

Lab 11

Contingency Tables

- What is a contingency table? It shows the relationship between two categorical variables:

	Outcome		Total
	Yes	No	
Group 1	a	b	a+b
Group 2	c	d	c+d
Total	a+c	b+d	a+b+c+d

Note: Sometimes in contingency tables the rows will be labeled “Disease” and “No Disease” and the columns will be labeled “Exposed” and “Not Exposed”. Be aware that the table presented above labels the *columns* with the outcome, with the outcome being “Disease”/“No Disease”. For odds ratios, \hat{p}_1 is $P(\text{Disease}|\text{Exposed})$ and \hat{p}_2 is $P(\text{Disease}|\text{Not Exposed})$, which is what is represented below. [To clarify, using the labels of the above table our probabilities would be written as $P(\text{Yes}|\text{Group 1})$ and $P(\text{Yes}|\text{Group 2})$].

- Estimation: $\hat{p}_1 = \frac{a}{a+b}$ and $\hat{p}_2 = \frac{c}{c+d}$

Odds Ratios

- If p is the probability of outcome, $\frac{p}{1-p}$ are the odds of outcome.
- The odds ratio compares the odds of a given outcome in two groups. It is a measure of the *strength of association* between the outcome and group membership.
- Estimated odds ratio: $\hat{OR} = \frac{\hat{p}_1/(1-\hat{p}_1)}{\hat{p}_2/(1-\hat{p}_2)} = \frac{ad}{bc}$
- When there is no association $OR=1$.
- If $OR=4$ the interpretation would be that the odds of the disease (the outcome being “Yes”) in the exposed (Group 1) is 4 times that in the unexposed (Group 2).
- Note that $e^{\ln(X)} = X$ thus $e^{\ln(OR)} = OR$
- The expected cell counts should be ≥ 5 for normal approximation to be reasonable.

Hypotheses	$H_0 : OR = 1$ $H_A : OR \neq 1$
Estimated OR	$\hat{OR} = \frac{\hat{p}_1 / (1 - \hat{p}_1)}{\hat{p}_2 / (1 - \hat{p}_2)} = \frac{ad}{bc}$
$se \left[\ln \left(\hat{OR} \right) \right]$	$\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
CI for $\ln \left(OR \right)$	$\ln \left(\hat{OR} \right) \pm z_{\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
CI for OR	$e^{\left(\ln \left(\hat{OR} \right) \pm z_{\alpha/2} \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}} \right)}$

Multiple 2x2 Tables

- Multiple 2x2 tables can arise either from multiple studies or from one study stratified by a third variable.
- If the association is the same in each of the 2x2 tables, we would like to report a single odds ratio (rather than separate odds ratios).
- To do this, we need to answer three questions:
 1. Should we combine the information in the multiple tables? (i.e., Is the association (OR) the same in each study or across strata?)
 - Test of homogeneity (Stata calls this the “test of heterogeneity”)
 - $H_0: OR_1 = OR_2 = \dots = OR_g$ (where g is the number of contingency tables)
 - H_A : at least one OR is different
 - If we reject H_0 then we *cannot* report a combined OR. We need to report study/stratum specific results.
 2. If the test of homogeneity is NOT rejected, how should we combine the information in the multiple tables to yield a single summary measure of the strength of the association (OR)?
 - What can happen if we simply collapse the tables into one, ignoring the third factor? (Simpson’s Paradox)

- Calculate the Mantel-Haenszel common odds ratio, OR_{MH}
- If all expected cell counts summed over strata/studies ≥ 5 , then it is okay to use the normal approximation for calculating the confidence interval for OR_{MH} .

3. Is there an association between the two variables?

- Test whether the common odds ratio (estimated by OR_{MH}) = 1 using the confidence interval.
- $H_0: OR_{MH} = 1$ vs. $H_A: OR_{MH} \neq 1$

Problems

Open and name a log file and name it if you'd like to save your work.

1. In a 1985 study of the relationship between contraceptive use and infertility, 89 out of 283 infertile women, compared with 640 out of 3833 control women, had used an IUD at some time in their lives.

a.) Create a contingency table of this data. (In this case, what is our “disease” and what is the “exposure”?)

