THE ROYAL STATISTICAL SOCIETY

2002 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.

Higher Certificate, Paper I, 2002. Question 1

(i) $P(\text{all 13 from 39 non-spades}) = \dfrac{\binom{39}{13}}{\binom{52}{13}} = \dfrac{39!\ 13!\ 39!}{13!\ 26!\ 52!}$

$$= \frac{39\times 38\times \ldots\ldots \times 28\times 27}{52\times 51\times \ldots\ldots \times 41\times 40} = 0.01279.$$

(ii) There are 32 "low" cards (2 up to 9) and 20 "high" (10 to Ace). Hence all 13 are selected from 32, so probability is

$$\frac{\binom{32}{13}}{\binom{52}{13}} = \frac{32!13!39!}{13!19!52!} = \frac{32!39!}{19!52!} \qquad \text{(which is 0.000547).}$$

(iii) 6 of the 13 Hearts, and 7 of the remaining 39, must be selected.

Probability is $\dfrac{\binom{13}{6}\binom{39}{7}}{\binom{52}{13}} = \dfrac{13!39!13!39!}{6!7!7!32!52!} = \dfrac{(13!)^2(39!)^2}{(6!)(7!)^2(32!)(52!)}$ (= 0.04156).

(iv) The division of Clubs must be 3, 3, 3, 4 between players, any one of whom can be the person who receives 4; there are 4 ways for this. If the first player receives 3, this can happen in $\binom{13}{3}$ ways, then $\binom{10}{3}$, $\binom{7}{3}$ and $\binom{4}{4}$ for the other players. Their other cards are then selected from the remaining ones in $\binom{39}{10}$, $\binom{29}{10}$, $\binom{19}{10}$ and $\binom{9}{9}$ ways respectively. Therefore the probability is

$$4\frac{\binom{13}{3}\binom{39}{10}\cdot\binom{10}{3}\binom{29}{10}\cdot\binom{7}{3}\binom{19}{10}\cdot\binom{4}{4}\binom{9}{9}}{\left\{\dfrac{52!}{(13!)^4}\right\}},$$ the divisor being the number of ways of splitting 52 cards into 4 hands of 13 each.

Probability is $4\dfrac{13!39!}{3!10!10!29!}\cdot\dfrac{10!29!}{3!7!10!19!}\cdot\dfrac{7!19!}{3!4!10!9!}\cdot 1\times\dfrac{13!13!13!13!}{52!}$

$= 4\dfrac{(39!)(13!)^5}{(52!)(10!)^3(9!)(3!)^3(4!)} = \dfrac{39!(13!)^5}{52!(10!)^3(9!)(3!)^4}$ (which is 0.1054).

Higher Certificate, Paper I, 2002. Question 2

$$H \sim \mathrm{N}(160, 16)$$

(i) $P(156 < H \le 164) = P\left(\dfrac{156-160}{4} < Z \le \dfrac{164-160}{4}\right)$ [where $Z \sim \mathrm{N}(0,1)$]
$= P(-1 < Z \le 1) = \Phi(1) - \Phi(-1) = \Phi(1) - \{1 - \Phi(1)\}$ by symmetry, i.e. $2(\Phi(1)) - 1$
$= 0.6826$.

$P(H > 168) = P(Z > 2) = 1 - \Phi(2) = 0.0228$.

(ii) (a) The relevant portion of the Normal distribution of H is that beginning at 168, which corresponds to the value $Z = 2$. The median value m within this portion has $\frac{1}{2}(0.0228)$ probability above it, i.e. 0.0114, so $\Phi(m)$ = $1 - 0.0114 = 0.9886$, corresponding to $Z = 2.277$. The corresponding value of H is $\mu + \sigma Z$ which is $160 + (4 \times 2.277) = 169.1$ cm.

(b) H = 170 corresponds to $Z = \dfrac{10}{4} = 2.5$; we have $\Phi(2.5) = 0.9938$, so $1 - \Phi(2.5) = 0.0062$. Conditional on $H > 168$, the probability is $\dfrac{P(H > 170)}{P(H > 168)}$
$= \dfrac{0.0062}{0.0228} = 0.272$.

