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:

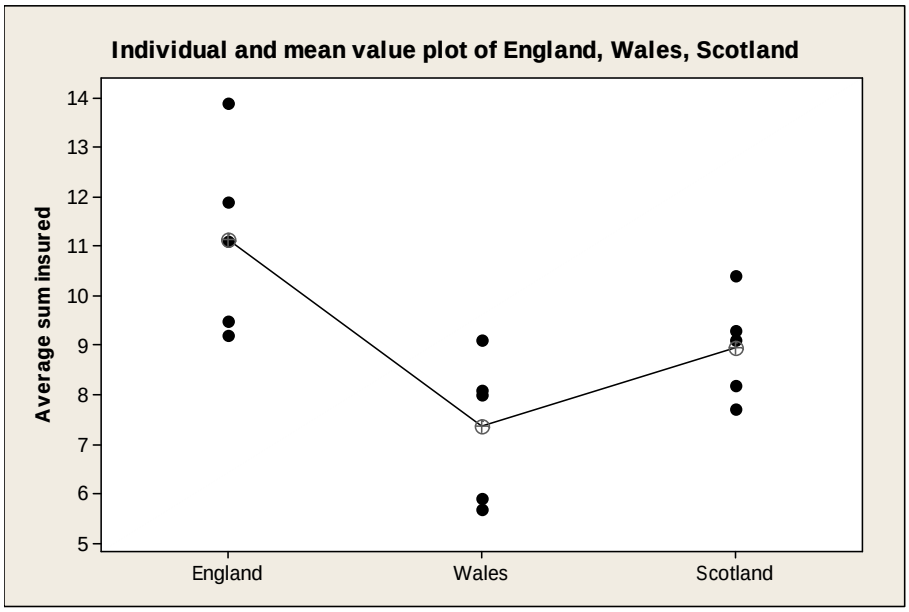
	England					*Wales*					*Scotland*				
y	11.9	11.1	9.5	9.2	13.9	5.9	9.1	8.0	5.7	8.1	9.3	9.1	7.7	8.2	10.4

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]

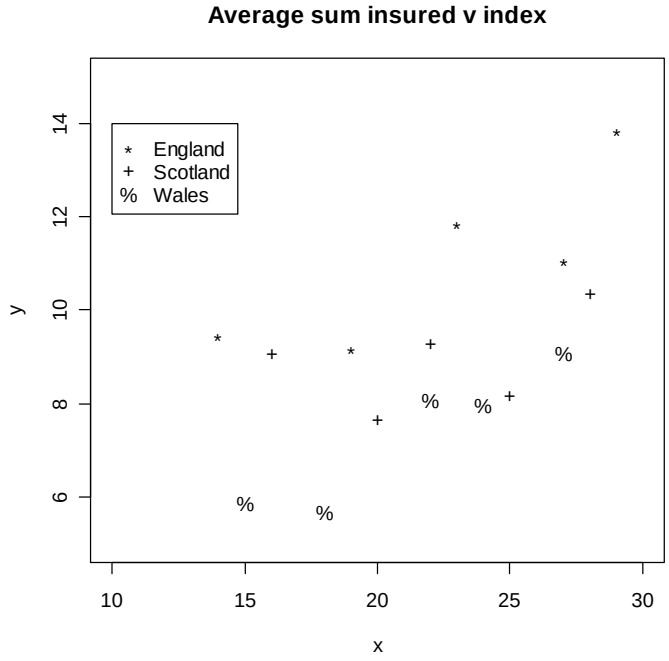
The lecturer of the module decides to provide further information. It has been suggested that the value of a UK index of the town's prosperity (X) might also be a useful explanatory variable (in addition to the country in which the town is situated).

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