Unit 10 & 11 HW

Rashmi Patel

7/10/2021

## Proportion Conceptual questions

1. State under what circumstance that a difference in proportion confidence interval or test should not be used in favor of an odds ratio metric.

\* If the data is sampled retrospectively, then only the odds ratio can be estimated.

1. Under what sampling schemes can a hypothesis test be generally worded as a “test for association”, rather than a test for difference in proportion, odds, or relative risk?

\* In a double-blind study, such as the vitamin C example from the book.

1. It might be tempting in a predictive model setting (using a test set) to obtain a count of correct vs incorrect prediction for two or more modeling approaches. To be clear here, imagine that the classification accuracy on a test set of 200 observations for model 1 is .82 while for model 2 it was .80. To determine that one model is actually predicting “better”" than the other in terms of overall accuracy, you produce a confidence interval for the difference in proportion of correctly predicted observations and the interval does not contain 0. Why is this type of analysis using a two by two count table inappropriate?

\* This is an apples to oranges comparison, different modeling approaches are fundamentally different and as an example one modeling approach may be better suited to one thing than the other, or one might be more sensitive to different data types.

1. Give an example of study that could not be conducted in a prospective manner due to either logistical or ethical considerations.

\* If for example we wanted to study if eating batteries causes cancer, so we assign one group of subjects to eat batteries and the other as a control.

1. What metric allows for the following interpretation: “The chances of a person getting the flu who has not taken the flu vaccine is thirty percent (1.3 times) higher than someone who has taken the flu vaccine”?

\* This is a retrospective study and the metric is an odds ratio.

## Exercise #1 Basic Tools

The statistical sleuth only considers statistical inference and estimation for two by two count tables in which a comparison is needed. However, it is often useful to perfom inference on a single population proportion (think 1 sample t-test for a binary outcome). The technical details and formulas can be easily obtained on line or in a text book, the following example illustrates how to produce a z-test and confidence interval for a single proportion. The same function can be used for the difference in proportions.

Consider the vitamin C study on flu occurences in which study members either took a placebo or vitamin C tablets prior to a flu season and they continued to take the supplement throughout the study. Does Vitamin C have any effect on the occurence of flu? That is the main question here. Suppose for the moment a study was conducted just on a placebo group to understand what the general rate of flu outbreak is for an “average” person. The prop.test allows us to estimate the proportion of this population and generate a confidence interval. Additional z-testing can be done by specifying a null hypothesized value but we will not worry about that here. Since 335 out of 411 people obtained the flu, the following code generates inference for a single population proportion:

prop.test(335,411,p=.7,correct=TRUE) #correct is the continuity correction option

##   
## 1-sample proportions test with continuity correction  
##   
## data: 335 out of 411, null probability 0.7  
## X-squared = 25.376, df = 1, p-value = 4.716e-07  
## alternative hypothesis: true p is not equal to 0.7  
## 95 percent confidence interval:  
## 0.7733959 0.8507513  
## sample estimates:  
## p   
## 0.8150852

Similar to SAS proc freq, categorical analysis in R essentially requires information on the counts of key variables of interest (response and explanatory). Usually, this requires an additional step when working from a data file that has a traditional format (each row is an observation, columns are variables). To illustrate how this would work lets create a traditional data set out of the vitamin c study. First we will create a data set, and use table to aggregate the data into a 2x2 count matrix. I will then give an example of how to directly create a count table manually.

vit.c<-data.frame(Supp=rep(c("Placebo","Vitamin C"),times=c(411,407)),  
 Cold=rep(c("Yes","No","Yes","No"),times=c(335,76,302,105)))  
head(vit.c)

## Supp Cold  
## 1 Placebo Yes  
## 2 Placebo Yes  
## 3 Placebo Yes  
## 4 Placebo Yes  
## 5 Placebo Yes  
## 6 Placebo Yes

mymat<-table(vit.c) # or table(vit.c$Supp,vit.c$Cold)  
mymat

## Cold  
## Supp No Yes  
## Placebo 76 335  
## Vitamin C 105 302

Note here we could have created the table ourselves by just creating a matrix.

mymat<-matrix(c(76,335,105,302),2,2,byrow=T,dimnames=list(c("Placebo","Vitamin C"),c("No","Yes")))

As discussed in the analysis plan of live session, a common approach to streamline the analysis of 2x2 contingency tables is to run either a fishers exact or chisquare test to answer the key hypothesis question at hand. The following is an example of how to do both. At this stage the format of my mat doesn’t matter, the p-value will be the same regardles of how the rows and columns are situated.

