THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 1

SPECIMEN MATERIAL

The syllabus for Module 1 in the modular scheme is very similar to that for Statistical Theory and Methods Paper I ("STM I") in the old Graduate Diploma. Therefore many questions on past papers for STM I can be regarded as specimen material for Module 1. 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 4 and those parts of question 7 that use the multinomial distribution (note that solutions to the 2008 paper are not available)
2007 paper	All questions except question 8
2006 paper	All questions except question 8 parts (ii) and (iii)

Specimen questions, with solutions, covering material that is in the syllabus for Module 1 but was not in that for STM I are set out below.

The time for the examination for Module 1 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 I.

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 1 – Specimen Question 1

Explain briefly why it might be reasonable to use a Poisson distribution to model the number of claims made over a year on a particular type of insurance policy.

(4)

An insurance company assumes that the total number of claims its clients make over a year follows a Poisson distribution with mean λ, and that the sizes of individual claims (in £) are independent and identically distributed with mean μ and variance σ^2.

(i) Derive expressions for the mean and variance of the total cost (£) of all claims over a year under this model.

[You may use without proof the result $\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$.]

(8)

(ii) For each of the following fields of insurance, state whether or not the assumption that the sizes of all claims are independent is a reasonable one, justifying your answer.

(a) UK motor car insurance, with claims for accident repairs.

(b) Property damage claims in a particular US city liable to hurricanes.

(8)

Graduate Diploma Module 1 – Specimen Question 1 – Solution

It is reasonable to assume that there is a large number of clients, each with a small chance of making a claim during a year, with claims arising largely at random. The numbers of claims in non-overlapping short time periods can reasonably be expected to be independent, and in a very short time period the number can be expected to be, to first order, proportional to the length of that period. This is a standard situation where the Poisson distribution is likely to apply.

(i) Let X denote the number of claims and Y the total cost of all claims. μ and σ^2 denote respectively the mean and variance of the size of any individual claim. As X follows a Poisson distribution, λ is both the mean and the variance of X.

Clearly $E(Y|X = n) = n\mu$ and thus $E(Y|X) = X\mu$. Therefore the formula $E(Y) = E(E(Y|X))$ gives $E(Y) = E(X\mu) = E(X).\mu = \lambda\mu$, and this is the mean of the total cost.

For the variance of the total cost, we use the formula (given in the question) $\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$.

We have $\text{Var}(Y|X = n) = n\sigma^2$, so $\text{Var}(Y|X) = X\sigma^2$. $\therefore$ $E(\text{Var}(Y|X)) = \text{E}(X).\sigma^2 = \lambda\sigma^2$.

Also, $E(Y|X) = X\mu$ [see above], so $\text{Var}(E(Y|X)) = \text{Var}(X\mu) = \mu^2\text{Var}(X) = \mu^2\lambda$.

Therefore $\text{Var}(Y) = \lambda(\sigma^2 + \mu^2)$; this is the variance of the total cost.

(ii) (a) For the case of UK motor car insurance, with claims for accident repairs, the assumption is reasonable. The number and geographical and time spreads of claims are such that the size of any one claim will be unrelated to the size of almost all other claims.

(b) For the case of property damage claims in a particular US city liable to hurricanes, the assumption is unreasonable. It seems likely that there would *either* be hardly any claims (and mostly small and unrelated ones) *or*, if a hurricane did strike, a large number of large (and often inter-related) claims.

Graduate Diploma Module 1 – Specimen Question 2

(i) Suppose that the random variable U has the continuous uniform distribution on the interval (0, 1). Show that the random variable $X = -\log U$ has the standard exponential distribution with density function $f(x) = \exp(-x)$ on $x > 0$.

(4)

(ii) Let $g(x)$ be the density function of $X = |Z|$ where Z has the standard Normal distribution, so that $g(x) = \sqrt{(2/\pi)} \exp(-x^2/2)$ on $x \geq 0$.

Show that $\exp\{x - (x^2/2)\}$ takes its maximum value when $x = 1$. Deduce that, for $k = \sqrt{(2e/\pi)}$, $kf(x) \geq g(x)$ for all $x \geq 0$.

(6)

(iii) Suppose you have a supply of values $u_1, u_2, u_3, \ldots$ from independent random variables $U_1, U_2, U_3, \ldots$ respectively, each of which has the distribution of U in part (i). Using the value of k and the inequality $kf(x) \geq g(x)$ shown in part (ii), describe a rejection method for generating values from the distribution of $|Z|$.

(6)

(iv) Describe further steps that would lead to the generation of values from a Normal distribution with mean μ and variance σ^2.

(4)

Graduate Diploma Module 1 – Specimen Question 2 – Solution

(i) For $x > 0$, $P(X < x) = P(-\log U < x) = P(U > e^{-x}) = 1 - P(U < e^{-x}) = 1 - e^{-x}$. This is the cdf of X, so the pdf is the derivative, i.e. e^{-x} as required.