Disease	Exposure		Total
	Yes	No	
Yes			
No			
Total			

- b.) What are the estimated odds of infertility among the women who used an IUD?
- c.) What are the estimated odds of infertility among the women who did not use an IUD?
- d.) What is the estimated odds ratio for infertility comparing women who used in IUD versus those who didn't?
- e.) Provide a 95% confidence interval for the true odds ratio. Interpret the confidence interval. You can use the command: `cci 89 194 640 3193`
- f.) At the 0.05 level of significance, test the null hypothesis that there is no association between IUD use and infertility. State your null and alternative hypotheses. Report the p-value and make a conclusion.

2. Suppose we are presented with a pair of 2×2 contingency tables, both of which provide information about the same dichotomous random variables representing exposure and disease, but that originate from two distinct studies. For example, the following data come from two studies, both conducted in San Francisco, which investigated risk factors for epithelial ovarian cancer.

Study 1			
Disease Status	Term Pregnancies		Total
	None	One or more	
Cancer	31	80	111
No Cancer	93	379	472
Total	124	459	583

Study 2			
Disease Status	Term Pregnancies		Total
	None	One or more	
Cancer	39	149	188
No Cancer	74	465	539
Total	113	614	727

- The data are contained in the data set “ovarian.dta”. Open this in Stata.
- Quick note: if you wanted to check the numbers in the contingency tables with the Stata data set (your numbers should be the same), try the commands:

by study: tabulate cancer

by study: tabulate term

a.) What are the relative odds of developing ovarian cancer for women who have **never had a term pregnancy** versus women **who have had one or more** for each of the two studies?

b.) Interpret each odds ratio. Are the two estimates similar?

c.) For each study, test the null hypothesis that there is no association between term pregnancy and ovarian cancer. Within each study, what do you conclude?

d.) It is possible that each study is estimating the same population value. Can we combine the evidence collected in the two different studies to make a single overall statement about the association between the number of term pregnancies a woman has had and the occurrence of epithelial ovarian cancer?

We need to test whether the ORs are the same.

What are the null and alternative hypotheses for this test? Conduct the test.

What are the steps to follow if you reject?

We need to consider the ORs for each study separately.

What are the steps to follow if you do not reject?

We can combine the ORs using the Mantel-Haenszel method.

e.) If you do not reject the null in part d, calculate the combined OR.

f.) Test that the combined $OR=1$. Give the p-value and CI.

3. Circle the designs that use matched pairs:

- a.) We want to compare 2 lotions for the treatment of poison ivy. People with poison ivy on both arms are recruited. One arm is randomly assigned to lotion 1, the other to lotion 2.
 - b.) We want to study the association between perinatal mortality and maternal smoking. We randomly select 200 mothers to participate, and it turns out 100 of the mothers smoked during pregnancy and 100 did not.
 - c.) We want to examine the association between sleeping habits and passing a pop quiz. We randomly select 1,000 students to give the pop quiz to.
 - d.) We want to study the association between retirement and heart disease. We recruit 250 people who have heart disease and match them on age, gender, and health status with 250 healthy control subjects.
 - e.) We want to study participants' plaque index (each person is categorized as having a plaque index of high or low) over time. 70 participants are recruited and we measure their plaque index at baseline and then 4 weeks later.
4. We recruit 218 people who have poison ivy on both arms. We want to test the null that there is no association between the type of lotion and relief. (i.e. We want to test the null that the probability that the arm with lotion 1 feels relief is the same as the probability the arm with lotion 2 feels relief.)

Arm with lotion 1			
Arm with lotion 2	Relief	No Relief	
	Relief	28	55
	No Relief	41	94
		69	149
			218

(extra problem)

5. In the 2x2 contingency tables below, the data from a German study investigating the relationship between smoking status and invasive cervical cancer have been stratified by the number of sexual partners that a woman has had.

≤ 1 Partner				≥ 2 Partners			
Cancer				Cancer			
Smoke	Yes	No	Total	Smoke	Yes	No	Total
Yes	7	12	19	Yes	96	142	238
No	18	112	130	No	92	150	242
Total	25	124	149	Total	188	292	480

- a.) How could the number of sexual partners potentially affect the measurement of the relationship between smoking and cervical cancer?
- b.) For each group, test the null hypothesis that there is no association between cervical cancer and smoking. Within each study, what do you conclude?
- c.) Estimate the odds of cervical cancer for smoker relative to nonsmokers for women who have had at most one sexual partner.

d.) Estimate the odds ratio for women who have had two or more sexual partners.

f.) What test would we use to test the null that the OR within each group is the same?

g.) Say we conduct the homogeneity test and get a p-value of .041. Then what can we do?