P8130: Biostatistical Methods I

Lecture 10: Contingency Tables - Tests for Categorical Data

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Lectures 9: Recap

- One and two-sample proportions (Normal Approximation)
  - Point estimates
  - Confidence intervals
  - Hypothesis testing

Lectures 10: Outline

- ‘R x C’ Contingency Tables
  - Chi-Squared Test of Independence/Homogeneity
  - Fisher’s Exact Test
- McNemar’s Test for matched-pair data (Normal Approximation)

R x C Contingency Tables

- Thus far, we compared two independent proportions (using or not the Normal Approximation).
- What if we want to compare 3+ independent proportions?
- Data from observations made on two different categorical variables (with 2 or more levels) can be summarized in a tabular format.
- An ***R x C Contingency Table*** is a table with R rows and C columns:
  - It displays the relationship between two variables, where:
  - The variable depicted in the rows has R categories
  - The variable depicted in the columns has C categories
  - 2 x 2 table is a special case with 2 rows and 2 columns

General *R x C* Contingency Tables

Each cell represents the observed frequency (count) in that row/column combination.

R X C Table	Column Variable					
Row Variable	1	2	3	…	C	Row Total
1	n_{11}	n_{12}	n_{13}	…	n_{1C}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	…	n_{2C}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	…	n_{3C}	$n_{3.}$
…	…	…	…	…	…	…
R	n_{R1}	n_{R2}	n_{R3}	…	n_{RC}	$n_{R.}$
Column Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	…	$n_{.C}$	$n_{..}$

General $R \times C$ Contingency Tables

Each cell represents the observed frequency in that row/column combination.

For each cell, we need to compare what we observed to what would be expected under the null hypothesis.

	Column Variable					
Row Variable	1	2	3	…	C	Row Total
1	n_{11} $E_{11} = \frac{n_{1.}n_{.1}}{n_{..}}$	n_{12} $E_{12} = \frac{n_{1.}n_{.2}}{n_{..}}$	n_{13} $E_{13} = \frac{n_{1.}n_{.3}}{n_{..}}$	…	n_{1C} $E_{1C} = \frac{n_{1.}n_{.C}}{n_{..}}$	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	…	n_{2C}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	…	n_{3C}	$n_{3.}$
…	…	…	…	…	…	…
R	n_{R1}	n_{R2}	n_{R3}	…	n_{RC}	$n_{R.}$
Column Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	…	$n_{.C}$	$n_{..}$

Chi-Squared: Test of Homogeneity

- In the general case where the row totals are fixed, let p_{ij} represent the probability that an individual in the i^{th} row falls in the j^{th} column.

$\boldsymbol{H_0}$: states that the probability of falling in the j^{th} column is the same for all rows

$: p_{1j} = p_{2j} = \ldots = p_{Rj} = p_{.j}, j = 1, 2, \ldots, C.$

$\boldsymbol{H_1}$: states that for at least one column there are two rows i and i' where the probabilities are not the same.

$: p_{ij} \neq p_{i'j}, j = 1, 2, \ldots, C.$

- This assumes that the row totals are fixed before the sample is drawn and that columns are observed.

Chi-Squared: Test of Homogeneity

- Under the null, the Chi-Squared test statistic is given by:

$$\chi^2 = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi^2_{(R-1)\times(C-1)}, \text{where } df = (R-1)\times(C-1)$$

Where:

O_{ij} represents the observed counts in the ij^{th} cell

E_{ij} represents the expected counts in the ij^{th} cell and it can be calculated as:

$$E_{ij} = \frac{n_{i.}n_{.j}}{n_{..}} = \frac{(total\ number\ in\ the\ i^{th}\ row)(total\ number\ in\ the\ j^{th}\ column)}{grand\ total}$$

Chi-Squared: Test of Homogeneity

- Under the null, the test statistic follows a Chi-Squared distribution with $(R-1) \times (C-1)$ degrees of freedom.

- For an α level test:

  Reject H_0: if $\chi^2 > \chi^2_{(R-1)\times(C-1),1-\alpha}$

  Fail to reject H_0: $\chi^2 \leq \chi^2_{(R-1)\times(C-1),1-\alpha}$

- Assumptions:
  - Independent random samples
  - No expected cell counts are 0, and no more than 20% of the cells have an expected count less than 5.

Chi-Squared: Test of Homogeneity

Example: A researcher is studying the extent of marijuana usage among college students.

He selected 150 freshmen, 135 sophomores, 125 juniors and 100 seniors, and asked them to complete a questionnaire to indicate the extent of marijuana use.

Class	**Extent of Drug Usage**			
	Experimental	Casual	Moderate/Heavy	Total
Freshman	57	50	43	150
Sophomore	57	58	20	135
Junior	56	45	24	125
Senior	45	22	33	100
Total	215	175	120	510

Chi-Squared: Test of Homogeneity

Example: Marijuana use.

