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

ASSIGNMENT 1B – Due on Sunday 11:55pm of Week 5.

TOPICS TESTED: EXPECTATION & ENTROPY

Do’ & Don’ts:

*** All answers 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 guidelines 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 a 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 follows: If you get 70% on this assignment and you are late for 2 days (you submit on Tuesday), your score 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 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 reveal your working to others as this might be counted as plagiarism on your part. A good format for inquiry topic would be “Assignment 1a – 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 “Assignment1B_StudentId.pdf”. File with wrong format incurs 30% penalty ***

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

QUESTIONS:

A. EXPECTATION:

(7 Marks Total)

1) Let X be a continuous random variable with probability density function:

Given that **E(X) = 0.5**, find out **Var(X)**? **(2 Marks)**

ANSWERS:

A1). Given:

$$p(x) = \begin{cases} ax + bx^2 & 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

$E(x) = 0.5$

To find: $Var(x)$

solution:

We know, $E(x) = \int_{-\infty}^{\infty} x f(x)\, dx$

$$\Rightarrow \int_0^1 x(ax+bx^2)\, dx = 0.5$$

$$\Rightarrow \int_0^1 (ax^2+bx^3)\, dx = 0.5$$

$$\Rightarrow a\frac{x^3}{3}\Big|_0^1 + b\frac{x^4}{4}\Big|_0^1 = 0.5$$

$$\Rightarrow \frac{a}{3} + \frac{b}{4} = 0.5$$

$$\Rightarrow 4a + 3b = 6 \text{ —— } ①$$

Illy, $\int_0^1 p(x)\, dx = 1$

$$\Rightarrow \int_0^1 (ax+bx^2) = 1$$

$$\Rightarrow a\frac{x^2}{2}\Big|_0^1 + b\frac{x^3}{3}\Big|_0^1 = 1 \Rightarrow \frac{a}{2} + \frac{b}{3} = 1 \Rightarrow 3a + 2b = 6 \text{ —— } ②$$

①

From 1 and 2,

$$4a + 3b = 6 \times \cancel{3}2$$
$$3a + 2b = 6 \times \cancel{2}3$$

$$\Rightarrow \quad 8a + 6\cancel{b} = 12$$
$$-9\cancel{12}a + \cancel{6b} = 18$$

$$-4a = -6$$
$$a = 6$$

and, $18 + 2b = 6$
$$2b = -12$$
$$b = -6$$

Thus, $a = 6$ — 3
$$b = -6$$

Now, $Var(x) = E(x^2) - [E(x)]^2$ — 4

$$E(x^2) = \int_{-\infty}^{\infty} x^2 f(x)\, dx$$
$$= \int_{-\infty}^{\infty} x^2 (ax + bx^2)\, dx$$
$$= \int_0^1 x^2 (6x + (-6x^2))\, dx$$
$$= 6 \int_0^1 (x^3 - x^4)\, dx$$
$$= 6 \left[\frac{x^4}{4} \Big|_0^1 - \frac{x^5}{5} \Big|_0^1 \right]$$

$$= 6(1/4 - 1/5) = 6 \times \frac{1}{20} = 3/10 = 0.3$$ — 5

From 4 and 5

$$Var(x) = E(x^2) - [E(x)]^2$$
$$= 0.3 - (0.5)^2$$
$$= 0.3 - 0.25$$
$$= 0.05 //.$$

ANSWER: $Var(x) = 0.05 //.$

2) Let X be a continuous random variable with probability density function:

Given that P(-2 < x < 1) = P(1 < x < 2) , find out the cumulative distribution function / **CDF(x). (2 Marks)**

ANSWERS:

A2) Given:

$$f(x) = \begin{cases} A & -2 <= x < 1 \\ Bx & 1 <= x < 2 \\ 0 & \text{else} \end{cases}$$

$$P(-2<x<1) = P(1<x<2)$$

To find: $CDF(x)$

Solution:

