EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2011

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 7 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. Astronomers are interested in the distance μ in light years from earth to a certain star. A random sample of measurements of this distance, X_1, X_2, …, X_n, is available, where X_i has a Normal distribution with mean μ and known variance σ_i^2 (> 0), for $i = 1, 2, \ldots, n$.

(i) Show that $\sum_{i=1}^{n} \frac{X_i}{\sigma_i^2}$ is a sufficient statistic for μ.

(5)

(ii) Find $\hat{\mu}$, an unbiased estimator of μ based on the sufficient statistic given in part (i).

(2)

(iii) Find the distribution of $\hat{\mu}$.

(3)

(iv) Find the Cramér-Rao lower bound for the variance of unbiased estimators of μ and hence find the efficiency of $\hat{\mu}$.

(5)

(v) An alternative estimator $\tilde{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ has been proposed. Find the relative efficiency of $\tilde{\mu}$ compared to $\hat{\mu}$. By using the result $(\sum a_i b_i)^2 \le \sum a_i^2 \sum b_i^2$ for any values $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$, show that this relative efficiency cannot exceed one.

(5)

2. Independent random samples of failure times of two devices have been observed, $S_1, S_2, \ldots, S_m$ for Device 1 and $T_1, T_2, \ldots, T_n$ for Device 2. The probability density function of the failure times for Device 1 is $f(s) = \frac{1}{\alpha} e^{-s/\alpha}$ $(s > 0)$, while for Device 2 the probability density function is $g(t) = \frac{1}{\alpha + \beta} e^{-t/(\alpha+\beta)}$ $(t > 0)$, where α and β are unknown parameters ($\alpha > 0$, $\alpha + \beta > 0$).

(i) Using integration, or otherwise, show that the variance of the first of these distributions is α^2. Hence deduce the variance of the other distribution.

(5)

(ii) Find the maximum likelihood estimators of α and β. [You need only consider the first partial derivatives of the log likelihood.]

(6)

(iii) Determine the variance of the maximum likelihood estimator of β.

(3)

(iv) Suppose now that the total sample size, $m + n$, is fixed at N and that interest lies in estimating β as precisely as possible. If it can be assumed that α and β are approximately equal, find values that you would recommend for m and n, giving reasons for this choice.

(6)

3. (a) Explain what are meant by the *size* of a statistical test, a *test statistic* and a *confidence set*.

(4)

(b) Observations $X_1, X_2, \ldots, X_n$ constitute a random sample from a distribution with unknown parameter θ, and it is required to test the hypothesis $H_0: \theta = \theta_0$ against $H_1: \theta \neq \theta_0$. A test statistic $h(X_1, \ldots, X_n; \theta_0)$ is available which can be used to find a test for any θ_0. Show how a 95% confidence interval for θ can be constructed based on this test statistic.

(3)

(c) Waiting times $Y_1, Y_2, \ldots, Y_n$ constitute a random sample from the gamma distribution with probability density function

$$f(y) = \frac{y^k e^{-y/\lambda}}{k!\lambda^{k+1}} \quad (y > 0),$$

where k is a known non-negative integer and λ (> 0) is an unknown parameter. You are given that the moment generating function of this distribution is $(1-\lambda t)^{-k-1}$ and that the above distribution is χ^2_{2k+2} when $\lambda = 2$.

(i) Using moment generating functions, show that $W = \sum_{i=1}^{n} Y_i$ has a gamma distribution, with parameters that you should identify.

(3)

(ii) Show that $\frac{2W}{\lambda}$ is a pivotal quantity for λ.

(5)

(iii) Use your answer to part (ii) to find a 95% confidence interval for λ when $n = 5$, $\sum_{i=1}^{n} Y_i = 8.0$ and $k = 1$.

(5)

4. Let $X_1, X_2, \ldots, X_n$ be a random sample from a Normal distribution with unknown mean μ and known variance σ^2 (> 0).

(i) Use the Neyman-Pearson method to find the uniformly most powerful level α test of the null hypothesis $\mu = \mu_0$ against the alternative $\mu > \mu_0$, where μ_0 is known and $0 < \alpha < 1$.

(8)

(ii) Draw a graph of the operating characteristic of the test for the hypotheses in part (i) in the case $\mu_0 = 50$, $\alpha = 0.05$, $\sigma^2 = 4.0$ and $n = 10$. You should show the calculation of at least three points on your graph.

