ST 411/511 Homework 4

Pavan Sai Nallagoni

Summer 2022

Instructions

This assignment is due by 11:59 PM, August 5th on Canvas via Gradescope. **You should submit your assignment as a PDF which you can compile using the provide .Rmd (R Markdown) template.**

> Note: Create a PDF by either compiling a PDF directly (via LaTeX) or from a Word Doc. Do not submit a PDF of the HTML output as they're cumbersome and difficult to read.

Include your code in your solutions and indicate where the solutions for individual problems are located when uploading into Gradescope (Failure to do so will result in point deductions). You should also write complete, grammatically correct sentences for your solutions.

Once you've completed the assignment, and before you submit it to Gradescope, you should read the document to ensure that (1) The computed values show up in the document, (2) the document "looks nice" (i.e. doesn't have extraneous code/outputs and includes *just* the essentials), and (3) ensure the document is "easy" to read.

Goals:

1. Practice using, what we often think of as, alternatives to the t-tests.
2. Familiarize ourselves with how the underlying hypotheses are different from the t-test, and amongst the alternatives, to better understand what our conclusions mean.
3. Practice hypothesis testing for equality of "scales" using Levene's test.

Question 1 (14 points): One Sample Tests

Consider the following data ($n = 18$ observations) (Assume $\alpha = 0.05$ unless otherwise specified)

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

- State the hypotheses you are testing.
- 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.
- Write a conclusion for your test.

Null Hypothesis: The population median is equal to 0 ($H_0 : m = 0$) Alternative Hypothesis: The population median is not equal to 0 ($H_A : m \neq 0$)

($H_0 : m = 0$) vs. ($H_A : m \neq 0$)

Step 1: How many observations are greater than 0?

```
K <- length(data1[data1 > 0])
K
```

```
## [1] 8
```

Step 2: With $K = 8$, we next compute the z-statistic:

```
n <- length(data1)
z <- (K - n/2) / sqrt(n/4)
z
```

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

Step 3: The two-sided p-value using the normal approximation to the null distribution of K.

```
p_value <-2 * (1 - pnorm(abs(z)))
p_value
```

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

Conclusion:

We fail to reject the Null Hypothesis ($H_0 : m = 0$) since the p-value (0.6373519) is greater than the given significance level of $\alpha = 0.05$ (assumption). Therefore, we have a plausible evidence that the population median can be equal to 0.

(b) (6 points) Use the Wilcoxon Signed-Rank test to test that the "center" of this distribution is 0 using the following steps:

- State the hypotheses you are testing.
- 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`.
- Write a conclusion for your test.

Null hypothesis: The center of the distribution is equal to 0 ($H_0 : m_o = 0$) Alternative hypothesis: The center of the distribution is equal to 0 ($H_A : m_o \neq 0$)

Step 1: Compute the signed-rank test statistic S by hand.

```
# Find distance of each distance from m0
m0 = 0 # based on null hypothesis
distance <- abs(data1 - m0)
distance
```

```
##  [1]  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
## [16]  8.0 16.5 15.3
```

```
# Rank the distances from smallest to largest
rank <- rank(distance)
rank
```

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

```
# Add up ranks corresponding to differences greater than m0 = 0
S <- sum(rank[data1 > 0])
S
```

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

Sum of ranks corresponding to differences greater than $m_o = 0$ is 52

Step 2: Compute the z-statistic corresponding to the value of S that you obtained

```
Numerator <- S - ((n*(n+1))/4)
Denominator <- sqrt((n*(n+1)*(2*n+1))/24)
new_Z <- Numerator / Denominator
new_Z
```

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

The z-statistic value is -1.458937

Step 3: Compute the two-sided p-value using the normal approximation to the null distribution of S

```
new_p_value <- 2 * (1 - pnorm(abs(new_Z)))
new_p_value
```

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

Step 4: Verify the answers match those of the wilcox.test() function

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

```
# REF: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test
```

The value of V in the output is the same as our S we computed by hand i.e., we get the same value for the test statistic (52) and the same p-value as before

Conclusion:

We fail to reject the Null Hypothesis ($H_0 : m_o = 0$) since the p-value (0.1445825) is greater than the given significance level of $\alpha = 0.05$ (assumption). Therefore, we have a plausible evidence that the center of the distribution can be equal to 0 ($H_0 : m_o = 0$).

(c) (3 points) Discuss the differences between the results in parts (a) and (b). Specifically, address:

- Do both tests come to similar conclusions (i.e. both reject or both fail to reject)?
- What are the differences between the set of hypotheses, (or are they testing the same thing)?
- In general, describe why you might choose to use one test versus the other (You might consider how you would explain the results of the test to someone who is not well versed in statistics).

