Categorical Data Analysis for Survey Data

Professor Ron Fricker
Naval Postgraduate School
Monterey, California

Goals for this Lecture

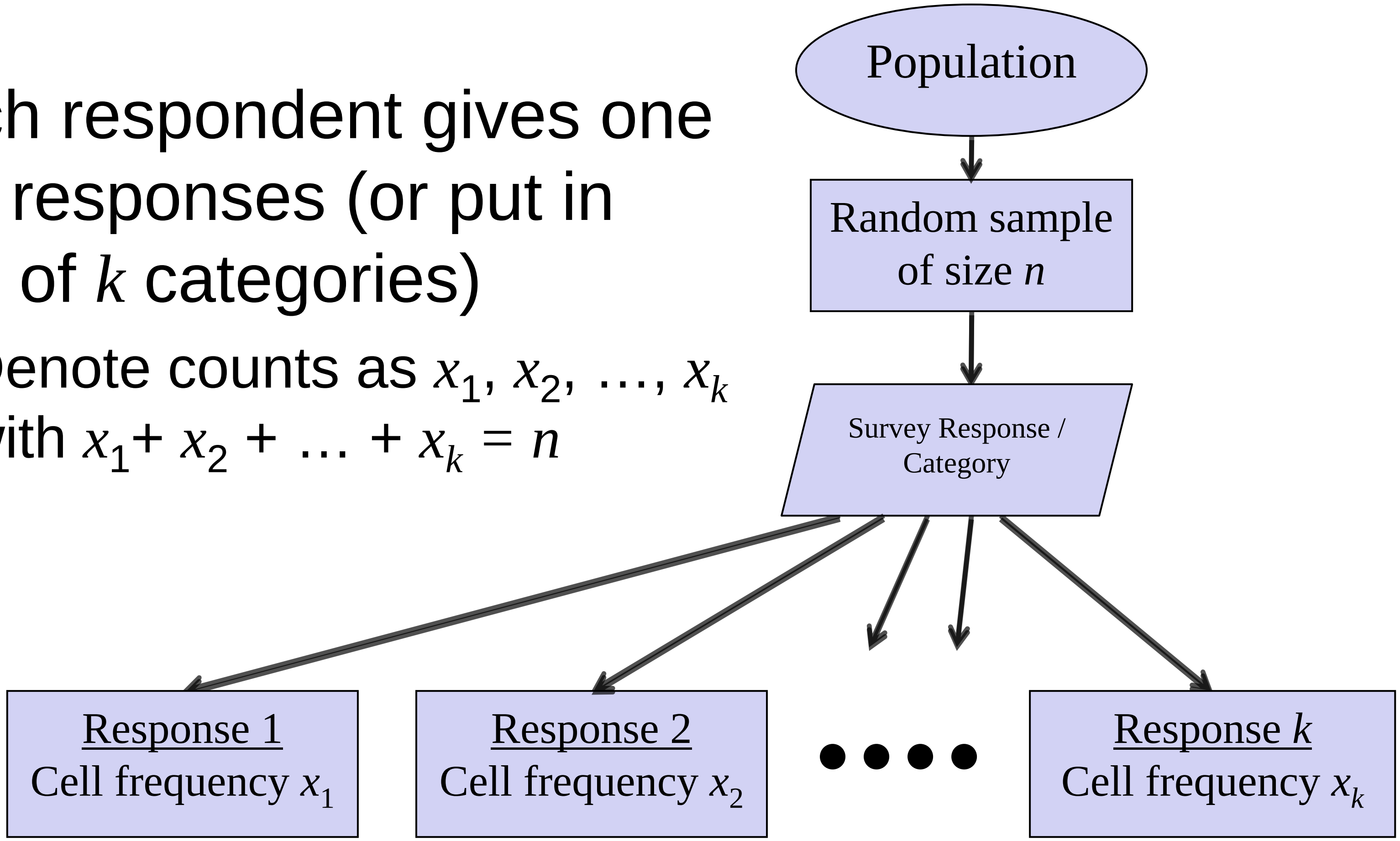
- Under SRS, be able to conduct tests for discrete contingency table data
  - One-way chi-squared goodness-of-fit tests
  - Two-way chi-squared tests of independence
- Understand how complex survey designs affect chi-squared tests
  - Discuss some ways to correct

