Statistics for Data Science - 2

Week 8 Graded Assignment Solution

1. Let $X_1, X_2, \dots, X_n$ be i.i.d. samples from a distribution X with mean μ and standard deviation σ. Let $\hat{\mu} = 6\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)$ be an estimator of μ.

   i) Is the estimator unbiased?

   a) Yes

   b) No

   Solution:

$$\begin{aligned} E[\hat{\mu}] &= E\left[6\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)\right] \\ &= \frac{6}{n}(n\mu) \\ &= 6\mu \end{aligned}$$

   And
   Bias$(\hat{\mu}, \mu) = E[\hat{\mu}] - \mu = 6\mu - \mu = 5\mu$
   Since, Bias$(\hat{\mu}, \mu) \neq 0$, therefore the estimator is not unbiased.

   ii) Find the risk of $\hat{\mu}$.

   (a) $\frac{36\sigma^2}{n} + 25\mu^2$

   (b) $\frac{36\sigma^2}{n} + 5\mu$

   (c) $\frac{6\sigma^2}{n} + 25\mu^2$

   (d) $\frac{6\sigma^2}{n} + 5\mu$

   Solution:

$$\begin{aligned} \text{Var}(\hat{\mu}) &= \text{Var}\left[6\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)\right] \\ &= \frac{36}{n^2}(n\sigma^2) \\ &= \frac{36\sigma^2}{n} \end{aligned}$$

$$\begin{aligned}\text{Risk}(\hat{\mu}) &= \text{Bias}(\hat{\mu}, \mu)^2 + \text{Var}(\hat{\mu}) \\ &= (5\mu)^2 + \frac{36\sigma^2}{n} \\ &= 25\mu^2 + \frac{36\sigma^2}{n}\end{aligned}$$

2. Consider a sample of iid random variables $X_1, X_2, \ldots, X_n$, where $n > 20, E[X_i] = \mu, \text{Var}(X_i) = \sigma^2$ and the estimator of μ, $\hat{\mu}_n = \dfrac{1}{n-20}\sum_{i=21}^{n} X_i$. Find the MSE of $\hat{\mu}_n$.

   a) $\frac{\sigma}{n-20}$
   b) $\frac{\sigma^2}{n-20}$
   c) $\frac{\sigma^2}{n-21}$
   d) $\frac{\sigma}{n}$

   Solution:

$$\begin{aligned}E[\hat{\mu}_n] &= E\left[\frac{1}{n-20}\sum_{i=21}^{n} X_i\right] \\ &= \frac{(n-20)\mu}{n-20} \\ &= \mu\end{aligned}$$

   This implies that
   $\text{Bias}(\hat{\mu}_n, \mu) = E[\hat{\mu}_n] - \mu = \mu - \mu = 0$

$$\begin{aligned}\text{Var}(\hat{\mu}_n) &= \text{Var}\left[\frac{1}{n-20}\sum_{i=21}^{n} X_i\right] \\ &= \frac{1}{(n-20)^2}\sum_{i=21}^{n}\text{Var}(X_i) \\ &= \frac{1}{(n-20)^2}[(n-20)\sigma^2] \\ &= \frac{\sigma^2}{(n-20)}\end{aligned}$$

$$\begin{aligned}\text{Risk}(\hat{\mu}) &= \text{Bias}(\hat{\mu}, \mu)^2 + \text{Var}(\hat{\mu}) \\ &= 0 + \frac{\sigma^2}{(n-20)} \\ &= \frac{\sigma^2}{(n-20)}\end{aligned}$$

3. Let $X_1, X_2, \ldots, X_n \sim$ iid X, where X is a random variable with density function

$$f_X(x) = \begin{cases} \frac{\theta}{x^{\theta+1}}, & x > 1, \\ 0, & \text{otherwise.} \end{cases}$$

The mean of the random variable X is $\frac{\theta}{\theta-1}$. Find an estimator of θ using method of moments.

