QDA Lab 8 part 2 - a Solution

Isabel Sassoon

November 2020

Part 1 - ARM data

```
arm.folding<-read.csv("arm.csv")
```

This data contains two columns only

```
summary(arm.folding)
```

```
##     gender            armcross
##  Length:54          Length:54
##  Class :character   Class :character
##  Mode  :character   Mode  :character
```

The two categorical variables are the gender of the participant and the arm crossed on top.

```
table(arm.folding$gender, arm.folding$armcross)
```

```
##
##      L  R
##   F  5  9
##   M 17 23
```

We can test for independence using χ^2, as the expected number in the cells is more than 5. Note that if that were not the case R would complain!

```
arm.table<-table(arm.folding$gender, arm.folding$armcross)
chisq.test(arm.table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  arm.table
## X-squared = 0.016574, df = 1, p-value = 0.8976
```

We can see from here that this is not significant.

If you wanted to run Fisher's exact:

```
fisher.test(arm.table)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  arm.table
## p-value = 0.7582
```

```
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1669855 3.0781098
## sample estimates:
## odds ratio
##  0.7555721
```

Note: it is not necessary in this case but wanted to show that it would return the similar conclusion. ***

Part 2 - back to titanic

```
titanic<-read.csv("titanic-all-cols.csv")
```

Recall from week 7 - we need to make sure that the data is read in correctly:

```
summary(titanic)
```

```
##   PassengerId       Survived          Pclass          Name
##  Min.   :  1.0   Min.   :0.0000   Min.   :1.000   Length:891
##  1st Qu.:223.5   1st Qu.:0.0000   1st Qu.:2.000   Class :character
##  Median :446.0   Median :0.0000   Median :3.000   Mode  :character
##  Mean   :446.0   Mean   :0.3838   Mean   :2.309
##  3rd Qu.:668.5   3rd Qu.:1.0000   3rd Qu.:3.000
##  Max.   :891.0   Max.   :1.0000   Max.   :3.000
##
##      Sex                 Age            SibSp           Parch
##  Length:891         Min.   : 0.42   Min.   :0.000   Min.   :0.0000
##  Class :character   1st Qu.:20.12   1st Qu.:0.000   1st Qu.:0.0000
##  Mode  :character   Median :28.00   Median :0.000   Median :0.0000
##                     Mean   :29.70   Mean   :0.523   Mean   :0.3816
##                     3rd Qu.:38.00   3rd Qu.:1.000   3rd Qu.:0.0000
##                     Max.   :80.00   Max.   :8.000   Max.   :6.0000
##                     NA's   :177
##     Ticket               Fare           Cabin             Embarked
##  Length:891         Min.   :  0.00   Length:891         Length:891
##  Class :character   1st Qu.:  7.91   Class :character   Class :character
##  Mode  :character   Median : 14.45   Mode  :character   Mode  :character
##                     Mean   : 32.20
##                     3rd Qu.: 31.00
##                     Max.   :512.33
##
```

We need to make sure that both Survived and Pclass are treated as categorical, so I use as.factor to make sure that is the case.

```
titanic$Survived<-as.factor(titanic$Survived)
titanic$Pclass<-as.factor(titanic$Pclass)
```

Lets look at table analysis for Survived vs Pclass

Are survival status and Passenger class independent? THat is the H_0 that we are testing here

```
table(titanic$Survived, titanic$Pclass)
```

```
##
##       1   2   3
```

```
##   0  80  97 372
##   1 136  87 119
```

Lets use Chisq

```
chisq.test(table(titanic$Survived, titanic$Pclass))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(titanic$Survived, titanic$Pclass)
## X-squared = 102.89, df = 2, p-value < 2.2e-16
```

We can see that there is a significant relationship between Passenger travel class and survival status. In other words the probability of surviviing depends on the passenger's travel class.

If we wanted to model this probability then this is when logistic regression can help - and this is what we did in Lab 7 part 2.

OPTIONAL: Using the gmodels library:

The Gmodels library is useful if you are used to SAS PROC means or PROC summary.

```
library(gmodels)
```

```
CrossTable(titanic$Survived, titanic$Pclass, digits=2, prop.r = TRUE, prop.c = TRUE,  prop.chisq = FALSE
           chisq = TRUE, fisher = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  891
##
##
##                  | titanic$Pclass
## titanic$Survived |         1 |         2 |         3 | Row Total |
## -----------------|-----------|-----------|-----------|-----------|
##                0 |        80 |        97 |       372 |       549 |
##                  |      0.15 |      0.18 |      0.68 |      0.62 |
##                  |      0.37 |      0.53 |      0.76 |           |
##                  |      0.09 |      0.11 |      0.42 |           |
## -----------------|-----------|-----------|-----------|-----------|
##                1 |       136 |        87 |       119 |       342 |
##                  |      0.40 |      0.25 |      0.35 |      0.38 |
##                  |      0.63 |      0.47 |      0.24 |           |
##                  |      0.15 |      0.10 |      0.13 |           |
## -----------------|-----------|-----------|-----------|-----------|
```

```
##     Column Total |       216 |       184 |       491 |       891 |
##                  |      0.24 |      0.21 |      0.55 |           |
## -----------------|-----------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  102.889     d.f. =  2     p =  4.549252e-23
##
##
##
## Fisher's Exact Test for Count Data
## ------------------------------------------------------------
## Alternative hypothesis: two.sided
## p =  3.306641e-23
##
##
```

As you can see it provides both tests. When you are reporting your findings you should still comment on the one that is appropriate to the situation.