B.2 Life expectancy, graduation, and 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. 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

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

   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

$$\begin{aligned} q_0 &\approx (33-25+1)/(1/2+33) = 9/33.5 = 0.269. \\ {}_4q_1 &\approx (43-33+4)/(2+43) = 14/45 = 0.311. \\ {}_5q_5 &\approx (41-43+5)/(2.5+41) = 3/43.5 = 0.069. \end{aligned}$$

   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 $x = 5, 6, 7, 8, 9$

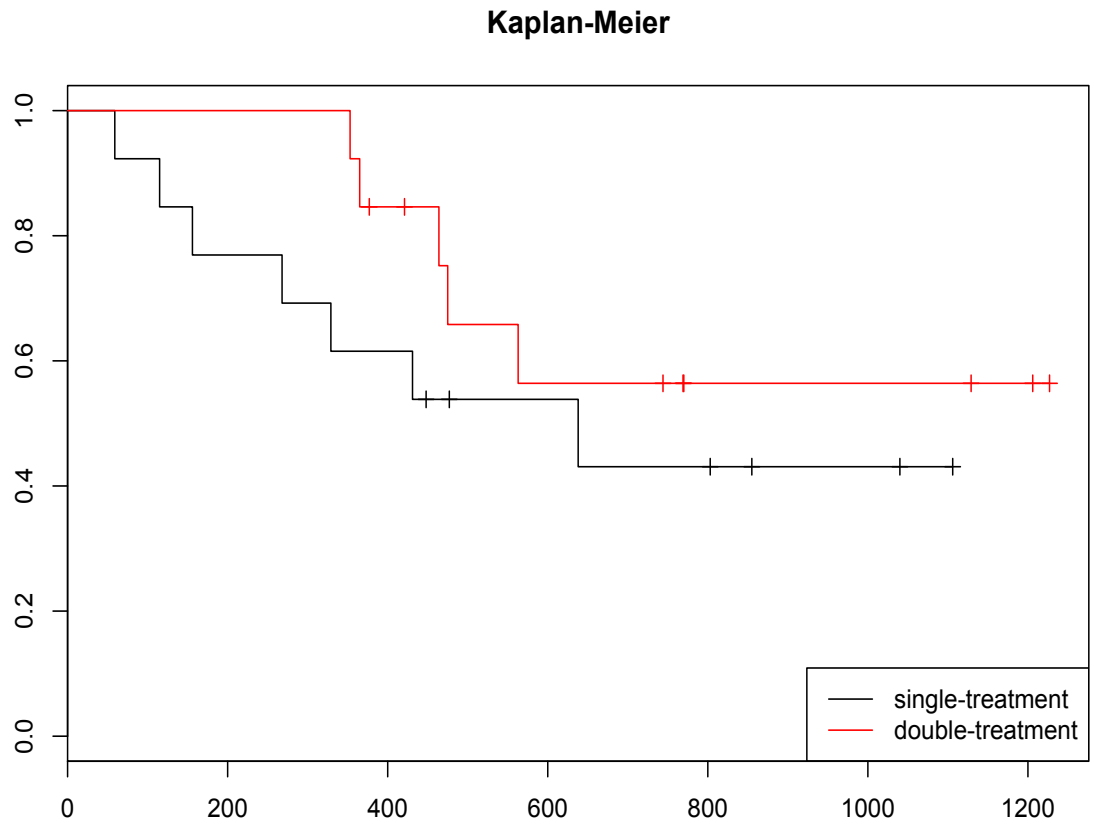
$$q_x = 1 - (1-{}_5q_5)^{1/5}\,,$$

   leading to $q_x = 0.089$ for $x = 1, 2, 3, 4$ and $q_x = 0.014$ for $x = 5, 6, 7, 8, 9$.

