ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2014

MODULE 2 : Statistical inference

Time allowed: Three hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

The notation log *denotes logarithm to base* ***e****.*
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^nC_r$.

This examination paper consists of 8 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. In a quality control department, the number V of items that has to be tested until a faulty item is found has the geometric distribution given by

$$P(V = k) = p(1-p)^{k-1} \qquad k = 1, 2, \ldots,$$

where p ($0 < p < 1$) is an unknown parameter. Independent observations $V_1, V_2, \ldots, V_n$ have been made, each with the same distribution as V, and it is required to estimate p. Let $W = \sum_{i=1}^{n} V_i$.

(i) You are **given** that the negative binomial distribution with parameters n and p has probability distribution

$$p(k) = \binom{k-1}{n-1} p^n (1-p)^{k-n} \quad \text{for } k = n, n+1, n+2, \ldots,$$

probability generating function $\left(\frac{pt}{1-t(1-p)}\right)^n$ for $t < (1-p)^{-1}$, and mean $\frac{n}{p}$. Write down the probability generating function of the geometric distribution and hence show that W has a negative binomial distribution.

(3)

(ii) Show that the Cramér-Rao lower bound for the variance of unbiased estimators of p is equal to $\frac{p^2(1-p)}{n}$.

(6)

(iii) For $i = 1, 2, \ldots, n$, define the random variable X_i to take the value 1 when $V_i = 1$ and the value 0 when $V_i > 1$. Show that $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of p and find its variance. (Any general result that you use need not be proved but must be stated clearly.)

(4)

(iv) By defining $T = \sum_{j \neq i} V_j$ and considering the joint distribution of V_i and T, show that $P(X_i = 1 | W = w) = \frac{n-1}{w-1}$.

(5)

(v) Hence show that $\tilde{p} = \frac{n-1}{W-1}$ is an unbiased estimator of p.

[You may assume the result that $E(X_i) = \sum_{w=1}^{\infty} P(X_i = 1 | W = w) P(W = w)$.]

(2)

2. Explain what is meant by the *invariance property* of maximum likelihood estimators.

(2)

In a search where there are M (> 0) objects to be found, the time T_1 until the first find has the exponential distribution with mean $\frac{\mu}{M}$. The first find is then removed and the time until the next find is T_2, which is independent of T_1 and has the exponential distribution with mean $\frac{\mu}{M-1}$. This continues in the same way so that for $i = 1, 2, \ldots, M$ the time T_i between the $(i-1)$th and ith finds has the exponential distribution with mean $\frac{\mu}{M+1-i}$, independent of $T_1, T_2, \ldots, T_{i-1}$.

(i) Assume that M is known and μ is unknown. Show the log likelihood based on $T_1, T_2, \ldots, T_n$ $(n \leq M)$ can be written as

$$\text{constant} - n\log\mu - \frac{1}{\mu}\sum_{i=1}^{n}(M+1-i)T_i$$

and that the maximum likelihood estimator of μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}(M+1-i)T_i\,.$$

(7)

(ii) Find the mean and variance of $\hat{\mu}$.

[You may assume the result that the variance of the exponential distribution is equal to the square of its mean.]

(4)

(iii) Now assume that μ is known and M is unknown. By considering the ratio of the likelihood at M and the likelihood at $M+1$, show that the maximum likelihood estimator of M based on $T_1, T_2, \ldots, T_n$ is the largest integer less than or equal to

$$\hat{M} = \frac{n\exp\left(\mu^{-1}\sum_{i=1}^{n}T_i\right)}{\exp\left(\mu^{-1}\sum_{i=1}^{n}T_i\right)-1}\,.$$

(5)

(iv) By considering the behaviour of $\hat{M}$ when $\exp\left(\mu^{-1}\sum_{i=1}^{n}T_i\right)$ is large, comment on whether you think this is a very satisfactory estimator.

