Statistics Assignment Group Hi6007

Holmes Institute

Faculty of Higher Education

SUBJECT: STATISTICS

WORD COUNT: 2430

DEADLINE: 22/01/2018

REFERENCE STYLE: APA

COUNTRY: AUSTRALIA

Question 1

Table 1: F-test

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Between treatments	90	3	*__30___?*	*__5___?*
Within treatments (Error)	120	20	*__6___?*	
Total	*__210___?*	*__23___?*		

a)

Mean of squares (Between treatments) = 90 ÷3

= 30

Mean of squares (Within treatments) = 120 ÷20

= 6

Total sum of squares = 90 + 120 = 120

Total sum of squares = 3 + 20 = 23

b) F- value= (Mean square between treatments) ÷ (Mean square within treatments)

F- value = 30 ÷ 6

= 5

c) What has been the total number of observations?

Number of observations = Degrees of freedom + 1

= (3 + 20) + 1

= 27 observations.

Question 2

Develop a linear trend expression and project the sales (the number of cars sold) for time period $t = 11$.

Let the number of cars sold be represented x, then, the linear trend will be expressed as, $y = a + bt$, where a is a constant, t is the time period in years and x is the number of cars cold in

a thousand units in a particular period and y is the total number of cars sold in a given time period (Lun 2017). Therefore, the estimated linear equation will be represented as;

$\hat{y} = \alpha + \beta t$

Table 2: Trend line estimation

ti	yi	$ti - \bar{t}$	$yi - \bar{y}$	$(ti - \bar{t})(yi - \bar{y})$	$(ti - \bar{t})^2$	$\hat{y} = 136 + 39.181818ti$
1	195	-4.5	-156.5	704.25	20.25	175.1818
2	200	-3.5	-151.5	530.25	12.25	214.3636
3	250	-2.5	-101.5	253.75	6.25	253.5455
4	270	-1.5	-81.5	122.25	2.25	292.7273
5	320	-0.5	-31.5	15.75	0.25	331.9091
6	380	0.5	28.5	14.25	0.25	371.0909
7	440	1.5	88.5	132.75	2.25	410.2727
8	460	2.5	108.5	271.25	6.25	449.4545
9	500	3.5	148.5	519.75	12.25	488.6364
10	500	4.5	148.5	668.25	20.25	527.8182
$\sum ti$ **=55**	$\sum yi$ **=3515**	$\sum ti - \bar{t}$ **=0**	$\sum yi - \bar{y}$ **=0**	$\sum(ti - \bar{t})(yi - \bar{y})$ **=3232.50**	$\sum (ti - \bar{t})^2$ **=82.50**	$\sum \hat{y}$ **=3515**

$$\bar{t} = \frac{\sum ti}{n} = 55/10 = 5.5$$

$\bar{y} = \frac{\sum yi}{n}$, where n is the number of observations (the sample)

$$\bar{y} = \frac{\sum yi}{n} = \frac{3515}{10} = 351.5$$

$yi = a + bti$ - Actual trend line equation

$\hat{y} = \alpha + \beta t$ - Actual trend line equation.

a in the actual trend line is the constant that does not have any effect on the number of cars sold as time changes.

$$b = \frac{\sum(ti - \bar{t})(yi - \bar{y})}{\sum(ti - \bar{t})^2}$$

$$= 3232.50 / 82.50 = 39.181818$$

$$a = \bar{y} - b\bar{t}$$

$$= 351.5 - 39.181818(5.5)$$

$$= 351.5 - 215.5$$

$$a = 136$$

$$\hat{y} = 136 + 39.181818ti$$

The estimated equation determines how the number of cars sold relates with time in years.

yi is the observed y while $\hat{y}$ is the estimated y.

According to the table generated above, for every value of t, there is an estimated y. The summation of the actual number of vehicles sold and summation of the number of vehicles estimated to be sold are the same. However, these values vary each year.

