THE ROYAL STATISTICAL SOCIETY

2005 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

STATISTICAL THEORY AND METHODS

PAPER II

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.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Graduate Diploma, Statistical Theory & Methods, Paper II, 2005. Question 1

$$f(x) = \sqrt{\frac{2}{\pi\theta}} \exp\left(-\frac{x^2}{2\theta}\right), \qquad x > 0.$$

(i) $E\left(X^2\right) = \int_0^\infty x^2 f(x)\,dx = \sqrt{\frac{2}{\pi\theta}} \int_0^\infty x^2 e^{-x^2/2\theta} dx$

$= \sqrt{\frac{2}{\pi\theta}} \int_0^\infty x\left(xe^{-x^2/2\theta}\right) dx$

Integrate by parts, taking the parts as x and $xe^{-x^2/2\theta}$

$= \sqrt{\frac{2}{\pi\theta}} \left\{ x\left[-\theta e^{-x^2/2\theta}\right]_0^\infty + \int_0^\infty \theta e^{-x^2/2\theta} d\theta \right\}$

$= \sqrt{\frac{2}{\pi\theta}} \left\{[0-0]\right\} \; + \; \theta \int_0^\infty f(x)\,dx$

$= \theta$.

(ii) $\log L = \log\left\{\left(\frac{2}{\pi\theta}\right)^{n/2} \exp\left(-\frac{\Sigma x_i^2}{2\theta}\right)\right\} = \frac{n}{2}\log\left(\frac{2}{\pi\theta}\right) - \frac{\Sigma x_i^2}{2\theta}$.

$\therefore \frac{d \log L}{d\theta} = -\frac{n}{2\theta} + \frac{1}{2\theta^2} \sum x_i^2$.

Setting this equal to zero gives $n\hat{\theta} = \Sigma x_i^2$, i.e. $\hat{\theta} = \frac{\Sigma x_i^2}{n}$.

It may be verified (e.g. by considering the second derivative) that this is indeed a maximum, and so it is the MLE of θ.

(iii) Since $E(X^2) = \theta$, we immediately have $E(\hat{\theta}) = \theta$, i.e. $\hat{\theta}$ is unbiased for θ.

To find the Cramér-Rao lower bound, we first find $\frac{d^2 \log L}{d\theta^2} = \frac{n}{2\theta^2} - \frac{\Sigma x_i^2}{\theta^3}$, from which $E\left(-\frac{d^2 \log L}{d\theta^2}\right) = -\frac{n}{2\theta^2} + \frac{n\theta}{\theta^3} = \frac{n}{2\theta^2}$, so the Cramér-Rao lower bound is $\frac{2\theta^2}{n}$.

Now, $\mathrm{Var}(\hat{\theta}) = 2\theta^2/n$, using the result (given in the question) $\mathrm{Var}(X^2) = 2\theta^2$, and so the variance of $\hat{\theta}$ attains the bound.

(iv) $\phi = \sqrt{\theta}$; so MLE of ϕ is $\hat{\phi} = \sqrt{\text{MLE of } \theta} = \sqrt{\Sigma x_i^2 / n}$. Because ϕ is a non-linear transformation of θ, and $\hat{\theta}$ is unbiased for θ, $\hat{\phi}$ cannot be unbiased for ϕ.

Graduate Diploma, Statistical Theory & Methods, Paper II, 2005. Question 2

$$f(x) = \frac{1}{\theta}, \qquad 0 < x < \theta .$$

(i) $\mu = E(X) = \theta/2$ (from symmetry or by simple integration), so $\theta = 2\mu$.

Thus the method of moments estimator is $\tilde{\theta} = 2\bar{X}$ where $\bar{X}$ is the sample mean.

$E(\tilde{\theta}) = 2E(\bar{X}) = 2E(X) = \theta$, i.e. $\tilde{\theta}$ is unbiased.

(ii) With the given sample, $\bar{x} = 0.4$ and so the value of $\tilde{\theta}$ is 0.8. This is not valid as an estimate of θ because the largest observation in the sample is 1.0 and therefore we know that θ must be ≥ 1.

(iii) $P(Y \leq y) = P(\text{all } X_i \text{ are} \leq y) = \prod_{i=1}^{n} P(X_i \leq y) = (y/\theta)^n$. The pdf of Y is the derivative of this, i.e. $g(y) = \frac{ny^{n-1}}{\theta^n}$, for $0 < y < \theta$.

