^2.

```
proc freq data=table order=data; weight count;
 tables use*outcome / chisq norow nocol expected;
 exact chisq;
run;
```

SAS output: table and asymptotic tests for independence

The FREQ Procedure

Table of use by outcome

use	outcome		
Frequency | Expected | Percent	fatal	nonfatal	Total
no	54	10325	10379
	13.184	10366	
	0.09	16.60	16.69
yes	25	51790	51815
	65.816	51749	
	0.04	83.27	83.31
Total	79	62115	62194
	0.13	99.87	100.00

Statistics for Table of use by outcome

Statistic	DF	Value	Prob
Chi-Square	1	151.8729	<.0001
Likelihood Ratio Chi-Square	1	104.0746	<.0001

SAS output: exact tests for independence

Pearson Chi-Square Test	
Chi-Square	151.8729
DF	1
Asymptotic Pr > ChiSq	<.0001
Exact Pr >= ChiSq	2.663E-24

Likelihood Ratio Chi-Square Test	
Chi-Square	104.0746
DF	1
Asymptotic Pr > ChiSq	<.0001
Exact Pr >= ChiSq	2.663E-24

These test the null H_0 that wearing a seat belt is independent of living. What do we conclude?

Obtaining p-values for exact tests are discussed in detail in Section 16.5.

3.2.2 Belief in God, a 3×6 table

Highest degree	Belief in God					
	Don't believe	No way to find out	Some higher power	Believe sometimes	Believe but doubts	Know God exists
Less than high school	9	8	27	8	47	236
High school or junior college	23	39	88	49	179	706
Bachelor or graduate	28	48	89	19	104	293

General Social Survey data cross-classifies opinion on whether God exists by highest education degree obtained.

SAS code, belief in God data

```
data table;
input degree$ belief$ count @@;
datalines;
1 1   9 1 2   8 1 3  27 1 4   8 1 5  47 1 6 236
2 1  23 2 2  39 2 3  88 2 4  49 2 5 179 2 6 706
3 1  28 3 2  48 3 3  89 3 4  19 3 5 104 3 6 293
;
proc format; value $dc
 '1' = 'less than high school'
 '2' = 'high school or junior college'
 '3' = 'bachelors or graduate';
value $bc
 '1' = 'dont believe'
 '2' = 'no way to find out'
 '3' = 'some higher power'
 '4' = 'believe sometimes'
 '5' = 'believe but doubts'
 '6' = 'know God exists';
run;
proc freq data=table order=data; weight count;
 format degree $dc. belief $bc.;
 tables degree*belief / chisq expected norow nocol;
run;
```

Annotated output from proc freq

degree	belief					
Frequency Expected Percent	dont believe	no way to find out	some higher power	believe sometimes	believe but doubts	know God exists	Total
less than high school	9 10.05 0.45	8 15.913 0.40	27 34.17 1.35	8 12.73 0.40	47 55.275 2.35	236 206.86 11.80	335 16.75
high school or junior college	23 32.52 1.15	39 51.49 1.95	88 110.57 4.40	49 41.192 2.45	179 178.86 8.95	706 669.37 35.30	1084 54.20
bachelors or graduate	28 17.43 1.40	48 27.598 2.40	89 59.262 4.45	19 22.078 0.95	104 95.865 5.20	293 358.77 14.65	581 29.05
Total	60 3.00	95 4.75	204 10.20	76 3.80	330 16.50	1235 61.75	2000 100.00

Statistics for Table of degree by belief

Statistic	DF	Value	Prob
Chi-Square	10	76.1483	<.0001
Likelihood Ratio Chi-Square	10	73.1879	<.0001

Statistic	Value	ASE
Gamma	-0.2483	0.0334

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

3.3.1 Pearson and standardized residuals

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 $e_{ij} \overset{\bullet}{\sim} N(0, v)$, where $v < 1$, in large samples.

Note that $\sum_{i=1}^{I} \sum_{j=1}^{J} e_{ij}^2 = X^2$.

Standardized Pearson residuals

Standardized Pearson residuals are Pearson residuals divided by their standard error under multinomial sampling (see Chapter 14).

