EXAMINATION

22 April 2010 (am)

Subject CT3 — Probability and Mathematical Statistics Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 12 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is NOT required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 The mean height of the women in a large population is 1.671m while the mean height of the men in the population is 1.758m. The mean height of all the members of the population is 1.712m.

Calculate the percentage of the population who are women. [2]

2 Consider a group of 10 life insurance policies, seven of which are on male lives and three of which are on female lives. Three of the 10 policies are chosen at random (one after the other, without replacement).

Find the probability that the three selected policies are all on male lives. [2]

3 Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with mean μ and variance σ^2.

Let the sample mean be $\bar{X}$ and the sample variance be $S^2 = \frac{1}{n-1}\{\Sigma X_i^2 - n\bar{X}^2\}$.

You may assume that $E\left[\bar{X}\right] = \mu$ and $V\left[\bar{X}\right] = \frac{\sigma^2}{n}$.

Show that $E\left[S^2\right] = \sigma^2$. [3]

4 It is assumed that the numbers of claims arising in one year from motor insurance policies for young male drivers and young female drivers are distributed as Poisson random variables with parameters λ_m and λ_f respectively.

Independent random samples of 120 policies for young male drivers and 80 policies for young female drivers were examined and yielded the following mean number of claims per policy in the last calendar year: $\bar{x}_m = 0.24$ and $\bar{x}_f = 0.15$.

Calculate an approximate 95% confidence interval for $\lambda_m - \lambda_f$, the difference between the respective Poisson parameters. [3]

5 A computer routine selects one of the integers 1, 2, 3, 4, 5 at random and replicates the process a total of 100 times. Let S denote the sum of the 100 numbers selected.

Calculate the approximate probability that S assumes a value between 280 and 320 inclusive. [5]

6 Let $X_1, X_2, \ldots, X_n$ be a random sample of claim amounts which are modelled using a gamma distribution with known parameter $\alpha = 4$ and unknown parameter λ.

(i) (a) Specify the distribution of $\sum_{i=1}^{n} X_i$.

(b) Justify the fact that $2n\lambda\bar{X}$ has a χ^2_k distribution, where $\bar{X}$ is the mean of the sample, by using a suitable relationship between the gamma and the χ^2 distribution, and specify the degrees of freedom k.

[3]

A random sample of five such claim amounts yields a mean of $\bar{x} = 17.5$.

(ii) Use the pivotal method with the χ^2 result from part (i)(b) to obtain a 95% confidence interval for λ. [3]

[Total 6]

7 An employment survey is carried out in order to determine the percentage, p, of unemployed people in a certain population in a way such that the estimation has a margin of error less than 0.5% with probability at least 0.95. In a similar study conducted a year ago it was found that the percentage of unemployed people in the population was 6%.

Calculate the sample size, n, that is required to achieve this margin of error, by constructing an appropriate confidence interval (or otherwise). [6]

8 For a sample of 100 insurance policies the following frequency distribution gives the number of policies, f, which resulted in x claims during the last year:

x:	0	1	2	3
f:	76	22	1	1

(i) Calculate the sample mean, standard deviation and coefficient of skewness for these data on the number of claims per policy.

[4]

A Poisson model has been suggested as appropriate for the number of claims per policy.

(ii) (a) State the value of the estimated parameter when a Poisson distribution is fitted to these data using the method of maximum likelihood.

(b) Verify that the coefficient of skewness for the fitted model is 1.92, and hence comment on the shape of the frequency distribution relative to that of the corresponding fitted Poisson distribution.

[3]

[Total 7]

9 The number of claims, N, arising over a period of five years for a particular policy is assumed to follow a "Type 2" negative binomial distribution (as in the book of Formulae and Tables page 9) with mean $E[N] = \frac{k(1-p)}{p}$ and variance $V[N] = \frac{k(1-p)}{p^2}$.

Each claim amount, X (in units of £1,000), is assumed to follow an exponential distribution with parameter λ independently of each other claim amount and of the number of claims.

Let S be the total of the claim amounts for the period of five years, in the case $k = 2$, $p = 0.8$ and $\lambda = 2$.

(i) Calculate the mean and the standard deviation of S based on the above assumptions. [4]

Now assume that:

N follows a Poisson distribution with parameter $\mu = 0.5$, that is, with the same mean as N above;

X follows a gamma distribution with parameters $\alpha = 2$ and $\lambda = 4$, that is, with the same mean as X above.

(ii) Calculate the mean and the standard deviation of S based on these assumptions. [3]

(iii) Compare the two sets of answers in (i) and (ii) above. [2]

[Total 9]

10 The size of claims (in units of £1,000) arising from a portfolio of house contents insurance policies can be modelled using a random variable X with probability density function (pdf) given by:

$$f_X(x) = \frac{ac^a}{x^{a+1}}, \quad x \geq c$$

where $a > 0$ and $c > 0$ are the parameters of the distribution.

(i) Show that the expected value of X is $E[X] = \frac{ac}{a-1}$, for $a > 1$. [2]

(ii) Verify that the cumulative distribution function of X is given by

$$F_X(x) = 1 - \left(\frac{c}{x}\right)^a, \quad x \geq c \quad \text{(and = 0 for } x < c\text{)}.$$ [2]

Suppose that for the distribution of claim sizes X it is known that $c = 2.5$, but a is unknown and needs to be estimated given a random sample $x_1, x_2, \ldots, x_n$.

(iii) Show that the maximum likelihood estimate (MLE) of a is given by:

$$\hat{a} = \frac{n}{\sum_{i=1}^{n} \log\left(\frac{x_i}{2.5}\right)}.$$ [3]

(iv) Derive the asymptotic variance of the MLE $\hat{a}$, and hence determine its approximate asymptotic distribution. [4]

Consider a sample of 30 observations from this distribution, for which:

$$\sum_{i=1}^{30} \log(x_i) = 32.9.$$

(v) Calculate the MLE $\hat{a}$ in this case, together with an approximate 95% confidence interval for a. [5]

In the current year, claim sizes are assumed to follow the distribution of X with $a = 6$, $c = 2.5$. Inflation for the following year is expected to be 5%.

(vi) Calculate the probability that the size of a claim arising from this portfolio in the following year will exceed £4,000. [3]

[Total 19]

11 Consider the following three independent random samples from a normally distributed population with unknown mean μ:

Sample 1:

19.9 20.4 20.3 22.3 16.7 18.7 20.5 19.0 20.1 16.4 21.5 21.4 17.8 22.5 15.2

For these data: $n = 15$, $\sum x_i = 292.7$, $\sum x_i^2 = 5,778.69$

Sample 2:

20.8 25.9 22.1 21.7 16.0 12.1 27.6 16.1 16.8 17.1 21.3 18.6 24.9 14.8 22.2

For these data: $n = 15$, $\sum x_i = 298.0$, $\sum x_i^2 = 6,192.32$
sample mean = 19.867, sample variance = 19.432

Sample 3:

20.6 18.5 21.5 16.9 21.5 21.2 20.9 22.4 14.5 22.0 20.2 17.0 20.3 23.0 19.3
18.9 20.6 20.9 15.3 21.5 16.8 18.5 21.6 16.8 20.4

For these data: $n = 25$, $\sum x_i = 491.1$, $\sum x_i^2 = 9,773.77$
sample mean = 19.644, sample variance = 5.275

Consider *t*-tests of the hypotheses H_0: $\mu = 18$ v H_1: $\mu > 18$.

(i) (a) Calculate the sample mean and variance for Sample 1.

(b) Carry out a *t*-test of the stated hypotheses using the Sample 1 data (stating the approximate *P*–value) and show that H_0 can be rejected at the 1% level of testing.

[6]

(ii) (a) Carry out a *t*-test of the stated hypotheses using the Sample 2 data (stating the approximate *P*–value and the conclusion clearly).

(b) Discuss the comparison of the results with those based on Sample 1 (include reasons for any difference or similarity in the test conclusions).

[6]

(iii) (a) Carry out a *t*-test of the stated hypotheses using the Sample 3 data (stating the approximate *P*–value and your conclusion clearly).

(b) Discuss the comparison of the results with those based on Sample 1 (include reasons for any difference or similarity in the test conclusions).

[6]
[Total 18]

12 As part of a project in a modelling module, a statistics student is required to submit a report on the sums insured on home contents insurance policies based on samples of such policies covering risks in five medium-sized towns in each of England, Wales, and Scotland. Data are provided on the average sum insured (Y, in units of £1,000) for each of the 15 towns and are as follows:

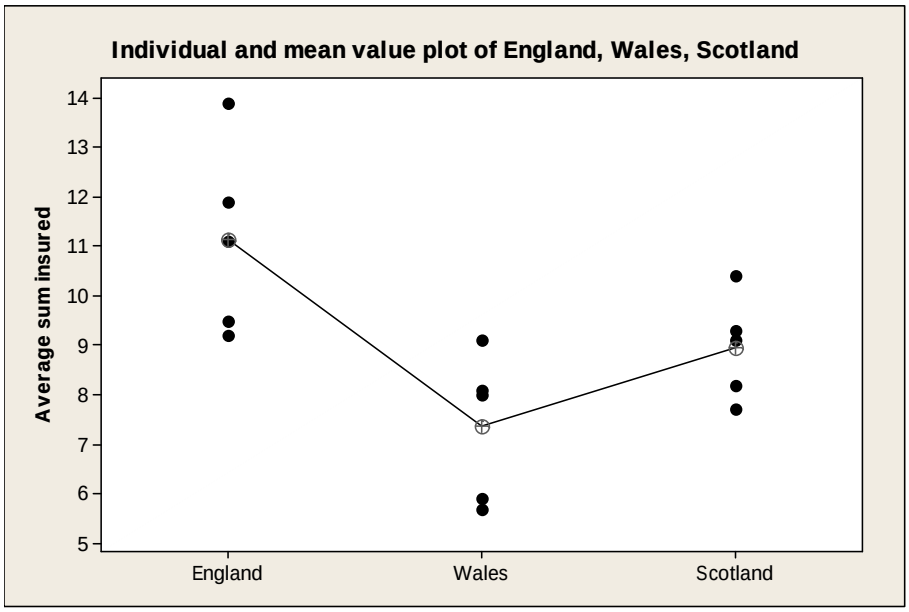
	England					*Wales*					*Scotland*				
y	11.9	11.1	9.5	9.2	13.9	5.9	9.1	8.0	5.7	8.1	9.3	9.1	7.7	8.2	10.4

For these data: $\sum y$ = 55.6 (England), 36.8 (Wales), 44.7 (Scotland)

overall $\sum y = 137.1$, $\sum y^2 = 1{,}316.63$

The student decides to use an analysis of variance approach.

(i) Suggest brief comments the student should make on the basis of the plot below:



[2]

(ii) (a) Carry out the analysis of variance on the average sums insured.
(b) Comment on your conclusions.

[6]

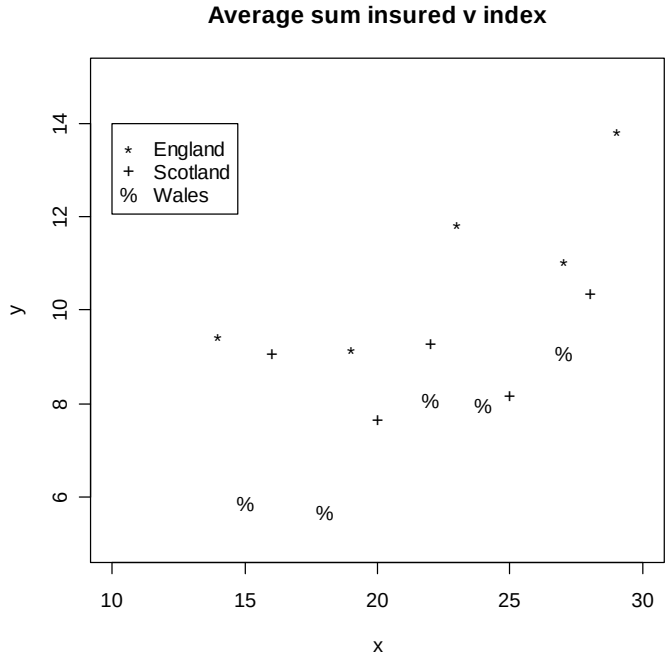
The lecturer of the module decides to provide further information. It has been suggested that the value of a UK index of the town's prosperity (X) might also be a useful explanatory variable (in addition to the country in which the town is situated).

The data on the index are as follows (for the towns in the same order as in the first table):

	England					*Wales*					*Scotland*				
x	23	27	14	19	29	15	27	24	18	22	22	16	20	25	28

For these data: overall $\sum x = 329$, $\sum x^2 = 7{,}543$, $\sum xy = 3{,}091.7$

A graph of average sum insured against index (with country identified) is given below:



The student decides to add the results of a regression approach to her report, using "index" as an explanatory variable, so she fits the regression model

$$Y = a + bX + e$$

using the least squares criterion.

Part of the output from fitting the model using a statistics package on a computer is as follows:

Coefficients:	*Estimate*		*Std. Error*	*t value*	*Pr(>|t|)*
	(Intercept)	3.46166	2.21919	1.560	0.1428
	x	0.25889	0.09896	2.616	0.0213*
	Residual standard error:		1.789 on 13 degrees of freedom		
	R-Squared: 0.3449				

(iii) Verify (by performing your own calculations) the following results for the fitted model as given in the output above:

(a) the fitted regression line is $y = 3.462 + 0.2589x$

(b) the percentage of the variation in the response (y) explained by the model (x) is 34.5%

(c) the standard error of the slope estimate is 0.09896 [8]

(iv) Comment briefly on the usefulness of "index" as a predictor for the average sum insured. [2]

(v) Suggest another model which you think might be more successful in explaining the variability in the values of the average sum insured and provide a better predictor. [2]

[Total 20]

END OF PAPER

EXAMINERS' REPORT

April 2010 Examinations

Subject CT3 — Probability and Mathematical Statistics Core Technical

Introduction

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

R D Muckart
Chairman of the Board of Examiners

July 2010

The paper was answered well overall and there is only one question that stands out as being very poorly attempted (and poorly answered by those who did attempt it), namely question 7. The question is about the precision of estimation - it involves applying the standard error of estimation of a sample proportion to find the sample size required to ensure a stated margin of error. The standard error of estimation of a sample proportion is an important and widely-used quantity.

1 Let p be the proportion of women.

Then, using a weighted average, $1.671p + 1.758(1-p) = 1.712$
$\Rightarrow 0.087p = 0.046 \Rightarrow p = 0.529$ so percentage is 52.9%

2 $P(\text{all 3 on male lives}) = \frac{7}{10}\times\frac{6}{9}\times\frac{5}{8} = \frac{7}{24} = 0.292$

[OR $\binom{7}{3}/\binom{10}{3} = 35/120 = 7/24$]

3
$$\begin{aligned} E\left[S^2\right] &= \frac{1}{n-1}\{\Sigma E(X_i^2) - nE(\bar{X}^2)\} \\ &= \frac{1}{n-1}\{\Sigma(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)\} \\ &= \frac{1}{n-1}\{n(\sigma^2+\mu^2) - \sigma^2 - n\mu^2\} \\ &= \frac{1}{n-1}\{(n-1)\sigma^2\} = \sigma^2 \end{aligned}$$

4 Approximate 95% CI for $\lambda_m - \lambda_f$ is $(\bar{x}_m - \bar{x}_f) \pm 1.96\sqrt{\frac{\bar{x}_m}{120} + \frac{\bar{x}_f}{80}}$

$\Rightarrow \quad (0.24-0.15) \pm 1.96\sqrt{\frac{0.24}{120} + \frac{0.15}{80}}$

$\Rightarrow \quad 0.09 \pm 1.96(0.062) \quad \Rightarrow \quad 0.09 \pm 0.122 \quad \text{or} \quad \Rightarrow \quad (-0.032, 0.212)$

5 $S = \Sigma X_i$ where X_i has a uniform distribution on 1, 2, 3, 4, 5, with mean 3 and variance $(25 - 1)/12 = 2$ (result known, or calculated via $E[X^2] = 11$, or from book of formulae, p10, with $a = 1$, $b = 5$, $h = 1$).

So $S \sim N(300, 200)$ approximately

$$P(280 \le S \le 320) = P\left(\frac{279.5 - 300}{\sqrt{200}} < Z < \frac{320.5 - 300}{\sqrt{200}}\right)$$

$$= P(-1.450 < Z < 1.450) = 0.853$$

6 (i) (a) $\Sigma X_i \sim$ gamma$(4n, \lambda)$

(b) If $Y \sim$ gamma(α, λ) and 2α is an integer, then $2\lambda Y \sim \chi^2_{2\alpha}$ (from book of formulae, p12)

So $2\lambda n\bar{X} \sim \chi^2$ with df $8n$.

(ii) $P(\chi^2_{40}(97.5) < 10\lambda\bar{X} < \chi^2_{40}(2.5)) = 0.95$

giving the 95% CI as $\left(\frac{\chi^2_{40}(97.5)}{10\bar{X}}, \frac{\chi^2_{40}(2.5)}{10\bar{X}}\right)$

Data $\Rightarrow$ $\left(\frac{24.43}{10(17.5)}, \frac{59.34}{10(17.5)}\right) = (0.140, 0.339)$

7 The 95% CI for the population percentage p is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

giving $|\, p - \hat{p} \,| \leq 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

For the margin of error to be less than 0.5% we need to solve

$$0.005 = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow n = \frac{1.96^2\,\hat{p}(1-\hat{p})}{0.005^2} \ .$$

Using the percentage from the previous study as the value for $\hat{p}$, i.e. $\hat{p} = 0.06$, we obtain n = 8,666.6.

So we need a sample of (at least) 8667 people.

(OR, solution can be based on $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ and

$P(-0.005 < \hat{p} - p < 0.005) > 0.95$ without referring to the CI.)

8 (i) $\Sigma f = 100, \quad \Sigma fx = 27, \quad \Sigma fx^2 = 35$

$$\bar{x} = \frac{27}{100} = 0.27$$

$$s^2 = \frac{1}{99}\{35 - \frac{27^2}{100}\} = 0.2799 \quad \therefore s = 0.529$$

Third moment about mean is

$$m_3 = \frac{1}{100}\{76(0-0.27)^3 + 22(1-0.27)^3 + (2-0.27)^3 + (3-0.27)^3\} = 0.3259$$

[OR: using $\Sigma fx^3 = 57, \quad m_3 = \frac{1}{100}\{57 - 3(0.27)(35) + 2(100)(0.27)^3\}$]

So coefficient of skewness is $\frac{0.3259}{(0.2799)^{3/2}} = 2.20$

[OR: can use $m_2 = 0.2771$ in denominator to give 2.23]

(ii) (a) $\hat{\mu} = \bar{x} = 0.27$

(b) Coefficient of skewness is $\frac{1}{\sqrt{0.27}} = 1.92$ (from book of formulae, p7)

so, the data distribution is slightly more positively skewed than the fitted Poisson.

9 (i) $E[N] = \frac{k(1-p)}{p} = \frac{2(0.2)}{0.8} = 0.5$ and $V[N] = \frac{k(1-p)}{p^2} = \frac{2(0.2)}{0.8^2} = 0.625$

$E[X] = \frac{1}{\lambda} = \frac{1}{2} = 0.5$ and $V[X] = \frac{1}{\lambda^2} = \frac{1}{2^2} = 0.25$

$E[S] = E[N]E[X] = 0.5 \times 0.5 = 0.25$, i.e. £250

$V[S] = E[N]V[X] + V[N]\{E[X]\}^2 = 0.5 \times 0.25 + 0.625 \times 0.5^2 = 0.28125$

$\therefore \quad SD[S] = 0.530$, i.e. £530

(ii) $E[N] = V[N] = \mu = 0.5$

$E[X] = \frac{\alpha}{\lambda} = \frac{2}{4} = 0.5$ and $V[X] = \frac{\alpha}{\lambda^2} = \frac{2}{4^2} = 0.125$

$E[S] = E[N]E[X] = 0.5 \times 0.5 = 0.25$, i.e. £250

$V[S] = E[N]V[X] + V[N]\{E[X]\}^2 = 0.5 \times 0.125 + 0.5 \times 0.5^2 = 0.1875$

$\therefore \quad SD[S] = 0.433$, i.e. £433

(iii) As expected the means are the same,
but the standard deviation in (i) is larger than that in (ii) due to the fact that both *N* and *X* have larger variances.

10 (i) We have:

$$E[X] = \int_c^\infty x f_X(x)dx = \int_c^\infty x\frac{ac^a}{x^{a+1}}dx = ac^a\int_c^\infty x^{-a}dx = -\frac{ac^a}{a-1}\left[x^{-(a-1)}\right]_c^\infty$$

and for $a > 1$

$$E[X] = -\frac{ac^a}{a-1}(0-c^{-a+1}) = \frac{ac}{a-1}.$$

(ii) $$F_X(x) = \int_c^x f_X(t)dt = \int_c^x \frac{ac^a}{t^{a+1}}dt$$

which gives

$$F_X(x) = -c^a\left[t^{-a}\right]_c^x = -c^a(x^{-a}-c^{-a}) = 1-\left(\frac{c}{x}\right)^a, \quad x \geq c$$

[*OR* differentiate $F_X(x)$ to obtain $f_X(x)$]

(iii) The likelihood function is given by:

$$L(a) = \prod_{i=1}^n f_X(x_i) = \prod_{i=1}^n \frac{ac^a}{x_i^{a+1}} = a^n c^{na}\prod_{i=1}^n x_i^{-(a+1)}$$

and

$$l(a) = n\log(a) + na\log(c) - (a+1)\sum_{i=1}^n \log(x_i)$$

For the MLE:

$$l'(a) = 0 \Rightarrow \frac{n}{a} + n\log(c) - \sum_{i=1}^n \log(x_i) = 0$$

$$\Rightarrow \hat{a} = \frac{n}{\sum_{i=1}^n \log(x_i) - n\log(c)} = \frac{n}{\sum_{i=1}^n \log\left(\frac{x_i}{c}\right)},$$

and for $c = 2.5$, $\hat{a} = \dfrac{n}{\sum_{i=1}^n \log\left(\frac{x_i}{2.5}\right)}$

(iv) For the asymptotic variance we use the Cramer-Rao lower bound:

$$l''(a) = -\frac{n}{a^2}, \text{ and } E\left[l''(a)\right] = -\frac{n}{a^2}$$

giving

$$V\left[\hat{a}\right] = -\left\{E\left[l''(a)\right]\right\}^{-1} = \frac{a^2}{n}.$$

Hence, asymptotically, $\hat{a} \sim N(a, a^2/n)$.

(v) MLE is

$$\hat{a} = \frac{n}{\sum_{i=1}^{n} \log\left(\frac{x_i}{c}\right)} = \frac{n}{\sum_{i=1}^{n} \log(x_i) - n\log(c)} = \frac{30}{32.9 - 30 \times \log(2.5)} = 5.544.$$

Using the asymptotic normal distribution given above, an approximate 95% CI is given by

$$\hat{a} \pm 1.96\sqrt{\frac{a^2}{n}} = \hat{a} \pm 1.96\frac{\hat{a}}{\sqrt{n}}$$

i.e. $5.544 \pm 1.96\frac{5.544}{\sqrt{30}}$, giving (3.560, 7.528).

(vi) Size of claim in the following year will be given by $1.05X$

So we want $P(1.05X > 4) = P\left(X > \frac{4}{1.05}\right) = 1 - F_X\left(\frac{4}{1.05}\right)$

and using F_X given in the question

$$P(1.05X > 4) = \left(\frac{1.05 \times 2.5}{4}\right)^6 = 0.0799.$$

11 (i) (a) $\bar{x} = 19.513,\ s^2 = \frac{1}{14}\left(5778.69 - \frac{292.7^2}{15}\right) = 4.7955$

(b) Test statistic is $\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$

Here $t = (19.513 - 18)/(4.7955/15)^{1/2} = 2.68$

P-value = $P(t_{14} > 2.68)$, which is just less than 0.01 (1%)

We reject H_0 and accept "$\mu > 18$" at the 1% level of testing.

(ii) (a) Here $t = (19.867 - 18)/(19.432/15)^{1/2} = 1.64$

P-value = $P(t_{14} > 1.64)$, which is between 0.05 and 0.1.

P-value exceeds 5% and so we cannot reject H_0, so "$\mu = 18$" can stand.

(b) Sample 2 does not provide enough evidence to justify rejecting H_0, despite having the same size and a similar mean to Sample 1.

The reason for the loss of significance is the much greater variation in the data in Sample 2 – the variance is four times bigger than in Sample 1 (19.432 v 4.7955)

– this greatly increases the standard error of estimation and reduces the value of the t-statistic (1.64 v 2.68).

(iii) (a) Here $t = (19.644 - 18)/(5.275/25)^{1/2} = 3.58$

P-value = $P(t_{24} > 3.58)$, which is less than 0.001 (0.1%)

We reject H_0 and accept "$\mu > 18$" at a level lower than 0.1%.

(b) Sample 3 provides even stronger evidence against H_0, despite having a similar mean and variance to Sample 1.

The main reason for the much greater level of significance is the increased sample size (25 v 15)
– this decreases the standard error of estimation and increases the value of the t-statistic considerably (3.58 v 2.68).

12 (i)

- the three sets of points are positioned at different levels (the means are shown), so there is a prima facie case for suggesting that the underlying means are different (i.e. there are country effects)
- the means are in the order England (highest), Scotland, Wales (lowest)
- the variation in the data for Scotland is perhaps lower than that for England, but with only 5 observations for each country, we cannot be sure that there is a real underlying difference in variance

(ii) (a) $SS_T = 1316.63 - 137.1^2/15 = 63.536$, $SS_B = (55.6^2 + 36.8^2 + 44.7^2)/5 - 137.1^2/15 = 35.644$

$\therefore\ SS_R = 63.536 - 35.644 = 27.892$

Source of variation	*Df*	*SS*	*MSS*
Between countries	2	35.644	17.82
Residual	12	27.892	2.324
Total	14		

Under H_0: no country effects $F = 17.82/2.324 = 7.67$ on (2,12) df

P-value of $F = 7.67$ is less than 0.01, so we reject H_0 and conclude that there are differences among the population means of the average sum insured

(b) We have strong evidence that country effects exist – the means appear to be in the order England (highest), Scotland, Wales (lowest).

(iii) (a) $S_{xx} = 7543 - 329^2/15 = 326.9333$, $S_{yy} = 63.536$ (from (i)(b) above)

$S_{xy} = 3091.7 - 329 \times 137.1/15 = 84.64$

$\hat{\beta} = 84.64 / 326.9333 = 0.25889$, $\hat{\alpha} = 137.1/15 - \hat{\beta} \times (329/15) = 3.4617$

So fitted line is $y = 3.462 + 0.2589x$

(b) $R^2 = S_{xy}^2/(S_{xx}S_{yy}) = 84.64^2/(326.9333 \times 63.536) = 0.34488$ so 34.5%

(c) SSRES $= S_{yy} - S_{xy}^2/S_{xx} = 63.536 - 84.64^2/326.9333 = 41.62349$

$\Rightarrow \hat{\sigma}^2 = 41.62349/13 = 3.201807$

$\Rightarrow s.e.(\hat{\beta}) = (3.201807/326.9333)^{1/2} = 0.09896$

(iv) From the plot we see that the relationship between "index" and "average sum insured" is weak, positive (and possibly linear) – the percentage of the variation in "average sum insured" explained by the relationship with "index" is only 34.5%.

So "index" is of some, but limited, use as a predictor of "average sum insured".

(v) We should try a "multiple regression" model which includes "country" and "index" in the model.

[*Note:* although not explicitly in the syllabus, a comment to the effect that "Country" should be included as a qualitative variable (a "factor") e.g. by using a text vector (with entries "*E*", "*W*", "*S*" say) or a pair of (Bernoulli) dummy variables, may attract a bonus for a borderline candidate.]

END OF EXAMINERS' REPORT

EXAMINATION

4 October 2010 (am)

Subject CT3 — Probability and Mathematical Statistics Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 12 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 The marks of a sample of 25 students from a large class in a recent test have sample mean 57.2 and standard deviation 7.3. The marks are subsequently adjusted: each mark is multiplied by 1.1 and the result is then increased by 8.

Calculate the sample mean and standard deviation of the adjusted marks. [2]

2 In a survey, a sample of 10 policies is selected from the records of an insurance company. The following data give, in ascending order, the time (in days) from the start date of the policy until a claim has arisen from each of the policies in the sample.

297 301 312 317 355 379 404 419 432+ 463+

Some of the policies have not yet resulted in any claims at the time of the survey, so the times until they each give rise to a claim are said to be censored. These values are represented with a plus sign in the above data.

(i) Calculate the median of this sample. [2]

(ii) State what you can conclude about the mean time until claims arise from the policies in this sample. [2]

[Total 4]

3 Suppose that in a group of insurance policies (which are independent as regards occurrence of claims), 20% of the policies have incurred claims during the last year. An auditor is examining the policies in the group one by one in random order until two policies with claims are found.

(i) Determine the probability that exactly five policies have to be examined until two policies with claims are found. [2]

(ii) Find the expected number of policies that have to be examined until two policies with claims are found. [1]

[Total 3]

4 For a certain class of business, claim amounts are independent of one another and are distributed about a mean of μ = £4,000 and with standard deviation σ = £500.

Calculate an approximate value for the probability that the sum of 100 such claim amounts is less than £407,500. [3]

5 A random sample of 200 travel insurance policies contains 29 on which the policyholders made claims in their most recent year of cover.

Calculate a 99% confidence interval for the proportion of policyholders who make claims in a given year of cover. [3]

6 The random variable X has a Poisson distribution with mean Y, where Y itself is considered to be a random variable. The distribution of Y is lognormal with parameters μ and σ^2.

Derive the unconditional mean $E[X]$ and variance $V[X]$ using appropriate conditional moments. (You may use any standard results without proof, including results from the book of Formulae and Tables.) [4]

7 Let X be a discrete random variable with the following probability distribution:

X:	0	1	2	3
$P(X = x)$:	0.4	0.3	0.2	0.1

(i) Simulate three observations of X using the following three random numbers from a uniform distribution on (0,1) (you should explain your method briefly and clearly).

Random numbers: 0.4936, 0.7269, 0.1652 [3]

Let X be a random variable with cumulative distribution function:

$$F_X(x) = \frac{1}{1-e^{-1}}\left(1-e^{-x^2}\right), \quad 0 < x < 1 \quad (F_X(x) = 0 \text{ for } x \le 0 \text{ and } F_X(x) = 1 \text{ for } x \ge 1).$$

(ii) Derive an expression for a simulated value of X using a random number u from a uniform distribution on (0,1) and hence simulate an observation of X using the random number $u = 0.8149$. [3]

[Total 6]

8 A certain type of claim amount (in units of £1,000) is modelled as an exponential random variable with parameter $\lambda = 1.25$. An analyst is interested in S, the total of 10 such independent claim amounts. In particular he wishes to calculate the probability that S exceeds £10,000.

(i) (a) Show, using moment generating functions, that:

(1) S has a gamma distribution, and

(2) $2.5S$ has a χ^2_{20} distribution.

(b) Use tables to calculate the required probability.

[5]

(ii) (a) Specify an approximate normal distribution for S by applying the central limit theorem, and use this to calculate an approximate value for the required probability.

(b) Comment briefly on the use of this approximation and on the result.

[3]

[Total 8]

9 Let the random variable X have the Poisson distribution with probability function:

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots$$

(i) Show that $P(X = k+1) = \frac{\lambda}{k+1} P(X = k), \quad k = 0, 1, 2, \ldots$ [2]

It is believed that the distribution of the number of claims which arise on insurance policies of a certain class is Poisson. A random sample of 1,000 policies is taken from all the policies in this class which have been in force throughout the past year. The table below gives the number of claims per policy in this sample.

No. of claims, k:	0	1	2	3	4	5	6	7	8 or more
No. of policies, f_k:	310	365	202	88	26	6	2	1	0

For these data the maximum likelihood estimate (MLE) of the Poisson parameter λ is $\hat{\lambda} = 1.186$.

