Inference for two proportions

Put your name here

Put the date here

Introduction

In this assignment, we revisit the idea of inference for two proportions, but this time using a normal model as the sampling distribution model.

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 as well as the `MASS` package for the `Melanoma` data. The `gmodels` package gives us nice contingency tables. The `broom` package gives us tidy output.

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

Research question

In an earlier module, we used the data set `Melanoma` from the `MASS` package to explore the possibility of a sex bias among patients with melanoma. A related question is whether male or females are more likely to die from melanoma. In this case, we are thinking of `sex` as the explanatory variable and `status` as the response variable.

The sampling distribution model for two proportions

When we simulated using shuffling, it looked like the simulated sampling distribution was roughly normal. Therefore, we should be able to use a normal model in place of simulation when we want to perform statistical inference.

The question is, "Which normal model?" In other words, what is the mean and standard deviation we should use?

Since we have two groups, let's call the true proportion of success p_1 for group 1 and p_2 for group 2. Therefore, the true difference between groups 1 and 2 in the population is $p_1 - p_2$. If we sample repeatedly from groups 1 and 2 and form many sample differences $\hat{p}_1 - \hat{p}_2$, we should expect most of the values $\hat{p}_1 - \hat{p}_2$ to be close to the true difference $p_1 - p_2$. In other words, the sampling distribution is centered at a mean of $p_1 - p_2$.

What about the standard deviation? This is much more technical and complicated. Here is the formula that you'll have to take on faith:

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

So the somewhat complicated normal model is

$$N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

Pooling

When we ran hypothesis tests for one proportion, the true proportion p was assumed to be known, set equal to some null value. Therefore, we could calculate the standard error $\sqrt{\frac{p(1-p)}{n}}$ under the assumption of the null.

We also have a null hypothesis for two proportions. When comparing two groups, the default assumption is that the two groups are the same. This translates into the mathematical statement $p_1 - p_2 = 0$ (i.e., there is no difference between p_1 and p_2).

But there is a problem here. Although we are assuming something about the difference $p_1 - p_2$, we are not assuming anything about the actual values of p_1 and p_2. For example, both groups could be 0.3, or 0.6, or 0.92, or whatever, and the difference between the groups would still be zero.

Without values of p_1 and p_2, we cannot plug anything into the ugly standard error formula above.

There is, however, one assumption we can still salvage. Since we're assuming the two groups are the same, let's compute a single overall success rate for both samples together. In other words, if the two groups aren't different, let's just pool them into one single group and calculate the successes for the whole group.

This is called a *pooled proportion*. It's straightforward to compute: just take the total number of successes in both groups and divide by the total size of both groups. Here is the formula:

$$\hat{p}_{pooled} = \frac{successes_1 + successes_2}{n_1 + n_2}.$$

Occasionally, we are not give the raw number of successes in each group, but rather, the proportion of successes in each group, $\hat{p}_1$ and $\hat{p}_2$. The simple fix is to recompute the raw count of successes as $n_1\hat{p}_1$ and $n_2\hat{p}_2$. Here is what it looks like in the formula:

$$\hat{p}_{pooled} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}.$$

The normal model can still have a mean of $p_1 - p_2$. (We are assuming this is 0 in the null hypothesis.) But its standard error will use the pooled proportion:

$$N\left(p_1 - p_2, \sqrt{\frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_1} + \frac{\hat{p}_{pooled}(1-\hat{p}_{pooled})}{n_2}}\right).$$

Not only do we use the pooled proportion in the standard error, but in fact we use it anywhere we assume the null. For example, the success/failure condition is also subject to the assumption of the null, so we will use the pooled proportion there too.

For a confidence interval, things are different. There is no null hypothesis in effect while computing a confidence interval, so there is no assumption that would justify pooling.

The standard error in the one-proportion interval is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, which just substitutes $\hat{p}$ for p. We can do the same for the standard error in the two-proportion case:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Inference for two proportions

Below is a fully-worked example of inference (hypothesis test and confidence interval) for two proportions. Some of the steps have commentary. When you work your own example, you can thoughtfully copy and paste the R code, making changes as necessary. Do not copy the commentary; that is for your benefit, but not part of the formal inference process. Please do **not** copy and paste any of the full sentence responses. You are responsible for putting everything into your own words.

