SOLUTIONS TO
MATH10282
INTRO TO STATISTICS

Solutions to Question 1

ILOs addressed: present numerical summaries of a data set.

(a) Let $x_1, x_2, \ldots, x_n$ denote a data set and let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ denote the order statistics in ascending order.

(i) sample median $= Q(1/2) = x_{([\frac{n+1}{2}])} + \left\{\frac{n+1}{2} - \left[\frac{n+1}{2}\right]\right\}\left\{x_{([\frac{n+1}{2}]+1)} - x_{([\frac{n+1}{2}])}\right\}$. (1 marks)
Level of difficulty = low

(ii) sample first quartile $= Q(1/4) = x_{([\frac{n+1}{4}])} + \left\{\frac{n+1}{4} - \left[\frac{n+1}{4}\right]\right\}\left\{x_{([\frac{n+1}{4}]+1)} - x_{([\frac{n+1}{4}])}\right\}$. (1 marks)
Level of difficulty = low

(iii) sample third quartile $= Q(3/4) = x_{([\frac{3(n+1)}{4}])} + \left\{\frac{3(n+1)}{4} - \left[\frac{3(n+1)}{4}\right]\right\}\left\{x_{([\frac{3(n+1)}{4}]+1)} - x_{([\frac{3(n+1)}{4}])}\right\}$. (1 marks)
Level of difficulty = low

(iv) sample inter quartile range $= Q(3/4) - Q(1/2)$. (1 marks)
Level of difficulty = low

SEEN

First, we calculate $Q(1/4)$. Note that $r = p(n+1)$ and $r' = [p(n+1)]$ are

$$r = \begin{cases} m + \frac{1}{4}, & \text{if } n = 4m, \\ m, & \text{if } n = 4m-1, \\ m - \frac{1}{4}, & \text{if } n = 4m-2, \\ m - \frac{1}{2}, & \text{if } n = 4m-3 \end{cases}$$

and

$$r' = \begin{cases} m, & \text{if } n = 4m, \\ m, & \text{if } n = 4m-1, \\ m-1, & \text{if } n = 4m-2, \\ m-1, & \text{if } n = 4m-3, \end{cases}$$

respectively. So,

$$r - r' = \begin{cases} \frac{1}{4}, & \text{if } n = 4m, \\ 0, & \text{if } n = 4m-1, \\ \frac{3}{4}, & \text{if } n = 4m-2, \\ \frac{1}{2}, & \text{if } n = 4m-3. \end{cases}$$

Hence,

$$\text{first quartile} = \begin{cases} x_{(m)} + \frac{1}{4}\left[x_{(m+1)} - x_{(m)}\right], & \text{if } n = 4m, \\ x_{(m)}, & \text{if } n = 4m-1, \\ x_{(m-1)} + \frac{3}{4}\left[x_{(m)} - x_{(m-1)}\right], & \text{if } n = 4m-2, \\ x_{(m-1)} + \frac{1}{2}\left[x_{(m)} - x_{(m-1)}\right], & \text{if } n = 4m-3. \end{cases} \tag{1}$$

Next we calculate $Q(3/4)$. Note that $r = p(n+1)$ and $r^{'} = [p(n+1)]$ are

$$r = \begin{cases} 3m + \frac{3}{4}, & \text{if } n = 4m, \\ 3m, & \text{if } n = 4m-1, \\ 3m - \frac{3}{4}, & \text{if } n = 4m-2, \\ 3m - \frac{6}{4}, & \text{if } n = 4m-3 \end{cases}$$

and

$$r^{'} = \begin{cases} 3m, & \text{if } n = 4m, \\ 3m, & \text{if } n = 4m-1, \\ 3m-1, & \text{if } n = 4m-2, \\ 3m-2, & \text{if } n = 4m-3, \end{cases}$$

respectively. So,

$$r - r^{'} = \begin{cases} \frac{3}{4}, & \text{if } n = 4m, \\ 0, & \text{if } n = 4m-1, \\ \frac{1}{4}, & \text{if } n = 4m-2, \\ \frac{1}{2}, & \text{if } n = 4m-3. \end{cases}$$