(iii) Mean height of 25 members $\sim \mathrm{N}\left(169.5, \dfrac{1.352^2}{25}\right)$.

$$P(\text{mean} > 170) = P\left(Z > \frac{170 - 169.5}{(1.352/5)}\right) = 1 - \Phi\left(\frac{0.5 \times 5}{1.352}\right) = 1 - \Phi(1.849)$$

$$= 1 - 0.9677 = 0.0323.$$

<u>Higher Certificate, Paper I, 2002. Question 3</u>

(i) The taster does not know how many there are of each sort, so with random guessing $p = \frac{1}{2}$ is the probability of being correct. Binomial $\left(10, \frac{1}{2}\right)$ is a satisfactory model if the samples are independent (i.e. presented randomly). Hence the mean is $np = 5$, and variance $np(1-p) = 2.5$.

(ii)

		SAMPLE		
		Butter	*Margarine*	*TOTAL*
GUESS	Butter	x	$5-x$	5
	Margarine	$5-x$	x	5
	TOTAL	5	5	10

The correct value of X would be 5, but with random guessing the actual value of X may be 0, 1, 2, 3, 4 or 5. With random guessing, all of the $\binom{10}{5}$ ways of guessing 5 of each type will be equally likely. The number of ways of guessing x out of 5 butter samples is $\binom{5}{x}$, and of guessing $(5-x)$ out of 5 margarine is $\binom{5}{5-x}$. The total number of ways of generating the above table is $\binom{5}{x}\binom{5}{5-x}$, and each has probability $1/\binom{10}{5}$, so the required probability distribution is $p(x) = \binom{5}{x}\binom{5}{5-x}\Big/\binom{10}{5}$, for $x = 0, 1, \ldots, 5$.

$\binom{10}{5} = \frac{10!}{5!5!} = 252$. For $x = 0$ or 5, the numerator in $p(x)$ is 1. For $x = 1$ or 4, the numerator is $\binom{5}{1}\binom{5}{4} = 5\times 5 = 25$; and for $x = 2$ or 3 it is $\binom{5}{2}\binom{5}{3} = 100$. Therefore the probability mass function for x is

x	0	1	2	3	4	5
$p(x)$	1/252	25/252	100/252	100/252	25/252	1/252

By symmetry the mean is 2.5.

$$\sum x^2 p(x) = \frac{1}{252}\left((25\times 17) + (100\times 13) + (1\times 25)\right) = \frac{1750}{252} = \frac{125}{18}.$$

Therefore Var(X) = $E\left[X^2\right] - \left(E[X]\right)^2 = \frac{125}{18} - \frac{25}{4} = \frac{250-225}{36} = \frac{25}{36}$.

The number of samples <u>out of all 10</u> that are guessed correctly is $2X$, which has mean = 5 and variance = $\frac{25}{9}$.

<u>Higher Certificate, Paper I, 2002. Question 4</u>

(i) $P(X=r)=\dfrac{e^{-4}4^r}{r!}, \quad r = 0, 1, 2, \dots .$

$P(0)=e^{-4}=0.0183$

$P(1)=4e^{-4}=0.0733$

$P(2)=2P(1)=0.1465$

$P(3)=\dfrac{4}{3}P(2)=0.1954$

$P(4)=P(3)=0.1954$

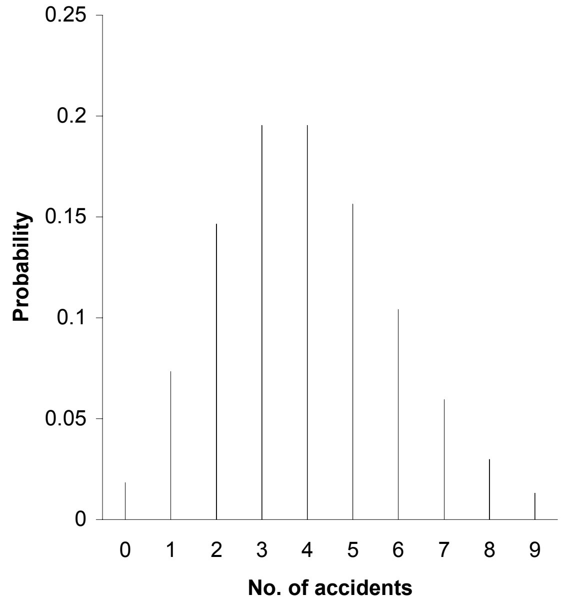
$P(5)=\dfrac{4}{5}P(4)=0.1563$

$P(6)=\dfrac{2}{3}P(5)=0.1042$

$P(7)=\dfrac{4}{7}P(6)=0.0595$

$P(8)=\dfrac{1}{2}P(7)=0.0298$

$P(9)=\dfrac{4}{9}P(8)=0.0132$



and so on (probabilities beyond $r = 9$ have not been shown on the diagram).