(6)

(iii) In the case $\mu_0 = 50$, $\alpha = 0.05$ and $\sigma^2 = 4.0$, calculate the minimum sample size required for the power at $\mu = 50.5$ to be at least 0.90.

(6)

5. (a) A random sample has been drawn from a distribution with unknown parameters α_1, α_2 and α_3 and it is required to test the null hypothesis $\alpha_1 = \alpha_2 = \alpha_3$ against the alternative hypothesis that the α_j are not all equal ($j = 1, 2, 3$). Describe how the test statistic, Λ, of the generalised likelihood ratio test is found and state the approximate distribution (assuming regularity conditions hold) of $-2\log\Lambda$ when the sample size is large.

(6)

(b) Mangoes are sold in packs of 3 in a chain of supermarkets. There is a constant probability θ that any mango will be unripe at the time of sale. As part of a quality control procedure, 200 packs of mangoes are inspected, and it is found that x_k packs contain k unripe mangoes for $k = 0, 1, 2$ and 3 (so that $x_0 + x_1 + x_2 + x_3 = 200$). You may assume that the numbers of unripe mangoes in different packs are independent.

(i) Find the form of the generalised likelihood ratio test of $\theta = \theta_0$ against $\theta \neq \theta_0$, where $0 < \theta_0 < 1$.

(8)

(ii) Carry out the generalised likelihood ratio test in the case $\theta_0 = 0.25$, $x_0 = 102$, $x_1 = 79$, $x_2 = 16$, $x_3 = 3$, giving the approximate p-value.

(6)

6. (a) A random sample $V_1, V_2, \ldots, V_n$ is available from a symmetric distribution with mean μ and it is required to test the null hypothesis $\mu = 20$ against the alternative $\mu < 20$. An experimenter wishes to devise a non-parametric test based on the test statistic $T = \dfrac{\text{sample mean} - 20}{\text{standard error of the sample mean}}$, making use of the fact that the distribution of $V_i - 20$ is symmetric about zero under the null hypothesis. [Note: the experimenter does not want to assume that T has a t distribution under the null hypothesis.] Find such a test and say how the exact p-value may be found when n is small.

(5)

(b) Independent random samples $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$, where $n_1 \leq n_2$, have been drawn from two distributions. An experimenter has decided to carry out a Wilcoxon rank sum test to see whether there is significant evidence that values from the second distribution tend to be larger.

(i) Explain how this test is carried out if there are no ties.

(5)

(ii) In the Royal Statistical Society tables, Table 10, 0.025 level, the critical value corresponding to $n_1 = 5$, $n_2 = 5$ is given as 17. Show how this value has been obtained.

(6)

(iii) Table 10 only gives the critical values for $5 \leq n_i \leq 20$, $i = 1, 2$. Find the approximate critical value at the 0.05 level corresponding to $n_1 = 10$, $n_2 = 25$ by making use of the asymptotic distribution of T, given at the bottom of Table 10.

(4)

7. Explain what is meant by a *conjugate family of distributions* in Bayesian inference.

(2)

The numbers of accidents in a mine for the last n years are $X_1, X_2, \ldots, X_n$, where accidents in different years are independent and X_i has the Poisson distribution, mean λ, for $i = 1, \ldots, n$. The prior distribution of λ is gamma, parameters k and ν, both known positive constants. [This gamma distribution has probability density function $f(y) = \frac{\nu^k y^{k-1} e^{-\nu y}}{\Gamma(k)}$ for $y > 0$, mean $\frac{k}{\nu}$ and variance $\frac{k}{\nu^2}$.]

(i) Find the posterior distribution of λ.

(5)

(ii) Show how a Normal approximation can be used to find an approximate 95% Bayesian interval estimate for λ.

(3)

(iii) Let X_{n+1} be the number of accidents in the mine next year. Find the predictive distribution of X_{n+1}.

(6)

(iv) Suppose now that a different prior distribution has been judged appropriate, and for this prior there is no simple formula for the posterior distribution of λ. However, a computer program is available to simulate values $L_1, L_2, L_3, \ldots$ which behave as if they are a random sample from the posterior distribution. Explain how these simulated values may be used to find an approximate 95% Bayesian interval for λ and the predictive distribution of X_{n+1}.

(4)

8. Explain the rationale of *likelihood-based inference* and describe the useful properties that estimators and tests based on the likelihood function possess. Under what circumstances does likelihood-based inference yield similar results to Bayesian and frequentist inference?

(20)