UNIVERSITY OF LONDON

[C245 2003]

B.ENG. AND M.ENG. EXAMINATIONS 2003

For Internal Students of Imperial College London

This paper is also taken for the relevant examination for the Associateship of the City & Guilds of London Institute.

COMPUTING C245

STATISTICS

Date Wednesday 14th May 2003 10.00 - 11.30 am

Answer THREE questions

[Before starting, please make sure that the paper is complete. There should be a total of FOUR questions. Ask the invigilator for a replacement if this copy is faulty.]

1. (i) I find that, on average, I am unable to log on to a particular web site once in every 5 attempts.

   (a) Assuming the successes or failures on different attempts are independent, what is the probability that I will first succeed at the 4th attempt?

   (b) On average, how many failures will I make before I first succeed?

   (c) What is the probability that I will have 3 or more failures before succeeding?

   (ii) For a different web site, I find that the variance of the number of failed attempts I make before I manage to log on successfully is 2. Use this information to calculate the probability that I will successfully log on at any attempt, and hence calculate the probability that my first successful attempt will be the fifth.

   (iii) A large organisation buys 200 PCs from manufacturer A and 1800 PCs from manufacturer B. It is known, from past experience, that 5% of those from A will be expected to fail within the first month of use. After the machines have been running for a month, the records are examined and it is seen that 50% of those that failed were provided by manufacturer A. Using these figures,

   (a) Fill in the numbers of machines at each X in the following cross-classification table

	A	B	Total
Not fail	X	X	X
Fail	X	X	X
Total	X	X	X

   (b) What is the joint probability that a randomly selected machine is from manufacturer A and does not fail?

   (c) Create a table showing the conditional probabilities of failing and not failing for each manufacturer, and the probabilities of failing and not failing for all machines.

   (d) Show mathematically why the overall probability of a machine failing is not simply the average of the conditional probabilities of failure from manufacturer A and manufacturer B.

2. A probability density function on two variables is defined by

$$f(x,\,y) \;=\; \begin{cases} k(x+y^2), & \text{when } 0<x<1, \quad 0<y<1, \\ 0, & \text{otherwise}\,, \end{cases}$$

for some value of k.

(i) Find the value of k which makes this a legitimate probability density function.

(ii) Find the marginal density functions $f_X(x)$ and $f_Y(y)$.

(iii) Determine whether the random variables X and Y are independent or not.

(iv) Find the conditional density of X when $Y = 1/2$ and evaluate the mean of X, subject to the condition $Y = 1/2$.

3. (i) Observation shows that the distribution of lifetimes of the PCs in a computer laboratory has a survivor function $R(t) = e^{-\lambda t}$, $t > 0$.

(a) Find the probability density function, $f(t)$, for the distribution of lifetimes.

(b) Find the hazard function, $r(t)$, for this distribution.

(c) Find the probability density function of further life for those components which have already survived up until time s.

(d) Using the result in (c), find the probability density function of total life for those components which have already survived up until time s.

(ii) A system of three components is arranged such that it will only fail if one or both of C1 and C2 fails at the same time as C3 fails.

(a) Draw a diagram illustrating the series/parallel configuration of these components.

(b) If the three components are all independently drawn from a population of components which has the survivor function given in part (i), derive the survivor distribution of the overall system.

(c) Hence calculate the probability that the system will survive for longer than 3 months if the mean survival time of components is 6 months.

4. (i) (a) Define the following terms, as used in estimation and hypothesis testing:
   - an unbiased estimator;
   - type I error;
   - type II error.

   (b) If $x_1, \ldots, x_n$ is an independent random sample of observations from some distribution with unknown mean, show that the estimator $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is an unbiased estimator of the population mean.

   (c) Show that the maximum likelihood estimator of the parameter of a Poisson distribution is simply the sample mean.

