THE ROYAL STATISTICAL SOCIETY

2001 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

STATISTICAL THEORY AND METHODS

PAPER II

Graduate Diploma, Statistical Theory & Methods, Paper II, 2001. Question 1

(i) The median, η, is such that $P(X \le \eta) = 0.5$. Suppose we are to test the null hypothesis $H_0 : \eta = \eta_0$ against $H_1 : \eta \ne \eta_0$, given a random sample of observations $X_1, X_2, \ldots X_n$. Let T be the number of X_i that are $\le \eta_0$. Under H_0, T is B(n, ½), and in this case n = 30. The sign test rejects H_0 if $\left|T - \frac{1}{2}n\right| \ge k$, where k is the smallest value such that $P\left(T \le \frac{1}{2}n - k\right) \le \frac{\alpha}{2}$ on H_0, and α is the significance level. If n is not too small, T has on H_0 an approximately Normal distribution $N\left(\frac{1}{2}n, \frac{1}{4}n\right)$.

So $P\left(T \le \frac{n}{2} - k\right) \approx \Phi\left\{\dfrac{\frac{n}{2} - k + \frac{1}{2} - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right\} = \Phi\left(\dfrac{1-2k}{\sqrt{n}}\right)$ where Φ is the cdf of N(0,1). It is now easy to find k. If required, a one-tail test can be constructed in the same way.

(ii) For n = 30 and α = 0.05, we require $\Phi\left(\dfrac{1-2k}{\sqrt{30}}\right) \le 0.025$ so that $\dfrac{1-2k}{\sqrt{30}} \le -1.96$, and therefore $k \ge 5.9$, i.e. use k = 6 since it must be an integer. Take T as the number of observations ≤ 10, and reject H_0 if $T \le 9$ or $T \ge 21$.

(iii) A non-parametric confidence interval is a confidence interval that requires few assumptions about the form of the distribution function. It is a random interval with a specified probability of including a parameter of the distribution, valid for any value the parameter can take.

(iv) The required interval should contain all those values of η which would not be rejected by a test at the 5% level. Consider testing $H_0 : \eta = \eta_0$ against $H_1 : \eta \ne \eta_0$. Let T_{η_0} be the number of observations less than or equal to η_0. Then H_0 is not rejected if $\left|T_{\eta_0} - 15\right| < 5.9$, i.e. $10 \le T_{\eta_0} \le 20$, equivalent to $X_{(10)} \le \eta_0 < X_{(21)}$ as given.

<u>Graduate Diploma, Statistical Theory & Methods, Paper II, 2001. Question 2</u>

The exponential parameter is $1/\nu$; $E[X]=\nu$, $\text{Var}(X)=\nu^2$.

(i) $E[\hat{\nu}]=\frac{1}{n}\sum_{i=1}^{n}E[X_i]=\frac{1}{n}n\nu=\nu$ for all ν.

Also $\text{Var}(\hat{\nu})=\frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(X_i)=\frac{1}{n^2}n\nu^2=\frac{\nu^2}{n}$.

(ii) The likelihood $L=\left(\frac{1}{\nu}\right)^n\prod_{i=1}^{n}e^{-x_i/\nu}$ and the log of this is $l(\nu)=-n\ln\nu-\frac{1}{\nu}\sum_{i=1}^{n}x_i$.

$\frac{dl}{d\nu}=-\frac{n}{\nu}+\frac{1}{\nu^2}\sum x_i$ and $\frac{d^2l}{d\nu^2}=\frac{n}{\nu^2}-\frac{2}{\nu^3}\sum x_i$.

$$I(\nu)=E\left[-\frac{d^2l}{d\nu^2}\right]=-\frac{n}{\nu^2}+\frac{2}{\nu^3}E\left[\sum x_i\right]=-\frac{n}{\nu^2}+\frac{2}{\nu^3}n\nu=\frac{n}{\nu^2}.$$

Hence the Cramér-Rao lower bound is $\frac{\nu^2}{n}$, so that $\hat{\nu}$ is efficient.

