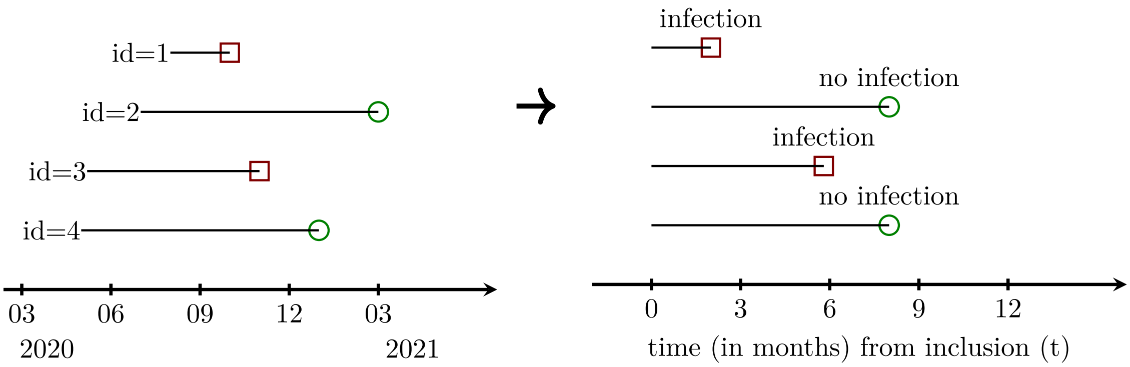
Lecture 2: Measures of disease frequency and association

Key concepts

1 Data representation

Typical epidemiological data correspond to n individuals (here $n = 4$) followed over time until a specific event (here infection) or until the end of study (here $\tau = 8$ months):



In an ideal study, each subject may experience 3 states:

- **not at risk**: the subject cannot experience the event (before the solid line)
  (e.g. COVID infection: COVID did not exist yet, pregnancy: outside fertility age)
- **at risk**: the subject may experience the event at any time (solid line)
  (ideally the start of the at risk time coincide with the start of the study)
- **event**: the subject experiences the event (squares). Thereafter he is no longer at risk.
  ⚠ The subject can also experience competing events preventing the event of interest to occur (e.g. death) or we might lose track of him at a certain time (censoring, here circles)

Notations: $\mathbb{1}_x$ indicator function of x being true and $a \wedge b$ minimum between a and b

- time to event: T
- right-censored time: $\tilde{T} = T \wedge \tau$ time to death stopped at the end of study
- event occurence: $N(t) = \mathbb{1}_{T \leq t}$ i.e. 0 (healthy/survival before time T)
  1 (sick/death/failure after time T)
- observable occurence: $\tilde{N}(t) = \mathbb{1}_{T \leq t \wedge \tau}$
- health status: $H(t) = 1$ if sick (still affect by the event) otherwise 0

On the computer there are various (sometimes equivalent) way to store the data:

Individual data: one line per subject

```
patient  inclusion        end time status
    id1 01-08-2020 01-10-2020  2.0   dead
    id2 01-07-2020 01-03-2021  8.0  alive
    id3 02-05-2020 01-11-2021  5.9   dead
    id4 01-05-2020 01-01-2021  8.0  alive
```

Aggregated data: one line per timepoint

```
time n.atRisk dead n-D D    Y
 0.0        4    0   4 0  0.0
 2.0        4    1   3 1  8.0
 5.9        3    1   2 2 19.7
 8.0        2    0   2 2 23.9
```

The aggregated format provides a much more compact data representation. This can be critical for real life dataset that may consider millions of individuals, like for the COVID pandemic[1]:

```
           date country  atRisk cases
  1: 2019-12-30 Denmark 5840045    10
  2: 2020-01-06 Denmark 5840035    12
  3: 2020-01-13 Denmark 5840023     8
  4: 2020-01-20 Denmark 5840015    15
  5: 2020-01-27 Denmark 5840000    13
 ---
130: 2022-06-20 Denmark 3004251  8696
131: 2022-06-27 Denmark 2995555 10720
132: 2022-07-04 Denmark 2984835 12264
133: 2022-07-11 Denmark 2972571 11965
134: 2022-07-18 Denmark 2960606 10171
```

```
           date country   atRisk  cases
  1: 2019-12-30  France 67656682      0
  2: 2020-01-06  France 67656682      0
  3: 2020-01-13  France 67656682      0
  4: 2020-01-20  France 67656682      3
  5: 2020-01-27  France 67656679      3
 ---
130: 2022-06-20  France 37432876 468726
131: 2022-06-27  France 36964150 742395
132: 2022-07-04  France 36221755 916068
133: 2022-07-11  France 35305687 694877
134: 2022-07-18  France 34610810 530397
```

⚠ Generally this compact representation is performed at the expense of a loss of information about subject's covariates or inclusion time. It makes it difficult to account for heterogeneity between populations (e.g. age unbalance) or time trends.

A simple (but crude) way to compare two populations is via a **2 by 2 table**, containing the data relative to each population at a specific timepoint, e.g. week 29 in 2022:

```
         at risk infected
Denmark  2960606    10171
 France 34610810   530397
```

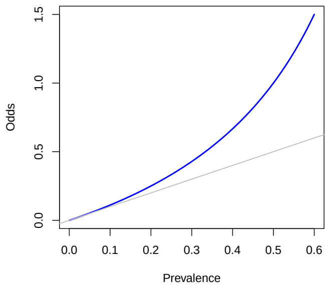
Notations:

- cumulative number of events: $D(t) = \sum_{i=1}^{n} N_i(t)$
- remaining event-free individuals: $n - D(t)$
- cumulative follow-up time: $Y(t) = \sum_{i=1}^{n} \tilde{T}_i \wedge t$

[1] these numbers are given for illustrative purpose and may not match the official numbers.