(iv) $MSE(cY) = E\left[(cY - \theta)^2\right] = c^2 E(Y^2) - 2c\theta E(Y) + \theta^2$.

$\therefore \frac{d(MSE)}{dc} = 2cE(Y^2) - 2\theta E(Y)$, which equals zero for $c = \frac{\theta E(Y)}{E(Y^2)}$. Further,

$\frac{d^2(MSE)}{dc^2} = 2E(Y^2) > 0$, so this is a minimum.

We have $E(Y) = \frac{n}{\theta^n}\int_0^\theta y^n dy = \frac{n}{\theta^n}\left[\frac{y^{n+1}}{n+1}\right]_0^\theta = \frac{n\theta}{n+1}$ and $E(Y^2) = \frac{n}{\theta^n}\int_0^\theta y^{n+1} dy = \frac{n\theta^2}{n+2}$.

$$\therefore c = \frac{\frac{n\theta^2}{n+1}}{\frac{n\theta^2}{n+2}} = \frac{n+2}{n+1} .$$

Graduate Diploma, Statistical Theory & Methods, Paper II, 2005. Question 3

(i) The formal definition of a sufficient statistic (S) is that the conditional distribution of the sample $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ given the value of S (= s) does not depend on θ.

(ii) If $f(\boldsymbol{x} \mid \theta)$ is the joint pdf of the sample $\boldsymbol{X}$, the statistic S is sufficient for θ if and only if there exist functions $g(s \mid \theta)$ and $h(\boldsymbol{x})$ such that, for all sample points $\{x_i\}$ and all θ,

$$f(\boldsymbol{x} \mid \theta) = g(s \mid \theta)\, h(\boldsymbol{x}) .$$

(iii) $$f(\boldsymbol{x} \mid \theta) = \frac{\prod_{i=1}^{n} x_i^{\theta-1} e^{-x_i}}{\{\Gamma(\theta)\}^n}$$

$$= \underbrace{\frac{\exp\{(\theta-1)\Sigma \log x_i\}}{\{\Gamma(\theta)\}^n}}_{\uparrow \; g(\Sigma \log x_i \mid \theta)} \times \underbrace{e^{-\Sigma x_i}}_{\uparrow \; h(\boldsymbol{x})} \qquad \text{[= } L\text{, for use in part (iv)]}$$

Thus, by the factorisation theorem, $S = \sum_{i=1}^{n} \log X_i$ is sufficient for θ.

(iv) L_0, i.e. the value of L when $\theta = 1$, is $e^{-\Sigma x_i}$. L_1, the value when $\theta = 2$, is $\exp(\Sigma \log x_i) \times e^{-\Sigma x_i}$. (Note that $\Gamma(1) = \Gamma(2) = 1$.)

Thus $\lambda = \dfrac{L_1}{L_0} = \exp(\Sigma \log x_i)$, which is an increasing function. So the Neyman-Pearson method rejects H_0 when $\lambda > c$, where c is a constant depending on the chosen rejection probability α.

So H_0 is rejected when $\Sigma \log x_i > c$ (not the same c), i.e. the rejection criterion is $S > c$.

(i) The median is that value of X below which half of the distribution lies, i.e. θ where $P(X \le \theta) = ½$. Suppose we have the null hypothesis H_0 that $\theta = \theta_0$ and alternative hypothesis $\theta \ne \theta_0$. A two-sided test is thus required (the work carries through in the obvious way for one-sided situations). The sign test counts the number, T, of the 100 observations which are less than θ (or, equivalently, more than θ; any observations which are (to the limits of accuracy of measurement) exactly equal to θ are commonly simply ignored). If this is "near" to half the number of observations in the sample, H_0 is supported; if not, H_0 is discredited. Under H_0, the distribution of T is binomial with $n = 100$ and $p = ½$.

A test of H_0 against H_1 can be based directly on the binomial distribution by cutting off "tails" at each end of the distribution. This is a tractable approach for small values of n, either using binomial tables or by calculating binomial probabilities. A Normal approximation is adequate for reasonably large values of n (particularly for $p = ½$), certainly for $n = 100$.

Thus we take T to be approximately $N(np, np(1-p))$, i.e. $N(\tfrac{1}{2}n, \tfrac{1}{4}n)$.