(2)

3. Observations $X_1, X_2, \ldots, X_{n_1}, Y_1, Y_2, \ldots, Y_{n_2}$ are independent, with X_i having the $\mathrm{N}(\mu, a_i\sigma_1^2)$ distribution for $i = 1, 2, \ldots, n_1$ and Y_i having the $\mathrm{N}(\mu, b_i\sigma_2^2)$ distribution for $i = 1, 2, \ldots, n_2$. Here $a_1, a_2, \ldots, a_{n_1}, b_1, b_2, \ldots, b_{n_2}$ are known positive constants. The parameters σ_1^2 and σ_2^2 are unknown, but μ is known.

(i) Write down the distribution of $\dfrac{X_i - \mu}{\sqrt{a_i\sigma_1^2}}$ and find the distributions of

$$V = \sum_{i=1}^{n_1} \frac{(X_i - \mu)^2}{a_i\sigma_1^2} \quad \text{and} \quad W = \sum_{i=1}^{n_2} \frac{(Y_i - \mu)^2}{b_i\sigma_2^2}.$$

[Here and below standard results concerning distributions can be assumed without proof, but should be stated clearly.]

(4)

(ii) Let $U = \dfrac{\sigma_1^2(V/n_1)}{\sigma_2^2(W/n_2)}$. State the distribution of $U\dfrac{\sigma_2^2}{\sigma_1^2}$ and explain why this quantity is pivotal for $\dfrac{\sigma_1^2}{\sigma_2^2}$.

(4)

(iii) Explain how a 90% confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$ can be found using part (ii), and use your method to find the 90% confidence interval when $n_1 = 8$, $n_2 = 12$, $\displaystyle\sum_{i=1}^{n_1} \frac{(X_i - \mu)^2}{a_i} = 15$ and $\displaystyle\sum_{i=1}^{n_2} \frac{(Y_i - \mu)^2}{b_i} = 5$.

(6)

(iv) It is now required to test H_0: $\sigma_1^2 = \sigma_2^2$ against H_1: $\sigma_1^2 < \sigma_2^2$ at the 5% level. Find such a test in the case $n_1 = 8$, $n_2 = 12$.

(3)

(v) Obtain the approximate power of the test found in part (iv) at $7.35\sigma_1^2 = \sigma_2^2$.

(3)

4. State the *Neyman-Pearson lemma*.

(3)

A random sample of observations X_1, X_2, …, X_n comes from a distribution with probability density function $f(x)$. The null hypothesis is that $f(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2}$ ($x > 0$) and the alternative hypothesis is that $f(x) = e^{-x}$ ($x > 0$). Under both hypotheses, $f(x) = 0$ for $x \leq 0$.

(i) Show that the most powerful test has a critical region depending on the value of $\sum_{i=1}^{n} (X_i - 1)^2$.

(5)

(ii) Show that, under the null hypothesis, $E(X) = \sqrt{\frac{2}{\pi}}$.

(2)

(iii) Noting that under the null hypothesis X has the same distribution as $|Z|$, where Z has the N(0, 1) distribution, and using the result that $E(Z^4) = 3$, show that $E(X_i^2) = 1$ and $E(X_i^4) = 3$ under the null hypothesis.

(2)

(iv) Using the results of parts (ii) and (iii) and the result that under the null hypothesis $E(X_i^3) = 2\sqrt{\frac{2}{\pi}}$, evaluate $E\left(\sum_{i=1}^{n} (X_i - 1)^2\right)$ and $\mathrm{Var}\left(\sum_{i=1}^{n} (X_i - 1)^2\right)$ under the null hypothesis.

(4)

(v) Use the central limit theorem to find the critical region of a test with size approximately equal to 0.10 when $n = 200$.

(4)

5. Independent discrete random variables $X_1, X_2, \ldots, X_n$ have probability distribution

$$P(X_j = k) = \begin{cases} (1-p)^3 & \text{for } k = 0 \\ 3p(1-p) & \text{for } k = 1 \\ p^3 & \text{for } k = 2 \\ 0 & \text{otherwise} \end{cases}$$

where p $(0 < p < 1)$ is an unknown parameter. Let Y_k $(k = 0, 1, 2)$ be the number of the variables $X_1, X_2, \ldots, X_n$ that take the value k. It is required to test the null hypothesis $p = p_0$ against the alternative $p \neq p_0$, where p_0 $(0 < p_0 < 1)$ is a known value.

