Chi-Sq Tests vs Proportion Tests vs Logistic Regression: Binary Responses and Binary/Categorical Predictors

Solutions

PART I. PREDICTORS WITH 2 CATEGORIES (BINARY PREDICTORS)

Students at a small liberal arts college took a placement exam prior to entry in order to provide them guidance when selecting their first math course. The dataset **MathPlacement** (in Stat2Data) contains the placement scores for 2696 students, along with whether they took the recommended course, and several admissions variables (GPA, SAT score, etc.). We want to use this data to decide how well the math placement process is working. If they take the recommended course, do they succeed (where "success" is defined as a grade of "B" or above)?

We already investigated this question using Logistic regression in the "Multiple Logistic Regression" activity. But **there are other statistical methods/tests** that we could use to answer this question.

Some data cleaning: Before we go further, I'm going to remove the students who are missing values for "CourseSuccess".

```
MathPlacement <- filter(MathPlacement, !is.na(CourseSuccess))
```

A. Logistic Regression

Summarize your results from part D of "Multiple Logistic Regression". If students take the recommended course, are they signficiantly more likely to succeed than those who don't take the recommendation? *How much* more likely to succeed?

According to "Multiple Logistic Regression", we can say the following:

- Of those that did take the recommendation and we know their grade, 1045/(441+1045) = 70.3% were successful; of those that didn't take the recommendation and we know their grade, 396/(247+396) = 61.6% were successful. So the difference in probabilities is about 9%.
- The odds of success for someone who did take the recommended course is 148% the odds of success for someone who didn't listen to the recommendation.

B. 2-Sample Proportion Test

We have binary response variable (CourseSuccess) and a binary explanatory variable (RecTaken). Thus, we could investigate the relationship between them using a 2-sample proportion test.

1. Hypotheses. If $\pi_{RecTaken}$ is the probability of success for those who took the recommended course and $\pi_{NotRecTaken}$ is the probability of success for those who didn't take the recommended course, the hypotheses are (write in words and in symbols):

$H_0 : \pi_{RecTaken} = \pi_{NotRecTaken}$ versus $H_a : \pi_{RecTaken} \neq \pi_{NotRecTaken}$

2. Assumptions/Conditions. A 2-sample proportion test has the same conditions as the one-sample proportion test, applied to both samples.

- number of successes for both groups is at least 10
- number of failures for both groups is at least 10
- samples are independent of each other

Are the conditions met in this case? Make sure you discuss all 3 conditions. (In this case, our two groups are: 1) those who took the recommended course; and 2) those who didn't take the recommended course.)

```
tally(CourseSuccess ~ RecTaken, data=MathPlacement)
```

```
##              RecTaken
## CourseSuccess    0    1
##             0  247  441
##             1  396 1045
```

We see from the table above that there are (many) more than 10 successes and failures for both groups. It is reasonable to believe that these individuals (those who did take the rec and those who didn't) are independent.

3. Using R for Inference. Just as in the one-sample case, `prop.test()` will find the p-value and confidence interval.

Syntax:

`prop.test(response var ~ explanatory var)`

Use prop.test to test the the hypotheses in #1.

```
prop.test(CourseSuccess~RecTaken, data=MathPlacement)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(CourseSuccess ~ RecTaken)
## X-squared = 15.265, df = 1, p-value = 9.342e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.04206161 0.13267240
## sample estimates:
##    prop 1    prop 2
## 0.3841369 0.2967699
```

Notice that the "sample estimates" at the bottom of the output are the estimates of the probability of course *failure* as opposed to course *success*. This is because R is choosing "CourseSuccess=0" to be "success". Of course, we want "CourseSuccess=1" to be the success, so we can change the code like this:

```
prop.test(CourseSuccess~RecTaken, data=MathPlacement, success=1) #the "success" argument tells R what t
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(CourseSuccess ~ RecTaken)
## X-squared = 15.265, df = 1, p-value = 9.342e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.13267240 -0.04206161
## sample estimates:
##    prop 1    prop 2
## 0.6158631 0.7032301
```

If you have a **summary 2-way table** (as opposed to untabulated data), you use this alternative syntax

```
prop.test(x=c(successes_in_group1,successes_in_group2), n=c(total_#_in_group1,total_#_in_group1))
```

```
prop.test(x=c(396,1045), n=c((396+247),(1045+441))) #396 successes among those (396+247) who didn't tak
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(396, 1045) out of c((396 + 247), (1045 + 441))
## X-squared = 15.265, df = 1, p-value = 9.342e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.13267240 -0.04206161
## sample estimates:
##    prop 1    prop 2
## 0.6158631 0.7032301
```

4. Conclusion Make a conclusion about H_0 in the context of the problem.

This very small p-value indicates that we have strong evidence that the success rates of those who do take the recommendation are significantly different from those who don't.

