STAT 515 hw 10

Two-sample testing, comparative experiments and ANOVA

Attach a sheet with any R plots and R code printed on it. You may write out your other answers by hand if you want. Just try to make it easy for me grade!!

1. In a study of how different types of greetings transmit bacteria, a sterile glove was donned, dipped in bacteria, and then used in a handshake, a high five, or a fist bump with a hand wearing a sterile glove. Afterwards the bacteria on the sterile glove were counted. These data come from exercise 9.24 of McClave and Sincich (2016). Read the data into R using

```
handshake <- c(131,74,129,96,92)
highfive <- c(44,70,69,43,53)
fistbump <- c(15,14,21,29,21)
```

   It is of interest to study differences among the mean bacteria counts expected from these types of greeting, which we may denote by $\mu_{\text{handshake}}$, μ_{highfive}, and μ_{fistbump}.

   (a) Use R to get a 99% confidence interval for $\mu_{\text{handshake}} - \mu_{\text{highfive}}$ assuming $\sigma^2_{\text{handshake}} = \sigma^2_{\text{highfive}}$. Use the command

```
t.test(handshake,highfive,conf.level=.99,var.equal=TRUE)
```

   The output is

```
        Two Sample t-test

data:  handshake and highfive
t = 3.8738, df = 8, p-value = 0.004716
alternative hypothesis: true difference in means is not equal to 0
99 percent confidence interval:
  6.503594 90.696406
sample estimates:
mean of x mean of y
    104.4      55.8
```

   so the 99% confidence interval for $\mu_1 - \mu_2$ is $(6.503594, 90.696406)$.

   (b) Use R to get a 90% confidence interval for $\mu_{\text{highfive}} - \mu_{\text{fistbump}}$ under the assumption that $\sigma^2_{\text{highfive}} \neq \sigma^2_{\text{fistbump}}$.

   Use the command

```
t.test(highfive,fistbump,conf.level=.90,var.equal=FALSE)
```

   This gives the output

```
        Welch Two Sample t-test

data:  highfive and fistbump
t = 5.5546, df = 5.6067, p-value = 0.0018
alternative hypothesis: true difference in means is not equal to 0
90 percent confidence interval:
 23.1168 48.4832
sample estimates:
mean of x mean of y
     55.8      20.0
```

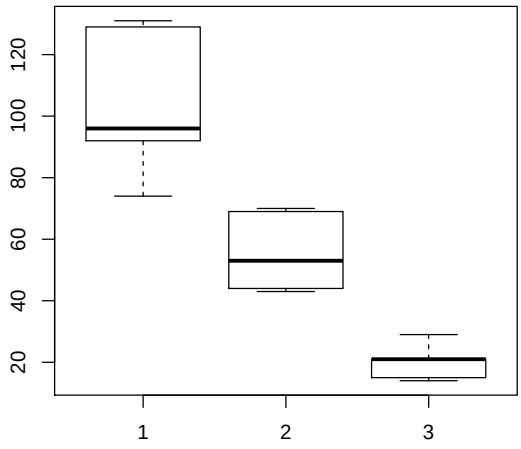
so the 90% confidence interval for $\mu_1 - \mu_2$ is $(23.1168, 48.4832)$.

(c) Use the command

`boxplot(handshake,highfive,fistbump)`

to get boxplots of the data. Turn in this plot.

The plot is the following:



(d) Based on the boxplots, comment on whether you should assume $\sigma^2_{\text{highfive}} = \sigma^2_{\text{fistbump}}$.

It does not seem reasonable to assume equal variances because the interquartile ranges—the heights of the rectangle parts of the boxplots—are very different.

(e) Use R to test the hypotheses

$$H_0\text{: } \mu_{\text{handshake}} - \mu_{\text{fistbump}} = 0 \text{ versus } H_1\text{: } \mu_{\text{handshake}} - \mu_{\text{fistbump}} \neq 0$$

at the $\alpha = .05$ significance level. You must decide whether to put `var.equal=TRUE` or `var.equal=FALSE`. Say whether you reject H_0 and why based on the output.

Use the command

```
t.test(handshake,fistbump,var.equal=FALSE)
```

We should put `var.equal=FALSE` because the variances do not appear to be equal (from looking at the boxplots). The above command gives the output

```
        Welch Two Sample t-test

data:  handshake and fistbump
t = 7.395, df = 4.4665, p-value = 0.001145
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  53.97547 114.82453
sample estimates:
mean of x mean of y
    104.4      20.0
```

(f) Suppose an investigator wanted to do an ANOVA for these data, where `handshake`, `highfive`, and `fistbump` are considered treatments. Which one of the ANOVA assumptions does not appear to be satisfied for these data?

