Statistics for Data Science - 2

Week 8 Practice assignment

1. Let $X_1, \ldots, X_n$ be n i.i.d. samples from a random variable X with mean μ and variance σ^2. Let $\bar{X}^2$ be an estimator of μ^2 where $\bar{X}$(sample mean) is an unbiased estimator of μ. Is the estimator $\bar{X}^2$ unbiased always?

   (a) Yes

   (b) No

   Solution:

   $\bar{X} = \dfrac{X_1 + \ldots + X_n}{n}$

   Given $\bar{X}$ is an unbiased estimator of μ and $\bar{X}^2$ is an estimator of μ^2.

   $\implies E[\bar{X}] = \mu$

   Now,

$$\begin{aligned} E[\bar{X}^2] =& \text{Var}(\bar{X}) + (E[\bar{X}])^2 \\ =& \frac{\sigma^2}{n} + \mu^2 \\ \neq& \mu^2 \end{aligned}$$

   Therefore, estimator $\bar{X}^2$ is not an unbiased estimator of μ^2.

2. Let $X_1, X_2, \ldots, X_n$ be n i.i.d. samples from a distribution with PDF

$$f_X(x) = \frac{1 + \theta x}{2}, \quad -1 < x < 1$$

   Let $\hat{\theta} = 3\bar{X}$ be an estimator of θ. Find the mean squared error of $\hat{\theta}$.

   (a) $\dfrac{(3 - \theta^2)}{n}$

   (b) $\dfrac{(3 + \theta^2)}{n}$

   (c) $\dfrac{(3 + \theta)}{n}$

   (d) $\dfrac{(3 - \theta)}{n}$

Solution:

Given $\hat{\theta} = 3\bar{X}$ an estimator of θ.
Expectation of X is given by

$$\begin{aligned}
E[X] &= \int_{-1}^{1} x f_X(x) dx \\
&= \int_{-1}^{1} x \left(\frac{1 + \theta x}{2} \right) dx \\
&= \frac{1}{2} \int_{-1}^{1} (x + \theta x^2) dx \\
&= \frac{x^2}{4} + \frac{\theta x^3}{6} \Bigg|_{-1}^{1} = \frac{\theta}{3}
\end{aligned}$$

$$\begin{aligned}
Bias(\hat{\theta}, \theta) &= E[\hat{\theta} - \theta] \\
&= E\left[3 \left(\frac{X_1 + \ldots + X_n}{n} \right) - \theta \right] \\
&= 3 \left(\frac{n\theta}{3n} \right) - E[\theta] = 0
\end{aligned}$$

Therefore, estimator $\hat{\theta}$ is unbiased.

$$\begin{aligned}
E[X^2] &= \int_{-1}^{1} x^2 f_X(x) dx \\
&= \int_{-1}^{1} x^2 \left(\frac{1 + \theta x}{2} \right) dx \\
&= \frac{1}{2} \int_{-1}^{1} (x^2 + \theta x^3) dx \\
&= \frac{x^3}{6} + \frac{\theta x^4}{8} \Bigg|_{-1}^{1} = \frac{1}{3}
\end{aligned}$$

Therefore, $\text{Var}[X] = \frac{1}{3} - \frac{\theta^2}{9}$

$$\begin{aligned}
\text{Var}(\hat{\theta}) =& \text{Var}\left[3\left(\frac{X_1 + \ldots + X_n}{n}\right)\right] \\
=& \frac{9}{n^2}(n\text{Var}[X]) \\
=& \frac{9}{n^2}\left[n\left(\frac{1}{3} - \frac{\theta^2}{9}\right)\right] \\
=& \frac{3-\theta^2}{n}
\end{aligned}$$

$\text{MSE}(\hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \text{Var}[\hat{\theta}] = \frac{3-\theta^2}{n}$.

3. Consider 100 samples $X_1, X_2, \ldots, X_{100}$ from a random variable X whose distribution has mean μ and variance σ^2. Let $\sum_{i=1}^{100} X_i = 150$ and $\sum_{i=1}^{100} X_i^2 = 1999$. Find an unbiased estimate for $\text{Var}(X)$.

   (a) 17.74
   (b) 17.91
   (c) 1.5
   (d) 2.25

   Solution:

   Given the distribution of X has mean equal to μ and variance equal to σ^2.

   Also, $\sum_{i=1}^{100} X_i = 150$ and $\sum_{i=1}^{100} X_i^2 = 1999$

   We know that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an unbiased estimator of $\text{Var}[X]$.