(ii) Calculate the frequencies expected under the Poisson model with parameter given by the MLE above, using the recurrence formula of part (i) (or otherwise). [3]

(iii) Perform an appropriate statistical test to investigate the assumption that the numbers of claims arising from this particular class of policies follow a Poisson distribution. [5]

[Total 10]

10 In the collection of questionnaire data, randomised response sampling is a method which is used to obtain answers to sensitive questions. For example a company is interested in estimating the proportion, p, of its employees who falsely take days off sick. Employees are unlikely to answer a direct question truthfully and so the company uses the following approach.

Each employee selected in the survey is given a fair six-sided die and asked to throw it. If it comes up as a 5 or 6, then the employee answers yes or no to the question "have you falsely taken any days off sick during the last year?". If it comes up as a 1, 2, 3 or 4, then the employee is instructed to toss a coin and answer yes or no to the question "did you obtain heads?". So an individual's answer is either yes or no, but it is not known which question the individual has answered.

For the purpose of the following analysis you should assume that each employee answers the question truthfully.

(i) Show that the probability that an individual answers yes is $\frac{1}{3}(p+1)$. [2]

Suppose that 100 employees are surveyed and that this results in 56 yes answers.

(ii) (a) Show that the likelihood function $L(p)$ can be expressed in the form:

$$L(p) \propto (p+1)^{56}(2-p)^{44}.$$

(b) Hence show that the maximum likelihood estimate (MLE) of p is $\hat{p} = 0.68$.

[5]

Let $\theta = \frac{1}{3}(p+1)$ and note that, using binomial results, the MLE of θ is $\hat{\theta} = \frac{56}{100}$.

(iii) Explain why $\hat{p}$ can be obtained as the solution of $\frac{1}{3}(\hat{p}+1) = \hat{\theta}$, and hence verify that $\hat{p} = 0.68$. [2]

(iv) (a) Determine the second derivative of the log likelihood for p and evaluate it at $\hat{p} = 0.68$.

(b) State an approximate large-sample distribution for the MLE $\hat{p}$.

(c) Hence calculate approximate 95% confidence limits for p.

[5]

Now suppose that the same numerical estimate, that is $\hat{p} = 0.68$, had been obtained from a sample of the same size, that is 100, without using the randomised response method but relying on truthful answers. So the number of yes answers was 68 and $\hat{p} = \frac{68}{100}$ using binomial results.

(v) (a) Calculate approximate 95% confidence limits for p for this situation.

(b) Suggest why the confidence limits in part (iv)(c) are wider than these limits.

[3]
[Total 17]

11 A life insurance company issuing critical illness insurance wants to compare the delay times from the date when a claim is made until it is settled, for different causes of illness covered. Random samples of 12 claims associated with two types of illness (A and B) related to heart disease have been collected. The logarithms of the delay times are given below (where the original times were measured in days):

Cause A, y_A:	4.0	5.4	4.6	3.5	4.2	4.5	4.2	4.9	5.1	5.2	5.1	5.4
Cause B, y_B:	5.7	5.6	4.2	5.1	4.4	5.9	5.4	3.9	5.7	4.5	4.8	3.9

For these data: $\sum y_A = 56.1, \quad \sum y_A^2 = 266.33, \quad \sum y_B = 59.1, \quad \sum y_B^2 = 297.03$

(i) Use a suitable *t*-test to investigate the hypothesis that the mean delay time is the same for claims related to the two causes of illness and state clearly your conclusion. [6]

(ii) Give a possible reason why the logarithms of the original delay time observations are used in this analysis. [2]

(iii) (a) Calculate an equal-tailed 95% confidence interval for σ_A^2 / σ_B^2, the ratio of the variances of the delay times for the two causes of illness.

(b) Comment on the validity of the test in part (i) based on this confidence interval.

[4]

The company collects a third sample of 12 claims associated with an illness (C) related to brain disease, and the logarithms of the delay times are given below:

Cause C, y_C:	5.6	6.2	6.0	5.6	7.1	5.0	4.5	6.4	4.6	6.0	5.5	5.3

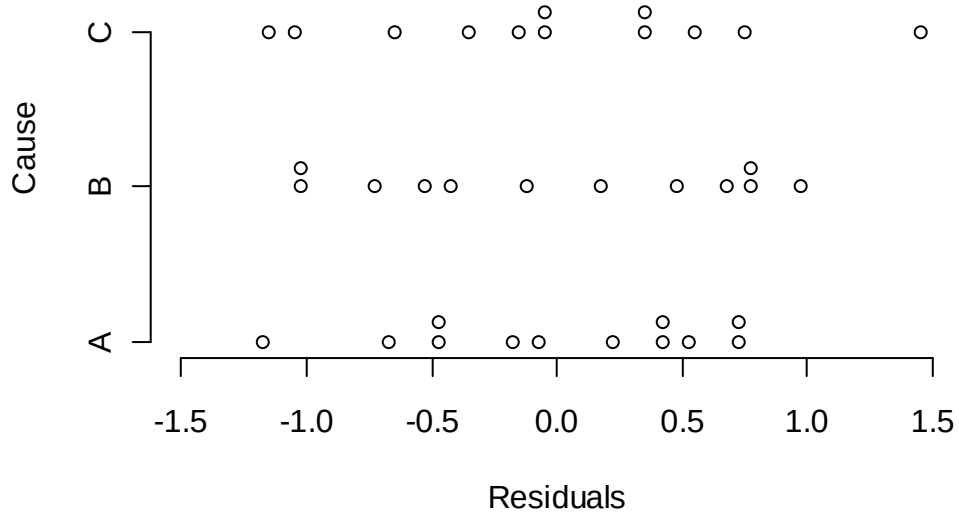
For these data: $\sum y_C = 67.8, \quad \sum y_C^2 = 389.28$

For data in all three samples: $\sum\sum y = 183.0, \quad \sum\sum y^2 = 952.64$

(iv) Use analysis of variance to test the hypothesis that the mean delay times are the same for all three causes of illness. [6]

(v) State the assumptions made for this analysis of variance. [2]

(vi) Comment briefly on the validity of the test in (iv), using the plot of the residuals of the analysis given below. [2]



[Total 22]

12 An investigation concerning the improvement in the average performance of female track athletes relative to male track athletes was conducted using data from various international athletics meetings over a period of 16 years in the 1950s and 1960s. For each year and each selected track distance the observation y was recorded as the average of the ratios of the twenty best male times to the corresponding twenty best female times.

The data for the 100 metres event are given below together with some summaries.

year t:	1	2	3	4	5	6	7	8
ratio y:	0.882	0.879	0.876	0.888	0.890	0.882	0.885	0.886

year t:	9	10	11	12	13	14	15	16
ratio y:	0.885	0.887	0.882	0.893	0.878	0.889	0.888	0.890

$$\Sigma t = 136, \quad \Sigma t^2 = 1496, \quad \Sigma y = 14.160, \quad \Sigma y^2 = 12.531946, \quad \Sigma ty = 120.518$$

(i) Draw a scatterplot of these data and comment briefly on any relationship between ratio and year. [3]

(ii) Verify that the equation of the least squares fitted regression line of ratio on year is given by:

$$y = 0.88105 + 0.000465t.$$ [4]

(iii) (a) Calculate the standard error of the estimated slope coefficient in part (ii).

(b) Determine whether the null hypothesis of "no linear relationship" would be accepted or rejected at the 5% level.

(c) Calculate a 95% confidence interval for the underlying slope coefficient for the linear model.

[5]

Corresponding data for the 200 metres event resulted in an estimated slope coefficient of:

$\hat{\beta} = 0.000487$ with standard error 0.000220.

(iv) (a) Determine whether the "no linear relationship" hypothesis would be accepted or rejected at the 5% level.

(b) Calculate a 95% confidence interval for the underlying slope coefficient for the linear model and comment on whether or not the underlying slope coefficients for the two events, 100m and 200m, can be regarded as being equal.

(c) Discuss why the results of the tests in parts (iii)(b) and (iv)(a) seem to contradict the conclusion in part (iv)(b).

[6]

[Total 18]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2010 examinations

Subject CT3 — Probability and Mathematical Statistics
Core Technical

Introduction

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

T J Birse
Chairman of the Board of Examiners

December 2010

Comments

The paper was answered well and overall performance was satisfactory. However, some questions were poorly attempted. A number of candidates could not answer Question 1 correctly and efficiently – the question required basic knowledge of data summaries. Question 8 was not answered very well – answers to questions that require candidates to "show" a particular statement, need to demonstrate intermediate steps clearly and accurately. The same applies to Question 10, where deriving specific results regarding maximum likelihood estimation was not performed accurately by many candidates.

1 Sample mean = $(1.1 \times 57.2) + 8 = 70.92$
Sample standard deviation = $1.1 \times 7.3 = 8.03$

2 (i) Sample median is not affected by the fact that the last two observations are censored.

It is therefore given by the 5.5^{th} ranked observation, i.e. $(355 + 379)/2 = 367$ days.

(ii) We know that the last two observations have minimum values 432 and 463.

Using these two values the sample mean would be equal to 3679/10 = 367.9.

So, the sample mean is at least equal to 367.9 days.

3 (i) Using the negative binomial distribution, or from first principles,

$P(5 \text{ policies required}) = \binom{5-1}{2-1}(0.2)^2(0.8)^3 = 0.0819$

(ii) Expected number = mean of negative binomial distribution = $\frac{2}{0.2} = 10$

4 Working in units of £1,000, sum of 100 claim amounts S has $E[S] = 100 \times 4 = 400$ and $V[S] = 100 \times 0.5^2 = 25$, and so $S \sim N(400, 5^2)$ approximately.

$P(S < 407.5) = P(Z < 1.5) = 0.933$

5 Sample proportion = 29/200 = 0.145

99% CI is given by $0.145 \pm 2.5758\sqrt{\frac{0.145 \times 0.855}{200}}$ i.e. 0.145 ± 0.064

i.e. (0.081, 0.209).

6 $E[X] = E[E(X \mid Y)] = E[Y] = e^{\mu+\sigma^2/2}$

$V[X] = E[V(X \mid Y)] + V[E(X \mid Y)] = E[Y] + V[Y] = e^{\mu+\sigma^2/2} + e^{2\mu+\sigma^2}(e^{\sigma^2} - 1)$

7 (i) Method

$0 < u \le 0.4 \Rightarrow x = 0$
$0.4 < u \le 0.7 \Rightarrow x = 1$
$0.7 < u \le 0.9 \Rightarrow x = 2$
$0.9 < u \le 1 \Rightarrow x = 3$

We get $x = 1, 2, 0$

(ii) Setting $u = \frac{1}{1-e^{-1}}\left(1 - e^{-x^2}\right) \Rightarrow e^{-x^2} = 1 - \left(1 - e^{-1}\right)u$

$\Rightarrow x = \left[-\log\left[1 - \left(1 - e^{-1}\right)u\right]\right]^{1/2}$

$u = 0.8149 \Rightarrow x = 0.851$

8 (i) (a) (1) Let X_i be a claim amount.

Mgf of X_i is $M_X(t) = \left(1 - \frac{t}{1.25}\right)^{-1}$

Mgf of $S = \sum_{i=1}^{10} X_i$ is $M_S(t) = [M_X(t)]^{10} = \left(1 - \frac{t}{1.25}\right)^{-10}$,

which is the mgf of a gamma(10, 1.25) variable.

(2) Mgf of $2.5S$ is $E[e^{t(2.5S)}] = E[e^{(2.5t)S}] = M_S(2.5t) = (1-2t)^{-10}$,

which is the mgf of a gamma(10, ½) variable, i.e. χ^2_{20}.

(b) $P(\text{total} > £10{,}000) = P(S > 10) = P(\chi^2_{20} > 25) = 1 - 0.7986 = 0.2014$

(ii) (a) S has mean $\frac{10}{1.25} = 8$ and variance $\frac{10}{1.25^2} = 6.4$. So $S \approx N(8, 6.4)$

$P(S > 10) \cong P(Z > \frac{10-8}{\sqrt{6.4}} = 0.791) = 1 - 0.786 = 0.214$

(b) n is not particularly large for the use of the CLT, but the approximation is still quite close to the true probability.

9 (i) $P(X = k+1) = e^{-\lambda}\dfrac{\lambda^{k+1}}{(k+1)!} = e^{-\lambda}\dfrac{\lambda^k}{k!}\dfrac{\lambda}{k+1} = \dfrac{\lambda}{k+1}P(X = k)$.

(ii) Using $P(X = 0) = e^{-1.186}$, $P(X = 8 \text{ or more}) = 1 - \sum_{i=0}^{7} P(X = i)$, and the recurrent formula, we obtain:

K	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8 or more*
$P(X = k)$	0.3054	0.3623	0.2148	0.0849	0.0252	0.0060	0.0012	0.0002	4×10^{-5}
Expected, e_k	305.4	362.3	214.8	84.9	25.2	6.0	1.2	0.2	0.0

(iii) Combining the last 4 categories to obtain expected frequencies greater than 5, we have:

k	*0*	*1*	*2*	*3*	*4*	*5 or more*
No. of policies, f_k	310	365	202	88	26	9
Expected, e_k	305.4	362.3	214.8	84.9	25.2	7.4

This gives

$$\chi^2 = \sum \frac{(f_k - e_k)^2}{e_k}$$

$$= 0.0693 + 0.0201 + 0.7628 + 0.1132 + 0.0254 + 0.3459 = 1.3367$$

DF = 6 – 1 – 1 = 4, and from statistical tables, $\chi^2_{0.05,4} = 9.488$.

Therefore, we do not have evidence against the hypothesis that the number of claims comes from a Poisson(1.186) distribution.

(Alternatively if we only combine the last 3 categories, the expected frequencies for 5 and 6 or more policies are 6 and 1.4, with observed frequencies 6 and 3 respectively. These give $\chi^2 = 2.819$ on 5 DF, and with $\chi^2_{0.05,5} = 11.071$ the conclusion is the same as before.)

10 (i) $P(\text{yes}) = P(5,6)P(\text{yes}\,|\,\text{main question}) + P(1,2,3,4)P(\text{yes}\,|\,\text{coin question})$

$$= \frac{2}{6}p + \frac{4}{6}\cdot\frac{1}{2} = \frac{1}{3}(p+1)$$

(ii) (a) $L(p) \propto [\frac{1}{3}(p+1)]^{56}[1-\frac{1}{3}(p+1)]^{100-56}$

$$\propto (p+1)^{56}(2-p)^{44}$$

(b) $\log L = 56\log(p+1) + 44\log(2-p) + const.$

$$\frac{d}{dp}\log L = \frac{56}{p+1} - \frac{44}{2-p}$$

$$= \frac{56(2-p)-44(p+1)}{(p+1)(2-p)} = \frac{68-100p}{(p+1)(2-p)}$$

Equate to zero $\Rightarrow$ $68-100p = 0$ $\therefore \hat{p} = \frac{68}{100} = 0.68$

(iii) Due to the invariance property of MLEs $\frac{1}{3}(\hat{p}+1) = \hat{\theta}$

$$\therefore \frac{1}{3}(\hat{p}+1) = \frac{56}{100} \quad \therefore \hat{p} = \frac{168}{100} - 1 = \frac{68}{100} = 0.68$$

(iv) (a) $$\frac{d^2}{dp^2}\log L = -\frac{56}{(p+1)^2} - \frac{44}{(2-p)^2}$$

at $\hat{p} = 0.68$, $\frac{d^2}{dp^2}\log L = -\frac{56}{1.68^2} - \frac{44}{1.32^2} = -45.0938$

(b) $CRlb = \frac{-1}{-45.0938} = 0.02218$ and $\hat{p} \approx N(p, 0.02218)$

(c) Approximate 95% CI for p is $\hat{p} \pm 1.96\sqrt{0.02218}$
giving 0.68 ± 0.292

(v) (a) Approximate 95% CI for p is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{100}}$

giving

$$0.68 \pm 1.96\sqrt{\frac{0.68(1-0.68)}{100}} \Rightarrow 0.68 \pm 1.96(0.0466) \Rightarrow 0.68 \pm 0.091$$

(b) Less accurate estimation is the penalty paid for using the randomised response method.

11 (i) We want to test H_0: $\mu_A = \mu_B$ against H_1: $\mu_A \neq \mu_B$.

Data give: $\bar{y}_A = 56.1/12 = 4.675$, $\bar{y}_B = 59.1/12 = 4.925$

$s_A^2 = (266.33 - 56.1^2/12)/11 = 0.36932$,
$s_B^2 = (297.03 - 59.1^2/12)/11 = 0.54205$

Assuming that the two samples come from normal distributions with the same variance,

we first compute the pooled variance as $s_p^2 = \dfrac{11s_A^2 + 11s_B^2}{22} = 0.455685$

which gives $t = \dfrac{\bar{y}_A - \bar{y}_B}{s_p\sqrt{2/12}} = -0.907$.

Critical values at 5% level are $t_{22}(0.025) = -2.074$ and $t_{22}(0.975) = 2.074$ so we don't have evidence against H_0 and conclude that the mean delay time is the same for claims associated with the two causes of illness.

(ii) Distribution of times can be skewed to the right, and we need a log transformation to normalise the data (for test to be valid).

(iii) (a) CI is given by $\left(\dfrac{s_A^2/s_B^2}{F_{11,11}(0.025)}, (s_A^2/s_B^2) * F_{11,11}(0.025) \right)$

$F_{11,11}(0.025) = 3.478$ (using interpolation in the tables)
giving CI as (0.68134/3.478, 0.68134*3.478) = (0.196, 2.370).

(b) The value "1" is included in the 95% CI, meaning that the assumption of common variance made for the test is valid.

(iv) $SS_T = 952.64 - 183^2/36 = 22.39$
$SS_B = (56.1^2 + 59.1^2 + 67.8^2)/12 - 183^2/36 = 6.155$
$\Rightarrow SS_R = 22.39 - 6.155 = 16.235$

Source of variation	*d.f.*	*SS*	*MSS*
Between	2	6.155	3.078
Residual	33	16.235	0.492
Total	35	22.390	

F = 3.078/0.492 = 6.256 on (2,33) degrees of freedom.

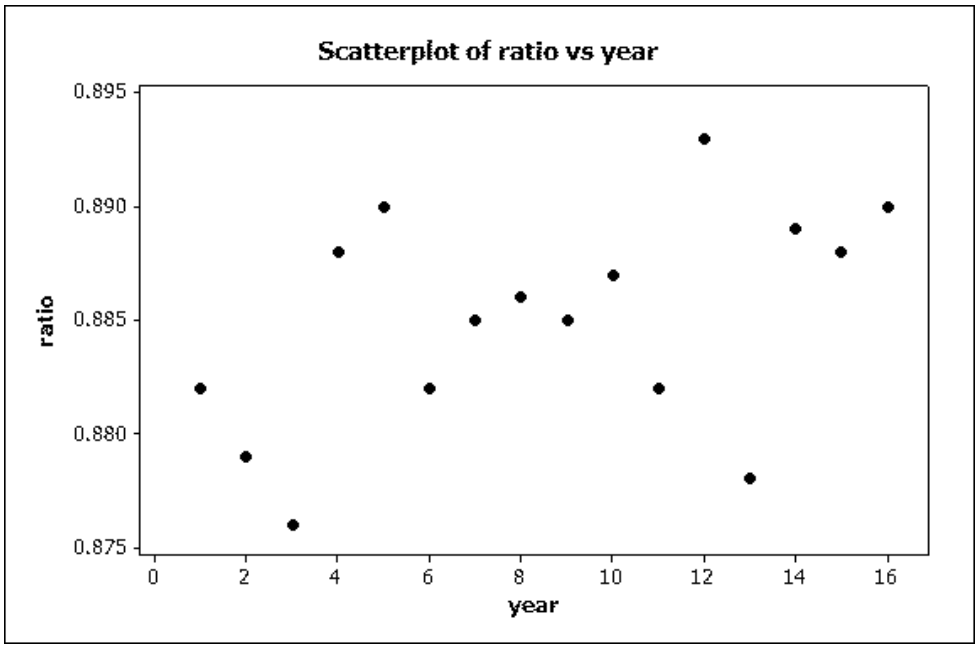
From tables, $F_{2,33}(0.05)$ is between 3.276 and 3.295, and $F_{2,33}(0.01)$ is between 5.289 and 5.336.

We have strong evidence against the null hypothesis and conclude that the three mean delay times are not equal.

(v) The assumptions are that the data come from normal populations with constant variance.

(vi) The plot suggests that the normality assumption is reasonable and that variance does not depend on cause. Test seems valid.

12 (i) Scatterplot with suitable axes and clearly labelled:



There does not appear to be much of a relationship, perhaps a slight increasing linear relationship but it is weak with quite a bit of scatter.

(ii) $n = 16$

$$S_{tt} = 1496 - \frac{136^2}{16} = 340$$

$$S_{yy} = 12.531946 - \frac{14.160^2}{16} = 0.000346$$

$$S_{ty} = 120.518 - \frac{(136)(14.160)}{16} = 0.158$$

$$\hat{\beta} = \frac{S_{ty}}{S_{tt}} = \frac{0.158}{340} = 0.0004647$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{t} = \frac{14.160}{16} - (0.0004647)\frac{136}{16} = 0.88105$$

Fitted line is $y = 0.88105 + 0.000465t$

(iii) (a) s.e. $(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{tt}}}$ where $\hat{\sigma}^2 = \frac{1}{n-2}(S_{yy} - \frac{S_{ty}^2}{S_{tt}})$

$$\hat{\sigma}^2 = \frac{1}{14}(0.000346 - \frac{0.158^2}{340}) = 0.0000195$$

$$\therefore s.e.(\hat{\beta}) = \sqrt{\frac{0.0000195}{340}} = 0.000239$$

(b) Null hypothesis of "no linear relationship" is equivalent to H_0: $\beta = 0$

We use $t = \frac{\hat{\beta}}{s.e.(\hat{\beta})} \sim t_{14}$ under H_0: $\beta = 0$

Observed $t = \frac{0.000465}{0.000239} = 1.95$ and $t_{0.025,14} = 2.145$

So we must accept H_0: no linear relationship at the 5% level.

(c) 95% CI is $0.000465 \pm 2.145 \times 0.000239$
giving 0.000465 ± 0.000513 or (–0.000048, 0.000978)

(iv) (a) Observed $t = \frac{0.000487}{0.000220} = 2.21$ – this is greater than $t_{0.025,14} = 2.145$

So we reject H_0: no linear relationship at the 5% level.

(b) 95% CI is $0.000487 \pm 2.145 \times 0.000220$
giving 0.000487 ± 0.000472 or (0.000015, 0.000959)

The two CIs overlap substantially, so there is no evidence to suggest that the slopes are different.

(c) Although the tests have different conclusions at the 5% level, the 100m observed t is only just inside the critical value of 2.145 and the 200m one is just outside. This in fact agrees with, rather than contradicts, the conclusion that the slopes are not different.

END OF EXAMINERS' REPORT

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

15 April 2011 (pm)

Subject CT3 — Probability and Mathematical Statistics Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 10 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 The numbers of claims which have arisen in the last twelve years on 60 motor policies (continuously in force over this period) are shown (sorted) below:

0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1
1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 3 3 3
3 3 3 3 3 3 3 4 4 4 4 4 5 5 5 5 6 6 6 7

Derive:

(i) The sample median, mode and mean of the number of claims. [3]
(ii) The sample inter-quartile range of the number of claims. [2]
(iii) The sample standard deviation of the number of claims. [3]
[Total 8]

2 A random sample of size $n = 36$ has sample standard deviation $s = 7$.

Calculate, approximately, the probability that the mean of this sample is greater than 44.5 when the mean of the population is $\mu = 42$. [3]

3 In a large population, 35% of voters intend to vote for party A at the next election. A random sample of 200 voters is selected from this population and asked which party they will vote for.

Calculate, approximately, the probability that 80 or more of the people in this sample intend to vote for party A. [4]

4 Let N be the random variable that describes the number of claims that an insurer receives per month for one of its claim portfolios. We assume that N has a Poisson distribution with $E[N] = 50$. The amount X_i of each claim in the portfolio is normally distributed with mean $\mu = 1{,}000$ and variance $\sigma^2 = 200^2$. The total amount of all claims received during one month is

$$S = \sum_{i=1}^{N} X_i$$

with $S = 0$ for $N = 0$. We assume that $N, X_1, X_2, \ldots$ are all independent of each other.

(i) Specify the type of the distribution of S. [1]

(ii) Calculate the mean and standard deviation of S. [3]
[Total 4]

5 Let X_1, X_2, X_3, X_4, and X_5 be independent random variables, such that $X_i \sim$ gamma with parameters i and λ for $i = 1, 2, 3, 4, 5$. Let $S = 2\lambda\sum_{i=1}^{5} X_i$.

(i) Derive the mean and variance of S using standard results for the mean and variance of linear combinations of random variables. [3]

(ii) Show that S has a chi-square distribution using moment generating functions and state the degrees of freedom of this distribution. [4]

(iii) Verify the values found in part (i) using the results of part (ii). . [1]

[Total 8]

6 Consider two random variables X and Y, for which the variances satisfy $V[X] = 5V[Y]$ and the covariance $Cov[X,Y]$ satisfies $Cov[X,Y] = V[Y]$.

Let $S = X + Y$ and $D = X - Y$.

(i) Show that the covariance between S and D satisfies $Cov[S,D] = 4V[Y]$. [3]

(ii) Calculate the correlation coefficient between S and D. [3]

[Total 6]

7 An insurance company distinguishes between three types of fraudulent claims:

Type 1: legitimate claims that are slightly exaggerated
Type 2: legitimate claims that are strongly exaggerated
Type 3: false claims

Every fraudulent claim is characterised as exactly one of the three types. Assume that the probability of a newly submitted claim being a fraudulent claim of type 1 is 0.1. For type 2 this probability is 0.02, and for type 3 it is 0.003.

(i) Calculate the probability that a newly submitted claim is not fraudulent. [1]

The insurer uses a statistical software package to identify suspicious claims. If a claim is fraudulent of type 1, it is identified as suspicious by the software with probability 0.5. For a type 2 claim this probability is 0.7, and for type 3 it is 0.9.

Of all newly submitted claims, 20% are identified by the software as suspicious.

(ii) Calculate the probability that a claim that has been identified by the software as suspicious is:

(a) a fraudulent claim of type 1,
(b) a fraudulent claim of any type.

[5]

(iii) Calculate the probability that a claim which has NOT been identified as suspicious by the software is in fact fraudulent. [3]

[Total 9]

8 Two medications, labelled A and B, were being investigated using a group of twelve patients each of whom was approximately at the same stage of suffering from a severe cough. The patients were divided randomly into two groups of six and medication A was administered to each patient in the first group while medication B was administered to each patient in the second group. Over the next three days the total number of coughs was recorded for each patient, with the following results:

A:	321	585	468	619	447	532
B:	478	381	596	552	358	426

$\Sigma x_A = 2,972$, $\Sigma x_A^2 = 1,530,284$, $\Sigma x_B = 2,791$, $\Sigma x_B^2 = 1,343,205$

(i) Apply an appropriate t-test to determine whether these two medications differ in their effectiveness for the relief of coughing, assuming that the two samples are independent and come from normal populations with equal variances. [7]

(ii) In relation to the test performed in part (i):

(a) Comment on the required assumption of independence.

(b) Present the data graphically and hence comment on the required assumption of normality.

(c) Apply an appropriate F-test to comment on the required assumption of equal variances.

[6]

Suppose that the investigators had used a total of eighteen such patients divided into three groups of six and that a placebo (an inactive substance) was administered to each patient in the third group, labelled C. Suppose that the resulting data for medications A and B were as above together with the following results for the placebo group.

C: 691 827 785 531 603 714

$$\Sigma x_C = 4{,}151, \quad \Sigma x_C^2 = 2{,}933{,}001$$

(iii) Perform an analysis of variance to test the hypothesis that there is no difference among the three groups as regards coughing. [8]

(iv) Comment briefly on any difference among the three groups. [1]

[Total 22]

9 Claims on a certain type of policy are such that the claim amounts are approximately normally distributed.

(i) A sample of 101 such claim amounts (in £) yields a sample mean of £416 and sample standard deviation of £72. For this type of policy:

(a) Obtain a 95% confidence interval for the mean of the claim amounts.

(b) Obtain a 95% confidence interval for the standard deviation of the claim amounts.

[8]

The company makes various alterations to its policy conditions and thinks that these changes may result in a change in the mean, but not the standard deviation, of the claim amounts. It wants to take a random sample of claims in order to estimate the new mean amount with a 95% confidence interval equal to

sample mean ± £10.

(ii) Determine how large a sample must be taken, using the following as an estimate of the standard deviation:

(a) The sample standard deviation from part (i).

(b) The upper limit of the confidence interval for the standard deviation from part (i)(b).

[6]

(iii) Comment briefly on your two answers in (ii)(a) and (ii)(b). [2]

[Total 16]

10 A life insurance company runs a statistical analysis of mortality rates. The company considers a population of 100,000 individuals. It assumes that the number of deaths X during one year has a Poisson distribution with expectation $E[X] = \mu$. Over four years the company has observed the following realisations of X (number of deaths).

Year	1	2	3	4
Number of deaths (*per 100,000 lives*)	1,140	1,200	1,170	1,190

The maximum likelihood estimator for the parameter μ of the Poisson distribution is given by $\bar{X}$.

(i) Obtain the maximum likelihood estimate of the parameter μ using these data. [1]

To obtain a more realistic model, it is proposed that the number of deaths should depend on the age of the population. To this end the total population is divided into four age groups of equal size and the number of deaths in each group during the following year is counted. The observed values are given in the following table. The total number of deaths is again per 100,000 lives.

Middle age (t) in group	25	35	45	55
Number of deaths (*x*) *in age group*	84	113	255	727

For these data we obtain: $\Sigma t = 160$, $\Sigma t^2 = 6,900$, $\Sigma x = 1,179$, $\Sigma x^2 = 613,379$ and $\Sigma xt = 57,515$

(ii) (a) Calculate the correlation coefficient between the middle age t in a group and the number of deaths x in that group, and comment briefly on its value.

(b) Perform a linear regression of the number of deaths x as a function of the middle age t of the group.

[10]

A statistician suggests using a Poisson distribution for the number of deaths per year in each group, where the parameter μ depends on the middle age in that group. Under the suggested model the number of deaths in the group with middle age t_i is given by $X_i \sim \text{Poisson}(\mu_i)$ with $\mu_i = wt_i$, where t_i is the middle age of the group that the individual belongs to at the time of death.

(iii) Derive a maximum likelihood estimator for the parameter w and estimate the value of w from the data in the above table. [9]

[Total 20]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2011 examinations

Subject CT3 — Probability and Mathematical Statistics
Core Technical

Introduction

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

T J Birse
Chairman of the Board of Examiners

July 2011

General comments

The paper was answered very well and overall performance was satisfactory. Some problems were encountered with specific questions. Many candidates did not attempt Question 6 at all, while for those who did, manipulation of covariance terms often proved problematic. In Question 7 part (ii), a number of candidates failed to use the information given in the question regarding the proportion of claims identified as suspicious (i.e. 0.2) – instead they tried to compute this using the total probability theorem. This erroneously assumes a zero false positive rate for the software. In Question 10 part (iii), the non-standard form of the likelihood caused some poor answers. Also, some candidates inserted the data directly into the likelihood derivation, which resulted to only obtaining the ML estimate rather than also deriving the ML estimator as instructed.

1 (i) 30^{th} and 31^{st} observations in order are both 2 $\Rightarrow$ median = 2

mode = value with highest frequency = 1

$\Sigma x = 1(14) + 2(11) + 3(10) + 4(5) + 5(4) + 6(3) + 7(1) = 131$
$\Rightarrow$ mean = 131/60 = 2.18

(ii) Lower quartile is 15.5^{th} observation counting from below = 1

Upper quartile is 15.5^{th} observation counting from above = 3

$\Rightarrow$ IQR = 2

(iii) $\Sigma x^2 = 1(14) + 4(11) + 9(10) + 16(5) + 25(4) + 36(3) + 49(1) = 485$

$\Rightarrow$ standard deviation = $[(485 - 131^2/60)/59]^{1/2} = 3.3726^{1/2} = 1.84$

2 $$P(\bar{X} > 44.5) = P\left(\frac{\bar{X} - \mu}{S/\sqrt{n}} > \frac{44.5 - 42}{7/\sqrt{36}}\right)$$

$\Rightarrow P(\bar{X} > 44.5) \approx P(Z > 2.143)$, where $Z \sim N(0,1)$,

and from tables,

$P(\bar{X} > 44.5) = 1 - 0.984 = 0.016$

(A t_{35} distribution can also be used if a normal distribution is assumed for the data.)