Hence,

$$\text{thirdquartile} = \begin{cases} x_{(3m)} + \frac{3}{4}\left[x_{(3m+1)} - x_{(3m)}\right], & \text{if } n = 4m, \\ x_{(3m)}, & \text{if } n = 4m-1, \\ x_{(3m-1)} + \frac{1}{4}\left[x_{(3m)} - x_{(3m-1)}\right], & \text{if } n = 4m-2, \\ x_{(3m-2)} + \frac{1}{2}\left[x_{(3m-1)} - x_{(3m-2)}\right], & \text{if } n = 4m-3. \end{cases} \tag{2}$$

The result follows from (1) and (2). (6 marks)

Level of difficulty = medium

UNSEEN

Solutions to Question 2

ILOs addressed: define elementary statistical concepts and terminology such as unbiasedness; analyse and compare statistical properties of simple estimators.

(a) Suppose $\widehat{\theta}$ is an estimator of θ based on a random sample of size n. Define what is meant by the following:

(i) the bias of $\widehat{\theta}$ is $E\left(\widehat{\theta}\right) - \theta$; (1 marks)

Level of difficulty = low

(ii) the mean squared error of $\widehat{\theta}$ is $E\left[\left(\widehat{\theta} - \theta\right)^2\right]$; (1 marks)

Level of difficulty = low

(iii) $\widehat{\theta}$ is a consistent estimator of θ if $\lim_{n \to \infty} E\left[\left(\widehat{\theta} - \theta\right)^2\right] = 0$. (1 marks)

Level of difficulty = low

UP TO THIS BOOK WORK.

(b) Suppose $X_1, \ldots, X_n$ are independent Uniform$(-\theta, \theta)$ random variables. Let $\widehat{\theta} = \max\left(|X_1|, \ldots, |X_n|\right)$ denote a possible estimator of θ.

(i) Let $Z = \max\left(|X_1|, \ldots, |X_n|\right)$. The cdf of Z is

$$
\begin{aligned}
F_Z(z) &= \Pr\left[\max\left(|X_1|, \ldots, |X_n|\right) \leq z\right] \\
&= \Pr\left[|X_1| \leq z, \ldots, |X_n| \leq z\right] \\
&= \Pr\left[|X_1| \leq z\right] \cdots \Pr\left[|X_n| \leq z\right] \\
&= \left\{\Pr\left[|X| \leq z\right]\right\}^n \\
&= \left\{\Pr\left[-z \leq X \leq z\right]\right\}^n \\
&= \left\{F_X(z) - F_X(-z)\right\}^n \\
&= \left\{\frac{z + \theta}{2\theta} - \frac{-z + \theta}{2\theta}\right\}^n \\
&= \frac{z^n}{\theta^n}
\end{aligned}
$$

for $0 < z < \theta$. The corresponding pdf is

$$
f_Z(z) = \frac{n z^{n-1}}{\theta^n}
$$

for $0 < z < \theta$. Hence, the bias is

$$\begin{aligned}
\text{Bias}\left(\widehat{\theta}\right) &= E(Z) - \theta \\
&= \frac{n}{\theta^n} \int_0^\theta z^n dz - \theta \\
&= \frac{n}{\theta^n} \left[\frac{z^n}{n+1} \right]_0^\theta dz - \theta \\
&= \frac{n\theta}{n+1} - \theta \\
&= -\frac{\theta}{n+1}.
\end{aligned}$$

(3 marks)

Level of difficulty = medium

UNSEEN

(ii) The MSE is

$$\begin{aligned}
\text{MSE}\left(\widehat{\theta}\right) &= \text{Var}\left(\widehat{\theta}\right) + \left[\text{Bias}\left(\widehat{\theta}\right)\right]^2 \\
&= E\left(Z^2\right) - \frac{n^2\theta^2}{(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\
&= \frac{n}{\theta^n} \int_0^\theta z^{n+1} dz - \frac{n^2\theta^2}{(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\
&= \frac{n}{\theta^n} \left[\frac{z^{n+2}}{n+2} \right]_0^\theta - \frac{n^2\theta^2}{(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\
&= \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2} + \frac{\theta^2}{(n+1)^2} \\
&= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2}.
\end{aligned}$$

(2 marks)

Level of difficulty = medium

UNSEEN

(iii) $\widehat{\theta}$ is a biased since its bias is not zero. (1 marks)

Level of difficulty = low

UNSEEN

(iv) $\widehat{\theta}$ is a consistent since its MSE approaches 0 as $n \to \infty$. (1 marks)

Level of difficulty = low

UNSEEN

Solutions to Question 3

ILOs addressed: define elementary statistical concepts and terminology such as confidence intervals and hypothesis tests.

