THE ROYAL STATISTICAL SOCIETY

2004 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

PAPER II – STATISTICAL THEORY & METHODS

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

(i) Since the mean of the random variable is $\theta\lambda$, the method of moments estimator $\tilde{\lambda}_\theta$ is found from $\theta\tilde{\lambda}_\theta = \bar{X}$, giving $\tilde{\lambda}_\theta = \bar{X}/\theta$.

$E(\bar{X}) = E(X) = \theta\lambda$, so $E(\tilde{\lambda}_\theta) = \frac{1}{\theta}.\theta\lambda = \lambda$, i.e. $\tilde{\lambda}_\theta$ is unbiased for λ.

(ii) [Note that "log" denotes logarithm to base *e* throughout the solutions to this paper, in accordance with the Society's usual notation.]

We have

$$\log f(x) = (\theta-1)\log x - \frac{x}{\lambda} - \log k_\theta - \theta\log\lambda .$$

$$\therefore \frac{d}{d\lambda}\log f(x) = \frac{x}{\lambda^2} - \frac{\theta}{\lambda} \qquad \text{and} \qquad \frac{d^2}{d\lambda^2}\log f(x) = -\frac{2x}{\lambda^3} + \frac{\theta}{\lambda^2}.$$

$\therefore E\left[-\frac{d^2}{d\lambda^2}\log f(X)\right] = \frac{2\theta\lambda}{\lambda^3} - \frac{\theta}{\lambda^2} = \frac{\theta}{\lambda^2}$, so the Cramér-Rao lower bound is $\frac{\lambda^2}{n\theta}$.

Now,

$$\mathrm{Var}(\tilde{\lambda}_\theta) = \mathrm{Var}\left(\frac{\bar{X}}{\theta}\right) = \frac{1}{\theta^2}\mathrm{Var}(\bar{X}) = \frac{1}{\theta^2}.\frac{\theta\lambda^2}{n} = \frac{\lambda^2}{n\theta},$$

and so the variance of $\tilde{\lambda}_\theta$ attains the lower bound.

(iii) We now have, instead, that λ is known and θ is not. The likelihood function is $L(\theta) = \prod_{i=1}^{n} \frac{x_i^{\theta-1}}{k_\theta \lambda^\theta} e^{-x_i/\lambda}$, from which we have

$$\log L(\theta) = (\theta-1)\sum_{i=1}^{n}\log x_i - n\log k_\theta - n\theta\log\lambda - \frac{1}{\lambda}\sum_{i=1}^{n} x_i ,$$

in which the final term is a function of the data (recall that λ is known) and the remaining three terms form a function of θ and $\Sigma\log x_i$. Therefore by the Neyman-Fisher factorisation theorem, $\Sigma\log x_i$ is a sufficient statistic for θ.

(iv) We can argue directly that the method of moments estimator must be a function of $\bar{X}$, i.e. of ΣX_i. But Σx_i is <u>not</u> a sufficient statistic, so an estimator based on it will not be fully efficient.

To find the estimator, we have $E(\bar{X}) = \theta\lambda$, so the estimator is $\bar{X}/\lambda$.

Graduate Diploma, Statistical Theory & Methods, Paper II, 2004. Question 2

$$f(x) = \frac{cx^{c-1}}{\lambda^c} \exp\left\{ -\left(\frac{x}{\lambda} \right)^c \right\} \qquad (x, c, \lambda \text{ all} > 0)$$

(i) $\log f(x) = \log c + (c-1)\log x - \left(\frac{x}{\lambda} \right)^c - c\log\lambda$, giving

$$\frac{d}{d\lambda}\log f(x) = \frac{cx^c}{\lambda^{c+1}} - \frac{c}{\lambda}.$$

Thus, from $E\left[\frac{d}{d\lambda}\log f(X) \right] = 0$, we have $\frac{c}{\lambda^{c+1}} E\left[X^c \right] = \frac{c}{\lambda}$ so that $E\left[X^c \right] = \lambda^c$.

(ii) From part (i), $\frac{d\log L}{d\lambda} = \frac{c}{\lambda^{c+1}} \Sigma x_i^c - \frac{nc}{\lambda}$. Setting this equal to zero gives

$\hat{\lambda}^c = \frac{1}{n}\sum_{i=1}^{n} x_i^c$, i.e. $\hat{\lambda} = \left\{ \frac{1}{n}\sum_{i=1}^{n} x_i^c \right\}^{1/c}$. [It may easily be verified from the second derivative that this is indeed a maximum.]

