THE ROYAL STATISTICAL SOCIETY

2006 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

PAPER I – STATISTICAL THEORY

The Society provides these solutions to assist candidates preparing for the examinations in future years and for the information of any other persons using the examinations.

The solutions should NOT be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids.

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

Note. In accordance with the convention used in the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Higher Certificate, Paper I, 2006. Question 1

(i) The first place can be occupied by 9 different digits, 1 to 9. Each of the other three places can be occupied by 10 digits, 0 to 9.

Hence there are $9 \times 10 \times 10 \times 10 = 9000$ possible PINs.

(ii) All of the combinations in (i) are allowed except 1111, 2222, …, 9999, so there are $9000 - 9 = 8991$ possibilities.

(iii) Only the 9 digits 1 to 9 can be used. The first place can be filled in 9 ways, the second in 8, the third in 7 and the last in 6. So there are $9 \times 8 \times 7 \times 6 = 3024$ possibilities.

(iv) With all 10 digits possible in any position, there would be 10^4 PINs. There are 7 increasing sequences (0123, 1234, …, 6789) and 7 decreasing sequences (9876, 8765, …, 3210), which are not allowed. The number of possible PINs is therefore $10^4 - 14 = 9986$.

(v) All of the 10^4 combinations are allowed except:

(a) the 10 where all 4 digits are the same: 0000, 1111, …, 9999;

(b) those where one digit occurs three times and another just once. There are $10 \times 9 = 90$ ways of choosing the two digits. But note that, for example, 2333, 3233, 3323 and 3332 are four different PINs; whichever two digits occur, the odd one out can be in any of the 4 places in the PIN. Therefore there are $4 \times 90 = 360$ PINs of this sort.

The number of possible PINs is therefore $10^4 - 10 - 360 = 9630$.

Higher Certificate, Paper I, 2006. Question 2

(i) A: (a) $P(0 \text{ entries}) = \left(\frac{1}{2}\right)^2 = \frac{1}{4} = 0.25\,.$

(b) $P(1 \text{ entry}) = 2 \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{2} = 0.5.$

B: (a) $P(0 \text{ entries}) = \left(\frac{3}{4}\right)^3 = \frac{27}{64} = 0.4219\,.$

(b) $P(1 \text{ entry}) = 3 \times \frac{1}{4} \times \left(\frac{3}{4}\right)^2 = \frac{27}{64} = 0.4219.$

C: (a) $P(0 \text{ entries}) = \left(\frac{4}{5}\right)^5 = \frac{1024}{3125} = 0.3277\,.$

(b) $P(1 \text{ entry}) = 5\left(\frac{1}{5}\right)\left(\frac{4}{5}\right)^4 = \frac{256}{625} = 0.4096\,.$

(ii) P(1 entry in total)

= P(1 from A, 0 from B and C) + P(1 from B, 0 from A and C)
+ P(1 from C, 0 from A and B)

$$= \frac{1}{2} \times \frac{27}{64} \times \frac{1024}{3125} + \frac{27}{64} \times \frac{1}{4} \times \frac{1024}{3125} + \frac{256}{625} \times \frac{1}{4} \times \frac{27}{64} = \frac{459}{3125}\,.$$

[If worked in decimals, this is 0.1469.]

P(1 from A | 1 in total) = P(1 from A and 1 in total) / P(1 in total)

= P(1 from A, 0 from B and C) / P(1 in total)

$$= \frac{\frac{1}{2} \times \frac{27}{64} \times \frac{1024}{3125}}{\frac{459}{3125}} = \frac{8}{17}\,.$$

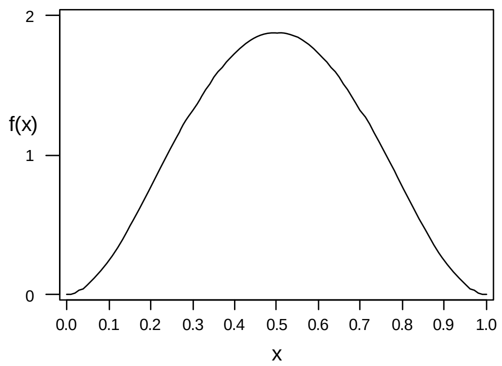
(iii) Denote the numbers of entries from A, B, C as (0, 0, 0) etc. Then we need $P(2, 0, 0) + P(0, 2, 0) + P(0, 0, 2) + P(1, 1, 0) + P(1, 0, 1) + P(0, 1, 1)$. Since entries from each group are independent, we have, as an example, $P(1, 1, 0) =$ P(1 from A).P(1 from B).P(0 from C).

