# HW7

Robert Chandler

2024-07-28

## Problem 1

Consider a model to predict the average length of a physician's stay based on the four regions. The one-way ANOVA model $\log_1 0Y \sim$ Region can be represented with different factor effects model, depending on the baseline. Hence, the mean responses, $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$ are the means computed on $\log_1 0Y$.

Build a model summary using the first region (1) as the baseline, the R output shows table as the following. The values in the corresponding cells are denoted by $A_i, \quad i = \{1, 2, ..., 16\}$.

### 1.a

Based on Table 1, The means, $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$ can be represented as:

$$\begin{aligned}
\mu_1 &: A_1 + 0 = A_1 \\
\mu_2 &: A_1 + A_2 \\
\mu_3 &: A_1 + A_3 \\
\mu_4 &: A_1 + A_4
\end{aligned}$$

### 1.b

Based on Table 2, The means, $\mu_1$, $\mu_2$, $\mu_3$, and $\mu_4$ can be represented as:

$$\begin{aligned}
\mu_1 &: B_1 + B_2 \\
\mu_2 &: B_1 + 0 = B_1 \\
\mu_3 &: B_1 + B_3 \\
\mu_4 &: B_1 + B_4
\end{aligned}$$

### 1.c

To perform the hypothesis $H_0 : \mu_2 = \mu_1$, the point estimate $\hat{Y}_2 - \hat{Y}_1 = \mu_2 - \mu_1 = A_1 + A_2 - A_1 = A_2$ and the p-value is $A_{14} = B_{14}$.

### 1.d

To perform the hypothesis $H_0 : \mu_3 = \mu_1$, the point estimate $\hat{Y}_3 - \hat{Y}_1 = \mu_3 - \mu_1 = A_1 + A_3 - A_1 = A_3$ and the p-value is $A_{15}$.

## Problem 2

According to the following data summary, compute the following confidence intervals:

```r
df <- data.frame(
  group = c(1, 2, 3, 4), mean = c(25, 46, 18, 62), n = c(108, 103, 152, 77)
)
```

### 2.a

The difference between the average $Y$ in Groups 1 and 2 and in Groups 3 and 4, i.e.

$$\frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

```r
alpha <- 0.05
n <- sum(df$n)
r <- length(unique(df$group))
mse <- 100
t_crit <- qt(1 - alpha / 2, n - r)
l_hat <- 0.5 * (sum(df$mean[1:2]) - sum(df$mean[3:4]))
stderr <- sqrt(mse * 0.5^2 * sum(1 / df$n))
interval_lower <- l_hat - stderr * t_crit
interval_upper <- l_hat + stderr * t_crit
```

$$L = \frac{\mu_1 + \mu_2}{2} - \frac{\mu_3 + \mu_4}{2}$$

so $c_1 = \frac{1}{2}, c_2 = \frac{1}{2}, c_3 = -\frac{1}{2}, c_4 = -\frac{1}{2}$.

$$\hat{L} = \frac{\bar{Y}_1 + \bar{Y}_2}{2} - \frac{\bar{Y}_3 + \bar{Y}_4}{2} = \frac{25 + 46}{2} - \frac{18 + 62}{2} = -4.5$$

Using the $MSE$ and $n_i$ from the table:

$$s^2[L] = MSE \sum_{i=1}^{r} \frac{c_i^2}{n_i} = 100 \left( \frac{(1/2)^2}{108} + \frac{(1/2)^2}{103} + \frac{(-1/2)^2}{152} + \frac{(-1/2)^2}{77} \right) = 0.9633489$$

so

$$s[\hat{L}] = 0.9815034$$

and with $df_E = n_T - r = 108 + 103 + 152 + 77 - 4 = 436$, the critical t-value is:

$$t(1 - \alpha/2, n - r) = t(0.975, 436) = 1.9654199$$

Therefore, the 95% confidence interval on the difference is:

$$CI = \hat{L} \pm t(1 - \alpha/2, n - r) s[\hat{L}] = (-6.4290663, -2.5709337)$$

### 2.b

The simultaneous confidence intervals for the differences of the following:

$$\mu_1 - \mu_2, \quad \mu_1 - \mu_3, \quad \mu_2 - \mu_3$$

```
group_names <- c("1-2", "1-3", "2-3")
g <- length(group_names)
bonf <- qt(1 - alpha / (2 * g), n - r)
df_bonf <- data.frame()
coeffs <- list(
  c(1, -1, 0, 0),
  c(1, 0, -1, 0),
  c(0, 1, -1, 0)
)

df_bonf[group_names, "l_hat"] <- sapply(
  coeffs,
  function(coef) sum(df$mean * coef)
)

df_bonf[group_names, "stderr"] <- sapply(
  coeffs,
  function(coef) sqrt(mse * sum(coef^2 / df$n))
)

intervals <- t(rbind(
  df_bonf$l_hat - bonf * df_bonf$stderr,
  df_bonf$l_hat + bonf * df_bonf$stderr
))
rownames(intervals) <- group_names
colnames(intervals) <- c(
  paste(100 * alpha / 2, "%"), paste(100 * (1 - alpha / 2), "%")
)
```

Calculating the critical Bonferroni value:

$$B = t(1 - \alpha/(2g), n_T - r)$$

The $\hat{L}$ for each case are:

$$\begin{aligned}\bar{Y}_1 - \bar{Y}_2 &= 25 - 46 = -21 \\ \bar{Y}_1 - \bar{Y}_3 &= 25 - 18 = 7 \\ \bar{Y}_2 - \bar{Y}_3 &= 46 - 18 = 28\end{aligned}$$

The $s[\hat{L}]$ for each case are:

$$\sqrt{MSE\sum_{i=1}^{r} c_i^2/n_i} = \sqrt{100\left(\frac{1}{108}+\frac{1}{103}\right)} = 1.3772435$$
$$\sqrt{MSE\sum_{i=1}^{r} c_i^2/n_i} = \sqrt{100\left(\frac{1}{108}+\frac{1}{152}\right)} = 1.2584994$$
$$\sqrt{MSE\sum_{i=1}^{r} c_i^2/n_i} = \sqrt{100\left(\frac{1}{103}+\frac{1}{152}\right)} = 1.2762322$$

And using the formula

$$\hat{L} \pm Bs[\hat{L}]$$

for each group

$$\begin{aligned} -21 &\pm 2.4032537 \cdot 1.3772435 \\ 7 &\pm 2.4032537 \cdot 1.2584994 \\ 28 &\pm 2.4032537 \cdot 1.2762322 \end{aligned}$$

the intervals are

$$\begin{aligned} \mu_1 - \mu_2 &: (-24.3098655, -17.6901345) \\ \mu_1 - \mu_3 &: (3.9755068, 10.0244932) \\ \mu_2 - \mu_3 &: (24.9328904, 31.0671096) \end{aligned}$$