

ST 411/511 Homework 5

Due on February 19

Aashish Adhikari

Winter 2020

Instructions

This assignment is due by 11:59 PM, February 19, 2020 on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile (should you choose – recommended) using the provide .Rmd (R Markdown) template.** If you opt to not use R Markdown, please format your solutions in a similar manner as provided in this document. Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope. You should also use complete, grammatically correct sentences for your solutions.

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)
```

(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 .

Since there are 8 values in the sample greater than 0, $k = 8$. No. of values in the sample = 18. Hence, $n = 18$

```
z_stat <- (8 - 9)/sqrt(18/4)
```

```
z_stat
```

```
## [1] -0.4714045
```

The z -statistic is -0.4714.

```
two_sided_p_val <- 2 * (1 - pnorm(abs(z_stat)))  
two_sided_p_val
```

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

The two-sided p -value is 0.6373.

We fail to reject the null hypothesis because the p value is more than the significance level.

(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`.

The signed-rank test statistic $s = 52$

```
z_stat <- (52 - ((18 * (18 + 1)) / 4)) / (sqrt((18 * (18 + 1) * ((2 * 18) + 1)) / (24)))
z_stat
```

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

The required z -statistic is -1.458937

```
two_sided <- 2 * (1 - pnorm(abs(z_stat)))
two_sided
```

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

The two-sided p -value is 0.1445825 \sim 0.1446.

```
wilcox.test(data1, exact = FALSE, correct = FALSE)
```

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

The p -value obtained here is the same as the p -value obtained above, i.e., 0.1446

We fail to reject the null hypothesis because the p -value is greater than the significance level.

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

While we fail to reject the null hypotheses in both the cases, there is a difference in how we reached to these conclusions.

In part (a), we are testing the null hypothesis that the median of the population equals some value.

But in part (b), we are testing the null hypothesis that the pseudomedian equals some value. It does not test hypotheses about the population median.

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.test()` 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 $\alpha = 0.01$?

```

C<-ex0211[ex0211$Group=="Control",]
B<-ex0211[ex0211$Group=="Bacilli",]

C_lifetime<-C$Lifetime
B_lifetime<-B$Lifetime

wilcox.test(B_lifetime, C_lifetime, alternative = "less", correct = F)

```

```

##
## Wilcoxon rank sum test
##
## data: B_lifetime and C_lifetime
## W = 1478.5, p-value = 0.02647
## alternative hypothesis: true location shift is less than 0

```

Since the p-value is greater than the significance level 0.1, thus we fail to reject the null hypothesis that the true location shift between the two distributions is equal to zero.

(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 $\alpha = 0.01$?

```

t.test(B_lifetime, C_lifetime, alternative = "less")

