Anmol Saraf
Email: 200070007@iitb.ac.in
Course: DS203 - Programming for Data Science Minor
Instructor: Dr. Manjesh K Hanawal

Assignment - 3
Roll No.: 200070007
Semester: Autumn 2021
Due Date: 29^{th} August, 2021

Exercise 1
For any random variables X_1, X_2, X_3 defined on the same sample space, show that

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

Solution: We know that the covarriance of two random variables A and B is defined as $Cov(A, B) = E(AB) - E(A)E(B)$.

Note, $E(A + B) = E(A) + E(B)$, where A and B are random variables. Thus using this definition we solve the problem as follows,

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

Exercise 2
Let $X_1, X_2, \ldots, X_n$ are i.i.d. such that $\mu = E(X_i)$ and $\sigma^2 = Var(X_i)$ for all $1 \leq i \leq n$. Define

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

Show that $E(Y) = 0$ and $Var(Y) = 1$.

Solution: FIrstly to find $E(Y)$ we take the expectation of the given expression in the question,

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

Now, since $\mu = E(X_i)$ for all $1 \le i \le n$, $\sum_{i=1}^{n} E(X_i) = n\mu$.

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

Now for the second part we take the Variance of Y. Note, $Var(A) = E[(A - E(A))^2]$. Therefore,

$$\begin{aligned} Var(Y) &= E[(Y - E(Y))^2] \\ &= E[(Y - 0)^2] \\ &= E(Y^2) \\ &= E\left[\left(\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n\sigma^2}} \right)^2 \right] \\ &= E\left[\left(\frac{\sum_{i=1}^{n}(X_i) - n\mu}{\sqrt{n\sigma^2}} \right)^2 \right] \\ &= \frac{1}{n\sigma^2} E\left[\sum_{i=1}^{n} X_i^2 + 2 \cdot \sum\sum_{1 \le i < j \le n} X_i X_j - 2n\mu \cdot \sum_{i=1}^{n} X_i + n^2\mu^2 \right] \\ &= \frac{1}{n\sigma^2} \left[\sum_{i=1}^{n} E(X_i^2) + 2 \cdot \sum\sum_{1 \le i < j \le n} E(X_i X_j) - 2n\mu \cdot \sum_{i=1}^{n} E(X_i) + n^2\mu^2 \right] \end{aligned}$$

Since, $Var(X_i) = E(X_i^2) - [E(X_i)]^2 = E(X_i^2) - \mu^2$ for all $1 \le i \le n$. Therefore, $E(X_i^2)$ can be written as $Var(X_i) + \mu^2 = \sigma^2 + \mu^2$ for all $1 \le i \le n$. Thus,

$$\begin{aligned} Var(Y) &= \frac{1}{n\sigma^2} \cdot \left[\sum_{i=1}^{n} (\sigma^2 + \mu^2) + 2 \cdot \sum\sum_{1 \le i < j \le n} E(X_i X_j) - 2n\mu \cdot \sum_{i=1}^{n} \mu + n^2\mu^2 \right] \\ &= \frac{1}{n\sigma^2} \cdot \left[n\sigma^2 + n\mu^2 - 2n^2\mu^2 + n^2\mu^2 + 2 \cdot \sum\sum_{1 \le i < j \le n} E(X_i X_j) \right] \\ &= \frac{1}{n\sigma^2} \cdot \left[n\sigma^2 + n\mu^2 - n^2\mu^2 + 2 \cdot \sum\sum_{1 \le i < j \le n} E(X_i X_j) \right] \end{aligned}$$

Since all X_i's are i.i.d.s, $Cov(X_i, X_j) = 0$ for all $1 \le i, j \le n$ and $i \ne j$. Also Note, $Cov(A, B) = E(AB) - E(A)E(B)$. Thus, we can write each $E(X_i X_j) = Cov(X_i, X_j) + E(X_i)E(X_j) =$

$E(X_i)E(X_j)$. Thus,

$$\begin{aligned}
Var(Y) &= \frac{1}{n\sigma^2} \cdot \left[n\sigma^2 + n\mu^2 - n^2\mu^2 + 2 \cdot \sum_{1 \le i < j \le n}\sum E(X_i X_j) \right] \\
&= \frac{1}{n\sigma^2} \cdot \left[n\sigma^2 + n\mu^2 - n^2\mu^2 + 2 \cdot \sum_{1 \le i < j \le n}\sum E(X_i)E(X_j) \right] \\
&= \frac{1}{n\sigma^2} \cdot \left[n\sigma^2 + n\mu^2 - n^2\mu^2 + 2 \cdot \sum_{i=1}^{n} E(X_i) \sum_{j=i+1}^{n} E(X_j) \right] \\
&= \frac{1}{n\sigma^2} \cdot \left[n\sigma^2 + n\mu^2 - n^2\mu^2 + 2\mu \cdot \sum_{i=1}^{n} E(X_i)(n-i) \right] \\
&= \frac{1}{n\sigma^2} \cdot \left[n\sigma^2 + n\mu^2 - n^2\mu^2 + 2\left(n^2 - \frac{n(n+1)}{2} \right) \mu^2 \right] \\
&= \frac{1}{n\sigma^2} \cdot \left[n\sigma^2 + n\mu^2 - n^2\mu^2 + n^2\mu^2 - n\mu^2 \right] \\
&= \frac{1}{n\sigma^2} \cdot [n\sigma^2] = 1
\end{aligned}$$

