EXAMINATIONS

April 2000

Subject 101 — Statistical Modelling

EXAMINERS' REPORT

1 As $n = 14$ the median is half way between the 7^{th} and 8^{th} value
i.e. $m = (1.8 + 1.9)/2 = 1.85$.

The quartiles are the 4^{th} and 11^{th} values, so $Q_1 = 1.5$ and $Q_3 = 2.7$.

OR: Using the definition of the quartiles as the $15/4^{th}$ and $45/4^{th}$ value gives $Q_1 = 1.5$ and $Q_3 = 2.8$.

2 The total claim, T, will be normally distributed with mean $50 \times 1870 = 93500$ and variance $50 \times 610^2 = 18{,}605{,}000 = 4313^2$.

(Alternatively, we can work with the mean claim.)

Thus, the probability that the total claim is greater than £100,000 is

$$1 - \Phi\left(\frac{100{,}000 - 93{,}500}{4313}\right) = 1 - \Phi(1.507) = 0.066.$$

3 $$P(\theta > 0.2) = \int_{0.2}^{1} 9\,(1-\theta)^8\, d\theta$$

$$= \left[-(1-\theta)^9\right]_{0.2}^{1}$$

$$= 0 + (1 - 0.2)^9 = 0.8^9 = 0.13$$

4 For (a) to be true, 0.250 must be lower 5% pt of $F_{6,12}$ i.e. reciprocal of upper 5% pt of $F_{12,6}$ which is $\frac{1}{4.000} = 0.250$ $\therefore$ true.

For (b) to be true, 4.821 must be upper 1% pt of $F_{6,12}$ which is 4.821 $\therefore$ true.

For (c) to be true, 0.130 must be lower 1% pt of $F_{6,12}$ i.e. reciprocal of upper 1% pt of $F_{12,6}$ which is $\frac{1}{7.718} = 0.130$ $\therefore$ true.

5 As claims are independent the number of claims by inexperienced drivers will follow a Poisson distribution with mean $20 \times 0.15 = 3$, and the number of claims made by experienced drivers will follow a Poisson distribution with mean $40 \times 0.1 = 4$. Again using the independence assumption, the total number of claims, X, is Poisson with mean 7.

Thus, $P(X \le 3) = \sum_{i=0}^{3} e^{-7} \frac{7^i}{i!} = 0.082.$

(The answer can also be taken directly from the Green Book, which gives 0.08177.)

6 Total number of claims $X \sim$ Poisson(600λ)

Under H_0: $X \sim$ Poisson(84) ~ $N(84,84)$ approximately

Prob. value = $P(X \le 72) = P[Z < (72.5 - 84)/\sqrt{84}] = P(Z < -1.255) = 0.105$

7 $$E(X) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)dxdy = \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} xf(x \mid y)dx\right\} f_{Y'}(y)dy = \int_{-\infty}^{\infty} E(X \mid Y = y) f_{Y'}(y)dy$$

8 $T = \sum_{i=1}^{100} X_i$ has mean $100(3.6) = 360$ hours

and s.d. $\sqrt{100}(2.6) = 26$ hours.

Central limit theorem $\Rightarrow T$ is approximately normal as n is large.

$$\therefore P(T > 400) \doteqdot P\left(Z > \frac{400 - 360}{26} = 1.54\right)$$

$$= 1 - 0.93822 = 0.062$$

9 (i) This will be a probability function provided the specified probabilities are non-negative; i.e. if and only if $-0.1 \le a \le 0.1$.

(ii) The method of moments estimate of a is obtained by equating the sample mean to the population mean. To do this note that

$$\mu = \Sigma_{-2}^{2}\ i(0.2 + ai) = a\Sigma_{-2}^{2} i^2 = 10a.$$

Thus, the method of moments estimate is $\bar{X}/10$.

As $\bar{X}$ can take any value between –2 and +2, the method of moments estimate can take any value between –0.2 and +0.2. Thus it can be outside the range (–0.1, 0.1).

10 (i) $G'_{X_t}(s) = (1 - \lambda t)\{1 + \lambda t(1 - s)\}^{-1} - \{s + \lambda t(1 - s)\}\{-\lambda t\}\{1 + \lambda t(1 - s)\}^{-2}$

$\therefore \mu = G'_{X_t}(1) = 1 - \lambda t + \lambda t = 1$

(ii) P(extinct by time t) = P(population size at time t is zero)
$= G_{X_t}(0) = \lambda t / (1 + \lambda t)$

(iii) P(extinct by time t) $\to 1$ as $t \to \infty$, so eventual extinction is certain.

11 Complete the table of Expected values:

Expected	*Education*	*Health*	*Poverty*	*Total*
Donate	24.25	24.25	48.5	97
Don't donate	175.75	175.75	351.5	703
	200	200	400	800

Calculate $\chi^2 = 3.98$.

The 5% point of a χ^2 random variable on 2 degrees of freedom is 5.991, so the χ^2 test is not significant at the 5% level.

On the basis of the data collected, it is plausible that the three packs are equally effective.

