STAT 135 CONCEPTS OF STATISTICS
HOMEWORK 6

Assigned August 7, 2021, due August 14, 2021

This homework pertains to materials covered in Lecture 15, 16 and 17. The assignment can be typed or handwritten, with your name on the document, and with properly labeled input code and computer output for those problems that require it. If not specified, **please try to perform the hypothesis testings from scratch** without using the built-in `anova` function in `R` to obtain full credit. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the Week 6 Assignment on bCourses.

Note in this homework, we use the following abbreviations: Analysis of Variance (ANOVA), confidence intervals (CIs).

Problem 1. For any numbers y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, k$,

(1) prove that

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2,$$

where $\bar{y}_{i\cdot} = n_i^{-1}\sum_j y_{ij}$, $\bar{y}_{\cdot\cdot} = \sum_i n_i\bar{y}_{i\cdot}/\sum_i n_i$;

(2) prove that for any $\mu \in \mathbb{R}$,

$$\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \mu)^2 - n(\bar{y}_{\cdot\cdot} - \mu)^2.$$

Solution.

(1) We have

$$\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + 2\sum_{i=1}^{k}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot}) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 + \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2.
\end{aligned}$$

(2) We also have

$$\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \mu + \mu - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \mu)^2 + n(\bar{y}_{\cdot\cdot} - \mu)^2 - 2(\bar{y}_{\cdot\cdot} - \mu)\sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \mu)$$
$$= \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \mu)^2 - n(\bar{y}_{\cdot\cdot} - \mu)^2.$$

Problem 2. Prove the sums of squares identity for the two-way layout:

$$SS_{\text{Tot}} = SS_A + SS_B + SS_{AB} + SS_E.$$

Also show that the sums of squares on the right-hand side of the above equation are mutually independent under the two-way ANOVA model assumption in Lecture 16.

Solution.

(1) Under the assumption that $n_{ij} = M$,

$$\begin{aligned}
SS_{\text{Tot}} &= \sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \overline{Y}_{\cdots}\right)^2 \\
&= \sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \overline{Y}_{ij\cdot} + \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots} + \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots} + \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)^2 \\
&= SS_A + SS_B + SS_{AB} + SS_E + \\
&+ 2\sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \overline{Y}_{ij\cdot}\right)\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots} + \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots} + \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right) \\
&+ 2\sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots} + \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right) \\
&+ 2\sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}\right)\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right).
\end{aligned}$$

Now we prove the sums of cross-products are all zero:

$$\sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \overline{Y}_{ij\cdot}\right)\left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots} + \overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots} + \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right) = \sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \overline{Y}_{ij\cdot}\right)\left(\overline{Y}_{ij\cdot} - \overline{Y}_{\cdots}\right)$$
$$= \sum_i \sum_j \left(\overline{Y}_{ij\cdot} - \overline{Y}_{\cdots}\right) \sum_{l=1}^{n_{ij}} \left(Y_{ijl} - \overline{Y}_{ij\cdot}\right) = 0.$$

$$\sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)\left(\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots} + \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right) = \sum_i \sum_j \sum_{l=1}^{n_{ij}} \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right)\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right)$$
$$= \sum_i \left(\overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdots}\right) \sum_j n_{ij}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot}\right) = 0.$$

$$\sum_i \sum_j \sum_{l=1}^{n_{ij}} (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right) = \sum_i \sum_j n_{ij} (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)$$

$$= \sum_j (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots}) \sum_i n_{ij}\left(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{Y}_{\cdot j\cdot} + \overline{Y}_{\cdots}\right)$$

$$= \sum_j (\overline{Y}_{\cdot j\cdot} - \overline{Y}_{\cdots})\left(Y_{\cdot j\cdot} - \sum_i n_{ij}\overline{Y}_{i\cdot\cdot} - \sum_i n_{ij}\overline{Y}_{\cdot j\cdot} + n_{\cdot j}\overline{Y}_{\cdots}\right) = 0$$

Therefore, $SS_{\text{Tot}} = SS_A + SS_B + SS_{AB} + SS_E$.

In Corollary D of Lecture 17, we prove that SS_{Tot} is independent of each cell mean $\overline{Y}_{ij\cdot}$, and SS_A, SS_B and SS_{AB} are simply functions of the cell means. Therefore, SS_A, SS_B and SS_{AB} are independent of SS_E.

Problem 3. We have seen the Bonferroni corrections for the simultaneous confidence intervals a couple of times but we never explained why that works. Now let's take a look.

(1) Consider two events A and B. Prove that $P(A \cap B) \geq P(A) + P(B) - 1$. Demonstrate using a Venn diagram will suffice.
(2) Consider m events $A_1, \ldots, A_m$. Show that the probability that they happen simultaneously can be bounded as follows:
$$P\left(\bigcap_{i=1}^m A_i\right) \geq \sum_{i=1}^m P(A_i) - (m-1).$$
(3) Suppose we have m confidence intervals for m different parameters. Each confidence interval statement can be construed as an event. Show that to have an overall confidence level of $1-\alpha$, each individual confidence interval can simply be adjusted to the level of $1 - \alpha/m$.
(4) Compared to Tukey's method, does the Bonferroni procedure produce more conservative or more confident CIs?

