DS 203 Assignment 3

Name - Rohan Rajesh Kalbag

Roll no - 20D170033

Exercise 1 : let X_1, X_2, X_3 are random variables

w.k.t $\quad Cov(X_1, X_2) = E\left[(X_1 - E[X_1])(X_2 - E[X_2])\right]$

Claim : $\quad Cov(X_1 + X_2, X_3) = Cov(X_1, X_3) + Cov(X_2, X_3)$

[Result used here is
$E[aX_1 + bX_2] = aE[X_1] + bE(X_2)$]

$$Cov(X_1 + X_2, X_3) = E\left[\left(X_1 + X_3 - E(X_1 + X_3)\right)\left(X_2 - E(X_2)\right)\right]$$

$$= E\left[\left(X_1 - E(X_1)\right) + \left(X_3 - E(X_3)\right)\left(X_2 - E(X_2)\right)\right]$$

$$= E\left[\left(X_1 - E(X_1)\right)\left(X_2 - E(X_2)\right)\right] + E\left[\left(X_3 - E(X_3)\right)\left(X_2 - E(X_2)\right)\right]$$

$$= Cov(X_1, X_2) + Cov(X_3, X_2)$$

Hence we are done.

Exercise 2:

$X_1, X_2, \cdots X_n$ are i.id

$\mu = E(X_i)$ and $\sigma^2 = Var(X_i)$

$\forall \ 1 \leq i \leq n$

$$Y = \frac{\sum_{i=1}^{n} (X_i - \mu)}{\sqrt{n\sigma^2}}$$

Claim 1: $E(Y) = 0$

proof : $$E(Y) = \frac{E\left(\sum_{i=1}^{n} (X_i - \mu)\right)}{\sqrt{n\sigma^2}}$$

$$E(Y) = \frac{E\left(\sum_{i=1}^{n} X_i\right) - E\left(\sum_{i=1}^{n} \mu\right)}{\sqrt{n\sigma^2}}$$

$$= \frac{\sum_{i=1}^{n} E(X_i) - n\mu}{\sqrt{n\sigma^2}}$$

$$= \frac{\sum_{i=1}^{n} \mu - n\mu}{\sqrt{n\sigma^2}} = \frac{n\mu - n\mu}{\sqrt{n\sigma^2}}$$

$= 0$, Hence we are done

Claim 2: $\operatorname{Var}(Y) = 1$

$$\operatorname{Var}(Y) = \operatorname{Var}\left(\frac{\sum_{i=1}^{n} (X_i - \mu)}{\sqrt{n\sigma^2}} \right)$$

Results used

$\operatorname{Var}(kX) = k^2 \operatorname{Var}(X) \quad k \in \mathbb{R}$

$\operatorname{Var}(X + k) = \operatorname{Var}(X) \quad k \in \mathbb{R}$

$$\operatorname{Var}(Y) = \operatorname{Var}\left(\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \right)$$

$$= \frac{\operatorname{Var}\left(\sum_{i=1}^{n} X_i - n\mu \right)}{n\sigma^2} = \frac{\operatorname{Var}\left(\sum_{i=1}^{n} X_i \right)}{n\sigma^2}$$

$$= \frac{\sum_{i=1}^{n} \operatorname{Var}(X_i)}{n\sigma^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n\sigma^2}$$

$$= \frac{n\sigma^2}{n\sigma^2} = 1$$

Hence $\operatorname{Var}(Y) = 1$

Hence we are done.