Hence Proved that $E(Y) = 0$ and $Var(Y) = 1$.

Exercise 3

A random sample of a population of size 2000 yields the following values 25 values:

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

- Calculate sample mean, sample variance, and sample standard deviantions.
- Calculate sample range, median, lower and upper quartiles.
- Give approximate 95% confidence intervals for the population mean.

Solution:

Part a The Sample mean $\bar{X}$ is,

$$\begin{aligned}
\bar{X} &= \frac{1}{n} \cdot \sum_{i=1}^{n} X_i \\
&= \frac{1}{25} \cdot [2451] \\
&= 98.04
\end{aligned}$$

The Sample Variance S^2 is given as,

$$\begin{aligned} S^2 &= \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \bar{X})^2 \\ &= \frac{1}{24} \cdot \sum_{i=1}^{25} (X_i - 98.04)^2 \\ &= \frac{1}{24} \cdot [3208.96] \\ &\approx 133.707 \end{aligned}$$

Thus using the Sample Variance, the Sample standard deviation S is,

$$\begin{aligned} S &= \sqrt{S^2} \\ &= \sqrt{133.707} \\ &= 11.56 \end{aligned}$$

Part b For the second part of this question we use the ordered statistics in the increasing order. Thus, the 25 values will be ordered as:

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

For the sample range, it is given as $X_{(n)} - X_{(1)}$, where $X_{(n)}$ and $X_{(1)}$ are the largest and the smallest in the sample.
Therefore, Sample Space = 122 - 79 = 43.

Median is the middle value of the ordered sample, so for 25 values the 13^{th} value would be the median, which is 98 here.

$$(100p)^{th} sample\ percentile\ is = \begin{cases} X_{(\lceil np \rceil)} & if\ p < 0.5 \\ X_{(n+1-\lceil n(1-p) \rceil)} & if\ p \geq 0.5 \end{cases}$$

Thus, the lower and upper quartiles are the 25^{th} and the 75^{th} sample percentile respectively.

Therefore,
Lower Quartile = $\lceil 25 \times 0.25 \rceil = \lceil 6.25 \rceil = 7^{th}$term of the orser statistics, which is 88.
Upper Quartile = $25+1-\lceil 25 \times 0.25 \rceil = 26 - \lceil 6.25 \rceil = 26 - 7 = 19^{th}$ term of the order statistiics, which is 107.

Therefore, the Sample space is 43, the Median is 98 and the lower and upper quartiles are 88 and 107 respectively.

Part c To have 95 % confidence interval, we set $P(|\hat{\mu} - \mu| \leq \epsilon) = 0.95$. Thus, using the Hoeffding's Inequality $P(|\hat{\mu} - \mu| > \epsilon) \leq 2e^{-n\epsilon^2}$, we get,

$$0.05 \leq 2e^{-n\epsilon^2}$$

Thus to find the confidence interval we equate them and find the biggest ϵ allowed,

$$\begin{aligned}
2e^{-n\epsilon^2} &= 0.05 \\
e^{-n\epsilon^2} &= 0.025 \\
n\epsilon^2 &= -\ln(0.025) = 3.689 \\
\epsilon^2 &= \frac{3.689}{25} = 0.14756 \\
\epsilon &= \sqrt{0.14756} = 0.384
\end{aligned}$$

The population mean $\mu = E(\bar{X})$ will have a confidence interval with 95 % as $(\mu - \epsilon, \mu + \epsilon)$, where $\epsilon = 0.384$.

Exercise 4
Two populations are surveyed with random samples. A sample of size n_1 is used for population I, which has a population standard deviation σ_1; a sample of size $n_2 = 2n_1$ is used for population II, which has a population standard deviation $\sigma_2 = 2\sigma_1$. Ignoring finite population corrections, in which of the two samples would you expect the estimate of the population mean to be more accurate? Justify your answer!

