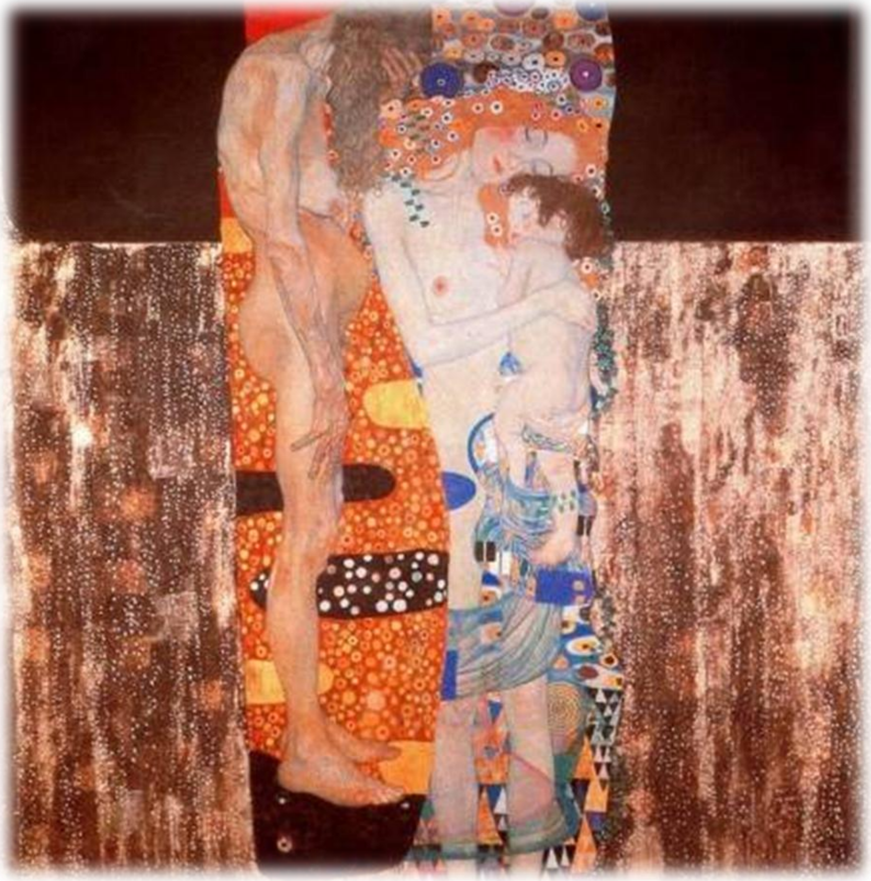
LECTURE 13

Survival analysis

Survival analysis

- Survival analysis is the analysis of *time to an event*: its mean, distribution, differences across groups, effect of covariates

- The most commonly analysed event is death (hence *survival analysis*); but menopause, birth, mechanical failure and other events can be studied too



Survival analysis

- A common problem: right censoring (=when date of a possible event is unknown)
  - lost to follow up
  - study ends before subjects are dead

- Such cases must be *censored* but not excluded
  - censored data can still provide information on survival and risk of death

- Survival analysis is the preferred method when data are censored

Survival analysis

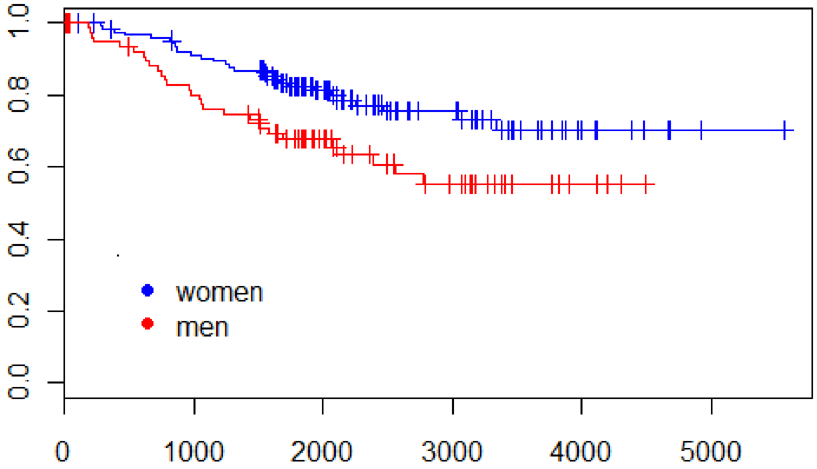
- Some techniques used in survival/mortality studies:

- ***Life tables***: preliminary analysis of survival and probability of death by age

- ***Kaplan-Meier method***: estimation of survival curves from individual death dates; visual comparison of group survival

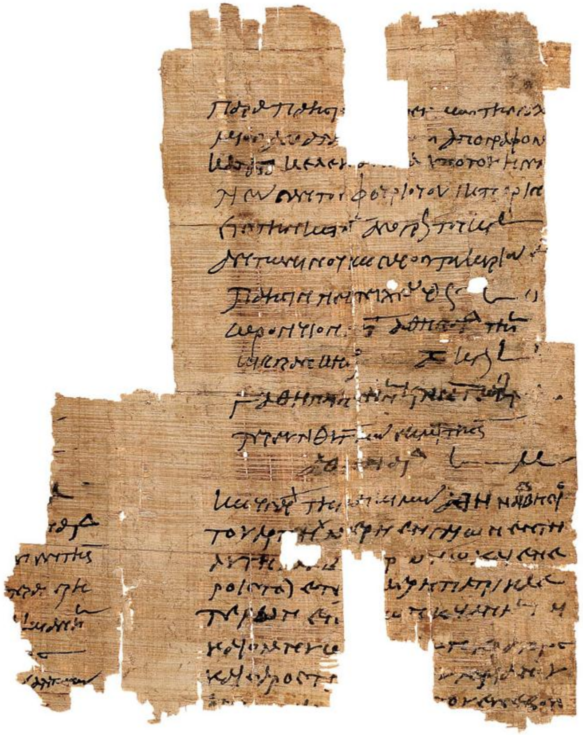
- ***Log-rank test***: tests of statistical differences between two survival curves

- ***Cox (proportional hazard) regression***: estimating effect of factors on between-group differences in survival curves



Life tables

- Life tables are useful for summarising survival patterns and calculating parameters such as age-dependent mortality rates and life expectancy
- They typically display age-dependent probabilities of survival decreasing from 1 to 0 in population
- Ideally based on cohort (longitudinal) data
  - all individuals start at time 0 (initial event: birth, cancer diagnosis etc.), and are followed until they are all dead
- But often we only have current (cross-sectional or census) data
  - it would take over 100 years to study one human cohort!