2 Measures of disease frequency

Quantity of interest (deterministic)



- prevalence: $\pi(t) = \mathbb{P}\left[H(t) = 1\right] \in [0, 1]$
- odds: $\Omega(t) = \frac{\pi(t)}{1-\pi(t)} \in [0, +\infty[$.
  Rare event: $\Omega \approx \pi$ when prevalence is small.
- hazard or rate: $\lambda(t) = \lim\limits_{\tau \rightarrow 0} \frac{\mathbb{P}[t<T\leq t+\tau, N(t+\tau)=1 | T>t]}{\tau}$
- t-years risk: $r(t) = \mathbb{P}\left[0 < T \leq t, N(t) = 1\right]$

Estimation (stochastic)

- cross sectional study a time t with a single group: $(N_i(t))_{i\in\{1,\ldots,n\}}$
  - prevalence: $\widehat{\pi}(t) = \frac{1}{n}\sum_{i=1}^n H_i(t) = \frac{\text{number of person sick at time } t}{\text{number of people in the population}}$
  - odds: $\widehat{\Omega}(t) = \frac{\widehat{\pi}(t)}{1-\widehat{\pi}(t)}$.
- τ-years cohort study with a single group: $\left(\widetilde{N}_i(t), \widetilde{T}_i\right)_{i\in\{1,\ldots,n\}, t\in[0,\tau]}$
  - incidence rate: $\widehat{\lambda}(\tau) = \frac{\sum_{i=1}^n \widetilde{N}_i(\tau)}{\sum_{i=1}^n \widetilde{T}_i} = \frac{\widetilde{D}(\tau)}{\widetilde{Y}} = \frac{\text{number of events up to time } \tau}{\text{number of person time at risk}}$ ⚠ has a unit!
  - τ-years risk: $\widehat{r}(\tau) = \frac{1}{n}\sum_{i=1}^n \widetilde{N}_i(\tau) = \frac{\widetilde{D}(\tau)}{n} = \frac{\text{number of events up to time } \tau}{\text{number of persons at risk}}$

Uncertainty (stochastic, the dependency on time is dropped to simplify the expressions)

1. prevalence: $\mathrm{CI}_{\widehat{\pi},95\%} = [\widehat{\pi} - 1.96\sigma_{\widehat{\pi}}\,,\, \widehat{\pi} + 1.96\,\sigma_{\widehat{\pi}}]$ where $\sigma_{\widehat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}}$
2. odds: $\mathrm{CI}_{\widehat{\Omega},95\%} = \left[\widehat{\Omega}\exp\left(-1.96\,\sigma_{\log\widehat{\Omega}}\right)\,,\, \widehat{\Omega}\exp\left(1.96\,\sigma_{\log\widehat{\Omega}}\right)\right]$ where $\sigma_{\log\widehat{\Omega}} = \sqrt{\frac{1}{D} + \frac{1}{n-D}}$
3. incidence rate: $\mathrm{CI}_{\widehat{\lambda},95\%} = \left[\widehat{\lambda}_\tau\exp\left(-1.96\sigma_{\log\widehat{\lambda}}\right)\,,\, \widehat{\lambda}\exp\left(1.96\sigma_{\log\widehat{\lambda}}\right)\right]$ where $\sigma_{\log\widehat{\lambda}} = \frac{1}{\sqrt{\widetilde{D}}}$
4. τ-years risk: $\mathrm{CI}_{\widehat{r},95\%} = [\widehat{r} - 1.96\,\sigma_{\widehat{r}}, \widehat{r} + 1.96\,\sigma_{\widehat{r}}]$ where $\sigma_{\widehat{r}} = \sqrt{\frac{r(1-r)}{n}}$

Confidence intervals:

Even though the estimate $\widehat{\bullet}$ will not exactly match the quantity of interest $\bullet$, we can derive an interval $\mathrm{CI}_{\widehat{\bullet},95\%}$ containing the target with high probability (here 0.95), e.g.:

$[\widehat{\bullet} - 1.96\,\sigma_{\widehat{\bullet}}, \widehat{\bullet} + 1.96\,\sigma_{\widehat{\bullet}}]$ (original scale)
$\left[\widehat{\bullet}\exp\left(-1.96\,\sigma_{\log\widehat{\bullet}}\right)\,,\, \widehat{\bullet}\exp\left(1.96\,\sigma_{\log\widehat{\bullet}}\right)\right]$ (log-scale)

where $\sigma_{\widehat{\bullet}}$ denotes the standard deviation of the estimate. As shown here different CIs exists and which one to use only really matters with small samples.

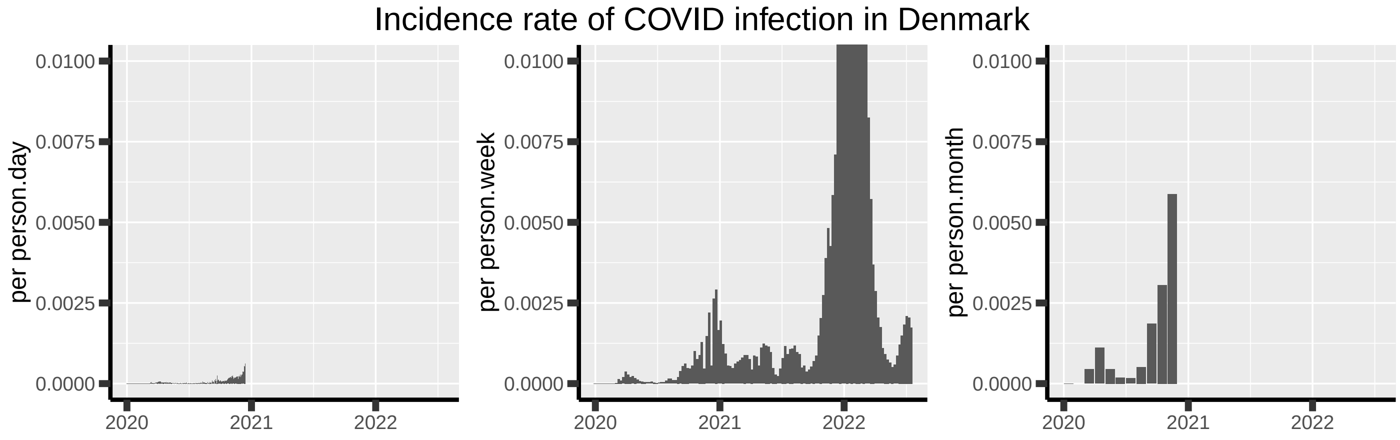
Illustration:

- Toy example with 4 patients:

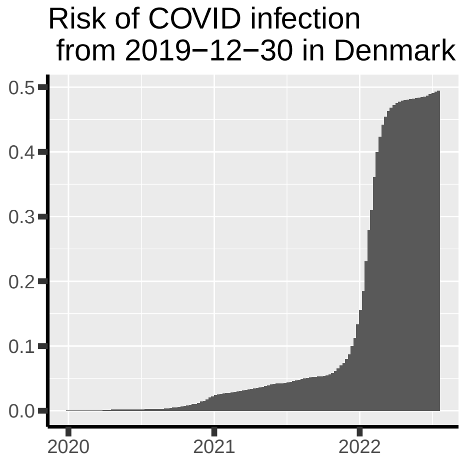
$\lambda = \frac{1+0+1+0}{2+8+5.9+8} = 0.08368$ per person.month or 83.68 per 1000 person.month
or
$\lambda = \frac{1+0+1+0}{2/12+8/12+5.9/12+8/12} = 1.004$ per person.year

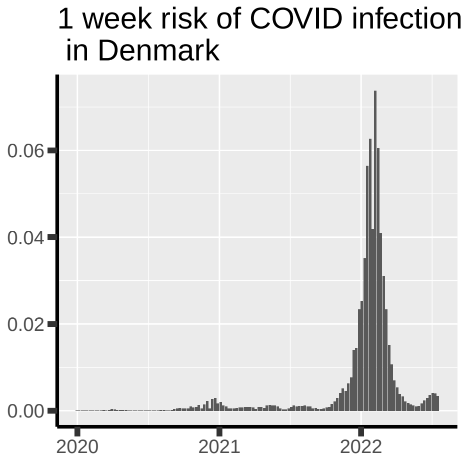
8 months risk: $r(8 \text{ months}) = 2/4 = 0.5 = 50\%$

- COVID dataset


Incidence rate of COVID infection in Denmark

Cumulated number of cases/Number of individuals (here 5840035)
vs. Number of cases for the week/Number of individual never infected (from 5840035 to 2950435)

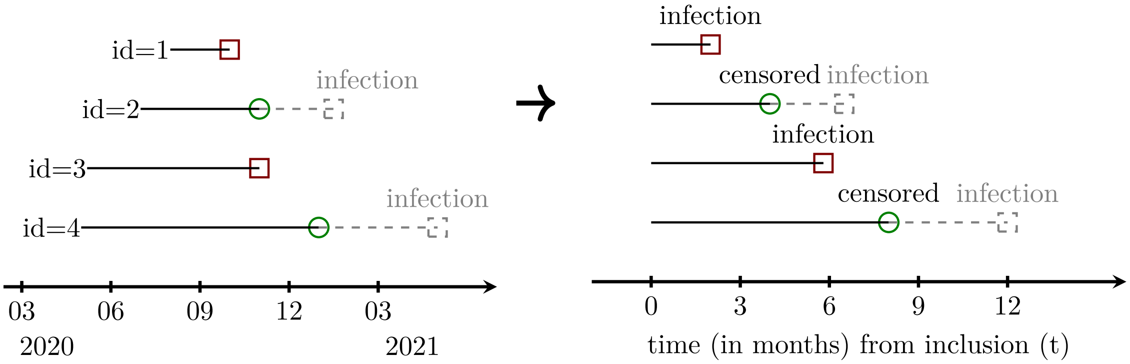

Risk of COVID infection
from 2019−12−30 in Denmark


1 week risk of COVID infection
in Denmark

3 Risk-rate relationship

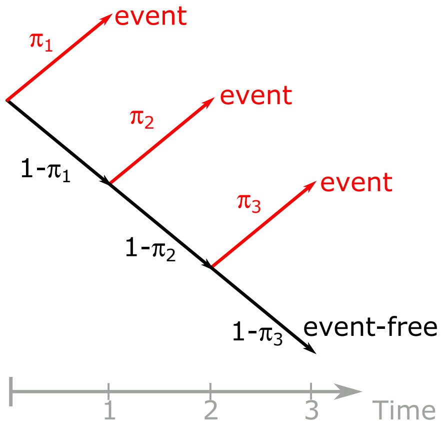
Motivations:

- we are usually interested in the effect of an exposure on the risk of an event. Covariates, such as vaccine or wearing a mask, generally impact the instantenous risk, i.e. the rate/hazard.
- in presence of right-censoring, modeling the rate and deducing the risk avoids to model the censoring mechanism



Sequence of **binary probability models**

- simple model assuming piecewise constant hazard, e.g. hazard may be week-dependent



- π_1 is the probability to have the event in the first time interval
- π_2 is the probability to have the event in the second time interval *for those who did not experience the event in the first time interval.*
- π_2 is the probability to have the event in the third time interval *for those who did not experience the event in the first two time intervals.*

Probability of getting the event:

$$\begin{aligned}\mathbb{P}\left[N=1,T\leq 3\right] &= \mathbb{P}\left[N=1,T=1\right]+\mathbb{P}\left[N=1,T=2\right]+\mathbb{P}\left[N=1,T=3\right]\\ &= \pi_1+(1-\pi_1)\pi_2+(1-\pi_1)(1-\pi_2)\pi_3\end{aligned}$$

Probability of not getting the event:

$$1-\mathbb{P}\left[N=1,T\leq 3\right]=(1-\pi_1)(1-\pi_2)(1-\pi_3)\approx\exp(-(\pi_1+\pi_2+\pi_3))=\exp(-\int_0^3\lambda(t)dt)$$

- assuming constant hazard: $r(\tau)\approx 1-\exp(-\lambda\tau)$

⚠ Approximation $r(\tau)\approx 1-\exp(-\int_{t=0}^{\tau}\lambda(t)dt)$ valid with $\lambda << 1$, e.g. small time intervals.

