THE ROYAL STATISTICAL SOCIETY

2004 EXAMINATIONS – SOLUTIONS

GRADUATE DIPLOMA

PAPER I – STATISTICAL THEORY & METHODS

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.

Graduate Diploma, Statistical Theory & Methods, Paper I, 2004. Question 1

X and Y are independent, each with probability function given by $P(W = w) = (1 - \theta)^{w-1}\theta$, for $w = 1, 2, \ldots$ and where $0 < \theta < 1$.

(i) Let $U = X + Y$. U may take the values 2, 3, … . By the independence of X and Y,

$$P(U=u)=\sum_{x=1}^{u-1}P(X=x \text{ and } Y=u-x) \qquad \text{for } u = 2, 3, \ldots .$$

Hence $P(U=u)=\sum_{x=1}^{u-1}\left\{(1-\theta)^{x-1}\theta\right\}\left\{(1-\theta)^{u-x-1}\theta\right\}=\sum_{x=1}^{u-1}\theta^2(1-\theta)^{u-2}=(u-1)\theta^2(1-\theta)^{u-2}$,

for $u = 2, 3, \ldots$ [this is a negative binomial distribution].

(ii) If $U\ (= X + Y) = u$, X must take values in the range $1, 2, \ldots, u - 1$. Then

$$P(X=x|U=u)=\frac{P(X=x \text{ and } Y=u-x)}{P(U=u)}$$

$$=\frac{\left\{(1-\theta)^{x-1}\theta\right\}\left\{(1-\theta)^{u-x-1}\theta\right\}}{(u-1)\theta^2(1-\theta)^{u-2}}=\frac{1}{u-1} \quad (\text{for } x = 1, 2, \ldots, u-1).$$

This is a discrete uniform distribution on the integers $1, 2, \ldots, u - 1$.

(iii) Suppose Andy takes X attempts and Bob takes Y attempts. Then X and Y are independent geometric random variables with $\theta = 0.4$.

Thus from part (i), $P(X + Y = 6) = (6 - 1)(0.4)^2(0.6)^4 = 0.1037$.

From part (ii), $P(X < Y \mid X + Y = 6) = P(X = 1 \text{ or } 2 \mid X + Y = 6)$

$$=\frac{1}{6-1}+\frac{1}{6-1} = \frac{2}{5}.$$

Graduate Diploma, Statistical Theory & Methods, Paper I, 2004. Question 2

(i) (a) We have $P(U = u) = 1/m$, for $u = 1, 2, \ldots, m$.

So $E[U] = \sum uP(u) = \frac{1}{m}(1+2+\ldots+m) = \frac{1}{m}\frac{m(m+1)}{2} = \frac{m+1}{2}$.

Also,

$$E\left[U^2\right] = \sum u^2 P(u) = \frac{1}{m}(1^2+2^2+\ldots+m^2) = \frac{1}{m}\frac{m(m+1)(2m+1)}{6} = \frac{(m+1)(2m+1)}{6}$$

and therefore $\mathrm{Var}(U) = E[U^2] - \{E[U]\}^2$

$$= \frac{(m+1)(2m+1)}{6} - \left(\frac{m+1}{2}\right)^2 = \frac{m+1}{12}(4m+2-3m-3) = \frac{(m+1)(m-1)}{12} = \frac{m^2-1}{12}.$$

(b) $V = U + k$, so $E[V] = \frac{m+1}{2} + k$ and $\mathrm{Var}(V) = \frac{m^2-1}{12}$.

(ii) **Given that $Y = y$**, the first $y - 1$ rolls must have been 6s and the final roll not a 6.

Thus X is a discrete uniform random variable with, using the notation of part (i)(b), $k = 6(y-1)$ and $m = 5$. Hence $E[X] = \frac{5+1}{2} + 6(y-1) = 6y - 3$ and $\mathrm{Var}(X) = \frac{5^2-1}{12} = 2$.

These are the values **conditional on $Y = y$**.