We assume that the row totals are fixed, i.e., number of students selected by class level.

$\boldsymbol{H_0}$: $p_{11} = p_{21} = p_{31} = p_{41}$, the proportions of 'experimental' users among class levels are equal

AND

$p_{12} = p_{22} = p_{32} = p_{42}$, the proportions of 'casual' users among class levels are equal

AND

$p_{13} = p_{23} = p_{33} = p_{43}$, the proportions of 'moderate/heavy' users among class levels are equal

$\boldsymbol{H_1}$: not all proportions are equal.

Chi-Squared: Test of Homogeneity

Example: Marijuana use: observed and the expected cell counts

Class	Extent of Drug Usage			
	Experimental	Casual	Moderate/Heavy	Total
Freshman	57 $E_{11} = \frac{215 \cdot 150}{510} = 63.24$	50 $E_{12} = \frac{175 \cdot 150}{510} = 51.47$	43 $E_{13} = \frac{120 \cdot 150}{510} = 35.29$	150
Sophomore	57 $E_{21} = 56.91$	58 $E_{22} = 46.32$	20 $E_{23} = 31.76$	135
Junior	56 $E_{31} = 52.70$	45 $E_{32} = 42.89$	24 $E_{33} = 29.41$	125
Senior	45 $E_{41} = 42.16$	22 $E_{42} = 34.31$	33 $E_{43} = 23.53$	100
Total	215	175	120	510

Chi-Squared: Test of Homogeneity

Example: Marijuana use. Calculate the test statistic.

$$\chi^2 = \sum_{i=1}^{4}\sum_{j=1}^{3}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(57-63.24)^2}{63.24} + \frac{(50-51.47)^2}{51.47} + \cdots + \frac{(33-23.53)^2}{23.53} = 19.4$$

Under the null hypothesis, this test statistic $\chi^2 \sim \chi^2_{(4-1)\times(3-1)=6\ df}$

Because $\chi^2 > \chi^2_{6,0.95}$ = 12.94, we reject the null hypothesis at 0.05 significance level, and conclude that the extent of marijuana use is significantly different by class.

Chi-Squared: Test of Independence

- The same Chi-squared test, with the same steps, can be used for testing the independence/association of the row and column variables.
- Example: participants in a study might be categorized with respect to both disease status: (Diseased/Not Diseased) and exposure status (Exposed/Not Exposed).
- The question of interest is whether knowledge of one variable's value provides any information about the value of the other variable, i.e. are the two variables independent?
- Setting: 2 x 2 table:

$\boldsymbol{H_0}$: disease and exposure are independent ($p_1 = p_2$)

$\boldsymbol{H_1}$: disease and exposure are associated/dependent

Case-Control:

$$p_1 = P(E|D) = P(E)$$

$$p_2 = P\left(E\middle|\overline{D}\right) = P(E)$$

Cohort Study:

$$p_1 = P(D|E) = P(D)$$

$$p_2 = P\left(D\middle|\overline{E}\right) = P(D)$$

Chi-Squared: Test of Independence

Example: A lab is testing the potential risk factors for ectopic pregnancy. Out of 279 women who had experienced ectopic pregnancy, 28 had suffered from pelvic inflammatory disease (PID). Of the women who did not experience ectopic pregnancy, 6 had suffered from PID.

Is there an association between pelvic inflammatory disease and ectopic pregnancy?

	Ectopic Pregnancy	**No Ectopic Pregnancy**	**Total**
PID	28 Expected value: (17)	6 (17)	34
No PID	251 (262)	273 (262)	524
Total	279	279	558

Chi-Squared: Test of Independence

H_0: PID and ectopic pregnancy are independent

H_1: PID and ectopic pregnancy are dependent/associated

Compute the test statistic with Yates' Continuity Correction (only for 2 x 2 table):

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(|O_{ij}-E_{ij}|-0.5)^2}{E_{ij}} = \frac{(|28-17|-0.5)^2}{63.24} + \cdots + \frac{(|273-262|-0.5)^2}{262} = 13.8$$

Under the null: $\chi^2 \sim \chi_1^2$.

At 0.05 significance level, $\chi^2 > \chi^2_{1,0.95} = 3.84$, we reject the null and conclude that there is sufficient evidence that PID and ectopic pregnancy are associated.

Assumptions? All satisfied: independent samples, all expected values ≥ 5.

Fisher’s Exact Test for 2 x 2 Table

- The methods for comparing two proportions using either normal approximation or contingency tables are equivalent.
- But … both of them are normal-theory based.
- What if this normal approximation is not valid, usually the case when the cell counts are small (< 5)?
- Use an ‘exact method’ called Fisher’s Exact test.