#Conducting a fishers exact test  
fisher.test(mymat)

##   
## Fisher's Exact Test for Count Data  
##   
## data: mymat  
## p-value = 0.01444  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.4603013 0.9229415  
## sample estimates:  
## odds ratio   
## 0.6528601

#Conducting a chi square test  
chisq.test(mymat,correct=TRUE)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: mymat  
## X-squared = 5.9196, df = 1, p-value = 0.01497

Note here that the sample sizes are quite big for the two groups, the p-value for both tests correspond quite nicely here because of this.

Another helpful tool is the ability to look at the proportions of the table rather than the raw counts. This is where it is helpful to have a good understanding of how your matrix is formatted. Since my rows are the explanatory variable I want the proportions to be obtained via each row (margin=1 for rows, margin=2 for columns).

prop.table(mymat,margin=1)

## No Yes  
## Placebo 0.1849148 0.8150852  
## Vitamin C 0.2579853 0.7420147

Make note the proportion of No’s for each group. This will come in handy later for interpretation of Confidence Intervals.

Turning our attention now to confidence intervals, these can be obtained from various packages. Its just a matter of preference. Differences in proportions are easily obtained using the previous prop.test function. For the odds ratio and relative risk metrics, we will obtain these from the epitools package.

#CI's for proportions  
prop.test(mymat,correct=TRUE)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: mymat  
## X-squared = 5.9196, df = 1, p-value = 0.01497  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.13222111 -0.01391972  
## sample estimates:  
## prop 1 prop 2   
## 0.1849148 0.2579853

Here is where we have to be careful with output. Note the proportion table we made previously. The confidence intervals and estimates are for the proportions of people who did not get the cold rather than the proportion of people who did get the cold. To swap it around we can do the following.

prop.test(mymat[,c(2,1)],correct=TRUE)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: mymat[, c(2, 1)]  
## X-squared = 5.9196, df = 1, p-value = 0.01497  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## 0.01391972 0.13222111  
## sample estimates:  
## prop 1 prop 2   
## 0.8150852 0.7420147

#this is the proportion of people who got the cold

Further, if we wanted to compare the second group to the first rather than the first to the second, we could simply swap the rows. This is made easier for the other metrics using the epitools package.

prop.test(mymat[c(2,1),c(2,1)],correct=TRUE)

##   
## 2-sample test for equality of proportions with continuity correction  
##   
## data: mymat[c(2, 1), c(2, 1)]  
## X-squared = 5.9196, df = 1, p-value = 0.01497  
## alternative hypothesis: two.sided  
## 95 percent confidence interval:  
## -0.13222111 -0.01391972  
## sample estimates:  
## prop 1 prop 2   
## 0.7420147 0.8150852

As you dive through the output from these calls, you should be coming to a realization that it doesn’t really matter if you define proportion of yes or no’s, group a versus group b or vice-versa, the results all tell you the same thing. (See Sleuth Ch 18 Exercise 5 for a detailed look)

Confidence intervals for odds ratios and relative risk can be obtained using the epitools package. The choice of which metric to use is up to you.

library(epitools)  
#Another way to format a count matrix  
mymat<-matrix(c(76,335,105,302),2,2,byrow=T)  
dimnames(mymat)<-list("Treatment"=c("Plac","Vit C"),"Response"=c("No","Yes"))  
mymat

## Response  
## Treatment No Yes  
## Plac 76 335  
## Vit C 105 302

#Odds Ratio Intervals  
oddsratio.wald(mymat,rev = "columns")

## $data  
## Response  
## Treatment Yes No Total  
## Plac 335 76 411  
## Vit C 302 105 407  
## Total 637 181 818  
##   
## $measure  
## odds ratio with 95% C.I.  
## Treatment estimate lower upper  
## Plac 1.000000 NA NA  
## Vit C 1.532546 1.097777 2.139504  
##   
## $p.value  
## two-sided  
## Treatment midp.exact fisher.exact chi.square  
## Plac NA NA NA  
## Vit C 0.01201242 0.01444212 0.01182729  
##   
## $correction  
## [1] FALSE  
##   
## attr(,"method")  
## [1] "Unconditional MLE & normal approximation (Wald) CI"

#Relative Risk Intervals  
riskratio.wald(mymat, rev="columns")

