Stat/Math 415 Homework 8

Due Friday Nov 22; Joseph Sepich (jps6444)

1 Problem 9.3-1

Problem Constraints:

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_1: not all equal
- $\alpha = 0.05$

This problem relates to the analysis of variacne (ANOVA). We need to use various error sources to determine if we should reject the null hypothesis. First of all let's define three error sources (where n and m are the sample size and number of groups respectively):

$$SS_E = SS_{TO} - SS_T = \Sigma_{i=1}^{m}\Sigma_{j=1}^{n}(x_{ij} - \overline{x_{i.}})^2; MS_E = \frac{SS_E}{n-m}$$

$$SS_T = \Sigma_{i=1}^{m}(\overline{x_{i.}} - \overline{x_{..}})^2 n_i; MS_{TE} = \frac{SS_T}{m-1}$$

$$SS_{TO} = \Sigma_{i=1}^{m}\Sigma_{j=1}^{n}(x_{ij} - \overline{x_{..}})^2$$

We can then use these error sources to define the F Distribution:

$$F = \frac{MS_T}{MS_E}$$

And we reject the null hypothesis if $F \geq F_\alpha(m-1, n-m)$. Let's now calculate the required values:

```
x_1 <- c(5, 9, 6, 8)
x_2 <- c(11, 13, 10, 12)
x_3 <- c(10, 6, 9, 9)

x_matrix <- data.frame(x_1 = x_1, x_2 = x_2, x_3 = x_3)

n <- 4
m <- 3

mean_1 <- sum(x_1) / n
mean_2 <- sum(x_2) / n
mean_3 <- sum(x_3) / n

means <- c(mean_1, mean_2, mean_3)
total_mean <- sum(means) / m

SS_T <- 0
SS_E <- 0
SS_TO <- 0
```

```r
for (i in 1:m) {
  SS_T <- SS_T + n * (means[i] - total_mean)^2
}

for (i in 1:m) {
  for (j in 1:n) {
    SS_E <- SS_E + (x_matrix[j, i] - means[i])^2
  }
}

for (i in 1:m) {
  for (j in 1:n) {
    SS_TO <- SS_TO + (x_matrix[j, i] - total_mean)^2
  }
}

MS_T <- SS_T / (m-1)
MS_E <- SS_E / (n * m - m)

F_var <- MS_T / MS_E

print(paste0('SS(TO) is ', SS_TO))
```

```
## [1] "SS(TO) is 66"
```

```r
print(paste0('SS(T) is ', SS_T))
```

```
## [1] "SS(T) is 42"
```

```r
print(paste0('SS(E) is ', SS_E))
```

```
## [1] "SS(E) is 24"
```

```r
print(paste0('MS(T) is ', MS_T))
```

```
## [1] "MS(T) is 21"
```

```r
print(paste0('MS(E) is ', MS_E))
```

```
## [1] "MS(E) is 2.66666666666667"
```

```r
print(paste0('F test statistic is ', F_var))
```

```
## [1] "F test statistic is 7.875"
```

Using a significance level of $\alpha = 0.05$ we can reference the value of $F_\alpha(m-1, n-m) = F_{0.05}(2, 9) = 4.2565$ in the F distribution table reference. Since our F test statistic is $7.875 > 4.2565$, we can **reject** our null hypothesis (H_0) of $\mu_1 = \mu_2 = \mu_3$.

2 Problem 9.3-15

Problem Constraints:

- H_0: $\mu_1 = \mu_2 = \mu_3$
- H_1: not all equal
- $\alpha = 0.05$

This problem relates to the analysis of variacne (ANOVA). We need to use various error sources to determine if we should reject the null hypothesis. First of all let's define three error sources (where n and m are the sample size and number of groups respectively):

$$SS_E = SS_{TO} - SS_T = \Sigma_{i=1}^{m}\Sigma_{j=1}^{n}(x_{ij} - \overline{x_{i.}})^2; MS_E = \frac{SS_E}{n-m}$$

$$SS_T = \Sigma_{i=1}^{m}(\overline{x_{i.}} - \overline{x_{..}})^2 n_i; MS_{TE} = \frac{SS_T}{m-1}$$

$$SS_{TO} = \Sigma_{i=1}^{m}\Sigma_{j=1}^{n}(x_{ij} - \overline{x_{..}})^2$$

We can then use these error sources to define the F Distribution:

$$F = \frac{MS_T}{MS_E}$$

And we reject the null hypothesis if $F \geq F_\alpha(m-1, n-m)$. Let's now calculate the required values:

```
x_1 <- c(500, 650, 530, 680, NA)
x_2 <- c(700, 620, 780, 830, 860)
x_3 <- c(500, 520, 400, 580, 410)

x_matrix <- data.frame(x_1 = x_1, x_2 = x_2, x_3 = x_3)

n <- c(length(x_1) - 1, length(x_2), length(x_3))
m <- 3

mean_1 <- sum(x_1, na.rm = TRUE) / n[1]
mean_2 <- sum(x_2, na.rm = TRUE) / n[2]
mean_3 <- sum(x_3, na.rm = TRUE) / n[3]

means <- c(mean_1, mean_2, mean_3)

total_mean <- 0
for (i in 1:m) {
  total_mean <- total_mean + sum(x_matrix[,i], na.rm = TRUE)
}
total_mean <- total_mean / sum(n)

SS_T <- 0
SS_E <- 0
SS_TO <- 0

for (i in 1:m) {
  SS_T <- SS_T + n[i] * (means[i] - total_mean)^2
```

```
}

for (i in 1:m) {
  for (j in 1:n[i]) {
    if (is.na(x_matrix[j, i])) {
      break
    }
    SS_E <- SS_E + (x_matrix[j, i] - means[i])^2
  }
}

for (i in 1:m) {
  for (j in 1:n[i]) {
    if (is.na(x_matrix[j, i])) {
      break
    }
    SS_TO <- SS_TO + (x_matrix[j, i] - total_mean)^2
  }
}

MS_T <- SS_T / (m-1)
MS_E <- SS_E / (sum(n) - m)

F_var <- MS_T / MS_E

print(paste0('SS(TO) is ', SS_TO))
```

```
## [1] "SS(TO) is 278171.428571429"
```

```
print(paste0('SS(T) is ', SS_T))
```

```
## [1] "SS(T) is 193011.428571429"
```

```
print(paste0('SS(E) is ', SS_E))
```

```
## [1] "SS(E) is 85160"
```

```
print(paste0('MS(T) is ', MS_T))
```

```
## [1] "MS(T) is 96505.7142857143"
```

```
print(paste0('MS(E) is ', MS_E))
```

```
## [1] "MS(E) is 7741.81818181818"
```

```
print(paste0('F test statistic is ', F_var))
```

```
## [1] "F test statistic is 12.4655102999396"
```

Using a significance level of $\alpha = 0.05$ we can reference the value of $F_\alpha(m-1, n-m) = F_{0.05}(2, 11) = 3.9823$ in the F distribution table reference. Since our F test statistic is $12.47 > 3.9823$, we can **reject** our null hypothesis (H_0) of $\mu_1 = \mu_2 = \mu_3$. This implies that different feed supplements do **not** have the same affect on cow weight.

3 Problem 8.4-3

Problem Constraints

- H_0: $m = 5.900$
- H_1: $m > 5.900$
- $n = 25$
- $\alpha = 0.05$

```
weight <- c(5.625, 5.665, 5.697, 5.837, 5.863, 5.870, 5.878, 5.884, 5.908, 5.967,
            6.019, 6.020, 6.029, 6.032, 6.037, 6.045, 6.049, 6.050, 6.079, 6.116,
            6.159, 6.186, 6.199, 6.307, 6.387)
n <- 25
```

3.1 Part a

We are going to use the sign test to test our hypothesis. The sign test inspects the sign value of $x_i - m_0$. After we create this, we simply have a binomial variable Y, an indicator of whether the data point is above or below our m_0. Y has the parameters $n = 25$ and $p = 0.5$, since our null hypothesis is about the median, Y would have half the values negative and half positive if it is the true median. Let's calculate our values:

```
calculations <- data.frame(weight = weight)

calculations$med <- 5.9
calculations$Difference <- calculations$weight - calculations$med
calculations$Sign <- if_else(calculations$Difference > 0, TRUE, FALSE)

calculations %>%
  head()
```

```
##   weight med Difference  Sign
## 1  5.625 5.9     -0.275 FALSE
## 2  5.665 5.9     -0.235 FALSE
## 3  5.697 5.9     -0.203 FALSE
## 4  5.837 5.9     -0.063 FALSE
## 5  5.863 5.9     -0.037 FALSE
## 6  5.870 5.9     -0.030 FALSE
```

Since the weight vector is ordered, and 8 values are negative sign in a 25 row table, we know the value of Y = 17 in this case (the binomial distribution). Let's calculate our test statistic:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{17}{25} - 0.5}{\sqrt{\frac{0.25}{25}}}$$

```
Z <- (17 / 25 - 0.5) / (sqrt((0.5 * 0.5)/n))
paste0('The value of Z is ', Z)
```

```
## [1] "The value of Z is 1.8"
```

Using the critical region approach we have $Z > Z_\alpha = Z_{0.05} = 1.645$. Since our test statistic $1.8 > 1.645$ we **can** reject the null hypothesis that the median is 5.900 meaning we conclude the median weight of a Grape Jolly Rancher is not 5.900 grams.

