HW week 8

w203: Statistics for Data Science

Tako Hisada

10/29/2017

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.

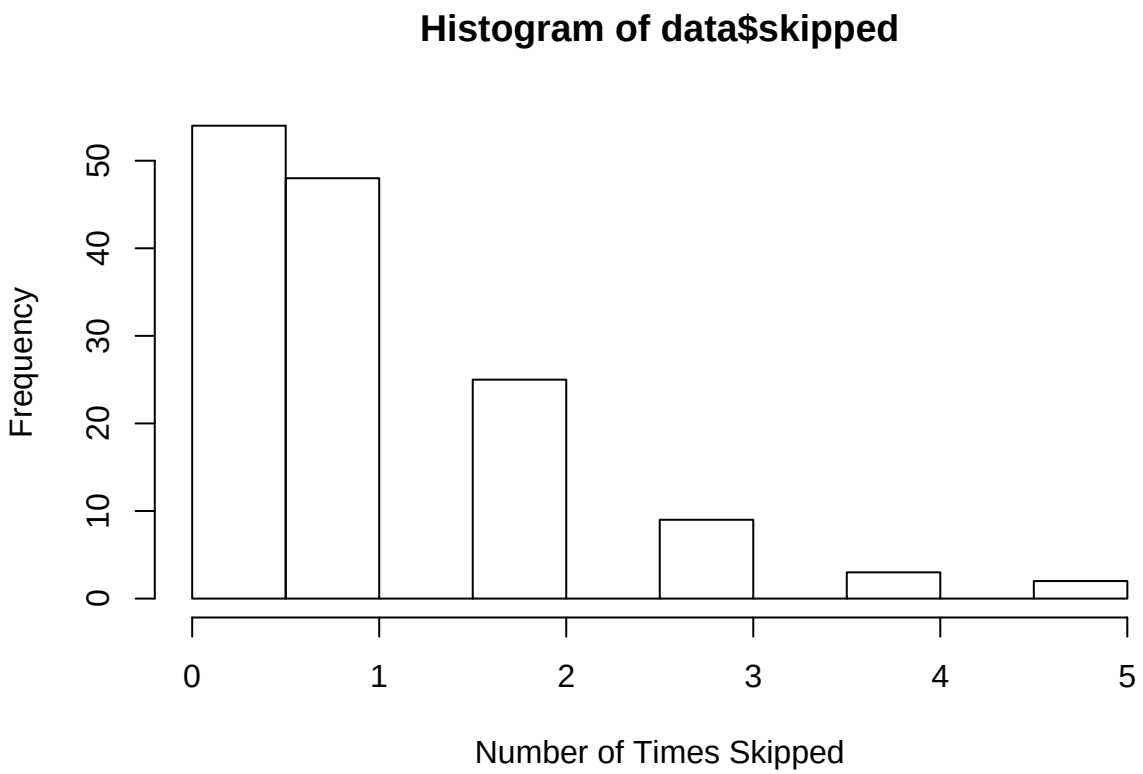
```
summary(data$skipped)
```

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

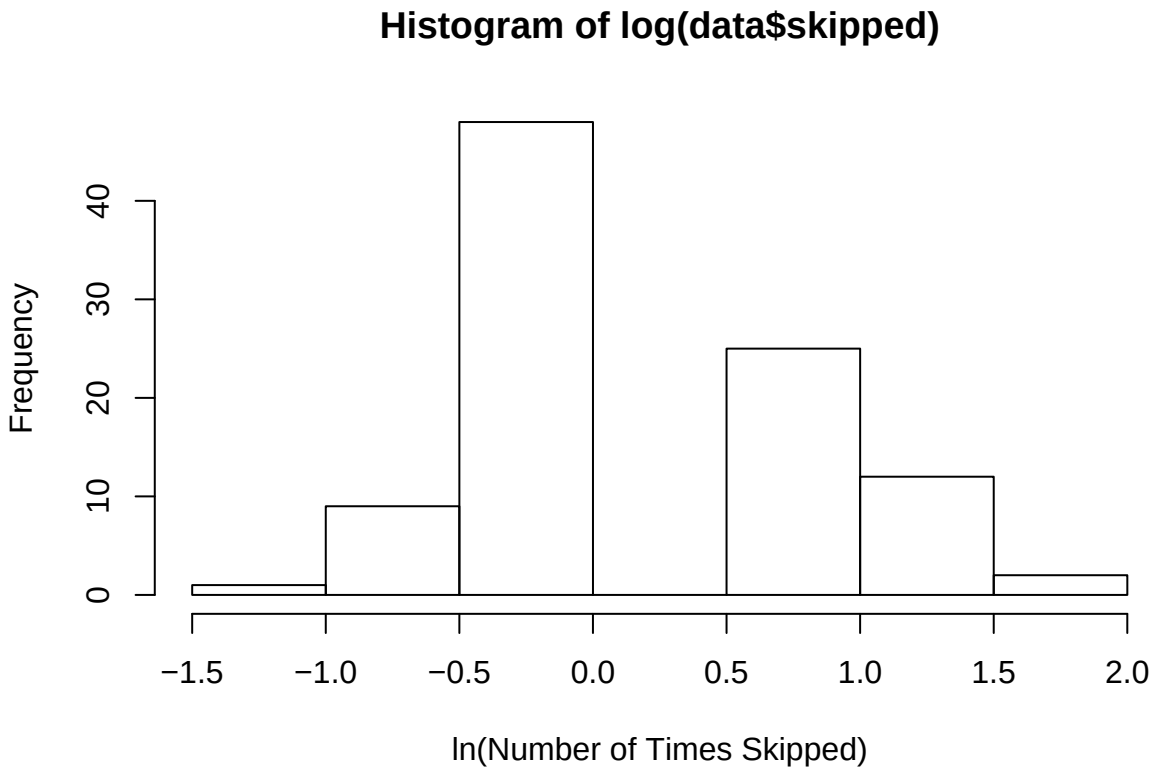
```
length(data$skipped[!is.na(data$skipped)])
```

