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

ASSIGNMENT 1C – Due on Sunday 11:55pm of Week 7.

TOPICS TESTED: MLE, DISTRIBUTION & INFERENCE

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 CENTRALLY VIA MONASH CONNECT.** 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 1F – Tutorial 10 (your tutorial slot) – Question about median "***

*** Assignments need to be submitted in **PDF** format. Failure to comply will result in 30% penalty. Furthermore, please make sure your assignment is working with Turnitin properly***

*** Filename format for submitting assignment "Assignment1c_StudentId.pdf". File with wrong format incurs 30% penalty ***

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

QUESTIONS:

A. DISTRIBUTION & INFERENCE: (7 Marks Total)

1) A small micro-loan bank has 1000 loan customers. If the total annual loan repayments made by an individual is a random variable with mean \$850 and standard deviation \$900, approximate the probability that the average total annual repayments made across all customers is greater than \$865. **(1 Marks)**

ANSWERS:

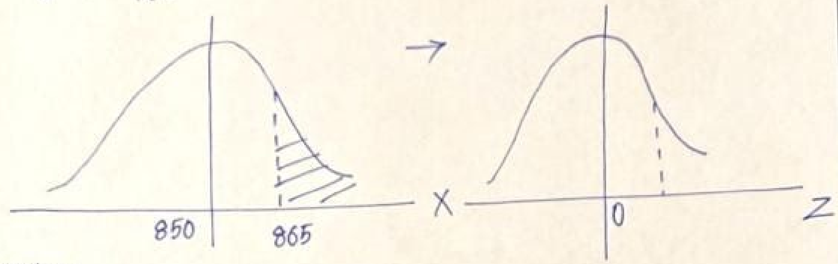
A1). Given:

$n = 1000$

$u = 850$

$\sigma = 900$

$\sim N(0,1)$



solution:

$$\text{New } \sigma = \frac{900}{\sqrt{n}} = \frac{900}{\sqrt{1000}} = 9\sqrt{10}$$

$$\Rightarrow P\left[X > 865\right]$$

$$= P\left[\frac{X-u}{\sigma} > \frac{865-u}{\sigma}\right]$$

$$= P\left[\frac{X-850}{(900/\sqrt{1000})} > \frac{865-850}{9\sqrt{10}}\right]$$

$$= P\left[Z > 0.53\right]$$

$$= 1 - \phi(0.53)$$

$$= 0.2981 //.$$

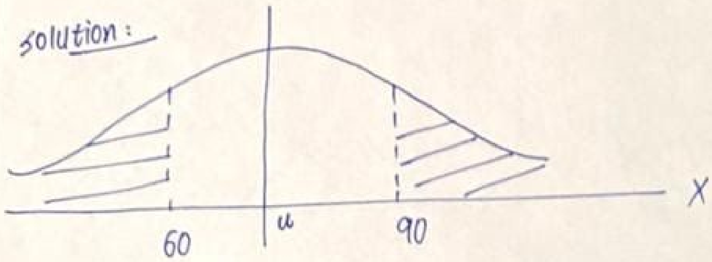
2) There are 10000 students participate in an exam and the exam score approximately follow a normal distribution. Given that 359 students get scores above 90 in the exam and 1151 students get lower than 60. If there are 2500 students pass the exam, find out the lowest score to pass. **(2 Marks)**

ANSWERS:

2) <u>Given:</u> $n = 10000$

Follow Normal Distribution

<u>Solution:</u>



$$P(X \geq 90) = \frac{359}{10000} = 0.0359$$

$$P(X \leq 60) = \frac{1151}{10000} = 0.1151$$

Now, Let the mean be u and standard deviation be σ.

$\Rightarrow$ $P(X \geq 90) \Rightarrow$

$$\frac{90 - u}{\sigma} = 1.8 \text{ — } (1) \quad : \quad u + 1.8\sigma = 90$$

$P(X \leq 60) \Rightarrow$

$$\frac{60 - u}{\sigma} = -1.2 \text{ — } (2) \quad u - 1.2\sigma = 60$$

solving (1), (2) we get:

$$\begin{aligned} u + 1.8\sigma &= 90 \\ - \; u - 1.2\sigma &= 60 \\ \hline 3\sigma &= 30 \\ \sigma &= 10 \text{ — } (3) \end{aligned}$$

putting (3) in (1)

$$\begin{aligned} u + 1.8\sigma &= 90 \\ u + 18 &= 90 \\ u &= 72 \text{ — } (4) \end{aligned}$$

