MH2814 – Probability & Statistics
AY 20/21 Semester 1 Solutions

(1ai)

$\underbrace{\text{D E C E M B E R}}_{8 \text{ letters}}$

# of	
B	: 1
C	: 1
D	: 1
E	: 3
M	: 1
R	: 1

$\therefore$ No. of ways $= \dfrac{8!}{3!\,(1!)^5}$

$= 6720$ //

(1aii)

fixed ↓ $\boxed{E}$ $\underbrace{\text{DCEMBER}}_{7 \text{ letters}}$ ⇒ # of E : 2

$\therefore$ No of ways $= \dfrac{7!}{2!\,(1!)^5}$

$= 2520$

Required probability $= \dfrac{2520}{6720}$

$= \dfrac{3}{8}$ //

(1b)

$$P(X=0) = \frac{\binom{4}{0}\binom{3}{2}}{\binom{7}{2}} = \frac{1}{7}$$

$$P(X=1) = \frac{\binom{4}{1}\binom{3}{1}}{\binom{7}{2}} = \frac{4}{7}$$

$$P(X=2) = \frac{\binom{4}{2}\binom{3}{0}}{\binom{7}{2}} = \frac{2}{7}$$

(1c)

we know that the area under the graph refers to a probability.

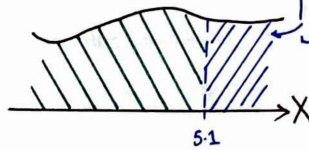
↳ any graph shape, for illustration purposes only.

↳ total area under the graph = 1

$P(X<5.1)$

$= 1-\frac{1}{16}$

$= \frac{15}{16}$



Need to prove $P(X>5.1) \leq \frac{1}{16}$

Chebyshev's Theorem:

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

Let $\mu + k\sigma = 5.1$, sub $\mu_X = 3$ & $\sigma_X = \frac{1}{2}$:

$3 + k(\frac{1}{2}) = 5.1$

$k = 4.2$

$\therefore P(X>5.1) \leq 1 - P\left(3-(4.2)(0.5) < X < 3+(4.2)(0.5)\right)$

$\leq 1 - P(0.9 < X < 5.1)$

$\leq 1 - \frac{1}{4.2^2}$

$\leq \frac{1}{16}$ (shown) //

Cumulative dist function: ↓ Present as a range even though X is discrete (ie an integer)

$$F(X) = \begin{cases} 0, & \text{if } X<0 \\ \frac{1}{7}, & \text{if } 0 \leq X < 1 \\ \frac{5}{7}, & \text{if } 1 \leq X < 2 \\ 1, & \text{if } X \geq 2 \end{cases}$$

FYI:
→ the no. of balls cannot be negative
→ $P(X=0)$
→ $P(X=0) + P(X=1)$
→ $P(X=0) + P(X=1) + P(X=2)$

(1d) Correlation coefficient, $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x \sigma_y}$ → given in the question ☺

If there is no information stating that X and Y are independent, use this formula:

$$\boxed{\sigma^2_{aX+bY+c} = a^2\sigma_X^2 + b^2\sigma_y^2 + 2ab\sigma_{xy}} \Rightarrow$$

NOTE!
If X and Y are independent,

$$\sigma^2_{aX+bY+c} = a^2\sigma_X^2 + b^2\sigma_y^2$$

If $Z = X - 2Y$,

$a = 1$

$b = -2$

From the question,

$\sigma_X^2 = 2$

$\sigma_y^2 = 2.5$

$\sigma_z^2 = 8$

$\sigma_z^2 = 8$

$\sigma^2_{X-2y} = 8$

Simply substitute the "ingredients" into the equation/formula

$$(1)^2(2) + (-2)^2(2.5) + 2(1)(-2)\sigma_{xy} = 8$$

$$12 - 4\sigma_{xy} = 8$$

$$4\sigma_{xy} = 4$$

$$\sigma_{xy} = 1$$

Hence, $\rho_{xy} = \dfrac{1}{\sqrt{2}\sqrt{2.5}}$

$= 0.4472135955$

≈ 0.4472 (4 s.f.) //

(2a) Binomial Probability:

$$P(X=x) = b(x;n,p)$$
$$= \binom{n}{x} p^x (1-p)^{n-x}$$

From the question,

$n = 5$

$p = 20\% = 0.2$

$\therefore$ P(at least two families have no children)

$$= P(X \geq 2)$$
$$= 1 - P(X=1) - P(X=0)$$
$$= 1 - \binom{5}{1}(0.2)^1(0.8)^4 - \binom{5}{0}(0.2)^0(0.8)^5$$
$$= 0.26272$$
$$\approx 0.2627 \text{ (4s.f.)}$$

