Sunandan Adhikary

21CS91R14

Stats Assignment 2

(1) $x_1 \dots x_n \overset{iid}{\sim} N(\mu, \sigma^2)$

(a) $T_1 = \dfrac{x_1 - x_2 \dots + (-1)^n x_n}{n} \neq \sum_{i=1}^{n} (-1)$

$\therefore$ MGF $M_{x_i}(t) = \exp(\mu t + \frac{\sigma^2 t^2}{2})$ $\because x_i \sim N(\mu, \sigma^2)$

$\therefore M_{T_1}(t) = \mathbb{E}\left[\exp\left(\frac{x_1 - x_2 \dots (-1)^n x_n}{n} t\right)\right]$

$= \prod_{i=1}^{n} \mathbb{E}\left[\exp\left(\frac{x_i}{n}(-1)^i\right)\right]$ $[x_i, x_j \; \forall i,j \le n \text{ iid}]$

$= \prod_{i=1}^{n} \exp\left(\frac{(-1)^i}{n}\mu t + \frac{\sigma^2 t^2}{2}\left(\frac{(-1)^i}{n}\right)^2\right)$

$= \exp\left(\frac{\mu t}{n} - \frac{\mu t}{n} + \dots n \text{ times}\right) \cdot \left(\exp\left(\frac{\sigma^2 t^2}{2}\frac{1}{n^2}\right)\right)^n$

if n = odd $\Rightarrow M_{T_1}(t) = \exp\left(\frac{\mu t}{n} + \frac{\sigma^2/n \; t^2}{2}\right)$

if n = even $\Rightarrow M_{T_1}(t) = \exp\left(0 \cdot t + \frac{\sigma^2/n \; t^2}{2}\right)$

$\therefore$ if n = odd $T_1 \sim N\left(\frac{\mu}{n}, \sigma^2/n\right)$

if n = even $T_1 \sim N(0, \sigma^2/n)$

(b) $T_2 = \dfrac{x_1 + \dots + x_m}{m}$ $1 < m < n$

$\Rightarrow M_{T_2}(t) = \mathbb{E}\left[\exp\left(\sum_{i=1}^{m} x_i/m\right)\right] = \prod_{i=1}^{m} \left[\exp\left(\frac{\mu}{m}t - \frac{\sigma^2/m^2 \; t^2}{2}\right)\right]$

$[\because x_i \sim N(\mu, \sigma^2)]$ and iid

$= \left[\exp\left(\frac{\mu}{m}t - \frac{\sigma^2/m^2 \; t^2}{2}\right)\right]^m$

$= \exp\left(\mu t - \frac{\sigma^2/m \; t^2}{2}\right)$

$\therefore T_2 \sim N\left(\mu, \frac{\sigma^2}{m}\right)$

Q1c. $T_3 = \frac{1}{\sigma^2}\sum_{i=1}^{m}(x_i - T_2)^2 = \sum_{i=1}^{m}\left(\frac{x_i - T_2}{\sigma}\right)^2 = \sum_{i=1}^{m}\left(z_i^2/\sigma^2\right)$

s.t. $z_i = x_i - T_2$, $1 \le i \le m$

Now $x_i \sim N(\mu, \sigma^2)$, $T_2 \sim N(\mu, \sigma^2/m) \Rightarrow z_i \sim N\left(0, \sigma^2 \frac{m+1}{m}\right)$

$\Rightarrow \frac{z_i - 0}{\sqrt{\frac{m+1}{m}}\,\sigma} \sim N(0,1) \Rightarrow z_i^* = \frac{z_i}{\sqrt{\frac{m+1}{m}}\,\sigma} \sim N(0,1)$

$\Rightarrow$ ~~So, $z_i^{*2} \sim \chi^2_m$ ... $f(z_i^{*2}) = \chi^2_m$~~

Since $T_3 = \left[\sum_{i=1}^{m}(z_i^*)^2\left(\frac{m+1}{m}\right)\right] \sim \left(\frac{m+1}{m}\right)^{-1}\chi^2_m$ $\left[\because \frac{dT_3}{d\left(\sum z_i^{*2}\right)} = \frac{m+1}{m}\cdot\frac{1}{\square}\right.$

