STA 106 HW2

Andrew Jowe

2024-01-26

1

a. $\hat{\mu}_1 = 38$, $\hat{\mu}_2 = 32$, $\hat{\mu}_3 = 24$
b. $E[Y_{1j}] = 38$, $E[Y_{2j}] = 32$, $E[Y_{3j}] = 24$
c. $SSTO = s_Y^2(n_T - 1) = 1088.691$ $SSA = 672$ $SSE = SSTO - SSA = 416.6912$

```
## [1] 1088.691
```

```
## [1] 672
```

```
## [1] 416.6912
```

d. $d.f.[SSTO] = n_T - 1 = 23$ $d.f.[SSA] = a - 1 = 2$ $d.f.[SSE] = n_T - a = 21$

```
## [1] 23
```

```
## [1] 2
```

```
## [1] 21
```

e. $MSTO = \frac{SSTO}{d.f.[SSTO]} = 47.3344$ $MSA = \frac{SSA}{d.f.[SSA]} = 336$ $MSE = \frac{SSE}{d.f.[SSE]} = 19.84244$

```
## [1] 47.3344
```

```
## [1] 336
```

```
## [1] 19.84244
```

2

a. The variance for each group is not constant because the variances between groups are different.
b. Seems like group 3 has the fastest recovery time. I do believe that group 3 is statistically significantly faster than group 1 because the mean of group 3 is around 3 standard deviations away from the mean of group 1.
c. $H_0 : \mu_1 = \mu_2 = \mu_3$ H_A : at least one population average is different (not equal) for the training methods
d. $F_s = 16.9334$, $p < 0.0001$

```
## [1] 16.9334
```

e. Since $p < \alpha$ when $\alpha = 0.01$, we reject the null hypothesis. Therefore, group 3 has a significantly faster recovery time than group 1 since the means are not equal and these groups have the biggest difference in means.

3

a. $\mu_M - \mu = 27.75 - 23.56 = 4.19$ $\mu_O - \mu = 21.42 - 23.56 = -2.14$ $\mu_Y - \mu = 21.50 - 23.56 = -2.06$
b. It represents how far the group mean is from the overall mean.
c. $SSTO = 399.854$, $SSA = 316.5516$, $SSE = 83.3024$

```
## [1] 399.854
```

```
## [1] 316.5516
```

```
## [1] 83.3024
```

d. $d.f.[SSTO] = 35$, $d.f.[SSA] = 2$, $d.f.[SSE] = 33$

```
## [1] 35
```

```
## [1] 2
```

```
## [1] 33
```

e. $MSTO = 11.4244$, $MSA = 158.2758$, $MSE = 2.524315$

```
## [1] 11.4244
```

```
## [1] 158.2758
```

```
## [1] 2.524315
```

4

a. It appears that group M differs significantly from group O and Y because the mean of group M is more than three standard deviations away from the mean of group O and Y.
b. The distribution of $\bar{Y}_2$ is normal where $\bar{Y}_2 \sim N(21.50, 1.73)$
c. $H_0 : \mu_M = \mu_O = \mu_Y$ H_A : at least one population average is different (not equal) for the training methods $F_s = 62.70049$

```
## [1] 62.70049
```

d. $p < 0.0001$ This is the area under the two tails of the normal distribution curve.
e. Since $p < \alpha$ when $\alpha = 0.01$, we reject the null hypothesis. Therefore, the groups differ significantly.
f. The groups that are most likely to differ significantly is group M and O and group M and Y because their means are more than 3 standard deviations apart.