3.2 Part b

The Wilcoxon Test uses the same methodology, but also looks at the magnitude of our data. We are under the assumption that our data is somewhat symmetrically distributed. We already know the sign, but we also want to know rank: the magnitude of the difference from the null hypothesis. Let's calculate this:

```
calculations$AbsDiff <- abs(calculations$Difference)
calculations$Rank <- rank(calculations$AbsDiff)

calculations %>%
  select(weight, med, AbsDiff, Rank, Sign) %>%
  arrange(Rank)
```

```
##    weight med AbsDiff Rank  Sign
## 1   5.908 5.9   0.008    1  TRUE
## 2   5.884 5.9   0.016    2 FALSE
## 3   5.878 5.9   0.022    3 FALSE
## 4   5.870 5.9   0.030    4 FALSE
## 5   5.863 5.9   0.037    5 FALSE
## 6   5.837 5.9   0.063    6 FALSE
## 7   5.967 5.9   0.067    7  TRUE
## 8   6.019 5.9   0.119    8  TRUE
## 9   6.020 5.9   0.120    9  TRUE
## 10  6.029 5.9   0.129   10  TRUE
## 11  6.032 5.9   0.132   11  TRUE
## 12  6.037 5.9   0.137   12  TRUE
## 13  6.045 5.9   0.145   13  TRUE
## 14  6.049 5.9   0.149   14  TRUE
## 15  6.050 5.9   0.150   15  TRUE
## 16  6.079 5.9   0.179   16  TRUE
## 17  5.697 5.9   0.203   17 FALSE
## 18  6.116 5.9   0.216   18  TRUE
## 19  5.665 5.9   0.235   19 FALSE
## 20  6.159 5.9   0.259   20  TRUE
## 21  5.625 5.9   0.275   21 FALSE
## 22  6.186 5.9   0.286   22  TRUE
## 23  6.199 5.9   0.299   23  TRUE
## 24  6.307 5.9   0.407   24  TRUE
## 25  6.387 5.9   0.487   25  TRUE
```

We can now use the rank to calculate the random variable W. This random variable will enable use to get our test statistic. W is defined:

$$W = \Sigma_{i=1}^{n} rank_i * sign_i$$

Since our variance for W is $n(n+1)(2n+2)/6$, we can use this to create our test statistic:

$$Z = \frac{W - 0}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}$$

Let's calculate these values:

```r
W <- sum(calculations$Rank * if_else(calculations$Sign == TRUE, 1, -1))
Z <- W / (sqrt((n * (n + 1) * ( 2 * n + 1)) / 6))

paste0('The value of W is ', W)
```

```
## [1] "The value of W is 171"
```

```r
paste0('The value of Z is ', Z)
```

```
## [1] "The value of Z is 2.30054095546739"
```

Using the critical region approach at $\alpha = 0.05$ significance level, we reject if $Z > Z_\alpha = Z_{0.05} = 1.645$. Given our test statistic of $2.301 > 1.645$ we **can** reject the null hypothesis H_0 that $m = 5.900$ meaning we conclude the median weight of a Grape Jolly Rancher is not 5.900 grams.

3.3 Part c

Since we are using the t test to conduct a hypothesis test, we are assuming a normal distribution of the data. Recall the difference values for each pair we found before:

```r
calculations %>%
  select(weight, med, Difference)
```

```
##    weight med Difference
## 1   5.625 5.9     -0.275
## 2   5.665 5.9     -0.235
## 3   5.697 5.9     -0.203
## 4   5.837 5.9     -0.063
## 5   5.863 5.9     -0.037
## 6   5.870 5.9     -0.030
## 7   5.878 5.9     -0.022
## 8   5.884 5.9     -0.016
## 9   5.908 5.9      0.008
## 10  5.967 5.9      0.067
## 11  6.019 5.9      0.119
## 12  6.020 5.9      0.120
## 13  6.029 5.9      0.129
## 14  6.032 5.9      0.132
## 15  6.037 5.9      0.137
## 16  6.045 5.9      0.145
## 17  6.049 5.9      0.149
## 18  6.050 5.9      0.150
## 19  6.079 5.9      0.179
## 20  6.116 5.9      0.216
## 21  6.159 5.9      0.259
## 22  6.186 5.9      0.286
## 23  6.199 5.9      0.299
## 24  6.307 5.9      0.407
## 25  6.387 5.9      0.487
```