(ii) Let $h(x) = \exp\{x - (x^2/2)\}$, so that $h'(x) = (1 - x)h(x)$. Since $h(x) > 0$ for all x, its only turning point is at $x = 1$. This is a maximum (eg note that $h'(x) > 0$ for $x < 1$ whereas $h'(x) < 0$ for $x > 1$). Plainly this maximum value is $\sqrt{e}$.

When $k = \sqrt{(2e/\pi)}$, we have

$$kf(x) - g(x) = \sqrt{(2e/\pi)}.\exp(-x) - \sqrt{(2/\pi)}.\exp(-x^2/2) = \exp(-x)\sqrt{(2/\pi)}\{\sqrt{e} - h(x)\}.$$

But we have seen that $h(x) \le \sqrt{e}$ for all x. So $kf(x) \ge g(x)$ for all $x \ge 0$, as required.

(iii) The required rejection method is as follows. First, use the value u_1 to generate a value x from the standard exponential distribution by the method shown in part (i). Then set $y = ku_2\exp(-x)$; accept x as a value from the distribution of $|Z|$ if $y < g(x) = \sqrt{(2/\pi)}\exp(-x^2/2)$, otherwise return to the first step, taking the next u_i value.

(iv) To obtain values from $N(\mu, \sigma^2)$, first generate a value z (note that $z \ge 0$) from $|Z|$ as described in part (iii). Now let v be the next (unused) value from the stream U_i. Replace z by $-z$ if $v < 0.5$, otherwise keep the current value of z. Then $\mu + \sigma z$ is a value from the desired $N(\mu, \sigma^2)$ distribution.

Graduate Diploma Module 1 – Specimen Question 3

Suppose that a sequence of n independent Bernoulli trials is carried out, each with success probability p. Let $W_i = 1$ if the ith trial is a success and $W_i = 0$ if the ith trial is not a success, for $i = 1, 2, \ldots, n$. Let X denote the total number of successes. Write down the expected value and variance of W_i. Hence show that $E(X/n) = p$ and find the variance of X/n.

(4)

Write $Y = \arcsin(\sqrt{(X/n)})$. Use a first order Taylor series method to show that, approximately, the variance of Y is $1/(4n)$, whatever the value of p.

(11)

In the case when $n = 4$, find the exact distribution of Y.

(5)

Graduate Diploma Module 1 – Specimen Question 3 – Solution

Write $X = W_1 + W_2 + \ldots + W_n$ where the W_i are as defined in the question.

For each i, $E(W_i) = [1 \times p] + [0 \times (1 - p)] = p$ and similarly $E(W_i^2) = p$, so $\text{Var}(W_i) = p - p^2$.

So $E(X) = np$ and therefore $E(X/n) = p$, and $\text{Var}(X) = n(p - p^2) = np(1 - p)$ and therefore $\text{Var}(X/n) = (1/n^2)\text{Var}(X) = p(1 - p)/n$.

Consider the Taylor series about p of the function $f(t) = \arcsin\left(\sqrt{t}\right)$, which begins with

$$f(t) = f(p) + (t-p)f'(p) \quad \text{with} \quad f'(t) = \frac{1}{\sqrt{1-t}}.\frac{1}{2}t^{-1/2} \quad \text{so that} \quad f'(p) = \frac{1}{2\sqrt{p(1-p)}}.$$

We have, approximately,

$$f(t) - f(p) = (t-p)f'(p).$$

Now taking t as a random variable (T) with mean p, we have, approximately,

$$E\left(f(T) - f(p)\right) = f'(p)E(T-p) = 0, \text{ so that } E\left(f(T)\right) \approx f(p).$$

Now squaring both sides of the approximation $f(t) - f(p) = (t-p)f'(p)$ and taking expectations, we get

$$E\left[\left(f(T) - f(p)\right)^2\right] \approx E\left[(T-p)^2.\left(f'(p)\right)^2\right],$$

i.e. $\text{Var}\left(f(T)\right) \approx \left[f'(p)\right]^2 \text{Var}(T)$.

Applying this with $T = X/n$ gives $\text{Var}\left(\arcsin\sqrt{X/n}\right) \approx \dfrac{1}{4p(1-p)}\dfrac{p(1-p)}{n} = \dfrac{1}{4n}$, which is the required result.

When $n = 4$, the possible values taken by Y are $\arcsin(\sqrt{(i/4)})$ for $i = 0, 1, 2, 3, 4$, which are 0, $\pi/6$, $\pi/4$, $\pi/3$ and $\pi/2$.

The corresponding probabilities are immediately found from the binomial distribution with parameters 4 and p and are respectively as follows (with q denoting $1 - p$): q^4, $4pq^3$, $6p^2q^2$, $4p^3q$, p^4.

Graduate Diploma Module 1 – Specimen Question 4

(i) Suppose that U has the continuous uniform distribution over the interval [0, 1] and that f is a continuous function defined over the same interval. Explain briefly why

$$E\left(f(U)\right) \equiv I = \int_0^1 f\left(u\right) du$$

and

$$\text{Var}\left(f(U)\right) \equiv \sigma^2 = \int_0^1 \left(f(u)\right)^2 du - I^2 \,.$$

(4)

(ii) Deduce that

$$E\left(\frac{f(U) + f(1-U)}{2} \right) = I \,,$$

and give an expression for the variance of $\dfrac{f(U) + f(1-U)}{2}$ in terms of σ^2 and τ, the covariance of $f(U)$ and $f(1-U)$.