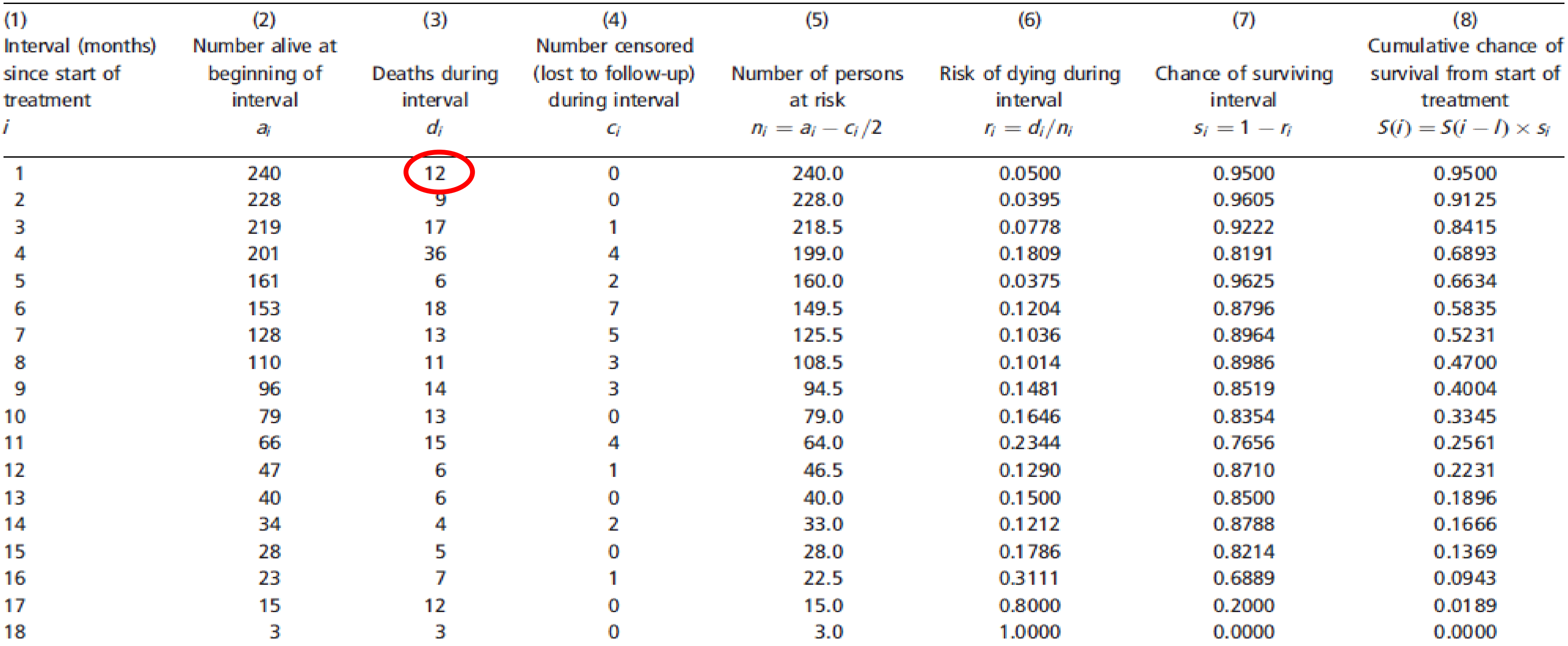


Example: survival of 240 cancer patients

At time t=0 (start of treatment), population is n=240

In interval 1 (t=0 to t=1 month)

- number of deaths = 12
- risk of death (mortality) = 12/240 = 0.05 = 5%
- survival = s_1 = 1 – 0.05 = 95%

