MTH 326 - Spring 2022
Assignment #9
Due: Friday, April 29 2022 (11:59pm)

1. The results that were obtained from an analysis of data obtained in a study to assess the relationship between percent increase in yield (Y) and base saturation (x_1, pounds/acre), phosphate saturation (x_2, BEC%), and spoil pH (x_3). Fifteen responses were analyzed in the study. The least-squares equation and other useful information follow.

$$\widehat{y} = 38.83 - 0.0092x_1 - 0.92x_2 + 11.56x_3, \qquad S_{yy} = 10965.46, \qquad \text{SSE} = 1107.01$$

$$10{,}000(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 151401.8 & 2.6 & 100.5 & -28082.9 \\ 2.6 & 1 & 0 & 0.4 \\ 100.5 & 0 & 8.1 & 5.2 \\ -28082.9 & 0.4 & 5.2 & 6038.2 \end{bmatrix}$$

(a) Assuming that all the x_i's are independent, is there sufficient evidence that $\beta_2 < 0$? What can you conclude about the usefulness of the x_2 term in the model?

Solution: Let $H_0 : \beta_2 = 0$ and $H_a : \beta < 0$. Then, from the data provided we have $c_{2\,2} \equiv (\mathbf{X}^T\mathbf{X})^{-1}_{3\,3} = 8.1/10000$. Note that our degrees of freedom are $15 - 4 = 11$. Hence

$$S = \sqrt{\frac{\text{SSE}}{\text{df}}} = \sqrt{\frac{1107.01}{11}} \approx 10.0318$$

and $\sqrt{\text{Var}\left(\widehat{\beta}_2\right)} = S\sqrt{c_{2\,2}} \approx 10.0318\sqrt{0.00081} \approx 0.28551.$

$$\text{Then } T = \frac{\widehat{\beta}_2 - 0}{\sqrt{\text{Var}\left(\widehat{\beta}_2\right)}} \approx \frac{-0.92}{0.28551} \approx -3.2223.$$

For $\alpha = 0.01$, $p = \Pr\left(T < -3.2232 \mid \text{df} = 11\right) = 0.004056$ by WebAssign's "technology." Since $p < \alpha$ we reject H_0. That is, the x_2 term is useful to this model.

(b) Give a 95% confidence interval for the mean percent increase in yield if $x_1 = 914$, $x_2 = 65$, and $x_3 = 6$.

Solution: The interval we are looking for is $\text{E}\left(Y\right) \pm t_{0.025}(11) \cdot S\sqrt{a^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}}$ where $a = (1,\ 914,\ 65,\ 6)^T$. Then

$$\begin{aligned} a^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1} &= \begin{pmatrix} 1 & 914 & 65 & 6 \end{pmatrix} \frac{1}{10000} \begin{bmatrix} 151401.8 & 2.6 & 100.5 & -28082.9 \\ 2.6 & 1 & 0 & 0.4 \\ 100.5 & 0 & 8.1 & 5.2 \\ -28082.9 & 0.4 & 5.2 & 6038.2 \end{bmatrix} \begin{pmatrix} 1 \\ 914 \\ 65 \\ 6 \end{pmatrix} \\ &= 92.9772 \\ \iff \sqrt{a^T\left(\mathbf{X}^T\mathbf{X}\right)^{-1}} &\approx 9.64246856 \end{aligned}$$

Then $t_{0.025}(11) = 2.201$ by table and $S = 10.0318$ from (a). Lastly $\mathrm{E}(Y) = \widehat{y}(a) = 39.9812$. Hence the confidence interval is

$$39.9812 \pm 2.201 \cdot 10.0318 \cdot 9.64246856 \equiv (-172.92442,\ 252.88682)$$

2. Data is collected from three populations (A, B, C) which have normal distributions with a common variance. The data is as follows:

A	1	3	2
B	6	4	
C	6	8	4

(a) This data set is small enough to do by hand. Determine $\widehat{\mu}_i$'s, $\widehat{\mu}_0$, SSA and SSW.

