Inference for Categorical Data

[Put your name here]

In this assignment we will learn how to run hypothesis tests for categorical data in R.

If we have one categorical variable with two categories (a "success" condition and a "failure" condition), we will run a test for a single proportion. In the presence of another categorical explanatory variable with two categories that we want to compare, we will run a test for the difference between two proportions.

Things change a little when there are three or more categories. In this case we run a chi-square goodness-of-fit test for one categorical variables and a chi-square test of independence for two categorical variables.

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 as well as the `MASS` package. At the same time, we'll also load the `gmodels` package to make nice contingency tables. (The `CrossTable` command will also have a side benefit in that it also runs the mechanics of a chi-square test)

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

Inference for a single proportion

We know that certain types of cancer are more common among females or males. The `Melanoma` data set in the `MASS` package describes 205 patients in Denmark with malignant melanoma. Is there a sex bias among these patients?

First, let's do some exploratory data analysis. We need to be careful because another package has a data set called `melanoma` with a lower-case "m".

```
table(Melanoma$sex)
```

```
##
##   0   1
## 126  79
```

It looks like it will be easier to work with this variable if it is re-coded. The help file says "1 = male, 0 = female".

```
sex <- factor(Melanoma$sex,
              levels = c(1, 0),
              labels = c("male", "female"))
```

The rubric is below with some commentary mixed in.

Hypotheses

Identify the sample and a reasonable population of interest.

The sample consists of 205 patients in Denmark with malignant melanoma. In order for this sample to be representative of the population, we should probably restrict our conclusions to the population of all people in Denmark with malignant melanoma.

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

H_0 : The null hypothesis states that females comprise 50% of all patients in Denmark with malignant melanoma.

H_A : The alternate hypothesis states that females do not comprise 50% of all patients in Denmark with malignant melanoma.

Express the null and alternative hypotheses in symbols.

$H_0 : p = 0$

$H_A : p \neq 0$

Model

Identify the correct sampling distribution model.

We will use a normal model.

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

- Random
  - We have no information about how this sample was obtained. We hope these 205 patients are representative of other Danish patients with malignant melanoma.
- 10%
  - I don't know exactly how many people in Denmark suffer from malignant melanoma, but I imagine over time it's more than 2000.
- Success/Failure

$np = 102.5$

$n(1 - p) = 102.5$

These are both larger than 10.

Mechanics

Compute the test statistic.

We'll run the `binom.test` command.

```
test <- binom.test(sex, p = 0.5, success = "female")
```

Unfortunately, this command does not give the z-score and there doesn't seem to be an easy way to get it short of just computing it manually. The sample proportion $\hat{p}$ is computed in the `binom.test` command and will be stored in `test$estimate`.

```
z <- (test$estimate - 0.5)/sqrt(0.5*(1 - 0.5)/length(sex))
z
```

```
## probability of success
##               3.282622
```

Because `test$estimate` comes with a label called "probability of success", unfortunately, our z-score inherits the name as well, which is super-misleading since the z-score is *not* a probability of success. Let's take care of that now:

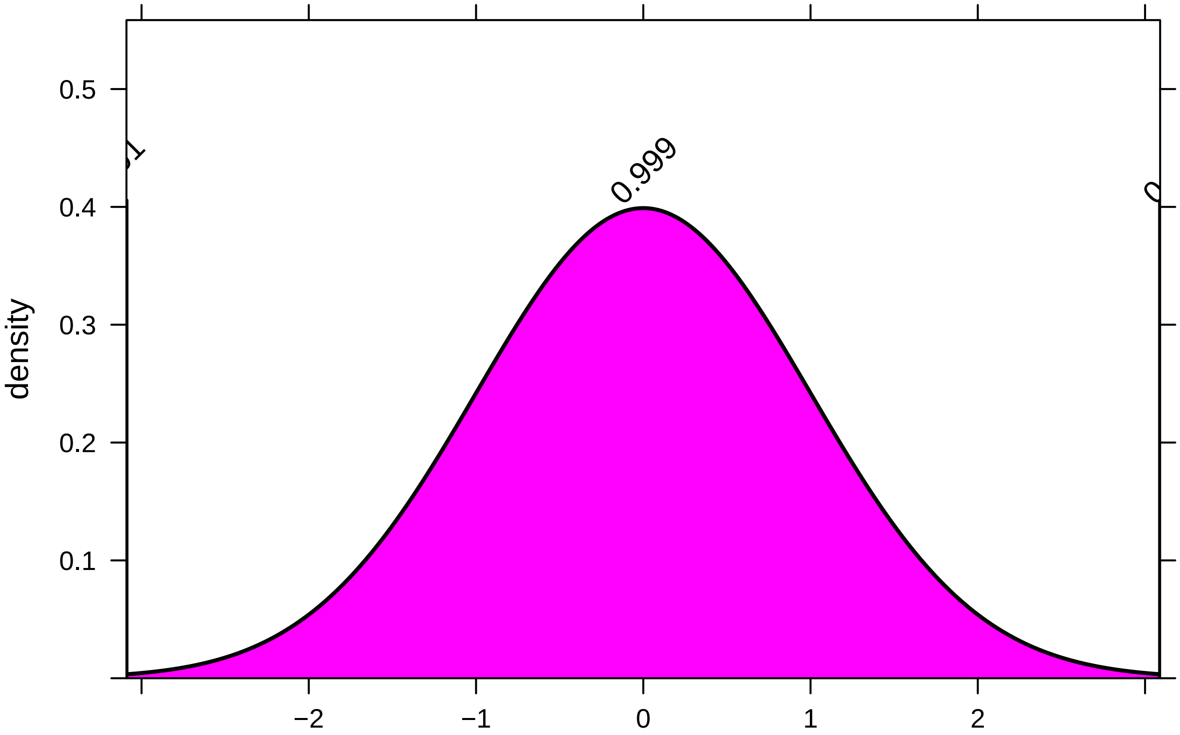
```
z <- unname(z)
z
```

```
## [1] 3.282622
```