Higher Certificate, Paper I, 2006. Question 3

(i) We have $k\int_0^1 x^2(1-x)^2dx = 1$, so $k\int_0^1\left(x^2-2x^3+x^4\right)dx = 1$. This gives

$$1 = k\left[\frac{1}{3}x^3 - \frac{1}{2}x^4 + \frac{1}{5}x^5\right]_0^1 = k\left(\frac{1}{3}-\frac{1}{2}+\frac{1}{5}\right), \quad \text{so } k = 30.$$

$f(x) = 0$ at $x = 0$ and at $x = 1$. $f(x)$ is symmetrical about $x = ½$. The sketch is as follows.



(ii) $E(X) = \frac{1}{2}$ by symmetry [or by direct integration: $\int_0^1 xf(x)dx$].

$$E(X^2) = 30\int_0^1 x^4(1-x)^2dx = 30\int_0^1\left(x^4-2x^5+x^6\right)dx$$

$$= 30\left[\frac{1}{5}x^5 - \frac{1}{3}x^6 + \frac{1}{7}x^6\right]_0^1 = 30\left(\frac{1}{5}-\frac{1}{3}+\frac{1}{7}\right) = 30\times\frac{1}{105} = \frac{2}{7}.$$

$$\therefore \text{Var}(X) = E(X^2) - \{E(X)\}^2 = \frac{2}{7} - \left(\frac{1}{2}\right)^2 = \frac{1}{28}.$$

$$P\left(X \le \frac{1}{3}\right) = \int_0^{1/3} 30\left(x^2-2x^3+x^4\right)dx = 30\left[\frac{1}{3}x^3 - \frac{1}{2}x^4 + \frac{1}{5}x^5\right]_0^{1/3}$$

$$= 30\left(\frac{1}{3^4} - \frac{1}{2}.\frac{1}{3^4} + \frac{1}{5}.\frac{1}{3^5}\right) = \frac{30}{81}\left(1-\frac{1}{2}+\frac{1}{15}\right) = \frac{30}{81}\times\frac{17}{30} = \frac{17}{81} \quad (= 0.2099).$$

(iii) The required probability is $\left(1-\frac{17}{81}\right)^5 = \left(\frac{64}{81}\right)^5 = 0.3079.$

(iv) The variance of $\bar{X}$ for a sample of size 5 is $\frac{\text{Var}(X)}{5} = \frac{1/28}{5} = \frac{1}{140} = 0.00714.$

Higher Certificate, Paper I, 2006. Question 4

Let X represent cycling time without delays: $X \sim N(15, 1)$.

(i) $P(X \le 17) = \Phi\left(\frac{17-15}{1}\right) = \Phi(2) = 0.9772$.

[Φ denotes the cdf of the standard Normal distribution as usual.]

(ii) Adding in the delay times, also Normally distributed [N(0.7, 0.09)], and letting T denote the total time:

(a) $T \sim N(15.7, 1.09)$, so $P(T \le 17) = \Phi\left(\frac{17-15.7}{\sqrt{1.09}}\right) = \Phi(1.245) = 0.8934$;

(b) $T \sim N(16.4, 1.18)$, so $P(T \le 17) = \Phi\left(\frac{17-16.4}{\sqrt{1.18}}\right) = \Phi(0.552) = 0.7096$;

(c) $T \sim N(17.1, 1.27)$, so $P(T \le 17) = \Phi\left(\frac{17-17.1}{\sqrt{1.27}}\right) = \Phi(-0.0887) = 0.4646$.

