Using Visualizations to Understand Intro Statistics

Using the CLT, Building a Sampling Distribution, and One-sided Hypothesis Testing for Population Mean

This R Markdown File will walk you through the steps of your analysis. There are multiple “chunks” of code below. Follow the instructions to make changes to the code. When you are ready to run the code, press the green arrow button on the right side of the code chunk.

Note: If you save your work and come back to it later

You can save and make edits to this file at any time. Remember that if you close and re-open the file, you need need to first re-import your dataset and re-run the code from all the steps above the step you’re currently working on. So for example, if we want to make edits to Step 6, we need to first import the dataset (directions in Step 2) then go to the first the code chunk for Step 6. To run all the code above Step 6, click the symbol (next to the usual green arrow) that looks like a grey triange with a line. This will run all the code above Step 6 we can access any values that we have previously calculated.

Step 0: Setup

First we need to install some packages that we will use in this analysis. Packages are collections of functions and tools that can be used in R. (Note that you will need to be connected to internet to install the packages.) Copy and paste the following in the Console window, then hit enter for the packages to start installing:

install.packages(c(“tidyverse”,“ggplot2”, “readr”, “BSDA”))

Great! Now that you’ve installed these packages, they are stored on your computer for you to load anytime, so you won’t need to install them again.

After we’ve installed the packages, next we need to load the packages so we use them in our current R Session. Simply press the green arrow button in the code chunk below. You should see a green bar on the left side of the code that indicates your commands are running, but other than that, you shouldn’t see much else happening here.

We’re all set up! Now let’s get to some statistics.

Step 1: Make your hypothesis and get sample

Visit https://learn.concord.org/eresources/1239.run_resource_html (https://learn.concord.org/eresources/1239.run_resource_html). This website has information about Americans of all ages from the 2003 National Health and Nutrition Examination Survey (NHANES).

Click the Options tab in the pane on the left side of the page. Specify the age range you’re interested in learning about, then click on Demography, Body measurements, Blood pressure, Biochemistry, and Sexual behavior to expand all the sub-attributes. Select ONE quantitative variable of your choice, and make sure all other variables

are unchecked.

What is your hypothesis about the population? For example, I might hypothesize that the average BMI (Body Mass Index) for adults aged 18-100 in 2003 was greater than 25. (The CDC considers a BMI between 18.5-24.9 to be “healthy weight”, a BMI between 25-29.9 to be “overweight”, and BMI over 30 to be “obese”.)

Question 1a: Please form a one-sided hypothesis about the population mean, and write it below:

After you have formed your hypothesis, it’s time to take a sample. At the top of the page where it says “How many people?”, specify the number of people you want in the sample. Make sure this number is at least 40, click “get # people”. The app will take a random sample of people from the specified population in the NHANES database.

To download your dataset, click the ruler next to your data, then select “Export Case Data”. Then click “Local File”. (Leave the name of the file as “2003 NHANES Data Portal”) then click Download.

Step 2: Import dataset into R and determine sample size

Now we will open our sample data in our R session. In RStudio, find the window that says Environment at the top. Click “Import Dataset” at the top of the Environment window, then select “From Text (readr)”. In the window that pops up, click “Browse” to select the file from wherever it is saved on your computer. After selecting the file, you should see a preview of your data. Make sure there is a check mark by “First rows as names” and “Trim Spaces”. Click “Import”.

After your data is imported, you can always click on the name of your dataset in the Environment window. You will see a tab appear that displays your data.

There may be some missing values in your dataset, so we want to figure out the size of sample without those missing values. In the first line below, simply replace the word variable (immediately after the $) with the name of your variable, *exactly* as the variable name appears in your dataset. (Pay attention to any capital letters or underscores!) For example, if I’m studying body mass index, I would replace the word variable with BMI. Then press the green arrow. The output, a variable we called n, tells us the size of our sample.

```
fullsample <- X2003_NHANES_Data_Portal$BMI
mysample <- fullsample[!is.na(fullsample)]

n <- length(mysample)
n
```

```
## [1] 77
```

Step 3: Checking for understanding

Based on your hypothesis in Step 1, please answer the following:

Question 3a: What is the population?

Question 3b: What is the sample?

Question 3c: What is the null hypothesis?

Type the number corresponding to your null hypothesis below after mu=___. For example, if my null hypothesis is that the mean BMI of American adults is 25, then I'd type mu = 25 Then press the green arrow.

```
mu = 25
```

Question 3d: What is the alternative hypothesis?

