Stat Methods I - Homework 2

Justin Reising

September 25, 2018

1) A random sample of 796 teenagers revealed that in this sample, the mean number of hours per week of TV watching was 13.2, with a standard deviation of 1.6. Find (AND INTERPRET) a 90% confidence interval for the true mean weekly TV-watching time for teenagers. Why can we use a t-CI procedure in this problem?

Note, a $100(1-\alpha)\%$ Confidence interval for μ is given by,

$$\bar{Y} \pm t_{1-\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}$$

```
n<-796
y_bar<- 13.2
s<-1.6
t_stat<- qt(0.9,n-1)
CI_Lowerbound<- y_bar-t_stat*(s/sqrt(n))
CI_upperbound<- y_bar+t_stat*(s/sqrt(n))
CI_Lowerbound
```

```
## [1] 13.12726
```

```
CI_upperbound
```

```
## [1] 13.27274
```

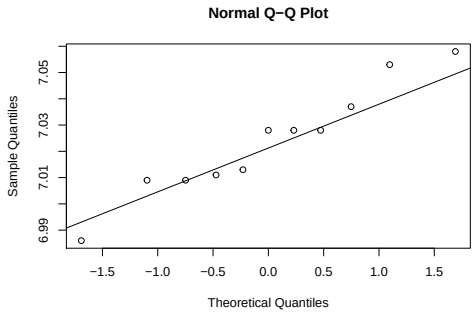
With 90% confidence, the true mean number of hours per week of TV watching by teenagers is between 13.127 hours and 13.273 hours.

According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. So when we know the standard deviation of the population, we can compute a z-score, and use the normal distribution to evaluate probabilities with the sample mean. But, since we do not know the standard deviation of the population, we invoke the use of the distribution of the t statistic. Also note that with a sample size this large, the t distribution is practically identical to the normal distribution.(Assuming the distribution of this sample is normal!)

2) An engineer wants to calibrate a pH meter. She wants to measure the pH in 14 neutral substances (pH = 7.0), obtaining the following data:

```
ph_data<- c(6.986,7.009,7.028,7.028,7.009,7.053,7.028,7.011,7.037,7.058,7.013)
```

(A) Use a graph and a formal test ($\alpha = 0.05$) to determine whether the assumption of normality for these data is reasonable. For the formal test, make sure you state the null and alternative hypotheses, p-value, and conclusion.



Note that the qqplot above indicates that the data is relatively normal. But to verify our assumptions, we can use the Shapiro Wilks Test to formally assess the normality of our data since it is of relative small sample size. Below we see that the p-value of the Shapiro-Wilk Test as $0.6783 > \alpha = 0.05$, which indicates that our data is normal and we can proceed with the one-sample t-test of hypotheses:

```
shapiro.test(ph_data)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ph_data
## W = 0.95268, p-value = 0.6783
```

```
t.test(ph_data, alternative = "two.sided", mu=7,conf.level = 0.95)
```

```
##
##  One Sample t-test
##
## data:  ph_data
## t = 3.7387, df = 10, p-value = 0.003855
## alternative hypothesis: true mean is not equal to 7
## 95 percent confidence interval:
##  7.009550 7.037723
## sample estimates:
## mean of x
##  7.023636
```

$$H_0 : \mu_0 = 7 \quad H_1 : \mu_0 \neq 7 \quad \text{p-value} = 0.003855$$

Our conclusion is that With 95% confidence, there is statistically significant evevidence that the true measure of the pH level of the 14 substances is between 7.0096 and 7.0377.

