Statistics for Data Science-2

Week 8 Solve with us

Table of contents

1. Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples from a distribution X with mean μ and standard deviation σ. Let $\hat{\mu} = 20\left(\dfrac{X_1 + X_2 + \ldots + X_n}{n}\right)$ be an estimator of μ. Find the risk of $\hat{\mu}$.
   (a) $\dfrac{400\sigma^2}{n} + 400\mu^2$
   (b) $\dfrac{20\sigma^2}{n} + 19\mu^2$
   (c) $\dfrac{400\sigma^2}{n} + 361\mu^2$
   (d) $\dfrac{20\sigma^2}{n} + 20\mu$

Solution:

$$
\begin{aligned}
E[\hat{\mu}] &= E\left[20\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)\right] \\
&= \frac{20}{n}(n\mu) \\
&= 20\mu
\end{aligned}
$$

And

$\text{Bias}(\hat{\mu}, \mu) = E[\hat{\mu}] - \mu = 20\mu - \mu = 19\mu$

$$\begin{aligned}
\text{Var}(\hat{\mu}) &= \text{Var}\left[20\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right)\right] \\
&= \frac{400}{n^2}(n\sigma^2) \\
&= \frac{400\sigma^2}{n}
\end{aligned}$$

$$\begin{aligned}
\text{Risk}(\hat{\mu}) &= \text{Bias}(\hat{\mu}, \mu)^2 + \text{Var}(\hat{\mu}) \\
&= (19\mu)^2 + \frac{400\sigma^2}{n} \\
&= 361\mu^2 + \frac{400\sigma^2}{n}
\end{aligned}$$

2. 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 sample from this distribution is taken:
3, 6, 2, 7, 8, 10.
What is the maximum likelihood estimate of θ for the sample.

a) 1.16
b) 1.39
c) 3.85
d) 0.72

Solution:

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

$\Rightarrow \log(L(x_1, x_2, \ldots, x_n)) =$
$n\log(\theta) - (\theta+1)[\log(x_1) + \log(x_2) + \ldots + \log(x_n)]$

Therefore, ML estimator for θ is given by
$\theta^\star = \arg\max_\theta[n\log(\theta) - (\theta+1)[\log(x_1) + \log(x_2) + \ldots + \log(x_n)]]$

$$\text{Let} Y = n\log(\theta) - (\theta+1)(\log(x_1) + \log(x_2) + \ldots + \log(x_n))$$
$$\Rightarrow \frac{dY}{d\theta} = \frac{n}{\theta} - (\log(x_1) + \log(x_2) + \ldots + \log(x_n))$$

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

$$\frac{n}{\theta} - (\log(x_1) + \log(x_2) + \ldots + \log(x_n)) = 0$$
$$\Rightarrow \theta = \frac{n}{\log(x_1) + \log(x_2) + \ldots + \log(x_n)}$$

This implies that $\theta^\star = \frac{n}{\log(x_1)+\log(x_2)+\ldots+\log(x_n)}$

Therefore the ML estimator of θ for the given sample 3, 6, 2, 7, 8, 10 will be

$$\begin{aligned}\theta^{\star} &= \frac{6}{\log(3)+\log(6)+\log(2)+\log(7)+\log(8)+\log(10)} \\ &= \frac{6}{\log(20160)} \\ &= 1.39\end{aligned}$$

3. Let $X_1, X_2, X_3, X_4 \sim$ iid Binomial$(4, p)$. Given a random sample $(2, 0, 4, 3)$, find the maximum likelihood estimate of p.

   a) $\frac{3}{4}$

   b) $\frac{9}{16}$

   c) $\frac{1}{2}$

   d) $\frac{1}{4}$

Solution:

$X_i \sim \text{Binomial}(4, p)$

$\Rightarrow f_{X_i}(x) = {}^4C_x p^x (1-p)^{4-x}$

Likelihood function is given by

$$L(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} f_{X_i}(x_i)$$

$$\Rightarrow L(x_1, x_2, x_3, x_4) = {}^4C_{x_1} p^{x_1} (1-p)^{4-x_1} \times {}^4C_{x_2} p^{x_2} (1-p)^{4-x_2} \times {}^4C_{x_3} p^{x_3} (1-p)^{4-x_3} \times {}^4C_{x_4} p^{x_4} (1-p)^{4-x_4}$$

$$L(2,0,4,3) = {}^4C_2\,{}^4C_0\,{}^4C_4\,{}^4C_3\,p^{2+0+4+3}(1-p)^{16-(2+0+4+3)}$$
$$= 24p^9(1-p)^7$$
$$\Rightarrow \log(L(2,0,4,3)) = \log(24) + 9\log(p) + 7\log(1-p)$$

