THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 2

SPECIMEN MATERIAL

The syllabus for Module 2 in the modular scheme is very similar to that for Statistical Theory and Methods Paper II ("STM II") in the old Graduate Diploma. Therefore many questions on past papers for STM II can be regarded as specimen material for Module 2. Past papers and solutions may be downloaded from the website at http://www.rss.org.uk/main.asp?page=1834 and in particular the following are suggested as being suitable:

2008 paper	All questions except question 7 (note that solutions to the 2008 paper are not available)
2007 paper	All questions except question 6
2006 paper	All questions except question 5 [a question of this nature could still be set, with appropriate guidance within the question]
2005 paper	All questions except question 4 parts (iii) and (iv) and question 7 part (a)

Specimen questions, with solutions, covering material that is in the syllabus for Module 2 but was not in that for STM II are set out below.

The time for the examination for Module 2 is 3 hours. The paper contains eight questions, of which candidates are to attempt **five**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%. All this also applies to examinations for STM II.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in 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 the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

Note. In accordance with the convention used in all 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 Module 2 – Specimen Question 1

A random sample $X_1, X_2, \ldots, X_n$ is drawn from a Bernoulli distribution for which

$$P(X_i = k) = \begin{cases} 1-p & k = 0 \\ p & k = 1 \\ 0 & \text{otherwise} \end{cases}$$

where p ($0 < p < 1$) is an unknown parameter.

(i) Show that the maximum likelihood estimator of p^2 is $\left(\sum_{i=1}^{n} X_i / n \right)^2$ ($= \hat{\theta}$, say).

(5)

(ii) Show that $\hat{\theta}$ is a biased estimator of p^2.

[Hint: $\left(\sum_{i=1}^{n} X_i \right)^2 = \sum_{i=1}^{n} X_i^2 + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} X_i X_j$.]

(4)

(iii) Show that the jack-knife estimator of p^2 based on $\hat{\theta}$ is $\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} X_i X_j$.

(8)

(iv) Show that this jack-knife estimator is unbiased for p^2.

(3)

Graduate Diploma Module 2 – Specimen Question 1 – Solution

(i) The likelihood is $L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\Sigma x_i}(1-p)^{n-\Sigma x_i}$.

$\therefore \log L(p) = \Sigma x_i \log p + (n - \Sigma x_i)\log(1-p)$. $\quad \therefore \frac{d}{dp}(\log L(p)) = \frac{\Sigma x_i}{p} - \frac{n-\Sigma x_i}{1-p}$.

Setting this equal to zero gives $\Sigma x_i - \hat{p}\Sigma x_i = n\hat{p} - \hat{p}\Sigma x_i$, i.e. $\hat{p} = \frac{\Sigma x_i}{n}$. To check that this is indeed a maximum, consider the second derivative of log$L(p)$:

$$\frac{d^2}{dp^2}(\log L(p)) = -\frac{\Sigma x_i}{p^2} - \frac{n-\Sigma x_i}{(1-p)^2}, \text{ which is clearly} < 0.$$

By the invariance property, it follows that the maximum likelihood estimator of p^2 is

$$(\hat{p})^2 = \left(\frac{\Sigma x_i}{n}\right)^2 \left(= \hat{\theta}, \text{ in the notation of the question}\right).$$

(ii) [Note. This part of the question can also be answered by observing that $\Sigma X_i \sim B(n, p)$.]

$E(X_i) = (0\times(1-p)) + (1\times p) = p$.

$E(X_i^2) = (0^2\times(1-p)) + (1^2\times p) = p$.

$$\therefore E(\hat{\theta}) = \frac{1}{n^2}E(\Sigma X_i)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n} E(X_i^2) + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} E(X_i X_j)\right)$$

[Note that X_i and X_j are independent for $i \neq j$]

$$= \frac{1}{n^2}(np + n(n-1)p^2) = p^2 + \frac{p(1-p)}{n} \neq p^2,$$

so $\hat{\theta}$ is a biased estimator of p^2.

