THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2007

MODULE 5

SPECIMEN PAPER B

AND SOLUTIONS

The time for the examination is 1½ hours. The paper contains four questions, of which candidates are to attempt **three**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

Note. In accordance with the convention used in all the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

1. The lengths X of offcuts of timber in a carpenter's workshop follow the continuous uniform distribution with probability density function (pdf)

$$f(x) = \frac{1}{\theta}, \qquad 0 \le x \le \theta,$$

where θ (> 0) is an unknown parameter.

(i) Find the mean and variance of X.

(5)

(ii) The carpenter takes a random sample of offcuts with lengths $X_1, X_2, ..., X_n$. Explain why

$$P\left(\text{length of longest offcut in sample } \le x\right) = \left(\frac{x}{\theta}\right)^n, \quad 0 \le x \le \theta,$$

and deduce the pdf of the sample maximum, $X_{(n)}$ say. Show that

$$E\left(X_{(n)}\right) = \frac{n\theta}{n+1}$$

and

$$\text{Var}\left(X_{(n)}\right) = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

Write down a multiple of $X_{(n)}$ which is an unbiased estimator of θ, and obtain its variance.

(11)

(iii) Show that $\frac{2}{n}\sum_{i=1}^{n} X_i$ is the method of moments estimator of θ, and obtain the variance of this estimator.

(4)

2. (i) The continuous random variable T has probability density function (pdf) $f(t)$ given by

$$f(t) = \lambda e^{-\lambda t}, \qquad t > 0; \quad \lambda > 0.$$

(a) Sketch the graph of $f(t)$.

(b) Show that the cumulative distribution function is given by

$$F(t) = 1 - e^{-\lambda t}, \qquad t > 0; \quad \lambda > 0.$$

(c) Deduce that

$$P(a < T \leq b) = e^{-\lambda a} - e^{-\lambda b}, \quad 0 < a < b.$$

(7)

(ii) The accounts manager for a building firm assumes that the time T taken to settle an invoice is a random variable with the above pdf $f(t)$ for some unknown value of λ. For a random sample of 100 invoices, he finds that 50 are settled within one week, 35 are settled during the second week and 15 are settled after 2 weeks. Explain clearly why the likelihood of these data may be written as

$$L(\lambda) = k\left(1 - e^{-\lambda}\right)^{50}\left(e^{-\lambda} - e^{-2\lambda}\right)^{35}\left(e^{-2\lambda}\right)^{15},$$

where k is a constant.

Hence show that

$$\log L(\lambda) = \log(k) + 85\log\left(1 - e^{-\lambda}\right) - 65\lambda \ .$$

Deduce that the maximum likelihood estimate of λ is approximately 0.836.

(8)

(iii) Assuming that $\lambda = 0.836$, calculate each of the expected numbers of invoices settled within the first week, settled during the second week, and settled after 2 weeks. Hence comment briefly on how well the model pdf $f(t)$ fits the data.

(5)

3. [In this question you may **use** the result $\int_0^\infty u^m e^{-u}\,du = m!$ for any non-negative integer m.]

The random variable X has probability density function

$$f(x) = \begin{cases} \dfrac{\lambda^{k+1} x^k e^{-\lambda x}}{k!}, & x > 0, \\ 0, & \text{elsewhere,} \end{cases}$$

where $\lambda > 0$ and k is a non-negative integer.

(i) Show that the moment generating function of X is $\left(\dfrac{\lambda}{\lambda - \theta}\right)^{k+1}$ for $\theta < \lambda$.

(8)

(ii) The random variable Y is the sum of n independent random variables each distributed as X. Find the moment generating function of Y and hence obtain the mean and variance of Y.

(9)

(iii) State the probability density function of Y.

(3)

4. (a) (i) Explain the terms *unbiased* and *consistent* as applied to point estimators of parameters. Discuss briefly how these properties apply to the maximum likelihood estimator of μ in the N(μ, σ^2) distribution, and to the maximum likelihood estimator of σ^2 in this distribution in the case where neither parameter is known. You may **use without proof** the expressions for these estimators:

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\Sigma(X_i - \bar{X})^2 .$$

You may also a**ssume without proof** that Var($\hat{\sigma}^2$) is O(1/n).

(ii) Let X, Y be random variables having finite variances. Show that Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y).

(iii) Observations X_1, X_2 have the same mean μ and the same variance σ^2, and the correlation coefficient between X_1 and X_2 is ρ. Show that $\bar{X} = (X_1 + X_2)/2$ is an unbiased estimator of μ and find its variance.

Under which of the following conditions would $\bar{X}$ estimate μ most precisely:

(A) if X_1, X_2 were uncorrelated;

(B) if X_1, X_2 were positively correlated;

(C) if X_1, X_2 were negatively correlated?