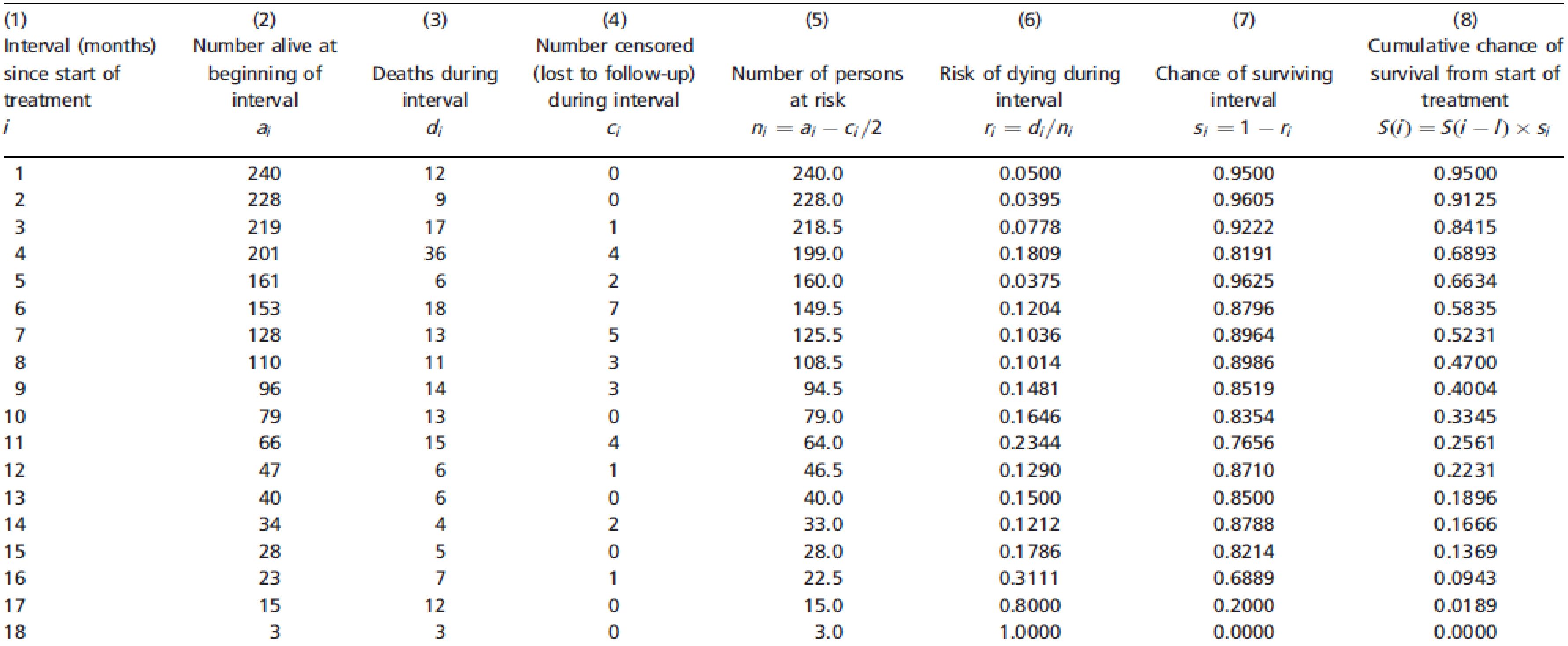


(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Interval (months) since start of treatment	Number alive at beginning of interval	Deaths during interval	Number censored (lost to follow-up) during interval	Number of persons at risk	Risk of dying during interval	Chance of surviving interval	Cumulative chance of survival from start of treatment
i	a_i	d_i	c_i	$n_i = a_i - c_i/2$	$r_i = d_i/n_i$	$s_i = 1 - r_i$	$S(i) = S(i-l) \times s_i$
1	240	12	0	240.0	0.0500	0.9500	0.9500
2	228	9	0	228.0	0.0395	0.9605	0.9125
3	219	17	1	218.5	0.0778	0.9222	0.8415
4	201	36	4	199.0	0.1809	0.8191	0.6893
5	161	6	2	160.0	0.0375	0.9625	0.6634
6	153	18	7	149.5	0.1204	0.8796	0.5835
7	128	13	5	125.5	0.1036	0.8964	0.5231
8	110	11	3	108.5	0.1014	0.8986	0.4700
9	96	14	3	94.5	0.1481	0.8519	0.4004
10	79	13	0	79.0	0.1646	0.8354	0.3345
11	66	15	4	64.0	0.2344	0.7656	0.2561
12	47	6	1	46.5	0.1290	0.8710	0.2231
13	40	6	0	40.0	0.1500	0.8500	0.1896
14	34	4	2	33.0	0.1212	0.8788	0.1666
15	28	5	0	28.0	0.1786	0.8214	0.1369
16	23	7	1	22.5	0.3111	0.6889	0.0943
17	15	12	0	15.0	0.8000	0.2000	0.0189
18	3	3	0	3.0	1.0000	0.0000	0.0000

Example: survival of 240 cancer patients

In interval 2 (t=1 to t=2),

- number at risk (alive at start of interval) = 228; number of deaths = 9
- risk (mortality) = 9/228 = 0.0395 = 3.95%
- survival at interval 2 = s_2= 1 – 0.0395 = 0.9605 = 96.05%
- survival to end of interval 2 = $S_2 = s_1$ x s_2 = 0.95 x 0.9605 = 0.9125 = 91.25%

