A.2 Estimation of lifetime distributions and Markov models

1. We have discussed the 22 skeletons of *A. sarcophagus* analysed by Erickson *et al.*. The observed (curtate) ages at death in years were 2,4,6,8,9,11,12,13,14,14,15,15,16, 17,17,18,19,19,20,21,23,28.

   (a) Estimate directly the life expectancy of this population.

   (b) Estimate a 95% confidence interval for the life expectancy.

   (c) We estimated survival probabilities by $\hat{q}^d_x = d_x/\ell_x$. Show that the life expectancy predicted from this estimated distribution must be the same as that computed directly from the observed lifetimes.

   (d) Estimate survival probabilities ($\hat{q}^c_x$) for this population, using the continuous method, grouping the lifetime by periods of five years. (The relevant mathematical generalisation of the continuous method is explored in the problem below.) Based on this estimated life-table, estimate the life expectancy for the population. Why is it different from the life expectancy estimated above?

   (e) As we explained in the lecture notes, it is reasonable to add 1/2 year to the life expectancy estimated from averaging curtate lifetime observations to estimate the full life expectancy $\mathring{e}_x$, on the assumption that individuals who died between age x and $x+1$ probably lived on average an extra half year. We have estimated a hazard rate ("force of mortality") of 0.333 for ages 20 and above. Supposing this is true, what is the true expected length of life of an individual whose curtate lifetime is reported as 25 years? What does this suggest about the validity of the $+\frac{1}{2}$ rule when the mortality is high.

   (f) Suppose a population has Gompertz hazard rate given by $h(x) = Be^{\theta x}$ at age x, for $x \geq 0$, where B and θ are assumed nonnegative. We observe n individuals, with deaths at ages $x_1, \ldots, x_n$. Define $Q(\theta) := \frac{1}{n}\sum e^{\theta x_i}$, $\bar{x} := \frac{1}{n}\sum x_i$. The equation

   $$\frac{Q'(\hat{\theta})}{Q(\hat{\theta}) - 1} - \frac{1}{\hat{\theta}} = \bar{x}$$

   has a unique solution (optional extra: Find conditions under which such a solution must exist). Show that $\hat{\theta}$ and $\hat{B} := \hat{\theta}/(Q(\hat{\theta}) - 1)$ are the maximum-likelihood estimator for (θ, B). Compute the maximum likelihood estimate for fitting Gompertz parameters to the dinosaur population. Use the asymptotic theory to compute a 95% confidence interval for B, assuming the Gompertz model. (As an extra optional challenge: Use the bivariate distribution to compute a 95% confidence interval for the mortality rate at age 20.)

   (g) The exponential integral functions are defined by

   $$E_n(z) := \int_z^\infty t^{-n} e^{-t} dt.$$

   Express the life expectancy at age x of this population, in terms of the function E_1. Using either a table or computer software, estimate the life expectancy for the dinosaur population from the Gompertz model. (For instance, `Maple` uses the command `Ei`.)

2. [**Optional computer exercise**] In the context of question 1, estimate the 95% confidence intervals by a *bootstrap* analysis. The idea is, we want to see how different the parameter estimates might have been if we'd happened by chance to pick a different sample of skeletons. So take 1000 samples of 22 skeletons from the same distribution and fit the Gompertz model. Collect all the parameter estimates you get, and use those to estimate the errors. But where do we get these 1000 samples from? We don't have any more skeletons...There are two approaches. In the *parametric bootstrap*, you simulate 22 deaths as though they were drawn from the Gompertz survival function that you estimated in part (a). In the *nonparametric bootstrap*, you simulate 22 deaths by sampling them at random (with replacement!) from the 22 observations that we already have. Thus, in our resampling, some data points will appear twice (or more), and others not at all. Play around with this!

3. In practice, not all data are in the form that is most convenient to apply the basic statistical methods. In a mortality study, suppose that lifetimes are not observed directly, but only counts of subjects at risk on 1 January and numbers of deaths are available, over N years $[K, K+N+1]$. One then has to use approximations based on some assumptions. These techniques are known as the "census approximation". 1 January is then called the "census date".

   (a) Denote by $P_{x,t}$ the number of lives under observation, aged x (last birthday), at any time t.

   i. Given n individuals at risk and aged x between times a_i and b_i, show that the total time at risk is

   $$E_x^c = \sum_{i=1}^{n}(b_i - a_i) = \int_K^{K+N+1} P_{x,t}dt.$$

   ii. Assume that $P_{x,t}$ is linear between census dates $t = K, K+1, \dots, K+N+1$. Calculate E_x^c in terms of $P_{x,t}$, $t = K, K+1, \dots, K+N+1$. Explain why the assumption cannot hold exactly. However, it provides a simple approximation used in practice.

   (b) Instead of $d_x^{(1)}$ = # deaths with x last birthday before death (leading to $\hat{\mu}_{x+\frac{1}{2}} = d_x/E_x^c$, assuming $\mu_t = \mu_{x+\frac{1}{2}}$, $x \le t < x+1$), some insurance companies record $d_x^{(2)}$ = # deaths in calendar year of the xth birthday.

   Describe the resulting estimate of the force of mortality. Explain the definition that you need to use for $P_{x,t}$. State any further assumptions you make.

4. Suppose $X_1, \dots, X_n$ are independent and have exponential distribution with parameter λ. In an earlier sheet you showed that the MLE is $\hat{\lambda} = n/\sum X_i$, and that $2n\lambda/\hat{\lambda} \sim \chi^2_{2n}$.

   (a) **(Review)** Suppose n is ∞. What is the distribution of $\#\{k : X_1 + \cdots X_k \le t\}$, for fixed t? Describe the connection of this to the Poisson process.

   (b) Use this fact to show that if x is a single observation from a Poisson distribution with parameter μ, then an exact $(1-\alpha)$-confidence interval for μ is

   $$\left(\frac{1}{2}c_{\alpha/2}(2x), \frac{1}{2}c_{1-\alpha/2}(2x+2)\right),$$

where $c_\alpha(d)$ is the α quantile of the χ^2 distribution; that is, $P\{X \le c_\alpha(d)\} = \alpha$ where X has the χ^2 distribution with d degrees of freedom.

(c) Suppose we observe n individuals with independent exponential lifetimes, with unknown parameter λ. During a time-period $[0, t]$, k of the n die. Show that an approximate $(1-\alpha)$-confidence interval for λ is

$$\left(\frac{1}{2tn} c_{\alpha/2}(2k), \frac{1}{2tn} c_{1-\alpha/2}(2k+2) \right).$$

Why is it not exact? Under what circumstances would you expect it to be better than the confidence interval derived from the normal approximation?

(d) Apply the above results to the *Albertosaurus* data from Table 1.1, to compute approximate 95% confidence intervals for the mortality rates in each 5-year period (under the modelling assumption that mortality rates are constant during these intervals). Compare these to the confidence intervals that you get from the asymptotic normal approximation.

5. A large investigation has been carried out into mortality among people of working age. They are to be compared with a well-known standard table.

Age	Exposed to risk	Observed deaths	standard mortality
	E_x	d_x	$q_x^s \times 10^5$
20–24	35000	35	97
25–29	33000	30	88
30–34	30000	31	117
35–39	30000	45	173
40–44	31000	84	260
45–49	28000	138	460
50–54	25000	229	850
55–59	23000	360	1500
60–64	20000	522	2500

Perform the following three tests, finding the p-values and the test statistic (where appropriate): a) χ^2 b) sign test c) cumulative-deviations test, commenting on the outcomes.