HW week 8

w203: Statistics for Data Science

Ted Phamds

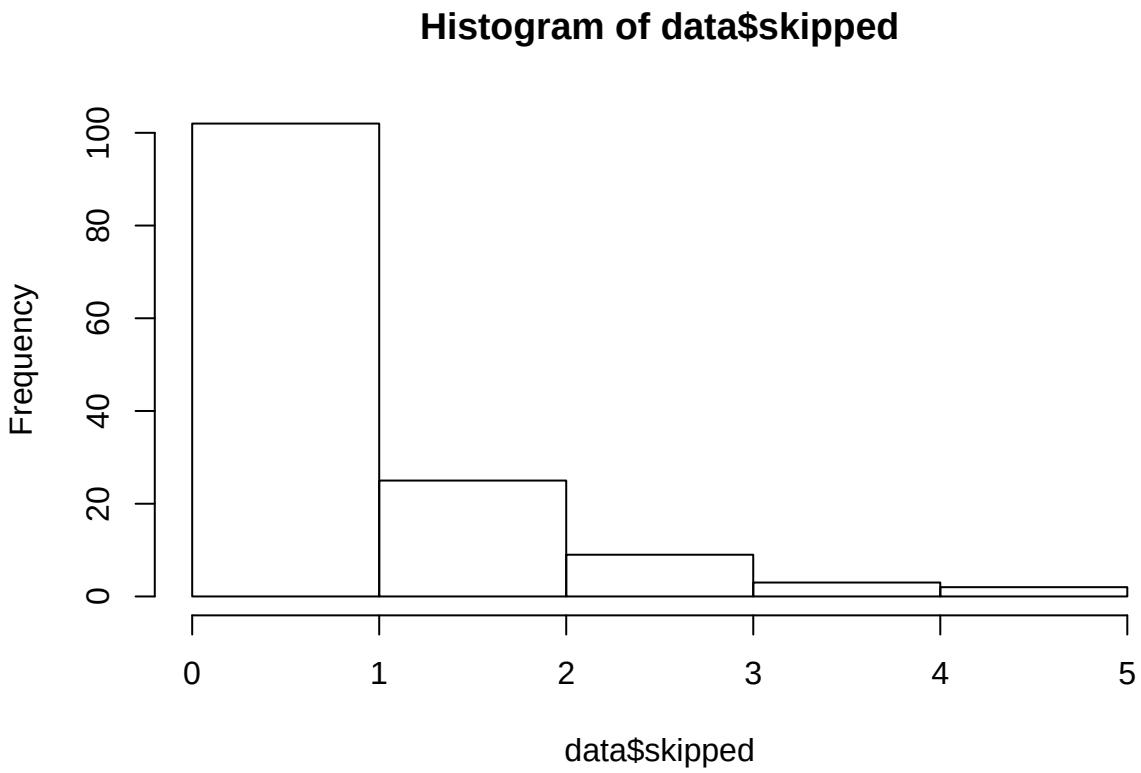
The file GPA1.RData contains data from a 1994 survey of MSU students. The survey was conducted by Christopher Lemmon, a former MSU undergraduate, and provided by Wooldridge.

```
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.

```
hist(data$skipped,breaks=5)
```



The sampling distribution is skewed to the right (not normal), and given that we don't know whether the sampling was done randomly we cannot apply a t-test in this scenario.

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?

Given that the sampling was done randomly and all occupants in the randomly seletected rooms were interviewed (meaning selected sample = actual sample), we can apply a t-test assumming that the number of skipped class follow a normal distribution. The assumption is reasonable because students inevitably skip class but they also do not want to fail. It's also reasonable to assume that the roommates are most likely have different class schedules hence the probability of all of them skipping classes at the same time can be negligible.

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.

We don't know which direction the # of lectures students skip will go, more or fewer than 1. So a 2-tailed test is more appropriate. Doing one-tail test would be cheating and biased.

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
```

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

```
#compute the sample stdev
s = sd(data$skipped)

#compute the sample mean
mu = mean(data$skipped)

#compute t statistic
t = (mu-1)/s*sqrt(141)

#compute p-value for a two-tailed t-test

p_value= (1-pt(t,140))*2

p_value
```

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

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

```
alpha = 0.99
alpha_2 = alpha + (1-alpha)/2
t_99 = qt(alpha_2,140)
upper_bound = mu + abs(t_99*s/sqrt(141))
lower_bound = mu - abs(t_99*s/sqrt(141))
```

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. But we can say that 99% of similarly constructed intervals contain the population mean.