Solution.

(1) We know that
$$1 \geq P(A \cup B) = P(A) + P(B) - P(A \cap B),$$
which is exactly what we want to prove.
(2) Through induction,
$$P\left(\bigcap_{i=1}^m A_i\right) \geq P\left(\bigcap_{i=1}^{m-1} A_i\right) + P(A_m) - 1$$
$$\geq P\left(\bigcap_{i=1}^{m-2} A_i\right) + P(A_{m-1}) + P(A_m) - 2$$
$$\vdots$$
$$\geq \sum_{i=1}^m P(A_i) - (m-1).$$

(3) For each parameter θ_i, $i = 1, \cdots, m$, denote the event $A_i = \{\theta_i \in [L_i(\mathbf{X}_n), U_i(\mathbf{X}_n)]\}$, in which $[L_i(\mathbf{X}_n), U_i(\mathbf{X}_n)]$ is a CI of θ_i with a confidence level β. To achieve an overall confidence level of $1-\alpha$ for all CIs,

$$1-\alpha = P\left(\bigcap_{i=1}^{m} A_i\right) \geq \sum_{i=1}^{m} P(A_i) - (m-1) = m\beta - m + 1,$$

which is equivalent to have $\beta \leq 1 - \alpha/m$.

(4) The inequality in (2) was relaxed $m-1$ times to ensure $P\left(\bigcap_{i=1}^{m} A_i\right)$ is sufficiently large. Thus, the actual confidence level of a set of Bonferroni CIs would in fact be much higher than the advertised $1-\alpha$, which is why Bonferroni CIs tend to be more conservative than the Tukey's CIs.

Problem 4. One researcher collected data to see whether there exists difference in the energy use of four gas ranges for seven menu days. He was in the process of performing the one-way ANOVA analysis when his laptop crashed and all his data were erased. Help him restore the lost information and complete the following ANOVA table for him:

Source	**Sum Sq**	**Df**	**Mean Sq**	F **statistic**	p**-value**
Treatment	64.42	?	?	8.98	?
Residual	?	?	2.39		
Total	?	20			

Solution. Since there are 4 gas ranges under inspection, $k = 4$ and $n - k = 20 + 1 - 4 = 17$. Thus,

$MS_B = 64.42/3 = 21.47$, $SS_R = 2.39 * 17 = 40.63$, $SS_{Tot} = 40.63 + 64.42 = 105.05$,

and the p-value can be calculated via

```
pf(8.98, 3, 17, lower.tail=FALSE)
[1] 0.0008663231
```

Now the table is completed:

Source	**Sum Sq**	**Df**	**Mean Sq**	F **statistic**	p**-value**
Treatment	64.42	3	21.47	8.98	0.00087
Residual	40.63	17	2.39		
Total	105.05	20			

Problem 5. To determine diet quality, male weanling rats were fed diets with various protein levels. Each of 18 rats was randomly assigned to one of three diets, and their weight gain in grams was recorded in Table 1.

(1) Calculate the sums of squares and fill out a one-way ANOVA table;

Diet protein level		
Low	Medium	high
3.89	8.54	20.39
3.87	9.32	24.22
3.26	8.76	39.91
2.70	9.30	22.78
3.82	10.45	26.33
3.23	8.94	
	10.37	

TABLE 1. Weanling rat diet data set

(2) Denote the unique effects of the protein levels by α_1, α_2 and α_3. Test
$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$
with significance level $\alpha = 0.05$.

(3) Derive the Bonferroni simultaneous 95% CIs for α_1, α_2 and α_3. Do these CIs overlap with each other? What information can we learn if they don't overlap?

Solution.

(1) We have 3 treatments and 18 rats. The ANOVA table can be calculated as follows (Code attach right beneath the table):

Source	**Sum Sq**	**Df**	**Mean Sq**	F **statistic**	p**-value**
Treatment	1568.715	2	784.3577	48.91187	2.6772e-07
Residual	240.5421	15	16.03614		
Total	1809.258	17			

```
# ---- Generate One-way ANOVA table ----
# Read in data
Low <- c(3.89, 3.87, 3.26, 2.70, 3.82, 3.23)
Med <- c(8.54, 9.32, 8.76, 9.30, 10.45, 8.94, 10.37)
High <- c(20.39, 24.22, 39.91, 22.78, 26.33)

#Combine into a list
Observations <- list(Low, Med, High)
k <- length(Observations); n <- length(unlist(Observations))

# Squares within groups
S_W <- lapply(Observations, function(vec)
    sum((vec-mean(vec))^2))

# Squares between groups
tot_mean <- mean(unlist(Observations))
```

```
S_B <- lapply(Observations, function(vec)
    length(vec)*(mean(vec)-tot_mean)^2)

# F statistic
SS_B <- sum(unlist(S_B))
SS_B
[1] 1568.715

SS_W <- sum(unlist(S_W))
SS_W
[1] 240.5421

{SS_B/(k-1)}/{SS_W/(n-k)}
[1] 48.91187

# p-value
pf(48.91187, k-1, n-k, lower.tail = FALSE)
[1] 2.6772e-07
```

(2) To test the main effects of the three treatments, we can utilize the F-statistic and the p-value=2.6772e-07$<$ 0.05. Thus, we reject H_0 and conclude that there is very significant evidence that the weight gains under different diets are different.