Now, Y is a geometric random variable with $E[Y] = \frac{1}{5/6} = \frac{6}{5}$ and $\mathrm{Var}(Y) = \frac{1-\frac{5}{6}}{\left(\frac{5}{6}\right)^2} = \frac{6}{25}$.

$$\therefore E[X] = E\left[E(X|Y)\right] = E[6Y-3] = 6E[Y]-3 = \frac{36}{5}-3 = \frac{21}{5}.$$

We also need

$$E\left[\mathrm{Var}(X|Y)\right] = E[2] = 2 \text{ and } \mathrm{Var}\left(E\left[X|Y\right]\right) = \mathrm{Var}(6Y-3) = 36\,\mathrm{Var}(Y) = 216/25,$$

from which (using the result quoted in the question)

$$\mathrm{Var}(X) = 2 + \frac{216}{25} = \frac{266}{25}.$$

<u>Graduate Diploma, Statistical Theory & Methods, Paper I, 2004. Question 3</u>

(i) $$E\left[U^m\right] = \int_0^\infty u^m \frac{\theta^\alpha u^{\alpha-1}e^{-\theta u}}{\Gamma(\alpha)}du = \frac{\theta^\alpha}{\Gamma(\alpha)}\int_0^\infty u^{\alpha+m-1}e^{-\theta u}du \qquad [\text{put } t = \theta u]$$

$$= \frac{\theta^\alpha}{\Gamma(\alpha)}\frac{1}{\theta^{\alpha+m}}\int_0^\infty t^{\alpha+m-1}e^{-t}dt = \frac{\theta^\alpha}{\Gamma(a)}\frac{\Gamma(\alpha+m)}{\theta^{\alpha+m}} .$$

Hence $E[U]=\frac{\alpha}{\theta}$. Also $E\left[U^2\right]=\frac{\alpha(\alpha+1)}{\theta^2}$ and so Var(U) = $E[U^2] - \{E[U]\}^2 = \frac{\alpha}{\theta^2}$.

(ii) $$f_X(x)=\int_{y=x}^\infty 8xe^{-2y}dy = 8x\left[\frac{e^{-2y}}{-2}\right]_x^\infty = 4xe^{-2x} \qquad (\text{for } x > 0).$$

This is gamma with $\alpha = 2$ and $\theta = 2$.

$$f_Y(y)=\int_0^y 8xe^{-2y}dx = 8e^{-2y}\left[\tfrac{1}{2}x^2\right]_0^y = 4y^2e^{-2y} \qquad (\text{for } y > 0).$$

This is gamma with $\alpha = 3$ and $\theta = 2$.

Therefore from part (i), $E[X]$ =1 and Var(X) = 1/2, and $E[Y]$ = 3/2 and Var(Y) = 3/4.

$$E[XY]=\int_{y=0}^\infty\int_{x=0}^y xy.8xe^{-2y}dxdy = \int_0^\infty ye^{-2y}\left(\int_0^y 8x^2dx\right)dy = \frac{8}{3}\int_0^\infty y^4e^{-2y}dy \qquad [\text{put } t = 2y]$$

$$= \frac{8}{3}\int_0^\infty \frac{t^4}{16}e^{-t}\frac{dt}{2} = \frac{1}{12}\Gamma(5) = \frac{4!}{12} = 2 .$$

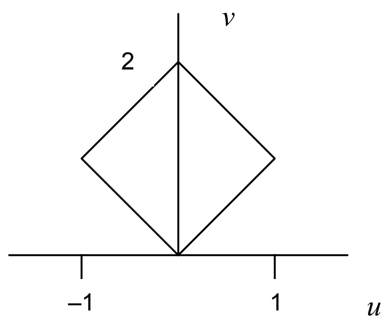
$\therefore \text{Cov}(X,Y) = 2 - E(X)E(Y) = 2-\frac{3}{2}=\frac{1}{2}$, and the correlation between X and Y is

$$\frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} = \frac{\frac{1}{2}}{\sqrt{\frac{1}{2}\cdot\frac{3}{4}}} = \sqrt{\frac{2}{3}}$$

Graduate Diploma, Statistical Theory & Methods, Paper I, 2004. Question 4

X and Y are uniform(0, 1). $U = X - Y$, $V = X + Y$. So $X = ½(U + V)$ and $Y = ½(V - U)$.

