Hypothesis testing with simulation, Part 1

Put your name here

Put the date here

Introduction

Using a sample to deduce something about a population is called "statistical inference". In this module, we'll learn about one form of statistical inference called "hypothesis testing". The focus will be on walking through the example from Part 2 of "Introduction to simulation" and recasting it here as a formal hypothesis test.

There are no new R commands here, but there are many new ideas that will require careful reading. You are not expected to be an expert on hypothesis testing after this one module. However, within the next few modules, as we learn more about hypothesis testing and work through many more examples, the hope is that you will begin to assimilate and internalize the logic of inference and the steps of a hypothesis test.

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 `MASS` package to access the `birthwt` data on risk factors associated with low birth weight. We also load the `mosaic` package for simulation tools and the `gmodels` package for nice contingency tables using the `CrossTable` command.

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

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

```
set.seed(9999)
```

Our research question

We are interested in finding out if there is an association between smoking during pregnancy and low birth weight.

Data preparation

You will recall that we need to convert the two variables of interest to factor variables and put them in a new data frame.

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

Hypothesis testing

The approach we used in Part 2 of "Introduction to Simulation" was to assume that the two variables `smoke` and `low` were independent. From that assumption, we were able to compare the observed difference in low birth weight percentages between smokers and non-smokers from the actual data to the distribution of random values obtained by simulation. When the observed difference was far enough away from zero, we concluded that the assumption of independence was probably false, giving us evidence that the two variables were associated after all.

This logic is formalized into a sequence of steps known as a *hypothesis test*. In this section, we will introduce a rubric for conducting a full and complete hypothesis test for the smoking and low birth weight example. Please locate the file `Rubric_for_inference.pdf` to see the steps for each part of the rubric.

A hypothesis test can be organized into four parts:

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

Below, I'll address each of these steps.

Hypotheses

We are trying to ask some question about a population of interest. However, all we have in our data is a sample of that population. The word inference comes from the verb "infer": we are trying to infer what might be true of a population just from examining a sample. It's also possible that our question involves comparing two or more populations to each other. In this case, we'll have multiple samples, one from each of our populations. For example, in our birth weight question, we are comparing two populations: women who smoked during pregnancy and women who didn't. Our data gives us two samples (again smokers and non-smokers) who form only a part of the larger populations of interest.

To convince our audience that our analysis is correct, it makes sense to take a skeptical position. If we are trying to prove that there is an association between smoking during pregnancy and low birth weight, we

don't just declare it to be so. We start with a "null hypothesis", or an expression of the belief that there is no association. A null hypotheses always represent the "default" position that a skeptic might take. It codifies the idea that "there's nothing to see here."

Our job is to gather evidence to show that there is something interesting going on. The statement of interest to us is called the "alternative hypothesis". This is usually the thing we're trying to prove.

We can perform *one-sided* tests or *two-sided* tests. A one-sided test is when we have a specific direction in mind for the effect. For example, if we are trying to prove that smoking mothers are *more* likely to have low birth weight babies, then we would perform a one-sided test. On the other hand, if we only care about proving an association, then that could be that smoking mothers are either more likely or less likely to have low birth weight babies. (This is contrasted to the null that states that smoking mothers are equally likely as non-smoking mothers to have low birth weight babies.) It seems weird that we would run a two-sided test, but I want to give my statistical analysis a chance to prove an association regardless of the direction of the association. Wouldn't you be interested to know if it turned out that smoking mothers were, in fact, less likely to have low birth weight babies?

You can't cheat and look at the data first. In a normal research study out there in the real world, you develop hypotheses long before you collect data. So you have to decide to do a one-sided or two-sided test before you have the luxury of seeing your data pointing in one direction or the other.

Using two-sided tests is a good default option. Again, this is because our analysis will allow us to show interesting effects in any direction.

