Measures of disease frequency and association

Standard errors and confidence intervals

1 Illustrative dataset

To illustrate the estimation of the measures of disease frequency, associated standard errors (se) and confidence intervals (CI), we will use the `BrCa` dataset from the Epi package:

```
library(Epi)
data(BrCa, package = "Epi")
## only consider some of the variables
BrCaR <- BrCa[,c("pid","age","grade","tox","xst")]
## give more intuitive name
names(BrCaR) <- c("id","age","grade","time","status")
## display
str(BrCaR)
```

```
'data.frame':       2982 obs. of  5 variables:
 $ id    : int  1264 1150 838 1214 1130 1118 386 1417 927 489 ...
 $ age   : int  54 55 34 42 35 50 46 40 36 42 ...
 $ grade : Factor w/ 2 levels "2","3": 1 1 1 1 1 1 1 1 1 1 ...
 $ time  : num  12.97 8.78 9.41 10.47 10.35 ...
 $ status: Factor w/ 2 levels "Alive","Dead": 1 1 1 1 1 2 1 1 1 1 ...
```

It contains $n = 2982$ women with breast cancer (with grade indicated by the variable `grade`) followed from diagnosis until death or loss to follow-up. Summing `time` and `status` values over participants gives the number of events and total follow-up time:

```
c("riskTime" = sum(BrCaR$time), "person.year" = sum(BrCaR$status=="Dead"))
```

```
riskTime    event
21270.74  1272.00
```

The same calculation can be performed per group using `xtabs` (instead of subsetting the data set):

```
t23 <- xtabs(cbind(n=1,N. = status=="Dead",person.year = time) ~ grade, data = BrCaR)
t23
```

```
grade         n        N. person.year
    2   794.000   262.000    6323.439
    3  2188.000  1010.000   14947.300
```

2 Theory

Denote by:

- n the number of persons in the population under study.
- $H_\bullet(t)$ the number of persons sick at time t.
- $N_\bullet(t)$ the number of persons who contracted the disease at some point between time 0 and time t.
- $\int_0^t Y_\bullet(s)ds = \sum_{i=1}^n \min(T_i, t)$ the cumulative at risk time up to time t (`person.year` in the example dataset). For a given subject, it is the time during which he is under study and has not yet contracted the disease.

statistic	estimate	standard error (se)	confidence interval [lower, upper]
prevalence	$\widehat{\pi}(t) = \frac{H_\bullet(t)}{n}$	$\widehat{\sigma}_{\widehat{\pi}(t)} = \sqrt{\frac{\widehat{\pi}(t)(1-\widehat{\pi}(t))}{n}}$	lower=$\widehat{\pi}(t) - 1.96\widehat{\sigma}_{\widehat{\pi}}$
			upper=$\widehat{\pi}(t) + 1.96\,\widehat{\sigma}_{\widehat{\pi}(t)}$
odds	$\widehat{\Omega}(t) = \frac{H_\bullet(t)}{n-H_\bullet(t)}$	$\sigma_{\log\widehat{\Omega}(t)} = \sqrt{\frac{1}{H_\bullet(t)} + \frac{1}{n-H_\bullet(t)}}$	lower=$\widehat{\Omega}(t)\exp\left(-1.96\,\sigma_{\log\widehat{\Omega}(t)}\right)$
			upper=$\widehat{\Omega}(t)\exp\left(1.96\,\sigma_{\log\widehat{\Omega}(t)}\right)$
incidence rate	$\widehat{\lambda}(t) = \frac{N_\bullet(t)}{\int_0^t Y_\bullet(s)ds}$	$\sigma_{\log\widehat{\lambda}(t)} = \frac{1}{\sqrt{N_\bullet(t)}}$	lower=$\widehat{\lambda}(t)\exp\left(-1.96\sigma_{\log\widehat{\lambda}(t)}\right)$
			upper=$\widehat{\lambda}(t)\exp\left(1.96\sigma_{\log\widehat{\lambda}(t)}\right)$
risk	$\widehat{r}(t) = \frac{N_\bullet(t)}{n}$	$\widehat{\sigma}_{\widehat{r}(t)} = \sqrt{\frac{\widehat{r}(t)(1-\widehat{r}(t))}{n}}$	lower=$\widehat{r}(t) - 1.96\widehat{\sigma}_{\widehat{r}(t)}$
			upper=$\widehat{r}(t) + 1.96\,\widehat{\sigma}_{\widehat{r}(t)}$