(i) The limits of U are (–1, 1) and of V are (0, 2). The diagram shows the square region where $f(u, v) \neq 0$.



The Jacobian of the transformation is $J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{vmatrix} = \frac{1}{2}$.

Therefore, for u, v in the region shown, and using the independence of X and Y,

$$f(u,v) = |J| f(x,y) = \tfrac{1}{2} f(x) f(y) = \tfrac{1}{2}.1.1 = \tfrac{1}{2} .$$

(ii) For $-1 \leq u \leq 0$, $f(u) = \frac{1}{2}\int_{-u}^{2+u} dv = 1 + u$.

For $0 \leq u \leq 1$, $f(u) = \frac{1}{2}\int_{u}^{2-u} dv = 1 - u$.

For $0 \leq v \leq 1$, $f(v) = \frac{1}{2}\int_{-v}^{v} du = v$.

For $1 \leq v \leq 2$, $f(v) = \frac{1}{2}\int_{v-2}^{2-v} du = 2 - v$.

($f(u)$ and $f(v)$ are zero elsewhere.) These distributions are identical except for a shift of one unit in the x-direction.

(iii) We now have $W = a + tX$ and $Z = b + tY$ ($t > 0$). So $W + Z = (a + b) + t(X + Y)$ $= (a + b) + tV$ where V is as in parts (i) and (ii). So, using the result for V in part (ii),

$$P\left(W + Z < a + b + \frac{t}{2}\right) = P\left(tV < \frac{t}{2}\right) = P\left(V < \frac{1}{2}\right) = \int_0^{1/2} v dv = \left[\tfrac{1}{2}v^2\right]_0^{1/2} = \frac{1}{8} .$$

Graduate Diploma, Statistical Theory & Methods, Paper I, 2004. Question 5

(i) The method is to complete the square in the exponent in the integral and then recognise that this re-creates the pdf of a Normal distribution so that the value of the integral is simply 1.

$M_X(t) = E\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} . \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} dx$. Thus the exponent is

$$tx - \frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 = -\frac{1}{2\sigma^2}\left(x^2 - 2\mu x + \mu^2 - 2\sigma^2 tx\right) = -\frac{1}{2\sigma^2}\left\{\left[x - \left(\mu + \sigma^2 t\right)\right]^2 - 2\mu\sigma^2 t - \sigma^4 t^2\right\}$$

Hence $M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x - \left[\mu + \sigma^2 t\right]}{\sigma}\right)^2\right\} dx = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right)$.

(ii) $M_{aX+b}(t) = E\left[e^{t(aX+b)}\right] = e^{bt} E\left[e^{(at)X}\right] = e^{bt} M_X(at)$.

Taking $a = 1/\sigma$ and $b = \mu/\sigma$ we have Z, and so

$$M_Z(t) = e^{-\mu t/\sigma} \exp\left(\frac{\mu t}{\sigma} + \frac{\sigma^2 t^2}{2\sigma^2}\right) = e^{t^2/2} .$$

Putting $\mu = 0$ and $\sigma = 1$ in the result of part (i), we see that this is the moment generating function of the standard Normal distribution. Therefore (assuming without proof the uniqueness property between distributions and their moment generating functions) Z follows the standard Normal distribution.

(iii) We expand $M_Z(t)$ as a power series in t and use the result that the rth moment of Z about the origin, $E[Z^r]$, is the coefficient of $t^r/r!$ in this expansion. [Alternatively, use the result that it is given by the rth derivative evaluated at $t = 0$.]

$$M_Z(t) = 1 + \tfrac{1}{2}t^2 + \frac{\left(\tfrac{1}{2}t^2\right)^2}{2!} + \frac{\left(\tfrac{1}{2}t^2\right)^3}{3!} + \ldots = 1 + \frac{t^2}{2} + \frac{t^4}{8} + \frac{t^6}{48} + \ldots .$$

We need $E[Z^2]$ and $E[Z^4]$, which we see are 1 and 3 respectively.

So $\mathrm{Var}(Z^2) = E[Z^4] - \{E[Z^2]\}^2 = 3 - 1 = 2$.