3) Suppose a sample of 10 types of compact cars reveals the following one-day rental prices (in dollars) for Hertz and Thrifty:

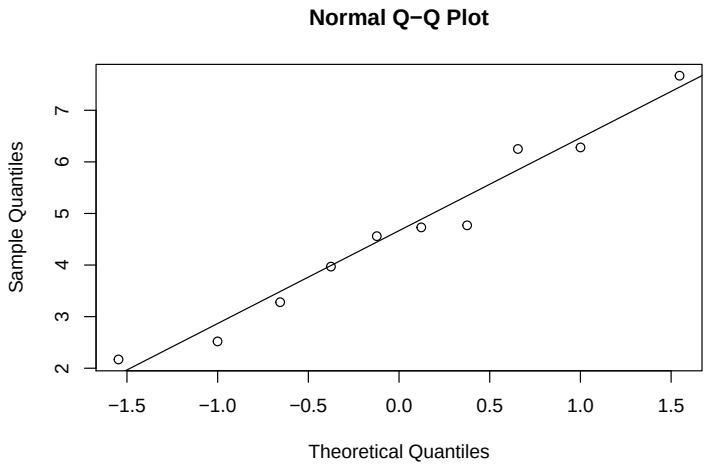
```
Hertz<-c(37.16,14.36,17.59,19.73,30.77,26.29,30.03,29.02,22.63,39.21)
Thrifty<-c(29.49,12.19,15.07,15.17,24.52,22.32,25.3,22.74,19.35,34.44)
```

(A) Explain why this is a paired problem.

This is a paired problem because we are comparing samples between two entities where each data point is the rental price for a type of compact car for each of the companies.

(B) Use a graph to determine whether the assumption of normality is reasonable.

```
diffs<- Hertz - Thrifty
qqnorm(diffs)
qqline(diffs)
```



From the qqplot above, we can see that the differences between Hertz and Thrifty appear to be normal as the points tend to be linear.

(C) Test (at a 0.05 significance level) with a t-test whether Thrifty has a lower true mean rental rate than Hertz. What is the conclusion of the test in the context of the problem?

$$H_0 : \mu_{diff} = 0 \quad H_1 : \mu_{diff} > 0$$

```
t.test(diffs, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  diffs
## t = 8.3756, df = 9, p-value = 7.657e-06
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  3.608852      Inf
## sample estimates:
## mean of x
##      4.62
```

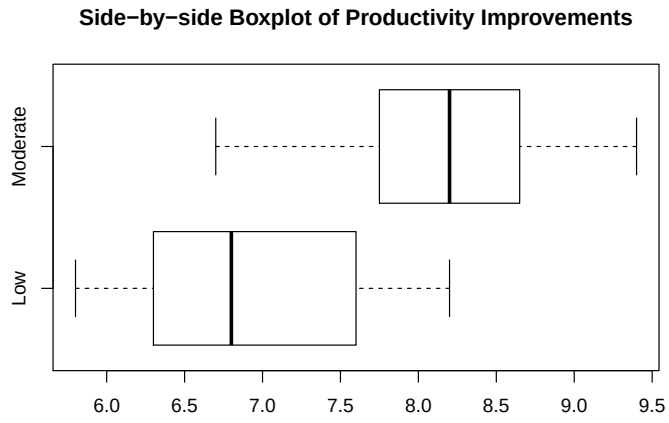
With a p-value of 0.0000007, we reject the null hypothesis of no difference between the true means of the car rental companies and conclude with 95% confidence that there is statistically significant evidence to indicate that Thrifty has a lower true mean rental rate than Hertz.

4) Examine the data in Problem 16.7 on page 723 of the PDF textbook found on D2L. We will only deal with the data on the first two line ("Low" and "Moderate").

```
Low<- c(7.6,8.2,6.8,5.8,6.9,6.6,6.3,7.7,6)
Moderate<-c(6.7,8.1,9.4,8.6,7.8,7.7,8.9,7.9,8.3,8.7,7.1,8.4)
```

(A) Prepare side-by-side box plots for the two samples. Do the spreads seem to differ accross samples? Also, perform a formal test ($\alpha = 0.05$) to test for equal variances. For the formal test, make sure you state the null and alternative hypotheses, p-value, and conclusion.

```
boxplot(Low,Moderate, names = c("Low","Moderate"),horizontal = T)
title(main = "Side-by-side Boxplot of Productivity Improvements")
```



