Hypothesis testing with simulation, Part 2

Put your name here

Put the date here

Introduction

Now that we have learned about hypothesis testing, we'll explore a different example. Although the rubric for performing the hypothesis test will not change, the individual steps will be implemented in a different way due to the research question we're asking and the type of data used to answer it.

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 `MASS` package to access the `Melanoma` data on patients in Denmark with malignant melanoma, and the `mosaic` package for simulation tools.

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

As explained in an earlier module, we will set the seed so that our results are reproducible.

```
set.seed(42)
```

Our research question

We know that certain types of cancer are more common among females or males. Is there a sex bias among patients with malignant melanoma?

Let's jump into the "Exploratory data analysis" part of the rubric first.

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.

[You can look at the help file by typing `?Melanoma` at the Console. However, do not put that command here in a code chunk. The R Markdown file has no way of displaying a help file when it's processed. The same holds for the `View` command. Be careful: there's another data set called `melanoma` with a lower-case "m". Make sure you are using an uppercase "M".]

Use `str` to examine the structure 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 ...
```

Prepare the data for analysis.

It appears that `sex` is coded an an integer. You will recall that we need to convert it to a factor variable since it is categorical, not numerical. If we have a preconceived idea of which category is considered the "success" category, we should list it first in the `factor` command. It doesn't matter in this case, so since the first number 0 is "female", we will take that to be our "success" and measure the proportion of malignant melanoma patients who are female.

Although it seems silly to create a data frame with only one variable in it, our analysis in R will work better if it's set up like that.

```
sex <- factor(Melanoma$sex, levels = c(0, 1), labels = c("Female", "Male"))
sex_df <- data.frame(sex)
head(sex_df)
```

```
##      sex
## 1   Male
## 2   Male
## 3   Male
## 4 Female
## 5   Male
## 6   Male
```

Make tables or plots to explore the data visually.

We only have one categorical variable, so we only need a frequency table. Since we are concerned with proportions, we'll also look at a relative frequency table.