Now, $P(-2<x<1) = P(1<x<2)$

$$\Rightarrow \int_{-2}^{1} f(x)\,dx = \int_{1}^{2} f(x)\,dx$$

$$\Rightarrow \int_{-2}^{1} A\,dx = \int_{1}^{2} Bx\,dx$$

$$\Rightarrow A\,x\Big|_{-2}^{1} = B\frac{x^2}{2}\Big|_{1}^{2}$$

$$\Rightarrow A(1+2) = \frac{B}{2}(4-1)$$

$$\Rightarrow 3A = \frac{3B}{2}$$

$$A = B/2 \quad \text{— (1)}$$

Also, $\int_{-\infty}^{\infty} f(x)\,dx = 1$

$$\Rightarrow \int_{-2}^{1} A\,dx + \int_{1}^{2} Bx\,dx = 1$$

$$\Rightarrow A\,x\Big|_{-2}^{1} + B\frac{x^2}{2}\Big|_{1}^{2} = 1$$

5

$$\Rightarrow \quad 3A + \frac{3B}{2} = 1 \quad \text{——— ②}$$

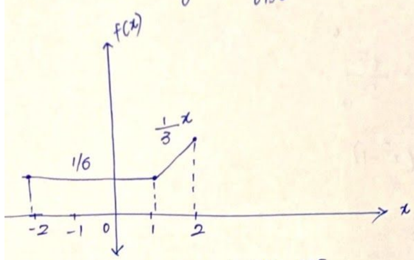
From ① and ②

$$3A + 3A = 1$$
$$A = 1/6$$
$$B = 2A = 1/3$$

Thus,

$$f(x) = \begin{cases} 1/6 & -2 \le x < 1 \\ \frac{1}{3}x & 1 \le x < 2 \\ 0 & \text{else.} \end{cases}$$

$$f(x) = \begin{cases} \frac{1}{\cancel{6}6} & -2 \le x < 1 \\ \frac{1}{3}x & 1 \le x < 2 \\ 0 & \text{Else} \end{cases}$$



Now, In order to calculate CDF:

PART-1:

Here $-2 \le t < 1$

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$
$$= \int_{-2}^{x} \frac{1}{6}\,dt$$
$$= \frac{1}{6}\, t \Big|_{-2}^{x}$$
$$= \frac{1}{6}(x+2) \quad \text{——— ①}$$

PART-2:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt \qquad 1 \le t < 2$$
$$= \int_{-2}^{1} f(t)\,dt + \int^{x} f(t)\,dt$$

$F(x)$

$$= \int_{-2}^{1} \frac{1}{6}\,dt + \int_{1}^{x} \frac{1}{3} t\,dt$$

$$= \frac{1}{6} t \Big|_{-2}^{1} + \frac{1}{3} \frac{t^2}{2} \Big|_{1}^{x}$$

$$= \frac{1}{6}(3) + \frac{1}{6}(x^2-1)$$

$$= \frac{3}{6} + \frac{x^2-1}{6}$$

$$= \frac{x^2+2}{6}$$

Now $CDF(x) = \begin{cases} 0 & x \le -2 \\ \frac{x+2}{6} & -2 \le x < 1 \\ \frac{x^2+2}{6} & 1 \le x < 2 \\ 1 & x > 2. \end{cases}$

3) The wealth of an individual is a random variable with probability density function:

a. What is the value of C to make the distribution normalise to 1? **(1 Marks)**

b. What is the mean of x? **(1 Marks)**

ANSWERS:

A3.

Given:

pdf: $f(x) = \frac{C}{x^{\alpha+1}}$ $\quad x \in [2, \infty)$

$\alpha > 1$

a) For the distribution to be normal,

$$\int_2^\infty f(x)\,dx = 1$$

$$\Rightarrow \int_2^\infty \frac{C}{x^{\alpha+1}}\,dx = 1$$

$$\Rightarrow \int_2^\infty x^{-(\alpha+1)}\,dx = \frac{1}{C}$$

$$\Rightarrow \left.\frac{x^{-(\alpha+1)+1}}{-(\alpha+1)+1}\right|_2^\infty = \frac{1}{C}$$

$$\Rightarrow \left.\frac{x^{-\alpha}}{-\alpha}\right|_2^\infty = \frac{1}{C}$$

$$\Rightarrow \left.\frac{1}{x^{\alpha}}\right|_2^\infty = \frac{-\alpha}{C}$$

$$\Rightarrow \frac{\cancel{-}1}{2^{\alpha}} = \frac{\cancel{-}\alpha}{C}$$

$$C = \alpha 2^{\alpha} \quad (1)$$

Thus, value of C to make the distribution normal,

$C = \alpha 2^{\alpha}$ //

(3)

b) Mean of x:

$$E[x] = \int_{-\infty}^{\infty} x f(x) \, dx$$

$$= \int_{2}^{\infty} \frac{x \; c}{x^{\alpha+1}} \, dx$$

We know, $c = \alpha 2^{\alpha}$

$$\Rightarrow \int_{2}^{\infty} \frac{x}{x^{\alpha+1}} \alpha 2^{\alpha} \, dx$$

$$\Rightarrow \alpha 2^{\alpha} \int_{2}^{\infty} x^{-\alpha} \, dx$$

$$\Rightarrow \alpha 2^{\alpha} \left. \frac{x^{-\alpha+1}}{-\alpha+1} \right|_{2}^{\infty}$$

$$\Rightarrow \frac{\alpha 2^{\alpha}}{(1-\alpha)} \left. \frac{1}{x^{\alpha-1}} \right|_{2}^{\infty}$$

$$= \frac{\alpha 2^{\alpha}}{(1-\alpha)} \left[0 - \frac{1}{2^{\alpha-1}} \right]$$

$$= \frac{-\alpha 2^{\alpha}}{(1-\alpha) \, 2^{\alpha-1}}$$

$$= \frac{2\alpha}{(\alpha-1)} //.$$

Thus, Mean of x = $\frac{2\alpha}{(\alpha-1)}$ //.

