HW week 8

Adam Yang

6/28/2018

The file GPA1.RData contains data from a 1994 survey of MSU students. The survey was conducted by Christopher Lemmon, a former MSU undergraduate, and provided by Wooldridge.

```
load("GPA1.RData")
```

The skipped variable represents the average number of lectures each respondent skips per week. You are interested in testing whether MSU students skip over 1 lecture per week on the average.

a. Examine the skipped variable and argue whether or not a t-test is valid for this scenario.

```
length(data$skipped)
```

```
## [1] 141
```

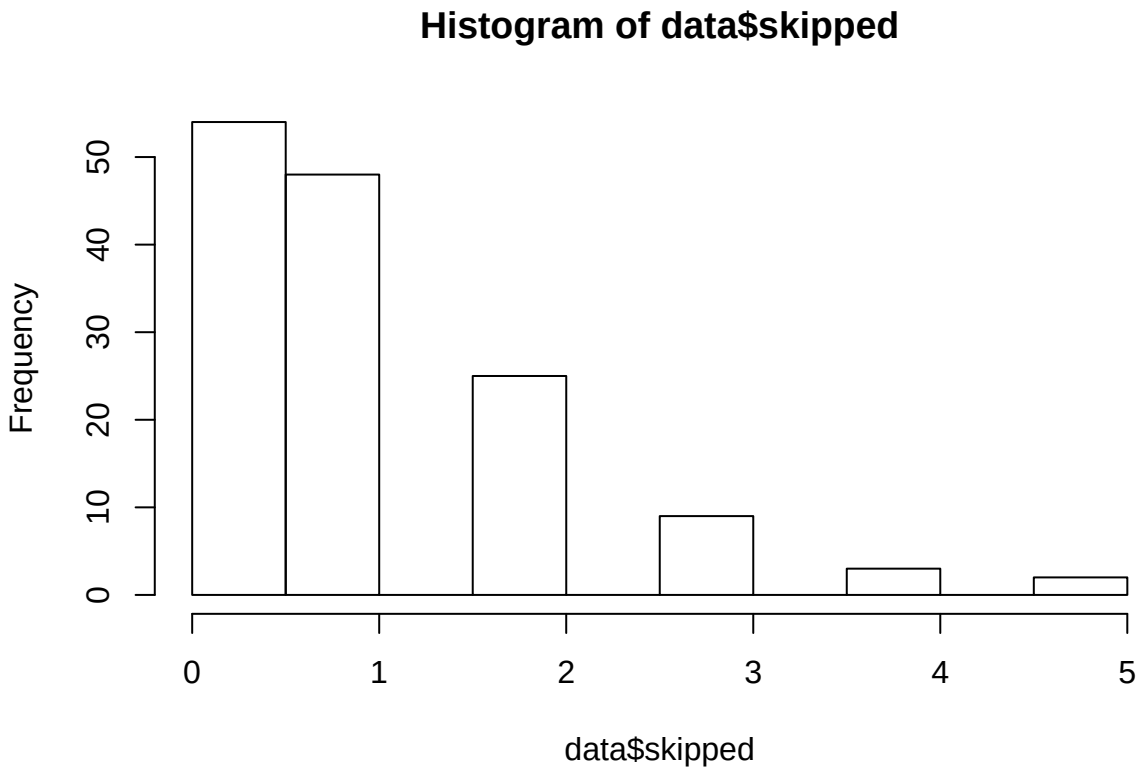
Our sample size is 141.

```
summary(data$skipped)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   1.000   1.076   2.000   5.000
```

The summary suggests a right skew because the mean is much closer to the Min than the Max and the Min and 1st quartile are equivalent.

```
hist(data$skipped)
```



In the Histogram and summary shown above, our sample distribution does not look normal at all, but actually skews towards the right. The Central Limit Theorem claims that if we have a large enough sample size, we can start approximating a normal sampling distribution. The general rule of thumb is that a sample size 30 would be large enough, but that is assuming there isn't a strong skew in the data. Because our data is so skewed toward the right, a sample size of 30 is most likely not large enough to approximate a normal sampling distribution. However, our sample size is actually 141 which is much larger than 30. Therefore, it may be possible that a t-test is valid in our scenario given that the skew isn't very strong. I cannot say that with 100% certainty, however, because the skew of the data might require a larger sample size. I would be more confident using a t-test if our sample size was larger but it is very possible that 141 samples is more than enough.

b. How would your answer to part a change if Mr. Lemmon selected dormitory rooms at random, then interviewed all occupants in the rooms he selected?