That’s better.

Plot the null distribution.

```
pdist(dist = "norm", q = c(-z, z))
```



```
## [1] 0.0005142316 0.9994857684
```

```
# Notice that this is using z which was defined in the previous step.
```

Calculate the P-value.

```
test$p.value
```

```
## [1] 0.001258942
```

Conclusion

State the statistical conclusion.

We reject the null.

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

There is sufficient evidence to suggest that females do not comprise 50% of all patients in Denmark with malignant melanoma.

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.

If we've made a Type I error, then this means that females do comprise 50% of all patients in Denmark with malignant melanoma and we've obtained an unusual sample.

Confidence interval

Conditions

The only condition that changes for a confidence interval is the success/failure condition. Instead of np and $n(1-p)$, we use $n\hat{p}$ and $n(1-\hat{p})$, which are whole numbers that represent the raw number of successes and failures in the data. We can use the `table` command to check this.

```r
table(sex)
```

```
## sex
##   male female
##     79    126
```

Indeed, the counts are each more than 10.

Calculation

This is part of the `binom.test` output as explained in a previous homework assignment. (In that assignment, you also learned how to change the confidence interval if needed.)

```r
test$conf.int
```

```
## [1] 0.5443045 0.6815971
## attr(,"conf.level")
## [1] 0.95
## attr(,"method")
## [1] "Score"
```

(Note that `binom.test` uses a slightly different method for computing a confidence interval than the normal model method we use. Therefore, if you do the computation by hand, you might get slightly different results.)

Conclusion

We are 95% confident that the true proportion of Danish patients suffering from malignant melanoma who are female is captured in the interval (54.4304456%, 68.1597108%).

Now it's your turn!

Go through the rubric to determine if half of the patients in Denmark with malignant melanoma have an ulcer. This time, I'm not going to give you the whole rubric as an outline. Thoughtfully copy and paste from the example above, making the necessary changes as you go along.

Inference for two proportions

Suppose we are interested in learning about the gender distribution among the patients who had died from melanoma at the time the Danish study was published. In this case, we are thinking of `sex` as the explanatory variable and `status` as the response variable.

As `status` is a variable we have not yet considered, let's take a look at it:

```
table(Melanoma$status)
```

```
## 
##   1   2   3 
##  57 134  14
```

You'll note that there are three categories here. However, we are going to take "died from melanoma" as the "success" category and everything else as a failure. We will re-code our data, not just to replace the numbers with words, but also to consolidate three categories into two.

```
status <- factor(Melanoma$status,
                 levels = c(1, 2, 3))
levels(status) <- c("Died from melanoma", "Other", "Other")
table(status)
```

```
## status
## Died from melanoma              Other 
##                 57                148
```

Now let's look at the two variables, `sex` and `status` together.

```
CrossTable(sex, status,
           prop.c = FALSE,  prop.t = FALSE, prop.chisq = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
## 
```

```
## 
## Total Observations in Table:  205 
## 
## 
##              | status 
##          sex | Died from melanoma |              Other |          Row Total | 
## -------------|--------------------|--------------------|--------------------|
##         male |                 29 |                 50 |                 79 | 
##              |              0.367 |              0.633 |              0.385 | 
## -------------|--------------------|--------------------|--------------------|
##       female |                 28 |                 98 |                126 | 
##              |              0.222 |              0.778 |              0.615 | 
## -------------|--------------------|--------------------|--------------------|
## Column Total |                 57 |                148 |                205 | 
## -------------|--------------------|--------------------|--------------------|
## 
## 
```

So the question is, among Danish patients with malignant melanoma, are men or women more like to die from the disease?

