Binary Responses and Binary/Categorical Predictors

Tam Nguyen

The data in the file **ICU** (from Stat2) show information for a sample of 200 patients who were parts of larger study conducted in a hospital’s Intensive Care Unit (ICU). Since an ICU often deals with serious, life-threatening cases, a key variable to study is patient survival, which is the (binary) response variable we will be investigating. Survival is coded in the *Survive* variable as 1 if the patient lived to be discharged and 0 if the patient died. Among the possible explanatory variables are

Age = age (in years)

AgeGroup = 1 if young (under 50), 2 if middle (50-69), 3 if old (70+)

Sex = 1 for female, 0 for male

Infection = 1 if infection is suspected, 0 if no infection

SysBP = systolic blood pressure (in mm of Hg)

Pulse = heart rate (in beats per minute)

Emergency = 1 if emergency admission, 0 if elective

library(mosaic)

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## Loading required package: lattice

## Loading required package: ggformula

## Loading required package: ggplot2

## Loading required package: ggstance

##   
## Attaching package: 'ggstance'

## The following objects are masked from 'package:ggplot2':  
##   
## geom\_errorbarh, GeomErrorbarh

##   
## New to ggformula? Try the tutorials:   
## learnr::run\_tutorial("introduction", package = "ggformula")  
## learnr::run\_tutorial("refining", package = "ggformula")

## Loading required package: mosaicData

## Loading required package: Matrix

##   
## The 'mosaic' package masks several functions from core packages in order to add   
## additional features. The original behavior of these functions should not be affected by this.  
##   
## Note: If you use the Matrix package, be sure to load it BEFORE loading mosaic.

##   
## Attaching package: 'mosaic'

## The following object is masked from 'package:Matrix':  
##   
## mean

## The following object is masked from 'package:ggplot2':  
##   
## stat

## The following objects are masked from 'package:dplyr':  
##   
## count, do, tally

## The following objects are masked from 'package:stats':  
##   
## binom.test, cor, cor.test, cov, fivenum, IQR, median,  
## prop.test, quantile, sd, t.test, var

## The following objects are masked from 'package:base':  
##   
## max, mean, min, prod, range, sample, sum

library(Stat2Data)  
data(ICU)

## PART I. PREDICTORS WITH 2 CATEGORIES (BINARY PREDICTORS)

It is well-known that women who have heart attacks are more likely to die, even in the hospital, than men who have heart attacks. We’d like to use this data set to test if this true for ICU patients in general.

## A. Data Collection

1. Make a two-way table of survival by gender.

tab = tally (~Sex+Survive, data = ICU) #Sex = 1 for female, 0 for male  
tab

## Survive  
## Sex 0 1  
## 0 24 100  
## 1 16 60

1. The proportion of males that survived is: 80.64%

100/(100+24)

## [1] 0.8064516

The proportion of females that survived is: 78.94%

60/(60+16)

## [1] 0.7894737

1. Based only on the proportions above, does it seem that our hypothesis is reasonable?

* Not really, since the proportion of male and female surviving are almost equal.

## B. 2-Sample Proportion Test

**1. Hypotheses.** If is the probability of survival for the men and is the probability of survival for the women, the hypotheses are (write in words and in symbols): - : p1 = p2 versus : p1 ≠ p2

* >

**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 selected independently of each other

Are the conditions met in this case? Make sure you discuss all 3 conditions.

* Yes, in both group, number of success and failure are at least 10.
* Selected independently: Not really, since people in the same unit has different caring system…

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

**Using prop.test() with 2 samples has a few different syntax options…**

1. *If you don’t have the mosaic package loaded*, the syntax is:

prop.test(table, correct=F) or prop.test(successes, total, correct=F)

where “table” is a 2x2 table (or matrix), where the first column is successes and the second column is failures, and there is one row for each gender (just like in A1). To get this table, just create a matrix yourself:

table1 <- matrix(data=c(100, 60, 24, 16), nrow=2, ncol=2)  
prop.test(table1, correct=FALSE, alternative="greater")

## Warning in stats::prop.test(x = count, n = n, p = p, alternative =  
## alternative, : Chi-squared approximation may be incorrect

##   
## 1-sample proportions test without continuity correction  
##   
## data: table1 [with success = 16]  
## X-squared = 1, df = 1, p-value = 0.8413  
## alternative hypothesis: true p is greater than 0.5  
## 95 percent confidence interval:  
## 0.05790732 1.00000000  
## sample estimates:  
## p   
## 0.25