Now, it's given that No. of questions students who pass = 2500

Let the lowest score required to pass be m.

$$\Rightarrow \frac{m - u}{\sigma} = 0.67 \Rightarrow \frac{m - 72}{10} = 0.67$$

$$m = 78.7$$

3) 360 random numbers are generated from the interval [0, 1]. Use Chebyshev's inequality to find a lower bound for the probability that the sum of the numbers lies between 160 and 200. **(1 Marks)**

ANSWERS:

A3). Given:

$n = 360$

No.s are generated in the interval: $[0,1]$ $[a,b]$

The distribution is UNIFORM.

$\Rightarrow \quad u = \frac{b+a}{2} = 0.5 \qquad \sigma^2 = \frac{(b-a)^2}{12} = 1/12$

Now, For n such samples,

$u' = nu = 360 \times 0.5 = 180$

$\sigma'^2 = n\sigma = 360 \times \frac{1}{12} = 30$

To find: Lower Bound for sum between 160 and 200

solution:

$P[160 < X < 200]$

We know, Now $P[\,|X-180| \geq 20] \leq \frac{\sigma^2}{a^2}$

[$\because$ From Chebyshev's Inequality]

$X - 180 \geq 20$

$X \geq 200$ — ①

$-(X-180) \geq 20$

$X - 180 \leq -20$

$X \leq 160$ — ②

$P[\,|X-180| \geq 20] \leq \frac{30}{(20)(20)}$

≤ 0.075 //.

Thus, for X to lie in the given interval,

$\Rightarrow \quad 1 - P[\,|X-180| \geq 20]$

$\Rightarrow \quad 1 - 0.075$

$\Rightarrow \quad 0.925$ //.

ANS: 0.925 //.

③

4) A biased coin with P(Head) = 0.4 is tossed 1000 times. Let X be the counts of heads in the tossing.

a. Find the upper bound for $P(X \leq 300 \text{ or } X \geq 500)$. **(1 Marks)**

b**.** Use Gaussian approximation to compute $P(X \leq 300 \text{ or } X \geq 500)$. **(1 Marks)**

ANSWERS:

4) Given:

$P(H) = 0.4 = \theta$

(probability of success)

n = no. of trials = 1000

a) Upper Bound for $P(X \leq 300 \text{ or } X \geqslant 500)$:

Now, for large values of n,

$$Bin(n, \theta) \sim N(n\theta, n\theta(1-\theta))$$

$$N(u, \sigma^2)$$

$$\sim N(400, 240)$$

Now, $P(X \leq 300 \text{ or } X \geqslant 500)$

$$\downarrow$$

$$P(X - 400 \leq -100 \quad \text{or} \quad X - 400 \geqslant 100)$$

$$\Rightarrow P[|X - 400| \geqslant 100]$$

From Chebyshev's Inequality, we know that,

$$P[|X - u| \geqslant a] \leq \frac{\sigma^2}{a^2}$$

$$\Rightarrow P[|X - 400| \geqslant 100] \leq \frac{240}{(100)(100)}$$

$$\leq 0.024$$

Thus, upper bound

$= 0.024$

4b) $P(X \leq 300 \text{ or } X \geqslant 500)$

$= P(X \leq 300) + P(X \geqslant 500)$ —— ①

$=$ $\downarrow$

$P(X \leq 300)$

$P\left[\frac{X-u}{\sigma} \leq \frac{300-u}{\sigma}\right]$

$= P\left[\frac{X-400}{\sqrt{240}} \leq \frac{300-400}{\sqrt{240}}\right]$

$= P[Z \leq -6.45]$

$= P[Z > 6.45]$ —— ②

IIIly, $P(X \geqslant 500)$

$= P\left[\frac{X-u}{\sigma} \geqslant \frac{500-u}{\sigma}\right]$

$= P\left[\frac{X-400}{\sqrt{240}} \geqslant \frac{500-400}{\sqrt{240}}\right]$

$= P[Z \geqslant 6.45]$ —— ③

$\Rightarrow$ ANS: ② + ③

$= 5.5 \times 10^{-11} + 5.5 \times 10^{-11}$

$= 11 \times 10^{-11}$

$\approx 0\%$

5) A fair coin is tossed 1000 times. Use the Central Limit Theorem to approximate the probability that between 470 and 530 heads are obtained. How does this compare to Chebyshev's bound? **(1 Marks)**

ANSWERS:

5) Given: $n = 1000 \rightarrow$ No. of trials.