The both tests come to similar conclusions that we fail to reject the Null Hypothesis.

They are not testing the same thing. The sign-test tests whether the hypothesis that the population median is equal to 0 or not whereas the wilcoxon signed rank test is used to test whether the center of the distribution can be equal to 0 or not (pseudo-median).

In general, both of these tests are used to assess the "center" of a population or of paired differences. If we want to test quantiles of a distribution or of differences we can use sign test which gives population median. But, if we want a more powerful test to test the "center" of a distribution or of differences we can use Wilcoxon signed-rank test which gives the pseudo-median.

Question 2 (16 points): 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) (4 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 $\alpha = 0.01$?

Null hypothesis: $H_0 : P(X_{Bacilli-Group} > Y_{Control-Group}) = 0.5$. The probability of the population distributions of lifetime in the Bacilli group greater than the Control group is equal to 0.5 i.e., the population distributions of lifetime in the Control and Bacilli groups are the same.

Alternative hypothesis: $H_A : P(X_{Bacilli-Group} > Y_{Control-Group}) < 0.5$. The probability of the population distributions of lifetime in the Bacilli group greater than the Control group is less than 0.5 i.e., the population distributions of lifetime in the Bacilli group is less than the control group.

```
# Load the data
data(ex0211)

# formula method:
wilcox.test(Lifetime ~ Group, data = ex0211, alternative = "less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Lifetime by Group
## W = 1478.5, p-value = 0.02663
## alternative hypothesis: true location shift is less than 0
```

Conclusion:

We fail to reject the Null hypothesis $H_0 : P(X_{Bacilli-Group} > Y_{Control-Group}) = 0.5$ since the p-value (0.02663) is greater than the significance level $\alpha = 0.01$. Therefore, we have sufficient evidence to conclude that the population distributions of lifetime in the Control and Bacilli groups can be the same.

(b) (4 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 $\alpha = 0.01$?

Null hypothesis: The difference between the population means of lifetime in the Bacilli and Control groups is equal to zero $H_0 : \mu_{Bacilli} - \mu_{Control} = 0$ i.e., the population means of lifetime in the Control and Bacilli groups are the same.

Alternative hypothesis: The difference between the population means of lifetime in the Bacilli and Control groups is less than zero $H_A : \mu_{Bacilli} - \mu_{Control} < 0$ i.e., the population means of lifetime in the Bacilli group is less than the Control group.

```
Bacilli <- (ex0211$Lifetime[ex0211$Group == "Bacilli"])
Control <- (ex0211$Lifetime[ex0211$Group == "Control"])

t.test(Bacilli, Control, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  Bacilli and 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
```

Conclusion:

We reject the Null hypothesis $H_0 : \mu_{Bacilli} - \mu_{Control} = 0$ since the p-value is less than the significance level $\alpha = 0.01$ in favor of alternative hypothesis. Therefore, we have sufficient evidence to conclude that the difference between the population means of lifetime in the Bacilli and Control groups is less than zero $H_A : \mu_{Bacilli} - \mu_{Control} < 0$ i.e., the population means of lifetime in the Control and Bacilli groups are not the same.

(c) (3 points) Discuss the differences between the results in parts (a) and (b). Specifically, address:

- Do both tests come to similar conclusions (i.e. both reject or both fail to reject)?
- What are the differences between the set of hypotheses, (or are they testing the same thing)?
- For this particular experiment, describe why you might choose to use one test versus the other.

The both tests came to different conclusions. Performing a Wilcoxon rank-sum test fail to reject the null hypothesis and concluded that the population distributions of lifetime in the Control and Bacilli groups can be the same whereas the Welch's two-sample t-test rejected the null hypothesis and concluded that the difference between the population means of lifetime in the Bacilli and Control groups is less than zero i.e., the population means of lifetime in the Control and Bacilli groups are not the same.

The Wilcoxon rank-sum test is used to test whether the the population distributions of two groups (Bacilli and Control Groups) is same or not i.e, it checks for any shift which lets us know if there're any additive effects. Whereas the Welch's two-sample t-test is used to test the difference between two independent population means (Bacilli and Control Groups).

For this particular experiment, if we want to answer the questions about the value or difference of the population means between the two groups (Bacilli and Control Groups), we can choose Welch's two-sample t-test over Wilcoxon rank-sum test. Welch's two-sample t-test is easier to interpret/explain the desired results.

However, if the population is not normal and has outliers with fewer sample sizes, we can perform the Wilcoxon rank-sum test to know if there is an additive effect by comparing the population distribution for two groups (Bacilli and Control Groups – non-parametric).