Note that the variances appear to be fairly close to being equal, but we will formally test this claim using the Levene Test with the following:

$$H_0 : \sigma_L = \sigma_M \quad H_1 : \sigma_L \neq \sigma_M$$

```
#library(lawstat)
comp<- c(rep(0,length(Low)),rep(1,length(Moderate)))
combined<-c(Low,Moderate)
leveneTest(combined, comp)
```

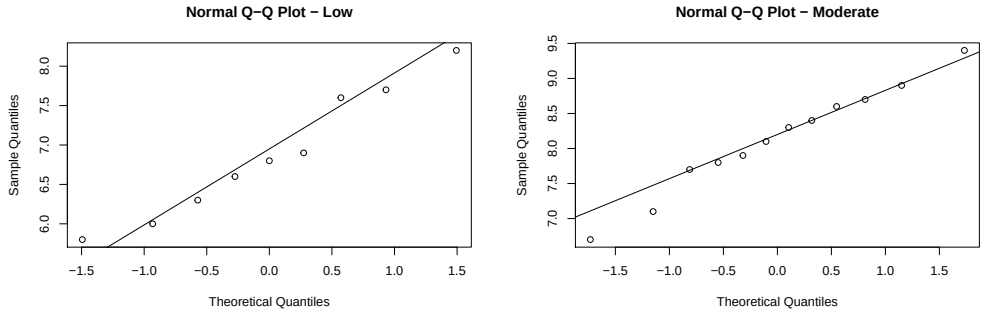
```
## Warning in leveneTest.default(combined, comp): comp coerced to factor.
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.0608 0.8079
##       19
```

By the Levene Test we can see that with a p-value of $0.8079 > 0.05$, we fail to reject the null hypothesis and the variances between the groups "Low" and "Moderate" have statistically significant homogeniety.

(B) Use a graph(s) to determine whether the assumption of normality is reasonable.

```
qqnorm(Low, main = "Normal Q-Q Plot - Low")
qqline(Low)
qqnorm(Moderate, main = "Normal Q-Q Plot - Moderate")
qqline(Moderate)
```



From the qqplots above, we can see that both data "Low" and "Moderate" for economic imporvements are normal, so we can proceed with t-test to compare the means of the two groups.

(C) Test (at a 0.05 significance level) with a t-test whether the firms rated "Moderate" have a significantly higher mean productivity improvement than those rated "Low". What is the conclusion of your test in the context of the problem?

$$H_0 : \mu_L = \mu_M \quad H_1 : \mu_L < \mu_M$$

```
t.test(Moderate, Low, var.equal = T, alternative = "greater")
```

```
##
##  Two Sample t-test
##
## data:  Moderate and Low
## t = 3.6437, df = 19, p-value = 0.0008638
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.6597289       Inf
## sample estimates:
## mean of x mean of y
##  8.133333  6.877778
```

```
t.test(Moderate, Low, var.equal = T)$conf.int
```

```
## [1] 0.5343388 1.9767723
## attr(,"conf.level")
## [1] 0.95
```

From the t-test prcoedure, we can conclude with 95% confidence that there is statistically significant evidence indicating the true mean of the firms rated "Moderate" is higher than the true mean of firms rated "Low" and that the difference between the two groups is between 0.534 and 1.977.

(D) Find and interpret a 90% confidence interval for the difference in the mean productivity improvement between firms rated "Moderate" and those rated "Low".

```
t.test(Moderate, Low, var.equal = T, conf.level = 0.9)
```

```
##
##  Two Sample t-test
##
## data:  Moderate and Low
## t = 3.6437, df = 19, p-value = 0.001728
## alternative hypothesis: true difference in means is not equal to 0
## 90 percent confidence interval:
##  0.6597289 1.8513822
## sample estimates:
## mean of x mean of y
##  8.133333  6.877778
```

From the t-test prcoedure, we can conclude with 90% confidence that there is statistically significant evidence indicating the true mean of the firms rated "Moderate" is higher than the true mean of firms rated "Low" and that the difference between the two groups is between 0.6597 and 1.8514.