Exercise 3:

Size of population = 2000.

let a be the sorted array

a = [79, 80, 83, 86, 87, 87,
88, 91, 92, 94, 96, 97, 98, 98,
99, 103, 104, 104, 107, 108, 109,
109, 111, 119, 112]

1) Sample mean $(\bar{x}) = \frac{1}{25}\sum_{i=1}^{25} x_i = 98.04$

2) Sample variance $(\sigma^2) = \frac{1}{24}\sum_{i=1}^{25}(x_i - \bar{x})^2 = 133.70$

3) sample stdev $(\sigma) = \sqrt{\sigma^2} = 11.56$

4) $x_{(1)} = \min x_i = 79 = a[0]$

$x_{(n)} = \max x_i = 122 = a[24]$

$\therefore$ Sample range $= 122 - 79 = 43$

5) Sample median (since $N = 25$) will be the $\frac{N+1}{2}$th term which is the 13^{th} term when X_i's are sorted in ascending order

The median after sorting the array is the 12^{th} index $= a[12] = 98$.

6) The 25^{th} sample percentile is lower quartile $(p = \frac{1}{4})$

lower quartile $= X_{\lceil 25 \times \frac{1}{4} \rceil} = X_7$

which is the 6^{th} index in the sorted array $= a[6] = 88$

upper quartile $= X_{\lceil 25 \times \frac{3}{4} \rceil} = X_{19}$

which in the 18^{th} index in the sorted array $= a[18] = 108$

7) 95% confidence intervals for the population mean.

→ using the confidence interval for t - distribution

$$CI = \bar{x} \pm \frac{sz}{\sqrt{n}}$$

∴ for $s = 11.56$, $n = 25$, $dof = 24$

z from t-table is 2.064

$$CI \text{ in } \left[\bar{x} - \frac{sz}{\sqrt{n}}, \; \bar{x} + \frac{sz}{\sqrt{n}} \right]$$

$$\frac{sz}{\sqrt{n}} = \frac{11.56 \times 2.064}{\sqrt{25}} = \frac{11.56 \times 2.064}{5}$$

$$= 4.7719$$

∴ the confidence interval is thus

$$[98.04 - 4.7719, \; 98.04 + 4.7719]$$

$$[93.268, \; 102.812]$$

Exercise 4:

Sample 1: (n_1, σ_1)

Sample 2: $(2n_1, 2\sigma_1)$

Population ① has a better estimate to the population mean than Population II if it has a smaller confidence interval

$$\frac{z_1 \sigma_1}{\sqrt{n_1}} < \frac{z_2 \sigma_2}{\sqrt{n_2}}$$

It is given that finite population corrections are to be ignored

$\Rightarrow$ n_1 is sufficiently large that $z_1 = z_2 = z$

$\therefore$ for population I confidence interval is proportional to $\frac{\sigma_1}{\sqrt{n_1}}$ and for population II confidence interval is proportional to $\frac{2\sigma_1}{\sqrt{2n_1}} = \frac{\sqrt{2}\sigma_1}{\sqrt{n_1}} > \frac{\sigma_1}{\sqrt{n_1}}$

Hence Sample I is better than sample II.

Exercise 5:

(X_1, X_2) are random samples from the distribution $N(0, \sigma^2)$

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} dx$$

$$X_{(1)} = \min(X_1, X_2)$$

$$X_{(1)} > K \Rightarrow X_1 > K \ \& \ X_2 > K$$

$$P(X_{(1)} > K) = P(X_1 > K, X_2 > K)$$

$$P(X_{(1)} > K) = P(X_1 > K) \cdot P(X_2 > K)$$

$$1 - P(X_{(1)} \le K) = (1 - P(X_1 \le K)) \cdot (1 - P(X_2 \le K))$$

$$1 - F_{X(1)}(x) = (1 - F_{X_1}(x))(1 - F_{X_2}(x))$$

since X_1 and X_2 are RVs of the same dist

$$1 - F_{X_{(1)}}(x) = (1 - F_X(x))^2$$

$$f_{X_{(1)}}(x) = 2(1 - F_X(x)) f_X(x)$$

$$f_{X_{(1)}}(x) = 2\left(1 - F_X(x)\right) f_X(x)$$

$$f_{X_{(1)}}(x) = -\left(2 F_X(x) f_X(x) - 2 f_X(x)\right)$$

$$\mathbb{E}\left(X_{(1)}\right) = \int_{-\infty}^{\infty} x \, f_{X_{(1)}}(x)\, dx$$

