STAT S251F

2020 Spring Take-home Assignment

(Final Examination UG)

Name: Jiawei Wang

Student Number: S12395870

Question 1 (12marks)

(a) (5(i) 2(ii))

(i) Because the sample size ($n=100>30$), and we need to get the standard deviation of the population we use z test:

$$\left(\bar{x}-Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{x}+Z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)=(172.3,\ 174.1) \quad ①$$

from the z-table:

$Z_{0.025}=1.96$ ②

we can get from the Question that:

$n=100$ ③

Combine ① ② ③, we can get

$$\begin{cases} \bar{x}=\dfrac{172.3+174.1}{2}=173.2 \\ \sigma=\dfrac{9}{1.96}\approx 4.5918 \end{cases}$$

(ii) from the z-table, when $CI=99\%$, $Z_{0.005}=2.575$

the 99% confidence interval for the mean hight is: $\left(\bar{x}-Z_{0.005}\frac{\sigma}{\sqrt{n}},\ \bar{x}+Z_{0.005}\frac{\sigma}{\sqrt{n}}\right)$.

$$CI=(172.0176,\ 174.3824) \quad =\left(173.2-2.575\cdot\frac{4.5918}{\sqrt{100}},\ 173.2+2.575\cdot\frac{4.5918}{\sqrt{100}}\right)$$

(b) (2)

from the question, we can know that:

$$2\times Z_{0.05}\cdot\frac{\sigma}{\sqrt{n}}<15 \quad ①$$

$$\sigma=12 \quad ②$$

from z-table:

$$Z_{0.05}=1.645 \quad ③$$

Combine ① ② ③: $n>6.9274$ $\therefore$ the least number of tests is 7.

Question 1 (c) (3)

the sample mean $\overline{x} = \frac{\sum x_i}{n} = \frac{973.5}{110} = 8.85$

the sample standard deviation $S = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 + n\cdot(\overline{x})^2 - 2\cdot\overline{x}\cdot\sum_{i=1}^{n} x_i}{n-1}} = \sqrt{1.5165} \approx 1.2315$

Since we don't know the population standard deviation (σ), and the sample size $n > 30$.

We use t-test

from the t-table $t(0.005, 109) = 2.6217$ ①

$\therefore$ the 99% confidence interval for the population mean μ is $\left(-t_{0.005,109}\cdot\frac{S}{\sqrt{n}} + \overline{x},\ t_{0.005,109}\cdot\frac{S}{\sqrt{n}} + \overline{x}\right)$

$$CI_{99\%} = (8.85 - 0.3078,\ 8.85 + 0.3078) = (8.5422,\ 9.1578)$$

Question 2 (4 marks)

(a) It is a random sampling method.
Because every alumni who attend this gathering has the same probability to be choosed from the box and get the gifts.

(b) I think it is not a random sampling method.
Because this questionaire is only for the person who take the MTR, not for the all Hong kong citizens.
In other words, all the Hong Kong citizens do not have the same probability to poll about the HK goverment's performance during COVID-19.

Question 3 (5 marks)

(a) (4)

Brandy Band	Rank (1st)	Rank (2nd)	$\lvert d\rvert = \lvert R_x - R_y\rvert$	d^2
1	2	3	1	1
2	3.5	2	1.5	2.25
3	1	4.5	3.5	12.25
4	5	4.5	0.5	0.25
5	3.5	1	2.5	6.25
			$\sum d^2$	22

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6 \cdot 22}{5(25-1)} = -0.1$$

(b) (1)

the r_s is -0.1, which means that by normal standards, the association betwee the two variables would not be considered statistically significant.

Question 4 (18 marks)

(a) (1)

$$\bar{x} = \frac{\sum_{i}^{n} x_i}{n} = \frac{495}{15} = 33$$

$$S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2} = \sqrt{\frac{2066}{14}} \approx 12.1479$$

(b) (1)

the population mean is defined as the average difference between the recorded amount and the audited amount.

(C) (1) <Question 4>

$H_0: \mu = 25$

$H_1: \mu > 25$

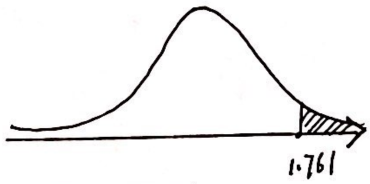
(d) (2)

since σ^2 is not known and $n = 15 < 30$, we may resort to T test.

(e) (1)

$\alpha = 0.05$

from the t-table: $t_{0.05;14} = 1.761$



from the graph above, the rejection region is $T > 1.761$.

(f) (2)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{33 - 25}{12.1479/\sqrt{15}} \approx 2.5506$$

(g) (3): from (f), we can know that on 5% level of significance. $t = 2.5506 > 1.753$.

$\therefore$ reject H_0 at the 5% level.

$\therefore$ we conclude that, at the 5% level, the average difference is greater than \$25.

