Chi-square test for independence

Put your name here

Put the date here

Introduction

In this assignment we will learn how to run the chi-square test for independence.

A chi-square test for independence tests the relationship between two categorical variables. This is an extension of the test for two proportions, except now applied in situations where either the explanatory or response variables (or both) have three or more categories.

When there are only two categories, one can still run a chi-square test and the results are nearly equivalent to the two-proportion test.

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document and work back and forth between this R Markdown file and the PDF output as you work through this module.

When you are finished with the assignment, knit to PDF one last time, proofread the PDF file **carefully**, export the PDF file to your computer, and then submit your assignment.

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
## Add code here to [do some task]...
```

Be sure to remove the line `## Add code here to [do some task]...` when you have added your own code.

Sometimes you will be asked to type up your thoughts. That will appear in the document as follows:

Please write up your answer here.

Again, please be sure to remove the line "Please write up your answer here" when you have written up your answer. In these areas of the assignment, please use contextually meaningful full sentences/paragraphs (unless otherwise indicated) and proper spelling, grammar, punctuation, etc. This is not R code, but rather a free response section where you talk about your analysis and conclusions. If you need to use some R code as well, you can use inline R code inside the block between `\begin{answer}` and `\end{answer}`, or if you need an R code chunk, please go outside the `answer` block and start a new code chunk.

Load Packages

We load the standard `mosaic` package. We also use the `gmodels` package for the `CrossTable` command, the `MASS` package for the `birthwt` data, and the `openintro` package for the `smoking` data.

```r
library(gmodels)
library(MASS)
library(openintro)
library(mosaic)
```

Research question

Are mothers from certain races more or less likely to have low birth weight babies? In other words, are race and low birth weight associated?

Chi-square test for independence

In a previous module, we learned about the chi-square goodness-of-fit test. With a single categorical variable, we summarized data in a frequency table. Each cell of the table had an observed count from the data which we compared to an expected count from the assumption of a null hypothesis. The chi-square statistic measured the discrepancy between observed and expected.

With two categorical variables, we use a contingency table instead of a frequency table. But the principle of the chi-square statistic is the same: each cell in the contingency table has an observed count and an expected count. This forms the basis of a chi-square test for independence.

A test for independence has a simple null hypothesis: the two variables are independent. This makes it easy to compute expected counts. Here's how it works.

After converting to factor variables, look at the contingency table for the relationship between `race` and `low` in the `birthwt` data from the `MASS` package. The new option in the `CrossTable` command is `expected = TRUE`:

