

Faculty Development Program

Clinical Epidemiology and Clinical Research

TOPIC: Biostatistics 4: Categorical Data Analysis

DATE: May 2 (1:30 PM – 5:00 PM)

LEADERS: Art Evans

OBJECTIVES:

1. Calculate and interpret a chi-square test;
2. Tell their friends that they have actually worked a Fisher's exact test by hand;
3. Describe when it is appropriate to use Fisher's exact test and a McNemar's test;
4. Describe the utility of a Mantel Haenszel test;
5. Interpret a printout of a Mantel Haenszel test.

REQUIRED READINGS:

Read enough to accomplish the objectives above. The topics are covered in the two books below. Pick the one that best fits your needs.

Norman and Streiner. PDQ Statistics. 2nd Ed. Pages 81-95; 107-110.

Norman and Streiner. Biostatistics: The Bare Essentials. Chapters 1-4. Pages: 150-157.

QUESTIONS:

1. Consider the study by Seckler and colleagues (Ann Intern Med 1991;115:92), which assessed the agreement between outpatients and their primary care physicians on whether the patient desired CPR in the event of sudden cardiac arrest.

		Patient		
		CPR	DNR	Total
Physician	CPR	47	5	52
	DNR	14	3	17
	Total	61	8	69

Is there a statistically significant association between the decisions of patients and their physicians? Justify your approach to answering this question.

2. Use the Fisher's exact test to calculate a P value for the following data:

	Live	Die	Total
Drug	4	1	5
Placebo	1	4	5
Total	5	5	10

Fisher's Exact Test:

- Only for 2x2 tables.
- Use if any *expected* count is < 5.
- P-value: Assuming independence, the probability that the observed frequencies, or frequencies more extreme, could occur by chance.
- P-value = Probability that observed distribution is due to chance *plus* probability that other distributions—as extreme or more extreme—are due to chance.
- Probability that a particular distribution is due to chance =

$$\frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! N!}$$

Categorical Data Analysis

Chi-square, Fisher's Exact, McNemar, and Stratified Analysis

A Evans

- **Chi-square** = sum of: $(\text{observed} - \text{expected})^2 / \text{expected}$
Numerator: Observed - expected = "signal"
Denominator: Expected (equivalent to variance) = "noise"
- **Chi-square formula (easier if doing by hand):**
 $N(ad - bc)^2 / \text{product of all marginals}$
- **Degrees of freedom** = (number of rows - 1) x (number of columns - 1)
- **Yates continuity correction:** sum of: $[(\text{absolute value of } O - E) - 1/2]^2 / E$
Yates continuity correction adjusts for the uncertainty associated with small numbers in some cells; statisticians disagree on whether it's helpful or harmful; if there is an important difference between the plain chi-square and the Yates corrected chi-square, then consider abandoning both and using an "exact" method (eg, Fisher's exact; but there are others).
- **Small numbers:**
 1. For 2x2 tables, use an "exact" method, eg, Fisher's exact test, if:
 - ▶ the total sample is less than 20, or
 - ▶ any cell has an *expected* number (not an *observed* number) less than 5.
 2. For tables larger than 2x2, the chi-square test should *not be trusted* if:
 - ▶ more than 20% of cells have expected values less than 5, or
 - ▶ the minimum expected value is less than 1.
- **Fisher's exact test:** assumes that the marginals are fixed. For example, in a randomized trial, you "fix" the proportion of all subjects who are in each group, usually 50% experimental and 50% control. The proportions in the two outcome categories, however, are **not** "fixed" (proportion alive and proportion dead for the entire study population). This is a conservative assumption and gives a P value that is greater than it should be if only one set of marginals is fixed.
- **Rule of "3" for confidence intervals:** in estimating the 95% CI for a percentage of 0% (0/n), the upper limit of the 95%CI can be estimated by adding 3 to the numerator.
- **Paired, matched, or pre/post data:** should **not** be analyzed with a Pearson's chi-squared test, but instead with McNemar's: $(|b - c| - 1)^2 / (b + c)$. Also, the OR for matched data is not ad/bc , but b/c .
- **Estimating P values:** if the chi-square test (or McNemar) has a value greater than the number of cells, then the P value will be < 0.05 . (Note: a 2x2 table has 4 cells.)

Stratified Analysis

- Stratified analysis is used to assess interaction and control for confounding.
- For categorical variables (X, Y, and Z):
 - Is the association between X and Y the same, regardless of the value of a third variable Z? (Interaction or Effect Modification)
 - Is the crude (unadjusted) association between X and Y false because of *confounding* by Z?