We typically express hypotheses in two ways. First, we write down full sentences that express in the context of the problem what our null and alternative hypotheses are stating. Then, we express the same ideas as mathematical statements. This translation from words to math is important as it gives us the connection to the quantitative statistical analysis we need to perform. The null hypothesis will always be that some quantity is equal to ($=$) the null value. The alternative hypothesis depends on whether we are conducting a one-sided test or a two-sided test. A one-sided test is mathematically saying that the quantity of interest is either greater than ($>$) or less than ($<$) the null value. A two-sided test always states that the quantity of interest is not equal to ($\neq$) the null value.

The most important thing to know is that the entire hypothesis test up until you reach the conclusion is conducted **under the assumption that the null hypothesis is true**. In other words, we pretend the whole time that our alternative hypothesis is false, and we carry out our analysis working under that assumption. This may seem odd, but it makes sense when you remember that the goal of inference is to try to convince a skeptic. Others will only believe your claim after you present evidence that suggests that the data is inconsistent with the claims made in the null.

Model

A model is an approximation—usually a simplification—of reality. In a hypothesis test, when we say "model" we are talking specifically about the "null model". In other words, what is true about the population under the assumption of the null? If we sample from the population repeatedly, we find that there is some kind of distribution of values that can occur by pure chance alone. This is called the *sampling distribution model*. We have been learning about how to use simulation to understand the sampling distribution model and how much sampling variability to expect, even when the null hypothesis is true.

Building a model is contingent upon certain assumptions being true. We cannot usually demonstrate directly that these assumptions are conclusively met; however, there are often conditions that can be checked with our data that can give us some confidence in saying that the assumptions are probably met. For example, there is no hope that we can infer from our sample unless that sample is close to a random sample of the population. There is rarely any direct evidence of having a properly random sample, and often, random samples are too much to ask for. There is almost never such a thing as a truly random sample of the population. Nevertheless, it is up to us to make the case that our sample is as representative of the population as possible. Additionally, we have to know that our sample comprises less than 10% of the size of the population. The reasons for this

are somewhat technical and the 10% figure is just a rough guideline, but we should think carefully about this whenever we want our inference to be correct.

Those are just two examples. For the simulation tests we are running, those are the only two conditions we need to check. For other hypothesis tests in the future that use different types of models, we will need to check more conditions that correspond to the modeling assumptions we will need to make.

Mechanics

This is the nitty-gritty, nuts-and-bolts part of a hypothesis test. Once we have a model that tells us how data should behave under the assumption of the null hypothesis, we need to check how our data actually behaved. The measure of where our data is relative to the null model is called the *test statistic*. For example, if the null hypothesis states that there should be a difference of zero between mothers who smoke and mothers who don't smoke, then the test statistic would be the actual observed difference in our data between smokers and non-smokers.

Once we have a test statistic, we can plot it in the same graph as the null model. This gives us a visual sense of how rare or unusual our observed data is. The further our test statistic is from the center of the null model, the more evidence we have that our data would be very unusual if the null model were true. And that, in turn, gives us a reason not to believe the null model. When conducting a two-sided test, we will actually graph the test statistic and the negative of the test statistic to acknowledge that we're interested in evidence of an effect in either direction.

Finally, we convert the visual evidence explained in the previous paragraph to a number called a *P-value*. This measures how likely it is to see our observed data—or data even more extreme—under the assumption of the null. A small P-value, then, means that if the null were really true, we wouldn't be very likely at all to see data like ours. That leaves us with little confidence that the null model is really true. (After all, we *did* see the data we gathered!) If the P-value is large—in other words, if the test statistic is closer to the middle of the null distribution—then our data is perfectly consistent with the null hypothesis. That doesn't mean the null is true, but it certainly does not give us evidence against the null.