(iii) $\frac{d^2}{d\lambda^2}\log L = -\frac{c(c+1)}{\lambda^{c+2}} \sum x_i^c + \frac{nc}{\lambda^2}$.

$$\therefore E\left[-\frac{d^2}{d\lambda^2}\log L \right] = \frac{c(c+1)}{\lambda^{c+2}} n\lambda^c - \frac{nc}{\lambda^2} = \frac{nc^2}{\lambda^2}.$$

Hence the large sample variance of $\hat{\lambda}$ is $\frac{\lambda^2}{nc^2}$.

(iv) An approximate 95% confidence interval for λ is $\hat{\lambda} \pm \frac{1.96\hat{\lambda}}{c\sqrt{n}}$, giving $4 \pm \frac{1.96 \times 4}{20}$, i.e. 4 ± 0.39.

[Also acceptable is $4 \pm \frac{2 \times 4}{20}$, i.e. 4 ± 0.4.]

Graduate Diploma, Statistical Theory & Methods, Paper II, 2004. Question 3

(i) $L = \dfrac{\theta^{\Sigma x_i} e^{-n\theta}}{\Pi\, x_i!}$, giving $\log L = (\Sigma x_i)\log\theta - n\theta - \log(\Pi\, x_i!)$.

$\therefore \dfrac{d \log L}{d\theta} = \dfrac{\Sigma x_i}{\theta} - n$, so the maximum likelihood estimator of θ is $\hat{\theta} = \dfrac{1}{n}\Sigma X_i = \bar{X}$.

(Note that $\dfrac{d^2 \log L}{d\theta^2} = -\dfrac{\Sigma x_i}{\theta^2}$, so this is indeed a maximum.)

By the "invariance property" of maximum likelihood estimators, the ML estimator of $\lambda = e^{-\theta}$ is $\hat{\lambda} = e^{-\hat{\theta}} = e^{-\bar{X}}$.

(ii) For the Poisson distribution with mean θ, Var(X) = θ. Hence $\mathrm{Var}(\bar{X}) = \theta/n$, i.e. $\mathrm{Var}(\hat{\theta}) = \theta/n$.

The delta method gives that the variance of $g(\hat{\theta})$ is approximated by $\left(\dfrac{dg}{d\theta}\right)^2 \mathrm{Var}(\hat{\theta})$ evaluated at the mean of the distribution which here is simply θ. So we need to obtain $\dfrac{\theta}{n}\left(\dfrac{dg}{d\theta}\right)^2$ with $g(\theta) = e^{-\theta}$. This immediately gives $dg/d\theta = -e^{-\theta}$, so the approximate variance is $\dfrac{\theta}{n}\left(-e^{-\theta}\right)^2 = \dfrac{\theta e^{-2\theta}}{n}$.

(iii) The number of zero observations is binomially distributed with $p = e^{-\theta} = \lambda$, i.e. B($n$, λ). Thus $\tilde{\lambda}$, the proportion of zeros, has expected value λ, i.e. it is unbiased. Also we have

$$\mathrm{Var}(\tilde{\lambda}) = \frac{\lambda(1-\lambda)}{n} = \frac{e^{-\theta}\left(1-e^{-\theta}\right)}{n}.$$

(iv) Using the approximate variance from part (ii), the efficiency of $\tilde{\lambda}$ relative to $\hat{\lambda}$ is given approximately by $\dfrac{\theta e^{-2\theta}}{n}.\dfrac{n}{e^{-\theta}\left(1-e^{-\theta}\right)} = \dfrac{\theta}{e^{\theta}-1}$. Hence if θ is small, the efficiency is near (but less than) 1; as θ increases, the efficiency decreases; as θ becomes large, the efficiency tends to zero.

[Candidates were expected to provide a rough sketch accordingly.]

Graduate Diploma, Statistical Theory & Methods, Paper II, 2004. Question 4

H_0: $\mu = 0$. H_1: $\mu = 1$. The likelihood is $L(\mu) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$.

(i) $\frac{L_1}{L_0} = \exp\left\{\frac{1}{2}\Sigma x_i^2 - \frac{1}{2}\Sigma(x_i-1)^2\right\} = \exp\left(\Sigma x_i - \tfrac{1}{2}n\right) = \exp\left(n\bar{x} - \tfrac{1}{2}n\right)$.

