B.2 Life expectancy, multiple decrements model, and introduction to survival analysis

1. (a) See lecture notes.

   (b) There is right censoring: The depression may not have recurred at the time that the study ended, or the patient died or dropped out. There is left truncation: The first episode of depression made the patients eligible for the study, but not immediately. Thus, the event of interest — the recurrence of depression — could already have happened before the patient was enrolled in the study.

   (c) This study design involves right truncation: The entire study population has already experienced the event of interest (AIDS diagnosis). Any individual whose incubation period extended beyond the truncation time would not have appeared in the study.

2. (a) We make the approximation that those who die between x and $x+t$ survive for $t/2$ years on average, or $(t-1)/2$ complete years. Then the contribution to the life expectancy e_x from those who die before $x+t$ is ${}_tq_x \cdot t/2$, and that from those who survive to $x+t$ is $(1-{}_tq_x)\,(t+e_{x+t})$. We have

$$e_x \approx {}_tq_x \cdot \frac{t-1}{2} + (1-{}_tq_x)\,(t+e_{x+t})\,,$$

   and rearranging leads to the given formula for ${}_tq_x$. The approximation is reasonable if for example deaths are approximately uniform across the interval $(x, x+t)$, which would occur if mortality is constant and low.

   (b) Applying the approximations we get

$$\begin{aligned} q_0 &\approx (33-25+1)/(0+33) = 9/33 = 0.273. \\ {}_4q_1 &\approx (43-33+4)/(1.5+43) = 14/44.5 = 0.315. \\ {}_5q_5 &\approx (41-43+5)/(2+41) = 3/43 = 0.0698. \end{aligned}$$

   Under the assumption of constant mortality on these intervals, then for $x = 1, 2, 3, 4$ we have

$$q_x = 1 - p_x = 1 - ({}_4p_1)^{1/4} = 1 - (1 - {}_4q_1)^{1/4}\,,$$

   and similarly for $x = 5, 6, 7, 8, 9$

$$q_x = 1 - (1 - {}_5q_5)^{1/5}\,,$$

   leading to $q_x = 0.0901$ for $x = 1, 2, 3, 4$ and $q_x = 0.0144$ for $x = 5, 6, 7, 8, 9$.

   (c) We have estimated that $q_0 = 27.2\%$ of infants died in their first year. Assuming mortality constant across the year, this corresponds to a mortality rate of $\mu = -\log(1-q_0) = 0.3175$. We are told that 20% of all infants died from dysentery, which is a proportion $20/27.2 = 0.7353$ of all the deaths. Again, we assume that this proportion is constant across the year. Under the competing risks framework, we then have a rate $\mu^D = 0.7353\mu = 0.2334$ for deaths from dysentery, and $\mu^O = \mu - \mu^D = 0.0840$ for deaths from other causes.

   If we can eliminate the risk from dysentery, we would be left with the mortality rate μ^O. This leads to a new first-year mortality probability of $\tilde{q}_0 = 1 - \exp(-\mu^O) = 0.08060$.

   If we ignore the further gains due to absence of dysentery later in life, the figure for e_1 remains unchanged. Then we can approximate the revised life expectancy at birth as

