FIT5197 – Statistical Data Modelling – Semester 1, 2020 Monash Clayton Campus

ASSIGNMENT 1D – Due on Sunday 11:55pm of Week 8.

TOPICS TESTED: HYPOTHESIS TESTING & CONFIDENCE INTERVAL

Do's & Don'ts:

*** All answer in this assignment can be done in any colour **except for red colour** as this is the colour your tutor will use to mark your assignment. Thus, answers in red colour will not be graded (even correct ones) ***

*** If you choose to handwrite your answer (scan and submit it electronically), make sure your **handwriting is readable** – this is a good practice for your exam. Failure to comply will result in a lack of marks awarded if your writing is unreadable and there will be no exception for this. ***

*** **Each sub-assignment will be worth 2.5%** towards your total score. Sub-assignments will have different point distributions within them as we aim to focus your attention to the important areas; however, the score for a sub-assignment remains 2.5%. ***

*** These questions are meant for you to solve **independently**, we encourage students to figure out the questions themselves as it would be good for their understanding of the topics; however, please feel free to consult your tutors if needed. **Plagiarism** (either from using online sources or copying the answers from your classmates) will be punished accordingly. ***

*** As this is considered to be an assignment (albeit a sub assignment), requests for special consideration or extension must be submitted at least **2 days BEFORE THE DEADLINE.** The due date is on Sunday, so the latest day you can ask for extension is on Friday (the last official working day of the week for the teaching team). Please follow Monash guideline to request for extensions (medical certificates, doctor or GP letter, etc). Emergencies are to be adjusted individually. ***

*** **No R or any other programming languages should be used in solving these questions**. All work for this assignment needs to be done manually, less the use of **non-programmable calculator** (this also applies to your Final Exam). Tutors are not required to answer questions in the difference between manual calculation and programmed calculation***

*** **Late submission is 10% per day, after 5 days you will be given no marks**. Late submission is calculated as following: If you get 70% on this assignment and you are late for 2 days (you submit on Tuesday), your scores is now 70% -20% (2x10% per day) = 50%. This is done to ensure that the teaching team can release your result as soon as possible so that you can review on your mistakes and have a better study experience. ***

*** Please **show all working** in answering questions, your score will be **halved** if you don't comply***

*** Assignments shall be marked completely in **two weeks' time** according to Monash Policies. If there are any changes to the marking time, we will duly inform you. Solutions **will not be released** for this assignment; you can come to the tutorial and ask for an explanation about how to solve the questions after scores are released. ***

*** Please don’t send emails to tutors asking for suggestions, we have Moodle and consultations for that, In writing your inquiries on Moodle please try to be clear in your problem and not revealing your working to others as this might be counted as plagiarism on your part. A good format for inquiry topic would be “Assignment 1D – Tutorial 10 (your tutorial slot) – Question about median “***

*** Assignments need to be submitted in **PDF** format. Failure to comply will result in 30% penalty***

*** Filename format for submitting assignment “Assignment1D_StudentId.pdf”. File with wrong format incurs 30% penalty ***

FIT5197 – Statistical Data Modelling – Semester 1, 2020 Monash Clayton Campus

QUESTIONS:

HYPOTHESIS TESTING & CONFIDENCE INTERVAL: (10 Marks Total)

1) Study show that the contaminants in farmed salmon follows a normal distribution $x \sim N(\mu,\sigma^2)$. 6 fishes are randomly chosen, and the contaminants data (ppm) is recorded as followed:

2.06, 1.93, 2.12, 2.16, 1.98, 1.95

a. If the standard deviation is 0.1, what is the 95% confidence interval for the mean contamination? **(1 Marks)**

b. If the standard deviation is unknown, what is the 95% confidence interval for the mean contamination? **(1 Marks)**

ANSWERS:

1a) Given: $x \sim N(u, \sigma^2)$

$n = 6$

Values: 2.06, 1.93, 2.12, 2.16, 1.98, 1.95

$\sigma = 0.1$

To find:
95% confidence interval,

$\alpha = 0.05$

Now, $\hat{u}_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i = 2.03$ //.