I am not very sure what is being changed by this new interviewing process. If by randomly selecting dormitory rooms means that the sample size is decreased, then it would make me less likely to use a t-test. If, however, he ends up sampling more students, I might feel a bit more confident with the t-test. Furthermore, by selecting students at random to ask them the question, assuming they all answer truthfully, I would feel more comfortable with the data than a self selected survey group. On the other hand, maybe students who are being asked the question rather than giving up the information voluntarily would feel the urge to lie in order to make themselves look like better students. A survey can have the advantage of anonymity which may make it more likely for students to tell the truth compared to if the student was asked directly. Finally, it also depends if the student is being asked alone, or in the presence of their roommate. The answer may not be truly independent if the students can be influenced by their roommate's answer. Considering all these factors, I think I would be less willing to trust the data and do a t-test to verify the hypothesis.

c. Provide an argument for why you should choose a 2-tailed test in this instance, even if you are hoping to demonstrate that MSU students skip more than 1 lecture per week.

The main reason to choose a 2-tailed test even though we only plan on demonstrating that MSU students skip more than 1 lecture per week, is to cover the bases. It is possible that the data behaves the opposite way than we expected. In that case, we cannot go back and change the side of our 1-tail test, or change into a 2-tailed test because that would be considered cheating and would abuse the trust of our readers. A 2-tailed test would allow us to prepare for such a scenario so that we can avoid cheating or at least cause suspicion from our readers.

d. Conduct the t-test using the t.test function and interpret every component of the results.

```
t.test(data$skipped, mu = 1, alternative = "two.sided")
```

```
##
##  One Sample t-test
##
## data:  data$skipped
## t = 0.83142, df = 140, p-value = 0.4072
## alternative hypothesis: true mean is not equal to 1
## 95 percent confidence interval:
##  0.8949445 1.2575377
## sample estimates:
## mean of x
##  1.076241
```

The result of this t-test tells us that our t statistic is 0.83142, our degree of freedom is 140, and our p-value is 0.4072. Because our p-value is much greater than 0.1 (ideally, we want less than 0.05), we cannot reject the null hypothesis. Another way to look at it is that 1 lies firmly within the 95% confidence interval given as (0.895, 1.258) which tells us that the null hypothesis cannot be rejected.

e. Show how you would compute the t-statistic and p-value manually (without using t.test), using the pt function in R.

The t-statistic can be found by the following equation:

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

```
mu <- mean(data$skipped)
sd <- sd(data$skipped)
n <- length(data$skipped)
x_bar <- 1

t <- (x_bar - mu)/(sd/sqrt(n))
abs(t)
```

```
## [1] 0.8314156
```

As shown, our t-statistic is 0.8314156. Our p-value would be the probability from negative infinity to -0.8314 combined with the probability from 0.8314 to positive infinity.

```
# our p-value would be double the probability from -infinity to -0.8314
2*pt(t,n-1)
```

```
## [1] 0.4071547
```

As shown, our 2-tailed p-value is 0.407 which matches what we got in part d.

f. Construct a 99% confidence interval for the mean number classes skipped by MSU students in a week.

To find the 99% confidence interval, we have to find the t-value for 0.5% (one for the left side of the distribution and one for the right side)

```
tleft <- qt(0.005, n-1)
tright <- qt(0.995, n-1)
CI = c(mu+tleft*sd/sqrt(n), mu+tright*sd/sqrt(n))
CI
```

```
## [1] 0.8367745 1.3157078
```

The 99% confidence interval is (0.8367745, 1.3157078)

g. Can you say that there is a 99% chance the population mean falls inside your confidence interval?

No we cannot, because the population mean is fixed. What we can say instead, is that there is a 99% chance our confidence interval contains the population mean. Or, we can say that if we calcuate many confidence intervals with many different samples, 99% of the confidence intervals will contain our population mean.