SC 617

Quiz-Week1

Gyandev Satyaram Gupta, 190100051

1. For any random variables X_1, X_2, X_3 defined on the same sample space, show that $\text{Cov}(X_1 + X_2, X_3\ ) = \text{Cov}(X_1, X_3\ ) + \text{Cov}(X_2, X_3\ )$

Solution: We know that E(X+Y) = E(X) + E(Y) and $E(aX+b) = aE(X)+b$

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

Hence proved

2. Let $X_1, X_2, ..., X_n$ 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: We use $Var(aX + b) = a^2Var(X)$

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

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

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

Solution: i. Sample mean = $\sum_{i=1}^{n} \frac{x_i}{n} = \sum_{i=1}^{2} 5\frac{x_i}{25} = 98.04$
Sample variance $s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1} = 133.707$
Sample standard deviation s = $\sqrt{133.707} = 11.563$

ii. Sample range = 122 - 79 = 43
Median = 98
Lower Quartile = 88
Upper Quartile = 107

iii. Take μ to be the actual mean

$$P(|\mu - \hat{\mu}| \leq \epsilon) \geq 0.95$$

We know that $P(|\mu - \hat{\mu}| > \epsilon) \leq \delta$, where $\delta = 2e^{-n\epsilon^2}$
Consider the marginal case,

$$P(|\mu - \hat{\mu}| > \epsilon) = \delta$$
$$P(|\mu - \hat{\mu}| \leq \epsilon) = 1 - \delta$$
$$\implies 1 - \delta \geq 0.95$$
$$\implies 0.05 \geq= 2e^{-25\epsilon^2}$$
$$\implies \epsilon \geq \sqrt{\frac{-ln(0.025)}{25}}$$
$$\implies \epsilon \geq 0.3841$$

So confidence interval (CI) = $[98.04-0.3841, 98.04+0.3841] = [97.6559, 98.4241]$

4. Two populations are surveyed with random samples. A sample of size n1 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: We know that standard deviation of sample mean = σ/n
So for case I , its $\sigma_1' = \sigma_1/\sqrt{n_1}$
So for case II, its $\sigma_2' = \sigma_2/\sqrt{n_2} = \sqrt{2}\sigma$
It means that we would get better accuracy in estimating sample mean for case I compared to case II

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., $E(X_{(1)})$.

Solution: The CDF of gaussian distribution for $\mathcal{N}(0, \sigma^2)$

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

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

$$\begin{aligned} P(min(X_1, X_2) > m) &= P(X_1 > m, X_2 > m) \\ &= P(X_1 > m)P(X_2 > m) \\ &= (1 - F_X(m))^2 \\ &= (1 - \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{m} exp(-x^2/2\sigma^2)dx)^2 \end{aligned}$$

$\implies P(min(X, Y) \leq m) = 1 - (1 - \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{m} exp(-x^2/2\sigma^2)dx)^2$
$\implies F_Z(z) = 1 - (1 - \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{z} exp(-x^2/2\sigma^2)dx)^2$
$\implies f_Z(z) = dF/dz = (1 - \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{z} exp(-x^2/2\sigma^2)dx) * \frac{2}{\sqrt{2\pi\sigma^2}} exp(-z^2/2\sigma^2)$

Now for expectation

$$\begin{aligned} E(Z) = E(X_{(1)}) &= \int_{-\infty}^{\infty} z f_Z(z)dz \\ &= \int_{-\infty}^{\infty} z\frac{2}{\sqrt{2\pi\sigma^2}} exp(-z^2/2\sigma^2)dz \\ &- \int_{-\infty}^{\infty} z\frac{2}{\sqrt{2\pi\sigma^2}} exp(-z^2/2\sigma^2)\frac{1}{\sqrt{2\pi\sigma^2}}(\int_{-\infty}^{z} exp(-x^2/2\sigma^2)dx)dz \\ &\text{Using fubini theorem and integral of odd fn from } -\infty \text{to} \infty \text{is } 0 \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-x^2/2\sigma^2)\frac{1}{\sqrt{2\pi\sigma^2}}(\int_{-\infty}^{x} 2z exp(-z^2/2\sigma^2)dz)dx \\ &= -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp(-x^2/2\sigma^2)\frac{1}{\sqrt{2\pi\sigma^2}} * 2\sigma^2 * e^{-x^2/2\sigma^2}dx \\ &= -\sigma/\sqrt{\pi} \end{aligned}$$

6. Suppose $\overline{X}$ and S^2 are sample mean and sample variance calculated from a random sample $X_1, X_2, ..., X_n$ drawn from a population with finite mean μ and variance σ^2. We know that $\overline{X}$ 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: We already know that $E(S^2) = \sigma^2$

$$Var(S) = E(S^2) - E(S)^2 > 0$$
$$E(S) < \sqrt{E(S^2)} = \sigma$$

So we got that E(S) is strictly less than σ so standard deviation (S) is not an unbiased estimator of σ

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

$$P(X = x) = \begin{cases} 3\theta/5 & \text{if } x = 0 \\ 2\theta/5 & \text{if } x = 1 \\ 3(1-\theta)/5 & \text{if } x = 2 \\ 2(1-\theta)/5 & \text{if } x = 3 \end{cases}$$

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: i.

