UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday 12 June 2006 9 to 11

PAPER 44

BIOSTATISTICS

*Attempt **THREE** questions.*

*There are **FIVE** questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

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

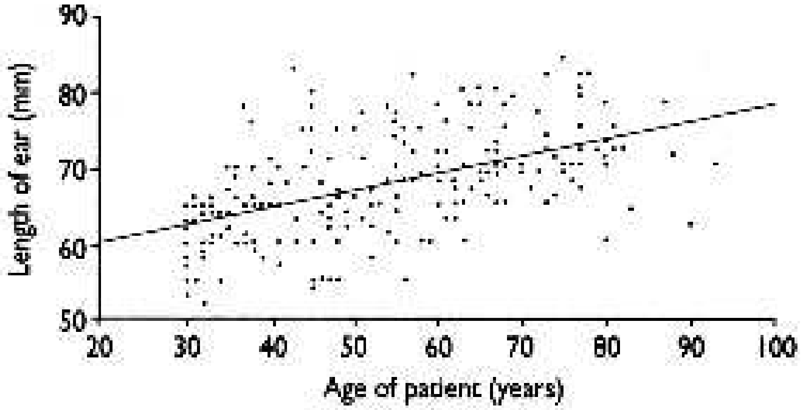
1 Statistics in Medical Practice

Heathcote (1995) reported the following cross-sectional study:

"Four ordinary general practitioners agreed to ask patients attending for routine surgery consultations for permission to measure the size of their ears... the aim was to ask consecutive patients aged 30 or over, of either sex, and of any racial group."

"In all, 206 patients were studied (mean age 53.75 (range 30-93; median age 53) years). The mean ear length was 675 mm (range 520-840 mm), and the linear regression equation was: ear length=55.9+(0.22 x patient's age) (95% confidence intervals for age co-efficient 0.17 to 0.27). The figure shows a scatter plot of the relation between length of ear and age."

We do not know the sex of the participants, but might assume around 60% were female.



He concluded "It seems therefore that as we get older our ears get bigger (on average by 0.22 mm a year)".

(a) Suggest an alternative study design specifically to check Heathcote's conclusion.

(b) Would you consider this a 'strong' relationship in terms of (i) statistical significance, (ii) predictive ability? What other factors might improve predictability?

(c) Give at least three reasons (even if not particularly plausible) why Heathcote's conclusion may not be appropriate, and for each reason give a possible extension of the study that might be appropriate to check alternative explanations for this observed relationship.

Reference

Heathcote JA (1995) Why do old men have big ears?, *British Medical Journal*, **311**, p1668.

2 Statistics in Medical Practice

(a) Consider the use of a gamma distribution as a model for incubation times associated with an infection disease. If T is the random variable representing an incubation time, with an observed value of $T = t$, then a gamma distribution for T is specified by the probability density function

$$g(t) = \frac{1}{s^a \Gamma(a)} t^{(a-1)} e^{-(t/s)},$$

where $t > 0, a > 0$ and $s > 0$. The associated distribution function is denoted $G(t)$.

Under the further assumption that this distribution is truncated at some time M, so that $0 < T < M$, write down a likelihood function for the estimation of a, s and M based on on n incubation times $t_1, t_2, \ldots, t_n$. In addition, define the profile likelihood for the parameter M and find the maximum likelihood estimate of M.

(b) Consider now the use of a truncated log-gamma model for incubation times.

With $\alpha, q \in \mathbb{R}$ and $\sigma > 0$, the log-gamma model can be written as a location scale model $y = log(t) = \alpha + \sigma w$, where the density for w, $f(w; q)$ depends on the parameter q. For appropriate values of q, the log-gamma distribution includes the gamma and log-normal distributions as special cases.

Based on the incubation times for 67 SARS cases, Figure 1 overleaf presents the profile likelihoods for the estimation of the truncation parameter M based on a truncated log-gamma model and the special cases of truncated gamma and truncated log-normal models. These profile likelihoods have been standardized to have a maximum value of one, and are referred to as relative likelihoods.

Explain what can be learned from such profile likelihoods and compare the information in the three profile likelihoods.

