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.

12 January 2022

Epidemiology (very short!)

Description of disease frequency:

- outcome: generally binary or time to event (Y, T)
- measure: **prevalence**, odds, **incidence rate**, **risk**.

Find causes/remedies to the disease (E):

- compare exposed and non-exposed with respect to the measure.
- interpretation and consequences

In any case, target a meaningful parameter of interest

- not just something 'easy' to estimate from your data

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),
  dynamic treatment regimes (switch of treatment), ...

Working with finite samples:

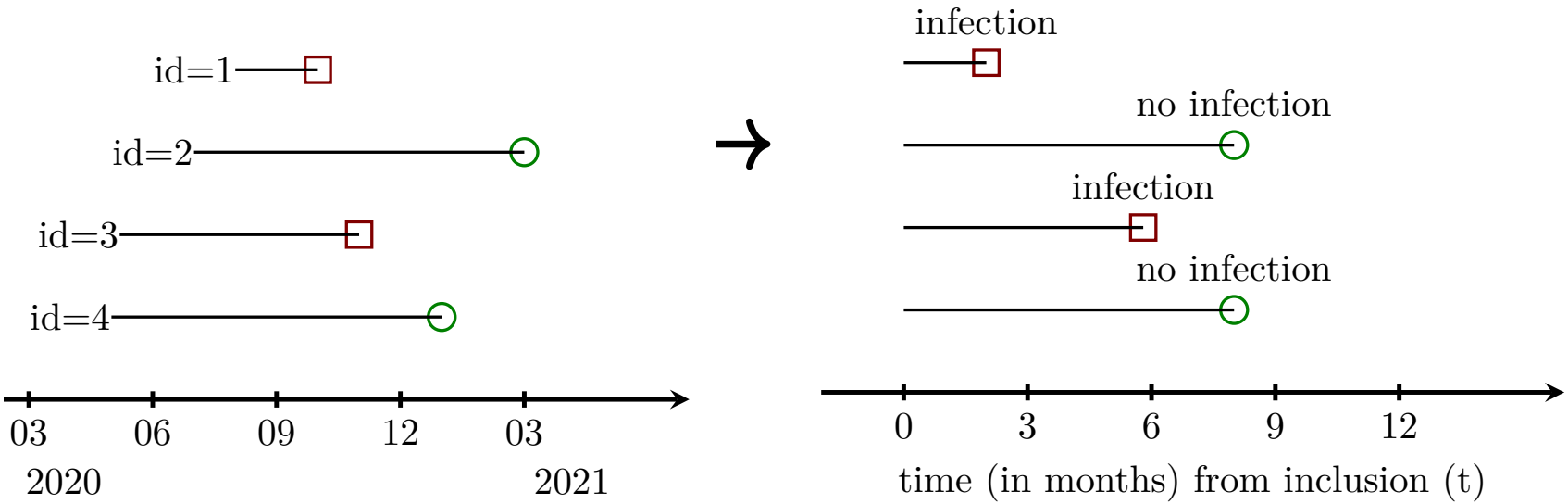
- quantitying uncertainty

Prediction:

- guess what would happen for **a** new patient?