(2) $Y_{n_i} \sim \chi^2_{n_i} \quad \forall\, i \in [1, 10]$

$\therefore$ MGF of $\chi^2_{n_i} = \mathbb{E}(e^{t}) = \displaystyle\int_0^\infty \frac{x^{n_i/2-1} e^{-x/2}}{2^{n_i/2}\,\Gamma(n_i/2)}\, e^{tx}\, dx$

$$= \cancel{(1-2t)^{-n_i/2}} \int_0^\infty \frac{x^{n_i/2-1} e^{-(1/2-t)x}}{2^{n_i/2}\,\Gamma(n_i/2)}$$

$[t < 1/2]$

$$= \frac{1}{(1-2t)^{n_i/2}} \int_0^\infty \left(\frac{1-2t}{2}\right)^{n_i/2} \frac{x^{n_i/2-1} \cdot e^{-(1-2t)x/2}}{\Gamma(n_i/2)}\, dx$$

$$= (1-2t)^{-n_i/2} \underbrace{\int_0^\infty \text{Gamma}\left(\frac{n_i}{2}, \frac{1-2t}{2}\right)}_{=1}$$

$$= (1-2t)^{n_i/2}$$

$\therefore$ (i) $Y_1 + Y_3 + \dots + Y_9 = Y$

$[\because Y_i \text{ iid}]$

$\therefore M_Y(t) = M_{Y_1}(t)\, M_{Y_2}(t) \dots$

$= (1-2t)^{-\sum \frac{n_i}{2}} \Rightarrow Y \sim \chi^2_{(n_1+n_2+\dots+n_9)}$

(2) (ii) As, $Y = \dfrac{Y'}{Y_5} = \dfrac{Y_1 + Y_4 + Y_6}{Y_5}$ From (i) we know $Y' \sim \chi^2_{(n_1+n_4+n_6)}$

since $Y^* = \dfrac{Y'/(n_1+n_4+n_6)}{Y_5/n_5} \sim F(n_1+n_4+n_6,\ n_5)$ $\because$ Y' and Y_5 iid χ^2

$\therefore Y^* = \dfrac{n_5}{n_1+n_4+n_6}\, Y \Rightarrow \left|\dfrac{dY^*}{dY}\right| = \dfrac{n_5}{n_1+n_4+n_6} = |J|^{-1}$

$\therefore f_Y(Y) = \dfrac{1}{|J|}\, f_{Y^*}(Y^*)$

$\Rightarrow Y \sim \left(\dfrac{n_1+n_4+n_6}{n_5}\right)^{-1} F(n_1+n_4+n_6,\ n_5)$

Q3 $Z = [z_1 \ldots z_n]^T$, $z_i \in \{0, 1\}$, $\sum_{i=1}^{n} z_i = 2$

(i) Z is discrete random vector since there are countably finite number of values in its range

(ii) z_i accepts 2 values i.e. 0, 1 in its range and there ~~should be $i = m, n$ s.t $z_m = z_n = 1$ and $z_{i \neq m,n} = 0$~~ exists $1 \leq p, q \leq n$ s.t $z_p = z_q = 1$ and $z_{i \neq p,q} = 0 \ \forall p, q$.

$\therefore$ Such $\langle p, q \rangle$ among n values of i can be found in ${}^nC_2$ ways. So there are ${}^nC_2$ possible sets of values within the range that the random vector Z considers

(iii) Yes, ~~there exi~~ a uniform pmf can be designed for Z if we denote $Z_{pq} = Z_{pq}$ s.t. $z_p = z_q = 1$ and $z_{i \neq p,q} = 0$ for any $p, q \in [1, n]$ (integers) then the uniform pdf

$$P(Z = Z_{pq}) = \frac{1}{{}^nC_2} \quad \forall \ 1 \leq p, q \leq n \text{ and } Z = Z_{pq}$$

$$= 0 \quad \text{elsewhere}$$

Q4 $X \sim N(3,2) \Rightarrow z = \frac{X-3}{\sqrt{2}} \sim N(0,1) \Rightarrow z^2 \sim \chi^2(1)$