```r
race <- factor(birthwt$race, level = c(1, 2, 3),
               labels = c("White",  "Black",  "Other"))
low <- factor(birthwt$low, levels = c(1, 0), labels = c("Yes",  "No"))
race_low <- data.frame(race, low)
CrossTable(race_low$race, race_low$low,
           prop.r = TRUE, prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE,
           expected = TRUE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## |           N / Row Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  189 
## 
## 
##               | race_low$low 
## race_low$race |       Yes |        No | Row Total | 
## --------------|-----------|-----------|-----------|
##         White |        23 |        73 |        96 | 
##               |    29.968 |    66.032 |           | 
##               |     0.240 |     0.760 |     0.508 | 
## --------------|-----------|-----------|-----------|
##         Black |        11 |        15 |        26 | 
##               |     8.116 |    17.884 |           | 
##               |     0.423 |     0.577 |     0.138 | 
## --------------|-----------|-----------|-----------|
##         Other |        25 |        42 |        67 | 
##               |    20.915 |    46.085 |           | 
##               |     0.373 |     0.627 |     0.354 | 
## --------------|-----------|-----------|-----------|
##  Column Total |        59 |       130 |       189 | 
## --------------|-----------|-----------|-----------|
## 
## 
## Statistics for All Table Factors
## 
## 
## Pearson's Chi-squared test 
## ------------------------------------------------------------
## Chi^2 =  5.004813     d.f. =  2     p =  0.0818877 
## 
## 
## 
```

To see how these expected counts are computed, look at the marginal distribution at the bottom. If race is ignored, there were 59 low birth weight babies and 130 normal weight babies out of 189 total babies. 59 of 189 is 0.3121693 or 31.2169312%, and 130 of 189 is 0.6878307 or 68.7830688%.

Now, if race and low birth weight are truly independent, it shouldn't matter if the mothers were white, black, or some other race. In other words, of 96 white mothers, we should still expect 31.2169312% of them to have low birth weight babies and 68.7830688% of them to have normal weight babies. 31.2169312% of 96 is 29.968254. Look in the contingency table above and note that this is the expected cell count for white women with low birth weight babies. Also, 68.7830688% of 96 is 66.031746. Again, this is listed in the contingency table in the cell for white mothers with normal weight babies. The same analysis can be done for the next two rows as well.

Unlike the goodness-of-fit test that requires one to specify expected counts for each cell, the test for independence uses only the data to determine the expected counts. For any given cell, if C is the column total, R is the row total, and n is the grand total (the sample size), the expected count is simply

$$E = \frac{CR}{n}.$$

This is equivalent to the explanation in the previous paragraph. Using white mothers with low birth weight babies as an example, C/n is 59/189 which is 0.3121693. Then we multiply this by the row total $R = 96$ to get

$$\left(\frac{C}{n}\right) R = \frac{CR}{n} = \frac{59 * 96}{189} = 29.96825.$$

Everything else works almost the same as it did for a chi-square goodness-of-fit test. We still compute χ^2 by adding up deviations across all cells:

$$\chi^2 = \sum \frac{(O - E)^2}{E}.$$

Even under the assumption of the null, there will still be some sampling variability. Like any hypothesis test, our job is to determine whether the deviations we see are possible due to pure chance alone. The random values of χ^2 that result from sampling variability will follow a chi-square model. But how many degrees of freedom are there? This is a little different from the goodness-of-fit test. Instead of the number of cells minus one, we use the following formula:

$$df = (\#rows - 1)(\#columns - 1).$$

In our example we have 3 rows ("White", "Black", "Other") and 2 columns ("Yes", "No"); therefore,

$$df = (3 - 1)(2 - 1) = 2 * 1 = 2$$

and we have 2 degrees of freedom (even though there are 6 cells).

We will run the chi-square test for independence with the `chisq.test` command like before, but note that all the information you need is actually in the contingency table above. Let's look at the table one more time, and we'll turn on one more argument (setting `prop.chisq = TRUE`):

```r
CrossTable(race_low$race, race_low$low,
           prop.r = TRUE, prop.c = FALSE, prop.t = FALSE, prop.chisq = TRUE,
           expected = TRUE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  189 
## 
## 
##               | race_low$low 
## race_low$race |       Yes |        No | Row Total | 
## --------------|-----------|-----------|-----------|
##         White |        23 |        73 |        96 | 
##               |    29.968 |    66.032 |           | 
##               |     1.620 |     0.735 |           | 
##               |     0.240 |     0.760 |     0.508 | 
## --------------|-----------|-----------|-----------|
##         Black |        11 |        15 |        26 | 
##               |     8.116 |    17.884 |           | 
##               |     1.024 |     0.465 |           | 
##               |     0.423 |     0.577 |     0.138 | 
## --------------|-----------|-----------|-----------|
##         Other |        25 |        42 |        67 | 
##               |    20.915 |    46.085 |           | 
##               |     0.798 |     0.362 |           | 
##               |     0.373 |     0.627 |     0.354 | 
## --------------|-----------|-----------|-----------|
##  Column Total |        59 |       130 |       189 | 
## --------------|-----------|-----------|-----------|
## 
## 
## Statistics for All Table Factors
## 
## 
## Pearson's Chi-squared test 
## ------------------------------------------------------------
## Chi^2 =  5.004813     d.f. =  2     p =  0.0818877 
## 
## 
## 
```

Not only does this show the chi-square component $\frac{(O-E)^2}{E}$ in each cell, but it sums them all up in a summary right below the table, along with the degrees of freedom and a P-value.

Let's run through the rubric in its entirety.

Exploratory data analysis

Use data documentaton (help files, code books, Google, etc.), the str command, and other summary functions to understand the data.

[Type `library(MASS)` then `?birthwt` at the Console to read the help file.]