Cohort study - example 1

A group of n persons is followed over time



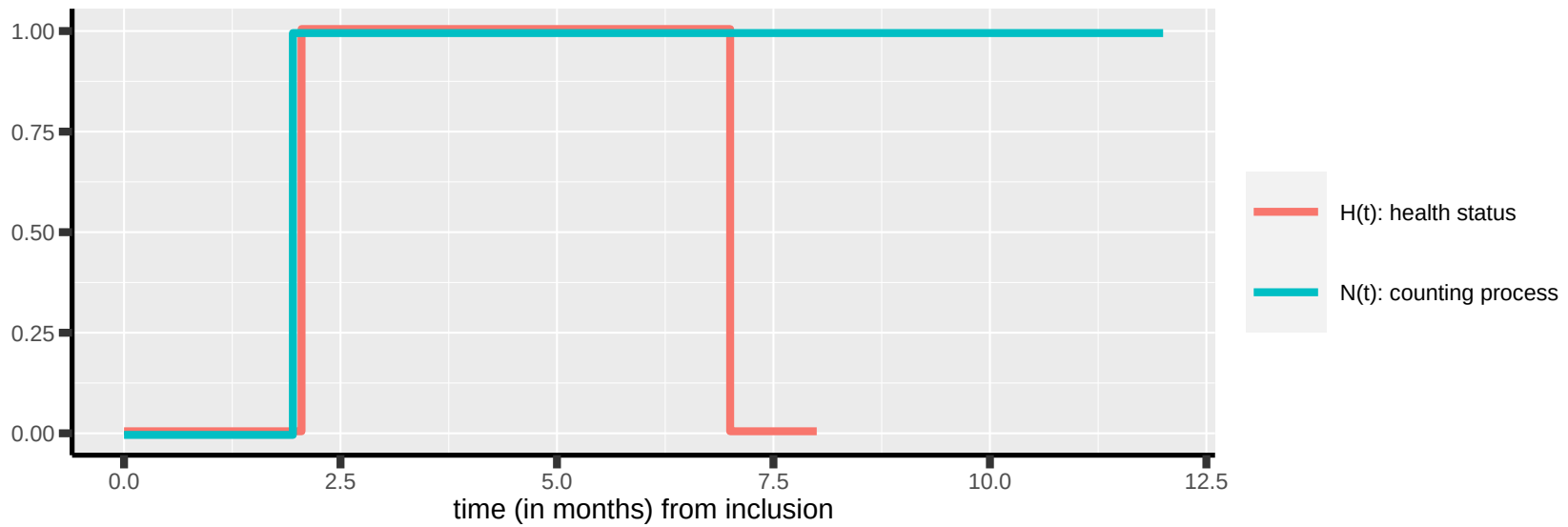
- $T_i \in [0, +\infty[$ time to event for subject i
  (in months, or years, or . . .)
- $N_i(t) \in \{0, 1\}$ event occurence by time t for subject i
  (e.g. death, death due to COVID, first COVID infection, . . .)

Note: counting process vs. health status

$N_i(t)$ is also refered to as a counting process

- indicates whether an event has occured
- not whether the patient is still affected by the event, $H_i(t)$

Illustration when the infection lasts 5 months:



Individual vs. aggregated 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	D	n-D	Y
0.0	4	0	0	4	0.0
2.0	4	1	1	3	8.0
5.9	3	1	2	2	19.7
8.0	2	0	2	2	23.9

Individual vs. aggregated 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	D	n-D	Y
0.0	4	0	0	4	0.0
2.0	4	1	1	3	8.0
5.9	3	1	2	2	19.7
8.0	2	0	2	2	23.9

- $D(t) = \sum_{i=1}^{n} N_i(t)$ events, $n - D(t)$ event-free.
- $Y(t) = \sum_{i=1}^{n} T_i \wedge t$ total follow-up time.

Example 2 (COVID)

From https://github.com/kjhealy/covdata:

- "weekly national-level ECDC data on COVID-19"

```
          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
```

Measures of frequency

Prevalence

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

$$\pi = \mathbb{P}[H = 1] \qquad \text{or} \qquad \pi(t) = \mathbb{P}[H(t) = 1]$$

