Ph.D. course: Epidemiological methods in medical research

# Lecture 2: Measures of disease frequency and association

Brice Ozenne[1,2] - brice.ozenne@nru.dk

[1] Section of Biostatistics, Department of Public Health, University of Copenhagen
[2] Neurobiology Research Unit, University Hospital of Copenhagen, Rigshospitalet.

9th January 2025

# Epidemiology (very short!)

Study of distribution and determinants of *disease frequency* in human populations.

The outcome is typically a **time varying binary** variable
(e.g. alive/dead, healthy/infected, . . . )

Measures of disease frequency:

- **prevalence**, **incidence rate**, **hazard rate**, **risk**

Comparison of frequency between exposure groups:

- **difference**, **ratio**, **odds**

# Need for statistical tools

Making exposed and non-exposed comparable

- e.g. adjustment for covariates in observational studies

Handling complications

- missing values (e.g. due to drop-out),
  competing events (e.g. death),
- time varying effects (e.g. seasonal variations)
  dynamic treatment regimes (switch of treatment), . . .

Understand complex effects
(e.g. treatment effect dependent on baseline covariates)

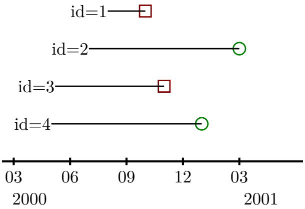
Working with finite samples (quantitying uncertainty)

# Case study (Beyersmann et al., 2014)

**Aim**: assess the impact of pneumonia diagnosis on ICU mortality

**Design**: cohort of 1876 patients admitted in ICU (time 0) are followed until death or discharged (no censoring)

**Data**: for each group we observe something like
(follow-up time has been artificially increased to ease visualization)

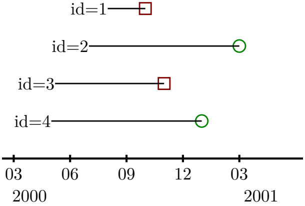


# Case study (Beyersmann et al., 2014)

**Aim**: assess the impact of pneumonia diagnosis on ICU mortality

**Design**: cohort of 1876 patients admitted in ICU (time 0) are followed until death or discharged (no censoring)

**Data**: for each group we observe something like (follow-up time has been artificially increased to ease visualization)



What can we do with this data

# Data representation

General case:

- status: alive/dead, **healthy/sick**, 0/1
- group: no pneumonia/pneumonia, **unexposed/exposed**, 0/1

# Individual data (artifical example)

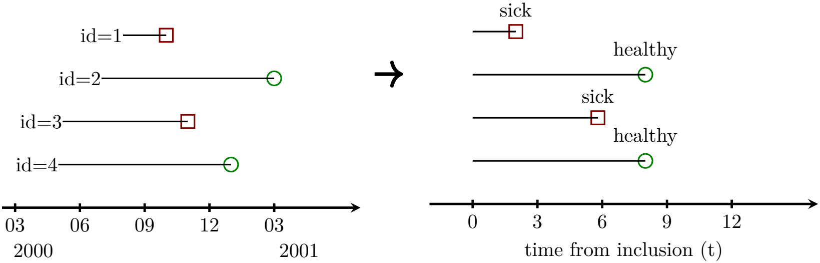
**Individual data**: one line per subject

```
patient  inclusion         end  status exposed
    id1 01-08-2000 01-10-2000    sick      no
    id2 01-07-2000 01-03-2001 healthy      no
    id3 02-05-2000 01-11-2001    sick      no
    id4 01-05-2000 01-01-2001 healthy      no
    id5 01-04-2000 01-08-2000    sick     yes
    id6 01-03-2000 01-09-2000 healthy     yes
    id7 02-06-2000 01-02-2001 healthy     yes
    id8 01-08-2000 01-03-2001    sick     yes
```

Compare disease frequency between exposure groups

→ for convenience, focus on the non-exposed individuals

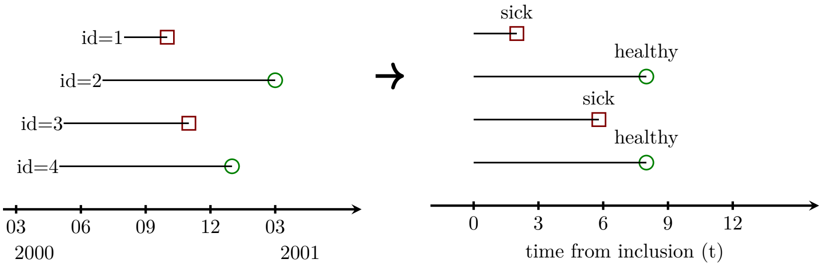
# Representation of individual data



For subject $i \in \{1, \dots, n\}$:

- $T_i^* \in [0, +\infty[$ time to event (in months, years, ...)
- $T_i$ observed time to event, typically $T_i = \min(T_i^*, \tau)$ where $\tau$ is the study time (here 8 months).
- $\Delta_i = \mathbb{1}_{T_i = T_i^*} \in \{0, 1\}$ event indicator (healthy/sick, alive/dead, ...)

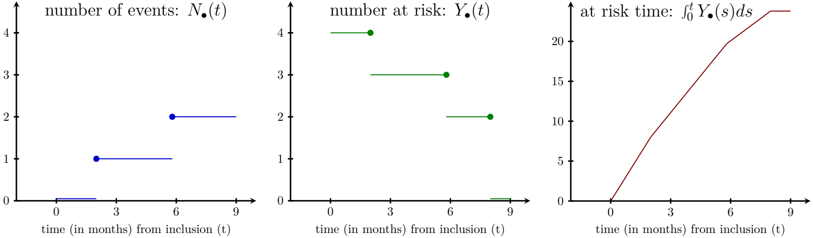
# Representation of individual data



- $T_1^* = 2,\ T_2^* =? \geq 8,\ T_3^* = 5.9,\ T_4^* =? \geq 8$
- $T_1 = 2,\ T_2 = 8,\ T_3 = 5.9,\ T_4 = 8$
- $\Delta_1 = 1,\ \Delta_2 = 0,\ \Delta_3 = 1,\ \Delta_4 = 0$

# Counting process representation

The data can be summarized using a counting process:



Bivariate outcome:

- $N_\bullet(t) = \sum_{i=1}^n \mathbb{1}_{T_i \leq t, \Delta = 1}$ number of events by time $t$.
- $Y_\bullet(t) = \sum_{i=1}^n \mathbb{1}_{T_i \geq t}$ number of individuals at risk at time $t$.
- $\int_0^t Y_\bullet(s)ds$ cumulated time at risk (in months).