Classical Statistical Assumptions…

- …apply for SRS with large population:
  - Very large (infinite) population
  - Sample is small fraction of population
  - Sample is drawn from population via SRS
- …but not with complex sampling
- Hence, standard statistical software generally works for SRS survey designs, but not for complex designs

One-Way Classifications

- Each respondent gives one of k responses (or put in one of k categories)
  - Denote counts as $x_1, x_2, \ldots, x_k$ with $x_1 + x_2 + \ldots + x_k = n$



One-Way Goodness-of-Fit Test

- Have counts for k categories, $x_1, x_2, \ldots, x_k$, with $x_1 + x_2 + \ldots + x_k = n$
- (Unknown) population cell probabilities denoted $p_1, p_2, \ldots, p_k$ with $p_1 + p_2 + \ldots + p_k = 1$
- Estimate each cell probability from the observed counts: $\hat{p}_i = x_i / n,\ i = 1, 2, \ldots, k$
- The hypotheses to be tested are

$$H_0 : p_1 = p_1^*, p_2 = p_2^*, \ldots, p_k = p_k^*$$

$$H_a : \text{at least one } p_i \neq p_i^*$$

One-Way Goodness-of-Fit Test for Homogeneity

- Null hypothesis is the probability of each category is equally likely: $p_i^* = 1/k,\ i = 1, 2, \ldots, k$
  - I.e., the distribution of category characteristics is homogeneous in the population
- If the null is true, in each cell (in a perfect world) we would expect to observe $e_i = np_i^*$ counts
- To do a statistical test, must assess how "far away" the e_i expected counts are from the x_i observed counts

Chi-squared Test

- Idea: Look at how far off table counts are from what is expected under the null
- Pearson chi-square test statistic:

$$X^2 = \sum_{i=1}^{k} \frac{(\text{observed - expected})^2}{\text{expected}}$$

$$= \sum_{i=1}^{k} \frac{(x_i - n/k)^2}{n/k}$$

Alternate Test Statistics

- Likelihood ratio test statistic:

$$G^2 = 2\sum_{i=1}^{k} \text{observed} \times \ln\left(\frac{\text{observed}}{\text{expected}}\right)$$

$$= 2\sum_{i=1}^{k} x_i \times \ln\left(\frac{x_i}{n/k}\right)$$

- Pearson and likelihood ratio test statistics asymptotically equivalent
- For either statistic, reject if too large
  – Assess "too large" using chi-squared dist'n

