UNIVERSITY OF
CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Wednesday 5 June 2002 9 to 12

PAPER 37

BIOSTATISTICS

*Attempt **FOUR** questions*

*There are **six** questions in this paper*

The questions carry equal weight

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Survival Data

Derive the maximum likelihood estimator of the rate parameter θ of an exponential(θ) distribution in the presence of censoring.

Distinguish between *uninformative* and *informative* censoring. Give an example of each. Why is the distinction important?

A researcher wants to simulate a sample from the exponential(1) distribution in which half the observations are censored. He uses a random number generator to obtain 1000 simulated observations from the exponential(1) distribution. He then divides the set of numbers into two halves; one half he declares to represent censored observations and the other half he decides to represent observed failures.

Given that the sum of the 1000 observations is 1031.6, show how to find an estimate for the rate parameter using the dataset *after* the simulated censoring was applied. Compare your estimate with that obtained by treating all the observations as uncensored.

State with a reason whether the simulated censoring is informative or uninformative. How would you simulate a dataset where the probability that an observation is censored is $\frac{1}{2}$ and the censoring is uninformative?

[*All censoring in this question is ordinary right-censoring; if an individual is censored at time t it is known only that failure would have occurred strictly later than t.*]

2 Survival Data

An individual is exposed to two mutually exclusive events A and B with hazard functions $h_A(t)$ and $h_B(t)$ respectively.

Find expressions for the probability that

(i) no event occurs by time t;

(ii) event A occurs by time t;

(iii) either A or B occurs at some time

in terms of the hazard functions and the integrated hazard functions.

Data has been obtained from a set of n identical individuals exposed to the two mutually exclusive events A and B.

Using the following information:-

(1) the estimated probability of event A occurring at or before $t = 7$ hours is 0.285 and the corresponding estimate for event B is 0.241;

(2) no events occurred in the time interval 7 hours $< t <$ 7.8 hours;

(3) immediately before $t = 7.8$ hours the risk set contained 20 individuals;

(4) at $t = 7.8$ hours exactly, one individual was censored, two individuals experienced event A and one individual experienced event B,

estimate the probability that event A occurs at or before $t = 7.8$ hours.

3 Statistics in Medical Practice

Table 1 shows estimated human dietary exposure to BSE from UK consumption of bovine meat products during 1980-89 and 1990-96 for two birth-cohorts. To date, there have been 72 UK onsets of vCJD in the post-1969 birth cohort and 36 in the 1940-1969 birth cohort.

Table 1: Estimated human dietary exposure to BSE from UK consumption of bovine meat products [cited as bovine infectious disease units]

		1980-89	**1990-1996**	**Total**
Birth cohort	**Post 1969**	85,000	170,000	255,000
	1940 - 1969	170,000	170,000	340,000

(i) Assuming age-independent susceptibility and incubation period, are UK's 72 and 36 vCJD onsets in post-1969 and 1940-1969 birth cohorts respectively consistent with their having arisen solely from human dietary exposure to BSE from consumption of bovine meat products in 1980-89?

(ii) Besides susceptibility and incubation period, suggest two other considerations that might affect your conclusion at (i).

You may note that BSE in sheep and scrapie in sheep look the same clinically, but can be distinguished by post-mortem tests even before the clinical disease has become apparent.

(iii) UK slaughters 4 million adult sheep per annum, up to 2000 of which are expected to test positive for scrapie. The prevalence of BSE among adult sheep with scrapie is known to be less than 2%; and is expected to be less than 2% also in adult sheep testing positive for scrapie. If BSE prevalence is indeed 2% among scrapie-test-positive adult sheep, how would you compute the probability of finding no BSE positive in: (a) 50,000 slaughtered adult sheet, (b) 500,000 slaughtered adult sheep. [You should assume that the probability that a sheep tests positive for BSE, given that it does NOT test positive for scrapie, is zero.]

(iv) Testing costs £40 per slaughtered adult sheep. BSE test positivity in 1 per 100,000 slaughtered adult cattle is a concern. BSE test positivity in slaughtered adult sheep is more worrying so that risk managers would be concerned if BSE test positivity in slaughtered adult sheep was 1 per 200,000. How would you advise the UK on the question of whether BSE surveillance in 50,000 slaughtered adult sheep is sufficient?

4 Statistics in Medical Practice

A tabular CUSUM monitoring scheme is characterised by the updating formula

$$X_t = \max(0, X_{t-1} + W_t), \qquad t = 1, 2, 3, \ldots$$

where X_t represents the value of the process at time t, $X_0 = 0$, and W_t is the sample score which is assigned to the t^{th} subgroup of data.

Such a CUSUM procedure is to be used to monitor deaths in a medical practice. Information is available on a yearly basis and is summarised by the number of observed deaths, denoted y_t, for the t^{th} year. Also for the t^{th} year there is calculated an expected number of deaths λ_{0t} based on national mortality rates and the patient mix in the practice to be monitored.

Assume that the number of deaths for a single year arises as an observation from a Poisson distribution with mean λ_t.

(i) Define the average run length (ARL) for a CUSUM process and explain how ARLs, along with null and alternative hypotheses H_0 and H_A respectively, are used to characterise a specific monitoring procedure. Compare and contrast the use of ARLs with the role of Type I and Type II errors in the testing of hypotheses.

(ii) Two possible definitions for W_t can be based (a) on $O - E$ which compares the observed mortality (O) with expected mortality (E), and (b) on log likelihood ratios.