Solution:
95% confidence interval is given by:

$$\left[\hat{u}_{ML} - Z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right), \hat{u}_{ML} + Z_{\frac{\alpha}{2}} \left(\frac{\sigma}{\sqrt{n}} \right) \right]$$

$$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = Z_{0.025} = 1.96 \text{ //.}$$

$$\Rightarrow \left[2.03 - \frac{(1.96)(0.1)}{\sqrt{6}}, 2.03 + \frac{(1.96)(0.1)}{\sqrt{6}} \right]$$

$$\Rightarrow \left[2.03 - 0.08, 2.03 + 0.08 \right]$$

$$\Rightarrow [1.95, 2.11] \text{ //.}$$

CS Scanned with CamScanner

1b) Given: $\alpha = 0.05$

σ = unknown

μ =

$n = 6$

To find: 95% confidence interval

Solution:

95% CI is given by:

$$\left[\hat{\mu}_{ML} - t_{\frac{\alpha}{2},n-1} \frac{\hat{\sigma}_u}{\sqrt{n}}, \hat{\mu}_{ML} + t_{\frac{\alpha}{2},n-1} \frac{\hat{\sigma}_u}{\sqrt{n}} \right] \quad \text{---- (1)}$$

$$\hat{\sigma}_u^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$= \frac{1}{5} \left[(0.03)^2 + (0.1)^2 + (0.09)^2 + (0.13)^2 + (0.05)^2 + (0.08)^2 \right]$$

$$= 0.00896$$

$$\sigma_u = 0.0946 \quad \text{---- (2)}$$

$$t_{\frac{\alpha}{2},n-1} = t_{\frac{0.05}{2},5} = t_{0.025,5} = 2.57 \quad \text{---- (3)}$$

$\Rightarrow$ CI:

$$\left[2.03 - \frac{(2.57)(0.0946)}{\sqrt{6}}, 2.03 + \frac{(2.57)(0.0946)}{\sqrt{6}} \right]$$

$$= [2.03 - 0.099, 2.03 + 0.099]$$

$$= [1.93, 2.13] \text{ //.}$$

Scanned with CamScanner

FIT5197 – Statistical Data Modelling – Semester 1, 2020 Monash Clayton Campus

2) Choose two independent samples $n_1 = 10$, $n_2 = 15$ from population $X \sim N(\mu_1, \sigma^2_1)$ and population $Y \sim N(\mu_2, \sigma^2_2)$. Given that $x = 82$, $s_x^2 = 56.5$, $\bar{y} = 76$, $s_y^2 = 52.4$

a. If $\sigma^2_1 = 64$, $\sigma^2_2 = 49$, what does a 95% confidence interval say about the difference $\mu_1 - \mu_2$? **(1.5 Marks)**

b. If $\sigma^2_1 = \sigma^2_2$ but unknown, what is the 95% confidence interval of $\mu_1 - \mu_2$? **(1.5 Marks)**

ANSWERS:

2) a) Given:

$n_1 = 10$ $\quad n_2 = 15$

$X \sim N(u_1, \sigma_1^2)$ $\quad Y \sim N(u_2, \sigma_2^2)$

$\bar{x} = 82$ $\quad \bar{y} = 76$

$s_x^2 = 56.5$ $\quad s_y^2 = 52.4$

To find: If $\sigma_1^2 = 64$

$\sigma_2^2 = 49$

$u_1 - u_2 = ?$ 95% confidence interval:

Solution:

95% confidence interval indicates,

$\alpha = 0.05$

$\Rightarrow u_1 - u_2 :$

$$\left[\hat{u_1} - \hat{u_2} \mp Z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

$\hat{u_1} = \bar{x} = 82$

$\hat{u_2} = \bar{y} = 76$

$Z_{\frac{\alpha}{2}} = Z_{\frac{0.05}{2}} = 1.96$

$$\Rightarrow u_1 - u_2 : \left[6 \mp 1.96 \sqrt{\frac{64}{10} + \frac{76}{15}} \right]$$

$$\Rightarrow \left[6 \mp 6.09 \right]$$

$\Rightarrow (-0.09, 12.09)$ //.

Scanned with CamScanner

2b) Given: $\sigma_1^2 = \sigma_2^2$ [unknown]

To find: 95% confidence interval of $\mu_1 - \mu_2$

Solution:

CI is given by:

$$(\hat{\mu_1} - \hat{\mu_2}) \pm t_{\frac{\alpha}{2}, n_1+n_2-2} \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{(n_1+n_2-2)}}$$

$$t_{\frac{\alpha}{2}, n_1+n_2-2} = t_{\frac{0.05}{2}, 10+16-2}$$

$$= t_{0.025, 23}$$

$$= t_{(1-0.975), 23}$$

$$= 2.069 \quad [\text{From table}]$$

$\Rightarrow$ CI:

$$\left[(82-76) \pm (2.069) \sqrt{\left(\frac{1}{10} + \frac{1}{15}\right) \frac{9(56.5) + 14(52.4)}{23}} \right]$$