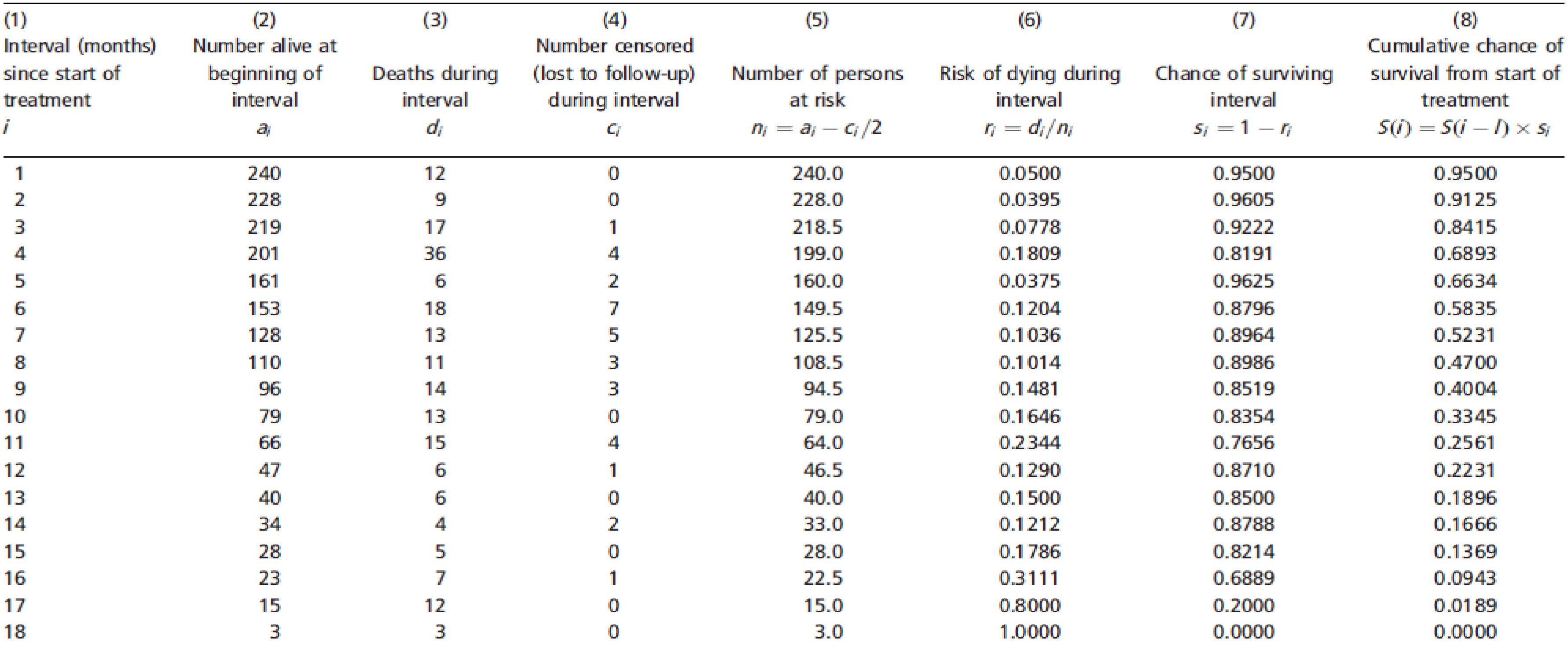


(1) Interval (months) since start of treatment i	(2) Number alive at beginning of interval a_i	(3) Deaths during interval d_i	(4) Number censored (lost to follow-up) during interval c_i	(5) Number of persons at risk $n_i = a_i - c_i/2$	(6) Risk of dying during interval $r_i = d_i/n_i$	(7) Chance of surviving interval $s_i = 1 - r_i$	(8) Cumulative chance of survival from start of treatment $S(i) = S(i - l) \times s_i$
1	240	12	0	240.0	0.0500	0.9500	0.9500
2	228	9	0	228.0	0.0395	0.9605	0.9125
3	219	17	1	218.5	0.0778	0.9222	0.8415
4	201	36	4	199.0	0.1809	0.8191	0.6893
5	161	6	2	160.0	0.0375	0.9625	0.6634
6	153	18	7	149.5	0.1204	0.8796	0.5835
7	128	13	5	125.5	0.1036	0.8964	0.5231
8	110	11	3	108.5	0.1014	0.8986	0.4700
9	96	14	3	94.5	0.1481	0.8519	0.4004
10	79	13	0	79.0	0.1646	0.8354	0.3345
11	66	15	4	64.0	0.2344	0.7656	0.2561
12	47	6	1	46.5	0.1290	0.8710	0.2231
13	40	6	0	40.0	0.1500	0.8500	0.1896
14	34	4	2	33.0	0.1212	0.8788	0.1666
15	28	5	0	28.0	0.1786	0.8214	0.1369
16	23	7	1	22.5	0.3111	0.6889	0.0943
17	15	12	0	15.0	0.8000	0.2000	0.0189
18	3	3	0	3.0	1.0000	0.0000	0.0000

Example: survival of 240 cancer patients

In interval 3 (t=2 to t=3), there is one *censored* individual (lost to follow up)

- including it in population at risk underestimates risk (it may not have died)
- excluding it from population at risk would overestimate risk (it may have died)
- compromise: subtract half the censored individuals from population at risk (i.e. assume they were lost half-way through interval)
- so at start of interval 3, number alive = 219, censored people = 1, people at risk = 219 – 0.5 = 218.5
  - 218.5 is the number used to calculate risk (=17/218.5) and survival s_3



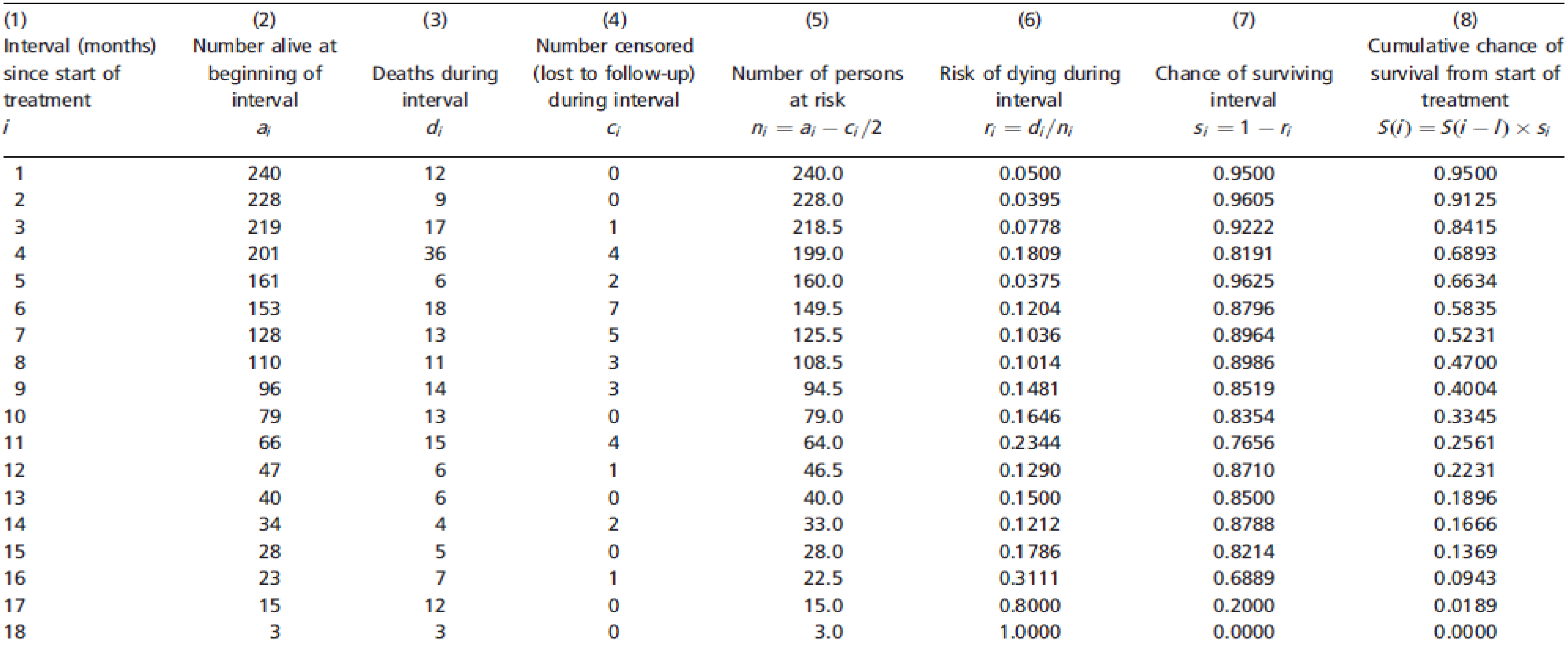
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Interval (months) since start of treatment	Number alive at beginning of interval	Deaths during interval	Number censored (lost to follow-up) during interval	Number of persons at risk	Risk of dying during interval	Chance of surviving interval	Cumulative chance of survival from start of treatment
i	a_i	d_i	c_i	$n_i = a_i - c_i/2$	$r_i = d_i/n_i$	$s_i = 1 - r_i$	$S(i) = S(i-l) \times s_i$
1	240	12	0	240.0	0.0500	0.9500	0.9500
2	228	9	0	228.0	0.0395	0.9605	0.9125
3	219	17	1	218.5	0.0778	0.9222	0.8415
4	201	36	4	199.0	0.1809	0.8191	0.6893
5	161	6	2	160.0	0.0375	0.9625	0.6634
6	153	18	7	149.5	0.1204	0.8796	0.5835
7	128	13	5	125.5	0.1036	0.8964	0.5231
8	110	11	3	108.5	0.1014	0.8986	0.4700
9	96	14	3	94.5	0.1481	0.8519	0.4004
10	79	13	0	79.0	0.1646	0.8354	0.3345
11	66	15	4	64.0	0.2344	0.7656	0.2561
12	47	6	1	46.5	0.1290	0.8710	0.2231
13	40	6	0	40.0	0.1500	0.8500	0.1896
14	34	4	2	33.0	0.1212	0.8788	0.1666
15	28	5	0	28.0	0.1786	0.8214	0.1369
16	23	7	1	22.5	0.3111	0.6889	0.0943
17	15	12	0	15.0	0.8000	0.2000	0.0189
18	3	3	0	3.0	1.0000	0.0000	0.0000