Illustration:

- Toy example with 4 patients: only two timepoints where infection can occur
  - at 2 months: 1 infected and 3 healthy, i.e. 4 at risk $\pi_1 = 1/4$
  - at 5.9 months: 1 infected and 1 healthy, i.e. 2 at risk $\pi_2 = 1/2$

  So the risk at 8 months is $(1 - \pi_1)(1 - \pi_2) = 3/8$

- COVID dataset: risk of infection or death between the start and the end of the period (771 days)

As the cumulated number of deaths divided by the population size

```
  infection       death
0.494792420 0.001129957
```

As 1 minus the product of 1 minus the weekly incidence rate

```
  infection       cases
0.494792420 0.001129957
```

which can be approximated by 1 minus the exponential of minus cumulated weekly incidence rate

```
  infection       cases
0.488263990 0.001129944
```

The approximation is very accurate for death as all incidence rate are small compared to one (max `0.00005154`) but less accurate for infection as some incidence rate are not small compared to 1 (max `0.0509600`).

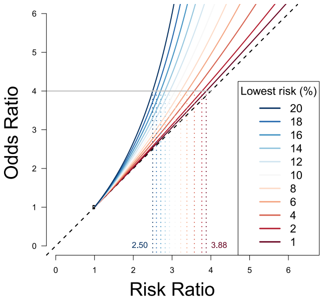
4 Measures of association

We denote by:

- r_E risk in the exposed group (n_E individuals, D_E failures)
  $\Omega_E = \frac{r_E}{1-r_E}$ odds in the exposed group
- $r_{\overline{E}}$ risk in the unexposed group ($n_{\overline{E}}$ individuals, $D_{\overline{E}}$ failures)
  $\Omega_{\overline{E}} = \frac{r_{\overline{E}}}{1-r_{\overline{E}}}$ odds in the unexposed group

Definition[2]

- risk difference: $RD = r_E - r_{\overline{E}}$
- relative risk / risk ratio: $RR = \frac{r_E}{r_{\overline{E}}}$
- risk odds ratio $OR = \frac{r_E}{r_{\overline{E}}}$



(graph courtesy of Paul Blanche)

– $RD > 0, RR > 1, OR > 1$: "harmful"
  (increased occurence of the outcome when exposed)

– $RD = 0, RR = 0, OR = 0$: "independent"
  (same occurence of the outcome for exposed and no-exposed)

– $RD < 0, RR < 1, OR < 1$ "protective"
  (decreased occurence of the outcome when exposed)

Estimation:

- risk difference: $\widehat{RD} = \hat{r}_E - \hat{r}_{\overline{E}} = \frac{D_E}{n_E} - \frac{D_{\overline{E}}}{n_{\overline{E}}}$
- relative risk / risk ratio: $\widehat{RR} = \frac{\hat{r}_E}{\hat{r}_{\overline{E}}} = \frac{D_E}{n_E} \Bigg/ \frac{D_{\overline{E}}}{n_{\overline{E}}}$
- odds ratio $\widehat{OR} = \frac{\hat{\Omega}_E}{\hat{\Omega}_{\overline{E}}} = \frac{D_E}{n_E - D_E} \Bigg/ \frac{D_{\overline{E}}}{n_{\overline{E}} - D_{\overline{E}}}$

Uncertainty:

- risk difference: $\sigma_{\widehat{RD}} = \sqrt{\sigma^2_{\hat{r}_E} + \sigma^2_{\hat{r}_{\overline{E}}}} = \sqrt{\frac{D_E(n_E - D_E)}{n_E^3} + \frac{D_{\overline{E}}(n_{\overline{E}} - D_{\overline{E}})}{n_{\overline{E}}^3}}$
- relative risk / risk ratio: $\sigma_{\log \widehat{RR}} = \sqrt{\frac{\sigma^2_{\hat{r}_E}}{\hat{r}^2_E} + \frac{\sigma^2_{\hat{r}_{\overline{E}}}}{\hat{r}^2_{\overline{E}}}} = \sqrt{\frac{1}{D_E} - \frac{1}{n_E} + \frac{1}{D_{\overline{E}}} - \frac{1}{n_{\overline{E}}}}$
- odds ratio $\sigma_{\log \widehat{OR}} = \sqrt{\frac{1}{D_E} + \frac{1}{n_E - D_E} + \frac{1}{D_{\overline{E}}} + \frac{1}{n_{\overline{E}} - D_{\overline{E}}}}$

[2] similar definitions hold for the prevalence and the incidence rate

5 Test of association: Chi-square test

As in the previous section we consider 2 groups and a binary outcome. We can summarize the data using the following 2x2 table:

Outcome Group	Survival	Failure	Total
Non-Exposed	$a = n_{\overline{E}} - D_{\overline{E}}$	$b = D_{\overline{E}}$	$n_{\overline{E}}$
Exposed	$c = n_E - D_E$	$d = D_E$	n_E
Total			n

As also mentioned in the previous section, the association between the outcome and the group variable can be assessed using an odds ratio (θ). It can be estimated by:

$$\widehat{\theta} = \frac{D_E}{n_E - D_E} \bigg/ \frac{D_{\overline{E}}}{n_{\overline{E}} - D_{\overline{E}}}$$

Testing the independence between the outcome and the group variable (i.e. $\theta = 1$) can be performed using a chi-squared test statistic:

$$\begin{aligned} t_{\chi^2} &= (n_{\overline{E}} + n_E)\frac{(n_{\overline{E}} - D_{\overline{E}})D_E - (n_E - D_E)D_{\overline{E}}}{n_{\overline{E}} n_E (D_E + D_{\overline{E}})(n_E + n_{\overline{E}} - D_{\overline{E}} - D_E)} \\ &= n\frac{(ad - bc)}{(a+b)(c+d)(a+c)(b+d)} \end{aligned}$$

which under $\theta = 1$ follows asymptotically a chi-square distribution with 1 degree of freedom.

