Jaideep Singh Chawla
190110030
jdchawla@iitb.ac.in

DS203 Programming for Data Sciences
Assignment 3

August 29, 2021

Exercise 1.

$$\begin{aligned}
Cov(X_1 + X_2, X_3) &= E[(X_1 + X_2)X_3] - E[X_1 + X_2]E[X_3] \\
&= E[X_1X_3 + X_2X_3] - (E[X_1] + E[X_2])E_(X_3) \\
&= E[X_1X_3] + E[X_2X_3] - E[X_1]E[X_3] - E[X_2]E[X_3] \\
&= E[X_1X_3] - E[X_1]E[X_3] + E[X_2X_3] - E[X_2]E[X_3] \\
&= Cov(X_1, X_3) + Cov(X_2 + X_3)
\end{aligned}$$

where, we make use of the linearity properties of the $E[X]$ function, i.e.,

$$\begin{aligned}
E[X + Y] &= E[X] + E[Y] \\
E[kX] &= kE[X]
\end{aligned}$$

Exercise 2.

$$\begin{aligned}
E[Y] = E\left[\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n\sigma^2}}\right] &= \frac{1}{\sqrt{n\sigma^2}}E\Big[n(\bar{X} - \mu)\Big] \\
&= \sqrt{\frac{n}{\sigma^2}}E[\bar{X} - \mu] \\
&= \sqrt{\frac{n}{\sigma^2}}(E[\bar{X}] - \mu) \\
&= \sqrt{\frac{n}{\sigma^2}}(\mu - \mu) \\
&= 0
\end{aligned}$$

These steps follow from the properties of $E[X]$ described in Exercise 1. Also, note that $\bar{X} = \sum X_i/n$
The expected value of sample mean $E[\bar{X}]$ is equal to the population mean μ i.e. $E[X_i]$

$$\begin{aligned}
Var[Y] = E[Y^2] - (E[Y])^2 = E\left[\left(\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n\sigma^2}}\right)^2\right] &- (0)^2 \\
&= \frac{n}{\sigma^2}E\Big[(\bar{X} - \mu)^2\Big] \\
&= \frac{n}{\sigma^2}E[\bar{X}^2 - 2\mu\bar{X} + \mu^2] \\
&= \frac{n}{\sigma^2}(E[\bar{X}^2] - 2\mu E[\bar{X}] + \mu^2) \\
&= \frac{n}{\sigma^2}(E[\bar{X}^2] - E[\bar{X}]^2) \\
&= \frac{n}{\sigma^2}Var[\bar{X}] \\
&= 1
\end{aligned}$$

Again we make use of the fact that $E[\bar{X}] = \mu$ and also that $Var(\bar{X}) = \sigma^2/n$

Exercise 3.

Sample:

104	109	111	109	87
86	80	119	88	122
91	103	99	108	96
104	98	98	83	107
79	87	94	92	97

Sample Mean: $\bar{X} = \frac{1}{25}\sum_{i=1}^{25} X_i = \mathbf{98.04}$
Sample Variance: $S^2 = \frac{1}{24}\sum_{i=1}^{25}(X_i - \bar{X})^2 = \mathbf{133.71}$
Sample Standard Deviation: $S = \sqrt{133.71} = \mathbf{11.56}$

Order Statistic:

79	80	83	86	87
87	88	91	92	94
96	97	98	98	99
103	104	104	107	108
109	109	111	119	122

Sample Range: **[79, 112]**
Sample Median: $X_{13} = \mathbf{98}$
Lower Quartile: $X_7 = \mathbf{88}$
Upper Quartile: $X_{19} = \mathbf{107}$

Confidence Interval:

$$1 - 2\exp(-25\epsilon^2) = 0.95 \implies \epsilon = \sqrt{-\frac{\ln 0.025}{25}} = 0.384$$

$$\text{The interval is } [98.04 - 0.384, 98.04 + 0.384] = [97.656, 98.424]$$

Alt. The statistic $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ has Student's t-distribution with $n-1$ degrees of freedom. So a 95% confidence interval can be built as, $\bar{X} \pm t_{0.05,24}\frac{s}{\sqrt{n}} = 98.04 \pm 2.064\frac{11.56}{\sqrt{25}} = 98.04 \pm 4.77 = [\mathbf{93.24, 102.81}]$

Exercise 4.