(2b) When using Poisson approximation,

$$\mu \approx np \qquad P(x;\mu) = \frac{e^{-\mu}(\mu)^x}{x!}$$
$$= 20(0.2)$$
$$= 4$$

$\therefore$ P(exactly 10 families have no children)

$$= P(X=10)$$
$$= \frac{(e^{-4})(4)^{10}}{10!}$$
$$= 0.005292476 7$$
$$= 0.005292 \text{ (4s.f.)}$$

(3a) For a joint (probability) density function:

interchangeable

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

IDENTIFYING LIMITS:

If $0 \leq x \leq 2y \leq 2$,

divide by 2 throughout

$0 \leq x \leq 2$ $\qquad$ $\frac{x}{2} \leq y \leq 1$

NOTE THAT THE LOWER LIMIT OF y IS NOT 0.

Using the above formula:

$$\int_{\frac{x}{2}}^{1}\int_{0}^{2} kxy\,dx\,dy = 1$$

"left hand side"

$$LHS = \int_{\frac{x}{2}}^{1}\left[\frac{k}{2}yx^2\right]_0^2 dy$$
$$= \int_{\frac{x}{2}}^{1}\left[\frac{k}{2}y(2)^2\right] dy$$
$$= \int_{\frac{x}{2}}^{1}[2ky]\,dy$$
$$= \left[\frac{\cancel{2}k(y^2)}{\cancel{2}}\right]_{\frac{x}{2}}^{1}$$
$$= k\left[y^2\right]_{\frac{x}{2}}^{1}$$
$$= k\left[1^2 - \left(\frac{x}{2}\right)^2\right]$$
$$= \left(1 - \frac{1}{4}x^2\right)k$$

Unfortunately, using x as the inner integral and y as the outer integral gives us two unknowns x and k.

Since we can interchange dx and dy in the formula, let us try using y as the inner integral & x as the outer instead.

(3a) Continued

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

$$LHS = \int_0^2\int_{\frac{x}{2}}^1 kxy\,dy\,dx$$

$$= \int_0^2 \left[\frac{kxy^2}{2}\right]_{\frac{x}{2}}^1 dx$$

$$= \int_0^2 \frac{kx}{2}\left[y^2\right]_{\frac{x}{2}}^1 dx$$

$$= \int_0^2 \frac{kx}{2}\left[1^2-\left(\frac{x}{2}\right)^2\right] dx$$

$$= \frac{k}{2}\int_0^2 x\left(1-\frac{1}{4}x^2\right) dx$$

$$= \frac{k}{2}\int_0^2\left(x-\frac{1}{4}x^3\right) dx$$

$$= \frac{k}{2}\left[\frac{1}{2}x^2-\frac{1}{16}x^4\right]_0^2$$

$$= \frac{k}{2}\left[\frac{1}{2}(2)^2-\frac{1}{16}(2^4)\right]$$

$$= \frac{k}{2}(1)$$

$$= \frac{k}{2}$$

Let LHS = RHS:

$$\frac{k}{2} = 1$$

$$k = 2 \;//$$

(3b)

$$f(x,y) = \begin{cases} 2xy, & 0 \le x \le 2y < 2 \\ 0, & \text{elsewhere} \end{cases}$$

Marginal dist. of X

$0 \le x \le 2y$

$x \le 2y < 2$) divide by 2

$\frac{x}{2} \le y < 1$

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

$$g(x) = \int_{\frac{x}{2}}^{1} 2xy\,dy$$

$$= \left[\frac{2xy^2}{2}\right]_{\frac{x}{2}}^1$$

$$= \left[xy^2\right]_{\frac{x}{2}}^1$$

$$= (x)(1)^2 - (x)\left(\frac{x}{2}\right)^2$$

$$= x - \frac{x^3}{4} \quad \text{for } 0 \le x \le 2$$

and $g(x) = 0$ elsewhere. //

$$h(y) = \int_0^{2y} 2xy\,dx$$

$$= \left[\frac{2x^2y}{2}\right]_0^{2y}$$

$$= \left[x^2y\right]_0^{2y}$$

$$= 4y^3 \quad \text{for } 0 \le y \le 1$$

and $h(y) = 0$ elsewhere. //

remember to state the (range of each variable) and the value of the marginal dist. for ("elsewhere")

