Simulations and Hypothesis Tests

[Put your name here]

In this assignment we will learn how to use the `mosaic` package to run simulations to test hypotheses. Then we will incorporate this computational technique into a full hypothesis test. This assignment will also introduce you to the rubric for hypothesis testing that will be used throughout the course.

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 so that we can continue to work with the birth weight data `birthwt`. At the same time, we'll also load the `gmodels` package to make nice contingency tables.

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

One more bit of technical detail. Since there will be some randomness involved here, I will need to include an R command to ensure that we all get the same results every time this document is knit. This is called "setting the seed". Don't worry too much about what this is doing. In particular, the number `1234` in the command below is totally irrelevant. It could have been any number at all. If you chance the number, you will get different answers, but the actual value of the number does not matter.

```
set.seed(1234)
```

Exploratory data analysis

We use the birth weight data from `birthwt` with which we are already familiar. In particular, we'll investigate a possible association between a mother's smoking status during pregnancy and low birth weight of the baby. Rather than using the actual birth weight of the baby (we'll do that in a future assignment), let's use the categorical variable `low` that is simply an indicator (yes/no) of whether the birth weight is less than 2.5 kg.

We can see below that neither `smoke` nor `low` are factor variables as we need them to be:

```
str(birthwt)
```

```
## 'data.frame':    189 obs. of  10 variables:
##  $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
##  $ race : int  2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...
```

We'll fix that by creating new factor variables called `smoke` and `low`.

```
smoke <- factor(birthwt$smoke, levels = c(0, 1), labels = c("No", "Yes"))
low <- factor(birthwt$low, levels = c(0, 1), labels = c("No", "Yes"))
```

Now create a contingency table with `smoke` as the row variable and `low` as the column variable with only counts and row percentages displayed. (In other words, as in the last assignment, use extra options in the `CrossTable` command to get rid of all the extra cruft that appears by default.)

```
## Add code here to create a contingency table with smoke as the
## row variable and low as the column variable.
```

Question: By placing `smoke` as the row variable and `low` as the column variable, and then looking at row percentages, we are implying that one variable is explanatory and one is response. Which variable are we treating as explanatory and which are we treating as response? Do you agree with that choice? Why or why not?

Although we can read off the percentages in the contingency table, we need a command to extract the proportions for use in further calculations. First, the `prop` command. This command uses the "tilde" notation. The variable on the left of the tilde (`low`) is the variable of interest, and we're dividing up the data into groups based on the variable on the right of the tilde (`smoke`). You can remember this by thinking, "Calculate low birth weight **by** smoking status." Note that this is backwards from the order we used in the `CrossTable` command.

```
prop(low ~ smoke)
```

```
##     No.No    No.Yes
## 0.7478261 0.5945946
```

Ignore the confusing column headers in the resulting output. The idea here is that the two percentages listed are the two percentages of interest that we're comparing.

Question: Interpret these percentages in the context of the data. In other words, what do these percentages say about the women who do not smoke during pregnancy versus the women who do?

The real statistic of interest to us, though, is the difference between these percentages.

```
diff(prop(low ~ smoke))
```

```
##     No.Yes
## -0.1532315
```

Let's store this value for future use as well. We can use any name we want, but I've chosen `obs_diff` here for "observed difference".

```
obs_diff <- diff(prop(low ~ smoke))
```

Shuffling

One way to see if there is evidence of an association between smoking and low birth weight is to assume, temporarily, that there is no association. If there is truly no association, then the difference between the non-smoking group and the smoking group should be 0%.

Now, we saw a difference of -15.3231492% between the two groups in the data. Then again, non-zero differences can just come about by pure chance alone. We may have accidentally sampled more smokers who also just happened to have babies with low birth weight, even if there were no association.

So how do test the range of values that could arise from just sampling variability? One way to force the variables to be independent is to use the method described in your textbook of "shuffling" the values of `smoke`.

If instead of measuring whether women actually smoke or not, we just randomly and arbitrarily label them as "smokers" or "non-smokers", we know for sure that such an assignment is random and not due to any actual evidence of smoking. In that case, low birth weight babies are equally likely to occur in both groups.