3.
```r
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
```

```
77      13  475        6        8     0.00     0.12     0.58      0.39  0.06  0.04
78      15  563        5        7     0.00     0.14     0.58      0.53  0.06  0.06
79      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. (a) Crude estimates from the data are subject to stochastic fluctuation. Smoothing (graduating) the estimates may make more reliable predictions.

   (b) $\mu_x = a + be^{\alpha x}$ 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, α 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 ≥ 5 expected deaths in each row. The last row becomes

   $$99 \qquad 17.5 \qquad 5 \qquad 0.2857 \qquad 0.3027 \qquad -0.1293$$

(We interpolate by weighting the two rows by their central exposed to risk.) The χ^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!

5. (a) Let us write e_x and p_x for the figures in the table, and $\tilde{e}_x$ and $\tilde{p}_x$ for the figures after we change the rates in the first two years.

We have

$$e_0 = p_0 \left(1 + p_1 \left(1 + e_2\right)\right),$$
$$\tilde{e}_0 = \tilde{p}_0 \left(1 + \tilde{p}_1 \left(1 + \tilde{e}_2\right)\right)$$

Since we change only the rates before year 2, we have $e_2 = \tilde{e}_2$. Then solving for $\tilde{e}_0$, we have

$$\tilde{e}_0 = \tilde{p}_0 \left(1 + \tilde{p}_1 \left(\frac{e_0/p_0 - 1}{p_1}\right)\right).$$

With $e_0 = 44.83$, $p_0 = 0.839$, $p_1 = 0.946$, $\tilde{p}_0 = 0.995$, $\tilde{p}_1 = 0.9996$ we obtain $\tilde{e}_0 = 56.12$, an increase of 11.29 years.

(b) It's clear from the table that the rates q_x $(x = 19, \ldots, 25)$ are much larger than we would normally expect. Comparing them with the rates just before and after, it looks like a plausible first approximation would be to replace all these rates by a rate around 0.0035. So we could work with a model where the mortality is constant with $q = 0.0035 = 1 - 0.9965$ for those 7 years. (Of course this is very rough, and there are all sorts of things we ignore including the various effects of the war on mortality, even after it had finished).

We can represent e_0 as a sum $A + B + C$ of three terms:

$A =$ expected number of whole years lived up to age 19,
$B =$ expected number of whole years lived up between 19 and 26,
$C =$ expected number of whole years lived after age 26.

Let us write $\widetilde{A}$, $\widetilde{B}$, and $\widetilde{C}$ for the new values once we change the rates.

The change to the rates between ages 19 and 26 makes no difference to the first term, so $\widetilde{A} = A$.

We have

$$\begin{aligned} C &= {}_{26}p_0 e_{26} \\ &= \frac{\ell_{26}}{\ell_0} e_{26} \\ &= 0.6014 \times 42.75 \\ &= 25.71. \end{aligned}$$

The change to the rates makes no difference to e_{26}, but we have a new value for the probability of surviving to age 26, giving

$$\begin{aligned} \widetilde{C} &= (0.9965)^7 \frac{\ell_{19}}{\ell_0} e_{26} \\ &= (0.9965)^7 \times 0.7309 \times 42.75 \\ &= 30.49. \end{aligned}$$

Finally, we can find B using $B + C = e_{19}\ell_{19}/\ell_0$ as in the previous calculation, so

$$\begin{aligned} B &= \frac{\ell_{19}}{\ell_0} e_{19} - C \\ &= 0.7309 \times 41.57 - 25.71 \\ &= 4.67. \end{aligned}$$

In the new model with constant rate between age 19 and age 26, we can use

$$\begin{aligned} \widetilde{B} &= {}_{19}\tilde{p}_0 \left({}_1\tilde{p}_{19} + {}_2\tilde{p}_{19} + \cdots + {}_7\tilde{p}_{19}\right) \\ &\frac{\ell_{19}}{\ell_0} \sum_{k=1}^{7} 0.9965^k \\ &= \frac{\ell_{19}}{\ell_0} \frac{0.9965 - 0.9965^8}{0.0035} \\ &= 0.7381 \times 6.903 \\ &= 5.10. \end{aligned}$$

The total change in life expectancy at birth is $\widetilde{B} + \widetilde{C} - B - C$ which comes to $30.49 + 5.10 - 25.71 - 4.67 = 5.21$, giving a new life expectancy of around 50.

6. 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

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

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.

7. (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

   We are planning to use the actuarial estimator — so we 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, we get the following counts:

Age	65	66	67	68	69	70	71	72	73	74	75
# at risk	1.5	6.0	9.5	9.5	11.5	13.0	11.5	10.5	9.0	6.0	3.5

   (c) Again, counting whole years at risk for those who enter, die, or are right-censored, we have

Age	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

   The actuarial estimate gives us

Age	n_i	d_i	h_i	$\hat{S}(t_i)$	$\widetilde{S}(t_i)$
70	13.0	4	0.308	0.692	0.735
72	10.5	1	0.095	0.626	0.668
73	9.0	3	0.333	0.418	0.479
74	6.0	4	0.667	0.139	0.246
75	3.5	1	0.286	0.099	0.185

   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 use the whole-year method, rather than the actuarial estimate. Our central estimate for the probability of surviving from age 70 to age 75 is $\hat{S}(74) = 0.343$. Using Greenwood's estimate, we estimate the variance of $\log \hat{S}(74)$ to be

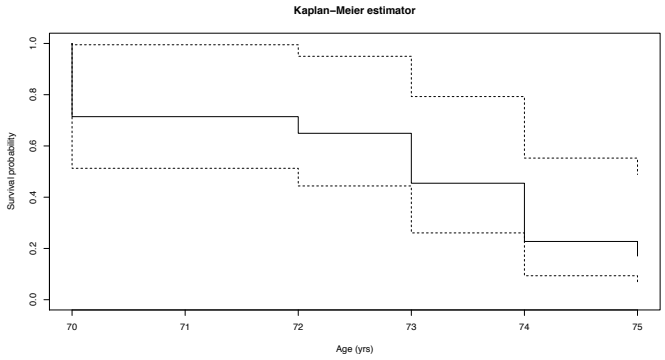
$$\sum_{t_i \leq 74} \frac{d_i}{n_i(n_i - d_i)} = \frac{4}{17 \cdot 13} + \frac{1}{12 \cdot 11} + \frac{3}{12 \cdot 9} + \frac{4}{8 \cdot 4}$$
$$= 0.178,$$

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

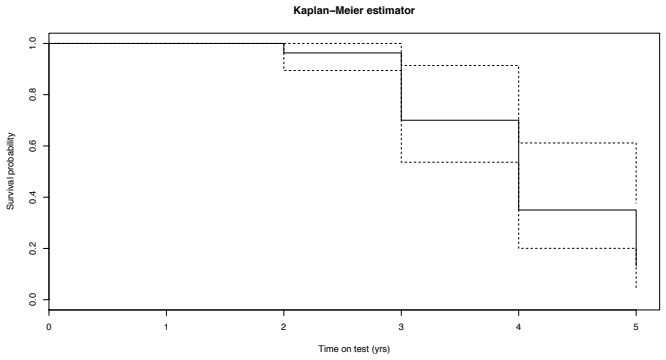
$$\left(0.343e^{-0.422 \cdot 1.96}, 0.343e^{0.422 \cdot 1.96}\right) = (0.150, 0.784).$$

(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)
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