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.

Instructions

Presumably, you have already created a new project and downloaded this file into it. From the `Run` menu above, select `Run All` to run all existing code chunks.

When prompted to complete an exercise or demonstrate skills, you will see the following lines in the document:

ANSWER

These lines demarcate the region of the R Markdown document in which you are to show your work.

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
```

Be sure to remove the line `# Add code here` when you have added your own code. You should run each new code chunk you create by clicking on the dark green arrow in the upper-right corner of the code chunk.

Sometimes you will be asked to type up your thoughts. That will appear in the document with the words, "Please write up your answer here." 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. You may need to use inline R code in these sections.

When you are finished with the assignment, knit to PDF and proofread the PDF file **carefully**. Do not download the PDF file from the PDF viewer; rather, you should export the PDF file to your computer by selecting the check box next to the PDF file in the Files pane, clicking the `More` menu, and then clicking `Export`. Submit your assignment according to your professor's instructions.

Load Packages

We load the standard `mosaic` package. We also use the `MASS` package for the `birthwt` data, and the `openintro` package for the `smoking` data. The `broom` package will give us tidy output.

```
library(MASS)
library(openintro)
library(broom)
library(mosaic)
```

Research question

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

Recall from previous modules (or from using `str`) that both `low` and `race` are not coded as factor variables. We rectify that here:

```
low <- factor(birthwt$low, levels = c(1, 0), labels = c("Yes", "No"))
race <- factor(birthwt$race, level = c(1, 2, 3),
               labels = c("White", "Black", "Other"))
low_race <- data.frame(low, race)
head(low_race)
```

```
##   low  race
## 1  No Black
## 2  No Other
## 3  No White
## 4  No White
## 5  No White
## 6  No Other
```

First, let's create a contingency table for the relationship between `low` and `race` and assign it to a variable for future use.

Very important: You must *not* include `margins = TRUE` in the `tally` command before running a chi-square test. R is not quite smart enough to figure out that the "Total" isn't really part of the data.

```
low_race_tally <- tally(low ~ race, data = low_race)
low_race_tally
```

```
##      race
## low   White Black Other
##   Yes    23    11    25
##   No     73    15    42
```

Now we are going to run a chi-square test. Just like the goodness-of-fit test, we run the command on the table (`low_race_tally`) and not the raw data.

```
low_race_test <- chisq.test(low_race_tally)
```

With the test run, we can access the expected counts as follows:

```
low_race_test$expected
```

```
##      race
## low       White     Black    Other
##   Yes 29.96825  8.116402 20.91534
##   No  66.03175 17.883598 46.08466
```

To see how these expected counts are computed, look at the sum of all the low birth weight babies $(23 + 11 + 25 = 59)$ and normal weight babies $(73 + 15 + 42 = 130)$. In other words, 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 0.31217 or 31.217%, and 130 of 189 is 0.68783 or 68.783%.

Now, if low birth weight and race 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 31.217% of them to have low birth weight babies and 68.783% of them to have normal weight babies. 31.217% of 96 is 29.968. Look in the table above and note that this is the expected cell count for low birth weight babies among white women. Also, 68.783% of 96 is 66.032. Again, this is listed in the table as the expected count for the cell corresponding to normal weight babies among white mothers. The same analysis can be done for the next two columns 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 R is the row total, C is the column total, and n is the grand total (the sample size), the expected count is simply

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

This is equivalent to the explanation in the previous paragraph. Using low birth weight babies among white mothers as an example, R/n is 59/189 which is 0.31217. Then we multiply this by the column total $C = 96$ to get

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

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 2 rows ("Yes", "No") and 3 columns ("White", "Black", "Other"); therefore,

$$df = (2-1)(3-1) = 1 \times 2 = 2$$

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

Let's run through the rubric in its entirety.

Exploratory data analysis

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

[Type `?birthwt` at the Console to read the help file and use `View` to look at the spreadsheet view of the data.]

```
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 ...
```

```
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

Prepare the data for analysis.

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

```
##   low  race
## 1  No Black
## 2  No Other
## 3  No White
## 4  No White
```

```
## 5  No White
## 6  No Other
```

Make tables or plots to explore the data visually.

```
low_race_tally <- tally(low ~ race, data = low_race)
low_race_tally
```

```
##      race
## low   White Black Other
##   Yes    23    11    25
##   No     73    15    42
```

Commentary: Again, be sure to save the results of the `tally` command to feed into the `chisq.test` command later. Also be sure *not* to use `margins = TRUE`.

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 : Low birth weight and race are independent.

H_A : Low birth weight and race are associated.