Therefore,

$$
\begin{aligned}
S^2 =& \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\
=& \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 + \bar{X}^2 - 2X_i\bar{X}) \\
=& \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 + n\bar{X}^2 - 2\bar{X}\sum_{i=1}^{n} X_i\right) \\
=& \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 + n\bar{X}^2 - 2n\bar{X}^2\right) \\
=& \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \\
=& \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\left(\frac{\sum\limits_{i=1}^{n} X_i}{n}\right)^2\right) = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum\limits_{i=1}^{n} X_i\right)^2}{n}\right)
\end{aligned}
$$

Therefore, $S^2 = \dfrac{1}{100-1}\left(1999 - \dfrac{150^2}{100}\right) = 17.91$

4. Let $X_1, X_2, \ldots, X_n \sim$ i.i.d. X. Let $a_1, \ldots, a_n \geq 0$ such that $\sum\limits_{i=1}^{n} a_i = 1$. Define the estimator for mean as $\bar{X} = \sum\limits_{i=1}^{n} a_i x_i$. Define the estimator for the variance as $S^2 = \sum\limits_{i=1}^{n} a_i (X_i - \bar{X})^2$ with $E[X] = \mu$ and $\text{Var}(X) = \sigma^2$. Choose the correct option(s) from the following:

   (a) $\bar{X}$ is an unbiased estimator.

   (b) $E[S^2] = \left(\dfrac{n-1}{n}\right)\sigma^2$

   (c) $E[S^2] = \left(1 - \sum\limits_{i=1}^{n} a_i^2\right)\sigma^2$

   (d) $E[S^2] = \sum\limits_{i=1}^{n} a_i^2 \sigma^2$

   (e) S^2 is an unbiased estimator for $\text{Var}(X)$.

Solution:

Given $X_1, X_2, \ldots, X_n \sim$ i.i.d. X, $E[X] = \mu, \text{Var}[X] = \sigma^2$
$\bar{X} = \sum_{i=1}^{n} a_i x_i$ is an estimator of μ, where $\sum_{i=1}^{n} a_i = 1$.

(a) $E[\bar{X}] = E[a_1X_1 + \cdots + a_nX_n] = \sum_{i=1}^{n} a_i E[X] = \mu$ (since $\sum_{i=1}^{n} a_i = 1$)
Bias$(\bar{X}) = E[\bar{X}] - E[X] = \mu - \mu = 0$
Therefore, $\bar{X}$ is an unbiased estimator of μ.

(b) $\text{Var}[\bar{X}] = \text{Var}[a_1X_1 + \cdots + a_nX_n] = \sum_{i=1}^{n} a_i^2 \text{Var}[X] = \sigma^2 \sum_{i=1}^{n} a_i^2$

$$E[\bar{X}] = \mu \tag{1}$$

$$\text{Var}[\bar{X}] = \sigma^2 \sum_{i=1}^{n} a_i^2 \tag{2}$$

$$\begin{aligned}
S^2 &= \sum_{i=1}^{n} a_i (X_i - \bar{X})^2 \\
&= \sum_{i=1}^{n} (a_i X_i^2 + a_i \bar{X}^2 - 2a_i X_i \bar{X}) \\
&= \sum_{i=1}^{n} a_i X_i^2 + \sum_{i=1}^{n} a_i \bar{X}^2 - \sum_{i=1}^{n} 2a_i \bar{X} X_i \\
&= \sum_{i=1}^{n} a_i X_i^2 + \bar{X}^2 - 2\bar{X}^2 = \sum_{i=1}^{n} a_i X_i^2 - \bar{X}^2
\end{aligned}$$

Now,

$$
\begin{aligned}
E[S^2] = E\left(\sum_{i=1}^{n} a_i X_i^2 - \bar{X}^2\right) &= \sum_{i=1}^{n} E[a_i X_i^2] - E[\bar{X}^2] \\
&= \sum_{i=1}^{n} a_i E[X_i^2] - E[\bar{X}^2] \\
&= \sum_{i=1}^{n} a_i(\sigma^2 + \mu^2) - (\mathrm{Var}[\bar{X}] + \mu^2) \\
&= \sigma^2 + \mu^2 - \sigma^2 \sum_{i=1}^{n} a_i^2 - \mu^2 \quad \text{[From(2)]} \\
&= \sigma^2 - \sigma^2 \sum_{i=1}^{n} a_i^2 \\
&= \left(1 - \sum_{i=1}^{n} a_i^2\right) \sigma^2
\end{aligned}
$$

Therefore, (b) is not true.

(c) Since $E[S^2] = \left(1 - \sum_{i=1}^{n} a_i^2\right) \sigma^2$, therefore, (c) is true.

(d) (d) is not the correct option.