Stratified Analysis in SPSS

1. SPSS: Analyze: Descriptive: Crosstabs:
 - insert Row (X) and Column (Y) variables;
 - insert layer variable (Z).
2. Association between X and Y will be measured for all levels (strata) of Z.
3. Compare measures of association (OR or RR) for the different strata. (You'll have to calculate ARR by hand.)
4. Make a judgment about possible interaction and confounding.
5. Test if the association between X and Y is the same across strata (interaction) by using the Mantel-Haenszel chi-square test (available only in SPSS version 10 or higher).
6. Recall: this only tests for interaction on a relative odds scale. Interaction on an absolute risk scale might be very different.
7. Calculate an *adjusted* OR—the association between X and Y after adjusting for Z—using another Mantel-Haenszel Chi-square test.

What's the association between X and Y after controlling for Z?

Example: What's the strength of the association between coffee drinking (yes/no) and Gastric CA after controlling for smoking (current/former/never)?

Current Smoker	Coffee	Gastric CA		
		+	-	
	+	a ₁	b ₁	n ₁
	-	c ₁	d ₁	
Former Smoker	+	a ₂	b ₂	n ₂
	-	c ₂	d ₂	
Never Smoked	+	a ₃	b ₃	n ₃
	-	c ₃	d ₃	

- Mantel-Haenszel OR: weighted average of the OR's across strata (assuming it's appropriate to combine them, ie, no important differences/interactions).

$$\text{Summary OR}_{MH} = \frac{\text{Sum: } (a_i)(d_i)/n_i}{\text{Sum: } (b_i)(c_i)/n_i}$$

- M-H Chi-squared test: Is the summary OR_{MH} different from 1?

$$\frac{\text{Sum for all "a" cells: } (O_{ai} - E_{ai})^2}{\text{Sum over all strata: } \frac{\text{product of all marginals}}{n^2(n-1)}}$$

degrees of freedom = $k-1$
(k is number of strata)

SPSS: Categorical Data Analysis

- **Examine a single proportion.**
Is it different from some known or hypothesized proportion?
SPSS: Analyze: Nonparametric: Binomial: select the variable and the expected percentage; set the cutpoint (cutpoint value goes with the first group). (Unfortunately, it doesn't give confidence intervals, only P values.)
- **Examine a nominal variable with more than two categories.**
Are the proportions among the different categories different from a set of hypothesized proportions?
SPSS: Analyze: Nonparametric: Chi-square: select the variable; under Expected Values, indicate the percentages from the 1st group to the last, eg, enter 25 (Add), 50 (Add), 25 (Add), if you expect the distribution to be 25%, 50%, and 25% in categories 1, 2, and 3, respectively.
- **Examine the association between two nominal variables.**
Is there an association?
SPSS: Analyze: Descriptive: Crosstabs: select your independent variable for Row, and select dependent (outcome) variable for Column; click on Statistics button: select Chi-square and select Risk; click on Cells button: select Observed and select Row percentages. (Unfortunately, it gives you confidence intervals for the OR and the RR, but not for the risk difference.)
- **Examine a dichotomous variable at two different times (before/after) or for two different observers (two physicians examining the same set of patients) or using two different methods.**
Are the paired proportions the same (Time 1 vs. 2, or Observer 1 vs. 2, or Method 1 vs. 2)?
SPSS: Analyze: Descriptive: Crosstabs: select one variable for Row, and select second variable for Column; click on Statistics button: select Chi-square and select McNemar; click on Cells button: select Observed.
Alternative: **SPSS: Analyze: Nonparametric: 2 Related Samples:** check McNemar: select the pair of variables (must be dichotomous and must have identical response options).
- **Examine the relationship between two categorical variables (paired or unpaired) among different subgroups of subjects.**
Are the relationships the same across strata (interaction present)?
Is the average relationship between X and Y across the strata different from the relationship of X and Y in the crude/pooled/unstratified data (confounding present)?
SPSS: Analyze: Summarize: Crosstabs: select one variable for *Row*, and select second variable for *Column*; select a third variable that defines strata for *Layer* (subgroup variable); click on Statistics button: select Chi-square (and Risk) or select McNemar, depending on whether the first two variables are unpaired

(independent) or paired (dependent); select Cochran's and Mantel-Haenszel statistics (with common OR=1); click on Cells button: select Observed and Row percentages.

- **Examine the effects of dichotomizing an ordinal or continuous variable.**

Does dichotomizing an ordinal or continuous variable always "waste information" and therefore always give a conservative statistical result?

Examine the database "Heart Rate." Fifty subjects were randomized into two groups of 25 subjects; Group 1 got one treatment and Group 2 another. The outcome was *change in heart rate* (pre minus post), with positive numbers indicating a favorable drop in heart rate. The hypothesis was that Group 1 would have more of a drop than Group 2 (somewhere between 1 and 10 beats per minute).

1. Using your best clinical judgment, define a clinically meaningful drop in heart rate. Using that cutpoint, dichotomize the data, produce a 2x2 table, and decide whether the proportion with an important drop in HR is the same for the two groups. Hopefully, several of you will select different cutpoints and we can compare results.
2. Justify your use of the chi-square test, Fisher's exact test, or McNemar's test.