Express the null and alternative hypotheses in symbols (when possible).

For a chi-square test for independence, this section is not applicable. With multiple categories in the response and explanatory 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
  - We'll have to run the hypothesis test a little early here in order to check the expected counts:

```
low_race_test <- chisq.test(low_race_tally)
low_race_test$expected
```

```
##      race
## low      White     Black    Other
##   Yes 29.96825  8.116402 20.91534
##   No  66.03175 17.883598 46.08466
```

All expected cell counts are larger than 5, so the condition is met.

Commentary: The `chisq.test` function is not as well supported by the `mosaic` package, which means that we can't use the "tilde" notation as we've done in the past. In other words, the command

```
low_race_test <- chisq.test(low ~ race, data = low_race)
```

will not work.

Mechanics

Compute and report the test statistic.

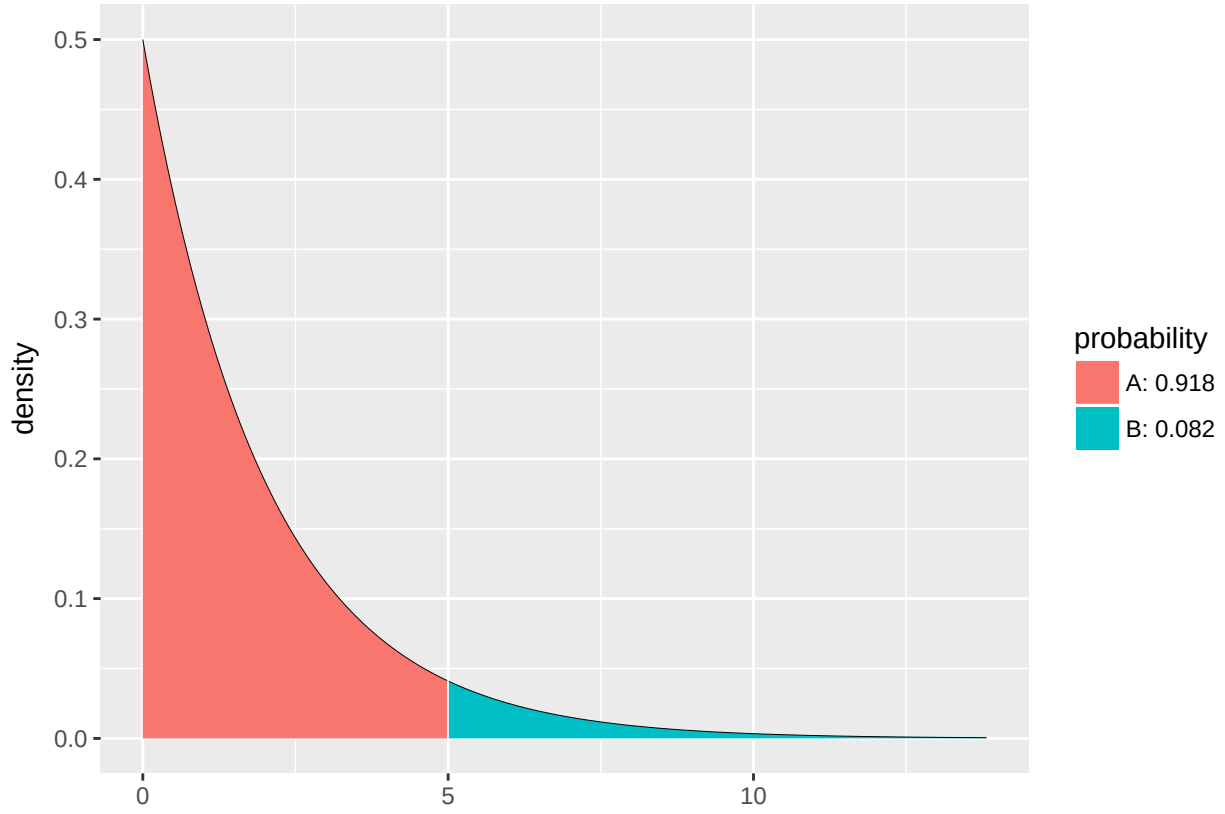
```
low_race_test_tidy <- tidy(low_race_test)
low_race_test_tidy$statistic
```

```
## [1] 5.004813
```

The value of χ^2 is 5.004813.

Plot the null distribution.

```
pdist("chisq", df = low_race_test_tidy$parameter,
      q = low_race_test_tidy$statistic,
      invisible = TRUE)
```



Commentary: We use `pdist`, but now we need to use "chisq" instead of "norm". Also, since the chi-square distribution requires the specification of degrees of freedom, there is a new argument to `pdist` called `df`. We could type `df = 2` since we know there are 2 degrees of freedom; however, the degrees of freedom are also stored in the output `low_race_test_tidy` in the `parameter` variable.

