

BIOS 600: Principles of Statistical Inference

Contingency Tables: Part II

Fall 2012

Reading

- ▶ Pagano and Gauvreau, Chapter 16

Multiple Tables

Previously we talked about how to analyze data from $r \times c$ tables to quantify a potential association between two factors. We will continue with this concept but now concentrate on the relationship between two factors in the presence of a third factor.

When might this be of interest?

- ▶ We have a 2×2 table at each site in a multi-site study
- ▶ We wish to combine results of several published studies in a meta-analysis
- ▶ The value of a third factor affects our estimates of the association between two factors

Combining Multiple Tables

Sometimes, the presence of a third factor can affect the relationship between the two factors of interest. We will discuss how to proceed in the presence of a third factor, and in particular will examine the consequences of aggregating (or not aggregating) data over levels of a third factor. (Note: whether or not to aggregate often depends on your subject matter knowledge as much as on statistical considerations)

To start, we will consider cases of Simpson's paradox, in which aggregating or not aggregating data can lead to drastically different conclusions.

Simpson's Paradox

Simpson's paradox occurs when the direction of an association between two variables is reversed after stratification upon a third variable.

Kidney Stone Treatment

In 1986 *British Medical Journal* reported the results of a study comparing open surgical treatment to percutaneous nephrolithotomy (PN) for removal of kidney stones. They first examined success rates of the procedure stratified by the size of the kidney stone.

Kidney Stone Treatment: Small Stones

	Surgery Successful?	
	Yes	No
Open	81	6
PN	234	36

The estimated probability of success for small stones using open procedures was $\frac{81}{87} = 0.93$, and the corresponding probability using PN was $\frac{234}{234+36} = 0.87$. Open procedures were better: the corresponding OR and 95% CI were 2.08 (0.82, 6.24).

Kidney Stone Treatment: Large Stones

	Surgery Successful?	
	Yes	No
Open	192	71
PN	55	25

The estimated probability of success for large stones using open procedures was $\frac{192}{192+71} = 0.73$, and the corresponding probability using PN was $\frac{55}{55+25} = 0.69$. Open procedures were better: the corresponding OR and 95% CI were 1.23 (0.68, 2.18).

Kidney Stone Treatment: All Stones Combined

	Surgery Successful?	
	Yes	No
Open	273	77
PN	289	61

The estimated probability of success for all stones using open procedures was $\frac{273}{273+77} = 0.78$, and the corresponding probability using PN was $\frac{289}{289+61} = 0.83$. The corresponding OR and 95% CI were 0.75 (0.50, 1.11).

WHAT HAPPENED? PN is better now?

Kidney Stone Treatment: What Happened?

- ▶ Group sizes are very different
- ▶ Doctors tended to give the harder-to-cure cases (large stones) the better treatment, and small stones the inferior treatment.
- ▶ The success rate is much more strongly influenced by the size of the stone than by the treatment type.

Birth to 10 Study

See Morrell, *Journal of Statistics Education* 7(3) (1999) for additional details.

The Birth to Ten study (BTT) commenced in the Johannesburg/Soweto metropolitan area of South Africa during 1990. A birth cohort was formed from all singleton births during a seven-week period, with 4029 enrolled. The BTT study collected prenatal, birth, and early development information on these children. The aim of the study was to identify factors related to the emergence of cardiovascular disease risk factors in children living in an urban environment in South Africa. In 1995, when the children were five years old, the children and caregivers were invited to attend a screening exam that included interviews to gather information on exposure to tobacco smoke and additional health-related issues. The five-year sample consisted of 964 children.

Birth to 10 Study

If the five-year sample is to be used to draw conclusions about the entire birth cohort, the five-year group should have characteristics similar to those who were not traced from the initial group. Thus, the five-year group was compared to those who did not participate in the five-year interview on a number of factors. One of the factors was whether the mother had medical aid (similar to health insurance) at the time of the birth of the child.

Birth to 10 Study

	Child Included at Age 5?	
	No	Yes
Medical Aid	195	46
No Medical Aid	979	370
Total	1174	416

Separate problem: severe missing data on medical aid status at baseline (started with 4029)

```
tabi 195 46 \\ 979 370, exact
```

Conducting Fisher's exact test, we find that there is a significant association between medical aid at baseline and inclusion at age 5 ($p=0.007$). Specifically, 19% of children with medical aid were included at age 5, while 27% of children without medical aid were included.