Note 1: while commonly used and easy to explain, there are (often) better alternatives:

- testing whether the difference between probability of failure is 0 (or the ratio is 1). It is usually more interpretable (probability are easier to understand than odds ratios) but is not valid in case control studies and may not always be feasible when adjusting for covariates.

- in small samples: the Fisher's exact test is preferable
  (better type 1 error control)

Note 2: The chi-squared test is identical to a score test from a logistic regression[3]

[3] more on that later in the course

6 In R

For illustration, we will use the dataset BrCa which originates from a study about survival after breast cancer. The dataset contains information about the age and grade of the tumor, survival time after surgery, and outcome (alive or death) at end of follow-up.

6.1 Data management

We start by loading packages

```
library(ggplot2)
library(Epi)
library(survival)
```

and then load the dataset BrCa:

```
data(BrCa, package = "Epi")
```

To facilitate data visualization we restrict the dataset to columns useful for this demonstration:

```
BrCaR <- BrCa[,c("pid","age","grade","tox","xst")]
```

and rename the columns with more intuitive names:

```
names(BrCaR) <- c("id","age","grade","time","status")
```

We will re-order the dataset by time:

```
BrCaR <- BrCaR[order(BrCaR$time),]
```

and convert categorical variables to factor

```
BrCaR$id <- as.factor(BrCaR$id)
BrCaR$grade <- as.factor(BrCaR$grade)
BrCaR$status <- factor(BrCaR$status, levels = c("Alive","Dead"))
```

while keeping a binary version of the outcome (i.e. taking value 0 or 1):

```
BrCaR$status.bin <- as.numeric(BrCaR$status=="Dead")
```

We will also consider a subset of the dataset. To do so we identify the lines in the dataset corresponding to patient of grade 2:

```
index.grade2 <- which(BrCaR$grade == 2)
str(index.grade2)
```

```
int [1:794] 1 7 15 20 24 32 44 50 51 60 ...
```

for which the patient was alive at the end of the study:

```
index.Alive <- which(BrCaR$status == "Alive")
index.grade2Alive <- intersect(index.grade2,index.Alive)
```

Repeating the operation we can select the 5 patients for each combination of grade 2/3 and alive/dead:

```
index.grade3 <- which(BrCaR$grade == 3)
index.Dead <- which(BrCaR$status == "Dead")
BrCaR.subset <- rbind(BrCaR[intersect(index.grade2,index.Alive)[50:54],],
                      BrCaR[intersect(index.grade2,index.Dead)[50:54],],
                      BrCaR[intersect(index.grade3,index.Alive)[50:54],],
                      BrCaR[intersect(index.grade3,index.Dead)[50:54],])
```

6.2 Data visualization

We can display the first 6 lines with the `head` method (and the last 6 lines with the `tail` method)

```
head(BrCaR)
## head(BrCaR, 10) ## first 10 lines
## tail(BrCaR) ## last 6 lines
```

```
       id age grade       time status status.bin
1905  407  54     2 0.09856263  Alive          0
2945 3004  75     3 0.12320329   Dead          1
2334 2962  66     3 0.17522246   Dead          1
2949 2956  87     3 0.20260096   Dead          1
2815 2979  75     3 0.26557153   Dead          1
1956  537  58     3 0.27652293  Alive          0
```

The `str` method provides a concise display of the data:

```
str(BrCaR)
## summary(BrCaR) ## alternative
```

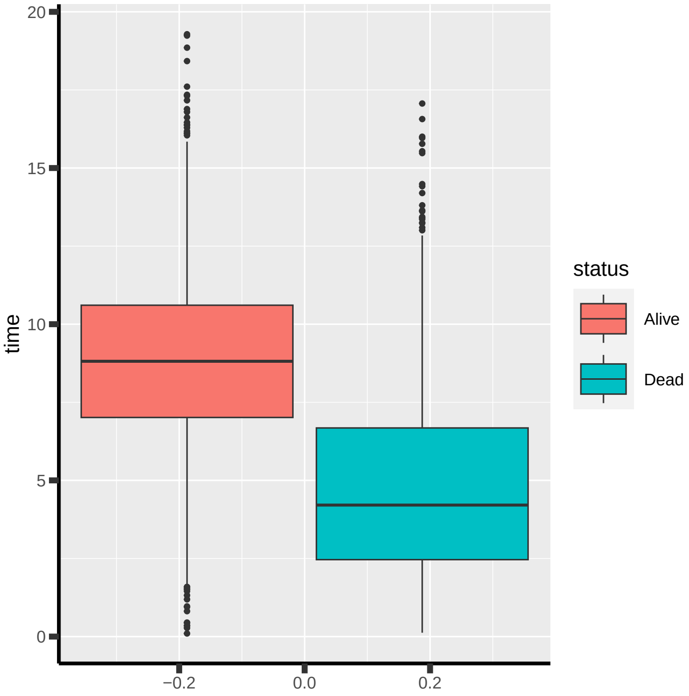
```
'data.frame':        2982 obs. of  6 variables:
 $ id        : Factor w/ 2982 levels "1","2","3","4",..: 402 2979 2937 2931 2954 531 2942 629 838 16
 $ age       : int  54 75 66 87 75 58 73 77 77 53 ...
 $ grade     : Factor w/ 2 levels "2","3": 1 2 2 2 2 2 1 2 2 2 ...
 $ time      : num  0.0986 0.1232 0.1752 0.2026 0.2656 ...
 $ status    : Factor w/ 2 levels "Alive","Dead": 1 2 2 2 2 1 2 2 1 1 ...
 $ status.bin: num  0 1 1 1 1 0 1 1 0 0 ...
```