When contrasting two groups composed of distinct individuals, the standard error of the difference is the square root of the sum of squared standard errors. For instance if we estimate:

- a risk $\widehat{r}_1(t)$ based on n_1 persons in one group, with standard error $\widehat{\sigma}_{\widehat{r}_1(t)} = \sqrt{\frac{\widehat{r}_1(t)(1-\widehat{r}_1(t))}{n_1}}$
- a risk $\widehat{r}_2(t)$ based on n_2 persons in the other group, with standard error $\widehat{\sigma}_{\widehat{r}_2(t)} = \sqrt{\frac{\widehat{r}_2(t)(1-\widehat{r}_2(t))}{n_2}}$

The standard error for the risk difference can be estimate with $\widehat{\sigma}_{\widehat{r}_2(t)-\widehat{r}_1(t)} = \sqrt{\frac{\widehat{r}_2(t)(1-\widehat{r}_2(t))}{n_2} + \frac{\widehat{r}_1(t)(1-\widehat{r}_1(t))}{n_1}}$.

3 Implementation

3.1 'by hand': using the entire follow-up

Consider the table `t23` previously created which contains the number of women (`n`), events (`N.`), and at risk time (`person.year`) for each group (grade 2 and grade 3). We can add an additional line for the overall cohort by adding the group specific values using `addmargin`:

```
t33 <- addmargins(t23, margin = 1) ## 1: sum over rows, 2 over columns
t33
```

```
grade         n        N. person.year
  2     794.000   262.000    6323.439
  3    2188.000  1010.000   14947.300
  Sum  2982.000  1272.000   21270.738
```

Then we can estimate the incidence rates (in each group and for the whole cohort) as:

```
lambda <- t33[,"N."]/t33[,"person.year"]
unname(lambda)
```

```
[1] 0.04143315 0.06757073 0.05980046
```

Confidence intervals can then be obtained using:

```
se.loglambda <- 1/sqrt(t33[,"N."])
cbind(estimate = lambda,
      lower = lambda * exp(-1.96 * se.loglambda), upper = lambda * exp(1.96 * se.loglambda))
```

```
      estimate      lower      upper
2   0.04143315 0.03670790 0.04676666
3   0.06757073 0.06352934 0.07186921
Sum 0.05980046 0.05660276 0.06317882
```

Having incomplete follow-up for many women

```
quantile(BrCaR$time[BrCaR$status=="Alive"])
```

```
        0%         25%         50%         75%        100%
0.09856263  7.01437394  8.81314150 10.60780271 19.28268305
```

complicates the evaluation of the risk. We could use the risk-rate relationship to evaluate the 2 year risk (in each group and for the whole cohort):

```
1 - exp(-lambda * 2)
```

```
        2         3       Sum
0.0795258 0.1264077 0.1127255
```

Both the estimation of the rate and of the risk assume that the rate is constant over the entire follow period (about 20 years).

3.2 'by hand': on a restricted follow-up time

Consider only the first 2 years of follow-up. We can re-compute the summary statistics (number of individual, number of events, total follow-up time) either by adding new columns to the dataset:

```
BrCaR$time2 <- pmin(BrCaR$time,2)
BrCaR$status2 <- ifelse(BrCaR$time<=2,as.character(BrCaR$status),"Alive")
xtabs(cbind(n=1,N. = status2=="Dead", person.year = time2) ~ grade, data = BrCaR)
```

or directly do the modification in `xtabs`:

```
t23.y2 <- xtabs(cbind(n=1,
                      N. = (status=="Dead")*(time<=2),
                      person.year = pmin(time,2)) ~ grade, data = BrCaR)
t33.y2 <- addmargins(t23.y2, margin = 1)
t33.y2
```

```
grade        n       N. person.year
  2     794.000   27.000    1563.507
  3    2188.000  191.000    4229.844
  Sum 2982.000  218.000    5793.351
```