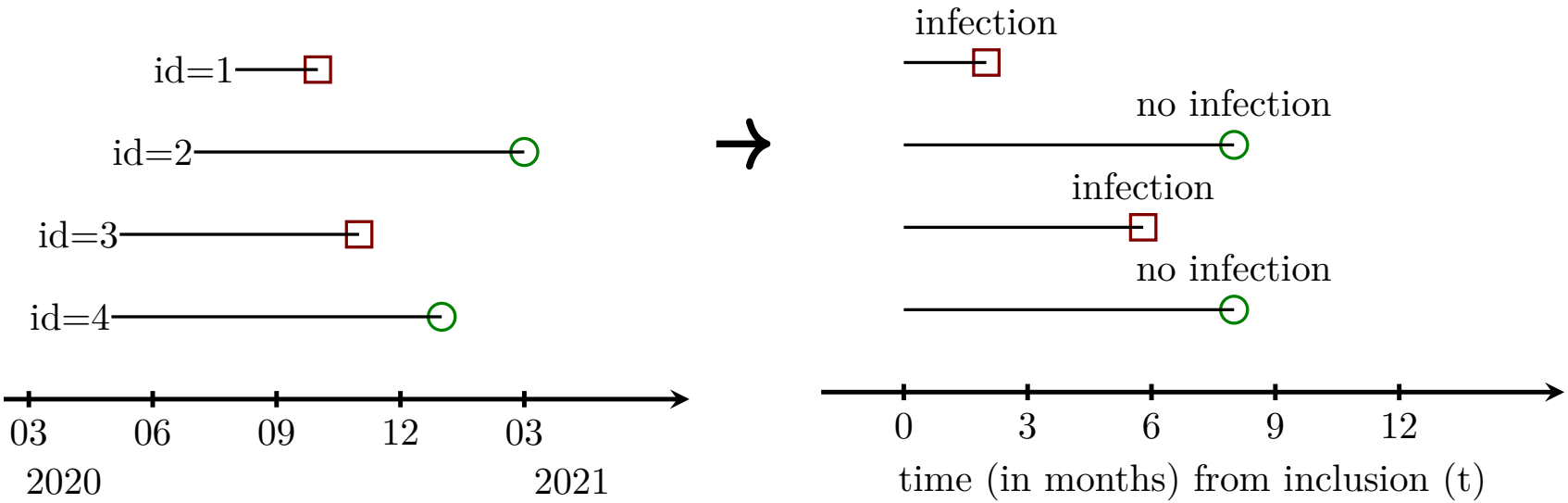
- $\pi \in [0, 1]$, $\pi = \begin{cases} 0 \text{ nobody has the disease} \\ 1 \text{ everybody has the disease} \end{cases}$

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

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

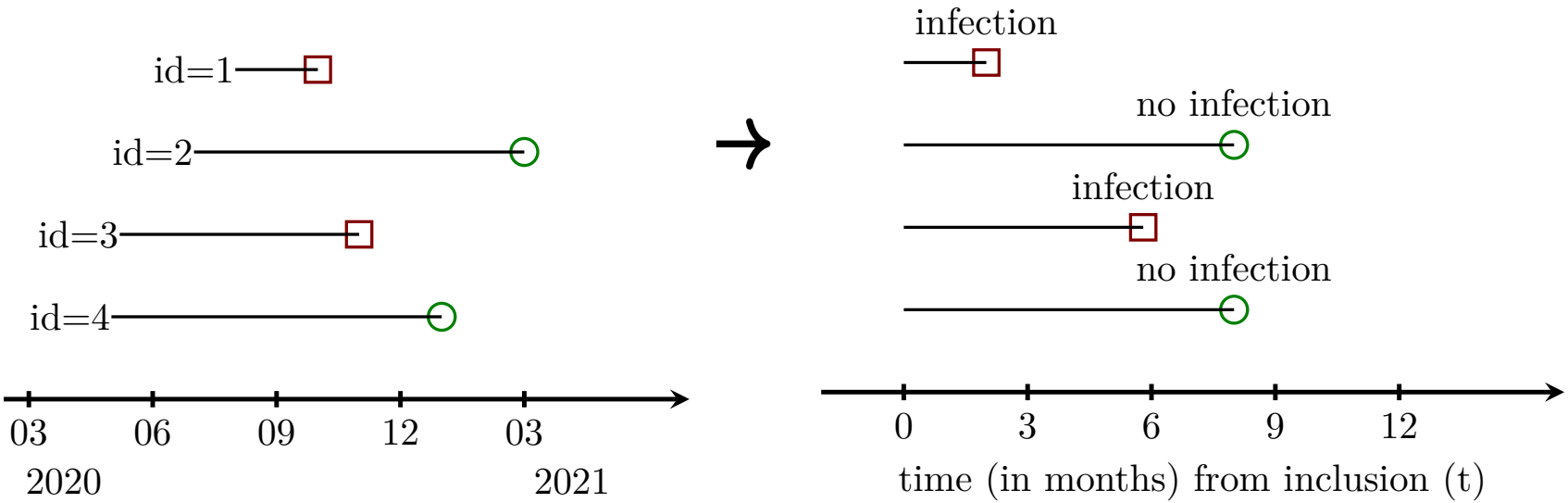
where $\overline{\bullet}$ denote the empirical average of $\bullet$.