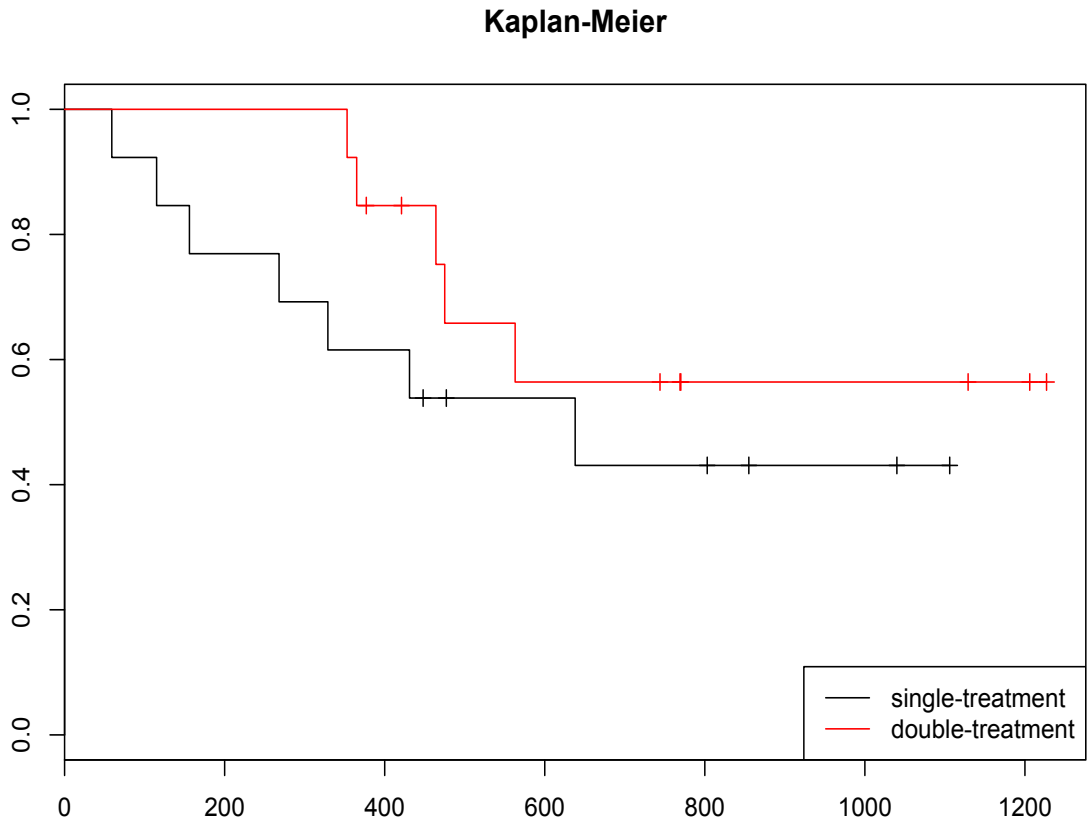
$$\tilde{e}_0 \approx (1+e_1)(1-q_0)$$

   to give a new figure of 31.26 for the (curtate) life expectancy at birth.

3.

```
library(survival)

## a ##

surv_object <- Surv(ovarian$futime, ovarian$fustat)

# To have a look at what has been computed about survival


## b ##
plot(survfit(surv_object~ovarian$rx), main="Kaplan-Meier")

> summary(surv_object)
Call: survfit(formula = Surv(futime, fustat) ~ rx)

#                rx=1
# time n.risk n.event survival std.err lower 95\% CI upper 95\% CI
#   59     13       1    0.923  0.0739        0.789        1.000
#  115     12       1    0.846  0.1001        0.671        1.000
#  156     11       1    0.769  0.1169        0.571        1.000
#  268     10       1    0.692  0.1280        0.482        0.995
#  329      9       1    0.615  0.1349        0.400        0.946
#  431      8       1    0.538  0.1383        0.326        0.891
#  638      5       1    0.431  0.1467        0.221        0.840
#
#                rx=2
# time n.risk n.event survival std.err lower 95\% CI upper 95\% CI
#  353     13       1    0.923  0.0739        0.789        1.000
#  365     12       1    0.846  0.1001        0.671        1.000
#  464      9       1    0.752  0.1256        0.542        1.000
#  475      8       1    0.658  0.1407        0.433        1.000
#  563      7       1    0.564  0.1488        0.336        0.946

#for extra challenge:
plot(survfit(surv_object~ovarian$rx) ,
col=c("black","red"),
main="Kaplan-Meier")
legend("bottomright",
c( "single-treatment", "double-treatment"),
col=c("black","red") , lty=1 )

## c ##

plot(survfit(surv_object~ovarian$rx, type='fleming-harrington'),main="
    Nelson-Aalen")


### The rest is to do this more 'by hand', computing the relevant
    quantities
### and directly computing the Nelson-Aalen estimator.

attach(ovarian)

x=order(futime)
futime=futime[x]
fustat=fustat[x]
rx=rx[x]
```

```
ns=rev(cumsum(rev(rx==1)))
nd=rev(cumsum(rev(rx==2)))
hs=round(fustat*(rx==1)/ns,2)
hd=round(fustat*(rx==2)/nd,2)

NelsonAalenTable =
subset(data.frame(t_i=futime,  n_single=ns, n_double=nd,
h_single=hs,h_double=hd,A_single=cumsum(hs),A_double=cumsum(hd)), h_
    single+h_double>0)

> NelsonAalenTable
t_i n_single n_double h_single h_double A_single A_double vars vard
1    59        13       13       0.08     0.00     0.08     0.00 0.01 0.00
2   115        12       13       0.08     0.00     0.16     0.00 0.01 0.00
3   156        11       13       0.09     0.00     0.25     0.00 0.02 0.00
4   268        10       13       0.10     0.00     0.35     0.00 0.03 0.00
5   329         9       13       0.11     0.00     0.46     0.00 0.04 0.00
6   353         8       13       0.00     0.08     0.46     0.08 0.04 0.01
7   365         8       12       0.00     0.08     0.46     0.16 0.04 0.01
10  431         8        9       0.12     0.00     0.58     0.16 0.06 0.01
12  464         6        9       0.00     0.11     0.58     0.27 0.06 0.03
13  475         6        8       0.00     0.12     0.58     0.39 0.06 0.04
15  563         5        7       0.00     0.14     0.58     0.53 0.06 0.06
16  638         5        6       0.20     0.00     0.78     0.53 0.10 0.06
```



