Unit 8 Homework

W203 - Section 4

Brad Andersen

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.

Context: 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.

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

Given the abovementioned context, we will assert that our null hypothesis is that *the true mean is equal to 1 (lecture skipped per week)*. Because we are interested in testing whether students skip more than 1 lecture per week, we will assert that our alternative hypothesis is that *the true mean is greater than 1 (lecture skipped per week)*.

Loading the data to examine the *skipped* variable, a summary of values is helpful in determining whether a t-test is valid:

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

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

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

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

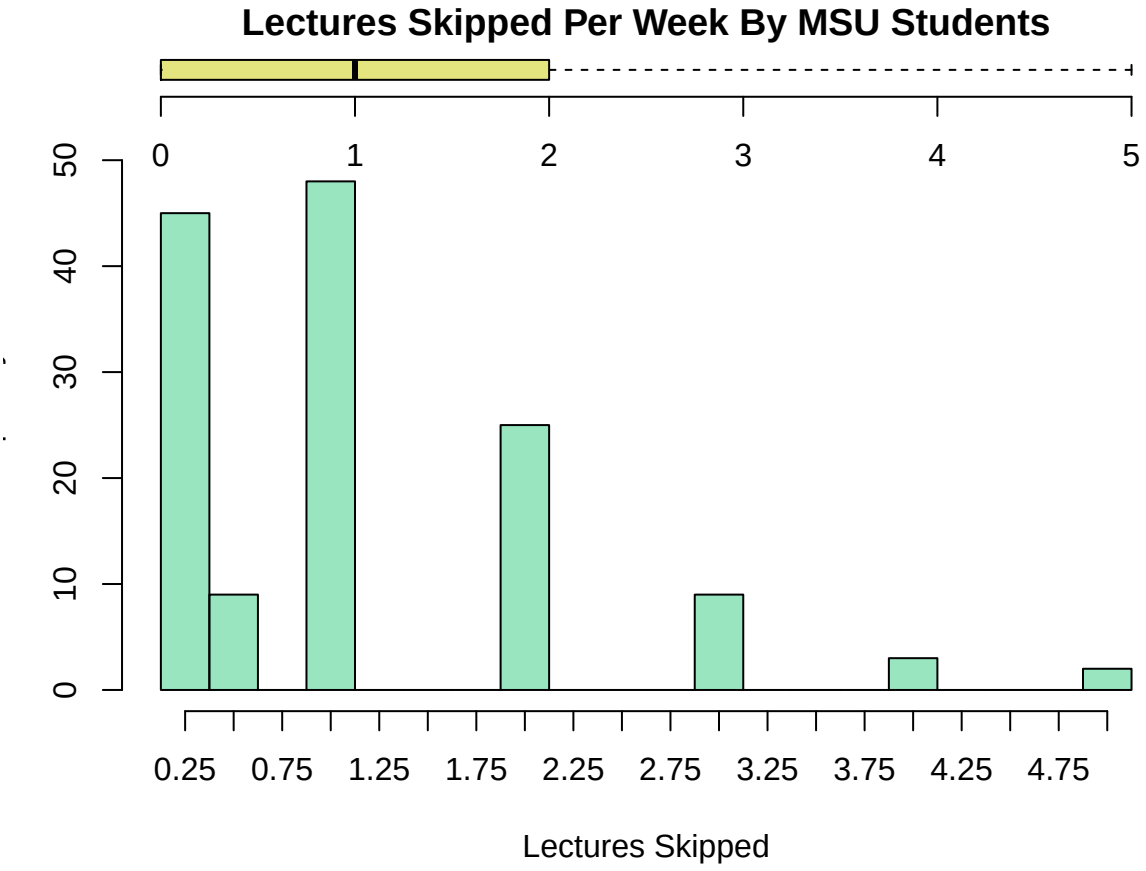
```
sd(data$skipped)
```

```
## [1] 1.088882
```

Because of the values in the *skipped* variable's summary (specifically the close proximity of the median to the minimum value when compared with the maximum value, as well as the sample's mean and median values being in close proximity to the mean value used in the null hypothesis), a visual representation of the data is helpful to better understand the skew:

```
layout(mat = matrix(c(1,2),2,1, byrow=TRUE),  height = c(1,8))
par(mar=c(0, 3.1, 1.1, 2.1))
boxplot(
  data$skipped,
  ylab=FALSE,
  horizontal=TRUE,
  main="Lectures Skipped Per Week By MSU Students",
  col=rgb(0.8,0.8,0,0.5),
  frame=FALSE
)
```

```r
par(mar=c(4, 3.1, 1.1, 2.1))
skipped_hist <- hist(
  data$skipped,
  breaks=seq(0,5,by=.25),
  axes=F,
  ylim=c(0,50),
  xlab="Lectures Skipped",
  main=NULL,
  col=rgb(0.2,0.8,0.5,0.5)
)
axis(side=1, at=skipped_hist$mids, labels=seq(0.25, 5,by=.25))
axis(side=2)
```