(iii) $P(Y>y)\equiv P(\min X_i>y)=P(X_1,X_2,...,X_n>y)=\prod_{i=1}^{n}P(X_i>y)$ by independence of $\{X_i\}$. And $P(X>y)=\int_y^\infty\frac{1}{\nu}e^{-x/\nu}dx=\left[-e^{-x/\nu}\right]_y^\infty=e^{-y/\nu}$ for $y\geq 0$.

So $P(Y>y)=e^{-ny/\nu}$, $y\geq 0$.

Hence $P(Y\leq y)\equiv F(y)=1-e^{-ny/\nu}$, and by differentiation $f(y)=\frac{n}{\nu}e^{-ny/\nu}$, $y\geq 0$.

This is exponential with parameter $\frac{n}{\nu}$ so $E[Y]=\frac{\nu}{n}$, $\text{Var}(Y)=\frac{\nu^2}{n^2}$.

(iv) $\tilde{\nu}=nY$ is clearly unbiased for ν. $\text{Var}(\tilde{\nu})=n^2\text{Var}(Y)=\nu^2$, for any *n*.

Its efficiency relative to $\hat{\nu}$ is $\frac{\text{Var}(\hat{\nu})}{\text{Var}(\tilde{\nu})}=\frac{1}{n}$.

Since $\text{Var}(\hat{\nu})\to 0$ as $n\to\infty$, $\hat{\nu}$ is consistent. ($\tilde{\nu}$ is not.)

Graduate Diploma, Statistical Theory & Methods, Paper II, 2001. Question 3

(i) The likelihood function is proportional to $\theta^{2x}\{2\theta(1-\theta)\}^{y}(1-\theta)^{2z}$;
and so $l = \ln L$ is $l=(2x+y)\ln\theta+(y+2z)\ln(1-\theta) \qquad (0<\theta<1)$.

$$\frac{dl}{d\theta}=\frac{2x+y}{\theta}-\frac{y+2z}{1-\theta} \quad \text{and} \quad \frac{d^2l}{d\theta^2}=-\frac{(2x+y)}{\theta^2}-\frac{(y+2z)}{(1-\theta)^2}<0$$

so the solution of $\frac{dl}{d\theta}=0$ will be a maximum.

$$\therefore \frac{2x+y}{\hat{\theta}}=\frac{y+2z}{1-\hat{\theta}} \quad \text{i.e.} \quad 2x+y=(2x+2y+2z)\hat{\theta} \quad \text{or} \quad \hat{\theta}=\frac{2x+y}{2n}.$$

(ii) $H_0: \theta=\theta_0$, $H_1: \theta=\theta_1<\theta_0$. Critical region size α. Reject H_0 if $t \le k$ where k satisfies $P(T\le k \mid \theta=\theta_0)\approx\alpha$; i.e. $\sum_{t=0}^{k}\binom{2n}{t}\theta_0^t(1-\theta_0)^{2n-t}\approx\alpha$.

(iii) The likelihood ratio for testing H_0 against H_1 is

$$\lambda=\frac{L(\theta_0)}{L(\theta_1)}=\left(\frac{\theta_0}{\theta_1}\right)^{2x+y}\left(\frac{1-\theta_0}{1-\theta_1}\right)^{y+2z}=\left(\frac{1-\theta_0}{1-\theta_1}\right)^{2n}\left(\frac{\theta_0(1-\theta_1)}{\theta_1(1-\theta_0)}\right)^{t}.$$

Since $\theta_0(1-\theta_1)>\theta_1(1-\theta_0)$, λ increases with t. The Neyman-Pearson lemma then shows that the most powerful test of size α rejects H_0 if $t\le k$ with k as in (ii).