```r
str(birthwt)
```

```
## 'data.frame':    189 obs. of  10 variables:
##  $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
##  $ race : int  2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...
```

Prepare the data for analysis.

```r
# Although we've already done this above,
# we include it here again for completeness.
race <- factor(birthwt$race, level = c(1, 2, 3),
               labels = c("White",  "Black",  "Other"))
low <- factor(birthwt$low, levels = c(1, 0), labels = c("Yes",  "No"))
race_low <- data.frame(race, low)
```

Make tables or plots to explore the data visually.

```
CrossTable(race_low$race, race_low$low,
           prop.r = TRUE, prop.c = FALSE, prop.t = FALSE, prop.chisq = TRUE,
           expected = TRUE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  189
##
##
##               | race_low$low
## race_low$race |       Yes |        No | Row Total |
## --------------|-----------|-----------|-----------|
##         White |        23 |        73 |        96 |
##               |    29.968 |    66.032 |           |
##               |     1.620 |     0.735 |           |
##               |     0.240 |     0.760 |     0.508 |
## --------------|-----------|-----------|-----------|
##         Black |        11 |        15 |        26 |
##               |     8.116 |    17.884 |           |
##               |     1.024 |     0.465 |           |
##               |     0.423 |     0.577 |     0.138 |
## --------------|-----------|-----------|-----------|
##         Other |        25 |        42 |        67 |
##               |    20.915 |    46.085 |           |
##               |     0.798 |     0.362 |           |
##               |     0.373 |     0.627 |     0.354 |
## --------------|-----------|-----------|-----------|
##  Column Total |        59 |       130 |       189 |
## --------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  5.004813     d.f. =  2     p =  0.0818877
##
##
##
```

Hypotheses

Identify the sample (or samples) and a reasonable population (or populations) of interest.

The sample consists of 189 mothers who gave birth at the Baystate Medical Center in Springfield, Massachusetts in 1986. The population is presumably all mothers, although it's safest to conclude only about mothers who gave birth at this hospital.

Express the null and alternative hypotheses as contextually meaningful full sentences.

H_0: Race and low birth weight are independent.

H_A: Race and low birth weight are associated.

Express the null and alternative hypotheses in symbols.

For a chi-square test for independence, this section is not applicable. With multiple categories in the explanatory and response variables, there are no specific parameters of interest to express symbolically.

Model

Identify the sampling distribution model.

We will use a chi-square model with 2 degrees of freedom.

Check the relevant conditions to ensure that model assumptions are met.

- Random
  - We hope that these 189 women are representative of all women who gave birth in this hospital.
- 10%
  - We don't know how many women gave birth at this hospital, but perhaps over many years we might have more than 1890 women.
- Expected cell counts
  - The contingency table above shows the expected counts. We can verify that they are all greater than 5.

Mechanics

Compute the test statistic.

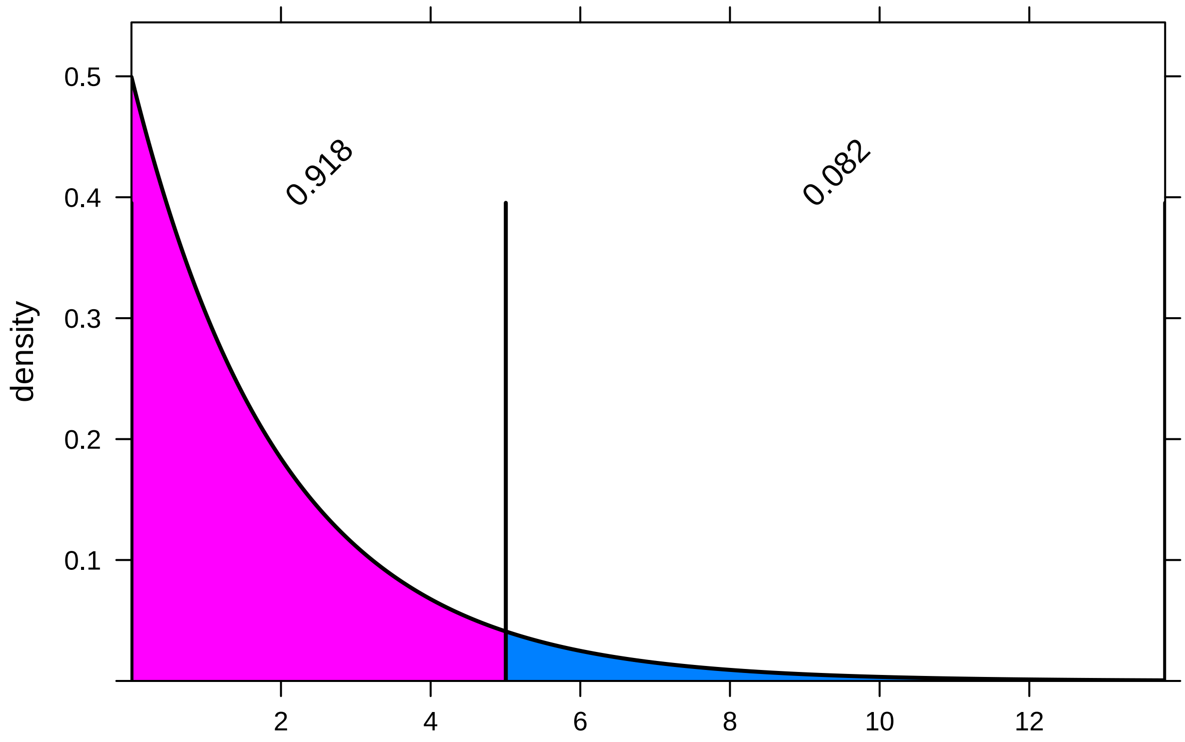
```
race_low_test <- chisq.test(race_low$race, race_low$low)
race_low_test$statistic
```