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 `?Melanoma` at the console to read the help file.]

```
str(Melanoma)
```

```
## 'data.frame':    205 obs. of  7 variables:
##  $ time     : int  10 30 35 99 185 204 210 232 232 279 ...
##  $ status   : int  3 3 2 3 1 1 1 3 1 1 ...
##  $ sex      : int  1 1 1 0 1 1 1 0 1 0 ...
```

```
##  $ age      : int  76 56 41 71 52 28 77 60 49 68 ...
##  $ year     : int  1972 1968 1977 1968 1965 1971 1972 1974 1968 1971 ...
##  $ thickness: num  6.76 0.65 1.34 2.9 12.08 ...
##  $ ulcer    : int  1 0 0 0 1 1 1 1 1 1 ...
```

Prepare the data for analysis.

The two variables of interest are `sex` and `status`. We are considering them as categorical variables, but they are recorded numerically in the data frame. We convert them to proper factor variables and put them in their own data frame using the help file to identify the levels and labels we need.

There is a hitch with `status`. The help file shows three categories: 1. died from melanoma, 2. alive, 3. dead from other causes. For two-proportion inference, it would be better to have two categories only, a success category and a failure category. Since our research question asks about deaths due to melanoma, the "success" condition is the one numbered 1 in the help file, "died from melanoma". That means we need to combine the other two categories into a single failure category. Perhaps we should call it "other". The `factor` command is not quite capable of doing this, so we'll need one more line of code to set it up manually.

```
sex <- factor(Melanoma$sex, levels = c(0, 1), labels = c("female", "male"))
status <- factor(Melanoma$status, levels = c(1, 2, 3))
levels(status) <- c("died from melanoma", "other", "other")
sex_status <- data.frame(sex, status)
```

Make tables or plots to explore the data visually.

As these are two categorical variables, we should look at a contingency table. The variable `sex` is explanatory and `status` is response.

```
CrossTable(sex_status$sex, sex_status$status,
           prop.r = TRUE, prop.c = FALSE,  prop.t = FALSE, prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  205
##
##
##                | sex_status$status
## sex_status$sex | died from melanoma |              other |          Row Total |
## ---------------|--------------------|--------------------|--------------------|
##         female |                 28 |                 98 |                126 |
##                |              0.222 |              0.778 |              0.615 |
## ---------------|--------------------|--------------------|--------------------|
##           male |                 29 |                 50 |                 79 |
##                |              0.367 |              0.633 |              0.385 |
## ---------------|--------------------|--------------------|--------------------|
```

```
##   Column Total |                 57 |                148 |                205 |
## ---------------|--------------------|--------------------|--------------------|
##
##
```

It will be helpful for us to extract all the counts from this table and assign them to variables that can be accessed later. 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_table <- CrossTable(sex_status$sex, sex_status$status,
           prop.r = TRUE, prop.c = FALSE,  prop.t = FALSE, prop.chisq = FALSE)
```

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

Then we'll grab the stuff we need from the `t` variable in `sex_status_table`:

```
female_dead <- sex_status_table$t["female", "died from melanoma"]
male_dead <- sex_status_table$t["male", "died from melanoma"]
female_other <- sex_status_table$t["female", "other"]
male_other <- sex_status_table$t["male", "other"]
n_F <- female_dead + female_other  ## This counts all females
n_M <- male_dead + male_other      ## This counts all males
```

Commentary: You can see why row percentages are necessary in a contingency table. There are 28 females and 29 males who died from melanoma, almost a tie. However, there are more females (126) than there are males (79) who have melanoma in this data set. So the *proportion* of males who died from melanoma is quite a bit larger.

Hypotheses

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

There are two samples: 126 female patients and 79 male 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, although we might be able to make the case that these males and females could be representative of people in other countries who have 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.

$H_0 : p_F - p_M = 0$

$H_A : p_F - p_M \neq 0$

Commentary: The order in which you subtract is irrelevant to the inferential process. However, you should be sure that any future steps respect the order you choose here. A good bet is to look back when you made the factor variables. The first condition listed in the labels of your explanatory variable is going to be the one that gets processed first by the `prop.test` function. In the variable `sex`, we listed "female" first. Therefore, it's safest to subtract $p_F - p_M$.

Model

Identify the sampling distribution model.

We will use a normal model.

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

- Random
  - We have no information about how these samples were obtained. We hope the 126 female patients and 79 male 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 1260 females and 790 males.
- Success/Failure
  - We need to use the pooled proportion of successes to check this. We calculate $\hat{p}_{pooled}$ as described earlier in the module.

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

```
## [1] 0.2780488
```

Here are the success/failure computations:

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

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

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

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

These are all greater than 10.

Commentary: You can check that the value you get for $\hat{p}_{pooled}$ makes sense. It should be somewhere in between the sample proportions for females ($\hat{p}_F = 0.2222222$) and the sample proportion for males ($\hat{p}_M = 0.3670886$). In fact, since there are more females in our sample, the pooled proportion is a little closer to the female value than the male value. It all checks out.

Mechanics

Compute the test statistic.

We will use the `prop.test` command for this, just as we did in the one-proportion case. 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".