Let's see how shuffling works in R.

```
head(smoke, 20)
```

```
##  [1] No  No  Yes Yes Yes No  No  No  Yes Yes No  No  No  No  Yes Yes No
## [18] Yes No  Yes
## Levels: No Yes
```

```
head(shuffle(smoke), 20)
```

```
##  [1] No  No  No  No  Yes No  No  No  No  Yes No  Yes No  No  No  Yes Yes
## [18] Yes Yes Yes
## Levels: No Yes
```

Simulation

The idea here is to keep the low birth weight status the same for each woman, but randomly shuffle the smoking labels. There will still be the same number of women who smoke, but now they will be randomly assigned such a designation. In other words, if smoking is truly not associated with low birth weight, then shuffling the women into different smoking statuses should not result in percentages vastly different from the observed percentages calculated earlier.

Again, note that the expected difference under the assumption of independent variables is 0%. In other words, if there is truly no association, then the percentage of women having low birth weight babies should be independent of smoking. However, sampling variability means that we are not likely to see an exact difference of 0%. (In fact, due to the sample sizes in each group, it is impossible to get a difference of exactly 0%.) The real question, then, is how different could the difference be from 0% and still be reasonably possible due to random chance.

Here are a few random simulations. (The randomness is built into the `shuffle` command.)

```
diff(prop(low ~ shuffle(smoke)))
```

```
##      No.Yes
## -0.06439483
```

```
diff(prop(low ~ shuffle(smoke)))
```

```
##      No.Yes
## -0.04218566
```

```
diff(prop(low ~ shuffle(smoke)))
```

```
##      No.Yes
## 0.002232667
```

The `do` command from the `mosaic` package allows us to repeat this simulation process any number of times. The results will be gathered together in a data frame. (Also, the assigned name of the resulting variable is terrible, so we'll rename it.)

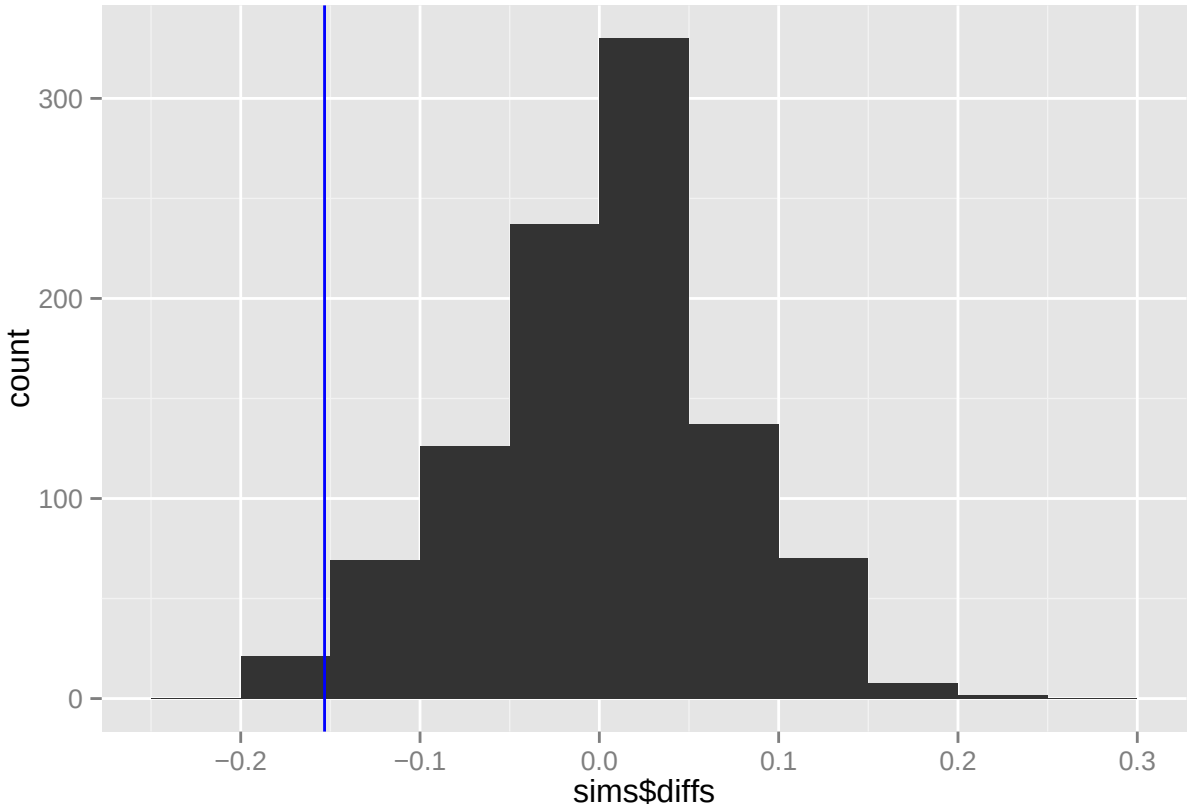
```
sims <- do(1000) * diff(prop(low ~ shuffle(smoke)))
sims <- rename(sims, diffs = No.Yes)
head(sims, 20)
```

```
##           diffs
## 1   0.002232667
## 2   0.024441833
## 3  -0.086603995
## 4  -0.019976498
## 5   0.068860165
## 6   0.068860165
## 7   0.002232667
## 8  -0.064394830
## 9  -0.019976498
## 10  0.068860165
## 11 -0.019976498
## 12 -0.019976498
## 13  0.046650999
## 14  0.002232667
## 15  0.068860165
## 16  0.091069330
## 17  0.024441833
## 18  0.024441833
## 19  0.046650999
## 20 -0.131022327
```

