UTM UNIVERSITI TEKNOLOGI MALAYSIA RESEARCH UNIVERSITY	**UNIVERSITI TEKNOLOGI MALAYSIA**

Probability & Statistical Data Analysis (SECI 1143)

ASSIGNMENT 4

CHAPTER 7 & CHAPTER 8

Student Name : 1. Angela Ngu Xin Yi (A24CS0226)
2. Tan Xin Tian (A24CS0198)

Lecturer Name : Dr Shamini A/P Raja Kumaran

Section : 2

1. (a)

X	Y	XY	X^2	Y^2
35	48	1680	1225	2304
50	65	3250	2500	4225
45	60	2700	2025	3600
60	72	4320	3600	5184
70	83	5810	4900	6889
55	62	3410	3025	3844
40	50	2000	1600	2500
65	75	4875	4225	5625
75	90	6750	5625	8100
80	95	7600	6400	9025
Total: 575	700	42395	35125	51296

$$r = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\sqrt{[(\Sigma x^2) - (\Sigma x)^2/n][(\Sigma y^2) - (\Sigma y)^2/n]}}$$

$$= \frac{42395 - (575 \times 700)/10}{\sqrt{\left(35125 - \frac{575^2}{10}\right)\left(51296 - \frac{700^2}{10}\right)}}$$

$$= 0.986$$

(b) $r = 0.986$, relatively strong positive linear association between the number of pastries produced and the production cost.

2. (a)

No.	Engagement Score	Sentiment Score	Ranks of Engagement Score	Ranks of Sentiment Score	d_i	d_i^2
1	85	4	4	4	0	0
2	70	3	2	2.5	-0.5	0.25
3	90	5	6	5.5	0.5	0.25
4	60	2	1	1	0	0
5	88	5	5	5.5	-0.5	0.25
6	75	3	3	2.5	0.5	0.25

$\Sigma d_i^2 = 1$

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{(6)(1)}{6(6^2-1)}$$

$$= 0.971$$

$\therefore$ $r_s = 0.971$, relatively strong positive linear association between Engagement and Sentiment Score.

(b)

	(x) Likes	(y) Shares	XY	X^2	Y^2
	150	30	4500	22500	900
	100	20	2000	10000	400
	200	40	8000	40000	1600
	80	15	1200	6400	225
	170	35	5950	28900	1225
	120	25	3000	14400	625
Total :	820	165	24650	122200	4975

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

$$= \frac{24650 - (820 \times 165)/6}{\sqrt{\left(122200 - \frac{820^2}{6}\right)\left(4975 - \frac{165^2}{6}\right)}}$$

$$= 0.997$$

$r = 0.997$, nearly perfect strong positive linear relationship between Like and Share

(c) $H_0 : \rho = 0$ (no linear correlation)

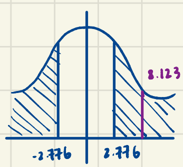
$H_1 : \rho \neq 0$ (linear correlation exists)

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.971}{\sqrt{\frac{1-0.971^2}{6-2}}} = 8.123$$

P-value = 0.01

critical value : $\alpha = 0.05$, d.f = 6 - 2 = 4

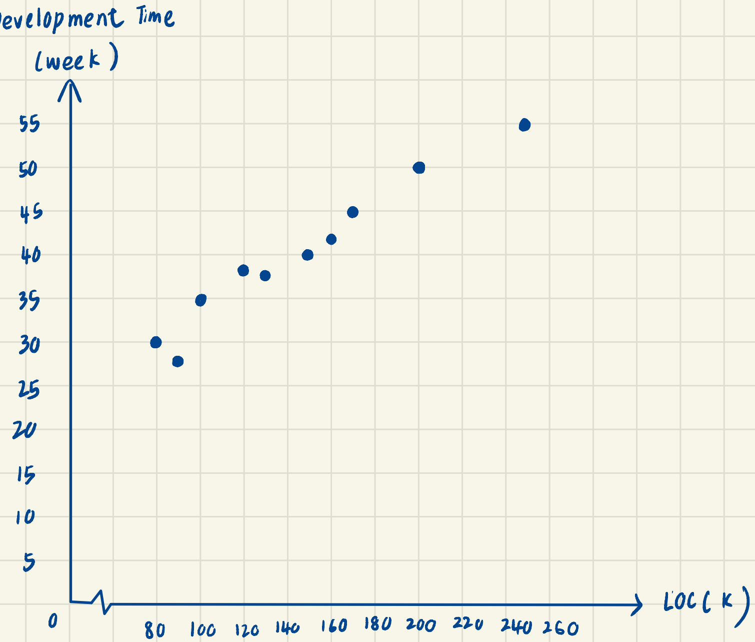
$t_{\alpha/2} = t_{0.025,4} = 2.776$



$\therefore$ Since 8.123 > 2.776, reject H_0.

There is sufficient evidence to support the claim that there is a significant linear correlation between Engagement and Sentiment Score.