3 If X is the number of voters in the sample voting for party A, we have

$X \sim$ Binomial(200, 0.35) and using the CLT $X \sim N(70, 45.5)$ approximately.

Using continuity correction

$$P(X \geq 80) = P\left(Z > \frac{79.5 - 70}{\sqrt{45.5}}\right) = P(Z > 1.408)$$

$= 1 - P(Z < 1.408) = 1 - 0.920 = 0.08.$

4 (i) Compound Poisson distribution

(ii) $E[S] = 50 * 1000 = 50{,}000$

$V[S] = 50 * E[S^2] = 50 * \{V[X] + (E[X])^2\} = 50 * \{200^2 + 1000^2\} = 52{,}000{,}000$

$SD[S] = 7{,}211.10$

5 (i) $$E[S] = 2\lambda \sum_{i=1}^{5} \frac{i}{\lambda} = 30$$

$$V[S] = 4\lambda^2 \sum_{i=1}^{5} \frac{i}{\lambda^2} = 60$$

(ii) $M_{X_i}(t) = \left(1 - \frac{t}{\lambda}\right)^{-i}$ (from book of formulae)

$$M_S(t) = E\left[\exp(tS)\right] = E\left[\exp\left(2\lambda t \sum_{i=1}^{5} X_i\right)\right] = \prod_{i=1}^{5} E\left[\exp(2\lambda t X_i)\right]$$

$$= \prod_{i=1}^{5} M_{X_i}(2\lambda t) = \prod_{i=1}^{5} (1 - 2t)^{-i} = (1 - 2t)^{-15}$$ so $S \sim \chi^2$, with 30 df

(iii) χ^2_{30} has mean 30 and variance 60 , as found in part (i).

6 (i) $Cov[S,D] = Cov[X + Y\,,\, X - Y] = Cov[X,X] - Cov[X,Y] + Cov[Y,X] - Cov[Y,Y]$
$= V[X] - V[Y] = 4V[Y]$

(ii) $V[S] = V[X] + V[Y] + 2Cov[X,Y] = 8V[Y]$

$V[D] = V[X] + V[Y] - 2Cov[X,Y] = 4V[Y]$

$\Rightarrow Corr[S,D] = 4V[Y]/\{8V[Y] \times 4V[Y]\}^{1/2} = +1/\sqrt{2} = +0.707$

7 (i) $1 - P[T1 \cup T2 \cup T3] = 1 - (0.1 + 0.02 + 0.003) = 1 - 0.123 = 0.877$

(ii) (a) $P[T1 \mid S] = \dfrac{P[T1 \cap S]}{P[S]} = \dfrac{P[S \mid T1]P[T1]}{P[S]} = \dfrac{0.5 * 0.1}{0.2} = 0.25$

(b) $P[T1 \cup T2 \cup T3 \mid S] = \dfrac{1}{P[S]}\,(P[T1 \cap S] + P[T2 \cap S] + P[T3 \cap S])$

$= \dfrac{1}{P[S]}\,(P[S \mid T1]\,P[T1] + P[S \mid T2]\,P[T2] + P[S \mid T3]\,P[T3])$

$= \dfrac{1}{0.2}(0.5 * 0.1 + 0.7 * 0.02 + 0.9 * 0.003) = \dfrac{0.0667}{0.2} = 0.3335$

(iii) $P[\mathrm{T1} \cup T2 \cup T3 \mid S^C] = \dfrac{1}{0.8}(\mathrm{P}[\mathrm{T1} \cup T2 \cup T3] - P[\{T1 \cup T2 \cup T3\} \cap S])$

$= \dfrac{1}{0.8}(0.123 - 0.5 * 0.1 - 0.7 * 0.02 - 0.9 * 0.003) = \dfrac{0.0563}{0.8} = 0.0704$

8 (i) $\bar{x}_A = \dfrac{2972}{6} = 495.33 \qquad s_A^2 = \dfrac{1}{5}\{1530284 - \dfrac{2972^2}{6}\} = 11630.67$

$\bar{x}_B = \dfrac{2791}{6} = 465.17 \qquad s_B^2 = \dfrac{1}{5}\{1343205 - \dfrac{2791^2}{6}\} = 8984.97$

$s_P^2 = \dfrac{5(11630.67) + 5(8984.97)}{10} = 10307.82 \qquad \therefore s_P = 101.527$

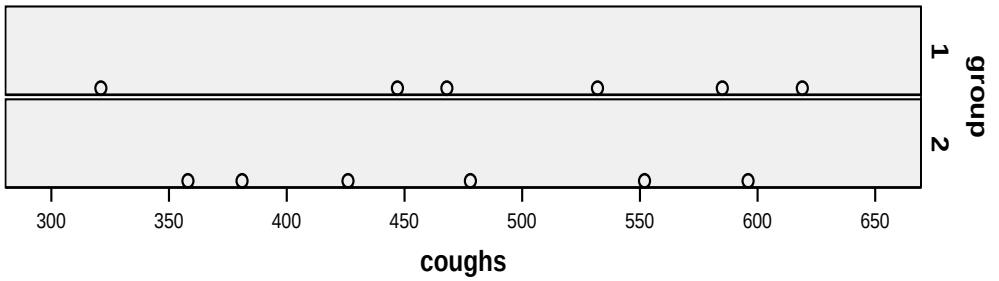
$t = \dfrac{495.33 - 465.17}{101.527\sqrt{\dfrac{1}{6} + \dfrac{1}{6}}} = \dfrac{30.16}{58.62} = 0.51$ on 10 df

without needing to look up tables (although candidates can do so, e.g. $t_{10}(2.5\%) = 2.228$)

there is clearly no evidence of a difference between medications *A* and *B* as regards their effectiveness for the relief of coughing.

(ii) (a) As the 12 patients were split at random into the two groups, the two samples are independent.
(Valid comments on the *need* for this assumption will also receive full credit.)

(b) The most appropriate graphical representation is two dotplots (or boxplots):



These show that there is nothing that suggests lack of normality in each case.
(Valid comments on the *need* for this assumption will also receive full credit.)

(c) $F = \dfrac{s_A^2}{s_B^2} = \dfrac{11630.67}{8984.97} = 1.29$ on 5,5 df

$F_{5,5}(10\%) = 3.453$. So no evidence against the assumption of equal variances.

(iii) $\Sigma x = 2972 + 2791 + 4151 = 9914$,
$\Sigma x^2 = 1530284 + 1343205 + 2933001 = 5806490$

$$SS_T = 5806490 - \frac{9914^2}{18} = 346079$$

$$SS_B = \frac{1}{6}(2972^2 + 2791^2 + 4151^2) - \frac{9914^2}{18} = 181800$$

$$SS_R = SS_T - SS_B = 164279$$

giving the ANOVA table:

Source of variation	*df*	*SS*	*MSS*
Between groups	2	181800	90900
Residual	15	164279	10952
Total	17	346079	

$$F = \frac{90900}{10952} = 8.30 \quad \text{on } 2, 15 \text{ df}$$

$F_{2,15}(5\%) = 3.682$ and $F_{2,15}(1\%) = 6.359$. So P-value < 0.01

So there is very strong evidence of a difference between medications *A* and *B* and the placebo as regards their effectiveness for the relief of coughing.

(iv) It would appear that both medications have a more beneficial effect on the level of coughing as compared to the placebo, but that they are equally beneficial.

9 (i) (a) With n large we use normal approximation to t_{100}.

$$416 \pm 1.96\frac{72}{\sqrt{101}}$$

$$= 416 \pm 14.04 = (402.0, 430.0)$$

(b) Using $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{100}$

a 95% CI for σ^2 is $\frac{(n-1)S^2}{\chi^2_{100}(0.025)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{100}(0.975)}$

which gives $\left(\frac{100 \times 72^2}{129.6}, \frac{100 \times 72^2}{74.22}\right) = (4000, 6985).$

95% CI for standard deviation σ is therefore

$$\left(\sqrt{4000}, \sqrt{6985}\right) = (63.2, 83.6).$$

(ii) (a) 95% CI is $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$ and with $s = 72$ we have

$$\frac{1.96 \times 72}{\sqrt{n}} = 10 \Rightarrow n = 199.15$$

So $n \geq 200$.

(b) Taking s = 83.57 gives

$$\frac{1.96 \times 83.57}{\sqrt{n}} = 10 \Rightarrow n = 268.30, \text{ , so } n \geq 269.$$

(iii) Assuming a larger value of *s* results in a larger standard error, so a larger sample size is required to achieve the same width of confidence interval.

10 (i) $\hat{\mu} = \bar{X} = 1{,}175$

(ii) (a) $$S_{tt} = 6900 - \frac{160^2}{4} = 500$$

$$S_{xx} = 613379 - \frac{1179^2}{4} = 265{,}868.75$$

$$S_{tx} = 57515 - \frac{160 * 1179}{4} = 10{,}355$$

$$Corr(t, x) = \frac{S_{tx}}{\sqrt{S_{tt} * S_{xx}}} = \frac{10355}{\sqrt{500 * 265868.75}} = 0.898114$$

This implies that there is a strong linear relationship between age and number of deaths.

(b) Model: $\hat{x} = \hat{\alpha} + \hat{\beta}t$

$$\hat{\beta} = \frac{S_{tx}}{S_{tt}} = \frac{10355}{500} = 20.71,$$

$$\hat{\alpha} = \bar{x} - \hat{\beta}\bar{t} = \frac{1179}{4} - 20.71 * \frac{160}{4} = -533.65$$

Estimated model: $\hat{x} = 20.71t - 533.65$

(iii) $p(x_i, w, t_i) = \dfrac{\exp(-wt_i)(wt_i)^{x_i}}{x_i!}$

$\log p(x_i, w, t_i) = -wt_i + x_i(\log w + \log t_i) - \log(x_i!)$

$$\frac{\partial}{\partial w}\log p(x_i, w, t_i) = -t_i + \frac{x_i}{w}$$

$$\Sigma\frac{\partial}{\partial w}\log p(x_i, w, t_i) = -\Sigma t_i + \frac{1}{w}\Sigma x_i = 0$$

$$\hat{w} = \frac{\Sigma x_i}{\Sigma t_i}$$

(Second derivative gives $-\sum x_i / w^2 < 0$ which confirms maximum.)

For the observed values we obtain $\hat{w} = \dfrac{1179}{160} = 7.36875$

END OF EXAMINERS' REPORT

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

7 October 2011 (pm)

Subject CT3 — Probability and Mathematical Statistics Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 10 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is NOT required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 The first 20 claims that were paid out under a group of policies were for the following amounts (in units of £1,000):

3.2	2.1	6.3	4.0	3.8	4.4	6.5	7.8	2.8	5.2
7.0	8.1	4.4	5.8	1.7	2.8	5.0	3.2	3.7	4.4

For these data $\sum x = 92.2$.

(i) Calculate the mean of these 20 claim amounts. [1]

The next 80 claims paid out had a mean amount of £5,025.

(ii) Calculate the mean amount for the first 100 claims. [2]

[Total 3]

2 The claims which arose in a sample of policies of a certain class gave the following frequency distribution for the number of claims per policy in the last year:

Number of claims x	0	1	2	3	4 or more
Number of policies f	15	20	10	5	0

Calculate the third order moment about the origin for these data. [3]

3 A random sample of 60 adult men who live in Leeds includes 21 who have visited Majorca. An independent random sample of 70 adult women who live in Leeds includes 28 who have visited Majorca.

Calculate a 98% confidence interval for the proportion of adults who live in Leeds who have visited Majorca. [4]

4 The random variables X and Y are related as follows:

X conditional on $Y = y$ has a $N(2y, y^2)$ distribution.
Y has a $N(200, 100)$ distribution.

Derive the unconditional variance of X, $V[X]$. [3]

5 Consider the random variable X taking the value $X = 1$ if a randomly selected person is a smoker, or $X = 0$ otherwise. The random variable Y describes the amount of physical exercise per week for this randomly selected person. It can take the values 0 (less than one hour of exercise per week), 1 (one to two hours) and 2 (more than two hours of exercise per week). The random variable $R = (3 - Y)^2(X + 1)$ is used as a risk index for a particular heart disease.

The joint distribution of X and Y is given by the joint probability function in the following table.

	Y		
X	0	1	2
0	0.2	0.3	0.25
1	0.1	0.1	0.05

(i) Calculate the probability that a randomly selected person does more than two hours of exercise per week. [1]

(ii) Decide whether X and Y are independent or not and justify your answer. [2]

(iii) Derive the probability function of R. [3]

(iv) Calculate the expectation of R. [2]

[Total 8]

6 The number of claims made by each policyholder in a certain class of business is modelled as having a Poisson distribution with mean λ.

(i) Derive an expression for the probability, p, that a policyholder in this class has made at least one claim. [2]

The claims records of 20 randomly chosen policyholders were examined and the number of policyholders that made at least one claim in a year, X, was recorded.

(ii) (a) State the distribution of the random variable X and its parameters.

(b) Derive an expression for the maximum likelihood estimator of the probability p given in (i) using your answer in (ii)(a).

[4]

(iii) Show that, in the case $X = 5$, the maximum likelihood estimate (MLE) of p is $\hat{p} = 0.25$ and hence calculate the MLE of λ. [3]

It is now found that of the five policyholders who had made at least one claim there were four who had made exactly one claim and one who had made two claims.

(iv) Calculate the MLE of λ given this additional information. [4]

[Total 13]

7 The total amounts y_{ij} (in £ millions) paid out under a certain type of policy issued by four different companies A, B, C, D in each of six consecutive years were as follows:

Company							*Total*
A	2.870	3.125	3.000	2.865	2.890	3.060	17.810
B	3.105	3.200	3.300	2.975	3.210	3.150	18.940
C	2.800	2.985	3.060	2.900	2.920	3.050	17.715
D	2.830	2.600	2.765	2.690	2.600	2.700	16.185

For these data, $\Sigma_i \Sigma_j y_{ij} = 70.650$ and $\Sigma_i \Sigma_j y_{ij}^2 = 208.828$.

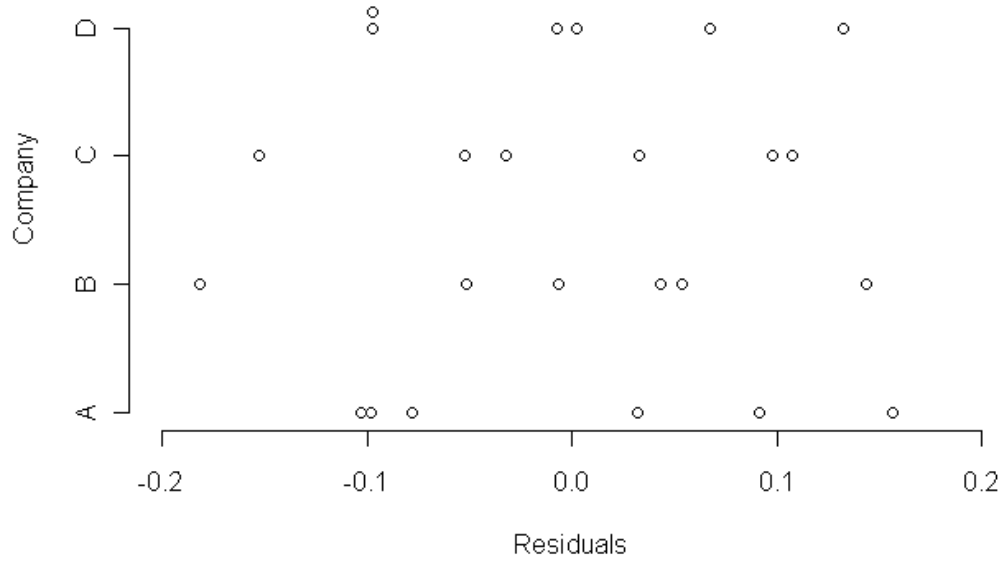
Consider the ANOVA model $Y_{ij} = \mu + \tau_i + e_{ij}$, $i = 1, \ldots, 4$, $j = 1, \ldots, 6$, where Y_{ij} is the jth amount paid out by company i, and $e_{ij} \sim N(0, \sigma^2)$ are independent errors.

The ANOVA table for these data is given below.

Source	*DF*	*SS*	*MS*
Company (between treatments)	3	0.640	0.213
Residual	20	0.212	0.0106
Total	23	0.852	

(i) Test the hypothesis that there are no differences in the means of the amounts paid out under such policies by the four companies (the company means), stating your conclusions clearly. [2]

(ii) Comment briefly on the validity of the test performed in (i), using the plot of the residuals given below. [2]



(iii) (a) Calculate the least significant difference between pairs of company means using a 5% significance level.

(b) List the company means in order, illustrate the non-significant pairs using suitable underlining, and comment briefly.

[6]
[Total 10]

8 Consider a random sample $X_1, \ldots, X_n$ from a Poisson distribution with expectation $E[X_i] = \lambda$. An estimator $\hat{\lambda}$ for the parameter λ is given by the observed mean of the sample, that is:

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

(i) Derive formulae for the expected value and the variance of $\hat{\lambda}$ in terms of λ and n. [3]

Assume in parts (ii) to (v) that the true parameter value is $\lambda = 0.25$.

(ii) Calculate the exact probability that $0.2 \leq \hat{\lambda} \leq 0.3$ if the sample size is $n = 10$. [3]

(iii) Calculate the approximate probability that $0.2 \leq \hat{\lambda} \leq 0.3$ if the sample size is $n = 10$ using the following:

(a) the normal approximation to $\sum_{i=1}^{n} X_i$ with continuity correction

(b) the normal approximation to $\sum_{i=1}^{n} X_i$ without continuity correction. [6]

(iv) Comment on the differences in your answers in parts (ii) and (iii). [2]

(v) Calculate the minimal required sample size n for which the probability that $0.2 \leq \hat{\lambda} \leq 0.3$ is at least 0.95, using the normal approximation without continuity correction [4]

Suppose a random sample of size $n = 400$ gives the estimate $\hat{\lambda} = 0.27$.

(vi) Calculate a 95% confidence interval for λ. [3]

[Total 21]

9 In a recent study of attitudes to a proposed new piece of consumer legislation ("proposal X") independent random samples of 200 men and 200 women were asked to state simply whether they were "for" (in favour of) , or "against", the proposal. The resulting frequencies, as reported by the consultants who carried out the survey, are given in the following table:

	Men	*Women*
For	138	130
Against	62	70

(i) Carry out a formal chi-squared test to investigate whether or not an association exists between gender and attitude to proposal X.

Note: in this and any later such tests in this question you should state the *P-value* of the data and your conclusion clearly. [6]

At a subsequent meeting to discuss these and other results, the consultants revealed that they had in fact stratified the survey, sampling 100 men and 100 women in England and 100 men and 100 women in Wales. The resulting frequencies were as follows:

	England		*Wales*	
	Men	*Women*	*Men*	*Women*
For	82	66	56	64
Against	18	34	44	36

A chi-squared test to investigate whether or not an association exists between gender and attitude to proposal X in England gives $\chi^2 = 6.653$, while an equivalent test for Wales gives $\chi^2 = 1.333$.

(ii) (a) Find the *P-value* for each of the chi-squared tests mentioned above and state your conclusions regarding possible association between gender and attitude to proposal X in England and in Wales.

(b) Discuss the results of the survey for England and Wales separately and together, quoting relevant percentages to support your comments. [9]

(iii) A different survey of 200 people conducted in each of England, Wales, and Scotland gave the following percentages in favour of another proposal:

	England	*Wales*	*Scotland*
% in favour of proposal	62%	53%	58%

A chi-squared test of association between country and attitude to the proposal gives $\chi^2 = 3.332$ on 2 degrees of freedom, with *P-value* 0.189.

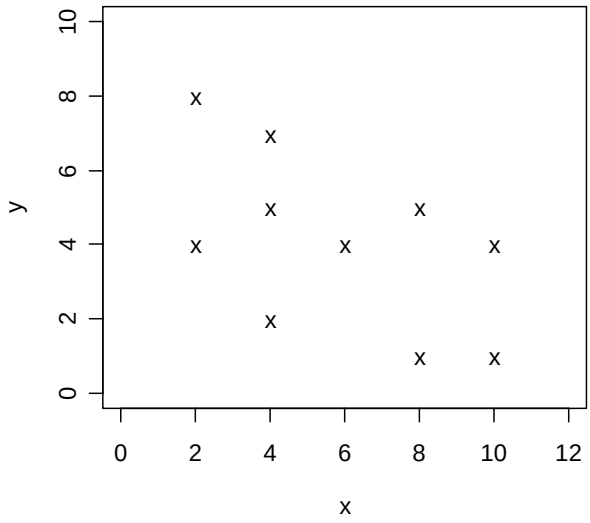
Suppose a second survey of the same size is conducted in the three countries and results in the same percentages in favour of the proposal as in the first survey. The results of the two surveys are now combined, giving a survey based on the attitudes of 1,200 people.

(a) State (or find) the results of a second chi-squared test for an association between country and attitude to the proposal, based on the overall survey of 1,200 people. [3]

(b) Comment briefly on the results. [1]

[Total 19]

10 Consider a situation in which integer-valued responses (y) are recorded at ten values of an integer-valued explanatory variable (x). The data are presented in the following scatter plot:



For these data: $\Sigma x = 58$, $\Sigma x^2 = 420$, $\Sigma y = 41$, $\Sigma y^2 = 217$, $\Sigma xy = 202$

(i) (a) Calculate the value of the coefficient of determination (R^2) for the data.

(b) Determine the equation of the fitted least-squares line of regression of y on x.

[7]

(ii) Calculate a 95% confidence interval for the slope of the underlying line of regression of y on x. [4]

(iii) (a) Calculate an estimate of the expected response in the case $x = 9$.

(b) Calculate the standard error of this estimate.

[3]

Suppose the observation ($x = 10$, $y = 8$) is added to the existing data. The coefficient of determination is now $R^2 = 0.07$.

(iv) Comment briefly on the effect of the new observation on the fit of the linear model. [2]

[Total 16]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2011 examinations

Subject CT3 — Probability and Mathematical Statistics
Core Technical

Purpose of Examiners' Reports

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and who are using past papers as a revision aid, and also those who have previously failed the subject. The Examiners are charged by Council with examining the published syllabus. Although Examiners have access to the Core Reading, which is designed to interpret the syllabus, the Examiners are not required to examine the content of Core Reading. Notwithstanding that, the questions set, and the following comments, will generally be based on Core Reading.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report. Other valid approaches are always given appropriate credit; where there is a commonly used alternative approach, this is also noted in the report. For essay-style questions, and particularly the open-ended questions in the later subjects, this report contains all the points for which the Examiners awarded marks. This is much more than a model solution – it would be impossible to write down all the points in the report in the time allowed for the question.

T J Birse
Chairman of the Board of Examiners

December 2011

General comments on Subject CT3

All valid alternative solutions receive credit as appropriate. Rounding errors are not penalised, unless if excessive rounding has led to significantly different answers. In cases where the same error is carried forward to later parts of the question, candidates are not penalised twice. In questions where comments are required, reasonable comments that are different than those provided in the solutions also receive credit.

Comments on the September 2011 paper

In general the paper was not answered as well as in recent diets. However, the overall performance was satisfactory with a number of candidates achieving notably high scores. Questions 8 and 9 were on topics frequently present in CT3 papers but perhaps examined from a slightly different angle, and as a result were answered less well. The comments on individual questions that follow concern specific parts that candidates answered poorly and important frequent errors.

1 (i) mean $\bar{x} = 4.61$ or £4,610.

(ii) mean of the whole 100 = $\frac{20(4610)+80(5025)}{100} = \frac{494200}{100} = £4,942$

Generally very well answered.

2 Required moment $= \frac{1}{\sum f}\sum_x fx^3$

$= \frac{1}{50}\left(15\times 0^3 + 20\times 1^3 + 10\times 2^3 + 5\times 3^3\right) = \frac{235}{50} = 4.7$

Many candidates calculated the third central moment (around the sample mean), rather than the moment around zero as required in the question.

3 Sample proportion = 49/130

Upper 1% normal percentage point = 2.326

98% CI is
$49/130 \pm 2.326*[(49/130)(81/130)/130]^{1/2}$ i.e. 0.3769 ± 0.0989 i.e. (0.278, 0.476)

Answers here were generally satisfactory. Some candidates erroneously computed CIs based on men and women separately.

4 $V[X] = E[V(X|Y)] + V[E(X|Y)]$
$= E[Y^2] + V[2Y] = (100 + 200^2) + 4*100 = 40,500$

Generally very well answered.

5 (i) $P[Y = 2] = 0.25 + 0.05 = 0.3$

(ii) $P[X = 0] = 0.75$ and

$P[\{X = 0\} \cap \{Y = 2\}] = 0.25 \neq 0.225 = 0.3 * 0.75 = P[X = 0] * P[Y = 2]$

Therefore X and Y are not independent.

(Any other joint probability can be used.)

(iii) The probability function is

r	1	2	4	8	9	18
$P(R = r)$	0.25	0.05	0.3	0.1	0.2	0.1

(iv) $E[R] = 0.2 * 9 + 0.3 * 4 + 0.25 * 1 + 0.1 * 18 + 0.1 * 8 + 0.05 * 2$
$= 1.8 + 1.2 + 0.25 + 1.8 + 0.8 + 0.1 = 5.95$

In part (ii) notice that one example of P(XY) = P(X)P(Y) is not sufficient for showing independence (it needs to hold for all cases). Also, some candidates failed to provide the probability function in (iii).

6 (i) $p = \Pr(N \geq 1) = 1 - \Pr(N = 0) = 1 - e^{-\lambda}$.

(ii) (a) $X \sim \text{Bin}(20,p)$

(b) $L(p) \propto p^X (1-p)^{20-X}$

$\Rightarrow l(p) = \log(L(p)) = X \log p + (20 - X)\log(1-p)$

and $l'(p) = 0 \Rightarrow \frac{X}{\hat{p}} - \frac{20-X}{1-\hat{p}} = 0 \Rightarrow X - 20\hat{p} = 0 \Rightarrow \hat{p} = \frac{X}{20}$

(and $l''(p) = -\frac{X}{\hat{p}^2} - \frac{20-X}{(1-\hat{p})^2} \leq 0$)

(iii) $\hat{p} = \frac{5}{20} = 0.25$

Then, using the invariance property of the MLE:

$\hat{p} = 1 - e^{-\hat{\lambda}} \Rightarrow \hat{\lambda} = -\log(1-\hat{p}) = -\log(0.75) = 0.288$

Likelihood function now is:

(iv) $L(\lambda) \propto P(X = 0)^{15} \times P(X = 1)^4 \times P(X = 2)$
$\propto \left(e^{-\lambda}\right)^{15} \left(\lambda e^{-\lambda}\right)^4 \left(\lambda^2 e^{-\lambda}\right)$

$\Rightarrow l(\lambda) \propto -15\lambda + 4\log\lambda - 4\lambda + 2\log\lambda - \lambda = -20\lambda + 6\log\lambda$

and $l'(\lambda) = 0 \Rightarrow -20 + \frac{6}{\hat{\lambda}} = 0 \Rightarrow \hat{\lambda} = 0.3$

(Also $l''(\lambda) = -\frac{6}{\hat{\lambda}^2} < 0$, hence max)

Notice that part (ii)(b) requires the use of the binomial distribution from (ii)(a). In part (iii) the invariance property must be used and mentioned for full credit.

7 (i) $F = 0.213/0.0106 = 20.094$ and at the 5% significance level, $F_{3,20}(0.05) = 3.098$.
Since $F = 20.094 > 3.098$, there is strong evidence against the null hypothesis, and we conclude that there are differences in the mean amounts paid out by the companies.

(ii) The variance of the residuals seems to be similar for the four companies; this is consistent with the assumption of constant variance in the response variable. Also there are no obvious patterns or outliers. The analysis seems valid.

(iii) (a) $LSD = t_{20}(0.025)\sqrt{\sigma^2\left(\frac{1}{6}+\frac{1}{6}\right)}$

$= 2.086\ \sqrt{0.0106/3} = 0.124$

(b) The four company (treatment) means are:

$$\bar{y}_{1\bullet} = \frac{17.810}{6} = 2.968,\ \bar{y}_{2\bullet} = \frac{18.940}{6} = 3.157,\ \bar{y}_{3\bullet} = \frac{17.715}{6} = 2.953$$

$$\bar{y}_{4\bullet} = \frac{16.185}{6} = 2.698$$

which are given in order and underlined as

$$\bar{y}_{4\bullet} < \underline{\bar{y}_{3\bullet} < \bar{y}_{1\bullet}} < \bar{y}_{2\bullet}$$

Amounts paid out by companies 2 and 4 are significantly different from those paid out by the other two companies. Company 4 seems to pay out significantly lower amounts, with Company 2 paying significantly higher.

Parts (i) and (ii) were generally well answered. In part (iii) many candidates did not use the correct formula for LSD and then performed pair-wise comparisons using the wrong statistic.

8 (i) $E[\hat{\lambda}] = \frac{1}{n}\sum_i E[X_i] = \lambda$

$V[\hat{\lambda}] = \frac{1}{n^2}\sum_i V[X_i] = \frac{1}{n}\lambda$ (using independence of X_i)

(ii) $P[0.2 \le \hat{\lambda} \le 0.3] = P[2 \le 10\hat{\lambda} \le 3]$

$= F(3; \lambda = 2.5) - F(1; \lambda = 2.5) = 0.75758 - 0.28730 = 0.47028$

(iii) (a) $10\hat{\lambda} = \sum_{i=1}^{10} X_i \sim N(2.5, 2.5)$ approximately.

With continuity correction:

$$P[0.2 \le \hat{\lambda} \le 0.3] = P[2 \le 10\hat{\lambda} \le 3] \approx P[2 - 0.5 \le 10\hat{\lambda} \le 3 + 0.5]$$

$$= P\left[\frac{1.5 - 2.5}{\sqrt{2.5}} \le Z \le \frac{3.5 - 2.5}{\sqrt{2.5}}\right] = 2 * F_Z\left(\frac{1}{\sqrt{2.5}}\right) - 1$$

$= 2^* F_Z(0.63246) - 1 = 2 * 0.73565 - 1 = 0.4713$

(b) $10\hat{\lambda} = \sum_{i=1}^{10} X_i = Y \approx N(2.5, 2.5)$ approximately.

Without continuity correction:

$$P[0.2 \le \hat{\lambda} \le 0.3] = P[2 \le 10\hat{\lambda} \le 3] \approx P\left[\frac{2 - 2.5}{\sqrt{2.5}} \le Z \le \frac{3 - 2.5}{\sqrt{2.5}}\right]$$

$$= 2^* F_Z\left(\frac{0.5}{\sqrt{2.5}}\right) - 1 = 2^* F_Z(0.32) - 1 = 2^* 0.62552 - 1 = 0.2510$$

(iv) When compared to the exact probability in (ii) the results in (iii) (a) and (b) show that the continuity correction reduces the approximation error significantly for this small sample size.

(v)

$$P[0.2 \le \hat{\lambda} \le 0.3] \approx P\left[\frac{0.2 - 0.25}{\sqrt{0.25/n}} \le z \le \frac{0.3 - 0.25}{\sqrt{0.25/n}}\right] = 2 * F_Z\left(\frac{0.05}{\sqrt{0.25/n}}\right) - 1 = 0.95$$

$$\frac{1.95}{2} = F(z), \text{ then } z = 1.96 = \frac{0.05}{\sqrt{0.25}}\sqrt{n}, \text{ and } \sqrt{n} = 1.96\frac{0.5}{0.05} = 1.96, \text{ and } n \approx 384$$

(vi) Using the normal approximation we find:

$$0.27 \pm z_{0.975}\sqrt{\frac{\hat{\lambda}}{n}} = 0.27 \pm 1.96\frac{\sqrt{0.27}}{20} = 0.27 \pm 0.05092 = [021908, 0.32092]$$

In part (i) independence must be mentioned for full marks in the derivation of the variance. In (ii) most candidates either went straight to a normal approximation, or incorrectly calculated the Poisson probability. In part (iii) many candidates applied the continuity correction wrongly.