Plot results

A histogram will show us the range of possible values under the assumption of independence of the two variables. On the same plot, we graph a line at the value of the observed difference in proportions to see if that value could have reasonably occurred by chance alone.

```
qplot(sims$diffs, binwidth = 0.05) +
    geom_vline(x = obs_diff, color = "blue")
```



By chance?

How likely is it that the observed difference (or a difference even more extreme) could have resulted from chance alone?

```
prop( ~ sims$diffs <= obs_diff)
```

```
##  TRUE
## 0.021
```

This percentage is small. This shows us that if there were truly no association between low birth weight and smoking, then our data is a rare event. (An observed difference this extreme or more extreme would only occur 2.1% of the time.)

Conclusion

Because the probability above is so small, it seems unlikely that our variables are independent. Therefore, it seems more likely that there is an association between a low birth weight and smoking. We have evidence of a statistically significant difference between the chance of having a baby with low birth weight among women who smoke versus women who don't smoke.

Keep in mind that this data is from an observational study, so we cannot conclude that smoking *causes* low birth weight.

Hypothesis testing

The exposition above can be formalized into what is known as a *hypothesis test*. In this section, we will walk through the rubric posted on Canvas for conducting a full and complete hypothesis test for our smoking and low birth weight example above.

A hypothesis test can be organized into four parts:

1. Hypotheses
2. Model
3. Mechanics
4. Conclusion

Below, I'll model the process of walking through a complete hypothesis test, showing how I would address each step. Then, you'll have a turn at doing the same thing for a different question.

Note that there is some mathematical formatting. This is done by enclosing such math in dollar signs. Don't worry too much about the syntax; just mimic what you see in the example.

Hypotheses

Identify the sample and a reasonable population of interest.

The sample consists of 189 births from Baystate Medical Center, Springfield, Massachusetts during 1986. The population of interest is probably all births, maybe in the U.S., although we are only really safe coming to conclusions about all births at this particular hospital.

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

H_0: The null hypothesis states that there is no association between smoking status during pregnancy and low birth weight.

H_A: The alternative hypothesis states that there is an association between smoking status during pregnancy and low birth weight.

Express the null and alternative hypotheses in symbols.

$H_0 : p_{smoker} - p_{nonsmoker} = 0$

$H_A : p_{smoker} - p_{nonsmoker} \neq 0$

Model

Identify the correct sampling distribution model.

This step is not applicable.

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

- Random
  - We have no evidence that this is a random sample of births from the the Baystate Medical Center. We hope that it is a representative sample. If the population of interest is all births in the U.S., for example, then I have some doubts as to how representative the sample is. (It is possible that the women who go to this hospital may be different from other women in the U.S. Perhaps this hospital serves women from certain backgrounds or socioeconomic statuses.)
- 10%
  - Regardless of the intended population, 189 births is surely less than 10% of all births under consideration.

Mechanics

Compute the test statistic.