Solution continued on next page

(iii) Let $\hat{\theta}_{\setminus i}$ be the above estimator based on data with X_i missing, i.e.

$$\hat{\theta}_{\setminus i} = \frac{1}{(n-1)^2}\left\{\sum_{j\neq i} X_j{}^2 + \sum_{j\neq i}\sum_{k\neq j \text{ or } i} X_j X_k\right\}.$$

The jack-knife estimator is then given by $\tilde{\theta}_J = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i$ where $\tilde{\theta}_i = n\hat{\theta} - (n-1)\hat{\theta}_{\setminus i}$.

We have $\tilde{\theta}_i = \frac{1}{n}\left(\sum X_j{}^2 + \sum_j\sum_{k\neq j} X_j X_k\right) - \frac{1}{n-1}\left(\sum_{j\neq i} X_j{}^2 + \sum_{j\neq i}\sum_{k\neq j \text{ or } i} X_j X_k\right)$,

and so $\tilde{\theta}_J = \frac{1}{n}\sum_{i=1}^{n}\tilde{\theta}_i$

$$= \frac{1}{n}\left(\sum X_j{}^2 + \sum_j\sum_{k\neq j} X_j X_k\right) - \frac{1}{n(n-1)}\left((n-1)\sum X_j{}^2 + (n-2)\sum_j\sum_{k\neq j} X_j X_k\right)$$

$$= \left(\frac{1}{n} - \frac{n-2}{n(n-1)}\right)\sum_j\sum_{k\neq j} X_j X_k = \frac{1}{n(n-1)}\sum_j\sum_{k\neq j} X_j X_k .$$

(iv) $$E\left(\hat{\theta}_J\right) = \frac{1}{n(n-1)}\sum_j\sum_{k\neq j} E\left(X_j X_k\right) = \frac{1}{n(n-1)}\, n(n-1)\, p^2 = p^2,$$

so $\hat{\theta}_J$ is an unbiased estimator of p^2.

Graduate Diploma Module 2 – Specimen Question 2

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent random samples from two populations. It is required to test the null hypothesis that the distributions of *X* and *Y* are the same against the alternative hypothesis that $P(X > Y) > \frac{1}{2}$.

(i) Explain how a *permutation test* might be carried out based on the difference of the two sample means.

(5)

(ii) Carry out the test described in part (i) at the 10% level for the case $m = 4$, $n = 3$, $X_1 = 8$, $X_2 = 13$, $X_3 = 6$, $X_4 = 9$, $Y_1 = 3$, $Y_2 = 7$ and $Y_3 = 2$.

(4)

(iii) Describe two other test statistics which are commonly used when a permutation test is carried out in this situation.

(4)

(iv) Describe how the permutation test referred to in part (i) could be modified to become a *randomisation test*. When might the randomisation test be preferred to the permutation test?

(3)

(v) Explain how *bootstrap sampling* might be used to carry out the test.

Suppose that the estimate of the *p*-value of the test based on *N* bootstrap samples is $\hat{p}$. Obtain an approximate 95% confidence interval for the true *p*-value when *N*, *m* and *n* are large. Given that the test was to have been carried out at a pre-specified significance level, describe briefly how this confidence interval might be used.

(4)

Graduate Diploma Module 2 – Specimen Question 2 – Solution

(i) Let $\bar{X}_0$ and $\bar{Y}_0$ be the sample means for the actual data and let $D_0 = \bar{X}_0 - \bar{Y}_0$.

Consider the data combined into a single sample of size $m + n$, i.e. (X_1, …, X_m, Y_1, …, Y_n). Altogether there are $\binom{m+n}{m}$ ways of choosing a sample of size m (without replacement).

Let $\bar{X}_i$ be the mean of the ith such sample and $\bar{Y}_i$ be the mean of the unselected items, with $D_i = \bar{X}_i - \bar{Y}_i$. Count the number of times D_i is greater than or equal to D_0, say ν times.

Then the p-value for the test is $\dfrac{\nu}{\binom{m+n}{m}}$.

As usual, if this is small then there is significant evidence against the null hypothesis.

(ii) We have $\bar{X}_0 = 9$ and $\bar{Y}_0 = 4$, so $D_0 = 5$.

The data are

*X*s			6		8	9	13
*Y*s	2	3		7			

.

By inspection, the only permutations that lead to $D_i \geq 5$ are as follows.

(1) the observed data (6, 8, 9, 13) and (2, 3, 7), for which $D_i = 5$

(2) the more extreme case (7, 8, 9, 13) and (2, 3, 6), for which $D_i = 5.6$ (5.5833)

Therefore the p-value is $\dfrac{2}{\binom{7}{4}} = \dfrac{2}{35} = 0.057$.

