ST 411/511 Homework 5

Due on February 20

Vijay Tadimeti

Winter 2019

Instructions

You should submit your assignment as either a PDF or Word document, which you can compile (should you choose – recommended) from the provided .Rmd (R Markdown) template. Please include your code.

Problems (25 points total)

Question 1: One Sample Tests

Consider the following data ($n = 18$ observations)

```
data1 <- c(5.4, -13.2, -4.3, 1.3, -14.3, -12.3, -11.4, 9.1, 2.5,
           -6.8, 2.6, 6.2, 10.4, -17.0, -17.1, -8.0, 16.5, -15.3)
sort(abs(data1))
```

```
##  [1]  1.3  2.5  2.6  4.3  5.4  6.2  6.8  8.0  9.1 10.4 11.4 12.3 13.2 14.3
## [15] 15.3 16.5 17.0 17.1
```

(a) (3 points) Use the sign test to test the null hypothesis that the population median is 0 ($H_0 : m = 0$) using the following steps:

- Compute the sign test statistic K by hand.
- Compute the z-statistic corresponding to the value of K that you obtained.
- Find the two-sided p-value using the normal approximation to the null distribution of K.

Here after by calculating by hand we find that the values bigger than 0 are 8. So K is 8.We calculate the Value of Z using K.

```
k <- 8
n <- 18
Z = (k - (n/2))/(sqrt(n/4))
p = 2*(1-pnorm(abs(Z)))
p
```

```
## [1] 0.6373519
```

(b) (5 points) Use the Wilcoxon Signed-Rank test to test that the "center" of this distribution is 0 using the following steps:

- Compute the signed-rank test statistic S by hand
- Compute the z-statistic corresponding to the value of S that you obtained.
- Find the two-sided p-value using the normal approximation to the null distribution of S.

- Verify that you get the same result by using the `wilcox.test()` function in R with `exact=FALSE` and `correct=FALSE`.

```
mu0 <- 0
distRanks <- rank(abs(data1 - mu0))
distRanks
```

```
##  [1]  5 13  4  1 14 12 11  9  2  7  3  6 10 17 18  8 16 15
```

```
Sstat <- sum(distRanks[data1 > mu0])
Sstat
```

```
## [1] 52
```

```
Z2 <- (Sstat-((n*(n+1))/4))/sqrt(n*(n+1)*((2*n)+1)/24)
Z2
```

```
## [1] -1.458937
```

```
p = 2*(1-pnorm(abs(Z2)))
p
```

```
## [1] 0.1445825
```

```
wilcox.test(data1, paired=FALSE, mu=0, exact=FALSE, correct=FALSE,alternative = "two.sided")
```

```
##
##  Wilcoxon signed rank test
##
## data:  data1
## V = 52, p-value = 0.1446
## alternative hypothesis: true location is not equal to 0
```

(c) (2 points) Discuss the differences between the results in parts a) and b). How would you interpret your analysis in each case?

We know that the Wilcoxon signed rank test is used for testing the center of a distribution in the population whereas the sign test tests if the population median is 0. Here we test different values so we cannot consider or interpret them as to be equal or not equal.

Question 2: Two-Sample Tests

Consider the guinea pig lifetime data we looked at in Homework 3 (`ex0211` in the `Sleuth3` package). Previously, we considered an equal variance two-sample t-test. Now we will consider some alternatives.

(a) (2 points) Perform a Wilcoxon rank-sum test to test whether the population distributions of lifetime in the Control and Bacilli groups are the same using the `wilcox.text()` function. Consider a one-sided lesser (Bacilli minus Control) alternative hypothesis. You do not need to use the continuity correction. What do you conclude at significance level $\alpha = 0.01$?

```
library(Sleuth3)
data(ex0211)
wilcox.test(ex0211$Lifetime[ex0211$Group == "Bacilli"], ex0211$Lifetime[ex0211$Group == "Control"], alt
```

```
##
##  Wilcoxon rank sum test with continuity correction
```

```
##
## data:  ex0211$Lifetime[ex0211$Group == "Bacilli"] and ex0211$Lifetime[ex0211$Group == "Control"]
## W = 1478.5, p-value = 0.02663
## alternative hypothesis: true location shift is less than 0
```

For the above we fail to reject the null hypothesis that the population distributions of lifetime in the Control and Bacilli groups are the same at significance level $\alpha = 0.01$? vs one-sided lesser (Bacilli minus Control) alternative hypothesis as p value is greater than alpha.

(b) (2 points) Perform a Welch's two-sample t-test to test the null hypothesis that the population means of lifetime in the Control and Bacilli groups are the same using the `t.test()` function. Consider a one-sided lesser (Bacilli minus Control) alternative hypothesis. What do you conclude at significance level $\alpha = 0.01$?

```
data(ex0211)
t.test(ex0211$Lifetime[ex0211$Group == "Bacilli"], ex0211$Lifetime[ex0211$Group == "Control"], alternati
```

```
##
##  Welch Two Sample t-test
##
## data:  ex0211$Lifetime[ex0211$Group == "Bacilli"] and ex0211$Lifetime[ex0211$Group == "Control"]
## t = -3.2296, df = 97.807, p-value = 0.0008443
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -49.89389
## sample estimates:
## mean of x mean of y
##  242.5345  345.2344
```

For this we reject the null hypothesis that the population distributions of lifetime in the Control and Bacilli groups are the same at significance level $\alpha = 0.01$? vs one-sided lesser (Bacilli minus Control) alternative hypothesis as p value is lesser than alpha.