# Individual vs. aggregated data

**Individual data**: one line per subject

```
patient  inclusion        end time  status
    id1 01-08-2000 01-10-2000  2.0    sick
    id2 01-07-2000 01-03-2001  8.0 healthy
    id3 02-05-2000 01-11-2001  5.9    sick
    id4 01-05-2000 01-01-2001  8.0 healthy
```

**Aggregated data**: one line per timepoint:

```
interval start time N Y risk.time dN drisk.time
       1   0.0  2.0 1 4       8.0  1        8.0
       2   2.0  5.9 2 3      19.7  1       11.7
       3   5.9  8.0 2 2      23.9  0        4.2
```

# In R (1/2)

```
dtL.toy <- survSplit(Surv(time,status=="sick")~patient,
                     data = dt.toy[exposed=="no",],
                     cut = c(2,5.9,8), episode = "interval")
dtL.toy
```

```
  patient tstart time event interval
1     id1    0.0  2.0     1        1
2     id2    0.0  2.0     0        1
3     id2    2.0  5.9     0        2
4     id2    5.9  8.0     0        3
5     id3    0.0  2.0     0        1
6     id3    2.0  5.9     1        2
7     id4    0.0  2.0     0        1
8     id4    2.0  5.9     0        2
9     id4    5.9  8.0     0        3
```

# In R (2/2)

```
dtS.toy <- aggregate(cbind(dN = event,
                           drtime = time-tstart,
                           Y = 1)~interval,
                     data = dtL.toy, FUN = "sum")
dtS.toy
```

```
  interval dN drtime Y
1        1  1    8.0 4
2        2  1   11.7 3
3        3  0    4.2 2
```

```
dtS.toy$N <- cumsum(dtS.toy$dN)
dtS.toy$risk.time <- cumsum(dtS.toy$drtime)
dtS.toy
```

```
  interval dN drtime Y N risk.time
1        1  1    8.0 4 1       8.0
2        2  1   11.7 3 2      19.7
3        3  0    4.2 2 2      23.9
```

## Historical (!) example

Weekly national-level ECDC data on COVID-19 (https://github.com/kjhealy/covdata)

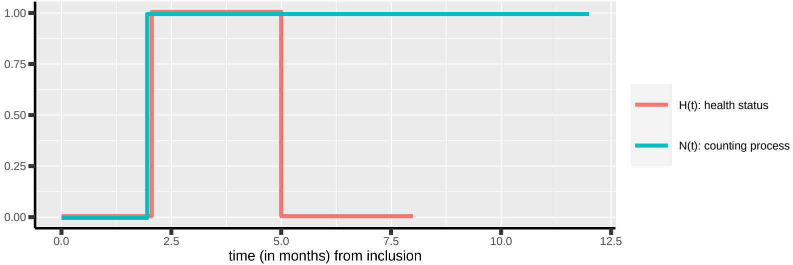
```
           date country population cases deaths
  1: 2019-12-30 Denmark    5840045    10      0
  2: 2020-01-06 Denmark    5840045    12      0
  3: 2020-01-13 Denmark    5840045     8      0
  4: 2020-01-20 Denmark    5840045    15      0
  5: 2020-01-27 Denmark    5840045    13      0
 ---
130: 2022-06-20 Denmark    5840045  8696     17
131: 2022-06-27 Denmark    5840045 10720     33
132: 2022-07-04 Denmark    5840045 12264     32
133: 2022-07-11 Denmark    5840045 11965     41
134: 2022-07-18 Denmark    5840045 10171     40
```

# Counting process vs. health status

$N_\bullet(t)$

- indicates whether an event has occured
- not the number of patients still affected by the event, (this will be denoted $H_\bullet(t)$)

Illustration when the infection lasts 3 months:



# Back to the case study (Beyersmann et al., 2014)

**Aim**: assess the impact of pneumonia diagnosis on ICU mortality

**Design**: cohort of 1876 patients admitted in ICU (time 0) are followed until death or discharged (no censoring)

**Data**:

- 220 patients with pneumonia: 6161 days at ICU  
  48 died before discharge
- 1656 patients without pneumonia: 22 337 days at ICU  
  166 died before discharge

What can we do with this data

# Measures of disease frequency

(under no or only administrative censoring)

# Prevalence

**Definition**: proportion of people with a disease (at a given time $t$)

$$\pi(t) = \mathbb{P}\left[H(t) = 1\right]$$

- $\pi \in [0,1]$, $\pi = \begin{cases} 0 \text{ nobody has the disease} \\ 1 \text{ everybody has the disease} \end{cases}$

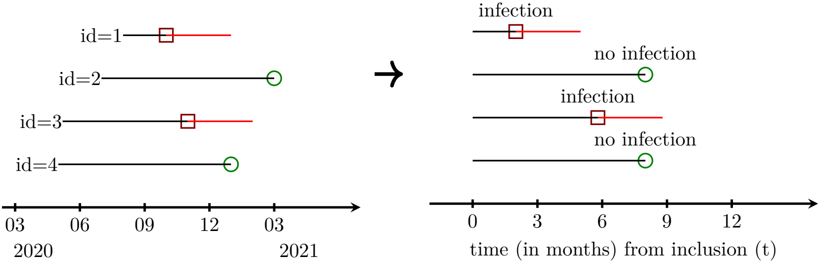
**Estimation**: $\frac{\text{"number of people with the disease"}}{\text{"number of people"}}$

$$\hat{\pi}(t) = \frac{H_\bullet(t)}{n} = \frac{1}{n}\sum_{i=1}^{n} H_i(t) \text{ when } H_i \text{ is binary } 0/1$$

# Prevalence - example 1

Assumes that:

- the infection lasts 3 months for everybody
- no re-infection

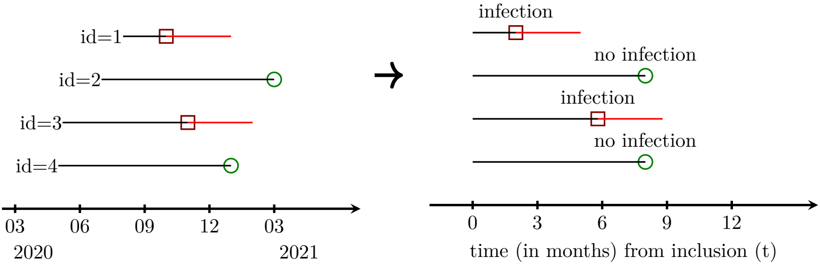


- $\widehat{\pi}(0) =$ at baseline
- $\widehat{\pi}(3) =$ after 3 months
- $\widehat{\pi}(8) =$ after 8 months

# Prevalence - example 1

Assumes that:

- the infection lasts 3 months for everybody
- no re-infection

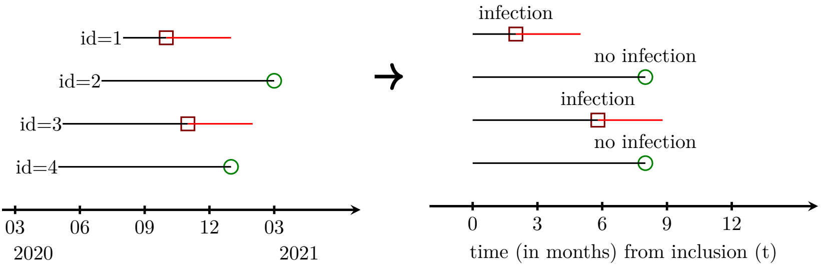


- $\widehat{\pi}(0) =$ at baseline
- $\widehat{\pi}(3) =$ after 3 months
- $\widehat{\pi}(8) =$ after 8 months

