

### B.3 Sampling theory for Life Table estimation; Census approximation

1. (a) i. Just write the quantities as integrals and sums and interchange the order of integration and summation:

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

- ii. Under the assumption of piecewise linearity we calculate

$$E_x^c = \sum_{k=K}^{K+N} \int_0^1 (rP_{x,k} + (1-r)P_{x,k+1}) dr = \sum_{k=K}^{K+N} \frac{P_{x,k} + P_{x,k+1}}{2}.$$

The assumption of piecewise linear  $P_{x,t}$  cannot hold exactly since  $P_{x,t} \in \mathbb{N}$ , but for large  $n$  this is negligible.

- (b) Denote by  $E_{[t,t+1]}^c$  the total time exposed to risk between ages  $t$  and  $t+1$  for  $t \in \mathbb{R}_+$ . Then, for the first three, clearly,

$$\begin{aligned} d_x^{(1)} / E_{[x,x+1]}^c & \text{ estimates } \mu_{x+\frac{1}{2}} & \mu_t \text{ constant for } t \in [x, x+1]; \\ d_x^{(2)} / E_{[x-\frac{1}{2}, x+\frac{1}{2}]}^c & \text{ estimates } \mu_x & \text{ if } \mu_t \text{ constant for } t \in [x-\frac{1}{2}, x+\frac{1}{2}]; \\ d_x^{(3)} / E_{[x-1, x]}^c & \text{ estimates } \mu_{x-\frac{1}{2}} & \mu_t \text{ constant for } t \in [x-1, x]; \end{aligned}$$

The expressions of  $E_x^c$  in (a) yield the first and the third, for the second we base calculations on

$$P_{x,t}^{(2)} = \# \text{ lives at risk with } x \text{ nearest birthday at time } t,$$

which can be approximated by  $\tilde{P}_{x,t}^{(2)} = \frac{1}{2}(P_{x-1,t} + P_{x,t})$ .

The last two are more tricky. A death contributing  $d_x^{(4)}$  can be due to somebody aged anywhere in  $(x-1, x+1)$ , so there is overlap between the one-year age groups. We can define

$$P_{x,t}^{(4)} = \# \text{ lives at risk at time } t \text{ with } x\text{th birthday in calendar year } [t],$$

but also ought to adjust

$$E_x^{c,4} = \int_K^{K+N+1} P_{x,t}^{(4)} dt \approx \sum_{k=K}^{K+N} \frac{P_{x,k} + P_{x+1,k+1}}{2}$$

since it is more natural to assume that the cohort of lives  $P_{x,k}$  with  $x$ th birthday in calendar year  $k$  changes linearly to  $P_{x+1,k+1}$  since this will count the same people (having their  $x+1$ st birthday in calendar year  $k+1$ ).

Similarly,

$$P_{x,t}^{(5)} = \# \text{ lives at risk age } x \text{ last birthday at last policy anniversary}$$

and

$$E_x^{c,5} = \int_K^{K+N+1} P_{x,t}^{(5)} dt = \sum_{k=K}^{K+N} \frac{P_{x,k} + P_{x,k+1}}{2},$$

assuming that  $P_{x,k}$  changes linearly to  $P_{x,k+1}$ , which amounts to assuming that policy anniversaries are spread uniformly over the calendar year.

The Lexis diagrams for  $d_x^{(1)}$  and  $d_x^{(4)}$  are precisely the ones that appear in Figures 5.3 and 5.4. For  $d_x^{(2)}$  the squares in Figure 5.3 are simply shifted down by 1/2 year, and for  $d_x^{(3)}$  they are shifted down by a full year.  $d_x^{(5)}$  doesn't represent a count over any simple geometric form.

2. (a) The times between arrivals in a Poisson process with intensity  $\lambda$  are independent exponential with parameter  $\lambda$ . Thus, the number of cumulative sums that are  $\leq t$  is precisely the number of arrivals on the interval  $[0, t]$ , which has a Poisson distribution with parameter  $\lambda t$ .
- (b) By definition, the interval  $(a(x), b(x))$  is a  $(1 - \alpha)100\%$  confidence interval for the parameter  $\mu$  if and only if  $P\{x : \mu \notin [a(x), b(x)]\} = \alpha$ . For simplicity, let's say we're looking for a symmetric confidence interval, so with  $P\{x : \mu < a(x)\} = P\{x : \mu > b(x)\} = \alpha/2$ .