(12)

(b) The joint probability density function of the continuous random variables X, Y is $f(x, y) = e^{-x-y}$, $x > 0$, $y > 0$.

Find the marginal distributions of X and Y, and their means and variances. Are X and Y independent (justify your answer)?

(8)

SOLUTIONS

Question 1

(i) $E(X) = \int_0^\theta \frac{x}{\theta} dx = \frac{1}{\theta}\left[\frac{1}{2}x^2\right]_0^\theta = \frac{1}{2}\theta$.

$E(X^2) = \int_0^\theta \frac{x^2}{\theta} dx = \frac{1}{\theta}\left[\frac{1}{3}x^3\right]_0^\theta = \frac{1}{3}\theta^2$.

$\therefore \text{Var}(X) = E(X^2) - \{E(X)\}^2 = \frac{1}{3}\theta^2 - \left(\frac{1}{2}\theta\right)^2 = \frac{1}{12}\theta^2$.

(ii) P(longest offcut is $\le x$) = P(all n offcuts are $\le x$).

The c.d.f. for each X_i is $F(x) = P(X \le x) = \int_0^x \frac{du}{\theta} = \left[\frac{u}{\theta}\right]_0^x = \frac{x}{\theta}$, and the X_i are all independent. Therefore P(all n offcuts are $\le x$) = $\{F(x)\}^n = \left(\frac{x}{\theta}\right)^n$, and this is also P(longest offcut is $\le x$), i.e. the c.d.f. of the sample maximum $X_{(n)}$. Thus the p.d.f. of $X_{(n)}$ is the derivative of this, i.e. nx^{n-1}/θ^n. This is for the interval (0, θ).

$\therefore E(X_{(n)}) = \int_0^\theta \frac{nx^n}{\theta^n} dx = \frac{n}{\theta^n}\left[\frac{x^{n+1}}{n+1}\right]_0^\theta = \frac{n\theta}{n+1}$.

$E\left(X_{(n)}^{\,2}\right) = \int_0^\theta \frac{nx^{n+1}}{\theta^n} dx = \frac{n}{\theta^n}\left[\frac{x^{n+2}}{n+2}\right]_0^\theta = \frac{n\theta^2}{n+2}$.

$\therefore \text{Var}(X_{(n)}) = E\left(X_{(n)}^{\,2}\right) - \left\{E(X_{(n)})\right\}^2 = \frac{n\theta^2}{n+2} - \frac{n^2\theta^2}{(n+1)^2}$

$= n\theta^2\left(\frac{(n+1)^2 - n(n+2)}{(n+2)(n+1)^2}\right) = \frac{n\theta^2}{(n+1)^2(n+2)}$.

Immediately we have $E\left(\frac{n+1}{n}X_{(n)}\right) = \theta$, so $\frac{n+1}{n}X_{(n)}$ is an unbiased estimator of θ.

$\text{Var}\left(\frac{n+1}{n}X(n)\right) = \frac{(n+1)^2}{n^2}\text{Var}(X_{(n)}) = \frac{(n+1)^2}{n^2}\frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$.

Solution continued on next page

(iii) We have (see part (i)) that $E(X) = \theta/2$. Thus the method of moments estimator of $\theta/2$ is $\bar{X}$, and so the method of moments estimator of θ is $2\bar{X}$ or $\frac{2}{n}\sum X_i$ as required.

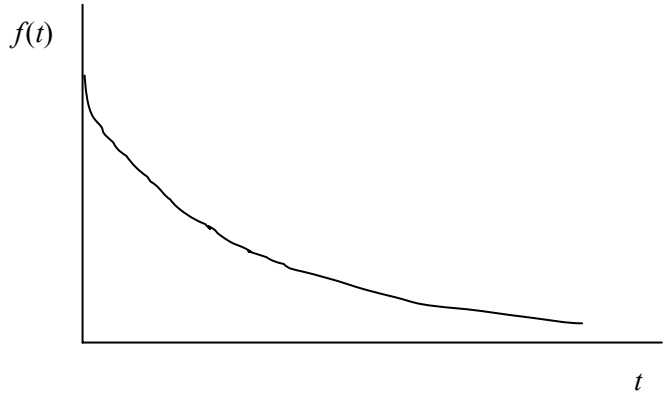
$$\mathrm{Var}\left(\frac{2}{n}\sum X_i\right) = \mathrm{Var}\left(2\bar{X}\right) = 4\mathrm{Var}\left(\bar{X}\right) = \frac{4}{n}\mathrm{Var}\left(X\right) = \frac{4}{n}.\frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

Question 2

(i) $f(t) = \lambda e^{-\lambda t}, \qquad t > 0; \quad \lambda > 0$

(a) Sketch of $f(t)$.

