UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday 29 May 2008 1.30 to 4.30

PAPER 42

STATISTICAL THEORY

*Attempt no more than **FOUR** questions.*

*There are **SIX** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Outline the *Fisherian approach* to inference in a parametric statistical model, including a description of *ancillary statistics*, both when there is no nuisance parameter and when a nuisance parameter is present.

Give an example of a model where there is a problem of non-uniqueness of ancillary statistics, stating the distributions of both ancillary statistics.

Let $f_0(y)$ denote a known probability density function, and let $Y_1, \ldots, Y_n$ be independent with density $f(y; \mu, \sigma) = \frac{1}{\sigma} f_0\big((y-\mu)/\sigma\big)$ for some $\mu \in \mathbb{R}$ and $\sigma > 0$. If $\hat{\mu} \equiv \bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ and $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$, show that

$$A = \left(\frac{Y_1 - \hat{\mu}}{\hat{\sigma}}, \ldots, \frac{Y_n - \hat{\mu}}{\hat{\sigma}} \right)$$

is ancillary. What distribution would be used for inference about (μ, σ) in the Fisherian approach?

2 Let F and G be non-degenerate distribution functions. In the context of extreme value theory for maxima, what does it mean for F to belong to the domain of attraction of G? Now suppose G is continuous, and suppose that $a_n > 0$ and b_n are such that $F^n(a_n x + b_n) \to G(x)$ for all $x \in \mathbb{R}$, and also that $\alpha_n > 0$ and β_n are such that $\frac{\alpha_n}{a_n} \to a > 0$ and $\frac{\beta_n - b_n}{a_n} \to b$. Use Slutsky's theorem to prove that $F^n(\alpha_n x + \beta_n) \to G(ax + b)$ for all $x \in \mathbb{R}$.

By giving appropriate definitions and quoting any required results carefully, identify a non-degenerate distribution function G such that F belongs to the domain of attraction of G, in both cases below:

(i) $F(x) = \big(1 - \frac{C}{x^2 \log x \log\log x}\big) \mathbb{1}_{\{x \geqslant 10\}}$, where $C > 0$ is a normalisation constant

(ii) $F(x) = (1 - e^{-x^{1/2}}) \mathbb{1}_{\{x > 0\}}$.

In case (ii) above, find constants $\alpha_n > 0$ and β_n, in terms of standard elementary functions, such that $F^n(\alpha_n x + \beta_n) \to G(x)$ for all $x \in \mathbb{R}$.

[*You may assume that if $\ell(x)$ is a continuous, slowly varying function, then for each $r > 0$,*

$$\int_n^\infty \frac{1}{x^{r+1}} \ell(x)\, dx = \frac{1}{rn^r} \ell(n) \{1 + o(1)\}$$

as $n \to \infty$.]

3 Let (Y_n) be a sequence of independent $N(0,1)$ random variables. Fix $\rho \in (0,1)$, and define a sequence (X_n), by $X_1 = Y_1$ and

$$X_j = \rho X_{j-1} + (1-\rho^2)^{1/2} Y_j, \quad \text{for } j \geqslant 2.$$

By first writing X_j in terms of $Y_1, \ldots, Y_j$, find the distribution of the vector $(X_1, \ldots, X_n)^T$.

It is decided to estimate the marginal density $f(x)$ of X_1 using a kernel density estimator $\hat{f}_h(x) = \frac{1}{nh}\sum_{j=1}^n K\Big(\frac{x-X_j}{h}\Big)$. Let $\psi(t) = \mathbb{E}(e^{itX_1})$, let $\psi_K(t) = \int_{-\infty}^{\infty} e^{itx} K(x)\,dx$, and let Re z denote the real part of a complex number z. Furthermore, let $\hat{f}_h^*(x) = \frac{1}{nh}\sum_{j=1}^n K\Big(\frac{x-X_j^*}{h}\Big)$, where $X_1^*, \ldots, X_n^*$ are independent with density $f(x)$. For data with this type of dependence, it may be shown that

$$\int_{-\infty}^{\infty} \mathrm{Var}\{\hat{f}_h(x)\}\,dx = \int_{-\infty}^{\infty} \mathrm{Var}\{\hat{f}_h^*(x)\}\,dx + \frac{1}{\pi n}\sum_{j=1}^{n-1}\Big(1-\frac{j}{n}\Big)g(j),$$

where

$$g(j) = \int_{-\infty}^{\infty} |\psi_K(ht)|^2 \big[\mathrm{Re}\ \mathbb{E}\{e^{it(X_1-X_{j+1})}\} - |\psi(t)|^2\big]\,dt.$$

Suppose that $K(x)$ is the standard normal density. Evaluate $g(j)$.