12 $\Sigma x = 50(-2) + 0 + 60(2) = 20$ $\Sigma x^2 = 50(4) + 0 + 60(4) = 440$

$\Sigma y = 50(2) + 0 + 60(-1) = 40$ $\Sigma y^2 = 50(4) + 0 + 60(1) = 260$

$\Sigma xy = 50(-4) + 0 + 60(-2) = -320$

so $r = [-320 - (20 \times 40)/200]/[(440 - 20^2/200)(260 - 40^2/200)]^{1/2} = -0.975$

13 (i) $F_Y(y) = P(Y < y) = P(1/X < y + 1) = P[X > 1/(y + 1)] = 1 - 1/(y + 1)$
so $f_Y(y) = dF_Y(y)/dy = 1/(y+1)^2$, $y > 0$

(ii) $E(Y) = \int_0^\infty y(1+y)^{-2}dy = \int_0^\infty (1+y)^{-1}dy - \int_0^\infty (1+y)^{-2}dy$

The integral of $(1 + y)^{-1}$ gives $\log(1 + y)$ which $\to \infty$ as $y \to \infty$ so this integral is not finite. So $E(Y)$ does not exist.

14 (i) (a) $\bar{x}_A = 247.9,\ s_A^2 = \frac{1}{9}\left\{617163 - \frac{2479^2}{10}\right\} = 290.99$

$\bar{x}_B = 261.9,\ s_B^2 = \frac{1}{9}\left\{687467 - \frac{2619^2}{10}\right\} = 172.32$

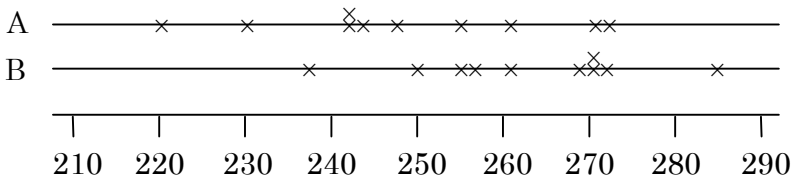
$s_p^2 = \frac{290.99 + 172.32}{2} = 231.66$

Obs $t = \frac{247.9 - 261.9}{\sqrt{231.66\left(\frac{1}{10} + \frac{1}{10}\right)}} = -2.06$

For two-sided test, $t_{18}(2.5\%) = 2.101$.

As $2.06 < 2.101$, there is no evidence at the 5% level of a difference between regions A and B.

(b)



Normality — OK in both cases.

Equal variances — OK.

(c) $\dfrac{18S_p^2}{\sigma^2} \sim \chi^2_{18}$

$\therefore$ 95% CI for σ^2 is $\left(\dfrac{18S_p^2}{\chi^2_{0.975,18}}, \dfrac{18S_p^2}{\chi^2_{0.025,18}} \right)$

$= \left(\dfrac{18(231.66)}{31.53}, \dfrac{18(231.66)}{8.231} \right) = (132.25,\ 506.6)$

$\therefore$ 95% CI for σ is (11.5, 22.5)

(ii) (a) $\Sigma x = 10216$, $\Sigma x^2 = 2621210$

$SS_T = 2621210 - \dfrac{10216^2}{40} = 12043.6$

$SS_B = \dfrac{2479^2 + 2619^2 + 2441^2 + 2677^2}{10} - \dfrac{10216^2}{40} = 3774.8$

$\therefore$ $SS_R = 8268.8$ by subtraction.

Source	*df*	*SS*	*MS*
Regions	3	3774.8	1258.3
Residual	36	8268.8	229.7
Total	39	12043.6	

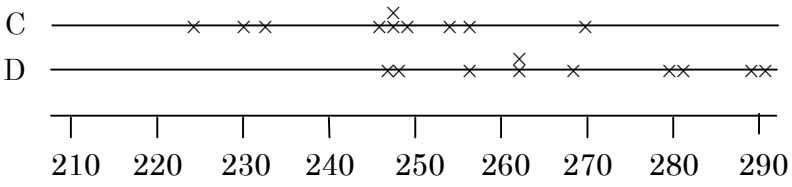
$F = \dfrac{1258.3}{229.7} = 5.48$ on (3,36) df

$F_{3,36}(5\%) \doteqdot 2.9$ by interpolation

Clearly reject $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ at the 5% level

$\therefore$ Strong evidence of a difference between regions A–D.

(b)



[same scale!]

Normality — OK.

Equal variances for A, B, C, D — OK.

(c) $\dfrac{36\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{36}$ where $36\hat{\sigma}^2 = SS_R$

$\therefore$ 95% CI for σ^2 is $\left(\dfrac{SS_R}{\chi^2_{0.975,36}}, \dfrac{SS_R}{\chi^2_{0.025,36}} \right)$

$= \left(\dfrac{8268.8}{54.4}, \dfrac{8268.8}{21.37} \right) = (152.0, 386.9)$ [interpolate in tables]

$\therefore$ 95% CI for σ is (12.3, 19.7)

(iii) Second CI is **narrower** as it is based on **more data**.