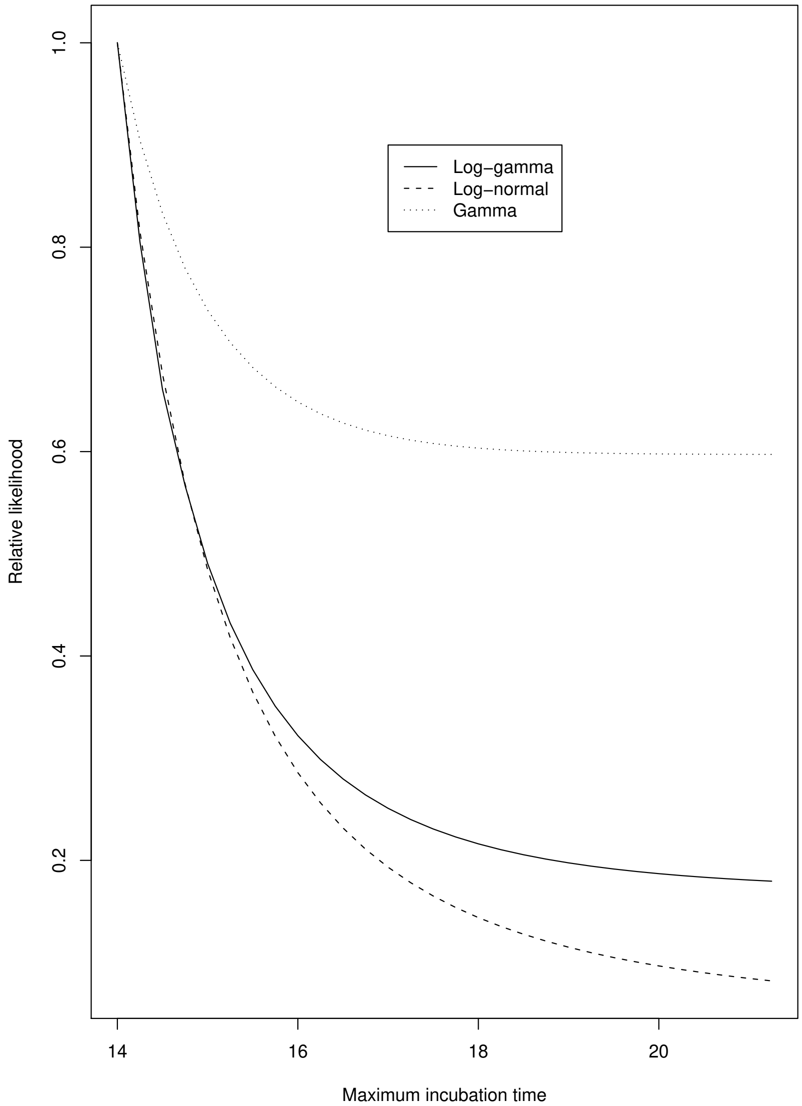


Figure 1: Profile likelihoods based on 67 SARS cases

(c) Explain briefly how quarantine is used to control the spread of infectious diseases. Discuss, in addition, how models for incubation times based on (1) truncated and (2) untruncated distributions can be used to set the length of quarantine periods and when the use of each might be appropriate.

3 Survival Data Analysis

What is meant by a *frailty* model?

The individual hazard in a proportional frailty model is given by $Uh_0(t)$, where U is a random variable, density $g(u)$, such that $U \geqslant 0$, $\mathbb{E}(U) = 1$ and $h_0(t)$ is the baseline hazard. Show that the overall survivor function is given by $\tilde{g}(H_0(t))$, where $H_0(t)$ is the integrated baseline hazard and $\tilde{g}(s)$ is the Laplace transform of $g(u)$.

A *cure* model is a survival model such that an individual has probability π of having a zero hazard function and a probability $1 - \pi$ of having a hazard function $h^*(t)$. Express the cure model as a frailty model (taking particular care of the definition of $h_0(t)$). Obtain, via the Laplace transform of an appropriate $g(u)$, the overall survivior function for a cure model.

4 Survival Data Analysis

Explain the principles of the *log-rank* test. Describe how you would construct a *stratified* version of the test.

The effect of two treatments, A and B, on the time to a certain event is being compared by a log-rank test. The individuals are stratified so that there are exactly two individuals in each stratum, one receiving treatment A and the other receiving treatment B. Show that each stratum provides at most one informative contingency table. In what circumstances does a stratum not provide an informative contingency table?

Write down the contingency table in the case that the individual receiving treatment A is known to have the event before the individual receiving treatment B. Calculate the contribution to the log-rank statistic from that table.

Hence obtain the overall log-rank statistic (do not attempt to normalize to unit standard deviation).

Let d_A be the number of strata in which the individual receiving treatment A is known to have the event first and d_B the number of strata in which the individual receiving treatment B is known to have the event first. Show that, by conditioning on the sum $d_A + d_B$, the log-rank statistic can be transformed to a statistic with a known distribution under the null hypothesis.

5 Survival Data Analysis

Outline how the parameters of a *proportional hazards* model are estimated. How can an estimate of the *baseline hazard function* be obtained?

A proportional hazards model has been fitted to the dataset x_i, v_i: $i \in \{1, \ldots, m\}$ where x_i are ordered, distinct times ($x_i < x_{i+1}$) and the v_i are the visibility indicators ($v_i = 1$) if x_i represents an observed event and $v_i = 0$ represents a censored observation).

Show that

(a) the parameter estimates $\hat{\beta}$ are unaffected by the value v_m of the visibility indictor of the last observation;

(b) the increase in the estimated integrated hazard for the mth subject between x_{m-1} and x_m is equal to v_m.

A cancer time-to-death dataset's largest observations (110 days) is a censored observation. A proportional hazards model is fitted and the Martingale residuals calculated. It was subsequently discovered that the subject last known to be alive at 110 days in fact died at 300 days. Show that the Martingale residuals are unaffected by the additional information.

END OF PAPER