(d) (5 points) Previously, 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.

```
# Using example from lecture 11 slides.
Lifetime <- ex0211$Lifetime
Groups <- ex0211$Group

ggplot(ex0211, aes(x = Lifetime)) + geom_histogram(color = "white", bins = "30") +
facet_wrap(~Group, ncol = 1)
```
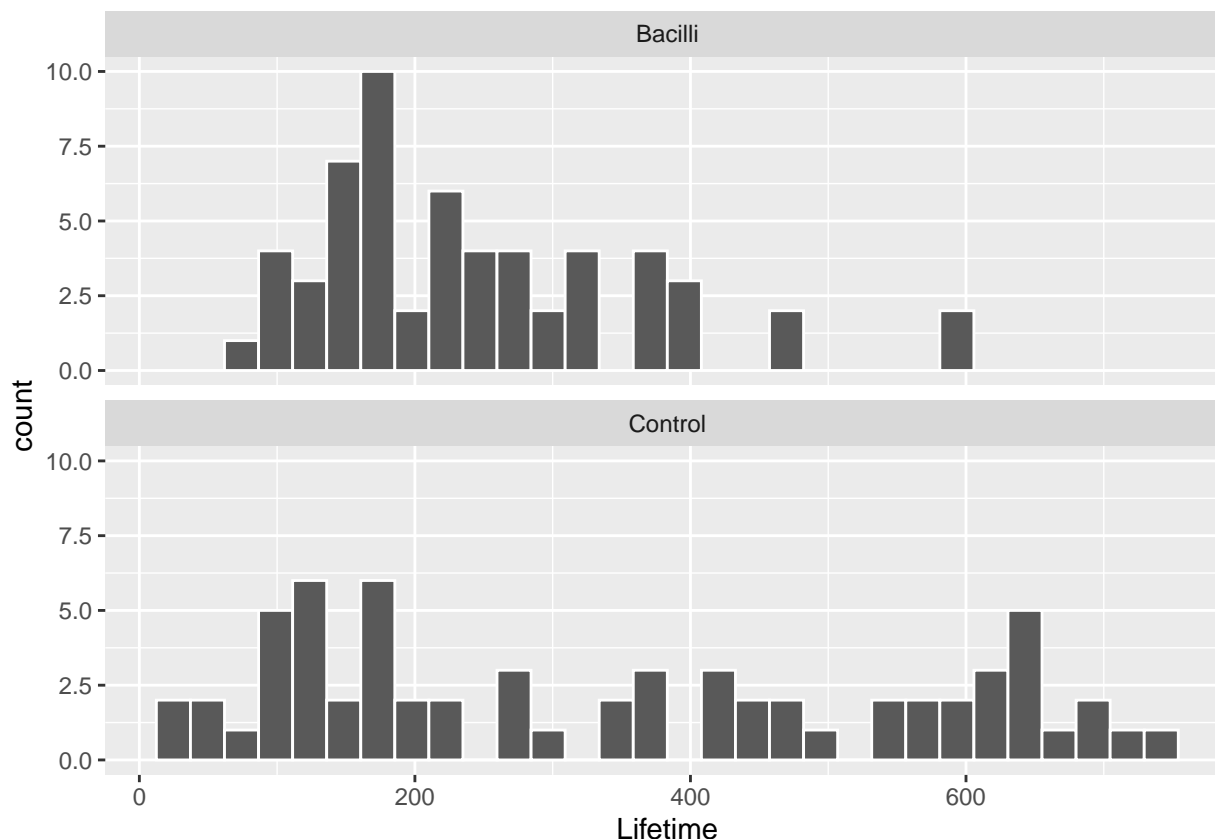
By observing the above two graphs, it is clear that the variability of two groups (Bacilli and Control) are different i.e., the spread of Bacilli Group is lesser than the spread of Control Group. Hence, we have an evidence that they do not have the equal variances. Since, the t-test assumes to have equal population variances, we cannot use a equal variance t-test and there's a need to use t-test alternatives. Therefore, we can use Welch's t-test which is similar to the equal variance two-sample t-test but by dropping the equal variance assumption.

```r
# two-sample (equal variance) t-test, and two.sided since nothing is mentioned in the question
t.test( Lifetime ~ Groups, data = ex0211, var.equal = TRUE, alternative = "two.sided")
```

```
##
##  Two Sample t-test
##
## data:  Lifetime by Groups
## t = -3.1411, df = 120, p-value = 0.00212
## alternative hypothesis: true difference in means between group Bacilli and group Control is not equal
## 95 percent confidence interval:
##  -167.43548  -37.96431
## sample estimates:
## mean in group Bacilli mean in group Control
##              242.5345              345.2344
```

```r
# Welch's two-sample t-test
t.test( Lifetime ~ Groups, data = ex0211, var.equal = FALSE, alternative = "two.sided")
```

```
##
```

