Chi-squared tests

[Put your name here]

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

A chi-square goodness-of-fit test is similar to a test for a single proportion, except instead of two categories (success/failure), we now try to understand the distribution of proportions among three or more categories.

A chi-square test for independence is for 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.

It is interesting to note that when there are only two categories, one can still run a chi-square test and the results are equivalent to the one-proportion or two-proportion tests (for the goodness-of-fit test and test for independence, respectively).

Instructions

Presumably, you have already created a new project and downloaded this file into it. Please knit the document (probably to HTML as you're working) and work back and forth between this R Markdown file and the knit output to answer the following questions.

When you are finished with the assignment, knit to PDF, export the PDF file to your computer, and then submit to the corresponding assignment in Canvas. Be sure to submit before the deadline!

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]...
```

When you see that in a code chunk, you need to type some R code to complete a task.

Sometimes you will be asked to type up your thoughts. Instructions to do that will be labeled as follows. If you are currently reading this in the knit output, please look back at the R Markdown file to see the following text:

(When you knit the document, you can't see the text from the line above. That's because the crazy notation surrounding that text marks it as a "comment", and therefore it doesn't appear in the output.) 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 (although it may include inline R code when necessary), but rather a free response section where you talk about your analysis and conclusions.

Getting started

Make sure you're in a project

If you're looking at this document, you should have already created a project and uploaded this R Markdown file to that project folder.

Save your file!

The first thing we **always** do is save our file. You'll probably want to save this under a new name. Go to the "File" menu and then "Save As". Once you've saved the file with the new name, from them on it's easier to just hit Ctrl-S (or Cmd-S on a Mac) to keep saving it periodically.

Remember that file names should not have any spaces in them. (In fact, you should avoid other kinds of special characters as well, like periods, commas, number signs, etc. Stick to letters and numerals and you should be just fine.) If you want a multiword file name, I recommend using underscores like this: `this_filename_has_spaces_in_it`.

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 `hsb2` and `smoking` data.

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

Chi-square goodness-of-fit test

Let's look at the `mtcars` data set with the 1974 Motor Trend magazine car data. In particular, let's focus on the `cyl` variable that indicates how many cylinders there are in the engine (with values 4, 6, or 8).

In order to conduct a hypothesis test, we need to propose a null distribution. In other words, what is the "default" distribution to which we are going to compare our data? Sometimes, this default null comes from substantive expert knowledge. (For example, we might compare the 1974 distribution to a known distribution from another year.) Sometimes we're interested to see if our data deviates from a null distribution that predicts an equal number of observations in each category.

For this test, we will see if there are equal numbers of cars with 4, 6, and 8 cylinders. Let's walk through the rubric.

Hypotheses

Identify the sample and a reasonable population of interest.

The sample consists of 32 types of car from 1974. The population is, presumably, all models from 1974. (I don't think we can reasonably infer anything about the number or cylinders in cars today, for example, given data from 1974.)

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

H_0: The numbers of cars in 1974 with 4, 6, and 8 cylinders are the same.

H_A: The numbers of cars in 1974 with 4, 6, and 8 cylinders are different.

Express the null and alternative hypotheses in symbols.

$H_0 : p_4 = p_6 = p_8$

There is no easy way to express the alternate hypothesis in symbols because any deviation in any of the categories can lead to rejection of the null. So the only requirement here is to express the null in symbols.

Model

Identify the correct sampling distribution model.

We will use a chi-square model. The degrees of freedom are determined by the number of categories minus one. In our data, we have three categories (4, 6, or 8 cylinders), so there are 2 degrees of freedom.

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

- Random
  - It is not apparent why Motor Trend chose these 32 automobiles, so it's not clear that this is even representative of cars from 1974. We should be cautious in our conclusions.
- 10%
  - I do not know how many types of cars were manufactured in 1974, so I do not know if this condition is met. Again, we need to be careful. (Also note that the population is not all automobiles manufactured in 1974. It is all *types* of automobile manufactured in 1974. There's a big difference.)
- Expected cell counts
  - This condition says that under the null, we should see at least 5 cars in each category. The easiest way to check this is to run the chi-square test (a bit prematurely) because its output contains the expected cell counts. One quirk of the chi-square goodness of fit test is that it will not work with the raw data. We have to get a table of the cell counts for each category first and feed that into the `chisq.test` command.