$$y = 136 + 39.181818ti$$

When t = 11

$$y = 136 + 136 + 39.181818(11)$$

$$y = 136 + 490.9999$$

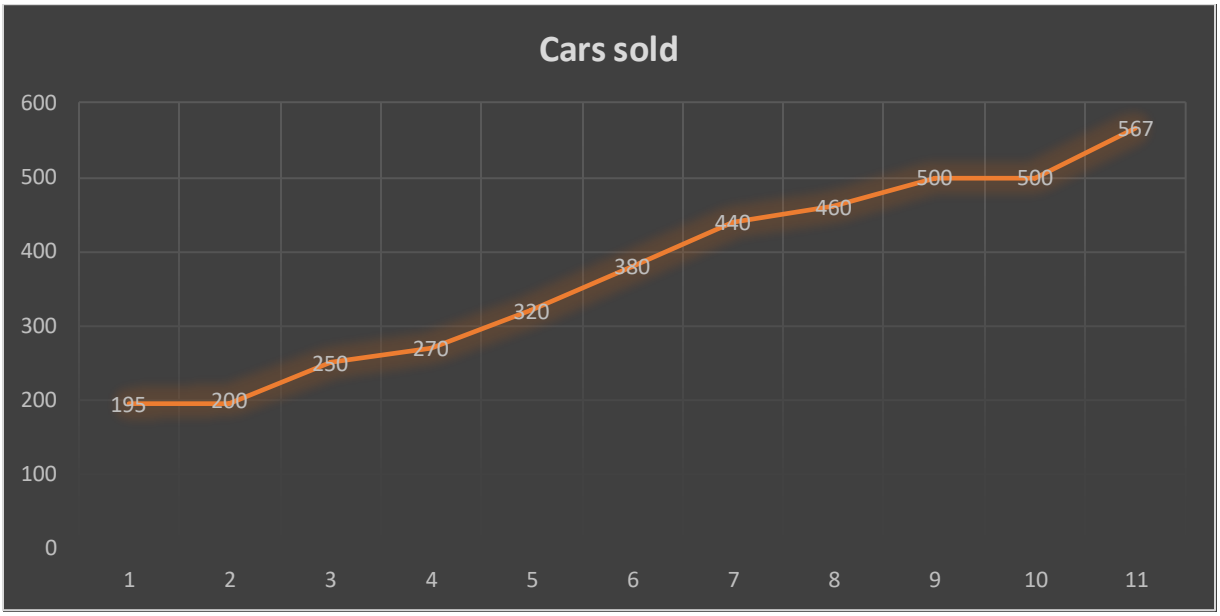
$$= 566.9999$$

$$= 566.9999 \times 1000$$

$$y = 566999.9$$

$$y = 567000$$

Therefore, the number of vehicles sold for the time period t = 11, will be 567000 vehicles as shown in the trend line below. A plot of the number of vehicles sold against time in years.



At the 11th year, 567 cars will be sold.

Question 3

a) F-test

Table 3: F-Test Two-Sample for Variances

	Variable 1	*Variable 2*
Mean	35	4.285714286
Variance	11.66666667	7.238095238
Observations	7	7
df	6	6
F	1.611842105	
P(F<=f) one-tail	0.288286332	
F Critical one-tail	8.46612534	

The calculated F-value is 1.611 is < the critical value (8.466) at .01 significance level. Thus, we fail reject the null hypothesis and conclude that price and the number of flash drives sold are not related. This is further confirmed by the p-value (0.288)

a) T-test

Table 4: T-test: Paired Two Sample for Means

	Variable 1	*Variable 2*
Mean	35	4.285714286
Variance	11.66666667	7.238095238
Observations	7	7
Pearson Correlation	-0.924982219	
Hypothesized Mean Difference	0	
Df	6	
t Stat	13.56167756	
P(T<=t) one-tail	4.98633E-06	
t Critical one-tail	3.142668403	
P(T<=t) two-tail	9.97267E-06	
t Critical two-tail	3.707428021	

The calculated t Stat is > than the t Critical. Thus, the null hypothesis that the price and number of flash disks sold are not related is rejected. Therefore, it can be concluded that the price number of flash drives sold are related. This is further confirmed by p-value (<.01) which is significant at 1%.

Question 4.

Table 5: F-statistics

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	*F*
Between treatments	*3,200*	*4*	800.00	*0.9730*
Within treatments (Error)	*7,400*	9	*822.22*	
Total	10,600	13		

Since the levels of the factors were five, then the degrees of freedom between treatments were 4. Considering the number of experimental units, it means that the degrees of freedom

within treatments will be nine since the total number of degrees of freedom are supposed to be 13. According to Kenny et al. (2015) as degrees of freedom will be calculated as (n – 1)

= 14 -1 = 13.

