EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2016

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* ${}^{n}C_{r}$.

This examination paper consists of 12 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. $X_1, X_2, ..., X_n$ is a random sample from the uniform distribution between θ and 1 (i.e. $f(x) = (1-\theta)^{-1}$ for $\theta < x < 1$), where θ (< 1) is an unknown parameter. Denote the sample mean by $\bar{X}$.

(i) Show that the method of moments estimator, $\hat{\theta}$, of θ is $2\bar{X} - 1$.

(3)

(ii) Show that $\hat{\theta}$ is unbiased and find its variance.

(4)

(iii) State the *Cramér-Rao lower bound* for the variance of unbiased estimators, and explain why it is not applicable when estimating θ defined above.

(3)

(iv) Let $Y = \min\{X_1, X_2, ..., X_n\}$. By finding $P(Y > y)$, show that the probability density function of Y is $f(y) = \dfrac{n(1-y)^{n-1}}{(1-\theta)^n}$ for $\theta < y < 1$.

(4)

(v) An estimator $\tilde{\theta} = 1 - c(1-Y)$ is to be used to estimate θ, where c is a constant to be chosen. Show that the mean square error of $\tilde{\theta}$, $E[(\tilde{\theta} - \theta)^2]$, is minimised when $c = \dfrac{n+2}{n+1}$.

(6)

[You are given that $E(1-Y) = \dfrac{n(1-\theta)}{n+1}$ and $E[(1-Y)^2] = \dfrac{n(1-\theta)^2}{n+2}$.]

2. (i) Explain what is meant by the *invariance property* of maximum likelihood estimators (MLEs).

(2)

(ii) Consider a binomial experiment with n trials and probability of success π in which y successes are observed. Find the MLE of π.

(5)

(iii) A random sample $X_1, X_2, ..., X_n$ of observations is taken from an exponential distribution with probability density function $f(x) = \frac{1}{\theta} e^{-\frac{x}{\theta}},\ x > 0$. The actual values of the observations are not available; all that is known is that y of the n values are less than a threshold T and the remaining $(n - y)$ values are greater than T. Use part (ii) and the invariance property of MLEs to show that the MLE $\hat{\theta}$ of θ, given this restricted information, is $-T\left[\log\left(1-\frac{y}{n}\right)\right]^{-1}$.

(8)

(iv) State the asymptotic distribution of $\hat{\theta}$ and hence write down an expression for an approximate 95% confidence interval for θ, defining any terms that appear in your expression.

(5)

3. (i) A random sample $X_1, X_2, ..., X_n$ is available from a distribution with probability density function (pdf) $f(x;\theta)$, where θ is a single unknown parameter. Define what is meant by a *statistic* and by a *sufficient statistic* and explain what sufficiency means.

(3)

(ii) A random variable X is said to belong to the one-parameter exponential family of distributions if its pdf can be written in the form

$$f(x;\theta) = \exp\{A(\theta)B(x) + C(x) + D(\theta)\}$$

where $A(\theta)$, $D(\theta)$ are functions of the single parameter θ (but not x) and $B(x)$, $C(x)$ are functions of x (but not θ). Write down the likelihood function, given a random sample $X_1, X_2, ..., X_n$ from a distribution with such a pdf.

(2)

(iii) If the likelihood can be expressed as the product of a function which depends on θ and which depends on the data only through a statistic $T(X_1, X_2, ..., X_n)$ and a function that does not depend on θ, then it can be shown that T is sufficient for θ. Use this result to show that $\sum_{i=1}^{n} B(x_i)$ is a sufficient statistic for θ in the one-parameter exponential family of part (ii).

(3)

(iv) The Rayleigh distribution has pdf

$$f(x;\theta) = \frac{x}{\theta}\exp\left(-\frac{x^2}{2\theta}\right), \quad x > 0, \quad \theta > 0.$$

Show that this distribution is a member of the one-parameter exponential family and hence find a sufficient statistic for θ.

