STATS251F

Statistical Data Analysis

2020 - 21

Take-Home Assignment 2 (Mid-term Test).

Student Name: Angold, Jiawei Wang

Student ID: 51239587O Mark: _____

Question 1 (24 marks)

Part (a) (15 marks)

(i) Standard deviation: $S' = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2} = \sqrt{\frac{(10.7-10.7)^2+\cdots+(10.6-10.7)^2}{5}} = \sqrt{\frac{0.025}{5}} = \sqrt{0.005} = 0.7071$.

Sample mean: $\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = 10.7$ (seconds).

(ii). H_0: $\mu = \mu_0 = 10.6$ (seconds). (is equal).

H_1: $\mu > \mu_0 = 10.6$ (seconds). (is greater).

(iii). t test, because the .6 of the population is known, but SD is unknown. and the sample is $5 < 30$, but follows normal distribution

(iv) From the t-table: $df = 5-1 = 4$

Take $\alpha = 5\%$ $t_{0.05} = 2.132$ ∴ rejection region: $t < -2.132$.

(v) test statistic $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{10.7-10.6}{s/\sqrt{5}}$.

$\because s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} = \sqrt{\frac{(10.7-10.7)^2+\cdots+(10.6-10.7)^2}{5-1}} = \sqrt{0.006\dot{2}5} = 0.0791$

$= \frac{0.1}{0.0791\cdot\sqrt{5}} = 0.5828$.

(vi) ∵ $t > -2.132$ ∴ We conclute that, at 5% lerel, the athtete 'runing time. is not greater than his past mean valne.

Scanned with CamScanner

Part (b) (9 marks)

(i) Because the sample size is $15 < 30$.

∴ Using unpaired t-test.

(ii) $H_0: \mu_{Home} = \mu_{minor}$ (μ_{Home} represents the average score of living in Home Enrionment).

$H_1: \mu_{Home} < \mu_{minor}$

(iii). $\bar{X}_{Home} = \frac{1075}{15} = 71.6667$ $\quad S_{Home} = \sqrt{\frac{273.33}{15}} = 4.269$

$\bar{X}_{minor} = \frac{1145}{15} = 76.3333$ $\quad S_{minor} = \sqrt{\frac{349.33}{15}} = 4.826$

From the Question, we can know that the population variances are not equal

∴ $\hat{\sigma}_{HM} = \sqrt{\frac{S_H^2 + S_M^2}{n_H n_M}} = \sqrt{\frac{622.66}{15 \times 15}} = 1.6635$

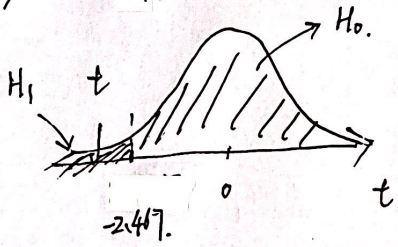
Under H_0, the test stastic is

$t = \frac{(\bar{X}_H - \bar{X}_M) - (\mu_H - \mu_M)}{\hat{\sigma}_{HM}} = \frac{71.6667 - 76.3333}{1.6635} = -2.8053$.

the critical value is: $-t(28, 0.02) = -2.467$.

(iv). ∵ $t = -2.8053$

∵ $df = n_H + n_M - 2 = 28$.



∴ $t(0.01, 28) = 2.763$.

∵ $t < -t(0.02, 28) = -2.467$.

∴ reject H_0 at the 1% level.

∴ At 1% level, there is insufficient evidence to support the claim that the μ_{Home} is equal to μ_{minor}.

Which means, at 1% level, the living in the minority environment can lead to higher mean scores of social attitudes.

Scanned with CamScanner

Question 2 (33 marks).

(a) $H_0: \mu_{July} = \mu_{sep}$ (μ_{July} means the average price (\$) of rice in July).

$H_1: \mu_{July} \neq \mu_{sep}$.

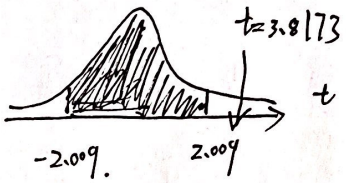
(b). t-test

Because ① the sample size is 27 or 25 < 30. ② We do not know the standard deviations of the population (both).

(c) (i) the critical value:

$df = 27+25-2 = 50$.

$t_0 = t(0.025, 50) = 2.009$.



rejection:

if t in $(-2.009, 2.009)$. H_0 is false

(ii). Because the pop variances are equal but unknown.

the standard error is

$$\hat{\sigma}_{JS} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= \sqrt{\frac{26\cdot(0.26)^2+24\cdot(0.27)^2}{50}} \cdot \sqrt{\frac{1}{27}+\frac{1}{25}} \approx 0.2548 \sqrt{0.07703} \approx 0.07073$$

Under H_0. the test stastic is.

$$t = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\hat{\sigma}_{JS}} = \frac{0.27}{0.07073} \approx 3.8173$$

(iii), $\because 3.8173 > 2.009$.