Prevalence - example 1



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

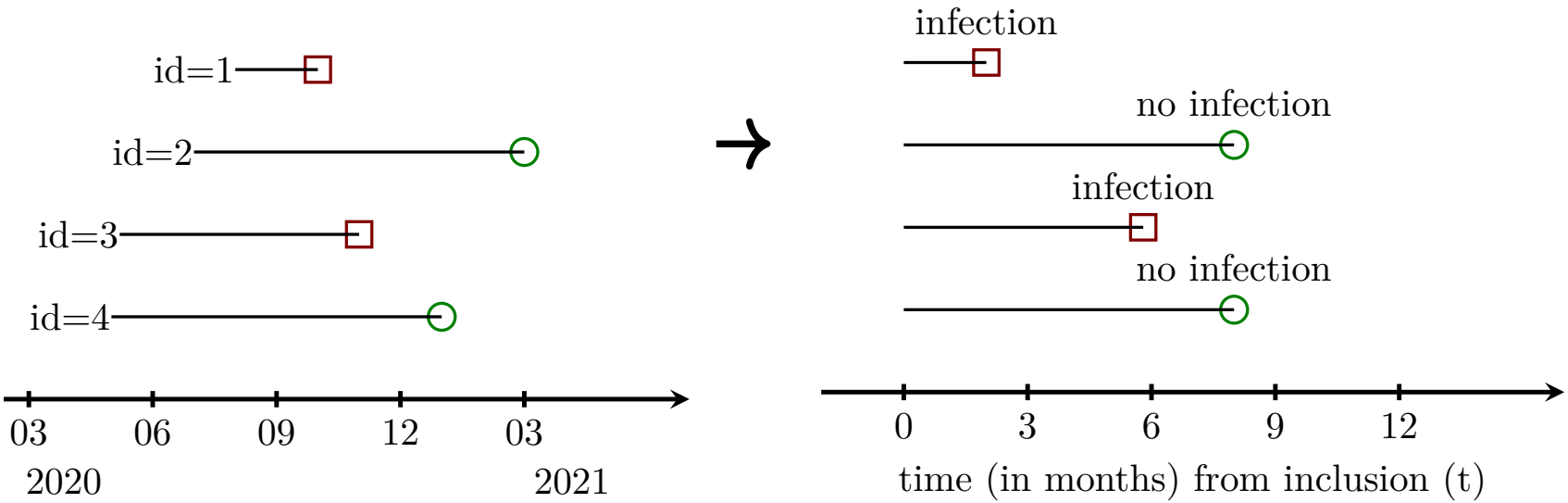
Prevalence - example 1



Assumes that the infection lasts 5 months for everybody and 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 5 months for everybody and 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 3 [1]: 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:

- ?
- ?
- ?

[1] example 2.2 from Kestenbaum (2019)

Prevalence - limitation

Example 3 [1]: 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

[1] example 2.2 from Kestenbaum (2019)