9 (i) H_0: no association exists v. H_1: association exists

	men	*women*	
for	138	130	268
against	62	70	132
	200	200	400

Under H_0: expected frequencies: 134 134
66 66

O – E: 4 –4
–4 4

$$\chi^2 = 4^2\left(\frac{1}{134} + \frac{1}{134} + \frac{1}{66} + \frac{1}{66}\right) = 0.724$$

P-value = $P(\chi^2_1 > 0.724) = 0.395$

No evidence against H_0 – we conclude that no association exists between gender and attitude to proposal X.

[*Note:* using Yates' correction (not in the Core Reading)
P-value = $P(\chi^2_1 > 0.554) = 0.457$]

(ii) (a) For England:

P-value = $P(\chi^2_1 > 6.653) = 0.010$

Evidence against H_0 – we reject it at the 1% level of testing and conclude that an association exists between gender and attitude to proposal X in England.

For Wales:

P-value = $P(\chi^2_1 > 1.333) = 0.248$

No evidence against H_0 – we conclude that there is no association between gender and attitude to proposal X in Wales.

(b) England: there is evidence of an association – 82% of men and only 66% of women support proposal X – these proportions are significantly different.

Wales: there is no evidence of an association – 56% of men and 64% of women support proposal X – these proportions are not significantly different.

The effects are in different directions and cancel out to some extent when the data are combined: now there is no evidence of an association – overall 69% of men and 65% of women support proposal X – these proportions are not significantly different.

The combined data give a misleading message – they hide the effect of the factor "country" and fail to reveal that there is an association in England.

(iii) (a) The χ^2 value doubles to 6.664

P-value = $P(\chi^2_2 > 6.664) = 0.0357$

Conclusion: reject "no association" at the 3.6% level of testing and conclude that an association does exist.

(b) Comment: having more data with the same proportions provides strong enough evidence to justify claiming that an association exists.

Caution required with the null and alternative hypotheses in (i) – some candidates got these wrong. Also, the associated degrees of freedom were wrongly given in some cases. Part (ii)(b) required comments on the results, but very few candidates did this. Part (iii) was not well answered either.

10 (i) (a) $S_{xx} = 420 - 58^2/10 = 83.6$, $S_{yy} = 217 - 41^2/10 = 48.9$,
$S_{xy} = 202 - 58*41/10 = -35.8$

$SS_T = 48.9$, $SS_{REG} = (-35.8)^2/83.6 = 15.3306$
$\Rightarrow R^2 = 15.3306/48.9 = 0.3135$ (or 31.4%)

[OR using $R^2 = S_{xy}^2/S_{xx}S_{yy}$]

(b) Fitted line $y = \hat{\alpha} + \hat{\beta}x$:

$\hat{\beta} = -35.8 / 83.6 = -0.42823$, $\hat{\alpha} = 4.1 - (-0.42823 * 5.8) = 6.58373$

Fitted line is $y = 6.5837 - 0.4282x$

(ii) $\hat{\sigma}^2 = (48.9 - 15.3306) / 8 = 4.1962$

$s.e.(\hat{\beta}) = (4.1962 / 83.6)^{1/2} = 0.2240$

95% confidence interval for β is $\hat{\beta} \pm t_8 * s.e.(\hat{\beta})$

i.e. $-0.42823 \pm 2.306 * 0.2240$

i.e. $(-0.945, 0.088)$

(iii) (a) At $x = 9$, $\hat{y} = 6.5837 - 0.4282 * 9 = 2.7299$ i.e. 2.730

(b) $s.e.^2 = \left(\frac{1}{10} + \frac{(9 - 5.8)^2}{83.6} \right) 4.1962 = 0.93360 \Rightarrow s.e. = 0.9662$

(iv) Addition of new observation makes data more randomly scattered. The strength of the linear relationship is reduced from "weak" to "almost nothing".

Generally well answered. Some problems were encountered in part (iii)(b), where the wrong formula was used.

END OF EXAMINER'S REPORT

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

19 April 2012 (am)

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 13 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is NOT required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

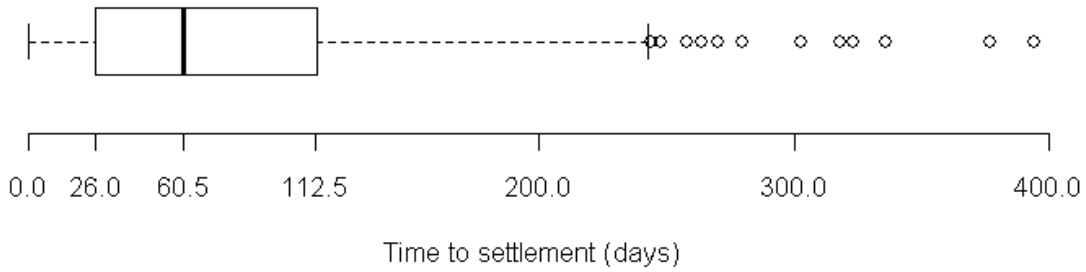
1 The following 24 observations give the length of time (in hours, ordered) for which a specific fully charged laptop computer will operate on battery before requiring recharging.

1.2	1.4	1.5	1.6	1.7	1.7	1.8	1.8	1.9	1.9	2.0	2.0
2.1	2.1	2.1	2.2	2.3	2.4	2.4	2.5	3.1	3.6	3.7	4.5

The owner of this computer is about to watch a film on his fully charged laptop.

Calculate from these data the longest showing time for a film that he can watch, so that the probability that the battery's lifetime will be sufficient for watching the entire film is 0.75. [3]

2 Data were collected on the time (in days) until each of 200 claims is settled by the insurer in a certain insurance portfolio. A boxplot of the data is shown below.



(i) Calculate the median and the interquartile range of the data using the plot. [2]

(ii) Comment on the distribution of the data as shown in the plot. [2]

[Total 4]

3 Two students are selected at random without replacement from a group of 100 students, of whom 64 are male and 36 are female.

Calculate the probability that the two selected students are of different genders. [3]

4 Claim amounts arising under a particular type of insurance policy are modelled as having a normal distribution with standard deviation £35. They are also assumed to be independent from each other.

Calculate the probability that two randomly selected claims differ by more than £100. [4]

5 Claims on a group of policies arise randomly and independently of each other through time at an average rate of 2 per month.

(i) Calculate the probability that no claims arise in a particular month. [2]

(ii) Calculate the probability that more than 30 claims arise in a period of one year. [2]

[Total 4]

6 In a random sample of 200 people taken from a large population of adults, 70 people intend to vote for party A at the next election.

(i) Calculate an approximate equal-tailed 95% confidence interval for θ, the true proportion of this population who intend to vote for party A at the next election. [3]

(ii) Give a brief interpretation of the interval calculated in part (i). [1]

[Total 4]

7 A coin has two sides, "heads" and "tails". Such a coin with $P(\text{heads}) = p$ is tossed repeatedly until it lands "heads" for the first time. Let X be the number of tosses required.

Suppose the process is repeated independently a total of n times, producing values of the variables $X_1, X_2, \ldots, X_n$, where each X_i has the same distribution as X.

Let $Y = \min(X_1, X_2, \ldots, X_n)$, so Y is the smallest number of tosses required to produce a "heads" in the n repetitions of the experiment.

(i) Explain why, for each $i = 1, 2, \ldots, n$, $P(X_i \geq x)$ is given by

$$P(X_i \geq x) = (1 - p)^{x-1}, \quad x = 1, 2, \ldots .$$ [2]

(ii) (a) Find an expression for $P(Y \geq y)$.

(b) Hence deduce the probability function of Y.

[5]

[Total 7]

8 In an analysis of variance investigation four treatments are compared using random samples each of size 10. The total sum of squares is calculated as $SS_T = 673.5$ and the between-treatments sum of squares as $SS_B = 148.3$.

(i) (a) Calculate an unbiased estimate of the error variance σ^2.

(b) State the number of degrees of freedom associated with the estimate in part (i)(a).

[3]

(ii) Suggest an unbiased estimator of σ^2 that is different from the one used in part (i). [1]

(iii) Comment on which of the two estimators should be used.

[2]

[Total 6]

9 A random sample of 200 email messages was selected from all messages delivered through an internet provider company. Each message is monitored for the presence of computer viruses. It is assumed that each message contains a virus with the same probability p, independently from all other messages.

Let Y_i, $i = 1,\ldots,200$ be indicator random variables taking the value 1 if message i contains a virus, and 0 otherwise. Also, let Y denote the total number of messages in this sample found to contain viruses, i.e. $Y = \sum_{i=1}^{200} Y_i$.

(i) Derive expressions for the expected value and the variance of Y in terms of the parameter p, using the indicator variables $Y_1, Y_2, \ldots, Y_{200}$. [4]

(ii) Explain why the approximate distribution of Y is $N(200p, 200p(1-p))$, using the indicator variables $Y_1, Y_2, \ldots, Y_{200}$. [3]

It is found that 38 email messages in this sample contained viruses.

(iii) Calculate an equal-tailed 90% confidence interval for the probability p using the approximate normal distribution from part (ii). [3]

[Total 10]

10 In a portfolio of car insurance policies, the number of accident-related claims, N, made by a policyholder in a year has the following distribution:

No. of claims, n	0	1	2
Probability	0.4	0.4	0.2

The number of cars, X, involved in each accident that results in a claim is distributed as follows:

No. of cars, x	1	2
Probability	0.7	0.3

It can be assumed that the occurrence of a claim and the number of cars involved in the accident are independent. Furthermore, claims made by a policyholder in any year are also independent of each other. Let S be the total number of cars involved in accidents related to such claims by a policyholder in a year.

(i) (a) Determine the probability function of S.
(b) Hence find $E(S)$.

[4]

The expectation $E(S)$ can also be calculated using the formula

$$E(S) = \sum_{n=0}^{2} E(S|N=n)\Pr(N=n).$$

(ii) (a) Find $E(S|N=n)$ for $n = 0,1,2$.
(b) Hence calculate $E(S)$.

[4]
[Total 8]

11 An experiment has three possible outcomes (A, B, C) and a model states that the probabilities of these outcomes are θ, θ^2, and $1 - \theta - \theta^2$ respectively, for some suitable value of $\theta > 0$.

Let n_A, n_B, and n_C be the number of occurrences of outcomes A, B, and C respectively in n ($= n_A + n_B + n_C$) repetitions of the experiment. Let $\ell(\theta)$ represent the log-likelihood function, and let $U(\theta) = \dfrac{\partial \ell(\theta)}{\partial \theta}$.

(i) (a) Show that

$$U(\theta) = \frac{n_A + 2n_B}{\theta} - \frac{n_C(1+2\theta)}{1-\theta-\theta^2}.$$

(b) Hence find a quadratic equation whose solution gives the maximum likelihood estimate of θ.

[5]

(ii) (a) Find an expression for $\dfrac{\partial U(\theta)}{\partial \theta}$.

(b) Hence show that

$$E\left[-\frac{\partial U(\theta)}{\partial \theta}\right] = \frac{n\left(1+4\theta-\theta^2\right)}{\theta\left(1-\theta-\theta^2\right)}.$$ [4]

The results of 100 repetitions of the experiment show that outcome A occurred 51times, outcome B occurred 16 times, and outcome C occurred 33 times.

(iii) (a) Show that the maximum likelihood estimate of θ is $\hat{\theta} = 0.4525$.

(b) Calculate an estimate of the asymptotic standard error of $\hat{\theta}$.

(c) Find an approximate 95% confidence interval for θ.

[6]

[Total 15]

12 Consider a random sample $X_1,\ldots,X_k$ of size $k = 400$. Statistician A wants to use a χ^2-test to test the hypothesis that the distribution of X_i is a binomial distribution with parameters $n = 2$ and unknown p based on the following observed frequencies of outcomes of X_i:

Possible realisation of X_i	0	1	2
Frequency	90	220	90

(i) Estimate the parameter p using the method of moments. [2]

(ii) Test the hypothesis that X_i has a binomial distribution at the 0.05 significance level using the data in the above table and the estimate of p obtained in part (i). [5]

Statistician B assumes that the data are from a binomial distribution and wants to test the hypothesis that the true parameter is $p_0 = 0.5$.

(iii) Explain whether there is any evidence against this hypothesis by using the estimate of p in part (i) and without performing any further calculations. [2]

Statistician C wants to test the hypothesis that the random variables X_i have a binomial distribution with known parameters $n = 2$ and $p = 0.5$.

(iv) Write down the null hypothesis and the alternative hypothesis for the test in this situation. [2]

(v) Carry out the test at the significance level of 0.05 stating your decision. [3]

(vi) Explain briefly the relationship between the test decisions in parts (ii), (iii) and (v), and in particular whether there is any contradiction. [3]

[Total 17]

13 The quality of primary schools in eight regions in the UK is measured by an index ranging from 1 (very poor) to 10 (excellent). In addition the value of a house price index for these eight regions is observed. The results are given in the following table:

Region i	1	2	3	4	5	6	7	8	Sum
School quality index x_i	7	8	5	8	4	9	6	9	56
House price index y_i	195	195	170	190	150	190	200	210	1500

The last column contains the sum of all eight columns.

From these values we obtain the following results:

$$\sum x_i y_i = 10,695; \quad \sum x_i^2 = 416; \quad \sum y_i^2 = 283,750$$

(i) Calculate the correlation coefficient between the index of school quality and the house price index. [4]

You can assume that the joint distribution of the two random variables is a bivariate normal distribution.

(ii) Perform a statistical test for the null hypothesis that the true correlation coefficient between the school quality index and the house price index is equal to 0.8 against the alternative that the correlation coefficient is smaller than 0.8, by calculating an approximate p-value. [6]

(iii) Fit a linear regression model to the data, by considering the school quality index as the explanatory variable. You should write down the model and estimate all parameters. [3]

(iv) Calculate the coefficient of determination R^2 for the regression model obtained in part (iii). [1]

(v) Provide a brief interpretation of the slope of the regression model obtained in part (iii). [1]

[Total 15]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2012 examinations

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Introduction

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and who are using past papers as a revision aid, and also those who have previously failed the subject. The Examiners are charged by Council with examining the published syllabus. Although Examiners have access to the Core Reading, which is designed to interpret the syllabus, the Examiners are not required to examine the content of Core Reading. Notwithstanding that, the questions set, and the following comments, will generally be based on Core Reading.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report. Other valid approaches are always given appropriate credit; where there is a commonly used alternative approach, this is also noted in the report. For essay-style questions, and particularly the open-ended questions in the later subjects, this report contains all the points for which the Examiners awarded marks. This is much more than a model solution – it would be impossible to write down all the points in the report in the time allowed for the question.

T J Birse
Chairman of the Board of Examiners

July 2012

General comments on Subject CT3

All valid alternative solutions receive credit as appropriate. Rounding errors are not penalised, unless if excessive rounding has led to significantly different answers. In cases where the same error is carried forward to later parts of the question, candidates are not penalised twice. In questions where comments are required, reasonable comments that are different than those provided in the solutions also receive credit.

Comments on the April 2012 paper

The performance of candidates was overall better than in the last session (September 2011), but generally not as strong as in previous diets. There were some excellent scripts achieving very high scores, but a few poor efforts were also recorded at the other end.

In general, answers were not as satisfactory as expected when questions deviated from the usual context, although dealing with commonly examined concepts – e.g. Q1, Q7, Q10. Also, many problems were encountered with straightforward algebraic manipulations, such as differentiation in Q11.

The comments on individual questions that follow concern specific parts that candidates answered poorly and important frequent errors.

1 We want the *first quartile* of the data.

$$Q_1 = \left(\frac{n+2}{4}\right)th \text{ observation counting from below} = 6.5\text{th observation}$$

$$= \frac{1.7+1.8}{2} = 1.75 \text{ hours.}$$

[With alternative definition:

$$Q_1 = \left(\frac{n+1}{4}\right)th \text{ observation counting from below} = 1.725\,]$$

Most answers were quite poor. Many candidates tried to work with a normal or t distribution, when this was not justified (or required). Only a small number of candidates realised that quartiles were required – but then a large proportion of them used the wrong quartile.

2 (i) From plot median = 60.5 days, *IQR* = 112.5 – 26 = 86.5 days.

(ii) The distribution is skewed to the right and a number of values appear to be outliers.

Well answered.

3 P(1st selected is male and 2nd selected is female) = $\frac{64}{100}.\frac{36}{99}$

P(1st selected is female and 2nd selected is male) = $\frac{36}{100}.\frac{64}{99}$

$\Rightarrow$ P(selected students are of different genders) = $2.\frac{64}{100}.\frac{36}{99} = \frac{128}{275} = 0.465$

[*OR* P(selected students are of different genders) = $\frac{\binom{64}{1}\binom{36}{1}}{\binom{100}{2}} = \frac{64\times 36\times 2}{100\times 99} = 0.465\,]$

Very well managed by most candidates. Some tried to calculate the probabilities with replacement.

4 Claim amount ~ $N(\mu, 35^2)$ $\Rightarrow$ difference between 2 claim amounts $D \sim N(0, 2\times35^2)$

i.e. $D \sim N(0, 2450)$

$\Rightarrow P(|D| > 100) = P(|Z| > 100/2450^{1/2}) = P(|Z| > 2.020) = 2 * 0.0217 = 0.043$

Performance was mixed here. There were some problems with specifying the correct variance, and a number of answers gave a one-sided probability.

5 (i) number of claims in a month $X \sim$ Poisson(2)

from tables: $P(X=0) = 0.1353$

[*alternative:* $P(X=0) = e^{-2}$]

(ii) number of claims in a year $X \sim$ Poisson(24)

from tables: $P(X > 30) = 1 - P(X \le 30) = 1 - 0.9042 = 0.0958$

[*alternative:* use normal approximation with continuity correction which gives 0.0923]

Well answered by majority of candidates.

6 (i) Use the normal approximation $\hat{\theta} \sim N\left(\theta, \frac{\theta(1-\theta)}{200}\right)$

to give the 95% confidence interval $\hat{\theta} \pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{200}}$.

With $\hat{\theta} = 70/200 = 0.35$ we obtain 0.35 ± 0.066, that is (0.284, 0.416).

(ii) If we take a large number of samples from this population, we expect 95% of the resulting CIs to include the true value of θ.

There were no problems with the first part. However, many candidates struggled with providing a reasonable interpretation in part (ii).

7 (i) "$X_i \geq x$" ≡ "no heads in first x–1 tosses" so $P(X_i \geq x) = (1-p)^{x-1}$, $x = 1,2,3, \ldots$

[*OR* Recognise (as geometric) and sum the probabilities

$(1-p)^{x-1}p + (1-p)^{x}p + (1-p)^{x+1}p + \ldots = p(1-p)^{x-1}\{1-(1-p)\}^{-1}$]

(ii) (a) "$Y \geq y$" ≡ "all X_i's are $\geq y$"

so $P(Y \geq y) = P(X_1 \geq y, \ldots, X_n \geq y) = P(X_1 \geq y) \ldots P(X_n \geq y)$ (independent)

$= ((1-p)^{y-1})^n = ((1-p)^n)^{y-1}$

(b) The probability in part (a) implies that Y has the same distribution as X, but with $1-(1-p)^n$ in place of p

i.e. $P(Y = y) = r(1-r)^{y-1}$, $y = 1, 2, 3, \ldots$ where $r = 1-(1-p)^n$.

[*OR* $P(Y = y) = P(Y \geq y) - P(Y \geq y+1) = ((1-p)^n)^{y-1} - ((1-p)^n)^y$

$= (1-p)^{n(y-1)}\{1-(1-p)^n\}$ as above]

Most candidates had problems with part (ii). Carefully expressed probability statements are required here. A common error was to try to differentiate the CDF, despite this being a discrete distribution.

8 (i) (a) $SS_R = SS_T - SS_B = 673.5 - 148.3 = 525.2$

$$\hat{\sigma}^2 = \frac{SS_R}{n-k} = \frac{525.2}{36} = 14.59$$

(b) Associated d.f. 36

(ii) Alternatively, an unbiased estimator could be given using only part of the data, e.g. responses from treatment i: $S_i^2 = \dfrac{\sum_j (Y_{ij} - \bar{Y}_{i.})^2}{n_i - 1}$

(iii) The estimator used in part (i) should be preferred as it is based on all data and is therefore more accurate.

Part (i) was well answered, although the df were wrongly given in many answers. Answers in part (ii) were very poor – this question required good understanding of ANOVA concepts and critical thinking.

9 (i) The variables Y_i are independent between them and have a Bernoulli(p) distribution with mean p and variance $p(1-p)$.

Therefore $E(Y) = E(Y_1 + \cdots + Y_{200}) = E(Y_1) + \cdots + E(Y_{200}) = 200p$

$V(Y) = V(Y_1 + \cdots + Y_{200}) = V(Y_1) + \cdots + V(Y_{200}) = 200p(1-p)$

(ii) Again, with Y_i being iid Bernoulli(p) random variables and n being sufficiently large,

the central limit theorem implies that $Y = \sum_{i=1}^{200} Y_i$ follows approximately a normal distribution with mean and variance given by the mean and variance of Y as derived in (i), i.e.

$Y \sim N(200p, 200p(1-p))$

(iii) From (ii) $\hat{P} = Y/200 \sim N(p, p(1-p)/200)$ approximately, which gives a 90% confidence interval of the form $\hat{p} \pm z_{0.05}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{200}}$

$\hat{p} = 0.19$ giving $0.19 \pm 1.6445 \times 0.02774$

i.e. (0.144, 0.236).

Generally well tackled. Some students failed to work with the indicator variables (Bernoulli), which was key to this question.

10 (i) (a) S takes values 0, 1, 2, 3, 4 and we have

$$P(S=0) = 0.4$$
$$P(S=1) = 0.4 \times 0.7 = 0.28$$
$$P(S=2) = 0.4 \times 0.3 + 0.2 \times 0.7^2 = 0.218$$
$$P(S=3) = 0.2 \times 2 \times 0.7 \times 0.3 = 0.084$$
$$P(S=4) = 0.2 \times 0.3^2 = 0.018$$

(b) Hence

$$E(S) = 0.28 + 2 \times 0.218 + 3 \times 0.084 + 4 \times 0.018 = 1.04$$

(ii) (a) $E(S|N=0) = 0, \quad E(S|N=1) = E(X) = 0.7 + 2 \times 0.3 = 1.3$
$E(S|N=2) = E(2X) = 2 \times 1.3 = 2.6$

(b) Hence, $E(S) = 1.3 \times 0.4 + 2.6 \times 0.2 = 1.04$ as before.

Most candidates encountered problems here, as they failed to work out the probability function from first principles in part (i). Also, many did not recognise this as a compound distribution type of question.

11 (i) (a) $L(\theta) = k\theta^{n_A}\left(\theta^2\right)^{n_B}\left(1-\theta-\theta^2\right)^{n_C}$

$$\ell(\theta) = (n_A + 2n_B)\log\theta + n_C \log\left(1-\theta-\theta^2\right) + c$$

$$U(\theta) = \frac{n_A + 2n_B}{\theta} - \frac{n_C(1+2\theta)}{1-\theta-\theta^2}$$

(b) Setting $U(\theta) = 0 \Rightarrow (n_A + 2n_B)(1 - \theta - \theta^2) = n_C\theta(1 + 2\theta)$

$\Rightarrow \hat{\theta}$ satisfies

$$(n_A + 2n_B + 2n_C)\theta^2 + (n_A + 2n_B + n_C)\theta - (n_A + 2n_B) = 0$$

(ii) (a) $$\frac{\partial U(\theta)}{\partial\theta} = -\frac{n_A + 2n_B}{\theta^2} - n_C \frac{2\left(1-\theta-\theta^2\right) - (1+2\theta)(-1-2\theta)}{\left(1-\theta-\theta^2\right)^2}$$

$$= -\frac{n_A + 2n_B}{\theta^2} - \frac{n_C\left(3+2\theta+2\theta^2\right)}{\left(1-\theta-\theta^2\right)^2}$$

(b) $$E\left[-\frac{\partial U(\theta)}{\partial\theta}\right] = \frac{n\theta + 2n\theta^2}{\theta^2} + \frac{n\left(1-\theta-\theta^2\right)\left(3+2\theta+2\theta^2\right)}{\left(1-\theta-\theta^2\right)^2}$$

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

(iii) (a) $\hat{\theta}$ satisfies $149\theta^2 + 116\theta - 83 = 0 \Rightarrow \hat{\theta} = 0.4525$

(b) Using the Cramer-Rao lower bound, estimate of asymptotic standard error is

$$\left[\frac{\hat{\theta}(1-\hat{\theta}-\hat{\theta}^2)}{100\left(1+4\hat{\theta}-\hat{\theta}^2\right)}\right]^{1/2} = 0.0244$$

(c) 95% CI for θ is 0.4525 ± (1.96×0.0244) i.e. 0.4525 ± 0.0478 i.e. (0.405, 0.500)

Part (i) was very well answered. The differentiation in part (ii) was problematic. Also, many candidates could not identify the random variable for which expectation was required in (ii)(b).

12 (i) $\hat{p} = \frac{\bar{X}}{n} = \frac{220+2*90}{400*2} = 0.5$

We obtain the following table to test H_0:

Possible realisation of X_i	0	1	2
Number of observations	90	220	90
expected frequency under H_0	100	200	100
$(f_j - e_j)^2$	100	400	100
$(f_j - e_j)^2 / e_j$	1	2	1

The test-statistic is $= \sum_{j=0}^{2} (f_j - e_j)^2 / e_j$. For the given data the value of C is $c = 4$.

C is χ^2-distributed with 3−1−1 = 1 degree of freedom.

H_0 is rejected since the $(1-\alpha)$-quantile (α = 0.05) of the χ^2-distribution with one degree of freedom is 3.841 < 4.

(iii) Since the estimated value is 0.5, any reasonable test will not reject that value, since the value 0.5 will always be in the acceptance region of the test. In other words, 0.5 will always be in any confidence interval around the estimate 0.5.

(iv) We now have: $H_0 : X_i \sim Bin(2,0.5)$ and

$H_1 : X_i$ does not follow $Bin(2,0.5)$ (emphasis on both Bin, $p = 0.5$)

(v) The value of the test-statistic is still $c = 4$ but the distribution of C is now a χ^2-distribution with 3−1 = 2 degrees of freedom.

Now H_0 is NOT rejected at a 5%-level since the $(1-\alpha)$-quantile ($\alpha = 0.05$) of the χ^2-distribution with two degrees of freedom is $5.991 > 4$.

(vi) The result in part (ii) states that a binomial distribution does not fit the data well and is rejected. However, in part (iii) we found that, under the assumption of a binomial distribution, $p_0 = 0.5$ cannot be rejected. A specific binomial distribution with parameter $p = 0.5$ is not rejected in part (v) for the same data. The reason is that the additional degree of freedom in part (v) allows for a larger value of the test-statistic under the null.

Most candidates answered very well the parts of this question that concerned "knowledge" and "application" aspects of the tests. However, there were problems with the comments and reasoning.

13 (i) $S_{xx} = 24,\ S_{yy} = 2500,\ S_{xy} = 195$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\ S_{yy}}} = 0.796084$$

(ii) $W = \frac{1}{2}\log\frac{1+r}{1-r}$ is normally distributed with mean $\frac{1}{2}\log\frac{1+\rho}{1-\rho}$ and standard deviation $1/\sqrt{n-3}$

observed value of W is $w = 1.087828$

Under H_0 the mean of W is 1.098612

And the standard deviation is 0.447214

p-value is $P[W < 1.087828] = P[Z < (1.087828 - 1.098612)/\ 0.447214]$
$= P[Z < -0.024113527] = 1 - F(0.024113527) > 0.49$

No evidence against the null hypothesis.

(iii) $Y_i = a + bX_i + \varepsilon_i$

For b we obtain: $\hat{b} = \left\{n\sum x_i y_i - \sum x_i \sum y_i\right\}\left\{n\sum x_i^2 - (\sum x_i)^2\right\}^{-1}$

And therefore: $\hat{b} = \frac{8*10695 - 56*1500}{8*416 - 56^2} = 8.125$

$$\text{And } \hat{a} = \frac{1}{8}\left(\sum y_i - \hat{b}\sum x_i\right) = 130.625$$

(iv) $R^2 = 0.796084^2 = 0.634$

(v) Any increase in school quality by 1 index-point, leads to an increase of 8.125 in the house price index.

Mostly very well answered.

END OF EXAMINERS' REPORT

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

28 September 2012 (am)

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 13 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is NOT required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 Calculate the mean, the median and the mode for the data in the following frequency table.

Observation	0	1	2	3	4
Frequency	20	54	58	28	0

[3]

2 The following data are sizes of claims (ordered) for a random sample of 20 recent claims submitted to an insurance company:

174	214	264	298	335	368	381	395	402	442
487	490	564	644	686	807	1092	1328	1655	2272

(i) Calculate the interquartile range for this sample of claim sizes. [3]

(ii) Give a brief interpretation of the interquartile range calculated in part (i). [1]

[Total 4]

3 Let X be a discrete random variable with the following probability distribution:

X	0	1	2	3
$P(X = x)$	0.4	0.3	0.2	0.1

Calculate the variance of Y, where $Y = 2X + 10$. [3]

4 Consider a random variable U that has a uniform distribution on (0,1) and a random variable X that has a standard normal distribution. Assume that U and X are independent.

Determine an expression for the probability density function of the random variable $Z = U + X$ in terms of the cumulative distribution function of X. [4]

5 A large portfolio consists of 20% class A policies, 50% class B policies and 30% class C policies. Ten policies are selected at random from the portfolio.

(i) Calculate the probability that there are no policies of class A among the randomly selected ten. [1]

(ii) (a) Calculate the expected number of class B policies among the randomly selected ten.

(b) Calculate the probability that there are more than five class B policies among the randomly selected ten.

[2]

[Total 3]

6 A random sample of size n is taken from a gamma distribution with parameters $\alpha = 8$ and $\lambda = 1/\theta$. The sample mean is $\bar{X}$ and θ is to be estimated.

(i) Determine the method of moments estimator (MME) of θ. [2]

(ii) Find the bias of the MME determined in part (i). [2]

(iii) (a) Determine the mean square error of the MME of θ.

(b) Comment on the efficiency of the MME of θ based on your answer in part (iii)(a).

[3]

[Total 7]

7 Analyst A collects a random sample of 30 claims from a large insurance portfolio and calculates a 95% confidence interval for the mean of the claim sizes in this portfolio. She then collects a different sample of 100 claims from the same portfolio and calculates a new 95% confidence interval for the mean claim size.

(i) Explain how the widths of the two confidence intervals will differ. [2]

Analyst B obtains a 95% confidence interval for the mean claim size of this portfolio based on a different sample of 30 claims. She subsequently realises that one of the claims in the sample has an extremely large value and can be considered as an outlier. She decides to replace this claim with a new randomly selected one, whose size is not an outlier, and obtains a new 95% confidence interval.

(ii) Explain how the two confidence intervals will differ in the case of Analyst B.

[3]

[Total 5]

8 The random variable S is given as $S = Y_1 + Y_2 + \ldots + Y_N$ (with $S = 0$ if $N = 0$) where the random variables Y_i are identically and independently distributed according to a lognormal distribution with parameters $\mu = 0.5$ and $\sigma^2 = 0.1$. N is also a random variable which is independent of Y_i, and its distribution given below.

N	0	1	2	3	4
$\Pr(N = n)$	0.1	0.3	0.3	0.2	0.1

Calculate the mean and the variance of the random variable S. [7]

9 An analyst is interested in using a gamma distribution with parameters $\alpha = 2$ and $\lambda = ½$, that is, with density function $f(x) = \frac{1}{4} x e^{-\frac{1}{2}x}$, $0 < x < \infty$.