(a) $\dfrac{X_1 + X_2 + \ldots + X_n}{X_1 + X_2 + \ldots + X_n - 1}$

(b) $\dfrac{X_1 + X_2 + \ldots + X_n}{1 - X_1 + X_2 + \ldots + X_n}$

(c) $\dfrac{X_1 + X_2 + \ldots + X_n}{X_1 + X_2 + \ldots + X_n - n}$

(d) $\dfrac{X_1 + X_2 + \ldots + X_n}{n - X_1 + X_2 + \ldots + X_n}$

Solution: The mean of the random variable X is $\frac{\theta}{\theta-1}$.
So,

$$\begin{aligned} M_1 &= \frac{\theta}{\theta - 1} \\ \Rightarrow M_1\theta - M_1 &= \theta \\ \Rightarrow \theta &= \frac{M_1}{M_1 - 1} \\ \Rightarrow \theta &= \frac{\frac{X_1+X_2+\ldots+X_n}{n}}{\frac{X_1+X_2+\ldots+X_n}{n} - 1} \\ \Rightarrow \theta &= \frac{X_1 + X_2 + \ldots + X_n}{X_1 + X_2 + \ldots + X_n - n} \end{aligned}$$

Therefore the estimator of θ is $\dfrac{X_1 + X_2 + \ldots + X_n}{X_1 + X_2 + \ldots + X_n - n}$.

4. Let $X_1, X_2, X_3 \sim$ iid Binomial$(4, \theta)$. Given a random sample $(1, 4, 2)$, find the maximum likelihood estimate of θ.

a) $\frac{2}{3}$

b) $\frac{7}{12}$

c) $\frac{1}{3}$

d) $\frac{5}{12}$

Solution: $X_i \sim$ Binomial$(4, \theta)$
$\Rightarrow f_{X_i}(x) = {}^4C_x \theta^x (1-\theta)^{4-x}$

Likelihood function is given by
$L(x_1, x_2, x_3) = \prod_{i=1}^{3} f_{X_i}(x_i)$
$\Rightarrow L(x_1, x_2, x_3) = {}^4C_{x_1}\theta^{x_1}(1-\theta)^{4-x_1} \times {}^4C_{x_2}\theta^{x_2}(1-\theta)^{4-x_2} \times {}^4C_{x_3}\theta^{x_3}(1-\theta)^{4-x_3}$

$$\begin{aligned} L(1,4,2) &= {}^4C_1 {}^4C_4 {}^4C_2 \theta^{(1+4+2)}(1-\theta)^{12-(1+4+2)} \\ &= 24\theta^7(1-\theta)^5 \\ \Rightarrow \log(L(1,4,2)) &= \log(24) + 7\log(\theta) + 5\log(1-\theta) \end{aligned}$$

Therefore, ML estimator for θ is given by
$\hat{\theta} = \arg\max_\theta[\log(24) + 7\log(\theta) + 5\log(1-\theta)]$

$$\begin{aligned} &\text{Let } Y = \log(24) + 7\log(\theta) + 5\log(1-\theta) \\ &\Rightarrow \frac{dY}{d\theta} = \frac{7}{\theta} - \frac{5}{1-\theta} \end{aligned}$$

Now we will equate this value to zero and find the value of θ

$$\begin{aligned} \frac{7}{\theta} - \frac{5}{1-\theta} = 0 \Rightarrow \theta &= \frac{7}{12} \\ \Rightarrow \hat{\theta}_{ML} &= \frac{7}{12} \end{aligned}$$

5. Let $X_1, X_2, \ldots, X_n \sim$ iid X, where X is a random variable with density function

$$f_X(x) = \begin{cases} e^{-(x-\theta)}, & x > \theta, \\ 0, & \text{otherwise.} \end{cases}$$

i) The mean of the distribution is $\theta + 1$. Find the estimator of θ using method of moments. [1 mark]

(a) $\dfrac{X_1 + X_2 + \ldots + X_n}{n}$

(b) $\dfrac{X_1 + X_2 + \ldots + X_n - n}{n}$

(c) $\dfrac{n}{X_1 + X_2 + \ldots + X_n - n}$

(d) $\dfrac{1}{n - X_1 + X_2 + \ldots + X_n}$

Solution: The mean of the random variable X is $\theta + 1$.
So,

$$\begin{aligned} M_1 &= \theta + 1 \\ \Rightarrow \theta &= M_1 - 1 \\ \Rightarrow \theta &= \frac{X_1 + X_2 + \ldots + X_n}{n} - 1 \\ \Rightarrow \theta &= \frac{X_1 + X_2 + \ldots + X_n - n}{n} \end{aligned}$$