The standard errors are in the code printout above. For type 1 the variance estimate for the Nelson–Aalen estimator is 0.04 at $t = 400$; for type 2 it is 0.01. So the corresponding standard errors for the cumulative hazard are 0.2 and 0.1. The standard errors for survival are obtained from multiplying these by $\hat{S}(400)^2$, obtaining 0.13 and 0.097. The standard errors computed by the `survfit` function for the Kaplan–Meier estimatorare in the printout above. They are 0.135 and 0.100.

4. The log likelihood is

$$\ell(p) = \log \binom{n}{x} + x \log p + (n - x) \log(1 - p).$$

This has solution $0 = \ell'(\hat{p}) = x/\hat{p} - (n - x)/(1 - \hat{p})$, implying $\hat{p} = x/n$. We know that the variance of a binomial random variable is $np(1 - p)$. Substituting $\hat{p}$ for p yields the estimate

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}(x/n) = n^{-2} \mathrm{Var}(x) = n^{-1} p(1 - p) = n^{-1} \frac{x}{n} \frac{n - x}{n} = \frac{x(n - x)}{n^3}.$$

If all the censoring occurs at $t = 0$ then the number of individuals at risk of dying in $(0, t)$ is actually $n(t) + d(t)$. Thus alive at time t is binomial with parameters $n = n(0) = n(t) + d(t)$ and $p = S(t)$. The MLE for p is thus

$$\hat{S}(t) = \hat{p} = \frac{n(t)}{n(t) + d(t)} = \frac{n(t)}{n(0)}.$$

(If the censoring all happens at time 0, then the number at risk at time 0+ will be the same as the sum of the number who die up to time t, and the number still at risk at time t.) The variance

estimate is

$$\frac{d(t)n(t)}{n(0)^3} = n(t)^{-1}\frac{d(t)}{n(0)}\frac{n(t)}{n(0)}\frac{n(t)}{n(0)} = n(t)^{-1}(1-\hat{S}(t))\hat{S}(t)^2.$$

Greenwood's estimate in the case of no censoring is

$$\begin{aligned}
\operatorname{Var}\hat{S}(t) &\approx \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)} \\
&= \hat{S}(t)^2 \sum_{t_i \le t} \frac{n_{i+1} - n_i}{n_i(n_{i+1})} \\
&= \hat{S}(t)^2 \sum_{t_i \le t} \left(\frac{1}{n_{i+1}} - \frac{1}{n_i}\right) \\
&= \hat{S}(t)^2 \left(\frac{1}{n_j} - \frac{1}{n_0}\right) \\
&= \hat{S}(t)^2 \frac{d(t)}{n(t)n(0)} \\
&= n(t)^{-1}\hat{S}(t)^2(1-\hat{S}(t))
\end{aligned}$$

as before.

5. The average mortality rate due to cancer between ages 90 and 95

$$ {}_5m_{90}^{\text{Cancer}} = -\frac{1}{5}\log(1 - {}_5q_{90}^{\text{Cancer}*}) = 0.0204.$$

The average mortality rate due to other causes between ages 90 and 95

$$ {}_5m_{90}^{\text{Other}} = -\frac{1}{5}\log(1 - {}_5q_{90}^{\text{Other}*}) = 0.0283.$$

The total mortality rate is thus .0487. Thus

$$ {}_5q_{90} = 1 - \mathrm{e}^{-5\cdot 0.0487} = 0.2161.$$

The fraction of deaths due to cancer is $0.0204/0.0487 = 0.4189$. We require the competing risks assumption — causes of death act independently — and that the mortality rates are fairly constant (in particular, rates for different causes remain in the same proportions) over the 5-year period.

6. (a) Right censoring and left truncation.

   (b) If individuals who enter at age x are considered immediately available to count at risk at age x, and those who die at age x are also at risk.