We reject H_0 if $\left|T - \tfrac{1}{2}n\right| \ge k$ where k is the smallest value having $P(T \le \tfrac{1}{2}n - k) \le \tfrac{\alpha}{2}$ if H_0 is true and the test is at the level α for the probability of type I error. Using the Normal approximation with inclusion of a continuity correction we have

$$P(T \le \tfrac{1}{2}n - k) \approx \Phi\left(\frac{\frac{n}{2} - k + \frac{1}{2} - \frac{n}{2}}{\sqrt{\frac{n}{4}}}\right) = \Phi\left(\frac{1-2k}{\sqrt{n}}\right).$$

(ii) With $n = 100$ and $\alpha = 0.05$, we want $\Phi\left(\dfrac{1-2k}{10}\right) \le 0.025$, so that $\dfrac{1-2k}{10} \le -1.96$ which gives $k \ge 10.3$. So we take $k = 11$ as it must be an integer, and the rule is to reject H_0 if $T \le 39$ or $T \ge 61$.

(iii) A point estimate of a parameter, based on data from a sample, may or may not be very precise. It is often more useful to have an interval estimate (L, U) where $L \le$ (parameter) $\le U$ with some specified probability. In the parametric case, exact values of L and U can be found by using the same mathematics as in hypothesis testing; for example, $P(\bar{X} - \frac{ts}{\sqrt{n}} \le \mu \le \bar{X} + \frac{ts}{\sqrt{n}}) = 0.95$ for a sample from a Normal distribution leads to an exact 95% confidence interval for μ. For non-parametric intervals, the confidence interval can only be set approximately at, say, 95% and it will contain all those values of the parameter which would not be rejected in a test at the 5% level.

(iv) The required interval contains all values of θ which would not be rejected at the 5% level. Consider testing H_0: $\theta = \theta^*$. Let T^* be the number of observations less than θ^*, then, as in (ii), H_0 is not rejected if $|T^* - 50| < 10.3$, leading to $40 \le T^* \le 60$. It follows that, when the sample has been ranked $X_{(1)}$ to $X_{(100)}$, the values $X_{(40)}$ and $X_{(60)}$ are the required ends of the interval.

Graduate Diploma, Statistical Theory & Methods, Paper II, 2005. Question 5

(i) Denoting the posterior pdf by $g(\theta|\mathbf{x})$, we have

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

(note that Σx_i is equal to the number of 1s that are observed).

We are given that beta(a, b) has pdf proportional to $y^{a-1}(1-y)^{b-1}$, so the posterior distribution is seen to be beta($2+\Sigma x_i$, $n+2-\Sigma x_i$).

(ii) With a squared error loss function, the Bayes estimator is the mean of the posterior distribution. The posterior pdf is

$$\frac{\Gamma(n+4)}{\Gamma(2+\Sigma x_i)\,\Gamma(n+2-\Sigma x_i)}\theta^{1+\Sigma x_i}(1-\theta)^{n+1-\Sigma x_i}$$

and so the posterior mean is

$$\frac{\Gamma(n+4)}{\Gamma(2+\Sigma x_i)\,\Gamma(n+2-\Sigma x_i)}\int_0^1 \theta^{2+\Sigma x_i}(1-\theta)^{n+1-\Sigma x_i}d\theta$$

$$= \frac{\Gamma(n+4)}{\Gamma(2+\Sigma x_i)\,\Gamma(n+2-\Sigma x_i)}\frac{\Gamma(3+\Sigma x_i)\,\Gamma(n+2-\Sigma x_i)}{\Gamma(n+5)}$$

(quoting standard results, or by explicitly evaluating the integral)

$$= \frac{\Gamma(n+4)\,\Gamma(3+\Sigma x_i)}{\Gamma(n+5)\,\Gamma(2+\Sigma x_i)} = \frac{2+\Sigma x_i}{4+n}$$

(using $\Gamma(p+1) = p\Gamma(p)$; the mean itself might also be quoted as a standard result).