(a) P(at most 1 fatal accident) = $P(0) + P(1) = 0.0916$.

(b) In half-year, mean = 2 giving $P(0)=e^{-2}=0.1353$.

(c) In $1\frac{1}{2}$ years, mean = 6 giving $P(0)=e^{-6}=0.00248$.

(ii) Given that X and Y are independent Poissons, $V = X + Y \sim$ Poisson(16).

$E[V]=\mathrm{Var}(V)=16$.

$$P(X=5\,|\,V=20)=\frac{P(X=5,Y=15)}{P(V=20)}=\frac{e^{-4}4^5}{5!}\cdot\frac{e^{-12}12^{15}}{15!}\cdot\frac{20!}{e^{-16}16^{20}}$$

$$=\frac{4^5\,12^{15}}{16^{20}}\cdot\frac{20!}{5!15!}=\frac{4^{20}\,3^{15}}{16^{20}}\cdot\frac{20\times19\times18\times17\times16}{5\times4\times3\times2}=\frac{3^{15}}{4^{20}}.19\times17\times48$$

$= 0.2023$.

(iii) W is Poisson with mean $16 \times 4 = 64$. Use Normal approximation N(64, 64).

$P(W>70)=P\left(Z>\dfrac{70.5-64}{8}\right)$, where $Z\sim\mathrm{N}(0,1)$ and a continuity correction is used. This is $P(Z>0.8125)$, which is $1-\Phi(0.8125)=1-0.7917=0.2083$.

<u>Higher Certificate, Paper I, 2002. Question 5</u>

Let A_1, …, A_k be a set of mutually exclusive and exhaustive events, and let B be any other event.

Then $P(A_i \mid B) = \dfrac{P(B \mid A_i)P(A_i)}{P(B)} = \dfrac{P(B \mid A_i)P(A_i)}{\sum_{j=1}^{k} P(B \mid A_j)P(A_j)}$.

It is useful in inference about an A_i which cannot be observed directly but is related to an observable event B.

$P(011100|A) = P(5 \text{ right, } 1 \text{ wrong}) = p(1-p)^5$.

$P(011100|R) = P(3 \text{ right, } 3 \text{ wrong}) = p^3(1-p)^3$.

$P(011100|S) = P(2 \text{ right, } 4 \text{ wrong}) = p^4(1-p)^2$.

This assumes all errors are independent.

Given that $P(A) = 0.1$, $P(R) = 0.4$, $P(S) = 0.5$, we have $P(A|011100) = \dfrac{0.1 \times p(1-p)^5}{P(B)}$,

where $P(B) = 0.1p(1-p)^5 + 0.4p^3(1-p)^3 + 0.5p^4(1-p)^2$.

So $P(A|011100) = \dfrac{p(1-p)^5}{\left(p(1-p)^5 + 4p^3(1-p)^3 + 5p^4(1-p)^2\right)}$

$$= \frac{(1-p)^3}{(1-p)^3 + 4p^2(1-p) + 5p^3}.$$

Similarly, $P(R|011100) = \dfrac{4p^3(1-p)^3}{\left(p(1-p)^5 + 4p^3(1-p)^3 + 5p^4(1-p)^2\right)}$

$$= \frac{4p^2(1-p)}{(1-p)^3 + 4p^2(1-p) + 5p^3}.$$

Also, $P(S|011100) = \dfrac{5p^4(1-p)^2}{\left(p(1-p)^5 + 4p^3(1-p)^3 + 5p^4(1-p)^2\right)}$

$$= \frac{5p^3}{(1-p)^3 + 4p^2(1-p) + 5p^3}.$$

Continued on next page

As $p \to 0$, $P(A|011100) \to 1$, while the others do not.

In general, $P(A|011100) > P(R|011100)$ if $(1-p)^3 > 4p^2(1-p)$, or $(1-p)^2 > 4p^2$, which requires $1-p > 2p$, i.e. $1 > 3p$ or $p < \frac{1}{3}$.