Since, the degrees of freedom between treatments are 4, the degrees of freedom within treatments will be calculated as follows, 13 – 4 = 9

Sum of squares = degrees of freedom × Mean sum of squares.

$$= 800 \times 4$$

$$= 3200$$

Total sum of squares = Sum of squares between treatments + Sum of squares within treatments. Let the sum of squares within groups be x,

$$10{,}600 = x + 3200$$

$$x = 10600 - 3200$$

$$x = 7400$$

F value = mean of square between treatments divided by mean square within treatments.

$$= 800 \div 822.22$$

$$= 0.9730$$

Question 5

Table 6: Oneway ANOVA table

Number of obs.	=	11	R-squared	=	0.8988
Root MSE	=	2.08167	Adj. R-squared	=	0.8735

Source	Partial SS	df	MS	F	Prob. > F
Model	307.87879	2	153.93939	35.52	0.0001
Store	307.87879	2	153.93939	35.52	0.0001
Residual	34.666667	8	4.3333333		
Total	342.54545	10	34.254545		

The results indicate significant (at better than the 1% level) differences in the average sales of the three stores. However, it cannot be ascertained whether the difference is between only two of the stores or all three stores. To unravel this, unpaired t-test is performed. Tables 2, 3 and 4 present the t-test results.

Table 7: Two-sample t test with equal variances between store 1 and store 2

Group	Obs.	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Store 1	5	45	0.83666	1.870829	42.67706	47.32294
Store 2	3	36.33333	1.452966	2.516611	30.08172	42.58494
Combined	8	41.75	1.729471	4.891684	37.66045	45.83955
Diff		8.666667	1.539601		4.899399	12.43393

diff = mean (Store 1) – mean (Store 2) t = 5.6292

H_o: diff = 0 degrees of freedom = 6

Ha: diff < 0	Ha: diff != 0	Ha: diff > 0
Pr(T < t) = 0.9993	Pr(|T| > |t|) = 0.0013	Pr(T > t) = 0.0007

Results in Table 2 indicates that average sales of store 1 and store 2 are significantly different at 5% significance level (Ha: diff != 0; p-value = 0.0013). Furthermore, results in Table 3 indicates that average sales of store 1 and store 3 are significantly different at 5% significance level (Ha: diff != 0; p-value = 0.0001). Lastly, results in Table 4 indicates that average sales of store 2 and store 3 are not significantly different at 5% significance level (Ha: diff ! = 0; p-value = 0.0001).

Table 8: Two-sample t test with equal variances between store 1 and store 3

Group	Obs.	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Store 1	5	45	0.83666	1.870829	42.67706	47.32294
Store 3	3	33	1.154701	2	28.03172	37.96828
Combined	8	40.5	2.283481	6.45866	35.10043	45.89957
Diff		12	1.398412		8.57821	15.42179

diff = mean (Store 1) – mean (Store 3) t = 8.5812

H_o: diff = 0 degrees of freedom = 6

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.9999 Pr(|T| > |t|) = 0.0001 Pr(T > t) = 0.0001

Table 9: Two-sample t test with equal variances between store 2 and store 3

Group	Obs.	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
Store 2	3	36.33333	1.452966	2.516611	30.08172	42.58494
Store 3	3	33	1.154701	2	28.03172	37.96828
Combined	6	34.66667	1.115547	2.73252	31.79906	37.53427
diff		3.333333	1.855921		-1.819531	8.486197

diff = mean (Store 1) – mean (Store 3) t = 1.7961

H_o: diff = 0 degrees of freedom = 4

Ha: diff < 0 Ha: diff ! = 0 Ha: diff > 0

Pr(T < t) = 0.9265 Pr(|T| > |t|) = 0.1469 Pr(T > t) = 0.0735

Question 6

a) Hypotheses

i. H_o: There is no significant difference in the average sales from the five store, that is, $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_a: There is significant difference in the average sales from the five store, that is $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4 \neq \mu_5$

ii. H_o: There is no significant difference in the average sales from three boxes, that is, $\mu_1 = \mu_2 = \mu_3$

H_a: There is significant differences in the average sales from the three boxes, that is $\mu_1 \neq \mu_2 \neq \mu_3$

b) Anova table

Table 10: Twoway ANOVA table

Number of obs.	=	15	R-squared	=	0.9260
Root MSE	=	15.8572	Adj. R-squared	=	0.8706

Source	Partial SS	df	MS	F	Prob. > F
Model	25188.8	6	4198.1333	16.70	0.0004
Store	721.6	4	180.4	0.72	0.6033
Box	23932.483	2	11966.242	47.59	0.0000
Residual	2011.6	8	251.45		
Total	27200.4	14	1942.8857		

c) Results from the factorial ANOVA indicates that the overall model that was used to fit the data is statistically significant (F = 16.70; p = 0.0004). Store is not statistically significant (F = 0.72; p = 0.6033), implying that average sales in the five different stores is not significantly different. In contrast, the average sales of boxes are statistically significant (F = 47.59; p = 0.0000), suggesting that the average sales differ among the three boxes.