(h) (3)

since we do not know σ, and both the recorded and audited amounts of inventory are both normally distributed, we may resort to T test.

from t table:

$t(0.025, 14) = 2.145$

the 95% confidence interval for the mean time μ is $\left(\bar{X} - t_{0.025;14} \cdot \frac{s}{\sqrt{n}}, \bar{X} + t_{0.025;14} \frac{s}{\sqrt{n}}\right)$

$$CI_{95\%} = \left(33 - 2.145 \cdot \frac{12.1479}{\sqrt{15}}, 33 + 2.145 \cdot \frac{12.1479}{\sqrt{15}}\right) = (26.2720, 39.7280)$$

Question 4.

(i) (2)

simplify the statement:

$$e = Z_{\frac{\alpha}{2}} \cdot \frac{6}{\sqrt{n}} = Z_{0.005} \cdot \frac{S}{\sqrt{n}} = 2.575 \cdot \frac{12.1479}{\sqrt{15}} \approx 8.0767$$

Because $e < 10$, we can use this formula to caculate the sample sizen, with 99% confidence within an error $e = \pm \$10$.

(j) (2)
there exists a relationship between confidence interval and the Level of significance.

like in (h), the range CI of μ is always greater than \$25 with 95% CI.

And in (g). we conclude that, at the 5% level, the μ is greter than \$25.

Question 5 (12 marks)

(a) (1)
$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{546}{30} = 18.2$$

$$S = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2} = \sqrt{\frac{3180.8}{29}} \approx 10.4730$$

(b) (2) the sample distribution of sample means is the distribution which formed by this sample of kong lung Airlines's days between the receipt of the complaints and the resolution of the complaints.

(C) (5) (Question 5)

(i) Since σ is unknown, but the sample size is $30 \geq 30$.

we use T-test to get CI.

from t-table

94%: $t_{0.03;29} = 1.957$

96%: $t_{0.02;29} = 2.150$

98%: $t_{0.01;29} = 2.462$

$$CI_{94\%} = \left(\bar{x} - t_{0.03;29}\frac{s}{\sqrt{n}}, \bar{x} + t_{0.03;29}\frac{s}{\sqrt{n}}\right)$$

$$= \left(18.2 - 1.957\cdot\frac{10.4730}{\sqrt{30}}, 18.2 + 1.957\cdot\frac{10.4730}{\sqrt{30}}\right)$$

$$CI_{96\%} = \left(18.2 - 2.150\cdot\frac{10.4730}{\sqrt{30}}, 18.2 + 2.150\cdot\frac{10.4730}{\sqrt{30}}\right).$$

$$CI_{98\%} = \left(18.2 - 2.462\cdot\frac{10.4730}{\sqrt{30}}, 18.2 + 2.462\frac{10.4730}{\sqrt{30}}\right).$$

$\therefore$

94% CI = (14.4580, 21.9420)   96% CI = (14.0890, 22.3110).

98% CI = (13.4924, 22.9076).

(ii) from C(i), we can know that: The wider the confidence interval, the higher the confidence.

(d) (4) Since we still do not know σ, because when confidence equal to 94% (from C(i)).

the d is equal to ± 3.7419. So the n must bigger than 30.

we can use t-test, Let the sample size be n_s

we can get that: $$t_{0.05;29}\cdot\frac{s}{\sqrt{n_s}} \leq 3$$

$$n_s \geq 35.1792$$

At least 36 samples should be given.

Question 6 (20 marks)

(a) (2)

from the table: $\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 50$ $\bar{y} = \frac{\sum_{i=1}^{10} y_i}{10} = 110$

$SS_x = 3400$ $SP = 6800$

$\begin{cases} b = \frac{SP}{SS_x} = 2 \\ a = -b\bar{x} + \bar{y} = 10 \end{cases}$ $\therefore$ the β_1 is equal to 2, β_0 is equal to 10

regression model: $\hat{y} = 2x + 10$

(b) (6) $MSB = \frac{SSB}{k-1} = \frac{18000}{1} = 18000$

$MSE = \frac{SSE}{N-k} = \frac{17060.0205}{18} = 947.7789.$

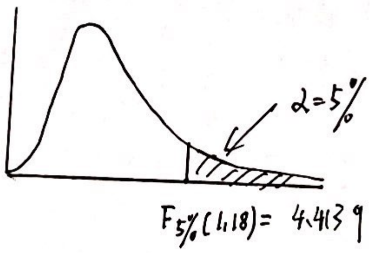
$F\text{-stat} = \frac{MSB}{MSE} = \frac{18000}{947.7789} = 18.9918.$

ANOVA

Sources	DF	SS	MS	F-stat	P-value
Between Groups	1	18000	18000	18.9918	0.0004
Within Groups	18	17060.0205	947.7789		
Total	19	35060.0205			

(C) (4) $H_0: \beta_1 = 0 \longleftrightarrow H_1: \beta_1 \neq 0$

At 5% Level, $F_{5\%}(1,18) = 4.4139$



$\because F_0 = 18.9918 > 4.4139$

$\therefore$ reject H_0 at the 5% level

we conclude that there exists a linear relationship between x and y.

(d) (2)

$R^2 = \frac{SSB}{SST} = \frac{18000}{35060.0205} \approx 0.5134$

Around 51.34% of the total variation in y can be explained by x.