Example: survival of 240 cancer patients

- Generalising: cumulative survival to end of interval *i* is s_1 x s_2 …x s_i

$$S(t) = \prod_{i=1}^{t} s_i$$

- that is, the ***product*** of survival probabilities (column 7 in table) in each of the intervals up to *i*



(1) Interval (months) since start of treatment i	(2) Number alive at beginning of interval a_i	(3) Deaths during interval d_i	(4) Number censored (lost to follow-up) during interval c_i	(5) Number of persons at risk $n_i = a_i - c_i/2$	(6) Risk of dying during interval $r_i = d_i/n_i$	(7) Chance of surviving interval $s_i = 1 - r_i$	(8) Cumulative chance of survival from start of treatment $S(i) = S(i - I) \times s_i$
1	240	12	0	240.0	0.0500	0.9500	0.9500
2	228	9	0	228.0	0.0395	0.9605	0.9125
3	219	17	1	218.5	0.0778	0.9222	0.8415
4	201	36	4	199.0	0.1809	0.8191	0.6893
5	161	6	2	160.0	0.0375	0.9625	0.6634
6	153	18	7	149.5	0.1204	0.8796	0.5835
7	128	13	5	125.5	0.1036	0.8964	0.5231
8	110	11	3	108.5	0.1014	0.8986	0.4700
9	96	14	3	94.5	0.1481	0.8519	0.4004
10	79	13	0	79.0	0.1646	0.8354	0.3345
11	66	15	4	64.0	0.2344	0.7656	0.2561
12	47	6	1	46.5	0.1290	0.8710	0.2231
13	40	6	0	40.0	0.1500	0.8500	0.1896
14	34	4	2	33.0	0.1212	0.8788	0.1666
15	28	5	0	28.0	0.1786	0.8214	0.1369
16	23	7	1	22.5	0.3111	0.6889	0.0943
17	15	12	0	15.0	0.8000	0.2000	0.0189
18	3	3	0	3.0	1.0000	0.0000	0.0000

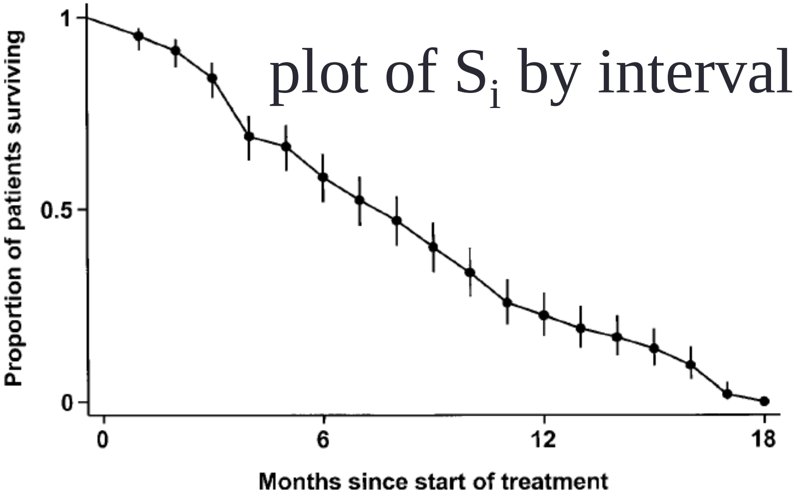
Life expectancy

- Life expectancy at time *t* is the ***sum*** of *cumulative probabilities* of surviving to the end of each of the remaining intervals, with interval duration as weights (column 8 in table):
  - some authors add 0.5 time units (since estimates are for the middle of each interval)
  - in most studies, cumulative probability is called l_x not S_x