(3) The Bonferroni CIs are visualized in Figure 1, and we see that they don't overlap with each other. By the duality of CIs and HTs, we basically reach the conclusion as (2) that $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ is not true.

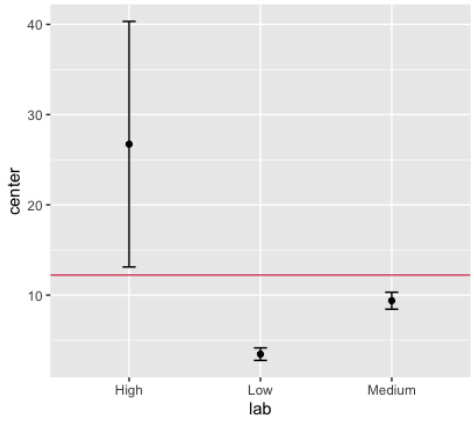

FIGURE 1. The Bonferroni CIs for α_1, α_2 and α_3 in the Weanling rat diet problem.

```
get_bound <- function(vec, k, alpha = 0.05){
```

```
  center = mean(vec)
  n = length(vec)
  halfwidth = qt(1 - (alpha/2)/k, df = n - 1)*sd(vec)/sqrt(n)

  return(c(center - halfwidth, center + halfwidth))
}

group_ind <- c('Low', 'Medium', 'High')
CIs <- sapply(Observations, get_bound, k=k, alpha=0.05)
CIs
         [,1]     [,2]    [,3]
[1,] 2.768800 8.444959 13.1211
[2,] 4.154534 10.320755 40.3309

library(ggplot2)
centers <- sapply(Observations, mean)
CI_df <- data.frame(lab = group_ind, center = centers,
                    lower = CIs[1,], upper = CIs[2,])
ggplot(CI_df, aes(x = lab, y = center)) + geom_point() +
  geom_errorbar(width = 0.1, aes(ymin = lower, ymax =
      upper)) +
  geom_hline(yintercept = mean(unlist(Observations)), col =
      2)
```

Problem 6. In Lecture 15, we derived the confidence intervals for the paired difference $\alpha_i - \alpha_r$ using the fact that

$$\frac{\left(\bar{Y}_{i\cdot} - \bar{Y}_{r\cdot}\right) - (\alpha_i - \alpha_j)}{\sqrt{S_{ir}^2\left(\frac{1}{n_i} + \frac{1}{n_r}\right)}} \sim t_{n_i+n_r-2}, \text{ for any pair } i \neq r.$$

Now let's generalize the above CIs and look at a *contrast* which is defined to be $\sum_{i=1}^k t_i\alpha_i$ in which $\sum_{i=1}^k t_i = 0$. You can see that $\alpha_i - \alpha_r$ is an specific example of a contrast.

(1) Prove that

$$\frac{\sum_{i=1}^k t_i\bar{Y}_{i\cdot} - \sum_{i=1}^k t_i\alpha_i}{\sqrt{MS_W\sum_{i=1}^k t_i^2/n_i}} \sim t_{n-k},$$

in which $MS_W = (n-k)^{-1}\sum_{i=1}^k\sum_{j=1}^{n_i}\left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2$.

(2) For the trout toxin example of Lecture 15, derive the Bonferroni simultaneous 95% CIs for the following contrasts:

$$\alpha_1 - \alpha_2,\ \alpha_1 - 2\alpha_2 + \alpha_4,\ 3\alpha_3 - \alpha_1 - 2\alpha_2,$$
$$5\alpha_2 - 4\alpha_3 - \alpha_4,\ 2\alpha_2 - \alpha_1 - \alpha_3.$$

(3) Scheffé (1959) coined a quite elegant approach to deriving simultaneous CIs on **all** contrasts. This procedure is valid for any number of contrast CIs concurrently via specifying a larger critical value $c = \sqrt{(k-1)F_{k-1,n-k}(\alpha)}$; that is, the overall probability is $1-\alpha$ that

$$\sum_{i=1}^{k} t_i\alpha_i \in \sum_{i=1}^{k} t_i\bar{Y}_{i\cdot} \pm c\sqrt{MS_W \sum_{i=1}^{k} t_i^2/n_i}$$

simultaneously for all $\mathbf{t} = (t_1, \ldots, t_k)$ such that $\sum t_i = 0$.

Calculate the Scheffé simultaneous 95% CIs for the contrasts in (2), and visualize the two sets of CIs side by side to compare.

Solution.