The key point is that the lower limit of the confidence interval depends on the upper tail of the distribution, and vice versa. Thus, if we want to find  $a(x)$ , we need to find  $\lambda$  small enough that the probability in the upper tail, of  $X$  as big as the  $x$  we observed, is below  $\alpha/2$ . That is, we need to solve the equation

$$P_a(X \geq x) = \alpha/2.$$

Of course, we could do that directly, numerically, though it is slightly awkward that we are searching for a fixed quantile of a distribution whose parameter is  $a$ , rather than a fixed distribution.

Letting  $T_1, \dots, T_x$  be i.i.d. exponential random variables with parameter 1, and  $Z_{2x}$  a random variable with  $\chi^2$  distribution with  $2x$  degrees of freedom, we know that

$$P_\lambda(X \geq x) = P(T_1 + \dots + T_x \leq \lambda) = P(Z_{2x} \leq 2\lambda).$$

Thus,  $a(x)$  must satisfy  $P(Z_{2x} \leq 2a) = \alpha/2$ , meaning that  $a = \frac{1}{2}c_{\alpha/2}(2x)$ . Similarly, we calculate  $b(x)$  from the bound  $P_b(X \leq x) = \alpha/2$ , which is equivalent to

$$P_b(X \geq x + 1) = 1 - \frac{\alpha}{2}.$$

- (c) Each individual has probability  $p = 1 - e^{-\lambda t}$  of dying during this time, and the events are independent. Since there are  $n$  individuals at risk,  $k$  may be seen as a sample from a binomial distribution with parameters  $(n, p)$ . If  $np$  is moderately small, and  $n$  is moderately large, then this distribution is approximately Poisson with parameter  $n(1 - e^{-\lambda t})$ , which for moderately small values of  $\lambda t$  is approximately  $nt\lambda$ . A confidence interval for  $\lambda$  will thus be  $1/nt$  times the corresponding confidence interval for the Poisson parameter. A slightly better approximation, will be

$$\left( \ln \left[ 1 - \frac{1}{2n} c_{\alpha/2}(2k) \right], \ln \left[ 1 - \frac{1}{2n} c_{1-\alpha/2}(2k + 2) \right] \right),$$

which avoids the approximation of  $e^{-x}$  by  $1 - x$ . This will still be only approximate, since it depends on replacing the binomial by Poisson distribution, and so will not be good when  $n$  is small. However, when  $n$  is large but  $np$  small — which we expect to be the case when  $n$  is large and  $k$  is not very large — the Poisson approximation beats the normal approximation. In particular, when  $k = 0$ , the normal approximation yields a confidence interval which extends into the negative, which makes little sense.

- (d) In table B.1 we see the (exact) confidence intervals for a Poisson random variable. In fact, since for 0 observed events there is no lower bound except 0, it makes sense to use a one-sided confidence interval, with upper bound  $c_{.95}(2)/2 = 3.00$  instead of  $c_{.975}(2)/2 = 3.69$ . (Of course, we always have the option of using asymmetric confidence intervals.)

In Table B.2 we take the corresponding row of the previous table (for the number of deaths in that period), and divide it by 10 ( $= 2 \times 5$  years) times the number at risk. The normal confidence intervals are based on the Standard Error computation

$$SE(\hat{\mu}) = \sqrt{\mu/\tilde{\ell}} \approx \sqrt{\hat{\mu}/\tilde{\ell}},$$

where  $\tilde{\ell}$  is the total time at risk. A symmetric 95% confidence interval is then  $\hat{\mu} \pm 1.96SE$ .