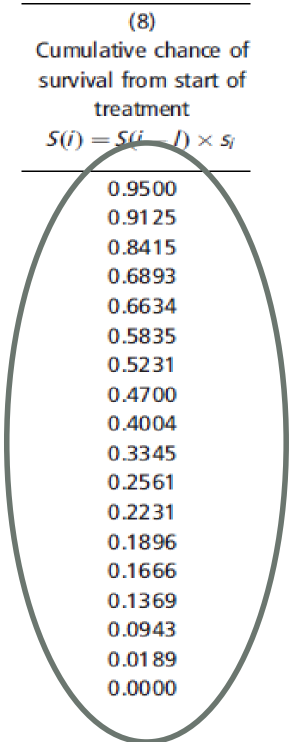
$$e_t = \sum_t^n S_i \times (\text{length of interval } i)$$

note that here we use S_i (cumulative survival, column 8) not s_i (interval survival, column 7)

- Example: life expectancy at birth is probability of surviving to t=1, plus probability of surviving to t=2, etc. until last interval
  - life expectancy of cancer patients from start of treatment (t=0) is the sum values in column 8 = 7.45 months



plot of S_i by interval



(8) Cumulative chance of survival from start of treatment $S(i) = S(i-1) \times s_i$
0.9500
0.9125
0.8415
0.6893
0.6634
0.5835
0.5231
0.4700
0.4004
0.3345
0.2561
0.2231
0.1896
0.1666
0.1369
0.0943
0.0189
0.0000

Longitudinal survival data

- Examining the longitudinal dataset *melanom*:
  - Load library *ISwR*
  - data on melanoma patients
  - but some patients are right censored
    - alive at the end of study, lost to follow up

- Variables:
  - *no*: patient ID
  - *status*:
    - **1=dead from melanoma**
    - 2=alive
    - 3=lost
  - *days*: time from start of study
  - *ulc*: ulceration (1=present; 2=absent)
  - *thick*: thickness of tumour
  - *sex*: 1=woman, 2=men
    - (ps not ideal! Better to use 0 and 1)

```
> melanom
     no   status days ulc thick sex
1   789     3    10   1   676   2
2    13     3    30   2    65   2
3    97     2    35   2   134   2
4    16     3    99   2   290   1
5    21     1   185   1  1208   2
6   469     1   204   1   484   2
7   685     1   210   1   516   2
8     7     1   232   1  1288   2
9   932     3   232   1   322   1
10  944     1   279   1   741   1
11  558     1   295   1   419   1
12  612     3   355   1    16   1
13    2     1   386   1   387   1
14  233     1   426   1   484   2
15  418     1   469   1   242   1
16  765     3   493   1  1256   2
…
202 798      2  4668  2   612   1
203 806      2  4688  2    48   1
204 606      2  4926  2   226   1
205 328      2  5565  2   290   1
```

Censoring data

- To start survival analysis, load package *survival*

- As a first step, we must indicate the right-censored observations

- To prepare an object for survival analysis, we single out the **dead (status=1)** individuals (the non-censored ones)
  - status 2 (alive) and 3 (lost) are censored

```
> Surv(melanom$days, melanom$status==1)
 [1]   10+   30+   35+   99+  185   204   210   232   232+  279   295   355+  386
 [14]  426   469   493+  529   621   629   659   667   718   752   779   793   817
 [27]  826+  833   858   869   872   967   977   982  1041  1055  1062  1075  1156
 [40] 1228  1252  1271  1312  1427+ 1435  1499+ 1506  1508+ 1510+ 1512+
…
[196] 4124+ 4207+ 4310+ 4390+ 4479+ 4492+ 4668+ 4688+ 4926+ 5565+
```

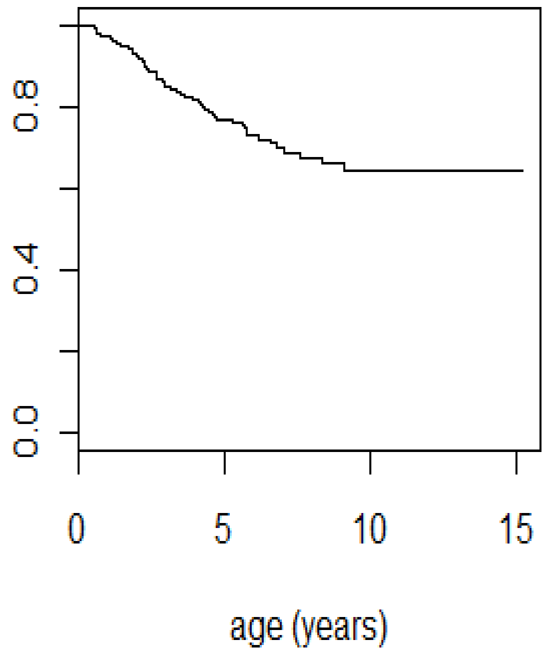
- all censored individuals (not “1”) are marked with +
- death events are on continuous time scale (days); no pre-set time intervals

Kaplan-Meier estimation

- K-M estimates a survival function S(t) from longitudinal data, taking right-censoring into account

- K-M uses the same product of s_i to calculate probability of survival to age *t*, i.e. S(t);
  - each death defines a new interval, and a new s_i and S(t) are calculated

- For this reason, K-M produces a step-function
  - each time there is a death, survival estimate is reduced by a factor 1-(1/N)
    - N is number of *uncensored* people at risk



- To estimate K-M curve, apply function *survfit* to censored data:

```
km1 <- survfit(Surv(melanom$days,
melanom$status==1) ~ 1)
```

("~1" estimates survival by *days)*

- Mortality is not displayed
  - intervals have different durations (the interval between two deaths)
  - Better to display survival to that time
  - death dates are interval transitions