Therefore, ML estimator for p is given by
$\hat{p} = \arg\max_p[\log(24) + 9\log(p) + 7\log(1-p)]$

$$\text{Let } Y = \log(24) + 9\log(p) + 7\log(1-p)$$
$$\Rightarrow \frac{dY}{dp} = \frac{9}{p} - \frac{7}{1-p}$$

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

$$\frac{9}{p} - \frac{7}{1-p} = 0 \Rightarrow p = \frac{9}{16}$$
$$\Rightarrow \hat{p}_{ML} = \frac{9}{16}$$

4. Suppose that we want to estimate the true average number of eggs a queen bee lays with 95% confidence. The margin of error we are willing to accept is 0.3. Suppose we also know that standard deviation is 9. What sample size should we use?

   a) 3457
   b) 3458
   c) 5144
   d) 5145

Solution:

Let X denote the number of eggs a queen bee lays.
Given that $\sigma = 9$
To find the value of n such that $P(|\hat{\mu} - \mu| \leq 0.3) = 0.95$

$$P(|\hat{\mu} - \mu| \leq 0.3) = 0.95$$
$$\Rightarrow P\left(|\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}| \leq \frac{0.3}{\sigma/\sqrt{n}}\right) = 0.95$$
$$\Rightarrow P\left(|Z| \leq \frac{0.3}{\sigma/\sqrt{n}}\right) = 0.95$$

$$\frac{0.3}{\sigma/\sqrt{n}} = 1.96$$
$$\Rightarrow \sqrt{n} = 9 \times \frac{1.96}{0.3}$$
$$\Rightarrow n = 3457.44$$

Therefore the sample size should be 3458.

5. Consider a sample of iid random variables $X_1, X_2, \ldots, X_n$, where $n > 30, E[X_i] = \mu, \mathrm{Var}(X_i) = \sigma^2$ and the estimator of μ, $\hat{\mu}_n = \dfrac{1}{n-30} \displaystyle\sum_{i=31}^{n} X_i$. Find the bias of $\hat{\mu}_n$.

   a) 30μ
   b) 0
   c) $\dfrac{30\mu}{n-30}$
   d) 31μ

Solution:

$$\begin{aligned} E[\hat{\mu}_n] &= E\left[\frac{1}{n-30}\sum_{i=31}^{n} X_i\right] \\ &= \frac{(n-30)\mu}{n-30} \\ &= \mu \end{aligned}$$

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

6. Suppose it is known that a sample consisting of the values 15, 16, 10, 8, 7, 9, 20 and 19 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.

a) 0.07
b) 13
c) 104
d) 8

Solution:

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

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

Therefore, ML estimator for θ is given by
$\hat{\theta} = \arg\max_{\theta}[-n\log(\theta) - \frac{(x_1+x_2+\ldots+x_n)}{\theta}]$

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

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

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

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

$$\begin{aligned}\hat{\theta} &= \frac{15+16+10+8+7+9+20+19}{8} \\ &= \frac{104}{8} \\ &= 13\end{aligned}$$

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

x	0	1	2	3
$f_X(x)$	$\dfrac{1-p}{2}$	$\dfrac{p}{2}$	$\dfrac{1-p}{2}$	$\dfrac{p}{2}$

Table 8.1: PMF of X

Suppose a sample consisting of the values 0, 2, 1, 3, 0, 2, 1 and 3 is taken from the random variable X. Find the estimate of p using method of moments.

a) 0.40
b) 0.50
c) 0.60
d) 0.75

Solution:

$$
\begin{aligned}
E[X] &= 0 \times \frac{1-p}{2} + 1 \times \frac{p}{2} + 2 \times \frac{1-p}{2} + 3 \times \frac{p}{2} \\
&= \frac{p + 2(1-p) + 3p}{2} \\
&= p + 1
\end{aligned}
$$

Now

$M_1 = E[X] = p + 1$

$\Rightarrow p = M_1 - 1$

Therefore, estimate of p will be
$\dfrac{X_1 + X_2 + \ldots + X_n}{n} - 1.$
So, the estimate of p for the given sample will be

$$\begin{aligned}\hat{p} &= \frac{0+2+1+3+0+2+1+3}{8} - 1 \\ &= \frac{12}{8} - 1 \\ &= 0.5\end{aligned}$$