(iii) The number of delays is distributed as B(3, ½). Hence the situations in (i), (ii)(a), (ii)(b) and (ii)(c) arise with probabilities 1/8, 3/8, 3/8 and 1/8 respectively, so the (unconditional) mean of the total journey time is

$$E(T) = \frac{1}{8}\times 15 + \frac{3}{8}\times 15.7 + \frac{3}{8}\times 16.4 + \frac{1}{8}\times 17.1 = \frac{128.4}{8} = 16.05 \text{ minutes.}$$

(iv) Mean time $\bar{T} \sim N\left(16.05, \frac{1.5025}{10}\right)$.

$$P(\bar{T} \le 17) = \Phi\left(\frac{17-16.05}{\sqrt{0.15025}}\right) = \Phi(2.451) = 0.9929.$$

Higher Certificate, Paper I, 2006. Question 5

(i) $E(X) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda e^{-\lambda} \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$.

$E(X^2) = E\left[X(X-1)+X\right] = E\left[X(X-1)\right] + E[X]$.

$$E\left[X(X-1)\right] = \sum_{x=0}^{\infty} x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = \lambda^2 e^{-\lambda} \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = \lambda^2 e^{-\lambda} e^{\lambda} = \lambda^2 .$$

Hence $E(X^2) = \lambda^2 + \lambda$, and $\text{Var}(X) = E(X^2) - \{E(X)\}^2 = \lambda$.

(ii) $L = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$, and hence $\log L = -n\lambda + \sum_{i=1}^{n} x_i \log\lambda + \text{constant}$.

$\therefore \frac{d\log L}{d\lambda} = -n + \frac{\Sigma x_i}{\lambda}$ which on setting equal to zero gives that the maximum likelihood estimate is $\hat{\lambda} = \frac{\Sigma x_i}{n} = \bar{x}$. [Consideration of $\frac{d^2 \log L}{d\lambda^2}$ confirms that this is a maximum: $\frac{d^2 \log L}{d\lambda^2} = \frac{-\Sigma x_i}{\lambda^2} < 0$.]

(iii) $\text{Var}\left(\hat{\lambda}\right) = \text{Var}\left(\bar{X}\right) = \frac{\text{Var}(X)}{n} = \frac{\lambda}{n}$.

Thus the maximum likelihood estimator of Var($\hat{\lambda}$) is $\frac{\hat{\lambda}}{n}$.

By the central limit theorem, $\hat{\lambda}\ (= \bar{X})$ is approximately Normally distributed with mean λ and variance λ/n. We estimate the variance by $\hat{\lambda}/n$, so that we have $\hat{\lambda} \sim \text{N}\left(\lambda,\ \frac{\hat{\lambda}}{n}\right)$, approximately.

Hence an approximate 95% confidence interval is given by

$$0.95 \approx P\left(-1.96 < \frac{\hat{\lambda}-\lambda}{\hat{\lambda}/\sqrt{n}} < 1.96\right),$$

leading to the interval $\left(\hat{\lambda} - 1.96\sqrt{\frac{\hat{\lambda}}{n}}\ ,\quad \hat{\lambda} + 1.96\sqrt{\frac{\hat{\lambda}}{n}}\right)$.

Solution continued on next page

(iv) For the given sample, we have $n = 12$ and $\Sigma x_i = 48$, leading to $\hat{\lambda} = \bar{x} = 4$. The approximate confidence interval is therefore

$$\left(4 - 1.96\sqrt{\frac{4}{12}} \quad \text{to} \quad 4 + 1.96\sqrt{\frac{4}{12}}\right), \quad \text{i.e. } 2.87 \text{ to } 5.13.$$

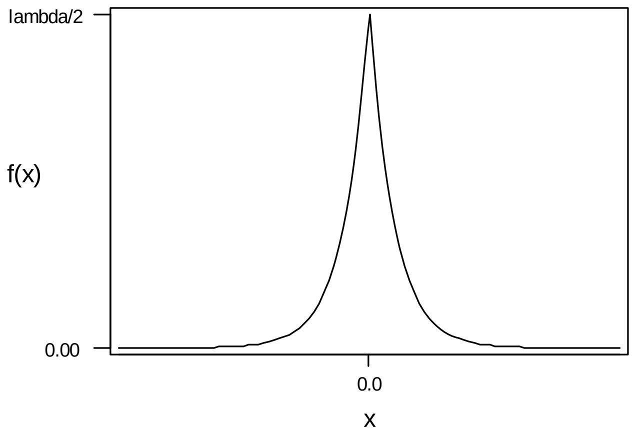
The sample also gives $\Sigma x_i^2 = 238$; so the sample variance is

$$s^2 = \frac{1}{11}\left(238 - \frac{48^2}{12}\right) = \frac{46}{11} = 4.182\,.$$

This is close to the sample mean (4), supporting a Poisson hypothesis for the underlying model.

