THE ROYAL STATISTICAL SOCIETY

2007 EXAMINATIONS – SOLUTIONS

HIGHER CERTIFICATE

(MODULAR FORMAT)

MODULE 2

PROBABILITY MODELS

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.

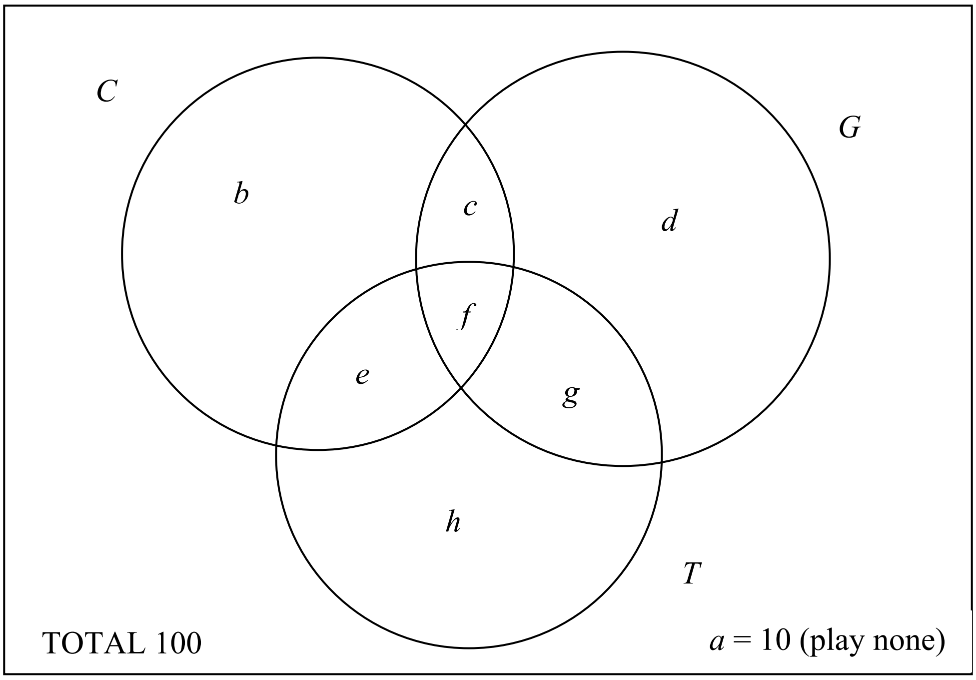
Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

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, Module 2, 2007. Question 1



From the information given, we have the following.

$f = 5$.

$b + c + d + e + f + g + h = 100 - a = 100 - 10 = 90$.

$b + c + e + f + g + h = 88$, But this equals $90 - d$. $\therefore d = 2$.

$b + c + d + e + f + g = 78$. But this equals $90 - h$. $\therefore h = 12$.

$g = 30$.

$c + d + f + g = 38$. $\therefore c = 38 - 2 - 5 - 30 = 1$.

$e + f + g + h = 74$. $\therefore e = 74 - 5 - 30 - 12 = 27$.

(i) At least one sport: $100 - a = 90$.

(ii) Exactly one sport: $b + d + h = 90 - (c + e + f + g) = 90 - 63 = 27$.

(iii) Exactly two sports: $c + e + g = 1 + 27 + 30 = 58$.

(iv) Number not playing golf is

$$a + b + \mathrm{e} + h = a + 90 - (c + d + f + g) = 100 - (1 + 2 + 5 + 30) = 62$$

$\therefore$ Required proportion is $\frac{b+e}{62} = \frac{(27-d-h)+e}{62} = \frac{13+27}{62} = \frac{40}{62} = 0.645$.

(v)

Number of sports played:	0	1	2	3	TOTAL
Number of men:	10	27	58	5	100

$\therefore$ Mean number of sports played = (0 + 27 + 116 + 15)/100 = 1.58.

