Homework 3

PSTAT 120C

Summer 2022 Session B

Reading

The purpose of this portion of the assignment is to guide your reading and help you generate concise reading notes that list the key concepts – generally, terminology, definitions, and theorems. For the submission, treat each bullet point as an exercise and submit your 'answers' as you would a problem set.

- Define the ANOVA procedure in your own words. How does the model detect a difference in means by comparing variances?

  The ANOVA model is essentially a statistical hypothesis test for the ratio of signal to noise. In other words, an ANOVA works by taking a measure of the total variation in Y – the total sum of squares – and partitioning it into at least two parts, one related to the indepndent variable and another related to error. The test statistic is then the ratio of the sum of squares related to the independent variable to the sum of squares due to error; the **signal** to the **noise**.

- Write equations for the following in a general one-way ANOVA:
  - The total sum of squares:
    $$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}\right)^2$$
  - The sum of squares for treatment:

    Let $\bar{Y}_{i*}=\left(\frac{1}{n_i}\right)\sum_{j=1}^{n_i}Y_{ij}$. Then the sum of squares for treatment is:

    $$\sum_{i=1}^{k}n_i\left(\bar{Y}_{i*}-\bar{Y}\right)^2$$
  - The sum of squares for error:
    $$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_{i*}\right)^2$$

- What are the null and alternative hypotheses for a one-way ANOVA?

  The null hypothesis is $H_0:\mu_1=\mu_2=...=\mu_k$. The alternative is $H_a:\mu_i\neq\mu_{i'}$ for some $i\neq i'$

- What assumptions should be met when we conduct an ANOVA F-test?
  - Normality: Each sample is obtained from a normally distributed population;
  - Independence: Each sample is independent of the others;
  - Variance equality: The population variances for each sample are equal;
  - Continuous DV: The outcome should be a continuous variable.

- Write the general statistical model for a one-way ANOVA.

  For $i = 1, 2, ..., k$ and $j = 1, 2, ..., n_i$:

  $$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

  where Y_{ij} = the jth observation from population (or treatment) i, μ = the overall mean, τ_i = the nonrandom effect of treatment i where $\sum_{i=1}^{k} \tau_i = 0$, and ϵ_{ij} = random error terms such that ϵ_{ij} are independent normally distributed random variables with $E[\epsilon_{ij}] = 0$ and $V(\epsilon_{ij}) = \sigma^2$.

- Write the general statistical model for a two-way ANOVA.

  For $i = 1, 2, ..., k$ and $j = 1, 2, ..., b$:

  $$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$

  where Y_{ij} = the jth observation from treatment i in block j, μ = the overall mean, τ_i = the nonrandom effect of treatment i where $\sum_{i=1}^{k} \tau_i = 0$, β_j = the nonrandom effect of block j where $\sum_{j=1}^{b} \beta_j = 0$,and ϵ_{ij} = random error terms such that ϵ_{ij} are independent normally distributed random variables with $E[\epsilon_{ij}] = 0$ and $V(\epsilon_{ij}) = \sigma^2$.

Practice

The purpose of this portion of the assignment is to help you practice applying concepts in the reading, and in some cases, establish results that will be used later on. Remember that you will be graded on problem attempts, not solutions; do your best and ask questions if you get stuck.

1. The Florida Game and Fish Commission desires to compare the amounts of residue from three chemicals found in the brain tissue of brown pelicans. Independent random samples of ten pelicans each yielded the accompanying results (measurements in parts per million). Is there evidence of sufficient differences among the mean residue amounts at the 5% level of significance?

	Chemical		
Statistic	DDE	DDD	DDT
Mean	.032	.022	.041
Standard deviation	.014	.008	.017

We know that $n = 10$ for each group, and that $SSE = \sum_{i=1}^{k} (n_i - 1)\, s_i^2$, so $SSE = 9(.014)^2 + 9(.008)^2 + 9(.017)^2$, or $SSE = .005$.

We also know that the grand mean, or overall mean, is $\bar{Y} = \frac{(.032+.022+.041)}{3} = .0317$. Therefore, the sum of squares treatment, or $SST = \sum_{i=1}^{k} n_i \left(\bar{Y}_i - \bar{Y}\right)^2$, so:

$$SST = 10(.032 - .0317)^2 + 10(.022 - .0317)^2 + 10(.041 - .0317)^2 \quad (1)$$

$$SST = .0018067 \quad (2)$$

Since $k = 3$ and $n = n_1 + n_2 + n_3 = 30$, $k - 1 = 2$ and $n - k = 27$. So $MST = \frac{.0018067}{2} = .00090335$ and $MSE = \frac{.005}{27} = .0001851852$. Then the F-statistic is $F = \frac{0.00090335}{0.0001851852}$, or $F = 4.88$ with 2 and 27 degrees of freedom.

The critical F value, F_α, is then:

```
qf(p = 0.05, df1 = 2, df2 = 27, lower.tail = F)
```

```
## [1] 3.354131
```

so the observed F-statistic falls in the rejection region. We reject the null hypothesis with $p < .05$ and conclude that there is a statistically significant difference among the mean chemical residues.

2. It has been hypothesized that treatments (after casting) of a plastic used in optic lenses will improve wear. Four different treatments are to be tested. To determine whether any differences in mean wear exist among treatments, 28 casting from a single formulation of the plastic were made and 7 castings were randomly assigned to each of the treatments. Wear was determined by measuring the increase in "haze'' after 200 cycles of abrasion (better wear being indicated by smaller increases). The data collected are reported in the accompanying table.