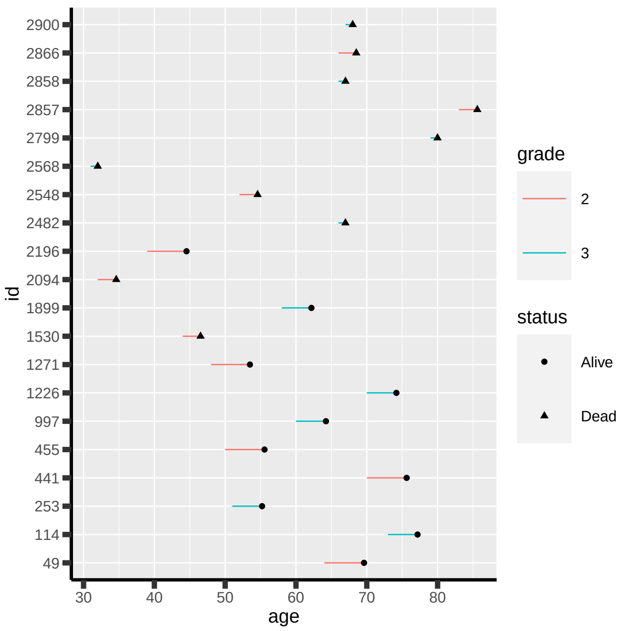
A boxplot of the time to event per type of event can be obtain using `ggplot`:

```
ggBox <- ggplot(BrCaR) + geom_boxplot(aes(y = time, fill = status))
ggBox
```

A display individual trajectories (here on a subset of the data) can also be obtained first displaying the time at risk using a segment and then points to indicate the type of event:

```
ggTraj <- ggplot(BrCaR.subset)
ggTraj <- ggTraj + geom_segment(aes(x = age, xend = age + time, y = id, yend = id,
                                    color = grade))
ggTraj <- ggTraj + geom_point(aes(x = age + time, y = id, shape = status), size = 2)
ggTraj
```





6.3 Contingency tables

The function `table` creates a 2 by 2 table, counting the number of occurences of the possible combinations between two variables. The function `rowSums` and `colSums` can be used to obtain, respectively, the total by row and by column:

```
t22 <- table(BrCaR$grade,
             outcome = BrCaR$status)
t22
```

```
 outcome
  Alive Dead
2   532  262
3  1178 1010
```

```
rowSums(t22)
```

```
  2    3
794 2188
```

```
colSums(t22)
```

```
Alive  Dead
 1710  1272
```

The functions `xtabs` can be used to sum values of variables per group:

```
t23.end <- xtabs(cbind(n=1, death = status.bin,person.year=time) ~ grade,
                 data = BrCaR)
t23.end
```

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

To restrict to 10 years follow-up, we should only count deaths happening within the first 10 years and limit to 10 the number of person.year for a given person:

```
t23.10 <- xtabs(cbind(n=1,
                      death = (time<=10)*status.bin,
                      person.year=pmin(time,10)) ~ grade,
                data = BrCaR)
t23.10
```

```
grade         n     death person.year
    2   794.000   231.000    5852.509
    3  2188.000   940.000   14149.946
```

6.4 Measures of disease frequency

Incidence rate: we estimate the incidence by dividing the number of deaths and person years:

```
D <- t23.10[,"death"]
Y <- t23.10[,"person.year"]
```

We can do that for each grade, leveraging that with `/` the division is performed element-wise when calling vectors:

```
lambda <- D/Y
lambda
```

```
         2          3
0.03947025 0.06643135
```

Confidence intervals can then be computed according to the previously seen formula:

```
qz <- qnorm(0.975) ## 1.96
sigma_loglambda <- 1/sqrt(D)
cbind(estimate = lambda,
      lower = lambda*exp(-qz*sigma_loglambda),
      upper = lambda*exp(qz*sigma_loglambda))
```

```
    estimate      lower      upper
2 0.03947025 0.03469484 0.04490294
3 0.06643135 0.06231749 0.07081679
```

Alternatively, one can use the `glm` function:
(the estimate is the same but the confidence intervals slightly difference as they are computed using a different method - profile likelihood)

```
## restrict to 10 years
BrCaR$status.bin10 <- BrCaR$status.bin*(BrCaR$time<=10)
BrCaR$time10 <- pmin(BrCaR$time,10)

e.pois <- glm(status.bin10 ~ grade-1, offset = log(time10),
              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 %
grade2 0.03947025 0.03459658 0.04478111
grade3 0.06643135 0.06227454 0.07076900
```

Risk: if there was no censoring (⚠) the risk could be computed by taking the ratio between the number of death and the population size:

```
n <- t23.10[,"n"]
r <- D/n
r
```

```
        2         3
0.2909320 0.4296161
```

Confidence intervals could then be computed using another previously mentioned formula:

```
qz <- qnorm(0.975) ## 1.96
sigma_r <- sqrt(r*(1-r)/n)
cbind(estimate = r,
      lower = r - qz*sigma_r,
      upper = r + qz*sigma_r)
```

```
   estimate     lower    upper
2 0.2909320 0.2593400 0.322524
3 0.4296161 0.4088742 0.450358
```

Alternatively, still if there was no censoring (⚠), an exact confidence interval could be obtained via the function `binom.test`:

```
stats::binom.test(x = D[1], n = n[1])
## stats::binom.test(x = c(D[1], n[1] - D[1])) ## same
```

```


        Exact binomial test

data:  D[1] and n[1]
number of successes = 231, number of trials = 794, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.2595358 0.3238894
sample estimates:
probability of success
              0.290932
```