[*Hint: Recall that if* $Z \sim N(\mu,\sigma^2)$, *then* $\mathbb{E}(e^{itZ}) = e^{it\mu - t^2\sigma^2/2}$.]

Hence show that if $h = h_n \to 0$ as $n \to \infty$ but $nh \to \infty$ as $n \to \infty$, then

$$\int_{-\infty}^{\infty} \mathrm{Var}\{\hat{f}_h(x)\}\,dx = \int_{-\infty}^{\infty} \mathrm{Var}\{\hat{f}_h^*(x)\}\,dx + O(n^{-1})$$

as $n \to \infty$. What does this result imply about the bandwidth that minimises the asymptotic mean integrated squared error, and the optimal rate of convergence of the mean integrated squared error to zero?

4 By defining appropriate notation and stating your assumptions, explain what is meant by an Edgeworth expansion for the density of a standardised sum S_n^* of independent, identically distributed random variables $Y_1, \ldots, Y_n$.

[*An explicit expression for the* $O(n^{-1})$ *term is not required.*]

State the corresponding expansion for the distribution function $F_{S_n^*}$ of S_n^*. Fix $\alpha \in (0,1)$, and let y_α and z_α satisfy $F_{S_n^*}(y_\alpha) = \alpha$ and $\Phi(z_\alpha) = \alpha$ respectively, where Φ is the standard normal distribution function. Assuming that $y_\alpha = p_0(z_\alpha) + p_1(z_\alpha)n^{-1/2} + O(n^{-1})$ as $n \to \infty$, find explicit expressions for the functions p_0 and p_1.

Now suppose that $Y_1, \ldots, Y_n$ are independent with density

$$f(y;\theta) = \frac{e^{y-\theta}}{(1+e^{y-\theta})^2}, \quad y \in \mathbb{R},\ \theta \in \mathbb{R}.$$

Show that

$$\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - n^{-1/2}\frac{\pi}{\sqrt{3}} z_{1-\alpha/2}\ ,\ \frac{1}{n}\sum_{i=1}^{n} Y_i - n^{-1/2}\frac{\pi}{\sqrt{3}} z_{\alpha/2} \right)$$

is a confidence interval for θ of asymptotic $(1-\alpha)$-level coverage.

[*Hint: You may use the fact that* $\int_0^\infty \frac{y^2 e^y}{(1+e^y)^2}\,dy = \pi^2/6$.]

Show further that the coverage error of this confidence interval is $O(n^{-1})$ as $n \to \infty$.

5 Explain what is meant by an *exponential dispersion family of order 1*, the *variance function* and a *generalised linear model.* Define the canonical link function, and state one advantage of its use.

Give three examples of exponential dispersion families of order 1, showing explicitly that they satisfy the definition, and identifying the canonical link function in each case.

Explain how, in general, you would test the hypothesis that one of the components of the parameter of interest was zero.

6 Outline briefly the methods of *local polynomial kernel* estimation and *natural cubic spline* estimation of a regression function in nonparametric regression.

[*You may restrict attention to the homoscedastic, fixed design case. Explicit formulae for the estimators are not required, but you should state carefully the minimisation problem being solved in each case.*]

The following model occurs in the oil industry. Let $n \geqslant 3$ and let $a < x_1 < \ldots < x_n < b$ be known. Suppose that $Y_1, \ldots, Y_{n-1}$ are independent, with

$$Y_i = \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} g(x)\, dx + \epsilon_i, \quad i = 1, \ldots, n-1,$$

where $\epsilon_i \sim N(0, \sigma^2)$ for some known $\sigma > 0$. Fixing $\lambda > 0$, we wish to estimate the regression function $g(x)$ by minimising

$$S_\lambda(\tilde{g}) = \sum_{i=1}^{n} \left\{ Y_i - \frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} \tilde{g}(x)\, dx \right\}^2 + \lambda \int_a^b \tilde{g}'(x)^2\, dx$$

over $\tilde{g} \in S_1[a, b]$, the set of real-valued functions on $[a, b]$ having one continuous derivative. Prove that any minimiser of $S_\lambda(\tilde{g})$ over $\tilde{g} \in S_1[a, b]$ must be a quadratic spline with knots at $x_1, \ldots, x_n$ that is constant on $[a, x_1]$ and $[x_n, b]$.

[*You may assume that, given any $y = (y_1, \ldots, y_{n-1}) \in \mathbb{R}^{n-1}$, there exists a unique quadratic spline $g : [a, b] \to \mathbb{R}$ with knots at $x_1, \ldots, x_n$ that is constant on $[a, x_1]$ and $[x_n, b]$ and satisfies $\frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} g(x)\, dx = y_i$, for $i = 1, \ldots, n-1$.*]

END OF PAPER