Morals of the story:

➡ You can interchange dx and dy in the equation, ensure that the corresponding limits match. Both results are mathematically the <u>same</u>.

I.e:

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

$$\Updownarrow$$

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

➡ If the upper or lower limit of y has x (like the question $\frac{x}{2} \le y < 1$), make dy the inner integral.

➡ It is important to know how to identify the upper limit of x and lower limit of x (same for y) from the given range in the question.

(3c)

$$f(x|y) = \frac{f(x,y)}{h(y)}$$

$$= \frac{2xy}{4y^3}$$

$$= \frac{x}{2y^2} \text{ for } 0 \leq x \leq 2$$

and $f(x|y) = 0$ elsewhere. //

Since the function $f(x|y)$ involves y, X and Y are not independent.

(3d)

$$E(X) = \int_{-\infty}^{\infty} x\, g(x)\, dx$$

$$= \int_0^2 x\left(x - \frac{x^3}{4}\right) dx$$

$$= \int_0^2 \left(x^2 - \frac{1}{4}x^4\right) dx$$

$$= \left[\frac{1}{3}x^3 - \frac{1}{20}x^5\right]_0^2$$

$$= \frac{8}{3} - \frac{32}{20}$$

$$= \frac{16}{15} \;//$$

$$E(Y) = \int_{-\infty}^{\infty} y\, h(y)\, dy$$

$$= \int_0^1 4y^4\, dy$$

$$= \left[\frac{4y^5}{5}\right]_0^1$$

$$= \frac{4}{5} \;//$$

(3e)

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f(x,y)\, dx\, dy$$

$$= \int_0^1\int_0^{2y} xy(2xy)\, dx\, dy$$

$$= 2\int_0^1\int_0^{2y} x^2y^2\, dx\, dy$$

$$= 2\int_0^1 \left[\frac{y^2x^3}{3}\right]_0^{2y} dy$$

$$= 2\int_0^1 \frac{y^2(2y)^3}{3}\, dy$$

$$= 2\int_0^1 \frac{8y^5}{3}\, dy$$

$$= \frac{16}{3}\int_0^1 y^5\, dy$$

$$= \frac{16}{3}\left[\frac{1}{6}y^6\right]_0^1$$

$$= \frac{16}{18}$$

$$= \frac{8}{9}$$

$$\sigma_{xy} = E(XY) - E(X)E(Y)$$

$$= \frac{8}{9} - \frac{16}{15}\left(\frac{4}{5}\right)$$

$$= \frac{8}{225} \;//$$

(4a)

x	1	2	3	Total
$P(X=x)$	0.3	0.3	0.4	1.0
$x\cdot P(X=x)$	0.3	0.6	1.2	2.1
$x^2\cdot P(X=x)$	0.3	1.2	3.6	5.1

$$E(X) = \mu$$
$$= 2.1 \;//$$

$$Var(X) = E(X^2) - [E(X)]^2$$
$$= 5.1 - 2.1^2$$
$$= 0.69 \;//$$

(4b) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

$$\bar{X} \sim N\left(2.1, \frac{0.69}{69}\right) //$$

(4c) $P(\bar{X} > 2.3)$

$$= P\left(Z > \frac{2.3-2.1}{\sqrt{\frac{0.69}{69}}}\right)$$
$$= P(Z > 2)$$
$$= 0.0228 \;//$$

(4d) $\alpha = 0.05$

$$Z_{\frac{\alpha}{2}} = Z_{0.025}$$
$$= 1.96$$

$$\frac{\sigma}{\sqrt{n}} = \frac{\sqrt{0.69}}{\sqrt{69}}$$
$$= 0.1$$

"based on huge historical data, the dist of X ..."