(E) Perform an appropriate nonparametric test (at 0.05 significance level) to test whether the firms rated "Moderate" have a significantly higher median productivity improvement than those rated "Low". What is the conclusion of your test in the context of the problem?

$$H_0 : \eta_M = \eta_L \quad H_1 : \eta_M > \eta_L$$

```
wilcox.test(Moderate, Low, conf.int = T, conf.level = 0.95, alternative = "greater")
```

```
## Warning in wilcox.test.default(Moderate, Low, conf.int = T, conf.level =
## 0.95, : cannot compute exact p-value with ties
```

```
## Warning in wilcox.test.default(Moderate, Low, conf.int = T, conf.level =
## 0.95, : cannot compute exact confidence intervals with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Moderate and Low
## W = 95.5, p-value = 0.00178
## alternative hypothesis: true location shift is greater than 0
## 95 percent confidence interval:
##  0.6999873       Inf
## sample estimates:
## difference in location
##               1.299905
```

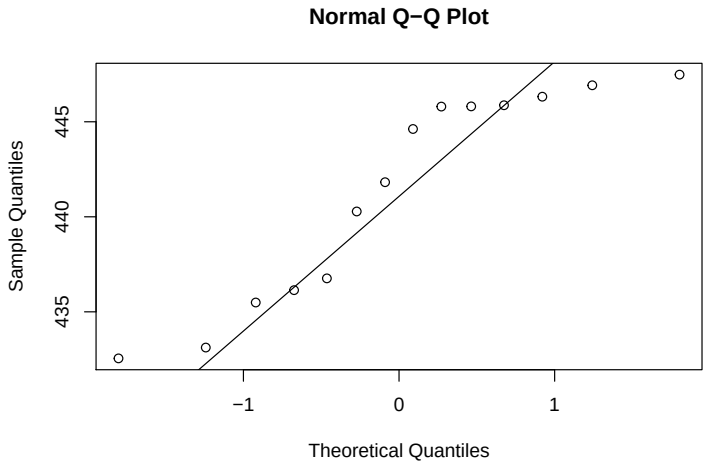
With a p value of 0.00178, we reject the null hypothesis and conclude that with 95% confidence, there is statistically significant evidence to suggest that the firms rated "Moderate" have a higher median productivity improvement than those rated "Low".

5) A cereal company claims its boxes contain 445 grams of cereal. A random sample of 15 boxes prduces the following measurements:

```
cereal<-c(446.92,447.48,436.14,441.82,445.8,435.49,445.87,445.81,440.28,433.12,
          436.76,432.55,446.32,444.62)
```

(A) Use a graph to determine whether the assumption of normality is reasonable.

```
qqnorm(cereal)
qqline(cereal)
```




```
shapiro.test(cereal)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cereal
## W = 0.86361, p-value = 0.03417
```

There appears to be some abnormal deviations in the Q-Q Plot above which may indicate that the cereal data is not normal, and with the formal Shapiro-Wilk test, we can verify that the data is not normal. Then we will proceed with a Nonparametric test, namely the Sign Test.

(B) Using an appropriate nonparametric test (at 0.05 significance level), determine whether the center of the distribution of cereal weights is 445 grams. What is the conclusion of your test in the context of the problem?

$$H_0 : \eta = 445 \quad H_1 : \eta \neq 445$$

```
quantile.test(cereal,eta = 445, quantile=0.5,alternative="two.sided")$p.value
```

```
## [1] 0.7905273
```

```
quantile.interval(cereal,quantile=0.5,conf.level=.95)$interval
```

```
## [1] 435.49 446.32
```

With a p-value of 0.79, we fail to reject the null hypothesis claim that true median of grams of cereal is 445 and conclude that with 95% confidence, there is statistically significant evidence to support that the true median grams of cereal in each box between 435.49 and 446.32 grams.