(a) Suppose we wish to test $H_0 : \mu = \mu_0$ versus $H_0 : \mu \neq \mu_0$.

(i) the Type I error occurs if H_0 is rejected when in fact $\mu = \mu_0$; (1 marks)

Level of difficulty = low

SEEN

(ii) the Type II error occurs if H_0 is accepted when in fact $\mu \neq \mu_0$; (1 marks)

Level of difficulty = low

SEEN

(iii) the significance level is the probability of type I error. (1 marks)

Level of difficulty = low

SEEN

(b) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from $N\left(\mu, \sigma^2\right)$, where σ is unknown. The rejection region for the following tests are

(i) reject $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ if $\sqrt{n}\left|\overline{X} - \mu_0\right|/S > t_{n-1,1-\frac{\alpha}{2}}$; (1 marks)

Level of difficulty = low

SEEN

(ii) reject $H_0 : \mu = \mu_0$ versus $H_1 : \mu < \mu_0$ if $\sqrt{n}\left(\overline{X} - \mu_0\right)/S < t_{n-1,\alpha}$. (1 marks)

Level of difficulty = low

SEEN

(c) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from $N\left(\mu, \sigma^2\right)$, where σ is unknown. Then,

(i) the required probability is

$$
\begin{aligned}
& \Pr\left(\text{Reject } H_0 \mid H_1 \text{ is true}\right) \\
= \; & \Pr\left(\frac{\sqrt{n}\left|\overline{X}-\mu_0\right|}{S} > t_{n-1,1-\frac{\alpha}{2}} \mid \mu \neq \mu_0\right) \\
= \; & \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{S} > t_{n-1,1-\frac{\alpha}{2}} \text{ or } \frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{S} < -t_{n-1,1-\frac{\alpha}{2}} \mid \mu \neq \mu_0\right) \\
= \; & \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{S} > t_{n-1,1-\frac{\alpha}{2}} \mid \mu \neq \mu_0\right) + \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{S} < -t_{n-1,1-\frac{\alpha}{2}} \mid \mu \neq \mu_0\right) \\
= \; & \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu+\mu-\mu_0\right)}{S} > t_{n-1,1-\frac{\alpha}{2}} \mid \mu \neq \mu_0\right) + \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu+\mu-\mu_0\right)}{S} < -t_{n-1,1-\frac{\alpha}{2}} \mid \mu\right. \\
= \; & \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{S} > t_{n-1,1-\frac{\alpha}{2}} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S} \mid \mu \neq \mu_0\right) + \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{S} < -t_{n-1,1-\frac{\alpha}{2}} - \sqrt{n}\right. \\
= \; & \Pr\left(T_{n-1} > t_{n-1,1-\frac{\alpha}{2}} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S}\right) + \Pr\left(T_{n-1} < -t_{n-1,1-\frac{\alpha}{2}} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S}\right) \\
= \; & 1 - \Pr\left(T_{n-1} \leq t_{n-1,1-\frac{\alpha}{2}} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S}\right) + \Pr\left(T_{n-1} < -t_{n-1,1-\frac{\alpha}{2}} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S}\right) \\
= \; & 1 - F_{T_{n-1}}\left(t_{n-1,1-\frac{\alpha}{2}} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S}\right) + F_{T_{n-1}}\left(-t_{n-1,1-\frac{\alpha}{2}} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S}\right).
\end{aligned}
$$

(3 marks)

Level of difficulty = medium

UNSEEN

(ii) the required probability is

$$
\begin{aligned}
& \Pr\left(\text{Reject } H_0 \mid H_1 \text{ is true}\right) \\
= \; & \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu_0\right)}{S} < t_{n-1,\alpha} \mid \mu < \mu_0\right) \\
= \; & \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu+\mu-\mu_0\right)}{S} < t_{n-1,\alpha} \mid \mu < \mu_0\right) \\
= \; & \Pr\left(\frac{\sqrt{n}\left(\overline{X}-\mu\right)}{S} < t_{n-1,\alpha} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S} \mid \mu < \mu_0\right) \\
= \; & \Pr\left(T_{n-1} < t_{n-1,\alpha} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S}\right) \\
= \; & F_{T_{n-1}}\left(t_{n-1,\alpha} - \frac{\sqrt{n}\left(\mu-\mu_0\right)}{S}\right).
\end{aligned}
$$