Age	65	66	67	68	69	70	71	72	73	74	75
# at risk	3	9	11	13	14	17	14	12	12	8	4

   (c) The survival-analysis framework is designed for continuous distribution of event times, so that there are few ties, and the observed event times are not at any regular intervals. We treat the recorded event times as exact observations. We compute the Kaplan–Meier $\hat{S}$ and Nelson–Aalen $\widetilde{S}$ estimators for survival from time 70.

Age (t_i)	n_i	d_i	h_i	$\hat{S}(t_i)$	$\widetilde{S}(t_i)$
70	17	4	0.235	0.765	0.790
72	12	1	0.083	0.701	0.727
73	12	3	0.250	0.526	0.566
74	8	4	0.500	0.263	0.343
75	4	1	0.250	0.197	0.268

Since we are observing the population age in discrete intervals it seems natural to think of it as a life table. The discrete approach to life-table estimation requires that we know the number initially at risk E_x^0, the number present in the population on their x-th birthday. We don't actually have those counts, but treating these as equivalent to the n_i, counts of individuals at risk — essentially the same as pretending that individuals all enter and exit the population on their birthdays — is equivalent to the Kaplan–Meier estimates.

Age (x)	E_x^0	d_x^{obs}	q_x	d_x	ℓ_x
70	17	4	0.235	235	1000
71	14	0	0.000	0	765
72	12	1	0.083	64	765
73	12	3	0.250	175	701
74	8	4	0.500	263	526
75	4	1	0.250	66	263

The left-hand side shows the data, including observed numbers of deaths (denoted d_x^{obs}), and the right-hand two columns show the computed life table, starting with a radix of 1000. The one notable difference, which is purely a matter of convention, is that the estimated survival probability $\hat{S}(t_i)$ is replaced by the number surviving ℓ_x, so all multiplied by the radix 1000; and these are shifted down by one row, because of the convention that survival probabilities are probabilities of survival **past** a certain age, whereas life-table survival numbers are survival **to** that age. Of course, this distinction is irrelevant when survival times are continuous, but seem different when times are grouped into large intervals where a significant fraction of events occur.

The continuous method requires that we estimate the central exposed to risk $\mathbb{E}_x^c$. The actuarial estimator tells us to count those who are censored or died as having had half a year at risk, and count those who entered at a given age as having half a year at risk in that year. The effect is that E_x^c is the mean of the numbers we estimated at risk at ages x and $x+1$ — effectively, the actuarial estimator. This yields the table

Age (x)	E_x^c	d_x^{obs}	m_x	q_x	d_x	ℓ_x
70	15.5	4	0.258	0.227	227	1000
71	13	0	0.000	0.000	0	773
72	12	1	0.0833	0.0800	62	773
73	10	3	0.3	0.259	184	711
74	6	4	0.667	0.487	257	527
75	3.5	1	0.286	0.249	67	270

The continuous method is in principle the same as the Nelson–Aalen estimator. We note, though, that the estimates we obtained from Nelson–Aalen were quite different. This is a result of overestimating the time at risk, by ignoring the partial years at risk. We observe (as discussed in section 2.8) that the discrete method (and Kaplan–Meier) produces very similar results to the continuous method with the actuarial estimator. This suggests that the Nelson–Aalen estimator is not very good for highly discretised data. It is like calculating life tables with the continuous method, but neglecting to apply the actuarial estimator.

Note that we might reasonably suggest that age is not a sensible time variable here, since mortality is largely determined by time since diagnosis. We see that the estimator of survival past age 78 is 0, since the single individual who happened to be in the study at that age died. This despite the fact that there are other individuals who entered later and survived to much older ages. We might reasonably look instead at the *time-on-test* as time variable. We would then get the following calculation:

t_j	n_j	d_j	h_j	$\widehat{S}(t_j)$	$\widetilde{S}(t_j)$
2	27	1	0.04	0.96	0.96
3	22	6	0.27	0.70	0.73
4	16	8	0.50	0.35	0.44
5	8	5	0.62	0.13	0.24