(iv) Using the Central Limit Theorem, $T \sim$ approx $\mathrm{N}(2n\theta,\ 2n\theta(1-\theta))$. Therefore

$$\Phi\left(\frac{k+\frac{1}{2}-2n\theta_0}{\sqrt{2n\theta_0(1-\theta_0)}}\right)=0.05; \quad \therefore \frac{k+\frac{1}{2}-2n\theta_0}{\sqrt{2n\theta_0(1-\theta_0)}}=-1.645 \quad \text{or} \quad \frac{k+0.5-0.8n}{\sqrt{0.48n}}=-1.645.$$

Likewise
$$\Phi\left(\frac{k+\frac{1}{2}-2n\theta_1}{\sqrt{2n\theta_1(1-\theta_1)}}\right)=0.9, \quad \text{so} \quad \frac{k+0.5-0.6n}{\sqrt{0.42n}}=1.282.$$

From these results, $k+0.5=0.8n-1.645\sqrt{0.48n}=0.6n+1.282\sqrt{0.42n}$

which gives $0.2\sqrt{n}=1.645\sqrt{0.48}+1.282\sqrt{0.42}$

$$\text{or} \quad n=\frac{1}{0.04}\left(1.645\sqrt{0.48}+1.282\sqrt{0.42}\right)^2=97.1.$$

Since n must be an integer, take $n = 98$.

Graduate Diploma, Statistical Theory & Methods, Paper II, 2001. Question 4

A decision rule is a function from the sample space to the action space. The risk of a decision rule δ at parameter value θ, $R_\delta(\theta)$, is the expected loss. A decision rule δ is dominated by a decision rule δ^* if $R_{\delta^*}(\theta) \le R_\delta(\theta)$ for all θ in the parameter space Θ. An admissible decision rule is one that is not dominated by any other decision rule.

(i) $\theta \sim \mathrm{N}(u_0, v_0)$ and $L \propto \prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\theta)^2\right\}$

$$= \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}, \quad -\infty < \theta < \infty .$$

Hence the posterior distribution of θ is

$$\pi(\theta \mid \underline{X}) \propto \exp\left\{-\frac{1}{2v_0}(\theta-u_0)^2\right\}.\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\theta)^2\right\}$$

$$= \exp\left\{-\frac{1}{2}\left[\left(\frac{1}{v_0}+\frac{n}{\sigma^2}\right)\theta^2 - 2\theta\left(\frac{u_0}{v_0}+\frac{n\bar{x}}{\sigma^2}\right)+\left(\frac{u_0^2}{v_0}+\frac{\sum x_i^2}{\sigma^2}\right)\right]\right\}$$

$$\propto \exp\left\{-\frac{1}{2v_n}(\theta-u_n)^2\right\}, \quad -\infty < \theta < \infty ,$$

where $\frac{1}{v_n} = \frac{1}{v_0} + \frac{n}{\sigma^2}$ and $\frac{u_n}{v_n} = \frac{u_0}{v_0} + \frac{n\bar{x}}{\sigma^2}$.

Thus $\theta \mid X \sim \mathrm{N}(u_{n,} v_n)$.

(ii) Using quadratic loss, the Bayes estimator of θ is $E[\theta]$ in the posterior distribution. This is u_n.

$$u_n = \frac{\frac{u_n}{v_n}}{\frac{1}{v_n}} = \frac{\frac{u_0}{v_0}+\frac{n\bar{x}}{\sigma^2}}{\frac{1}{v_0}+\frac{n}{\sigma^2}} = \frac{\bar{x}+\frac{\sigma^2}{nv_0}u_0}{1+\frac{\sigma^2}{nv_0}} .$$

A 95% confidence interval for θ is "estimate $\pm 1.96 \times$ its SE",

i.e. $$\frac{\bar{x}+\frac{\sigma^2}{nv_0}u_0}{1+\frac{\sigma^2}{nv_0}} \pm 1.96\sqrt{\frac{\frac{\sigma^2}{n}}{1+\frac{\sigma^2}{nv_0}}}, \quad \text{since } \frac{1}{v_n} = \frac{\sigma^2+nv_0}{\sigma^2 v_0} .$$