Alternatively,

prop.test(c(100,60),c(124,76), correct=FALSE, alternative="greater") #the 1st vector is successes for men & women; 2nd vector is total # of men & women

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: c(100, 60) out of c(124, 76)  
## X-squared = 0.08489, df = 1, p-value = 0.3854  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## -0.0795748 1.0000000  
## sample estimates:  
## prop 1 prop 2   
## 0.8064516 0.7894737

1. *If you do have the mosaic package loaded*, the syntax is:

prop.test( ~ response var, groups=grouping var, correct=F) or or prop.test(successes, total, correct=F)

prop.test(~Survive, groups=Sex,data=ICU,correct=F,alternative="g")

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: tally(Survive ~ Sex)  
## X-squared = 0.08489, df = 1, p-value = 0.6146  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## -0.1135307 1.0000000  
## sample estimates:  
## prop 1 prop 2   
## 0.1935484 0.2105263

Hey! This gave you a different answer than in part (a)! Even though the mosaic version of prop.test is supposed to treat 0 and failure and 1 as success, and test that the probabilities of success are equal, it seems to be testing that the probabilities of *failure* are equal, instead. There are a couple of possible fixes to this problem, so that we get the test we want. How could you change the code (given again below) to get the same p-value as in part (a)?

c = tally(Sex ~ Survive, data = ICU)  
  
prop.test(c[,2],(c[,1]+c[,2]),correct=F,alternative="greater") #change this so the p-value is the same as in part (a)

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: c[, 2] out of (c[, 1] + c[, 2])  
## X-squared = 0.08489, df = 1, p-value = 0.3854  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## -0.0795748 1.0000000  
## sample estimates:  
## prop 1 prop 2   
## 0.8064516 0.7894737

Alternatively, you could use:

prop.test(c(100,60),c(124,76), correct=FALSE, alternative="greater") #notice that this syntax still works, whether or not mosaic is loaded

##   
## 2-sample test for equality of proportions without continuity  
## correction  
##   
## data: c(100, 60) out of c(124, 76)  
## X-squared = 0.08489, df = 1, p-value = 0.3854  
## alternative hypothesis: greater  
## 95 percent confidence interval:  
## -0.0795748 1.0000000  
## sample estimates:  
## prop 1 prop 2   
## 0.8064516 0.7894737

Note that prop.test doesn’t give you the z test statistic; instead it reports the chi-squared statistic (“X-squared”). Since the chi-squared is just the standard normal squared, “X-squared” is just z squared.

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

* p-value is 0.3854 > 0.05, so we accept the null.

**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 90% confidence interval for the difference in survival between men and women:

#CI FOR (MEN - WOMAN (MINUS)) - CI: [-0.0795748,1.0000000]

* We are 90% confidence that, based on these samples, that the proportion of men having heart attack surviving is between -7.96% and 100% more than the proportion of women having heart attack surviving.

## C. Chi-Squared Test for Association

Another way of testing whether the two genders have significantly different probabilities of survival 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…

: the two variables (gender and survival) are independent vs. : not independent

OR

: the proportion of survival is the same for both populations (men and women) vs. : 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 in A1. Under the null hypothesis of the Chi-square Test, we’d expect men and women to survive 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, , compares the observed cell counts to the expected cell counts under this assumption.

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

*Under the null hypothesis*, we’d expect 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 are an indication that is false. Under , follows a chi-squared distribution with 1 degree of freedom.

**4. Using R to find C3.** Again, there are two ways to compute the Chi-sq test statistic and p-value in R.

**a. Using chisq.test():** If you have a 2-way table of successes and failures,

sex.chisq <- chisq.test(table1, correct=F); sex.chisq

##   
## Pearson's Chi-squared test  
##   
## data: table1  
## X-squared = 0.08489, df = 1, p-value = 0.7708

**b. Using summary():** If you created a matrix of successes/failures using xtabs() or table(), you can ask for the summary() of that matrix, and this will include the Chi-square test statistic and p-value.

summary(xtabs(~Sex+Survive,data=ICU))

## Call: xtabs(formula = ~Sex + Survive, data = ICU)  
## Number of cases in table: 200   
## Number of factors: 2   
## Test for independence of all factors:  
## Chisq = 0.08489, df = 1, p-value = 0.7708

Notice that neither of these methods care which column is successes and which column is failures! (Unlike with prop.test.)

