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

```
library(MASS)
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 `status` as the response variable and `sex` as the explanatory 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

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

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. One easy "cheat" is to just use the sample values $\hat{p}_1$ and $\hat{p}_2$:

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

There is a more sophisticated way to address this called "pooling". This more advanced concept is covered in an optional appendix to this module.

Inference for two proportions

Below is a fully-worked example of inference (hypothesis test and confidence interval) for two proportions. When you work your own example, you can thoughtfully copy and paste the R code, making changes as necessary.

The example below will pause frequently for commentary on the steps, especially where their execution will be different from what you've seen before when you used simulation. When it's your turn to work through another example on your own, you should follow the outline of the rubric, but you should **not** copy and paste the commentary that accompanies it.

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

```
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 ...
```

```
head(Melanoma)
```

```
##   time status sex age year thickness ulcer
## 1   10      3   1  76 1972      6.76     1
## 2   30      3   1  56 1968      0.65     0
## 3   35      2   1  41 1977      1.34     0
## 4   99      3   0  71 1968      2.90     0
## 5  185      1   1  52 1965     12.08     1
## 6  204      1   1  28 1971      4.84     1
```

Prepare the data for analysis.

The two variables of interest are `status` and `sex`. 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.

CAUTION: If you are copying and pasting from this example to use for another research question, some of the following code chuck is specific to this research question and not applicable in other contexts.

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

```
##                status    sex
## 1              other   male
## 2              other   male
## 3              other   male
## 4              other female
## 5 died from melanoma   male
## 6 died from melanoma   male
```

Make tables or plots to explore the data visually.

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

```r
tally(status ~ sex, data = status_sex, margins = TRUE)
```

```
##                       sex
## status               female male
##   died from melanoma     28   29
##   other                  98   50
##   Total                 126   79
```

```r
tally(status ~ sex, data = status_sex, margins = TRUE, format = "percent")
```

```
##                       sex
## status                  female      male
##   died from melanoma  22.22222  36.70886
##   other               77.77778  63.29114
##   Total              100.00000 100.00000
```

Commentary: You can see why column 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 females and males in Denmark with malignant melanoma, although we might be able to make the case that these females and males 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 women and men in Denmark die from malignant melanoma.

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

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

$H_0 : p_F - p_M = 0$

$H_A : p_F - p_M \neq 0$

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 to 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%
  - We don't know exactly how many people in Denmark suffer from malignant melanoma, but we could imagine over time it's more than 1260 females and 790 males.
- Success/Failure
  - Checking the contingency table above (the one with counts), we see the numbers 28 and 98 (the successes and failures among females), and 29 and 50 (the successes and failures among males). These are all larger than 10.

Commentary: Ideally, for the success/failure condition we would like to check $n_1 p_1$, $n_1(1 - p_1)$, $n_2 p_2$, and $n_2(1 - p_2)$; however, the null makes no claim about the values of p_1 and p_2. We do the next best thing and estimate these by substituting the sample proportions $\hat{p}_1$ and $\hat{p}_2$. But $n_1\hat{p}_1$ and $n_2\hat{p}_2$ are just the raw counts of successes in each group. Likewise, $n_1(1 - \hat{p}_1)$ and $n_2(1 - \hat{p}_2)$ are just the raw counts of failures in each group. That's why we can just read them off the contingency table.

For a more sophisticated approach, one could also use "pooled proportions". See the optional appendix for more information.

Mechanics

Compute and report the test statistic.

```
status_sex_test <- prop.test(status ~ sex, data = status_sex)
status_sex_test_tidy <- tidy(status_sex_test)
status_sex_test_tidy
```

```
##   estimate1 estimate2 statistic    p.value parameter   conf.low
## 1 0.2222222 0.3670886  4.380312 0.03635633         1 -0.2838766
##      conf.high
## 1 -0.005856138
##                                                                  method
## 1 2-sample test for equality of proportions with continuity correction
##   alternative
## 1   two.sided
```

The test statistic is the difference of proportions in the sample, $\hat{p}_1 - \hat{p}_2$:

```
status_sex_test_tidy$estimate1 - status_sex_test_tidy$estimate2
```

```
## [1] -0.1448664
```

We also compute a z-score.