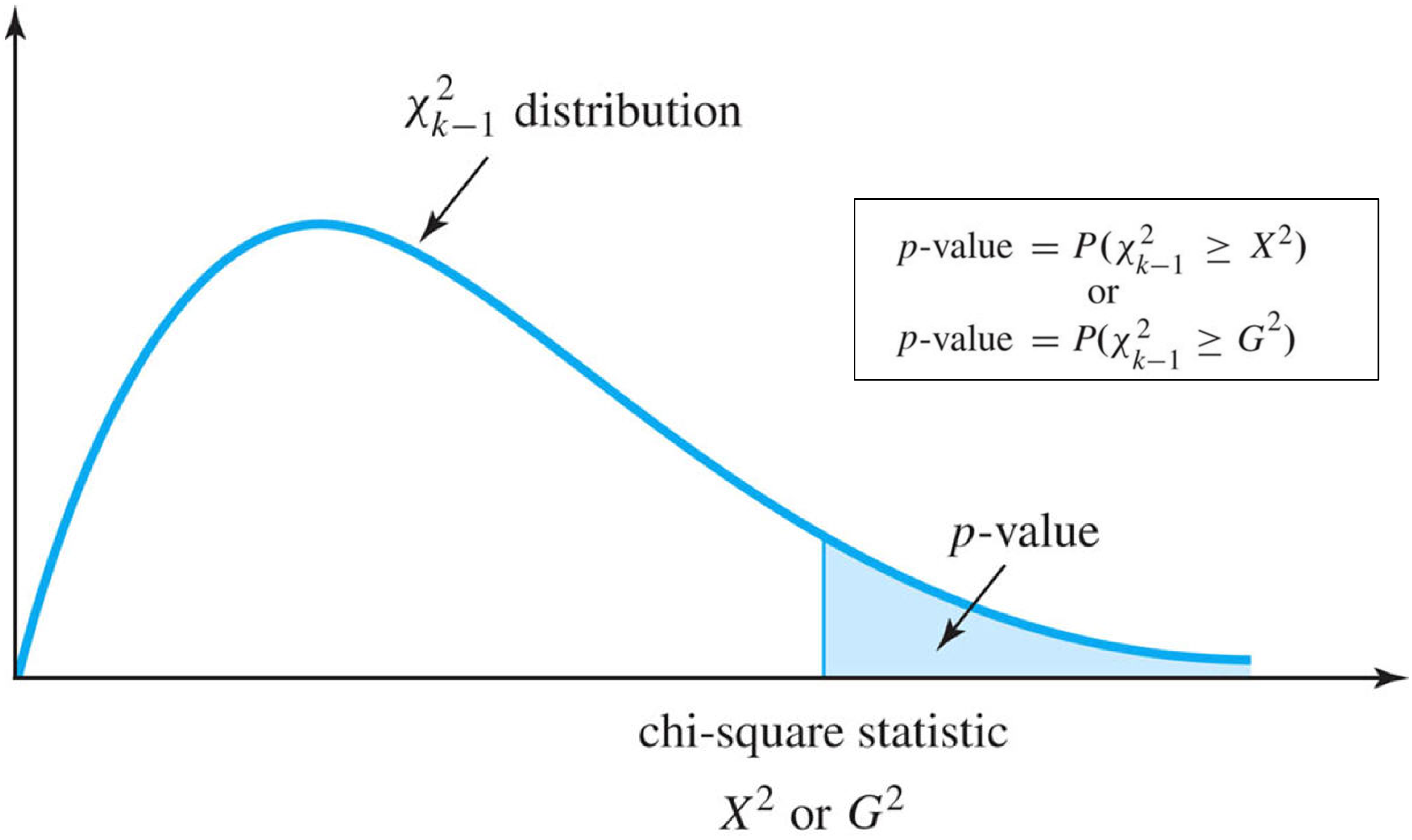
Conducting the Statistical Test

- First calculate X^2 or G^2 statistic
- Then calculate the p-value; e.g.,

$$p\text{-value} = \Pr(\chi^2_{k-1} \geq X^2)$$

- χ^2_{k-1} is the chi-squared distribution with k-1 degrees of freedom
- Reject null if p-value < α, for some pre-determined significance level α

p-value Calculation for Chi-square Goodness-of-Fit Tests



Simple Example

- Respondents were equally likely to choose any answer on a 5-point Likert scale question
  - Survey results:

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
21	15	19	20	17

n=92

  - Expected under the null:

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
18.4	18.4	18.4	18.4	18.4

  - Pearson test statistic:

$$X^2 = \sum_{i=1}^{5} \frac{(x_i - 18.4)^2}{18.4} = 1.26$$

  - p-value:

$$\Pr(\chi^2_{\nu=4} \geq 1.26) = 0.87$$

Goodness-of-Fit Tests for Other Distributions

- Homogeneity is just a special case
- Can test whether the p_i^*s are anything so long as $$\sum_{i=1}^{k} p_i^* = 1$$
- Might have some theory that says what the distribution should be, for example
- Remember, don't look at that data first and then specify the probabilities…

Simple Example

- Theoretical response distribution for question:

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
0.05	0.2	0.5	0.2	0.05

  - Survey results:

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
20	60	132	47	17

$n=276$

  - Expected under the null:

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
13.8	55.2	138	55.2	13.8

  - Pearson test statistic:

$$X^2 = \sum_{i=1}^{5} \frac{(x_i - e_i)^2}{e_i} = 5.42$$

  - p-value:

$$\Pr(\chi^2_{\nu=4} \geq 5.42) = 0.25$$

Doing the Test in JMP and Excel

- Analyze > Distribution
  - Put nominal variable in “Y, Columns” > OK
  - Red triangle > “Test Probabilities” > fill in probabilities to test
- Note that if you look in the JMP help, you will find “goodness of fit” tests
  - Different test for regression – don’t use
- Chi-square tests also easy to do in Excel
  - CHIDIST function useful for calculating p-values

A Couple of Notes

- Likelihood ratio and Pearson test statistics usually very close
  - I tend to focus on Pearson
  - JMP gives both as output
- Note that Pearson test depends on all cells having sufficiently large expected counts: $e_i = np_i^* \geq 5$
  - If not, collapse across some categories

Chi-square Test of Independence

- Survey of 500 households
  - Two of the questions:
    - Do you own at least one personal computer?
    - Do you subscribe to cable television