# Prevalence - example 1

Assumes that:

- the infection lasts 3 months for everybody
- no re-infection



- $\widehat{\pi}(0) = 0$ at baseline
- $\widehat{\pi}(3) = 1/4$ after 3 months
- $\widehat{\pi}(8) = 1/4$ after 8 months

# Prevalence - limitation

**Example 2.2 from Kestenbaum (2019)**:
Prevalence of multiple sclerosis (MS):

- vitamin D deficient individuals (VD-): $\hat{\pi}_{VD-} = 0.3\%$
- vitamin D sufficient individuals (VD+): $\hat{\pi}_{VD+} = 0.1\%$

**Interpretation**:

- ?
- ?
- ?

# Prevalence - limitation

**Example 2.2 from Kestenbaum (2019)**:
Prevalence of multiple sclerosis (MS):

- vitamin D deficient individuals (VD-): $\hat{\pi}_{VD-} = 0.3\%$
- vitamin D sufficient individuals (VD+): $\hat{\pi}_{VD+} = 0.1\%$

**Interpretation**:

- VD- causes MS
- MS causes VD-
- VD- and MS have a common cause

⚠ Prevalence data **alone** are insufficient for establishing a temporal relationship between outcome and exposure

## Risk / cumulative incidence

**Definition**: proportion of people *becoming* sick by time $t$

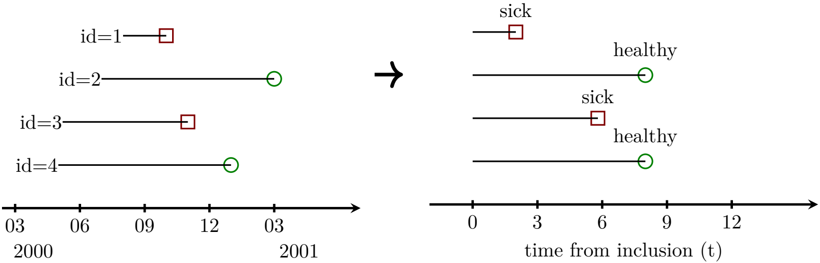
$$r(t) = \mathbb{P}\left[T^* \leq t, \Delta = 1\right]$$

- $r(0) = 0$ i.e. $T^* > 0$
- $r \in [0,1]$, $r = \begin{cases} 0 \text{ nobody will get the disease} \\ 1 \text{ everybody will get the disease} \end{cases}$
- $r(t)$ is non-decreasing with $t$

**Estimation (no censoring)**: $\frac{\text{"number of new cases"}}{\text{"number of persons at risk"}}$

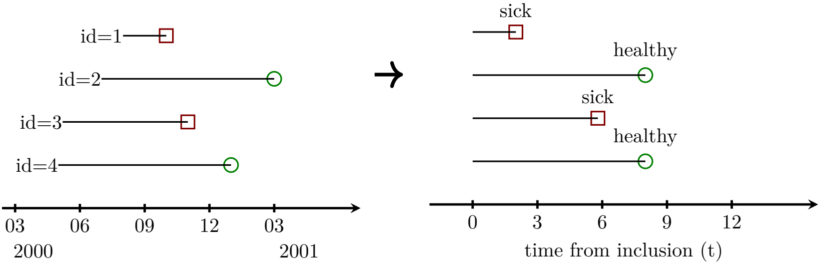
$$\hat{r}(t) = \frac{N_\bullet(t)}{n} = \frac{1}{n}\sum_{i=1}^{n} N_i(t) \text{ when } N_i \text{ is binary } 0/1$$

# Risk - example 1



- $\widehat{r}(0) =$ at baseline
- $\widehat{r}(3) =$ after 3 months
- $\widehat{r}(8) =$ after 8 months

# Risk - example 1



- $\widehat{r}(0) = 0$ at baseline
- $\widehat{r}(3) = 1/4$ after 3 months
- $\widehat{r}(8) = 2/4$ after 8 months

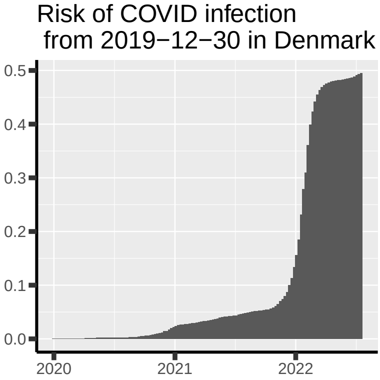
# Risk - example 2

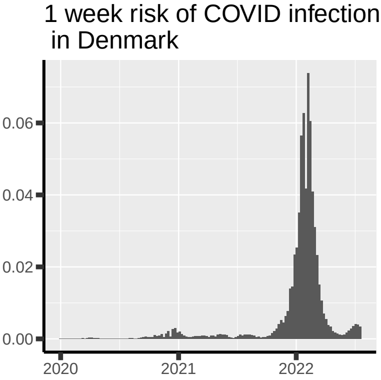
- `population`: population size at the start of COVID
- `atRisk`: (approximate) number of COVID naive people
- `cases` number COVID cases detected during the week
- `cu_cases` cumulative number of COVID cases

```
           date country population  atRisk cu_cases cases
  1: 2019-12-30 Denmark    5840045 5840045       10    10
  2: 2020-01-06 Denmark    5840045 5840035       22    12
  3: 2020-01-13 Denmark    5840045 5840023       30     8
 ---
132: 2022-07-04 Denmark    5840045 2984835  2867474 12264
133: 2022-07-11 Denmark    5840045 2972571  2879439 11965
134: 2022-07-18 Denmark    5840045 2960606  2889610 10171
```

Risk as `cu_cases/population` or `cases/atRisks`

# Example 2 - illustration





There is no such thing as 'the risk'!