(iii) For $u_0 = 0$, the risk function is given by

$$R_{\delta_\pi}(0) = E\left[\left\{\delta_\pi(x) - \theta\right\}^2\right] \qquad \text{where } \delta_\pi(x) \text{ is } E[\theta \mid X]$$

$$= \text{Var}\left\{\delta_\pi(x)\right\} + \left[E\left\{\delta_\pi(x)\right\} - \theta\right]^2$$

$$= \frac{\dfrac{\sigma^2}{n}}{\left(1+\dfrac{\sigma^2}{nv_0}\right)^2} + \left(\frac{\theta}{1+\dfrac{\sigma^2}{nv_0}} - \theta\right)^2 = \frac{\dfrac{\sigma^2}{n}}{\left(1+\dfrac{\sigma^2}{nv_0}\right)^2} + \frac{\theta^2\left(\dfrac{\sigma^2}{nv_0}\right)^2}{\left(1+\dfrac{\sigma^2}{nv_0}\right)^2}$$

$$= \frac{\dfrac{\sigma^2}{n}}{\left(1+\dfrac{\sigma^2}{nv_0}\right)^2}\left(1+\frac{\sigma^2}{n{v_0}^2}\theta^2\right).$$

If $\delta(X) = \bar{X}$, then $R_\delta(\theta) = \text{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$.

$$R_{\delta_\pi}(\theta) < R_\delta(\theta) \Leftrightarrow 1+\frac{\sigma^2\theta^2}{nv_0^2} < \left(1+\frac{\sigma^2}{nv_0}\right)^2 = 1+\frac{2\sigma^2}{nv_0}+\frac{\sigma^4}{n^2{v_0}^2}$$

$$\text{i.e.} \quad \theta^2 < 2v_0 + \frac{\sigma^2}{n}.$$

(i) The likelihood function is

$$L(\theta)=\prod_{i=1}^{n}\left(\frac{\theta v^{\theta}}{x_i^{\theta+1}}\right) \quad \text{for } v \le x < \infty$$

$$=\frac{\theta^n v^{n\theta}}{\left(\prod_{i=1}^{n} x_i\right)^{\theta+1}} \quad \text{for } \theta > 0.$$

$$\therefore \ln L \equiv l(\theta)=n\ln\theta+n\theta\ln v-(\theta+1)\sum\ln(x_i).$$

$$\therefore \frac{dl}{d\theta}=\frac{n}{\theta}+n\ln v-\sum\ln x_i \quad \text{and} \quad \frac{d^2l}{d\theta^2}=-\frac{n}{\theta^2}<0.$$

$\hat{\theta}$ is found from $\dfrac{dl}{d\theta}=0$, i.e. $\dfrac{n}{\hat{\theta}}=\sum\ln x_i-n\ln v=\sum\ln\left(\dfrac{x_i}{v}\right)$

so that $$\hat{\theta}=\frac{n}{\sum_{i=1}^{n}\ln\left(\frac{x_i}{v}\right)}.$$

(ii) For null hypothesis $\theta=1$, the generalised likelihood ratio is $\lambda=\dfrac{L(1)}{L(\hat{\theta})}$

so that $\ln(\lambda(\underline{x}))=l(1)-l(\hat{\theta})$.

Thus
$$\ln\{\lambda(\underline{x})\}=n\ln v-2\sum_{i=1}^{n}\ln(x_i)-n\ln(\hat{\theta})-n\hat{\theta}\ln v+(\hat{\theta}+1)\sum_{i=1}^{n}\ln(x_i)$$

$$=n\ln v+(\hat{\theta}-1)\sum\ln(x_i)-n\ln(\hat{\theta})-n\hat{\theta}\ln v$$

$$=-\frac{n}{\hat{\theta}}+\hat{\theta}\sum\ln(x_i)-n\ln(\hat{\theta})-n\hat{\theta}\ln v, \quad \text{using } \frac{n}{\hat{\theta}}=\sum\ln(x_i)-n\ln v$$

$$=-\frac{n}{\hat{\theta}}+n-n\ln(\hat{\theta}), \quad \text{again using } \frac{n}{\hat{\theta}}=\sum\ln(x_i)-n\ln v$$