The standard error in estimating population mean with known standard deviation is given by $\frac{\sigma}{\sqrt{n}}$.
So, for population I, if the standard error is $\frac{\sigma_1}{\sqrt{n_1}}$, the standard error for population II is $\frac{\sigma_2}{\sqrt{n_2}} = \frac{2\sigma_1}{\sqrt{2n_1}} = \frac{\sqrt{2}\sigma_1}{\sqrt{n_1}}$.
If we ignore finite-population corrections, the populations can be considered to varying normally (CLT).
In that case, the 95% confidence interval falls within twice the standard error. From our earlier calculations, the 95% confidence interval will be $\sqrt{2}$ tighter for **Population I** and hence, it would be **more accurate.**

Exercise 5.

We have random sample $X_1, X_2 \sim \mathcal{N}(0, \sigma^2)$

$$F_{X_{(1)}}(x) = P(X_{(1)} \le x) = P(min(X_1, X_2) \le x) = 1 - P(X_1, X_2 > x) = 1 - [P(X > x)]^2 = 1 - \Big[1 - \Phi\Big(\frac{x}{\sigma}\Big)\Big]^2$$

Here, $\Phi(z)$ is the standard normal CDF.

$$\text{Now,} \quad f_{X_{(1)}}(x) = \frac{dF_{X_{(1)}}(x)}{dx} = \frac{2}{\sigma}\phi\Big(\frac{x}{\sigma}\Big)\Bigg(1 - \Phi\Big(\frac{x}{\sigma}\Big)\Bigg)$$

$\phi(z)$ is the standard normal PDF.

$$\begin{aligned}
\text{Finally,} \quad E[X_{(1)}] = \int_{-\infty}^{\infty} x f_{X_{(1)}}(x)dx &= \int_{-\infty}^{\infty} \frac{2x}{\sigma}\phi\Big(\frac{x}{\sigma}\Big)\Bigg(1 - \Phi\Big(\frac{x}{\sigma}\Big)\Bigg)\,dx \\
&= 2\sigma\int_{-\infty}^{\infty} t\phi(t)(1 - \Phi(t))\,dt \qquad [t = x/\sigma] \\
&= 2\sigma\int_{-\infty}^{\infty} t\phi(t)\,dt - \int_{-\infty}^{\infty} t\phi(t)\Phi(t)\,dt \\
&= 2\sigma\Bigg(0 - \frac{1}{\sqrt{1+1^2}}\phi(0)\Bigg) = -\boldsymbol{\frac{\sigma}{\sqrt{\pi}}} \qquad \text{[Ref.]}
\end{aligned}$$

Exercise 6.

Since, sampling is random, $Var[S] > 0$. Now, $Var[S] = E[S^2] - E[S]^2 \implies \sigma^2 - E[S]^2 > 0$ or $\sigma > E[S]$
So, the estimaotr is **biased**.

Exercise 7.

Let $\mathbf{x} = (2, 3, 2, 1, 0, 0, 3, 2, 1, 1)$, then,

$$\begin{aligned}
\mathcal{L}(\theta|\mathbf{x}) &= \frac{3}{5}(1-\theta).\frac{2}{5}(1-\theta).\frac{3}{5}(1-\theta).\frac{2}{5}\theta.\frac{3}{5}\theta.\frac{3}{5}\theta.\frac{2}{5}(1-\theta).\frac{3}{5}(1-\theta).\frac{2}{5}\theta.\frac{2}{5}\theta \\
&= \frac{2^5.3^3}{5^8}\theta^4(1-\theta)^4 \\
\frac{d\mathcal{L}(\theta|\mathbf{x})}{dx} &= \frac{2^5.3^3}{5^8}\left[4\theta^3(1-\theta)^4 - 4\theta^4(1-\theta)^3\right] = 0 \implies \hat{\theta} = \boldsymbol{\frac{1}{2}}
\end{aligned}$$

Exercise 8.

$$\mathcal{L}(\theta|x) = \frac{\theta}{(1+x)^{\theta+1}}$$

$$\frac{d\mathcal{L}(\theta|x)}{d\theta} = \frac{d}{d\theta}\theta(1+x)^{-\theta-1} = (1+x)^{-\theta-1}(1-\theta\log(x+1)) = 0 \implies \hat{\theta} = \frac{\mathbf{1}}{\mathbf{log(x+1)}}$$

Exercise 9.

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{x=1}^{n}\theta x^{\theta-1} = \theta^n(\Pi\, x_i)^{\theta-1}$$

$$\frac{d\mathcal{L}(\theta|x)}{d\theta} = \frac{d}{d\theta}\theta^n(\Pi\, x_i)^{\theta-1} = (\Pi\, x_i)^{\theta-1}\theta^n(\theta\,\Sigma\log(x_i) + n) = 0 \implies \hat{\theta} = -\frac{\boldsymbol{n}}{\boldsymbol{\Sigma\log(x_i)}}$$