5. Confidence intervals. `prop.test()` will give you the CI for the difference between the two proportions, as long as the "alternative" is two-sided (which is the default). Report and interpret the 95% confidence interval for the difference in success rates between those who did and didn't take the recommended course.

We are 95% confident that those who do take the rec have a success rate between 4% and 13% higher than those who don't.

C. Chi-Squared Test for Association

Another way of testing whether the two groups have significantly different probabilities of success is to conduct a Chi-square Test for a 2x2 table. You may have seen this in your Intro Stats class, where it was called the "Test for Independence", the "Test for Association", or the "Test for Homogeneity".

1. Hypotheses. 2 ways to state the hypotheses…

H_0: the two variables (CourseSuccess and RecTaken) are independent vs. H_a: not independent

OR

H_0: the probability of success is the same for both populations (those who took the recommendation and those who didn't) vs. H_a: the proportions are not the same

2. Assumptions/Conditions. The only conditions for the Chi-square Test are that:

- all expected cell counts are at least 5
- the sample (or samples) are randomly selected, or representative of the population of interest

You've already thought about the second condition; we'll discuss whether the first condition is met later.

3. Test Statistic & Distribution. Consider the table below.

```
tally(CourseSuccess ~ RecTaken, data=MathPlacement)
```

```
##              RecTaken
## CourseSuccess    0    1
##             0  247  441
##             1  396 1045
```

Under the null hypothesis of the Chi-square Test, we'd expect our two groups (those who took the recommendation and those who didn't) to success at about the same rate. Thus, the frequencies in each cell should correspond to the overall proportions of successes and failures for the entire sample. The Chi-square test statistic, χ^2, compares the observed cell counts to the expected cell counts under this assumption.

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

where the sum is taken over all the cells in the table (in this case, four).

Under the null hypothesis, we'd expect χ^2 to be quite small, since Observed and Expected should be nearly equal. How far apart these numbers are tells us how unusual our data is; thus, big values of χ^2 are an indication that H_0 is false. Under H_0, χ^2 follows a chi-squared distribution with 1 degree of freedom.

4. Using R to find the test stat & p-value.

Syntax: First, you must create a 2-way table of successes and failures. Then you run `chisq.test` on that table.

`chisq.test(table)`

Create the 2-way table below (using tally, xtabs, or table), then run `chisq.test` on the table, and save the test as `rec.chisq`.

```
table2 <- tally(CourseSuccess ~ RecTaken, data=MathPlacement) #make your 2-way table here
rec.chisq <- chisq.test(table2); rec.chisq
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table2
## X-squared = 15.265, df = 1, p-value = 9.342e-05
```

Report the test statistic and p-value here:

Test stat = 15.265; p-value = 0.0000934

5. Back to the assumptions. The `chisq.test()` object `rec.chisq` contains the expected cell counts in `rec.chisq$expected`: you can check that they are all at least 5. As a nice bonus, however, `chisq.test()` will actually warn you during the test if one or more of the expected values is less than 5.

Is this condition met in this case?

```
rec.chisq$expected
```

```
##              RecTaken
## CourseSuccess        0         1
##             0 207.7896  480.2104
##             1 435.2104 1005.7896
```

Yes.

6. Conclusion. Make a conclusion about H_0 in the context of the problem.

We have strong evidence that Course success is not independent of taking the course recommended.

D. Synthesis

We've seen three methods of looking at the relationship between a binary response (CourseSuccess) and a binary predictor (RecTaken).

1. What are the similarities/difference in the conclusions between these three methods?

The conclusions are all the same: those who take the rec are more likely to succeed, and the difference is signficiant. The p-values for the proportion test and the chi-sq test are identical, but *slightly* different than the p-value from the logistic regression (which was 0.0000791).

2. Discuss the pros and cons of each method. Think about what information the different methods provide, what conclusions we can make with each, and the differences between assumptions/conditions. Are there situations where you feel one or the other method would be best?

PART II. CATEGORICAL PREDICTORS (with > 2 categories)

We saw that those who didn't take the recommended course are *significantly* less likely to succeed than those who do take the recommended course. But of course, these students are being placed into different courses depending on their placement tests. Is the actual course they end up taking a significant factor in whether they succeed? Let's use the chi-sq test for association and logistic regression to investigate this.