$$\int_{-\infty}^{\infty} x f_{X_{(1)}}(x)\, dx = -2\int_{-\infty}^{\infty} x \, F_X(x) f_X(x)\, dx + 2 \cancel{\int_{-\infty}^{\infty} x f_X(x)\, dx}^{\,0}$$

$$\mathbb{E}\left(X_{(1)}\right) = -2\int_{-\infty}^{\infty} x \, F_X(x) f_X(x)\, dx$$

$$= -2\int_{-\infty}^{\infty} x \, F_X(x) f_X(x)\, dx$$

$$= -2\left[F_X(x) \cancel{\int x f_X(x)\, dx}^{\,0} - \int f_X(x) \int x f_X(x)\, dx \right]_{-\infty}^{\infty}$$

$$= 2 \int_{-\infty}^{\infty} f_X(x) \int x f_X(x)\, dx$$

$$= 2 \int_{-\infty}^{\infty} \frac{1}{\sigma^2 2\pi} e^{-\frac{x^2}{2\sigma^2}} \cdot \int x \, e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= 2 \int_{-\infty}^{\infty} \frac{1}{\sigma^2 2\pi} e^{-\frac{x^2}{2\sigma^2}} \cdot \int x \, e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= -\frac{1}{\pi} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} \int \left(-\frac{x}{\sigma^2}\right) e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= -\frac{1}{\pi} \int_{-\infty}^{\infty} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx$$

$$= -\frac{1}{\pi} \int_{-\infty}^{\infty} e^{-\frac{x^2}{\sigma^2}} dx$$

using $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$

$$= -\frac{1}{\pi} \sqrt{\frac{\pi}{\left(\frac{1}{\sigma^2}\right)}}$$

$$= -\frac{1}{\pi} \sigma \sqrt{\pi} = \frac{-\sigma}{\sqrt{\pi}}$$

$$\boxed{\therefore \mathbb{E}\left(X_1(x)\right) = \frac{-\sigma}{\sqrt{\pi}}}$$

Exercise 6:

given $\bar{x}$ and s^2 are sample mean and variance calculated from $x_1, x_2, x_3 \cdots \cdots x_n$

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$E(s) = \frac{1}{\sqrt{n-1}} \mathbb{E}\left(\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

Also w.k.t

$$(n-1) s^2 \sim \sigma^2 \chi^2_{n-1}$$

χ_{n-1}^2 - chi-squared distribution with n-1 degrees of freedom

PDF of χ^2_{n-1} in

$$P_{\chi^2_{n-1}}(x) = \frac{(1/2)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} x^{\left(\frac{n-1}{2}\right)-1} e^{-\frac{x}{2}}$$

$$\chi^2_{n-1} \sim \frac{(n-1)s^2}{\sigma^2} = Y$$

$$\Rightarrow Y \sim \chi^2_{n-1}$$

$$\mathbb{E}(s) = \sqrt{\frac{\sigma^2}{n-1}}\, \mathbb{E}(\chi_{n-1})$$

$$\mathbb{E}(s) = \sqrt{\frac{\sigma^2}{n-1}}\, \mathbb{E}(\sqrt{Y})$$

$$\mathbb{E}(s) = \frac{\sigma}{\sqrt{n-1}}\, \mathbb{E}(\sqrt{Y})$$

$$\mathbb{E}(s) = \frac{\sigma}{\sqrt{n-1}} \int_0^\infty \sqrt{y}\, f_Y(y)\, dy$$

$\rightarrow$ where $f_Y(y)$ is the PDF of χ^2_{n-1}.

