MATH3085/6143 Survival Models – Worksheet 6 Solutions

1. i)
   - The central exposed to risk at age x, E_x^C, is the observed waiting time in the age range $[x, x+1)$ of all individuals under study, where observation ends either with death, censoring or the end of study period.
   - In aggregate data, the central exposed to risk can be estimated by (an estimate of) the number of lives exposed to risk at the mid-point of the rate interval.
   - The initial exposed to risk at age x, E_x^0, is the number of survivors to exact age x, and therefore does not account for the exact timing of deaths.
   - The calculation of initial exposed to risk in aggregate data typically requires adjustments (from central exposed to risk) for those lives who die.
   - It may be approximated as $E_x^C + d_x/2$, where d_x is the number of deaths of individuals aged x through the assumption that deaths are uniformly distributed (UDD) across each year of age.

   ii) The age definition used for both deaths and exposed to risk is the same since the standard approaches we have seen previously use age at last birthday, so no adjustment is necessary. Using the census approximation, we have that

$$E_x^C = \int_{t_0}^{t_1} P_x(t)\mathrm{d}t \approx \frac{1}{2}[P_x(t_0) + P_x(t_1)].$$

   The initial exposed to risk, E_x^0, can then be obtained using the approximation

$$E_x^C + 0.5d_x,$$

   through the assumption of UDD across each year of age. Therefore, at age 22 we have

$$E_{22}^0 = \frac{1}{2}(P_{22}(2005) + P_{22}(2006)) + \frac{1}{2} \times d_{22} = \frac{150+160}{2} + \frac{20}{2} = 165,$$

   and for age 23,

$$E_{23}^0 = \frac{1}{2}(P_{23}(2005) + P_{23}(2006)) + \frac{1}{2} \times d_{23} = \frac{160+155}{2} + \frac{25}{2} = 170.$$

   Hence,

$$\tilde{q}_{22} = \frac{d_{22}}{E_{22}^0} = \frac{20}{165} = 0.1212$$

   and

$$\tilde{q}_{23} = \frac{d_{23}}{E_{23}^0} = \frac{25}{170} = 0.1471.$$

   iii) Using the notation from Chapter 14, that is suppose $x + a_i$ is the age at which individual i enters the study (the latest of the date of individual i's 22^{nd} birthday and the start of calendar year 2005), and $x + b_i$ denotes the age at which individual ceased to be observed if death had not happened (the earliest of the date of individual i's 23^{rd} birthday and the date of individual i's exit from study for reasons other than death). The initial exposed to risk is then easily calculated as $\sum_i (b_i - a_i)$.

   Alternatively, one can also start with the approximation of central exposed to risk using $\sum_i (b_i' - a_i)$, where $x + b_i'$ is the age at which individual i leaves the study including deaths. The initial exposed to risk can then be approximated as central exposed to risk plus half the number of individuals who died during the study.

2. i) The census approximation is used when the exact dates of entry and exit from observation are not recorded, but rather policies in-force or census data are available at particular dates (for example total number of individuals aged x at the end of each year). In particular, define $P_x(t)$ to be the number of lives under observation aged x last birthday, at time t, the central exposed to risk over the interval considered $[t_0, t_f]$ can be computed as

$$E_x^C = \int_{t_0}^{t_f} P_x(t)\mathrm{d}t,$$

where t_0 and t_f are the earliest and latest census dates respectively. Moreover, suppose also that we have census data at t_1, t_2, ..., t_{f-1}, then we may partition the integral into sum of integrals,

$$E_x^C = \int_{t_0}^{t_f} P_x(t)\mathrm{d}t = \int_{t_0}^{t_1} P_x(t)\mathrm{d}t + \int_{t_1}^{t_2} P_x(t)\mathrm{d}t + \cdots + \int_{t_{f-1}}^{t_f} P_x(t)\mathrm{d}t.$$

Now by employing the key assumption that $P_x(t)$ can be well approximated by a linear function between census dates, the census approximation of E_x^C can be expressed as

$$\begin{aligned} E_x^C &= \frac{t_1 - t_0}{2}[P_x(t_0) + P_x(t_1)] + \frac{t_2 - t_1}{2}[P_x(t_1) + P_x(t_2)] + \cdots + \frac{t_f - t_{f-1}}{2}[P_x(t_{f-1}) + P_x(t_f)] \\ &= \sum_{i=1}^{f} \frac{t_i - t_{i-1}}{2}[p_x(t_{i-1}) + P_x(t_i)] \end{aligned}$$

through repeated application of trapezium rule.

ii) Denote E_{45}^C as the exposed to risk for 1 January 2011 - 31 December 2011, age 45 last birthday and $P_x(t)$ as the number of policies in-force age x last birthday on 1 January in year t. If the number of policies in-force is recorded on 1 January, then the required E_{45}^C is easily calculated as

$$E_x^C = \frac{1}{2}[P_{45}(2011) + P_{45}(2012)]$$

through the census formula. However, as we shall see in a moment, the above calculation is complicated by different company's age definition and the timing at which the number of policies is recorded.