It will be helpful for us in the work below to extract all the counts we need and assign them to variables. Fortunately for us, the `CrossTable` command gives us a way to access all the numbers in the contingency table. So first, we'll give the contingency table a name:

```
sex_status <- CrossTable(sex, status,
                         prop.c = FALSE,  prop.t = FALSE, prop.chisq = FALSE)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
## 
## 
## Total Observations in Table:  205 
## 
## 
##              | status 
##          sex | Died from melanoma |              Other |          Row Total | 
## -------------|--------------------|--------------------|--------------------|
##         male |                 29 |                 50 |                 79 | 
##              |              0.367 |              0.633 |              0.385 | 
## -------------|--------------------|--------------------|--------------------|
##       female |                 28 |                 98 |                126 | 
##              |              0.222 |              0.778 |              0.615 | 
## -------------|--------------------|--------------------|--------------------|
## Column Total |                 57 |                148 |                205 | 
## -------------|--------------------|--------------------|--------------------|
## 
## 
```

Then we'll grab the stuff we need:

```r
male_dead <- sex_status$t["male", "Died from melanoma"]
female_dead <- sex_status$t["female", "Died from melanoma"]
male_other <- sex_status$t["male", "Other"]
female_other <- sex_status$t["female", "Other"]
n_M <- male_dead + male_other
n_F <- female_dead + female_other
```

Hypotheses

Identify the sample and a reasonable population of interest.

There are two samples: 79 male patients and 126 female patients in Denmark with malignant melanoma. In order for these samples to be representative of their respective populations, we should probably restrict our conclusions to the population of all males and females in Denmark with malignant melanoma.

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

H_0 : There is no difference between the rate at which men and women in Denmark die from malignant melanoma.

H_A : There is a difference between the rate at which men and women in Denmark die from malignant melanoma.

Express the null and alternative hypotheses in symbols.

$H_0 : p_M - p_F = 0$

$H_A : p_M - p_F \neq 0$

Model

Identify the correct sampling distribution model.

We will use the normal model.

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

- Random
  - We have no information about how this sample was obtained. We hope these 205 patients are representative of other Danish patients with malignant melanoma.
- 10%
  - I don't know exactly how many people in Denmark suffer from malignant melanoma, but I imagine over time it's more than 2000.
- Success/Failure
  - We need to use the pooled proportion of successes to check this. We calculate $\hat{p}_{pooled}$ as shown in class.

```
phat_pooled <- (male_dead + female_dead)/(n_M + n_F)
```

Here are the success/failure computations:

$n_M\hat{p}_{pooled} = 21.9658537$

$n_M(1 - \hat{p}_{pooled}) = 57.0341463$

$n_F\hat{p}_{pooled} = 35.0341463$

$n_F(1 - \hat{p}_{pooled}) = 90.9658537$

Mechanics

Compute the test statistic.

We will use the `prop.test` command for this. Now that we are working with two variables, we can use the "formula" notation with the tilde that we have seen before. The only tricky thing to remember is the order of the variables. Remember that the tilde is pronounced "by", so we want to measure "status by sex" or "status grouped by sex".

```
test2 <- prop.test(status ~ sex)
```