$\therefore$ reject H_0 at 5% level.

The data support the contention that the mean prices in July and September are not the same at 5% level.

Scanned with CamScanner

(d) (i). the critical value

$\cdot df = n_1 + n_2 - 2 = 50.$

$t_0 = t(0.05, 50) = 2.009.$

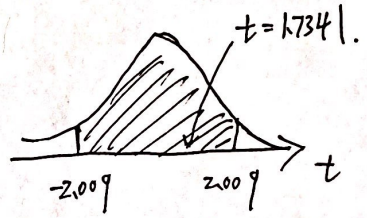
$\therefore$ Desision rule: Reject if $T_0 < -2.009$ or $T_0 > 2.009$.

(ii) Because the σ is unknown and unequal

$$\therefore \hat{\sigma}_{JS} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = \sqrt{\frac{(0.24)^2}{27} + \frac{(0.27)^2}{25}} \approx 0.1557$$

under H_0, the test stastic is.

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma_{12}} = \frac{0.27}{0.1557} \approx 1.7341$$



(iii). $\because 1.7341 < 2.009.$

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

The data do not support the contention that the means in july and september are not the same.

Question 3 (43 marks).

(a). independent variable (X): Test Score

dependent variable (Y): No. of Units sold.

(b). It can be seen from the scatter diagram:

Because except for a few adjacent two points, the rest of y increases with the increase of X.

we can think that. X and y are roughly positively.

(c) $$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}} = \frac{292.6}{\sqrt{7.02 \cdot 1315.69}} = 0.9628.$$

① the r is positive., which means. X and y are roughly correlated

② the r is bigger than 0.75. and very close to 1.0. which means. X and y has a strong Positive. correlation.

Scanned with CamScanner

(d) (i). Sample mean $\bar{x} = 3.4$. $\bar{y} = 137.1$.

$\therefore a = \bar{x} = 3.4$ $b = \bar{y} = 137.1$.

(ii). Let the fitted regression equation be like:

$$y = \beta_1 X + \beta_0.$$

which $\begin{cases} \beta_0 = b - a\cdot\beta_1 = 137.1 - 3.4 \cdot 41.68 = -46151. \\ \beta_1 = \dfrac{n\Sigma xy - (\Sigma x)(\Sigma y)}{n\Sigma x^2 - (\Sigma x)^2} = 41.68091 \approx 41.68 \end{cases}$

$\therefore$ the equation should be: $\hat{y} = 41.68091X - 4.6151$.

(e) (i) When $X = 3.2$:

$\hat{y} = 41.68091(3.2) - 4.6151 = 128.763811999997$.

When $X = 6.0$.

$\hat{y} = 41.68091(6) - 4.6151 = 245.4703599999997$.

(ii) from. Q_3c, we can know that the r is 0.9628.

which means the coefficient of determination $R^2 = r^2 = 0.92698384$.

which means 92.69838% of the total variation in y can be explained by the fitted regression equation.

So It is not reliable, because X and y are not completely positively corelated. there must exist deviation.

(f). ANOVA Table

SV	DF	SS	MS	F-ratio
Between Groups	$k-1=1$	89378.45	89378.45	122.2138.
Within Groups.	$N-k=18$	13163.9133.	731.3285	
Total	$N-1=19$.	102542.3633		

Scanned with CamScanner

$SS_B = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2 \approx 89378.45$

$SS_W = \sum_{i=1}^{k}(n_i - 1)s_i^2 \approx 13163.9133$

$SS_T = SS_B + SS_W = 13163.9133 + 89378.45 = 102542.3633$.

(g).

test for the linearity of the regression:

suppose $\hat{y} = \hat{\beta}_1 \bar{x} + \hat{\beta}_0$

(i) $H_0: \beta_1 = 0 \iff H_1: \beta_1 \neq 0$.

(ii). $F_{5\%}(1, 18) = 4.4139$.

(iii) if $F \leq 4.4139$. then, accept H_0.

if $F > 4.4139$ then, reject H_0.

(iv). From the ANOVA. table. we can know that

$$F = \frac{MS_B}{MS_W} = 122.2138.$$

(v) $\because 122.2138 > 4.4139$

$\therefore$ reject H_0 at 5% level.

(vi). We conclude that there exists a linear relationship between x and y.

(h). From Q3 c, we can know that the r is 0.9628.

$\therefore R^2 = r^2 = 0.92698384$.

which means 92.698384% of the total variation in y can be. explained by the fitted regression equation.

Scanned with CamScanner