EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2009

(Modular format)

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 9 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. The weights of product delivered by a packaging machine on successive occasions are $W_1, W_2, \ldots, W_n$. Here W_i has the $N(\mu, \sigma^2)$ distribution for i = 1, 2, …, n and $\text{corr}(W_i, W_{i+1}) = \rho$ for i = 1, 2, …, n – 1, and μ, σ^2 and ρ are unknown parameters with $\sigma^2 > 0$ and $-1 < \rho < 1$.

(i) Obtain the expected values of $\sum_{i=1}^{n} W_i$ and $\sum_{i=1}^{n} W_i^2$ and show that the expected value of $\sum_{i=1}^{n-1} W_i W_{i+1}$ is $(n-1)(\rho\sigma^2 + \mu^2)$.

(5)

(ii) Use the results of (i) to find estimators of μ, σ^2 and ρ based on the above three statistics.

(5)

For the remainder of this question let n = 2.

(iii) Find the bias of the estimator of σ^2.

(4)

(iv) Find the variance of the estimator of μ.

(3)

(v) Evaluate the estimator of ρ. Comment on this estimate.

(3)

2. The number N of species of insect caught in a trap during one night in a certain region is modelled by a distribution of the form

$$P(N = n) = \frac{p^n}{-n \log(1-p)}$$

for $n = 1, 2, 3, \ldots$, where the unknown parameter p must lie between 0 and 1. Forty independent observations $N_1, N_2, \ldots, N_{40}$ are made.

(i) Show that the mean of this distribution is $E(N) = -p\left[(1-p)\log(1-p)\right]^{-1}$.

(3)

(ii) Find an equation that determines the maximum likelihood estimator, $\hat{p}$, of p. (Do not attempt to solve this equation.)

(7)

(iii) The second derivative of the log likelihood is given by

$$-\frac{\sum N_i}{p^2} + \frac{40(1+\log(1-p))}{\left[(1-p)\log(1-p)\right]^2}.$$

Derive the Fisher information and hence find an approximate 95% confidence interval for p, assuming that the maximum likelihood estimator is asymptotically efficient. Evaluate this confidence interval for the case $\hat{p} = 0.8$.

(6)

(iv) Suppose now that $\sum N_i = 100$. Describe an iterative method for finding the maximum likelihood estimate. Demonstrate one iteration of this method, using a starting value of $\hat{p} = 0.75$.

(4)

3. What is meant by a *95% confidence interval* for a parameter θ?

(2)

Let $X_1, X_2, \ldots, X_n$ be independent observations from a distribution with probability density function

$$f(x) = \frac{x^{k-1}e^{-x/\alpha}}{(k-1)!\alpha^k} \qquad (x > 0)$$

and moment generating function $m(t) = (1-\alpha t)^{-k}$ for $t < \alpha^{-1}$, where k is a known positive integer and α (> 0) is an unknown parameter. (Note: in the special case $\alpha = 2$, this distribution is the chi-squared distribution with 2k degrees of freedom.)

(i) Show that $Y = \sum_{i=1}^{n} X_i$ is a sufficient statistic for α.

(4)

(ii) Show that $2Y/\alpha$ has the χ^2_{2nk} distribution.

(5)

(iii) Using your answer to (ii), describe how standard tables could be used to find a 95% confidence interval for α.

(4)

(iv) Hence find a formula for a 95% confidence interval for α when n = 10 and k = 3. Find the expected length of this interval.

(5)

Turn over

4. Explain the *Neyman-Pearson* approach to hypothesis testing when the null and alternative hypotheses are simple.

(4)

Independent observations $X_1, X_2, \ldots, X_n$ are taken from Poisson distributions with means $\mu(1+a_i\theta)$ for $i = 1, 2, \ldots, n$, where $a_1, a_2, \ldots, a_n$ are known positive constants and μ (> 0) and θ (≥ 0) are unknown parameters.

(i) Show that the critical region of the most powerful test of the null hypothesis $\mu = 10, \theta = 0$ against the alternative hypothesis $\mu = 10, \theta = 1$ is of the form $\sum x_i \log(1+a_i) \geq c$, for some constant c.

(5)

(ii) Using a Normal approximation, find c to give an approximate 5% significance level.

(5)

(iii) Say, with reasons, whether a uniformly most powerful test exists in the following cases; in each case, give the form of the most powerful test if one exists.

(a) Null hypothesis $\mu = 10, \theta = 0$, alternative hypothesis $\mu > 10, \theta = 0$.

(3)

(b) Null hypothesis $\mu = 10, \theta = 0$, alternative hypothesis $\mu = 10, \theta > 0$.

(3)

5. (i) The random variables $X_1, X_2, \ldots$ are independent observations of a distribution with parameter θ, for which only positive values are possible. $\hat{\theta}_n$ denotes an estimator of θ based on the n variables $X_1, X_2, \ldots, X_n$, and the sequence of estimators $\hat{\theta}_n$ for $n = 1, 2, \ldots$ satisfies $E(\hat{\theta}_n) = \theta + \frac{k}{n}$, where k is a constant not depending on n. Describe the *jack-knife* estimator of θ based on $\hat{\theta}_n$ and show that it is an unbiased estimator of θ.