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

Some Notation for Two-Way Contingency Tables

- Table has r rows and c columns
- Observed cell counts are x_{ij}, with

$$\sum_{i=1}^{r}\sum_{j=1}^{c} x_{ij} = n$$

- Denote row sums: $x_{i+} = \sum_{j=1}^{c} x_{ij}, \quad i = 1, \ldots, r$
- Denote column sums: $x_{+j} = \sum_{i=1}^{r} x_{ij}, \; j = 1, \ldots, c$

The Hypotheses

- Independence means, for all cells in the table, $p_{ij} = p_{i+} p_{+j}$ where
  - p_{i+} is the probability of having row i characteristic
  - p_{+j} is the probability of having column j characteristic
- The hypotheses to be tested are

$$H_0 : p_{ij} = p_{i+} p_{+j}, \; i = 1, 2, \ldots, r; \; j = 1, 2, \ldots, c$$

$$H_a : p_{ij} \neq p_{i+} p_{+j}, \text{ for some } i \text{ and } j$$

Pearson Chi-square Test Statistic

- Test statistic:

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(x_{ij} - e_{ij})^2}{e_{ij}}$$

- Under the null, the expected count is calculated as

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

$$= \frac{x_{i+} \times x_{+j}}{n}$$

Back to the Example

- Assuming independence, we have

Observed counts:

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

<u>Example</u>

$$\hat{p}_{\text{Yes},+} = \frac{x_{\text{Yes},+}}{n} = \frac{307}{500} = 0.614$$

$$\hat{p}_{+,\text{No}} = \frac{x_{+,\text{No}}}{n} = \frac{293}{500} = 0.586$$

$$\begin{aligned} e_{i=\text{Yes},\, j=\text{No}} &= n\hat{p}_{\text{Yes},+}\,\hat{p}_{+,\text{No}} \\ &= 500 \times 0.614 \times 0.586 \\ &= 179.9 \end{aligned}$$

Expected counts:

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

Doing the Calculations

- Proceed as with the goodness-of-fit test
  - Except degrees of freedom are $\nu = (r-1)(c-1)$
- Large values of the chi-squared statistic are evidence that the null is false
- JMP does the p-value calculation (as do all stat software packages
  - Reject null if p-value < α, for some pre-determined significance level α

Conducting the Test in JMP

- Ensure variables coded as nominal
- JMP > Analyze > Fit Y by X, put:
  – One variable in as X
  – Other variable in as Y

Tests

N	DF	-LogLike	RSquare (U)
500	1	1.1374807	0.0034

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	2.275	0.1315
Pearson	2.281	0.1310

p-values

Complex Surveys: What's the Problem?

- Chi-square distribution of test statistic results from SRS assumption
- Complex survey designs result in incorrect p-values
  - E.g., Clustered sample designs can result in incorrectly low p-values
    - With high intra-class correlation (ICC) it's as if the sample size has been artificially inflated

Revisiting Computer/Cable Example

- What if interviewed two individuals in each house and got same answers?

New data:

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

$$X^2 = 4.562$$
$$p\text{-value} = 0.03$$

Original data:

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

$$X^2 = 2.281$$
$$p\text{-value} = 0.13$$

The Issue

- In complex surveys table counts unlikely to reflect relative frequencies of the categories in the population
  - Unless sample is self-weighting
- I.e., can't just plug counts into X^2 or G^2 calculations:

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

Effects of Stratified Sampling Design on Hypothesis Tests and CIs (1)

- If rows in contingency table correspond to strata, usual chi-square test of homogeneity fine
  - But may want to test association between other (non-strata) factors
- In general, stratification increases precision of estimates
  - E.g., stratified sample of size n gives same precision for estimating p_{ij} as a SRS of size $n \,/\, d_{ij}$, where d_{ij} is the design effect

Effects of Stratified Sampling Design on Hypothesis Tests and CIs (2)

- Thus p-values for chi-square tests with stratification are conservative
  - E.g., actual p-value will be smaller than calculated p-value
  - Means if null rejected, it is appropriate
    - However, if don't reject but close, how to tell if null should be rejected?