(1) We know that for each group mean,

$$\bar{Y}_{i\cdot} \sim N(\mu + \alpha_i, \frac{\sigma^2}{n_i}).$$

By the independence between groups,

$$\sum_{i=1}^{k} t_i\bar{Y}_{i\cdot} \sim N\left(\sum_{i=1}^{k} t_i\alpha_i, \sigma^2 \sum_{i=1}^{k} \frac{t_i^2}{n_i}\right).$$

We proved in Lecture 15 that $SS_W/\sigma^2 \sim \chi^2_{n-k}$ and it's independent of each group mean. Therefore, by the definition of a t distribution,

$$\frac{\left(\sum_{i=1}^{k} t_i\bar{Y}_{i\cdot} - \sum_{i=1}^{k} t_i\alpha_i\right) \Big/ \sqrt{\sigma^2 \sum_{i=1}^{k} \frac{t_i^2}{n_i}}}{\sqrt{MS_W/\sigma^2}} = \frac{\sum_{i=1}^{k} t_i\bar{Y}_{i\cdot} - \sum_{i=1}^{k} t_i\alpha_i}{\sqrt{MS_W \sum_{i=1}^{k} t_i^2/n_i}} \sim t_{n-k}.$$

(2) For each contrast $\sum_{i=1}^{k} t_i\alpha_i$, the CI should be

$$\sum_{i=1}^{k} t_i\bar{Y}_{i\cdot} \pm t_{n-k}(\alpha/(2m))\sqrt{MS_W \sum_{i=1}^{k} t_i^2/n_i},$$

in which $m = 5$.

We can calculate them in `R` as follows:

```
group1 <- c(28, 23, 14, 27)
group2 <- c(33, 36, 34, 29, 31, 34)
group3 <- c(18, 21, 20, 22, 24)
group4 <- c(11, 14, 11, 16)
#Combine into a list
Observations <- list(group1, group2, group3, group4)
k <- length(Observations); n <-
    length(unlist(Observations))
group_means <- sapply(Observations, function(vec)
    mean(vec))
```

```
group_lengths <- sapply(Observations, function(vec)
    length(vec))

# Calculate MS_W
S_W <- lapply(Observations, function(vec)
    sum((vec-mean(vec))^2))
SS_W <- sum(unlist(S_W))
MS_W <- SS_W/(n-k)

# Combine the t vectors of the contrasts into a matrix
t_vec_rbind <- rbind(c(1,-1,0,0), c(1,-2,0,1),
    c(-1,-2,3,0),
                    c(0,5,-4,-1), c(-1,2,-1,0))

get_contrast_bounds <- function(t_vec, alpha=0.05, m=5){
    half_width <- qt(alpha/(2*m), n-k,
        lower.tail=FALSE)*sqrt(MS_W*sum(t_vec^2/group_lengths))
    return(c(sum(t_vec*group_means),
        sum(t_vec*group_means) - half_width,
        sum(t_vec*group_means)+half_width))
}
Bonferroni_CIs <- apply(t_vec_rbind, 1,
    get_contrast_bounds, alpha=0.05, m=5)
Bonferroni_CIs
         [,1]     [,2]      [,3]    [,4]    [,5]
[1,] -9.833333 -29.66667 -25.666667 67.16667 21.66667
[2,] -16.617768 -41.01920 -42.990240 38.15974 10.56007
[3,] -3.048898 -18.31414 -8.343093 96.17360 32.77327
```

(3) We only need to change the critical value in (2) from $t_{n-k}(\alpha/(2m))$ to $\sqrt{(k-1)F_{k-1,n-k}(\alpha)}$:

```
get_contrast_bounds_scheffe <- function(t_vec,
    alpha=0.05){
    half_width <- sqrt((k-1)*qf(alpha, k-1, n-k,
        lower.tail=FALSE)*MS_W*sum(t_vec^2/group_lengths))
    return(c(sum(t_vec*group_means),
        sum(t_vec*group_means) - half_width,
        sum(t_vec*group_means)+half_width))
}
Sheffe_CIs <- apply(t_vec_rbind, 1,
    get_contrast_bounds_scheffe, alpha=0.05)
Sheffe_CIs
          [,1]     [,2]      [,3]    [,4]     [,5]
[1,] -9.833333 -29.66667 -25.666667 67.16667 21.666667
```

```
[2,] -17.063721 -41.76542 -44.128949 36.25306 9.830012
[3,]  -2.602946 -17.56791 -7.204384 98.08027 33.503322
```

Visualize the two sets of CIs in one plot:

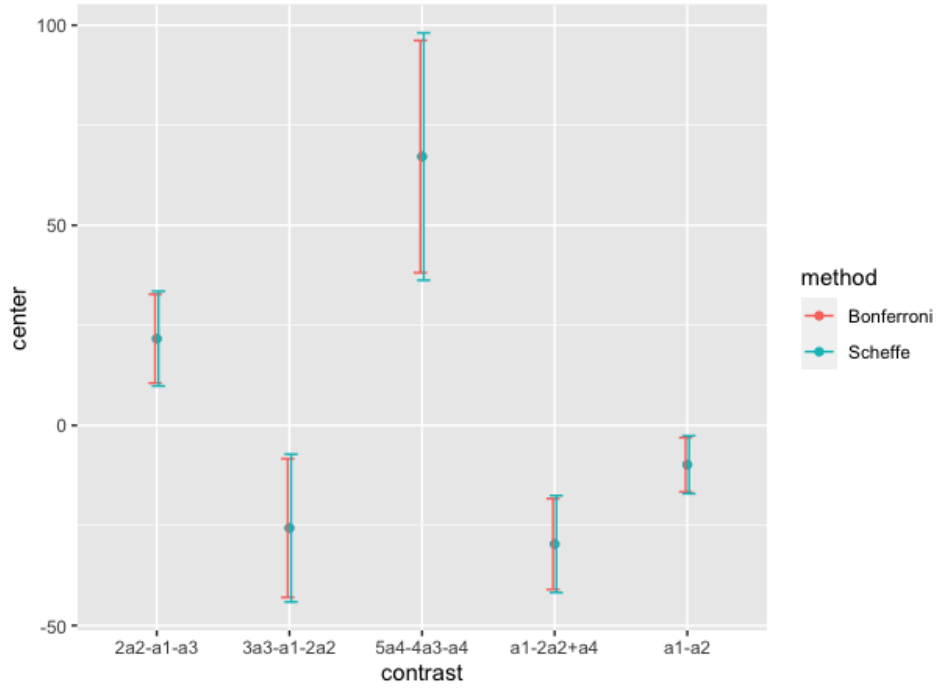

FIGURE 2. Comparison between Bonferroni and Scheffé simultaneous 95% CIs. Scheffé CIs are slightly less accurate than Bonferroni CIs.

```
# Combine CIs into a data frame
contrast_names <- c('a1-a2','a1-2a2+a4','3a3-a1-2a2',
       '5a4-4a3-a4','2a2-a1-a3')
CIs <- rbind(t(Bonferroni_CIs), t(Sheffe_CIs))
CIs <- data.frame(CIs, c(contrast_names, contrast_names),
    c(rep('Bonferroni',5), rep('Scheffe', 5)))
colnames(CIs) <- c('center', 'lower','upper', 'contrast',
    'method')

library(ggplot2)
ggplot(CIs, aes(x=contrast, y=center, group=method,
    color=method)) +
    geom_point()+
    geom_errorbar(aes(ymin=lower, ymax=upper), width=.2,
            position=position_dodge(0.05))
```

Problem 7. In Section 12.2, there is an example data from Kirchhoefer (1979), who studied the measurement of chlorpheniramine maleate in tablets. Now we look at the measurement data from another manufacturer. (You saw this dataset in Lab 11. We are only considering the first three columns here.)

Lab1	Lab2	Lab3
4.15	3.93	4.1
4.08	3.92	4.1
4.09	4.08	4.05
4.08	4.09	4.07
4.01	4.06	4.06
4.01	4.06	4.03
4	4.02	4.04
4.09	4	4.03
4.08	4.01	4.03
4	4.01	4.06

Derive three sets of simultaneous CIs for all pairwise differences using the Bonferroni method, Tukey's method and Scheffé's method respectively. Discuss the widths of the CIs calculated from different methods. Which method is the least accurate and why is that?

Solution. To derive the three sets of CIs, see the `R` script below. The results are plotted in the Figure 3. We see that in most cases, Bonferroni is the least accurate because the way Bonferroni CIs are constructed make them more conservative.

```
# Read in data
group1 <- c(4.15, 4.08, 4.09, 4.08, 4.01, 4.01, 4.00, 4.09, 4.08,
    4.00)
group2 <- c(3.93, 3.92, 4.08, 4.09, 4.06, 4.06, 4.02, 4.00, 4.01,
    4.01)
group3 <- c(4.10, 4.10, 4.05, 4.07, 4.06, 4.03, 4.04, 4.03, 4.03,
    4.06)
Observations <- list(group1, group2, group3)
k <- length(Observations); n <- length(unlist(Observations))

# Generate function to obtain Bonferroni and Scheffe at once
get_pair_CI_BS <- function(pair, Observations, alpha=0.05){
  vec1 = Observations[[pair[1]]]
  vec2 = Observations[[pair[2]]]
```