(d) We refer back to the Kaplan–Meier estimator, based on whole-year counts. Our central estimate for the probability of surviving from age 70 to age 75 is $\hat{S}(74) = 0.263$. We estimate the variance of $\log \hat{S}(74)$ by equation (4.6) to be

$$\begin{aligned} \text{Ƶ}(74) = \sum_{t_i \le 74} \frac{d_i}{n_i(n_i - d_i)} &= \frac{4}{17 \cdot 13} + \frac{1}{12 \cdot 11} + \frac{3}{12 \cdot 9} + \frac{4}{8 \cdot 4} \\ &= 0.178, \end{aligned}$$

so the standard error is $\sqrt{0.178} = 0.422$. Thus an approximate 95% confidence interval for $\log S(74)$ is

$$\log(0.263) \pm 0.422 \cdot 1.96.$$

Exponentiating the lower and upper bounds yields a 95% confidence interval for $S(74)$

$$\left(0.263\mathrm{e}^{-0.422 \cdot 1.96}, 0.263\mathrm{e}^{0.422 \cdot 1.96}\right) = (0.115, 0.601).$$

Alternatively, we can use Greenwood's formula $\mathrm{Var}(\hat{S}(74)) \approx \sigma_G^2(74) = \hat{\sigma}^2(74)\hat{S}(74)^2$, yielding the 95% confidence interval for $S(74)$

$$\big(0.263(1 - 0.422 \cdot 1.96), 0.263(1 + 0.422 \cdot 1.96)\big) = (0.0455, 0.481).$$

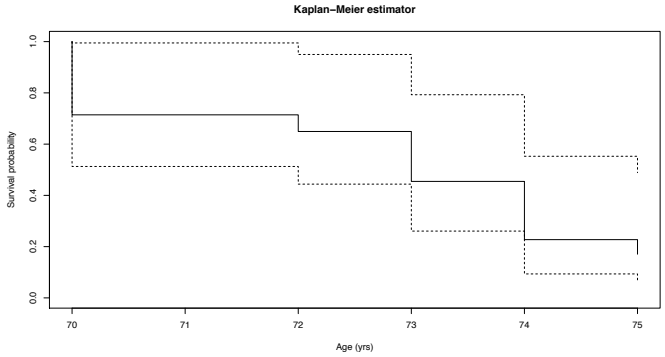
When the standard error is small these confidence intervals will be similar, as $\mathrm{e}^x \approx 1 + x$ for x small. When the standard error is not small (as in this case) because the estimates are based on small samples, all of these methods are somewhat dubious — in particular, the assumption of normality — but the first method is to be preferred, because it avoids one step of approximation (based on the delta method).

We could also use the estimator (4.4) to estimate a standard error for the cumulative hazard based on the Nelson–Aalen estimator, and from there compute a confidence interval for the survival probability. Given the above-mentioned problems with the Nelson–Aalen estimator in this setting, this approach is probably not to be preferred.

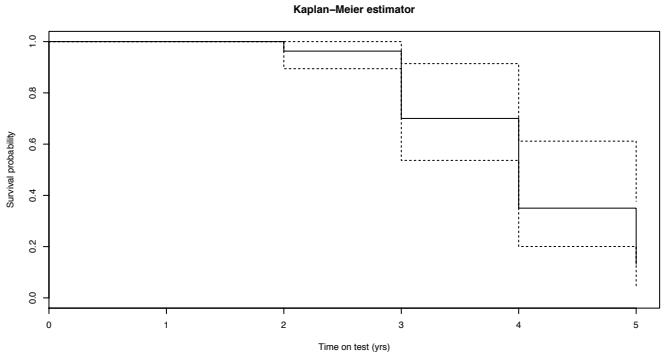
(e)

```
require('survival')
age.entry=c(67,70,70,65,65,73,69,76,66,72,65,71,69,71,68,69,69,66,
73,67,66,69,66,78,66,68,70,66,89,68)
age.exit=c
    (72,71,73,70,68,78,74,78,67,76,70,75,71,74,73,74,71,68,76,68,70,73,

70,81,70,73,74,68,92,72)
delta=c(0,0,1,0,1,1,1,1,0,1,1,1,0,1,0,1,0,0,1,0,1,1,1,1,1,1,1,0,1,1)

clinic.surv=Surv(time=age.entry,time2=age.exit,event=delta) # left-
    truncated, right-censored is default
KM.fit=survfit(clinic.surv~1,subset=(age.exit>=70)) # Survival of
    those present after age 70
plot(KM.fit, firstx=70,xmax=75,ylab='Survival probability',main='
    Kaplan-Meier estimator',xlab='Age (yrs)')

TOT.surv=Surv(time=time.on.test,event=delta)
TOT.fit=survfit(TOT.surv~1)
```

```
14    plot(TOT.fit,ylab='Survival probability',main='Kaplan-Meier
          estimator',xlab='Time on test (yrs)')
```



(a) Survival by age



(b) Survival by time on test