(iii) The estimator is $\tilde{\theta} = \dfrac{2+\Sigma X_i}{4+n}$, where ΣX_i is the sum of Bernoulli distributions and so is binomial with parameters n and θ. So $E\left(\tilde{\theta}\right) = \dfrac{2+n\theta}{4+n}$, and thus

$$\text{bias}\left(\tilde{\theta}\right) = E\left(\tilde{\theta}\right) - \theta = \frac{2+n\theta}{4+n} - \theta = \frac{2(1-2\theta)}{4+n}.$$

To find the MSE, we use MSE = (bias)^2 + Variance. As $\Sigma X_i \sim$ B(n, θ), we have $\text{Var}\left(\tilde{\theta}\right) = \dfrac{1}{(4+n)^2}n\theta(1-\theta)$ and thus

$$MSE(\tilde{\theta}) = \frac{1}{(4+n)^2}\left\{4(1-2\theta)^2 + n\theta(1-\theta)\right\}$$

$$= \frac{4-16\theta+16\theta^2+n\theta-n\theta^2}{(4+n)^2} = \frac{4+(n-16)\theta(1-\theta)}{(4+n)^2}.$$

Graduate Diploma, Statistical Theory & Methods, Paper II, 2005. Question 6

(i) Denoting the posterior pdf by $g(\theta|\mathbf{x})$, we have

$$g(\theta|\mathbf{x}) \propto L(\mathbf{x}|\theta)g(\theta) \propto \theta^n e^{-\theta\Sigma x_i^2}.\theta^{a-1}e^{-b\theta} = \theta^{a+n-1}\exp\left\{-\theta(b+\Sigma x_i^2)\right\}$$

which we see to be of the form of the gamma pdf quoted in the question, merely with a replaced by $a + n$ and b replaced by $b + \Sigma x_i^2$. Since the prior and posterior are both gamma distributions, the gamma distribution is a conjugate prior.

(ii) Using the expressions given in the question,

$$E\left(\theta|\mathbf{x}\right) = \frac{a+n}{b+\Sigma x_i^2} \qquad \text{and} \qquad \mathrm{SD}\left(\theta|\mathbf{x}\right) = \frac{\sqrt{a+n}}{\left(b+\Sigma x_i^2\right)} \ .$$

(iii) $a = b = 1$; $n = 48$; $\Sigma x_i^2 = 48.0$.

Because $a = b = 1$, the prior pdf is simply $g(\theta) = e^{-\theta}$, i.e. the exponential distribution with parameter 1 (so mean = 1 and SD = variance = 1).

The posterior distribution is gamma with parameters 1 + 48 = 49 and 1 + 48.0 = 49, so the mean is 1, the variance is $49/49^2 = 1/49$ and the standard deviation is 1/7. This will be approximately Normal with mean 1 and variance 1/49.

[Note that the mean of both prior and posterior is 1, but the standard deviation of the posterior is only (1/7)th that of the prior.]

Sketches have not been included here because of the limitations of electronic reproduction.

(iv) Using the Normal approximation, a Bayesian 95% posterior interval for θ is "mean $\pm$ 1.96 SD", i.e. $1 \pm (1.96 \times (1/7))$, i.e. (0.72, 1.28).

Part (a)

The Neyman-Pearson method uses the likelihood ratio principle. If a random sample of n observations from a distribution with pdf $f(x, \theta)$ is available, the method is based on the ratio

$$\lambda = \frac{\prod_{i=1}^{n} f(x_i, \theta_1)}{\prod_{i=1}^{n} f(x_i, \theta_2)} ,$$

where θ_1 and θ_2 are the values of θ on the null and alternative hypotheses respectively.

The critical region R of size α (probability of Type I error) which maximises the power for testing $\theta = \theta_1$ against $\theta = \theta_2$ is that region for which

$$\lambda < k \quad \text{and} \quad \int_R \prod_{i=1}^{n} f(x_i, \theta_1)\, dx = \alpha .$$

As an example [but a *clear* description of the necessary steps, without an example, would be sufficient in the examination], suppose the hypotheses are $\theta = 0$ and $\theta = 1$ in N(θ, 1). Then

$$\lambda = \frac{e^{-\frac{1}{2}\Sigma x_i^2}}{e^{-\frac{1}{2}\Sigma (x_i - 1)^2}} = e^{\frac{1}{2}(\Sigma x_i^2 - 2\Sigma x_i + n - \Sigma x_i^2)} = e^{-n\bar{x} + \frac{1}{2}n}$$

which is to be $< k$ for a Neyman-Pearson test. This gives

$$-n\bar{x} + \frac{1}{2}n < \log k$$

$$-n\bar{x} < -\frac{n}{2} + \log k$$

$$\bar{x} > \frac{1}{2} - \frac{1}{n}\log k , \qquad \text{i.e. say} \quad \bar{x} > c .$$