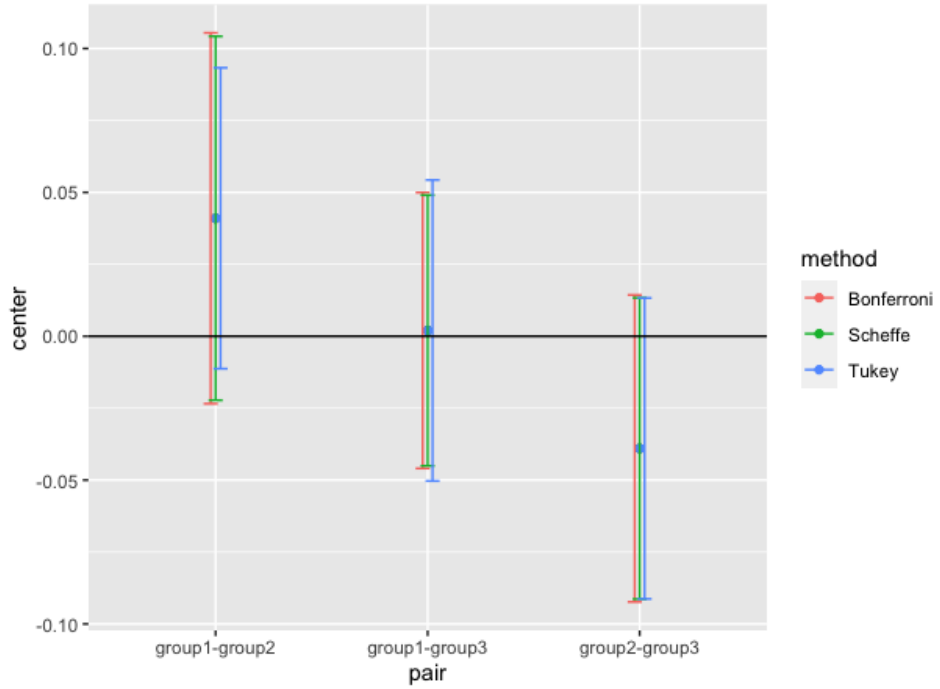

FIGURE 3. Comparison between Bonferroni, Scheffé and Tukey simultaneous 95% CIs.

```
  # bookkeeping
  center1 = mean(vec1)
  center2 = mean(vec2)
  n1 = length(vec1)
  n2 = length(vec2)
  k = length(Observations); m = k*(k-1)/2

  # Get CI bounds
  Sp2 = {(n1-1)*var(vec1) + (n2-1)*var(vec2)}/(n1+n2-2)
  halfwidth_Bon = qt(1-(alpha/2)/m, df =
      n1+n2-2)*sqrt(Sp2*(1/n1+1/n2))
  halfwidth_Sch = sqrt((k-1)*qf(alpha, k-1, n-k,
      lower.tail=FALSE)*Sp2*(1/n1+1/n2))
  return(c(center1-center2,
           center1-center2-halfwidth_Bon,
               center1-center2+halfwidth_Bon,
           center1-center2-halfwidth_Sch,
               center1-center2+halfwidth_Sch))
}
```

```
# Calculate the Bonferroni and Scheffe CIs
CIs = t(apply(combn(k, 2), 2, get_pair_CI_BS, Observations))
CIs

# Calculate Tukey's HSD CIs
group_ind <- c('group1', 'group2', 'group3')
input <- stack(setNames(Observations, group_ind))
anova_fit <- aov(values ~ ind, data = input)
posthoc <- TukeyHSD(x=anova_fit, which = 'ind', conf.level=0.95)

# Visualize in one plot
pair_names = apply(combn(k, 2), 2,
     function(pair) paste0(group_ind[pair[1]],'-',
         group_ind[pair[2]]))
Bonferonni <- data.frame(center = CIs[,1], lower=CIs[,2],
    upper=CIs[,3], pair = pair_names, method='Bonferroni')
Scheffe <- data.frame(center = CIs[,1], lower=CIs[,4],
    upper=CIs[,5], pair = pair_names, method='Scheffe')
Tukey <- data.frame(center = CIs[,1], lower=-posthoc$ind[,3],
    upper=-posthoc$ind[,2], pair = pair_names, method='Tukey')
CIs_tot <- rbind(Bonferonni, Scheffe, Tukey)
library(ggplot2)
ggplot(CIs_tot, aes(x=pair, y=center, group=method, color=method)) +
       geom_point()+
       geom_errorbar(aes(ymin=lower, ymax=upper), width=.2,
                position=position_dodge(0.07))+
       geom_hline(yintercept = 0)
```

Problem 8. A researcher ran a two-factor experiment to compare 3 different species (Species A, B and C) under different fertilizer treatments in a greenhouse. He assigned combinations of fertilizer and species levels to 72 pots to have 6 replications in the greenhouse. This would be a referred to as 3×4 factorial treatment design.

The data is in Table 2 (You can read this data set via copying and pasting in R):.

(1) Make a Treatment Mean Plot to visually examine whether there is interaction between the two factors;
(2) Calculate the sums of squares and fill out a two-way ANOVA table. Is your table the same as the table output from `anova()` in R?
(3) Denote the differential effects of species as α_1, α_2 and α_3. Test

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$$

with significance level $\alpha = 0.05$.

	Fertilizer Treatment			
	Control	**F1**	**F2**	**F3**
Species A	21.0	32.0	22.5	28.0
	19.5	30.5	26.0	27.5
	22.5	25.0	28.0	31.0
	21.5	27.5	27.0	29.5
	20.5	28.0	26.5	30.0
	21.0	28.6	25.2	29.2
Species B	23.7	30.1	30.6	36.1
	23.8	28.9	31.1	36.6
	23.7	34.4	34.9	37.1
	22.8	32.7	30.1	36.8
	22.8	32.7	30.1	36.8
	24.4	32.7	25.5	37.1
Species C	25.1	28.4	22.8	32.8
	22.6	26.4	23.2	34.3
	24.5	27.8	26.4	33.3
	23.7	26.7	23.8	31.9
	22.6	25.3	25.4	32.6
	23.9	25.9	22.7	30.6

TABLE 2. Greenhouse data set

(4) Denote the differential effects of fertilizer treatments as β_1, β_2, β_3 and β_4. Test

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

with significance level $\alpha = 0.05$.

(5) Denote the interactive effects between species and fertilizer treatments as δ_{ij}, $i = 1, 2, 3,\ j = 1, 2, 3, 4$. Test

$$H_0 : \delta_{ij} = 0$$

with significance level $\alpha = 0.05$.

Solution.