Therefore the estimator of θ is $\dfrac{X_1 + X_2 + \ldots + X_n - n}{n}$.

ii) Is the method of moments estimator unbiased?

a) Yes

b) No

Solution:
Estimator of θ is

$$\hat{\theta} = \frac{X_1 + X_2 + \ldots + X_n - n}{n}$$

$$\begin{aligned} E[\hat{\theta}] &= E[\frac{X_1 + X_2 + \ldots + X_n - n}{n}] \\ &= \frac{1}{n}(E[X_1] + E[X_2] + \ldots + E[X_n] - n) \\ &= \frac{1}{n}(n\theta + n - n) \\ &= \theta \end{aligned}$$

And
$\text{Bias}(\hat{\theta}, \theta) = E[\hat{\theta}] - \theta = \theta - \theta = 0$
Since, $\text{Bias}(\hat{\theta}, \theta) = 0$, therefore the estimator is unbiased.

6. Suppose it is known that a sample consisting of the values 10, 12, 15, 16.5, 18, 19, 20 and 21.5 comes from a population with the density function

$$f(x) = \begin{cases} \frac{1}{\theta} e^{\frac{-x}{\theta}}, & x > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Find the maximum likelihood estimate of θ. Enter your answer correct to one decimal.

Solution:

$$\begin{aligned}
L(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} f_X(x_i) \\
&= \prod_{i=1}^{n} \frac{1}{\theta} e^{\frac{-x_i}{\theta}} \\
&= \frac{1}{\theta^n} \left(e^{\frac{-x_1}{\theta}} e^{\frac{-x_2}{\theta}} \ldots e^{\frac{-x_n}{\theta}} \right) \\
&= \frac{1}{\theta^n} \left(e^{\frac{-(x_1+x_2+\ldots+x_n)}{\theta}} \right)
\end{aligned}$$

$\Rightarrow \log(L(x_1, x_2, \ldots, x_n)) = -n\log(\theta) - \dfrac{(x_1 + x_2 + \ldots + x_n)}{\theta}$

Therefore, ML estimator for θ is given by

$\hat{\theta} = \arg\ \max_\theta[-n\log(\theta) - \dfrac{(x_1 + x_2 + \ldots + x_n)}{\theta}]$

$$\begin{aligned}
\text{Let } Y &= -n\log(\theta) - \frac{(x_1 + x_2 + \ldots + x_n)}{\theta} \\
\Rightarrow \frac{dY}{d\theta} &= -\frac{n}{\theta} + \frac{(x_1 + x_2 + \ldots + x_n)}{\theta^2}
\end{aligned}$$

Now we will equate this value to zero and find the value of θ.

$$\begin{aligned}
\Rightarrow -\frac{n}{\theta} + \frac{(x_1 + x_2 + \ldots + x_n)}{\theta^2} &= 0 \\
\Rightarrow \theta &= \frac{x_1 + x_2 + \ldots + x_n}{n} \\
\Rightarrow \hat{\theta} &= \frac{x_1 + x_2 + \ldots + x_n}{n}
\end{aligned}$$

Therefore, maximum likelihood estimate of θ for the given sample will be

$$\begin{aligned}
\hat{\theta} &= \frac{10 + 12 + 15 + 16.5 + 18 + 19 + 20 + 21.5}{8} \\
&= \frac{132}{8} \\
&= 16.5
\end{aligned}$$

7. Let X be a discrete random variable with the following probability mass function

x	1	2	3	4
$f_X(x)$	$\frac{1-p}{2}$	$\frac{p}{2}$	$\frac{1-p}{2}$	$\frac{p}{2}$

Table 8.1.G: PMF of X

Suppose a sample consisting of the values 2, 2, 4, 3, 1, 3, 1 and 2 is taken from the random variable X. Find the estimate of p using method of moments. Enter your answer correct to two decimals accuracy.