Question 7

The data is unpaired because only one method is used to measure the tyres for tread ware. Therefore, two sample *t*-test can be used to test the mean mileage for the three brands of tyres.

$$t = \frac{\bar{x}_1 - \bar{x}_2 - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Letting μ_1, μ_2 and μ_3 represent the mean sample mean for Brand A, Brand B and Brand C tyres, respectively.

Null hypothesis 1: $H_0 : \mu_1 = \mu_2$

$$t = \frac{37 - 38 - 0}{\sqrt{\frac{3^2}{9} + \frac{4^2}{9}}} = \frac{-1}{1.667} = -0.6$$

Null hypothesis 2: $H_0 : \mu_1 = \mu_3$

$$t = \frac{37 - 33 - 0}{\sqrt{\frac{3^2}{9} + \frac{2^2}{9}}} = \frac{4}{1.20} = 3.333$$

Null hypothesis 3: $H_0 : \mu_2 = \mu_3$

$$t = \frac{38 - 33 - 0}{\sqrt{\frac{4^2}{9} + \frac{2^2}{9}}} = \frac{5}{1.491} = 3.354$$

The $t_{0.5,9}$ critical value from the *t*-table is 1.833. The t-calculated values for a test between Brands A and B, Brands A and C, and Brands B and C, are -0.6, 3.333, and 3.354, respectively. The calculated it value for the difference in the average mileage for Brands A and B does not exceed the *t*-critical value, hence we fail to reject the null hypothesis that average mileage for wear characteristics of Brand A and Brand B are equal. Conversely, the calculated *t*-values for the differences between the average mileage for Brands A and C and Brands B and C exceeds the *t*-critical of 1.833, so the null hypothesis that the difference in the average mileage between

the Brands of tyres is reject and alternative hypothesis that the difference between average mileage of Brands A and C and Brands B and C are statistically different is accepted.

Question 8

Table 11: Moving avaerage

Day	Tips	Moving average
1	18	
2	22	
3	17	19
4	18	19
5	28	21
6	20	22
7	12	20

1^{st} set = (18 + 22 + 17)/3 = 19

2^{nd} set = (22 + 17 + 18)/3 = 19

3^{rd} set = (17 + 18 + 28)/3 = 21

4^{th} set = (18 + 28 + 20)/3 = 22

5^{th} set = (28 + 20 + 12)/3 = 20

a. Compute the mean square error for the forecasts.

Let days be represented by x while tips y.

$y = a + bx$ - actual trend equation.

$\hat{y} = \hat{\alpha} + \hat{\beta} x$ - ,estimated trend equation

$\hat{\beta} = \frac{\text{cov} x, y}{\text{var } x}$, Where covx,y, is the covariance of x and y and varx is the variance of x

$$\hat{\beta} = -11/28$$

$$= -0.393$$

Table 12: Coefficient of determination

xi	yi	$xi-\bar{x}$	$yi-\bar{y}$	$(xi-\bar{x})^2$	$(xi-\bar{x})(yi-\bar{y})$	$\hat{y}$	$(yi-\hat{y})$	$(yi-\hat{y})^2$	$(yi-\bar{y})^2$	$\hat{y}-\bar{y}$	$(\hat{y}-\bar{y})$
1	18	-3	-1.29	9	3.87	20.47	-2.47	6.10	1.66	1.18	1.39
2	22	-2	2.71	4	-5.42	20.08	1.92	3.69	7.34	0.79	0.62
3	17	-1	-2.29	1	2.29	19.68	-2.68	7.18	5.24	0.39	0.15
4	18	0	-1.29	0	0	19.29	-1.29	1.66	1.66	0	0
5	28	1	8.71	1	8.71	18.88	9.12	83.17	75.86	-0.41	0.17
6	20	2	0.71	4	1.42	18.50	1.50	2.25	0.50	-0.79	0.62
7	12	3	7.29	9	-21.87	18.11	-6.11	37.33	53.14	-1.18	1.39
28	**135**	**0**	**-0.03**	**28**	**-11**	**135.01**	**-0.01**	**141.38**	**145.40**	**-0.02**	**4.34**