(6)

(iii) A statistics teacher notes that $\int_0^1 \sin(\pi x/2) dx = 2/\pi$, and decides to illustrate the notions of simulation and Monte Carlo methods by using the properties above to get an approximate value of $2/\pi$. Thus, suppose that $\{U_i,\ i = 1, 2, \ldots\}$ are independent, all having the same distribution as U, and, with $f(u) = \sin(\pi u/2)$, write

$$J_n = \frac{\sum_{i=1}^{2n} f(U_i)}{2n} \quad \text{and} \quad K_n = \frac{\sum_{i=1}^{n} \left(f(U_i) + f(1-U_i)\right)}{2n} \,.$$

Show that $E\left(J_n\right) = E\left(K_n\right) = 2/\pi$. Prove that Var($K_n$) is less than Var($J_n$). What is the practical implication of these results?

(10)

Graduate Diploma Module 1 – Specimen Question 4 – Solution

(i) For a general random variable X with probability density function $g(x)$, by definition $E\left(f(X)\right) = \int f(x)g(x)dx$ between appropriate limits. Here, X is the random variable U with density 1 over [0, 1] (and zero elsewhere). So $E\left(f(U)\right) = \int_0^1 f(u)du$ which is I as defined in the question.

Similarly, as a general result $E\left[\left(f(X)\right)^2\right] = \int \left(f(x)\right)^2 g(x)dx$, so here we have $E\left[\left(f(U)\right)^2\right] = \int_0^1 \left(f(u)\right)^2 du$. Therefore we have

$$\text{Var}\left(f(U)\right) = E\left[\left(f(U)\right)^2\right] - \left\{E\left(f(U)\right)\right\}^2 = \int_0^1 \left(f(u)\right)^2 du - I^2 .$$

(ii) As U is uniformly distributed over [0, 1], it is clear that 1 – U has the same distribution as U. Therefore

$$E\left(\frac{f(U) + f(1-U)}{2}\right) = \frac{I}{2} + \frac{I}{2} = I$$

and

$$\text{Var}\left(\frac{f(U) + f(1-U)}{2}\right) = \frac{1}{4}\left\{\text{Var}\left(f(U)\right) + \text{Var}\left(f(1-U)\right) + 2\text{Cov}\left(f(U),\ f(1-U)\right)\right\}$$

$$= \frac{1}{4}\left\{\sigma^2 + \sigma^2 + 2\tau\right\} = \frac{\sigma^2 + \tau}{2} .$$

Solution continued on next page

(iii) Clearly

$$E\left(J_n\right)=\frac{1}{2n}\sum_{i=1}^{2n}E\left(f(U_i)\right)=\frac{1}{2n}.2n\,E\left(f(U)\right)=\frac{2nI}{2n}=I=\frac{2}{\pi}$$

and similarly

$$E\left(K_n\right)=\frac{1}{2n}\sum_{i=1}^{n}\left(E\left(f(U_i)\right)+E\left(f(1-U_i)\right)\right)=\frac{1}{2n}\left(nI+nI\right)=I=\frac{2}{\pi}.$$

So $E\left(J_n\right)=E\left(K_n\right)=2/\pi$, as required.

We also have, from work above,

$$\mathrm{Var}\left(J_n\right)=\frac{\sigma^2}{2n} \quad \text{and} \quad \mathrm{Var}\left(K_n\right)=\frac{\sigma^2+\tau}{2n}.$$

We need to show that the second of these is less than the first, which requires us to show that τ, the covariance of $f(U)$ and $f(1-U)$, is negative.

We have

$$\tau \;=\; E\left(f(U)f(1-U)\right)-E\left(f(U)\right).E\left(f(1-U)\right) \;=\; E\left(f(U)f(1-U)\right)-\left(\frac{2}{\pi}\right)^2$$

and

$$E\left(f(U)f(1-U)\right) \;=\; \int_0^1 \sin\left(\frac{\pi u}{2}\right)\sin\left(\frac{\pi(1-u)}{2}\right)du$$

Use the identity
$2\sin\theta\sin\phi=\cos\left(\theta-\phi\right)-\cos\left(\theta+\phi\right)$
and note that $\cos(\pi/2)=0$

$$=\;\frac{1}{2}\int_0^1\cos\left(\frac{\pi}{2}-\pi u\right)du \;=\; \frac{1}{2}\int_0^1\sin(\pi u)du \;=\frac{1}{2}\frac{2}{\pi}\;.$$

Hence $\tau=\dfrac{1}{\pi}-\dfrac{4}{\pi^2}$ which is clearly negative.

Thus, while J_n and K_n both give the value $2/\pi$ "on average", values of K_n should be closer to $2/\pi$ than values of J_n.