9

```
##  Welch Two Sample t-test
##
## data:  Lifetime by Groups
## t = -3.2296, df = 97.807, p-value = 0.001689
## alternative hypothesis: true difference in means between group Bacilli and group Control is not equal
## 95 percent confidence interval:
##  -165.80689  -39.59289
## sample estimates:
## mean in group Bacilli mean in group Control
##              242.5345              345.2344
```

```
# t.test( Bacilli, Control, alternative = "less") -- other method
```

The Two Sample t-test has a greater t-statistic value (-3.1411), degrees of freedom (120) and p-value (0.00212) when compared to the t-statistic value (-3.2296), degrees of freedom (97.807) and p-value (0.001689) of Welch Two Sample t-test. This difference is due to the change in the formula of Standard Error (the population variance is assumed to be not equal) and the degrees of freedom (calculated based on the Satterthwaite's approximation) which effects the p-value in the Welch Two Sample t-test. Due to this, the confidence interval converges and gives more significant results of finding the difference between two independent population means with unequal variance.

Question 3 (5 points)

(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 (See Lab 6 for help with creating the data for this question). 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. Note: You might also consider using the `set.seed()` function.

```
# Given
m <- 10
n <- 20

set.seed(4511)

Sample_A <- rnorm(m, mean = 0, sd = sqrt(1))
Sample_A
```

```
##  [1] -0.79704927 -0.45185342 -1.25762656  0.08502648 -0.19207027  0.50772183
##  [7] -0.59041369 -0.17272625  0.47606346 -1.55258971
```

```
Sample_B <- rnorm(n, mean = 0, sd = sqrt(4))
Sample_B
```

```
##  [1]  2.6770219  1.5554764  0.3137570  2.1384107 -0.6501526 -1.3602756
##  [7]  0.6569371  1.4686715 -1.5248644 -0.6450761 -0.5518855 -2.7907902
## [13]  0.5788747  1.8445739  1.8223832 -0.7454333 -0.5715903 -1.7755913
## [19] -0.2951931  0.4273261
```

```
sampComb <- c(Sample_A, Sample_B)
sampGrp <- as.factor(rep(c(1,2), c(m,n)))
df <- data.frame(sampComb, sampGrp)
df
```

```
##        sampComb sampGrp
## 1   -0.79704927       1
## 2   -0.45185342       1
## 3   -1.25762656       1
## 4    0.08502648       1
## 5   -0.19207027       1
## 6    0.50772183       1
## 7   -0.59041369       1
## 8   -0.17272625       1
## 9    0.47606346       1
## 10  -1.55258971       1
## 11   2.67702188       2
## 12   1.55547644       2
## 13   0.31375703       2
## 14   2.13841072       2
## 15  -0.65015261       2
```

```
## 16 -1.36027555       2
## 17  0.65693712       2
## 18  1.46867153       2
## 19 -1.52486436       2
## 20 -0.64507606       2
## 21 -0.55188548       2
## 22 -2.79079020       2
## 23  0.57887467       2
## 24  1.84457392       2
## 25  1.82238321       2
## 26 -0.74543325       2
## 27 -0.57159033       2
## 28 -1.77559133       2
## 29 -0.29519310       2
## 30  0.42732605       2
```

(b) (3 points) Perform Levene's test in R using the `leveneTest()` function in the `car` library. Note: you will need to install/load the `car` package using `install.packages("car")` [Run the `install.packages()` function in the console and not the R Markdown] `library(car)` [Include this command in the R Markdown]. Report the resulting p-value and summarize your findings (state the hypothesis tested, the results of your analysis, and your conclusions).

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

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  6.7849 0.01455 *
##        28
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Null hypothesis: $H_0 : \sigma^2_{SampleA} = \sigma^2_{SampleB}$ The population variances across the Sample_A and Sample_B are equal.

Alternative hypothesis: $H_A : \sigma^2_{SampleA} \neq \sigma^2_{SampleB}$ The population variances across the Sample_A and Sample_B are not equal.

The value of the test statistic is 6.7849 (F value), the significance level is 0.05 (alpha value '*').

We reject the null hypothesis since the p-value 0.01455 is less than the significance level $\alpha = 0.05$. Therefore, we have sufficient evidence that the population variances across the Sample_A and Sample_B can not be equal.