Report the test statistic and p-value here:

* Chisq = 0.08489, df = 1, p-value = 0.7708

**5.** Why don’t the p-values in C4 and B3 match up?

* B3: p-value = 0.3854
* The chi-square test of independence to test for equality of proportions between populations. In this case, it is called a test for equality of proportions rather than a test for independence. The null and alternative hypotheses are written differently, but the rest of the test procedure remains the same

**6. Back to the assumptions.** The chisq.test() object sex.chisq contains the expected cell counts in sex.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?

* Yes

**7. Conclusion.** Make a conclusion about in the context of the problem.

* p-value = 0.7708 > 0.05, accept the null

## D. Logistic Regression

(A review, since we saw this with the MathPlacement data when ‘RecTaken’ was our single predictor.)

The final way of investigating the relationship between a binary predictor (male/female) and a binary response (survival/death) is using logistic regression.

**1. Hypotheses (Wald test).** vs.

The hypotheses above are the standard Wald test hypotheses. But remember that in this case, we want to test whether women are *less* likely to survive. So what should our hypotheses be? (Hint: Think about what has to be if women are less likely to survive.)

vs.

logit(π) =β0 +β1(Sex)

**2. Assumptions/Conditions.** The conditions are the same as we discussed earlier in Chapter 9:

* Randomness – you’ve already discussed this
* Independence – you’ve already discussed this
* Linearity – This is a given with binary predictor variables! There are only two points on your logits vs. X plot, and two points always make a line, by definition.

**3. The Logistic Regression Model** Use R to fit a logistic regression model with “Survive” as the response variable and “Sex” as the explanatory variable. Save this as sex.log and write the fitted model below.

sex.log = glm(Survive ~ Sex, data = ICU, family = binomial)  
summary(sex.log)

##   
## Call:  
## glm(formula = Survive ~ Sex, family = binomial, data = ICU)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.8123 0.6559 0.6559 0.6876 0.6876   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 1.4271 0.2273 6.278 3.42e-10 \*\*\*  
## Sex -0.1054 0.3617 -0.291 0.771   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 200.16 on 199 degrees of freedom  
## Residual deviance: 200.08 on 198 degrees of freedom  
## AIC: 204.08  
##   
## Number of Fisher Scoring iterations: 4

Model : logit(π) = 1.4271 -0.1054Sex

**4. Test statistic & p-value.**

The fitted slope is: 3.42e-10 \*\*\*

…and the Wald test statistic is:

…which has a p-value of:

Also note the G-test statistic is:200.16 - 200.08 = 0.08/1 unit

…which has a p-value of:

**5. Conclusion.** Make a conclusion about in the context of the problem.

**6. Interpretation of slope/odds ratio.** Calculate and interpret the odds ratio between men and women:

oddsRatio(tab)

## [1] 1.111111

* The odds of a male heart attack patient surviving is 111% of the odds of a female heart attack patient surviving.

**7. Confidence interval for the odds ratio.** Use confint() to calculate a 90% confidence interval for the slope of this model.

#OR(Female/Male) : cai gi can so sanh thi dat len truoc

exp(confint(sex.log))

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 2.7180428 6.653023  
## Sex 0.4455739 1.854424

* Interpret the CI in terms of odds ratios:
* We are 90% confident that the odd ratios of picking a male heart attack patient surviving is between 44.55% and 185.44% of the odd of a female heart attack patient surviving.

## E. Synthesis

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

It seems logical that older people would be more likely to not survive the ICU. 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 survival by age group.

tab\_age = tally (~AgeGroup+Survive, data = ICU)   
tab\_age

## Survive  
## AgeGroup 0 1  
## 1 5 54  
## 2 17 60  
## 3 18 46

1. The proportion of younger people that survived is:

54/59

## [1] 0.9152542

The proportion of middle-range people that survived is:

60/77

## [1] 0.7792208

The proportion of older people that survived is:

46/(46+18)

## [1] 0.71875

1. Based only on the proportions above, does it seem that there is a relationship between age group and survival?

* Yes, since the middle-range and older people has lower survival rate than the younger age.

## B. ~~2-Sample Proportion Test~~

We can’t do a 2-sample proportion test here because we have 3 proportions!

