Question 1

In an experiment to investigate the performance of four brands of spark plugs, five plugs of each brand were tested for the amount of time until failure. A partial ANOVA table for the data is given in *Excel*.

(a) Find the missing entries in the table.
(b) State the null and alternative hypotheses.
(c) Carry out the hypothesis test using $\alpha = 0.05$.

(Hints: please show all relevant steps, and round your answers to 3 decimal places; attach an *Excel* screenshot if it helps, but no need to submit *Excel*.)

. . . . . .

(a) **Answer:**
Each cell in the table refer to a certain statistic, as shown below.

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	SSA	k-1	MSA	MSA/MSE	=1-f.dist(F,k-1,N-k,1)	=f.inv(1-0.05,k-1,N-k,1)
Within Groups	SSE	N-k	MSE = 14713.69			
Total	SST = 310500.76	N-1				

Thus, using the formulas given in class, we can deduce the values of these statistics. Given $N = 20$ plugs and $k = 4$ brands, we have $k - 1 = 3$, $N - k = 16$, and $N - 1 = 19$. Now we deduce that:

$$\begin{aligned} \text{MSE} &= \text{SSE}/(N-k) \\ 14713.69 &= \text{SSE}/16 \\ \boxed{\text{SSE} = 235419.04} & \end{aligned}$$

Since $\text{SST} = \text{SSA} + \text{SSE}$, we have:

$$\begin{aligned} \text{SSA} &= \text{SST} - \text{SSE} \\ &= 310500.76 - 235419.04 \\ \boxed{\text{SSA} = 75081.72} & \end{aligned}$$

Now we can find the MSA since $\text{MSA} = \text{SSA}/(k-1)$.

$$\begin{aligned} \text{MSA} &= \text{SSA}/(k-1) \\ &= 75081.72/3 \\ \boxed{\text{MSA} = 25027.24} & \end{aligned}$$

Since $F = \text{MSA}/\text{MSE}$:

$$\begin{aligned} F &= \text{MSA}/\text{MSE} \\ &= 25027.24/14713.69 \\ \boxed{F = 1.701} & \end{aligned}$$

Thus, we now simply use the *Excel* formulas given above and find the p-value and the F critical value. The table is now complete.

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	75081.72	3	25027.24	1.701	0.207	3.239
Within Groups	235419.04	16	14713.69			
Total	310500.76	19				

(b) **Answer:** Let μ_i denote the performance of the spark plugs for $i = 1, 2, 3, 4$. The null and alternate hypotheses are:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1 : \neg H_0$$

If H_1 is true, it means that at least one of the spark plugs have a different mean performance.

(c) **Answer:** To test the hypothesis, simply compare the value of α with the p-value. Here, we see that $p = 0.207$ and $\alpha = 0.05$. Thus $p > \alpha$ and we do not have sufficient evidence at the 5% significance level to reject H_0.

Question 2

Water salinity measurements (in parts per thousand) at four sites are given in *Excel*, and the corresponding ANOVA table is also given. However, one data entry (marked by x) has been accidentally deleted.

(a) If you were to apply the Bonferroni method here, then what would be the value of m?
(b) Find x, by setting up a quadratic equation involving SSA.

(Hints: for (a), you only need to find the value of m, no need to carry out the actual pairwise comparisons. For (b), you may use technology to solve the quadratic (please describe the method used); you may get 2 solutions, but it is possible to figure out which is the correct solution.)

. . . . . .

(a) **Answer:** Given $k = 4$ sites, we will have $m = \binom{4}{2} = 6$ pairs of test if we apply the Bonferroni method here.

(b) **Answer:** The formula for SSA is given by:

$$\text{SSA} = \sum_{i=1}^{4} n_i(\bar{y}_i - \bar{\bar{y}})^2$$

From the table, we have SSA = 40.46382259. We can now use this information along with some *Excel* calculations to form the following quadratic equation:

	Water salinity at four sites			
	site I	**site II**	**site III**	**site IV**
	37.54	40.17	39.04	38.94
	37.01	40.8	39.21	39.53
	36.71	39.76	39.05	39.18
	37.03	39.7	38.24	38.38
	37.32	40.79	38.53	38.92
	37.01	40.44	38.71	38.65
	37.03	39.79	38.89	39.96
	37.7	39.38	38.66	38.65
	37.36		38.51	39.38
	36.75		40.08	
	37.45			
	x			
Sum	408.91+x	320.83	388.92	351.59
Count	12	8	10	9
Mean	(408.91+x) / 12	40.10375	38.892	39.06556

$$40.46382259 = 12\left(\frac{408.91+x}{12} - \bar{\bar{y}}\right)^2 + 8(40.10375 - \bar{\bar{y}})^2 + 10(38.892 - \bar{\bar{y}})^2 + 9(39.06555556 - \bar{\bar{y}})^2$$