A. Data Collection

1. Make a two-way table of CourseSuccess by Course.

```
tally(CourseSuccess~Course, data=MathPlacement)
```

```
##              Course
## CourseSuccess 109 114 117 120 122 126 128 210 220 398
##             0  25   3  63 440  57  19  69   9   3   0
##             1  24   4 159 583 315  22 305  11  17   1
```

2. The success rate in Math109 is: 49%

The success rate in Math120 is: 57%

The success rate in Math220 is: 85%

```
prop.table(tally(CourseSuccess~Course, data=MathPlacement), margin=2)
```

```
##              Course
## CourseSuccess       109       114       117       120       122       126
##             0 0.5102041 0.4285714 0.2837838 0.4301075 0.1532258 0.4634146
##             1 0.4897959 0.5714286 0.7162162 0.5698925 0.8467742 0.5365854
##              Course
## CourseSuccess       128       210       220       398
##             0 0.1844920 0.4500000 0.1500000 0.0000000
##             1 0.8155080 0.5500000 0.8500000 1.0000000
```

3. Based only on the proportions above, does it seem that there is a relationship between success and course?

Yes!

4. You see that very few students took Math114 or Math398. I'm going to filter those students out for the rest of the analysis…

```
MathPlacement <- filter(MathPlacement, Course %in% c(109,117,120,122,126,128,210,220))
```

B. ~~2-Sample Proportion Test~~

We can't do a 2-sample proportion test here because we have more than 2 groups!

C. Chi-Squared Test: Test for Association

1. Hypotheses

H_0: the probability of success is the same for all courses vs. H_a: the proportions are not the same

2. Conduct the Chi-sq test and save it as `course.chisq`. Report the Test Stat and p-value:

```
course.chisq <- chisq.test(tally(CourseSuccess~Course, data=MathPlacement)); course.chisq
```

```
## 
##  Pearson's Chi-squared test
## 
## data:  tally(CourseSuccess ~ Course, data = MathPlacement)
## X-squared = 152.85, df = 7, p-value < 2.2e-16
```

3. Assumptions of the Chi-sq test Is this condition that all expected cell counts are at least 5 met in this case?

```
course.chisq$expected
```

```
##              Course
## CourseSuccess      109       117     120      122     126      128       210
##             0 15.82508  71.69731 330.389 120.1414 13.2414 120.7874  6.459217
##             1 33.17492 150.30269 692.611 251.8586 27.7586 253.2126 13.540783
##              Course
## CourseSuccess       220
##             0  6.459217
##             1 13.540783
```

Yes, since all expected counts are > 5.

4. Conclusion. Make a conclusion about H_0 in the context of the problem.

Since the p-value is so small, the probability of success is NOT the same for all courses.

5. . . . Can we say more?

Since we have rejected the null hypothesis, it would nice to pinpoint which cells are significantly different than expected. In this way, we can say something more than just "there is an association".

We can do this by finding the "contribution to the chi-sq" for each cell. That is, we calculate the amount that each cell "contributed" to the chi-sq test statistic you found above. The larger the contribution, the farther that cell's observed value was from what was expected (the closer to 0, the closer to what was expected).

The contributions are equivalent to the "residuals" from each cell, which you can find in `course.chisq$residuals`. Cells with larger absolute residuals have larger differences between what's observed and what's expected, and thus have "contributed" most to the rejection of the null.

a. In this case, what three courses have the largest contribution?

```
course.chisq$residuals
```

```
##              Course
## CourseSuccess        109        117        120        122        126        128
##             0  2.3063710 -1.0271495  6.0303359 -5.7606048  1.5825242 -4.7120844
##             1 -1.5929320  0.7094172 -4.1649478  3.9786537 -1.0929956  3.2544764
##               Course
## CourseSuccess        210        220
##             0  0.9997181 -1.3610933
##             1 -0.6904713  0.9400609
```

Math 120, 122, aand 128

b. A negative "residual" means that cell had fewer successes than what was expected under the null; a positive "residual" means that cell had more successes than expected. Using this information, what can you say about the success/failure rate in those three courses with the largest contribution, and how it compares to what was expected under the null?

From the negative/postive residuals in the "CourseSuccess=1" row, we see that students in 120 succeeded *less* often than expected, while students in 122 and 128 succeeded *more* often than expected.

D. Logistic Regression

You can absolutely use a logistic regression model to come to a similar conclusion as we did in Part C above. However, the output looks a little different than it did in the case of a binary predictor variable (or a numerical predictor)…