Birth to 10 Study

One of the researchers noted an additional problem with the follow-up study, which was that 28% of black children in the birth cohort were included at age 5, while only 9% of white children were included. Because of this imbalance, the researchers looked at the association between medical aid and inclusion stratified by race.

	Child Included at Age 5?			
	White		Black	
	No	Yes	No	Yes
Medical Aid	104	10	91	36
No Medical Aid	22	2	957	368
Total	126	12	1048	404

After stratifying on race, there is no longer an association between inclusion based on medical aid status. Among whites, 9% of those with medical aid and 8% of those without were included. Among blacks, 28% in each group were included.

Birth to 10 Study

How did this happen?

- ▶ The sizes of the groups, which are combined when we ignore the lurking variable, are very different. Many more black than white babies were enrolled in the study from the beginning.
- ▶ The lurking variable has a very strong effect on the outcome probabilities. The probability of re-enrollment is more strongly influenced by race (28% vs 9%) than by insurance status (27% vs 19%).

What do we conclude? The investigators concluded that within race, medical aid status had no effect on the decision to participate in the 5 year old visit. However, black children were much more likely to participate than white children. There is still a problem with generalizability of the results of the 5 year old study, but it is due not to aid status but to race.

Simpson's Paradox in Graduate School Admissions

A math and English double major (like me!), you are looking at graduate programs and discover during your interview that your dream university last year admitted 30.0% of men but only 21.3% of women. You are shocked by the gender bias and decide to apply elsewhere, but during your meetings with the directors of graduate studies of the two departments (math and English are the only departments in this dream university), you see some startling figures.

English Department

The English department's admissions statistics are below.

	Admitted?	
	No	Yes
Female	29	21
Male	60	40
Total	89	61

So English admitted $\frac{21}{50} = 42\%$ of women, and $\frac{40}{60} = 40\%$ of men. Maybe it's just the sexist math professors who are driving the abysmal admissions statistics.

Math Department

Because math is such an amazing major, you decide to meet with them despite your reservations. Their admissions director provides you with the following admissions statistics.

	Admitted?	
	No	Yes
Female	89	11
Male	45	5
Total	89	61

Math admits $\frac{11}{100} = 11\%$ of women and $\frac{5}{50} = 10\%$ of men!

Overall, your dream university admitted $\frac{21+11}{50+100} = 21.3\%$ of women, and $\frac{40+5}{100+50} = 30\%$ of men. However, $\frac{2}{3}$ of the women applied to the department that was harder to get into, and only $\frac{1}{3}$ of the men applied to the more competitive department.

The same thing happened at UC-Berkeley in Fall 1973, and the university was sued for gender discrimination based on the university-wide aggregate data. Essentially, Simpson's paradox was the culprit – on the whole, women had applied to more competitive departments. (In this case, they were the less well-funded departments that had fewer graduate fellowships to offer.)

Recommendations

Often it is appropriate to stratify and look at unit-specific analysis, but not always. Take EPID 600 or 710 for more guidance! There is also a nice discussion at [the Wikipedia site on Simpson's paradox](#). If you are an epidemiological methods guru, check out the Judea Pearl chapter on the subject.

Combining Multiple 2×2 Tables

- ▶ Should we combine tables?
- ▶ If so, how do we combine them?
- ▶ Once combined, how do we make inferences?
- ▶ The Mantel-Haenszel strategy potentially removes the confounding influence of explanatory variables that comprise the stratification and can provide increased power for detecting association.

Should We Combine Tables?

This is actually a pretty hard question to answer! You will learn more about how to approach this question in epidemiology when you discuss confounders, effect modifiers, and other important inferential issues.

Suppose we have two 2×2 tables.

- ▶ Calculate estimates separately for each table. If the association is the same in the two tables, it is usually fine to combine the two tables.
- ▶ If the association is different, then it is often (though not always) not a good idea to combine tables.
 - ▶ Want to combine if difference due to chance occurrence
 - ▶ Don't want to combine if there really is a different association (e.g., only individuals with a certain genotype are susceptible to exposure)

Mantel-Haenszel Strategy

- ▶ Determine whether strength of association is uniform across tables
 - ▶ If not, stop and report separate odds ratios
- ▶ If strength of association is similar across tables,
 - ▶ Calculate a combined OR
 - ▶ Test whether the overall association is significant

Should We Combine Tables?

If both tables have an $OR=1.2$, then our decision is easy!
However, what if in one table we have $OR=0.9$ and in the other have $OR=1.1$? How do we decide whether OR 's are different?