(ii) The data below were obtained from a study of the possible harmful effects of using VDUs. Test the null hypothesis that there is no difference in the proportions reporting eye strain in the four groups, using the 5% significance level. State clearly each step and the conclusion in your analysis.

Type of work	Number without eye strain	Number with eye strain
Data entry by VDU	42	11
Conversational use of VDU	79	30
Full-time typing	64	14
Clerical work	52	3

END OF PAPER

COMP 245: Probability and Statistics for Students of Computing
2002

This sheet contains important formulae you may need in the examination. It does not contain definitions, concepts, or other material, and it does not contain simple formulae you would be expected to be able to derive or remember yourselves.

(Arithmetic) mean $\bar{x} = \frac{1}{n}\sum x_i$

Median: order the sample values $\{x_1, x_{2,}, \ldots, x_n\}$ so that $x_{(1)}$ is the smallest, $x_{(2)}$ is the next smallest, and so on, then the median is the value $x_{(n+1)/2}$.

Quartiles: the first quartile is $x_{([n+1]/4)}$, using the same ideas as in defining the median.

Geometric mean: $x_G = \sqrt[n]{\prod x_i}$ and *Harmonic mean:* $x_H = \left(\frac{1}{n}\sum\frac{1}{x_i}\right)^{-1} = \frac{n}{\sum 1/x_i}$

Variance $s^2 = \frac{1}{(n-1)}\sum\left(x_i - \bar{x}\right)^2$, *standard deviation*: $\sqrt{\text{variance}}$

Skewness: $\frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s}\right)^3$

S the set of all possible events, ϕ the empty set

Notation: $s \in S$ $\quad A \subset B$

$\phi \subset A \subset S$ for all A

$A \cup B$ (A or B) $\quad A \cap B$ (A and B) $\quad$ both commutative

$A \cap B$ is the *joint event* of A and B

A and B are *disjoint events* if $A \cap B = \phi$

Complement of A denoted by : A' or $\bar{A}$

$$P(\phi) = 0 \qquad P(S) = 1 \qquad P(A) = 1 - P(A')$$

For two disjoint events A and B (i.e. events for which $A \cap B = \phi$) $P(A \cup B) = P(A) + P(B)$

For *any* two events A and B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Generalise: $P(\cup_i A_i) = \sum_i P(A_i) - \sum P(A_i \cap A_j) + \sum P(A_i \cap A_j \cap A_k) + \ldots\ldots$

Two events are said to be *independent* if the occurrence or non-occurrence of one is not affected by whether or not the other occurs

If two events A and B are independent, then $P(A \cap B) = P(A) . P(B)$

The probability that A will occur, *given that B has occurred* is denoted $P(A \mid B)$
If A and B are independent then $P(A \mid B) = P(A)$.

In general, $P(A \cap B) = P(A \mid B) . P(B)$ and $P(A \cap B) = P(B \mid A) . P(A)$

From this $P(A \mid B) = P(B \mid A)P(A)/P(B)$ *(Bayes theorem)*

Now $P(B) = P(B \mid A)P(A) + P(B \mid A')P(A')$ *(theorem of total probability)*

So Bayes theorem can also be written as $P(A|B) = \dfrac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A')P(A')}$

The mean or *expected value* of a random variable is $E(X) = \sum_x xP(x)$, often denoted μ.

The variance of a random variable is

$V(X) = \sum_x (x-\mu)^2 P(x) = E(X^2) - E(X)^2 = E\left[(X-E(X))^2\right]$, often denoted σ^2.