- dependents on the time horizon
- and on the initial time

# Incidence rate

**Definition**: frequency at which an event occurs per unit of time

**Estimation**: $\frac{\text{”number of new cases"}}{\text{"cumulative time at risk"}}$ (incidence rate)

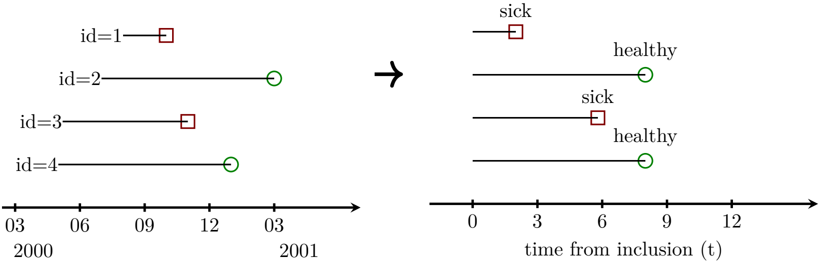
$$\widehat{\lambda} = \frac{N_\bullet(t)}{\int_0^t Y_\bullet(s)ds} = \frac{\sum_{i=1}^n N_i(t)}{\sum_{i=1}^n \min(T_i, t)}$$

⚠ unit (person.time $^{-1}$)

$\widehat{\lambda} = 0.001$ person.month $= 1$ per 1000 person.month
$= 12$ per 1000 person.year

$\widehat{\lambda} > 1$ is "un-natural" (for non-recurrent event)
typically due to extrapolation beyond the follow-up time

# Incidence rate - example with $\tau = 8$ months



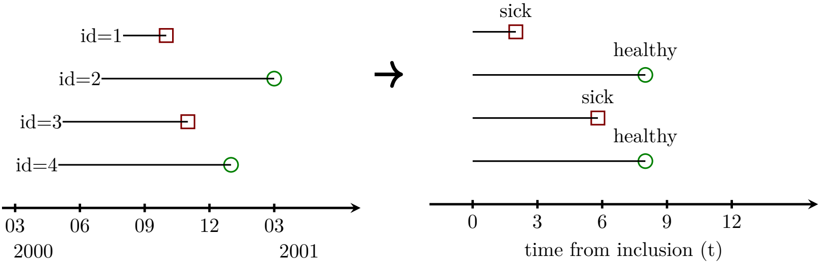
- $T_1 = 2$ months, $\Delta_1 = 1$
- $T_2 = 8$ months, $\Delta_2 = 0$
- $T_3 = 5.9$ months, $\Delta_3 = 1$
- $T_4 = 8$ months, $\Delta_4 = 0$

$\widehat{\lambda}(\tau) =$ $\approx$ per person-month

$\approx$ per 1000 person-month

$\approx$ per person-year

# Incidence rate - example with $\tau = 8$ months



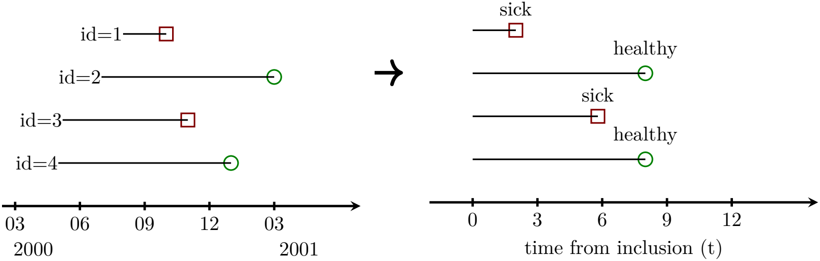
- $T_1 = 2$ months, $\Delta_1 = 1$
- $T_2 = 8$ months, $\Delta_2 = 0$
- $T_3 = 5.9$ months, $\Delta_3 = 1$
- $T_4 = 8$ months, $\Delta_4 = 0$

$$\widehat{\lambda}(\tau) = \frac{1+0+1+0}{2+8+5.9+8} = \frac{2 \text{ new cases}}{23.8 \text{ person-month}} \approx 0.084 \text{ per person-month}$$

$$\approx 84 \text{ per 1000 person-month}$$

$$\approx \qquad \text{per person-year}$$

# Incidence rate - example with $\tau = 8$ months



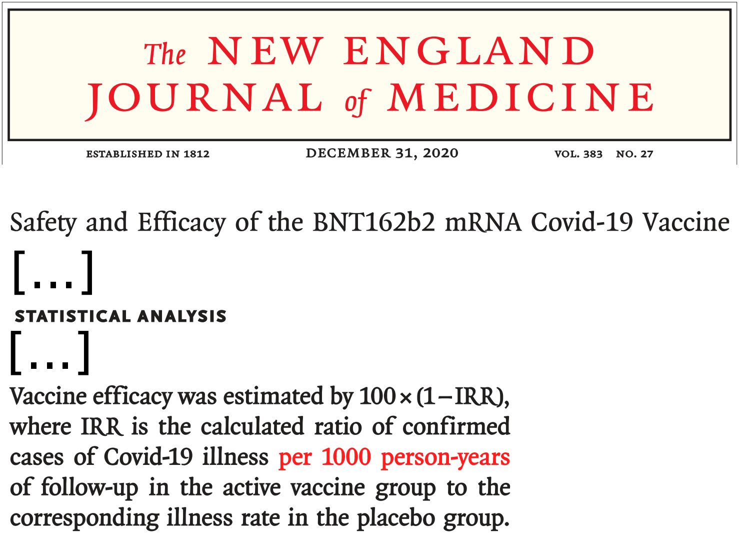
- $T_1 = 2$ months, $\Delta_1 = 1$
- $T_2 = 8$ months, $\Delta_2 = 0$
- $T_3 = 5.9$ months, $\Delta_3 = 1$
- $T_4 = 8$ months, $\Delta_4 = 0$

$$\widehat{\lambda}(\tau) = \frac{1+0+1+0}{2+8+5.9+8} = \frac{\text{2 new cases}}{\text{23.8 person-month}} \approx 0.084 \text{ per person-month}$$

$$\approx 84 \text{ per 1000 person-month}$$

$$\frac{\text{2 new cases}}{\text{23.8/12 person-year}} \approx 1.004 \text{ per person-year}$$

# Person-year in the litterature



The NEW ENGLAND JOURNAL of MEDICINE

ESTABLISHED IN 1812 DECEMBER 31, 2020 VOL. 383 NO. 27

Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine

[...]

STATISTICAL ANALYSIS

[...]

Vaccine efficacy was estimated by 100 × (1 − IRR), where IRR is the calculated ratio of confirmed cases of Covid-19 illness per 1000 person-years of follow-up in the active vaccine group to the corresponding illness rate in the placebo group.

## Hazard rate

The estimation of the incidence rate as $\frac{\text{”number of new cases"}}{\text{"cumulative time at risk"}}$ assumed a constant rate

# Hazard rate

The estimation of the incidence rate as $\frac{\text{"number of new cases"}}{\text{"cumulative time at risk"}}$ assumed a constant rate

- within a time interval

# Hazard rate

The estimation of the incidence rate as $\frac{\text{"number of new cases"}}{\text{"cumulative time at risk"}}$ assumed a constant rate

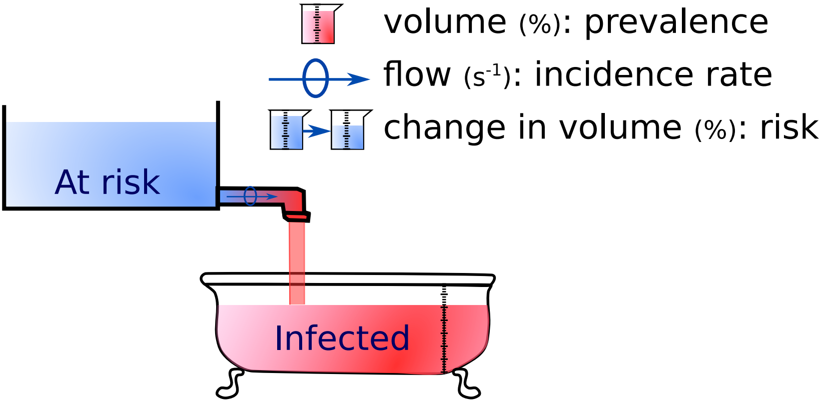
- within a time interval

A more general expression would be:

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}\left[t \leq T^* < t + dt, \Delta = 1 | T^* \geq t\right]}{dt}$$

- how likely an event is to occur in the next instant, given that it has not occurred yet
- called hazard rate
- $\lambda(t) \in [0, +\infty[$: higher values → higher disease frequency

## Graphical summary

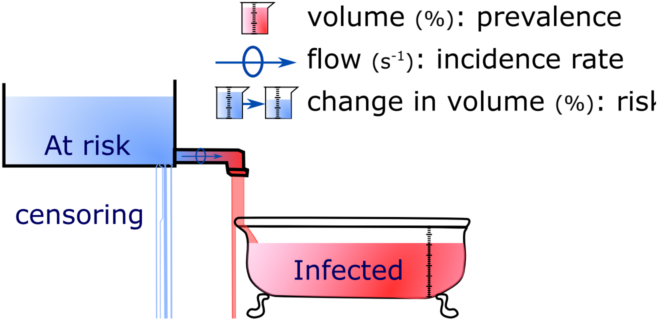


volume (%): prevalence

flow ($s^{-1}$): incidence rate

change in volume (%): risk

# Handling right-censoring

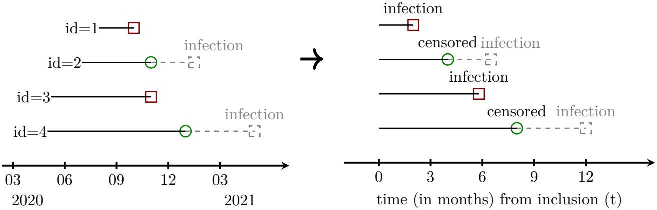


volume (%): prevalence

flow (s$^{-1}$): incidence rate

change in volume (%): risk

# Another cohort, with random right-censoring



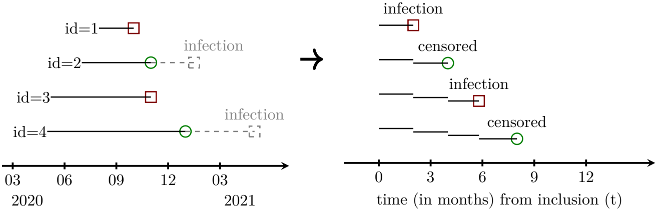
Risk after 8 months:

- $\widehat{r}(8) =$

Incidence:

- $\widehat{\lambda}_1 =$ $t \in [0; 2]$
- $\widehat{\lambda}_2 =$ $t \in [2; 4]$
- $\widehat{\lambda}_3 =$ $t \in [4; 5.9]$
- $\widehat{\lambda}_4 =$ $t \in [5.9; 8]$

# Another cohort, with random right-censoring



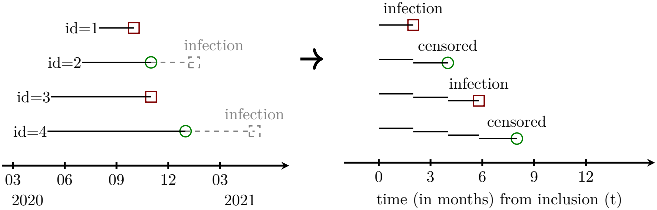
Risk after 8 months:

- $\widehat{r}(8) = (2+?)/4 = 0.5$ or $0.75$

Incidence:

- $\widehat{\lambda}_1 = 1/(2+2+2+2) = 1/8$ $\qquad t \in [0;2]$
- $\widehat{\lambda}_2 = 0/(2+2+2) = 0$ $\qquad t \in [2;4]$
- $\widehat{\lambda}_3 = 1/(1.9+1.9) = 1/3.8$ $\qquad t \in [4;5.9]$
- $\widehat{\lambda}_4 = 0/2.1 = 0$ $\qquad t \in [5.9;8]$

# Another cohort, with random right-censoring



Risk after 8 months:

- $\widehat{r}(8) = (2+?)/4 = 0.5$ or $0.75$

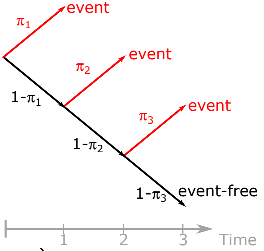
⚠ Removing censored individuals (complete case) → upward biased risk estimator

Incidence:

- $\widehat{\lambda}_1 = 1/(2+2+2+2) = 1/8$ $\quad t \in [0;2]$
- $\widehat{\lambda}_2 = 0/(2+2+2) = 0$ $\quad t \in [2;4]$
- $\widehat{\lambda}_3 = 1/(1.9+1.9) = 1/3.8$ $\quad t \in [4;5.9]$
- $\widehat{\lambda}_4 = 0/2.1 = 0$ $\quad t \in [5.9;8]$

# Binary probability models

Assuming piecewise constant hazard:



Survival (probability of not getting the event)

$$S(3) = \mathbb{P}\left[T^* > 3\right] = \mathbb{P}\left[T^* > 1\right]\mathbb{P}\left[T^* > 2 | T^* > 1\right]\mathbb{P}\left[T^* > 3 | T^* > 2\right]$$

$$=$$

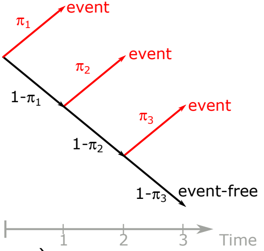
Risk (probability of getting the event)

$$r(3) = \mathbb{P}\left[T^* \leq 3\right] =$$

$$=$$

# Binary probability models

Assuming piecewise constant hazard:



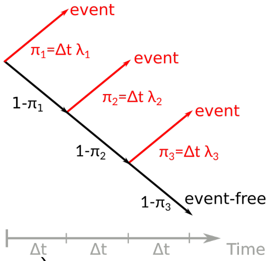
Survival (probability of not getting the event)

$$S(3) = \mathbb{P}\left[T^* > 3\right] = \mathbb{P}\left[T^* > 1\right]\mathbb{P}\left[T^* > 2 | T^* > 1\right]\mathbb{P}\left[T^* > 3 | T^* > 2\right]$$
$$= (1-\pi_1)(1-\pi_2)(1-\pi_3)$$

Risk (probability of getting the event)

$$r(3) = \mathbb{P}\left[T^* \leq 3\right] = 1 - S(3) = 1 - (1-\pi_1)(1-\pi_2)(1-\pi_3)$$
$$=$$

