EXAMINATIONS

1 April 2003 (pm)

Subject 101 — Statistical Modelling

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 14 questions, beginning your answer to each question on a separate sheet.*

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 Actuarial Tables and your own electronic calculator.

© Faculty of Actuaries

© Institute of Actuaries

1 Sickness and absence records were kept on 30 employees in a company over a 91-day period. These data are tabulated below:

Number of employees absent	0	1	2	3	4	5
Number of days	44	19	10	8	7	3

Calculate the sample mean and standard deviation of the number of employees absent per day. [3]

2 A set of claim amounts (£) is given below:

192	136	253	138	87
112	221	176	336	203
159	55	308	165	254

Present these data graphically using a boxplot. [3]

3 The probability that a car accident is due to faulty brakes is 0.02, the probability that a car accident is correctly attributed to faulty brakes is 0.95, and the probability that a car accident is incorrectly attributed to faulty brakes is 0.01.

Calculate the probability that a car accident which is attributed to faulty brakes was due to faulty brakes. [3]

4 The probability that information held on an individual record on a company's database is incorrect (e.g. information is out-of-date) is 0.13.

Calculate the probability that in a random sample of 200 records at most 20 contain incorrect information. [3]

5 Consider a random sample of size n from a normal distribution $N(\mu,\sigma^2)$ and let S^2 denote the sample variance.

(i) State the sampling distribution for $\dfrac{(n-1)S^2}{\sigma^2}$, and specify an approximate sampling distribution for this expression when n is large. [2]

(ii) For $n = 101$ calculate an approximate value for the probability that S^2 exceeds σ^2 by more than a factor of 10%, i.e. $P(S^2 > 1.1\ \sigma^2)$. [1]

[Total 3]

6 Calculate the maximum possible width of a symmetrical two-sided 95% confidence interval for the proportion of a population who possess a particular characteristic, based on the corresponding information in a random sample of size 1600 from the population. [3]

7 A group of 500 insurance policies gave rise to a total of 83 claims during the last year. Assuming a Poisson model for the occurrence of claims, calculate an approximate 95% confidence interval for λ, the claim rate per policy per year. [3]

8 The ratio of the standard deviation to the mean of a random variable is called the *coefficient of variation*.

For each of the following distributions, decide whether increasing the mean of the random variable increases, decreases, or has no effect on the value of the coefficient of variation:

(a) Poisson with mean λ
(b) exponential with mean μ
(c) chi-square with n degrees of freedom [6]

9 An insurance company specifies in its advertising literature that 75% of all small claims (less than £200) on household policies are fully settled within one month. As part of an internal audit a random sample of 200 small claims is examined.

(i) It is found that 146 of these small claims were fully settled within one month. Obtain an approximate one-sided 99% confidence interval for the true percentage of all small claims which are fully settled within one month. [3]

(ii) The mean and standard deviation of the individual claim amounts were calculated as $\bar{x} = £112.41$ and $s = £51.62$ respectively. Obtain an approximate two-sided 99% confidence interval for the population mean of all small claim amounts. [2]

[Total 5]

10 In an investigation into the comparison of claim amounts between three different regions, a random sample of 10 independent claim amounts was taken from each region and an analysis of variance was performed. The resulting ANOVA table is given below with some entries omitted:

Source of variation	*d.f.*	*SS*	*MSS*
Between regions	2	4439.7	2219.9
Residual	*	*	*
Total	29	15153.2	

(i) Calculate the missing entries in this table and perform the appropriate *F*-test to determine whether there are significant differences between the mean claim amounts for the three regions. [3]

(ii) The three sample means were given by:

Region	*A*	*B*	*C*
Sample mean	147.47	154.56	125.95

Given that it was of particular interest to compare regions A and B, calculate a 95% confidence interval for the difference in their means and comment on your answer in the light of the *F*-test performed in part (i). [4]

[Total 7]

11 Let $S = X_1 + X_2 + ... + X_N$ (and $S = 0$ if $N = 0$) where the X_i's are independent and identically distributed as exponential random variables with mean μ and are also independent of N, which has a Poisson distribution with mean λ.