(e) $\mathrm{Bias}(S^2) = E[S^2] - \sigma^2 \neq \sigma^2$.
Therefore, S^2 is not an unbiased estimator of $\mathrm{Var}[X]$.

5. Let $X_1, \ldots, X_n \sim$ i.i.d. Uniform$(-a, a)$. Find the ML estimator of a.

(a) $\hat{a}_{ML} = \max(| X_1 |, \ldots, | X_n |)$

(b) $\hat{a}_{ML} = \max(X_1, \ldots, X_n)$

(c) $\hat{a}_{ML} = \min(X_1, \ldots, X_n)$

(d) $\hat{a}_{ML} = \dfrac{1}{2^n} \min(X_1, \ldots, X_n)$

Solution:

$X_1, \cdots, X_n \sim$ Uniform$(-a, a)$.
$f_{X_i}(x_i)$ is given by

$$
f_{X_i}(x_i) = \begin{cases} \dfrac{1}{2a} & \text{for } -a < x_i < a \\ 0 & \text{otherwise} \end{cases}
$$

Likelihood function of a is given by

$$
L(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i) = \left(\frac{1}{2a}\right)^n
$$

In order to maximise the likelihood function, we need to minimize a.
Since $-a < x_i < a$ for all i and $| x_i |< a$, therefore, $a = \max(| x_1 |, \dots, | x_n |)$.

Therefore, the ML estimator of a is $\max(| X_1 |, \dots, | X_n |)$.

6. Let $X_1, X_2, X_3 \sim$ iid Normal(μ, σ^2). Given a random sample $(-1, 0, 1)$, find the maximum likelihood estimate of σ^2.

   a) $\frac{2}{3}$
   b) $\frac{7}{12}$
   c) $\frac{1}{3}$
   d) $\frac{5}{12}$

   Solution:

   ML estimator of σ^2 is $\dfrac{\sum\limits_{i=1}^{n}(X_i - \hat{\mu}_{ML})^2}{n}$, where $\hat{\mu}_{ML} = \bar{X}$.

   Given the samplings $-1, 0, 1$, $\bar{X} = \dfrac{-1+0+1}{3} = 0$

   Therefore, ML estimator of σ^2 is $\dfrac{(-1)^2 + 0^2 + 1^2}{3} = \dfrac{2}{3}$.

7. Let $X_1, \dots, X_n$ be n i.i.d. samples of a random variable X. Let X have the PDF $f(x) = (\alpha + 1)x^\alpha$, where $0 < x < 1$.

   (a) Find the ML estimator of α.

   i. $\hat{\alpha}_{ML} = 1 + \dfrac{n}{\sum\limits_{i=1}^{n} \log X_i}$

   ii. $\hat{\alpha}_{ML} = -1 - \dfrac{n}{\sum\limits_{i=1}^{n} \log X_i}$

   iii. $\hat{\alpha}_{ML} = 1 - \dfrac{n}{\sum\limits_{i=1}^{n} \log X_i}$

   iv. $\hat{\alpha}_{ML} = -1 + \dfrac{n}{\sum\limits_{i=1}^{n} \log X_i}$

   Solution:

   Given,

   $$f(x) = (\alpha + 1)x^\alpha, \quad 0 < x < 1$$

Likelihood function of a sampling $X_1, X_2, \ldots, X_n$ will be given by

$$\begin{aligned} L(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} f_X(x_i) \\ &= (\alpha+1)^n x_1^{\alpha} \cdots x_n^{\alpha} \\ \Rightarrow \log(L) &= n\log(\alpha+1) + \alpha(\log(x_1) + \cdots + \log(x_n)) \end{aligned}$$

Therefore, ML estimator for α is given by

$$\hat{\alpha} = \arg\max_{\alpha}[n\log(\alpha+1) + \alpha(\log(x_1) + \cdots + \log(x_n))]$$

Let $Y = n\log(\alpha+1) + \alpha(\log(x_1) + \cdots + \log(x_n))$
Now,

$$\begin{aligned} \frac{dY}{d\alpha} &= \frac{d}{d\alpha}[n\log(\alpha+1) + \alpha(\log(x_1) + \cdots + \log(x_n))] \\ &= \frac{n}{\alpha+1} + \log(x_1) + \cdots + \log(x_n) \end{aligned}$$

Now,

$$\begin{aligned} \frac{dY}{d\alpha} &= 0 \\ \Rightarrow \frac{n}{\alpha+1} &= -[\log(x_1) + \cdots + \log(x_n)] \\ \Rightarrow \hat{\alpha}_{ML} &= -1 - \frac{n}{\sum_{i=1}^{n} \log X_i} \end{aligned}$$

