HW 7

Testing Attendance

Data: 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.

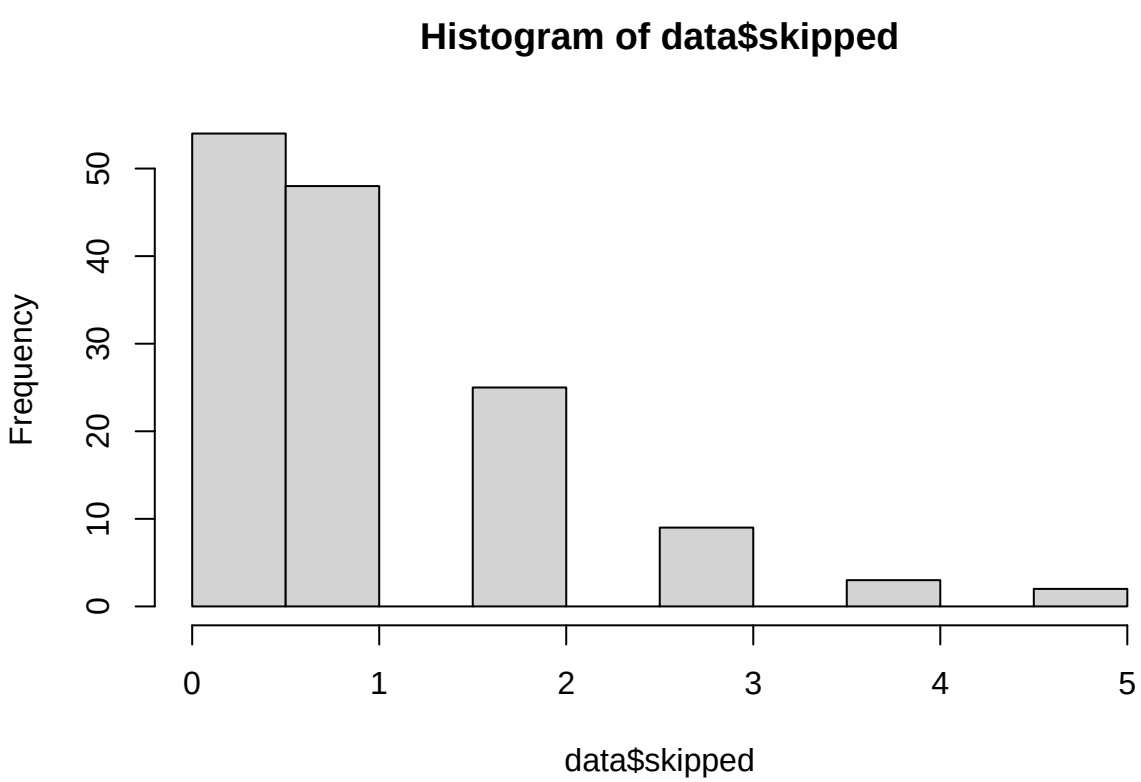
The variable skipped represents the average number of lectures each respondent skips per week. You are interested in testing whether MSU students skip over 1 lecture per week on the average.

1. Examine the skipped variable and argue whether or not a t-test is valid for this scenario. (3 points)

Conditions for a T-test

- Independent and Identically Distributed
  Without more information about how this sample was conducted, it is hard to determine if the data is i.i.d. The other columns of this dataset (gender, class year, etc) do not seem to match what could reasonably be assumed to be their population averages. This could indicate that the sample was not random. This condition is likely not met.
- Data is not skewed

```
load("gpa1.RData")
hist(data$skipped)
```



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

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

There is considerable skew to the data, but for the large number of observations this assumption can be considered satisfied.

- Metric Data
  There are clear differences and order between skipping 1 class, 2 classes, 3 classes, etc. This assumption is satisfied.
- Large number of observations
  141 observations is much large than the rule of thumb of 30 observations. This assumption is satisfied.

A T-Test is most likely valid for this scenario

2. 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? (3 points)

This information provides more information to evaluate the i.i.d. assumption and casts doubt on its validity. There is likely some dependence between members within the same dorms, thus causing a loss of independence.

3. 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. (3 points)

We believe that the data can vary on both sides of the mean. Even though we are hoping to demonstrate that MSU students skip more than 1 lecture per week, we recognize that the population may vary to be less than or greater than this value.

4. Conduct the t-test using the t.test function and interpret every component of the results. (3 points)

```r
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
```

I ran a t-test to generate a 95% confidence interval that the sample mean is equal to the popuation mean. The output determined that we can be 95% certain that the population mean falls between .895 and 1.258 skipped classes per week. This sample of 141 observations has 140 degrees of freedom and a sample mean of 1.076 skipped classes per week.

5. Show how you would compute the t-statistic and p-value manually (without using t.test), using the pt function in R. (3 points)

```r
skip = data$skipped

mu = mean(skip)
paste("Sample Mean: ", mu)
```

```
## [1] "Sample Mean:  1.07624113475177"
```

```r
sigma = sd(skip)
paste("Standard Deviation: ", sigma)
```

```
## [1] "Standard Deviation:  1.08888182423667"
```

```r
t_critical = qt(1-.025, df = length(skip)-1)
paste("Critical Value: ", t_critical)
```

```
## [1] "Critical Value:  1.97705371965709"
```

```r
t = (mu-1)/(sigma / sqrt(length(skip)))
paste("t statistic: ", t)
```

```
## [1] "t statistic:  0.831415581558769"
```

```r
p_vall = pt(.025,df=140)
p_valr = pt(.975,df=140)
paste(p_vall, p_valr)
```

```
## [1] "0.509954718764013 0.83437903456321"
```