$$\hat{\alpha} = \bar{y} - \beta \bar{x}$$

$$= \bar{y} \text{ --} 0.393\ (4)$$

$$= 19.29 + 0.393\ (4)$$

$$= 19.29 + 1.572$$

$$\hat{\alpha} = 20.862$$

$$\hat{y} = 20.862 - 0.393\,x$$

$$TSS = \sum(yi - \bar{y})^2$$

$$= 145.40$$

$$SSE = 141.38$$

$$MSE = TSS - SSE$$

$$= 145.40 - 141.38$$

$$= 4.02$$

$$OR$$

$$MSE = \sum(\hat{y} - \bar{y})^2$$

$$= 4.34$$

b. Compute the mean absolute deviation for the forecasts.

$$\text{Mean absolute deviation} = \sqrt{S^2}$$

$$\sqrt{S^2} = \frac{SSE}{n - p}$$

Where, p is the number of independent variables regressed on y. Hence p, = 2

$$\frac{141.38}{7 - 2} = 141.38/5$$

$$\delta = \sqrt{28.276}$$

$$= 5.3175$$

Question 9

Table 13: F-statistics

Source of Variation	Degrees of Freedom	Sum of Squares	Mean *F* Square

Regression	4	283,940.60	0.456690
Error	18	621,735.14	
Total			

a. Compute the coefficient of determination and fully interpret its meaning.

Coefficient of determination is also known as R^2, it explains the variations caused in the model by the explanatory variables (Miljkovic 2017).

R^2 = Residual sum of squares divided by the total sum of squares.

F value = SSR/SSE

= 283940.60 ÷ 621735.14

= 0.456690

TSS = SSR +SSE

TSS = 283940.60 + 621735.14

= 905675.74

$$R^2 = \frac{SSR}{SST} = 283940.60 / 905675.74$$

R^2 = 0.3135

It means that 31.35% of the variations in the model are explained by the explanatory variables while 68.65% of the variations are explained by other factors outside the model hence the error term.

b. Is the regression model significant? Explain what your answer implies. Let α = .05.

The model is not significant. This implies that at 5% significance level, the model does not fit the data well. The explanatory variable does not significantly account for the variations in the model.

c. What has been the sample size for this analysis?

The sample size for analysis = the total degrees of freedom +1

= (4 + 18) +1

=23

Question 10

a) Equation that can be used to predict the price of a stock

$\hat{y} = 118.51 - 0.02(x_1) - 1.57(x_2)$

Note: The intercept (constant) and slope estimates have been rounded to two decimal places

b) Interpretation the coefficients of the estimated regression equation

- An additional share of stocks sold decreases the price of Rawlston Inc. stock by 0.02 units.
- A unit increase in the volume exchange (x_2) on the New York stock Exchange reduces/decrease the price of the Rawlston Inc. stock by 1.57 units.

c) The volume of exchange (x_2) on the New York Stock exchange was statistically significant in influencing the price Rawlston Inc. stock at all confidence levels. It is significant at α = .05 because 0.0018 < 0.05. In contrast, the number of shares of the company's stocks sold (x_1) is insignificant since 0.6176 > 0.05

d) $\hat{y} = 118.51 - 0.0163(94{,}500) - 1.5726(16{,}000{,}000)$

$\hat{y} = 118.51 - 1890 - 25120000$

$\hat{y} = -25{,}163{,}022$

References

Kenny, D. A., Kaniskan, B., & McCoach, D. B. 2015. "The performance of RMSEA in models with small degrees of freedom." *Sociological Methods & Research* 44(3), 486-507.

Lun, A. T., & Smyth, G. K. 2017. "No counts, no variance: allowing for loss of degrees of freedom when assessing biological variability from RNA-seq data." *Statistical Applications in Genetics and Molecular Biology*. 34-46.

Miljkovic, T., & Orr, M. 2017. "An evaluation of the reconstructed coefficient of determination and potential adjustments. ." *Communications in Statistics-Simulation and Computation*, 1-14.