$$= \left[6 \pm (2.069) \sqrt{(0.166)(54)} \right]$$

$$= [6 \pm (2.069)(2.994)]$$

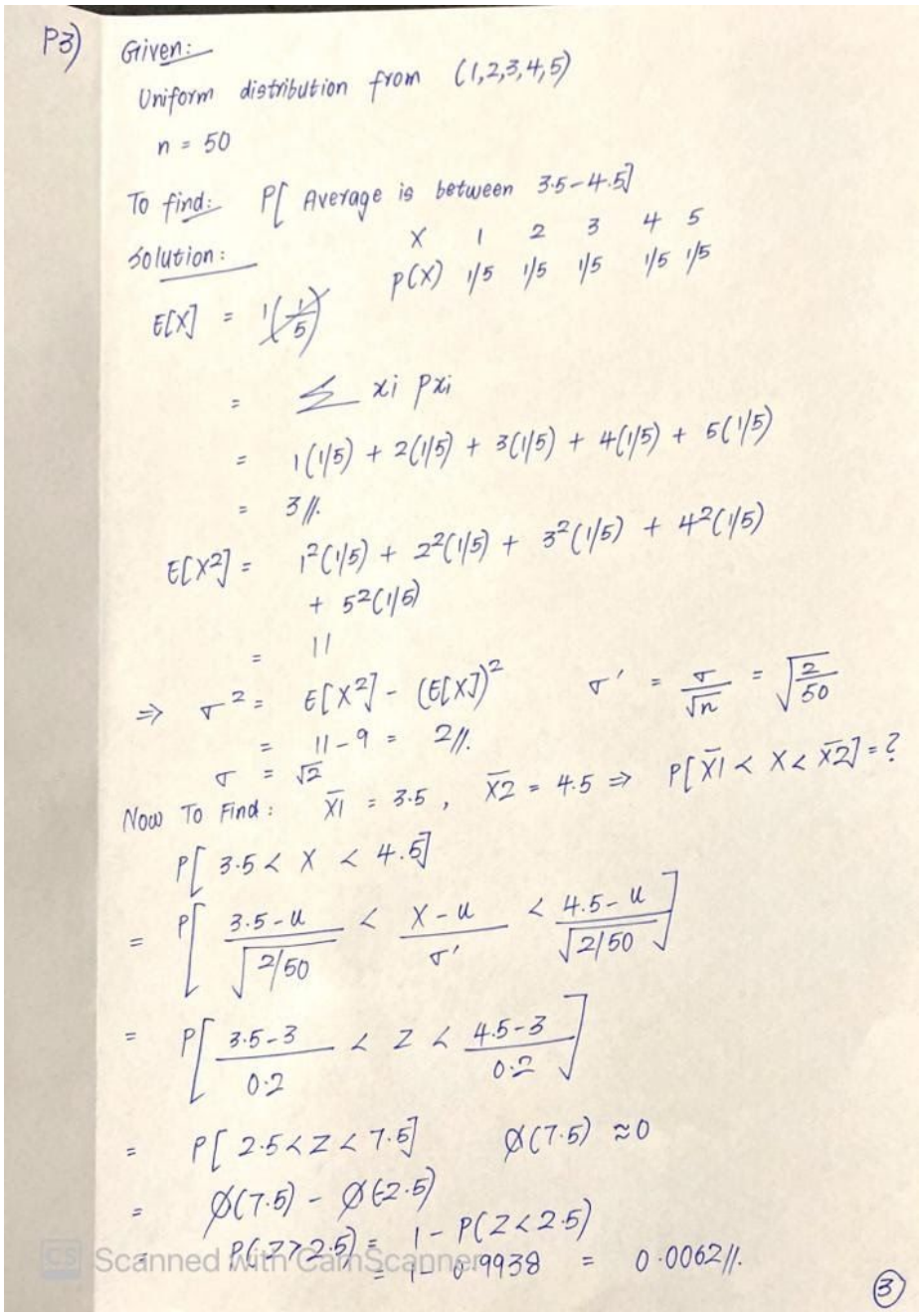
$$= 6 \pm 6.1945$$

$$= (-0.1945, 12.1945) //.$$

FIT5197 – Statistical Data Modelling – Semester 1, 2020 Monash Clayton Campus

3) Assume that the rating of customers for a restaurant is uniformly distributed from {1, 2, 3, 4, 5}. Given that you pick 50 of the ratings, estimate the probability that their average is between 3.5 - 4.5 stars. (**2 marks**)

ANSWERS:



P3) Given:

Uniform distribution from (1,2,3,4,5)

$n = 50$

To find: P[Average is between 3.5–4.5]

Solution:

X	1	2	3	4	5
P(X)	1/5	1/5	1/5	1/5	1/5

$E[X] = \sum x_i P x_i$

$= 1(1/5) + 2(1/5) + 3(1/5) + 4(1/5) + 5(1/5)$

$= 3$ //.