Higher Certificate, Paper I, 2006. Question 6

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x|}, \qquad -\infty < x < \infty$$



By symmetry, $E(X) = 0$.

Hence $\text{Var}(X) = E(X^2) - 0 = \frac{\lambda}{2}\int_{-\infty}^{\infty} x^2 e^{-\lambda|x|} dx = \frac{\lambda}{2}\left\{\int_{-\infty}^{0} x^2 e^{\lambda x} dx + \int_{0}^{\infty} x^2 e^{-\lambda x} dx\right\}$.

Substituting $u = -x$ in the first integral gives $\int_0^\infty u^2 e^{-\lambda u} du$, which is the same as the second. Hence we get, integrating by parts,

$$E(X^2) = \lambda \int_0^\infty x^2 e^{-\lambda x} dx$$

$$= \lambda \left\{ \left[x^2 \frac{e^{-\lambda x}}{-\lambda} \right]_0^\infty + \int_0^\infty \frac{e^{-\lambda x}}{\lambda} .2x\, dx \right\}$$

$$= [0-0] + \int_0^\infty 2x e^{-\lambda x} dx$$

$$= 2\left\{ \left[x \frac{e^{-\lambda x}}{-\lambda} \right]_0^\infty + \int_0^\infty \frac{e^{-\lambda x}}{\lambda} dx \right\}$$

$$= [0-0] + \frac{2}{\lambda}\left[\frac{e^{-\lambda x}}{-\lambda} \right]_0^\infty = \frac{2}{\lambda^2}.$$

Solution continued on next page

If Q, q are the upper and lower quartiles, we have $\int_0^Q \frac{1}{2}\lambda e^{-\lambda x}dx = \frac{1}{4}$, and q will be the same distance below 0 by symmetry.

$\therefore \frac{1}{4} = \left[-\frac{1}{2}e^{-\lambda x}\right]_0^Q = \frac{1}{2}\left(-e^{-\lambda Q}+1\right)$, giving $\frac{1}{2} = 1 - e^{-\lambda Q}$. Therefore $\lambda Q = \log 2$. Hence the semi-interquartile range is $(\log 2)/\lambda$.

$L = \prod_{i=1}^{n}\left(\frac{\lambda}{2}e^{-\lambda|x_i|}\right) = \left(\frac{\lambda}{2}\right)^n e^{-\lambda\sum|x_i|}$, and hence $\log L =$ constant $+ n\log\lambda - \lambda\sum_i |x_i|$.

$\therefore \frac{d\log L}{d\lambda} = \frac{n}{\lambda} - \sum_i |x_i|$ which on setting equal to zero gives that the maximum likelihood estimate is $\hat{\lambda} = \frac{n}{\sum_i |x_i|}$. [Consideration of $\frac{d^2\log L}{d\lambda^2}$ confirms that this is a maximum: $\frac{d^2\log L}{d\lambda^2} = \frac{-n}{\lambda^2} < 0$.]

Higher Certificate, Paper I, 2006. Question 7

(i) The sum of all 12 table entries is $30c$. These probabilities must add up to 1, so $c = 1/30$.

(ii) The marginal distributions are given by the row and column totals.

Hence: $P(X = 1) = 15c = 1/2$; $P(X = 2) = 10c = 1/3$; $P(X = 3) = 5c = 1/6$.

Similarly: $P(Y = 1) = 12c = 2/5$; $P(Y = 2) = 6c = 1/5$; $P(Y = 3) = 6c = 1/5$; $P(Y = 4) = 6c = 1/5$.