(2 marks)

Level of difficulty = medium

UNSEEN

Solutions to Question 4

ILOs addressed: define elementary statistical concepts and terminology such as confidence intervals and hypothesis tests; conduct statistical inferences, including confidence intervals and hypothesis tests, in simple one and two-sample situations; sampling distributions.

(a) Let $\mathbf{X} = (X_1, \ldots, X_n)$, with $X_1, \ldots, X_n$ an independent random sample from a distribution F_X with unknown parameter θ. Let $I(\mathbf{X}) = [a(\mathbf{X}), b(\mathbf{X})]$ denote an interval estimator for θ.

(i) $I(\mathbf{X})$ is a $100(1-\alpha)\%$ confidence interval if

$$\Pr\left(a(\mathbf{X}) < \theta < b(\mathbf{X})\right) = 1 - \alpha;$$

(1 marks)

Level of difficulty = low

SEEN

(ii) the coverage probability of $I(\mathbf{X})$ is

$$\Pr\left(a(\mathbf{X}) < \theta < b(\mathbf{X})\right);$$

(1 marks)

Level of difficulty = low

SEEN

(iii) the coverage length of $I(\mathbf{X})$ is $b(\mathbf{X}) - a(\mathbf{X})$. (1 marks)

Level of difficulty = low

SEEN

(b) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$. Then

$$
\begin{aligned}
& \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \\
\Leftrightarrow\quad & \Pr\left(\chi^2_{n-1,\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha \\
\Leftrightarrow\quad & \Pr\left(\frac{1}{\chi^2_{n-1,1-\alpha/2}} < \frac{\sigma^2}{(n-1)s^2} < \frac{1}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha \\
\Leftrightarrow\quad & \Pr\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha \\
\Leftrightarrow\quad & \Pr\left(\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}}\right) = 1-\alpha.
\end{aligned}
$$

Hence, a $100(1-\alpha)\%$ confidence interval for θ is

$$\left[\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}}, \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}}\right].$$

Level of difficulty = medium

SEEN

(c) Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution specified by the cumulative distribution function

$$F_X(x) = 1 - \exp(\theta - x)$$

for $x > \theta$.

(i) The cumulative distribution function $\min\left(X_1, X_2, \ldots, X_n\right) = Z$ say, is

$$\begin{aligned} F_Z(z) &= \Pr\left(Z \leq z\right) \\ &= 1 - \Pr\left(\min\left(X_1, X_2, \ldots, X_n\right) > z\right) \\ &= 1 - \Pr\left(X_1 > z, \ldots, X_n > z\right) \\ &= 1 - \Pr\left(X_1 > z\right) \cdots \Pr\left(X_n > z\right) \\ &= 1 - \left[\Pr\left(X > z\right)\right]^n \\ &= 1 - \left[1 - F_X(z)\right]^n \\ &= 1 - \exp(n\theta - nz) \end{aligned}$$

for $z > \theta$. (3 marks)

Level of difficulty = medium

UNSEEN

(ii) Set $U = Z - \theta$. The cumulative distribution function of U is

$$F_U(u) = 1 - \exp(-nz)$$

for $z > \theta$. The $\left(\frac{\alpha}{2}\right)$th and $\left(1 - \frac{\alpha}{2}\right)$th percentiles of U are $-\frac{1}{n}\log\left(1 - \frac{\alpha}{2}\right)$ and $-\frac{1}{n}\log\left(\frac{\alpha}{2}\right)$, respectively. So,

$$\Pr\left(-\frac{1}{n}\log\left(1 - \frac{\alpha}{2}\right) < Z - \theta < -\frac{1}{n}\log\left(\frac{\alpha}{2}\right)\right) = 1 - \alpha,$$

which can be rewritten as

$$\Pr\left(Z + \frac{1}{n}\log\left(\frac{\alpha}{2}\right) < \theta < Z + \frac{1}{n}\log\left(1 - \frac{\alpha}{2}\right)\right) = 1 - \alpha.$$

Hence, a $100(1-\alpha)\%$ confidence interval for a is

$$\left[Z + \frac{1}{n}\log\left(\frac{\alpha}{2}\right), Z + \frac{1}{n}\log\left(1 - \frac{\alpha}{2}\right)\right].$$

(3 marks)

Level of difficulty = medium

UNSEEN

Solutions to Question 5

ILOs addressed: analyse and compare statistical properties of simple estimators.