15 (i) Start by writing down the likelihood function

$$L(\alpha) = \frac{(\alpha - 1)^n}{\Pi(1 + x_i)^\alpha}.$$

The log-likelihood function is

$$l(\alpha) = \log L(\alpha) = n \log(\alpha - 1) - \alpha\Sigma \log(1 + x_i).$$

Differentiating gives

$$\frac{\partial l}{\partial \alpha} = \frac{n}{\alpha - 1} - \Sigma \log(1 + x_i).$$

It is easy to see that the log-likelihood has only one turning point, so this can be found by equating the derivative to zero. This gives that the maximum likelihood estimate is

$$\hat{\alpha} = 1 + \frac{n}{\Sigma \log(1 + x_i)}.$$

The second derivative is

$$\frac{\partial^2 l}{\partial \alpha^2} = \frac{-n}{(\alpha - 1)^2}.$$

So an approximate 95% confidence interval for α is $\hat{\alpha} \pm 1.96 \dfrac{\hat{\alpha} - 1}{\sqrt{n}}$.

(ii) (a) The probability a component will last less than 12 hours before failing can be estimated by the point estimate

$$1 - \frac{1}{13^{\hat{\alpha}-1}} = 1 - \frac{1}{13^{0.56}} = 0.762.$$

(b) An approximate 95% confidence upper bound for α is

$$\hat{\alpha} + 1.645\,\frac{\hat{\alpha}-1}{\sqrt{n}} = 1.56 + 1.645\,\frac{0.56}{\sqrt{80}} = 1.663.$$

(c) This gives an upper bound for the failure probability of

$$1 - \frac{1}{13^{0.663}} = 0.817.$$

(iii) (a) The endpoint of the binomial confidence interval is

$$0.7625 + 1.645 \times \sqrt{\frac{0.7625 \times (1 - 0.7625)}{80}} = 0.841.$$

(b) There is no single answer to this part. The main points are:

The second engineer's (binomial) method will be valid and doesn't need any parametric assumptions about the data.

The first engineer's method needs the data to follow the distribution specified, but in that case it will be more powerful than the binomial method, which does not use the data efficiently. This is illustrated in this data as the two point estimates are very close, but the first engineer's confidence interval is narrower.

16 (i) $S_{xx} = 91.3978 - \frac{34.023^2}{13} = 2.354$, and

$$S_{xy} = 286.6299 - \frac{110.679 \times 34.023}{13} = -3.0341.$$

The least squares estimates are $\hat{\beta} = \frac{-3.0341}{2.354} = -1.289$ and

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = 11.887.$

(ii) (a) The sum of squares of the residuals is 0.049019, so σ^2 is estimated by the residual mean square $0.049019/11 = 0.004456$.

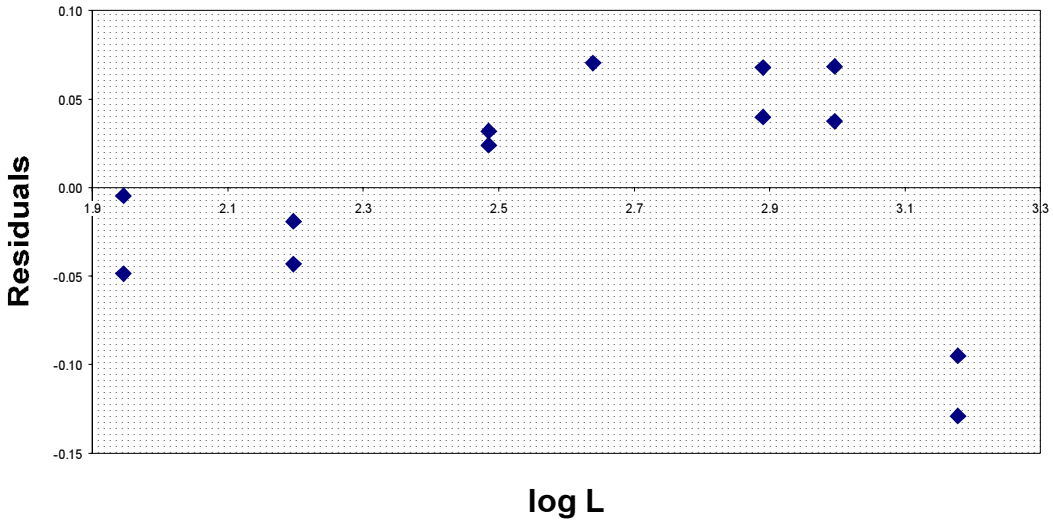
(b) The estimated variance of $\hat{\beta}$ is $0.004456/2.3544 = 0.00189$.

This leads to the following 95% confidence interval for β:

$$-1.289 \pm 2.201\sqrt{0.00189} = (-1.384, -1.193).$$

(c) If the relationship $P = k/L$ is correct the slope parameter of the regression line should be –1. As the upper end of the interval is less than –1, the data do not support the suggested relationship.

(iii)



The residuals show that the line underfits in the centre. A straight line doesn't fit the data very well.