A one-sided test will give you a P-value that only counts data more extreme than the observed data in the direction that we explicitly hypothesized. For example, if our alternative hypothesis was that smoking mothers were more likely to have low birth weight babies, then we would only look at the part of the model that showed differences with as many or more low birth weight babies as our data showed. A two-sided P-value, by contrast, will count data that is extreme in either direction. This will include values on both sides of the distribution, which is why it's called a two-sided test. Computationally, it is usually easiest to calculate the one-sided P-value and just double it.[1]

Remember the statement made earlier that throughout the hypothesis testing process, **we work under the assumption that the null hypothesis is true**. The P-value is no exception. It tells us **under the assumption of the null** how likely we are to to see data at least as extreme (if not even more extreme) as the data we actually saw.

Conclusion

The P-value we calculate in the Mechanics section allows us to determine what our decision will be relative to the null hypothesis. As explained above, when the P-value is small, that means we had data that would be very unlikely had the null been true. The sensible conclusion is then to "reject the null hypothesis". On the other hand, if the data is consistent with the null hypothesis, then we "fail to reject the null hypothesis."

How small does the P-value need to be before we are willing to reject the null hypothesis? That is a decision we have to make based on how much we are willing to risk an incorrect conclusion. A value that is widely

[1] This is not technically the most mathematically appropriate thing to do, but under certain assumptions that are usually checked in the "Conditions" section of the hypothesis test, it can be justified in many cases.

used is 0.05; in other words, if $P < 0.05$ we reject the null, and if $P \geq 0.05$, we fail to reject the null. However, for situations where we want to be conservative, we could choose this threshold to be much smaller. If we insist that the P-value be less than 0.01, for example, then we will only reject the null when we have a lot more evidence. The threshold we choose is called the "significance level", denoted by the Greek letter alpha: α. Ideally, α should be chosen long before we compute our P-value so that we're not tempted to cheat and change the value of α to suit our P-value (and by doing so, quite literally, move the goalposts).

Note that we never accept the null hypothesis. This procedure gives us no evidence in favor of the null. All we can say is that the evidence is either strong enough to warrant rejection of the null, or else it isn't, in which case we can conclude nothing. If we can't prove the null false, we are left not knowing much of anything at all.

The phrases "reject the null" or "fail to reject the null" are very statsy. Your audience may not be statistically trained. Besides, the *real* conclusion you care about concerns the question of interest you posed at the beginning of this process, and that is built into the alternative hypothesis, not the null. Therefore, we need some statement that addresses the alternative hypothesis in words that a general audience will understand. I recommend the following templates:

- When you reject the null, you can safely say, "We have sufficient evidence that [restate the alternative hypothesis]."
- When you fail to reject the null, you can safely say, "We have insufficient evidence that [restate the alternative hypothesis]."

The last part of your conclusion should be an acknowledgement of the uncertainty in this process. Statistics tries to tame randomness, but in the end, randomness is always somewhat unpredictable. It is possible that we came to the wrong conclusion, not because we made mistakes in our computation, but because statistics just can't be right 100% of the time when randomness is involved. Therefore, we need to explain to our audience that we may have made an error.

A *Type I* error is what happens when the null hypothesis is actually true, but our procedure rejects it anyway. This happens when we get a "freak" sample. For example, perhaps there really is no association between smoking and low birth weight. Even if that were true, we could accidentally survey a group of women who smoked and also—by pure chance alone—happen to have more babies with low birth weight. Our test statistic will be "accidentally" far from the null value, and we will mistakenly reject the null. Whenever we reject the null, we are at risk of making a Type I error. Given that we are conclusively stating a statistically significant finding, if that finding is wrong, this is a *false positive*, a term that is synonymous with a Type I error. The significance level α discussed above is, in fact, the probability of making a Type I error. (If the null is true, we will still reject the null if our P-value happens to be less than α.)

On the other hand, the null may actually be false, and yet, we may not manage to gather enough evidence to disprove it. This can also happen due to a "freak sample"—a sample that doesn't conform to the "truth". But there are other ways this can happen as well, most commonly when you have a small sample size (which doesn't allow you to prove much of anything at all) or when the effect you're trying to measure exists, but is so small that it is hard to distinguish from no effect at all (which is what the null postulates). In these cases, we are at risk of making a *Type II* error. Anytime we say that we fail to reject the null, we have to worry about the possibility of making a Type II error, also called a *false negative*.