```
## X-squared
##  5.004813
```

Commentary: This just reproduces the value of χ^2 we already saw in the summary below the contingency table.

Plot the null distribution.

```
pdist("chisq", df = race_low_test$parameter, q = race_low_test$statistic)
```



```
## X-squared
## 0.9181123
```

Commentary: We know there are 2 degrees of freedom, but since the result is stored in `race_low_test$parameter`, we might as well use that. It makes our code more portable than if we had manually typed `df = 2`. (The next time we need to run this test, we might not have two degrees of freedom anymore.)

Calculate the P-value.

```
1 - pdist("chisq", df = race_low_test$parameter, q = race_low_test$statistic,
          plot = FALSE)
```

```
## X-squared
## 0.0818877
```

```
race_low_test$p.value
```

```
## [1] 0.0818877
```

Commentary: Remember that if we get the P-value from the `pdist` command, we need to subtract from 1 because we are shading the right tail of the distribution.

Conclusion

State the statistical conclusion.

We fail to reject the null hypothesis.

State (but do not overstate) a contextually meaningful conclusion.

There is insufficient evidence that race and low birth weight are associated.

Identify the possibility of either a Type I or Type II error and state what making such an error means in the context of the hypotheses.

It's possible that we have made a Type II error. It may be that race and low birth weight are associated, but our sample has not given enough evidence of such an association.

Confidence interval

There are no parameters of interest in a chi-square test, so there is no confidence interval to report.

Post hoc analysis

Had we rejected the null, we would look at the residuals to determine which cells were contributing the most to the chi-square statistic. But since we didn't, there's not much to say about residuals.

```
race_low_test$residuals
```

```
##              race_low$low
## race_low$race        Yes         No
##         White -1.2728970  0.8575266
##         Black  1.0121687 -0.6818789
##         Other  0.8931471 -0.6016963
```

If the overall test is not statistically significant, it does not make much sense to try to interpret any of these residuals. They are all too small to distinguish from chance variability.

Your turn

Use the `smoking` data set from the `openintro` package. Run a chi-square test for independence to determine if smoking status is associated with marital status. If you reject the null, run a post hoc analysis and comment on the cells that seem to be contributing the most to the discrepancy between observed and expected counts.