Solution: We have $k = 3$, $n_1 = n_3 = 3$, $n_2 = 2$ and $n = 8$. Then

$$\widehat{\mu}_1 = \frac{1}{3}\left[1+3+2\right] = 2$$
$$\widehat{\mu}_2 = \frac{1}{2}\left[6+4\right] = 5$$
$$\widehat{\mu}_3 = \frac{1}{3}\left[6+8+4\right] = 6$$
$$\widehat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{k} n_i\widehat{\mu}_i = \frac{1}{8}\left[3(2)+2(5)+3(6)\right] = 4.25$$
$$\mathrm{SSA} = \sum_{i=1}^{k} n_i(\widehat{\mu}_i - \widehat{\mu}_0)^2 = 3(2-4.25)^2 + 2(5-4.25)^2 + 3(6-4.25)^2 = 25.5$$
$$\mathrm{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \widehat{\mu}_i\right)^2$$
$$= \underbrace{(1-2)^2 + (3-2)^2 + (2-2)^2}_{i=1} + \underbrace{(6-5)^2 + (4-5)^2}_{i=2} + \underbrace{(6-6)^2 + (8-6)^2 + (4-6)^2}_{i=3} = 12$$

(b) Fill out the ANOVA table for this experiment.

Solution:

Source	SS	df	Mean Squares	F_{obs}	p-value
among	25.5	2	12.75	5.3125	0.057926
within	12	5	2.4		
total	37.5	7			

(c) Is there sufficient evidence at level $\alpha = 0.10$ that at least one of the means is different from the others? Briefly explain.

Solution: Since $0.058 = p < \alpha = 0.1$ then we reject the null hypothesis. Hence, there *is* evidence that at least one of the means are different.

(d) Give bounds for the p-value in this experiment.

Solution: $p \in (0.05, 0.1)$ by appendix table 7 since $F(2,5)_{.1} = 3.78 < F < F(2,5)_{.05} = 5.79$

3. Students at a large high-school preparing for the ACT exam have the option of attending an ACT tutorial session. Some students chose to do this and some do not. School administrators want to determine if the tutorial helps. The ACT scores of students at this school are summarized below:

attended tutorial:	$\bar{y} = 26$	$S = 3.5$	$n = 50$
did not attend:	$\bar{y} = 25$	$S = 3$	$n = 80$

(a) Determine $\widehat{\mu}_0$ and then compute SSA (or SST, using the book's notation) and SSW (or SSE) in the book.

Solution:

$$\widehat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{k} n_i\widehat{\mu}_i = \frac{1}{130}\big[50(26) + 80(25)\big] = 25.3846154$$

$$\text{SSA} = \sum_{i=1}^{k} n_i(\widehat{\mu}_i - \widehat{\mu}_0)^2 = 50(26 - 25.3846154)^2 + 80(25 - 25.3846154)^2 = 30.7692308$$

$$\text{SSW} = (50-1)(3.5^2) + (80-1)(3^2) = 1311.25$$

(b) Fill out the ANOVA table and test the administrators' conjecture at the $\alpha = 0.05$ significance level. Fully justify your conclusion.

Solution:

Source	SS	df	Mean Squares	F_{obs}	p-value
among	30.769	1	30.769	3.00361187	0.085488
within	1311.25	128	10.244		
total	1342.02	129			

Let $H_0 : \mu_1 = \mu_2$ and $H_a : \mu_1 \neq \mu_2$. Then for $\alpha = 0.05$, computing $F_{0.05}(1,128) = 3.9151$ by WebAssign technology. Since $F_{\text{obs}} < F = 3.9151$ then we accept H_0. There is not enough evidence to indicate a difference in ACT performance.

(c) Determine a 95% confidence interval for the difference in means.