# events observed	lower	upper
0	*	3.69
1	0.025	5.57
2	0.242	7.22
3	0.619	8.77
4	1.09	10.2
5	1.62	11.7
6	2.20	13.1
7	2.81	14.4
8	3.45	15.8

Table B.1: 95% confidence intervals for Poisson distribution.

age	# deaths	# at risk	lower	upper
0–4	2	22	.0022	.066
5–9	3	20	.0062	.088
10–14	5	17	.019	.137
15–19	8	12	.058	.263
20–24	3	4	.031	.438
25–29	1	1	.0051	1.11

Table B.2: 95% confidence intervals for mortality rates in dinosaur data.

age	$\tilde{\ell}_x$	$d_x$	$\hat{\mu}_x$	Standard Error	Confidence Interval
0–4	106	2	.019	.013	(-0.007,.045)
5–9	93	3	.032	.018	(-0.004,.068)
10–14	74	5	.068	.029	(-.010,.126)
15–19	36	8	.22	.07	(.08,.36)
20–24	9	3	.33	.16	(.01,.65)
25–29	3	1	.33	.29	(-.25,.91)

Table B.3: 95% confidence intervals based on the normal approximation.

3. We compute  $z_x$  for each age class as follows:

Age	Exposed to risk $E_x$	Observed deaths $d_x$	Expected deaths $E_x q_x^s$	$z_x$
20–24	35000	35	34	0.17150
25–29	33000	30	29	0.18569
30–34	30000	31	35	-0.67612
35–39	30000	45	52	-0.97072
40–44	31000	84	80	0.44721
45–49	28000	138	130	0.70165
50–54	25000	229	213	1.09630
55–59	23000	360	348	0.64327
60–64	20000	522	505	0.75649

Table B.4: Computing  $z_x$ 

- (a) The
- $X^2$
- statistic is of the form

$$X^2 = \sum z_x^2, \text{ where } z_x = \frac{d_x - E_x q_x^s}{\sqrt{E_x q_x^s (1 - q_x^s)}}. \quad (2)$$

Substituting in the values from Table B.4 we get  $X^2 = 4.344$ . Since this corresponds to  $\chi^2$  with 9 degrees of freedom, we get a p-value of 0.887.

- (b) The cumulative deviations test statistic is

$$Z = \frac{\sum d_x - E_x q_x^s}{\sqrt{\sum E_x q_x^s (1 - q_x^s)}},$$

which should have approximately  $N(0, 1)$  distribution under the null hypothesis. We compute  $Z = 1.27$ , yielding a p-value (for the 2-sided Z test) of 0.20. This lower p-value makes sense, since there are only two negative deviations, but it is still far too high to be considered any significant evidence that the data did not come from the standard table.

- (c) Sign test: We observe 7 positives out of 9 tries. Under the null hypothesis these 7 should be like
- $P = \text{Binom}(9, \frac{1}{2})$
- . The p-value is

$$P\{0, 1, 2, 7, 8, 9\} = \sum_{0,1,2,7,8,9} \left(\frac{1}{2}\right)^9 \binom{9}{k} = 0.18.$$

While there is no strong evidence that the population differs from the standard population, it looks as though there may under-estimation of mortality using the standard life table. This would be a bad thing for an insurance company, which would then set its premiums too low. Graduation, possibly using  $q_x^0 = b + a q_x^s$  may help.

4. (a) Crude estimates from the data are subject to stochastic fluctuation. Smoothing (graduating) the estimates may make more reliable predictions.
- (b)  $\mu_x = a + b e^{\alpha x}$  for Gompertz-Makeham. This is generally considered a reasonable model for the hazard rate (force of mortality) from middle age onward. Note, though, that the mortality rate doubling times (which would be approximately constant under Gompertz-Makeham) lengthen progressively. The parameters  $a, b, \alpha$  will have to be fitted from the data.

We apply the chi-squared test. To begin with, we combine the last two rows to have  $\geq 5$  expected deaths in each row. The last row becomes

99    17.5    5    0.2857    0.3027    - 0.1293

(We interpolate by weighting the two rows by their central exposed to risk.) The  $\chi^2$  statistic is then 4.96 on 8 observations. Since we have estimated 3 parameters, we compare this to the table with 5 degrees of freedom, obtaining p-value 0.42.

To test for bias we use the cumulative deviations test, obtaining  $Z = 0.96$ , and a p-value of 0.3375. Thus, the model seems to fit. Notice that graduated hazard is generally lower — it is strongly affected by the mortality plateau a very late ages — which would lead to an overestimate of benefits paid. This is a relatively good error to make, though it would be reversed if the company were selling life insurance!