Module 4 Lab

Loading Data

In this lab we will read in data, do a hypothesis test and find a confidence interval for a population mean, and write a summary of our findings.

The data for this lab comes from the book, *An Introduction to Statistical Methods and Data Analysis, 5th Edition* by Ott and Longnecker. Leaves from 36 different apple tree orchards are crushed and their percentage nitrogen content is analyzed. This will give scientists an idea of the mean nitrogen content of apple trees across all such orchards and whether there is too little or too much nitrogen. The best yield occurs when the nitrogen content is approximately 2.5%.

The data for this lab come in the form of a Comma Separate Values (.csv) file. This type of file is one plain text way to store rectangular data. We'll download this file right from R. Open RStudio, and open your ST516 project (File -> Open Project). Then run:

```
download.file("https://gist.github.com/cwickham/36654bf1483eccca8885e08eacb8df8c/raw",
  "nitrogen.csv")
```

All going well, you'll see some messages in the Console that indicate R has succesfully downloaded the file, and you'll now see the file `nitrogen.csv` in your Files pane. If you click on the file, RStudio will open it up. It's not a very complicated file, a column name, then observed values, one per line.

To read the data from the file into R's memory we use the `read.csv()` function:

```
nitrogen <- read.csv('nitrogen.csv', header = TRUE)
```

The first argument is the path to the data file on the computer. The other argument we've passed here `header`, is an instruction to tell R that there are column names in the first line of the file. We store the table as `nitrogen`.

Use `?read.csv` to learn more about the `read.csv()` function. Also, you will learn more about reading in data in future labs.

Test Statistics and p-values

Now that we have loaded our data, and named it `nitrogen`, let's look at it. The command `head()` returns the first few rows of our data, including column names. If the data file had multiple columns - each one for a different variable - there would be multiple variable names and each would be associated with a different column in the output of `head()`. Use `?head` to learn more about this function.

```
head(nitrogen) # Take a look at data
```

```
##   nitrogen_content
## 1             2.10
```

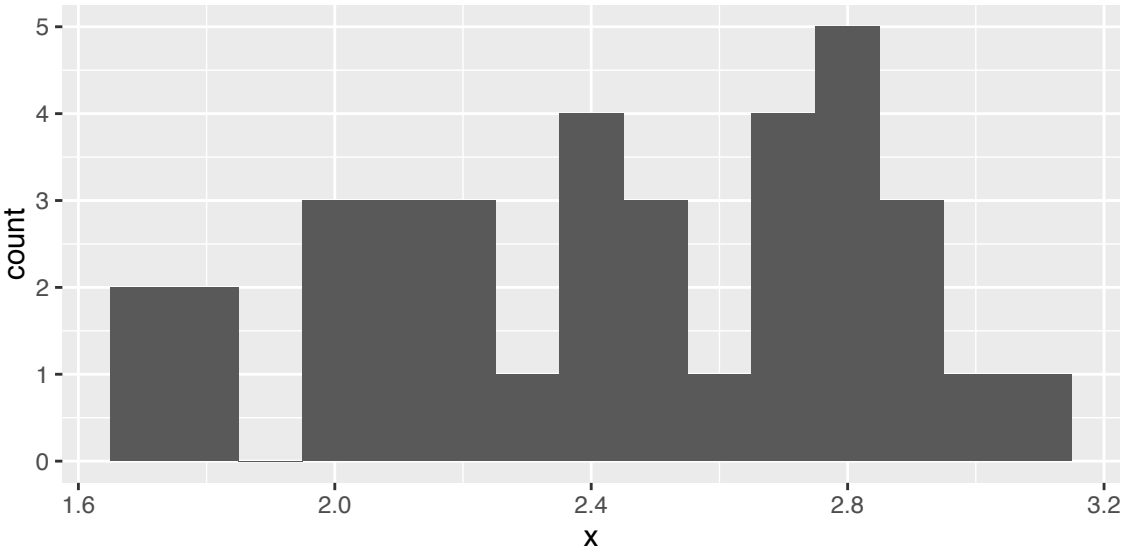
```
## 2            2.47
## 3            1.75
## 4            2.94
## 5            1.69
## 6            2.75
```

`nitrogen` is actually what we call a data frame (see `str(nitrogen)`) the most useful way of storing data in R, but we haven't talked about them yet, so for now let's pull out the data we need into a vector called `x`:

```
x <- nitrogen[ , 1] # Define x as *all rows, column 1* of "nitrogen"
```

Now we load the `ggplot2` package, and take a look at a histogram of our nitrogen contents to get a sense of our data.

```
library("ggplot2")
qplot(x, binwidth = 0.1)
```



The sample nitrogen contents have a mean of 2.4336. Now we will conduct a simple hypothesis test, where the null hypothesis is that the mean nitrogen content of all orchard trees is 2.5, and the alternative hypothesis is that mean nitrogen content of all orchard trees is different than 2.5. We choose this number because the best yield results happen when the nitrogen content is approximately 2.5%.

$$H_0 : \mu = 2.5$$
$$H_A : \mu \neq 2.5$$

With a sample size of 36 and no outliers and no indication of strong skew, we know $\overline{x}$ is approximately Normally distributed. Therefore, under H_0:

$$\frac{\overline{x} - \mu}{s/\sqrt{n}} \dot{\sim} N(0, 1)$$

For more details, see Module 4 lectures, or **Section 4.3.4** of *OpenIntro Statistics*. For now, we focus on the calculating the test statistic in R, and getting a p-value. We will use the `pnorm()` function to get a p-value. If you provide `pnorm()` with a quantile (our test statistic) it will return a probability.

```
# pg 178, OpenIntro

# Calculate z-statistic
Z <- (mean(x) - 2.5)/(sd(x)/sqrt(36))
Z
```