Hence, $Var(X) = \sigma^2 = 0.69$ in (a) is the true variance.

Since σ^2 is known,

$$\bar{X} - Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + Z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$2.0 - 1.96(0.1) < \mu < 2.0 + 1.96(0.1)$$

$$1.804 < \mu < 2.196 \;//$$

- can also present as $\Rightarrow$ $(1.804, 2.196)$

(5a) $H_0 : \mu_A \geq \mu_B$

$H_0 : \mu_A - \mu_B \geq 0$ //

(5b) $H_1 : \mu_A < \mu_B$

$H_1 : \mu_A - \mu_B < 0$ //

H_1 will always be the opposite of H_0, and there must be no "=" sign.

(5c) $\alpha = 0.05$ //

(5d) Since the variances of the distribution of A and B are unknown (NOTICE that the question writes S_A & S_B, not σ_A^2 and σ_B^2), we use the <u>T-statistic</u> whereby the population variances for both brands can be considered to be equal.

$$T = \frac{(\bar{x}_A - \bar{x}_B) - (\mu_A - \mu_B)}{\sqrt{\frac{S_p^2}{n_A} + \frac{S_p^2}{n_B}}}$$

$$S_p^2 = \frac{(n_A - 1)(S_A)^2 + (n_B - 1)(S_B)^2}{n_A + n_B - 2}$$

$$= \frac{(16-1)(4.5)^2 + (20-1)(5.5)^2}{16+20-2}$$

$$= 25.84$$

$$\therefore T = \frac{(35.5 - 38.0) - (0.0)}{\sqrt{\frac{25.84}{16} + \frac{25.84}{20}}}$$

$$= -1.466281965$$

$$= -1.466 \text{ (4 s.f.)}$$

$P(Z > z) = P(Z < -z)$

$\text{p-value} = P(T > t)$

$= P(T < -1.466)$

If $\nu = n_A + n_B - 2$
$= 16 + 20 - 2$
$= 34 > 30$, use the z-table to approximate the p-value. $P(T \leq t) \approx P(Z < z)$

Using normal approximation,

$= P(Z < -1.47)$

≈ 0.0708 // $> \alpha = 0.05$

(5e) Since p-value > α, we do not reject H_0.

There is insufficient evidence to conclude that the mean distance of Brand A is smaller than that of Brand B, at the 0.05 level of significance.

(6a)

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{yy}}$$

$$= \frac{94.79}{247.24}$$

$$= 0.383$$

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

$$= 1.78 - 0.383(3.96)$$

$$= 0.259$$

$$\therefore E(y) = 0.259 + 0.383x$$ //

(6b) When $x_0 = 2$,

$$E(y_0) = 0.259 + 0.383(2)$$

$$= 1.0255 \times \$1{,}000{,}000$$

$$s^2 = \frac{S_{yy} - \hat{\beta}_1 (S_{xy})}{n-2}$$

$$= \frac{37.54 - 0.383(94.79)}{13-2}$$

$$= 0.1089$$

$$= (0.330)^2$$

Using a t-distribution with

$\nu = n-2$

$= 11$,

a 95% confidence interval is given by:

$$E(y_0) \pm t_{\frac{\alpha}{2},11} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$$1.0255 \pm 2.201\,(0.330)\sqrt{\frac{1}{13} + \frac{(2-3.96)^2}{247.24}}$$

$$\Rightarrow (0.805, 1.246) \times \$1{,}000{,}000 \quad (4s.f.)$$ //

(6c) It is not advisable to use this model to predict what would be the sales when the advertising cost is $10,000,000 (i.e. $x = 100$) because 100 is very far from the observed $\bar{x}$ of the costs used to create this model.

FYI: standard deviation of observed $\bar{x} = \sqrt{\frac{S_{xx}}{n-1}}$

$$= \sqrt{\frac{247.24}{12}}$$

$= 4.54$, which is very small

NOTE :

Gotta pay attention to details !!

- This paper was rather <u>tricky</u>, don't be disheartened if you found it difficult. For math, practice makes perfect so do take some time to re-do tutorials and do past years' content. Clarify with your prof ASAP whenever in doubt and you'll be fine :)
- Once you've understood the content, try practicing under <u>timed</u> conditions

All the best ♡

Koh Roy-an

30/01/2021