[***NOTE.*** *The curve should of course appear as a smooth decaying exponential; it might not do so, due to the limits of electronic reproduction.*]



(b) C.d.f. is $F(t) = P(T \le t) = \int_0^t \lambda e^{-\lambda v} dv = \lambda \left[-\frac{1}{\lambda} e^{-\lambda v} \right]_0^t = 1 - e^{-\lambda t}$.

(c) $P(a < T \le b) = F(b) - F(a) = e^{-\lambda a} - e^{\lambda b}$.

(ii) Assume all settlements of invoices are independent.

P(50 in first week) = $\{F(1)\}^{50} = \left(1 - e^{-\lambda}\right)^{50}$, because $T \le 1$ for all these 50.

Likewise, $1 < T \le 2$ for the 35 in the second week, so we have P(35 in second week) = $\{F(2) - F(1)\}^{35} = \left(e^{-\lambda} - e^{-2\lambda}\right)^{35}$.

The remaining 15 have $T > 2$, which has probability $1 - P(T \le 2) = e^{-2\lambda}$, and thus P(15 after week 2) = $\left(e^{-2\lambda}\right)^{15}$.

The likelihood is therefore the product

$$L(\lambda) = k\left(1 - e^{-\lambda}\right)^{50} \left(e^{-\lambda} - e^{-2\lambda}\right)^{35} \left(e^{-2\lambda}\right)^{15}$$

where k is a constant of proportionality.

Solution continued on next page

Taking logarithms (base *e*),

$$\log L(\lambda) = \log k + 50\log\left(1-e^{-\lambda}\right) + 35\log\left\{e^{-\lambda}\left(1-e^{-\lambda}\right)\right\} + 15\log\left(e^{-2\lambda}\right)$$

$$= \log k + 85\log\left(1-e^{-\lambda}\right) - (35+30)\lambda \;=\; \log k + 85\log\left(1-e^{-\lambda}\right) - 65\lambda\,.$$

$$\therefore \frac{d}{d\lambda}\log L \;=\; \frac{85e^{-\lambda}}{1-e^{-\lambda}} - 65 \;=\; \frac{85}{e^{\lambda}-1} - 65\,.$$

Equating to zero, $85 = 65\left(e^{\lambda}-1\right)$ or $e^{\lambda} = 150/65$, so that $\hat{\lambda} = \log(150/65) = 0.836$.

[It is easy to check that this is indeed a maximum; e.g. $\dfrac{d^2}{d\lambda^2}\log L = -\dfrac{85}{\left(e^{\lambda}-1\right)^2} \;<\; 0$.]

(iii) $1-e^{-0.836} = 0.5666$; $e^{-0.836} - e^{-1.672} = 0.43344 - 0.18787 = 0.2456$. Hence, out of 100 invoices, 56.66, 24.56 and 18.78 would be expected to be paid, on this model, in weeks 1, 2 and later. The actual numbers were 50, 35 and 15. The prediction for the second week is a long way from what happened, balanced by smaller discrepancies in the other two periods. This does not seem very satisfactory.

Question 3

$$f(x)=\begin{cases}\dfrac{\lambda^{k+1}x^k e^{-\lambda x}}{k!}, & x>0\\ 0, & \text{elsewhere}\end{cases}$$

Given result: $\int_0^\infty u^m e^{-u}du = m!$

(i) Moment generating function of X is

$$E\left(e^{\theta X}\right)=\int_0^\infty \frac{\lambda^{k+1}}{k!}x^k e^{-(\lambda-\theta)x}dx \qquad \text{[Use substitution } (\lambda-\theta)x = u \text{]}$$

$$=\frac{\lambda^{k+1}}{k!(\lambda-\theta)^{k+1}}\int_0^\infty u^k e^{-u}du \qquad \text{[The integral here is } k! \text{ by the given result]}$$

$$=\left(\frac{\lambda}{\lambda-\theta}\right)^{k+1} \quad \text{as required.}$$

(Note that the condition $\theta < \lambda$ is required to ensure that u remains positive in the substitution.)

(ii) $Y = X_1 + X_2 + \ldots + X_n$, where all the X_i are independent. So by the convolution theorem for moment generating functions, the mgf of Y is simply the mgf of X raised to the nth power, i.e. it is

$$M(\theta)=\left(\frac{\lambda}{\lambda-\theta}\right)^{nk+n}=\lambda^{nk+n}(\lambda-\theta)^{-nk-n} \qquad (\text{for } \theta<\lambda).$$

The mean of Y is given by $M'(0)$ and the variance by $\{M''(0) - (\text{mean})^2\}$.