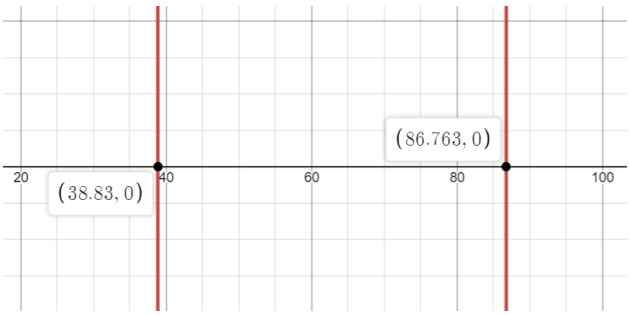
By definition, $\bar{\bar{y}}$ is the grand mean and thus:

$$\bar{\bar{y}} = \frac{1}{N}\sum_{i,j} y_{ij} = \frac{1470.25+x}{39}$$

Now, we can rearrange the equation and get:

$$40.46382259 = 12\left(\frac{408.91+x}{12} - \frac{1470.25+x}{39}\right)^2 + 8\left(40.10375 - \frac{1470.25+x}{39}\right)^2 + 10\left(38.892 - \frac{1470.25+x}{39}\right)^2 + 9\left(39.06555556 - \frac{1470.25+x}{39}\right)^2$$

which forms a quadratic equation. Using *Desmos* (https://www.desmos.com/calculator/jgy0knmdav), we find that $x = 38.83$ or $x = 86.763$.



As x only takes 1 value, we can plug in x to the table and run the ANOVA test again.

site I	site II	site III	site IV
37.54	40.17	39.04	38.94
37.01	40.8	39.21	39.53
36.71	39.76	39.05	39.18
37.03	39.7	38.24	38.38
37.32	40.79	38.53	38.92
37.01	40.44	38.71	38.65
37.03	39.79	38.89	39.96
37.7	39.38	38.66	38.65
37.36		38.51	39.38
36.75		40.08	
37.45			
38.83			

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	12	447.74	37.31166667	0.322542424
Column 2	8	320.83	40.10375	0.282369643
Column 3	10	388.92	38.892	0.260928889
Column 4	9	351.59	39.06555556	0.247502778

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	40.46382259	3	13.48794086	47.91241022	1.76215E-12	2.874187484
Within Groups	9.852936389	35	0.281512468			
Total	50.31675897	38				

Figure 1: ANOVA for $x = 38.83$

site I	site II	site III	site IV
37.54	40.17	39.04	38.94
37.01	40.8	39.21	39.53
36.71	39.76	39.05	39.18
37.03	39.7	38.24	38.38
37.32	40.79	38.53	38.92
37.01	40.44	38.71	38.65
37.03	39.79	38.89	39.96
37.7	39.38	38.66	38.65
37.36		38.51	39.38
36.75		40.08	
37.45			
86.763			

SUMMARY

Groups	Count	Sum	Average	Variance
Column 1	12	495.673	41.30608333	205.0193295
Column 2	8	320.83	40.10375	0.282369643
Column 3	10	388.92	38.892	0.260928889
Column 4	9	351.59	39.06555556	0.247502778

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	40.4629008	3	13.4876336	0.208739113	0.889663018	2.874187484
Within Groups	2261.517595	35	64.61478842			
Total	2301.980495	38				

Figure 2: ANOVA for $x = 86.763$

Evidently, the ANOVA table for $x = 38.83$ matches the given table whereas $x = 86.763$ doesn't. Thus, we can be sure that $x = 38.83$.

Question 3

The Olympic men's triple jump winning distances can be found in *Excel*. Compute a 90% two-sided prediction interval for the winning distance corresponding to the year 1940.

(Hints: you may use *Excel* to find s, $\hat{y}$, etc, but no need to submit *Excel*; please give your answer to 3 decimal places.)

. . . . . .

Answer: The $(100 - \alpha)\%$ prediction interval for the winning distance in 1940 is given by:

$$\text{PI} = \left[\hat{y}^* - t_{n-2,\alpha/2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^*_{1940} - \bar{x})^2}{(n-1)s_x^2}}, \hat{y}^* + t_{n-2,\alpha/2} \cdot s\sqrt{1 + \frac{1}{n} + \frac{(x^*_{1940} + \bar{x})^2}{(n-1)s_x^2}}\right]$$

where, after using *Excel* to do linear regression, we find:

$$\begin{aligned} \hat{y}^* &= 15.7911460591405 \\ n &= 29 \\ \alpha &= 0.1 \\ t_{n-2,\alpha/2} = t_{27,0.05} &= 1.703288446 \\ s &= 0.385404865 \\ x^* &= 1940 \\ \bar{x} &= 1960.58620689655 \\ s_x^2 &= 1474.82266009852 \end{aligned}$$

Substituting these values into the PI formula gives:

$$\text{PI} = \left[15.7911460591405 - 1.703288446 \cdot 0.385404865\sqrt{1 + \frac{1}{29} + \frac{(1940 - 1960.58620689655)^2}{(28)1474.82266009852}},\right.$$
$$\left. 15.7911460591405 + 1.703288446 \cdot 0.385404865\sqrt{1 + \frac{1}{29} + \frac{(1940 - 1960.58620689655)^2}{(28)1474.82266009852}}\right]$$

Hence our 2-sided 90% prediction interval:

$$\text{PI} = [15.65228544, 15.93000668]$$
$$\boxed{\text{PI} \approx [15.652, 15.930]}$$

Question 4

Please refer to the *Excel* sheet which contains (real) economic data for the USA. Our goal is to find the best way to model Y (unemployment) based on x_1 (interest rate) and/or x_2 (inflation). There are three linear regression models to consider: (1) Y vs x_1, (2) Y vs x_2, (3) Y vs x_1 and x_2.