(b) The mean of the random variable X is $\frac{\alpha+1}{\alpha+2}$. Find the estimator of α using method of moments.

i. $\hat{\alpha}_{MME} = \dfrac{1+2M_1}{M_1-1}$

ii. $\hat{\alpha}_{MME} = \dfrac{1-M_1}{M_1-1}$

iii. $\hat{\alpha}_{MME} = \dfrac{1+M_1}{M_1-1}$

iv. $\hat{\alpha}_{MME} = \dfrac{1-2M_1}{M_1-1}$

Solution:

The expected value of X, $E(X)$ is given as $\frac{\alpha+1}{\alpha+2}$.
Using method of moments,

$$\frac{\alpha+1}{\alpha+2} = m_1$$

$$\alpha = \frac{1-2m_1}{m_1-1}$$

The estimator is

$$\hat{\alpha}_{MME} = \frac{1-2M_1}{M_1-1}$$

8. Let X be a discrete random variable taking the values $-1, 0, 1$ with probabilities $P(X = -1) = \dfrac{p}{2}, P(X = 0) = \dfrac{p}{2}, P(X = 1) = 1-p$. Let $X_1, \ldots, X_n \sim$ i.i.d.$\{-1, 0, 1\}$. Find the estimator of p using the method of moments.

   (a) $\dfrac{2-2M_1}{3}$

   (b) $\dfrac{2+2M_1}{3}$

   (c) $\dfrac{1+2M_1}{3}$

   (d) $\dfrac{2+M_1}{3}$

   Solution:

   The expected value of X, $E(X)$ is given by

$$E[X] = \sum_x x p_X(x) = \left(-1 \times \frac{p}{2}\right) + \left(0 \times \frac{p}{2}\right) + (1 \times (1-p)) = \frac{(2-3p)}{2}$$

$$E[X] = \frac{(2-3p)}{2}$$

   Using method of moments,

$$\frac{(2-3p)}{2} = m_1$$

   The estimator is

$$\hat{p} = \frac{2-2m_1}{3}$$

$$\hat{p} = \frac{2-2M_1}{3}$$

9. Let X be a random variable with PDF

$$f_X(x) = (\lambda a)x^{\alpha-1}e^{-\lambda x^{\alpha}}, \quad x > 0.$$

   where α and a are constants. Find the maximum likelihood estimator of λ for n i.i.d. samples of X.

(a) $\dfrac{\sum\limits_{i=1}^{n} X_i^{\alpha}}{n}$

(b) $\dfrac{n}{\sum\limits_{i=1}^{n} X_i^{\alpha}}$

(c) $\dfrac{n}{\alpha \sum\limits_{i=1}^{n} X_i^{\alpha}}$

(d) $\dfrac{\sum\limits_{i=1}^{n} X_i^{\alpha}}{n\alpha}$

Solution:

Given,

$$f_X(x) = (\lambda a)x^{\alpha-1}e^{-\lambda x^{\alpha}}, \quad x > 0$$

Likelihood function of a sampling $X_1, X_2, \ldots, X_n$ will be given by

$$\begin{aligned} L(x_1, x_2, \ldots, x_n) &= \prod_{i=1}^{n} f_X(x_i) \\ &= (\lambda a)^n (x_1 \cdots x_n)^{\alpha-1} e^{-\lambda(x_1^{\alpha}+\cdots+x_n^{\alpha})} \end{aligned}$$

Likelihood is a function of the parameter so, we can ignore the constant terms in the likelihood function. Therefore,

$$\begin{aligned} L &= \lambda^n e^{-\lambda(x_1^{\alpha}+\cdots+x_n^{\alpha})} \\ \Rightarrow \log(L) &= n\log(\lambda) - \lambda(x_1^{\alpha}+\cdots+x_n^{\alpha}) \end{aligned}$$

Therefore, ML estimator for λ is given by

$$\hat{\lambda} = \arg\max_{\lambda}[n\log(\lambda) - \lambda(x_1^{\alpha}+\cdots+x_n^{\alpha})]$$

Let $Y = n\log(\lambda) - \lambda(x_1^{\alpha}+\cdots+x_n^{\alpha})$

Now,

$$\begin{aligned} \frac{dY}{d\lambda} &= \frac{d}{d\lambda}[n\log(\lambda) - \lambda(x_1^{\alpha}+\cdots+x_n^{\alpha})] \\ &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i^{\alpha} \end{aligned}$$

Now,

$$
\begin{aligned}
\frac{dY}{d\lambda} &= 0 \\
\Rightarrow \frac{n}{\lambda} &= \sum_{i=1}^{n} x_i^{\alpha} \\
\Rightarrow \lambda &= \frac{n}{\sum_{i=1}^{n} X_i^{\alpha}}
\end{aligned}
$$

10. A random sample of 1000 television screens taken from the household of a city shows that the average running time of television is 7 hours per day with a standard deviation of 2 hours. Assume the distribution of measurements to be approximately normal. Calculate a 99% confidence interval for the daily average television running hours.
Hint: Use $P(-2.58 < Z < 2.58) = 0.99$.