```
sex_status_test <- tidy(prop.test(status ~ sex))
```

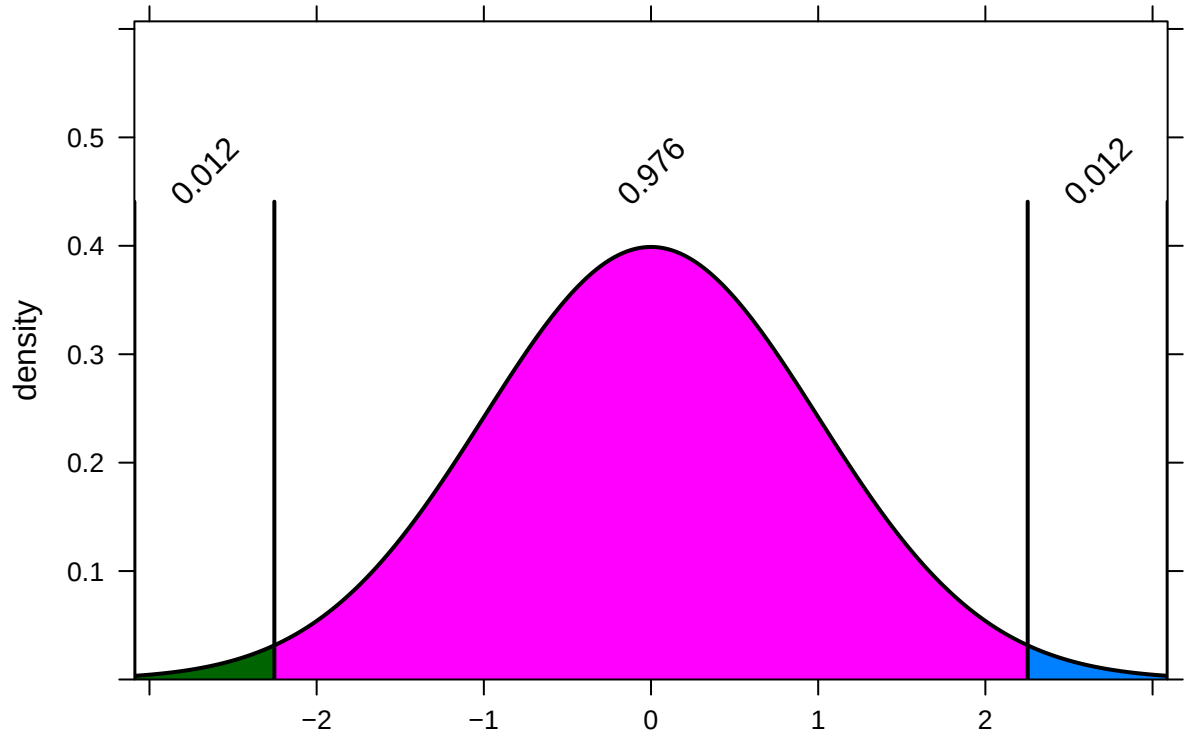
As with the single proportion test, the z-score is not part of the output, so we have to compute it directly. Unfortunately, this is much uglier:

```
z <- (sex_status_test$estimate1 - sex_status_test$estimate2)/
    sqrt(phat_pooled * (1 - phat_pooled)/ n_F +
         phat_pooled * (1 - phat_pooled)/ n_M)
```

The test statistic has a z-score of -2.2530721.

Plot the null distribution.

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




```
## [1] 0.9878727 0.0121273
```

Commentary: Remember that this is a two-sided test.

Calculate the P-value.

```
2 * pdist("norm", q = z,  plot = FALSE)
```

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

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

```
sex_status_test$conf.low
```

```
## [1] -0.2838766
```

```
sex_status_test$conf.high
```

```
## [1] -0.005856138
```

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 (-28.3876633%, -0.5856138%). (This difference is measured by calculating female minus male.)

Note the addition of that last sentence. If you are looking at a confidence interval for a difference, 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. Once we know that we are subtracting female minus male, then given the values are negative, we can infer that males die from malignant melanoma more often than females—at least according to this confidence interval.

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.

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.