For Company A: The data are recorded on 1 January in each year, but the age definition used is age nearest birthday, and so require adjustments to transform them into age last birthday. For example, to obtain $P_{45}(2011)$ (age range $[45, 46]$), we know:

- Half is contributed by the number of policies in-force at age 45 nearest birthday (age range $[44.5, 45.5]$).
- Another half is contributed by the number of policies in-force at age 46 nearest birthday (age range $[45.5, 46.5]$).

Hence, by assuming that birthdays are uniformly distributed through the calendar year, then

$$P_{45}(2011) = \frac{1}{2}(5920 + 5993) = 5956.5.$$

Similarly,

$$P_{45}(2012) = \frac{1}{2}(5911 + 5988) = 5949.5.$$

Finally,

$$E_{45}^C = \frac{1}{2}[P_{45}(2011) + P_{45}(2012)] = \frac{1}{2}(5956.5 + 5949.5) = 5953.$$

For Company B: The age definition used is age last birthday, which is consistent with the required E_{45}^C. But the data are recorded on 31 March in each year, so we need to adjust the data to match

the $P_x(t)$, which is defined on 1 January in each year. In order to obtain $P_{45}(2011)$, we need to perform linear interpolation by assuming linearity in the number of policies in-force across policy years (between 31 March in each year). Thus, by performing linear interpolation within the years 2010 and 2011, we obtain

$$P_{45}(2011) = \frac{1}{4} \times 3939 + \frac{3}{4} \times 3921 = 3925.5.$$

Similarly,

$$P_{45}(2012) = \frac{1}{4} \times 3921 + \frac{3}{4} \times 3930 = 3927.75.$$

Hence,

$$E_{45}^{C} = \frac{0.25}{2}(3925.5 + 3921) + \frac{0.75}{2}(3921 + 3927.75) = 3924.09375.$$

For Company C: In this case, both the age definition and the timing of recorded data are inconsistent with the exposure we are trying to estimate. However, the adjustments required are rather straightforward here since we only need to realign the age and time accordingly. Specifically,

$P_{45}(2011) =$ number of policies in-force age 46 next birthday at the end of 31 Dec 2010 $= 9237$,

$P_{45}(2012) =$ number of policies in-force age 46 next birthday at the end of 31 Dec 2011 $= 9252$.

Thus,

$$E_{45}^{C} = \frac{1}{2}[P_{45}(2011) + P_{45}(2012)] = \frac{1}{2}(9237 + 9252) = 9244.5$$

3. i) In order to test for the overall goodness of fit, the chi-squared test is used. In particular, we are interested to the null hypothesis

$$H_0 : \mu_x = m_x^S.$$

Under the null hypothesis that the observed rates come from a population with underlying mortality equal to the PMA92C20 rates, the test statistic

$$\sum_x \frac{(d_x - E_x^C m_x^S)^2}{E_x^C m_x^S} = \sum_x \frac{(\hat{m}_x - m_x^S)^2}{\frac{m_x^S}{E_x^C}}$$

follows χ_v^2, where v is the degree of freedom computed as the number of age groups. Thus, we extend the table in the question by including two more quantities.

Age, x	$\frac{\hat{m}_x - m_x^S}{\sqrt{\frac{m_x^S}{E_x^C}}}$	$\frac{(\hat{m}_x - m_x^S)^2}{\frac{m_x^S}{E_x^C}}$
70	1.243	1.546
71	0.062	0.004
72	-0.025	0.001
73	-0.0275	0.076
74	0.284	0.081
75	2.427	5.89
76	4.231	17.901
77	2.809	7.888
78	1.694	2.869
79	1.953	3.815
		40.070

The observed statistic is this 40.07. Since we have 10 age groups here, by comparing this value with the critical value of χ_{10}^2 at 5% significance level (18.31), we reject the null hypothesis since the observed test statistic is substantially greater than the critical value, and conclude that the PMA92C20 rates does not describe the experience of the observed population well.

ii)
- Small but consistent/systematic bias over the whole age range.
- Extreme standardised differences at one or two ages balanced by small deviations at the remaining ages.
- Runs of consecutive deviations of the same sign, indicating that the mortality distribution by age has a different shape from that of the rates being compared with.

4. i) We use the chi-squared test to assess the overall goodness of fit. In other words, we test the null hypothesis H_0 that the observed rates come from a population in which the standard mortality rates are the true underlying rates.