4) Let E[Z] = 1 and E[Z^2] = 6,E[Y] = −2 and E[Y^2] = 5,Z and Y are independent,then:

What is V[(3Z + 2Y)] **(1 Marks)**

ANSWERS:

A4). Given:

$E[Z] = 1$

$E[Z^2] = 6$

$E[Y] = -2$

$E[Y^2] = 5$

Z, Y are independent.

To find:

$V[3Z + 2Y]$

Solution:

$V[Z] = E[Z^2] - (E[Z])^2$

$= 6 - 1 = 5$ //

$V[Y] = E[Y^2] - [E(Y)]^2$

$= 5 - (-2)^{\cancel{4}2}$

$= 1$ //.

Now,

$V[aA + bB] = a^2 V[A] + b^2 V[B] - 2ab \operatorname{cov}(A, B)$

since Z, Y are independent,

$\operatorname{cov}(Z, Y) = 0$ //.

$\Rightarrow V[3Z + 2Y] = 3^2 V[Z] + 2^2 V[Y]$

$= 9(5) + 4(1)$

$= 45 + 4$

$= 49$ //.

B. ENTROPY (3 Marks Total)

This exercise is similar to the one given to you during the tutorial, please use **Kraft's Inequality** here to solve for the coding tree as we would do in the tutorial. (Other answers not using Kraft's Inequality will not be accepted) You have a probability distribution on six symbols, the letters 'A' to 'E'.

x A B C D E

f(x) - Probability 0.25 0.25 0.2 0.15 0.15

l(x) - Length

a. Please Fill in the table above with the appropriate values **(1 Marks)**

b. Please draw the coding tree as we did in the tutorial (an efficient coding tree) **(1 Marks)**

c. Please show that the length of your tree is within the boundary of the coding length. **(1 Marks)**

ANSWERS:

B). ENTROPY:

a)

x	A	B	C	D	E
$f(x)$ = probability	0.25	0.25	0.2	0.15	0.15
$l(x)$ = length	2	2	3	3	3

Now,

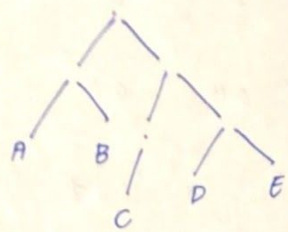
$$l_K = \log_2\left(\frac{1}{P_K}\right)$$

$$l_A = \log_2\left(\frac{1}{0.25}\right) = 2 ,\quad l_B = \log_2\left(\frac{1}{0.25}\right) = 2 ,\quad l_C = \log_2\left(\frac{1}{0.2}\right) = 2.3219 \sim 3$$

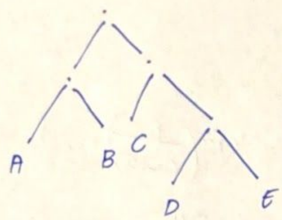
$$l_D = \log_2\left(\frac{1}{0.15}\right) = 2.7369 \sim 3 ,\quad l_E = \log_2\left(\frac{1}{0.15}\right) = 2.7369 \sim 3$$

[Using Kraft's Inequality]

b) CODING TREE:



Efficient coding tree:



c) To show that the length of my tree, is within the boundary of coding length:

$$\sum \frac{1}{2^{l_k}} \le 1$$

LHS: $\frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^3} + \frac{1}{2^3} = 0.875$

RHS: 1

Thus, LHS < RHS//. Hence Proved.

Also,

Expected code length $\le 1 + \sum_{k=1}^{n} P_k \log_2\left(\frac{1}{P_k}\right)$

LHS: $(2 \times 0.25) + (2 \times 0.25) + (2 \times 0.2) + (3 \times 0.15) + (3 \times 0.15)$

$= 2.3$//.

RHS $= 1 + [(0.25 \times 2) + (0.25 \times 2) + (0.2 \times 3) + (0.15 \times 3) + (0.15 \times 3)]$

$= 1 + 2.5 = 3.5$//

Clearly $2.3 \le 3.5$

LHS < RHS//. Hence Proved.