5

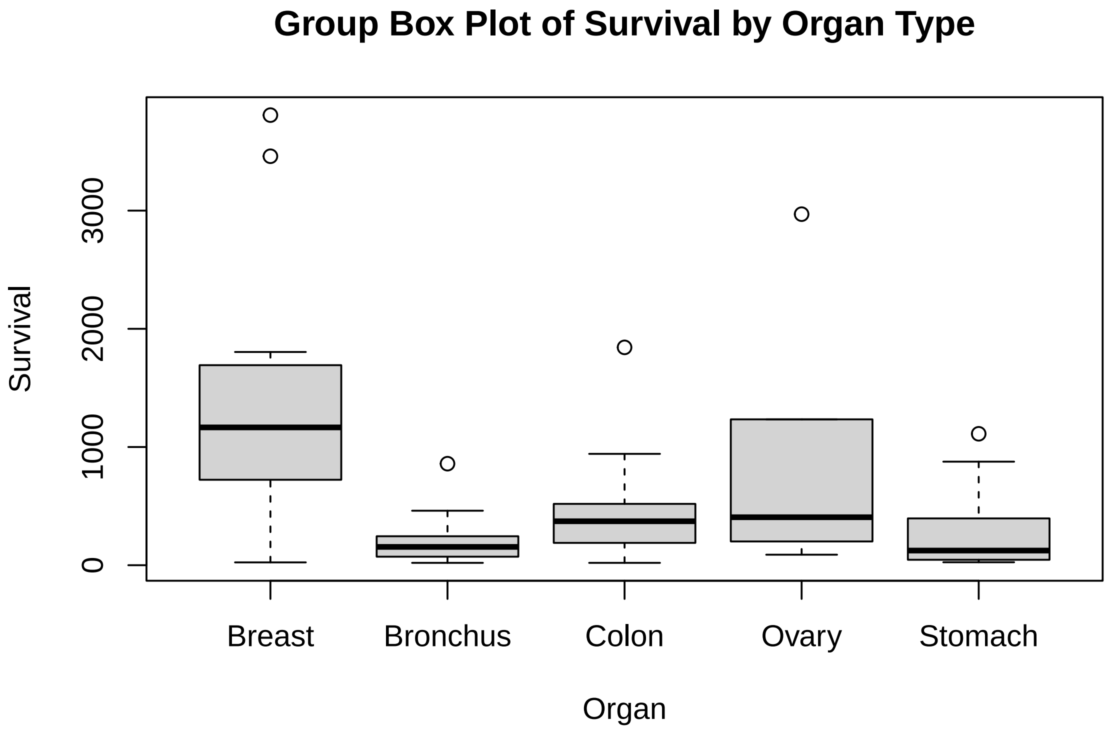
a. $\sum_j e_{ij} \rightarrow \sum_j (Y_{ij} - \mu_i) \rightarrow \sum_j Y_{ij} - \sum_j \mu_i \rightarrow \sum_j Y_{ij} - n_i * \mu_i \rightarrow \sum_j Y_{ij} - n_i * \frac{\sum_j Y_{ij}}{n_i} \rightarrow 0$
b. $\sum_i \sum_j e_{ij} = \sum_i 0 = 0$
c. $\sum_i \sum_j (\bar{Y}_{i.} - \bar{Y}_{..})(e_{ij}) \rightarrow \sum_i \sum_j (\bar{Y}_{i.} * e_{ij} - \bar{Y}_{..} * e_{ij}) \rightarrow \sum_i \sum_j (\bar{Y}_{i.})(e_{ij}) - \sum_i \sum_j (\bar{Y}_{..})(e_{ij}) \rightarrow \sum_i (\bar{Y}_{i.})(\sum_j e_{ij}) - \bar{Y}_{..} * \sum_i \sum_j e_{ij} \rightarrow \sum_i (\bar{Y}_{i.})(0) - \bar{Y}_{..} * 0 \rightarrow 0$
d. $E[\bar{Y}_{..}] = \frac{\sum_i \bar{Y}_{i.}}{n_T} = \frac{\sum_i \sum_j Y_{ij}}{n_T} = \frac{\sum_i n_i \mu_i}{n_T} = \mu.$
e. $\sigma^2(Y_{ij}) = \sigma^2(\mu_i + \epsilon) = \sigma^2(\mu_i) + \sigma^2(\epsilon) = 0 + \sigma^2_\epsilon = \sigma^2_\epsilon$

6

a. This is true, and is exactly the reason why we need to use $MSTO = \frac{SSTO}{d.f.[SSTO]}$ to normalize the result.
b. False, it depends on what the null hypothesis is. It does not always have to be that all means are equal.
c. True, the probability of a type 1 error is α.
d. True, in theory, it can be the case that all sample means are equal to the total mean, although this is unlikely in practice.