Higher Certificate, Module 2, 2007. Question 2

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

(i) $$E(X) = \int_0^\infty \lambda x e^{-\lambda x} dx = \left[-xe^{-\lambda x}\right]_0^\infty - \int_0^\infty (-e^{-\lambda x}) dx = 0 + \left[\frac{e^{-\lambda x}}{-\lambda}\right]_0^\infty = \frac{1}{\lambda}.$$

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

$$= 0 + \frac{2}{\lambda} E(X) = \frac{2}{\lambda^2}.$$

Hence $\mathrm{Var}(X) = E(X^2) - [E(X)]^2 = \dfrac{2}{\lambda^2} - \dfrac{1}{\lambda^2} = \dfrac{1}{\lambda^2}$, and $\mathrm{SD}(X) = \dfrac{1}{\lambda}$.

(ii) $$P(X > c) = \int_c^\infty \lambda e^{-\lambda x} dx = \left[-e^{-\lambda x}\right]_c^\infty = e^{-\lambda c}.$$

For $x > c$, $$P(X > x \mid X > c) = \frac{P\{(X > x) \cap (X > c)\}}{P(X > c)} = \frac{P(X > x)}{P(X > c)} = \frac{e^{-\lambda x}}{e^{-\lambda c}}$$

$$= e^{-\lambda(x-c)}, \text{ as required.}$$

Thus the conditional cdf of X, given that $X > c$, is $1 - e^{-\lambda(x-c)}$, and by differentiation we get that the conditional pdf is $\lambda e^{-\lambda(x-c)}$.

Therefore $X - c$ has an exponential distribution with parameter λ or, putting it another way, X has the exponential distribution but with the origin shifted to c.

(iii) The sample mean is 2.0 and the sample variance is $\dfrac{1}{9}\left(74.38 - \dfrac{20.0^2}{10}\right) = 3.82$, so the sample standard deviation is $\sqrt{3.82} = 1.95$.

The sample mean and standard deviation are very nearly equal. This supports the exponential model, as the exponential distribution has equal mean and standard deviation (see part (i)).

Higher Certificate, Module 2, 2007. Question 3

(i) (a) Suppose the coin is tossed n times and there are x heads and therefore $(n - x)$ tails. The number of possible orders in which this can happen is $\binom{n}{x}$ or $\frac{n!}{x!(n-x)!}$. The probability of each of these orders for independent tosses is $p^x \times (1-p)^{n-x}$, so the required overall probability is $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}$, for $x = 0, 1, \ldots, n$.

A Normal approximation with the same mean and variance as this binomial distribution, i.e. $N(np, np(1 - p))$, is satisfactory when n is fairly large and p is not near 0 or 1. As a guideline, the value of np (or of $n(1 - p)$ if p is near 1) should be at least 5. A continuity correction should be used.

(b) For $n = 20$ and $p = 0.2$, the Normal approximation is N(4, 3.2). Using a continuity correction, $P(X \leq 3)$ is approximated by the area under the pdf of the Normal distribution up to 3.5. Thus the required probability is

$$\Phi\left(\frac{3.5-4}{\sqrt{3.2}}\right) = \Phi(-0.2795) = 1 - 0.610 = 0.390.$$

The Society's "Statistical tables for use in examinations" give the exact probability from B(20, 0.2) as $P(X \leq 3) = 0.411$. Hence the percentage error is $100(0.390 - 0.411)/0.411$ which is 5.11% (in the negative direction). This is not very satisfactory. Here, n is not large and p is fairly small; the guideline $np \geq 5$ is not satisfied. The binomial distribution is not well approximated by the Normal distribution (it will be noticeably positively skew as $p < ½$).

(ii) There must be a head at the final toss and $(x - 1)$ heads in the other $(n - 1)$ tosses. So the probability is

$$P(N = n) = p \times \binom{n-1}{x-1} p^{x-1} (1-p)^{(n-1)-(x-1)} = \binom{n-1}{x-1} p^x (1-p)^{n-x},$$

for $n = x, x + 1, x + 2, \ldots$. For $p = 0.2$, $x = 3$, $n = 20$, this gives

$$\frac{19!}{2!17!}(0.2)^3(0.8)^{17} = \frac{19.18}{2.1}(0.2)^3(0.8)^{17} = 0.0308.$$

From the tables, $P(X = 3)$ for the B(20, 0.2) distribution is $0.4114 - 0.2061 = 0.2053$. The previous probability is x/n times this.