```
cyl <- factor(mtcars$cyl)
cyl_table <- table(cyl)
cyl_table
```

```
## cyl
##  4  6  8
## 11  7 14
```

```
test_gf <- chisq.test(cyl_table)
test_gf$expected
```

```
##        4        6        8
## 10.66667 10.66667 10.66667
```

It's easy to see how these numbers came about. Since there are three categories, and the null states that they should all be equally represented, we expect $32/3 = 10.6666667$ cars in each cell of the table.

Mechanics

Compute the test statistic.

The test statistic is called chi-square and it is calculated as follows:

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

where O is the observed count and E is the expected count in each cell, with the sum taken over all the cells.

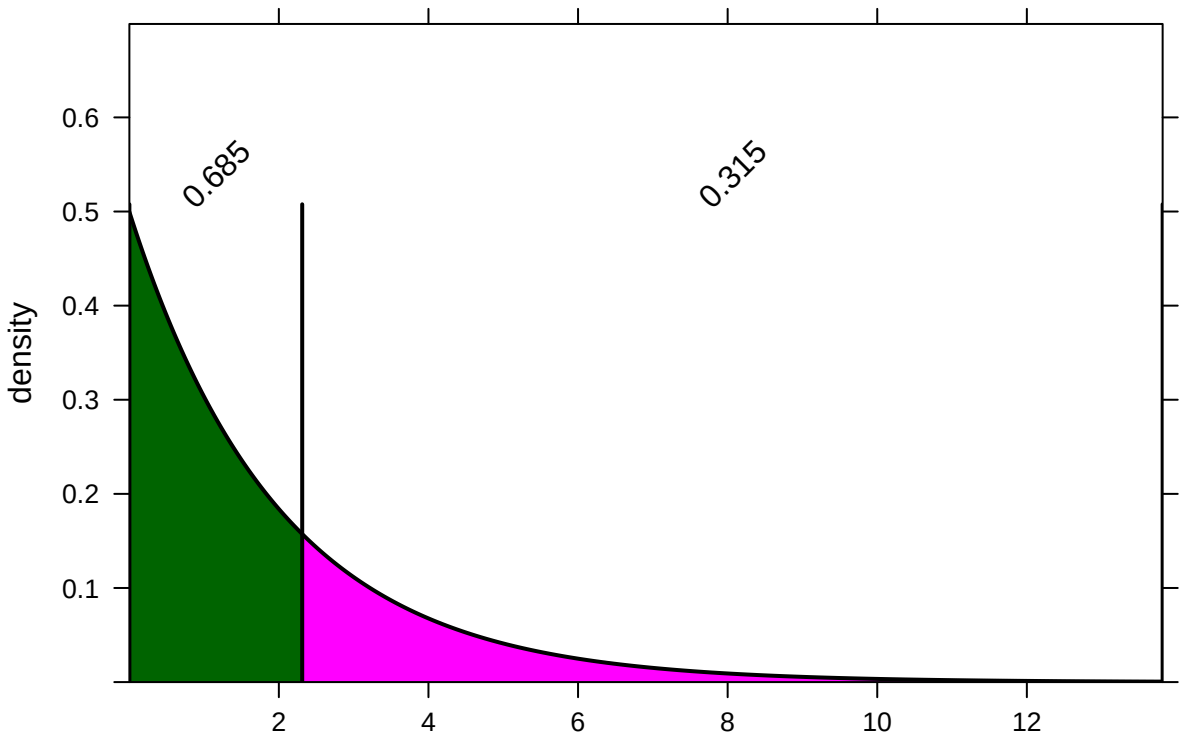
```
test_gf$statistic
```

```
## X-squared
##    2.3125
```

Plot the null distribution.

There are infinitely many chi-square distributions, one for each possible value of degrees of freedom. Here we plot the one with df = 2. All chi-square distributions are postive and skewed to the right. Our P-values will always be shaded to the right. (In other words, all chi-square tests are one-sided.)

```
1 - pdist(dist = "chisq", df = test_gf$parameter, q = test_gf$statistic)
```




```
## X-squared
##  0.314664
```

Calculate the P-value.

```
test_gf$p.value
```

```
## [1] 0.314664
```

Conclusion

State the statistical conclusion.

We fail to reject the null hypothesis.

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

We have insufficient evidence that there is a difference in the number of 4, 6, or 8 cylinder cars among 1974 automobiles.

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.

There is the possibility of making a Type II error here. If this were the case, then there would be a difference in the models with 4, 6, or 8 cylinders, but our sample didn't show enough of a difference to be statistically significant.

Confidence interval

There is no confidence interval for a chi-square test. Since our test is not about measuring some parameter of interest (like p or $p_1 - p_2$), there is no interval to produce.

However we will perform a different kind of analysis...

Post hoc analysis

When we reject the null (which we did not do above), we are left with a very vague alternative hypothesis: there is some difference somewhere in one or more categories. Often, we want to follow up to figure out which categories are the ones deviating from the null expectation.

The best way to do this is to look at the *residuals*. A residual for a cell measures how far away the observed count is from the expected count. It does no good just to calculate $O - E$ however; cells with a large count may be far away from their expected values only because they are large numbers. What we want is some kind of relative distance.

We could use the chi-sq component from each cell; in other words, we could look at

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

It is more traditional in statistics to look at the square root of this quantity:

$$\frac{(O-E)}{\sqrt{E}}.$$

Additionally, the above quantity can be positive or negative, and that gives us more information about the direction in which there is a deviation.

Because we failed to reject the null, we didn't have any evidence of a difference anywhere and so there's not much point in examining the residuals. We'll do it here just for practice.