$$\mathbb{E}(S) = \frac{\sigma}{\sqrt{n-1}} \int_0^\infty \sqrt{y}\, f_Y(y)\, dy$$

$$\mathbb{E}(S) = \frac{\sigma}{\sqrt{n-1}} \int_0^\infty \sqrt{t}\, \frac{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} t^{\left(\frac{n-1}{2}\right)-1} e^{-t/2} dt$$

$$\mathbb{E}(S) = \frac{\sigma}{\sqrt{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \frac{\left(\frac{1}{2}\right)^{\left(\frac{n-1}{2}\right)}}{\left(\frac{1}{2}\right)^{\left(\frac{n}{2}\right)}} \int_0^\infty \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}} t^{\frac{n}{2}-1} e^{-t/2}}{\Gamma\left(\frac{n}{2}\right)} dt$$

$$\mathbb{E}(S) = \frac{\sigma}{\sqrt{n-1}} \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \sqrt{2} \int_0^\infty p_{\chi^2_n}(t)\, dt$$

$$\mathbb{E}(S) = \sigma \sqrt{\frac{2}{n-1}} \left(\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \right)$$

$\therefore$ $\mathbb{E}(S) \neq \sigma \quad \forall\ n \in \mathbb{N}$

hence S is a biased estimator of σ

Hence we are done.

Exercise 7 :

$$P_\theta(X = x) = \begin{cases} \frac{3}{5}\theta , & x=0 \\ \frac{2}{5}\theta , & x=1 \\ \frac{3}{5}(1-\theta) , & x=2 \\ \frac{2}{5}(1-\theta) , & x=3 \end{cases}$$

observation $= (2, 3, 2, 1, 0, 0, 3, 2, 1, 1)$

$$L(\theta) = P_\theta(X=2) \cdot P_\theta(X=3)\, P_\theta(X=2) \cdot P_\theta(X=1) \cdot P_\theta(X=0)^2 \cdot P_\theta(X=3) \cdot P_\theta(X=2) \cdot P_\theta(X=1)^2$$

$$= (P_\theta(X=2))^3 \, (P_\theta(X=1))^3 \, (P_\theta(X=3))^2 \cdot (P_\theta(X=0))^2$$

$$= \left(\frac{3}{5}(1-\theta)\right)^3 \cdot \left(\frac{2}{5}\theta\right)^3 \cdot \left(\frac{2}{5}(1-\theta)\right)^2 \cdot \left(\frac{3}{5}\theta\right)^2$$

$$L(\theta) = \left(\frac{3}{5}\right)^5 \cdot \left(\frac{2}{5}\right)^5 \cdot \theta^5 (1-\theta)^5$$

This is the likelihood function for θ

$$L(\theta) = \left(\frac{3}{5}\right)^5 \cdot \left(\frac{2}{5}\right)^5 \cdot \theta^5 (1-\theta)^5$$

$$L'(\theta) = \left(\frac{6}{5}\right)^5 \left[5\theta^4 (1-\theta)^5 - 5\theta^5 (1-\theta)^4\right]$$

$$= \left(\frac{6}{5}\right)^5 \cdot 5\left[\theta^4 (1-\theta)^4 (1-2\theta)\right]$$

since $0 \le \theta \le 1$

Sign of $L'(\theta)$ $\qquad + \quad \underset{\frac{1}{2}}{|} \quad -$

hence $L(\theta)$ attains maxima at $\theta = \frac{1}{2}$, hence

$$\max(L(\theta)) = \left(\frac{6}{5}\right)^5 \cdot \left(\frac{1}{2}\right)^5 \cdot \left(\frac{1}{2}\right)^5$$

$$= \frac{\cancel{2^5} \cdot 3^5}{5^5 \cdot \cancel{2^5} \cdot 2^5} = \left(\frac{3}{10}\right)^5$$

$$= (0.3)^5 = 0.0024$$

Exercise 8 :

$$f(x|\theta) = \begin{cases} \dfrac{\theta}{(1+x)^{\theta+1}}, & 0 < x < \infty \\ 0, & \text{otherwise} \end{cases}$$

let n observations be made (i.i.d)

