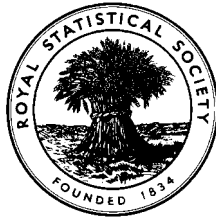# EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

(formerly the Examinations of the Institute of Statisticians)

GRADUATE DIPLOMA IN STATISTICS, 1997

Statistical Theory and Methods II

Time Allowed: Three Hours

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

All questions carry equal marks.

Graph paper and Official tables are provided.

Candidates may use silent, cordless, non-programmable electronic calculators.

*Where a calculator is used the **method** of calculation should be stated in full.*

Note that $\begin{pmatrix} n \\ r \end{pmatrix}$ *is the same as* $^{n}C_{r}$ *and that* $\ln$ *stands for* $\log_{e}$.

1. Explain what is meant by a *maximum likelihood estimator* and give the large - sample properties of this type of estimator.

   Let $X_1, X_2, ..., X_n$ be a random sample from a distribution with probability density function

   $$f(x) = \theta^{-1} x^{\frac{1-\theta}{\theta}}, \quad 0 < x < 1,$$

   where $\theta > 0$ is an unknown parameter.

   (i)   Show that the maximum likelihood estimator of θ is $\hat{\theta} = -\sum \ln(X_i)/n$.
   (ii) Show that $Y = -\ln(X)$ has an exponential distribution with mean θ.
   (iii) Verify that $\hat{\theta}$ is an unbiased estimator of θ and find its variance.
   (iv) Find the Cramér-Rao lower bound for the variance of unbiased estimators of θ, and deduce the efficiency of $\hat{\theta}$.

2. A botanist has the theory that certain seedlings fall into one of four categories, 1, 2, 3 or 4, with respective probabilities $(2+\theta)/4$, $(1-\theta)/4$, $(1-\theta)/4$ and $\theta/4$, where $0<\theta<1$ is an unknown parameter. In a random sample of n seedlings, the number of seedlings in category i is N_i for $i = 1,2,3,4$.

   (i)   Find the maximum likelihood estimator of θ.

   (ii) Assuming the asymptotic efficiency of the maximum likelihood estimator, find the approximate distribution of this estimator for large samples.

   For a random sample of 8404 seedlings, the frequencies of categories 1,2,3 and 4 are 4299, 1960, 2062 and 83, respectively. Calculate the maximum likelihood estimator of θ and find an approximate 95% confidence interval for θ.

3. Explain what is meant by a *sequential test*. Describe the advantages and disadvantages of sequential tests compared to fixed - sample size tests.

A manufacturing firm produces a chemical in large batches. Each batch is to be tested for excessive impurity using a sequential probability ratio test in which samples of one gram of the chemical are taken one by one, analysed, and the batch eventually either accepted or rejected. The amount of impurity, in mg per gram, is known from past experience to be normally distributed with a standard deviation of 0.51. The manufacturer considers a mean level of 1.0 mg/g to be acceptable and a mean level exceeding 1.7 mg/g to be unacceptable. The Type I and II errors should both be close to 0.05.

(i) Construct a sequential probability ratio test with approximately these error probabilities.

(ii) Show how a graph may be used to help carry out the test.

(iii) Find the approximate expected sample size when the true mean impurity level is 1.0 mg/g.

4. Describe the Neyman - Pearson approach to testing one simple hypothesis against another simple hypothesis. What is the optimal property that such Neyman-Pearson tests possess?

Suppose that $X_1, X_2, ..., X_n$ are independent random variables such that X_i has a normal distribution with mean θ_i and variance 1. It is required to test the null hypothesis that each θ_i is zero against the alternative hypothesis that $\theta_i = \frac{1}{2}$ for $i = 1, 2, ..., r$ and $\theta_i = -\frac{1}{2}$ for $i = r + 1, r + 2, ..., n$.

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

(ii) Find the most powerful test with size 0.05.

(iii) Evaluate the power of the test found in part (ii), and deduce how large n must be to ensure that the power is at least 0.9.

5. Explain what is meant by a *confidence region* in Bayesian inference.

The respective probabilities of recovering from a certain medical condition using two treatments are p_1 and p_2. The prior distributions of p_1 and p_2 are Beta (1,1) and Beta (2,1), respectively. Each treatment is tested on two patients and it is found that one of the patients using Treatment 1 recovers, while both patients using Treatment 2 recover.

(i) Find the posterior distributions of p_1 and p_2.

(ii) Show that the shortest 95% Bayesian confidence interval for p_1 is approximately (0.0943, 0.9057), and find the shortest 95% Bayesian confidence interval for p_2.

(iii) Find the posterior probability that Treatment 2 is better than Treatment 1.

[The beta distribution with parameters $\alpha > 0$ and $\beta > 0$ has probability density function

$$ f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} \, x^{\alpha-1} (1-x)^{\beta-1}, \; 0 < x < 1, $$

where $\Gamma(\cdot)$ denotes the gamma function.]

6. Give an account of the decision theory approach to statistical inference. In your answer, distinguish between its rôle in estimation and hypothesis testing problems, and also show that Neyman-Pearson tests may be interpreted as Bayes solutions to certain decision problems.

7. Describe the *generalised likelihood ratio test* and explain how the critical region of the test may be found for large samples.

   Suppose that the times to relief of two brands of aspirin, 1 and 2, are being compared in a small clinical trial. The difference between the true time to relief and the advertised time to relief is called the "error" of the time to relief. A reasonable model for the errors for aspirins of brand i, $i = 1, 2$, is the distribution with probability density function

   $$f(x) = \frac{1}{2} \theta_i \, e^{-\theta_i |x|}, \quad -\infty < x < \infty,$$

   where $\theta_1, \theta_2 > 0$ are unknown parameters.

   In the trial, aspirins of the two brands were allocated at random to 160 patients. Of these, 70 received brand 1, yielding $\sum |x_j| = 300$, and 90 received brand 2, yielding $\sum |x_j| = 250$. Use a generalised likelihood ratio test to test the null hypothesis that $\theta_1 = \theta_2$ against the alternative that $\theta_1 \neq \theta_2$. Carry out this test at approximately the 1% level.

8. Let $X_1, X_2, ..., X_n$ be a random sample from an exponential distribution with unknown mean $\theta > 0$.

   (i) Explain what is meant by a 90% *confidence interval* for θ.

   (ii) (a) State, with reasons, whether the sample mean, $\overline{X}$, is a consistent estimator of θ.

   (b) State the distribution of $Y = \sum_{i=1}^{n} X_i$ and say how this distribution is related to a chi-squared distribution.

   (c) Explain carefully why $2Y / \theta$ is a pivotal quantity.

   (d) Find a 90% confidence interval for θ when $n = 7$ and $\overline{X} = 93.6$.

4