We can evaluate the incidence rate and 2 year risk as:

```
lambda.y2 <- t33.y2[,"N."]/t33.y2[,"person.year"]
rbind(rate = lambda.y2, risk = 1 - exp(- lambda.y2 * 2))
```

```
              2          3        Sum
rate 0.01726887 0.04515533 0.03762934
risk 0.03394812 0.08635269 0.07249648
```

Compared to the estimation based on the entire follow-up, we use a weaker assumption (constant rate within the first two years only) but have less events to work with (i.e. larger statistical uncertainty):

```
se.loglambda.y2 <- 1/sqrt(t33.y2[,"N."])
cbind(estimate = lambda.y2,
      lower = lambda.y2 * exp(-1.96 * se.loglambda.y2),
      upper = lambda.y2 * exp(1.96 * se.loglambda.y2))
```

```
      estimate      lower      upper
2   0.01726887 0.01184260 0.02518145
3   0.04515533 0.03918475 0.05203564
Sum 0.03762934 0.03295148 0.04297128
```

Note that had we have had no loss to follow-up we could have computed the 2 year risk without making assumptions on the incidence rate doing:

```
t33.y2[,"N."]/t33.y2[,"n"]
```

```
         2          3        Sum
0.03400504 0.08729433 0.07310530
```

3.3 'glm': on a restricted follow-up time

The glm function can be used with the poisson family to estimate incidence rates:

```
e.pois <- glm(status2=="Dead" ~ grade, offset = log(time2),
              family = poisson(link="log"), data = BrCaR)
cbind(estimate = exp(coef(e.pois)), exp(confint(e.pois)))
```

```
Waiting for profiling to be done...
              estimate      2.5 %     97.5 %
(Intercept) 0.01726887 0.01154676 0.02462553
grade3      2.61484005 1.78076822 3.99981206
```

provides the incidence rate for the reference group (here grade2) and the rate ratio (grade3 3 vs. 2). Note that removing the intercept in the formula leads to the same model:

```
e.pois2 <- glm(status2=="Dead" ~ grade-1, offset = log(time2),
               family = poisson(link="log"), data = BrCaR)
logLik(e.pois2)
logLik(e.pois)
```

```
'log Lik.' -881.9157 (df=2)
'log Lik.' -881.9157 (df=2)
```

but parametrized differently: an incidence rate per group

```
cbind(estimate = exp(coef(e.pois2)), exp(confint(e.pois2)))
```

```
Waiting for profiling to be done...
         estimate      2.5 %     97.5 %
grade2 0.01726887 0.01154676 0.02462553
grade3 0.04515533 0.03905040 0.05186517
```

The incidence rate for the whole cohort can be obtained by:

```
e.pois <- glm(status2=="Dead" ~ 1, offset = log(time2),
              family = poisson(link="log"), data = BrCaR)
c(exp(coef(e.pois)), exp(confint(e.pois)))
```

```
Waiting for profiling to be done...
(Intercept)       2.5 %      97.5 %
 0.03762934  0.03285262  0.04284772
```

The CIs are slightly different as they are based on a different approximation (profile likelihood instead of delta method).

For the risk (⚠ in absence of censoring) one can use the `binomial` family:

- with an identity link to get risk difference (no back-transformation)

```
e.RD <- glm(status2=="Dead" ~ grade,  family = binomial(link="identity"), data = BrCaR)
cbind(coef(e.RD), confint(e.RD))
```

```
Waiting for profiling to be done...
                            2.5 %     97.5 %
(Intercept) 0.03400504 0.02286881 0.04814043
grade3      0.05328929 0.03530580 0.07012383
```

- with an log link to get risk ratio (back-transformation via `exp`)

```
e.RR <- glm(status2=="Dead" ~ grade,  family = binomial(link="log"), data = BrCaR)
cbind(exp(coef(e.RR)), exp(confint(e.RR)))
```

```
Waiting for profiling to be done...
                            2.5 %     97.5 %
(Intercept) 0.03400504 0.02286695 0.04813859
grade3      2.56709984 1.76421904 3.89779150
```