$\{x_1, x_2 \ldots x_n\}$

$$L(\theta|x) = \prod_{i=1}^{n} f(x = x_i | \theta)$$

$$L(\theta|x) = \theta^n \prod_{i=1}^{n} \frac{1}{(1+x_i)^{\theta+1}}$$

This is the likelihood function the observations are $\{x_1, x_2, x_3 \ldots x_n\}$

$$\log(L(\theta|x)) = n \log\theta - (\theta+1) \sum_{i=1}^{n} \log(1+x_i)$$

$$\log(L(\theta|x)) = n\log\theta - (\theta+1)\sum_{i=1}^{n}\log(1+X_i)$$

$$\frac{1}{L(\theta|x)}\frac{d(L(\theta|x))}{d\theta} = \frac{n}{\theta} - \sum_{i=1}^{n}\log(1+X_i)$$

For maximum likelihood estimate $\frac{d}{d\theta}(L(\theta|x)) = 0$.

$$\Rightarrow \frac{n}{\theta} - \sum_{i=1}^{n}\log(1+X_i) = 0$$

$$\Rightarrow \hat{\theta} = \frac{n}{\sum_{i=1}^{n}\log(1+X_i)}$$

$\therefore$ $\hat{\theta}$ is the maximum likelihood estimate of θ given r.v X and sample of n observations with $\{X_1, X_2 \ldots X_n\}$

Exercise 9:

if $X_1, X_2, X_3 \ldots \ldots X_n$ are i.i.d with pdf

$$f(x|\theta) = \theta x^{(\theta-1)} \quad 0 \le x \le 1 \;, 0 < \theta < \infty$$

$$L(\theta|x) = \prod_{i=1}^{n} \theta x_i^{(\theta-1)}$$

$$L(\theta|x) = \theta^n \prod_{i=1}^{n} X_i^{(\theta-1)}$$

$$\log L(\theta|x) = n \log \theta + (\theta-1) \sum_{i=1}^{n} \log(X_i)$$

$$\frac{1}{L(\theta|x)} \frac{d}{d\theta}\left(L(\theta|x)\right) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(X_i)$$

for MLE we have $\frac{d}{d\theta}\left(L(\theta|x)\right) = 0.$

$$\Rightarrow \quad \frac{n}{\hat{\theta}} + \sum_{i=1}^{n} \log(X_i)$$

$$\Rightarrow \quad \hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log(X_i)}$$

$$\hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log(X_i)}$$

$$\hat{\theta} = \frac{-n}{\sum_{i=1}^{n} \log(x_i)} \qquad n \in \mathbb{N}$$

The estimate of variance of a data x is got by using the Fisher information.

$$I(\theta) = -\mathbb{E}\left(\frac{\partial^2}{\partial \theta^2} \log L(\theta|x)\right)$$

$$\sigma^2 \geq \frac{-1}{\frac{\partial^2}{\partial \theta^2} \ln L(\theta|x)}, \quad \text{for large } n \quad \sigma^2 \approx \frac{-1}{\frac{\partial^2}{\partial \theta^2}[\ln L(\theta|x)]}$$

$$\log L(\theta|x) = n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log(x_i)$$

$$\frac{\partial}{\partial \theta}\left(\log L(\theta|x)\right) = \frac{n}{\theta} + \sum_{i=1}^{n} \log(x_i)$$

$$\frac{\partial^2}{\partial \theta^2}\left(\log L(\theta|x)\right) = \frac{-n}{\theta^2}$$

$$\frac{\theta^2}{n} = -\frac{1}{\frac{\partial^2}{\partial \theta^2}\left(\log L(\theta|x)\right)} \approx \sigma^2$$

$$\sigma^2 \approx \frac{\theta^2}{n}$$

Note that $\sigma^2 \approx \frac{\theta^2}{n}$

$$\lim_{n \to \infty} \sigma^2 \approx \lim_{n \to \infty} \frac{\theta^2}{n} = 0.$$

Hence the variance goes to zero as $n \to \infty$.

