Lab 18 - Chi-Square Test

Nick Sumpter (Edited by Eddie-Williams Owiredu & Guy Twa)

2023-10-25

Today’s Lab

This lab will introduce you to how to analyze frequency data. We will learn how to perform the chi-square test and the related Fisher’s exact test. The goal of these tests is to test for a correlation between two (or more) categorical variables.

Loading Packages and Data

For today’s lab, we will be using two new packages called `gmodels` and `epitools` , which you should go ahead and install prior to loading it in.

```
install.packages("gmodels")
install.packages("epitools")
```

```
library(gmodels)
library(epitools)
library(tidyverse)

theme_set(theme_bw())

setwd("~/Documents/GRD770/Lab 18 - Chi-Square Test")

load("Categorical-Data.RData")
```

Getting to Know Your Data

The dataset you will be using for this lab is called `bacteria` , which has 50 observations of the following three variables:

1. `drug.tx` (factor, 2 levels): current drug treatment, either treatment or placebo
2. `diet` (factor, 2 levels): vegetarian or meat eater
3. `bac.presence` (factor, 2 levels): presence or absence of bacteria

This is simulated data from an experiment testing whether an experimental drug has adverse effects on the presence of a certain bacteria. Additional data about the person's diet was also collected as it may play a role in the presence or absence of that bacteria. For this lab, we will be ignoring the `diet` variable.

We can get a quick summary of the variables using the `summary` function:

```
summary(bacteria)
```

```
##       drug.tx     diet         bac.presence
##  Placebo  :20   Veg :31   Not Present:30
##  Treatment:30   Meat:19   Present    :20
```

Let's also go ahead and set an order for these factors, as that will be important for interpreting our resutls later.

```
bacteria <- bacteria %>%
  mutate(drug.tx = factor(drug.tx, levels = c("Treatment", "Placebo")),
         bac.presence = factor(bac.presence, levels = c("Present", "Not Present")))
```

To visualize the relationship between these variables, a contingency table must be created, which essentially lists the frequency of all combinations of your categorical variables of interest. We will use the `xtabs` function for this:

```
conting_table <- xtabs(~ drug.tx + bac.presence, data = bacteria)

conting_table
```

```
##             bac.presence
## drug.tx     Present Not Present
##   Treatment       5          25
##   Placebo        15           5
```

Basically, we told the function to make a table "with respect to" both `drug.tx` and `bac.presence` based on the `bacteria` dataset. As you can see it has output a 2 by 2 table with the number of rows that fall into each combination of the two variables displayed within each cell. For example, the top left cell tells us that 5 individuals were on the placebo drug and didn't have bacteria present.

Assumptions of the Chi-Square Test

There are two assumptions for this test:

1. Independence: each individual datapoint should contribute to only one cell in the contingency table (i.e. they can't belong to more than one level of any categorical variable of interest)

2. The expected frequencies should be greater than 5. Up to 20% of larger tables can be below 5, but none should be lower than 1. If any are lower than 1, or there are many cells with expected frequencies below 5, use Fisher’s exact test.

Running the Chi-Square Test

We will be using the `CrossTable` function from the `gmodels` package to run the chi-square test on our data. This has a lot of different arguments, but we are interested in the following:

- `x =` your contingency table
- `chisq` : whether to perform a chi-square test
- `expected =` whether to include expected values for each unique group from the chi-square test
- `fisher` : whether to perform a Fisher’s exact test
- `resid` : whether to include residuals
- `sresid` : whether to include standardized residuals
- `format` : whether to print using “SAS” or “SPSS” format - use “SPSS”