Solution: The standard error of a sample is defined as the standard deviation of the sample mean $\bar{X}$. So, firstly we find $Var(\bar{X})$,

$$\begin{aligned}
Var(\bar{X}) &= Var\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \\
&= E\left(\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2\right) - \left[E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\right]^2 \\
&= \frac{1}{n^2} \cdot \left[\left(\sum_{i=1}^{n} E(X_i^2) + 2 \cdot \sum\sum_{1 \le i << j \le n} E(X_i X_j)\right) - \left(\sum_{i=1}^{n} [E(X_i)]^2 + 2 \cdot \sum\sum_{1 \le i << j \le n} E(X_i) \cdot E(X_j)\right)\right] \\
&= \frac{1}{n^2} \cdot \left[\sum_{i=1}^{n} Var(X_i) + 2 \cdot \sum\sum_{1 \le i << j \le n} Cov(X_i, X_j)\right] \\
&= \frac{1}{n^2} \cdot [n\sigma^2 + 0] = \frac{\sigma^2}{n}
\end{aligned}$$

Note, as $Cov(X_i, X_j) = 0$ as they are i.i.d.

Thus the standard error $= \sqrt{Var(\bar{X})} = \dfrac{\sigma}{\sqrt{n}}$.

Therefore using this as a meausre of error, we can get an estimate of how accurate the population mean would be. For Population-I, the standard error is $\dfrac{\sigma_1}{\sqrt{n_1}}$ while for the Population-II, the standard error is $\dfrac{\sigma_2}{\sqrt{n_2}} = \dfrac{\sqrt{2} \cdot \sigma_1}{\sqrt{n_1}}$.

Thus, the second sample would have a higher standard error, so we can expect the estimate of population mean of sample I to be more accurate.

Exercise 5
Let (X_1, X_2) denote random samples drawn from population distribution $\mathcal{N}(0, \sigma^2)$. Find mean of the first order statistics, i.e., $\mathbb{E}(X_{(1)})$.

Solution: The PDF of $\mathcal{N}(0, \sigma^2)$ is $f_X(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-x^2/2\sigma^2}$ as it is a Gaussian distribution by definition.
Now, $X_{(1)}$ denotes the minimum of X_1, X_2. So, we can write $E(\min\{X_1, X_2\})$ as an integral as follows,

$$E(\min\{X_1, X_2\}) = \int\int_{\mathbb{R}^2} \min\{x_1, x_2\} \cdot f_{X_1}(x_1) f_{X_2}(x_2) dx_1 \cdot dx_2$$

Thus, using the symmetry of the integral we can write it as,

$$\begin{aligned}
\mathbb{E}(X_{(1)}) &= 2 \int\int_{x_1 < x_2} x_1 f(x_1) f(x_2) dx_1 dx_2 \\
&= \frac{1}{\pi\sigma^2} \int_{-\infty}^{\infty} \int_{-\infty}^{x_2} x_1 e^{-x_1^2/2\sigma^2} e^{-x_2^2/2\sigma^2} dx_1 dx_2 \\
&= \frac{1}{\pi} \int_{-\infty}^{\infty} \left[-e^{-x_1^2/2\sigma^2} \right]_{x_1=-\infty}^{x_1=x_2} e^{-x_2^2/2\sigma^2} dx_2 \\
&= -\frac{1}{\pi} \int_{-\infty}^{\infty} e^{-x_2^2/\sigma^2} dx_2
\end{aligned}$$

Since we know the value of integral $\int_{-\infty}^{\infty} e^{-ax^2} dx = \sqrt{\frac{\pi}{a}}$. Therefore,

$$\mathbb{E}(X_{(1)}) = -\frac{1}{\pi} \cdot \sqrt{\pi\sigma^2} = -\frac{\sigma}{\sqrt{\pi}}$$

Exercise 6
Suppose $\bar{X}$ and S^2 are sample mean and sample variance calculated from a random sample $X_1, X_2, \ldots, X_n$ drawn from a population with finite mean μ and variance σ^2. We know that $\bar{X}$ and S^2 are unbiased estimator of mean and variance, respectively. Is the sample standard deviation (S) is an unbiased estimator of σ? Justify your claim.

Solution: For the Sample standard deviation S to be an unbiased estimator of σ, $\mathbb{E}(S) = \sigma$, for any type of distribution.
Since $Var(S) > 0$ and cannot be equal to zero as all the elements are not equal. Thus,

$$\begin{aligned}
Var(S) &= E(S^2) - [E(S)]^2 \\
0 &< E(S^2) - [E(S)]^2 \\
[E(S)]^2 &< E(S^2) \\
E(S) &< \sqrt{E(S^2)} = \sqrt{\sigma^2} \\
E(S) &< \sigma
\end{aligned}$$

Since, Variance is an Unbiased estimator $E(S^2) = \sigma^2$. Thus, $E(S) < \sigma$ and cannot be an unbiased estimator of standard deviation.