The skewness of a random variable is $S(X) = \sum \left(\dfrac{x-\mu}{\sigma}\right)^3 P(x) = \dfrac{E\left[(x-\mu)^3\right]}{\sigma^3}$

$E(aX + bY) = aE(X) + bE(Y)$

$V(aX + bY) = a^2V(X) + b^2V(Y)$ if *X* and *Y* are independent

$V(aX + bY) = a^2V(X) + b^2V(Y) + 2abCov(X,Y)$, always, with $Cov(X,Y)$ the *covariance* of *X* and *Y*, defined as $Cov(X,Y) = E[(X - E(X))(Y - E(Y))]$

More generally, the mean of the sum of a weighted combination of $X_1,...,X_n$, $\sum a_i X_i$ is $\mu = \sum a_i \mu_i$ and the variance of the sum of a weighted combination of *independent* $X_1,...,X_n$, $\sum a_i X_i$ is $\sigma^2 = \sum a_i^2 \sigma_i^2$. If they are not independent then $\sigma^2 = \sum a_i^2 \sigma_i^2 + \sum_{i \neq j} a_i a_j Cov(X_i, X_j)$

The discrete uniform distribution

Let *S* be the set of integers from 1 to *n*.

$P(X = x) = 1/n$ with $\mu = \dfrac{(n+1)}{2}$ and $\sigma^2 = \dfrac{1}{12}(n^2 - 1)$

Bernoulli distribution

Let *P*(*E*) = *P*(*X*=1) = *p* and *P*(*E'*) = *P*(*X*=0) = 1-*p* = *q*

$P(X = x) = p^x q^{1-x}$ with $\mu = p$ $\sigma^2 = pq$

Binomial distribution

$P(X = x) = \binom{n}{x} p^x q^{n-x}$, Notation: *B*(*n*,*p*) $\mu = np$ $\sigma^2 = npq$

Geometric (e.g. prob *x* failures before first success, parameter *q*)

$P(X = x) = q^x p$ $\mu = \dfrac{q}{p}$ $\sigma^2 = \dfrac{q}{p^2}$ $x = 0,1,2,3,....$

Poisson

$P(X = x) = \dfrac{e^{-\mu}\mu^x}{x!}$ Mean = variance = μ

The *probability distribution function* (or *cumulative distribution function*, the cdf) is $F(x) = P(X \le x)$. The *probability density function* or pdf is $f(x) = F'(x)$ (the derivative of *F*), so that $F(x) = \int_{-\infty}^{x} f(y)dy$

$$\mu = E(X) = \int x f(x) dx \qquad \sigma^2 = E(X^2) - E(X)^2$$

Uniform: Pdf $\quad f(x) = \begin{cases} 1/(b-a) & a \le x < b \\ 0 & otherwise \end{cases}$

The Exponential distribution: Pdf $\quad f(x) = \begin{cases} \lambda \exp(-\lambda x) & x > 0 \\ 0 & x \le 0 \end{cases}$

and cdf $\quad F(x) = 1 - \exp(-\lambda x)$ when $x > 0$

$$\mu = 1/\lambda \qquad \sigma^2 = 1/\lambda^2$$

The Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \quad \text{for } -\infty < x < \infty$$

μ is the mean and σ is the standard deviation

The *standard normal distribution* has $\mu = 0$ and $\sigma = 1$: $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$

If a random variable *X* follows a $N(\mu, \sigma^2)$ distribution, then the random variable $(X - \mu)/\sigma$ follows a standard normal distribution *N*(0,1), often denoted $\phi(x)$. The cdf of the standard normal distribution is often denoted $\Phi(x)$.

The area between two points *a* and *b* under a normal curve $N(\mu, \sigma^2)$ is the same as the area under a N(0,1) curve between points $(a - \mu)/\sigma$ and $(b - \mu)/\sigma$.

Joint, marginal, and conditional densities

$$f(x, y) \qquad f_X(x) = \int f(x, y) dy \qquad f(y \mid x) = \frac{f(x, y)}{f_X(x)}$$

Given a random sample $x_1, \ldots, x_n$ from a distribution $p(x; \theta)$, the likelihood function for θ is $L(\theta) = \prod_{i=1}^{n} p(x_i; \theta)$. A 95% confidence interval for the mean μ of a distribution is approximately given by $\bar{x} \pm 1.96 \times s/\sqrt{n}$.