Question 6. (cont'd)

(e) (3)

(i) from Q6(a), we can get the fitted regression model:

$$\hat{y} = 2x + 10$$

$e_1 = y_1 - \hat{y}_1 = 73 - (2\times30+10) = 3$ $\quad e_6 = y_6 - \hat{y}_6 = 108 - (2\times50+10) = -2$

$e_2 = y_2 - \hat{y}_2 = 50 - (2\times20+10) = 0$ $\quad e_7 = y_7 - \hat{y}_7 = 135 - (2\times60+10) = 5$

$e_3 = y_3 - \hat{y}_3 = 128 - (2\times60+10) = -2$ $\quad e_8 = y_8 - \hat{y}_8 = 69 - (2\times30+10) = -1$

$e_4 = y_4 - \hat{y}_4 = 170 - (2\times80+10) = 0$ $\quad e_9 = y_9 - \hat{y}_9 = 148 - (2\times70+10) = -2.$

$e_5 = y_5 - \hat{y}_5 = 87 - (2\times40+10) = -3$ $\quad e_{10} = y_{10} - \hat{y}_{10} = 132 - (2\times60+10) = 2.$

(ii) from (i):

$$\sum_{i=1}^{10} e_i = 3+0-2+0-3-2+5-1-2+2 = 0$$

f(3)

we need to find $E = \sum_{i=1}^{n} (y_i - ax_i - b)^2$ minimum

then it is an Quadratic function with b. when b is equal to: $-\dfrac{-2\sum_{i=1}^{n}(y_i - ax_i)}{2n} = \dfrac{\sum_{i=1}^{n}(y_i - ax_i)}{n}$

$= \bar{y} - a\bar{x}$.

the E get minimum value

For any regression line, if it do not pass $(\bar{x}, \bar{y})$, then the error ε would not be minimum, then we cannot get the accurate linear regression equation.

Question 7 (13 marks)

(a) (i)

$$\bar{X}_1 = \frac{\sum_{i=1}^{9} X_i}{9} = \frac{2260}{9} \approx 251.1111.$$

$$S_1 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(X_i-\bar{x})^2} = \sqrt{523.61} \approx 22.8826.$$

$$\bar{X}_2 = \frac{\sum_{i=1}^{10} X_i}{10} = \frac{2633}{10} = 263.3$$

$$S_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{N}(X_i-\bar{x})^2} = \sqrt{312.9} \approx 17.689$$

(b) $H_0: \mu_1 = \mu_2 \Leftrightarrow H_1: \mu_1 < \mu_2$.

(c) Since we do not know the value of σ, but samples are independent and the σ are equal but unknown. we use t-test (unpaired).

(d) the sample is normally distributied.

(e)

$$d.f = n_1 + n_2 - 2 = 9 + 10 - 2 = 17$$

$\therefore$ critical value

$$t(0.025; 17) = -2.110$$

(f) Because of the variance of X_1, X_2 were equal:

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} = \sqrt{\frac{8\times 523.6 + 9\times 312.9}{17}} \approx 20.2991$$

$$\sigma_{12} = S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} = 20.2991\cdot\sqrt{\frac{1}{9}+\frac{1}{10}} \approx 9.3268$$

Under H_0, the test statistic is

$$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sigma_{12}} = \frac{-12.1889}{9.3268} \approx -1.3069$$

(g)

$$t = -1.3069 > t(0.025; 17) = -2.110$$

$\therefore$ do not reject H_0 at 5% level

which means the new drug do not have significantly effect than the placebo in reducing cholesterol levels.

Question 8 (16 marks)

Assuming there exists a linear relationship between the Exercise level and Body weight.

$y = \beta_1 x + \beta_0 + \varepsilon$.

the Test is in order to test the significance of linearity of regression line.

We may perform the usual F-test <i>

The Hypotheses are

$H_0: \beta_1 = 0 \longleftrightarrow H_1: \beta_1 \neq 0$ <ii>

ANOVA

Sources	DF	SS	MS	F-stat
Between Groups	1	193442.7	193442.7	240.7332
Within Groups	28	22499.5785	803.5564	
Total	29	215942.2785		

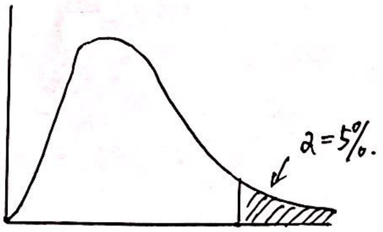
<v>

$MSB = \frac{SSB}{k-1} = \frac{193442.7}{1} = 193442.7$.

$F\text{-stat} = \frac{MSB}{MSE} = 240.7332$.

$MSE = \frac{SSE}{N-k} = \frac{22499.5785}{28} = 803.5564$

At 5% level, $F_{5\%}(1, 28) = 4.1960$



<iii>

the rejection region: F-stat > 4.1960.

<iv>

$F_{5\%}(1,28) = 4.1960$

$\because$ F-stat = 240.7332 > 4.1960.

$\therefore$ reject H_0 at the 5% level. <vi>

we conclude that there exist a relationship between the body weights and exercise level.

<vii>

END