A.2 Life expectancy, goodness of fit and graduation

1. Erickson *et al.* analysed 22 skeletons of *A. sarcophagus*. 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 (curtate) life expectancy of this population.

   (b) Construct an approximate 95% confidence interval for this life expectancy.

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

   (d) Using the approximation that the fractional part of the lifetimes are independent of the integer part and uniformly distributed, adapt your answers to (a) and (b) for the full life expectancy.

   (e) Estimate mortality rates for the population, under the assumption that these rates are constant over five-year intervals $[0, 5]$, $[5, 10]$, etc. (Consider two approaches: either use the "uniform fractional part" assumption to approximate the total exposed to risk $\tilde{\ell}_x$ given the curtate lifetimes, or consider directly the likelihood of the observed curtate data). Show how to estimate again the life expectancy for the population, based on the life table obtained.

2. Below is an excerpt from a cohort life table for men in England and Wales born in 1894, including curtate life expectancies. (Data from the Human Mortality Database at http://www.mortality.org). Using the given data:

   (a) Estimate the change to e_0, the curtate life expectancy at birth, if the mortality rate in the first two years of life were reduced to modern-day levels (say $q_0 = 0.005$, $q_1 = 0.0004$).

   (b) Make a rough estimate of the change to e_0 if the increases in mortality due to the 1914-18 war and the 1918-19 influenza pandemic had not occurred.

AGE x	ℓ_x	q_x	e_x
0	100000	0.16134	44.83
1	83866	0.05398	52.39
	⋮	⋮	
14	74067	0.0022	45.99
15	73904	0.00237	45.09
16	73729	0.0026	44.2
17	73538	0.00301	43.31
18	73316	0.00313	42.44
19	73087	0.00787	41.57
20	72512	0.01836	40.9
21	71181	0.03218	40.65
22	68890	0.04424	40.98

AGE x	ℓ_x	q_x	e_x
23	65842	0.06194	41.86
24	61764	0.02088	43.59
25	60474	0.00551	43.51
26	60141	0.00385	42.75
27	59910	0.00384	41.91
28	59680	0.00391	41.07
29	59446	0.00377	40.23
30	59222	0.00386	39.38
31	58994	0.00367	38.53
32	58777	0.0038	37.67
33	58554	0.00399	36.81
34	58320	0.00445	35.96
35	58061	0.0046	35.11

Cohort life table for male births in 1894 in England and Wales

3. Show how to represent on a Lexis diagram (i) the number of individuals in a population who are aged 20 at the start of 2014; (ii) the number of individuals in the population who turn 21 during 2014.

4. Consider the situation of estimating mortality rates by observing a population over the course of a time interval $[K, K+N]$.

   Let $d_x^{(1)}$ be the number of deaths at curtate age x, and $E_x^{c(1)}$ the corresponding central exposure to risk, given by

$$E_x^{c(1)} = \int_K^{K+N} P_{x,t}^{(1)} dt,$$

   where $P_{x,t}^{(1)}$ is the number of individuals under observation at time t with curtate age x. If we observe only census data at times $K, K+1, \ldots, K+N$, we can use linear interpolation to approximate

$$E_x^{c(1)} \approx \sum_{k=K+1}^{K+N} \frac{1}{2} \left(P_{x,k-1}^{(1)} + P_{x,k}^{(1)} \right).$$

   Then $d_x^{(1)}/E_x^{c(1)}$ gives an estimator of $\mu_{x+1/2}$, under the assumption that the mortality rate μ_t at age t is constant over $t \in [x, x+1]$.

   More generally, $d_x^{(1)}/E_x^{c(1)}$ will approximate $\int_{t=x}^{x+1} \mu_t dt$ as long as the number observed does not change too significantly over the relevant time (say, if mortality is low and there are not significant numbers of people entering or leaving the observed population for other reasons).

   (a) State the *Principle of Correspondence*. Adapt the properties above to the following different definitions of d_x (giving corresponding definitions and suitable approximations for E_x^c, and explaining what may be estimated by d_x/E_x^c). Where appropriate, explain what further assumptions are needed:

   $d_x^{(2)} = \#$ deaths with x nearest birthday to death;
   $d_x^{(3)} = \#$ deaths in calendar year of the xth birthday;
   $d_x^{(4)} = \#$ deaths with curtate age x at time of last annual policy renewal;

   (b) For cases of $d_x^{(1)}$, $d_x^{(2)}$ and $d_x^{(3)}$, indicate the areas relevant to the calculation of deaths and exposure on a Lexis diagram.