(i) Prove, from first principles (that is without quoting any general results for compound distributions), that the moment generating function of S, $M_S(t)$, is given by $M_S(t) = \exp\left[\lambda\left\{\left(1-\mu t\right)^{-1} - 1\right\}\right]$. [4]

(ii) Using the expression for the moment generating function of S in (i) (and without quoting any general results for compound distributions), derive an expression for the variance of S. [4]

[Total 8]

12 The following data give the invoiced amounts for work carried out on 12 jobs performed by a plumber in private customers' houses. The durations of the jobs are also given.

duration x (hrs)	1	1	2	3	4	4	5	6	7	8	9	10
amount y (£)	45	65	80	95	100	125	145	180	180	210	330	240

$$\sum x_i = 60,\ \sum x_i^2 = 402,\ \sum y_i = 1795,\ \sum y_i^2 = 343,725,\ \sum x_i y_i = 11,570$$

The plumber claims to calculate his total charge for each job on the basis of a single call-out charge plus an hourly rate for the time spent working on the job.

(i) (a) Draw a scatterplot of the data on graph paper and comment briefly on your plot.

(b) The equation of the fitted regression line of y on x is $y = 22.4 + 25.4x$ and the coefficient of determination is $R^2 = 87.8\%$ *(you are not asked to verify these results).*

Draw the fitted line on your scatterplot. [5]

(ii) (a) Calculate the fitted regression line of invoiced amount on duration of job using only the 11 pairs of values remaining after *excluding the invoice* for which $x = 9$ and $y = 330$.

(b) Calculate the coefficient of determination of the fit in (ii)(a) above.

(c) Add the second fitted line to your scatterplot, distinguishing it clearly from the first line you added (in part (i)(b) above).

(d) Comment on the effect of omitting the invoice for which $x = 9$ and $y = 330$.

(e) Carry out a test to establish whether or not the slope in the model fitted in (ii)(a) above is consistent with a rate of £25 per hour for work carried out. [13]

[Total 18]

13 The random variable, X, has a gamma distribution with probability density function given by:

$$f(x) = \frac{x^{m-1}\exp(-x/\beta)}{\beta^m \Gamma(m)} \quad (x > 0),$$

where m and β are positive constants. This distribution has mean $m\beta$ and variance $m\beta^2$. Let $x_1, \ldots, x_n$ denote a random sample of n observations on X.

(i) Suppose that m is known.

(a) Show that the maximum likelihood estimate of β is given by $\hat{\beta} = \sum_{i=1}^{n} x_i / mn$.

(b) Show that $\hat{\beta}$ is an unbiased estimator of β.

(c) Obtain the Cramer-Rao lower bound for estimators of β.

(d) Show that the maximum likelihood estimator of β has variance equal to the lower bound given in (i)(c). [10]

Suppose now that m is unknown, and is also to be estimated by maximum likelihood. It is assumed that m is large enough so that $\Gamma(m)$ is well approximated by $g(m) = \exp(-m)m^{(m-0.5)}(2\pi)^{0.5}$.

(ii) Determine the approximate maximum likelihood estimates of β and m, substituting $g(m)$ for $\Gamma(m)$ in the likelihood function. [7]

(iii) Suppose that the sample values are:

32 48 51 43 82 155

Obtain the approximate maximum likelihood estimates for β and m given in (ii). [2]

[Total 19]

14 A social researcher is interested in the gender distribution among children in families, and has collected data for her investigation as follows.

Three hundred families were selected at random. The table below shows frequency distributions of the numbers of girls in families of size 1, 2, 3, and 4, that is with 1, 2, 3, and 4 children.

	Number of girls in family					
Size of family	*0*	*1*	*2*	*3*	*4*	*Number of families*
1	23	27	-	-	-	50
2	30	46	24	-	-	100
3	9	36	43	12	-	100
4	4	17	15	11	3	50

(i) The researcher wants to investigate whether the proportion of girls within families is independent of family size.

(a) Construct a suitable 2×4 contingency table, and calculate the overall proportion of girls.

(b) State appropriate hypotheses to use in the researcher's investigation.

(c) Calculate the value of an appropriate test statistic and state whether or not its probability value exceeds 0.05.

(d) State your conclusion. [12]

(ii) (a) Suggest a model (with all parameter values stated or estimated) for the number of girls in a family, for each family size (1, 2, 3, and 4), using your conclusion from part (i).

(b) Suppose the researcher was to test the goodness-of-fit of the models you have suggested in part (ii)(a) for family sizes 2, 3, and 4 and that the models were rejected as being unsuitable. Discuss briefly how you would interpret this lack of fit. [4]

[Total 16]