$\therefore P(z^2 \leq a) = 0.7 \Rightarrow a \simeq 1.07$ from χ^2 table

$\therefore P(-1.07 \leq X \leq 1.07) = P((-1.07-3)/\sqrt{2} \leq z \leq (1.07-3)/\sqrt{2})$

$= P(-2.8779 \leq z \leq -1.3647)$

$= -\Phi(-2.8779) + \Phi(-1.3647)$

$\simeq 0.0862 - 0.002$

$= 0.0842$

5. $X_1 \dots X_n \overset{iid}{\sim} U(0,1) \Rightarrow f_X(x) = \frac{1}{1-0} = 1$ and $F_X(x) = \frac{x-0}{1-0} = x$

(ii) $X_{Max} = \max\{X_1, \dots, X_n\}$, ~~$P(X_{max})=$~~ pdf: $f_{MaxX}(\cdot)$, cdf: $F_{MaxX}(\cdot)$

~~$\therefore f_{Max}(x_{max}) = P((X_{max} \geq x_1)$ and $(X_{max} \geq x_2) \dots (X_{max} \geq x_n))$~~

$$P(X_{Max} \leq x_{max}) = F_{MaxX}(x_{max}) = P(X_1 \leq x_{max} \text{ and } X_2 \leq x_{max} \dots)$$

$$= F_X(x_{max})^n = (x_{max})^n \quad [\because X_1 \dots X_n \text{ are iid}]$$

$$\therefore f_{MaxX}(X_{Max}) = \frac{d}{dx_{max}}(x_{max})^n = n X_{Max}^{\,n}, \quad X_{Max} \in (0,1)$$

(i) $X_{Min} = \min\{X_1 \dots X_n\}$, pdf: $f_{MinX}(\cdot)$, cdf: $F_{MinX}(\cdot)$

~~$f_{MinX}(x_{min})$~~ $= P(X_1 \geq x_{min}$ and $\dots X_n \geq x_{min})$

$$P(X_{min} \geq x_{min}) = [1 - F_X(x_{min})]^n \quad [\because x_1 \dots x_n \text{ are iid}]$$

$$= (1 - x_{min})^n$$

$$\therefore F_{MinX}(x_{min}) = 1 - P(X_{Min} \geq x_{min}) = 1 - (1 - x_{min})^n$$

$$\therefore f_{MinX}(X_{Min}) = \frac{d}{dx_{Min}}\left[1 - (1 - x_{Min})^n\right] = \underline{n(1 - x_{min})^{n-1}} \quad \underline{X_{Min} \in (0,1)}$$

⑥ $X \sim U(0, \pi)$ $\quad Y = \sin X$, $\dfrac{dy}{dx} = \cos x = \sqrt{1-y^2}$, $R_Y = (0,1)$

$f_X = \dfrac{1}{-0+\pi} = \dfrac{1}{\pi}$

$\therefore$ Mediana $= m$ s.t. $\displaystyle\int_0^m f_Y(y)\,dy = 1/2$

$f_Y(y) = \dfrac{1}{\left|\frac{dy}{dx}\right|} f_X(x) = \dfrac{1}{\pi\sqrt{1-y^2}}$ (i.e. Cauchy dist.)

$\therefore \displaystyle\int_0^m \frac{1}{\pi\sqrt{1-y^2}}\,dy = 1/2 \Rightarrow \int_{\sin^{-1}0}^{\sin^{-1}(m)} dx = \pi/2 \Rightarrow \sin^{-1}(m) = \pi/2$

$[\because x \in (0,\pi)]$

$\therefore$ $\underline{m = \sin \pi/2 = 1 \text{ is median of } Y}$

Q7 $X \sim \text{Gama}(\mu, \sigma^2)$ $\because \mu = \frac{\alpha}{\beta} \quad \sigma^2 = \frac{\alpha}{\beta^2}$

$= \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$ where $\Rightarrow \beta = \mu/\sigma^2 \quad \alpha = \mu^2/\sigma^2$

$\because \mu = 2 \quad \sigma^2 = 4 \Rightarrow \alpha = \frac{4}{4} = 1, \beta = \frac{2}{4} = \frac{1}{2}$