This is an increasing function of $\bar{x}$, so the Neyman-Pearson test will reject H_0 when $\bar{x} > k$ for some suitable k.

(ii) Type I error is $P\left(\bar{X} > k \middle| \mu = 0\right)$, required to be ≤ 0.05.

Type II error is $P\left(\bar{X} < k \middle| \mu = 1\right)$, required to be ≤ 0.05.

If $\mu = 0$, we have $\bar{X} \sim N(0, 1/n)$; the Type I error criterion gives $1 - \Phi\left(k\sqrt{n}\right) \leq 0.05$, i.e. $k\sqrt{n} \geq 1.6449$.

If $\mu = 1$, we have $\bar{X} \sim N(1, 1/n)$; the Type II error criterion gives $\Phi\left(\frac{k-1}{1/\sqrt{n}}\right) \leq 0.05$, i.e. $(k-1)\sqrt{n} \leq -1.6449$.

Solving these two inequalities together gives $k = ½$ and $n \geq 1.6449^2 \div \left(\tfrac{1}{2}\right)^2$, i.e. $n \geq 10.82$, so use $n = 11$.

(iii) Let $L_0(m)$ and $L_1(m)$ denote the likelihoods after taking m observations, with likelihood ratio $\lambda_m = \frac{L_0(m)}{L_1(m)}$. Here we have $\lambda_m = \exp\left(\tfrac{1}{2}m - m\bar{x}\right)$.

The sequential probability ratio test rule is to continue sampling if $A < \lambda_m < B$, accept H_0 if $\lambda_m \geq B$ and reject H_0 (i.e. accept H_1) if $\lambda_m \leq A$. A and B are given (approximately) by $A = \frac{\alpha}{1-\beta} = \frac{0.05}{0.95} = \frac{1}{19}$, $B = \frac{1-\alpha}{\beta} = \frac{0.95}{0.05} = 19$.

The approximate expected sample size under H_0 is given by

$E\left(N \middle| H_0\right) = \frac{\alpha \log A + (1-\alpha)\log B}{E\left(Z_i \middle| H_0\right)}$, where $z_i = \tfrac{1}{2} - x_i$ since $\log \lambda_m = \tfrac{1}{2}m - \Sigma x_i$. This gives $E\left(Z_i \middle| H_0\right) = \tfrac{1}{2}$ and so $E\left(N \middle| H_0\right) \approx \left(0.05\log(1/19) + 0.95\log 19\right) \div 0.5 = 5.30$.

Similarly, the approximate expected sample size under H_1 is given by

$E\left(N \middle| H_1\right) = \frac{(1-\beta)\log A + \beta \log B}{E\left(Z_i \middle| H_1\right)}$, and $E\left(Z_i \middle| H_1\right) = -\tfrac{1}{2}$, giving

$E\left(N \middle| H_1\right) \approx \left(0.95\log(1/19) + 0.05\log 19\right) \div (-0.5) = 5.30$.

(i) $F_0(x)$ is specified and can therefore be evaluated at the n sample points $x_1, x_2, \ldots, x_n$. It is compared with the empirical distribution function $S(x)$ constructed by ranking the sample values (it is assumed that this has already been done for $x_1, x_2, \ldots, x_n$) and ascribing to $S(x)$ the values $1/n$ at x_1, $2/n$ at x_2, …, $(n-1)/n$ at x_{n-1} and 1 at x_n.

The Kolmogorov-Smirnov (KS) test is based on the absolute values $|F(x_i) - S(x_i)|$. If these are all "sufficiently small", the null hypothesis (that $F_0(x)$ is the correct underlying cumulative distribution function) cannot be rejected. The procedure uses the largest absolute value (commonly denoted by D_n) and compares it with a special table of critical values depending on sample size.

(ii) KS can be applied to any fully specified F_0. It operates by comparing cumulative distribution functions, whereas the familiar chi-squared test compares histograms constructed from the data and the pdf. Thus KS does not require the data to be grouped into intervals as does the chi-squared test. Also, it typically needs smaller sample sizes. However, the chi-squared test is much more straightforward to use.