Differentiating, $M'(\theta)=\lambda^{nk+n}(-nk-n)(\lambda-\theta)^{-nk-n-1}(-1)$.

Inserting $\theta = 0$ gives mean $= \dfrac{nk+n}{\lambda}$.

Differentiating again, $M''(\theta)=(nk+n)\lambda^{nk+n}(-nk-n-1)(\lambda-\theta)^{-nk-n-2}(-1)$.

Inserting $\theta = 0$ in this gives $M''(0)=(nk+n)(nk+n+1)/\lambda^2$ and so the variance of Y is

$$\frac{(nk+n)(nk+n+1)}{\lambda^2}-\frac{(nk+n)^2}{\lambda^2}=\frac{nk+n}{\lambda^2}.$$

Solution continued on next page

(iii) The moment generating function of Y is of the same functional form as that of X with $k + 1$ replaced by $nk + n$, i.e. k replaced by $nk + n - 1$. Because of the uniqueness of the relationship between a distribution and its moment generating function, this must also be true of the probability density function. So the pdf of Y is

$$\frac{\lambda^{nk+n}}{(nk+n-1)!} y^{nk+n-1} e^{-\lambda y} \qquad \text{(for } y > 0, \text{ and zero elsewhere).}$$

Question 4

(a) (i) An estimator is *unbiased* if the expectation (mean) of its sampling distribution is equal to the parameter being estimated. An estimator is *consistent* if the probability of it differing from the parameter being estimated by more than ε, a very small quantity, approaches 0 as sample size $\to \infty$. It is however easier to use a criterion based on variance: if the variance of the sampling distribution $\to 0$ as sample size $\to \infty$, the estimator is *consistent*. [Some care is needed in using this criterion for biased estimators, in case the estimator is "homing in" on the wrong place. Provided any bias itself $\to 0$ as the sample size $\to \infty$, the criterion is satisfactory.]

The estimator of μ in N(μ, σ^2) is $\bar{X}$, and we have the standard results $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \sigma^2/n$. So $\bar{X}$ is both unbiased and consistent.

The estimator of σ^2 is $\frac{1}{n}\Sigma(X_i - \bar{X})^2$. This is not unbiased: standard results give that divisor $n - 1$ is required for unbiasedness, whereas the expectation of this estimator (divisor n) is $[(n-1)/n]\sigma^2$. But it is consistent.

(ii) $$\text{Var}(X+Y) = E\left[\left\{X+Y-E(X+Y)\right\}^2\right]$$

$$= E\left[\left\{X-E(X)\right\}^2 + \left\{Y-E(Y)\right\}^2 + 2\left\{X-E(X)\right\}\left\{Y-E(Y)\right\}\right]$$

$$= \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y).$$

(iii) $E(\bar{X}) = E\left((X_1 + X_2)/2\right) = (\mu+\mu)/2 = \mu$, so $\bar{X}$ is unbiased.

$$\text{Var}(\bar{X}) = \tfrac{1}{4}\text{Var}(X_1 + X_2) = \tfrac{1}{4}\left\{\text{Var}(X_1) + \text{Var}(X_2) + 2\,\text{Cov}(X_1, X_2)\right\}$$

Now, $\text{Cov}(X_1, X_2) = \rho\sqrt{\text{Var}(X_1)\text{Var}(X_2)} = \rho\sigma^2$. So we have

$$\text{Var}(\bar{X}) = \tfrac{1}{4}\left(\sigma^2 + \sigma^2 + 2\rho\sigma^2\right) = \tfrac{1}{2}\sigma^2\left(1+\rho\right).$$

We now require the minimum variance for each of the three possible situations (*A*), (*B*), (*C*), and this is clearly when $\rho < 0$, i.e. negative correlation (situation (*C*)).

Solution continued on next page

(b) $f(x,y) = e^{-x-y}, \quad x > 0,\ y > 0.$

The marginal distribution of X is obtained by "integrating out y", i.e. it is

$$f_X(x) = \int_{y=0}^{y=\infty} e^{-x}e^{-y}dy = e^{-x}\left[-e^{-y}\right]_0^{\infty} = e^{-x}\left[0-(-1)\right] = e^{-x}.$$

This can be recognised as the exponential distribution with mean 1, and for which the variance is also 1. Alternatively, the mean and variance are easily obtained by routine integration.

By symmetry, Y has the same distribution (marginally) as X. [If the symmetry is not recognised, it is straightforward to obtain this marginal distribution in the same way as above, integrating out x.]

X and Y are independent. The joint probability density function $f(x, y)$ can be factorised as the product of the two marginal probability density functions.