(1) The Treatment Means Plot is shown in Figure 4 (Code attached below). We can see that the connected means are not parallel across fertilizer groups. Therefore, we can see interaction effects through visual inspection.

```
p <- 6
cell_names <- expand.grid(species = c('A','B','C'),
    fertilizer = c('Control', 'F1', 'F2','F3'))
repeats <- rep(1:nrow(cell_names), each = p)
```

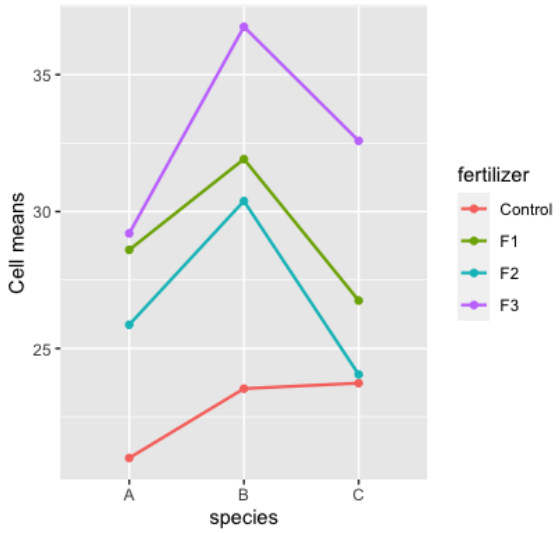

FIGURE 4. Treatment Means Plot for the greenhouse data set.

```
Cell_names <- cell_names[repeats, ]
Obs <- as.vector(matrix(c(21.0, 32.0, 22.5, 28.0, 19.5,
    30.5, 26.0, 27.5, 22.5, 25.0, 28.0, 31.0, 21.5, 27.5,
    27.0, 29.5, 20.5, 28.0, 26.5, 30.0, 21.0, 28.6, 25.2,
    29.2, 23.7, 30.1, 30.6, 36.1, 23.8, 28.9, 31.1, 36.6,
    23.7, 34.4, 34.9, 37.1, 22.8, 32.7, 30.1, 36.8, 22.8,
    32.7, 30.1, 36.8, 24.4, 32.7, 25.5, 37.1, 25.1, 28.4,
    22.8, 32.8, 22.6, 26.4, 23.2, 34.3, 24.5, 27.8, 26.4,
    33.3, 23.7, 26.7, 23.8, 31.9, 22.6, 25.3, 25.4, 32.6,
    23.9, 25.9, 22.7, 30.6), ncol=4, byrow=TRUE))
Obs_w_names <- data.frame(species =
    factor(Cell_names$species),
                fertilizer =
                    factor(Cell_names$fertilizer), Obs =
                    Obs)

Cell_means <- aggregate(Obs ~ species + fertilizer,
                    data = Obs_w_names, FUN = mean)
ggplot(data=Cell_means,
       aes(x=species, y=Obs, group=fertilizer,
           color=fertilizer)) +
    geom_line(size = 0.8) + geom_point() + ylab("Cell
        means")
```

(2) The two-way ANOVA table can be calculated as follows (Code attached below), which is exactly the same as the `anova()` output:

Source	**Sum Sq**	**Df**	**Mean Sq**	F **statistic**	p**-value**
Factor A	283.1136	2	141.5568	52.9812	5.5494e-14
Factor B	964.8994	3	321.6331	120.3792	2.3768e-25
Interaction A*B	126.2231	6	21.0372	7.8737	2.8345e-06
Residual	160.31	60	2.6718		
Total	1534.546	71			

```
## Bookkeeping
I = 3; J = 4; n = nrow(Obs_w_names)
all_mean = mean(Obs_w_names$Obs)

## Calculate SS_A
S_A = aggregate(Obs ~ species, data = Obs_w_names,
      FUN = function(vec)
          length(vec)*(mean(vec)-all_mean)^2)
SS_A = sum(S_A$Obs)
SS_A

## Calculate SS_B
S_B = aggregate(Obs ~ fertilizer, data = Obs_w_names,
           FUN = function(vec)
               length(vec)*(mean(vec)-all_mean)^2)
SS_B = sum(S_B$Obs)
SS_B

## Calculate SS_E
S_E = aggregate(Obs ~ species + fertilizer, data =
    Obs_w_names,
           FUN = function(vec) sum((vec-mean(vec))^2))
SS_E = sum(S_E$Obs)
SS_E


## Calculate SS_AB
cell_mean = aggregate(Obs ~ species + fertilizer, data =
    Obs_w_names,
                 FUN = mean)
cell_mean_vec = rep(cell_mean$Obs, each = p)
A_means = aggregate(Obs ~ species, data = Obs_w_names,
                FUN = mean)
```

```
A_means_vec = rep(rep(A_means$Obs, times = J), each = p)
B_means = aggregate(Obs ~ fertilizer, data = Obs_w_names,
                FUN = mean)
B_means_vec = rep(rep(B_means$Obs, each = I), each = p)
Tmp_data = cbind(cell_mean_vec, A_means_vec, B_means_vec,
    all_mean)
S_AB = apply(Tmp_data, 1, function(vec)
    (vec[1]-vec[2]-vec[3]+vec[4])^2)
SS_AB = sum(S_AB)
SS_AB


## Test main effect of A
F_stat = {SS_A/(I-1)}/{SS_E/(n-I*J)}
F_stat
pf(52.9812, I-1, n-I*J, lower.tail = FALSE)

## Test main effect of B
F_stat = {SS_B/(J-1)}/{SS_E/(n-I*J)}
F_stat
pf(120.3792, J-1, n-I*J, lower.tail = FALSE)

## Test interaction A*B
F_stat = {SS_AB/((I-1)*(J-1))}/{SS_E/(n-I*J)}
F_stat
pf(7.8737, (I-1)*(J-1), n-I*J, lower.tail = FALSE)

fit <- lm(Obs ~ species*fertilizer, data = Obs_w_names)
anova(fit)
```