3. (a)



(b)

	(x) LOC(K)	(y) Development Time(week)	xy	x^2	y^2
	150	40	6000	22500	1600
	100	35	3500	10000	1225
	200	50	10000	40000	2500
	80	30	2400	6400	900
	170	45	7650	28900	2025
	120	38	4560	14400	1444
	160	42	6720	25600	1764
	90	28	2520	8100	784
	250	55	13750	62500	3025
	130	37	4810	16900	1369
Total :	1450	400	61910	235300	16636

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - (\sum x)^2/n][(\sum y^2) - (\sum y)^2/n]}}$$

$$= \frac{61910 - (1450 \times 400)/10}{\sqrt{\left(235300 - \frac{1450^2}{10}\right)\left(16636 - \frac{400^2}{10}\right)}}$$

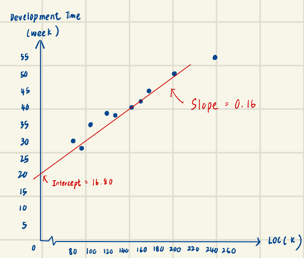
$$= 0.9795$$

$$= 0.98$$

∴ $r = 0.98$, relatively strong positive linear association between length of code and development time

(c) $$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}} = \frac{61910 - (1450 \times 400 \div 10)}{235300 - \frac{1450^2}{10}} = 0.16$$

$$b_0 = \bar{y} - b_1\bar{x} = \left(\frac{400}{10}\right) - (0.16)\left(\frac{1450}{10}\right) = 16.80$$



∴ The regression equation is : $\hat{y} = 16.80 + 0.16x$

∴ b_0 (16.80) is the estimated average value of development time when the value of LOC is 0

Here no had zero LOC, so b_0 (16.80) just indicates that, for LOC within the range of sizes observed, 16.80 week is the portion of development time not explained by LOC.

∴ b_1 (0.16) measures the estimated change in the average value of development time as a result of one-unit change in LOC.

For every 1K increase in LOC, development time increases by about 0.16 weeks.

(d) $x = 180K$

$\hat{y} = 16.80 + 0.16x$

$= 16.80 + 0.16(180)$

$= 45.60$

$\therefore$ The predicted time for a project with 180K LOC is 45.60 weeks

(e)

(x) LOC (K)	(y) Development Time (week)	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$
150	40	$16.80 + (0.16)(150) = 40.80$	0	0.64
100	35	$16.80 + (0.16)(100) = 32.80$	25	51.84
200	50	$16.80 + (0.16)(200) = 48.80$	100	77.44
80	30	$16.80 + (0.16)(80) = 29.60$	100	108.16
170	45	$16.80 + (0.16)(170) = 44.00$	25	16.00
120	38	$16.80 + (0.16)(120) = 36.00$	4	16.00
160	42	$16.80 + (0.16)(160) = 42.40$	4	5.76
90	28	$16.80 + (0.16)(90) = 31.20$	144	77.44
250	55	$16.80 + (0.16)(250) = 56.80$	225	282.24
130	37	$16.80 + (0.16)(130) = 37.60$	9	5.76
Total: 1450	400		SST = $\Sigma(y-\bar{y})^2 = 636$	SSR = $\Sigma(\hat{y}-\bar{y})^2 = 641.28$

$\bar{x} = 145$ $\quad \bar{y} = 40$

$\therefore$ SST = 636.00

$\therefore$ SSR = 641.28

$\therefore$ $R^2 = \dfrac{SSR}{SST} = \dfrac{641.28}{636} = 1.01 \approx 1$

$\therefore$ $R^2 = 1$ means that approximately 100% of variation in development time is explained by variation in LOC.

Not round off to 2d.p in steps

e)

x	$\hat{y}$
150	$17.37 + 0.16(150) = 40.780$
100	32.976
200	48.584
80	29.854
170	43.902
120	36.098
160	42.341
90	31.415
250	56.389
130	37.658

$$SSR = (40.780-40)^2 + (32.976-40)^2 + (48.584-40)^2 + (29.854-40)^2 + (43.902-40)^2 + (36.098-40)^2 + (42.341-40)^2 + (31.415-40)^2 + (56.389-40)^2 + (37.658-40)^2$$

$$= 610.289 \approx 610.29$$

$$SST = (40-40)^2 + (35-40)^2 + (50-40)^2 + (30-40)^2 + (45-40)^2 + (38-40)^2 + (42-40)^2 + (28-40)^2 + (55-40)^2 + (37-40)^2$$

$$= 636 \approx 636.00$$

$$R^2 = \frac{SSR}{SST}$$

$$= \frac{610.289}{636}$$

$$= 0.9596$$

$$\approx 0.96$$

$\therefore$ $R^2 = 0.96$

96% of variation in development time is explained by LOC.