(4)

(v) In Bayesian inference define what is meant by a *prior distribution* and by a *conjugate prior distribution*. Show that the family of distributions with pdfs proportional to $\theta^{-\alpha_1}\exp\left(-\frac{\alpha_2}{\theta}\right)$, $\alpha_1 > 0$, $\alpha_2 > 0$ is conjugate for the family of Rayleigh distributions defined in part (iv).

(8)

4. (i) Define what is meant by a *generalised likelihood ratio test*.

(3)

(ii) Show how the test statistic from a generalised likelihood ratio test, with a simple null hypothesis for a single parameter θ, can be used to construct an approximate confidence interval for the parameter of the form $l(\theta;\boldsymbol{x}) \geq l(\hat{\theta};\boldsymbol{x}) - \tfrac{1}{2}\chi^2_{1;\alpha}$, where $l(\theta;\boldsymbol{x})$ is the log-likelihood function and $\chi^2_{1;\alpha}$ is an appropriate critical value from the χ^2 distribution with 1 degree of freedom.

(6)

(iii) The number of minor accidents per month in a large factory can be assumed to have a Poisson distribution with mean μ. Obtain the maximum likelihood estimator of μ given a random sample of observations from this distribution. Over a long period, the mean number of minor accidents per month has been around 5. Some of the safety procedures in the factory have changed recently. In the first 9 months after the change, the numbers of minor accidents were 8, 3, 2, 2, 3, 1, 2, 5, 1. Use a generalised likelihood ratio test to test the null hypothesis that $\mu = 5$ against a two-sided alternative at the 5% significance level.

(8)

(iv) Show that the endpoints of an approximate 95% interval for μ are the solutions of the equation $3\log\mu - \mu = 0.0825$.

(3)

5. A random sample of observations $X_1, X_2, ..., X_n$ is available from a probability distribution with probability density function $f(x;\theta)$, where θ is a single unknown parameter.

(a) Define a *pivotal quantity* and explain how to use such a quantity to construct a (frequentist) confidence interval for θ.

(6)

(b) Explain how to construct a credible interval (Bayesian confidence interval).

(5)

(c) Explain how to construct a bootstrap percentile confidence interval.

(5)

Give interpretations for the intervals in (a) and (b), contrasting those interpretations.

(4)

6. Given a random sample $X_1, X_2, ..., X_n$ from a distribution with probability density function $f(x)$, it is required to test the hypothesis $\mathrm{H}_0 : f(x) = f_0(x)$, where $f_0(x)$ is a known function, against the general alternative $\mathrm{H}_1 : f(x) \neq f_0(x)$.

(i) Define the test statistic used in the one-sample Kolmogorov-Smirnov test of these hypotheses.

(3)

(ii) The data below are times to breakdown, x, of five mechanical components. It is required to test whether they could have been generated by an exponential distribution with mean 2. Use the one-sample Kolmogorov-Smirnov test to examine this hypothesis.

[You are given that the 5% critical value for this test is 0.563 and the 1% critical value is 0.669. You are also given $1 - e^{-\frac{1}{2}x}$ for each of the five values of x, as follows.

x	0.365	1.154	2.849	4.466	4.578
$1 - e^{-\frac{1}{2}x}$	0.167	0.438	0.759	0.893	0.899]

(8)

(iii) It is now required to test $\mathrm{H}_0 : m = m_0$ against $\mathrm{H}_1 : m \neq m_0$ using the data and assumptions in part (ii), for some measure of location m. It is suggested that either the t test or the sign test might be used. Discuss which measure of location would be used for each of these tests and which test it would be most appropriate to carry out.

(4)

(iv) Use the sign test to investigate whether the data in part (ii) could have come from a distribution with mean 2, assuming that the distribution is exponential. [You need not calculate the p-value associated with this test.]

(5)

7. (a) Given two simple hypotheses, $\mathrm{H}_0, \mathrm{H}_1$, define what is meant by the *odds* of H_0 and by the *Bayes factor* of H_0 compared to H_1. Show that in Bayesian inference when comparing two simple hypotheses, the posterior odds equals the product of the prior odds and the Bayes factor.