Example

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 slightly different question. Unless otherwise stated, we will always assume a significance level of $\alpha = 0.05$. (In other words, we will reject the null if our computed P-value is less than 0.05, and we will fail to reject the null if our P-value is greater than or equal to 0.05.)

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

There are technically two samples of interest here. All the data comes from mothers who gave birth at the Baystate Medical Center, Springfield, Massachusetts, during 1986, but one group of interest are mothers who smoked during pregnancy, and the other group of interest are mothers who did not smoke during pregnancy.

The following table shows the sample sizes for each group in the marginal distribution along the right side of the table (i.e., the row sums):

```
CrossTable(smoke_low$smoke, smoke_low$low,
           prop.r = TRUE, prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |           N / Row Total |
## |-------------------------|
##
##
## Total Observations in Table:  189
##
##
##                  | smoke_low$low
## smoke_low$smoke |       Yes |        No | Row Total |
## ----------------|-----------|-----------|-----------|
##             Yes |        30 |        44 |        74 |
##                 |     0.405 |     0.595 |     0.392 |
## ----------------|-----------|-----------|-----------|
##              No |        29 |        86 |       115 |
##                 |     0.252 |     0.748 |     0.608 |
## ----------------|-----------|-----------|-----------|
##    Column Total |        59 |       130 |       189 |
## ----------------|-----------|-----------|-----------|
##
##
```

So there are 74 mothers who smoked during pregnancy, and 115 mothers who did not.

The populations of interest are probably all mothers who smoke and all mothers who don't smoke, maybe in the U.S., although we are only really safe coming to conclusions about the 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.

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

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

(Note: pay close attention here to the order of the subtraction. While it doesn't matter conceptually, the `diffprop` command later on will do the subtraction a certain way. You need to make sure the order listed here in the hypotheses is consistent with the way `diffprop` works. You'll be able to tell based on whether `diffprop` gives you a negative number or a positive number.)

Model

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

- Random
  - We have no evidence that these are random samples of mothers from the the Baystate Medical Center. We hope that they are representative samples. If the populations of interest are all mothers (smokers and non-smokers) in the U.S., for example, then I have some doubts as to how representative these samples are. (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 populations, 74 smoking mothers and 115 non-smoking mothers are surely less than 10% of all mothers under consideration.