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

where $p_{ij} = n_{ij}/n$ are MLEs under the full (non-independence) 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 $H_0 : X \perp Y$.

Residuals, belief in God data

Annotated output from `proc genmod`:

```
proc genmod order=data; class degree belief;
 model count = degree belief / dist=poi link=log residuals;
run;
```

The GENMOD Procedure

Observation	Raw Residual	Pearson Residual	Deviance Residual	Std Deviance Residual	Std Pearson Residual	Likelihood Residual
1	-1.050027	-0.33122	-0.337255	-0.375301	-0.368586	-0.374018
2	-7.912722	-1.983598	-2.196043	-2.466133	-2.227559	-2.41867
3	-7.17002	-1.226585	-1.273736	-1.473157	-1.418624	-1.459585
4	-4.730002	-1.325706	-1.423967	-1.591184	-1.481383	-1.569931
5	-8.275002	-1.113022	-1.142684	-1.370537	-1.33496	-1.35979
6	29.137492	2.0258686	1.9809013	3.5103847	3.5900719	3.5648903
7	-9.520085	-1.669418	-1.762793	-2.644739	-2.504646	-2.567827
8	-12.49071	-1.740695	-1.819318	-2.754505	-2.635467	-2.688045
9	-22.56805	-2.146245	-2.226274	-3.471424	-3.346635	-3.398513
10	7.8079994	1.2165594	1.1808771	1.7790347	1.8327913	1.8093032
11	0.1400133	0.0104692	0.0104678	0.016927	0.0169292	0.0169284
12	36.630048	1.4158081	1.403181	3.3524702	3.3826387	3.3773731
13	10.56995	2.5317662	2.3247777	2.8023308	3.0518386	2.8824417
14	20.402111	3.883624	3.51114	4.2710987	4.724204	4.4230839
15	29.737956	3.862983	3.5931704	4.5015643	4.8395885	4.6270782
16	-3.078006	-0.655073	-0.671253	-0.812499	-0.792914	-0.806333
17	8.1349809	0.8308573	0.8195034	1.0647099	1.0794611	1.0707466
18	-65.76757	-3.472204	-3.587324	-6.88618	-6.665198	-6.725887

Direction and 'significance' of standardized Pearson residuals r_{ij}

$|r_{ij}| > 3$ indicate severe departures from independence; these are in boxes below.

−	−	−	−	−	+
−	−	−	+	+	+
+	+	+	−	+	−

Which cells are over-represented relative to independence? Which are under-represented? In general, what can one say about belief in God and education? Does this correspond with the γ statistic?

Also see mosaic plot on p. 82.

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 $Z_1, \ldots, Z_\nu$ are *iid* $N(0,1)$ & so $Z_1^2, \ldots, Z_\nu^2$ are *iid* χ^2_1.
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. 81-83) of the form:

$$\begin{array}{c|c} \sum_{a<i} \sum_{b<j} n_{ab} & \sum_{a<i} n_{aj} \\ \hline \sum_{b<j} n_{ij} & n_{ij} \end{array}$$

for $i = 2, \ldots, I$ and $j = 2, \ldots, J$. 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 (p. 83)

	Schizophrenia origin		
Psych school	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

	Bio	Env	$\hat{\theta}_{11} = 0.58$
Ecl	90	12	$G^2_{11} = 0.294$
Med	13	1	$p = 0.59$

	Bio+Env	Com	$\hat{\theta}_{12} = 0.56$
Ecl	102	78	$G^2_{12} = 1.359$
Med	14	6	$p = 0.24$

	Bio	Env	$\hat{\theta}_{21} = 5.4$
Ecl+Med	103	13	$G^2_{21} = 12.953$
Psy	19	13	$p = 0.0003$

	Bio+Env	Com	$\hat{\theta}_{22} = 2.2$
Ecl+Med	116	84	$G^2_{22} = 8.430$
Psy	32	50	$p = 0.004$

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

Analysis...

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. I included them on slide 18 to show how SAS returns the results.