MODERN EPIDEMIOLOGY

ASSIGNMENT 2

Hephaes Chuen Chau (M116 I54105029)

2022-03-16

Contents

Questions are adapted from *Epidemiology, Beyond the Basics*.

Question 1

1(a)

We import the data as follows:

```
follow_up_time <- c(2,4,7,8,12,15,17,19,20,23)
event <- c("death","censored","censored","death","censored","death",
           "death","death","censored","death")
dataset <- data.frame(follow_up_time,event)
dataset
```

```
##    follow_up_time    event
## 1               2    death
## 2               4 censored
## 3               7 censored
## 4               8    death
## 5               12 censored
## 6               15    death
## 7               17    death
## 8               19    death
## 9               20 censored
## 10              23    death
```

Using this dataset, we estimate the probability of death at a time i by the following formula:

$$\mathbb{P}(\text{death at time } i) = \frac{\text{number of deaths at } i}{\text{number of at-risk individuals at time } i}$$

while the probability of survival beyond time i is estimated simply to be

$$\mathbb{P}(\text{survival beyond time } i) = 1 - \mathbb{P}(\text{death at time } i)$$

Cumulative probabilities of survival at time j is thus estimated by:

$$\text{Cumulative probabilities of survival at time } j = \prod_{i \leq j} \mathbb{P}(\text{survival beyond time } i)$$

We apply the following function:

```
# function for calculating survival prob
Calculate_P <- function(x = 0, y = 0){
  # x is the counter of loop
  # y is counting the number of dead individuals
  prob_death <<- NA # prob_death is the probability of death at time i
  function(event) {
    if (event == "death") {
      x <<- x + 1
      y <<- y + 1
      message("iteration: ",x)
      message("accumulated number of dead: ",y)
      prob_death <<- 1/(20-x+1)
      return(prob_death)
    } else {
      x <<- x + 1
      message("iteration: ",x)
      message("accumulated number of dead: ",y)
      return(NA)
    }
  }
}

# function for calculating cumulative survival prob
Calculate_SP <- function() {
  x <<- 1
  function(y) {
    if (is.na(y)) {
      return(NA)
    } else {
      z <<- x
      x <<- z*y
      return(z*y)
    }
  }
}

# higher-order function initialisation
calculate_p <- Calculate_P()
calculate_sp <- Calculate_SP()

# calculations
prob_death <- sapply(dataset$event, calculate_p)
prob_surv <- 1 - prob_death
cum_prob_surv <- sapply(prob_surv, calculate_sp)

dataset$prob_death <- prob_death
dataset$prob_surv <- prob_surv
dataset$cum_prob_surv <- cum_prob_surv
```

The required answer is therefore

```
dataset
```

```
##    follow_up_time    event prob_death prob_surv cum_prob_surv
## 1               2    death 0.05000000 0.9500000     0.9500000
## 2               4 censored         NA        NA            NA
## 3               7 censored         NA        NA            NA
## 4               8    death 0.05882353 0.9411765     0.8941176
## 5              12 censored         NA        NA            NA
## 6              15    death 0.06666667 0.9333333     0.8345098
## 7              17    death 0.07142857 0.9285714     0.7749020
## 8              19    death 0.07692308 0.9230769     0.7152941
## 9              20 censored         NA        NA            NA
## 10             23    death 0.09090909 0.9090909     0.6502674
```

where `prob_death` is the probability of death, `prob_surv` is the probability of survival, and `cum_prob_surv` is the cumulative probability of survival.