(i) (a) State the mean and standard deviation of this distribution.

(b) Hence comment briefly on its shape.

[2]

(ii) Show that the cumulative distribution function is given by

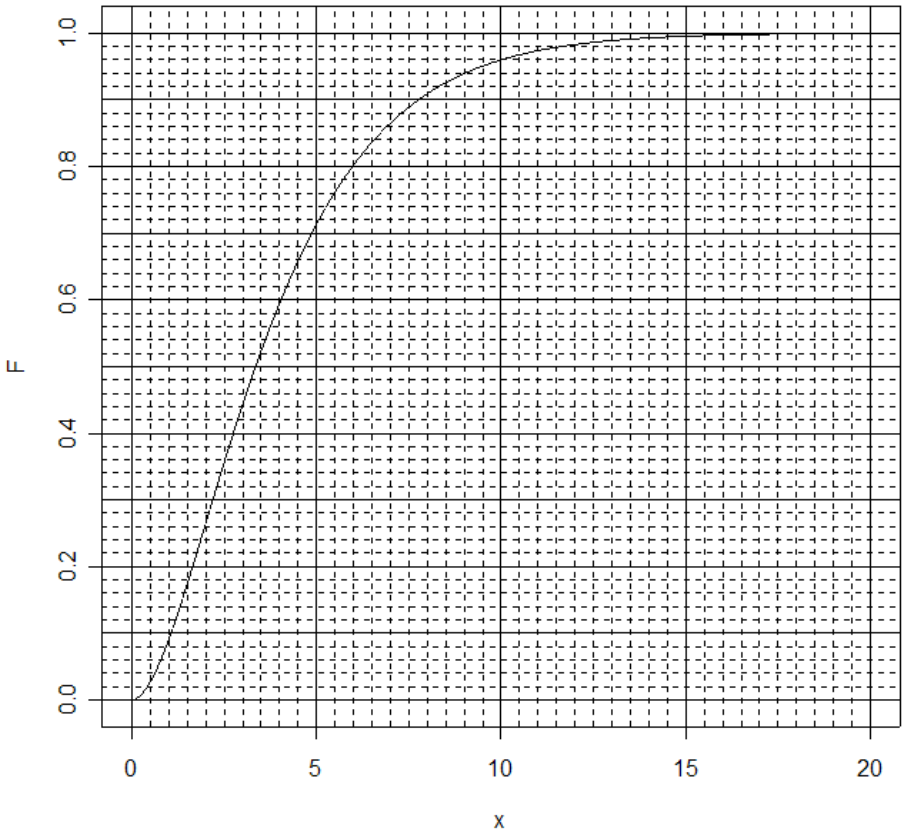
$$F(x) = 1 - (1 + \frac{1}{2}x)e^{-\frac{1}{2}x}, \quad 0 < x < \infty \quad \text{(zero otherwise).}$$ [3]

The analyst wishes to simulate values x from this gamma distribution and is able to generate random numbers u from a uniform distribution on (0,1).

(iii) (a) Specify an equation involving x and u, the solution of which will yield the simulated value x.

(b) Comment briefly on how this equation might be solved.

(c) The graph below gives $F(x)$ plotted against x. Use this graph to obtain the simulated value of x corresponding to the random number $u = 0.66$.



[3]
[Total 8]

10 The number of hours that people watch television per day is the subject of an empirical study that is carried out in four regions in a country. Five people are randomly selected in each of the regions and are asked about the average number of hours per day that they spent watching television during the last year. The results are shown in the following table, with the last column shows the average in each region.

						Average
Region 1	2.0	1.1	0.2	3.8	2.8	1.98
Region 2	1.2	1.0	0.9	1.1	1.6	1.16
Region 3	2.5	2.0	2.6	2.4	2.3	2.36
Region 4	1.2	1.7	1.0	1.8	1.3	1.40

Based on the above observations the following ANOVA table was obtained:

Source of variation	d.f.	SS	MSS
Between regions	...	4.4655	...
Residual	...	8.892	...

(i) State the mathematical model underlying the one-way analysis of variance together with all associated assumptions. [3]

(ii) Complete the ANOVA table. [1]

(iii) Carry out an analysis of variance to test the hypothesis that the region has no effect on the average time spent watching television. You should write down the null hypothesis, calculate the value of the test-statistic, state its distribution including any parameters, calculate the *p*-value approximately and state your conclusion. [4]

[Total 8]

11 In order to compare the effectiveness of two new vaccines, A and B, for a childhood disease, 11 infants were immunised with vaccine A and 9 infants were immunised with vaccine B. One month after immunisation the concentration of the disease antibodies in the blood of each infant was recorded in appropriate units. The sample mean and variance for each group is given below.

Vaccine A: $n_A = 11, \bar{x}_A = 4.05, s_A^2 = 0.692$

Vaccine B: $n_B = 9, \bar{x}_B = 4.36, s_B^2 = 0.813$

It is assumed that the distributions of the antibody concentration levels after immunisation with vaccine A and vaccine B are $N(\mu_A, \sigma_A^2)$ and $N(\mu_B, \sigma_B^2)$ respectively. You may assume that the samples are independent.

(i) State the distribution of the pivotal quantity $\dfrac{s_A^2 / \sigma_A^2}{s_B^2 / \sigma_B^2}$. [2]

(ii) Calculate an equal-tailed 95% confidence interval for the ratio $\dfrac{\sigma_A^2}{\sigma_B^2}$ using the pivotal quantity in part (i). (You are not required to show the derivation of the interval.) [4]

We now assume that $\sigma_A^2 = \sigma_B^2 = \sigma^2$. Under this assumption, you are given that the distribution of $\dfrac{18S_p^2}{\sigma^2}$ is χ_{18}^2, where S_p^2 is the pooled variance of the two samples and is independent from $\bar{x}_A$ and $\bar{x}_B$.

(iii) Explain why, under the above result, the sampling distribution of

$$\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{S_p\sqrt{\dfrac{1}{11} + \dfrac{1}{9}}}$$

is t_{18}. [4]

(iv) Calculate an equal-tailed 95% confidence interval for $\mu_A - \mu_B$ using the sampling distribution in part (iii). (You are not required to show the derivation of the interval.) [3]

(v) Comment on your results with regard to differences between vaccine A and vaccine B. [2]

[Total 15]

12 An insurer has collected data about the body mass index of 200 males between the age of 18 and 40. The results are shown in the following table.

Body mass index	< 18.5	18.5–25	25–30	>30
Observed frequency	6	114	62	18

A statistician suggests the following model for the distribution of the body mass index with an unknown parameter p.

Body mass index	< 18.5	18.5–25	25–30	>30
Relative frequency	p	$20p$	$10p$	$1-31p$

(i) Estimate the parameter p using the method of maximum likelihood. [4]

(ii) Perform a statistical test to decide whether the suggested distribution is appropriate for the observed data. You should state the null hypothesis for the test and your decision. [6]

To improve the description of the distribution of the body mass index, it is suggested that the marital status of the males in this study is also recorded. The results are shown in the following table.

Marital Status	Body mass index				Total
	< 18.5	18.5–25	25–30	>30	
Single	5	98	43	12	158
Married	1	16	19	6	42
Total	6	114	62	18	200

A life office has considered a sample of 10,000 men aged between 18 and 40 of which 50% are married and the other 50% are single.

(iii) Estimate the proportion of men with a body mass index of more than 30 in this sample, based on the data in the above table. [2]

(iv) Determine whether the body mass index is independent of the marital status or not, using an appropriate statistical test. You should state the null hypothesis for the test, calculate the value of the test statistic and the approximate p-value and state your decision. [8]

[Total 20]

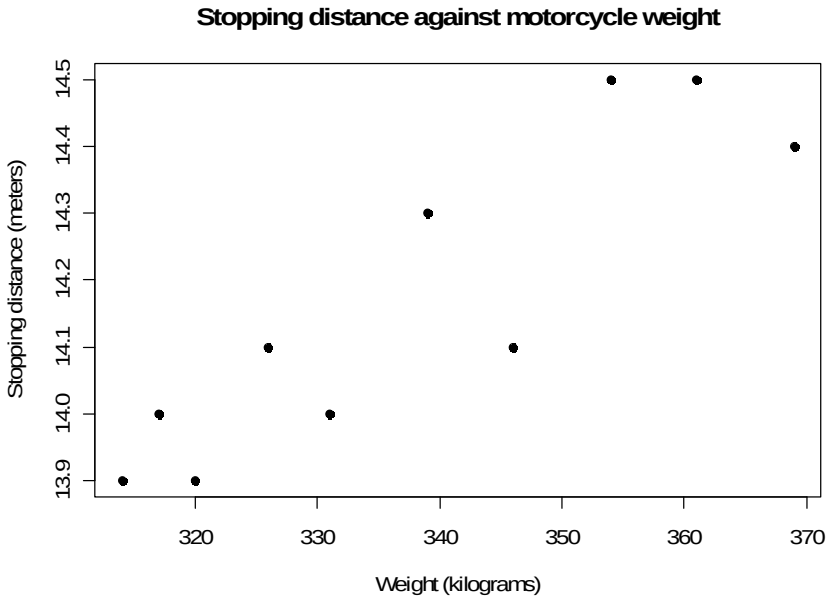
13 The following data give the weight, in kilograms, of a random sample of 10 different models of similar motorcycles and the distance, in metres, required to stop from a speed of 20 miles per hour.

Weight x	314	317	320	326	331	339	346	354	361	369
Distance y	13.9	14.0	13.9	14.1	14.0	14.3	14.1	14.5	14.5	14.4

For these data: $\sum x = 3{,}377$, $\sum x^2 = 1{,}143{,}757$, $\sum y = 141.7$, $\sum y^2 = 2{,}008.39$, $\sum xy = 47{,}888.6$

Also: $S_{xx} = 3{,}344.1$, $S_{yy} = 0.501$, $S_{xy} = 36.51$

A scatter plot of the data is shown below.



(i) (a) Comment briefly on the association between weight and stopping distance, based on the scatter plot.

(b) Calculate the correlation coefficient between the two variables. [2]

(ii) Investigate the hypothesis that there is positive correlation between the weight of the motorcycle and the stopping distance, using Fisher's transformation of the correlation coefficient. You should state clearly the hypotheses of your test and any assumption that you need to make for the test to be valid. [6]

(iii) (a) Fit a linear regression model to these data with stopping distance being the response variable and weight the explanatory variable.

(b) Calculate the coefficient of determination for this model and give its interpretation.

(c) Calculate the expected change in stopping distance for every additional 10 kilograms of motorcycle weight according to the model fitted in part (iii)(a).

[5]

[Total 13]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2012 examinations

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Introduction

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

D C Bowie
Chairman of the Board of Examiners

December 2012

General comments on Subject CT3

For CT3 exams some questions admit alternative solutions or different ways in which the provided answer can be determined. All valid alternative solutions or answers received credit as appropriate. Rounding errors were not penalised, unless excessive rounding led to significantly different answers. In cases where the same error was carried forward to later parts of the answer, candidates were not penalised twice. In questions where comments were required, reasonable comments that were different from those provided in the solutions also received full credit.

Comments on the September 2012 paper

The overall performance was similar to recent sessions, but not as strong as in the last diet (April 2012). A good number of candidates achieved very high scores, although the high end of the mark distribution was negatively affected by the inability of most candidates to tackle certain questions.

As in past sessions, questions corresponding to parts of the syllabus that had not been recently examined were generally poorly answered (e.g. Q4). This highlights the need for candidates to cover the whole syllabus when they revise for the exam and not only rely on themes appearing in past papers. Problems were also recorded in questions where basic algebraic manipulations were required, such as in Q9(ii) and Q12(i).

The comments on individual questions that follow concern specific parts that candidates answered poorly and important frequent errors.

1 $mean = \frac{1}{160}(54 + 2*58 + 3*28) = \frac{254}{160} = 1.5875$ 1

Median = value between 80^{th} and 81^{st} observation = 2 1

Mode = 2 1

Generally well answered. Note that the median is NOT the 80^{th} observation, as some candidates quoted.

2 (i) $Q_1 = \left(\frac{n+2}{4}\right)$th observation counting from below $= 5.5$th observation

$= \frac{335 + 368}{2} = 351.5$ 1

$Q_3 = \left(\frac{n+2}{4}\right)$th observation counting from above $= 5.5$th observation from above

$= \frac{807 + 686}{2} = 746.5$ 1

$IQR = Q_3 - Q_1 = 395$ 1

[With alternative definition:

$Q_1 = \left(\frac{n+1}{4}\right)$th observation counting from below $= 343.25$,

$Q_3 = \left(\frac{n+1}{4}\right)$th observation counting from above $= 776.75$, $IQR = 433.5$.]

(ii) The *length* of the interval containing the *central* half of the claim sizes is 395. 1

The vast majority of candidates calculated the quartiles correctly, although some were confused with their definition. Part (ii) was not very well answered.

3 $E[X] = 1\times0.3 + 2\times0.2 + 3\times0.1 = 1$

$\Rightarrow V[X] = (0-1)^2\times0.4 + (1-1)^2\times0.3 + (2-1)^2\times0.2 + (3-1)^2\times0.1$
$= 0.4 + 0.2 + 0.4 = 1$

(*OR* via $E[X^2] = 2$)

$V[Y] = 4V[X] = 4$

[*OR:* Directly from the distribution of Y, which is $Y = 10, 12, 14, 16$ with probabilities 0.4, 0.3, 0.2, 0.1 respectively.]

No particular problems encountered here. There are a variety of different methods for obtaining the correct answer.

4 Let $f_Z(z)$ be the density of $Z = U + X$.

$$f_Z(z) = \int_u f_U(u)\, f_X(z-u)\, du = \int_0^1 f_X(z-u)\, du$$

$$= \int_{z-1}^{z} f_X(x)\, dx = F_X(z) - F_X(z-1)$$

where we have used the substitution $u = z - x$, and where F_X is the distribution function of X.

This question was very poorly answered. A large number of candidates did not attempt it at all, while many others did not follow any reasonable approach. Note that this is based on standard bookwork, viz. Unit 6, Section 3 in the Core Reading.

5 (i) P(none of class *A*) = P(all 10 of class *B* or *C*) = $(0.8)^{10} = 0.1074$

(ii) (a) Let B = number of class *B*.

Note that $B \sim$ binomial (10, 0.5), so that E(B) = (10)(0.5) = 5

(b) P($B > 5$) = 1 – P($B \le 5$) = 1 – 0.6230 = 0.3770

[0.6230 is from tables; alternatively by evaluation]

This was generally very well answered. A common error in part (ii) (b) was to calculate P(B < or = 5) instead of P(B > 5).

6 (i) Population mean = 8θ

So MME is solution of $\bar{X} = 8\theta \quad \Rightarrow \quad \text{MME} = \dfrac{\bar{X}}{8}$

(ii) $E\left(\dfrac{\bar{X}}{8}\right) = \dfrac{1}{8}E(\bar{X}) = \dfrac{1}{8}(8\theta) = \theta$

Bias = $E\left(\dfrac{\bar{X}}{8}\right) - \theta = 0$ (i.e. MME is unbiased for θ).

(iii) (a) Since MME is unbiased, $MSE\left(\frac{\bar{X}}{8}\right) = var\left(\frac{\bar{X}}{8}\right) = \frac{8\theta^2}{64n} = \frac{\theta^2}{8n}$

(b) MME gets more efficient (MSE gets smaller) as sample size increases.

There was a mix of quality in the answers, especially in parts (ii) and (iii). Attention to detail is required when determining the expected value and variance of functions of sample statistics (here the sample mean).

7 (i) With the larger sample of 100 claims the standard error of the sample mean will be smaller, giving a narrower confidence interval.

(ii) The replacement of the extreme value will give a smaller sample mean, which means that the interval will be shifted to the left.

The variance of the sample will also be smaller, which will again give a narrower interval.

Many candidates recognised the correct effect on the interval, without being able to justify it properly. Note that reasonably accurate wording is important in providing the comments and justification required here.

8 $E[N] = \sum n \, \mathrm{P}(N = n) = 0.3 + 0.6 + 0.6 + 0.4 = 1.9$

$E[N^2] = \sum n^2 \, \mathrm{P}(N = n) = 0.3 + 1.2 + 1.8 + 1.6 = 4.9$

$V[N] = E[N^2] - (E[N])^2 = 1.29$

Also $E[Y] = \exp(\mu + \sigma^2 / 2) = e^{0.55} = 1.73325$

$V[Y] = (E[Y])^2 (\exp(\sigma^2) - 1) = 1.73325^2 * (e^{0.1} - 1) = 0.31595$

Using known results

$E[S] = E[N]\ E[Y] = 1.9 * 1.73325 = 3.293$

$V[S] = E[N]\ V[Y] + V[N]\ (E[Y])^2 = 0.60031 + 3.87536 = 4.476$

Some frequent errors were due to mis-interpretation of the mean and variance of the log-normal distribution.

9 (i) (a) mean $= \frac{\alpha}{\lambda} = 4$ and s.d. $= \sqrt{\frac{\alpha}{\lambda^2}} = \sqrt{8} = 2.8$

(b) As claims are non-negative and the s.d. is quite large relative to the mean, then the distribution will be quite positively skewed.

(ii) $$F(x) = \int_0^x \frac{1}{4} t e^{-\frac{1}{2}t}\, dt$$

$$= -\frac{1}{2}\int_0^x t\, d(e^{-\frac{1}{2}t})$$

$$= -\frac{1}{2}[te^{-\frac{1}{2}t}]_0^x + \frac{1}{2}\int_0^x e^{-\frac{1}{2}t}\, dt$$

$$= -\frac{1}{2}xe^{-\frac{1}{2}x} - [e^{-\frac{1}{2}t}]_0^x$$

$$= 1 - (1 + \frac{1}{2}x)e^{-\frac{1}{2}x}$$

(iii) (a) $F(x) = u$ i.e. $1 - (1 + \frac{1}{2}x)e^{-\frac{1}{2}x} = u$

(b) This equation would have to be solved numerically

(c) Using $u = 0.66$ on the vertical axis, we invert to get $x = 4.5$ on the horizontal axis.

In part (ii) many candidates failed to integrate correctly. A lot of problems were caused by not using the correct limits for the integral. In part (iii) a popular answer was to use "trial-and-error", which is not an appropriate approach here.

10 (i) $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$

with ε_{ij} being i.i.d. $N(0, \sigma^2)$

In particular, it is assumed that the variance is the same in all groups.

(ii)

Source of variation	d.f.	SS	MSS
Between regions	3	4.4655	1.4885
Residual	16	8.892	0.55575

(iii) $H_0 : \tau_i = 0$ for all groups i

$F = 2.6784$ should be from F distribution with 3,16 d.f.

From the tables we know that this gives a p-value of 0.086 (with interpolation).

Reject at 10%, not at 5%, some but very weak evidence against H_0

Mainly well answered. Care is required in calculating the p-values correctly. Also, a number of candidates had difficulties in writing down a sensible form of the ANOVA model in part (i).

11 (i) This is an F distribution with 10, 8 degrees of freedom.

(ii) The interval is given by $\left(\dfrac{S_A^2 / S_B^2}{F_{10,8,0.025}}, \dfrac{S_A^2 / S_B^2}{F_{10,8,0.975}} \right)$

From tables $F_{10,8,0.025} = 4.295$ and $F_{10,8,0.975} = 1 / F_{8,10,0.025} = 1 / 3.855$

giving $\left(\dfrac{0.692 / 0.813}{4.295}, \ (0.692 / 0.813) * 3.855 \right) = (0.198, 3.281)$

(iii) As the two samples are independent we have that

$$V\left(\bar{X}_A - \bar{X}_B\right) = \frac{V\left(X_A\right)}{11} + \frac{V\left(X_B\right)}{9} = \sigma^2 (1/11 + 1/9)$$

Normality of the data then gives that $Z = \dfrac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{\sigma\sqrt{\dfrac{1}{11} + \dfrac{1}{9}}} \sim N(0,1)$

We are also given that $Y = \frac{18S_p^2}{\sigma^2} \sim \chi^2_{18}$ and with Z and Y being independent we can use that $\frac{Z}{\sqrt{Y/18}} \sim t_{18}$ to obtain $\frac{\bar{X}_A - \bar{X}_B - (\mu_A - \mu_B)}{S_p\sqrt{\frac{1}{11}+\frac{1}{9}}} \sim t_{18}$.

(iv) First compute $s_p^2 = \frac{10*0.692 + 8*0.813}{18} = 0.74577 \Rightarrow s_p = 0.864$

Then with $t_{18,0.025} = 2.101$ the interval is given by $(4.05 - 4.36) \pm 2.101 * 0.864\ (1/11 + 1/9)^{1/2}$ i.e. $(-1.126, 0.506)$.

(v) The interval includes the value 0, suggesting that there is no difference in the mean effectiveness of the two vaccines.

Part (iii) was problematic for many candidates. Many candidates struggled to provide a 'proof' that had sufficient rigour. There were errors also in determining the endpoints of the CI in part (ii), often due to using the wrong percentiles of the F distribution.

12 (i) Likelihood function

$$L(p) = p^6 (20p)^{114} (10p)^{62} (1-31p)^{18} = Cp^{182}(1-31p)^{18}$$

$$\log L(p) = \log C + 182\log p + 18\log(1-31p)$$

$$\frac{\partial}{\partial p}\log L(p) = \frac{182}{p} + \frac{18}{1-31p}(-31) = 0$$

$$\frac{182}{p} = \frac{558}{1-31p} \Rightarrow \frac{p}{182} = \frac{1-31p}{558} \Rightarrow \frac{p}{182} + \frac{31p}{558} = \frac{1}{558} \Rightarrow p\left(\frac{1}{182} + \frac{31}{558}\right) = 1/558$$

$$\hat{p} = 0.02935$$

(ii) H_0 : The proposed distribution is the true distribution of the data with non-specified parameter p (it is important to mention that the parameter itself is not part of the null hypothesis)

Under H_0 and using $\hat{p} = 0.02935$ from (i)(a) we obtain the following expected frequencies

Body-Mass-Index	< 18.5	18.5–25	25–30	>30
Expected frequency	5.87	117.4	58.7	18.03

Test-statistic is 0.286915

from a Chi-square distribution with 2 d.f.

The test statistic has a very small value, and there is no evidence against the null.

(iii) $P\left[BMI > 30\right]$

$$= P\left[BMI > 30|single\right]P\left[single\right] + P\left[BMI > 30|married\right]P\left[married\right]$$

$$= \frac{12}{158} * 0.5 + \frac{6}{42} * 0.5 = 0.1094$$

(iv) H_0 : Marital status is independent of BMI

Under H_0 we have:

Marital Status	Body-Mass-Index				Total
	< 18.5	18.5–25	25–30	>30	
Single	4.74	90.06	48.98	14.22	158
Married	1.26	23.94	13.02	3.78	42
Total	6	114	62	18	200

Use χ^2 test.

Test-statistic: $C = \sum_{i=1}^{2}\sum_{j=1}^{4} \frac{(f_{ij} - \frac{f_{i.} * f_{.j}}{n})^2}{\frac{f_{i.} * f_{.j}}{n}} = 8.528399$

C is χ^2-distributed with $(2-1)(4-1)=3$ degrees of freedom.

p-value: $P\left[C > 8.528399\right] < 1 - 0.9616 = 0.0384$

Therefore, we reject H_0 at 5% level, but not at the 1% level.

There were errors in part (i) caused by failure to differentiate correctly. In part (iv) alternative solutions involving merging of adjacent categories were given full redit where correct. However note that merging the first and last column is not correct in this question.

13 (i) (a) The scatter plot suggests a positive linear association between weight and stopping distance.

(b) $r = \dfrac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0.892$

(ii) We want to test H_0: $\rho = 0$ against H_1: $\rho > 0$.

Need to assume that data come from a bivariate normal distribution.

Fisher's (standardised) transformation statistic is given by

$$\frac{\frac{1}{2}\log\left(\frac{1+r}{1-r}\right)}{\sqrt{1/(n-3)}} = \frac{\sqrt{7}}{2}\log\left(\frac{1.892}{0.108}\right) = 3.79$$

and under H_0 this should be a value from the $N(0,1)$ distribution.

This gives P-value $= \Pr(Z \geq 3.79) \approx 0.0001$, so there is very strong evidence against H_0 and we conclude that motorcycle weight and stopping distance are positively correlated.

[Or by considering critical values of N(0,1) distribution.]

(iii) (a) $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}} = \dfrac{36.51}{3344.1} = 0.01092$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 14.17 - 0.01092 * 337.7 = 10.4823$

Fitted line is $\hat{y} = 10.48 + 0.01092x$

(b) $R^2 = \dfrac{S_{xy}^2}{S_{xx}S_{yy}} = \dfrac{36.51^2}{3344.1 * 0.501} = 0.7956$

This gives the proportion of total variation explained by the model.

(Note that R^2 can also be computed as r^2.)

(c) For every additional unit (kilogram) of weight the stopping distance is expected to increase by $\hat{\beta} = 0.01092$ metres. So, for 10 kilograms of weight the distance is expected to increase by 0.109 meters.

Generally adequately answered. Identifying the correct hypotheses in part (ii) was problematic in some cases, while many candidates failed to assume bivariate normality.

END OF EXAMINERS' REPORT

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

25 April 2013 (pm)

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 11 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 The following data represent the number of claims for twenty policyholders made during a year.

0	0	0	0	0	0	0	1	1	1
1	1	1	1	1	1	2	2	2	3

Determine the sample mean, median, mode and standard deviation of these data. [5]

2 Consider a random variable U that has a uniform distribution on $[0,1]$ and let F be the cumulative distribution function of the standard normal distribution.

Show that the random variable $X = F^{-1}(U)$ has a standard normal distribution. [3]

3 A discrete random variable X has a cumulative distribution function (CDF) with the following values:

Observation	10	20	30	40	50
CDF	0.5	0.7	0.85	0.95	1

Calculate the probability that X takes a value:

(i) larger than 10. [1]
(ii) less than 30. [1]
(iii) exactly 40. [1]
(iv) larger than 20 but less than 50. [2]
(v) exactly 20 or exactly 40. [2]
[Total 7]

4 Consider a random sample, $X_1,\ldots,X_n$, from a normal $N(\mu, \sigma^2)$ distribution, with sample mean $\bar{X}$ and sample variance S^2.

(i) Define carefully what it means to say that $X_1,\ldots,X_n$ is a random sample from a normal distribution. [2]

(ii) State what is known about the distributions of $\bar{X}$ and S^2 in this case, including the dependencies between the two statistics. [3]

(iii) Define the t-distribution and explain its relationship with $\bar{X}$ and S^2. [2]
[Total 7]

5 Bank robberies in various countries are assumed to occur according to Poisson processes with rates that vary from year to year. It was reported that the number of robberies in a particular country in a specific year was 123. The number of robberies in a different country in the same year was 111. It can be assumed that each robbery is an independent event and that robberies occur independently in the two countries.

Determine an approximate 90% confidence interval for the difference between the true yearly robbery rates in the two countries. [6]

6 A survey is undertaken to investigate the proportion p of an adult population that support a certain government policy. A random sample of 100 adults is taken and contains 30 who support the policy.

(i) Calculate an approximate 95% confidence interval for p. [2]

(ii) Comment on the validity of the interval obtained in part (i). [1]

A different sample of 1,000 adults is taken and it contains 300 who support the policy.

(iii) Explain how the width of a 95% confidence interval for p in this case will compare to the width of the interval in part (i), without performing any calculations. [1]

[Total 4]

7 A regulator wishes to inspect a sample of an insurer's claims. The insurer estimates that 10% of policies have had one claim in the last year and no policies had more than one claim. All policies are assumed to be independent.

(i) Determine the number of policies that the regulator would expect to examine before finding 5 claims. [1]

On inspecting the sample claims, the regulator finds that actual payments exceeded initial estimates by the following amounts:

£35 £120 £48 £200 £76

(ii) Find the mean and variance of these extra amounts. [3]

It is assumed that these amounts follow a gamma distribution with parameters α and λ.

(iii) Estimate these parameters using the method of moments. [3]

[Total 7]

8 A random sample of 10 independent claim amounts was taken from each of three different regions and an analysis of variance was performed to compare the mean level of claims in these regions. The resulting ANOVA table is given below.

Source	*d.f.*	*SS*	*MSS*
Between regions	2	4,439.7	2,219.9
Residual	27	10,713.5	396.8
Total	29	15,153.2	

(i) Perform the appropriate F test to determine whether there are significant differences between the mean claim amounts for the three regions. You should state clearly the hypotheses of the test. [4]

The three sample means were:

Region	*A*	*B*	*C*
Sample mean	147.47	154.56	125.95

It was of particular interest to compare regions A and B.

(ii) (a) Calculate a 95% confidence interval for the difference between the means for regions A and B.

(b) Comment on your answer in part (ii)(a) given the result of the F test performed in part (i). [4]

[Total 8]

9 A behavioural scientist is observing a troop of monkeys and is investigating whether social status affects the amount of food that an individual takes. The monkeys are divided into two groups of different social rank and the scientist counts the number of bananas each individual takes. Each monkey can take a maximum of 7 bananas.

Social rank	A	B
Number of monkeys	6	11
Total bananas taken	33	37

(i) It is first suggested that the number of bananas taken by each individual of each group follows the same binomial distribution with common parameter p and n=7.

(a) Use the method of moments to estimate the parameter p.

(b) The scientist is unsure whether a common parameter is appropriate and wishes to compare p_A and p_B, the probability that a banana is taken by an individual in groups A and B respectively.

Test the hypothesis that $p_A = p_B$.

[7]

(ii) A statistician suggests an alternative model. The number of bananas taken by an individual still follows a binomial distribution with n=7, but for group A the parameter is 2θ and for group B the parameter is θ, where $\theta < 0.5$.

(a) Show that the log likelihood for θ is given by:

$$33\ln(2\theta) + 9\ln(1-2\theta) + 37\ln(\theta) + 40\ln(1-\theta) + \text{constant}$$

(b) Hence calculate the maximum likelihood estimate of θ.

[6]

(iii) (a) Compare the fit of the two suggested models in parts (i) (with common parameter p) and (ii) by considering the expected number of bananas taken in groups A and B under the two models. You are not required to perform a formal test.

(b) Comment on the above comparison in relation to your answer in part (i)(b).

[4]
[Total 17]

10 The random variable S represents the annual aggregate claims for an insurer from policies covering damage due to windstorms. S is modelled as follows:

$$S = \sum_{i=1}^{M} Y_i$$

where:

M denotes the number of windstorms each year and has a Poisson distribution with mean κ

Y_i denotes the aggregate claims from the ith windstorm and is modelled as

$$Y_i = \sum_{j=1}^{N_i} X_{ij}$$

where:

N_i	denotes the number of claims from the ith windstorm.
$N_1, N_2, \ldots, N_M$	are independent and identically distributed random variables, each with a Poisson distribution with rate λ.
X_{ij}	denotes the amount of the jth claim from the ith windstorm.
X_{ij} , $i = 1, \ldots, M, j = 1, \ldots, N_i$	is a sequence of independent and identically distributed random variables, each with mean μ and variance σ^2.

It is assumed that the random variables M, N_i and X_{ij} are independent of each other.

(i) Derive expressions for the mean and the variance of Y_i in terms of λ, μ and σ. [2]

(ii) Derive expressions for the mean and the variance of S in terms of κ, λ, μ and σ. [3]

Now suppose that X_{ij} has an exponential distribution with mean 1.

(iii) Show that for any positive numbers x and C

$$P(X_{ij} \le x + C | X_{ij} > C) = P(X_{ij} \le x).$$ [3]

Consider the new random variable S_R given as:

$$S_R = \sum_{i=1}^{M} \sum_{j=1}^{N_i} X_{ij}^*$$

where: $X_{ij}^* = \begin{cases} X_{ij} - 2 & \text{if } X_{ij} \geq 2 \\ 0 & \text{otherwise} \end{cases}$.

Let N_i^* be the number of non-zero X_{ij}^* amounts, i.e. the number of claim amounts from the *i*th windstorm that are greater than 2.