Thus, at the 10% level, there is significant evidence against the null hypothesis, i.e. significant evidence that the X_i tend to be larger.

Solution continued on next page

(iii) Permutation tests may be tailored to particular situations by choice of test statistic.

For example, if we are concerned about very extreme situations we might consider test statistics based on ranks. A common statistic of this kind is the (Wilcoxon) rank sum statistic. The $m + n$ observations are given ranks (1 for the smallest observation up to $m + n$ for the largest) and the test statistic is the sum of these ranks for the X observations. If the X_i tend to be larger than the Y_i, they will tend to get the larger ranks. So large values of this test statistic indicate this, i.e. they favour the alternative hypothesis that the X_i tend to be larger.

Another statistic that might be used is the familiar two-sample t statistic $\dfrac{\overline{X} - \overline{Y}}{s\sqrt{\frac{1}{m} + \frac{1}{n}}}$ where s is the pooled estimate of the assumed common variance. Again, large values favour the alternative hypothesis that the X_i tend to be larger.

(iv) First, D_0 is calculated as in part (i). Then select a simple random sample of the possible permutations and calculate D_i for each. The p-value is the proportion of the D_i that are greater than or equal to D_0.

This would be preferred to the permutation test when $\binom{m+n}{m}$ is so large as to make computation of the p-value in the permutation test problematic.

(v) Select m values at random, with replacement, from the observed data $X_1, X_2, \ldots, X_m$ and similarly n values at random, with replacement, from $Y_1, Y_2, \ldots, Y_n$. This gives the ith bootstrap sample. Calculate $D_i = \overline{X}_i - \overline{Y}_i$. Repeat this many times and proceed as above.

Let N be the number of bootstrap samples and $\hat{p}$ be the calculated p-value. Using the Normal approximation to the binomial, an approximate 95% confidence interval for the true p-value is given by $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/N}$.

If this confidence interval does not contain the pre-specified significance level, with the entire interval covering *smaller* values, this is an indication that the null hypothesis should be rejected. For example, suppose the significance level was 0.05 (5%) and the confidence interval turned out to be (0.015, 0.034): this would strongly suggest that the actual p-value was less than 0.05, i.e. the null hypothesis should be rejected. If not, the indication is that the null hypothesis should be accepted.

Graduate Diploma Module 2 – Specimen Question 3

Describe the relationship between *odds* and *probabilities*.

(3)

Let $X_1, X_2, \ldots, X_n$ be a random sample from the exponential distribution with density $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and $\lambda > 0$.

(i) We wish to test H_0: $\lambda = 1$ against H_1: $\lambda = 2$, where the prior odds of H_0 is ω (> 0).

(a) Find the Bayes factor and show how it can be used to obtain the posterior odds of H_0.

(5)

(b) Under what circumstances will the posterior odds of H_0 exceed ω?

(3)

(ii) Suppose now that we wish to test H_0: $\lambda = 1$ against H_1: $\lambda \neq 1$, where the prior distribution of λ under H_1 has density

$$\pi(\lambda) = \frac{\nu^\alpha \lambda^{\alpha-1}}{\Gamma(\alpha)} e^{-\nu\lambda}, \quad \lambda > 0, \alpha > 0 \text{ and } \nu > 0,$$

and where the prior odds of H_0 is ω (> 0). Find the Bayes factor and the posterior odds of H_0.

(9)

Graduate Diploma Module 2 – Specimen Question 3 – Solution

If p is the probability of an event, then the odds $\omega = p/(1-p)$. [Note: in statistical work, odds refers to "odds on"; the reciprocal, "odds against", is used in gambling.]

Alternatively, solving for p for given odds, we have $p = \omega/(1+\omega)$.

p and ω can be interpreted as follows:

p is the long-run proportion of times an event occurs

$$\omega = \frac{\text{proportion of times event occurs}}{\text{proportion of times event does not occur}}$$

Parts (i) and (ii) refer to Bayesian inference: $\pi(\theta|x) = \dfrac{f(x|\theta)\pi(\theta)}{h(x)}$.