(3) From the ANOVA table, we know the p-value for testing $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ is 5.5494e-14, which is significantly less than 0.05. Hence we reject H_0 and conclude that the main effect of species is statistically significant.

(4) From the ANOVA table, we know the p-value for testing $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ is 2.3768e-25, which is also significantly less than 0.05. Hence we reject H_0 and conclude that the main effect of fertilizer is more statistically significant.

(5) From the ANOVA table, we know the p-value for testing $H_0 : \delta_{ij} = 0$ is 2.8345e-06, which is significantly less than 0.05. Hence we again reject H_0 and conclude that the interaction effect between species and fertilizer is statistically significant.

Problem 9. Try to develop a parametrization for a balanced three-way layout. Define main effects and two-factor and three-factor interactions,

and discuss their interpretation. What linear constraints do the parameters satisfy?

Solution. In three-way ANOVA, we have three treatment factors A, B, C. Suppose A has I levels, B has J levels and C has K levels. Then we have IJK cells in total. Since we are designing a balanced layout, each cell would have M observations. The model assumption can be written as

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl},$$
$$l = 1, \ldots, M,\ i = 1, \ldots, I,\ j = 1, \ldots, J,\ k = 1, \ldots, K,$$

in which terms like $(\alpha\beta)_{ij}$ are abuses of notations to simply emphasize which factors are interacting. Simplarly to one-way and two-way ANOVA, we have to impose the following contraints so that the model is identifiable:

$$\sum_i \alpha_i = \sum_j \beta_j = \sum_k \gamma_k = 0,$$
$$\sum_i (\alpha\beta)_{ij} = \sum_j (\alpha\beta)_{ij} = 0, \text{and so forth for } (\alpha\gamma)_{ik} \text{ and } (\beta\gamma)_{jk},$$
$$\sum_i (\alpha\beta\gamma)_{ijk} = \sum_j (\alpha\beta\gamma)_{ijk} = \sum_k (\alpha\beta\gamma)_{ijk}.$$

The interpretations for the main effects and two-way interactions are the same as covered in one-way and two-way ANOVA. A three-way interaction means that the interaction among the two factors $A*B$ is different across the levels of the third factor C. If the interaction of $A * B$ differs a lot among the levels of C then it sounds reasonable that the three way interaction $A * B * C$ should appear as significant (Therefore you can also visually inspect the three-way interactions via plotting the treatment means plot of $A * B$ vs C).

To put it another way: A two way interaction $A * B$ exists in reality (not statistically) along with a three order interaction $A * B * C$ only if the way that the factors A and B interacts among the levels of the factor C is similar (i.e. parallel connected means).

Problem 10. Pottery shards are collected from four sites in the British Isles: Llanedyrn (L), Caldicot (C), Isle Thorns (I) and Ashley Rails (A). Each pottery sample was returned to the laboratory for chemical assay. In these assays the concentrations of five different chemicals were determined: Aluminum (Al), Iron (Fe), Magnesium (Mg), Calcium (Ca) and Sodium (Na). The data set can be downloaded from the *data_sets* directory under bCourses.

Try to follow the example from Lecture 17 to perform MANOVA in `R` to test

$$H_0 : \boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_3 = \boldsymbol{\alpha}_4 = \mathbf{0}$$

with $\alpha = 0.05$, in which $\boldsymbol{\alpha}_i$'s are the unique effects of different sites.

Solution. The MANOVA test results using different test statistics are summarized in the following table: We see from the table that Pillai trace

	Approx F	Df	p-value
Pillai trace	4.2984	15, 60	2.413e-05
Wilk's lambda	13.088	15, 50.091	1.84e-12
Hotelling-Lawley trace	39.376	15, 50	<2.2e-16
Roy's maximum root	136.64	5, 20	9.444e-15

is the most conservative out of all 4 tests. However, they all reach the same conclusion that H_0 should be rejected and that the unique effects of different sites are statistically significant.

```
dataset = read.table("/Users/LikunZhang/Downloads/pottery.txt",
    header=TRUE, row.names = NULL)

colnames(dataset)[1] <- 'Group'

# MANOVA test
res.man <- manova(cbind(Al, Fe, Mg, Ca, Na) ~ factor(Group),
                  data = dataset)
summary(res.man, 'Pillai')
summary(res.man, 'Wilks')
summary(res.man, 'Hotelling-Lawley')
summary(res.man, 'Roy')
```