Risk / cumulative incidence

Definition: proportion of people *becoming* sick within a period

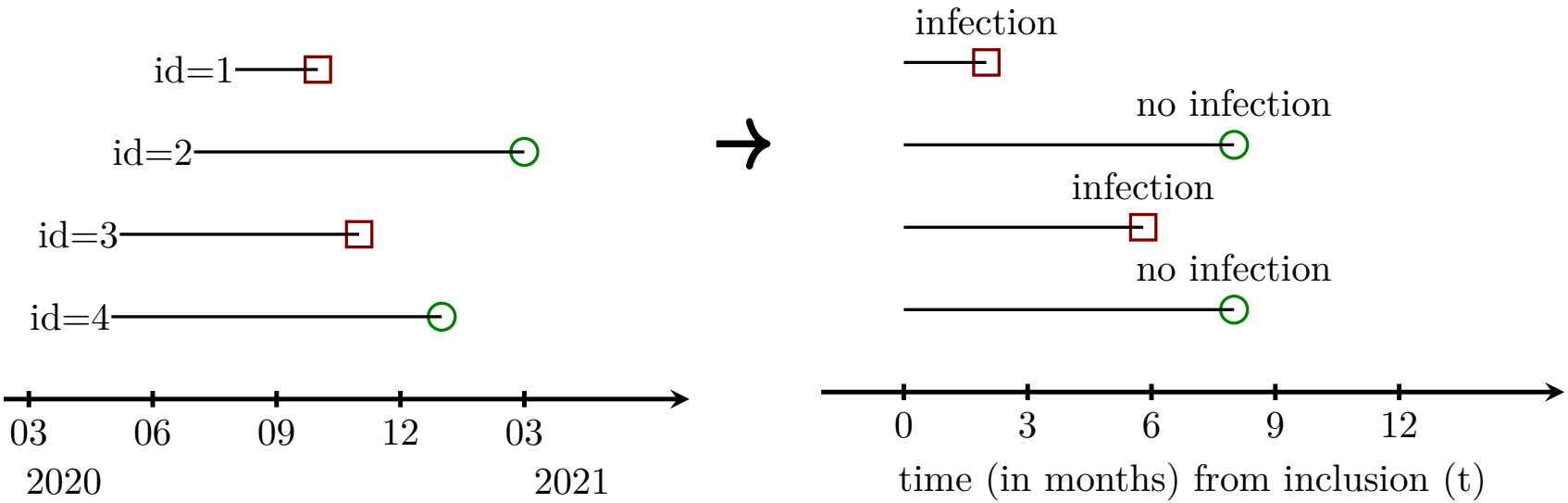
$$r(\tau) = \mathbb{P}\left[T \leq \tau, N(\tau) = 1 | T > 0\right]$$

- $r(0) = 0$
- $r \in [0, 1]$, $r = \begin{cases} 0 \text{ nobody will get the disease} \\ 1 \text{ everybody will get the disease} \end{cases}$
- $r(\tau)$ is non-decreasing with τ

Estimation: $\frac{\text{"number of new cases"}}{\text{"number of persons at risk"}}$

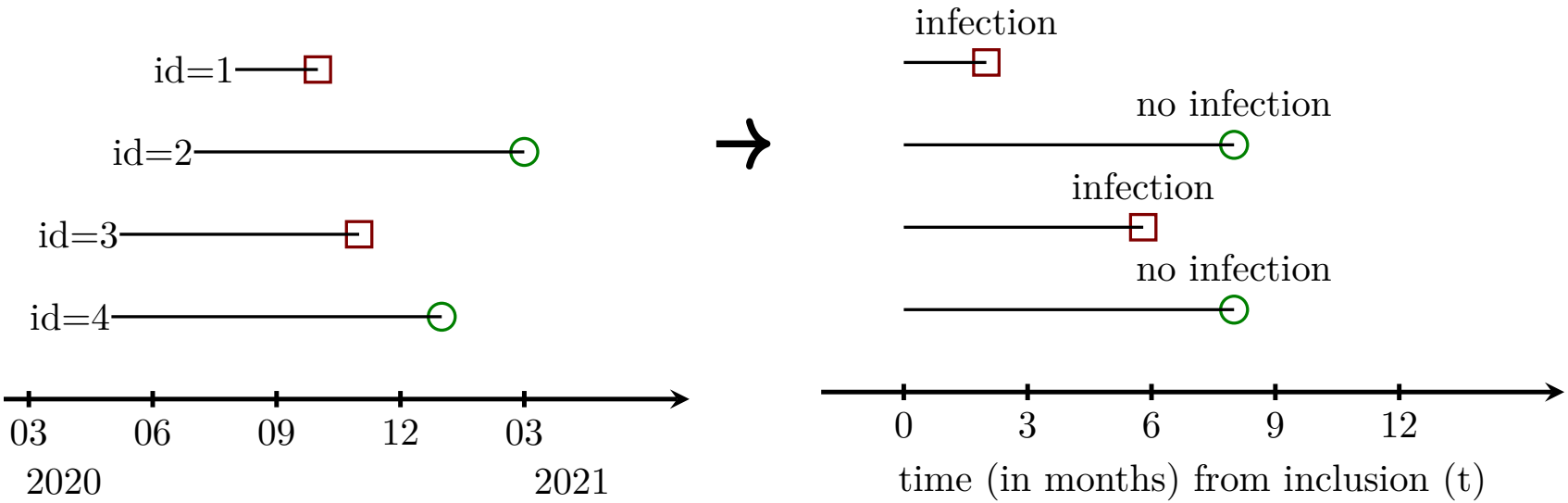
$$\hat{r}(\tau) = \frac{D(\tau)}{n} = \frac{1}{n}\sum_{i=1}^{n} N_i(\tau) = \overline{N} \text{ when } N_i \text{ is binary } 0/1$$