Mechanics

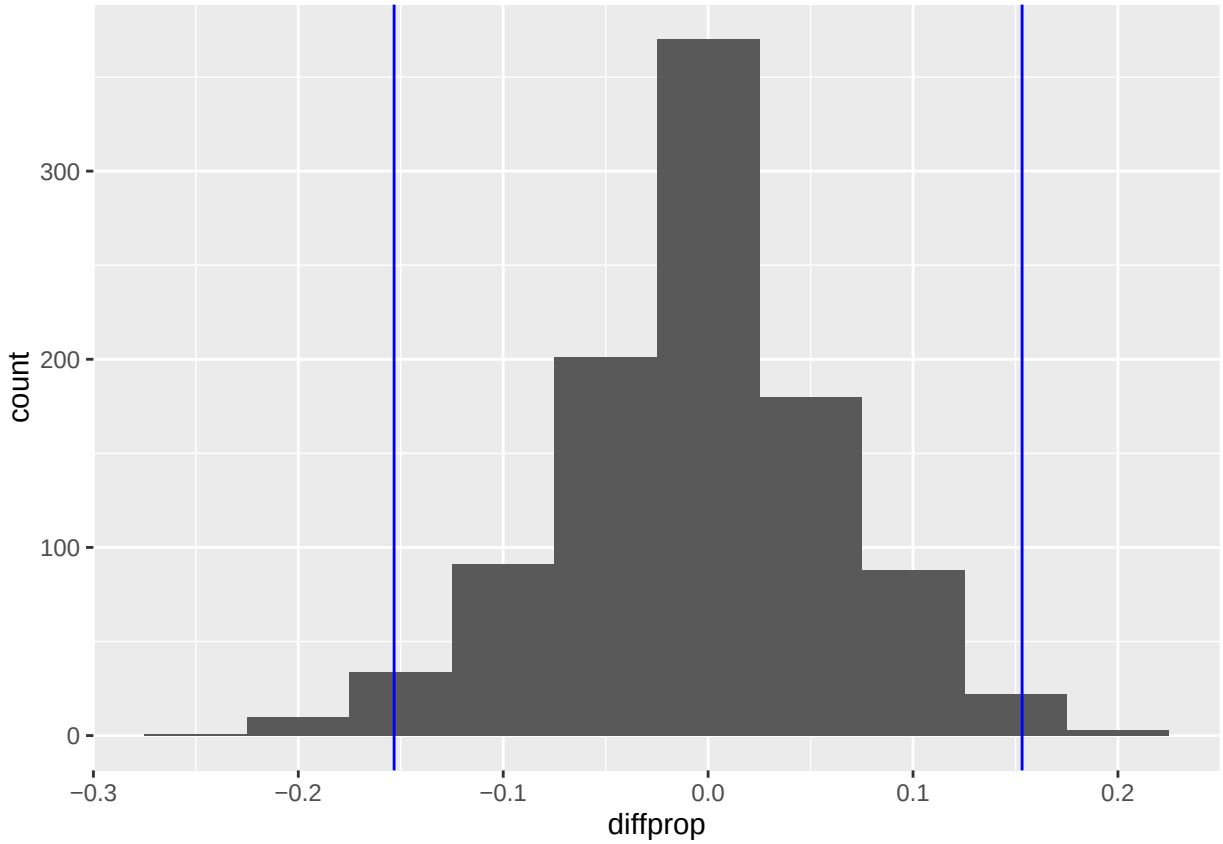
Compute the test statistic.

```
obs_diff <- diffprop(low ~ smoke, data = smoke_low)
obs_diff
```

```
##   diffprop
## -0.1532315
```

Plot the simulated values of the null distribution.

```
sims <- do(1000) * diffprop(low ~ shuffle(smoke), data = smoke_low)
ggplot(sims, aes(x = diffprop)) +
    geom_histogram(binwidth = 0.05) +
    geom_vline(xintercept = obs_diff, color = "blue") +
    geom_vline(xintercept = -obs_diff, color = "blue")
```



(You'll note that we added two blue vertical lines here. This is because we are conducting a two-sided test, which means that we're interested in values that are more extreme than our observed difference in *both* directions.)

Calculate the P-value.

```
2 * prop(sims$diffprop <= obs_diff)
```

```
## TRUE
## 0.05
```

(Note, we multiply here by two because we are conducting a two-sided test. We would be surprised by values that are unusually positive as well as unusually negative. Also, you can safely ignore the word `TRUE` in the output.)

Conclusion

State the statistical conclusion.

We fail to reject the null hypothesis.

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

There is insufficient 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 failed to reject the null, we run the risk of committing a Type II error. It is possible that there is an association between smoking and low birth weight, but our samples did not give us evidence that was conclusive enough.

After writing up your conclusions and acknowledging the possibility of a Type I or Type II error, the hypothesis test is complete. (At least for now. In the future, we will add one more step of computing a confidence interval.)

Failing to reject the null

If you're doing this assignment soon after finishing Part 2 of the "Introduction to simulation", you may have noticed something unusual. In the previous assignment we concluded that there seemed to be evidence of an assocation between smoking and low birth weight. And yet, here in this document, we have failed to reject the null, meaning that we have found insufficient evidence of such an association.