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

```
hist(data$skipped, xlab="Number of Times Skipped", ylab="Frequency")
```




```
hist(log(data$skipped), xlab="ln(Number of Times Skipped)", ylab="Frequency")
```



The sample size 141 is larger than 30 and is sufficiently large for a z-test. We however do not have the population μ nor σ as the survey only has responses from 141 students who were attending MSU in 1994. This makes therefore the use of a t-test more appropriate than z-test in this case.

Looking at the histogram, the distribution of the skipped variable is positively skewed. Taking a natural log of the vector makes the distribution more normalized. However, since our sample size of 141 is large enough, we should be able to carry out a t-test with the original sample values.

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?

It depends on how large the sample size Mr. Lemmon chooses. In the case of $t < 30$, a t-test may be valid. However, there could potentially be a correlation between the fact a given student living in a dormitory and his/her frequency of skipping lectures in which case the data may not serve as a valid random sample of the MSU student population.

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 use of a 2-tailed test is generally prepferred to a 1-tailed test unless you are absolutely sure that it makes sense for a given H_α. In this case, as the distribution we are working with is highly positively skewed, using a 1-tailed test could yield an undesirable result. Using a upper-tailed test is likely to result in a higher chance of triggering a type I error. Likewise, using a lower-tailed test would result in a higher chance of triggering a type II 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
```

t = 0.83142 - test statistic computed for the given sample with a t distribution with 140 df

df = 140 - This is derived by n = 141 - 1.

p-value = 0.4072. $0.4072 > 0.05$

This means that we fail to reject H_0 and the average frequency of MSU students skipping lectures ! = 1 per week.

95 percent confidence interval (0.8949445, 1.2575377) - This means that if this were to be conducted repeatedly, the interval $(0.8949445, 1.2575377)$ will contain the true μ of frequnecy of MSU students skipping lectures 95% of the time.

Sample estimates: mean of x 1.076241 - This matches the mean computed from data$skipped as shown below. The mean suggests that MSU students skip more than 1 lecture per week on average.

```
mean(data$skipped)
```

```
## [1] 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.

```
mu <- 1
xbar <- mean(data$skipped)
sd <- sd(data$skipped)
n <- length(data$skipped)
v <- n-1
(t <- (xbar - mu)/(sd/sqrt(n)))
```

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

t-value: 0.83142

```
(p <- 2*(1-pt(t, v)))
```

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

p-value: 0.4072

The values match what we got from t.test in d.

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

```
cl = .99
tstat = qt((1-cl)/2, df=v)
(ci_l = xbar + tstat * sd/sqrt(n))
```

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

```
(ci_r = xbar - tstat * sd/sqrt(n))
```

```
## [1] 1.315708
```

99% confidence interval = (0.8367745, 1.315708)

```r
t.test(data$skipped, mu=1, conf.level = .99)
```

```
##
##  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
## 99 percent confidence interval:
##  0.8367745 1.3157078
## sample estimates:
## mean of x
##  1.076241
```

Matches the values derived manually above.

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

No. This means that if this were to be conducted repeatedly, the interval $(0.8367745, 1.315708)$ will contain the true μ of frequnecy of MSU students skipping lectures 99% of the time.