Yet another way, still assuming no censoring (⚠), would be to use `glm`:

```
e.logitr <- glm(status.bin10 ∼ grade-1, family = binomial(link="identity"), data = BrCaR)
cbind(estimate = coef(e.logitr), confint(e.logitr))
```

```
Waiting for profiling to be done...
        estimate     2.5 %    97.5 %
grade2 0.2909320 0.2600490 0.3231635
grade3 0.4296161 0.4089657 0.4504314
```

A proper way to compute the risk in presence of right-censoring (assuming independent censoring) is to use the Kaplan-Meier estimator (more on that later on Day 7):

```
e.KM <- survfit(Surv(time, status) ∼ grade, data=BrCaR)
summary(e.KM, time = c(0.3, 0.55, 10))
```

```
Call: survfit(formula = Surv(time, status) ~ grade, data = BrCaR)

                grade=2
  time n.risk n.event Pr((s0)) Pr(Dead)
  0.30    792       1    0.999  0.00126
  0.55    790       1    0.997  0.00252
 10.00    232     229    0.665  0.33481
```

```
               grade=3 
  time n.risk n.event Pr((s0)) Pr(Dead)
  0.30   2182       5    0.998  0.00229
  0.55   2172       7    0.995  0.00549
 10.00    453     928    0.510  0.48984
```

We can see substantial underestimation of the risk when treating censoring as no event (about 5% too low). These results can easily be replicated 'by hand' for the first timepoints where there are no ties (i.e. single event per timepoint)

- select the patients with grade 2 and focus on the first 5 timepoitns

```
BrCaR.grade2 <- BrCaR[index.grade2,]
BrCaR.grade2 <- BrCaR.grade2[BrCaR.grade2$time %in% BrCaR.grade2$time[1:5],]
BrCaR.grade2
```

```
       id age grade       time status status.bin status.bin10     time10
1905  407  54     2 0.09856263  Alive          0            0 0.09856263
2095 2967  73     2 0.27652293   Dead          1            1 0.27652293
1055 1147  60     2 0.44079399  Alive          0            0 0.44079399
2818 2147  40     2 0.54209447   Dead          1            1 0.54209447
2904 2998  76     2 0.58863791   Dead          1            1 0.58863791
```

- compute the hazard

```
atRisk <- length(index.grade2)-0:4
haz <- (BrCaR.grade2$status=="Dead")/atRisk
haz
```

```
[1] 0.000000000 0.001261034 0.000000000 0.001264223 0.001265823
```

- use the product limit formula to deduce the risk:

```
1-cumprod(1-haz)
```

```
[1] 0.000000000 0.001261034 0.001261034 0.002523662 0.003786291
```

6.5 Measures of disease association

Incidence: we can compute the incidence difference / ratio directly from the estimated incidence:

```
ID <- lambda["3"]-lambda["2"]
ID

IR <- lambda["3"]/lambda["2"] ## select element named 3 or name 2
lambda[2]/lambda[1] ## same (select 2nd or 1st element)
```

```
        3
0.0269611
       3
1.683074
```

The `unname` function can be used to remove names:

```
IR <- unname(IR)
IR
```

```
[1] 1.683074
```

A confidence interval for the ratio can be obtained by summing the variance on the log scale:

```
sigma_logIR <- sqrt(sum(1/D))
c(estimate = IR,
  lower = IR / exp(qz * sigma_logIR),
  upper = IR * exp(qz * sigma_logIR))
```

```
estimate    lower    upper
1.683074 1.457453 1.943623
```

Alternatively the `glm` function can be used:

```
e.pois <- glm(status.bin10 ~ grade, offset = log(time10),
              family = poisson(link="log"), data = BrCaR)
cbind(exp(coef(e.pois)), exp(confint(e.pois)))
```

```
Waiting for profiling to be done...
                            2.5 %     97.5 %
(Intercept) 0.03947025 0.03459658 0.04478111
grade3      1.68307401 1.46040680 1.94783461
```

or the `effx` function from the Epi package:

```
effx(response = BrCaR$status.bin,
     exposure = BrCaR$grade,
     fup = BrCaR$time, type = "failure", eff = "RR")
## use eff = "RD" for incidence difference
```

```
---------------------------------------------------------------------------
response      :  BrCaR$status.bin
type          :  failure
exposure      :  BrCaR$grade

BrCaR$grade is a factor with levels: 2 / 3
baseline is  2
effects are measured as rate ratios
---------------------------------------------------------------------------

effect of BrCaR$grade on BrCaR$status.bin
number of observations  2982

Effect   2.5%  97.5%
  1.63   1.42   1.87

Test for no effects of exposure on 1 df: p-value= 1.56e-13
There were 50 or more warnings (use warnings() to see the first 50)
```

risk: similarly we can compute the risk difference or risk ratio directly from the estimated risk:

```
r[2]-r[1]
r[2]/r[1]
```

```
        3
0.1386841
       3
1.476689
```

or from `glm`, still assuming no censoring (⚠):

```
e.logitr <- glm(status.bin10 ~ grade, family = binomial(link="identity"), data = BrCaR)
cbind(coef(e.logitr), confint(e.logitr))[2,]
```

```
Waiting for profiling to be done...
              2.5 %    97.5 %
0.1386841 0.1004610 0.1760154
```

```
e.logitr <- glm(status.bin10 ~ grade, family = binomial(link="log"), data = BrCaR)
cbind(exp(coef(e.logitr)), exp(confint(e.logitr)))[2,]
```

```
Waiting for profiling to be done...
            2.5 %   97.5 %
1.476689 1.314824 1.667895
```

To account for censoring, we would need to extract the risks obtained with the Kaplan-Meier estimator:

```
eS.KM <- summary(e.KM,times=10)
RD <- diff(eS.KM$pstate[,2])
RD
```

```
[1] 0.1550246
```

the associated standard errors[4], and deduce the confidence interval:

```
sigma_RD <- sqrt(sum(eS.KM$std.err[,2]^2))
c(estimate = RD,
  lower = RD - qz * sigma_RD,
  upper = RD + qz * sigma_RD)
```

```
 estimate     lower     upper
0.1550246 0.1107735 0.1992758
```

[4] The confidence interval are computed using a transformation (log by default) but the standard error is the untransformed one. This can be verified by re-fitting the survfit object with the argument `conf.type="plain"`

7 Likelihood theory

Maximum likelihood: method using an iid[5] sample to estimate model parameters

- define a statistical model (blinded to the data)
  $\mathbb{P}[Y=1]=\pi$ and $\mathbb{P}[Y=0]=1-\pi$

- express the likelihood (probability of observing the data given the model)
  We followed n patients and D of them died. The likelihood of a death is π and of a survival is $1-\pi$.
  So the likelihood associated to the sample is $\mathcal{L}(\pi)=\prod_{i=1}^{n}\mathcal{L}_i(\pi)=\pi^D(1-\pi)^{n-D}$.