We can use a *test of homogeneity* of odds ratios, and several versions have been developed. They test $H_0 : OR_1 = OR_2 = \dots = OR_g$ for g 2×2 contingency tables (other methods are available for $r \times c$ tables).

This is in essence a test of interaction (e.g., effect modification) with the third factor.

Test of Homogeneity

The null hypothesis for this test is that the OR's are the same. If we see a small p-value, we reject the null hypothesis and conclude the OR's are different. If we fail to reject the null hypothesis, we often will calculate a combined odds ratio, collapsing over the levels of the third factor to obtain one larger contingency table.

Exposure	Stratum i	
	Disease	
	+	-
+	a_i	b_i
-	c_i	d_i

Handy formula:

$$y_i = \ln(\widehat{OR}_i) = \ln\left(\frac{a_i d_i}{b_i c_i}\right) \text{ with estimated variance of } y_i \text{ given by}$$
$$s_i^2 = \frac{1}{a_i} + \frac{1}{b_i} + \frac{1}{c_i} + \frac{1}{d_i}$$

Test of Homogeneity

Exposure	Stratum i	
	Disease	
	+	-
+	a_i	b_i
-	c_i	d_i

One estimate of an average $\ln(OR)$ is a simple weighted average, which gives more weight to better (low variance) estimates. It is given by

$$\bar{y} = \frac{\sum_{i=1}^g \frac{y_i}{s_i^2}}{\sum_{i=1}^g \frac{1}{s_i^2}}. \text{ (Note that Mantel}$$

& Haenszel proposed a slightly different estimate.)

One common test statistic for homogeneity is given by $X^2 = \sum_{i=1}^g \frac{1}{s_i^2} (y_i - \bar{y})^2$, which under H_0 is approximately χ^2 with $g - 1$ df.

Should We Combine Tables?

Consider a study of coffee drinkers and myocardial infarction (MI). Work in JAMA and other journals has linked high levels of coffee consumption with increased risk of MI. A third variable, smoking status, may be affecting our OR estimates.

MI	Smokers		Nonsmokers	
	Coffee	No Coffee	Coffee	No Coffee
Yes	1011	81	383	66
No	390	77	365	123
Total	1401	158	748	189

The OR for smokers is $\frac{1011 \times 77}{81 \times 390} = 2.46$ with 95% CI (1.74, 3.48).

The OR for nonsmokers is $\frac{383 \times 123}{66 \times 365} = 1.96$ with 95% CI (1.39, 2.77). Are these OR's significantly different?

Breslow-Day Test of Homogeneity

We use the Breslow-Day Test of Homogeneity for

$$H_0 : OR_{smoker} = OR_{nonsmoker}.$$

```
. cc MI coffee [freq=count] , by(smoke)
```

smoke	OR	[95% Conf. Interval]		M-H Weight	
0	1.955542	1.387234	2.769425	25.70971	(exact)
1	2.464292	1.739714	3.485125	20.26299	(exact)
Crude	2.512051	1.981955	3.186788		(exact)
M-H combined	2.179779	1.721225	2.760499		

Test of homogeneity (M-H) chi2(1) = **0.93** Pr>chi2 = **0.3342**

Test that combined OR = 1:

Mantel-Haenszel chi2(1) = **43.58**
Pr>chi2 = **0.0000**

Test Results for MI Data

Because our p-value for the test of homogeneity is large, we fail to reject H_0 . On the output you see the combined odds ratio estimate and 95% CI: 2.2 (1.7, 2.8). We conclude that heavy coffee drinkers have roughly twice the odds of MI as nondrinkers.

Mantel-Haenszel Test

How was the combined odds ratio obtained? We need to be careful when combining odds ratios given Simpson's paradox. The Mantel-Haenszel method provides a safe method for combining odds ratios across contingency tables and for testing whether the Mantel-Haenszel combined odds ratio is equal to 1 or not.

The summary odds ratio estimate for multiple 2×2 tables is given by

$$\widehat{OR} = \frac{\sum_{i=1}^g \frac{1}{T_i} (a_i d_i)}{\sum_{i=1}^g \frac{1}{T_i} (b_i c_i)},$$

where g is the number of 2×2 tables (e.g., 2 in the MI study) and T_i is the total number of subjects in table i .

Mantel-Haenszel Test

We can get a 95% confidence interval for this combined OR estimate as well as a chi-squared test of $H_0 : OR_{COMBINED} = 1$.

Because the distribution of the combined OR is heavily skewed, we usually calculate a 95% CI on the natural logarithm scale and transform it back to the OR scale. This is how most software packages calculate the CI.