Exercise 10 :

For m bernoulli random variables

$y_1, y_2 \, . \, . \, . \, . \, . \, y_m$

$$f(y_i|\theta) = \prod_{i=1}^{m} \theta^{y_i} (1-\theta)^{1-y_i}$$

$$L(\theta|y_i) = \theta^{K} (1-\theta)^{m-k}$$

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

$$\ln L(\theta|x) = k \ln \theta + (m-k) \ln (1-\theta)$$

$$\frac{1}{L(\theta|x)} \frac{d}{d\theta} L(\theta|x) = \frac{K}{\theta} + \frac{(K-m)}{1-\theta} = 0$$

$$\frac{1}{L(\theta|x)} \frac{d}{d\theta} L(\theta|x) = \frac{K}{\theta} + \frac{(K-m)}{1-\theta} = 0$$

$$\frac{K}{\theta} = \frac{(m-K)}{1-\theta}$$

$$K - \cancel{K\theta} = m\theta - \cancel{K\theta} \qquad m\theta_0 = K$$

$$\theta_0 = \frac{K}{m}$$

Likelihood Ratio for Ber (θ)

$$\lambda(K) = \left\{ \begin{array}{ll} \dfrac{\theta^k (1-\theta)^{m-k}}{\left(\frac{K}{m}\right)^K \left(1-\frac{K}{m}\right)^{m-K}}, & \theta < \theta_0 \\ 1, & \theta = \theta_0 \end{array} \right\}$$

H_0 will be rejected

$\iff \lambda(K)\,[\theta \le \theta_0] < \varepsilon$

where $\varepsilon > 0$ is a constant.

$\Longleftrightarrow \quad \lambda(K)$ decreases

$$\log \lambda(K) \; [\theta \le \theta_0] = K\left[\log \theta - \log \frac{K}{m}\right] + (m-K)\left[\log(1-\theta) - \log\left(1-\frac{K}{m}\right)\right]$$

$$\frac{1}{\lambda(K)} \frac{d\lambda(K)}{dK} = \log \theta - \log(1-\theta) - \log\left(\frac{K}{m}\right) + \log\left(\frac{m-K}{m}\right) - \frac{K}{K} + \frac{m-K}{m-K}$$

$$\frac{1}{\lambda(K)} \frac{d\lambda(K)}{dK} = \log\left(\frac{\theta}{(1-\theta)} \frac{(m-K)}{m \; K} \cdot m\right)$$

$$\frac{1}{\lambda(K)} \frac{d\lambda(K)}{dK} = \log\left(\frac{\theta}{1-\theta} \frac{(m-K)}{K}\right)$$

since $\theta < \frac{K}{m} \Rightarrow 1-\theta > \frac{m-K}{m}$

$\frac{m\theta}{K} < 1$ and $\frac{m}{(m-K)(1-\theta)} < 1$

$$\Rightarrow \frac{\theta}{K}\left(\frac{m-K}{1-\theta}\right) < 1$$

$$\Rightarrow \log\left(\frac{\theta}{K}\left(\frac{m-K}{1-\theta}\right)\right) < 0$$

$\frac{d}{dK} \lambda(K) < 0 \Rightarrow$

$\lambda(K)$ decreases $\Rightarrow$

$\lambda(K) < \varepsilon \Rightarrow$ the hypothesis H_0 is rejected.

$\lambda(K) > \varepsilon \Rightarrow K > b$

for some $b \in \mathbb{R}$ as $\lambda(K)$ is a decreasing function.

$$\Rightarrow \sum_{i=1}^{n} Y_i > b$$

Hence we are done.