I

a. There seems to be significant differnces between the different groups because the mean of Bronchus is far away from the mean of Breast (mean of Bronchus is below the 25th percentile of Breast).



b. The averages are displayed in the table below. I believe that we will reject the single factor ANOVA because the mean of Breast is far from the mean of Bronchus relative to the variance.

```
##    Breast  Bronchus     Colon     Ovary   Stomach
## 1395.9091  211.5882  457.4118  884.3333  286.0000
```

c. No, because the range between the 25th and 75th percentile of each organ differs significantly.
d. It would make sense to have a high α value to maximize chances of a type I error.

II

a.

```
## SSTO = 37983905
```

```
## SSA = 11535761
```

```
## SSE = 26448144
```

b.

```
## MSTO = 602919.1
```

```
## MSA = 2883940
```

```
## MSE = 448273.6
```

c.

```
## f_s = 6.433437
```

```
## df_ssa = 4
```

```
## df_sse = 59
```

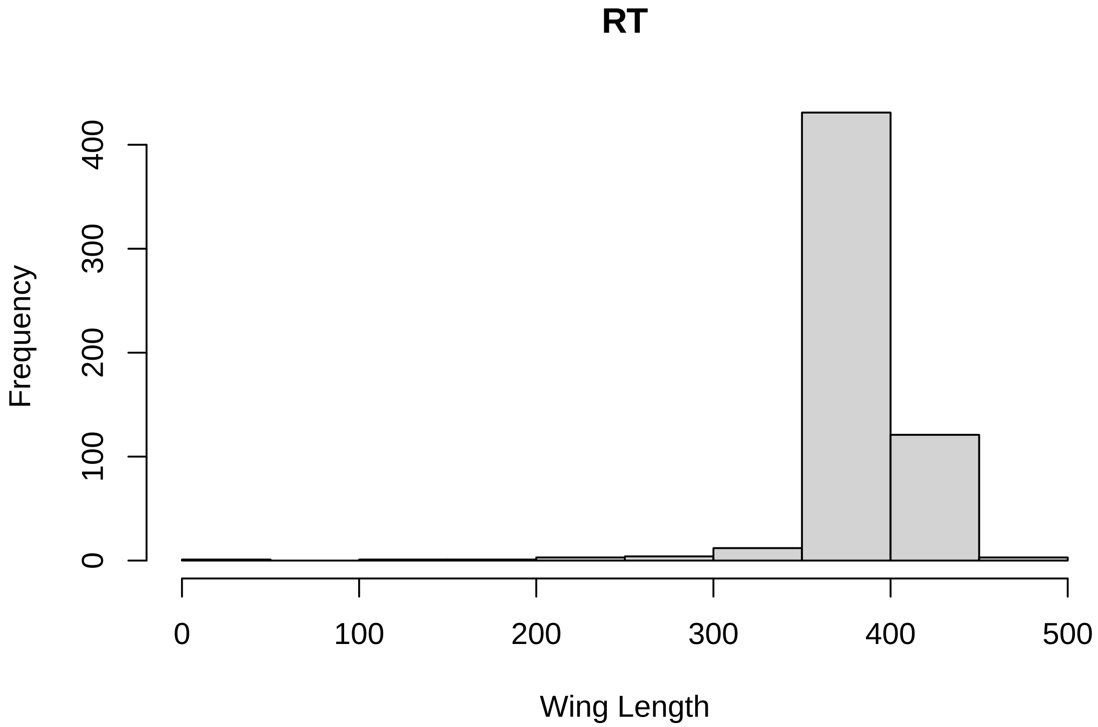
```
## p = 0.0002294532
```

d. Since $p < \alpha$ when $\alpha = 0.05$, we reject the null hypothesis. Therefore, the survival time is not the same across groups.