D_n is distribution-free so long as F is continuous. Its exact distribution is known, unlike that of the chi-squared test statistic which is only approximate. However, it requires F_0 to be fully specified. When F_0 is not fully specified, KS is not so easily applied, whereas the chi-squared test is very easily adjusted by reducing the number of degrees of freedom. KS is also not so easily applied to discrete distributions.

(iii) The ranked data are 3, 15, 30, 45, 57, 80, 145, 170, 251, 280.

$F_0(x) = 1 - e^{-x/100}$, for $x = 3, 15, \ldots, 280$. $S(x)$ takes the values 0.1, 0.2, …, 1.0.

x	3	15	30	45	57	80	145	170	251	280
$F_0(x)$	0.0296	0.1393	0.2592	0.3624	0.4345	0.5507	0.7654	0.8173	0.9187	0.9392
$S(x)$	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\lvert D \rvert$	0.070	0.061	0.041	0.038	0.066	0.049	0.065	0.017	0.019	0.061

The maximum absolute difference is 0.066, which is not signficant (5% critical point is 0.409). The null hypothesis is not rejected.

Graduate Diploma, Statistical Theory & Methods, Paper II, 2004. Question 6

(i) For an estimator T and loss function $l(t, \theta)$, the risk of T is $R(\theta) = E[l(t, \theta)]$.

T is inadmissible if there is an estimator U such that

$$R_U(\theta) \le R_T(\theta) \quad \text{for all } \theta$$

and $R_U(\theta) < R_T(\theta)$ for at least one value of θ.

(ii) For the given distribution, $E[X] = \theta$ and $\text{Var}(X) = \theta^2$.

Thus

$$E\left[\hat{\theta}\right] = \frac{1}{n} n\theta = \theta \text{ (so } \hat{\theta} \text{ is unbiased)},$$

$$\text{Var}\left(\hat{\theta}\right) = \left(\frac{1}{n}\right)^2 n\theta^2 = \frac{\theta^2}{n}.$$

Therefore

$$MSE\left(\hat{\theta}\right) = \text{variance} + (\text{bias})^2 = \frac{\theta^2}{n} + 0 = \frac{\theta^2}{n}.$$

Similarly,

$$E\left(\tilde{\theta}\right) = \frac{1}{n+1} n\theta \text{, so the bias of } \tilde{\theta} \text{ is } \frac{n\theta}{n+1} - \theta = \frac{-\theta}{n+1},$$

$$\text{Var}\left(\tilde{\theta}\right) = \left(\frac{1}{n+1}\right)^2 n\theta^2 = \frac{n\theta^2}{(n+1)^2}.$$

Therefore

$$MSE\left(\tilde{\theta}\right) = \frac{n\theta^2}{(n+1)^2} + \left(\frac{-\theta}{n+1}\right)^2 = \frac{\theta^2}{n+1}.$$

(iii) For squared error loss, $R(\theta) = E[(T - \theta)^2] = MSE(T)$.

Since $R_{\tilde{\theta}}(\theta) < R_{\hat{\theta}}(\theta)$ for all θ, $\hat{\theta}$ is inadmissible.

(iv) As $n \to \infty$, $E\left(\tilde{\theta}\right) \to \theta$ and $\text{Var}\left(\tilde{\theta}\right) \to 0$. Hence $\tilde{\theta}$ is a consistent estimator of θ.

Graduate Diploma, Statistical Theory & Methods, Paper II, 2004. Question 7

(i) The likelihood is $L(\mathbf{x}|\theta) = \theta^n (1-\theta)^{\Sigma x_i}$. Thus the posterior density is

$$p(\theta|\mathbf{x}) \propto p(\theta) L(\mathbf{x}|\theta) \propto \theta^{a-1}(1-\theta)^{b-1}.\theta^n(1-\theta)^{\Sigma x_i} = \theta^{a+n-1}(1-\theta)^{b+\Sigma x_i - 1}$$

which is beta with parameters $a + n$ and $b + \Sigma x_i$. Thus the beta distribution is a conjugate prior.

(ii) Mean = $\dfrac{a}{a+b} = \dfrac{1}{2}$; this gives $a = b$.

Variance = $\dfrac{ab}{(a+b+1)(a+b)^2} = \dfrac{a^2}{(2a+1)(4a^2)} = \dfrac{1}{100}$; thus $2a + 1 = 25$, so $a = 12$ and thus also $b = 12$.

(iii) The posterior distribution is beta with parameters 100 and 60.