    (a) $[6.02, 6.98]$
    (b) $[7.02, 8.19]$
    (c) $[6.12, 7.98]$
    (d) $[6.83, 7.17]$

    Solution:

    Given $\beta = 0.99$, $n = 1000$, $\bar{X} = 7$ and $\sigma = 2$.
    To find: $P(\mid \bar{X} - \mu \mid \leq \alpha) = 0.99$

$$
\begin{aligned}
P\left(\mid \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \mid \leq \frac{\alpha}{\sigma/\sqrt{n}}\right) &= 0.99 \\
\implies P\left(\mid Z \mid \leq \frac{\alpha}{\sigma/\sqrt{n}}\right) &= 0.99 \quad \text{where } Z \sim \text{Normal}(0,1) \\
\implies P\left(-\frac{\alpha}{\sigma/\sqrt{n}} \leq Z \leq \frac{\alpha}{\sigma/\sqrt{n}}\right) &= 0.99
\end{aligned}
$$

    It is given that $(-2.58 < Z < 2.58) = 0.99$, therefore,
    $\dfrac{\alpha}{\sigma/\sqrt{n}} = 2.58 \implies \alpha = 2.58 \times \dfrac{\sigma}{\sqrt{n}} = 2.58 \times \dfrac{2}{\sqrt{1000}} = 0.163$

    The confidence interval for μ is $[\bar{X} - \alpha, \bar{X} + \alpha]$.
    Therefore, 99% confidence interval for μ is $[6.83, 7.17]$.

11. The distribution of the diameter of screws produced by a certain machine is normally distributed with μ and σ unknown. We observe a random sample $9.8, 10.2, 10.4, 9.8, 10.0, 10.2$ and 9.6 (in cm).
Find a 95% confidence interval for the mean diameter of screws.
Hint: Use $P(-2.447 < t_6 < 2.447) = 0.95$ and S(sample standard deviation) $= 0.283$.

    (a) $[10.74, 11.26]$
    (b) $[9.74, 10.26]$
    (c) $[7.47, 8.26]$
    (d) $[7.98, 8.75]$

    Solution:

    Given that $S = 0.283$, $n = 7$, $\beta = 0.95$

    Now, $\bar{X} = \dfrac{9.8 + 10.2 + 10.4 + 9.8 + 10.0 + 10.2 + 9.6}{7}$=10
    Using t-distribution, $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$.

$$\begin{aligned}\frac{\alpha}{S/\sqrt{n}} &= 2.447\\ \alpha &= 2.447 \times \frac{0.283}{\sqrt{7}}\\ &= 0.26\end{aligned}$$

    $P(|\hat{\mu} - \mu| < 0.26) = 0.95$
    So, 95% confidence interval is $[10 - 0.26, 10 + 0.26] = [9.74, 10.26]$.

12. A data scientist wishes to determine the average time it takes to run one epoch of a machine learning model in her machine. How large a sample will she need to be 95% confident that her sample mean will be within 15 seconds of the true mean? Assume that it is known from previous studies that $\sigma = 40$ seconds.
Hint: Use $P(-1.96 < Z < 1.96) = 0.95$.
Answer: 28

    Let X denote the time taken to run epoch of a machine learning model.
    Given that $\sigma = 40$
    To find the value of n such that $P(|\hat{\mu} - \mu| \leq 15) = 0.95$

$$\begin{aligned} P(|\hat{\mu} - \mu| \leq 15) &= 0.95\\ \Rightarrow P\left(|\frac{\hat{\mu} - \mu}{\sigma/\sqrt{n}}| \leq \frac{15}{\sigma/\sqrt{n}}\right) &= 0.95\\ \Rightarrow P\left(|Z| \leq \frac{15}{\sigma/\sqrt{n}}\right) &= 0.95\end{aligned}$$

Now,

$$\frac{15}{\sigma/\sqrt{n}} = 1.96$$
$$\Rightarrow \sqrt{n} = 40 \times \frac{1.96}{15}$$
$$\Rightarrow n = 27.31$$

Therefore, the sample size should be 28.