Let’s go ahead and set all of these to TRUE and run the model as follows:

```
CrossTable(x = conting_table,
           chisq = TRUE,
           expected = TRUE,
           resid = TRUE,
           sresid = TRUE,
           format = "SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |                Residual |
## |            Std Residual |
## |-------------------------|
##
## Total Observations in Table:  50
##
##              | bac.presence
##      drug.tx |     Present  | Not Present  |   Row Total |
## -------------|-------------|-------------|-------------|
##    Treatment |          5  |         25  |         30  |
##              |     12.000  |     18.000  |             |
##              |      4.083  |      2.722  |             |
##              |     16.667% |     83.333% |     60.000% |
##              |     25.000% |     83.333% |             |
##              |     10.000% |     50.000% |             |
##              |     -7.000  |      7.000  |             |
##              |     -2.021  |      1.650  |             |
## -------------|-------------|-------------|-------------|
##      Placebo |         15  |          5  |         20  |
##              |      8.000  |     12.000  |             |
##              |      6.125  |      4.083  |             |
##              |     75.000% |     25.000% |     40.000% |
##              |     75.000% |     16.667% |             |
##              |     30.000% |     10.000% |             |
##              |      7.000  |     -7.000  |             |
##              |      2.475  |     -2.021  |             |
## -------------|-------------|-------------|-------------|
## Column Total |         20  |         30  |         50  |
##              |     40.000% |     60.000% |             |
## -------------|-------------|-------------|-------------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  17.01389     d.f. =  1     p =  3.710739e-05
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  14.67014     d.f. =  1     p =  0.0001280591
##
```

```
## 
##        Minimum expected frequency: 8
```

The top box shows how the following table is organized. We can see the expected frequencies for each group as the second row in each box. The table also gives the row-wise, column-wise, and total percentage of data in each cell. Raw and standardized residuals are reported below this. Results from the Chi-square test with and without the Yates' correction are shown below the table. We will just use the normal Chi-square test values, Yate's correction is beyond the scope of this class.

The first thing we need to check is our expected frequencies assumption. In our example, none of the expected values are below 5, therefore we can use the normal chi-square output. Based on this, there is a significant relationship between drug treatment status and the presence or absence of bacteria.

Fisher's Exact Test

If it does come out that cells have expected values below 5, or any of the cells have expected values below 1, use the results from the Fisher's Exact test only (ignore the Chi-square test results). We can run this test by adding the `fisher=TRUE` argument to `CrossTable()`

```
CrossTable(x = conting_table,
           chisq = FALSE,
           expected = TRUE,
           resid = TRUE,
           sresid = TRUE,
           fisher = TRUE,
           format = "SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |         Expected Values |
## | Chi-square contribution |
## |             Row Percent |
## |          Column Percent |
## |           Total Percent |
## |                Residual |
## |            Std Residual |
## |-------------------------|
##
## Total Observations in Table:  50
##
##              | bac.presence
##      drug.tx |     Present  | Not Present  |   Row Total |
## -------------|-------------|-------------|-------------|
##    Treatment |          5  |         25  |         30  |
##              |     12.000  |     18.000  |             |
##              |      4.083  |      2.722  |             |
##              |     16.667% |     83.333% |     60.000% |
##              |     25.000% |     83.333% |             |
##              |     10.000% |     50.000% |             |
##              |     -7.000  |      7.000  |             |
##              |     -2.021  |      1.650  |             |
## -------------|-------------|-------------|-------------|
##      Placebo |         15  |          5  |         20  |
##              |      8.000  |     12.000  |             |
##              |      6.125  |      4.083  |             |
##              |     75.000% |     25.000% |     40.000% |
##              |     75.000% |     16.667% |             |
##              |     30.000% |     10.000% |             |
##              |      7.000  |     -7.000  |             |
##              |      2.475  |     -2.021  |             |
## -------------|-------------|-------------|-------------|
## Column Total |         20  |         30  |         50  |
##              |     40.000% |     60.000% |             |
## -------------|-------------|-------------|-------------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  17.01389     d.f. =  1     p =  3.710739e-05
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  14.67014     d.f. =  1     p =  0.0001280591
##
```

```
## 
## Fisher's Exact Test for Count Data
## ------------------------------------------------------------
## Sample estimate odds ratio:  0.07174891 
## 
## Alternative hypothesis: true odds ratio is not equal to 1
## p =  7.361661e-05 
## 95% confidence interval:  0.01300179 0.3183426 
## 
## Alternative hypothesis: true odds ratio is less than 1
## p =  4.979718e-05 
## 95% confidence interval:  0 0.2627504 
## 
## Alternative hypothesis: true odds ratio is greater than 1
## p =  0.9999971 
## 95% confidence interval:  0.01679956 Inf 
## 
## 
## 
##        Minimum expected frequency: 8
```

You will see 3 entries under the Fisher’s Exact test section in the output table. For reporting the statistical test, use the first enty where the alternative hypothesis is that the true odds ratio is not equal to 1.