So the posterior mean is $\dfrac{100}{100+60} = \dfrac{5}{8} = 0.625$. Also, the posterior variance is $\dfrac{100 \times 60}{161 \times 160^2} = 0.001456$, so the posterior standard deviation is 0.038.

(iv) Using a Normal approximation with the same mean and variance as the posterior beta distribution, an approximate 90% interval for θ is

$$0.625 \pm (1.645 \times 0.038) = 0.625 \pm 0.063,$$

i.e. (0.562, 0.688).

(i) Let $f(x\,|\,\theta)$ denote the pdf of the given distribution. Then the likelihood function for a sample $x_1, x_2, \ldots, x_n$ is $L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta)$, viewed as a function of θ.

The maximum likelihood estimator $\hat{\theta}(\mathbf{x})$ is the value of θ at which $L(\theta|\mathbf{x})$ attains its maximum. This can be obtained (under suitable regularity conditions) as the solution of $dL/d\theta = 0$ or equivalently of $d\log L/d\theta = 0$.

(a) The likelihood ratio test statistic for testing H_0: $\theta = \theta_0$ against H_1: $\theta \neq \theta_0$ is

$$\lambda(\mathbf{x}) = \frac{L(\theta_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})}.$$

H_0 is rejected for "small" values of λ, since these indicate that the likelihood under H_0 is "too small" compared with its maximum possible value. Thus H_0 is rejected for $\lambda \leq c$ say, where c is a constant to be determined.

We illustrate for the case of the Normal distribution with unknown mean θ and known variance σ^2. Here the maximum likelihood estimator of θ is simply $\bar{x}$ and we have

$$\lambda(\mathbf{x}) = \frac{\left(\sigma\sqrt{2\pi}\right)^{-n} \exp\left\{-\sum(x_i-\theta_0)^2/2\sigma^2\right\}}{\left(\sigma\sqrt{2\pi}\right)^{-n} \exp\left\{-\sum(x_i-\bar{x})^2/2\sigma^2\right\}}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum(x_i-\theta_0)^2 - \sum(x_i-\bar{x})^2\right]\right\}.$$

Using $\sum(x_i-\theta_0)^2 = \sum(x_i-\bar{x})^2 + n(\bar{x}-\theta_0)^2$, this gives

$$\lambda(\mathbf{x}) = \exp\left\{-\frac{n}{2\sigma^2}(\bar{x}-\theta_0)^2\right\}, \qquad \text{or} \qquad \log\lambda(\mathbf{x}) = -\frac{n}{2\sigma^2}(\bar{x}-\theta_0)^2.$$

The rejection region is given by $\mathbf{x}$ such that $\lambda(\mathbf{x}) \leq c$, so H_0 is rejected when

$$|\bar{x}-\theta_0| \geq \sqrt{-\frac{2\sigma^2}{n}\log c}\,.$$

Now, c must lie between 0 and 1, so the test criterion is to reject H_0 when $|\bar{x}-\theta_0|$ is greater than some constant – i.e. we get the familiar two-tailed test comparing the sample mean $\bar{x}$ with the hypothesised value θ_0.

Solution continued on next page

(b) Under regularity conditions,

$$E\left[\frac{d\log L}{d\theta}\right]=0$$

and

$$E\left[\left(\frac{d\log L}{d\theta}\right)^2\right]=-E\left[\frac{d^2\log L}{d\theta^2}\right].$$

This leads to the Cramér-Rao lower bound for the variance of an unbiased estimator as

$$\frac{1}{E\left[\left(\frac{d\log L}{d\theta}\right)^2\right]} = -\frac{1}{E\left[\frac{d^2\log L}{d\theta^2}\right]}.$$

As sample size becomes very large, this bound applies to the variance of a maximum likelihood estimator.

Further, maximum likelihood estimators are asymptotically Normally distributed.

Hence a large-sample confidence interval for θ can be obtained as

$$\hat{\theta} \pm z\sqrt{-\frac{1}{E\left[\frac{d^2\log L}{d\theta^2}\right]}}$$

where z is the required percentage point from the N(0, 1) distribution (e.g. 1.96 for a 95% confidence interval).

(ii) If prior knowledge is vague, with a locally relatively flat prior distribution being used, the posterior distribution will be approximately proportional to the likelihood in the vicinity of the maximum likelihood estimator. Thus the method outlined above will give an approximation to a Bayesian interval for the parameter.