Using this data we must obtain the sample mean $\overline{D}$, which is the mean of the Difference values. This is then used in our test statistic:

$$t = \frac{\overline{D} - (5.900 - 5.900)}{\sqrt{S_D/n}}$$

$$S_D^2 = \Sigma_{i=1}^{n}(D_i - \overline{D})^2/(n-1)$$

```
meanD <- sum(calculations$Difference) / n
sd <- sum((calculations$Difference - meanD)^2) / (n-1)

t <- meanD/sqrt(sd/n)

paste0('Difference sample variance SD^2 is ', sd)
```

```
## [1] "Difference sample variance SD^2 is 0.0341068933333333"
```

```
paste0('Test statistic t is ', t)
```

```
## [1] "Test statistic t is 2.60774665355709"
```

Given a significance level of $\alpha = 0.05$, using the t table we have the critical region $t(n-1) > t_\alpha(n-1) = t_{0.05}(24) = 1.711$. Since our test statistic $2.608 > 1.711$ we **can** reject the null hypothesis that $m = 5.900$ meaning we conclude the median weight of a Grape Jolly Rancher is not 5.900 grams.

3.4 Part d

The sign test is the easiest test to do, because it requires no assumptions about the data, but the sign test is clearly also least likely to reject the null hypothesis, since we are less confident about the distribution of the data. The Wilcoxon test is reasonably powerful when assuming a distribution is symmetric, but the t test clearly works better if you know the data is normally distributed.

4 Problem 8.4-7

Problem Constraints

- m is median
- H_0: $m = 1.14$
- H_1: $m > 1.14$
- $n = 14$
- $\alpha \approx 0.10$

4.1 Part a

In a Wilcoxon test, we use a test statistic in the normal standard distribution table. This would make our critical region:

$$Z > Z_\alpha = Z_{0.10} = 1.28$$

4.2 Part b

The Wilcoxon Test looks at the magnitude of our data. We are under the assumption that our data is somewhat symmetrically distributed. We need to calculate the sign, but we also want to know rank: the magnitude of the difference from the null hypothesis. Let's calculate this:

```
weight <- c(1.12, 1.13, 1.19, 1.25, 1.06, 1.31, 1.12, 1.23, 1.29, 1.17, 1.20,
            1.11, 1.18, 1.23)
n <- 14

calculations <- data.frame(weight = weight)

calculations$med <- 1.14
calculations$Difference <- calculations$weight - calculations$med
calculations$Sign <- if_else(calculations$Difference > 0, TRUE, FALSE)
calculations$AbsDiff <- abs(calculations$Difference)

calculations <- calculations %>%
  select(weight, med, AbsDiff, Sign) %>%
  arrange(AbsDiff)

calculations
```

```
##    weight  med AbsDiff  Sign
## 1    1.13 1.14    0.01 FALSE
## 2    1.12 1.14    0.02 FALSE
## 3    1.12 1.14    0.02 FALSE
## 4    1.11 1.14    0.03 FALSE
## 5    1.17 1.14    0.03  TRUE
## 6    1.18 1.14    0.04  TRUE
## 7    1.19 1.14    0.05  TRUE
## 8    1.20 1.14    0.06  TRUE
## 9    1.06 1.14    0.08 FALSE
## 10   1.23 1.14    0.09  TRUE
## 11   1.23 1.14    0.09  TRUE
## 12   1.25 1.14    0.11  TRUE
## 13   1.29 1.14    0.15  TRUE
## 14   1.31 1.14    0.17  TRUE
```

Note that some of our values have the same magnitude. In ties we take the average range of ranks and insert that as each variables rank value:

```
calculations$Rank <- c(1, 2.5, 2.5, 4.5, 4.5, 6, 7, 8, 9, 10.5, 10.5, 12, 13, 14)

calculations %>%
  select(weight, Rank)
```

```
##    weight Rank
## 1    1.13  1.0
## 2    1.12  2.5
## 3    1.12  2.5
## 4    1.11  4.5
## 5    1.17  4.5
```

```
## 6    1.18  6.0
## 7    1.19  7.0
## 8    1.20  8.0
## 9    1.06  9.0
## 10   1.23 10.5
## 11   1.23 10.5
## 12   1.25 12.0
## 13   1.29 13.0
## 14   1.31 14.0
```