- express the log-likelihood (probability of observing the data given the model)
  $\ell(\pi)=D\log(\pi)+(n-D)\log(1-\pi)$.
  See left panel of figure 1 for an illustration.

- find the parameter value maximizing the likelihood (MLE), i.e. solve[6] $\frac{d\ell(\pi)}{d\pi}=0$
  The first derivative of the likelihood (called score) is $\frac{d\ell(\pi)}{d\pi}=\frac{D}{\pi}-\frac{n-D}{1-\pi}$. See the middle panel of figure 1 for an illustration. It take value 0 at $\widehat{\pi}=\frac{D}{n}$.

- quantify the variance of the MLE
  - express the second derivative of the likelihood
    $\frac{d^2\ell(\pi)}{d\pi^2}=-\frac{D}{\pi^2}-\frac{n-D}{(1-\pi)^2}=-\frac{D-2\pi D+\pi^2 n^2}{\pi^2(1-\pi)^2}=-\frac{n\widehat{\pi}(1-2\pi+\pi^2/\widehat{\pi})}{\pi^2(1-\pi)^2}$
  - evaluate the opposite of its inverse at the MLE
    $\widehat{\sigma}^2_{\widehat{\pi}}=-\left\{\left(\frac{d^2\ell(\pi)}{d\pi^2}\right)\Big|_{\pi=\widehat{\pi}}\right\}^{-1}=\frac{\widehat{\pi}(1-\widehat{\pi})}{n}$

- For large enough sample size, the MLE is unbiased and normally distributed
  $\widehat{\pi}\sim\mathcal{N}\left(\pi,\sigma^2_{\widehat{pi}}\right)$

Wald test: use the asymptotic distribution of the estimate to test an hypothesis

$\mathcal{H}_0:\pi=0.25$, Wald statistic $t_W=\frac{\widehat{\pi}-0.25}{\sigma_{\widehat{\pi}}}\sim\mathcal{N}(0,1)$, p-value $p_W=2(1-F_{\mathcal{N}}(|t_W|))$
where $F_{\mathcal{N}}$ is the cdf[7] of a standard normal distribution

Likelihood ratio test (LRT): use the log-likelihood under a specific restriction to test an hypothesis

$\mathcal{H}_0:\pi=0.25$, maximum log-likelihood $\ell(\widehat{\pi})$, log-likelihood under $\mathcal{H}_0$ $\ell(0.25)$,
LRT statistic $t_{LRT}=-2\left(\ell(0.25)-\ell(\widehat{\pi})\right)\sim\chi^2_1$, p-value $p_W=1-F_{\chi^2_1}(t_{LRT}))$
where $F_{\chi^2_1}$ is the cdf of a chi-squared distribution with one degree of freedom.

[5] independent and identically distributed
[6] one should also check that the second derivative of the likelihood is negative
[7] cumulative distribution function

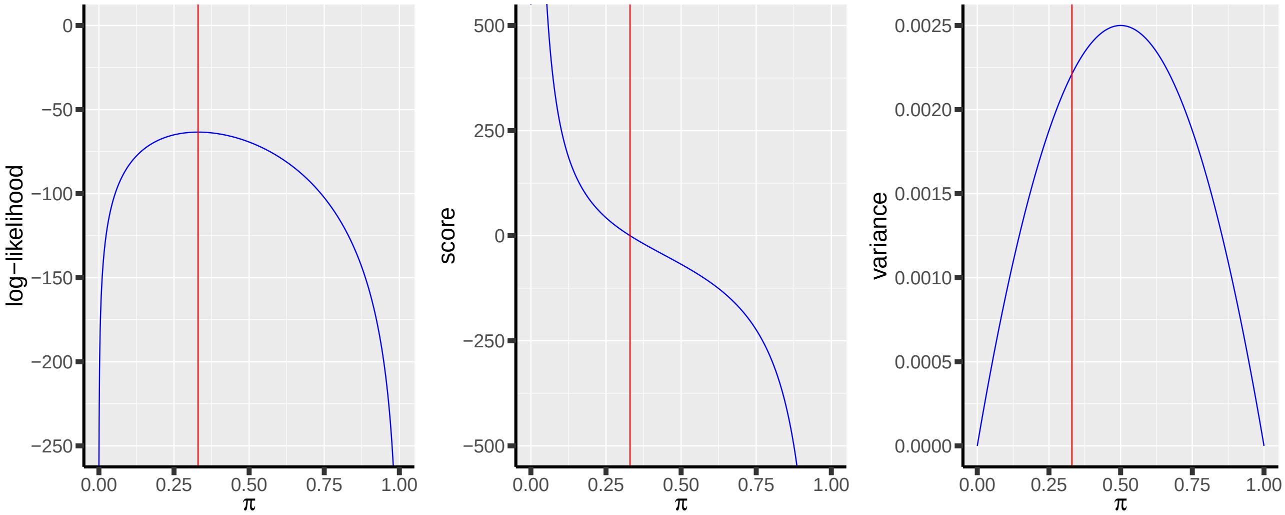


Figure 1: Illustration of the log-likelihood function, its first derivative, and the variance of the estimate for $n = 100$ and $D = 33$.