```
test_gf$residuals
```

```
## cyl
##          4          6          8
##  0.1020621 -1.1226828  1.0206207
```

These numbers don't mean anything in absolute terms; they are only interpretable relative to each other. For example, the first residual is positive, but tiny compared to the others. This means that the observed number of 4-cylinder cars is very close to the expected value. On the other hand, the number of observed 6-cylinder cars is somewhat less than expected, whereas the number of observed 8-cylinder cars is a bit more than expected. If you go back to the table we made earlier, you can verify that.

What if the null is not a uniform distribution?

Suppose we didn't expect an equal number of 4, 6, and 8 cylinder models. How would we run the test under a different null?

Suppose that we expected 35% 4-cylinder cars, 40% 6-cylinder cars, and 25% 8-cylinder cars. We would run the test as follows:

```
test_gf2 <- chisq.test(cyl_table, p = c(0.35, 0.4, 0.25))
```

The numbers defining the null have to add up to 1 (since they are percentages). This causes some trouble, for example, when you have a percentage that has an infinite number of decimal places. For example, what if we expected under the null a distribution of 6/13, 4/13, 3/13? You can't express any of these as a decimal without rounding. Well it turns out that `chisq.test` can handle any set of numbers as long as you set `rescale.p = TRUE`:

```
test_gf3 <- chisq.test(cyl_table, p = c(6, 4, 3), rescale.p = TRUE)
```

Also note that these change the statements of the null distribution (in sentences and in symbols). For example, in the example above where the null is 35% 4-cylinder cars, 40% 6-cylinder cars, and 25% 8-cylinder cars, our null in symbols would be

$H_0 : p_4 = 0.35, p_6 = 0.4, p_8 = 0.25$

Your turn!

Use the `hsb2` data and determine if the propotion of high school students who attend general programs, academic programs, and vocational programs is 25%, 50%, and 25% respectively. 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.

Chi-square test for independence

Suppose we have two categorical variables. For example, let's consider `race` and `low` from the `birthwt` data set. The `race` variable codes the mother's race and `low` is an indicator of a birth weight less than 2.5 kg. The natural question is if these two variables are associated or independent. In other words, are mothers from certain races more or less likely to have low birth weight babies?

We'll need to recode these two variables as factor variables with meaningful labels.

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

Now we can look a contingency table. We'll add some arguments to the `CrossTable` function that we haven't seen before that are related to the chi-square test. We are thinking about `race` as the explanatory variable and `low` as the response.

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

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |              Expected N |
## | Chi-square contribution |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  189
##
##
##              | low
##         race |        No |       Yes | Row Total |
## -------------|-----------|-----------|-----------|
##        White |        73 |        23 |        96 |
##              |    66.032 |    29.968 |           |
##              |     0.735 |     1.620 |           |
##              |     0.760 |     0.240 |     0.508 |
## -------------|-----------|-----------|-----------|
##        Black |        15 |        11 |        26 |
##              |    17.884 |     8.116 |           |
##              |     0.465 |     1.024 |           |
##              |     0.577 |     0.423 |     0.138 |
## -------------|-----------|-----------|-----------|
##        Other |        42 |        25 |        67 |
##              |    46.085 |    20.915 |           |
##              |     0.362 |     0.798 |           |
##              |     0.627 |     0.373 |     0.354 |
## -------------|-----------|-----------|-----------|
## Column Total |       130 |        59 |       189 |
## -------------|-----------|-----------|-----------|
```

```
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  5.004813     d.f. =  2     p =  0.0818877
##
##
##
```

Let's run through the rubric.

Hypotheses

Identify the sample and a reasonable population of interest.

The sample consists of 189 women who gave birth at Baystate Medical Center in Massachusetts during 1986. A reasonable population might be all women who gave birth in that hospital around that time.

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

H_0: The null hypothesis states that race and low birth weight are independent.

H_A: The alternative hypothesis states that 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 correct sampling distribution model.