Suppose $X \sim \text{Binomial}(m, p)$ and $Y \sim \text{Binomial}(n, p)$ are independent random variables. Consider the following estimators for p:

$$\widehat{p_1} = \frac{X}{2m} + \frac{Y}{2n}$$

and

$$\widehat{p_2} = \frac{X + Y}{m + n}.$$

(i) The bias of the first estimator is

$$\begin{aligned}
\text{Bias}\left(\widehat{p_1}\right) &= E\left(\widehat{p_1}\right) - p \\
&= E\left[\frac{1}{2}\left(\frac{X}{m} + \frac{Y}{n}\right)\right] - p \\
&= \frac{1}{2}\left[\frac{E(X)}{m} + \frac{E(Y)}{n}\right] - p \\
&= \frac{1}{2}\left(\frac{mp}{m} + \frac{np}{n}\right) - p \\
&= \frac{1}{2}(p + p) - p \\
&= 0.
\end{aligned}$$

(3 marks)

Level of difficulty = medium

UNSEEN

(ii) The bias of the second estimator is

$$\begin{aligned}
\text{Bias}\left(\widehat{p_2}\right) &= E\left(\widehat{p_2}\right) - p \\
&= E\left(\frac{X + Y}{m + n}\right) - p \\
&= \frac{E(X + Y)}{m + n} - p \\
&= \frac{E(X) + E(Y)}{m + n} - p \\
&= \frac{mp + np}{m + n} - p \\
&= p - p \\
&= 0.
\end{aligned}$$

(3 marks)

Level of difficulty = medium

UNSEEN

(iii) The mean squared error of the first estimator is

$$\begin{aligned}
\mathrm{MSE}\left(\widehat{p_1}\right) &= \mathrm{Var}\left(\widehat{p_1}\right) \\
&= \mathrm{Var}\left(\frac{1}{2}\left(\frac{X}{m}+\frac{Y}{n}\right)\right) \\
&= \frac{1}{4}\mathrm{Var}\left(\frac{X}{m}+\frac{Y}{n}\right) \\
&= \frac{1}{4}\left[\frac{\mathrm{Var}(X)}{m^2}+\frac{\mathrm{Var}(Y)}{n^2}\right] \\
&= \frac{1}{4}\left[\frac{mp(1-p)}{m^2}+\frac{np(1-p)}{n^2}\right] \\
&= \frac{1}{4}\left[\frac{p(1-p)}{m}+\frac{p(1-p)}{n}\right] \\
&= \frac{p(1-p)}{4}\left(\frac{1}{m}+\frac{1}{n}\right).
\end{aligned}$$

(4 marks)

Level of difficulty = medium

UNSEEN

(iv) The mean squared error of the second estimator is

$$\begin{aligned}
\mathrm{MSE}\left(\widehat{p_2}\right) &= \mathrm{Var}\left(\widehat{p_2}\right) \\
&= \mathrm{Var}\left(\frac{X+Y}{m+n}\right) \\
&= \frac{1}{(m+n)^2}\mathrm{Var}\left(X+Y\right) \\
&= \frac{1}{(m+n)^2}\left[\mathrm{Var}(X)+\mathrm{Var}(Y)\right] \\
&= \frac{1}{(m+n)^2}\left[mp(1-p)+np(1-p)\right] \\
&= \frac{p(1-p)}{m+n}.
\end{aligned}$$

(4 marks)

Level of difficulty = medium

UNSEEN

(v) Both estimators have zero bias, so they are equally good. (1 marks)