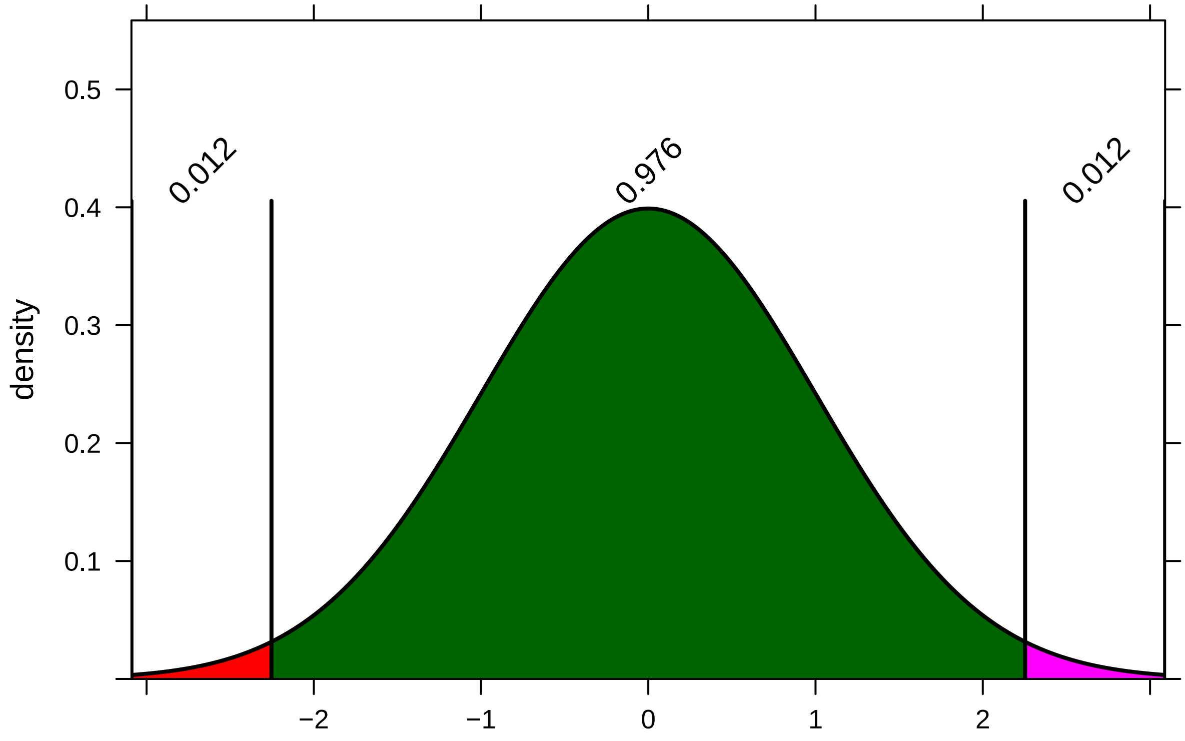
As with the single proportion test, the z-score is not part of the output, so we have to compute it directly. Unfortunately, this is much uglier:

```
z <- (test2$estimate[1] - test2$estimate[2])/
    sqrt(phat_pooled * (1 - phat_pooled)/ n_M +
             phat_pooled * (1 - phat_pooled)/ n_F)
z
```

```
##   prop 1
## 2.253072
```

Plot the null distribution.

```
pdist(dist = "norm", q = c(-z, z))
```



```
##    prop 1    prop 1
## 0.0121273 0.9878727
```

Calculate the P-value.

```
test2$p.value
```

```
## [1] 0.03635633
```

If you're paying close attention, you might notice that the P-value above is a little bigger than what you'd get if you doubled the number from the `pdist` command. That's because under the hood `prop.test` is actually running a more sophisticated test. If you add an extra argument to the `prop.test` command, you'll see the "right" P-value:

```
test2_alt <- prop.test(status ~ sex, correct = FALSE)
test2_alt$p.value
```

```
## [1] 0.0242546
```

Of course, which one is actually the "correct" P-value? You have to wonder when you are required to add an argument that literally says `correct = FALSE`! Well, it turns out that "correct" stands for "Yates's Continuity Correction", which most statisticians these days seem not to recommend. It tends to over-inflate the P-value.

Question: If we use a version of a test that inflates the P-value, which type of error (Type I or Type II) are we more likely to make?

So, the moral of the story here is that we should run our `prop.test` commands with `correct = FALSE`.

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

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

We have sufficient evidence to suggest that there is a difference between the rate at which men and women in Denmark die from malignant melanoma.

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.

If we have made a Type I error, then there would actually be no difference between the rate at which men and women in Denmark die from malignant melanoma, but our samples showed a significant difference.

Confidence interval

Conditions

Only the success/failure condition changes. We now need to check the raw counts. We won't reproduce the contingency table here. Just scroll up and check that for both males and females, the counts of those who died and the others are all greater than 10.

Calculation

As before, we should use the version without the "correction":

```
test2_alt$conf.int
```

```
## [1] 0.01615351 0.27357927
## attr(,"conf.level")
## [1] 0.95
```

Conclusion

We are 95% confident that the true difference between the rate at which men and women die from malignant melanoma is captured in the interval (1.6153506%, 27.3579265%). (This difference is measured by calculating male minus female.)

Note the addition of that last sentence. The rubric specifies that you must indicate the direction of the difference. Without that, we would know that there was a difference, but we would have no idea whether men or women die more from malignant melanoma.

Now it’s your turn!

Go through the rubric to determine if males and females in Denmark who are diagnosed with malignant melanoma suffer from ulcers at different rates.

As before, I’m not going to give you the whole rubric as an outline. Thoughtfully copy and paste from the example above, making the necessary changes as you go along.