$\therefore f_X(x) = \frac{1}{2} \cdot \frac{1}{\Gamma(1)} x^{1-1} \cdot e^{-x/2} = \frac{1}{2} e^{-x/2} \Rightarrow \text{exponential}(1/2)$

$$\int_0^n f_X(x)\,dx = \frac{1}{2}\int_0^n e^{-x/2}\,dx = \frac{1}{2}\cdot(-1/2)^{-1}\left[e^{-x/2}\right]_0^n$$

$$= -1 \left[e^{-n/2} - e^0\right]$$

$$= \left(1 - e^{-n/2}\right) = F_X(n) = P(x \le n)$$

$\therefore P(x \ge 8 \mid x \ge 5) = \frac{P(x \ge 8 \text{ and } x \ge 5)}{P(x \ge 5)}$

$$= \frac{P(x \ge 8)}{P(x \ge 5)} = \frac{1-(1-e^{-8/2})}{1-(1-e^{-5/2})} = \frac{e^{-4}}{e^{-2.5}} = e^{-1.5}$$

$= 0.2231$

Q8. Say, $X_1, X_2 \ldots X_n \overset{iid}{\sim} f_\theta(X)$ where θ is a parameter

The Maximum Likelihood Estimator helps in determining optimal parameter value that maximizes the likelihood function i.e. $f_X(x \mid \theta)$. θ can be any unknown parameter that can shape the pdf of X e.g. $\theta = \langle \mu, \sigma^2 \rangle$ for random normal samples, $\theta = \lambda$ for samples from exponential dist.

Steps: (i) Since Log is a monotonic and one to one function for easy computation we often maximize Log of the likelihood func. i.e. $\log f(x \mid \theta)$

(ii) To find the maxima of this function we simply find the first order derivative w.r.t. θ and find the θ value for which $\frac{d}{d\theta} \log f(x \mid \theta_{MLE}) = 0$

or $\left.\frac{d}{d\theta} \log f_\theta(x)\right|_{\theta = \theta_{MLE}} = 0 \Rightarrow \theta_{MLE} = \operatorname{argmax}(\log f(x \mid \theta))$

* This analytical method helps one choose the parameter value such that ~~the~~ is optimal ~~to~~ or most likely to generate desired set of samples from a given distribution function ~~fx~~. i.e. $X_1^* \ldots X_n^* \overset{iid}{\sim} f_{\theta_{MLE}}(X)$ where $x_1^* \ldots x_n^*$ are desired samples

Contd.

Q8 Contd.
MLE is an ~~unbiased~~ estimator like $\bar{x}$ when the sample size increases, However there are some challenges

(i) First obvious challange is we need a large number of samples to get an unbiased estimate since for ~~low~~ a less number of samples it works very poorly unlike $\bar{x}$.

(ii) For practical usecases, usually for mixture distrib-utions the analytical procedure might not always have a closed form which might lead to the use of sophisticated iterative gradient descent type ~~algo~~ numerical methods/algo which are often sensitive to the starting parameter value

(iii) It gives us a point estimate instead of an estimated probability distribution, so calculating sample variance and confidences are not trivi-al even though the actual method is easy to implement.

Q9 $X_1 \dots X_n \overset{iid}{\sim} f_x$ with mean μ variance σ^2, $n = 2k \geq 6$

$\therefore$ we can consider CLT to estimate $Y_1 - Y_2$

since n is large $Y_1 = X_1 + X_3 + \dots + X_{n-1}$

$Y_2 = X_2 + X_4 + \dots + X_n$

$$P\left(\frac{Y_1 - k\mu}{\sigma\sqrt{k}} \leq z\right) \approx \Phi(z) \text{ and } P\left(\frac{Y_2 - k\mu}{\sigma\sqrt{k}} \leq z\right) \approx \Phi(z$$

$\cancel{Y_3 = Y_1 - Y_2 = X_1 - X_2 + \dots - X_n \sim N(0, n\sigma^2)}$

$$\Rightarrow P(Y_1 \leq z) \approx \Phi\left(\frac{z - k\mu}{\sigma\sqrt{k}}\right) \text{ and } P(Y_2 \leq z) \approx \Phi\left(\frac{z - k\mu}{\sigma\sqrt{k}}\right)$$

$$\therefore Y_1, Y_2 \overset{approx}{\sim} N(k\mu, \sigma^2 k) \Rightarrow Y_3 = Y_2 - Y_1 \overset{approx}{\sim} N(k\mu - k\mu, 2k\sigma^2$$