Level of difficulty = low

UNSEEN

(vi) $\widehat{p_2}$ is the better since it has smaller MSE than $\widehat{p_1}$ since

$$
\begin{aligned}
& \frac{p(1-p)}{m+n} \le \frac{p(1-p)}{4}\left(\frac{1}{m}+\frac{1}{n}\right) \\
\iff \quad & \frac{1}{m+n} \le \frac{1}{4}\left(\frac{1}{m}+\frac{1}{n}\right) \\
\iff \quad & \frac{1}{m+n} \le \frac{1}{4}\frac{m+n}{mn} \\
\iff \quad & 4mn \le (m+n)^2 \\
\iff \quad & 4mn \le m^2+n^2+2mn \\
\iff \quad & 0 \le m^2+n^2-2mn \\
\iff \quad & 0 \le (m-n)^2.
\end{aligned}
$$

(5 marks)

Level of difficulty = medium

UNSEEN

Solutions to Question 6

ILOs addressed: analyse statistical properties of simple estimators.

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution specified by the probability density function $\frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right)$ for $x > 0$.

(i) The likelihood function of σ^2 is

$$\begin{aligned} L\left(\sigma^2\right) &= \prod_{i=1}^{n} \left[\frac{\sqrt{2}}{\sqrt{\pi}\sigma} \exp\left(-\frac{X_i^2}{2\sigma^2}\right) \right] \\ &= \frac{2^{n/2}}{\pi^{n/2}\sigma^n} \left(\prod_{i=1}^{n} X_i \right) \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} X_i^2 \right). \end{aligned}$$

(4 marks)

Level of difficulty = medium

UNSEEN

(ii) The log likelihood function of σ^2 is

$$\log L\left(\sigma^2\right) = \frac{n}{2} \log 2 - \frac{n}{2} \log \pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n} X_i^2.$$

The derivative with respect to σ is

$$\frac{d \log L\left(\sigma^2\right)}{d\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{i=1}^{n} X_i^2.$$

Setting this to zero gives

$$\widehat{\sigma^2} = \frac{1}{n} \sum_{i=1}^{n} X_i^2.$$

This is a maximum likelihood estimator since

$$\begin{aligned} \frac{d^2 \log L\left(\sigma^2\right)}{d\sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4} \sum_{i=1}^{n} X_i^2 \\ &= \frac{1}{\sigma^4} \left[n\sigma^2 - 3 \sum_{i=1}^{n} X_i^2 \right] \\ &= \frac{1}{\sigma^4} \left[n \frac{1}{n} \sum_{i=1}^{n} X_i^2 - 3 \sum_{i=1}^{n} X_i^2 \right] \\ &< 0 \end{aligned}$$

at $\sigma = \widehat{\sigma}$. (4 marks)

Level of difficulty = medium

UNSEEN

(iii) By the invariance principle, the maximum likelihood estimator of σ is

$$\widehat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}.$$

(4 marks)

Level of difficulty = medium

UNSEEN

(iv) The bias of $\widehat{\sigma^2}$ is

$$\begin{aligned}
\text{Bias}\left(\widehat{\sigma^2}\right) &= E\left(\widehat{\sigma^2}\right) - \sigma^2 \\
&= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \sigma^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i^2\right) - \sigma^2 \\
&= \frac{\sqrt{2}}{n\sqrt{\pi}\sigma}\sum_{i=1}^{n}\int_0^\infty x^2 \exp\left(-\frac{x^2}{2\sigma^2}\right) dx - \sigma^2 \\
&= \frac{2\sigma^2}{n\sqrt{\pi}}\sum_{i=1}^{n}\int_0^\infty \sqrt{y}\exp\left(-y\right) dy - \sigma^2 \\
&= \frac{2\sigma^2}{n\sqrt{\pi}}\sum_{i=1}^{n}\Gamma\left(\frac{3}{2}\right) - \sigma^2 \\
&= \frac{2\sigma^2}{n\sqrt{\pi}}\sum_{i=1}^{n}\frac{\pi}{2} - \sigma^2 \\
&= 0.
\end{aligned}$$

Hence, $\widehat{\sigma^2}$ is unbiased for σ^2. (4 marks)