Likewise $P(A|011100) > P(S|011100)$ if $(1-p)^3 > 5p^3$, or $(1-p) > p\sqrt[3]{5}$, which requires $1 > p\left(1+5^{1/3}\right) = 2.71p$, or $p < \dfrac{1}{2.71}$.

When $p \le 0.1$, both these conditions are satisfied so <u>choose A</u>.

[The probabilities are in the ratio $A{:}R{:}S \equiv 0.729 : 0.036 : 0.005$.]

<u>Higher Certificate, Paper I, 2002. Question 6</u>

$$f(x) = \lambda e^{-\lambda x} \qquad x > 0,\ \lambda > 0$$

(i) $\quad M_X(t) = E\left[e^{tX}\right] = \int_0^\infty \lambda e^{-\lambda x + tx} dx = \int_0^\infty \lambda e^{-(\lambda - t)x} dx$

$$= \left[-\frac{\lambda e^{-(\lambda-t)x}}{\lambda - t}\right]_0^\infty = \frac{1}{\lambda - t} = \left(1 - \frac{t}{\lambda}\right)^{-1} \qquad [|t| < \lambda].$$

$M_x(t) = 1 + \frac{t}{\lambda} + \frac{t^2}{\lambda^2} + \ldots$.

$E[X^k]$ = coefficient of $\frac{x^k}{k!}$ in the expansion.

Hence $E[X] = \frac{1}{\lambda}$; also, $E\left[X^2\right] = \frac{2}{\lambda^2}$, so $\mathrm{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$.

(ii) $\quad L(\lambda \mid x_1, \ldots, x_n) = \prod_{i=1}^n \left(\lambda e^{-\lambda x_i}\right) = \lambda^n \exp\left(-\lambda \sum_{i=1}^n x_i\right)$.

$\ln L = n \ln \lambda - \lambda \sum x_i = n\left(\ln \lambda - \lambda \bar{x}\right)$

$\frac{d(\ln L)}{d\lambda} = \frac{n}{\lambda} - n\bar{x} = 0$ for $\hat{\lambda} = \frac{1}{\bar{x}}$.

$\frac{d^2(\ln L)}{d\lambda^2} = -\frac{n}{\lambda^2}$ which confirms maximum of L.

(iii) The asymptotic variance of $\hat{\lambda}$ [Cramér-Rao lower bound for variance] is $\frac{1}{E\left[-\frac{d^2(\ln L)}{d\lambda^2}\right]} = \frac{\lambda^2}{n}$.

We have $\hat{\lambda} \sim$ approx $N\left(\lambda, \frac{\lambda^2}{n}\right)$, i.e. the estimate of $SE\left(\hat{\lambda}\right)$ is $\frac{\hat{\lambda}}{\sqrt{n}}$, so that $\hat{\lambda} - 1.96\frac{\hat{\lambda}}{\sqrt{n}} < \lambda < \hat{\lambda} + 1.96\frac{\hat{\lambda}}{\sqrt{n}}$ is an approximate 95% confidence interval for λ when n is large.

This is $\frac{1}{\bar{x}} - \frac{1.96}{\bar{x}\sqrt{n}} < \lambda < \frac{1}{\bar{x}} + \frac{1.96}{\bar{x}\sqrt{n}}$.

<u>Higher Certificate, Paper I, 2002. Question 7</u>

		Y			
		0	1	2	3
X	0	k	$6k$	$9k$	$4k$
	1	$8k$	$18k$	$12k$	$2k$
	2	k	$6k$	$9k$	$4k$

(i) The sum of all the entries in the table is $80k$. Hence $k = \frac{1}{80}$.

(ii) Row and column totals give the marginal distributions of X and Y:

X	0	1	2
$P(X)$	1/4	1/2	1/4

Y	0	1	2	3
$P(Y)$	1/8	3/8	3/8	1/8

(iii) For $P(X = x \mid Y = 2)$, use $\dfrac{P(X = x \text{ and } Y = 2)}{P(Y = 2)}$:

X	0	1	2
Probability	$\dfrac{9/80}{3/8}$ = 9/30 = 0.3	$\dfrac{12/80}{3/8}$ = 0.4	$\dfrac{9/80}{3/8}$ = 0.3

(iv) $E[X] = 1$, $E[Y] = 1.5$, by symmetry. The distribution of XY is:

XY	0	1	2	3	4	6
$P(XY)$	29/80	18/80	18/80	2/80	9/80	4/80

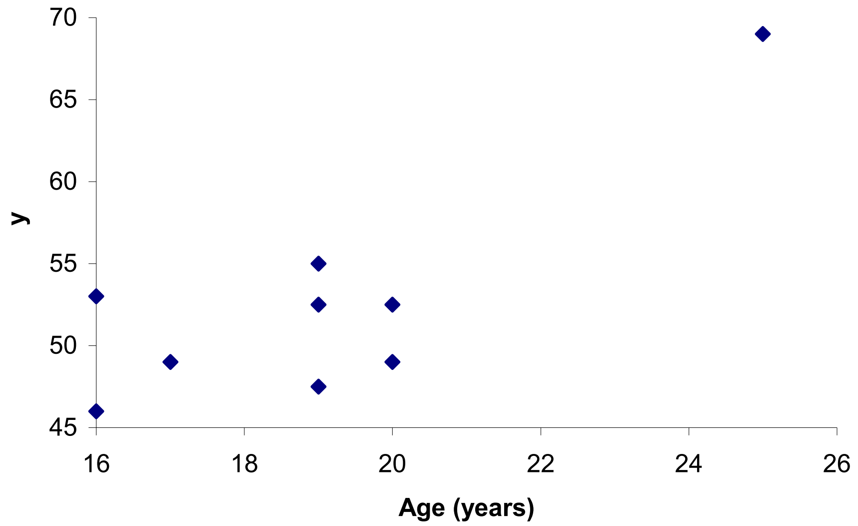
$E[XY] = \dfrac{120}{80} = 1.5$.

So $E[XY] = E[X]E[Y]$. Therefore $\text{Cov}(X,Y) = 0$, so correlation = 0.

(iv) Zero correlation is not sufficient. <u>Every</u> individual $P(X = x, Y = y)$ in the table must be the product of its two marginal probabilities.

Consider $x = y = 0$. We have $P(0,0) = k = \frac{1}{80}$. But $P_X(0)P_Y(0) = \frac{1}{4} \times \frac{1}{8} = \frac{1}{32}$. So there is <u>not</u> independence.

Higher Certificate, Paper I, 2002. Question 8



There appears to be an outlier (25, 69), without which correlation (and the regression gradient) would be near 0. (Rank correlation may be more appropriate.)

$\sum x = 171 \qquad \sum y = 477 \qquad n = 9 \qquad \bar{x} = 19 \qquad \bar{y} = 53$

$S_{XX} = 3309 - \frac{171^2}{9} = 60$. $\quad S_{YY} = 25633 - \frac{477^2}{9} = 352$. $\quad S_{XY} = 9183 - \frac{171 \times 477}{9} = 120$.

$y - \bar{y} = \hat{b}(x - \bar{x})$, and $\hat{b} = \frac{S_{XY}}{S_{XX}} = \frac{120}{60} = 2$.

So we have $y - 53 = 2(x - 19)$ or $\underline{y = 2x + 15}$

(a) For $x = 18$, $\underline{\hat{y} = 51}$. (b) For $x = 26$, $\underline{\hat{y} = 67}$.

Including the data for *I*, correlation $r_{xy} = \sqrt{\frac{{S_{XY}}^2}{S_{XX}S_{YY}}}$ = <u>0.826</u>. (Significant at 5% level.)

[Hence the null hypothesis "b = 0" is rejected.]

However, <u>without *I*</u>, $\sum x = 146$, $\bar{x} = 18.25$; $\sum y = 408$, $\bar{y} = 51.00$.

$\sum x^2 = 3309 - 25^2 = 2684$; $\sum y^2 = 25633 - 69^2 = 20872$;

$\sum xy = 9183 - (25 \times 69) = 7458$; $\hat{b} = \frac{7458 - (146 \times 408)/8}{2684 - 146^2/8} = \frac{12}{19.5} = 0.615$.

$S_{XX} = 19.5$, $S_{XY} = 12$, $S_{YY} = 20872 - \frac{408^2}{8} = 64$.

$\therefore r_{XY} = \frac{12}{\sqrt{64 \times 19.5}}$ = <u>0.340</u> which does not approach significance. The line still has positive gradient, but not significantly different from 0.