## $data  
## Response  
## Treatment Yes No Total  
## Plac 335 76 411  
## Vit C 302 105 407  
## Total 637 181 818  
##   
## $measure  
## risk ratio with 95% C.I.  
## Treatment estimate lower upper  
## Plac 1.000000 NA NA  
## Vit C 1.395157 1.074201 1.812011  
##   
## $p.value  
## two-sided  
## Treatment midp.exact fisher.exact chi.square  
## Plac NA NA NA  
## Vit C 0.01201242 0.01444212 0.01182729  
##   
## $correction  
## [1] FALSE  
##   
## attr(,"method")  
## [1] "Unconditional MLE & normal approximation (Wald) CI"

**HW Questions** 6. Consider the last bit of output from line 91, where the estimate and confidence interval for the odds ratio is reported. What is the ratio currently for? Odds (Y:N) or (N:Y) and is placebo on top or is placebo on bottom?

* Based on the above output, I believe the ratio is for odds (N:Y) and that the placebo is on the top.

1. Rerun the oddsratio.wald function but include the option (rev=“both”) or maybe try rev=“rows” or rev=“columns”. Determine which version(s) produces the odds ratio when we want the odds (Y:N) for the placebo group over the odds (Y:N) for the Vitamin C group.

* The version we are after is (mymat,rev=“columns”)

1. Rinse and repeat #1 and #2 for the relative risk calculation. What proportions (Yes or No) are being used to calculate the relative risk and which group is on top of the ratio. Use the rev option to swap things around so that the relative risk metric is (proportion of Yes placebo group over the proportion of yes for the vitamin c group).

* Based on the above output, I believe the ratio is for odds (N:Y) and that the placebo is on the top, so we are looking at the “No” proportion calcukating relative risk. And in response to the second step, The version we are after is (mymat,rev=“columns”).

Note: The whole point of these questions are to get you experienced with the idea of swapping estimates around to get at the most comfortable interpretation of the metric you would like to report. It’ll force you to really appreciate that you can’t just trust the output directly and don’t forget you can always do a hand calculation to make sure what is being reported makes sense.

1. Conduct an analysis of the study reported in the Sleuth book Ch 18 Exercice #7. Make sure to comment on the study design, which metrics are available to use, conduct your analysis and write a conclusion to answer the researchers question “Is alcoholic beverage consumption associated with breast cancer?”.
2. Which factor is the response?

* In a retrospective study subjects are selected from groups with specified response levels. Their explanatory factors are then determined.
* In this study the presence of breast cancer is the response.

1. Is this a retrospective or prospective study?

* This is a retrospective study as we are studying the habits of women and drinking. We will consider this information about their past.

1. Is this a randomized experiment or observational study?

* This is an observational study since we have selected two different groups of women to study, and have NOT placed them into treatment groups.

mymat<-matrix(c(330,658,204,386),2,2,byrow=T,dimnames=list(c("Fewer than four drinks per week ","4 or more drinks per week"),c("Cases","Controls")))  
mymat

## Cases Controls  
## Fewer than four drinks per week 330 658  
## 4 or more drinks per week 204 386

#Conducting a fishers exact test  
fisher.test(mymat)

##   
## Fisher's Exact Test for Count Data  
##   
## data: mymat  
## p-value = 0.6601  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.7611771 1.1842188  
## sample estimates:  
## odds ratio   
## 0.9489964

#Conducting a chi square test  
chisq.test(mymat,correct=TRUE)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: mymat  
## X-squared = 0.1785, df = 1, p-value = 0.6727

prop.table(mymat,margin=1)

## Cases Controls  
## Fewer than four drinks per week 0.3340081 0.6659919  
## 4 or more drinks per week 0.3457627 0.6542373

Step 1: H0: p1=p2 Ha: p1!=p2

Step 2:

Step 3: zstat

Step 4: p-value: 0.6601

Step 5: Fail To Reject

Step 6: Conclusion: We will fail to reject the null hypothesis and conclude that there isn’t a signigicant difference between the two groups. Step 7: Since this is an observational experiment, we cannot generalize this to the greater population. We also failed to reject the null hypothesis, and concluded that there isn’t a difference in breast cancer rates among women who drink and those who do not drink.