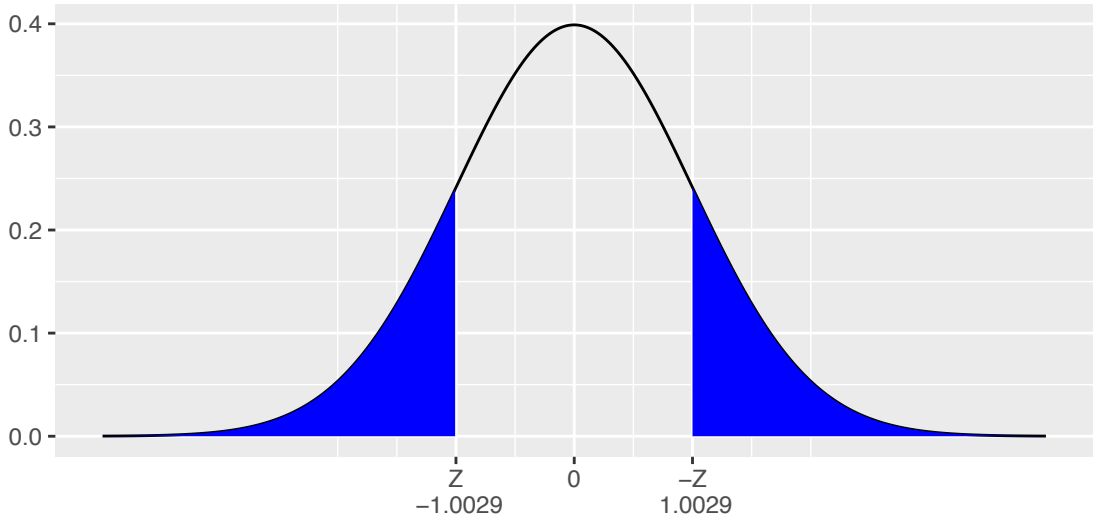
```
## [1] -1.00293
```

```
# Calculate p-value (two-sided test)
P <- 2 * pnorm(abs(Z), mean = 0, sd = 1, lower.tail = FALSE)
P
```

```
## [1] 0.3158947
```

The argument `lower.tail = FALSE` tells `pnorm()` to return the probability that a N(0,1) random variable is greater than the absolute value of Z, thus giving us the probability in the right tail. Since we are performing a two-sided test, and the Normal distribution is symmetric, we must multiply this probability by 2. This gives us the total probability of obtaining a sample mean at least as far from the center of the population as the sample mean that we have. Use `?pnorm()` to learn more about the function.

So what does this tell us? This tells us that if the null hypothesis (H_0) is true, the probability of observing a sample mean as unlikely (in this case–as far from the center of the population), or more unlikely (in this case–or farther away from the center of the population), than what we actually observed is 0.3159. That is fairly likely! How about a picture?



This is a N(0,1) density curve, with a blue shaded region representing the probability of observing a sample mean as unlikely as, or more unlikely than, what we actually observed.

Statistical Summary

Now we need to present our findings. As with most things, there are a few right ways, and many wrong ways to do this. Here is one example of the right way.

- There is no evidence to suggest that the mean nitrogen content of the orchard trees is different from 2.5 (Z-test, n $= 36$, p-value $= 0.32$).

There are a few things to notice:

- The sentence should translate the p-value to a level of evidence ($p > 0.1$ corresponds to "no" evidence)
- The sentence should be worded in terms of *evidence against the null*, e.g. "no evidence the mean is different from 2.5" or "no evidence the mean is not equal to 2.5". It should **never** be worded as evidence for the null e.g. "convincing evidence the mean is 2.5"
- Apart from the parenthesis, the sentence should be written in a way anyone can understand. Don't use the words "null", "alterative", or "hypothesis", translate the hypotheses into words in the context of the problem.
- You can and should include details of the actual analysis in parentheses. In general, what procedure was used, some indication of sample size, and the p-value.

Now for a few wrong ways. Here are some incorrect and/or incomplete summaries:

- We conclude that the mean nitrogen content of the orchard trees is not 2.5.
- The probability our conclusion is incorrect is 0.3159.
- We fail to reject the null hypothesis.
- We conclude the null hypothesis is true, based on a p-value of 0.3159.

Can you figure out what is wrong with each statement?

We will now obtain an interval estimate of the actual mean nitrogen content of the orchard trees. Note that we use $z_{\alpha/2}$ to denote the z-value with an area of $\alpha/2$ to the *right* of it. The form for such an estimate with 95% confidence is:

$$\left(\overline{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right)$$

Therefore our code is:

```
lower_bound <- mean(x) - qnorm(0.975) * sd(x)/sqrt(36)
upper_bound <- mean(x) + qnorm(0.975) * sd(x)/sqrt(36)
lower_bound
upper_bound
```

The confidence interval is: $(2.3, 2.56)$

Just like hypothesis test findings, there are a few right and wrong ways to present confidence interval findings. Here is one example of the right way:

- With 95% confidence, the true mean nitrogen content of orchard trees is between 2.3 and 2.56.

Here are a few wrong ways to include findings in a summary:

- There is a 95% probability that the true mean nitrogen content is between 2.3 and 2.56.
- The true mean nitrogen content is between 2.3 and 2.56.
- We are 95% confident that the sample mean nitrogen content is between 2.3 and 2.56.

Can you figure out what is wrong with each statement?

Notice that 2.5 *is* within the bounds of our confidence interval, so even if we only calculated the confidence interval, we know we would fail to reject the null hypothesis.

What would happen to the p-value if the hypothesized mean were barely outside the bounds of the confidence interval? What would happen as it approaches the middle of the confidence interval?