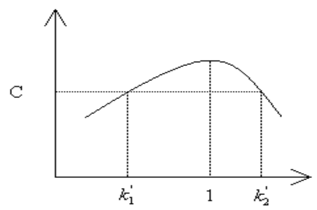
$$=n\left(1-\ln\hat{\theta}-\frac{1}{\hat{\theta}}\right).$$

(iii) Let $u = \frac{1}{\hat{\theta}}$. Then $\ln\{\lambda(\underline{x})\} = -n(u - 1 - \ln u)$

and $\frac{d}{du}(\ln\lambda) = -n\left(1 - \frac{1}{u}\right)$.

$\ln\lambda$ has a maximum at $u = 1$.

$H_0 : \theta = 1$ will be rejected if $\lambda(\underline{x}) \le c$, for some c; i.e. if $u \le k_1'$ or $u \ge k_2'$ as in the diagram which indicates the graph of $\lambda(\underline{x})$ against u.



From (i), reject H_0 if $\sum_{i=1}^{n} \ln(x_i) \le k_1$ or $\sum_{i=1}^{n} \ln(x_i) \ge k_2$,

where $k_1 = n\left\{k_1' + \ln v\right\}$ and $k_2 = n\left\{k_2' + \ln v\right\}$.

For a test of size α, choose k_1, k_2 to satisfy

$$P\left\{\sum_{i=1}^{n} \ln(x_i) \le k_1 \quad \text{or} \quad \sum_{i=1}^{n} \ln(x_i) \ge k_2 \ \mid\ \theta = 1\right\} = \alpha .$$

(i) Given observations $x_1, x_2, ..., x_n$ the likelihood function is

$$L_{(n)}(\theta) = \frac{\theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta-1}}{25^{n\theta}}, \quad \theta > 0.$$

Hence the likelihood ratio is $\lambda_n = \dfrac{L_{(n)}(3)}{L_{(n)}(6)} = \dfrac{3^n \left(\prod x_i\right)^2 25^{6n}}{6^n \left(\prod x_i\right)^5 25^{3n}}$

$$= \frac{25^{3n}}{2^n \left(\prod x_i\right)^3} = \frac{(7812.5)^n}{\left(\prod x_i\right)^3}.$$

In an SPR test, continue sampling if $A < \lambda_n < B$, accept H_0 if $\lambda_n \geq B$ and accept H_1 if $\lambda_n \leq A$ where $A = \dfrac{\alpha}{1-\beta} = \dfrac{0.05}{0.9} = \dfrac{1}{18}$ and $B = \dfrac{1-\alpha}{\beta} = \dfrac{0.95}{0.1} = 9.5$.

Continue sampling if

$$\ln A < n\ln(7812.5) - 3\sum_{i=1}^{n}\ln(x_i) < \ln B \Leftrightarrow 2.99n - 0.75 < \sum_{i=1}^{n}\ln(x_i) < 2.99n + 0.96.$$

Stop, and decide in favour of H_0, if $\sum \ln(x_i) \leq 2.99n - 0.75$;
Stop, and decide in favour of H_1, if $\sum \ln(x_i) \geq 2.99n + 0.96$.

(ii) $E\{\ln X\} = \displaystyle\int_0^{25} \frac{\ln x.\theta x^{\theta-1}}{25^\theta} dx = \left[\frac{x^\theta}{25^\theta}\ln x\right]_0^{25} - \int_0^{25} \frac{x^{\theta-1}}{25^\theta} dx = \ln 25 - \frac{1}{\theta}$.

For $\theta = 6$, $Z_i = \ln\left\{\dfrac{p_0(x_i)}{p_1(x_i)}\right\} = \ln(7812.5) - 3\ln X_i$ for $i = 1, 2, ..., n$.