Also assume that $N_1^*, N_2^*, \ldots, N_M^*$ are independent and identically distributed Poisson random variables, with parameter λ^*.

Let $\kappa = 4$, $\lambda = 1{,}000$.

(iv) (a) Show that $\lambda^* = 135.3$.

(b) Explain why the distribution of X_{ij}^* is exponential with mean 1.

(c) Calculate the mean and variance of S_R. [7]

[Total15]

11 The table below gives the frequency of a critical illness disease by age group in a certain study. The table also gives the age midpoint (x), the number of people in each group (n), and $y = \log\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right)$, where $\hat{\theta}$ denotes the proportion in an age group with the disease.

		Contracted disease?			
Age group	*x*	*Yes*	*No*	*n*	*y*
20–29	25	1	9	10	−2.19722
30–34	32.5	2	13	15	−1.87180
35–39	37.5	3	9	12	−1.09861
40–44	42.5	5	10	15	−0.69315
45–49	47.5	6	7	13	−0.15415
50–54	52.5	5	3	8	0.51083
55–59	57.5	13	4	17	1.17865
60–69	65	8	2	10	1.38629

$\sum x = 360$; $\sum x^2 = 17437.5$; $\sum y = -2.9392$; $\sum y^2 = 13.615$; $\sum xy = -9.0429$

(i) Calculate an estimate of the probability of having the disease under the assumption that the probability is the same for all age groups. [1]

Consider the hypothesis that there are no differences in the probability of having the disease for the different age groups.

(ii) (a) Construct an 8 × 2 contingency table which includes the expected frequencies under this hypothesis.

(b) Conduct a χ^2 test to investigate the hypothesis.

[6]

Consider the linear regression model $y = \alpha + \beta x + \varepsilon$, where the error terms (ε) are independent and identically distributed following a $N(0, \sigma^2)$ distribution.

(iii) (a) Draw a scatterplot of y against x and comment on the appropriateness of the considered model.

(b) Calculate the fitted regression line of y on x.

(c) Calculate a 99% confidence interval for the slope parameter.

(d) Interpret the result obtained in part (ii) with reference to the confidence interval obtained in part (iii)(c).

[14]
[Total 21]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2013 examinations

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Introduction

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

D C Bowie
Chairman of the Board of Examiners

July 2013

General comments on Subject CT3

For CT3 exams some questions admit alternative solutions or different ways in which the provided answer can be determined. All valid alternative solutions or answers received credit as appropriate. Rounding errors were not penalised, unless excessive rounding led to significantly different answers. In cases where the same error was carried forward to later parts of the answer, candidates were not penalised twice. In questions where comments were required, reasonable comments that were different than those provided in the solutions also received full credit.

Comments on the April 2013 paper

Performance was generally good, but overall not as strong as in the previous examination diet. There was a wide distribution of marks, with well prepared candidates achieving very high scores. On the other hand, less well prepared candidates struggled with questions that did not appear in very similar form in recent examination papers. This is a recurring issue, and candidates are advised to take a wider and more inclusive approach in their preparation for the subject, rather than overly rely on questions appearing in past papers.

The comments on individual questions that follow concern specific parts that candidates answered poorly and important frequent errors.

1 Mean = $(7\times 0+9\times 1+3\times 2+3)/20=(9+6+3)/20=18/20=0.9$

Median = 1 (observation with rank 10.5)

Mode = 1

VAR = $\dfrac{7\times 0.9^2+9\times 0.1^2+3\times 1.1^2+2.1^2}{19}=\dfrac{5.67+0.09+3.63+2.1^2}{19}=\dfrac{13.80}{19}=0.7263$

STD = 0.8522

Well answered. Some working needs to be shown for full marks.

2 For any number x we get

$$P[X\le x]=P\left[F^{-1}(U)\le x\right]=P[U\le F(x)]=F(x)$$

which shows that F is the distribution function of the random variable X, which proves the result.

Very poorly answered. Most candidates did not attempt this question and very few completed it correctly.

3
$$P[X>10]=1-P[X\le 10]=1-F(10)=0.5$$
$$P[X<30]=P[X\le 20]=F(20)=0.7$$
$$P[X=40]=F(40)-F(30)=0.1$$
$$P[20<X<50]=F(40)-F(20)=0.25$$
$$P\left[\{X=20\}\cup\{X=40\}\right]=P[X=20]+P[X=40]$$
$$=\left[F(20)-F(10)\right]+\left[F(40)-F(30)\right]=0.2+0.1=0.3$$

Some problems were encountered here involving understanding and distinguishing the need (or not) for strict inequalities for discrete variables, e.g. P[X<30] = P[X≤20].

4 (i) The random variables $X_1,\ldots,X_n$ are independent and identically distributed with $X_i\sim N(\mu,\sigma^2)$

(ii) $\bar{X}$ and S^2 are independent

$$\bar{X}\sim N(\mu,\sigma^2/n)$$

$$\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$$

(iii) $t_k = N(0,1)/\sqrt{\chi^2_k / k}$ where $N(0,1)$ and χ^2_k are independent

This result can be applied here, and we get $\dfrac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$

Mixed quality in the answers. Some candidates answered part (iii) in the process of answering part (ii) – this did not always show clear understanding, but full marks were given.

5 Under given assumptions $X_1 \sim \text{Poisson}(\lambda_1)$, $X_2 \sim \text{Poisson}(\lambda_2)$

and approximately $X_1 \sim N(\lambda_1, \lambda_1)$, $X_2 \sim N(\lambda_2, \lambda_2)$

giving $X_1 - X_2 \sim N(\lambda_1 - \lambda_2, \lambda_1 + \lambda_2)$, or $\dfrac{X_1 - X_2 - (\lambda_1 - \lambda_2)}{\sqrt{\lambda_1 + \lambda_2}} \sim N(0,1)$

Approximate 90% interval given as

$X_1 - X_2 \pm z_{0.05}\sqrt{\hat{\lambda}_1 + \hat{\lambda}_2} = X_1 - X_2 \pm z_{0.05}\sqrt{X_1 + X_2}$

$= 12 \pm 1.6449 \times (234)^{1/2} = 12 \pm 25.1621$ i.e. $(-13.162, 37.162)$

A common error here involved the normal approximation of the difference of the two variables – especially its variance.

6 (i) Using approximate normality, and with $\hat{p} = 0.3$ we can calculate the interval a

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = (0.21, 0.39)$$

(ii) Sample size is large (or np, or $np(1-p)$), so normal approximation is valid.

(iii) With larger sample size the standard error will be smaller, and therefore the interval will be narrower.

This was straightforward for most candidates. However, the explanation was often not clear or convincing.

7 (i) No of inspected policies ~ Negative binomial(5, 0.1).
Expected no of inspected policies = 5/0.1 =50

(ii) $\sum x = 479, \sum x^2 = 63705$
Mean = 479/5 = 95.8
Variance = $(63705 - 5*95.8^2)/4 = 4454.2$

(iii) $E[X] = \frac{\alpha}{\lambda} = 95.8,\ V[X] = \frac{\alpha}{\lambda^2} = 4454.2$

$$\Rightarrow \lambda = \frac{E[X]}{V[X]} = \frac{95.8}{4454.2} = 0.0215$$

$$\Rightarrow \alpha = \lambda E[X] = 0.0215 * 95.8 = 2.06$$

Generally very well answered with no particular issues.

8 (i) H_0: The means of the claims in the 3 regions are all equal; H_1: means are different for at least one pair.

F = 5.59 on 2 and 27 d.f. From tables the 1% critical point is 5.488.

Therefore, we have (strong) evidence against the null hypothesis, and conclude that there are differences in the means for the 3 regions.

(ii) (a) 95% CI for $\mu_A - \mu_B$ is given by

$$(\bar{y}_A - \bar{y}_B) \pm t_{0.025,27}\hat{\sigma}\sqrt{\frac{1}{10} + \frac{1}{10}} \text{ giving } (147.47 - 154.56) \pm 2.052\sqrt{396.8}\sqrt{\frac{1}{10} + \frac{1}{10}}$$

i.e -7.09 ± 18.28 or (–25.37, 11.19)

(b) The CI comfortably contains zero, suggesting no difference between the true means for regions A and B.

The significant result of the F test clearly comes from region C mean being much lower than the means for regions A and B.

Generally well answered. Some candidates failed to identify the connection between the conclusion of the ANOVA and that of the CI for regions A and B.

9 (i) (a) If X_i is the number of bananas for each monkey then $X_i \sim \text{Bin}(7, p)$

$$E(X_i) = \bar{x} \Rightarrow 7\hat{p} = \frac{33+37}{(6+11)} \Rightarrow \hat{p} = 0.588$$

(b) $$\hat{p}_A = \frac{33}{6*7} = 0.786, \hat{p}_B = \frac{37}{11*7} = 0.481$$

σ^2 = Variance of test statistic = 0.588 * (1 – .588) * (1/42 + 1/77)
= 0.00891

Test statistic = $\frac{\hat{p}_A - \hat{p}_B}{\sigma} = \frac{0.786 - 0.481}{\sqrt{0.00891}} = 3.23$

Test statistic has $N(0,1)$ distribution so p-value is 0.00124

i.e. reject $H_0 : p_A = p_B$

(ii) (a) Let n_i be the number of monkeys in group i and B_i be the total number of bananas taken by group i.

$$L(b;\theta) = (2\theta)^{B_A}(1-2\theta)^{7n_A - B_A}(\theta)^{B_B}(1-\theta)^{7n_B - B_B} \times \text{constant}$$

$$l(b;\theta) = \ln L(b;\theta)$$

$$= 33\ln(2\theta) + (42-33)\ln(1-2\theta) + 37\ln(\theta) + (77-37)\ln(1-\theta) + \text{constant}$$

$$= 33\ln(2\theta) + 9\ln(1-2\theta) + 37\ln(\theta) + 40\ln(1-\theta) + \text{constant}$$

(b) $$\frac{dl}{d\theta} = \frac{66}{2\theta} - \frac{18}{1-2\theta} + \frac{37}{\theta} - \frac{40}{1-\theta} = \frac{70}{\theta} - \frac{18}{1-2\theta} - \frac{40}{1-\theta}$$

Set equal to zero and solve

$$\frac{70(1-2\theta)(1-\theta) - 18\theta(1-\theta) - 40\theta(1-2\theta)}{\theta(1-2\theta)(1-\theta)} = 0$$

$$\Rightarrow 70 - 210\theta + 140\theta^2 - 18\theta + 18\theta^2 - 40\theta + 80\theta^2 = 0$$

$$\Rightarrow 238\theta^2 - 268\theta + 70 = 0$$

$$\Rightarrow \theta = 0.412 \text{ or } 0.714$$

As $\theta < 0.5$, $\hat{\theta} = 0.412$.

(iii) (a) Expected values under 2 models are:

	A	*B*
Model in (i)	42 * 0.588 = 24.7	77 * 0.588 = 45.3
Model in (ii)	42 * 2 * 0.412 = 34.6	77 * 0.412 = 31.7
Observed	33	37

Model in (ii) seems to provide a better fit as expected values are closer to observed.

(b) In part (i)(b) we rejected $p_A = p_B$ which suggests a model with a common value of p would not be appropriate. The comparison above suggests that an improved model can be used.

There were some common errors here, mainly involving part (i)(b) where many candidates failed to identify an appropriate test to perform. There were also basic errors with algebraic and calculus operations.

We note that in part (i)(b)an alternative solution can be given, using a chi-square test with 1 d.f. in a 2x2 table (4 cells). This is exactly equivalent to the test presented here and full credit was given when completed correctly.

10 (i) Y_i has a compound distribution, so

$$E(Y_i) = E(N_i)E(X_{ij}) = \lambda\mu$$

$$V(Y_i) = E(N_i)V(X_{ij}) + V(N_i)E(X_{ij})^2 = \lambda\sigma^2 + \lambda\mu^2$$

(ii) S also has a compound distribution.

$$E(S) = E(M)E(Y_i) = \kappa\lambda\mu$$

$$V(S) = E(M)V(Y_i) + V(M)E(Y_i)^2 = \kappa\lambda(\sigma^2 + \mu^2) + \kappa\lambda^2\mu^2 = \kappa\lambda(\sigma^2 + \mu^2 + \lambda\mu^2)$$

(iii) $$P(X_{ij} \le x + C | X_{ij} > C) = \frac{P(C < X_{ij} \le x + C)}{P(X_{ij} > C)}$$

$$= \frac{\left(1 - e^{-x-C} - 1 + e^{-C}\right)}{e^{-C}} = 1 - e^{-x}$$

$$= P(X_{ij} \le x)$$

(iv) (a) $\lambda^* = 1000 \times P\left(X_{ij} > 2\right) = 1000e^{-2} = 135.3$

(b) From definition of new variable and part (iii) we have that

$$P\left(X_{ij}^* \le x\right) = P(X_{ij} - 2 \le x | X_{ij} > 2) = P(X_{ij} \le x + 2 | X_{ij} > 2) = P(X_{ij} \le x)$$

meaning that X_{ij}^* has the same distribution as X_{ij}, i.e. Exp(1).

(c) $E\left(S_R\right) = \kappa\lambda^*\mu = 4 \times 135.3 \times 1 = 541.2$

$$V\left(S_R\right) = \kappa\lambda^*(\sigma^2 + \mu^2 + \lambda^*\mu^2) = 4 \times 135.3 \times \left(1 + 1 + 135.3\right) = 74306.8$$

Most candidates found this question challenging. Answers to the memoryless property of the exponential distribution (amply discussed in the CR) in part (iii) were often disappointing, and the relevant application in part (iv) was poorly attempted. These shortcomings highlight the issue of being prepared to tackle questions that deviate from the form that appears in past papers.

11 (i) Notation: n_i = number in group i; r_i = number with disease in group i.

$$\hat{\theta} = \frac{\sum r_i}{\sum n_i} = \frac{43}{100} = 0.43$$

(ii) (a) Expected frequencies (in brackets) are given assuming constant probability of disease for all groups, independently of age:

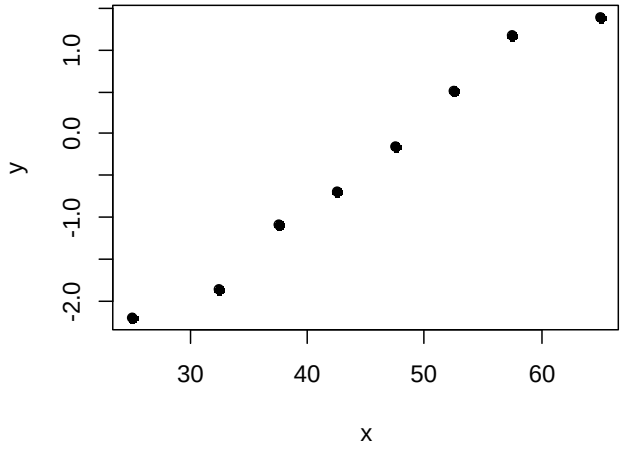
	Disease		
Age group	*Yes*	*No*	*Total*
20–29	1	9	10
	(4.30)	(5.70)	
30–34	2	13	15
	(6.45)	(8.55)	
35–39	3	9	12
	(5.16)	(6.84)	
40–44	5	10	15
	(6.45)	(8.55)	
45–49	6	7	13
	(5.59)	(7.41)	
50–54	5	3	8
	(3.44)	(4.56)	
55–59	13	4	17
	(7.31)	(9.69)	
60–69	8	2	10
	(4.30)	(5.70)	
Total	43	57	100

(b) $\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = \frac{(1-4.3)^2}{4.3} + \ldots + \frac{(2-5.7)^2}{5.7} = 26.6$ on 7 d.f.

From tables, $\chi^2_{7,0.01} = 18.48$

We have (strong) evidence against the hypothesis of no differences in probability of disease among age groups.

(iii) (a) Plot given below. Linear model seems appropriate for middle ages, but perhaps not for younger and older ages.



(b) $S_{xx} = 17437.5 - \frac{360^2}{8} = 1237.5$

$S_{yy} = 13.615 - \frac{(-2.9392)^2}{8} = 12.535$

$S_{xy} = -9.0429 - \frac{(360)(-2.9392)}{8} = 123.22$

Least squares estimates:

$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{123.22}{1237.5} = 0.09957$

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -0.3674 - 0.09957(45) = -4.85$

Fitted line: $\hat{y} = -4.85 + 0.09957x$

(c) $$\hat{\sigma}^2 = \frac{\left(S_{yy} - \frac{S_{xy}{}^2}{S_{xx}}\right)}{n-2} = \frac{\left(12.535 - \frac{123.22^2}{1237.5}\right)}{6} = 0.04430$$

$$se\left(\hat{\beta}\right) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \frac{0.210}{\sqrt{1237.5}} = 0.0060$$

and $t_{6,0.005} = 3.707$

99% CI for $\hat{\beta}$ is given by $0.09957 \pm 3.707(0.0060)$

i.e. 0.09957 ± 0.0222 or $(0.0774,\ 0.1218)$

(d) In (ii) it was found that the probability of having the disease is different for different age groups. In part (iii)(c) it was also found that the probability of disease depends on age, as zero was not included in the interval for the slope parameter.

The quality of the answers was mixed, with some common errors appearing in part (ii) where many candidates failed to produce an appropriate 8×2 table (both "yes" and "no" columns) and perform the correct chi-square test (with 7 d.f.).

It is noted that in part (ii)(b) the chi-square test can alternatively be performed by combining some of the age groups (both columns) to achieve expected frequencies greater than 5, with no change in the conclusion of the test. Although this is not strictly required in this case, full credit was given to candidates that combined groups sensibly and completed the question correctly.

END OF EXAMINERS' REPORT

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

3 October 2013 (pm)

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 10 questions, beginning your answer to each question on a separate sheet.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is NOT required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 The stem and leaf plot below shows 40 observations of an exchange rate.

1.21	9
1.22	4569
1.23	2679
1.24	3467889
1.25	011222345677778
1.26	00346688
1.27	
1.28	1

For these data, $\sum x = 50.000$.

(i) Find the mean, median and mode. [3]

(ii) State, with reasons, which measure of those considered in part (i) you would prefer to use to estimate the central point of the observations. [1]

[Total 4]

2 An insurance company experiences claims at a constant rate of 150 per year.

Find the approximate probability that the company receives more than 90 claims in a period of six months. [4]

3 The random variable X has a distribution with probability density function given by

$$f(x) = \begin{cases} \dfrac{2x}{\theta^2} & ; \quad 0 \le x \le \theta \\ 0 & ; \quad x < 0 \text{ or } x > \theta \end{cases}$$

where θ is the parameter of the distribution.

(i) Derive expressions in terms of θ for the expected value and the variance of X. [3]

Suppose that $X_1, X_2, ..., X_n$ is a random sample, with mean $\bar{X}$, from the distribution of X.

(ii) Show that the estimator $\hat{\theta} = \dfrac{3\bar{X}}{2}$ is an unbiased estimator of θ. [2]

[Total 5]

4 An actuary is considering statistical models for the observed number of claims, X, which occur in a year on a certain class of non-life policies. The actuary only considers policies on which claims do actually arise. Among the considered models is a model for which

$$P(X=x) = -\frac{1}{\log(1-\theta)}\frac{\theta^x}{x}, \quad x=1, 2, 3, \ldots$$

where θ is a parameter such that $0 < \theta < 1$.

Suppose that the actuary has available a random sample $X_1, X_2, \ldots, X_n$ with sample mean $\bar{X}$.

(i) Show that the method of moments estimator (MME), $\tilde{\theta}$, satisfies the equation

$$\bar{X}(1-\tilde{\theta})\log(1-\tilde{\theta}) + \tilde{\theta} = 0. \qquad [3]$$

(ii) (a) Show that the log likelihood of the data is given by

$$l(\theta) \propto -n\log\{-\log(1-\theta)\} + \sum_{i=1}^{n} x_i \log(\theta).$$

(b) Hence verify that the maximum likelihood estimator (MLE) of θ is the same as the MME. [4]

(iii) Suggest two ways in which the MLE of θ can be computed when a particular data set is given. [1]

[Total 8]

5 Consider a random sample consisting of the random variables $X_1, X_2, ..., X_n$ with mean μ and variance σ^2. The variables are independent of each other.

(i) Show that the sample variance, S^2, is an unbiased estimator of the true variance σ^2. [3]

Now consider in addition that the random sample comes from a normal distribution, in which case it is known that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$.

(ii) (a) Derive the variance of S^2 in terms of σ and n.

(b) Comment on the quality of the estimator S^2 with respect to the sample size n. [4]

[Total 7]

6 A researcher obtains samples of 25 items from normally distributed measurements from each of two factories. The sample variances are 2.86 and 9.21 respectively.

(i) Perform a test to determine if the true variances are the same. [3]

(ii) For each factory calculate central 95% confidence intervals for the true variances of the measurements. [4]

(iii) Comment on how your answers in parts (i) and (ii) relate to each other. [1]

[Total 8]

7 A motor insurance company has a portfolio of 100,000 policies. It distinguishes between three groups of policyholders depending on the geographical region in which they live. The probability p of a policyholder submitting at least one claim during a year is given in the following table together with the number, n, of policyholders belonging to each group.. Each policyholder belongs to exactly one group and it is assumed that they do not move from one group to another over time.

Group	*A*	*B*	*C*
p	0.15	0.1	0.05
n (in 1000s)	20	20	60

It is assumed that any individual policyholder submits a claim during any year independently of claims submitted by other policyholders. It is also assumed that whether a policyholder submits any claims in a year is independent of claims in previous years conditional on belonging to a particular group.

(i) Show that the probability that a randomly selected policyholder will submit a claim in a particular year is 0.08. [2]

(ii) Calculate the probability that a randomly selected policyholder will submit a claim in a particular year given that the policyholder is not in group C. [2]

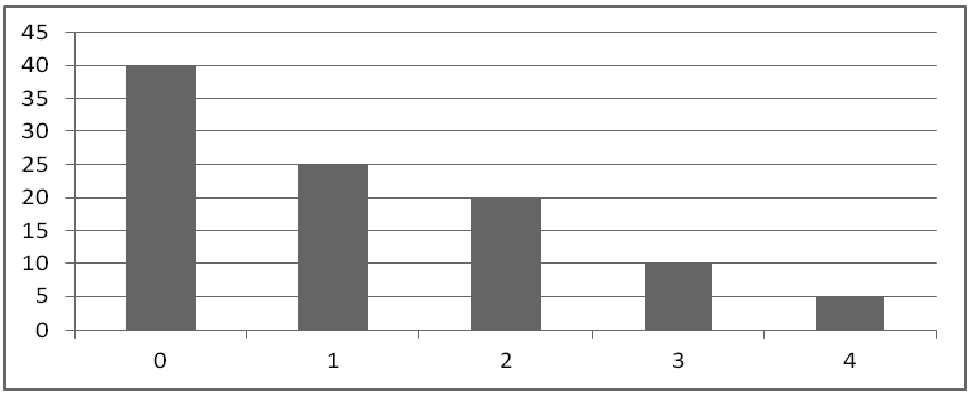
(iii) Calculate the probability for a randomly selected policyholder to belong to group A given that the policyholder submitted a claim last year. [2]

(iv) Calculate the probability that a randomly selected policyholder will submit a claim in a particular year given that the policyholder submitted a claim in the previous year. It is assumed that the insurance company does not know to which group the policyholder belongs. [3]

(v) Calculate the probability that a randomly selected policyholder will submit a claim in two consecutive years. [2]

[Total 11]

8 The following graph shows the number of policyholders who made 0, 1, 2, 3 or 4 claims during the last year in a group of 100 policyholders.



(i) Calculate the sample mean, median, mode and standard deviation of the number of claims per policyholder. [5]

Assume that the number of claims X per policyholder per annum from this group of policyholders has a Poisson distribution with unknown parameter λ.

(ii) Calculate an approximate 95% confidence interval for the unknown parameter λ using the data in the above graph, justifying the validity of your approach. [4]

The following table shows the average claim size for each group of number of claims that a policyholder made during the last year.

Number of claims per policyholder	0	1	2	3	4
Average claim size (£)	--	1000	1100	930	980

Assume that the claim size is independent of the number of claims, and that policyholders make claims independently. Also assume that the size of each claim is normally distributed with estimated standard deviation s = £120.

(iii) Estimate the expected size of a single claim. [2]

(iv) State the type of the distribution of the total amount claimed in the group of the 100 policyholders. [1]

Now assume that the number of claims per policyholder has a Poisson distribution with true parameter $\lambda = 1.15$ and that the true expected value of the size of a single claim is £1,010 and its true standard deviation is £120.

(v) Calculate the expected value of the total amount claimed in the group of the 100 policyholders and its standard deviation. [4]

[Total 16]

9 The random variables Y_A and Y_B describe the number of hours per month that a randomly selected household in Cities A and B, respectively, uses its car. Both cities recently decided to introduce measures to reduce road congestion. To investigate the effect of these measures ten households in each city were randomly selected and asked about the hours per month that they use their car before and after the measures were introduced. The random variables Z_A and Z_B describe the hours of car usage after the measures have been introduced, and $X_A = Y_A - Z_A$ and $X_B = Y_B - Z_B$ denote the reduction in car usage. The following table shows the summary statistics for the ten households in the two cities.

	Sample size n	$\bar{y}$	s_Y	$\bar{z}$	s_Z	s_X
City A	10	33	7.5	28.5	7	2
City B	10	29	8	28	7	2.5

Here, $\bar{y}$ and $\bar{z}$ denote the sample means of Y and Z in the two cities, and s_Y, s_Z and s_X denote the sample standard deviations for Y, Z and X respectively.

You can assume that the random variables Y_A and Y_B are independent and approximately normally distributed

(i) Perform a statistical test at a 5% significance level to test the null hypothesis that expected car usage in City A was the same as expected car usage in City B before the measures were introduced. State all other assumptions that you make and justify them. [6]

An actuary wishes to investigate whether the measures to reduce road congestion have been effective.

(ii) Perform a statistical test at the 5% significance level, where the alternative hypothesis is that car usage in City A has been reduced as a result of the measures. [4]

(iii) Calculate a 95% confidence interval for the expected reduction in car usage for City B. [3]

To investigate further the impact of measures to reduce road congestion, a third city, City C, is included in the study. The following table contains the data for 10 randomly selected households in City C:

	Sample size n	$\bar{y}$	s_Y	$\bar{z}$	s_Z	s_X
City C	10	37	9	33	8	3

Let x_{ij} denote the observed reduction in car usage in city i for household j.

(iv) Confirm that $\sum_{j=1}^{10} x_{Aj} = 45$ and $\sum_{j=1}^{10} x_{Aj}^2 = 238.5$. [2]

You are also given $\sum_{j=1}^{10} x_{Bj} = 10$, $\sum_{j=1}^{10} x_{Cj} = 40$, $\sum_{j=1}^{10} x_{Bj}^2 = 66.25$ and $\sum_{j=1}^{10} x_{Cj}^2 = 241$.

(v) Perform an analysis of variance to test at a 5% significance level the null hypothesis that there is no difference in the mean reduction in car usage between the three cities. [6]

[Total 21]

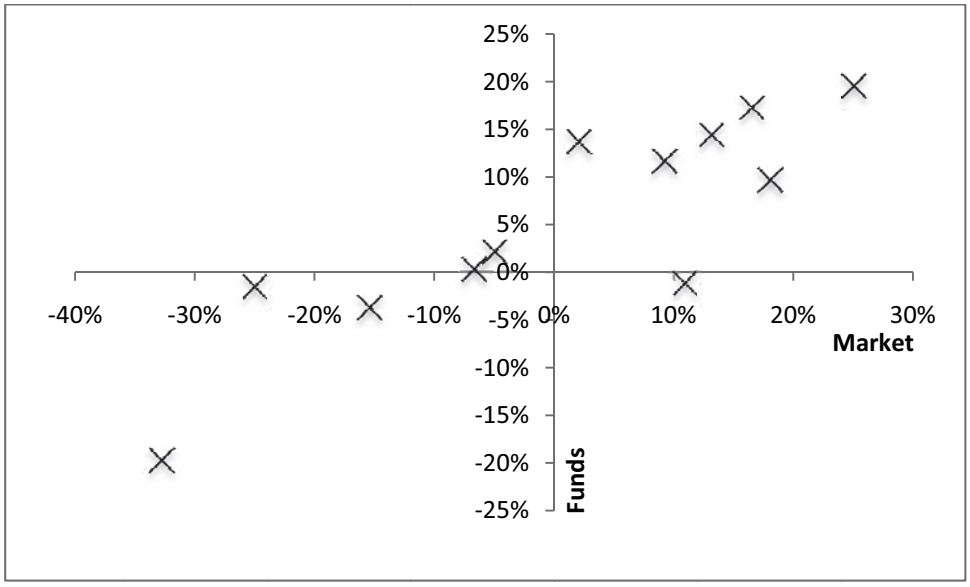
10 An analyst wishes to compare the results from investing in a certain category of hedge funds, f, with those from the stock market, x. She uses an appropriate index for each, which over 12 years each produced the following returns (in percentages to one decimal place).

Year	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Market (x)	–5.0	–15.4	–25.0	16.6	9.2	18.1	13.2	2.0	–32.8	25.0	10.9	–6.7
Funds (f)	2.1	–3.7	–1.6	17.3	11.6	9.7	14.4	13.7	–19.8	19.5	–1.2	0.3

$$\sum x = 0.101, \sum x^2 = 0.3612, \sum f = 0.622, \sum f^2 = 0.1710, \sum xf = 0.1989$$

It is assumed that observations from different years are independent of each other.

Below is a scatter plot of market returns against fund returns for each year.



(i) Comment on the relationship between the two series. [1]

The hedge fund industry often claims that hedge funds have low correlation with the stock market.

(ii) (a) Calculate the correlation coefficient between the two series.

(b) Test whether the correlation coefficient is significantly different from 0. [7]

(iii) Calculate the parameters for a linear regression of the fund index on the market index. [2]

(iv) Calculate a 95% confidence interval for the underlying slope coefficient for the linear model in part (iii). [4]

(v) Comment on your answers to parts (ii)(b) and (iv). [2]

[Total 16]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2013 examinations

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Introduction

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

D C Bowie
Chairman of the Board of Examiners

December 2013

General comments on Subject CT3

Some of the questions in this paper admit alternative solutions from these presented in the marking schedule, or different ways in which the provided answer can be determined. All mathematically correct and valid alternative solutions or answers received credit as appropriate. Rounding errors were not penalised, unless excessive rounding led to significantly different answers. In cases where the same error was carried forward to later parts of the answer, candidates were only penalised once. In questions where comments were required, reasonable comments that were different from those provided in the solutions also received full credit.

Comments on the September 2013 paper

Performance was overall satisfactory, resulting in high pass rate. Candidates that were sufficiently prepared were able to answer all questions, and there was a good proportion of very high marks. As in previous diets, questions that addressed topics that were not recently examined proved to be challenging for less well prepared candidates.

The comments on individual questions that follow cover important frequent errors, and specific parts that were not answered well.

1 (i) Mean $= {}^{50}\!/_{40} = 1.25$

Median = 1.252

Mode = 1.257

(ii) Mean. Distribution is roughly symmetrical with no outliers so no reason to use anything else.

Generally well answered. In part (ii), answers claiming that the median is preferred due to some skewness in the distribution were not penalised.

2 Annual claims ~ Poisson(150) so six-month claims, $X \sim$ Poisson(75)

CLT gives approximate distribution $N(75,75)$

$$P(X > 90) = P(X > 90.5) = P\left(\frac{X-75}{\sqrt{75}} > \frac{90.5-75}{\sqrt{75}}\right) = 1 - \Phi(1.790) = 1 - 0.963 = 0.037$$

[Without continuity correction $1 - \Phi(1.732) = 0.042$]

There were no particular problems with this question. Note that the continuity correction must be applied for full marks.