## Exercise #2 Mantel-Haenszel Tests- Don’t do this!

Like the table function previously discussed, the most realistic way to generate a tables for Mantel Haenszel Tests is from aggregating a data set from a more traditional format. Sticking with the theme of the Vitamin C study, Ch 19 exercise 16 provides a hypothetical extension to the study that also kept track of whether each patient knew they were in the placebo or treatment group ahead of time. Sometimes people can just figure it out. The question now becomes, is the conclusion of the single 2x2 analysis still remain consistent after adjusting for this additional factor? (Note: This is just a fabricated data set. Use the text to take a look at the tables of numbers)

This data set requires the knowledge of two 2x2 count tables (one for each knowledge group).

vit.c<-data.frame(Supp=rep(c("Placebo","Vitamin C","Placebo","Vitamin C"),times=c(312,168,88,232)),  
 Cold=rep(c("Yes","No","Yes","No","Yes","No","Yes","No"),times=c(266,46,139,29,62,26,157,75)),  
 Known=rep(c("Yes","No"),times=c(480,320)))  
head(vit.c)

## Supp Cold Known  
## 1 Placebo Yes Yes  
## 2 Placebo Yes Yes  
## 3 Placebo Yes Yes  
## 4 Placebo Yes Yes  
## 5 Placebo Yes Yes  
## 6 Placebo Yes Yes

#Going from a standard file to generate tables for each group  
mytable<-table(vit.c) #note here the order of the columns of vit.c matters. The last column should be the stratification variable.  
mytable

## , , Known = No  
##   
## Cold  
## Supp No Yes  
## Placebo 26 62  
## Vitamin C 75 157  
##   
## , , Known = Yes  
##   
## Cold  
## Supp No Yes  
## Placebo 46 266  
## Vitamin C 29 139

At this point, we may want to edit the tables here to get them formatted so the odds are produced in a format we would like to read the odds ratio in. Reordering the tables is easy, we are just using arrays instead of matrices. The 3rd dimension is the “Known” status which we do not care of the order that it is in.

#Rearranging the order of tables so that Yes's are first. Placebo is already on top.   
library(samplesizeCMH)

## Warning: package 'samplesizeCMH' was built under R version 4.0.5

mytable<-mytable[,c(2,1),] #Odds (Y:N) Placebo over   
mytable

## , , Known = No  
##   
## Cold  
## Supp Yes No  
## Placebo 62 26  
## Vitamin C 157 75  
##   
## , , Known = Yes  
##   
## Cold  
## Supp Yes No  
## Placebo 266 46  
## Vitamin C 139 29

#Comparing the odds for each known categoy (Yes they knew their group, no they didn't)  
apply(mytable,3,odds.ratio)

## No Yes   
## 1.139147 1.206444

Please note that the Yes/No table above is not the Yes/No variable for whether they have the flu or not. The apply function takes mytable and calculates the odds ratio for each of the 2 tables (the 3 index says use the 3rd column of table array which is what stratifies the tables to begin with.) Here we can see that the odds ratios are pretty consistent. A breslow-day test can confirm this. This test is available in various packages of which I have had a hard time installing. DescTools looks to be the most promising but I had troubles.

Assuming that a common odds ratio exists across the groups of patients who knew their treatment status, the MH test can be applied.

mantelhaen.test(mytable)

##   
## Mantel-Haenszel chi-squared test with continuity correction  
##   
## data: mytable  
## Mantel-Haenszel X-squared = 0.57345, df = 1, p-value = 0.4489  
## alternative hypothesis: true common odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.811711 1.696612  
## sample estimates:  
## common odds ratio   
## 1.173524

**HW Problems** 10. It is always interesting to go back and compare what you would have concluded if you had ignored the “known”" status variable that you adjusted the analysis with. Ignore the known status variable and conduct a standard 2x2 analysis from a collapsed table. Provide fishers exact test and a confidence interval for the ODDS ratio and compare it to the previous MH test results. The result of this comparison really shows how patients not being blinded to the treatment group can really impact the overall estimates of parameters like odds ratios as well as the hypothesis test conclusions.

* N/A Professor Turner explained not to do the MH problems.

1. Consider Chapter 19 Exercise 17 out of Statistical Sleuth. It is an extension to the alcohol/breast cancer study you previously did. Perform the MH test procedure after taking into account the individual womens body mass as this could be a confounding factor. (Use ?mantelhaen.test for ane example of writing the data set into an array table if you don’t want to make a big data set up front like I did in the example code.)

* N/A Professor Turner explained not to do the MH problems. {“mode”:“full”,“isActive”:false}