Interpretation

We have a significant test, but what does that really tell us. We can use the standardized residuals to tell us a little bit more. Basically, standardized residuals > 1.96 are significant at $p < 0.05$, standardized residuals > 2.58 are significant at $p < 0.01$, and standardized residuals > 3.29 are significant at $p < 0.001$. If we look through our table, we see that we have significant standardized residuals in 3 out of 4 cells. If a standardized residual for a cell is significant, that means the true value of that cell was significantly higher or lower than what was predicted (directionality based on the sign). For this example, the group with the placebo drug treatment showed a significantly lower frequency of individuals without bacteria than expected ($p < 0.05$) and a significantly higher frequency of individuals with bacteria than expected ($p < 0.05$). In the group treated with the drug there was a significantly lower frequency of those with bacteria present than expected ($p < 0.05$).

Odds Ratio

Further to the above interpretation, we can calculate an odds ratio, which is essentially an effect size for the association.

```
treatment.present <- conting_table[1,1]
treatment.notpresent <- conting_table[1,2]

placebo.present <- conting_table[2,1]
placebo.notpresent <- conting_table[2,2]

odds.present <-  treatment.present / placebo.present
odds.notpresent <-  treatment.notpresent / placebo.notpresent


odds.ratio <- odds.present / odds.notpresent
odds.ratio
```

```
## [1] 0.06666667
```

For this, we have an odds ratio of 0.067. This means that the odds of an individual treated with the drug having bacteria present was 0.067 times that of the odds of an individual treated with a placebo having bacteria present.

Be careful reading the odds ratio. It should be read in the exact order your table is in. In our example, the case where the individuals were given a placebo is 1st in the order followed by treatment with the drug, and the case where bacteria are absent is 1st followed by bacteria being present. If the level orders were reversed in either factor, the odds ratio would be inverted.

You'll notice the odds ratio returned by the Fisher's test is close to, but does not match, the odds ratio we calculated. This is because the Fisher's Exact Test function that `R` uses the conditional Maximum Likelihood Estimate (MLE) rather than the unconditional MLE (the sample odds ratio).

We can get the 95% confidence interval of our sample odds ratio using the `epitools` function `oddsratio.wald()`

```
oddsratio.wald(x = conting_table)
```

```
## $data
##            bac.presence
## drug.tx     Present Not Present Total
##   Treatment       5          25    30
##   Placebo        15           5    20
##   Total          20          30    50
##
## $measure
##            odds ratio with 95% C.I.
## drug.tx       estimate      lower     upper
##   Treatment 1.00000000         NA        NA
##   Placebo   0.06666667 0.01652034 0.2690286
##
## $p.value
##            two-sided
## drug.tx       midp.exact fisher.exact   chi.square
##   Treatment           NA           NA           NA
##   Placebo   5.271446e-05 7.361661e-05 3.710739e-05
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

The `$measure` section of the output contains the sample odds ratio with 95% confidence interval.

Reporting a Chi-Square Test

For this example we would report it like so:

There was a significant association between training reward and ability to dance χ2(1) = 17.01, p = 3.7e-5. This seems to represent the fact that, based on the odds ratio, the odds of individuals treated with the drug having bacteria present were 0.067 (0.016, 0.269) times those of individuals treated with a placebo.

If the Fisher's Exact Test was used instead, do not report the chi-square statistic. Only report the p-value under the Fisher's Exact test section and say the analysis was done with Fisher's Exact Test. Also report the odds ratio and the upper and lower confidence bounds like normal.

Independent Practice

For your independent practice, repeat the above analysis on the `cats` dataset.

1. Get to know your data: describe your variables and their distribution
2. Run the Chi-Square Test
3. Assumptions
    a. Determine whether the assumptions of both tests are met

    b. If assumptions are not met, describe what you will do to account for this. If possible, modify your models to meet the assumptions

4. Report your findings as you would describe them in the results section