(a) For each model, use *Excel* to compute SSE, then use a formula to compute AIC.
(b) Hence, determine the best model out of the three.

(Hint: no need to submit *Excel*, but please show enough steps so that a reader can understand how you arrived at your answers.)

. . . . . .

(a) **Answer:** For each of the three models, we can compute the SSE using *Excel*'s Data Analysis ToolPak. By performing a regression analysis, we get the ANOVA table which gives the SSE (3 decimal places).

$\text{SSE}_1 = 32.61954198$ $\quad$ $\text{SSE}_2 = 52.34313555$ $\quad$ $\text{SSE}_3 = 32.32.57028307$

$\boxed{\text{SSE}_1 = 32.620}$ $\quad$ $\boxed{\text{SSE}_2 = 52.343}$ $\quad$ $\boxed{\text{SSE}_3 = 32.570}$

To find the AIC of each model, use the following formula.

$$\text{AIC} = n \ln\left(\frac{\text{SSE}}{n}\right) + 2(m+1)$$

Given $n = 28$, $m_1 = m_2 = 1$, and $m_2 = 2$, we have:

$\text{AIC}_1 = 8.275797286$ $\text{AIC}_2 = 21.51725619$ $\text{AIC}_3 = 10.23348241$

$\boxed{\text{AIC}_1 = 8.276}$ $\boxed{\text{AIC}_2 = 21.517}$ $\boxed{\text{AIC}_3 = 10.233}$

(b) **Answer:** In terms of *information loss*, model 1 has the best performance as it has the lowest AIC.

Question 5

In *Excel*, you can find a data set containing experimental results from an early attempt to measure the speed of light. (We have worked with this data set in Modelling Uncertainty.) Resample 10^4 times, construct a histogram of the resample means, then find a 99% two-sided bootstrap confidence interval for the true mean.

(Hints: you can copy and paste the data into *Python*. Please suitably modify the *Python* code given in class, without using the '`bootstrap`' function itself. You only need to submit a screenshot of the code and the histogram, and the CI to 3 decimal places.)

Answer: Using *Python*, we get the 99% 2-sided CI:

$$\text{CI} = [22.151515151515152, 28.87878787878788]$$

$$\boxed{\text{CI} = [22.152, 28.879]}$$

```python
import math, statistics, random
import matplotlib.pyplot as plt
from scipy.stats import t, bootstrap

data = [28,26,33,24,34,-44,27,16,40,-2,29,22,24,21,25,30,23,29,31,19,24,20,36,
32,36,28,25,21,28,29,37,25,28,26,30,32,36,26,30,22,36,23,27,27,28,27,31,27,26,
33,26,32,32,24,39,28,24,25,32,25,29,27,28,29,16,23]
n = len(data)

bootstrap_mean = []
for i in range(10**4):
    resample = random.choices(data, k=n)
    resample_mean = statistics.mean(resample)
    bootstrap_mean.append(resample_mean)
bootstrap_mean.sort()

a = 0.01
x = 10**4
x_L, x_U = int(a/2 * x), int((1-a/2)*x  )
lower_bound = bootstrap_mean[x_L]
upper_bound = bootstrap_mean[x_U]
print(f"Lower Bound = {lower_bound}")
print(f"Upper Bound = {upper_bound}")
```

[17] ✓ 0.2s Python

```
Lower Bound = 22.151515151515152
Upper Bound = 28.87878787878788
```

```python
plt.hist(bootstrap_mean, bins = 20, ec = 'black')
plt.xticks([lower_bound, upper_bound], ['L', 'U'], minor = True)
plt.show()
```

[18] ✓ 0.1s Python

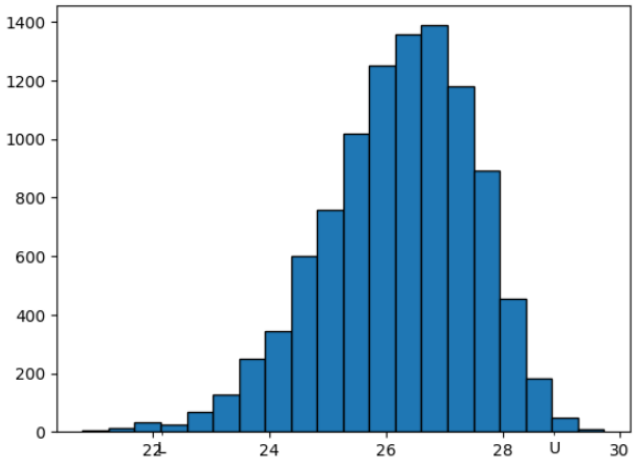


Figure 3: *Python* Code for Question 5