```
SE <- sqrt(status_sex_test_tidy$estimate1 *
               (1 - status_sex_test_tidy$estimate1)/126 +
           status_sex_test_tidy$estimate2 *
               (1 - status_sex_test_tidy$estimate2)/79)
SE
```

```
## [1] 0.06567104
```

```
z <- (status_sex_test_tidy$estimate1 - status_sex_test_tidy$estimate2)/SE
z
```

```
## [1] -2.20594
```

The test statistic has a z-score of -2.2059401.

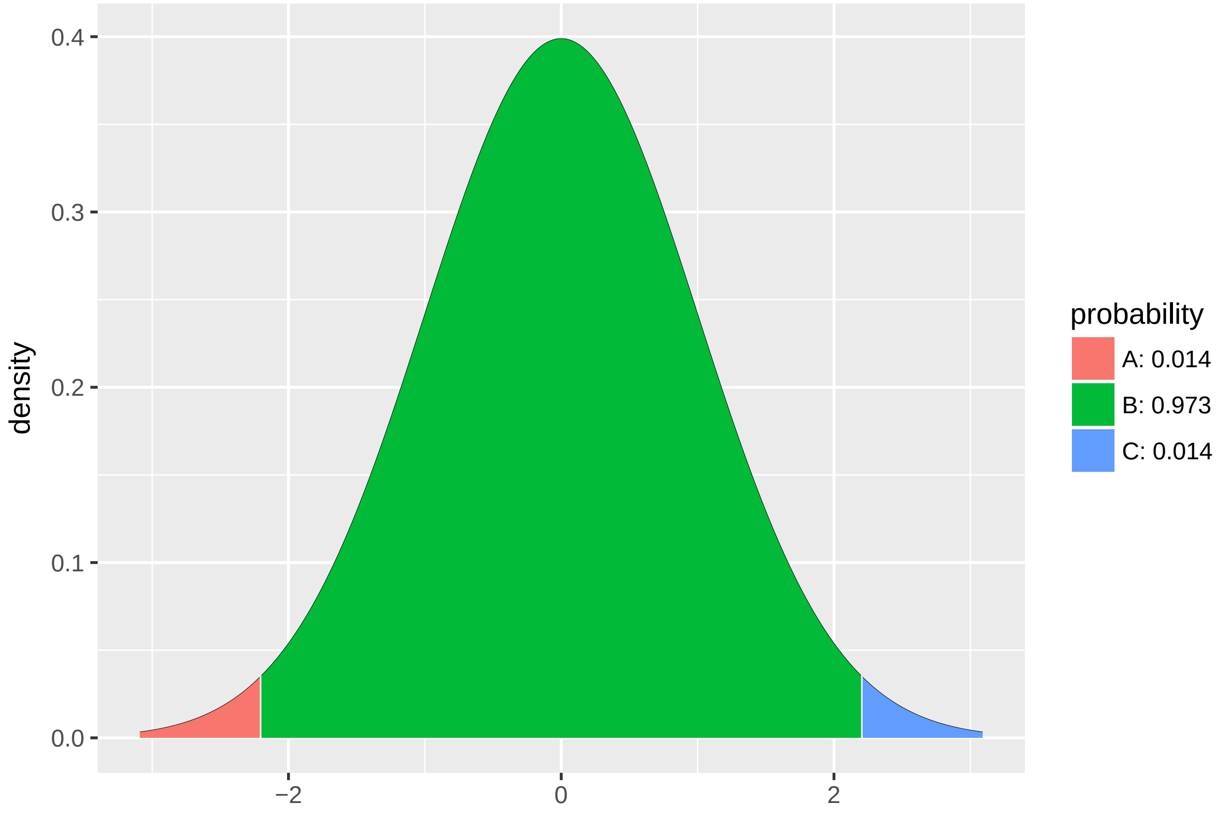
Commentary: We use the `prop.test` command for this. Now that we are working with two variables, we can use the "formula" notation with the tilde that we have seen before. The only tricky thing to remember is the order of the variables. Remember that the tilde is pronounced "by", so we want to measure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

As with the single proportion test, the z-score is not part of the output, so we have to compute it directly.[1] Unfortunately, this is a much uglier formula than it was for a test for one proportion.

Plot the null distribution.

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

[1] Ignore the test statistic from the tidy output. Under the hood, the prop.test command is doing something quite different, so this test statistic doesn't make sense in the context of a normal model.



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

Calculate and report the P-value.

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