Calculate and report the P-value.

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

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

The P-value is 0.0818877.

Commentary: Values that are as extreme or even more extreme than the test statistic are in the right tail. If we use `pdist`, remember that it always shades to the left by default, so we have to subtract the output from 1 to get the correct P-value. Also remember to add `plot = FALSE` as we don't really need to look at the same picture again.

The same P-value is also stored in the tidy output:

```
low_race_test_tidy$p.value
```

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

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 low birth weight and race 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 low birth weight and race 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. The residuals are not stored as part of the `tidy` output, so we'll have to grab them from the original `low_race_test` object:

```
low_race_test$residuals
```

```
##      race
## low        White      Black      Other
##   Yes -1.2728970  1.0121687  0.8931471
##   No   0.8575266 -0.6818789 -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.

Inference using a contingency table

In the previous example, we had access to the actual data frame. In some situations, you are not given the data; rather, all you have is a contingency table of the data. This certainly happens with homework problems from a textbook, but it can happen in "real life" too. If you're reading a research article, you will rarely have access to the original data used in the analysis. All you can see is what the researchers report in their paper.

Suppose all we know is the contingency table of low birth weight and race. Since the `chisq.test` command requires a table as input, we'll have to manually input the numbers into a table. This requires two steps:

1. Use the `rbind` ("row bind") command to enter each row of values. Each row requires the `c` notation to input them as vectors.
2. Use the `as.table` command to convert the resulting object to a table.

Observe:

```
low_race_table <- as.table(rbind(c(23, 11, 25),
                                 c(73, 15, 42)))
low_race_table
```

```
##    A  B  C
## A 23 11 25
## B 73 15 42
```

There is a way to change the row and column names to something more informative than "A", "B", and "C", but it's not important. The goal is to create a quick-and-dirty table just for purposes of getting `chisq.test` to work.

Now we use `chisq_test` as before.

```
low_race_test_manual <- chisq.test(low_race_table)
low_race_test_manual_tidy <- tidy(low_race_test_manual)
low_race_test_manual_tidy
```

```
##   statistic   p.value parameter                    method
## 1  5.004813 0.0818877         2 Pearson's Chi-squared test
```

Once this is done (in the step "Compute and report the test statistic"), all remaining steps of the rubric stay exactly the same except that you'll use `low_race_test_manual_tidy` instead of `low_race_test_manual_tidy`.

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.

The rubric outline is reproduced below. You may refer to the worked example above and modify it accordingly. Remember to strip out all the commentary. That is just exposition for your benefit in understanding the steps, but is not meant to form part of the formal inference process.

Another word of warning: the copy/paste process is not a substitute for your brain. You will often need to modify more than just the names of the data frames and variables to adapt the worked examples to your own work. Do not blindly copy and paste code without understanding what it does. And you should **never** copy and paste text. All the sentences and paragraphs you write are expressions of your own analysis. They must reflect your own understanding of the inferential process.

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.

Exploratory data analysis

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

ANSWER

```
# Add code here to understand the data.
```

Prepare the data for analysis. [Not always necessary.]

ANSWER

```
# Add code here to prepare the data for analysis.
```

Make tables or plots to explore the data visually.

ANSWER

```
# Add code here to make tables or plots.
```

Hypotheses

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

ANSWER

Please write up your answer here.

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

ANSWER

H_0 : Null hypothesis goes here.

H_A : Alternative hypothesis goes here.

Express the null and alternative hypotheses in symbols (when possible).

ANSWER

$H_0 : math$

$H_A : math$

Model

Identify the sampling distribution model.

ANSWER

Please write up your answer here.

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

ANSWER

Please write up your answer here. (Some conditions may require R code as well.)

Mechanics

Compute and report the test statistic.

ANSWER

```
# Add code here to compute the test statistic.
```

Please write up your answer here.

Plot the null distribution.

ANSWER

```
# Add code here to plot the null distribution.
```

Calculate and report the P-value.

ANSWER

```
# Add code here to calculate the P-value.
```

Please write up your answer here.

Conclusion

State the statistical conclusion.

ANSWER

Please write up your answer here.

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

ANSWER

Please write up your answer here.

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.

ANSWER

Please write up your answer here.

Post-hoc analysis (if null was rejected)

You only need to complete the following section if the null was rejected above.

ANSWER

```
# Add code here to calculate the residuals
```

Please write up your answer here.