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

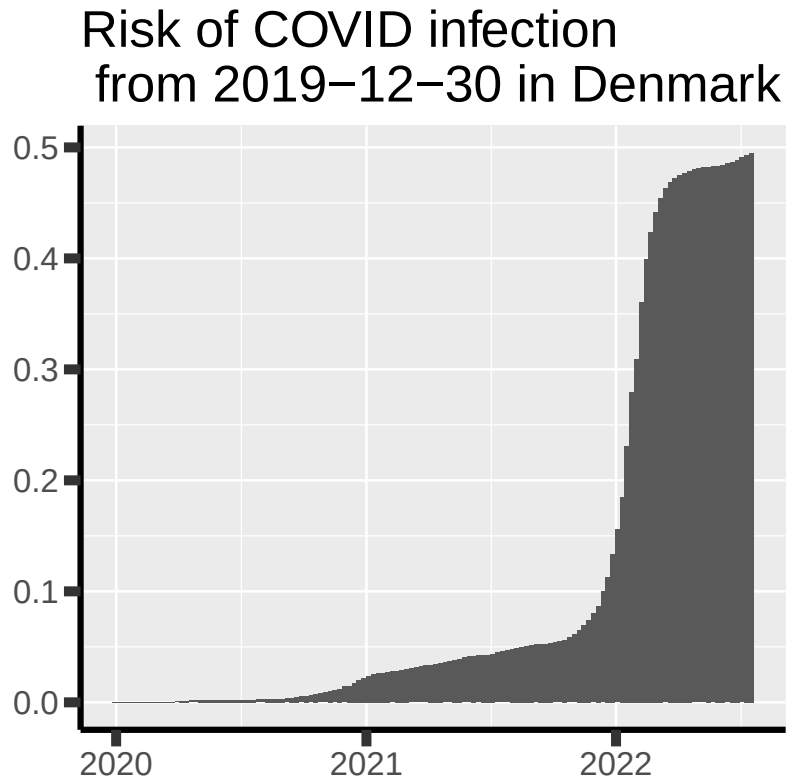
Risk - example 2

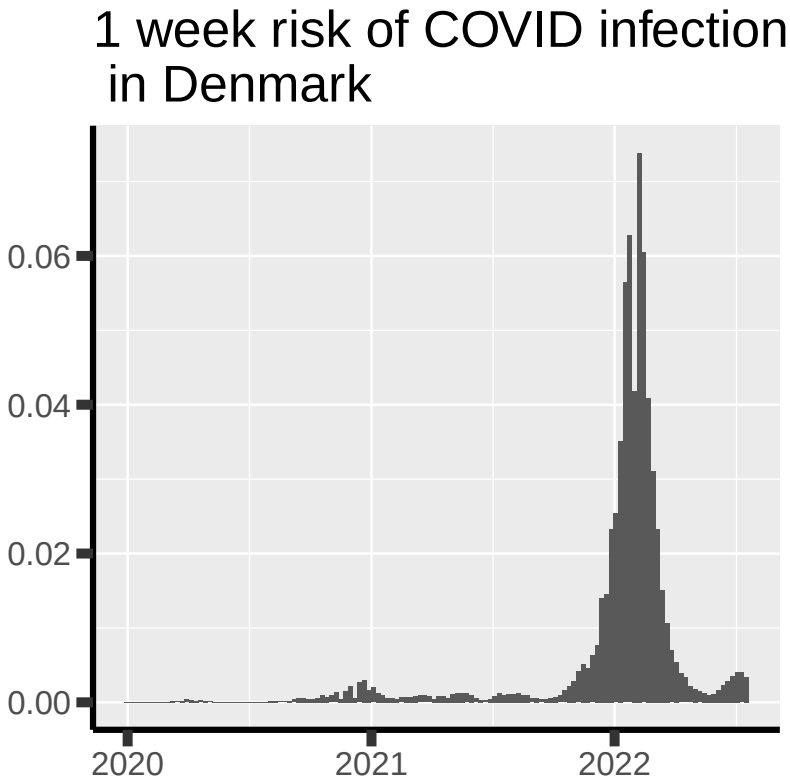
- `population`: population size at the start of COVID
- `atRisk`: (approximate) number of COVID naive people
- `cases` number COVID cases detected during the week
- `cu_cases` cumulative number of COVID cases

```
          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: risk of the event divided by exposure time

$$\lambda(0) = \frac{\mathbb{P}\left[T \leq \tau, N(\tau) = 1 | T > 0\right]}{\tau}$$

⚠ unit: time^{-1}

- $\lambda(t) \in [0, +\infty[$ higher values $\rightarrow$ higher disease frequency
- implicitely assume a constant disease frequency over the exposure time

Incidence rate

