A.3 Census approximation, graduation, goodness of fit

1. 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 $[K, K+N+1]$. In this exercise, we use the "census approximation" with "census date" 1 January, which gives useful approximations under some assumptions.

   (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 $t = K, K+1, \ldots, K+N+1$. Calculate E_x^c in terms of $P_{x,t}$, $t = K, K+1, \ldots, K+N+1$. 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 with x nearest birthday to death;
      or $d_x^{(3)}$ = # deaths with x next birthday after death;
      or $d_x^{(4)}$ = # deaths in calendar year of the xth birthday;
      or $d_x^{(5)}$ = # deaths with x last birthday before last policy anniversary.

      Draw a Lexis diagram for each case, showing where the deaths occur that are counted for each d_x, and where the corresponding lifelines are that are counted for E_x^c, and describe the resulting estimate of the force of mortality. Explain the definition that you need to use in each case for $P_{x,t}$. State any further assumptions you make.

2. Suppose $X_1, \ldots, X_n$ 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 $(1-\gamma)$-confidence interval for μ is
   $$\left(1/2c_{\gamma/2}(2x), 1/2c_{1-\gamma/2}(2x+2)\right),$$
   where $c_\gamma(d)$ is the γ quantile of the χ^2 distribution; that is, $P\{X \le c_\gamma(d)\} = \gamma$ 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 $(1-\gamma)$-confidence interval for λ is
   $$\left(\frac{1}{2tn}c_{\gamma/2}(2k), \frac{1}{2tn}c_{1-\gamma/2}(2k+2)\right).$$

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.

3. 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 E_x	Observed deaths d_x	standard mortality $q_x^s \times 10^5$
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-test b) sign test c) cumulative-deviations test, commenting on the outcomes.

4. The following is an investigation carried out by a (medium-sized) UK pension scheme into the mortality of its pensioners between 2000-2002.

(a) Explain why the crude rates are usually graduated.

(b) The data used to produce the crude rates and the proposed graduated rates are as follows.

Age x	Central ExpRisk E_x^c	Deaths d_x	crude hazard $\mu_{x+0.5}$	graduated hazard $\overset{\circ}{\mu}_{x+0.5}$	z_x
60–64	1388.9	10	0.0072	0.0061	0.5249
65–69	1188.8	17	0.0143	0.0131	0.3615
70–74	880.5	28	0.0318	0.0262	1.0266
75–79	841.6	34	0.0404	0.0487	-1.0912
80–84	402.8	41	0.1018	0.0839	1.2394
85–89	123.9	19	0.1533	0.1338	0.5949
90–94	27.9	7	0.2509	0.1975	0.6346
95–99	10.0	3	0.3000	0.2706	0.1787
100+	7.5	2	0.2666	0.3455	-0.3673

Assume the Gompertz-Makeham model has been used for graduation. Is this a sensible choice? Test the proposed graduation for

i) Overall goodness of fit; and ii) Bias.