Graduate Diploma, Statistical Theory & Methods, Paper I, 2004. Question 6

(i) $F(x) = P(X \le x) = P(X \le x | \text{boy})P(\text{boy}) + P(X \le x | \text{girl})P(\text{girl})$

$$= \frac{1}{2}\{F_1(x) + F_2(x)\} .$$

Hence the pdf of X is $f(x) = \frac{d}{dx}F(x) = \frac{1}{2}\{f_1(x) + f_2(x)\}$.

(ii) $E[X] = \int x f(x)\,dx = \frac{1}{2}\left(\int x f_1(x)\,dx + \int x f_2(x)\,dx\right) = \frac{1}{2}(\mu_1 + \mu_2)$.

$E[X^2] = \int x^2 f(x)\,dx = \frac{1}{2}\left(\int x^2 f_1(x)\,dx + \int x^2 f_2(x)\,dx\right)$ [now use $\sigma^2 = E[X^2] - \{E[X]\}^2$]

$$= \frac{1}{2}\left\{\left(\sigma^2 + \mu_1^{\ 2}\right) + \left(\sigma^2 + \mu_2^{\ 2}\right)\right\} = \sigma^2 + \frac{1}{2}\left(\mu_1^{\ 2} + \mu_2^{\ 2}\right)$$

and thus $\text{Var}(X) = \sigma^2 + \frac{1}{2}\left(\mu_1^{\ 2} + \mu_2^{\ 2}\right) - \frac{1}{4}(\mu_1 + \mu_2)^2 = \sigma^2 + \frac{1}{4}\left(\mu_1^{\ 2} + \mu_2^{\ 2}\right) - \frac{1}{2}\mu_1\mu_2$

$$= \sigma^2 + \frac{1}{4}(\mu_1 - \mu_2)^2 .$$

(iii) As the sample is large, we may use the central limit theorem to say that the distribution of $\bar{X}_r$ is approximately Normal, whatever the distributions f_1 and f_2.

The mean of $\bar{X}_r$ is equal to the mean of X, i.e. $\frac{1}{2}(\mu_1 + \mu_2)$.

$$\text{Var}\left(\bar{X}_r\right) = \frac{1}{2n}\text{Var}(X) = \frac{1}{2n}\left\{\sigma^2 + \frac{1}{4}(\mu_1 - \mu_2)^2\right\} .$$

(iv) Let the (independent) sample means for boys and girls be $\bar{X}_1$ and $\bar{X}_2$. Their means are μ_1 and μ_2, and their variances are each σ^2/n. So $\bar{X}_{st} = \frac{1}{2}\left(\bar{X}_1 + \bar{X}_2\right)$ has mean $\frac{1}{2}(\mu_1 + \mu_2)$ and variance $\frac{1}{4}\left(\frac{\sigma^2}{n} + \frac{\sigma^2}{n}\right) = \frac{\sigma^2}{2n}$. Thus both $\bar{X}_r$ and $\bar{X}_{st}$ are unbiased but the latter has smaller variance; it is more precise.

Graduate Diploma, Statistical Theory & Methods, Paper I, 2004. Question 7

(i) (a) The distribution of X is

x	0	1	2	3
$P(x)$	64/125	48/125	12/125	1/125
cdf $F(x)$	0.512	0.896	0.992	1.000

$0.3612 < 0.512$ and so corresponds to $x = 0$.

0.6789 is between 0.512 and 0.896 and so corresponds to $x = 1$.

$0.3552 < 0.512$ and so corresponds to $x = 0$.

$0.2898 < 0.512$ and so corresponds to $x = 0$.

So the sample is 0, 1, 0, 0.

(b) $F(x) = \int_0^x 2t^3 e^{-t^4/2} dt = \left[-e^{-t^4/2}\right]_0^x = 1 - e^{-x^4/2}$. Thus the inverse cdf method gives $u = 1 - e^{-x^4/2}$, where u is the random number, so $x = \left[-2\log(1-u)\right]^{1/4}$.

u	$1 - u$	$-2\log(1 - u)$	$[-2\log(1 - u)]^{1/4}$
0.3612	0.6388	0.8963	0.9730
0.6789	0.3211	2.2720	1.2277
0.3552	0.6448	0.8776	0.9679
0.2898	0.7102	0.6844	0.9096

↑

These are the required random numbers

(ii) Inter-arrival times, X, are exponential with parameter ½. Given u (as above), the inverse cdf method leads to $u = 1 - e^{-x/2}$, i.e. $x = -2\log(1 - u)$ [as is tabulated in column 3 above].