(c) (2 points) Discuss the differences between the results in parts (a) and (b). How would you interpret your analysis in each case?

We observe a difference between the parts a and b as we know that the Wilcoxon rank-sum test is used to test whether the population distributions of lifetime in the Control and Bacilli groups are the same whereas in Welch's two-sample t-test is used to test the null hypothesis that the population means of lifetime in the Control and Bacilli groups are the same. Here, we notice that both are testing two different concepts so its ok that the p values and conclusions are different. (In first case we fail to reject the Null Hypothesis and in second case we reject the null hypothesis)

Question 3: Levene's Test

(a) (2 points) Generate two samples using the `rnorm()` function. Combine the two samples into one vector, and create another vector that indicates which group the observations belong to. The two samples should be drawn as follows:

- Sample A: $m = 10$ observations from a Normal$(\mu = 0, \sigma^2 = 1)$ distribution.
- Sample B: $n = 20$ observations from a Normal$(\mu = 0, \sigma^2 = 4)$ distribution.

```r
m <- 10
n <- 20
samp1 <- rnorm(m, mean=0, sd=1)
samp1
```

```
##  [1] -1.51270671 -1.06635168  2.10624586  1.67752884  1.97929485
##  [6] -0.05837421 -0.02404473  0.85679000 -0.73776306 -0.23770600
```

```r
samp2 <- rnorm(n, mean=0, sd=4)
samp2
```

```
##  [1]  1.5012852  7.2179838 -3.6157056 -3.5842822 -3.4315972  6.7825136
##  [7] -3.3169338 -0.3452617  4.8468863  4.2930795 -1.5672744 -3.0597448
## [13] -2.5667338 -6.0348717 -4.2704468 -6.4735701  3.4369406 -5.4017375
## [19] -8.3006434 -0.4189486
```

```r
sampComb <- c(samp1, samp2)
sampGrp <- as.factor(rep(c(1,2), c(m,n)))
sampGrp
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
## Levels: 1 2
```

```r
df <- data.frame(sampComb, sampGrp)
df
```

```
##       sampComb sampGrp
## 1  -1.51270671       1
## 2  -1.06635168       1
## 3   2.10624586       1
## 4   1.67752884       1
## 5   1.97929485       1
## 6  -0.05837421       1
## 7  -0.02404473       1
## 8   0.85679000       1
## 9  -0.73776306       1
## 10 -0.23770600       1
## 11  1.50128520       2
## 12  7.21798380       2
## 13 -3.61570555       2
## 14 -3.58428224       2
## 15 -3.43159719       2
## 16  6.78251364       2
## 17 -3.31693381       2
## 18 -0.34526172       2
## 19  4.84688629       2
## 20  4.29307955       2
## 21 -1.56727441       2
## 22 -3.05974485       2
## 23 -2.56673383       2
## 24 -6.03487168       2
## 25 -4.27044679       2
## 26 -6.47357007       2
## 27  3.43694059       2
## 28 -5.40173746       2
## 29 -8.30064336       2
```

```
## 30 -0.41894861       2
```

(b) (3 points) Perform Levene's test in R using the `leveneTest()` function in the `car` library. Note: you will need to load the `car` package using `library(car)`. Report the resulting p-value and summarize your findings (state the hypothesis tested, the results of your analysis, and your conclusions).

```
library(car)
```

```
## Loading required package: carData
```

```
leveneTest(sampComb, group = sampGrp)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  6.0839 0.02003 *
##       28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ho : The population variance are the same p = 0.02958 (at the time of running the code chunk) Hence we reject the null hypothesis that the population variance is same against the alternate hypothesis that the population variance are not the same, as our p value is less than the significance value alpha 0.05.

Question 4: Intro to ANOVA

(a) (1 point) Why is s_p^2 not simply the average of the I sample variances?

Here it is a weighted average of the sample variances in the two groups, as the two samples can be of different sizes. By weighting, we notice that the larger group will have more influence on the pooled variance estimate.It is also expected that the larger group will influence the population more than the smaller group and so it has more weightage in pooled variance calculation. Hence we can understand that sp not simply the average of the I sample variances.

(b) (1 point) What does it mean if the F-statistic from ANOVA is so small that the chance of getting an F-statistic that small or smaller is only 0.001?

What it means is that; smaller F value depict that the variations between the group is smaller than the variations within group. So, we have stronger evidence to fail to reject the null hypothesis.

Refer to the following for parts (c) and (d): The following data are sample means of (wing length – tail length) in millimeters (mm) for 24 flycatchers in each of 10 different species of flycatcher. (Note: a flycatcher is a type of bird.)

Species	1	2	3	4	5	6	7	8	9	10
Avg (Wing - Tail)	13.6	15.4	14.7	12.4	9.2	13.7	10.3	7.0	9.5	9.5

(c) (1 point) Explain why a conclusion that this measurement tends to differ in the 10 species cannot be made from these averages alone. What additional piece of information is eneded to test for group differences and to evaluate the extent to which individuals from different species can be distinguished?

We cannot calculate variance or spread from the given data because we are only given sample averages. Also we cannot calculate the spread within groups as we do not have the individual values within each sample.

(d) (1 point) What is the grand mean for this collection of observations?

```
sum = 13.6+15.4+14.7+12.4+9.2+13.7+10.3+7.0+9.5+9.5
n = 10
GrandMean = sum/n
GrandMean
```

```
## [1] 11.53
```