Fisher's Exact Test for 2 x 2 Table

- Fisher's Exact test statistic and p-value can be calculated by following:

1. Enumerate all possible tables with the same margins as the observed table
   - Start with the table with 0 in cell (1,1); the other cells will be determined from the row and column margins.
2. Compute the 'exact' probability of each table from step 1:

   $P(a,b,c,d) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{n!a!b!c!d!}$; a, b, c, d, denote the cell counts of the 2 x 2 table
3. Calculate the 'exact' p-value based on cumulative probabilities of all tables as extreme or more extreme that what was observed.

- Exact tests are easily computed in SAS/R, but please feel free to challenge yourselves with some 'hand' calculations.

Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test)

Data type: paired data with binary outcome.

<u>Example:</u> A group of 75 patients are tested with two different diagnostic procedures: A and B, for determining the presence/absence of a disease. We want to test the hypothesis that the proportions of positives for the two procedures are equal.

	Procedure B	
Procedure A	Positive	Negative
Positive	41	8
Negative	14	12

Concordant pairs: pairs in which the outcome is the same for each member of the pair (53 concordant pairs).

Discordant pairs: pairs in which the outcome is different for each member of the pair (22 discordant pairs)

Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test)

- Discordant pairs: pairs in which the outcome is different for each member of the pair.

- Focus on discordant pairs as the concordant ones give no information about the differences.

- A <u>type A discordant pair</u>: is a discordant pair in which the treatment A member of the pair has the event and the treatment B member does not.

- A <u>type B discordant pair</u>: is a discordant pair in which the treatment B member of the pair has the event and the treatment A member does not.

Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test)

- Assume the following definitions:
- Let $\boldsymbol{p}$ be the probability that a discordant pair is of type A
- Let $\boldsymbol{n_D}$ the total number of discordant pairs
- Let $\boldsymbol{n_A}$ the total number of discordant pairs of type A
- If the probability of an event (positive test) is the same for both row and column variables, then there should be an equal number of type A and type B discordant pairs.

Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test)

- If the probability of an event (positive test) is the same for both row and column variables, then there should be an equal number of type A and type B discordant pairs, i.e., p=1/2.
- If an event is more likely for A than for B, then we would expect p >1/2.
- If an event is less likely for A than for B, then we would expect p <1/2.
- The hypothesis test focuses on the proportion of type A discordant pairs:

$$H_0: p = \frac{1}{2} \quad vs \quad H_1: p \neq \frac{1}{2}$$

- It follows that under the null, $n_A \sim Bin(n_D, 0.5)$

McNemar's Test: Normal Approximation

- Rule of thumb for normal approximation: $n_D \geq 20\ or\ \frac{n_D}{4} \geq 5.$
  - If normal approximation does not hold, use Exact McNemar's Test (based on binomial probabilities)
- Testing the hypotheses:

$$H_0: p = \frac{1}{2} \quad vs \quad H_1: p \neq \frac{1}{2}$$

- Calculate the test statistic (with continuity correction):

$$\chi^2 = \frac{\left(\left|n_A - \frac{n_D}{2}\right| - \frac{1}{2}\right)^2}{\frac{n_D}{4}} = \frac{(|n_A - n_B| - 1)^2}{n_A + n_B}$$

Where $\boldsymbol{n_D}$ represents the total number of discordant pairs
$\boldsymbol{n_A}$ represents the total number of discordant pairs of type A
$\boldsymbol{n_B}$ represents the total number of discordant pairs of type B

McNemar's Test: Normal Approximation

- Given the test statistic:

$$\chi^2 = \frac{\left(\left|n_A - \frac{n_D}{2}\right| - \frac{1}{2}\right)^2}{\frac{n_D}{4}} = \frac{(|n_A - n_B| - 1)^2}{n_A + n_B}$$

- For an α level test:

Reject H_0: if $\chi^2 > \chi^2_{1,1-\alpha}$

Fail to reject H_0: $\chi^2 \leq \chi^2_{1,1-\alpha}$

Two-Sample Test for Binomial Proportions for Matched-Pair Data (McNemar's Test)

Back to our example:

	Procedure B	
Procedure A	Positive	Negative
Positive	41	8
Negative	14	12

Test statistic: $\chi^2 = \frac{(|n_A - n_B| - 1)^2}{n_A + n_B} = \frac{(|8-14|-1)^2}{8+14} = 1.14$

At 0.05 significance level $\chi^2 < \chi^2_{1,0.95}$=3.84, fail to reject the null and conclude that the probability of a positive does not differ by procedure type.

Readings

Rosner, *Fundamentals of Biostatistics 8th Edition*

- Chapter 10, Sections 10.1-10.7
- Independent reading: Chapter 10, Section 10.8 (Kappa Statistic)