$$\begin{aligned} L(\theta|x) &= P(X = 2)P(X = 3)P(X = 2)P(X = 1)P(X = 0) \\ &P(X = 0)P(X = 3)P(X = 2)P(X = 1)P(X = 1) \\ &= (3(1-\theta)/5)^3 * (3\theta/5)^2 * (2\theta/5)^3 * (2(1-\theta)/5)^2 \end{aligned}$$

ii. Let's find log likelihood function

$$\begin{aligned} l(\theta|x) &= log(L(\theta|x)) \\ &= 3(log(3/5) + log(1-\theta)) + 2(log(3/5) + log(\theta)) + \\ &3(log(2/5) + log(\theta)) + 2(log(2/5) + log(1-\theta)) \\ &= C + 5log(\theta) + 5log(1-\theta) \end{aligned}$$

$\frac{dl}{d\theta} = 0 = 5/\theta - 5/(1-\theta) \implies \hat{\theta} = 0.5$
For $\theta < \hat{\theta}, log(L(\theta|x)$ is increasing and for $\theta > \hat{\theta}, log(L(\theta|x)$ is decreasing. so its a global maximum thus the MLE.

8. Let X is continuous random variable with the following pdf

$$f(x|\theta) = \begin{cases} \frac{\theta}{(1+x)^{(1+\theta)}} & \text{if } x > 0 \\ 0 & \text{if } otheriwse \end{cases}$$

where $\theta > 0$ is a parameter.
• Find the likelihood function for θ
• Find maximum likelihood estimate of θ

Solution:

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

Now calculate log likelihood

$$\begin{aligned} l(\theta|x) &= log(L(\theta|x)) \\ &= log(\theta) + (\theta+1)(log(1+x)) \end{aligned}$$

$\frac{dl}{d\theta} = 0 = 1/\theta + log(1+x) \implies \hat{\theta} = -1/log(1+x)$
For $\theta < \hat{\theta}, log(L(\theta|x)$ is increasing and for $\theta > \hat{\theta}, log(L(\theta|x)$ is decreasing. so its a global maximum thus the MLE.

9. Let $X_1, X_2, ..., X_n$ be iid with pdf

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

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

Solution:

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

Now calculate log likelihood

$$\begin{aligned} l(\theta|x) &= log(L(\theta|x)) \\ &= nlog(\theta) + (\theta-1)(log(\prod_{i=1}^{n} X_i)) \end{aligned}$$

$\frac{dl}{d\theta} = 0 = n/\theta + log(\prod_{i=1}^{n} X_i) \implies \hat{\theta} = -n/log(\prod_{i=1}^{n} X_i)$
For $\theta < \hat{\theta}, log(L(\theta|x)$ is increasing and for $\theta > \hat{\theta}, log(L(\theta|x)$ is decreasing. so its a global maximum thus the MLE.

$$Variance = \frac{-1}{\frac{d^2}{d\theta^2}(l(\theta|x)} = \frac{\theta^2}{n}$$

As $n \rightarrow \infty$ variance $\rightarrow 0$

10. **Solution:** $Y \sim Ber(\theta)$

$$
\begin{aligned}
L(\theta|y) &= \prod_{i=1}^{n} \theta^{y_i}(1-\theta)^{1-y_i} \\
&= \theta^{\sum y_i}(1-\theta)^{n-\sum y_i}
\end{aligned}
$$

$$
\begin{aligned}
l(\theta|x) &= log(L(\theta|x)) \\
&= \sum y_i log(\theta) + (n - \sum y_i)(log(1-\theta))
\end{aligned}
$$

$\frac{dl}{d\theta} = 0 = \sum y_i/\theta - (n - \sum y_i)/(1-\theta) \implies \hat{\theta} = \sum y_i/n$
For $\theta < \hat{\theta}, log(L(\theta|x)$ is increasing and for $\theta > \hat{\theta}, log(L(\theta|x)$ is decreasing. so its a global maximum thus the MLE.

Now take the LRT and take y $= \sum y_i$

$$
\lambda(y) = \frac{sup_{\theta<\theta_0} L(\theta|y_1,..y_n)}{sup_{\Theta} L(\theta|y_1,..y_n)}
$$

$$
\lambda(y) = \begin{cases} 1 & \text{if } \theta_0 \geq \hat{\theta} \\ \frac{\theta_0^y(1-\theta_0)^{n-y}}{\hat{\theta}^y(1-\hat{\theta})^{n-y}} & \text{if } otherwise \end{cases}
$$

Now consider log($\lambda(y)$) otherwise part
$log(\lambda(y)) = ylog(\theta_0) + (n-y)log((1-\theta_0) - ylog(\hat{\theta}) - (n-y)log((1-\hat{\theta})$
$d(log\lambda(y))/d\lambda = log(\theta_0) - log((1-\theta_0) - log(\hat{\theta}) - 1 + log((1-\hat{\theta}) + 1$
$d(log\lambda(y))/d\lambda = log(\theta_0 * (1-\hat{\theta})/(\hat{\theta} * (1-\theta_0)))$

If we have above expression as negative it means that its decreasing fn. and so its always going to be less than 1 and so we can always find a c$\in (0,1)$

$\theta_0 * (1-\hat{\theta})/(\hat{\theta} * (1-\theta_0)) < 1$
We get $\hat{\theta} > \theta_0 \implies \sum y_i > n\theta = b$