HW week 8

w203: Statistics for Data Science

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.

```
setwd("/Users/shanhe/Desktop/w203/Homework/Week 8/HW8")
load("GPA1.RData")
```

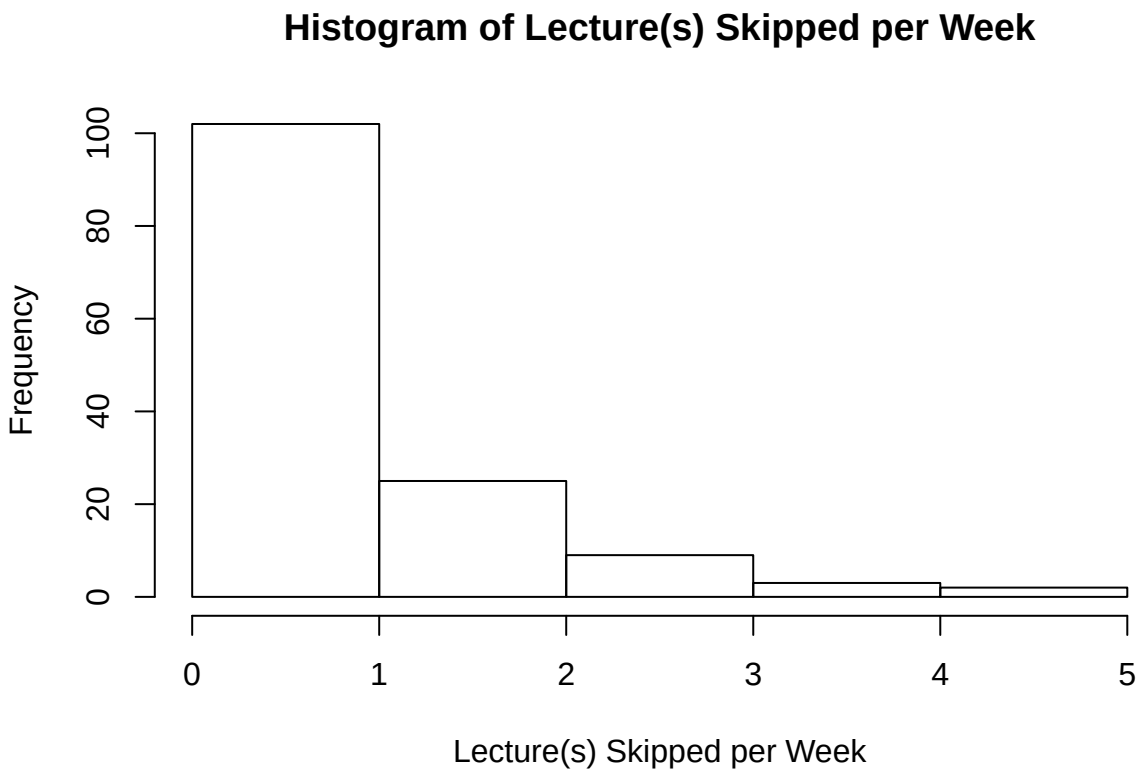
The skipped variable represents the average number of lectures each respondent skips per week. You are interested in testing whether MSU students skip over 1 lecture per week on the average.

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

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

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

```
hist(data$skipped, breaks = 5, main = 'Histogram of Lecture(s) Skipped per Week'
     , xlab = 'Lecture(s) Skipped per Week')
```



Typically, t-tests are used for small samples with an assumption of an approximately normal population distribution. Here, although the sample doesn't seem to be normally distributed, the large sample size, 141 (> 40), involkes Central Limit Theorem and hence validates the use of a t-test.

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?

In order to perform the t-test, it needs to be based on a random sample. Since there might be influences among roommates, this sample might not be independent (but coulld be identically distributed), which might undermine the validity of the t-test.

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.

Performing a one-tailed test requires a solid explanation on why the authors would treat a large observed difference in the unexpected direction no differently from a difference in the expected direction that was not strong enough to justify rejection of the null hypothesis. In this circumstance, large differences in both directions should have statistical significance.

Furthermore, if the sample mean is indeed larger than 1, a two-tailed test decreases a, the porbability of a Type I error.

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

```
t.test(data$skipped, mu = 1)
```

```
##
##  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
```

This t test was performed with the null hypothesis as that the true mean is equal to 1 and the alternative hypothesis as that the true mean is not equal to 1, a two-tailed test.

For our sample, we have a sample mean as 1.07 and a t statistic as 0.83, with 140 degrees of freedom, which yields a p-value of 0.41 in the t distribution. This p value can be interpreted as the possibility of getting a t statistic as extreme as ours (0.83) assuming the null hypotheis is true.

Moreover, a 95% Confidence Interval as (0.89, 1.26) means that if the mean in the null hypothesis is outside of this interval, we can reject the null hypothesis with 95% confidence.

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

```
t = (mean(data$skipped) - 1)/(sd(data$skipped) / sqrt(length(data$skipped)))
p_value = 2 * (1-pt(t, 140))
t
```

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

```
p_value
```

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

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

```
CI_99 = mean(data$skipped) + qt(c(0.005,0.995), 140)*(sd(data$skipped) / sqrt(length(data$skipped)))
CI_99
```

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

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

No; a 99% Confidence Interval means that if you repeat the calculation on multiple samples, 99% of the Confidence Intervals would capture the true population mean.