The reason for this is simple. In the previous assignment we did the "intuitive" thing and ran a one-sided test. We did not call it that in the previous assignment; we had not yet developed the vocabulary. But if you go back and look at the previous assignment, you'll find that we computed only simulated values that fell below the observed difference:

```
prop(sims$diffprop <= obs_diff)
```

```
##  TRUE
## 0.025
```

The justification was that we already suspected that smoking mothers were more likely to give birth to babies with low birth weight, and it appears that our evidence (the test statistic, or our observed difference) was pretty far in that direction.

(We did get a slightly different number in the previous assignment. Remember that we are simulating which involves randomness. Therefore, we won't expect to get the exact same numbers each time.)

By way of contrast, in this assignment we computed the two-sided P-value:

```
2 * prop(sims$diffprop <= obs_diff)
```

```
## TRUE
## 0.05
```

Our P-value in this assignment is twice as large as it could have been if we had run a one-sided test. And doubling the P-value means that it no longer falls under the significance threshold $\alpha = 0.05$.

This raises an obvious question: why use two-sided tests? If the P-values are higher, that makes it less likely that we will reject the null, which means we won't be able to prove our alternative hypothesis. Isn't that a bad thing?

As a matter of fact, there are many researchers in the world who do think it's a bad thing, and routinely do things like use one-sided tests to give them a better chance of getting small P-values. But this is not ethical.

The point of research is to do good science, not prove your pet theories correct. There are many incentives in the world for a researcher to prove their theories correct (money, awards, career advancement, fame and recognition, legacy, etc.), but these should be secondary to the ultimate purpose of advancing knowledge. Sadly, many researchers out there have these priorities reversed. I do not claim that researchers set out to cheat; I suspect that the vast majority of researchers act in good faith. Nevertheless, the rewards associated with "successful" research cause cognitive biases that are hard to overcome. And "success" is often very narrowly defined as research that produces small P-values.

A better approach is to be conservative. For example, a two-sided test is not only more conservative because it produces higher P-values, but also because it answers a more general question. That is, it is scientifically interesting when an association goes in either direction (e.g. smoking mothers have more babies with low birth weight, or smoking mothers have fewer babies with low birth weight). This is why we recommended above using two-sided tests by default, and only using a one-sided test when there is a very strong research hypothesis that justifies it.

Also remember that when we fail to reject the null hypothesis, we are not saying that the null hypothesis is true. Neither are we saying it's false. Failure to reject the null is really a failure to conclude anything at all. But rather than looking at it as a failure, a more productive viewpoint is to see it as an opportunity for more research, possibly with larger sample sizes.

Even when we do reject the null, it is important not to see that as the end of the conversation. Too many times, a researcher publishes a "statistically significant" finding in a peer-reviewed journal, and then that result is taken as "Truth". We should, instead, view statistical inference as incremental knowledge that works slowly to refine our state of scientific knowledge, as opposed to a collection of "facts" and "non-facts".

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. As we'll always assume unless otherwise stated, use a significance level of $\alpha = 0.05$.

I have copied the template below. You need to fill in each step. 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`.

Hypotheses

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

Please write up your answer here.

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

H_0 : Null hypothesis goes here.

H_A : Alternative hypothesis goes here.

Express the null and alternative hypotheses in symbols.

$H_0 : math$

$H_A : math$

Model

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

Please write up your answer here.

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.

Please write up your answer here.

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

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.

Please write up your answer here.

Conclusion

A hypothesis test is a formal set of steps—a procedure, if you will—for implementing the logic of inference. We take a skeptical position and assume a null hypothesis in contrast to the question of interest, the alternative hypothesis. We build a model under the assumption of the null hypothesis to see if our data is consistent with the null (in which case we fail to reject the null) or unusual/rare relative to the null (in which case we reject the null). We always work under the assumption of the null so that we can convince a skeptical audience using evidence. We also take care to acknowledge that statistical procedures can be wrong, and not to put too much credence in the results of any single set of data or single hypothesis test.