# Binary probability models



Assuming piecewise constant hazard:

- $\pi_t = \Delta t \lambda_t$: disease frequency equals rate times duration in each time interval

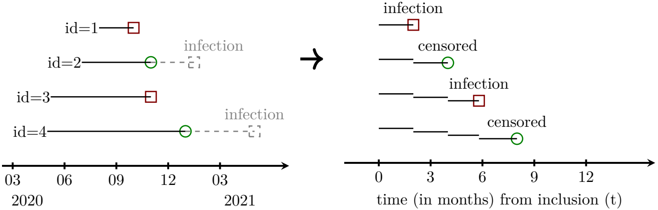
Survival (probability of not getting the event)

$$S(3) = \mathbb{P}\left[T^* > 3\right] = \mathbb{P}\left[T^* > 1\right]\mathbb{P}\left[T^* > 2 \middle| T^* > 1\right]\mathbb{P}\left[T^* > 3 \middle| T^* > 2\right]$$
$$= (1-\pi_1)(1-\pi_2)(1-\pi_3)$$

Risk (probability of getting the event)

$$r(3) = \mathbb{P}\left[T^* \leq 3\right] = 1 - S(3) = 1 - (1-\pi_1)(1-\pi_2)(1-\pi_3)$$
$$= 1 - (1-\Delta t\lambda_1)(1-\Delta t\lambda_2)(1-\Delta t\lambda_3)$$

# Cohort data: example 1 bis



Risk after 8 months:

- $\widehat{r}(8) = (2+?)/4 = 0.5$ or $0.75$
- $\widehat{r}(8) = 1 - (1 - \widehat{\lambda}_1 \Delta t_1)(1 - \widehat{\lambda}_2 \Delta t_2)(1 - \widehat{\lambda}_3 \Delta t_3)(1 - \widehat{\lambda}_4 \Delta t_4)$
  $= 1 - (1 - 1/8 * 2) * 1 * (1 - 1/3.8 * 1.9) * 1 = 0.625$

Incidence:

- $\widehat{\lambda}_1 = 1/8$ $t \in [0; 2]$
- $\widehat{\lambda}_2 = 0$ $t \in [2; 4]$
- $\widehat{\lambda}_3 = 1/3.8$ $t \in [4; 5.9]$
- $\widehat{\lambda}_4 = 0$ $t \in [5.9; 8]$

# From rate to risk

We just saw that the survival could be express as the product of 1 minus the rate:

$$S(t) = (1 - \lambda_1 \Delta t) \times (1 - \lambda_2 \Delta t) \times \dots$$

For $x \approx 0$, $\exp(x) \approx 1 + x$. So for short time intervals:

$$\begin{aligned} S(t) &\approx \exp(-\lambda_1 \Delta t) \exp(-\lambda_2 \Delta t) \dots \\ &\approx \exp(-\lambda_1 \Delta t - \lambda_2 \Delta t - \dots) \\ &\approx \exp(-\int_0^{t_1} \lambda_1 ds - \int_{t_1}^{t_2} \lambda_2 ds - \dots) \\ &\approx \exp(-\int_0^{t} \lambda(s) ds) \end{aligned}$$

(here assuming constant hazard rate within each interval)

# Application to example 2

Risk of infection/death within 771 days after start of COVID:

- via the number of events:

```
sum(covidDK$cases)/covidDK$population[1] # infection
```

```
  infection       death
0.494792420 0.001129957
```

- via the risk rate relationship

```
1-prod(1-covidDK$cases/covidDK$atRisk*1) # infection
```

```
  infection       death
0.494792420 0.001129957
```

- via an approximate risk rate relationship

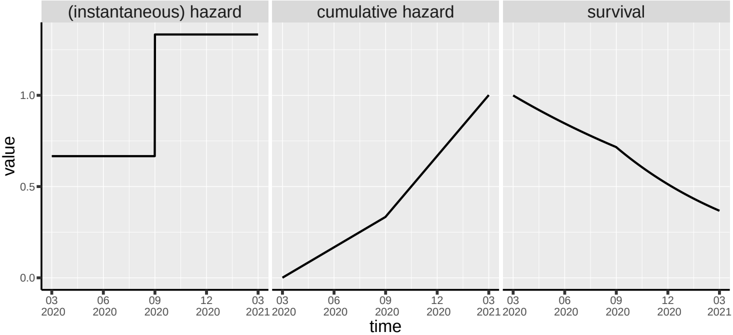
```
1-exp(-sum(covidDK$cases/covidDK$atRisk*1)) # infection
```

```
  infection       death
0.488263990 0.001129944
```

# Hazard, cumulative hazard, and survival

Special case: constant incidence rate

- $S(t) = \exp\left(-\int_0^\tau \lambda(t)dt\right) = \exp\left(-\lambda\tau\right)$
- $\Lambda(\tau) = \int_0^\tau \lambda(t)dt = \lambda\tau$ is called the cumulative hazard



# Summary

- **Prevalence**: proportion of people with a disease at time t

$$\hat{\pi} = \frac{\text{”number of people with the disease"}}{\text{"number of people"}} \in [0, 1]$$

- **Incidence rate**: frequency of disease occurrence over period $\tau$
⚠ unit: time$^{-1}$, e.g. person-year

$$\widehat{\lambda}(\tau) = \frac{\text{”number of new cases"}}{\text{"cumulative at-risk time"}} \in [0, +\infty[$$

- **Risk**: probability of experiencing the disease before time $\tau$

$$\widehat{r}(\tau) = \frac{\text{”number of new cases"}}{\text{"number of person at risk"}} \approx 1 - \exp\left(-\int_0^{\tau} \widehat{\lambda}(t)dt\right)$$

# Measures of association

# Example 2 at a specific timepoint

| Country \ Infection | No | Yes |
|---|---|---|
| Denmark (DEN) | $a = 2960606$ | $b = 2889610$ |
| Spain (SPA) | $c = 34224428$ | $d = 13231166$ |

Risk comparison: $\hat{r}_{DEN} = \frac{b}{a+b} = 49.48\%$ vs. $\hat{r}_{SPA} = \frac{d}{c+d} = 27.91\%$

# Example 2 at a specific timepoint

| Infection<br>Country | No | Yes |
|---|---|---|
| Denmark (DEN) | $a = 2960606$ | $b = 2889610$ |
| Spain (SPA) | $c = 34224428$ | $d = 13231166$ |

Risk comparison: $\hat{r}_{DEN} = \frac{b}{a+b} = 49.48\%$ vs. $\hat{r}_{SPA} = \frac{d}{c+d} = 27.91\%$

- **risk difference**: $RD(\tau) = r_{SPA}(\tau) - r_{DEN}(\tau) = -21.56\%$

## Example 2 at a specific timepoint

| Infection / Country | No | Yes |
|---|---|---|
| Denmark (DEN) | $a = 2960606$ | $b = 2889610$ |
| Spain (SPA) | $c = 34224428$ | $d = 13231166$ |