Exercise 7
Let X be a discrete random variable with the following PMF.

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

where $0 \leq \theta \leq 1$ is a parameter. The following 10 independent observation of X are made: (2,3,2,1,0,0,3,2,1,1).

- Find the likelihood function for θ
- Find maximum likelihood estimate of θ

Solution: The likelihood function $L(x_1, x_2, \ldots, x_{10}; \theta)$ will be

$$\begin{aligned} L(x_1, x_2, \ldots, x_{10}; \theta) &= P_{X_1 X_2 \ldots X_{10}}(x_1, x_2, \ldots, x_{10}; \theta) \\ &= P_{X_1 X_2 \ldots X_{10}}(2, 3, 2, 1, 0, 0, 3, 2, 1, 1; \theta) \\ &= \left(\frac{3}{5}\theta\right)^2 \left(\frac{2}{5}\theta\right)^3 \left(\frac{3}{5}(1-\theta)\right)^3 \left(\frac{2}{5}(1-\theta)\right)^2 \\ &= \frac{2^5 3^5}{5^{10}} \cdot \theta^5 \cdot (1-\theta)^5 \end{aligned}$$

Now, to find the Maximum Likelihood Estimate (MLE) of θ, we differentitate the likelihood function and equate it to 0.

$$\begin{aligned} 0 &= \frac{dL}{d\theta} \\ 0 &= \frac{2^5 3^5}{5^{10}} \left[5\theta^4 \cdot (1-\theta)^5 - 5\theta^5 \cdot (1-\theta)^4\right] \\ 0 &= \theta - (1-\theta) \end{aligned}$$

Eliminating the cases of $\theta = 0$ or 1, as that will be a minima. Thus, $\theta = 0.5$ will give the maximum.

Therefore, the Maximum Likelihood Estimate of θ is $\frac{1}{2}$.

Exercise 8
Let X is continuous random variable with the following PDF

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

where $\theta > 0$.

- Find the likelihood function for θ
- Find maximum likelihood estimate of θ.

Solution: The likelihood function $L(\theta|x)$ will be,

$$L(\theta|x) = f(x|\theta) = \begin{cases} \frac{\theta}{(1+x)^{\theta+1}} & for\ 0 < x < \infty \\ 0 & otherwise \end{cases}$$

To find the MLE of θ, we differentiate $L(\theta|x)$ and equate it to 0.

$$\begin{aligned}
\frac{dL(\theta|x)}{d\theta} &= 0 \\
\frac{d}{d\theta}\left(\frac{\theta}{(1+x)^{\theta+1}}\right) &= 0 \\
\frac{1}{(1+x)^{\theta+1}} - \frac{\theta}{(1+x)^{\theta+1}} \cdot \ln(1+x) &= 0 \\
\theta \cdot \ln(1+x) - 1 &= 0 \\
\theta &= \frac{1}{\ln(1+x)}
\end{aligned}$$

Thus, the Maximum Likelihood estimate of θ is $\dfrac{1}{\ln(1+x)}$.

Exercise 9
Let $X_1, X_2, \ldots, X_n$ be i.i.d. with PDF,

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

Find the MLE of θ, and show that its variance $\to 0$ as $n \to \infty$.

Solution: To find the MLE of θ we differentitate and equate $L(\theta|x)$ to 0. Here $L(\theta|x) = f(x|\theta) = \theta x^{(\theta-1)}$. Thus,

$$\begin{aligned}
\frac{dL(\theta|x)}{d\theta} &= 0 \\
\frac{d}{d\theta}(\theta x^{(\theta-1)}) &= 0 \\
x^{(\theta-1)} + \theta x^{(\theta-1)} \cdot \ln(x) &= 0 \\
-1 &= \theta \ln(x) \\
\theta &= -\frac{1}{\ln(x)}
\end{aligned}$$

Therefore the MLE of θ is,

$$\hat{\theta}(x) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\ln(x_i)}$$

Let $Y_i = -\dfrac{1}{\ln(X_i)}$ be an i.i.d. Therefore, its variance would be $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$.

Since, X lies between 0 and 1, $1/\ln(X)$ would decrease with less power than n and thus it will tend to 0 as n tends to ∞.

Exercise 10
Suppose we observe m i.i.d. Bernoulli(θ) random variables, denoted by $Y_1, Y_2, \ldots, Y_m$. Show that the LRT of $H_0 : \theta \le \theta_0$ versus $H_1 : \theta > \theta_0$ will reject H_0 if $\sum_{i=1}^{m} Y_i > b$.

Solution: (Could Not Solve. Skipped.)