```
## [1] 0.0273882
```

The P-value is 0.0273882.

Commentary: As in the one-proportion test, a two-sided P-value is stored in the output of the `prop.test` function:

```
status_sex_test$p.value
```

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

Because the `prop.test` function is using a slightly different method under the hood, this P-value will not agree exactly with the one we computed. Nevertheless, they should be somewhat close and lead to the same conclusion.

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 women and men 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 women and men in Denmark die from malignant melanoma, but our samples showed a significant difference.

Confidence interval

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

None of the conditions have changed, so they don't need to be rechecked.

Calculate the confidence interval.

```
status_sex_test_tidy$conf.low
```

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

```
status_sex_test_tidy$conf.high
```

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

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

We are 95% confident that the true difference between the rate at which women and men die from malignant melanoma is captured in the interval (-28.3876633%, -0.5856138%). (This difference is measured by calculating female minus male.)

Commentary: 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 women or men 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.

Inference using summary statistics

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 are summary statistics about 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. Depending on what kind of information you have, there are a couple of different ways of handling inference.

Method 1

You may just have a summary of the total number of successes and failures. In our melanoma example, among the females, 28 died from melanoma and 98 died from other causes, and among the males, 29 died from melanoma and 50 died from other causes. If that's all we know, we can run the `prop.test` command as follows:

```
status_sex_test_count <- prop.test(c(28, 29), n = c(126, 79))
status_sex_test_count_tidy <- tidy(status_sex_test_count)
status_sex_test_count_tidy
```

```
##   estimate1 estimate2 statistic    p.value parameter   conf.low
## 1 0.2222222 0.3670886  4.380312 0.03635633         1 -0.2838766
##      conf.high
## 1 -0.005856138
##                                                                  method
## 1 2-sample test for equality of proportions with continuity correction
##   alternative
## 1   two.sided
```

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

Method 2

If you are given the percentages of successes and/or failures in your data, you'll have to convert them to whole number totals. You might be told that of the 126 females, 22.2% died from melanoma, and of the 79 males, 36.7% died from melanoma. In that case, we can run the `prop.test` command as follows:

```
status_sex_test_prop <- prop.test(round(c(126*0.222, 79*0.367)),
                                       n = c(126, 79))
status_sex_test_prop_tidy <- tidy(status_sex_test_prop)
status_sex_test_prop_tidy
```

```
##   estimate1 estimate2 statistic    p.value parameter   conf.low
## 1 0.2222222 0.3670886  4.380312 0.03635633         1 -0.2838766
##      conf.high
## 1 -0.005856138
##                                                                  method
## 1 2-sample test for equality of proportions with continuity correction
##   alternative
## 1   two.sided
```

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

Your turn

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

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.

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.

Confidence interval

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.)

Calculate the confidence interval.

ANSWER

```
# Add code here to calculate the confidence interval.
```

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

ANSWER

Please write up your answer here.

Optional appendix: Pooling

Earlier, we mentioned that that we cannot calculate the standard error directly because the null hypothesis does not give us p_1 and p_2. (The null only addresses the value of the difference $p_1 - p_2$.) We dealt with this by simply substituting $\hat{p}_1$ for p_1 and $\hat{p}_2$ for p_2.

There is, however, one assumption from the null we can still salvage that will improve our test. Since the null hypothesis assumes that 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 given 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 usually assume 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 can we use the pooled proportion in the standard error, but in fact we can use it anywhere we assume the null. For example, the success/failure condition is also subject to the assumption of the null, so we could 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 do the same for the standard error in the two-proportion case:

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