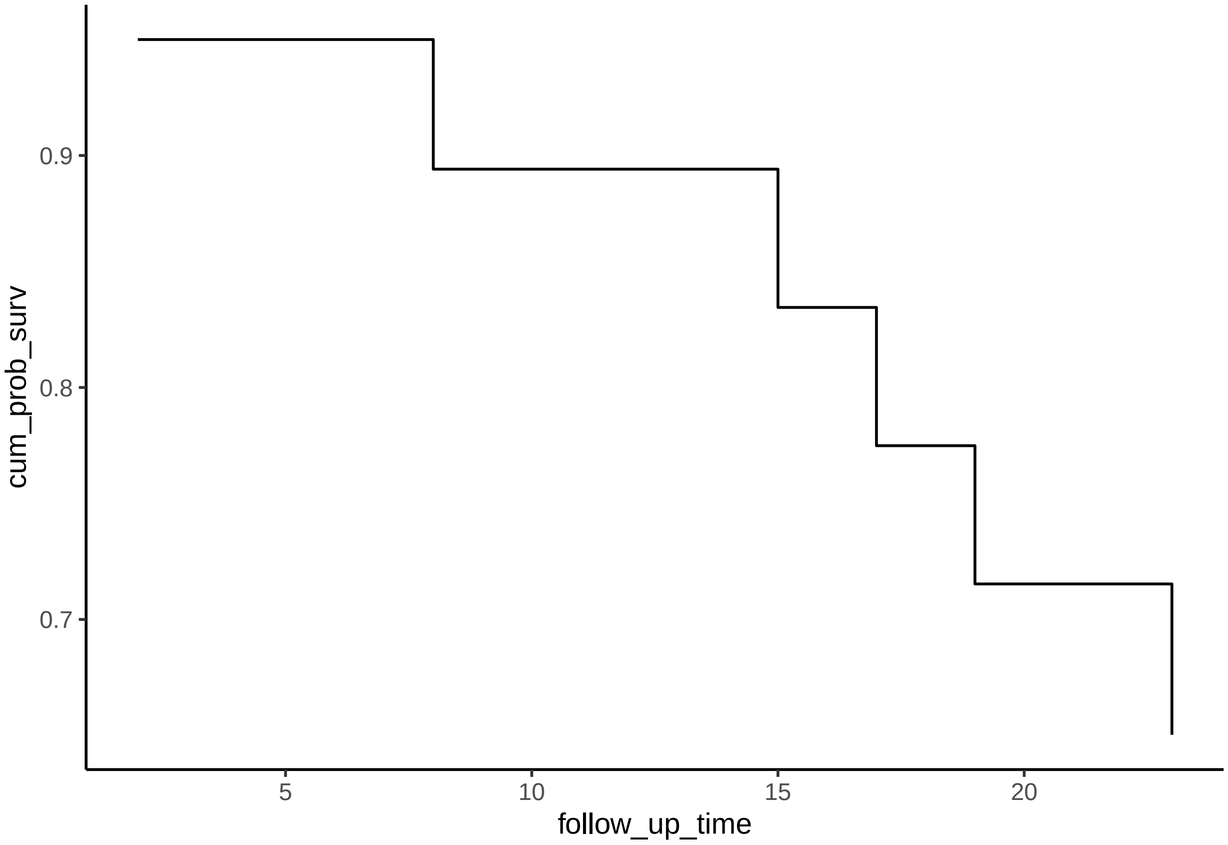
1(b)

The cumulative probability of survival at the end of the follow-up period is 0.65, as shown by the table in the previous answer.

1(c)

The cumulative survival probabilities is plotted as below:

```
library(ggplot2)
ggplot(dataset[!is.na(dataset$cum_prob_surv),], aes(x=follow_up_time, y=cum_prob_surv)) +
  geom_step() + theme_classic()
```



1(d)

We note that at the end of the study, 6 people have died. If we assume that the censored individuals have also survived, then the simple proportion required is given by $14/20 \times 100$.

1(e)

That is because we have simply assumed the censored individuals have survived, while in the life table method we applied, we ignore the censored individuals by subtracting them from the denominator in the calculation of the death probability. For example, at time 4, we calculate the death probability as $1/(20-4)$ as opposed to $1/(20-1)$.

1(f)

We need to calculate the total number of months of follow-up contributed by all 10 individuals with events/censoring in the data set. To do this we simply:

```
total_follow_up <- sum(dataset$follow_up_time)
total_follow_up
```

```
## [1] 127
```

Thus there are 127 months of observation, translating into 10.58333 years. For the other 10 individuals with no events and not lost to follow-up, they contribute a total of $2 \times 10 = 20$ years of observation. This implies that we have 30.58333 person-years. Therefore, we have

$$\text{Death rate per 100 person-years} = \frac{6}{30.58333/100} = 19.619$$

1(g)

To calculate the required rates, we need to determine the number of person-years observed in the first year and the number of person-years observed in the second year, to do this we calculate:

```
first_year_total_follow_up <- (20-5)*12 + sum(dataset$follow_up_time[1:5])
second_year_total_follow_up <- (20-10)*12 + sum(dataset$follow_up_time[6:10]-12)
sprintf("In the first year, total follow-up time in person-months is: %s ",
        first_year_total_follow_up)
```

```
## [1] "In the first year, total follow-up time in person-months is: 213 "
```

```
sprintf("In the second year, total follow-up time in person-months is: %s ",
        second_year_total_follow_up)
```

```
## [1] "In the second year, total follow-up time in person-months is: 154 "
```