Treatment			
A	B	C	D
9.16	11.95	11.47	11.35
13.29	15.15	9.54	8.73
12.07	14.75	11.26	10.00
11.97	14.79	13.66	9.75
13.31	15.48	11.18	11.71
12.32	13.47	15.03	12.45
11.78	13.06	14.86	12.38

a. Is there evidence of a difference in mean wear among the four treatments? Use $\alpha = 0.05$.

The null hypothesis here is $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$.

We can read in the data:

```
prob_2 <- tibble(
treatment = c(rep("A", 7), rep("B", 7), rep("C", 7), rep("D", 7)),
wear = c(9.16, 13.29, 12.07, 11.97, 13.31, 12.32, 11.78,
               11.95, 15.15, 14.75, 14.79, 15.48, 13.47, 13.06,
               11.47, 9.54, 11.26, 13.66, 11.18, 15.03, 14.86,
               11.35, 8.73, 10, 9.75, 11.71, 12.45, 12.38)
    )
```

Then we can fit a one-way ANOVA:

```
    aov(wear ~ treatment, data = prob_2) %>%
      summary()
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## treatment    3  36.75  12.250   4.877 0.00869 **
## Residuals   24  60.28   2.512
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is evidence of a difference in wear among the treatments; we can reject the null hypothesis with $F = 4.88$, $p < .05$.

b. Estimate the mean difference in haze increase between treatments B and C using a 99% confidence interval.

In a one-way ANOVA, the formula for a confidence interval around a difference between group means i and i' is $\left(\bar{Y}_i - \bar{Y}_{i'}\right) \pm t_{\frac{\alpha}{2}} S \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}}$.

The t-statistic is $t_{\frac{\alpha}{2}}$ with $n - k = 28 - 4 = 24$ degrees of freedom, or 2.79694.

The means of each treatment group are:

```
prob_2 %>% group_by(treatment) %>%
  summarise(means = mean(wear))
```

```
## # A tibble: 4 x 2
##   treatment means
##   <chr>     <dbl>
## 1 A          12.0
## 2 B          14.1
## 3 C          12.4
## 4 D          10.9
```

So $\bar{Y}_i - \bar{Y}_{i'} = 14.09286 - 12.42857 = 1.66429$.

We also know that $S = \sqrt{S^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n_A+n_B+n_C+n_D-k}}$. SSE is reported when we run the ANOVA with `aov()`; $SSE = 60.28$. Therefore, $S = \sqrt{\frac{60.28}{28-4}} = \sqrt{2.511667} = 1.58$.

Finally, we obtain a 99% CI for the group difference:

```
low_ci <- 1.66429 - (2.79694*1.58*sqrt(1/7 + 1/7))
high_ci <- 1.66429 + (2.79694*1.58*sqrt(1/7 + 1/7))
low_ci; high_ci
```

```
## [1] -0.6978532
```

```
## [1] 4.026433
```

The 99% CI for the difference between treatments B and C is then $(-0.6978532, 4.026433)$. It contains zero, which means there is not enough evidence to conclude that $\mu_B \neq \mu_C$.

c. Find a 90% confidence interval for the mean wear for lenses receiving treatment A.

The formula for a CI of a single group mean in a one-way ANOVA is $\bar{Y}_i \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n_i}}$. The mean for treatment A is 11.98571, as shown above; $S = 1.58$), and $n_i = 7$. The t-statistic is now 1.710882 because this is a 90% CI.

```
low_ci <- 11.98571 - (2.79694*(1.58/sqrt(7)))
high_ci <- 11.98571 + (2.79694*(1.58/sqrt(7)))
low_ci; high_ci
```

```
## [1] 10.31542
```

```
## [1] 13.656
```

The 90% CI for the mean of treatment A is then $(10.31542, 13.656)$. It does **not** contain zero, which means there is enough evidence to conclude that the population mean of treatment A, $\mu_A \neq 0$.

3. Fill in the blanks in the following two-way ANOVA table, using the information provided:

Source	SS	df	MS	F
Block		7		14.5
Treatment	5797.5		1932.5	
Interaction	11363.1			
Error	14841.6		154.6	
Total		127		

Here is the table with values filled in in bold:

Source	SS	df	MS	F
Block	**15691.9**	7	**2241.7**	14.5
Treatment	5797.5	**3**	1932.5	**12.5**
Interaction	11363.1	**21**	**541.1**	**3.5**
Error	14841.6	**96**	154.6	
Total	**47694.1**	127		

The easiest way to fill in this table is to first obtain the treatment and error degrees of freedom: $df_T = \frac{SS_T}{MS_T} = \frac{5797.5}{1932.5} = 3$ and $df_E = \frac{SS_E}{MS_E} = \frac{14841.6}{154.6} = 96$. Then the remaining degrees of freedom can be calculated by subtraction: $df_I = df_{total} - df_B - df_T - df_E$, so $df_I = 127 - 7 - 3 - 96$, or $df_I = 21$.

MS_I is then $= \frac{SS_I}{df_I} = \frac{11363.1}{21} = 541.1$. $MS_B = MS_E(F_B)$, or $MS_B = 154.6(14.5) = 2241.7$.

From here, it's relatively simple: $F_T = \frac{MS_T}{MS_E} = \frac{1932.5}{154.6} = 12.5$ and $F_I = \frac{MS_I}{MS_E} = \frac{541.1}{154.6} = 3.5$. $SS_B = MS_B(df_B) = 2241.7(7) = 15691.9$.

Finally, $SS_{total} = SS_B + SS_T + SS_I + SS_E$, so $SS_{total} = 15691.9 + 5797.5 + 11363.1 + 14841.6 = 47694.1$.