Risk comparison: $\hat{r}_{DEN} = \frac{b}{a+b} = 49.48\%$ vs. $\hat{r}_{SPA} = \frac{d}{c+d} = 27.91\%$

- **risk difference**: $RD(\tau) = r_{SPA}(\tau) - r_{DEN}(\tau) = -21.56\%$
- **relative risk**: $RR(\tau) = \frac{r_{SPA}(\tau)}{r_{DEN}(\tau)} = 0.5642$

# Example 2 at a specific timepoint

| Country \ Infection | No | Yes |
|---|---|---|
| Denmark (DEN) | $a = 2960606$ | $b = 2889610$ |
| Spain (SPA) | $c = 34224428$ | $d = 13231166$ |

Risk comparison: $\hat{r}_{DEN} = \frac{b}{a+b} = 49.48\%$ vs. $\hat{r}_{SPA} = \frac{d}{c+d} = 27.91\%$

- **risk difference**: $RD(\tau) = r_{SPA}(\tau) - r_{DEN}(\tau) = -21.56\%$
- **relative risk**: $RR(\tau) = \frac{r_{SPA}(\tau)}{r_{DEN}(\tau)} = 0.5642$
- **odds ratio**: $OR(\tau) = \left(\frac{r_{SPA}(\tau)}{1-r_{SPA}(\tau)}\right) \Big/ \left(\frac{r_{DEN}(\tau)}{1-r_{DEN}(\tau)}\right) = 0.3954$

# The 3 measures of associations

$RD(\tau) = -21.56\%$ $RR(\tau) = 0.5642$ $OR(\tau) = 0.3954$

Interpretation: the 771 days risk of being tested COVID positive

- **risk difference**: is about 0.2 lower in Spain vs. Denmark
- **relative risk**: is about half in Spain compared vs. Denmark
- **odds ratio**: ?

- **identical** risks: $RD$ $RR$ $OR$
- **higher risk** in SPA: $RD$ $RR$ $OR$
- **lower risk** in SPA: $RD$ $RR$ $OR$

# The 3 measures of associations

$RD(\tau) = -21.56\%$ $RR(\tau) = 0.5642$ $OR(\tau) = 0.3954$

Interpretation: the 771 days risk of being tested COVID positive

- **risk difference**: is about 0.2 lower in Spain vs. Denmark
- **relative risk**: is about half in Spain compared vs. Denmark
- **odds ratio**: ?

- **identical** risks: $RD = 0$ $RR = 1$ $OR = 1$
- **higher risk** in SPA: $RD > 0$ $RR > 1$ $OR > 1$
- **lower risk** in SPA: $RD < 0$ $RR < 1$ $OR < 1$

# Odds ratio

**odds**: $\Omega(\tau) = \frac{\text{"risk of an event"}}{\text{"risk of no event"}} = \frac{r(\tau)}{1-r(\tau)}$

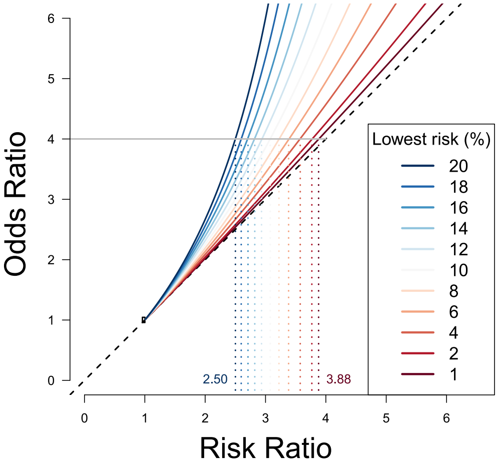
```
risk 0 0.01 0.10 0.25 0.3333333 0.5 0.75  0.99   1
odds  0 0.01 0.11 0.33 0.5000000 1.0 3.00 99.00 Inf
```

- $\Omega \in [0, \infty[$
- if risks are small $\Omega(\tau) \approx r(\tau)$ ("rare disease assumption")

**odds ratio**: $OR(\tau) = \left(\frac{r_{SPA}(\tau)}{1-r_{SPA}(\tau)}\right) \Big/ \left(\frac{r_{DEN}(\tau)}{1-r_{DEN}(\tau)}\right) = \frac{\Omega_{SPA}(\tau)}{\Omega_{DEN}(\tau)}$

- $RR(\tau) = \frac{OR(\tau)}{1-r_{SPA}+r_{SPA}OR(\tau)}$
- if risks are small $OR(\tau) \approx RR(\tau)$ ("rare disease assumption")
- needed for case-control studies / logistic regression

# Odds ratio vs. risk ratio



(graph courtesy of Paul Blanche)

# Test of association: chi-square test

| Infection<br>Country | No | Yes |
|---|---|---|
| Denmark (DEN) | $a = 2960606$ | $b = 2889610$ |
| Spain (SPA) | $c = 34224428$ | $d = 13231166$ |

Testing the independence between the outcome and the group variable is based on

$$t_{\chi^2} = (a+b+c+d)\frac{(ad-bc)}{(a+b)(c+d)(a+c)(b+d)}$$

which under independence follows* a $\chi^2_1$.

* chi-square distribution with 1 degree of freedom

# Interpretation

Consider the following result:

- $t_{\chi^2} = 4732$ and p-value $< 0.0001$

What can you conclude?

# Interpretation

Consider the following result:

- $t_{\chi^2} = 4732$ and p-value $< 0.0001$

What can you conclude?

*Personal opinion:* I don't like this test as it lacks an (intuitive) parameter of interest!

- better report risk difference or risk ratio with associated confidence intervals
  In R : function `binomMeld.test` of the exact2x2 package.

# Back to the case study (Beyersmann et al., 2014)

Risk of death in ICU:

- Pneumonia: $48/220 \approx 21.8\%$
- No pneumonia: $166/1656 \approx 10.0\%$

Incidence rate of death in ICU:

- Pneumonia: $48/6161 \approx 7.79$ death per 1000 patient-days
- No pneumonia: $166/22337 \approx 7.43$ death per 1000 patient-days

Apparent contradition?