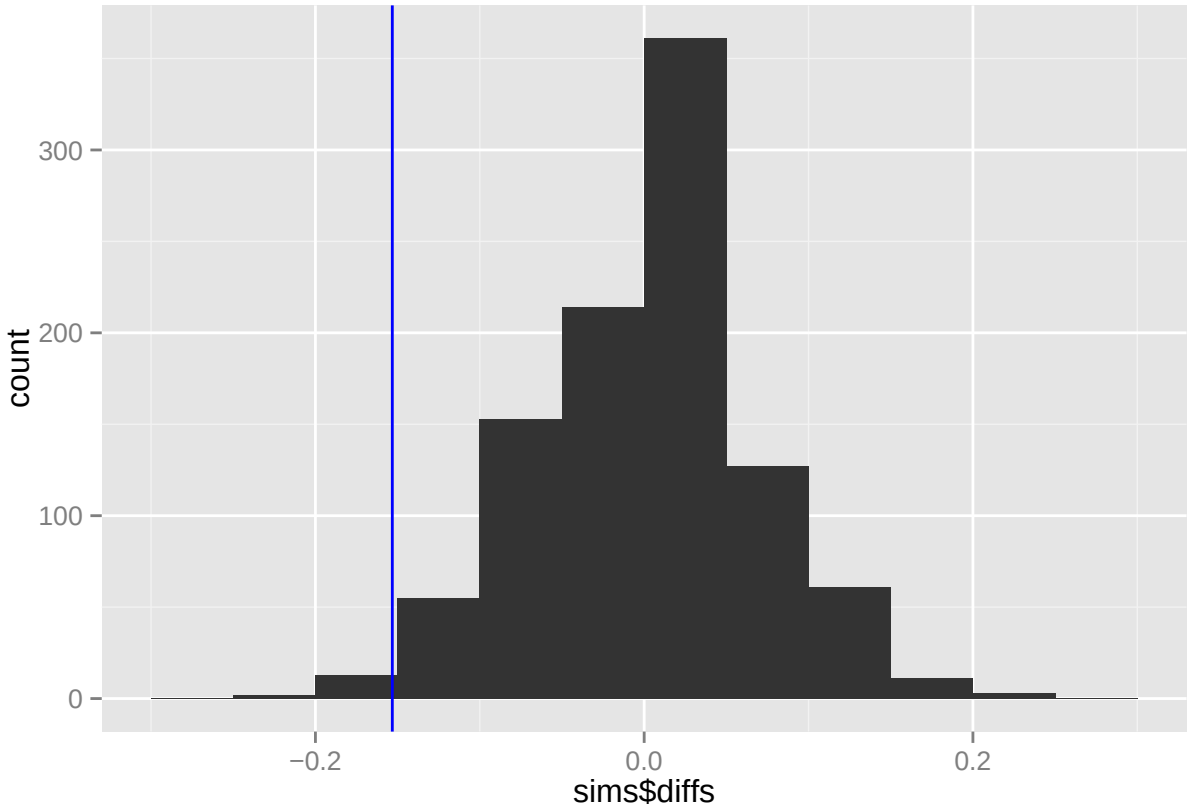
```
obs_diff <- diff(prop(low ~ smoke))
obs_diff
```

```
##     No.Yes
## -0.1532315
```

```
# We added that last line to print the result to the output file. If
# only the previous line were included, obs_diff would be defined,
# but the definition would take place silently, printing nothing to
# the screen.
```

Plot the simulated values of the null distribution.

```
sims <- do(1000) * diff(prop(low ~ shuffle(smoke)))
sims <- rename(sims, diffs = No.Yes)
qplot(sims$diffs, binwidth = 0.05) +
    geom_vline(x = obs_diff, color = "blue")
```



Calculate the P-value.

```
2 * prop( ~ sims$diffs <= obs_diff)
```

```
## TRUE
## 0.03
```

(Note, we multiply here by two because we are conducting a two-sided test.)

Conclusion

State the statistical conclusion.

We reject the null hypothesis.

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

There is sufficient evidence to suggest that there is an association between smoking and low birth weight.

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 committing a Type I error. It is possible that there is no association between smoking and low birth weight, but we got an unusual sample that led us to believe that there was such an association.

Your turn!

Now it's your turn to run a complete hypothesis test. Determine if there is evidence that the presence of uterine irritability is associated with low birth weight. I will give you the template below, then you need to fill in the gaps. Some of the steps will be the same or similar to steps in the example above. For this assignment, you can copy what I've written and modify where needed. In the future, you'll need to put everything in your own words.

Hypotheses

Identify the sample and a reasonable population of interest.

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

Express the null and alternative hypotheses in symbols.

Model

Identify the correct sampling distribution model.

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

Mechanics

Compute the test statistic.

```
## Add code here to compute the test statistic.
```

Plot simulated values of the null distribution.

```
## Add code here to plot simulated values of the null distribution.
```

Calculate the P-value.

```
## Add code here to calculate the P-value.
```

Conclusion

State the statistical conclusion.

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

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.