Step 4: Setting up the sampling distribution

Next we will build a sampling distribution based on your null hypothesis.

Question 4a: If the NULL hypothesis is true, and we took all possible samples from the population, what would the mean of these sample means be? Why?

(Answer should be the same value that students specified in the null hypothesis. Why? Central Limit Theorem. Answering this question correctly is important for the following steps, so the end of Step 4 might be a good checkpoint)

Question 4b: If the NULL hypothesis is true, what would the mean of the sampling distribution be?

Note that your answer for Question 4a and 4b should be the same because a sampling distribution is simply the distribution of all sample means!

Remember that the Central Limit Theorem tells us that the standard deviation of the sampling distribution is equal to the population standard deviation divided by the square root of n. Since we don't know the population standard deviation, we will need to use the standard deviation of our sample.

The code chunk below will help us find the standard deviation of our sample and sampling distribution. You don't need to edit the code; just press the green arrow for it to run.

```
sample_sd <- sd(mysample)
sample_sd
```

```
## [1] 5.819041
```

```
sampling_dist_sd <- sample_sd/sqrt(n)
sampling_dist_sd
```

```
## [1] 0.6631412
```

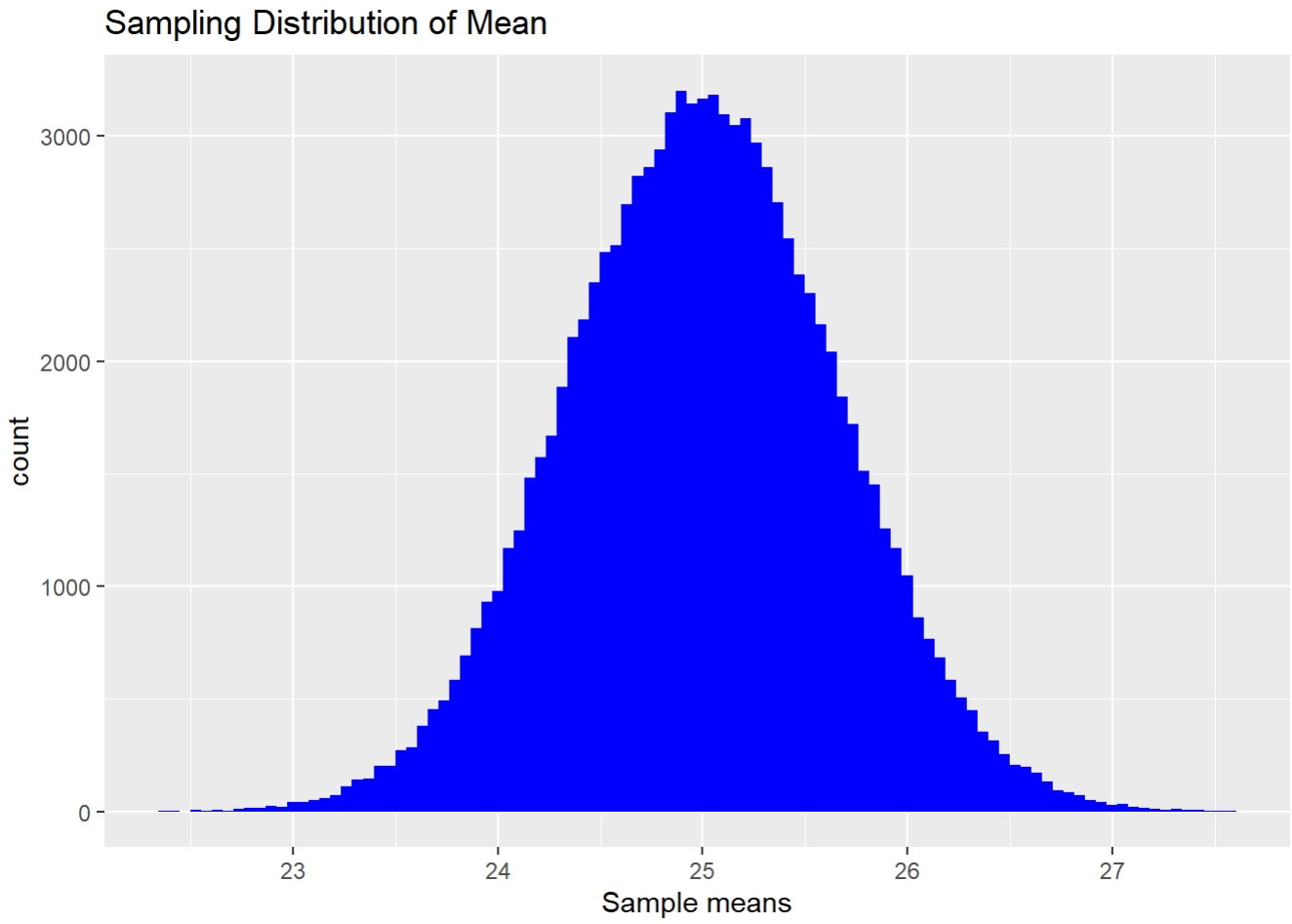
(Checkpoint!)