```
tally(~ sex, data = sex_df, margins = TRUE)
```

```
## sex
## Female   Male  Total
##    126     79    205
```

```
tally(~ sex, data = sex_df, margins = TRUE, format = "percent")
```

```
## sex
##    Female      Male     Total
##  61.46341  38.53659 100.00000
```

The logic of inference and simulation

This is a good place to pause and remember why statistical inference is important. There are certainly more females than males in this data set. So why don't we just show the table above, declare females are more likely to have malignant melanoma, and then go home?

Think back to coin flips. Even though there was a 50% chance of seeing heads, did that mean that exactly half of our flips came up heads? No. We have to acknowledge *sampling variability*: even if the truth were 50%, when we sample, we could accidentally get more or less than 50%, just by pure chance alone. Perhaps these 205 patients just happen to have more females than average.

The key, then, is to figure out if 61.4634146% is *significantly* larger than 50%, or if a number like 61.4634146% (or one even more extreme) could easily come about from random chance.

As we know from Part 1 of this assignment, we can run a formal hypothesis test to find out. As we do so, make note of the things that are the same and the things that have changed from the last hypothesis tests you ran. For example, we are not comparing two groups anymore. We have one group of patients, and all we're doing is measuring the percentage of this group that is female. It's tempting to think that we're comparing males and females, but that's not the case. We are not using `sex` to divide our data into two groups for the purpose of exploring whether some other variable differs between men and women. We just have one sample. "Female" and "Male" are simply categories in a single categorical variable. Also, because we are only asking about one variable (`sex`), the mathematical form of the hypotheses will look a little different.

Because this is no longer a question about two variables being independent or associated, the "shuffling" idea we've been using no longer makes sense. So what does make sense?

It helps to start by figuring out what our null hypothesis is. Remember, our question of interest is whether there is a sex bias in malignant melanoma. In other words, are there more or fewer females than males with malignant melanoma? As this is our research question, it will be the alternative hypothesis. So what is the null? What is the "default" situation in which nothing interesting is going on? Well, there would be no sex bias. In other words, there would be the same number of females and males with malignant melanoma. Or another way of saying that—with respect to the "success" condition of being female that we discussed earlier—is that females comprise 50% of all patients with malignant melanoma.

Okay, given our philosophy about the null hypothesis, let's take the skeptical position and assume that, indeed, 50% of all malignant melanoma patients in our population are female. Then let's take a sample of 205 patients. We can't get exactly 50% females from a sample of 205 (that would be 102.5 females!), so what numbers can we get?

Simulation will tell us. What kind of simulation? As we come across each patient in our sample, there is a 50% chance of them being female. So instead of sampling real patients, what if we just flipped a coin? A coin

flip will come up heads just as often as our patients will be female under the assumption of the null.

This brings us full circle, back to the first simulation idea we explored. We can simulate coin flips (using the `mosaic` package in R with the `rflip` function), graph our results, and calculate a P-value. More specifically, we'll flip a coin 205 times to represent sampling 205 patients. Then we'll use the `do` command to do this a bunch of times and establish a range of plausible percentages that can come about by chance from this procedure.

Let's dive back into the remaining steps of the formal hypothesis test.

Hypotheses

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

The sample consists of 205 patients from Denmark with malignant melanoma. Our population is presumably all patients with malignant melanoma, although in checking conditions below, we'll take care to discuss whether patients in Denmark are representative of patients elsewhere.

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

H_0 : Half of malignant melanoma patients are female.

H_A : There is a sex bias among patients with malignant melanoma (meaning that females are either overrepresented or underrepresented).

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

$H_0 : p = 0.5$

$H_A : p \neq 0.5$

Model

Identify the sampling distribution model.

We will simulate the sampling distribution.

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

- Random
  - As mentioned above, these 205 patients are not a random sample of all people with malignant melanoma. We don't even have any evidence that they are a random sample of melanoma patients in Denmark. Without such evidence, we have to hope that these 205 patients are representative of all patients who have malignant melanoma. Unless there's something special about Danes in terms of their genetics or diet or something like that, one could imagine that their physiology makes them just as susceptible to melanoma as anyone else. More specifically, though, our question is about females and males getting malignant melanoma. Perhaps there are more female sunbathers in Denmark than in other countries. That might make Danes unrepresentative in terms of the gender balance among melanoma patients. We should be cautious in interpreting any conclusion we might reach in light of these doubts.
- 10%

– Whether in Denmark or not, given that melanoma is a fairly common form of cancer, I assume 205 is less than 10% of all patients with malignant melanoma.

Mechanics

Compute and report the test statistic.

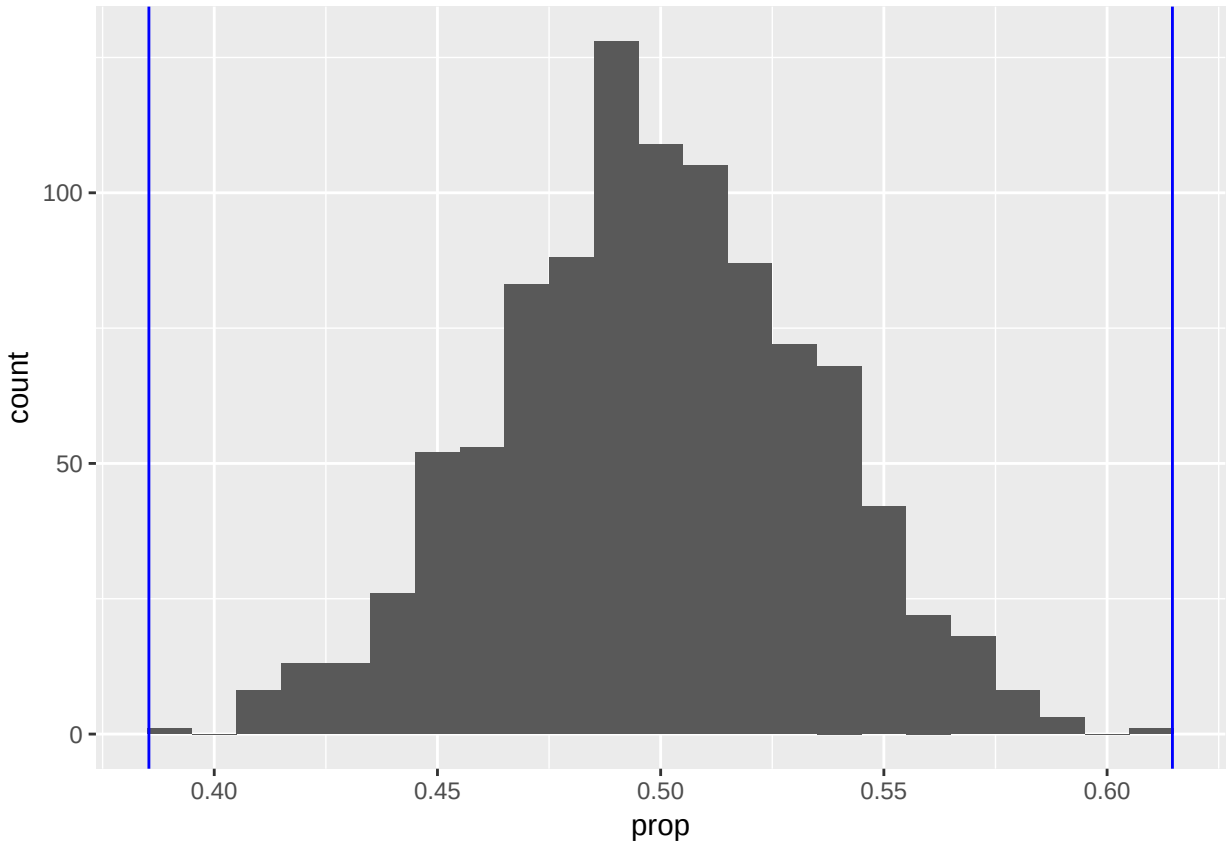
```
female_prop <- prop(sex, data = sex_df)
female_prop
```