Question 4

a) $H_0 : \mu_1 = \mu_2 = \mu_3$

H_1 : At least one mean is different

b) Fertilizer A

$$\bar{x} = \frac{262 + 246 + 266 + 288 + 244}{5}$$

$$= 261.200$$

$$s^2 = \frac{(262-261.2)^2 + (246-261.2)^2 + (266-261.2)^2 + (288-261.2)^2 + (244-261.2)^2}{5-1}$$

$$= 317.200$$

Fertilizer B

$$\bar{x} = \frac{235 + 271 + 255 + 230 + 250}{5}$$

$$= 248.200$$

$$s^2 = \frac{(235-248.2)^2 + (271-248.2)^2 + (255-248.2)^2 + (230-248.2)^2 + (250-248.2)^2}{5-1}$$

$$= 268.700$$

Fertilizer C

$$\bar{x} = \frac{223 + 223 + 233 + 201 + 204}{5}$$

$$= 216.800$$

$$s^2 = \frac{(223-216.8)^2 + (223-216.8)^2 + (233-216.8)^2 + (201-216.8)^2 + (204-216.8)^2}{5-1}$$

$$= 188.200$$

c)

$$\bar{\bar{x}} = \frac{261.2 + 248.2 + 216.8}{3}$$

$$= 242.067$$

$$\bar{\bar{s}}^2 = \frac{(261.2-242.067)^2 + (248.2-242.067)^2 + (216.8-242.067)^2}{3-1}$$

$$= 521.053$$

$$ns_{\bar{x}}^2 = 5(521.053)$$

$$= 2605.267$$

$$s_p^2 = \frac{317.2 + 268.7 + 188.2}{3}$$

$$= 258.033$$

$$F = \frac{ns_{\bar{x}}^2}{s_p^2}$$

$$= \frac{2605.267}{258.033}$$

$$= 10.097$$

d) numerator $= k-1$

$= 3-1$

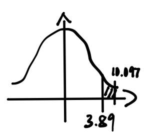
$= 2$

denominator $= k(n-1)$

$= 3(5-1)$

$= 12$

e) $F_{2,12,0.05} = 3.890$

f) 

$10.097 > 3.890$, so reject H_0.

There is no sufficient evidence to support the claim that different types of fertilizers have same mean for the plant height after 30 days.

Question 5

a) $H_0 : \mu_L = \mu_M = \mu_N = \mu_o$

H_1 : at least one mean is different

b)
$$\bar{x}_L = \frac{35.5 + 34.8 + 36.1 + 35.2 + 34.9}{5}$$

$= 35.300$

$$s_L^2 = \frac{(35.5-35.3)^2 + (34.8-35.3)^2 + (36.1-35.3)^2 + (35.2-35.3)^2 + (34.9-35.3)^2}{5-1}$$

$= 0.275$

$$\bar{x}_m = \frac{38.2 + 37.7 + 38.5 + 37.9 + 38.0}{5}$$

$= 38.060$

$$s_m^2 = \frac{(38.2-38.06)^2 + (37.7-38.06)^2 + (38.5-38.06)^2 + (37.9-38.06)^2 + (38-38.06)^2}{5-1}$$

$= 0.093$

$$\bar{x}_N = \frac{33.9 + 34.2 + 33.7 + 34.0 + 33.5}{5}$$

$= 33.860$

$$s_N^2 = \frac{(33.9-33.86)^2 + (34.2-33.86)^2 + (33.7-33.86)^2 + (34-33.86)^2 + (33.5-33.86)^2}{5-1}$$

$= 0.073$

$$\bar{x}_o = \frac{36.8 + 37.0 + 36.5 + 36.9 + 37.1}{5}$$

$= 36.860$

$$s_o^2 = \frac{(36.8-36.86)^2 + (37-36.86)^2 + (36.5-36.86)^2 + (36.9-36.86)^2 + (37.1-36.86)^2}{5-1}$$

$= 0.053$

c)
$$\bar{\bar{x}} = \frac{35.3 + 38.06 + 33.86 + 36.86}{4}$$

$= 36.020$

$$s_{\bar{x}}^2 = \frac{(35.3-36.02)^2 + (38.06-36.02)^2 + (33.86-36.02)^2 + (36.86-36.02)^2}{4-1}$$

$= 3.3504 \approx 3.350$

c) $ns_{\bar{x}}^2 = 5(3.3504)$

$= 16.752$

$s_p^2 = \frac{0.275 + 0.093 + 0.073 + 0.053}{4}$

$= 0.1235 \approx 0.124$

$F = \frac{ns_{\bar{x}}^2}{s_p^2}$

$= \frac{16.752}{0.1235}$

$= 135.644$

d) numerator $= k-1$

$= 4-1$

$= 3$

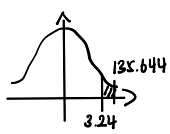
denominator $= k(n-1)$

$= 4(5-1)$

$= 16$

e) $F_{3,16,0.05} = 3.240$

f)



$135.644 > 3.240$, so reject H_0.

There is no sufficient evidence to support the claim that different brands of brake tires have the same mean for stopping distances.