Hence the required rates are calculated as follows:

$$\text{Death rate per 100 person-years in the first year} = \frac{2}{213/(12 \times 100)} = 11.268$$

$$\text{Death rate per 100 person-years in the second year} = \frac{4}{154/(12 \times 100)} = 31.169$$

1(h)

1(g)'s calculations told us that it was not appropriate to do so. The death rates differ between the first and second year, indicating that the probability distribution of event occurrence in the first year may be different from that of the second year (with the events becoming more frequent in the second year). Therefore, calculating the rate for the whole 2-year period will mask such heterogeneity.

1(i)

The missing data mechanism is not MNAR (-ie, missing not at random). This is equivalent to saying that the indicator variable "a person is censored" and the random variable "a person has an event at time T" are independent. Knowledge that a person is censored does not update our belief that such a person is more or less likely to have the event in question than those non-censored individuals.

1(j)

With the assumption in the question no individuals have been censored. Assuming that they all are followed till the end, we apply the incidence proportion calculation method to observe that the proportion of individuals who die is $\frac{6}{20} = 0.3$.

By definition, odds of a binary event is the ratio of its occurrence probability to that of its non-occurrence, therefore we have, assuming that the probability of occurrence is 0.3,

$$\text{Odds} = \frac{0.3}{1 - 0.3} = 0.428$$

1(k)

With proportion of individuals who die, we interpret the figure 0.3 as saying that we expect in a population with similar characteristics as this cohort's, 30% of them will have died at the end of 2-year follow-up. However, with odds, our interpretation is that with respect to a single individual in the cohort, the probability that he died is 42.8% of the probability that he did not die. Odds are simply used as a quantity to express the relative size of the probability of events (against that of non-event).

Question 2

2(a)

We note that

$$\text{Proportion of individuals who present hypertension} = \frac{70}{318}$$

and that

$$\text{Proportion of individuals who do notpresent hypertension} = 1 - \frac{70}{318} = \frac{248}{318}$$

Therefore the odds should be calculated as follows:

$$\text{Odds using absolute numbers for cases} = \frac{70/318}{248/318} = \frac{70}{248} = 0.282$$

By similar logic we can calculate odds for the control:

$$\text{Odds using absolute numbers for control} = \frac{30}{363} = 0.083$$

2(b)

We have

$$\text{Odds using percentages for cases} = \frac{22}{78} = 0.282$$

and

$$\text{Odds using percentages for control} = \frac{7.6}{92.4} = 0.082$$

2(c)

Odds can be calculated both by percentages or by absolute numbers, the numerical figures are almost equal (subject to correction of statistically significant figures).

2(d)

To explain this observation, first suppose we have a binary indicator variable $I \in \{0, 1\}$ such that $\mathbb{P}(I = 0) = p$ and $\mathbb{P}(I = 1) = 1 - p$. The odds are defined as

$$\text{Odds} = \frac{p}{1 - p}$$

Therefore for p very small (close to 0), we have $1 - p$ close to 1 and therefore Odds close to p. Now in the control, we observe that the proportion of those presenting with hypertensive history is 30/393, which is much lower than 70/318 the corresponding proportion in the cases. This means that for control, p is much smaller, and therefore the odds is much closer to p.

2(e)

Prevalence odds is calculated. In this study, which is a case-control study, the cases are those with uterine leiomyoma, and the control are those without (and possibly selected for some other matching criteria). Importantly, both the cases and the control are not followed-up prospectively from a point where no individuals in the groups have any hypertension episodes, and therefore it is NOT incidence that is being calculated when we compute the odds. The proportion we used (eg, 70/318 for cases) is the prevalence at a particular point in time (the time at which the cases and the control are selected)

Question 3

We used the formula provided by the textbook (equation 2.4, p.72):

$$\frac{\text{Point prevalence}}{1 - \text{ Point prevalence}} = \text{Incidence} \times \text{Duration}$$

Rearranging the terms, it is easy to see that duration is given by $\frac{0.56}{1-0.56} \times \frac{1}{0.05} = 25.45$ years.