```
##    Female
## 0.6146341
```

The observed percentage of females with melanoma in our sample is 61.4634146%.

Plot the null distribution.

```
sims <- do(1000) * rflip(205, prob = 0.5)
# Strictly speaking, you don't need prob = 0.5 because
# that's the default for rflip. However, I'm showing you
# the syntax because your null hypothesis won't always
# be p = 0.5.
ggplot(sims, aes(x = prop)) +
    geom_histogram(binwidth = 0.01) +
    geom_vline(xintercept = female_prop, color = "blue") +
    geom_vline(xintercept = 1 - female_prop, color = "blue")
```



(The last line of the code chunk above requires a little explanation. The proportion of successes is 0.6146341, but we're running a two-sided test. So we also need to plot another blue line at a value that's as extreme as 0.6146341, but on the other side of 50%. That number is 0.3853659. This trick of subtracting from 1 will only work, though, if you're working with a null of 50%. Otherwise, you'll have to calculate the value in the other tail manually.)

Calculate and report the P-value.

```
P <- 2 * prop(sims$prop >= female_prop)
P
```

```
## TRUE
##    0
```

$P < 0.001$.

Three observations here:

1. We need "greater than or equal to (`>=`) here because the simulated values that are more extreme than 0.6146341 are lying *above* 0.6146341, in the right tail of the simulated sampling distribution.
2. We multiplied by 2 because this is a two-sided test.
3. The P-value appears to be zero. Indeed, among the 1000 simulated values, we saw none that exceeded 0.6146341 and none that were less than 0.3853659. However, a true P-value can never be zero. If you did millions or billions of simulations (please don't try!), surely there would be one or two with even more extreme values.

In cases when the P-value is really, really tiny, it is traditional to report $P < 0.001$.

It is **incorrect** to say $P = 0$.

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

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

There is sufficient evidence that there is a sex bias in patients who suffer 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.

As we rejected the null, we run the risk of making a Type I error. If we have made such an error, that would mean that patients with malignant melanoma are equally likely to be male or female, but that we got a sample with an unusual number of female patients.

Your turn

Determine if the percentage of patients in Denmark with malignant melanoma who also have an ulcer is significantly different from 50%.

As before, you have the outline of the rubric for inference below. Some of the steps will be the same or similar to steps in the example above. It is perfectly okay to copy and paste R code, making the necessary changes. It is **not** okay to copy and paste text. You need to put everything into your own words.

The template below is exactly the same as in the file `Rubric_for_inference.pdf` up to the part about confidence intervals which we haven't learned yet.

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.

Conclusion

Now you have seen two fully-worked examples of hypothesis tests using simulation, and you have created two more examples on your own. Hopefully, the logic of inference and the process of running a formal hypothesis test are starting to make sense.

Keep in mind that the outline of steps will not change. However, the way each step is carried out will vary from problem to problem. Not only does the context change (one example involved smoking mothers and the other, melanoma patients), but the statistics you compute also change (one example compared proportions from two samples and the other only had one proportion from a single sample). Pay close attention to the research question and the data that will be used to answer that question. That will be the only information you have to help you know which hypothesis test applies.