5. In the setting of Question 2., assume that both birthdays and policy anniversaries are uniformly spread over the year. Consider

   $d_x^{(5)} = \#$ deaths with last annual policy renewal in the calendar year of xth birthday.

   (a) Show that the possible age range of deaths counted under $d_x^{(5)}$ is $(x-1, x+2)$.

   (b) For each $y \in (x-1, x+2)$, calculate the probability that a death aged y is counted by $d_x^{(5)}$.

   (c) Using the Principal of Correspondence, specify the denominator $E_x^{x(5)}$ for which $d_x^{(5)}/E_x^{c(5)}$ should be close to a weighted integral $\int_{-1}^{2} g(u)\mu_{x+u} du$ over the force of mortality between $x-1$ and $x+2$. Specify the weight function g.

6. (a) Suppose you are given estimates for a population of remaining life expectancy $\overset{\circ}{e}_x$ and $\overset{\circ}{e}_{x+t}$, corresponding to ages x and $x+t$ (years). You wish to compute the mortality probability ${}_tq_x$. Under the assumption that mortality rates are constant over this interval, explain how to derive the approximation

$$ {}_tq_x \approx \frac{t + \overset{\circ}{e}_{x+t} - \overset{\circ}{e}_x}{t/2 + \overset{\circ}{e}_{x+t}}. \qquad (*)$$

Under what conditions will this approximation be reasonable?

(b) The following is an estimated table of $\overset{\circ}{e}_x$ (in years) in ancient Rome, as computed by Tim Parkin *Demography and Roman Society* (Johns Hopkins University Press, 1992).

x	0	1	5	10	15	20	25	30	35	40	45	50	55	60	65	70
$\overset{\circ}{e}_x$	25	33	43	41	37	34	32	29	26	23	20	17	14	10	8	6

On the basis of these figures, and assuming the mortality rates to be constant over the relevant age intervals, use equation (*) to approximate the annual mortality probabilities ${}_1q_x$ over the age intervals 0–1, 1–5, 5–10.

(c) Suppose we know that 15% of Roman infants died of dysentry in their first year. Under the competing risks assumption, estimate the change in life expectancy at birth that would have resulted if this disease had been eliminated among infants. Note the assumptions you make in carrying out the calculation.

7. A large investigation has been carried out into mortality among people of working age, recording Observed deaths d_x in 5-year intervals and associated Exposed to risk E_x (5-year sums of annual initial exposed to risk). They are to be compared with a well-known standard table of one-year death probabilites q_x^s, constant within the 5-year intervals.

x	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

x	E_x	d_x	$q_x^s \times 10^5$
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-test (b) sign test (c) cumulative-deviations test, commenting on the outcomes.

8. A medium-sized UK pension scheme carried out an investigation into the mortality of its pensioners between 2000 and 2002.

(a) Explain why the graduation of crude rates obtained might be desirable.

(b) Graduation was done using a Gompertz-Makeham model. The crude rates $\mu_{x+2.5}$ and the proposed graduation $\overset{\circ}{\mu}_{x+2.5}$ are:

x	E_x^c	d_x	$\mu_{x+2.5}$	$\mathring{\mu}_{x+2.5}$	z_x
60–64	1388.9	10	0.0072	0.0070	0.100
65–69	1188.8	17	0.0143	0.0150	-0.198
70–74	880.5	28	0.0318	0.0287	0.542
75–79	841.6	34	0.0404	0.0521	-1.486
80–84	402.8	43	0.1068	0.0920	0.975
85–89	123.9	21	0.1695	0.1602	0.260
90–94	27.9	9	0.3226	0.2765	0.463
95–99	17.5	5	0.2857	0.4750	-1.149

Explain how the column z_x has been calculated.

The graduated rates were calculated by minimising $\sum z_x^2$. Explain why this corresponds (approximately) to a maximum likelihood procedure.

Under a null hypothesis that the underlying rates are indeed from the Gompertz-Makeham family, what should the distribution of the obtained $\sum z_x^2$ be?

Carry out an appropriate test for goodness of fit.