$E[X^2] = 1^2(1/5) + 2^2(1/5) + 3^2(1/5) + 4^2(1/5) + 5^2(1/5)$

$= 11$

$\Rightarrow \sigma^2 = E[X^2] - (E[X])^2 = 11 - 9 = 2$ //.

$\sigma = \sqrt{2}$

$\sigma' = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{2}{50}}$

Now To Find: $\bar{X1} = 3.5$, $\bar{X2} = 4.5 \Rightarrow P[\bar{X1} < X < \bar{X2}] = ?$

$P[3.5 < X < 4.5]$

$= P\left[\frac{3.5-u}{\sqrt{2/50}} < \frac{X-u}{\sigma'} < \frac{4.5-u}{\sqrt{2/50}}\right]$

$= P\left[\frac{3.5-3}{0.2} < Z < \frac{4.5-3}{0.2}\right]$

$= P[2.5 < Z < 7.5]$ $\quad \phi(7.5) \approx 0$

$= \phi(7.5) - \phi(2.5)$

$P(Z > 2.5) = 1 - P(Z < 2.5)$

$= 1 - 0.9938 = 0.0062$ //.

Scanned with CamScanner

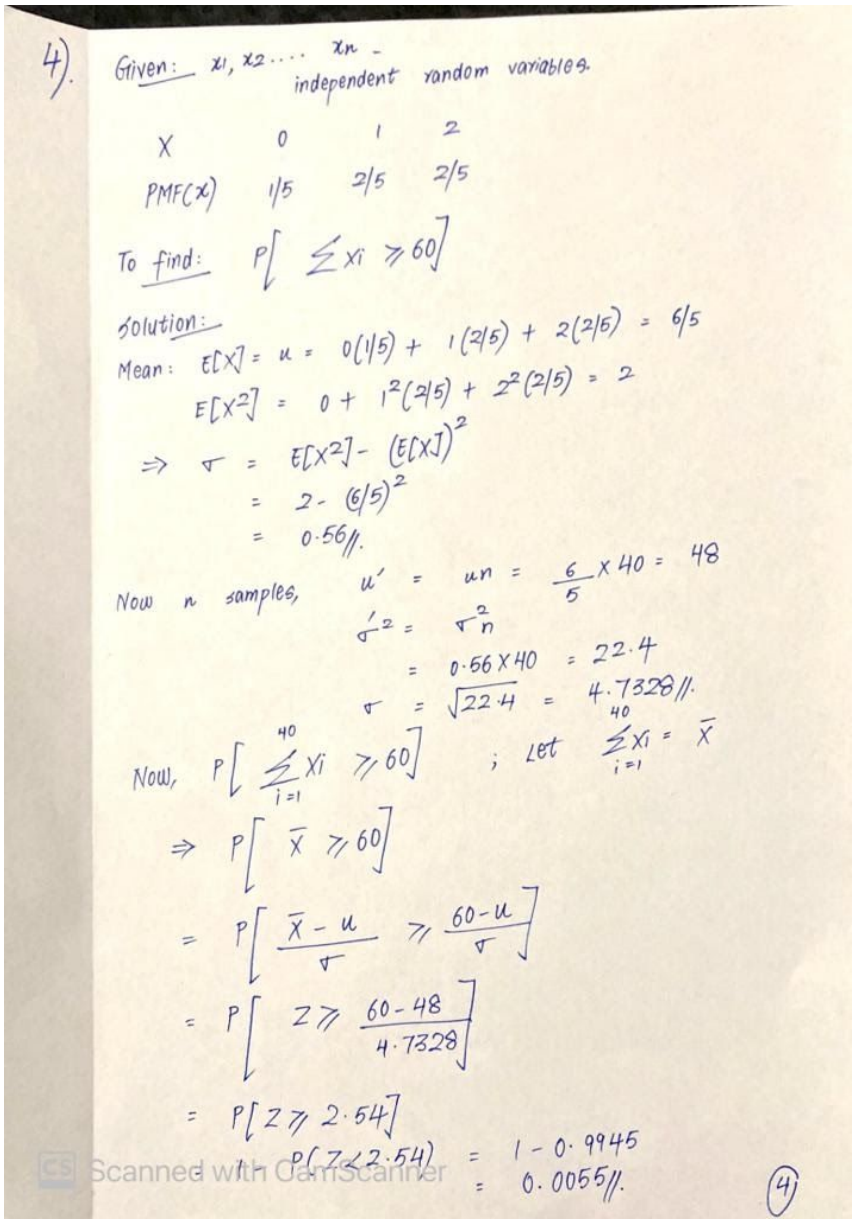
③

FIT5197 – Statistical Data Modelling – Semester 1, 2020 Monash Clayton Campus

4) Let $x_1, x_2, ..., x_{40}$ be independent random variables each having a discrete distribution with probability mass function:

X 0 1 2 PMF(X) 1/5 2/5 2/5 Estimate the probability that $x_1 + x_2 + \cdots + x_{40} \geq 60$ (1.5 marks)

ANSWERS:



4). Given: $x_1, x_2 \ldots x_n$ - independent random variables.

X	0	1	2
PMF(x)	1/5	2/5	2/5

To find: $P\left[\sum x_i \geq 60\right]$

Solution:

Mean: $E[X] = u = 0(1/5) + 1(2/5) + 2(2/5) = 6/5$

$E[X^2] = 0 + 1^2(2/5) + 2^2(2/5) = 2$

$\Rightarrow \sigma = E[X^2] - (E[X])^2$

$= 2 - (6/5)^2$

$= 0.56$ //.

Now n samples, $u' = un = \frac{6}{5} \times 40 = 48$

$\sigma'^2 = \frac{\sigma^2}{n}$

$= 0.56 \times 40 = 22.4$

$\sigma = \sqrt{22.4} = 4.7328$ //.

Now, $P\left[\sum_{i=1}^{40} x_i \geq 60\right]$; Let $\sum_{i=1}^{40} x_i = \bar{X}$

$\Rightarrow P\left[\bar{X} \geq 60\right]$

$= P\left[\frac{\bar{X} - u}{\sigma} \geq \frac{60 - u}{\sigma}\right]$

$= P\left[Z \geq \frac{60 - 48}{4.7328}\right]$

$= P[Z \geq 2.54]$

$= P(Z < 2.54) = 1 - 0.9945$

$= 0.0055$ //.

(4)

Scanned with CamScanner

FIT5197 – Statistical Data Modelling – Semester 1, 2020 Monash Clayton Campus

5) Monash stadium can tolerate an amount of weight believed to be W (unit of measurement is tonnes) without incurring any structural damage to the stadiums itself, $W \sim N(\mu = 10, \sigma^2 = 1.6)$ – W should not be a static value since it can be dependent on many factor of the day such as weather. Given this information, Monash security would like to be alerted if the probability of structural damage on the stadium is greater than 0.1 given that weight of every person getting into the stadium is a random variable with mean = 0.08 and variance = 0.00001 (unit in tons), find out the number of people that would alert Monash Security. (**1.5 Marks**)

ANSWERS:

5) Given:

Monash stadium's weight tolerance:

$$X \sim N(10, 1.6) \quad \text{—— ①}$$

y_i = weight of a person

$$Y \sim N(0.08, 0.00001) \quad \text{—— ②}$$

solution:

Let the no. of people that would alert the security be given by n.

$$\Rightarrow \frac{\sum y_i}{n} \sim N\left(0.08, \frac{0.00001}{n}\right)$$

$$\sum y_i \sim N(0.08n, 0.00001n)$$

Now, $P\left[\sum y_i > X\right] = 0.1$

Let $U = \left(\sum y_i - X\right) \Rightarrow U \sim N(10 - 0.08n, 1.6 + 0.00001n)$

$$\Rightarrow P\left[\left(\sum y_i - X\right) > 0\right] = 0.1$$

$$\Rightarrow P[U \geq 0] = 0.1$$

$$\Rightarrow P\left[\frac{U - (10 - 0.08n)}{\sqrt{1.6 + 0.00001n}} \geq \frac{(10 - 0.08n)}{\sqrt{1.6 + 0.00001n}}\right]$$

$$\Rightarrow P\left[Z \geq \left(\frac{10 - 0.08n}{\sqrt{1.6 + 0.00001n}}\right)\right] = 1.28$$

[Value from table]

Let's assume, $0.00001n \approx 0$

$$\Rightarrow \frac{10 - 0.08n}{\sqrt{1.6}} = 1.28 \Rightarrow n = 104.76$$

$\Rightarrow$ No. of people that would alert the security $= 105$.

Scanned with CamScanner