##
## Welch Two Sample t-test
##
## data: B_lifetime and C_lifetime
## 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

```

Since p-value 0.0008443 is less than the significance level 0.01, we reject the null hypothesis that the true difference in means is 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?

In (a), we are generating a distribution similar to permutation-test and calculating the p-value using the calculated statistic. We are then concluding on the basis of where the p value stands in comparison to the significance level. Wilcoxon rank-sum test does not test the difference in means. It assumes that the only difference between the two populations is the shift of the distribution i.e., the location of the center. In (a), we test the null hypothesis that the true location shift between the two distributions is 0. In short, it is testing additive effects.

In (b), we are conducting a two-sample tes. We are leveraging on the fact that the sampling distribution of the true difference in sample means is approximately normal and underlying our conclusion on the basis of the p value calculated thereafter. Welch two-sample test is equal to equal-variance two-sample t-test and assumes that the the sample variances are not equal. It tests about the difference of the population means. However, it is not resilient to outliers. In (b), we test the null hypothesis that the true difference in means equals 0. In short, Welch's two-sample test is testing the difference in the mean.

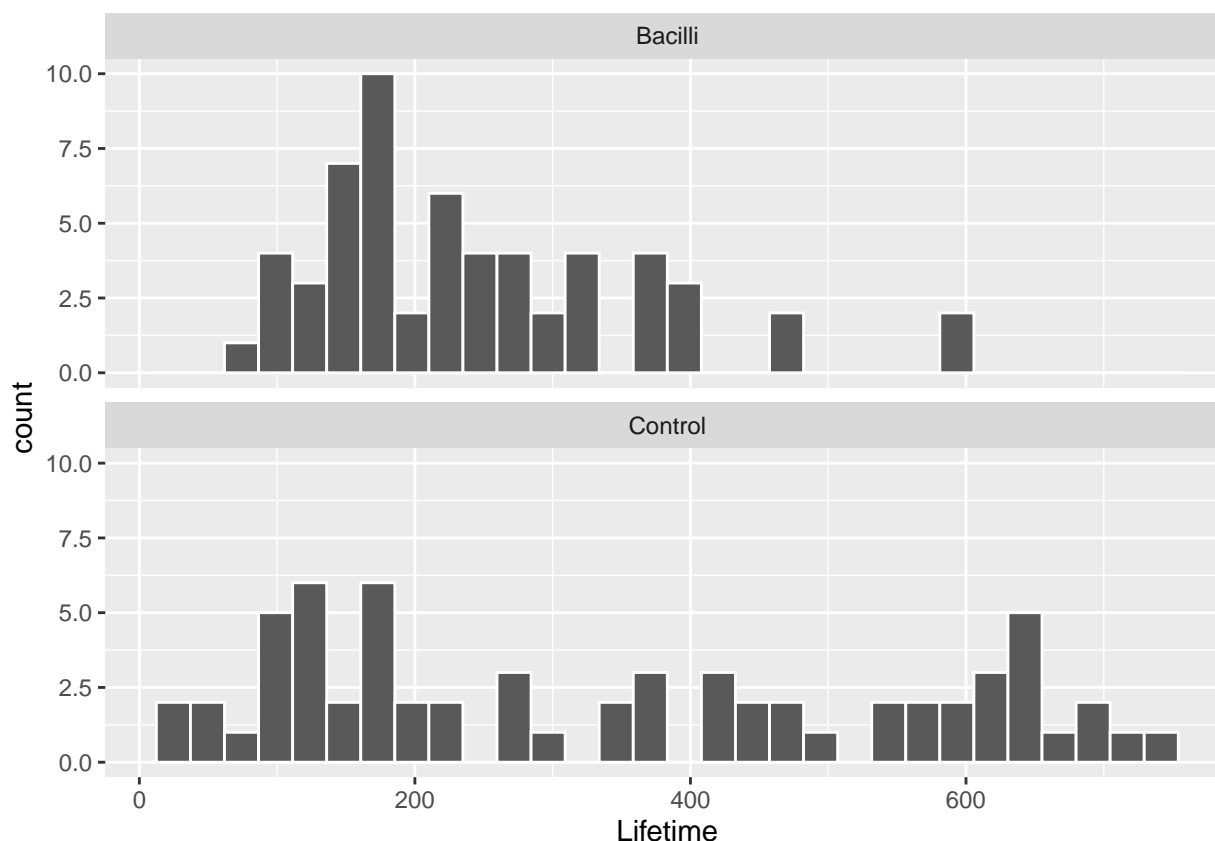
Part a tells us that we fail to reject that there is no additive effect while in part b, we reject that the difference between the means of the lifetime is zero.

(d) (4 points) Previously, in Homework 3, we analyzed the ex0211 data using the equal variance t -test. Create a faceted histogram of the Lifetime variable where the facets are based on the Group variable (Here, you should consider arranging the facets so that one histogram is on top of the other to help you better compare the difference in the amount of spread. To do this in ggplot2, include the `ncol = 1` argument within the `facet_wrap()` function).

###Based on the sample distributions of each group, argue why we should use Welch's t -test instead of the equal variance t -test. Then, compare the R output for Welch's t -test and the equal variance t -test and comment about how the two tests differ in terms of their computed test statistics, degrees of freedom, and p -values.

```
ggplot(ex0211, aes(x=Lifetime)) + geom_histogram(color = "white") + facet_wrap(~Group, ncol = 1)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We should use the Welch's t -test because as seen above, the "Control" group has a significantly higher variance while the "Bacilli" group is concentrated around 175. Thus, the variances of the two groups is not equal and this violates the assumption of equal variance taken in equal-variance two-sample t -test.

```
t.test(B_lifetime, C_lifetime, alternative = "less", var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data: B_lifetime and C_lifetime
```

```
## t = -3.1411, df = 120, p-value = 0.00106
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -48.50152
## sample estimates:
## mean of x mean of y
## 242.5345 345.2344

t.test(B_lifetime, C_lifetime, alternative = "less", var.equal = FALSE)
```

```
##
## Welch Two Sample t-test
##
## data: B_lifetime and C_lifetime
## 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
```

two-sample t-test: t statistic = -3.1411, df = 120, p-value = 0.00106

Welch's t-test: t statistic = -3.2296, df = 97.807, p-value = 0.0008443

We can see that the p-value for the two-sample t-test is much higher than the Welch's test and it would allow for more rejection than the Welch's test. This would lead to more Type-I errors. (Note: The question does not ask for any conclusions.)

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

```
A <- rnorm(10, mean = 0, sd = 1)
B <- rnorm(20, mean = 0, sd = 2)
combined <- c(A,B)
sampGrp <- as.factor(rep(c(1,2), c(10,20)))

df <- data.frame(combined, sampGrp)

df$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
## Levels: 1 2
```

Above is the required vector. Note that the question does not ask for the combined vector.

(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).

```
#install.packages("car")
library(car)
```

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

```
leveneTest(combined, group=sampGrp, center = mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##      Df F value  Pr(>F)
## group 1  11.848 0.001832 **
##      28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note: I used mean as the center. Null Hypothesis tested: The mean of the variances of the first population equals the mean of the variances of the second population.

Result: The degree of freedom is 1, the F value is 8.4499, and the corresponding significance/ p -value is 0.007063 .

Conclusion: Assuming we use a significance level of 0.05, this p -value is less than the significance level and thus we would reject the null hypothesis.