(i) (a) We have $\pi(H_0|x) = \dfrac{f(x|H_0)\pi(H_0)}{h(x)}$ and $\pi(H_1|x) = \dfrac{f(x|H_1)\pi(H_1)}{h(x)}$.

Dividing the first of these by the second,

$$\frac{\pi(H_0|x)}{\pi(H_1|x)} = \frac{\pi(H_0)}{\pi(H_1)} \times \frac{f(x|H_0)}{f(x|H_1)},$$

i.e. posterior odds = prior odds × Bayes factor.

In this case, the Bayes factor is $\dfrac{e^{-\Sigma x}}{2^n e^{-2\Sigma x}} = \dfrac{e^{\Sigma x}}{2^n}$.

(b) Posterior odds $> \omega$ if Bayes factor > 1.

Bayes factor = $\left(\dfrac{e^{\Sigma x/n}}{2}\right)^n$ which is > 1 if $e^{\Sigma x/n} > 2$ i.e. if $\bar{X} > \log 2 = 0.693$.

Solution continued on next page

(ii) We have $f\left(x\middle|\lambda\right)=\lambda^n e^{-\lambda\Sigma x_i}$, so $f\left(x\middle|H_0\right)=e^{-\Sigma x_i}$ [as in part (i)] and

$$f\left(x\middle|H_1\right)=\int_{\lambda=0}^{\infty}\lambda^n e^{-\lambda\Sigma x_i}\,\frac{\nu^\alpha\lambda^{\alpha-1}}{\Gamma(\alpha)}e^{-\nu\lambda}\,d\lambda$$

$$=\int_0^{\infty}\frac{\nu^\alpha\lambda^{n+\alpha-1}}{\Gamma(\alpha)}e^{-\lambda(\nu+\Sigma x_i)}\,d\lambda$$

$$=\frac{\nu^\alpha}{\Gamma\left(\alpha\right)}\times\frac{\Gamma\left(n+\alpha\right)}{\left(\nu+\Sigma x_i\right)^{n+\alpha}}\int_0^{\infty}\frac{\left(\nu+\Sigma x_i\right)^{n+\alpha}\,\lambda^{n+\alpha-1}}{\Gamma(n+\alpha)}e^{-\lambda(\nu+\Sigma x_i)}\,d\lambda$$

The integrand is the pdf of the gamma distribution with parameters $n+\alpha$ and $\nu+\Sigma x_i$

$$=\frac{\nu^\alpha}{\Gamma\left(\alpha\right)}\times\frac{\Gamma\left(n+\alpha\right)}{\left(\nu+\Sigma x_i\right)^{n+\alpha}}.$$

So the Bayes factor is $\dfrac{f\left(x\middle|H_0\right)}{f\left(x\middle|H_1\right)}=\dfrac{e^{-\Sigma x_i}\,\Gamma\left(\alpha\right)\left(\nu+\Sigma x_i\right)^{n+\alpha}}{\nu^\alpha\,\Gamma\left(n+\alpha\right)}$

and the posterior odds $=\dfrac{\omega\,e^{-\Sigma x_i}\,\Gamma\left(\alpha\right)\left(\nu+\Sigma x_i\right)^{n+\alpha}}{\nu^\alpha\,\Gamma\left(n+\alpha\right)}$.

Graduate Diploma Module 2 – Specimen Question 4

(a) The times in minutes T_1, T_2, …, T_n between calls arriving at a switchboard are independent and have an exponential distribution, mean $1/\lambda$, where λ is an unknown parameter whose prior density is assumed to be $\pi(\lambda) = e^{-\lambda}$ ($\lambda > 0$).

(i) Find the posterior distribution of λ.

(6)

(ii) Find $\pi(t_{n+1} \mid t_1, t_2, \ldots, t_n)$, the predictive distribution of a future time T_{n+1}.

(7)

[Note that the gamma distribution with parameters ν (> 0) and α (> 0) has probability density function $f(x) = \dfrac{\nu^{\alpha} x^{\alpha-1}}{\Gamma(\alpha)} e^{-\nu x}$, for $x > 0$.]