First arrival time is at 0.8963. Service starts immediately and lasts (uniform distribution on (1.5, 2.5)) 1.5 + 0.6789 = 2.1789, and thus ends at 0.8963 + 2.1789 = 3.0752.

Second arrival time is at 0.8963 + 0.8776 = 1.7739. Service cannot start until 3.0752 and then lasts 1.5 + 0.2898 = 1.7898, and thus ends at 3.0752 + 1.7898 = 4.8650.

Expressing the simulated results to the nearest 0.1 minute after 9.00 a.m., the first customer arrives at 0.9 and leaves at 3.1, the second arrives at 1.8 and leaves at 4.9.

Graduate Diploma, Statistical Theory & Methods, Paper I, 2004. Question 8

(i) Consider the number of balls in urn A, an integer from 0 to M. Define the transition probability $p_{ij} = P\left(j \text{ balls in urn A at step } n \middle| i \text{ balls in urn A at step } n-1 \right)$.

Then we have

$p_{01} = 1, \quad p_{0j} = 0$ for $j \neq 1$;

for $i = 1, 2, ..., M-1, \quad p_{i\,i-1} = \dfrac{i}{M}, \quad p_{i\,i+1} = \dfrac{M-i}{M}, \quad p_{ij} = 0$ for $j \neq i-1$ or $i+1$;

$p_{M\,M-1} = 1, \quad p_{Mj} = 0$ for $j \neq M-1$.

(ii) Let the stationary probabilities be $\pi_1, \pi_2, ..., \pi_M$. We have $\pi_j = \sum_{i=0}^{M} p_{ij}\pi_i$. Hence

$$\pi_0 = \frac{1}{M}\pi_1; \quad \text{for } j = 1, 2, ..., M-1, \quad \pi_j = \frac{M-j+1}{M}\pi_{j-1} + \frac{j+1}{M}\pi_{j+1}; \quad \pi_M = \frac{1}{M}\pi_{M-1}\ .$$

It is required to show that $\pi_j = \binom{M}{j}\left(\frac{1}{2}\right)^M$ ($j = 0, 1, \ldots, M$) satisfy these equations.

Immediately, $\pi_0 = \pi_M = (1/2)^M$ and $\pi_1 = \pi_{M-1} = M(1/2)^M$.

Thus $\pi_0 = \dfrac{1}{M}\pi_1$ and $\pi_M = \dfrac{1}{M}\pi_{M-1}$, as required.

To confirm that the general equation is satisfied, consider

$$\frac{M-j+1}{M}\binom{M}{j-1} + \frac{j+1}{M}\binom{M}{j+1} = \frac{(M-j+1)M!}{M(j-1)!(M-j+1)!} + \frac{(j+1)M!}{M(j+1)!(M-j-1)!}$$

$$= \frac{(M-1)!}{(j-1)!(M-j)!} + \frac{(M-1)!}{j!(M-j-1)!} = \frac{(M-1)!}{j!(M-j)!}(j+M-j) = \frac{M!}{j!(M-j)!} = \binom{M}{j}\ .$$

Multiplying by $(1/2)^M$, we see that the general equation ($j = 1, 2, \ldots, M-1$) is satisfied.

(iii) Let X be the number of balls in urn A. The given probabilities form a binomial distribution with parameters (60, ½). So $E(X) = 60 \times ½ = 30$ and $\text{Var}(X) = 60 \times ½ \times ½ = 15$. $P(X = 34)$ can be approximated by $P(33.5 < \text{N}(30, 15) < 34.5)$. Standardising to N(0, 1) gives $P(0.9037 < \text{N}(0, 1) < 1.1619) = 0.8774 - 0.8169 = 0.0605$.