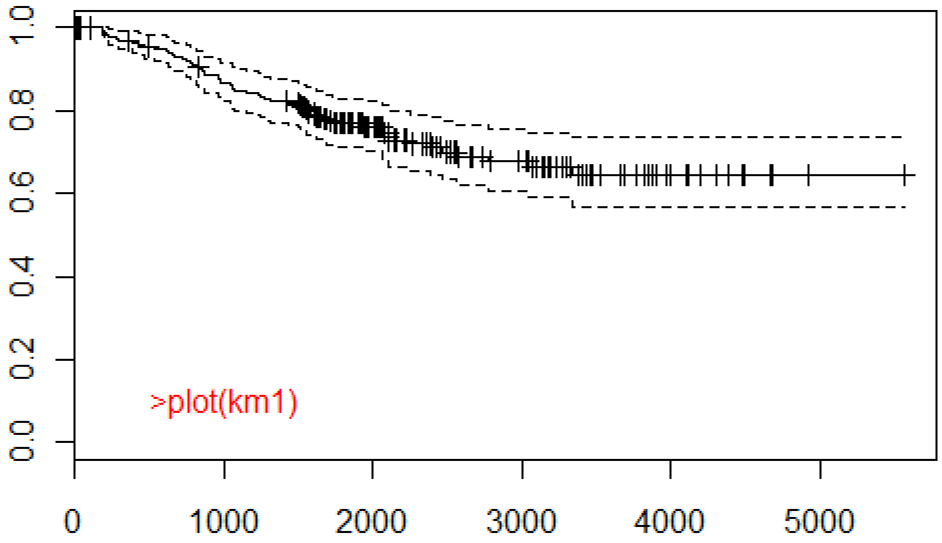
- Only uncensored times appear
  - for censored cases, add *censored=T*)
- To get only vector with S(t):
  - > summary(km1)$surv

```
> summary(km1)
Call: survfit(formula = Surv(days, status == 1) ~ 1)
 time n.risk n.event survival    std.err   lower 95% CI upper 95% CI
  185    201       1    0.995   0.00496        0.985        1.000
  204    200       1    0.990   0.00700        0.976        1.000
  210    199       1    0.985   0.00855        0.968        1.000
  232    198       1    0.980   0.00985        0.961        1.000
  279    196       1    0.975   0.01100        0.954        0.997
  295    195       1    0.970   0.01202        0.947        0.994
  386    193       1    0.965   0.01297        0.940        0.991
  426    192       1    0.960   0.01384        0.933        0.988
  469    191       1    0.955   0.01465        0.927        0.984
  529    189       1    0.950   0.01542        0.920        0.981
  621    188       1    0.945   0.01615        0.914        0.977
  629    187       1    0.940   0.01683        0.907        0.973
…
 2565     63       1    0.689   0.03729        0.620        0.766
 2782     57       1    0.677   0.03854        0.605        0.757
 3042     52       1    0.664   0.03994        0.590        0.747
 3338     35       1    0.645   0.04307        0.566        0.735
```

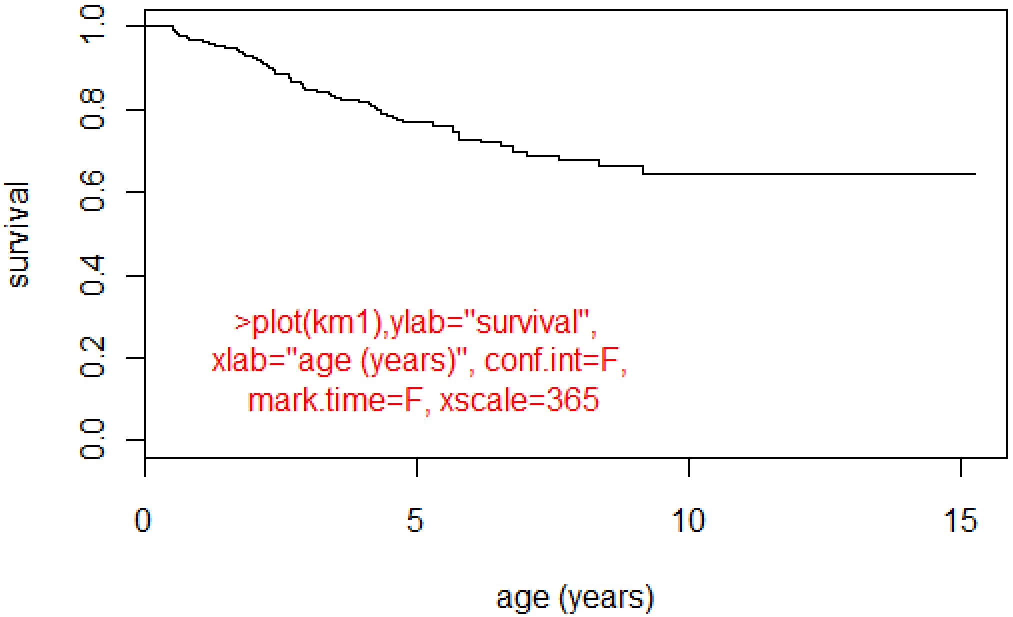
```
> summary(km1, censored=T)
Call: survfit(formula = Surv(days, status == 1) ~ 1)
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   10    205       0    1.000    0.00000    1.000        1.000
   30    204       0    1.000    0.00000    1.000        1.000
   35    203       0    1.000    0.00000    1.000        1.000
   99    202       0    1.000    0.00000    1.000        1.000
  185    201       1    0.995    0.00496    0.985        1.000
```

Survival plot

- We can plot survival object *km1* with *plot*
- In our example, S(t) does not drop to 0
  - there is no information on survival after 5565 days
  - 95% CI bands are plotted by default
- Confidence intervals are not symmetrical
  - calculated on a log scale; when unlogged, upper limit line is further from mean line
- Drops on S(t) are deaths
- Markings on S(t) curve are censored observations (time of loss)
  - to eliminate marks: *mark.time=F*



```
>plot(km1)
```



```
>plot(km1),ylab="survival",
xlab="age (years)", conf.int=F,
mark.time=F, xscale=365
```