The *skipped* variable data is far from normal, is right-skewed, and has values within the first three quartiles within one standard deviation from the mean. Because the number of subjects reporting *skipped* values ($n = 141$) is sufficiently large (i.e. $n >= 30$), it could be tempting to assume that the sample distribution is largely representative of the population distribution. However, because we don't know the values of the population's *skipped* variable's mean and standard deviation, a t-test is valid:

```r
t.test(data$skipped, mu=1, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  data$skipped
```

```
## t = 0.83142, df = 140, p-value = 0.2036
## alternative hypothesis: true mean is greater than 1
## 95 percent confidence interval:
##  0.9244027       Inf
## sample estimates:
## mean of x
##  1.076241
```

```
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
```

2.0: 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?

Whether or not a student lives in a dormitory is not a variable collected by Lemmon. Therefore, the population of subjects included in his test would narrow from *MSU students in 1994* to *MSU students in 1994 living in dormitories whose rooms were selected at random.*

The fact that a subject lived in a dormitory could influence whether or not – and how frequently – the subject skipped lectures. More information would then be known about the subject, as dormitory living could affect one's decision to attend or not attend lectures. For example, information such as the proximity of the dormitories to the locations of lectures, or the percentage of MSU students living in dormitories compared to other residences summcould influence a subject's likelihood to attend lectures.

The null hypothesis would need to change to *the true mean is equal to 1 (lecture skipped per week by dormitory resident).* The number of lectures skipped per week would not be considered conditional upon whether or not a subject lived in a dormitory, because the sample was known to consist of only dormitory residents.

3.0: 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.

When performing one- and two-tailed t-tests with the null hypothesis being *the true mean is equal to 1 (lecture skipped per week)*, probabilities that we should *not* reject the null hypothesis are calculated as 79.64% and 20.36% (in reference to *less-than* and *greater-than* one-tailed t-tests) and 40.72% (in reference to a two-tailed t-test). Therefore, there are strong probabilities that we should not reject the null hypothesis.

However, of import are that the sample mean and the mean assumed in the null hypothesis (1.076241 and 1, respectively) are close, given the range of the data in the sample, and that the sample data's distribution is heavily skewed. A two-tailed test would be most appropriate to observe potential effects on either side of the mean.

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

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

```
##
##  One Sample t-test
##
## data:  data$skipped
## t = 0.83142, df = 140, p-value = 0.2036
## alternative hypothesis: true mean is greater than 1
## 95 percent confidence interval:
##  0.9244027       Inf
## sample estimates:
## mean of x
##  1.076241
```

- *t = 0.83142* - The number of the sample standard deviations (s) that the mean assumed in the null hypothesis is away from the sample's mean.
- *df = 140* - Degrees of freedom, or the number of elements (minus one) in the sample.
- *p-value = 0.2036* - The probability that one should not reject the null hypothesis given the sample data. The higher the p-value, the greater the probability of not rejecting the null hypothesis.
- *alternative hypothesis: true mean is greater than 1* - A description of the effect that we are attempting to demonstrate, and from which the null hypothesis is derived.
- *95 percent confidence interval: 0.9244027, Inf* - The range of values in which a sample mean will lie given that the sample was retrieved from the population in an identical manner. One can be confident that these results can be repeated 95 out of 100 times. The 95% confidence interval is based upon a significance value of 0.05.
- *mean of x: 1.076241* - The sample mean.

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

```
tstat <- (mean(data$skipped) - 1) / (sd(data$skipped) / sqrt(length(data$skipped)))
pt(tstat, length(data$skipped) - 1, lower.tail = FALSE)
```

```
## [1] 0.2035773
```

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

```
ttest_out <- t.test(data$skipped, mu=1, alternative="greater", conf.level = 0.99)
paste("Confidence interval: ", ttest_out$conf.int[1], "-", ttest_out$conf.int[2])
```

```
## [1] "Confidence interval:  0.860444582452425 - Inf"
```

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

More accurately stated, there is a 99% chance that a sample of the population's *skipped* data obtained in the same manner will have a mean value within the range of the confidence interval. In other words, one can expect to receive the same results in 99 out of 100 samples.