We will use a chi-square model. If the number of rows in our contingency table (in other words, the number of categories of the explanatory variable) is called r and the number of columns in our contingency table (the number of categories of the response variable) is called c, then the degrees of freedom are calculated as follows:

$$df = (r-1)(c-1).$$

Here, there are three rows and two columns, so our chi-square model has $(3-1)(2-1) = 2$ degrees of freedom.

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

- Random
  - We hope that these women are representative of all women who gave birth in this hospital around 1986.

- 10%
  - We don't know how many women gave birth at this hospital, but perhaps over several years we might have 1890 women.
- Expected cell counts
  - There are two ways we can get expected counts. One is from the contingency table output of `CrossTable` above. By adding the argument `expected = TRUE`, we can see the expected cell counts and verify that they are all greater than 5.

The other place to find expected cell counts is in the output of the `chisq.test` command. As we have to run that command anyway, we might as well run it now.

```
test_ind <- chisq.test(race, low)
```

The expected counts are stored in `test_ind$expected`:

```
test_ind$expected
```

```
##        low
## race          No       Yes
##   White 66.03175 29.968254
##   Black 17.88360  8.116402
##   Other 46.08466 20.915344
```

You can compare these to the ones from the `CrossTable` command above. They are the same.

Mechanics

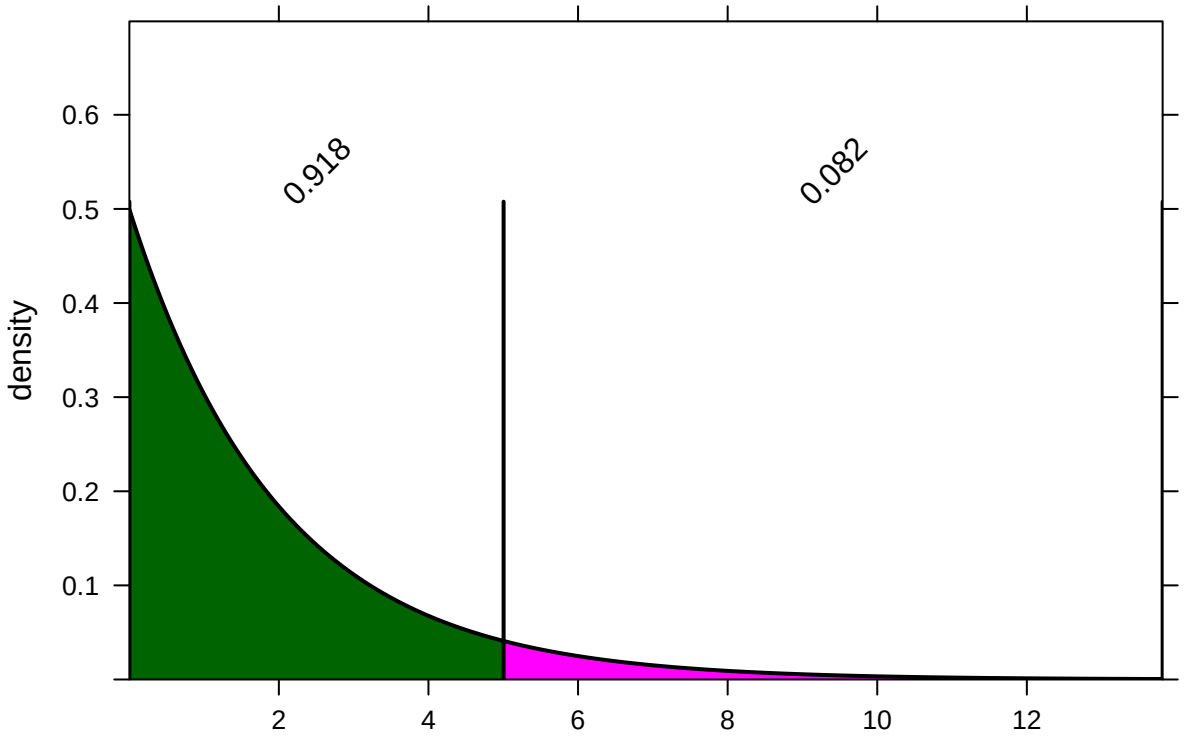
Compute the test statistic.

```
test_ind$statistic
```

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

Plot the null distribution.

```
1 - pdist(dist = "chisq", df = test_ind$parameter, q = test_ind$statistic)
```




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

Calculate the P-value.

```
test_ind$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.

We have insufficient evidence to suggest 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.

There is a possibility that we have made a Type II error, which would be if there really was an association between race and low birth weight, but our sample did not show enough of a difference to be statistically significant.

Confidence interval

As in the goodness-of-fit test, there is no single parameter to estimate with an interval, so this section is irrelevant.

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.

Your turn!

Use the `smoking` data set. Run a chi-square test for independence to determine if smoking status varies by 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.