Higher Certificate, Module 2, 2007. Question 4

(i) Since $A_1, A_2, \ldots, A_k$ are mutually exclusive, $P(B)=\sum_{j=1}^{k} P(B|A_j)P(A_j)$.

$$\therefore P(A_i|B)=\frac{P(A_i \cap B)}{P(B)}=\frac{P(B|A_i)P(A_i)}{P(B)}=\frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{k} P(B|A_j)P(A_j)}.$$

(ii) $P(\mathrm{P})$ = 1/6, $P(\mathrm{Q})$ = 1/3, $P(\mathrm{R})$ = 1/2. P, Q, and R are mutually exclusive and exhaustive purchases.

We have Poisson distributions with $\lambda_{\mathrm{P}}=3,\ \lambda_{\mathrm{Q}}=2,\ \lambda_{\mathrm{R}}=1$.

$$P(2\text{ flaws}|\mathrm{Q})=\frac{e^{-2}2^2}{2!}=2e^{-2} \quad \text{and} \quad P(2\text{ flaws}|\mathrm{R})=\frac{e^{-1}.1^2}{2!}=\frac{1}{2}e^{-1}.$$

We have $$P(\mathrm{Q}|2\text{ flaws})=\frac{P(2\text{ flaws}|\mathrm{Q})P(\mathrm{Q})}{\sum_{i=\mathrm{P,Q,R}} P(2\text{ flaws}|i)P(i)}$$

and similarly $$P(\mathrm{R}|2\text{ flaws})=\frac{P(2\text{ flaws}|\mathrm{R})P(\mathrm{R})}{\sum_{i=\mathrm{P,Q,R}} P(2\text{ flaws}|i)P(i)}.$$

So the ratio $$\frac{P(\mathrm{Q}|2\text{ flaws})}{P(\mathrm{R}|2\text{ flaws})}=\frac{P(2\text{ flaws}|\mathrm{Q})P(\mathrm{Q})}{P(2\text{ flaws}|\mathrm{R})P(\mathrm{R})}=\frac{2e^{-2}\times\frac{1}{3}}{\frac{1}{2}e^{-1}\times\frac{1}{2}}=\frac{8}{3e}=0.981.$$

so it is (very slightly) more likely to have come from R.

Solution continued on next page

(iii) $\{P(2 \text{ flaws}|\text{Q})\}^2 = (2e^{-2})^2$ is the probability that two (independent) lengths from Q are faulty, and similarly the corresponding probability for R is $\{P(2 \text{ flaws}|\text{R})\}^2 = \left(\frac{1}{2}e^{-1}\right)^2$.

Now we need the ratio

$$\frac{\{P(2 \text{ flaws}|\,Q)\}^2\,P(\text{both } Q|\text{ both the same})}{\{P(2 \text{ flaws}|\,R)\}^2\,P(\text{both } R|\text{ both the same})}$$

$$= \frac{4e^{-4} \times \dfrac{\frac{1}{9}}{\frac{1}{36}+\frac{1}{9}+\frac{1}{4}}}{\dfrac{1}{4}e^{-2} \times \dfrac{\frac{1}{4}}{\frac{1}{36}+\frac{1}{9}+\frac{1}{4}}} = \frac{4 \times 4 \times \dfrac{2}{7}}{e^2 \times \dfrac{9}{14}} = \frac{64}{9e^2} = 0.9624\,.$$

Again supplier R is slightly more likely to be the source than supplier Q in these circumstances. In fact R is slightly more likely to be the supplier in this case than in the case of part (ii).

These results might be counter-intuitive in that Q's material is more faulty than R's. However, this has to be balanced against the fact that more material comes from R than from Q.