The test statistic is

$$\sum_x \frac{(d_x - E_x^C m_x^S)^2}{E_x^C m_x^S}.$$

So the quantity is computed and illustrated in the table that follows.

x	d_x	$E_x^C m_x^S$	$\frac{d_x - E_x^C m_x^S}{\sqrt{E_x^C m_x^S}}$	$\frac{(d_x - E_x^C m_x^S)^2}{E_s^C m_x^S}$
14	3	4.86	-0.84	0.71
15	8	6.58	0.55	0.30
16	5	6.00	-0.41	0.17
17	14	7.95	2.15	4.62
18	17	9.74	2.33	5.43
19	9	6.67	0.90	0.81
20	15	9.47	1.80	3.24
21	10	11.37	-0.41	0.17
22	10	11.07	-0.32	0.10
				15.55

Now under the null hypothesis, the test statistic follows χ_9^2, which has a critical value of 16.92 at 5% significance level. Hence, since

$$\sum_x \frac{(d_x - E_x^C m_x^S)^2}{E_x^C m_x^S} = 15.55 < 16.92,$$

we do not reject the null hypothesis, i.e. there is no significance evidence against the hypothesis that the observed rates come from a population in which the standard mortality rates are the true underlying rates.

ii) To test for bias, we can use either the sign test or the cumulative deviations test.

Sign test

Denote U_x as an indicator to detect ages at which the observed rate is greater than the standard rate being compared:

$$U_x = \begin{cases} 1 \text{ if } \hat{m}_x \geq m_x^S \text{ (or equivalently } d_x \geq E_x^C m_x^S) \\ 0 \text{ if } \hat{m}_x < m_x^S \text{ (or equivalently } d_x < E_x^C m_x^S) \end{cases}.$$

Under the null hypothesis $H_0 : \mu_x = m_x^S$, we have the normal approximation,

$$S \sim N\left(\frac{9}{2}, \frac{9}{4}\right).$$

Hence, the test statistic is

$$Z = \frac{S \pm \frac{1}{2} - \frac{9}{2}}{\left(\frac{9}{4}\right)^{1/2}}.$$

H_0 is rejected if $|z| > 1.96$, where z is the observed value of Z. Here, observed S is 5, thus

$$z = \frac{5 - \frac{1}{2} - \frac{9}{2}}{\left(\frac{9}{4}\right)^{1/2}} = 0 < 1.96.$$

Therefore, we do not reject the null hypothesis.

Cumulative deviations test

The test statistic follows

$$Z = \frac{\sum_x (d_x - E_x^C m_x^S)}{\sqrt{\sum_x E_x^C m_x^S}} \sim N(0,1)$$

under $H_0 : \mu_x = m_x^S$. We reject the null hypothesis at 5% significance level if $|z| > 1.96$, where z is the observed value of Z. Using the data given, the test statistic is calculated as

$$z = \frac{17.29}{\sqrt{73.71}} = 2.01 > 1.96.$$

Hence, there is a significant evidence at 5% level that systematic bias exists in this part of the age range (as opposed to conclusion drawn from sign test). However, this conclusion should be treated with extreme caution, when applying multiple tests to the same hypothesis.

5. i) The chi-squared test is used to assess the overall goodness of fit. We test the null hypothesis H_0 the the observed rates come from a population in which the graduated mortality rates are the true underlying rates.

Denote $\hat{m}_x$ as the graduated rates, the test statistic is computed and illustrated in the table that follows.

Age x	d_x	$E_x^C \hat{m}_x$	$\frac{d_x - E_x^C \hat{m}_x}{\sqrt{E_x^C \hat{m}_x}}$	$\frac{(d - E_x^C \hat{m}_x)^2}{E_x^C \hat{m}_x}$
40	4	3.0816	0.5232	0.2737
41	4	5.4211	-0.6103	0.3725
42	12	5.7974	2.5760	6.6360
43	7	7.1646	-0.0615	0.0038
44	11	9.2008	0.5932	0.3518
45	7	6.0424	0.3896	0.1518
46	12	9.7440	0.7227	0.5223
47	16	9.6845	2.0294	4.1184
48	15	11.4765	1.0401	1.0818
49	10	11.0745	-0.3229	0.1043
				13.6163

Since the graduation is performed using a linear regression, we lose 2 degrees of freedom (two parameters corresponding to intercept and slope are estimated). Therefore, under H_0, the test statistic follows

$$\sum_x \frac{(d_x - E_x^C \hat{m}_x)^2}{E_x^C \hat{m}_x} \sim \chi_8^2.$$

Hence, we do not reject the null hypothesis as the observed test statistic (13.62) is smaller than the critical value of χ_8^2 at 5% significance level (15.51), and conclude that the graduated rates provide a satisfactory fit to the observed rates.

ii) Under the null hypothesis, the individual standardised deviations follow

$$\frac{d_x - E_x^C \hat{m}_x}{\sqrt{E_x^C \hat{m}_x}} \sim N(0,1).$$

Two moderately large standardised deviations (2.5760 and 2.0294) might give us cause for concern, as the probability of 2 observations outside the range $(-2, 2)$ for a sample of size 10 is only about 2%. However, a formal test is not recommended here.