3 (i) $$E(X) = \int_0^\theta x \frac{2x}{\theta^2}\,dx = \left[\frac{2x^3}{3\theta^2}\right]_0^\theta = \frac{2\theta}{3}$$

$$E(X^2) = \int_0^\theta x^2 \frac{2x}{\theta^2}\,dx = \left[\frac{2x^4}{4\theta^2}\right]_0^\theta = \frac{\theta^2}{2}$$

So, $\text{var}(X) = E(X^2) - [E(X)]^2 = \frac{\theta^2}{2} - \frac{4\theta^2}{9} = \frac{\theta^2}{18}$

(ii) $E(\hat{\theta}) = E\left(\frac{3\bar{X}}{2}\right) = \frac{3}{2}E(\bar{X}) = \frac{3}{2}E(X) = \theta$, so estimator is unbiased.

Generally well answered. A few problems were encountered when deriving the variance.

4 (i) First derive expected value:

$$E(X) = \sum_{x=1}^{\infty} -x \frac{1}{\log(1-\theta)} \frac{\theta^x}{x}$$

$$= \sum_{x=0}^{\infty} \left(-\frac{\theta^x}{\log(1-\theta)} \right) + \frac{1}{\log(1-\theta)}$$

$$= -\frac{\theta}{(1-\theta)\log(1-\theta)}$$

$$\bar{X} = E(X) \Rightarrow \bar{X} = -\frac{\tilde{\theta}}{(1-\theta)\log(1-\tilde{\theta})}$$

$$\Rightarrow \bar{X}\left(1-\tilde{\theta}\right)\log\left(1-\tilde{\theta}\right) + \tilde{\theta} = 0$$

(ii) (a) $$L(\theta) = \frac{\theta^{\sum_i x_i}}{(-\log(1-\theta))^n \prod_i x_i}$$

And $$l(\theta) = -n\log\left(-\log\left(1-\theta\right)\right) + \sum_i x_i \log(\theta) + C$$

(b) MLE given by:

$$\frac{dl(\theta)}{d\theta} = 0 \Rightarrow \frac{n}{\log\left(1-\hat{\theta}\right)\left(1-\hat{\theta}\right)} + \frac{\sum_i x_i}{\hat{\theta}} = 0$$

$$\Rightarrow \bar{X}\left(1-\hat{\theta}\right)\log\left(1-\hat{\theta}\right) + \hat{\theta} = 0$$

(iii) The equation above needs to be solved numerically. Alternatively, the likelihood (or log-likelihood) function can be plotted and the maximum can be identified from the graph.

In part (ii)(a) of the question the log-likelihood was shown as being equal, rather than proportional, to the given expression plus a constant (as given in the solution above). Candidates did not seem to be confused by this, but marking was adjusted in relevant cases.

In general the question was not particularly well answered, mainly due to difficulties in the mathematical operations involved in obtaining the log-likelihood function of non-standard densities. Candidates are advised to practise their calculus skills to deal with such questions.

5 (i) $$E\left(S^2\right)=E\left\{\frac{\left(\sum X_i^2-n\bar{X}^2\right)}{n-1}\right\}=\frac{1}{n-1}\sum E\left(X_i^2\right)-\frac{n}{n-1}E\left(\bar{X}^2\right)$$

and using $E\left(X^2\right)=var\left(X\right)+\left\{E(X)\right\}^2$

$$E\left(S^2\right)=\frac{n}{n-1}\left(\sigma^2+\mu^2\right)-\frac{n}{n-1}\left(\sigma^2/n+\mu^2\right)=\sigma^2$$

(ii) (a) $$\text{var}\left\{\frac{(n-1)S^2}{\sigma^2}\right\}=2(n-1)$$

$$\Rightarrow \text{var}\left(S^2\right)=\frac{2(n-1)\sigma^4}{(n-1)^2}=\frac{2\sigma^4}{(n-1)}$$

(b) Estimator gets better (more accurate) as n increases, as its variance reduces.

(MSE also gets smaller)

This question was generally well answered. There were a few problems with determining the expectation of the sample mean in part (i).

6 (i) H_0 = variances are the same, H_1 = variances are different

$S_2 / S_1 \sim F_{24,24}$

Test statistic = 9.21/2.86 = 3.22.

$F_{24,24,0.995}$ = 0.337 and $F_{24,24,0.005}$ = 2.967

i.e. reject H_0 at 1% significance level.

(ii) Confidence interval is given by $\left(\frac{(n-1)S^2}{X^2_{0.025,n-1}},\frac{(n-1)S^2}{X^2_{0.975,n-1}}\right)$

$X^2_{0.975,24}=12.40, X^2_{0.025,24}=39.36$

Confidence interval 1 = (1.74, 5.54)

Confidence interval 2 = (5.61, 17.83)

(iii) Confidence intervals don't overlap i.e. agree with result in (i) that variances are different.

Generally well answered. In part (i) some candidates worked with the S_1/S_2 ratio, which of course gives the same conclusion. Part (ii) requires the calculation of two CIs, but some candidates attempted to provide a CI for the ratio.

7 (i) $$\begin{aligned} P[\text{claim}] &= P[\text{claim}|A]P[A]+P[\text{claim}|B]P[B]+P[\text{claim}|C]P[C] \\ &= 0.15*0.2+0.1*0.2+0.05*0.6=0.08 \end{aligned}$$

(ii) $$P[\text{claim}\cap A]=P[\text{claim}|A]P[A]=0.15*0.2=0.03$$

$$P[\text{claim}\cap B]=P[\text{claim}|B]P[B]=0.1*0.2=0.02$$

$$P[\text{claim}\cap \bar{C}]=0.03+0.02=0.05$$

$$P[\text{claim}\,|\,\bar{C}]=\frac{P[\text{claim}\cap\bar{C}]}{P[\bar{C}]}=\frac{0.05}{0.4}=0.125$$

(iii) $$P[A|\text{claim last year}]=\frac{P[\text{claim}\cap A]}{P[\text{claim}]}=\frac{0.03}{0.08}=0.375$$

(iv) $$P[B|\text{claim last year}]=\frac{P[\text{claim}\cap B]}{P[\text{claim}]}=\frac{0.02}{0.08}=0.25$$

$$P[C|\text{claim last year}]=1-0.375-0.25=0.375$$

(CLY means "claim last year")

$$\begin{aligned} P[\text{claim}|\text{claim last year}] &= P[\text{claim}|A]P[A|CLY]+\ldots+P[\text{claim}|C]P[C|CLY] \\ &= 0.15*0.375+0.1*0.25+0.05*0.375=0.1 \end{aligned}$$

(v) Let Y be the event that a claim is submitted in two consecutive years

$$P[Y]=P[\text{claim in second year}|\text{claim in first year}]P[\text{claim in first year}]$$
$$=0.1*0.08=0.008$$

alternatively:

$$\begin{aligned} P[Y] &= P[Y|A]P[A]+P[Y|B]P[B]+P[Y|C]P[C] \\ &= 0.15*0.15*0.2+0.1*0.1*0.2+0.05*0.05*0.6=0.008 \end{aligned}$$

This turned out to be the most challenging question for the majority of candidates, with only a small number of "full mark" answers. Many candidates did not attempt parts (iv) and (v) at

all. The question deals with conditional probability concepts, starting with straightforward parts but building up to more complex calculations.

8 (i) Mean = (0*40 + 1*25 +2*20+3*10+4*5)/100 = (25+40+30+20)/100 = 1.15

Median = 1

Mode = 0

VAR = $[(-1.15)^2 * 40 + (-0.15)^2 * 25 + 0.85^2 * 20 + 1.85^2 * 10 + 2.85^2 * 5]/99$
= 1.4419

STD = 1.2

(ii) The estimate for the expectation of X is $\hat{\lambda}=1.15$
$n\hat{\lambda}=115$ is rather large and we can therefore, use a normal approximation to calculate the confidence interval.

$$1.15 \pm 1.96\sqrt{\frac{1.15}{100}} = (0.9398, 1.3602)$$

(iii) Total amount of claims = 25*1*1000 + 20*2*1100 +10*3*930 +5* 4*980
= 116,500

Average claim size = 116500/115 = 1,013.043

(iv) Compound Poisson

(v) Using standard results on compound distributions:

Expected value: $E = 100 * \lambda * 1010 = 115 * 1010 = 116{,}150.00$
Var = $115*120^2 + 115*1010^2 = 118{,}967{,}500$
STD = 10,907.22

Generally well done, but some mixed performance in parts (ii) and (v). Note that calculations refer to a group of 100 policyholders – some candidates failed to take this into account.

9 (i) Sample sizes are small, therefore, we need a t-test. We need to assume that the variances are equal, although the sample standard deviations are different. Since the sample size is small we can argue that equal variances is a reasonable assumption.

Test statistic $t = (\bar{Y}_A - \bar{Y}_B) / \left(S_P \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} \right) \sim t_{n_A + n_B - 2}$ under the null hypothesis that expected car usage is equal in both cities.

$$S_P^2 = \frac{9S_A^2 + 9S_B^2}{18} = \frac{1}{2}\left(7.5^2 + 8^2\right) = 60.125$$

$$t = \frac{33-29}{\sqrt{12.025}} = 1.1535$$

Two sided test, critical values are -2.101 and 2.101 from t_{18}.

The null hypothesis of equal car usage is not rejected.

(ii) $X_A = Y_A - Z_A \sim N\left(\mu_A, \sigma_A^2\right)$

$H_0 : \mu_A \le 0$ and $H_1 : \mu_A > 0$ (also full marks for $H_0 : \mu_A = 0$ vs. $H_1 : \mu_A > 0$)

$$t = \frac{33-28.5}{2/\sqrt{10}} = \frac{4.5}{0.6325} = 7.115$$

Critical values from t_9 at 5%: 1.833

This is clear evidence that the null hypothesis is rejected, and therefore, car usage has been reduced significantly in City A.

(iii) $X_B = Y_B - Z_B \sim N\left(\mu_B, \sigma_B^2\right)$

CI: $29 - 28 \pm \frac{t_{9,0.025}\, 2.5}{\sqrt{10}} = \left[1 - 2.262 \times 0.79,\ 1 + 2.262 \times 0.79\right] = \left[-0.788, 2.788\right]$

(marking: test statistic 1mark, critical value 1mark, correct answer 1mark)

(iv) Let x_{ij} be the difference in city i household j.

$$\sum_{j=1}^{10} x_{Aj} = 10\left(\bar{y} - \bar{z}\right) = 45,$$

$$\sum_{j=1}^{10} x_{Aj}^2 = (10-1)2^2 + 10(33-28.5)^2 = 238.5$$

(v) $$\sum_{i=A,B,C} \sum_{j=1}^{10} x_{ij}^2 = \sum_{j=1}^{10} x_{Aj}^2 + \sum_{j=1}^{10} x_{Bj}^2 + \sum_{j=1}^{10} x_{Cj}^2$$

$$SS_T = \left(238.5 + 66.25 + 241\right) - \frac{\left(45+10+40\right)^2}{10+10+10} = 545.75 - \frac{95^2}{30} = 244.92$$

$$SS_B = \left(\frac{45^2}{10} + \frac{10^2}{10} + \frac{40^2}{10} \right) - \frac{95^2}{30} = \frac{3 \times 3725 - 9025}{30} = 71.67$$

$$SS_R = 244.92 - 71.67 = 173.25$$

$$F = \frac{71.67 / 2}{173.25 / 27} = \frac{35.835}{6.417} = 5.58 \text{ on (2, 27) degrees of freedom}$$

Critical value at 5%: 3.354

The null hypothesis that reduction in car usage is equal in the three cities is rejected.

There were no particular problems with this question. However a number of candidates failed to justify the assumptions in part (i), while some seemed not to understand fully the different test (or CI) requirements in different parts of the question.

10 (i) There is a positive linear relationship between the two.

(ii) (a)
$$S_{xx} = 0.3612 - 0.101^2 / 12 = 0.360$$
$$S_{ff} = 0.1710 - 0.622^2 / 12 = 0.139$$
$$S_{xf} = 0.1989 - 0.101 * 0.622 / 12 = 0.194$$

$$r = \frac{S_{xf}}{\sqrt{S_{xx} S_{ff}}} = \frac{0.194}{\sqrt{0.360 * 0.139}} = 0.867$$

(b) $$\text{Statistic} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.867 * \sqrt{10}}{\sqrt{1-0.867^2}} = 5.50$$

$t_{10,0.995} = 3.169$

So reject H_0 that correlation coefficient = 0 at 1% level (2-sided test)

(iii) $$\hat{\beta} = S_{xf} / S_{xx} = 0.194 / 0.360 = 0.539$$

$$\hat{\alpha} = \bar{f} - \hat{\beta}\bar{x} = \frac{0.622}{12} - 0.539 * \frac{0.101}{12} = 0.0473$$

(iv) $\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{ff} - \frac{S_{xf}^2}{S_{xx}}\right) = \frac{1}{10}\left(0.139 - \frac{0.194^2}{0.360}\right) = 0.0034$

$t_{10,0.975} = 2.228$

$\text{C.I.} = \hat{\beta} \pm t_{10;0.975}\sqrt{\hat{\sigma}^2 / S_{xx}} = 0.539 \pm 2.228\sqrt{0.0034 / 0.36} = (0.321, 0.757)$

(v) C.I. does not contain zero. Consistent with correlation coefficient not equal to zero as the test is actually the same. Both suggest that the hedge industry's claim that correlation is low may not be correct.

Very well answered in general. This is a typical regression/correlation question and the only (minor) problems concerned errors with calculators. Note that part (ii)(b) can also be answered using Fisher's transformation, which results in the same conclusion.

END OF EXAMINERS' REPORT

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

30 April 2014 (pm)

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*
2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*
3. *Mark allocations are shown in brackets.*
4. *Attempt all 10 questions, beginning your answer to each question on a new page.*
5. *Candidates should show calculations where this is appropriate.*

Graph paper is required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 The following sample shows the durations x_i in minutes for 20 journeys from Edinburgh to Glasgow:

51 53 54 55 59 59 60 60 60 69 71 72 74 90 97 104 107 108 115 167

with $\sum_{i=1}^{20} x_i = 1{,}585$ and $\sum_{i=1}^{20} x_i^2 = 142{,}127$.

(i) Calculate the mean and the median of this sample. [2]

(ii) Calculate the standard deviation of this sample. [2]

[Total 4]

2 A set of data has mean 62 and standard deviation 6.

Derive a linear transformation for these data that will result in the new data having mean 50 and standard deviation 12. [3]

3 Sixty per cent of new drivers in a particular country have had additional driving education. During their first year of driving, new drivers who have not had additional driving education have a probability 0.09 of having an accident, while new drivers who have had additional driving education have a probability 0.05 of having an accident.

(a) Calculate the probability that a new driver does not have an accident during their first year of driving.

(b) Calculate the probability that a new driver has had additional driving education, given that the driver had no accidents in the first year. [5]

4 Let X be a random variable with probability density function:

$$f(x) = \begin{cases} \frac{1}{2}e^{x} & ; \quad x \le 0 \\ \frac{1}{2}e^{-x} & ; \quad x > 0 \end{cases}$$

(i) Show that the moment generating function of X is given by:

$$M_X(t) = (1 - t^2)^{-1},$$

for $|t| < 1$. [3]

(ii) Hence find the mean and the variance of X using the moment generating function in part (i). [3]

[Total 6]

5 Consider ten independent random variables $X_1,\ldots,X_{10}$ which are identically distributed with an exponential distribution with expectation 4.

(i) Specify the approximate distribution of $X = \sum_{i=1}^{10} X_i$, including all parameters, using the central limit theorem. [2]

(ii) Calculate the approximate value of the probability $P[X < 40]$ using the result in part (i). [1]

(iii) Calculate the exact probability $P[X < 40]$. [3]

(iv) Comment on the answers in parts (ii) and (iii). [1]

[Total 7]

6 In an opinion poll, a sample of 100 people from a large town were asked which candidate they would vote for in a forthcoming national election with the following results:

Candidate	A	B	C
Supporters	32	47	21

(i) Determine the approximate probability that candidate B will get more than 50% of the vote. [3]

A second opinion poll of 150 people was conducted in a different town with the following results:

Candidate	A	B	C
Supporters	57	56	37

(ii) Use an appropriate test to decide whether the two towns have significantly different voting intentions. [7]

[Total 10]

7 Let X and Y be two continuous random variables.

(i) Prove that $E[E[Y|X]] = E[Y]$. [3]

Suppose the number of claims, N, on a policy follows a Poisson distribution with mean μ, and the amount of the i^{th} claim, X_i, follows a Gamma distribution with parameters α and λ. Let S denote the total value of claims on a policy in a given year.

(ii) Derive the mean of S using the result in part (i). [2]

Suppose $\mu = 0.15$, $\alpha = 100$, and $\lambda = 0.1$.

(iii) Calculate the variance of S. [3]

[Total 8]

8 Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with parameter θ and density function:

$$f(x) = \begin{cases} \dfrac{2x}{\theta^2} & ; \quad 0 \le x \le \theta \\ 0 & ; \quad x < 0 \text{ or } x > \theta \end{cases} .$$

Suppose that $\underline{x} = (x_1, x_2, \ldots, x_n)$ is a realisation of $X_1, X_2, \ldots, X_n$.

(i) (a) Derive the likelihood function $L(\theta; \underline{x})$ and produce a rough sketch of its graph.

(b) Use the graph produced in part (i)(a) to explain why the maximum likelihood estimate of θ is given by $x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}$.

[4]

Let $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$ be the estimator of θ, that is the random variable corresponding to $x_{(n)}$.

(ii) (a) Show that the cumulative distribution function of the estimator $X_{(n)}$ is given by:

$$F_{X_{(n)}}(x) = \left(\frac{x}{\theta}\right)^{2n}$$

for $0 \le x \le \theta$.

(b) Hence, derive the probability density function of the estimator $X_{(n)}$.

(c) Determine the expected value $E(X_{(n)})$ and the variance $V(X_{(n)})$.

(d) Show that the estimator $\dfrac{2n+1}{2n} X_{(n)}$ is an unbiased estimator of θ.

[9]

(iii) (a) Derive the mean square error of the estimator given in part (ii)(d).

(b) Comment on the consistency of this estimator.

[5]
[Total 18]

9 The weekly amount spent on childcare for one child is believed to depend on the age of the child. We denote by X the random variable describing the cost per child for a randomly selected child of age one year, Y being the cost for a three year old child, and Z the cost for a five year old child. It is assumed that X, Y, and Z are normally distributed and that childcare costs are independent between children. Random samples of children of different ages are taken and the weekly childcare costs are recorded during the year 2012. A summary of the data is given in the following table:

Random variable	X	Y	Z
Age of child	1	3	5
Average cost per week per child	200	170	155
Sample standard deviation	30	30	20
Sample size	25	25	25

(i) Calculate the overall average weekly cost of childcare per child for the children in these samples. [1]

(ii) Calculate a 95% confidence interval for the expected childcare cost for a child aged one year. [2]

(iii) Calculate a 95% confidence interval for the expected childcare cost for a child aged five years. [2]

(iv) Calculate a 95% confidence interval for the ratio of the variances of X and Z. [3]

(v) Perform a test at 5% significance level for the null hypothesis that the variances of X and Z are equal based on your answer to part (iv). [1]

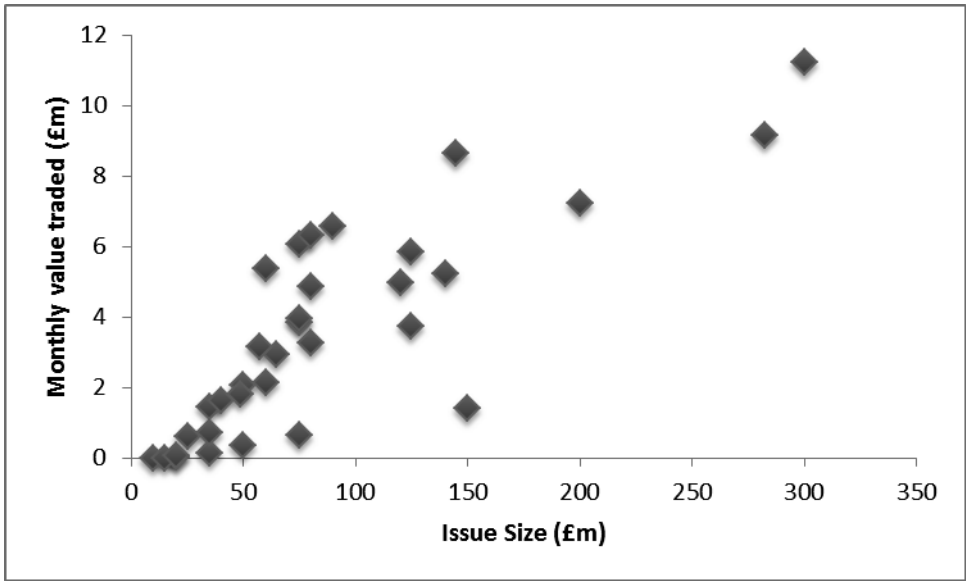
(vi) Calculate an approximate 95% confidence interval for the difference between the average weekly childcare cost per child for children aged one and for children aged five. Justify any assumptions that you make and explain any approximate values you use. [5]

(vii) Perform a test to decide if there is a difference between the expected weekly childcare cost per child spent for children aged one and for children aged five based on your answer to part (vi). [1]

(viii) Perform an analysis of variance to decide if the age of a child has an impact on the weekly amount spent on childcare. [7]

[Total 22]

10 An analyst is instructed to investigate the relationship between the size of a bond issue and its trading volumes (value traded). The data for 33 bonds are plotted in the following chart.



(i) Comment on the relationship between issue size and value traded. [2]

The analyst denotes issue size by s and monthly value traded by v. He calculates the following from the data:

$$\sum s_i = 2{,}843.7,\ \sum s_i^2 = 397{,}499.8, \sum v_i = 115.34, \sum v_i^2 = 689.37, \sum s_i v_i = 15{,}417.75$$

(ii) (a) Determine the correlation coefficient between s and v.

(b) Perform a statistical test to determine if the correlation coefficient is significantly different from 0.

[7]

(iii) Determine the parameters of a linear regression of v on s and state the fitted model equation. [3]

(iv) State the outcome of a statistical test to determine whether the slope parameter in part (iii) differs significantly from zero, justifying your answer. [2]

A colleague suggests that the central part of the data, with issue sizes between £50m and £150m, seem to have a greater spread of value traded and without the bonds in the upper and lower tails the linear relationship would be much weaker.

(v) Comment on the colleague's observation. [3]

[Total 17]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2014 examinations

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Introduction

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

D C Bowie
Chairman of the Board of Examiners

June 2014

General comments on Subject CT3

Some of the questions in this paper admit alternative solutions from these presented in the marking schedule, or different ways in which the provided answer can be determined. All mathematically correct and valid alternative solutions or answers received credit as appropriate. Rounding errors were not penalised, unless excessive rounding led to significantly different answers. In cases where the same error was carried forward to later parts of the answer, candidates were only penalised once. In questions where comments were required, reasonable comments that were different from those provided in the solutions also received full credit.

Comments on the April 2014 paper

The performance was generally good. The pass rate was in line with previous diets. Candidates that were sufficiently prepared were able to answer all questions and the best candidates scored close to full marks. As in previous diets, questions that covered topics that were not recently examined proved to be more challenging for less well prepared candidates.

The comments on individual questions that follow cover important frequent errors, and specific parts that were not answered well.

1 (i) Mean = $\frac{1585}{20} = 79.25$

Median = 70

(ii) Var = $\frac{142{,}127 - \frac{1585^2}{20}}{19} = 869.25$, SD = $\sqrt{869.25} = 29.48$

Well answered.

2 We want to find a and b for $y = a + bx$ such that

$\bar{y} = a + b\bar{x} = 50$ and $s_y = |b| s_x$

These give $b = 2$ and $a = 50 - 124 = -74$
or, $b = -2$ and $a = 50 + 124 = 174$

Generally well answered.

3 Consider the following events:

A: Driver has had additional education
B: Driver has *not* had additional education
C: Driver has *not* had accident in the first year.

(a) $P(\text{C}) = P(\text{C}|\text{A})\ \Pr(\text{A}) + P(\text{C}|\text{B})\ \Pr(\text{B}) = 0.95*0.6 + 0.91*0.4$

$= 0.934$

(b) $P(\text{A}\,|\,\text{C}) = \frac{P(\text{C}\,|\,\text{A})P(\text{A})}{P(\text{C})} = \frac{0.95 \times 0.6}{0.934} = 0.610$

Reasonably well answered. Some candidates did not realise that the answer from part (a) could be used in part (b).

4 (i) $M_X(t) = E(e^{tX}) = \frac{1}{2}\int_{-\infty}^{0} e^{(t+1)x}dx + \frac{1}{2}\int_{0}^{\infty} e^{(t-1)x}dx$

$= \frac{1}{2}\left[\frac{e^{(t+1)x}}{t+1}\right]_{-\infty}^{0} + \frac{1}{2}\left[\frac{e^{(t-1)x}}{t-1}\right]_{0}^{\infty}$

and for $|t| < 1$

$$M_X(t) = \frac{1}{2}\left(\frac{1}{t+1} - \frac{1}{t-1}\right) = \frac{1}{1-t^2}$$

(ii) $$M'_X(t) = \left\{(1-t^2)^{-1}\right\}' = -(1-t^2)^{-2}(-2t) = 2t(1-t^2)^{-2}$$

$$\Rightarrow E(X) = M'_X(0) = 0$$

$$M''_X(t) = \left\{2t(1-t^2)^{-2}\right\}' = 2(1-t^2)^{-2} + 2t(-2)(1-t^2)^{-3}(-2t)$$

$$= 2(1-t^2)^{-2} + 8t^2(1-t^2)^{-3}$$

$$\Rightarrow E(X^2) = M''_X(0) = 2$$

$$V(X) = E(X^2) - E^2(X) = 2$$

(Alternatively, based on a series expansion:

$M_X(t) = 1 + t^2 + t^4 + ... \Rightarrow E(X) = 0$ and $E(X^2) = 2$ and the variance follows.)

Generally well answered. In part (ii) most candidates were familiar with the method, but some showed poor differentiation skills.

5 (i) $X \sim N\left(\mu, \sigma^2\right)$ with $\mu = 10 \times 4 = 40$ and $\sigma^2 = 10 \times (4)^2 = 160$

(ii) X is symmetric so $P[X < 40] = 0.5$

(iii) The exact distribution of X is gamma(10, ¼)

$P[X < 40] = P\left[2 \times \frac{1}{4} \times X < 20\right] = P[Y < 20]$ where Y has a χ^2 distribution with 20 d.f.

$$P[X < 40] = P[Y < 20] = 0.5421$$

(iv) Although the sample size here is small, the CLT gives an answer which is close to the exact probability.

Mostly well answered. There were a few problems with the distributions in part (iii). In part (iv) comments should refer to the use of CLT with small samples for full marks.

6 (i) $$P(B > 0.5) = P\left(\frac{B - 0.47}{\sqrt{0.47 * \frac{1-0.47}{100}}} > \frac{0.5 - 0.47}{\sqrt{0.47 * \frac{1-0.47}{100}}} \right) = 1 - \Phi(0.601) = 0.274$$

(ii) H_0 = Towns have the same voting intentions

Actual	Candidate	A	B	C	Sum
	Town 1	32	47	21	100
	Town 2	57	56	37	150
		89	103	58	250

Expected	Candidate	A	B	C	
	Town 1	35.6	41.2	23.2	100
	Town 2	53.4	61.8	34.8	150
		89	103	58	250

$\frac{(f-e)^2}{e}$			
	0.364	0.817	0.209
	0.243	0.544	0.139

Test statistic = 2.315

Degrees of freedom = (3 – 1)*(2 – 1) = 2. Approximate *p*-value of X_2^2 distribution is between 0.30 and 0.32 (0.314 from interpolation.)

Therefore we fail to reject H_0 that towns have the same voting intentions.

The wording in part (i) of the question was not entirely clear, as the question should in fact refer to the probability that the candidate will get more than 50% of the vote in a different sample of the same size. However, there was very little evidence that candidates were confused by this, and marking was generous in cases where answers seemed to be affected.

In general the question was very well answered. Answers including a continuity correction were also given full marks in part (i).

7 (i) $$E\left[E\left[Y|X\right]\right] = \int E\left[Y|x\right] f_X(x)\,dx = \int \left(\int y f_{y|x}(y)\,dy\right) f_X(x)\,dx$$

$$= \int\int y f(x,y)\,dy\,dx = E\left[Y\right]$$

(ii) $$E\left[S\right] = E\left[E\left[S|N\right]\right] = E\left[E\left[X_1 + \ldots + X_N|N\right]\right] = E\left[N\alpha/\lambda\right] = \mu\alpha/\lambda$$

(iii) As S is compound Poisson,

$$V[S] = \mu E\left[X^2\right]$$

$$= \mu\left(\frac{\alpha}{\lambda^2} + \left(\frac{\alpha}{\lambda}\right)^2\right)$$

$$= 0.15*\left(\frac{100}{0.1^2} + \left(\frac{100}{0.1}\right)^2\right)$$

$$= 0.15*(1{,}010{,}000)$$

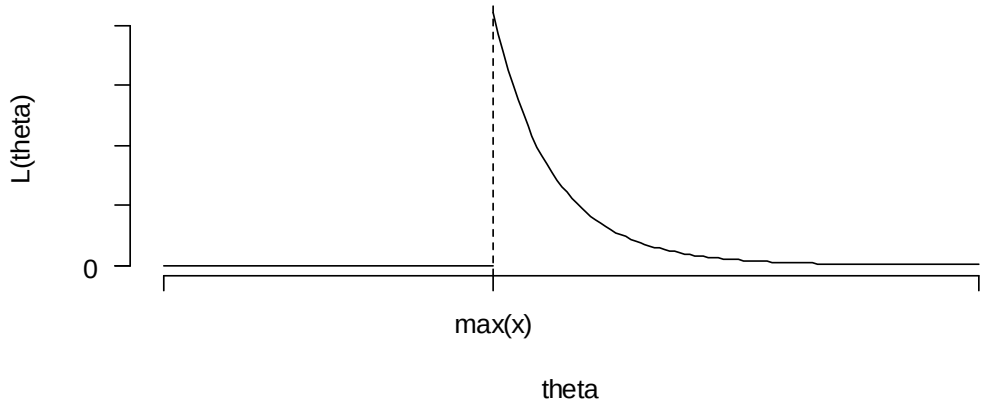
$$= 151{,}500$$

Some mixed performance in part (i), which suggests that some candidates struggled with basic integration skills. The rest of the question was answered well. Use of alternative formulae was given full credit where correct.

8 (i) (a) Likelihood is given as

$$L(\theta;\underline{x}) = \begin{cases} \Pi_{i=1}^{n} f(x_i;\theta) = \frac{2^n x_1 x_2 \cdots x_n}{\theta^{2n}} & \text{if } \theta \geq x_{(n)} = \max\{x_1, x_2, \cdots, x_n\} \\ 0 & \text{if } \theta < x_{(n)}. \end{cases}$$

Its graph is given below:



(b) From the graph, the likelihood is maximised at

$$\theta = x_{(n)} = \max\{x_1, x_2, \ldots, x_n\}.$$

(ii) (a) $F_{X_{(n)}}(x) = P\left\{X_{(n)} < x\right\} = P\left\{X_1 < x, X_2 < x, \ldots, X_n < x\right\}$

$= P\left(X_1 < x\right)P\left(X_2 < x\right)\cdots P(X_n < x)$ as X_i are independent

$= P\left(X_1 < x\right)^n$ since X_i are identically distributed

$$= \left\{\int_0^x \frac{2u}{\theta^2}\,du\right\}^n = \left\{\left[\frac{u^2}{\theta^2}\right]_0^x\right\}^n = \left(\frac{x}{\theta}\right)^{2n}$$

(b) Differentiating we obtain

$$f_{X_{(n)}}(x) = \begin{cases} \dfrac{2nx^{2n-1}}{\theta^{2n}} & \text{if } 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

(c) $$E\left(X_{(n)}\right) = \int_0^\theta \frac{2nx^{2n}}{\theta^{2n}}\,dx = \frac{2n\theta}{2n+1}$$

$$E\left(X_{(n)}^2\right) = \int_0^\theta \frac{2nx^{2n+1}}{\theta^{2n}}\,dx = \frac{n\theta^2}{n+1}$$

$$V\left(X_{(n)}\right) = E\left(X_{(n)}^2\right) - \left\{E\left(X_{(n)}\right)\right\}^2 = \frac{n\theta^2}{n+1} - \left(\frac{2n\theta}{2n+1}\right)^2 = \frac{n\theta^2}{(n+1)(2n+1)^2}$$

(d) $$E\left\{\frac{2n+1}{2n}X_{(n)}\right\} = \frac{2n+1}{2n}E\left(X_{(n)}\right) = \frac{2n+1}{2n}\frac{2n\theta}{2n+1} = \theta$$

(iii) (a) $$MSE\left\{\frac{2n+1}{2n}X_{(n)}\right\} = V\left\{\frac{2n+1}{2n}X_{(n)}\right\}$$

$$= \left(\frac{2n+1}{2n}\right)^2 V\left\{X_{(n)}\right\} = \left(\frac{2n+1}{2n}\right)^2 \frac{n\theta^2}{(n+1)(2n+1)^2} = \frac{\theta^2}{4n(n+1)}$$

(b) We have $MSE \to 0$ as $n \to \infty$, therefore the estimator is consistent.