Level of difficulty = medium

UNSEEN

(v) The mean squared error of $\widehat{\sigma^2}$ is

$$
\begin{aligned}
\mathrm{MSE}\left(\widehat{\sigma^2}\right) &= \mathrm{Var}\left(\widehat{\sigma^2}\right) \\
&= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left(X_i^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left\{E\left(X_i^4\right) - \left[E\left(X_i^2\right)\right]^2\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left\{E\left(X_i^4\right) - \left[\sigma^2\right]^2\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left\{\frac{4\sigma^4}{\sqrt{\pi}}\int_0^\infty y^{3/2}\exp\left(-y\right)dy - \sigma^4\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left\{\frac{4\sigma^4}{\sqrt{\pi}}\Gamma\left(\frac{5}{2}\right) - \sigma^4\right\} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left\{3\sigma^4 - \sigma^4\right\} \\
&= \frac{2\sigma^4}{n}.
\end{aligned}
$$

Hence, $\widehat{\sigma^2}$ is consistent σ^2. (4 marks)

Level of difficulty = medium

UNSEEN

Solutions to Question 7

ILOs addressed: analyse statistical properties of simple estimators.

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution specified by the probability mass function

$$p_X(x) = \binom{x+r-1}{x}(1-p)^r p^x$$

for $x = 0, 1, \ldots$ with the properties

$$E(X) = \frac{pr}{1-p}$$

and

$$\mathrm{Var}(X) = \frac{pr}{(1-p)^2}.$$

Furthermore, assume r is known but p is unknown.

(i) The likelihood function of p is

$$L(p) = \prod_{i=1}^{n}\left[\binom{x_i+r-1}{x_i}(1-p)^r p^{x_i}\right] = \prod_{i=1}^{n}\left[\binom{x_i+r-1}{x_i}\right](1-p)^{nr} p^{\sum_{i=1}^{n} x_i}.$$

(4 marks)

Level of difficulty = medium

UNSEEN

(ii) The log likelihood function of p is

$$\log L(p) = \sum_{i=1}^{n}\log\left[\binom{x_i+r-1}{x_i}\right] + nr\log(1-p)\sum_{i=1}^{n} x_i \log p.$$

The derivative with respect to p is

$$\frac{d\log L(p)}{dp} = -\frac{nr}{1-p} + \sum_{i=1}^{n} x_i \frac{1}{p}.$$

Setting this to zero gives

$$\widehat{p} = \frac{\displaystyle\sum_{i=1}^{n} X_i}{nr + \displaystyle\sum_{i=1}^{n} X_i}.$$

This is a maximum likelihood estimator since

$$\frac{d^2 \log L(p)}{dp^2} = -\frac{nr}{(1-p)^2} - \sum_{i=1}^{n} x_i \frac{1}{p^2} < 0.$$

(4 marks)

Level of difficulty = medium

UNSEEN

(iii) The maximum likelihood estimator of $p/(1-p) = \psi$ say is

$$\widehat{\psi} = \frac{1}{nr} \sum_{i=1}^{n} X_i.$$

(4 marks)

Level of difficulty = medium

UNSEEN

(iv) The estimator in part (iii) is an unbiased estimator of ψ since

$$\begin{aligned}
\text{Bias}\left(\widehat{\psi}\right) &= E\left(\widehat{\psi}\right) - \psi \\
&= E\left(\frac{1}{nr} \sum_{i=1}^{n} X_i\right) - \psi \\
&= \frac{1}{nr} \sum_{i=1}^{n} E\left(X_i\right) - \psi \\
&= \frac{1}{nr} \sum_{i=1}^{n} \frac{pr}{1-p} - \psi \\
&= \frac{p}{1-p} - \psi \\
&= 0.
\end{aligned}$$

(4 marks)

Level of difficulty = medium

UNSEEN

(v) The estimator in part (iii) is a consistent estimator of ψ since

$$\begin{aligned} \mathrm{MSE}\left(\widehat{\psi}\right) &= \mathrm{Var}\left(\widehat{\psi}\right) \\ &= \mathrm{Var}\left(\frac{1}{nr}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n^2r^2}\sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) \\ &= \frac{1}{n^2r^2}\sum_{i=1}^{n} \frac{pr}{(1-p)^2} \\ &= \frac{p}{nr(1-p)^2} \end{aligned}$$

approaches zero as $n \to \infty$. (4 marks)

Level of difficulty = medium

UNSEEN