Effects of Clustered Sampling Design on Hypothesis Tests and CIs

- Opposite effect from stratification: p-values artificially low
  - Means if fail to reject null, it is appropriate
    - However, if do reject null, how to tell if null really should be rejected?
- Clustering unaccounted for (in surveys or other data collection) can result in spurious "significant" results

Corrections to Chi-square Tests

- There are a number of ways to fix:
  - Wald tests
  - Bonferroni tests
  - Matching moments
  - Model-based methods
- See Lohr for the last two – we won't cover here

Think of Problem in Terms of Cell Probabilities (1)

- Use sampling weights to estimate population quantity

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

where

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

- Thus

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

Think of Problem in Terms of Cell Probabilities (2)

- So, using the $\hat{p}_{ij}$, construct the table
- Can express the test statistics as

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

$$X^2 = \sum_{\substack{\text{All}\\\text{cells}}} \frac{(\text{observed - expected})^2}{\text{expected}} = \sum_{\substack{\text{All}\\\text{cells}}} \frac{(n\hat{p}_{ij} - np_{ij})^2}{np_{ij}} = n \sum_{\substack{\text{All}\\\text{cells}}} \frac{(\hat{p}_{ij} - p_{ij})^2}{p_{ij}}$$

$$G^2 = 2 \sum_{\substack{\text{All}\\\text{cells}}} \text{observed} \times \ln\left(\frac{\text{observed}}{\text{expected}}\right) = 2n \sum_{\substack{\text{All}\\\text{cells}}} \hat{p}_{ij} \ln\left(\frac{\hat{p}_{ij}}{p_{ij}}\right)$$

Wald Tests (1)

- For a 2x2 table, null hypothesis of independence is $p_{ij} = p_{i+}p_{+j},\ 1 \le i, j \le 2$
- This is equivalent to testing

$$H_0 : p_{11}p_{22} - p_{12}p_{21} = 0$$

$$H_a : p_{11}p_{22} - p_{12}p_{21} \ne 0$$

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

Wald Tests (2)

- Then for large samples, under the null

$$\hat{\theta} \Big/ \sqrt{\hat{\mathrm{V}}\left(\hat{\theta}\right)}$$

follows an approximately standard normal distribution
- Equivalently, $\hat{\theta}^2 \Big/ \hat{\mathrm{V}}\left(\hat{\theta}\right)$ follows a chi-square distribution with 1 degree of freedom
- Must estimate the variance $\mathrm{V}\left(\hat{\theta}\right)$ appropriately

Example: Survey of Youth in Custody (1)

- Is there an association between:
  - "Was anyone in your family ever incarcerated?"
  - "Have you ever been put on probation or sent to a correctional institution for a violent offense?"
- Sample size: n=2,588 youths
- Table with sum of weights:

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

Example: Survey of Youth in Custody (2)

- Results in the following estimated proportions:

		Ever Violent? No	Ever Violent? Yes	
Family Member Incarcerated?	No	.1928	.2896	.4824
	Yes	.1959	.3217	.5176
		.3887	.6113	1.0000

- Test statistic: $\hat{\theta} = \hat{p}_{11}\hat{p}_{22} - \hat{p}_{12}\hat{p}_{21} = 0.0053$
- How to estimate the variance?

Example: Survey of Youth in Custody (3)

- Use resampling method:
- Thus, the standard error of $\hat{\theta}$ is $0.0158/\sqrt{7} = 0.006$
- So the test statistic is

$$t = \frac{\hat{\theta}}{\sqrt{\hat{\mathrm{V}}\left(\hat{\theta}\right)}} = \frac{0.0053}{0.0060} = 0.89$$

- p-value: $\Pr(|T| > t) = 2 \times \Pr(T_{\nu=6} > 0.89) = 0.41$
- Result: No evidence of association

Random Group	$\hat{\theta}$
1	0.0132
2	0.0147
3	0.0252
4	−0.0224
5	0.0073
6	−0.0057
7	0.0135
Average	0.0065
SD	0.0158

Wald Tests for Larger Tables

- Let $\boldsymbol{\theta} = \left[\theta_{11}, \theta_{12}, \ldots, \theta_{(r-1)(c-1)}\right]^T$
- Hypotheses are

$$H_0 : \boldsymbol{\theta} = \mathbf{0}$$

$$H_a : \boldsymbol{\theta} \neq \mathbf{0} \text{ for one or more cells}$$

- Wald test statistic is $X_W^2 = \hat{\boldsymbol{\theta}}^T \hat{\mathrm{V}}(\hat{\boldsymbol{\theta}})^{-1} \hat{\boldsymbol{\theta}}$ where $\hat{\mathrm{V}}(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix
- Problem is, need a large number of PSUs to estimate covariance matrix
  - E.g., 4x4 table results in 9x9 covariance matrix that requires estimation of 45 variance/covariances