We can now use the rank to calculate the random variable W. This random variable will enable use to get our test statistic. W is defined:

$$W = \Sigma_{i=1}^{n} rank_i * sign_i$$

Since our variance for W is $n(n+1)(2n+2)/6$, we can use this to create our test statistic:

$$Z = \frac{W - 0}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}$$

Let's calculate these values:

```
W <- sum(calculations$Rank * if_else(calculations$Sign == TRUE, 1, -1))
Z <- W / (sqrt(n * (n + 1) * ( 2 * n + 1) / 6))

paste0('The value of W is ', W)
```

```
## [1] "The value of W is 66"
```

```
paste0('The value of Z is ', Z)
```

```
## [1] "The value of Z is 2.07162390789534"
```

Since our test statistice $2.071 > 1.28$ we would therefore **reject** the null hypothesis that $m = 1.14$ meaning we conclude that the median weight of the one pound carrot bags is not 1.14 pounds.

4.3 Part c

The P-value of this test is the probability that the test statistic is more extreme than our observed test statistic:

$$P(Z > 2.07) = 1 - P(Z \leq 2.07) = 1 - 0.9808 = 0.0192$$

This makes our p-value **0.0192**.

5 Problem 8.4-15

Problem Constraints

- $\alpha = 0.05$
- H_0: $m_x = m_y$
- H_1: $m_x \neq m_y$

```
x <- c(-2.3864, -2.2171, -1.9148, -1.9097, -1.4883, -1.2007, -1.1077, -0.3601,
       0.4325, 1.0598, 1.3035, 1.5241, 1.7133, 1.7656, 2.4912)
y <- c(-1.7613, -0.9391, -0.7437, -0.5530, -0.2469, 0.0647, 0.2031, 0.3219, 0.3579,
       0.6431, 0.6557, 0.6724, 0.6762, 0.9041, 1.3571)
n <- 15
```

For our Wilcoxon test we must first assign ranks to every value in the dataset. Any ties will be settled by giving each tied value the average of their ranks.

```
wilcoxon <- data.frame(x = x, y = y)

wilcoxon
```

```
##          x       y
## 1  -2.3864 -1.7613
## 2  -2.2171 -0.9391
## 3  -1.9148 -0.7437
## 4  -1.9097 -0.5530
## 5  -1.4883 -0.2469
## 6  -1.2007  0.0647
## 7  -1.1077  0.2031
## 8  -0.3601  0.3219
## 9   0.4325  0.3579
## 10  1.0598  0.6431
## 11  1.3035  0.6557
## 12  1.5241  0.6724
## 13  1.7133  0.6762
## 14  1.7656  0.9041
## 15  2.4912  1.3571
```

```
wilcoxon$xRank <- 0
wilcoxon$yRank <- 0
wilcoxon$xRank <- c(1, 2, 3, 4, 6, 7, 8, 12, 18, 24, 25, 27, 28, 29, 30)
wilcoxon$yRank <- c(5, 9, 10, 11, 13, 14, 15, 16, 17, 19, 20, 21, 22, 23, 26)
```

Just like with the one-sided hypotheses we have a random variable W. We can use the wilcoxon statistic to find our test statistic. W is defined:

$$W = \Sigma_{i=1}^{n} rank(y_i); W - N(n_2(n_1 + n_2 + 1)/2, n_1 n_2(n_1 + n_2 + 1)/12)$$

$$Z = \frac{W - n_2(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2(n_1 + n_2 + 1)/12}}$$

```r
W <- sum(wilcoxon$yRank)
Z <- (W - n * (n + n + 1) / 2) / sqrt((n * n * (n + n + 1))/12)

paste0('The value of W is ', W)
```

```
## [1] "The value of W is 241"
```

```r
paste0('The value of Z is ', Z)
```

```
## [1] "The value of Z is 0.352563576208345"
```

Using the critical region approach we can test that with $\alpha = 0.05$ the critical region would be $|Z| > Z_\alpha = Z_{0.05} = 1.96$. Since our test statistic $0.35 < 1.96$ we **cannot** reject the null hypothesis that the medians are equal. That means that we cannot say with 95% confidence that the two medians are not equal.