(b) At another switchboard, the times in minutes S_1, S_2, …, S_n between calls arriving are also independent, but have a different continuous distribution with distribution function $F(s; \theta)$ where θ (> 0) is an unknown parameter with prior density $\pi(\theta)$. A simple formula for the posterior distribution of θ cannot be found, but a computer routine is available which will generate independent values θ_1, θ_2, θ_3, … from the posterior distribution. If this routine is used to generate 10 000 such values, explain how these values θ_1, θ_2, …, $\theta_{10\,000}$ may be used to

(i) find an approximate 95% Bayesian interval for θ,

(3)

(ii) find the Bayes estimate (assuming squared error loss) of the probability that a future time S_{n+1} will exceed 2 minutes.

(4)

Graduate Diploma Module 2 – Specimen Question 4 – Solution

[Solution continues on next page]

Part (a)

(i) $f(t_i;\lambda)=\lambda e^{-\lambda t_i}$ (for $t_i>0$, and $\lambda>0$).

$\therefore$ likelihood $L(\lambda)=\prod_{i=1}^{n} f(t_i;\lambda)=\lambda^n e^{-\lambda\Sigma t_i}$.

The posterior density is $\pi(\lambda|t_1,t_2,...,t_n)\propto\pi(\lambda)L(\lambda)=e^{-\lambda}\lambda^n e^{-\lambda\Sigma t_i}=\lambda^n e^{-\lambda(1+\Sigma t_i)}$.

From the form of the pdf quoted in the question, we see that $\lambda|t_1,t_2,...,t_n$ has the gamma distribution with parameters $\alpha=n+1$ and $\nu=1+\Sigma t_i$, so that its pdf is

$$\pi(\lambda|t_1,t_2,...,t_n)=\frac{\left(1+\sum_{i=1}^{n}t_i\right)^{n+1}}{n!}\lambda^n e^{-\lambda\left(1+\sum_{i=1}^{n}t_i\right)} \quad \text{(for } \lambda>0\text{)}.$$

(ii) The predictive distribution has

$$\pi(t_{n+1}|t_1,t_2,...,t_n)=\int_0^\infty f(t_{n+1};\lambda)\pi(\lambda|t_1,t_2,...,t_n)\,d\lambda$$

$$=\int_0^\infty \lambda e^{-\lambda t_{n+1}}\frac{\left(1+\sum_{i=1}^{n}t_i\right)^{n+1}}{n!}\lambda^n e^{-\lambda\left(1+\sum_{i=1}^{n}t_i\right)}\,d\lambda$$

$$=\left(1+\sum_{i=1}^{n}t_i\right)^{n+1}\int_0^\infty\frac{\lambda^{n+1}}{n!}e^{-\lambda t_{n+1}}e^{-\lambda\left(1+\sum_{i=1}^{n}t_i\right)}\,d\lambda$$

$$=\frac{\left(1+\sum_{i=1}^{n}t_i\right)^{n+1}(n+1)!}{\left(1+\sum_{i=1}^{n}t_i+t_{n+1}\right)^{n+2}n!}\int_0^\infty\frac{\left(1+\sum_{i=1}^{n}t_i+t_{n+1}\right)^{n+2}}{\Gamma(n+2)}\lambda^{n+1}e^{-\lambda\left(1+\sum_{i=1}^{n}t_i+t_{n+1}\right)}\,d\lambda$$

The integrand is the pdf of the gamma distribution with parameters $\alpha=n+2$ and $\nu=1+\sum_{i=1}^{n}t_i+t_{n+1}$

$$=\frac{(n+1)\left(1+\sum_{i=1}^{n}t_i\right)^{n+1}}{\left(1+\sum_{i=1}^{n}t_i+t_{n+1}\right)^{n+2}} \quad \text{(for } t_{n+1}>0\text{)}.$$

Part (b)

(i) Order the values (smallest to largest) $\theta_{(1)}, \theta_{(2)}, ..., \theta_{(10000)}$. Then an approximate 95% Bayesian interval for θ is $\theta_{(251)}$ to $\theta_{(9750)}$.

(ii) $P\left(S_{n+1} > 2 \mid \theta\right) = 1 - F\left(2; \theta\right)$.

With quadratic (i.e. squared error) loss, the Bayes estimate of this is the expected value with respect to the posterior distribution of θ given $s_1, s_2, ..., s_n$.

Using the simulated values, this expectation can be approximated by

$$\frac{\sum_{i=1}^{10000}\left(1 - F\left(2; \theta_i\right)\right)}{10000}.$$