First, note that the values of `Course` are numerical. So `glm()` will think that `Course` is a numerical variable unless you force it to be a factor (categorical variable). We've done this before: just use `as.factor(Course)` in the glm call.

```
course.log <- glm(CourseSuccess~as.factor(Course), family=binomial, data=MathPlacement); summary(course
```

```
##
## Call:
## glm(formula = CourseSuccess ~ as.factor(Course), family = binomial,
##     data = MathPlacement)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9479  -1.2990   0.6387   1.0605   1.1948
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.04082    0.28577  -0.143   0.8864
## as.factor(Course)117  0.96659    0.32223   3.000   0.0027 **
## as.factor(Course)120  0.32223    0.29267   1.101   0.2709
## as.factor(Course)122  1.75034    0.31998   5.470 4.49e-08 ***
## as.factor(Course)126  0.18743    0.42397   0.442   0.6584
## as.factor(Course)128  1.52703    0.31534   4.843 1.28e-06 ***
## as.factor(Course)210  0.24149    0.53262   0.453   0.6503
## as.factor(Course)220  1.77542    0.68835   2.579   0.0099 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2668.6  on 2120  degrees of freedom
## Residual deviance: 2508.2  on 2113  degrees of freedom
## AIC: 2524.2
##
## Number of Fisher Scoring iterations: 4
```

1. The Fitted Logistic Regression Model. Write down the fitted model here:

log(odds of success) = -0.04082 + 0.96659(Math117) + 0.32223(Math120) + 1.75034(Math122) + 0.18743(Math126) + 1.52703(Math 128) + 0.24149(Math210) + 1.77542(Math220)

2. Interpretation of Intercept. What does the intercept of -0.0408 mean? Interpret this in context.

This is the log(odds of success) for the reference group, which is Math109. So we can say that odds of succeeding in the course for those who take Math109 is e^(-0.04082)=0.96. Thus, for every 96 students who succeed in that class, we expect about 100 to fail.

3. All the course have positive coefficient estimates. What does this mean? How do we interpret a positive coefficient?

This means that all classes have a higher odds of success than the reference class of Math109.

4. Interpretation of slope/odds ratio.

The coefficient of Course=120 is **0.322**, which means the odds ratio for that group is **1.38**. Interpret this odds ratio: Those in Math120 have 138% the odds of success of those in Math109. Or, those in Math120 have 38% higher odds of success than those in Math109.

The coefficient of Course=220 is **1.775**, which means the odds ratio for that group is **5.903**. Interpret this odds ratio: Those in Math220 have 590% the odds of success of those in Math109.

5. What does it mean that only some courses are labelled as statistically significant? Whom are these groups significantly different from? (That is, who are they being compared to?)

All courses are compared to the reference course, Math109. So those labelled as significant are statistically different from Math109.

6. Confidence interval for the slope/odds ratio.

Use confint() to calculate a 90% confidence interval for the slope of Course=220:

```r
confint(course.log, level=0.9)
```

```
## Waiting for profiling to be done...

##                              5 %      95 %
## (Intercept)          -0.5138118 0.4306547
## as.factor(Course)117  0.4365217 1.4996742
## as.factor(Course)120 -0.1604574 0.8063933
## as.factor(Course)122  1.2243216 2.2801602
## as.factor(Course)126 -0.5096544 0.8887477
## as.factor(Course)128  1.0082009 2.0488786
## as.factor(Course)210 -0.6332658 1.1299389
## as.factor(Course)220  0.7252345 3.0382612
```

```r
exp(confint(course.log, level=0.9))
```

```
## Waiting for profiling to be done...

##                             5 %      95 %
## (Intercept)          0.5982110  1.538264
## as.factor(Course)117 1.5473158  4.480229
## as.factor(Course)120 0.8517541  2.239815
## as.factor(Course)122 3.4018575  9.778246
## as.factor(Course)126 0.6007031  2.432082
## as.factor(Course)128 2.7406658  7.759195
## as.factor(Course)210 0.5308553  3.095467
## as.factor(Course)220 2.0652154 20.868925
```

Interpret the CI in terms of odds ratios:

We are 95% confident that the odds of success in Math220 are between 2 times and 21 times the odds of success in Math109.

7. Hypotheses: drop-in-deviance test for model utility

Ho: Course is NOT a useful variable in predicting CourseSuccess vs. Ha: Course is useful in predicting CourseSuccess

8. Test statistic & p-value. The test statistic is as usual, the difference between the null deviance and the residual deviance, which follows a chi-sq distribution.

```
anova(course.log, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: CourseSuccess
##
## Terms added sequentially (first to last)
##
##
##                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                2120     2668.6
## as.factor(Course)  7   160.36      2113     2508.2 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Report the test stat and the p-value:

Test stat = 160.36; p-value is approx. 0

9. Conclusion.

We have very strong evidence that course is a useful variable in predicting course success.

E. Synthesis

We've seen two methods of looking at the relationship between a binary response (CourseSuccess) and a categorical predictor (Course).

1. How does your p-value and conclusion in part D (logistic regression) compare to your p-value and conclusion in part C (chi-sq test)?

The p-values are both approximately 0.

2. Discuss the pros and cons of the chi-sq test for association vs. the logistic model when using a categorical predictor. Think about what information the different methods provide, what conclusions we can make with each, and the differences between assumptions/conditions. Are there situations where you feel one or the other method would be best?