The assumption of equal variances in all of the treatment groups does not appear to be satisfied.

2. Execute the commands below in R to read in some data. The data points are the number of crashes (average per year) due to drivers' running red lights at thirteen intersections before and after the installation of red light cameras. These data come from exercise 9.53 of McClave and Sincich (2016). Read the data into R with the commands

```
before <- c(3.6,.27,.29,4.55,2.6,2.29,2.4,0.73,3.15,3.21,.88,1.35,7.35)
after <- c(1.36,0,0,1.79,2.04,3.14,2.72,0.24,1.57,0.43,0.28,1.09,4.92)
```

It is of interest to see whether the installation of a camera reduces the number of crashes due to running red lights.

(a) Compute the differences in the numbers of crashes at the intersections:

```
diff <- before - after
```

Give the mean before-minus-after difference from the sample.

The code

```
before <- c(3.6,.27,.29,4.55,2.6,2.29,2.4,0.73,3.15,3.21,.88,1.35,7.35)
after <- c(1.36,0,0,1.79,2.04,3.14,2.72,0.24,1.57,0.43,0.28,1.09,4.92)
diff <- before - after

mean(diff)
```

give a mean of 1.006923.

(b) Formulate a set of hypotheses for testing whether the installation of cameras is effective in reducing the number of accidents. Use μ_{diff} to denote the mean difference. Hint: This is not a two-sample problem but a one-sample problem, even though it may look like a two-sample problem because two sets of data have been given. Ask yourself, if the cameras are effective in reducing the number of crashes, should μ_{diff} be greater than or less than zero?

If cameras are effective in reducing the number of crashes, then the numbers of crashes before the installation of the cameras would tend to be higher than the numbers afterwards, so μ_{diff} would be greater than zero, since we are defining the differences as before-minus-after. So we are interested in testing the hypotheses

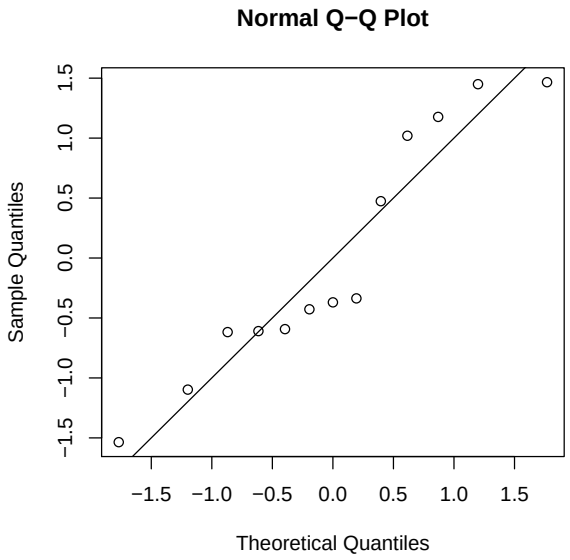
$$H_0\text{: } \mu_{\text{diff}} \leq 0 \text{ versus } H_1\text{: } \mu_{\text{diff}} > 0.$$

(c) Create a Normal quantile-quantile plot of the differences and comment on whether you think the differences are Normally distributed.

We can make a Normal quantile-quantile plot of the differences with the commands

```
qqnorm(scale(diff))
abline(0,1)
```

This produces the plot



(d) The `t.test()` function in R can be used in the one-sample setting too. Use the command

```
t.test(diff,alternative=" ")
```

to get a p-value for the test. You must decide whether to put `greater`, `less`, or `two.sided` in for the alternative.

The command

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

produces the output

```
        One Sample t-test

data:  diff
t = 3.0023, df = 12, p-value = 0.00551
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.4091686       Inf
sample estimates:
mean of x
 1.006923
```

The p-values for the test is 0.00551.

(e) What is your decision at the $\alpha = .05$ significance level? Are the cameras effective?

Based on the p-value computed in the previous part, which was 0.00551, we reject H_0 and conclude that the cameras *are* effective.

3. It is of interest whether the temperature has any effect on the mean ethanol concentration in bio-fuel produced in a fermentation process. An experiment is run under the temperatures 30°, 35°, 40°, and 45° degrees Celsius. Read the data, which come from exercise 10.39 of McClave and Sincich (2016), into R in preparation for ANOVA, with the commands

```
ethanol <- c( 103.3,103.4,101.0,101.7,102.0,101.1,97.2,96.9,96.2,55.0,56.4,54.9)
temp <- c(rep("30deg",3), rep("35deg",3),rep("40deg",3),rep("45deg",3))
temp <- as.factor(temp)
```

   (a) If it is of interest whether the temperature has any effect on the mean ethanol concentration, what are the relevant hypotheses in terms of μ_{30°, μ_{35°, μ_{40°, and μ_{45°?