# Uncertainty - risk

Exact:

```
binom.test(x = 48, n = 220)
```

```

        Exact binomial test

data:  48 and 220
number of successes = 48, number of trials = 220, p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1654696 0.2786567
sample estimates:
probability of success
             0.2181818
```

Approximate:

```
48/220 + c(-1.96,1.96) * sqrt(48/220*(1-48/220)/220)
```

```
[1] 0.1636052 0.2727585
```

# Uncertainty - risk difference

Nearly exact:

```
library(exact2x2)
binomMeld.test(x1 = 48, n1 = 220, x2 = 166, n2 = 1656)
```

```
        melded binomial test for difference

data:  sample 1:(48/220), sample 2:(166/1656)
proportion 1 = 0.21818, proportion 2 = 0.10024, p-value = 3.104e-06
alternative hypothesis: true difference is not equal to 0
95 percent confidence interval:
 -0.1801787 -0.0625695
sample estimates:
difference (p2-p1)
        -0.1179403
```

Approximate:

```
166/1656-48/220 + c(-1.96,1.96) * sqrt(48/220*(1-48/220)/220+166/1656*
    (1-166/1656)/1656)
```

```
[1] -0.1744012 -0.0614793
```

# Uncertainty - incidence rate difference

Approximate:

```
df <- data.frame(event = c(48,166), fup = c(6161,22337),
                 exposure = c(1,0))
e.glm <- glm(event ~ exposure + offset(log(fup)),
             family = poisson, data = df)
exp(cbind(coef(e.glm),confint(e.glm)))
```

```
Waiting for profiling to be done...
                              2.5 %      97.5 %
(Intercept) 0.007431616 0.006357651 0.008620128
exposure    1.048351171 0.752569456 1.432854368
```

Manually:

```
166/22337 * exp(c(-1.96,1.96)/sqrt(166))
```

```
[1] 0.006382870 0.008652677
```

# Resolving the paradox

Discharge [†]:

- Pneumonia: $(220 - 48)/6161 \approx 27.9$ per 1000 patient-day
- No pneumonia: $(1656 - 166)/22337 \approx 66.7$ per 1000 patient-day

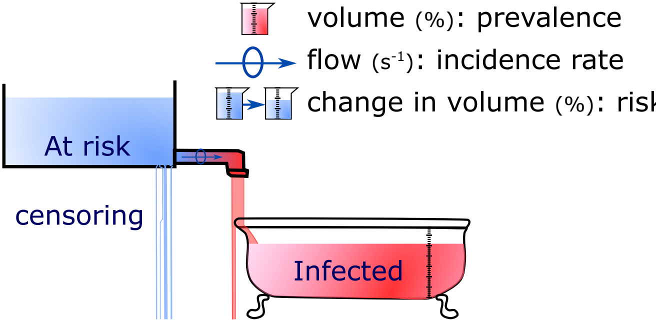
Pneumonia on admission prolongs ICU stays:

- patients with pneumonia are subject to the 'same' rate but for longer period of time
- they therefore have a larger 'risk' of death

⚠ Here risk of death during ICU stay is not very well defined as it corresponds to a time period that is patient dependent

[†] numbers slightly differ from the article due to censoring

# Conclusion



volume (%): prevalence

flow ($s^{-1}$): incidence rate

change in volume (%): risk

# What we have seen today

# What we have seen today

✓ Data representation:

- graphical representation of survival data
- 3 data formats: individual, aggregated, 2 by 2 table

✓ Measures of disease frequency:

- definition and estimation of **prevalence**, **incidence rate**, **risk**,
- unit: per **person.time** for incidence rates

✓ Handling right censoring

- risk-rate relationship
- complete case analysis (nearly) always biased!

✓ Measures of association

- **risk difference**, **relative risk**, odds ratio
- chi-squared test

A little bit about uncertainty quantification

# Reference I

Beyersmann, J., Gastmeier, P., and Schumacher, M. (2014). Incidence in icu populations: how to measure and report it? *Intensive Care Medicine*, 40:871–876.

Kestenbaum, B. (2019). *Epidemiology and Biostatistics: An Introduction to Clinical Research*.

# Interlude: high school physics

**Period** (T):

- time to complete one cycle
- unit: $s$ second

**Frequency** (f):

- the number of cycles per second
- $f = \frac{1}{T}$
- unit: $Hz = s^{-1}$ herts

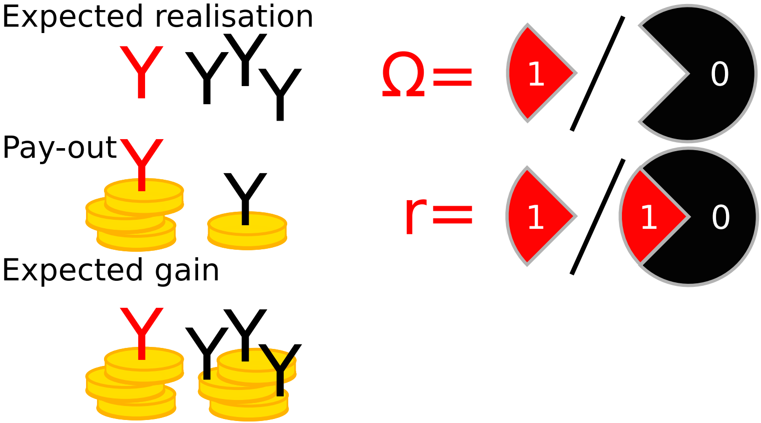
**Example**: Heart rate at 60 vs. 120 beats per minute

- $T = 1s$ vs $0.5s$
- $f = 1Hz$ vs $2Hz$

# Risk - hazard relationship

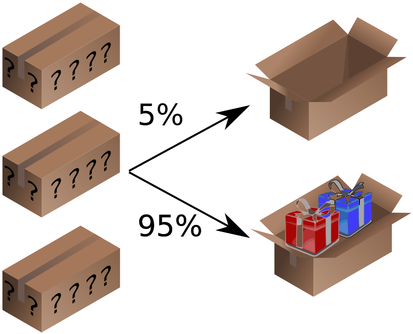
$$
\begin{aligned}
\lambda(t) &= \lim_{dt \to 0} \frac{\mathbb{P}\left[t < T \leq t + dt \middle| T > t\right]}{dt} \\
&= \lim_{dt \to 0} \frac{\frac{\mathbb{P}[t < T \leq t + dt]}{dt}}{\mathbb{P}\left[T > t\right]} = \lim_{dt \to 0} \frac{\frac{\mathbb{P}[T \leq t + dt] - \mathbb{P}[T \leq t]}{dt}}{\mathbb{P}\left[T > t\right]} \\
&= \lim_{dt \to 0} \frac{\frac{(1 - S(t + dt)) - (1 - S(t))}{dt}}{S(t)} = \frac{-\frac{\partial S(t)}{\partial t}}{S(t)} \\
\lambda(t) &= -\frac{\partial \log S(t)}{\partial t} \\
\Lambda(\tau) &= \int_0^\tau \lambda(t) dt = -\log S(\tau) \\
S(\tau) &= \exp(-\Lambda(\tau)) \\
r(\tau) &= 1 - \exp(-\Lambda(\tau))
\end{aligned}
$$

## Gambling at 1:3



# Interpretation of the CI - analogy

A machine generates boxes with 95% probability to contain a gift.



- 95% of the boxes I receive contain gifts.
- a specific box contains or not gifts

## Interpretation of the CI

Similar except that we are "blind"

- no able to precisely check the content of the box

✓ the calculation of the CI ensures that 95% of the time, it contains the (true) value.

$CI = [0.021; 0.336]$

✓ the (true) death rate may or may not be between 0.021 and 0.336

✓ the data at hand is concordant with a (true) death rate between 0.021 and 0.336