(5)

(b) A random sample $X_1, X_2, ..., X_n$ is available from a Poisson distribution with mean λ (> 0). It is required to test the null hypothesis $\mathrm{H}_0 : \lambda = 5$ against the alternative hypothesis $\mathrm{H}_1 : \lambda = 10$.

(i) Show that the Bayes factor is $e^{5n}(0.5)^{\Sigma x_i}$.

(4)

(ii) Show that the posterior odds of the null hypothesis will be greater than the prior odds if $\bar{x} < \dfrac{5}{\log 2}$, where $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$.

(4)

Suppose now that the null hypothesis is as above, but the alternative hypothesis is $\mathrm{H}_1 : \lambda \neq 5$, and the prior distribution of λ under H_1 is exponential with mean 5.

(iii) Show that the Bayes factor is $\dfrac{e^{-5n}(5n+1)^{1+\Sigma x_i}}{\Gamma(1+\Sigma x_i)}$.

[Hint: recall that the gamma function can be written as

$$\Gamma(\alpha) = \int_0^\infty v^\alpha x^{\alpha-1} e^{-vx}\, dx.]$$

(7)

8. A company that manufactures a narrow range of products is operating in an economy in economic recession and is faced with a persistent drop in sales. The managers are considering five possible strategies to cope with the crisis, as follows.

d_1: continue as at present

d_2: continue with present products alone, make 10% of the workforce redundant

d_3: continue with present products alone, make 20% of the workforce redundant

d_4: start small-scale production of new products, make 10% of the workforce redundant

d_5: start large-scale production of new products, keep present workforce

Four states of nature are envisaged, as follows.

θ_1: economic recession continues, small market for new products

θ_2: economic recession continues, large market for new products

θ_3: economic recovery, small market for new products

θ_4: economic recovery, large market for new products

By taking into account, for each state of nature, expected profits on the old and new products, together with redundancy costs, the management constructs a table of utilities as follows, where utility = expected profit in £ million.

	θ_1	θ_2	θ_3	θ_4
d_1	–5.0	–5.0	–2.0	–2.0
d_2	–2.5	–2.5	–0.5	–0.5
d_3	1.0	1.0	2.0	2.0
d_4	0.5	0.5	2.5	2.5
d_5	–4.5	1.0	–1.5	4.0

(i) Define what is meant by an *inadmissible strategy* and by a *maximin strategy*. Which of the five strategies are

(a) inadmissible,

(b) maximin?

(5)

(ii) Suppose that the prior probabilities for continued recession and for recovery are π_1 and $1-\pi_1$ respectively, that the prior probabilities for large and small demand are π_2 and $1-\pi_2$ respectively, and that the events 'continued recession' and 'large demand' are independent. Define the *Bayes strategy* and find the Bayes strategy if $\pi_1 = 0.8$ and $\pi_2 = 0.5$.

(5)

Question 8 continues on the next page

(iii) If π_2 is fixed at 0.5, for what values of π_1, if any, are d_3, d_4, d_5, respectively, Bayes strategies?

(3)

(iv) An economic consultancy firm offers to provide advice on the likely demand for the new product, for a fee of £50 000. If the firm advises that demand will be large, there is a probability of 0.95 that the advice is correct; if the advice is that demand will be small, there is a probability of 0.90 that it is correct. Assuming that $\pi_1 = 0.8$ and $\pi_2 = 0.5,$ as in part (ii), the posterior probabilities of θ_1, θ_2, θ_3, θ_4 when the advice is that there will be large demand are $\frac{0.02}{\phi}, \frac{0.38}{\phi}, \frac{0.005}{\phi}, \frac{0.095}{\phi}$ respectively, where ϕ is the probability that the advice is that demand will be large. When the advice is that there will be large demand, the Bayes strategy is d_5 with utility $\frac{0.6625}{\phi}$. For the case when the advice is that demand will be small, find the posterior probabilities of the 4 states of nature, and determine which strategy is Bayes in this case. [You need not evaluate ϕ.] Calculate whether the expected extra utility gained by using the consultancy firm exceeds their fee, and by how much.

(7)

BLANK PAGE

BLANK PAGE

BLANK PAGE