Step 5: Visualizing the sampling distribution

The code below simulates building a sampling distribution, so our commands tell R to take repeated samples from a population with the hypothesized mean mu. Then we are plotting all of those means to build a histogram of our sampling distribution. Simply press the green arrow; you don't need to make any edits to the code.

```
set.seed(0)
repeated_samples_df <- replicate(100000, mean(rnorm(n, mean = mu, sd= sample_sd))) %>% as_tibble
()

ggplot(repeated_samples_df) + geom_histogram(aes(value),fill = "blue", bins = 100) + ggtitle("Sa
mpling Distribution of Mean") + xlab("Sample means")
```



Question 5a: Great! From looking at the histogram you built, what do you estimate is the mean of the sample means (i.e. the mean of our sampling distribution)?

Question 5b: How does your answer to Question 5a compare to mu? (Remember that mu is the *population mean* under the null hypothesis!) How does your answer from Question 4c compare to what you expected to see in Question 4a?

Step 6: Find the test statistic and p value

Simply press the green arrow to find the mean of your sample:

```
sample_mean <- mean(mysample)
sample_mean
```

```
## [1] 26.33636
```

Next we will perform a hypothesis test to find our test statistic and p value.

Look again at your alternative hypothesis. If the alternative hypothesis is that the true population mean is greater than mu, then make sure the code below says alternative = “greater”. If the alternative hypothesis is that the true population mean is less than mu, then make sure the code below says alternative = “less”.

```
z.test(mysample, alternative = "greater", mu = mu, sigma.x = sample_sd, conf.level = .95)
```

```
##
##  One-sample z-Test
##
## data:  mysample
## z = 2.0152, p-value = 0.02194
## alternative hypothesis: true mean is greater than 25
## 95 percent confidence interval:
##  25.24559       NA
## sample estimates:
## mean of x
##  26.33636
```

Question 6a: In words, what probability does the p value represent?

Question 6b: In words, what probability does the p value represent *in this problem*?

Question 6c: Compare your p value to the significance level 0.05. Which is greater, your p value or 0.05? According to your data, do you reject or fail to reject the null hypothesis?