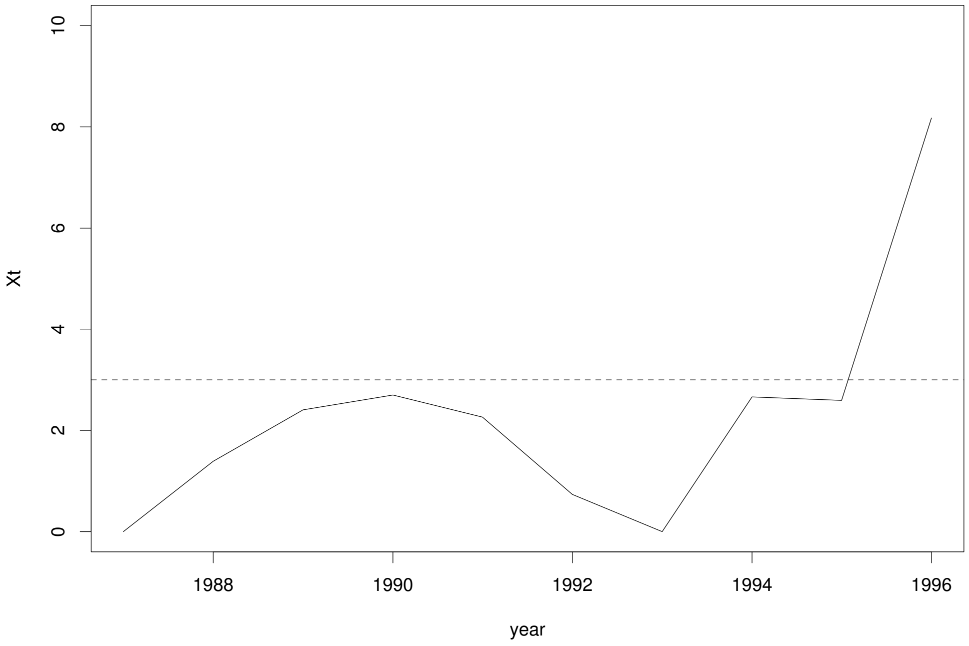
Which of these two is preferable and why?

(iii) Assume the expected mortality rates are taken to represent the null hypothesis. Assume further that alternative hypotheses are defined in terms of a parameter $\theta = \lambda_t/\lambda_{0t}$ for all t. For this situation, derive the sample score W_t based on the log likelihood ratio. Show that if $\theta = 2$, then W_t is approximately $O(\log_e 2) - E$.

(iv) Explain why a tabular CUSUM is not allowed to have a value below zero.

(v) Figure 1 gives a CUSUM plot which might have been used to monitor death rates per year in Harold Shipman's medical practice. The alternative of interest was taken to be $\theta = 2$. Assume a boundary level of 3 was chosen for monitoring purposes. Such a boundary has an ARL of 111 years under the null hypothesis and 5.2 years under the alternative. This boundary is crossed when the 1996 data are included in the plot. Suggest an appropriate course of action when this crossing is observed. As part of your answer comment particularly on the relevance or irrelevance of the observation that $X_{1993} = 0$.

Figure 1



CUSUM monitoring death rates per year under Harold Shipman, 1987 - 1996

5 Statistical Genetics

Consider a genetic locus with two alleles, a and b. The frequencies of these alleles in the general population are p and $1 - p$, respectively.

(i) Under Hardy-Weinberg equilibrium, what are the expected frequencies of individuals with genotypes a/a, a/b and b/b respectively, in terms of p?

(ii) Write down the six possible mating types (i.e. the possible combinations of genotypes that can occur) when two individuals from the population come together to produce offspring. For each mating type, what are the possible offspring that could occur?

(iii) Without assuming Hardy-Weinberg equilibrium in the initial population, show that Hardy-Weinberg equilibrium frequencies are achieved in the population after just a single generation of random mating.

6 Statistical Genetics

(i) In the multiplicative model for risk of a disease in relation to a diallelic locus, the penetrances for the 1/1, 1/2, and 2/2 genotypes are p, θp, and $\theta^2 p$ respectively. Assuming Hardy–Weinberg equilibrium in the population with a proportion ϕ of chromosomes carrying the 2 allele at the locus, show that the overall penetrance is $\alpha^2 p$, where

$$\alpha = 1 + \phi(\theta - 1).$$

(ii) When the trait is coded as 0 or 1, denoting absence or presence of disease respectively, give expressions for the total variance of the trait and for the variance within genotype. Hence show that the the genetic variance is $(\beta^2 - \alpha^4)p^2$, where

$$\beta = 1 + \phi(\theta^2 - 1).$$

(iii) Show that the covariance between trait values (again coded as 0 or 1) for two subjects who share genes 2-IBD at the trait locus is also equal to $(\beta^2 - \alpha^4)p^2$. Explain why this covariance is zero for two subjects who share no genes IBD at the trait locus.

(iv) Hence show that, under the model described in part (i), the probabilities that two subjects are both affected conditional upon their sharing, respectively, 2 or 0 genes IBD at the trait locus are $p^2\beta^2$ and $p^2\alpha^4$.

(v) What is the expected ratio of affected sibling pairs sharing 2 versus 0 genes IBD at the trait locus? Is the proportion of such pairs sharing 1 gene IBD,

$$0.5, \ > 0.5, \ \text{or} \ < 0.5?$$

Why?