So $E(Z_i) = \ln(7812.5) - 3\ln 25 + \dfrac{1}{2}$ and $E_{\mathrm{H}_1}(n) = \dfrac{(1-\beta)\ln A + \beta \ln B}{E_{\mathrm{H}_1}(Z_i)}$, which is

$$\frac{-0.9\ln 18 + 0.1\ln(9.5)}{-0.1931} = 12.3.$$

(iii)

n	$\sum \ln x_i$	$2.99n - 0.75$	$2.99n + 0.96$
1	3.09	2.24	3.95
2	6.26	5.23	6.94
3	9.30	8.22	9.93
4	12.47	11.21	12.92
5	15.69	14.20	15.91
6	18.87	17.19	18.90
7	21.95	20.18	21.89

– stop, and the decision is to accept H_1.

Graduate Diploma, Statistical Theory & Methods, Paper II, 2001. Question 7

If $\underline{X}$ is a set of data and $\underline{\theta}$ is an unknown parameter, then $q(\underline{X};\ \underline{\theta})$ is a pivotal quantity if

(i) $q(\underline{X};\ \underline{\theta})$ involves $\underline{\theta}$, but no other unknown parameters,

(ii) the distribution of q does not depend on $\underline{\theta}$ or on any other unknown parameters.

(a) The cdf of $Y = F(X)$ is given by

$$F_Y(y) = P(Y \le y) = P\{F(X) \le y\} = P\{X \le F^{-1}(y)\} = F\{F^{-1}(y)\} = y$$

for $0 < y < 1$.

Hence pdf of Y is $f_Y(y) = 1, \quad 0 < y < 1, \quad$ so $Y \sim \mathrm{U}(0,1)$.

(b) (i) Let $W = X - \theta$. Then $f_W(w) = \dfrac{e^w}{\left(1+e^w\right)^2}$.

$X - \theta$ is a function of θ whose distribution does not depend on θ, and so is a pivotal quantity.

Now, $f_W(w)$ is symmetric about zero, and so a $100(1-\alpha)\%$ confidence interval $[0 < \alpha < 1]$ for θ is $\{\theta:\ -c < X - \theta < c\}$, where c satisfies $P(W \le c) = \alpha/2$.

Hence $\displaystyle\int_{-\infty}^{c} \frac{e^w}{\left(1+e^w\right)^2}\,dw = \frac{\alpha}{2} \iff \left[-\frac{1}{1+e^w}\right]_{-\infty}^{c} = 1 - \frac{1}{1+e^c}$.

$\therefore c = \ln\left(\dfrac{\alpha/2}{1-(\alpha/2)}\right)$. When $\alpha = 0.05$, $c = -3.664$.

So when $X = 10$, the interval is (6.336, 13.664).

(ii) Let $\bar{X}$ denote the sample mean. CLT gives $\bar{X} \sim$ approx $\mathrm{N}\left(\theta, \dfrac{\pi^2}{3n}\right)$.

An approximate 95% confidence interval for θ is therefore

$$\left(\bar{X} - \frac{1.96\pi}{\sqrt{3n}},\ \bar{X} + \frac{1.96\pi}{\sqrt{3n}}\right).$$

Graduate Diploma, Statistical Theory & Methods, Paper II, 2001. Question 8

When it is not possible to find an analytical solution to a model used in practice, because it is too complex, the behaviour of the model may be mimicked or simulated using a computer to carry out a large number of runs of the model using generated data.

The accuracy of asymptotic results can be examined using finite sample sizes, giving approximations to sampling distributions.

Simulation can be used to study properties of estimators, such as bias, variance, distribution, shape, coverage probabilities, and to assist in finding robust estimators.

The relative performance of different inference procedures can be assessed, and the conditions under which particular procedures are superior can be identified.

Simulation is useful in checking assumptions, to see whether assumptions in a model, such as randomness, are reasonable before actually fitting it. If available data or knowledge are in line with simulation results, the model may be adequate.

Models where parameters are difficult to estimate by least squares methods can sometimes be handed by simulation.