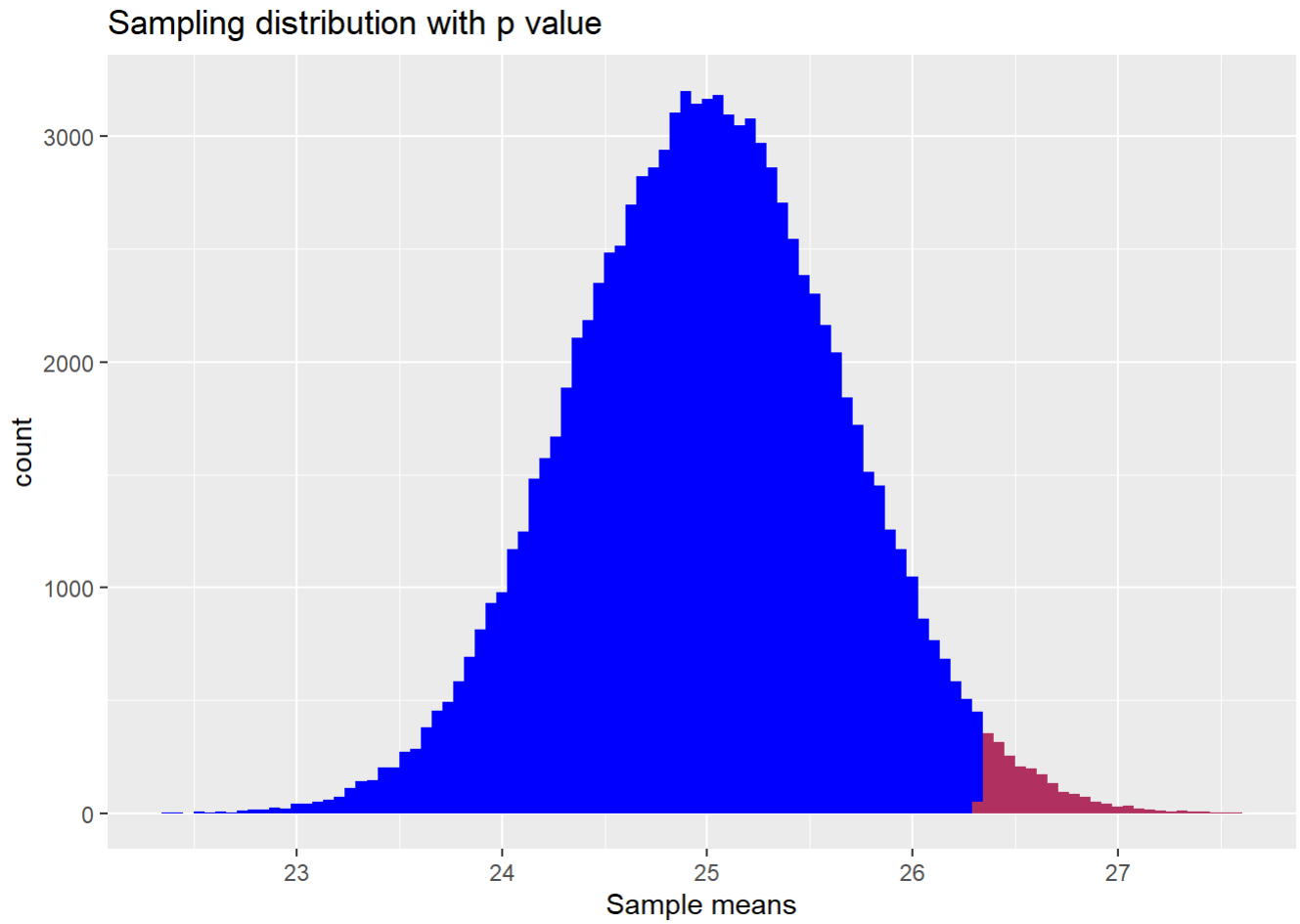
Question 6d: Please write 2-3 sentences describing the findings of your hypothesis test.

(Checkpoint after Section 6! Check understanding of p value in the context of this problem. Make sure they understand how p relates to the null hypothesis.)

Step 7: Adding the p value to our sampling distribution

Next we will show where our sample mean falls on the sampling distribution. Simply press the green arrow on the code below; no edits needed. All of the values above our sample mean will be colored in maroon. All the samples below our sample mean will be colored in blue.

```
ggplot(repeated_samples_df, aes(value)) + geom_histogram(aes(fill = value > sample_mean), bins = 100) + scale_fill_manual(values = c("blue", "maroon"), guide = F) + ggtitle("Sampling distribution with p value") + xlab("Sample means")
```



Note: If your sample mean is very small or very large, you might not even be able to see it on this plot!

Question 7a: What probability does the red area represent?

Question 7b: What probability does the blue area represent?

Question 7c: Which area is equal to p, the red area or the blue area? Why? (Hint: what is your alternative hypothesis?)