If we approximate Y_3 directly using CLT $Y_3 = X_1 - X_2 + X_3 \dots - X_n$

$$\therefore P\left(\frac{Y_3 - (k\mu - k\mu)}{\sigma\sqrt{2k}} \leq z\right) \approx \Phi(z)$$

$$\Rightarrow P(Y_3 \leq z) \approx \Phi\left(\frac{z - 0}{\sigma\sqrt{2k}}\right) \Rightarrow Y_3 \overset{approx}{\sim} N(0, \sigma^2 2k)$$

Therefore we can see $Y_1 - Y_2 \overset{approx}{\sim} N(0, \sigma^2 2k)$ using C

for $k \geq 3$

Q 10 a

In a random experiment of n barnoulli trials (independent) with p success probability in each trial, the number of success ~~portions~~ is a discrete random variable X that follows Binomial distribution with mean $= np$ and variance $= np(1-p)$

$\therefore$ $X \sim \text{Binomial}(np, np(1-p))$

$\therefore$ for a large n (e.g. $n \geq 20$, $[p = 0.5]$) according to CLT

$$P\left(\frac{X - np}{\sqrt{np(1-p)}} \leq z\right) \approx \Phi(z)$$ where Φ is standard normal CDF

$$\left[\Rightarrow \quad \frac{X - np}{\sqrt{np(1-p)}} \underset{\text{discrete}}{\overset{\text{approx}}{\sim}} N(0,1) \right]$$

But since we are approximating a discrete ~~proof~~ r.v using continuous pdf (i.e. normal CDF) we need to do continuity corrections, so, after continuity correction

$$P(X \leq z) \approx \Phi\left(\frac{z + 0.5 - np}{\sqrt{np(1-p)}}\right) \text{ or}$$

$$P(z_1 \leq X \leq z_2) \approx \Phi\left(\frac{z_1 - 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{z_2 + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

since z, z_1, z_2 values must be included in discrete CMF

Eg. For $n = 20$ $p = 1/2$ $X \sim$ Binomial$(10, 5)$

$\rightarrow P(8 \le X \le 10) = \frac{1}{2^{20}} \left({}^{20}C_8 + {}^{20}C_9 + {}^{20}C_{10} \right) = 0.46$ [Using Binomial]

and $P(8 \le X \le 10) \approx -\Phi\left(\frac{8-0.5-10}{\sqrt{5}}\right) + \Phi\left(\frac{10+0.5-10}{\sqrt{5}}\right)$

≈ 0.4567 [using CLT]

Q10 b.

We know poisson distribution expresses the probability of any event occurence expressed using r.v X given an expected rate of occurence λ ~~i.e X ~ expone~~

i.e. $X \sim$ poisson (λ)

If $\lambda = np \Rightarrow p = \lambda/n$ in a Binomial experiment

then we get $P(X = x) = {}^nC_x \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$

if $n \to \infty$ then $\lim_{n \to \infty} P(X = x) = \lim_{n \to \infty} {}^nC_x \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$

$$= \frac{\lambda^x}{x!} \lim_{n \to \infty} \frac{n!}{(n-x)!} \frac{1}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x}$$

$$= \frac{\lambda^x}{x!} \lim_{n \to \infty} \left(1 - \frac{\lambda}{n}\right)^n \cdot \frac{n!}{(n-x)!} \frac{1}{n^x}$$

$$= \frac{\lambda^x}{x!} \lim_{n \to \infty} \left[\left(1 - \frac{\lambda}{n}\right)^{-n/\lambda}\right]^{-\lambda} \frac{(n-x+1)(n-x+2)\dots n}{n.n.\dots x\text{times}} \cdot \left(1 - \frac{\lambda}{n}\right)^{-x}$$

$$= \frac{\lambda^x}{x!} e^{-\lambda}$$

$\therefore$ we can see for $n \to \infty$ ($p \to 0$ $\because p = \frac{\lambda}{n}$) the Binomial pmf can be approximated using poisson pmf.