(iii) $$E(X) = \left(1\times\frac{1}{2}\right)+\left(2\times\frac{1}{3}\right)+\left(3\times\frac{1}{6}\right) = \frac{1}{2}+\frac{2}{3}+\frac{1}{2} = \frac{5}{3}.$$

$$E(X^2) = \left(1\times\frac{1}{2}\right)+\left(4\times\frac{1}{3}\right)+\left(9\times\frac{1}{6}\right) = \frac{1}{2}+\frac{4}{3}+\frac{3}{2} = \frac{10}{3}.$$

$$\therefore \text{Var}(X) = \frac{10}{3}-\left(\frac{5}{3}\right)^2 = \frac{5}{9}.$$

We also need $E(Y)$ later: $E(Y) = \frac{2}{5}+\frac{2}{5}+\frac{3}{5}+\frac{4}{5} = \frac{11}{5}$.

Distribution of XY:

Values of xy	1	2	3	4	6	12	
Probability	$6c$	$7c$	$4c$	$6c$	$5c$	$2c$	[$c = 1/30$, see above]

$$E(XY) = \left(1\times\frac{6}{30}\right)+\left(2\times\frac{14}{30}\right)+\left(3\times\frac{4}{30}\right)+\left(4\times\frac{6}{30}\right)+\left(6\times\frac{5}{30}\right)+\left(12\times\frac{2}{30}\right) = \frac{110}{30} = \frac{11}{3}$$

Also we have $E(X)E(Y) = \frac{5}{3}\times\frac{11}{5} = \frac{11}{3}$.

$$\therefore \text{Cov}(X,Y) = E(XY) - E(X)E(Y) = 0.$$

(iv) X and Y are not independent [even though $\text{Cov}(X, Y) = 0$ and even though some cells have $P(X = x, Y = y) = P(X = x).P(Y = y)$]. For example, we have $P(X = 1, Y = 4) = 2/15$, but $P(X = 1).P(Y = 4) = 1/10$.

Solution continued on next page

(v) $U = 1$ if $X = 1$ or 3 $U = 0$ if $X = 2$

$V = 1$ if $Y = 1$ or 3 $V = 0$ if $Y = 2$ or 4

Table of joint distribution of U and V, with margins.

		Values of V		
		0	1	
Values of U	0	$2c = 1/15$	$8c = 4/15$	$10c = 1/3$
	1	$10c = 1/3$	$10c = 1/3$	$20c = 2/3$
		$12c = 2/5$	$18c = 3/5$	

Consider the cell with $(U, V) = (0, 0)$. The cell probability is 1/15 but the product of the marginal probabilities is 2/15. So U and V are not independent.

Higher Certificate, Paper I, 2006. Question 8

(i) $Y_i = a + bx_i + e_i$, $i = 1, 2, \ldots, n$.

The $\{e_i\}$ are uncorrelated random variables with mean 0 and constant variance σ^2.

The method of least squares is equivalent to the method of maximum likelihood for estimating the regression coefficients (a and b) if the $\{e_i\}$ are Normally distributed.

[If the analysis is to proceed to *inference* for the regression coefficients, Normality of the $\{e_i\}$ is required.]

(ii)(a) For $Y_i = \beta x_i + e_i$, we minimise $S = \sum e_i^2 = \sum (y_i - \beta x_i)^2$.

We have $\frac{dS}{d\beta} = -2\sum x_i (y_i - \beta x_i)$ which on setting equal to zero gives $\sum x_i y_i = \beta \sum x_i^2$, so the least squares estimate is $\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$.

[Consideration of $\frac{d^2 S}{d\beta^2}$ confirms that this is a minimum: $\frac{d^2 S}{d\beta^2} = 2\sum x_i^2 > 0$.]

(b) See scatter plot at foot of page. It shows an increasing trend, roughly linear; but there seems to be some increase in variability as x increases. There are not enough data points to be sure.

The usual summary statistics (not all required for the zero intercept model) are

$n = 10$, $\Sigma x_i = 180$, $\Sigma y_i = 40$, $\Sigma x_i^2 = 5150$, $\Sigma y_i^2 = 244$, $\Sigma x_i y_i = 1055$.

$\therefore \hat{\beta} = 1055/5150 = 0.205$. So the fitted line is $y = 0.205x$.

Hence the estimated expected number of violations for $x = 20$ is $0.205 \times 20 = 4.1$.

Logically, zero traffic flow should imply zero speed violations, so that y should be 0 when x is 0, i.e. the zero intercept model seems reasonable. The scatter plot does not contradict this.