This question was not well answered, and there were some poor efforts especially in part (i). In many cases, the plotted graph revealed inadequate understanding of the likelihood concept, with some candidates attempting to draw it as a function of x. Note that for full marks the likelihood needs to include the range of the parameter and the graph must indicate

the value of max(x) on the x-axis. In parts (ii) and (iii) some candidates did not cope well with the algebra.

9 (i) Overall average is (200 + 170 + 155)/3 = 175 since sample sizes are all equal.

(ii) $\bar{X} \pm t_{0.025,24} \frac{30}{5} = [200 - 2.064*6, 200 + 2.064*6] = [187.62,\ 212.38]$

(iii) $\bar{Z} \pm t_{0.025,24} \frac{20}{5} = [155 - 2.064*4, 155 + 2.064*4] = [146.74,\ 163.26]$

(iv) $\left[\frac{S_X^2}{S_Z^2} \times \frac{1}{F_{24,24}}, \frac{S_X^2}{S_Z^2} \times F_{24,24}\right] = \left[\frac{2.25}{2.269}, 2.25 \times 2.269\right] = [0.992, 5.105]$

(v) The ratio 1 is contained in the confidence interval, therefore the null hypothesis $\sigma_X^2 = \sigma_Z^2$ cannot be rejected.

(vi) Pooled variance: $s_p^2 = \frac{24 \times \left(30^2 + 20^2\right)}{48} = 650$.

Difference: 200 – 155 = 45

$$\left[45 - t_{0.025,48}\sqrt{650}\sqrt{\frac{2}{25}},\ 45 + t_{0.025,48}\sqrt{650}\sqrt{\frac{2}{25}}\right]$$

$$[45 - 2.01 \times 7.21,\ 45 + 2.01 \times 7.21] = [30.51, 59.49]$$

where we have used the approximation $t_{0.025,48} = 2.01$ (see tables, value for $t_{0.025,50} = 2.009$)

We made the assumption $\sigma_X^2 = \sigma_Z^2$ which is justified by the result in parts (iv) and (v).

(vii) The confidence interval does not contain 0, so there is a difference.

(viii) $SS_R = 24 \times \left(30^2 + 30^2 + 20^2\right) = 52800$

Alternative solution possible

$$SS_B = 25 \times \left[(200 - 175)^2 + (170 - 175)^2 + (155 - 175)^2\right] = 26250$$

$$\frac{SS_B/2}{SS_R/72} = \frac{13125}{733.33} = 17.9$$

This is clearly a very large value compared to $F_{2,72} < F_{2,60} = 4.977$ at the 1% level, so the age of the child has an impact on childcare cost.

Generally well answered. In part (iv) calculation of the ratio of the variance of Z over the variance of X was given full credit. In part (viii) many candidates attempted to calculate the SS values using the original data, rather than the "quick" formulae given in the answer. This was given full marks where appropriate, but was not the best use of time in the exam.

10 (i) There appears to be a positive linear relationship

(ii) (a) $S_{ss} = \sum s_i^2 - \left(\left(\sum s_i\right)^2\right)/n = 397499.8 - \left(2843.7\right)^2/33 = 152450.4$

$$S_{vv} = \sum v_i^2 - \left(\left(\sum v_i\right)^2\right)/n = 689.37 - \left(115.34\right)^2/33 = 286.24$$

$$S_{vs} = \sum v_i s_i - \left(\left(\sum v_i \sum s_i\right)\right)/n = 15417.75 - \left(2843.7 \times 115.34\right)/33 = 5478.6$$

$$r = \frac{S_{vs}}{\sqrt{S_{ss}S_{vv}}} = \frac{5478.6}{\sqrt{152450.4 \times 286.24}} = 0.8294$$

(ii) (b) $H_0 : r = 0, H_1 : r \neq 0$

Test statistic = $r\frac{\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.8294\sqrt{33-2}}{\sqrt{1-0.8294^2}} = 8.266$

At 0.5% level $t_{31} = 2.744$ which << test statistic

So reject H_0.

(iii) $\beta = \frac{S_{vs}}{S_{ss}} = \frac{5478.6}{152450.4} = 0.0359$

$$\alpha = \bar{v} - \beta\bar{s} = \frac{115.34}{33} - 0.0359\frac{2843.7}{33} = 0.398$$

$$v_i = 0.398 + 0.0359 s_i$$

(iv) Testing whether β is significantly different from zero is mathematically the same as testing whether the correlation coefficient is significantly different from zero.

As H_0 was rejected in (ii)(b), we can conclude testing $H_0 : \beta = 0$ would give the same result.

(v) It is true that extreme observations can determine the strength of a linear relationship. However, there are many more bonds in the central part of the data and we would consequently expect a greater range of value traded.

Generally well answered. In part (ii)(b) Fisher's z transformation method was also allowed. In part (v) other possible reasonable comments were given credit.

END OF EXAMINERS' REPORT

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINATION

29 September 2014 (am)

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Time allowed: Three hours

INSTRUCTIONS TO THE CANDIDATE

1. *Enter all the candidate and examination details as requested on the front of your answer booklet.*

2. *You must not start writing your answers in the booklet until instructed to do so by the supervisor.*

3. *Mark allocations are shown in brackets.*

4. *Attempt all 10 questions, beginning your answer to each question on a new page.*

5. *Candidates should show calculations where this is appropriate.*

Graph paper is NOT required for this paper.

AT THE END OF THE EXAMINATION

Hand in BOTH your answer booklet, with any additional sheets firmly attached, and this question paper.

In addition to this paper you should have available the 2002 edition of the Formulae and Tables and your own electronic calculator from the approved list.

1 A sample of marks from an exam has median 49 and interquartile range 19. The marks are rescaled by multiplying by 1.2 and adding 6.

Calculate the new median and interquartile range. [4]

2 Consider an insurer that offers two types of policy: home insurance and car insurance. 70% of all customers have a home insurance policy, and 80% of all customers have a car insurance policy. Every customer has at least one of the two types of policies.

Calculate the probability that a randomly selected customer:

(i) does not have a car insurance policy. [1]
(ii) has car insurance and home insurance. [1]
(iii) has home insurance, given that he has car insurance. [2]
(iv) does not have car insurance, given that he has home insurance. [2]
[Total 6]

3 Let N be a random variable describing the number of withdrawals from a bank branch each day. It is assumed that N is Poisson distributed with mean μ. Let X_i, the random variable describing the amount of each withdrawal, be exponentially distributed with mean $1/\lambda$. All X_i are independent and identically distributed. Let S denote the total amount withdrawn from that branch in a day i.e.

$$S = \sum_{i=1}^{N} X_i$$

with $S = 0$ if $N = 0$.

(i) Derive the moment generating function of S. [4]
(ii) Calculate the mean and variance of S if $\mu = 100$ and $\lambda = 0.025$. [3]
[Total 7]

4 Consider six life policies, each on one of six independent lives. Each of four of the policies has a probability of 2/3 of giving rise to a claim within the next five years, and each of the other two policies has a probability of 1/3 of giving rise to a claim within the next five years. It is assumed that only one claim can arise from each policy.

(i) Calculate the expected number of claims which will arise from the six policies within the next five years. [2]

(ii) Calculate the probability that exactly one claim will arise from the six policies within the next five years. [2]

(iii) Calculate the probability that two policies chosen at random from the six policies will both give rise to claims within the next five years. [4]
[Total 8]

5 Consider two random variables X and Y with $E[X] = 2$, $V[X] = 4$, $E[Y] = -3$, $V[Y] = 1$, and $\text{Cov}[X, Y] = 1.6$.

Calculate:

(a) the expected value of $5X + 20Y$.
(b) the correlation coefficient between X and Y.
(c) the expected value of the product XY.
(d) the variance of $X - Y$.

[4]

6 In a medical study conducted to test the suggestion that daily exercise has the effect of lowering blood pressure, a sample of eight patients with high blood pressure was selected. Their blood pressure was measured initially and then again a month later after they had participated in an exercise programme. The results are shown in the table below:

Patient	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Before	155	152	146	153	146	160	139	148
After	145	147	123	137	141	142	140	138

(i) Explain why a standard two-sample t-test would not be appropriate in this investigation to test the suggestion that daily exercise has the effect of lowering blood pressure. [1]

(ii) Perform a suitable t-test for this medical study. You should clearly state the null and alternative hypotheses. [7]

[Total 8]

7 Consider the following discrete distribution with an unknown parameter p for the distribution of the number of policies with 0, 1, 2, or more than 2 claims per year in a portfolio of n independent policies.

number of claims	*0*	*1*	*2*	*more than 2*
probability	$2p$	p	$0.25p$	$1-3.25p$

We denote by X_0 the number of policies with no claims, by X_1 the number of policies with one claim and by X_2 the number of policies with two claims per year. The random variable $X = X_0 + X_1 + X_2$ is then the number of policies with at most two claims.

(i) Derive an expression for the maximum likelihood estimator $\hat{p}$ of parameter p in terms of X and n. [5]

(ii) Show that the estimator obtained in part (i) is unbiased. [3]

The following frequencies are observed in a portfolio of $n = 200$ policies during the year 2012:

number of claims	*0*	*1*	*2*	*more than 2*
observed frequency	123	58	13	6

A statistician proposes that the parameter p can be estimated by $\tilde{p} = 58/200 = 0.29$ since p is the probability that a randomly chosen policy leads to one claim per year.

(iii) Estimate the parameter p using the estimator derived in part (i). [1]

(iv) Explain why your answer to part (iii) is different from the proposed estimated value of 0.29. [2]

An alternative model is proposed where the probability function has the form

number of claims	*0*	*1*	*2*	*more than 2*
probability	p	$2p$	$0.25p$	$1-3.25p$

(v) Explain how the maximum likelihood estimator suggested in part (i) needs to be adapted to estimate the parameter p in this new model. [1]

(vi) Suggest a suitable test to use to make a decision about which of the two models should be used based on empirical data. [1]

[Total 13]

8 The promoter of a touring dance show wishes to analyse how the price per ticket affects the size of its audiences. She tests two prices, £14 and £16, over 10 shows each which give rise to the following attendances.

Price										
£14	120	115	130	127	124	110	121	129	118	122
£16	111	107	101	115	111	105	99	104	110	98

(i) Calculate the mean and standard deviation of the attendance for each sample. [4]

(ii) Perform a statistical test to determine whether the variances of the attendance are equal under the two prices. [3]

(iii) Perform a t-test to determine whether the mean attendance is the same under the two prices. [4]

(iv) Calculate a 95% confidence interval for the difference between the mean show revenue under each price. [4]

(v) Comment on which price the promoter should choose. [3]

[Total 18]

9 The following data (x) give the acidity (in appropriate units) of three different varieties of grape.

Variety					*Mean*	*Variance*
A	8	7	18	15	12.0	28.7
B	90	74	200	122	121.5	3137.0
C	897	493	812	365	641.8	64284.9

A wine maker wants to test whether there are differences in the mean acidity level of the three varieties and wishes to use analysis of variance (ANOVA) methodology.

(i) Explain why ANOVA should not be used for the data as given in the table above. [2]

A statistician suggests two transformations of the original data:

- the natural logarithm, $y = \ln(x)$,
- and the square root, $z = \sqrt{x}$.

These give the following summary statistics:

	$y = \ln(x)$		$z = \sqrt{x}$	
Variety	*Mean*	*Variance*	*Mean*	*Variance*
A	2.4075	0.2136	3.3975	0.6046
B	4.7250	0.1892	10.8200	5.9242
C	6.4000	0.1800	24.9425	26.4567

The wine maker then decides to use the natural logarithm transformation (y) of the original data.

(ii) Justify the wine maker's choice of data transformation for performing the analysis. [1]

(iii) Perform ANOVA on the transformed data, y, to investigate possible differences in the mean acidity level of the three grape varieties and state your conclusions. [6]

(iv) Calculate 95% confidence intervals for the mean values of each of the three varieties on the original scale, based on the ANOVA performed on the transformed values. [6]

(v) Comment on the intervals obtained in part (iv) in relation to your conclusion in part (iii). [2]

[Total 17]

10 An insurer has collected data on average alcohol consumption (units per week) and cigarette smoking (average number of cigarettes per day) in eight regions in the UK.

Region, i	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*Average*
Alcohol units per week, x_i	15	25	21	29	13	18	21	17	19.875
Cigarettes per day, y_i	4	8	8	10	6	9	7	5	7.125

For these observations we obtain:

$$\sum x_i y_i = 1{,}190; \quad \sum x_i^2 = 3{,}355; \quad \sum y_i^2 = 435$$

(i) Calculate the coefficient of correlation between alcohol consumption and cigarette smoking. [4]

(ii) Calculate a 95% confidence interval for the true correlation coefficient. You may assume that the joint distribution of the two random variables is a bivariate normal distribution. [6]

(iii) Fit a linear regression model to the data, by considering alcohol consumption as the explanatory variable. You should write down the model and estimate the values of the intercept and slope parameters. [3]

(iv) Calculate the coefficient of determination R^2 for the regression model in part (iii). [1]

(v) Give an interpretation of R^2 calculated in part (iv). [1]

[Total 15]

END OF PAPER

INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 2014 examinations

Subject CT3 – Probability and Mathematical Statistics
Core Technical

Introduction

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and using past papers as a revision aid and also those who have previously failed the subject.

The Examiners are charged by Council with examining the published syllabus. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context at the date the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

F Layton
Chairman of the Board of Examiners

November 2014

General comments on Subject CT3

Some of the questions in this paper permit alternative solutions from these presented in this report. All mathematically correct and valid alternative solutions or answers received credit as appropriate. Rounding errors were not penalised, unless excessive rounding led to significantly different answers. In cases where the same error was carried forward to later parts of the answer, candidates were only penalised once. In questions where comments were required, reasonable comments that were different from those provided in the solutions also received full credit where appropriate.

Comments on the September 2014 paper

The performance was generally satisfactory and the pass rate was in line with previous diets. Candidates that were sufficiently prepared were able to answer all questions and the best candidates scored close to full marks.

Questions that required precise mathematical derivations, and questions that covered topics that were not recently examined proved to be more challenging. Some fundamental topics in probability and statistics at this level, such as conditional probability and the likelihood function, were not well addressed by candidates who were inadequately prepared.

The comments on individual questions that follow cover important frequent errors, and specific parts that were not answered well.

1 Let $\{X_1, \ldots, X_n\}$ be the existing marks and $\{Y_1, \ldots, Y_n\}$ denote the transformed marks. Then $Y_i = 1.2X_i + 6$.

Median is $(n+1)/2$ th observation so same transformation applies to median.
New median = 49*1.2 + 6 = 64.8.

As for the median, each transformed quartile, $QY_i = 1.2QX_i + 6$. Then the new interquartile range is, $IQR_Y = QY_3 - QY_1 = 1.2QX_3 + 6 - (1.2QX_1 + 6) = 1.2IQR_X$.

New IQR = 19*1.2 = 22.8. [4]

Generally well answered.

2 Note that each customer has at least one contract, that is, $P[\text{Car} \cup \text{Home}] = 1$.

(i) $P[\text{Car}^C] = 1 - P[\text{Car}] = 0.2 = 20\%$ [1]

(ii) $P[\text{Car} \cap \text{Home}] = P[\text{Car}] + P[\text{Home}] - P[\text{Car} \cup \text{Home}]$ [1]

$= 0.8 + 0.7 - 1 = 0.5$

(iii) $P[\text{Home}|\text{Car}] = \dfrac{P[\text{Car} \cap \text{Home}]}{P[\text{Car}]} = \dfrac{0.5}{0.8} = 0.625$ [2]

(iv) $P[\text{Car}^C \cap \text{Home}] = P(\text{Home}) - P(\text{Car} \cap \text{Home}) = 0.7 - 0.5 = 0.2$

$$P[\text{Car}^C|\text{Home}] = \frac{P[\text{Car}^C \cap \text{Home}]}{P[\text{Home}]} = \frac{0.2}{0.7} = 0.2857$$ [2]

[Total 6]

Reasonably well done, with the exception of part (iv). Note that events are not independent here. Alternative ways to arrive at the correct answer were given full credit.

3 (i) $E\left[e^{tS}|N=n\right]=E\left[\exp\left\{t\left(X_1+X_2+\ldots+X_N\right)\right\}|N=n\right]$

$$=E\left[\exp\left\{t\left(X_1+X_2+\ldots+X_n\right)\right\}\right]=\prod_{i=1}^{n}E\left[\exp\left(tX_i\right)\right]=\left\{M_X\left(t\right)\right\}^n=\left(1-\frac{t}{\lambda}\right)^{-n}$$

$$\therefore M_S\left(t\right)=E\left(e^{tS}\right)$$

$$=E\left[E\left(e^{tS}|N\right)\right]$$

$$=E\left[\left\{M_X\left(t\right)\right\}^N\right]$$

$$=E\left[\exp\left\{N\log M_X\left(t\right)\right\}\right]$$

$$=M_N\left\{-\log\left(1-\frac{t}{\lambda}\right)\right\}$$

$$=\exp\left[\mu\left\{\left(1-\frac{t}{\lambda}\right)^{-1}-1\right\}\right]$$ [4]

(ii) $E\left[X_i\right]=\frac{1}{0.025}=40,\ E\left[X_i^2\right]=\frac{1}{\left(0.025\right)^2}+40^2=3200$

$$E\left[S\right]=\mu E\left[X_i\right]=100*40=4000$$

$$V\left[S\right]=\mu E\left[X_i^2\right]=100*3200=320,000$$

(OR

$$V\left[S\right]=E\left[N\right]V\left[X_i\right]+V\left[N\right]\left[E\left[X\right]\right]^2=100*\frac{1}{\left(0.025\right)^2}+100*40^2=320,000\ )$$

[3]

[Total 7]

Part (i) required careful and precise derivation of the result, and many candidates struggled with it. Answers in questions involving work with MGF expressions have also been problematic in the past – more practice and better understanding is needed.

4 If X is the total number of claims, with X_1 from group 1 (G1, with probability 2/3) and $X_2$from group 2 (G2, with probability 1/3), we have

(i) $X_1 \sim$ Bin(4, 2/3) and $X_2 \sim$ Bin(2, 1/3) .

$$E(X) = E(X_1 + X_2) = E(X_1) + E(X_2)$$

$$= 4(2/3) + 2(1/3) = 10/3 = 3.333 \qquad [2]$$

(ii) $$P(X = 1) = P(X_1 = 1, X_2 = 0) + P(X_1 = 0, X_2 = 1)$$

$$= \binom{4}{1}(2/3)(1/3)^3 \times \binom{2}{0}(1/3)^0(2/3)^2 + \binom{4}{0}(2/3)^0(1/3)^4 \times \binom{2}{1}(1/3)^1(2/3)^1$$

$$= 4/81 = 0.0494 \qquad [2]$$

(iii) P(two randomly selected policies giving claims) =

P(both give claims | both from G1) * P(both from G1)
+P(both give claims | both from G2) * P(both from G2)
+ 2*P(both give claims | one from G1, one from G2) * P(one from G1, one from G2)

$$= \left(\frac{2}{3}\right)^2 \frac{4}{6}\frac{3}{5} + \left(\frac{1}{3}\right)^2 \frac{2}{6}\frac{1}{5} + 2 \times \left(\frac{2}{3}\right)\left(\frac{1}{3}\right)\frac{4}{6}\frac{2}{5} = \frac{41}{135} = 0.3037 \qquad [4]$$

[Total 8]

Mixed performance. Parts (i) and (ii) were answered well, but there were many inadequate attempts in part (iii). In many cases candidates failed to see the different combinations resulting in the required event, while there were also problems in calculating the correct probability for each combination.

5 (a) $$E\left[5X + 20Y\right] = 5*2 + 20*(-3) = 10 - 60 = -50$$

(b) $$Corr\left(X, Y\right) = \frac{1.6}{2} = 0.8$$

(c) $$E\left[XY\right] = Cov\left(X, Y\right) + E\left[X\right]E\left[Y\right] = 1.6 + 2*(-3) = -4.4$$

(d) $$V\left(X - Y\right) = V\left(X\right) + V\left(Y\right) - 2Cov\left(X, Y\right) = 4 + 1 - 3.2 = 1.8$$

[4]

Generally very well answered. There were only a few problems with using the correct expression for the variance (taking into account the covariance).

6 (i) The two samples are from the same patients, so they are clearly not independent. [1]

(ii) First calculate differences d = measurement before – measurement after :

d : 10 5 23 16 5 18 −1 10

For these we have $\sum d = 86, \sum d^2 = 1,360$

giving $\bar{d} = 86/8 = 10.75$ and $sd(d) = \sqrt{(1360 - 86^2/8)/7} = 7.8876$

H_0: mean difference = 0 v H_1: mean difference > 0

$$t = \frac{\bar{d}}{sd(d)/\sqrt{n}} = \frac{10.75}{7.8876/\sqrt{8}} = 3.855$$

From tables, $t_7(0.005) = 3.499$ and $t_7(0.001) = 4.785$

Therefore, we have strong evidence against H_0 (P-value < 0.5%), and conclude that daily exercise has the effect of lowering blood pressure. [7]

[Total 8]

Mixed performance in part (ii). The question clearly indicates that a standard two-sample t test is not appropriate here, and candidates should recognise the need for a paired test. In some cases, although the correct test was identified, its application was wrong.

7 We denote by X_0 the number of policies with no claims, by X_1 the number of policies with one claim and by X_2 the number of policies with two claims per year. Let $X = X_0 + X_1 + X_2$

(i) Likelihood function

$$L(p) \propto (2p)^{X_0} \, p^{X_1} (0.25p)^{X_2} (1 - 3.25p)^{n-X}$$

Log-likelihood

$$l(p) = X_0 \log(2p) + X_1 \log(p) + X_2 \log(0.25p) + (n - X)\log(1 - 3.25p) + \text{constant}$$

$$\frac{dl}{dp} = \frac{X_0}{p} + \frac{X_1}{p} + \frac{X_2}{p} - \frac{3.25(n-X)}{1-3.25p} = \frac{X}{p} - \frac{3.25(n-X)}{1-3.25p}$$

$\frac{dl}{dp} = 0$ gives $X\left(1-3.25p\right)-3.25\left(n-X\right)p = 0$

$X - 3.25Xp - 3.25np + 3.25Xp = X - 3.25np = 0$

$\hat{p} = \frac{X}{3.25n}$

[Alternative solution:

Set $\theta = 3.25p$ to be the probability of at most two claims.

$L(\theta) \propto \theta^X (1-\theta)^{n-X}$ and $l(\theta) = X\,ln(\theta) + (n-X)\ln(1-\theta) + \text{constant}$

$\frac{dl}{d\theta} = \frac{X}{\theta} - \frac{(n-X)}{1-\theta}$ and setting equal to zero: $\hat{\theta} = \frac{X}{n}$.

Using the invariance property of the MLE we obtain:

$\hat{\theta} = 3.25\hat{p} \Rightarrow \hat{p} = \frac{X}{3.25n}$] [5]

(ii) $E[\hat{p}] = \frac{1}{3.25n}E[X]$, X has Binomial dist. with parameters n and $2p + p + 0.25p$

$E[X] = n(2p + p + 0.25p) = 3.25pn$

and therefore $E[\hat{p}] = p$ [3]

(iii) $X = 194$, $\hat{p} = \frac{194}{3.25 \times 200} = 0.2985$ [1]

(iv) The MLE in part (iii) takes the structure of the entire probability function into account while the estimator 58/200 only considers the number of policies with one claim. [2]

(v) No change required, since the MLE $\hat{p}$ turns out to dependent only on the total number of policies with less than three claims. [1]

(vi) χ^2-test [1]

[Total 13]

The later parts of the question were well answered. However there was a considerable number of poor answers in parts (i) and (ii). Part (i) particularly, deals with the likelihood concept which is fundamental in statistics. The setting does not refer explicitly to a usual

distribution, but involves a standard model, and candidates at this level need to make sure that they can work with the likelihood function in a variety of standard models.

8 (i) Let $\{X_1,\ldots,X_{10}\}$ denote the sample at £14 and $\{Y_1,\ldots,Y_{10}\}$ the sample at £16.

$\Sigma x_i = 1216, \Sigma x_i^2 = 148220$

$\Rightarrow \bar{x} = \frac{1216}{10} = 121.6, s_x = \sqrt{\frac{148220 - 121.6^2 * 10}{9}} = 6.275$

$\Sigma y_i = 1061, \Sigma x_i^2 = 112863$

$\Rightarrow \bar{y} = \frac{1061}{10} = 106.1, s_y = \sqrt{\frac{112863 - 106.1^2 * 10}{9}} = 5.685$ [4]

(ii) $H_0 : \sigma_x^2 = \sigma_y^2, H_1 : \sigma_x^2 \neq \sigma_y^2$

Under H_0 $s_x^2 / s_y^2 \sim F_{9,9}$

$s_x^2 / s_y^2 = 6.275^2 / 5.685^2 = 1.22$

$F_{9,9;0.975} = \frac{1}{4.026} = 0.25, F_{9,9;0.025} = 4.026$ so we fail to reject H_0. [3]

(iii) Given (ii) we can assume that standard deviations are equal.

$s_P^2 = \frac{1}{10+10-2}\left(9 * 6.275^2 + 9 * 5.685^2\right) = 35.847$

test statistic $= \frac{121.6 - 106.1}{s_P\sqrt{\frac{2}{10}}} = \frac{15.5}{5.987\sqrt{\frac{1}{5}}} = 5.789$

test statistic $\sim t_{10+10-2} = t_{18} = 2.101$ at 2.5%.

So reject H_0: there is a significant difference between the means at 5% significance level. [4]

(iv) Difference in means = 14*121.6-16*106.1=4.8

$$s_P^2 = \frac{1}{10+10-2}\left(9*14^2*6.275^2 + 9*16^2*5.685^2\right) = 7995.7$$

Using t_{18} as before the confidence interval is

$$4.8 \pm 2.101*\sqrt{7995.7\left(\frac{1}{10}+\frac{1}{10}\right)} = (-79.22, 88.82) \qquad [4]$$

(v) There is a significant lower attendance with the higher price but, as the confidence interval contains zero, no significant difference in revenues. Financially it doesn't matter which price the promoter chooses, but the lower price would get more people to see the show. [3]

[Total 18]

Parts (i) – (iii) were well answered. In part (iv) some candidates did not realise that the required CI referred to revenue. In the same part, there were also many errors in calculating the common variance correctly. In part (v) other sensible comments were also given credit as appropriate.

9 (i) The original values vary in scale among the 3 varieties, resulting in large differences in the variances of the 3 groups. This violates the ANOVA requirement that the error variance should not depend on the treatment concerned. [2]

(ii) The logarithm transformation gives very similar variances for the 3 groups, as opposed to the square root which still produces large differences. [1]

(iii) First calculate relevant sums:

$$\sum y_A = 2.4075\times 4 = 9.63,\ \ \sum y_A^2 = 3\times 0.2136 + 4\times 2.4075^2 = 23.825$$
$$\sum y_B = 4.725\times 4 = 18.9,\ \ \sum y_B^2 = 3\times 0.1892 + 4\times 4.725^2 = 89.870$$
$$\sum y_C = 6.4\times 4 = 25.6,\ \ \sum y_C^2 = 3\times 0.18 + 4\times 6.4^2 = 164.38$$

SST = 23.825 + 89.87 + 164.38 – $(9.63+18.9+25.6)^2$ / 12 = 33.9036
SSB = $(9.63^2+18.9^2+25.6^2)/4$ – $(9.63+18.9+25.6)^2$ / 12 = 32.1553
SSR = SST – SSB = 1.7483

ANOVA table:

Source of variation	*df*	*SS*	*MSS*
Between groups	2	32.1553	16.0777
Residual	9	1.7483	0.1943
Total	11	33.9036	

$$F = \frac{16.0777}{0.1943} = 82.75 \text{ on } 2, 9 \text{ df}$$

$F_{2,9}(1\%) = 8.022$, so P-value << 0.01

There is overwhelming evidence against the null hypothesis. We conclude that there are differences in the mean level of acidity of the three grape varieties. [6]

(iv) The CIs are given by

$$\bar{y}_i \pm t_{9,0.975} \times \hat{\sigma} / \sqrt{n_i} \text{ with } t_{9,0.975} = 2.262 \text{ and } \hat{\sigma} = \sqrt{MSS_R} = 0.44079$$

For A: $2.4075 \pm 2.262 \times 0.44079 / 2$ i.e. (1.909, 2.906)
and on the original scale: $\left(e^{1.909}, e^{2.906}\right)$ = (6.75, 18.28)

For B: $4.725 \pm 2.262 \times 0.44079 / 2$ i.e. (4.226, 5.224)
and on the original scale (68.44, 185.68)

For C: $6.4 \pm 2.262 \times 0.44079 / 2$ i.e. (5.901, 6.899)
and on the original scale (365.40, 990.28) [6]

(v) The CIs do not overlap. This agrees with the ANOVA conclusion, and in addition shows differences between all 3 pairs of means. [2]

[Total 17]

There were no problems with the ANOVA part of this question. However, the explanation in part (i) was often unclear. In part (iv) some candidates failed to transform back to the original scale.

10 (i) $S_{xx} = 3,355 - 8*19.875^2 = 194.875,$

$S_{yy} = 435 - 8*7.125^2 = 28.875,$

$S_{xy} = 1190 - 8*19.875*7.125 = 57.125$

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\ S_{yy}}} = 0.76153$$ [4]

(ii) $W = \frac{1}{2}\log\frac{1+r}{1-r}$ is normally distributed with mean $\frac{1}{2}\log\frac{1+\rho}{1-\rho}$ and standard deviation $1/\sqrt{n-3}$

Confidence interval for the mean of W: $W \pm 1.96 / \sqrt{(n-3)}$

Using r from part (i), the estimated value of W is 0.999848.

This gives a confidence interval of

$$0.999848 \pm \frac{1.96}{\sqrt{5}} = [0.123309176,\ 1.87638647] \text{ for } W.$$

Since $r = \dfrac{e^{2W}-1}{e^{2W}+1}$ we obtain the C.I. for the true correlation ρ

$$\left[\frac{e^{2x0.123309176}-1}{e^{2x0.123309176}+1}, \frac{e^{2x1.87638647}-1}{e^{2x1.87638647}+1}\right] = [0.122688\ ,0.95417] \qquad [6]$$

(iii) $Y_i = a + bX_i + \varepsilon_i$

$$\hat{b} = S_{xy} / S_{xx} = \frac{57.125}{194.875} = 0.293137$$

$$\hat{a} = \frac{1}{8}\left(\sum y_i - \hat{b}\sum x_i\right) = 1.29891 \qquad [3]$$

(iv) $R^2 = 0.76153^2 = 0.58$ [1]

(v) About 58% of the total variability of the response "cigarettes per day" is statistically explained by alcohol consumption. [1]

[Total 15]

Generally well answered with some problems in part (ii), which involves the more demanding (and less frequently examined) CI for the correlation coefficient.

END OF EXAMINERS' REPORT