## C. Chi-Squared Test: Test for Association

**1. Hypotheses**

: the two variables (age and survival) are independent vs. : not independent

**2. Assumptions/Conditions** Remember that the 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

We’ll check the first condition later. Discuss whether you think the second condition is met:

**3. Conduct the Chi-sq test** and save it as agegp.chisq. Report the Test Stat and p-value:

agegp.chisq = chisq.test(ICU$Survive, ICU$AgeGroup, correct=FALSE)  
agegp.chisq

##   
## Pearson's Chi-squared test  
##   
## data: ICU$Survive and ICU$AgeGroup  
## X-squared = 7.7467, df = 2, p-value = 0.02079

* p-value = 0.02079 < 0.05

**4. Back to the assumptions.** Is this condition that all expected cell counts are at least 5 met in this case?

**5. Conclusion.** Make a conclusion about in the context of the problem.

**6. …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 in C3. 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 agegp.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.

agegp.chisq$residuals

## ICU$AgeGroup  
## ICU$Survive 1 2 3  
## 0 -1.9795565 0.4077178 1.4534442  
## 1 0.9897783 -0.2038589 -0.7267221

* Group 1 and 3: highest (|a| > 1) in group die
* Old people dies more, young people die less

**a.** In this case, what two cells have the largest contribution?

* Younger-rage age and older-age

**b.** What direction are the residuals for the two cells (positive or negative)? Does this mean the observed value of that cell was higher or lower than what was expected?

**c.** Interpret what this means in the context of the problem: How does it relate to our original research question (about whether old people are more likely to die)?

## 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 AgeGroup are numerical. So glm() will think that AgeGroup is a numerical variable unless you force it to be a factor (categorical variable). We’ve done this before: just use as.factor(AgeGroup) in the glm call.

agegp.log <- glm(Survive~as.factor(AgeGroup), family=binomial, data=ICU)  
summary(agegp.log)

##   
## Call:  
## glm(formula = Survive ~ as.factor(AgeGroup), family = binomial,   
## data = ICU)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.2218 0.4208 0.7063 0.7063 0.8127   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 2.3795 0.4675 5.090 3.57e-07 \*\*\*  
## as.factor(AgeGroup)2 -1.1184 0.5422 -2.063 0.03915 \*   
## as.factor(AgeGroup)3 -1.4413 0.5439 -2.650 0.00805 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 200.16 on 199 degrees of freedom  
## Residual deviance: 191.59 on 197 degrees of freedom  
## AIC: 197.59  
##   
## Number of Fisher Scoring iterations: 5

**1. The Fitted Logistic Regression Model:**

logit(oddsofSurvival) = 2.3795 - 1.1184as.factor(AgeGroup)2 - 1.4413as.factor(AgeGroup)3

**2.** This model looks different! Let’s unpack it a little bit…

The fitted model when AgeGroup=1 is: logit(oddsofSurvival) = 2.3795

The fitted model when AgeGroup=2 is: logit(oddsofSurvival) = - 1.1184as.factor(AgeGroup)2

The fitted model when AgeGroup=3 is: logit(oddsofSurvival) = - 1.4413as.factor(AgeGroup)3

\*\* Coefficient: negative: less likely to survive; the old: biggest coefficient ( |1.4413|) and negative - most likely to die.

**3. Interpretation of slope/odds ratio.**

The coefficient of AgeGroup=2 is -1.1184, which means the odds ratio for that group is \_\_\_\_\_\_\_\_\_\_.

Interpret this odds ratio:

The coefficient of AgeGroup=3 is -1.4413, which means the odds ratio for that group is \_\_\_\_\_\_\_\_\_\_. Interpret this odds ratio:

**4. Confidence interval for the slope/odds ratio.**

exp(2.3795 )

## [1] 10.7995

**a.** Use confint() to calculate a 90% confidence interval for the slope of AgeGroup=2:

exp(confint(agegp.log,'as.factor(AgeGroup)2', level = 0.9))

## Waiting for profiling to be done...

## 5 % 95 %   
## 0.1249490 0.7621619

Interpret the CI in terms of odds ratios: - The odds of a mid-age person surviving are between 12.5% and 76% the odds of a young person surviving with 90% confident

**b.** Use confint() to calculate a 90% confidence interval for the slope of AgeGroup=3:

exp(confint(agegp.log,'as.factor(AgeGroup)3', level = 0.9))

## Waiting for profiling to be done...

## 5 % 95 %   
## 0.09020755 0.55288535

Interpret the CI in terms of odds ratios:

**5. Hypotheses (G-test for model utility)**

vs. at least one , where is the slope for AgeGroup=2 and is the slope for AgeGroup=3.

**6. 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. But notice that this now has 2 degrees of freedom, not 1! (This is because there are essentially two terms in the model, AgeGroup2 and AgeGroup3, rather than one.)

Report the test stat and the p-value:

**7. Conclusion.**

## E. Synthesis

1. How does your p-value and conclusion in D6/D7 compare to your p-value and conclusion in C3/C5?

* Smaller

1. 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?