To achieve size α, we now use the distribution of $\bar{X}$ on the null hypothesis (i.e. with $\theta = 0$), which is N(0, $1/n$). This gives

$$Z = \frac{\bar{X} - 0}{\sqrt{1/n}} \sim \mathrm{N}(0, 1)$$

and $\bar{X} > c$ corresponds to $Z > c\sqrt{n}$. At (for example) the level $\alpha = 0.05$, the upper tail of the N(0, 1) distribution begins at $z = 1.645$. So $c\sqrt{n} = 1.645$, or $c = 1.645/\sqrt{n}$.

Solution continued on next page

A sequential test is one that is carried out on data that are collected and studied one at a time. Each time a new observation is obtained, it is added to the existing analysis and a decision is made as to whether to accept the null hypothesis, accept the alternative hypothesis, or continue sampling.

In the example above, the total Σx_i (or equivalently the mean) would be examined at each step. If it exceeded a particular value which depends on n (the number of observations collected so far), the hypothesis representing the larger value of θ would be accepted; if it was less than another critical value, the hypothesis representing the smaller value of θ would be accepted.

These critical values depend on the levels chosen for α and β, the error probabilities.

The analysis is often carried out graphically. Sequential tests can lead to smaller samples on average than corresponding Neyman-Pearson tests.

Part (b)

The first statistician is prepared to assume that the differences d_i for each person between "before" and "after" measurements are observations from a Normal distribution. The two sets ("before" and "after") need not themselves individually follow Normal distributions. The test is based on the mean of the differences and its standard deviation.

The second statistician is making no distributional assumptions about either of the individual sets of data *or* about the differences between "before" and "after". Differences d_i are again used. They are ranked in order of absolute value, but also given the appropriate sign (+ or –). A suitable test statistic is the total of the ranks of the positive-signed differences (or of the negative-signed differences). Special tables are needed when there are fewer than about 20 pairs of data items, but there is an adequate Normal approximation to the distribution of the total for larger samples. When the paired t test is valid, it is more powerful; but the Wilcoxon test is more robust to non-Normality.

The value of pairing is that it removes personal systematic differences, such as some people naturally having longer reaction times (*both* "before" *and* "after") than others. For both test procedures, a one-sided alternative hypothesis is often appropriate, it commonly being the case that *either* a decrease *or* an increase, but not a mixture of the two, is being looked for.

<u>Graduate Diploma, Statistical Theory & Methods, Paper II, 2005. Question 8</u>

(i) Let x_1, x_2, …, x_n be the first sample. We have $\sum x_i = 0$ and

$$\frac{1}{n-1}\sum(x_i-\bar{x})^2 = \frac{1}{n-1}\sum x_i^2 = 1 \quad \text{so that} \quad \sum x_i^2 = n-1.$$

For the n + 1 observations including y, the sum is y and so the mean is $\frac{y}{n+1}$. Thus the sample variance is

$$\frac{1}{n+1-1}\left(\sum x_i^2 + y^2 - \frac{y^2}{n+1}\right) = \frac{1}{n}\left(n-1+\frac{ny^2}{n+1}\right) = 1-\frac{1}{n}+\frac{y^2}{n+1}.$$

(ii) Denote the mean and variance for the augmented sample by $\bar{X}$ and S^2. Then the usual t statistic is

$$t = \frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n+1}}} = \frac{\sqrt{n+1}\left(\frac{y}{n+1}-\mu_0\right)}{\sqrt{1-\frac{1}{n}+\frac{y^2}{n+1}}} = \frac{\frac{y}{\sqrt{n+1}}-\mu_0\sqrt{n+1}}{\sqrt{\frac{y^2}{n+1}+1-\frac{1}{n}}}.$$

Hence, as $y \to +\infty$, $t \to +1$; and as $y \to -\infty$, $t \to -1$. That is, as $|y| \to \infty$, $|t| \to 1$.

(iii) Taking the original sample mean as 0 and sample standard deviation as 1 does not affect the generality of the result, because this can always be achieved by a suitable linear transformation.

Hence, if there is a very extreme outlier, $|t| \approx 1$ for any μ_0 and any n.

So the null hypothesis will not be rejected at any of the usual significance levels; the power of the test is very low.