   We are interested in the hypotheses

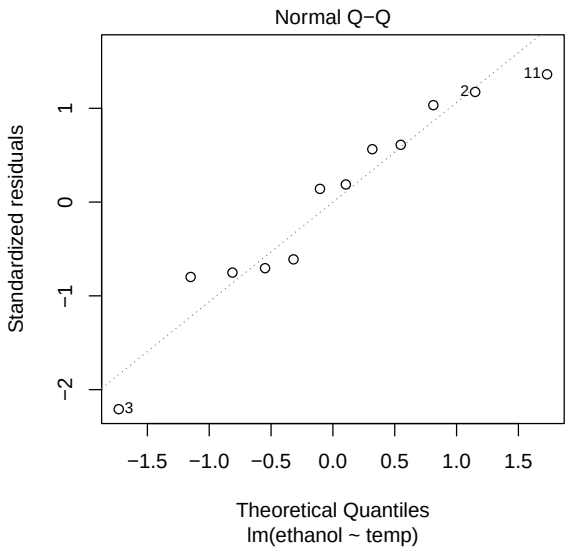
   $$H_0\text{: } \mu_{30^\circ} = \mu_{35^\circ} = \mu_{40^\circ} = \mu_{45^\circ} \text{ versus } H_1\text{: } \mu_{30^\circ}, \mu_{35^\circ}, \mu_{40^\circ}, \mu_{45^\circ} \text{ not all the same.}$$

   (b) Obtain a normal Q-Q plot of the residuals. Turn this in and comment on whether you think the responses are normally distributed around the treatment means.

   We can read in the data as follows:

```
ethanol <- c( 103.3,103.4,101.0,101.7,102.0,101.1,97.2,96.9,96.2,55.0,56.4,54.9)
temp <- c(rep("30deg",3), rep("35deg",3),rep("40deg",3),rep("45deg",3))
temp <- as.factor(temp)
```

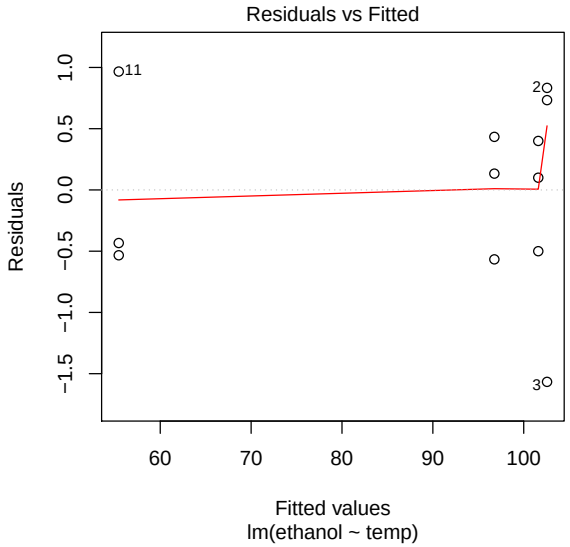
   Then we can use the command `plot(lm(ethanol ~ temp))` to obtain the plot



There appears not be be any significant departure from Normality.

(c) Obtain a residuals versus fitted values plot. Comment on whether you think the variance of the responses is the same across all groups.

We use the command `plot(lm(ethanol ~ temp))` to obtain the plot



There appears not be be any significant departure from the equal-variances assumption.

(d) Execute the command `anova(lm(ethanol ~ temp))` to generate the ANOVA table for the ethanol data. Turn in this table.

To get the ANOVA table we execute the command `anova(lm(ethanol ~ temp))`, which returns the output

```
        Analysis of Variance Table

Response: ethanol
          Df Sum Sq Mean Sq F value   Pr(>F)
temp       3 4589.5 1529.82  2026.3 7.34e-12 ***
Residuals  8    6.0    0.76
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(e) If the ANOVA assumptions are satisfied, what do you conclude about the effect of temperature on the ethanol concentration at the $\alpha = 0.01$ significance level?

Since the p-value is smaller than $\alpha = 0.01$, we reject H_0: at the $\alpha = 0.01$ significance level. We conclude that the temperature has some kind of effect on the mean ethonal concentration.

References

McClave, J. and Sincich, T. (2016). *Statistics*. Pearson Education.