Definition: risk of the event divided by exposure time

$$\lambda(0) = \frac{\mathbb{P}\left[T \leq \tau, N(\tau) = 1 | T > 0\right]}{\tau}$$

$$\lambda(t) = \frac{\mathbb{P}\left[T \leq t + \tau, N(\tau) = 1 | T > t\right]}{\tau}$$

⚠ unit: time^{-1}

- $\lambda(t) \in [0, +\infty[$ higher values $\rightarrow$ higher disease frequency
- implicitely assume a constant disease frequency over the exposure time

Incidence rate

Definition: risk of the event divided by exposure time

$$\lambda(0) = \frac{\mathbb{P}\left[T \leq \tau, N(\tau) = 1 \middle| T > 0\right]}{\tau}$$

$$\lambda(t) = \frac{\mathbb{P}\left[T \leq t + \tau, N(\tau) = 1 \middle| T > t\right]}{\tau}$$

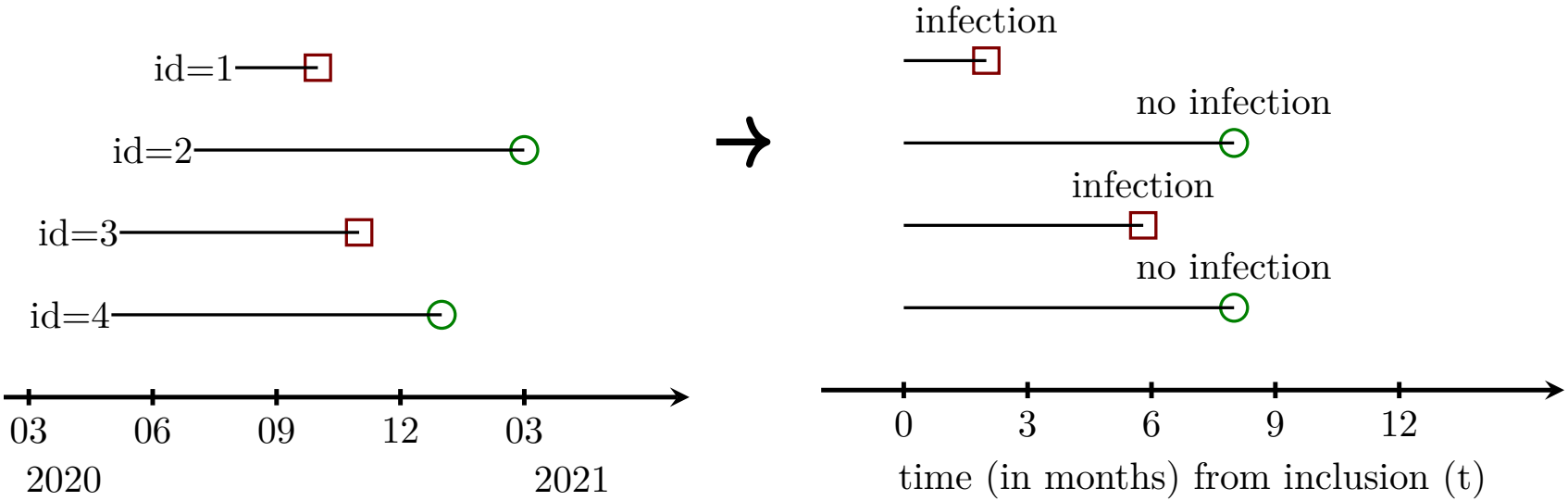
⚠ unit: time^{-1}