Solution: Using Chapter 9 techniques, CI $\equiv (\widehat{\mu}_1 - \widehat{\mu}_2) \pm t_{\alpha/2}(\text{df})\sqrt{\frac{S_1}{n_1} + \frac{S_2}{n_2}}$. Hence

$$\begin{aligned}\text{C.I.} &= (26 - 25) \pm t_{0.025}(129)\sqrt{\frac{3.5^2}{50} + \frac{3^2}{80}} \\ &= 1 \pm 1.960 \cdot 0.59791 \\ &\equiv (-0.1719,\ 2.1719) \text{ points}\end{aligned}$$

Using MSW $\equiv S^2$ yields (as it should) about the same decimals.

(d) Explain how your answer in (b) supports your conclusion in (a).

Solution: Interpretting this as how (b) supports (c): Since we accept the null hypothesis in (b) it would make logical sense that $0 \in$ CI from (c).

4. An experiment was conducted to compare the starch content of tomato plants grown in sandy soil supplemented by one of four different nutrient packages, A, B, C, and D.

A	45	47	67	61	56	72
B	39	39	43	46	59	27
C	60	42	55	44	30	18
D	56	61	60	47	69	42

(a) Using the technology of your choice, construct an ANOVA table.

Solution: We're really banking on the fact that I programmed this correctly (problem 2 was the only test case):

```
"C:\Program Files\Java\jdk-16.0.2\bin\java.exe" "-javaagent:C:\Program Files\Jet
+-------+-----------+---+-------------+-----------+-----------------------+
|Source |SS         |df |Mean Squares |Fobs       |p-value                |
+-------+-----------+---+-------------+-----------+-----------------------+
|among  |1380.4583333|3  |460.1527777777|3.2506230254|P{F(3, 20) > 3.250623025|
|within |2831.1666666|20 |141.5583333333|           |                       |
|total  |4211.625    |23 |             |           |                       |
+-------+-----------+---+-------------+-----------+-----------------------+

Process finished with exit code 0
```

(b) Report the p-value and interpret the p-value in the context of the problem

Solution: Computing the p-value (which I wasn't going to code or import an F-distibution) with the WebAssign technology, we get $p = 0.043377$. In the context of the problem we can be at most 95.66% confident that at least one of the nutrient packages differs significantly from the others.

5. In class we stated the MLE for $\widehat{\mu}_i$ and ${S_a}^2$ under the alternative hypothesis that all the means are not equal. Using the likelihood function $L(\text{all } Y_{i\,j} \mid \widehat{\mu}_1, \ldots, \widehat{\mu}_k, {S_a}^2)$ that we derived in class, show that the MLE is as stated in class. (Hint: Recall that when using this method to determine $\widehat{\theta}$, we differentiate with respect to the whole of the estimator $\frac{\partial}{\partial \widehat{\theta}}$)

6. In class we stated that if $X \sim \chi^2(n) = \Gamma\left(\frac{n}{2}, 2\right)$, so that the expected value of $\mathrm{E}\left(\frac{1}{X}\right) = \frac{1}{n-2}$. We seek to prove this.

   (a) Using the definition of the gamma function (the *function*, not distribution), write the following integral as a multiple of a gamma function:

$$\int_0^\infty x^{\frac{n}{2}-2} e^{-x/2}\, dx$$

$$\text{Goal: } c\Gamma(z) = c\int_0^\infty x^{z-1}e^{-x}dx \to \int_0^\infty x^{\frac{n}{2}-2}e^{-x/2}dx$$