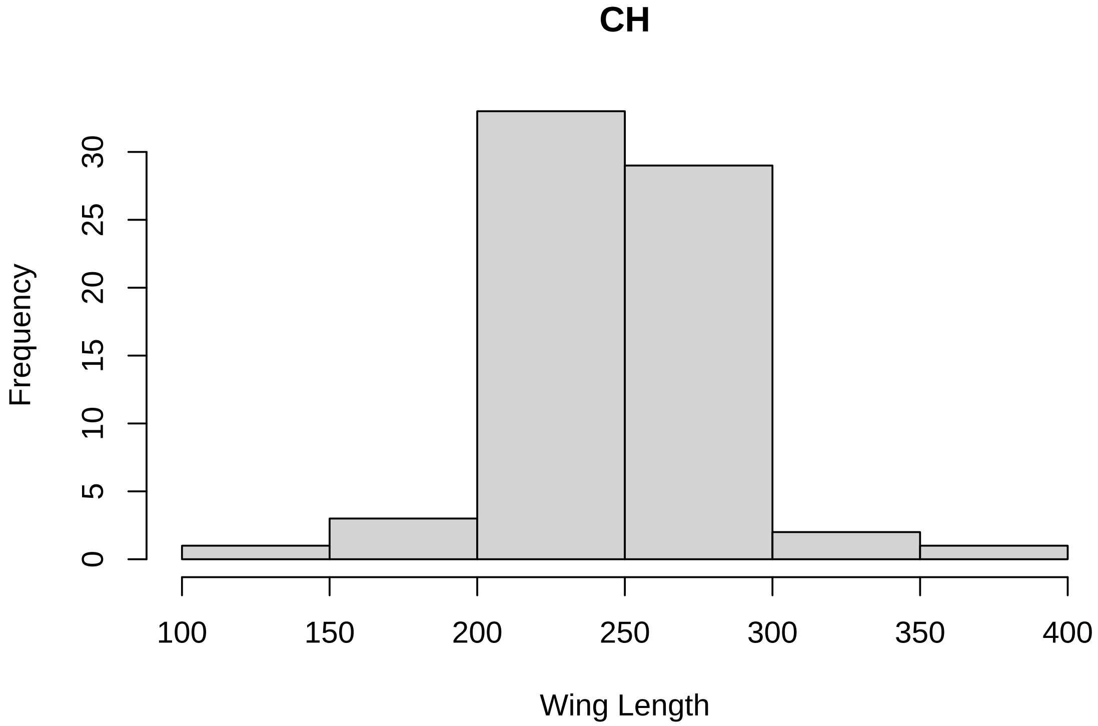
III

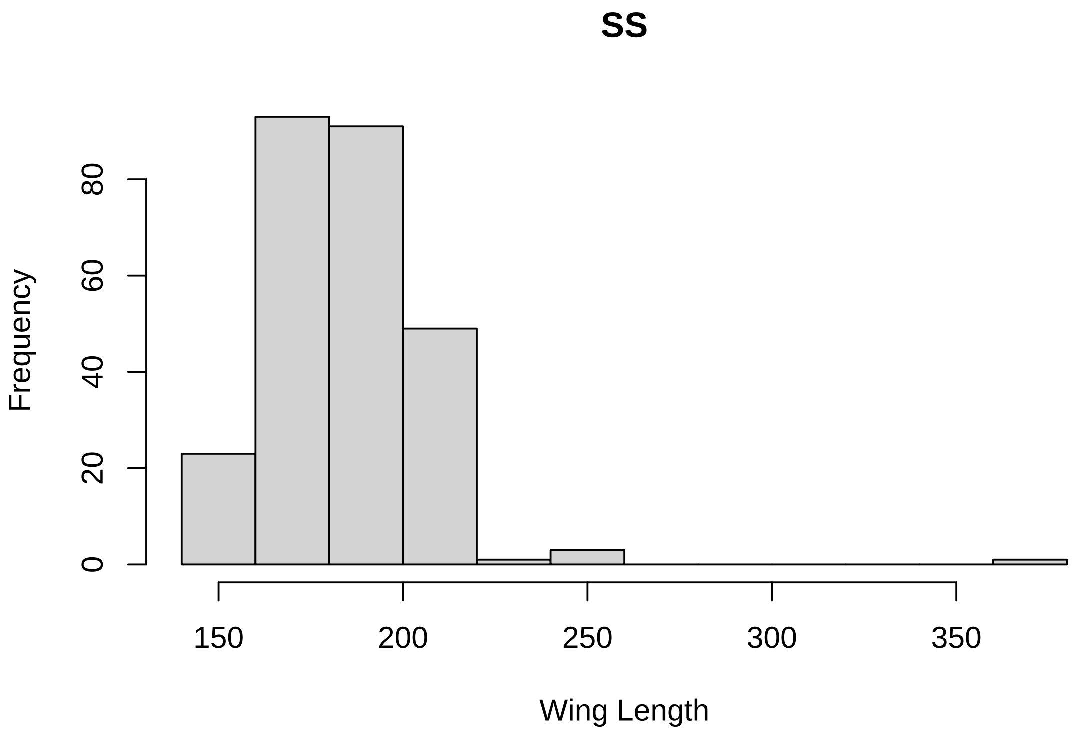
a. The table of standard deviations per group is as follows (where standard deviation is under the wing column).

```
##   Species     Wing
## 1      CH 32.13266
## 2      RT 31.48714
## 3      SS 22.42194
```


RT
Frequency
0
100
200
300
400
0
100
200
300
400
500
Wing Length

b.


CH
Frequency
0
5
10
15
20
25
30
100
150
200
250
300
350
400
Wing Length



c. I believe the assumption of equal variance is not met because the range of all three histograms (excluding outliers) are signficantly different (around 75 for SS, 300 for CH, and 200 for RT).
d. The RT group appears to have the largest wing feather length.
e. The SS group appears to have the smallest variance.

IV

a.

```
## SSTO = 8224504
```

```
## SSA = 7452511
```

```
## SSE = 771993.2
```

b.

```
## MSTO = 9077.819
```

```
## MSA = 3726256
```

```
## MSE = 853.9748
```

c.

```
## f_s = 4363.425
```

```
## df_ssa = 2
```

```
## df_sse = 904
```

```
## p = 0
```

d. Since $p < \alpha$ when $\alpha = 0.05$, we reject the null hypothesis. Therefore, the wing length is not the same across species.
e. The p-value is the area under the tails of the normal curve.