Solution:

$$
\begin{aligned}
E[X] &= 1 \times \frac{1-p}{2} + 2 \times \frac{p}{2} + 3 \times \frac{1-p}{2} + 4 \times \frac{p}{2} \\
&= \frac{(1-p) + 2p + 3(1-p) + 4p}{2} \\
&= p + 2
\end{aligned}
$$

Now
$M_1 = E[X] = p + 2$
$\Rightarrow p = M_1 - 2$

Therefore, estimate of p will be
$\dfrac{X_1 + X_2 + \ldots + X_n}{n} - 2.$

So, the estimate of p for the given sample will be

$$
\begin{aligned}
\hat{p} &= \frac{2+2+4+3+1+3+1+2}{8} - 2 \\
&= \frac{18}{8} - 2 \\
&= 0.25
\end{aligned}
$$

Use the following values of CDF of standard normal distribution to answer the questions:

$$F_Z(1.64) = 0.90, F_Z(1.96) = 0.95$$

8. The weights (in grams) of mangoes grown in a certain area are normally distributed with mean μ and standard deviation 40. The weights from a random sample of mangoes are as follows:
220, 210, 240, 260, 235, 225, 270, 300, 200.
Find a 95% confidence interval for the mean weight of mangoes.

   a) [203.87, 256.13]
   b) [213.87, 266.13]
   c) [230, 280]
   d) [215.13, 235.87]

Solution:
$n = 9, \hat{\mu} = 240$ and $\sigma = 40$.
$\beta = 0.95$, using CDF of Normal(0, 1),

$$\frac{\alpha}{\sigma/\sqrt{n}} = 1.96$$
$$\alpha = 1.96 \times \frac{40}{\sqrt{9}} = 26.13$$
$$P(|\hat{\mu} - \mu| < 26.13) = 0.95$$

So, 95% confidence interval is [240 - 26.13, 240 + 26.13] i.e. [213.87, 266.13]

9. From past experience it is known that the weights of seer fish grown at a commercial hatchery are normal with a mean that varies from season to season but with a standard deviation that remains fixed at 0.2 kilogram. If we want to be 90% certain that our estimate of the present season's mean weight of a seer fish is correct to within 0.01 kilograms, how large a sample is needed?
Solution:
Let X denote the weights of seer fish.
Given that $\sigma = 0.2$
To find the value of n such that $P(|\hat{\mu} - \mu| \leq 0.01) = 0.90$

$$P(|\hat{\mu} - \mu| \leq 0.01) = 0.90$$
$$\Rightarrow P\left(|\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}| \leq \frac{0.01}{\sigma/\sqrt{n}}\right) = 0.90$$
$$\Rightarrow P\left(|Z| \leq \frac{0.01}{\sigma/\sqrt{n}}\right) = 0.90$$

$$\frac{0.01}{\sigma/\sqrt{n}} = 1.64$$
$$\Rightarrow \sqrt{n} = 0.2 \times \frac{1.64}{0.01}$$
$$\Rightarrow n = 1075.84$$

Therefore the sample size should be 1076.

10. The distribution of heights of a certain population of women is normally distributed with μ unknown and σ unknown. We observe a random sample (in centimeters):
160, 155, 168, 167, 162, 150, 152, 148, 164.
Find a 95% confidence interval for μ. Use $P(-2.30 < T_8 < 2.30) = 0.95$ where T_8 is t-distribution with degree of freedom 8.

a) [152.73, 164.15]

b) [156.67, 160.2]

c) [160.28, 167.72]

d) [150.34, 165.66]

Solution:
$n = 9, \hat{\mu} = 158.44$ and $S^2 = 55.52 \Rightarrow S = 7.45$
Using $t-$distribution, $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

$$\frac{\alpha}{S/\sqrt{n}} = 2.30$$
$$\alpha = 2.30 \times \frac{7.45}{\sqrt{9}} = 5.71$$
$$P(|\hat{\mu} - \mu| < 5.71) = 0.95$$

So, 95% confidence interval is [158.44 - 5.71, 158.44 + 5.71] i.e. [152.73, 164.15].