Bonferroni Tests (1)

- Alternative to Wald test
- Idea is to separately (and conservatively) test each $\theta_{ij},\ 1 \le i \le r-1,\ 1 \le j \le c-1$
- Test each of m=(r-1)(c-1) tests separately at α/m significance level
- Reject null that variables are independent if any of the m separate tests reject

Bonferroni Tests (2)

- Specifically, reject $H_0 : \boldsymbol{\theta} = \mathbf{0}$ if

$$\left|\hat{\theta}_{ij}\right| \Big/ \sqrt{\hat{\mathrm{V}}\left(\hat{\theta}_{ij}\right)} > t_{\alpha/2m,\kappa}$$

  for any i and j, where κ is the appropriate degrees of freedom
  - Resampling: #resample groups – 1
  - Another method: #PSUs – #strata
- Lohr says method works well in practice

Example:
Survey of Youth in Custody (1)

- Is there a relationship between age and whether a youth was sent to an institution for a violent offense?

		Age Class			
		≤ 15	16 or 17	≥ 18	
Violent Offense?	No	.1698	.2616	.1275	.5589
	Yes	.1107	.1851	.1453	.4411
		.2805	.4467	.2728	1.0000

Example:
Survey of Youth in Custody (2)

- Hypotheses are $H_0 : \theta_{11} = p_{11} - p_{1+} p_{+1} = 0$

$$\theta_{12} = p_{12} - p_{1+} p_{+2} = 0$$

- What happens if clustering ignored?
  - With n=2,621, we have

$$X^2 = n \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(\hat{p}_{ij} - \hat{p}_{i+} \hat{p}_{+j})^2}{\hat{p}_{i+} \hat{p}_{+j}} = 34$$

which gives an (incorrect) p-value of ~0

- Compare to a Bonferroni test…

Example:
Survey of Youth in Custody (3)

- For these data, $\hat{\theta}_{11} = 0.013$ and $\hat{\theta}_{12} = 0.0119$
- Using resampling, we get the table:

Random Group	$\hat{\theta}_{11}$	$\hat{\theta}_{12}$
1	−0.0195	0.0140
2	0.0266	−0.0002
3	0.0052	0.0159
4	0.0340	0.0096
5	0.0197	0.0212
6	0.0025	0.0298
7	−0.0103	0.0143

- And from this, $\text{s.e.}\left(\hat{\theta}_{11}\right) = 0.0074,$ $\text{s.e.}\left(\hat{\theta}_{12}\right) = 0.0035$
- Thus

$$\frac{\left|\hat{\theta}_{11}\right|}{\text{s.e.}\left(\hat{\theta}_{11}\right)} = 1.8, \quad \frac{\left|\hat{\theta}_{12}\right|}{\text{s.e.}\left(\hat{\theta}_{12}\right)} = 3.4 \text{ and } t_{0.05/2\times 2,\nu=6} = 2.97$$

Reject null (more appropriately)

Using SAS to Conduct the Tests

- SAS v8 has some procedures for complex survey analysis (PROC SURVEYMEANS PROC SURVEYREG), but no PROCs for categorical data analysis
  - PROC FREQ and PROC CATMOD would incorrectly estimate standard errors
- SAS v9.1 has PROC SURVEYFREQ for categorical data analysis
  - Can specify various Wald and Rao-Scott tests
- See http://support.sas.com/onlinedoc/913/docMainpage.jsp

Using Stata to Conduct the Tests

- Stata 9: SVY procedures support both one-way and two-way tables
  - svy:tabulate oneway
  - svy:tabulate twoway
- Need to order manuals to see which methods used
- See www.stata.com/stata9/svy.html for more detail

Other Software…

- …designed for complex survey analysis include SUDAAN and WestVar
  - Don’t know if they can do these calculations or not
  - See their documentation
- JMP: Cannot do appropriate calculations

What We Have Just Learned

- Tests for contingency table data
  - For one variable, goodness-of-fit tests
  - For two variables, tests of independence
- Gained some insight into
  - What to do about categorical data analysis for complex designs
  - How complex designs affect chi-square hypothesis tests
- Learned about some methods to correct for the sampling design in chi-square tests