(i) Show that the maximum likelihood estimator of p is

$$\hat{p} = \frac{Y_1 + 3Y_2}{3Y_0 + 2Y_1 + 3Y_2}.$$

(7)

(ii) Find the form of the generalised likelihood ratio test.

(4)

(iii) Using an asymptotic result, find the critical region of this test if its size is to be approximately 0.05.

(3)

(iv) Explain how the generalised likelihood ratio test can be used to find an approximate 95% confidence interval for p.

(2)

(v) Show that in the case $Y_0 = Y_2$ this confidence interval can be written as

$$p(1-p) \geq 0.25 \exp\left(-\frac{1.92}{Y_1 + 3Y_2}\right).$$

(4)

6. (a) When reporting the results of a statistical test comparing the outcomes of two treatments, researchers sometimes say that they have found a "significant difference" or alternatively that "the difference was not significant". Explain why these two forms of words might be misleading.

(6)

(b) What is meant by *robustness* in statistics? What are the implications of robustness for estimation, confidence intervals and hypothesis testing?

(5)

(c) In Table 11 of the *statistical tables for use in examinations* (Wilcoxon signed rank test), the first entry against sample size 6 (i.e. under the 0.05 column) is "2". Explain carefully how this value is derived.

(9)

7. What is meant by a *sufficient statistic*? Suppose that a random sample is drawn from a distribution with parameter θ having a prior distribution $\pi(\theta)$ and that T is a sufficient statistic for θ. Show that the posterior distribution of θ given the random sample is identical to the posterior distribution of θ given T.

(6)

Suppose that $X_1, X_2, \ldots, X_{2n}$ constitute a random sample of $2n$ observations from a N(0, θ^{-1}) distribution, where the prior distribution of θ is exponential with known mean λ^{-1}.

(i) Find the posterior distribution of θ.

[You may use the results that the gamma distribution with parameters r and ν has probability density $f(y) = \dfrac{\nu^r y^{r-1} e^{-\nu y}}{\Gamma(r)}$ for $y > 0$, where $\Gamma(.)$ is the gamma function, and moment generating function $\left(\dfrac{\nu}{\nu - t}\right)^r$, for $t < \nu$.]

(5)

(ii) Use moment generating functions to show that the posterior distribution is also the distribution of the sum of $n + 1$ independent, identically distributed random variables each having an exponential distribution.

(3)

(iii) Find the mean and variance of the posterior distribution and use the central limit theorem to deduce an approximate 95% Bayesian confidence interval for θ when n is large.

(6)

8. Explain what is meant by a *decision rule* and a *minimax decision rule* in the context of decision theory.

(3)

Suppose that X is a discrete random variable with distribution given by

$$P(X = k) = (1 - p)p^k \quad \text{for } k = 0, 1, 2, \ldots$$

where p $(0 < p < 1)$ is unknown. After observing the value of X, one of two actions must be taken. If action 1 is taken, there is a loss of 1 if $p > 0.5$ or a loss of -3 if $p \leq 0.5$. If action 2 is taken, there is a loss of 0 if $p > 0.5$ or a loss of 2 if $p \leq 0.5$. Under decision rule δ_k, action 1 is taken if $X < k$ and otherwise action 2 is taken, for $k = 0, 1, 2, 3, \ldots$.

(i) Find the minimax decision rule among $\delta_0, \delta_1, \delta_2, \delta_3, \ldots$.

(9)

(ii) Suppose that the prior distribution of p is $\pi(p) = 2p$ $(0 < p < 1)$. Show that the Bayes risk, $B(k)$, of δ_k is given by

$$B(k) = \frac{(3 \times 0.5^k) - 2}{k + 2}.$$

(4)

(iii) Given that $B(k+1) - B(k) = \dfrac{2 - 0.5^k(1.5k + 6)}{(k + 2)(k + 3)}$, show that δ_3 is the decision rule which minimises the Bayes risk among $\delta_0, \delta_1, \delta_2, \delta_3, \ldots$.

(4)