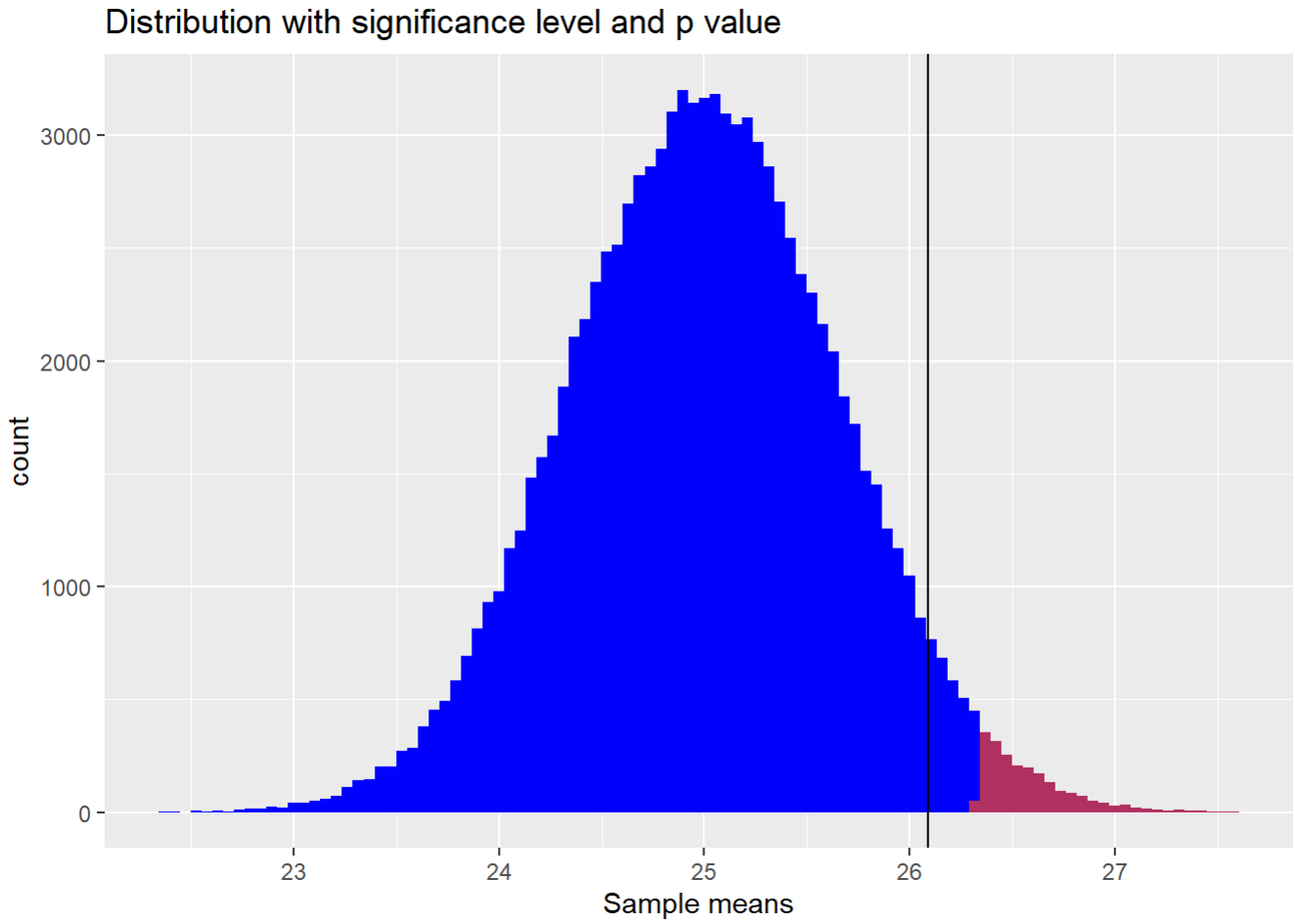
Step 8: Adding the significance level to the distribution

We are using a significance level of 0.05.

Look again at your alternative hypothesis. If the alternative hypothesis is that the true population mean is greater than mu, then make sure the first line of the code below says lower.tail = “FALSE”. If the alternative hypothesis is that the true population mean is less than mu, then make sure the code below says lower.tail = “TRUE”. Then run the code chunk to add a line representing the test significance level.

```
critical_value <- qnorm(.05, mean = mu, sd= sampling_dist_sd, lower.tail = FALSE)

ggplot(repeated_samples_df, aes(value)) + geom_histogram(aes(fill = value > sample_mean), bins = 100) + scale_fill_manual(values = c("blue", "maroon"), guide = F) + ggtitle("Distribution with significance level and p value") + xlab("Sample means")+ geom_vline(aes(xintercept = critical_value))
```



Question 8a: Go to either Part (1) or Part (2), depending on your alternative hypothesis:

1. If your alternative hypothesis was that the true population mean is *greater* than mu: In the plot we just made, look at the value where the significance line falls on the distribution. The probability of getting a sample with a mean that’s *greater* than that value is 5%. In other words, the area of the sampling distribution to the *right*

of the line is 5% of the total distribution. ### Is the area to the right of the line greater or less than the p value? How do you know?

2. If your alternative hypothesis was that the true population mean is *less* than mu: In the plot we just made, look at the value where the significance line falls on the distribution. The probability of getting a sample with a mean that's *less* than that value is 5%. In other words, the area of the sampling distribution to the *left* of the line is 5% of the total distribution. ### Is the area to the left of the line greater or less than the p value? How do you know?

Question 8b: How does your answer to Question 8a compare to your answer to question 6c?

Summary

Please write a 4-6 sentence statement summarizing your experiment, your hypotheses, the distributions that you plotted, and the conclusion of your analysis.