If *T* is a random variable denoting the lifetime of a component, with pdf $f(t)$ and cdf $F(t)$ the *survivor function* or *reliability function* is $R(t) = 1 - F(t)$ and the hazard function is $r(t) = f(t)/R(t)$. $\quad R(t) = \exp\left[-\int_0^t r(s) ds\right]$

The standard normal tables gives values of $\Phi(x) = F(x)$ for a N(0,1) distribution:

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
.0	.5	.9	.816	1.8	.964	2.8	.997
.1	.540	1.0	.841	1.9	.971	3.0	.998
.2	.579	1.1	.864	2.0	.977	3.5	.9998
.3	.618	1.2	.885	2.1	.982	1.282	.9
.4	.655	1.3	.903	2.2	.986	1.645	.95
.5	.691	1.4	.919	2.3	.989	1.96	.975
.6	.726	1.5	.933	2.4	.992	2.326	.99
.7	.758	1.6	.945	2.5	.994	2.576	.995
.8	.788	1.7	.955	2.6	.995	3.09	.999

The chi-squared table gives the values of x for which $\chi^2(k)$ has $P(X > x) = p$, where $\chi^2(k)$ is the chi-squared distribution with k degrees of freedom.

k	**.995**	**.975**	**.05**	**.025**	**.01**	k	**.995**	**.975**	**.05**	**.025**	**.01**
1	.000	.001	3.84	5.02	6.63	**18**	6.26	8.23	28.87	31.53	34.81
2	.010	.051	5.99	7.38	9.21	**20**	7.43	9.59	31.42	34.17	37.57
3	.072	.216	7.81	9.35	11.34	**22**	8.64	10.98	33.92	36.78	40.29
4	.207	.484	9.49	11.14	13.28	**24**	9.89	12.40	36.42	39.36	42.98
5	.412	.831	11.07	12.83	15.09	**26**	11.16	13.84	38.89	41.92	45.64
6	.676	1.24	12.59	14.45	16.81	**28**	12.46	15.31	41.34	44.46	48.28
7	.990	1.69	14.07	16.01	18.48	**30**	13.79	16.79	43.77	46.98	50.89
8	1.34	2.18	15.51	17.53	20.09	**40**	20.71	24.43	55.76	59.34	63.69
9	1.73	2.70	16.92	19.02	21.67	**50**	27.99	32.36	67.50	71.41	76.15
10	2.16	3.25	13.31	20.48	23.21	**60**	35.53	40.48	79.08	83.30	88.38
12	3.07	4.40	21.03	23.34	26.22	**70**	43.28	48.76	90.53	95.02	100.4
14	4.07	5.63	23.68	26.12	29.14	**80**	51.17	57.15	101.9	106.6	112.3
16	5.14	6.91	26.30	28.85	32.00	**100**	67.33	74.22	124.3	129.6	135.8

The Student's t table gives the values of x for which $t(v)$ has $P(|X| > x) = p$, where $t(v)$ is the Student t distribution with v degrees of freedom.

v	**.10**	**.05**	**.02**	**.01**	v	**.10**	**.05**	**.02**	**.01**
1	6.31	12.71	31.82	63.66	**9**	1.83	2.26	2.82	3.25
2	2.92	4.30	6.96	9.92	**10**	1.81	2.23	2.76	3.17
3	2.35	3.18	4.54	5.84	**12**	1.78	2.18	2.68	3.05
4	2.13	2.78	3.75	4.60	**15**	1.75	2.13	2.60	2.95
5	2.02	2.57	3.36	4.03	**20**	1.72	2.09	2.53	2.85
6	1.94	2.45	3.14	3.71	**25**	1.71	2.06	2.48	2.78
7	1.89	2.36	3.00	3.50	**40**	1.68	2.02	2.42	2.70
8	1.86	2.31	2.90	3.36	∞	1.645	1.96	2.326	2.576