$$\begin{aligned}
&\int_0^\infty x^{\frac{n}{2}-2} e^{-\frac{x}{2}} dx && (u = \tfrac{x}{2}\ du = \tfrac{1}{2}dx)\\
&= \int_0^\infty (2u)^{\frac{n}{2}-2} e^{-\frac{2u}{2}} (2du) && (\text{apply } u\text{-sub})\\
&= 2\int_0^\infty 2^{\frac{n}{2}-2} u^{\frac{n}{2}-2} e^{-u} du && (\text{pull constant \& distribute})\\
&= 2^{\frac{n}{2}-1}\int_0^\infty u^{\frac{n}{2}-2} e^{-u} du && (\text{pull new base 2 constant})\\
&= 2^{\frac{n}{2}-1}\int_0^\infty u^{(\frac{n}{2}-1)-1} e^{-u} du && (\text{linearity of addition})\\
&= 2^{\frac{n}{2}-1}\,\Gamma\left(\frac{n}{2}-1\right) && (\text{Gamma function substitution } z = \tfrac{n}{2}-1)
\end{aligned}$$

   Doing some factorial algebra, $\Gamma\left(\frac{n}{2}-1\right) = \frac{\Gamma\left(\frac{n}{2}\right)}{n/2}$, so the above equals

$$\frac{2^{\frac{n}{2}-1}}{n/2}\Gamma\left(\frac{n}{2}\right).$$

   (b) Via the definition, compute the expected value $\mathrm{E}\left(\frac{1}{X}\right)$.

Solution:

$$\begin{aligned}
\mathrm{E}\left(\frac{1}{X}\right) &= \int_0^\infty \frac{1}{x} f(x)dx \\
&= \int_0^\infty \frac{1}{x}\left(\frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}}\right) x^{\frac{n}{2}-1}e^{-\frac{x}{2}} \\
&= \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}}\int_0^\infty \frac{1}{x} x^{\frac{n}{2}-1}e^{-\frac{x}{2}} \\
&= \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}}\int_0^\infty x^{\frac{n}{2}-2}e^{-\frac{x}{2}} \\
&= \frac{1}{\Gamma(\frac{n}{2})2^{\frac{n}{2}}} \cdot 2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}-1\right) && \text{(substitution from (a))} \\
&= \frac{1}{(\frac{n}{2}-1)\Gamma(\frac{n}{2}-1)2^{\frac{n}{2}}} \cdot 2^{\frac{n}{2}-1}\Gamma\left(\frac{n}{2}-1\right) \\
&= \frac{1}{n-2} && \text{(“lots of killing” - (McAsey))}
\end{aligned}$$

(c) Let $X_1 \sim \chi^2(m)$ and $X_2 \sim \chi^2(n)$ be independent random variables. Define the random variable $F = \dfrac{X_1/m}{X_2/n}$. Use the previous results to compute the mean of F. (Note $F \sim F(m,n)$.)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The data is as follows:

x_1	2	1	1	-3	-1	0	-2	0	2
x_2	-1	1	-2	0	1	1	0	-2	2
y	3	6	5	8	12	7	4	11	10

(a) Determine the matrices $\mathbf{X}$, $\mathbf{X}^T$, $\mathbf{X}^T\mathbf{X}$, $\mathbf{X}^T\mathbf{Y}$, and $(\mathbf{X}^T\mathbf{X})^{-1}$

(b) Fit the model to the data

(c) Determine the residual for the data point $(1, 1, 6)$

(d) Use matrices to calculate SSE and determine S. (Note $\sum y^2 = 564$.)

(e) Test $H_0 : \beta_2 = 0$ versus $\beta_2 < 0$. Perform the test at $\alpha = 0.05$ level. What can you conclude about the usefulness of the x_2 term in the model?

(f) Find a 99% confidence interval for $\mathrm{E}(Y)$ when $x_1 = 1$ and $x_2 = 2$.

$$\frac{\left[(600 * 1) + (300 * 4) + (100 * 2)\right] * 10^3}{1{,}000{,}000} = 2$$

$$\frac{2 \times 10^9}{2 \times 10^6} = 10^3 = 1000$$

$$\underbrace{1{,}000{,}000}_{\text{IC}} \times \underbrace{\frac{2 \times 10^6}{2 \times 10^9}}_{\text{CPI / f}} = 1000\mu\text{s}$$