Survival curves by group

- There may be differences in survival among subgroups of sample

- We can use K-M estimation to calculate separate curves by sex:

```
kmsex <- survfit(Surv(melanom$days,
melanom$status==1) ~ melanom$sex)
```

- We obtain a separate table for men (sex=1) and women (sex=2)

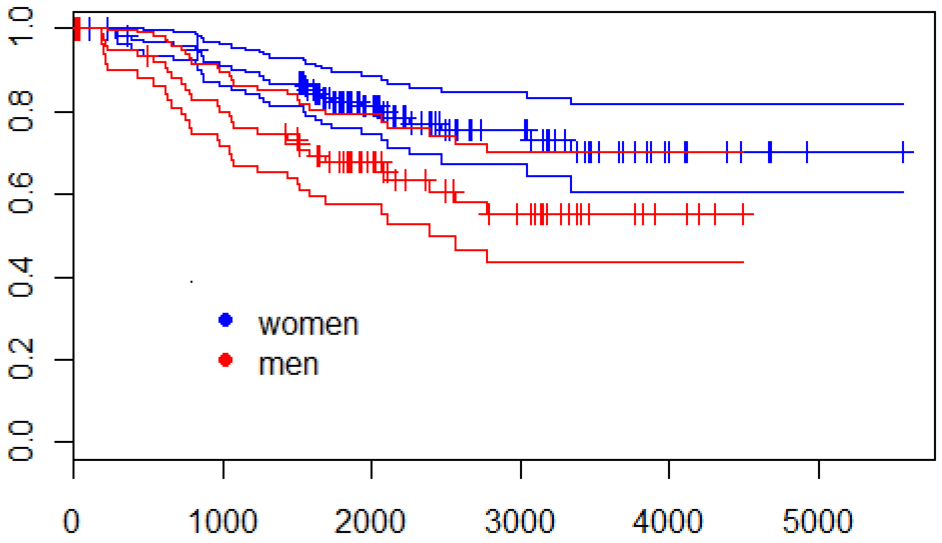
```
kmsex <- survfit(Surv(melanom$days,
melanom$status==1) ~ melanom$sex)
> summary(kmsex)
Call: survfit(formula = Surv(days, status == 1) ~ sex)

                sex=1
 time n.risk n.event survival std.err l   L95%CI  U95%CI
  279    124       1    0.992  0.00803     0.976    1.000
  295    123       1    0.984  0.01131     0.962    1.000
  386    121       1    0.976  0.01384     0.949    1.000
  469    120       1    0.968  0.01593     0.937    0.999
  667    119       1    0.959  0.01775     0.925    0.995
  817    118       1    0.951  0.01937     0.914    0.990
…
                sex=2
 time n.risk n.event survival std.err l   L95%CI  U95%CI
  185     76       1    0.987   0.0131     0.962    1.000
  204     75       1    0.974   0.0184     0.938    1.000
  210     74       1    0.961   0.0223     0.918    1.000
  232     73       1    0.947   0.0256     0.898    0.999
  426     72       1    0.934   0.0284     0.880    0.992
  529     70       1    0.921   0.0310     0.862    0.984
…
```

Plotting by group

- Now we can compare survival by sex visually

  >plot(kmsex, conf.int=T, col=c("blue", "red"))



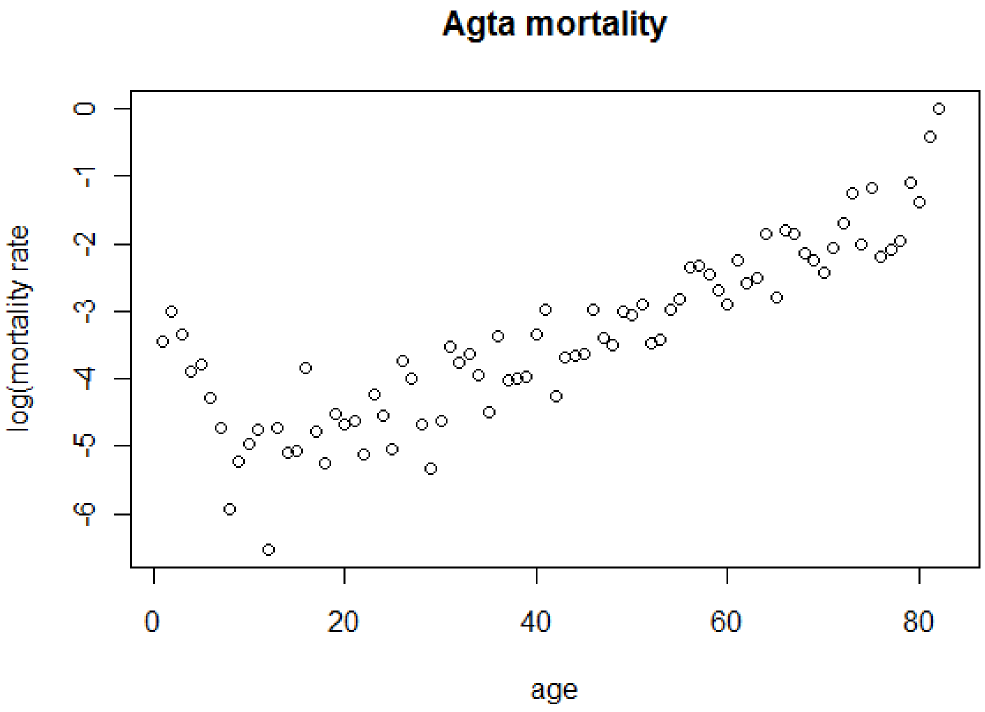
- When analysing by group, 95% CI bands must be added
  - colours, legend added too
  - (see *R* code script)

Example: Agta survival

- See R code
- File *Agta*: live and dead Agta people, by age at death or last observation

Gompertz mortality

- Gompertz proposed a mortality model
- $m = ae^{bx}$
  - m =mortality rate
  - a = baseline mortality
  - b = rate of ageing
- $\log(m) = \log(a) + bx$
  - log(mortality) increases linearly with age



Quiz: !Kung mortality

- See our Moodle page