Appendix

```r
knitr::opts_chunk$set(echo = FALSE)

# 1-c
s <- 6.88
n <- 24
n_i <- c(8, 10, 6)

ssto <- s ** 2 * (n - 1)

y_bar_i_dot <- c(38, 32, 24)
y_bar_dot_dot <- 32

ssa <- sum(n_i * (y_bar_i_dot - y_bar_dot_dot) ** 2)
sse <- ssto - ssa

ssto
ssa
sse

# 1-d
a <- 3

df_ssto <- n - 1
df_ssa <- a - 1
df_sse <- n - a

df_ssto
df_ssa
df_sse

# 1-e
msto <- ssto / df_ssto
msa <- ssa / df_ssa
mse <- sse / df_sse

msto
```

```r
msa
mse

# 2-d
msa / mse

# 3-c
means <- c(27.75, 21.42, 21.50)
overall_mean <- 23.56
variance <- 3.38 ** 2
n <- c(12, 12, 12)

n_t <- sum(n)
ssto <- variance * (n_t - 1)
ssa <- sum(n * (means - overall_mean) ** 2)
sse <- ssto - ssa

ssto
ssa
sse

# 3-d
a <- 3

df_ssto <- n_t - 1
df_ssa <- a - 1
df_sse <- n_t - a

df_ssto
df_ssa
df_sse

# 3-e
msto <- ssto / df_ssto
msa <- ssa / df_ssa
mse <- sse / df_sse

msto
msa
mse

# 4-c
msa / mse

# I-a
data <- read.csv("Cancer.csv")

boxplot(Survival ~ Organ, data = data,
        main = "Group Box Plot of Survival by Organ Type",
        xlab = "Organ",
        ylab = "Survival")

# I-b
avg_by_organ <- tapply(data$Survival, data$Organ, mean)
```

```r
avg_by_organ

# II-a
n_t <- nrow(data)
overall_mean <- mean(data$Survival)

ssto <- sum((data$Survival - overall_mean) ** 2)

mean_per_group <- aggregate(Survival ~ Organ, data = data, FUN = mean)
mean_per_group <- mean_per_group$Survival
count_per_group <- aggregate(Survival ~ Organ, data = data, FUN = length)
count_per_group <- count_per_group$Survival

ssa <- sum(count_per_group * (mean_per_group - overall_mean) ** 2)
sse <- ssto - ssa

cat("SSTO =", ssto)
cat("SSA =", ssa)
cat("SSE =", sse)

# II-b
a <- nrow(aggregate(Survival ~ Organ, data = data, FUN = length))

df_ssto <- n_t - 1
df_ssa <- a - 1
df_sse <- n_t - a

msto <- ssto / df_ssto
msa <- ssa / df_ssa
mse <- sse / df_sse

cat("MSTO =", msto)
cat("MSA =", msa)
cat("MSE =", mse)

# II-c
f_s <- msa / mse
p <- pf(f_s, df_ssa, df_sse, lower.tail = FALSE)

cat("f_s =", f_s)
cat("df_ssa =", df_ssa)
cat("df_sse =", df_sse)
cat("p =", p)
data <- read.csv("Hawk.csv")

# III-a
sd_by_group <- aggregate(Wing ~ Species, data = data, FUN = sd)
sd_by_group

# III-b
for (specie in unique(data$Species)) {
  group_data <- data$Wing[data$Species == specie]
  hist(group_data, main = specie, xlab = "Wing Length")
}

# IV-a
n_t <- nrow(data)
overall_mean <- mean(data$Wing)

ssto <- sum((data$Wing - overall_mean) ** 2)

mean_per_group <- aggregate(Wing ~ Species, data = data, FUN = mean)
mean_per_group <- mean_per_group$Wing
count_per_group <- aggregate(Wing ~ Species, data = data, FUN = length)
count_per_group <- count_per_group$Wing

ssa <- sum(count_per_group * (mean_per_group - overall_mean) ** 2)
sse <- ssto - ssa

cat("SSTO =", ssto)
cat("SSA =", ssa)
cat("SSE =", sse)

# IV-b
a <- nrow(aggregate(Wing ~ Species, data = data, FUN = length))

df_ssto <- n_t - 1
df_ssa <- a - 1
df_sse <- n_t - a

msto <- ssto / df_ssto
msa <- ssa / df_ssa
mse <- sse / df_sse

cat("MSTO =", msto)
cat("MSA =", msa)
cat("MSE =", mse)

# IV-c
f_s <- msa / mse
p <- pf(f_s, df_ssa, df_sse, lower.tail = FALSE)

cat("f_s =", f_s)
cat("df_ssa =", df_ssa)
cat("df_sse =", df_sse)
cat("p =", p)
```