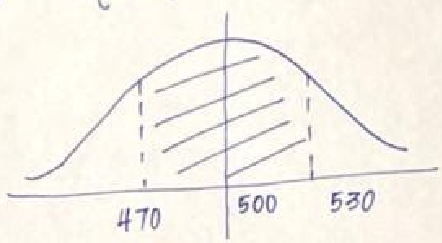
$\theta = 0.5 = P(H) \rightarrow$ probability of success.

Solution:

For large values of n,

$Bin(n, \theta) \sim N(n\theta, n\theta(1-\theta))$

$\Rightarrow N(500, 250) = N(u, \sigma^2)$



$P[470 \le X \le 530]$

$= P\left[\frac{470-u}{\sigma} \le \frac{X-u}{\sigma} \le \frac{530-u}{\sigma}\right]$

$= P\left[\frac{470-500}{\sqrt{250}} \le Z \le \frac{530-500}{\sqrt{250}}\right]$

$= P\left[-1.9 \le Z \le 1.9\right]$

$= \phi(1.9) - \phi(-1.9)$

$= \phi(1.9) - [1 - \phi(1.9)]$

$= 0.9426$ //.

Comparing with Chebyshev's bound:

$P[|X - 500| \geq 30] \le \frac{\sigma^2}{a^2}$

$\hookrightarrow P[X \geq 530$ and $X \le 470]$

$\Rightarrow P[|X-500| \geq 30] \le \frac{250}{(30)(30)} \le 0.2778$ //.

Thus,

$$P[470 \le X \le 530]$$

$$= 1 - P[|X-500| \geq 30]$$

$$= 1 - 0.2778$$

$$= 0.7222$$ //.

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

B. MLE (3 Marks Total)

A random variable Y is said to follow a geometric distribution with probability p if $P(Y = y|p) = (1 - p)^y p$ where $y \in \{0, 1, 2, . . .\}$ is a non-negative integer. Imagine we observe a sample of n non-negative integers y = (y1, . . . , yn) and want to model them using a geometric distribution. (hint: remember that the data is independently and identically distributed).

a. Write down the geometric distribution likelihood function for the data y (i.e., the joint probability of the data under a geometric distribution with probability parameter p). **(1 Marks)**

b. Write down the negative log-likelihood function of the data y under a geometric distribution with probability parameter p. **(1 Marks)**

c. Derive the maximum likelihood estimator for p. **(1 Marks)**

ANSWERS:

B). Given:

a) $P[Y = y|p] = (1-p)^y p$

$y \in \{0, 1, 2 \dots\}$

$y = (y_1, y_2 \dots y_n)$

<u>Likelihood function:</u>

$=$ product of individual probabilities

$= (1-p)^{y_1} p \cdot (1-p)^{y_2} p \dots \dots (1-p)^{y_n} p$

$= p^n (1-p)^{y_1} (1-p)^{y_2} \dots \dots (1-p)^{y_n}$

$= p^n (1-p)^{y_1 + y_2 \dots y_n}$

$= \boxed{p^n (1-p)^m}$

where $m = \sum_{i=1}^{n} y_i$

b) <u>Negative log-likelihood function:</u>

$l(y:p) = -\ln[p^n (1-p)^m]$

$= [\ln(AB) = \ln(A) + \ln(B)]$

$= -\ln[p^n] - \ln[(1-p)^m]$

$= \boxed{-n \ln(p) - m \ln(1-p)}$

c) <u>Maximum likelihood estimator for p:</u>

step1: $\frac{d}{dp}[-n \ln(p) - m \ln(1-p)]$

$\Rightarrow \frac{-n}{p} - \frac{m(-1)}{(1-p)}$

step2: $\frac{-n}{p} + \frac{m}{(1-p)} = 0$

step3: $\frac{-n}{p} + \frac{m}{(1-p)} = 0$

$-n(1-p) + mp = 0$

$-n + np + mp = 0$

$(n+m)p = n$

$\boxed{p = \frac{n}{n+m}}$ //

6