(5)

(ii) An estimator of the coefficient of variation of a random sample $X_1, X_2, \ldots, X_n$ is

$$\hat{c} = \frac{S}{\bar{X}}$$

where $\bar{X} = \frac{T}{n}$, $S^2 = \frac{U - \frac{T^2}{n}}{n-1}$, $T = \sum_{i=1}^{n} X_i$ and $U = \sum_{i=1}^{n} X_i^2$.

(a) Show that the jack-knife estimator of the coefficient of variation based on this estimator is given by

$$\tilde{c} = \frac{n^2}{\sqrt{n-1}}\sqrt{\frac{U}{T^2} - \frac{1}{n}} - \frac{(n-1)^2}{n\sqrt{n-2}}\sum_{i=1}^{n}\sqrt{\frac{U - X_i^2}{(T - X_i)^2} - \frac{1}{n-1}}.$$

(7)

(b) Describe how an approximate jack-knife 95% confidence interval for the coefficient of variation can be found.

(4)

(c) Describe how to find an approximate 95% confidence interval for the coefficient of variation using a *bootstrap* method and the estimator $\hat{c}$.

(4)

Turn over

6. The random variable Y is the percentage change in price of a commodity between consecutive trading days and has the double exponential distribution, with probability density

$$f(y) = \frac{1}{2}\alpha e^{-\alpha|y|} \quad \text{for } -\infty < y < \infty,$$

where α (> 0) is an unknown parameter. Independent observations $Y_1, Y_2, \ldots, Y_m$ have been made of this variable.

(i) Find the form of the generalised likelihood ratio test of the null hypothesis $\alpha = 2$ against the alternative $\alpha \neq 2$.

(7)

(ii) Explain how the approximate critical value of this test at the 5% significance level can be found when *m* is large.

(4)

The daily percentage change, W, of another commodity also has the double exponential distribution, but with unknown parameter β (> 0). Independent observations $W_1, W_2, \ldots, W_n$ are available for this variable.

(iii) Find the form of the generalised likelihood ratio test of the null hypothesis $\alpha = \beta$ against the alternative $\alpha \neq \beta$.

(6)

(iv) Carry out this test at approximately the 5% level when $m = 100$, $\sum |y_i| = 40$, $n = 200$ and $\sum |w_i| = 100$.

(3)

7. [Note the information about the beta and Dirichlet distributions given at the end of this question.]

An election is shortly to take place, with two candidates, 1 and 2. A proportion p_1 $(0 < p_1 < 1)$ of the electorate currently supports Candidate 1, a further proportion p_2 $(0 < p_2 < 1 - p_1)$ supports Candidate 2, and a proportion $p_3 = 1 - p_1 - p_2$ is undecided. In a random sample of *n* electors, x_1 support Candidate 1, x_2 support Candidate 2 and x_3 voters are undecided (so that $x_1 + x_2 + x_3 = n$).

(a) The prior probability density function of p_1 is

$$\pi(p_1) = 6p_1(1 - p_1) \qquad (\text{for } 0 < p_1 < 1).$$

(i) Show that the posterior probability density function of p_1 is proportional to $p_1^{x_1+1}(1 - p_1)^{n-x_1+1}$ (for $0 < p_1 < 1$). (5)

(ii) Find an approximate 95% Bayesian interval for p_1. (You may assume that *n* is large.) (6)

(b) The joint prior distribution of (p_1, p_2, p_3) is Dirichlet with parameters (2, 1, 1).

(i) Find the posterior joint distribution of (p_1, p_2, p_3). (4)

(ii) Assuming that the posterior distribution of $p_1 - p_2$ is approximately Normal, find an approximate 95% Bayesian interval for $p_1 - p_2$. (5)

[The beta distribution with parameters (α_1, α_2) has probability density proportional to

$$p_1^{\alpha_1 - 1}(1 - p_1)^{\alpha_2 - 1} \qquad (0 \le p_1 \le 1)$$

and has $E(p_1) = \dfrac{\alpha_1}{\alpha_1 + \alpha_2}$ and $\text{Var}(p_1) = \dfrac{\alpha_1\alpha_2}{(\alpha_1 + \alpha_2 + 1)(\alpha_1 + \alpha_2)^2}$.

The Dirichlet distribution with parameters $(\alpha_1, \alpha_2, \alpha_3)$ has probability density proportional to

$$p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} p_3^{\alpha_3 - 1} \qquad (\text{where } p_1 + p_2 + p_3 = 1,\ 0 \le p_i \le 1 \text{ for each } i),$$

and has $E(p_i) = \dfrac{\alpha_i}{\alpha_0}$, $\text{Var}(p_i) = \dfrac{\alpha_i(\alpha_0 - \alpha_i)}{\alpha_0^2(\alpha_0 + 1)}$ and $\text{Cov}(p_i, p_j) = -\dfrac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0 + 1)}$ for $i, j = 1, 2, 3$, $i \ne j$, where $\alpha_0 = \alpha_1 + \alpha_2 + \alpha_3$.]

8. Hypothesis testing is sometimes described as a method of decision making and sometimes as a method of measuring the strength of the evidence against a hypothesis. Explain and compare these two viewpoints, citing applications where each seems to be the more appropriate. Describe the factors affecting the choice of significance level and sample size in each case.

(20)