- $\lambda(t) \in [0, +\infty[$ higher values $\rightarrow$ higher disease frequency
- implicitely assume a constant disease frequency over the exposure time

Estimation: $\frac{\text{"number of new cases"}}{\text{"number of person-time at risk"}}$

$$\widehat{\lambda}(\tau) = \frac{D(\tau)}{Y(\tau)} = \frac{\sum_{i=1}^{n} N_i(\tau)}{\sum_{i=1}^{n} T_i \wedge \tau}$$

Incidence rate - example



- $\widetilde{T}_1 = 2$ months, $\widetilde{Y}_1 = 1$
- $\widetilde{T}_2 = 8$ months, $\widetilde{Y}_2 = 0$
- $\widetilde{T}_3 = 5.9$ months, $\widetilde{Y}_3 = 1$
- $\widetilde{T}_4 = 8$ months, $\widetilde{Y}_4 = 0$

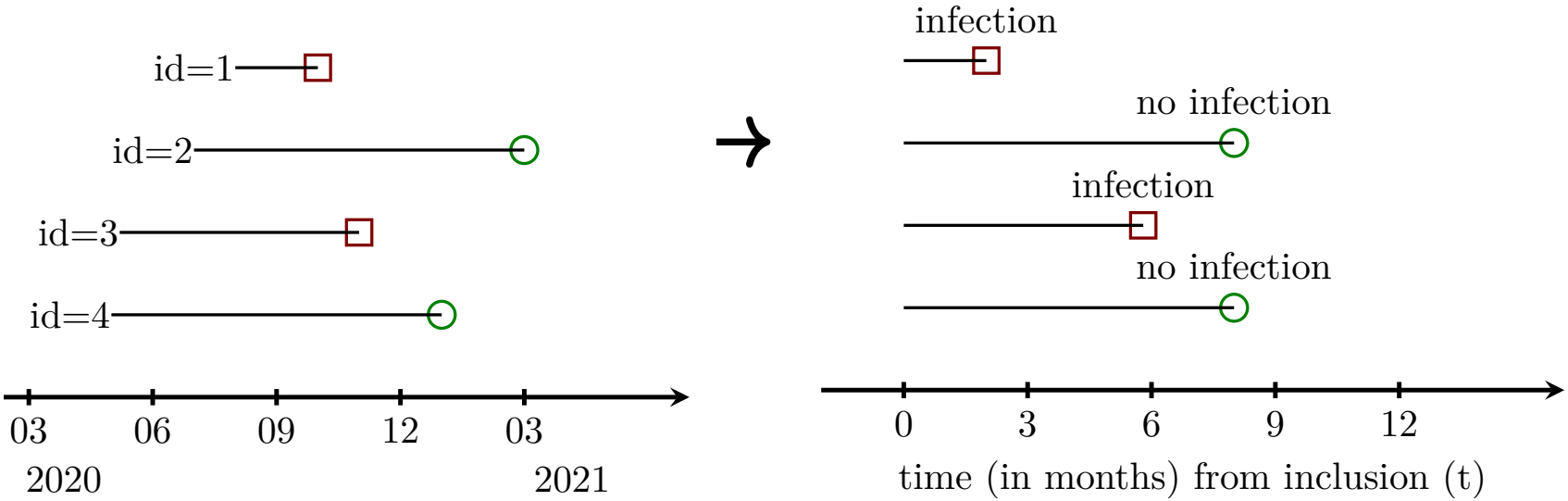
$\widehat{\lambda}_\tau =$

$\approx$ per person-month

$\approx$ per 1000 person-month

$\approx$ per person-year

Incidence rate - example



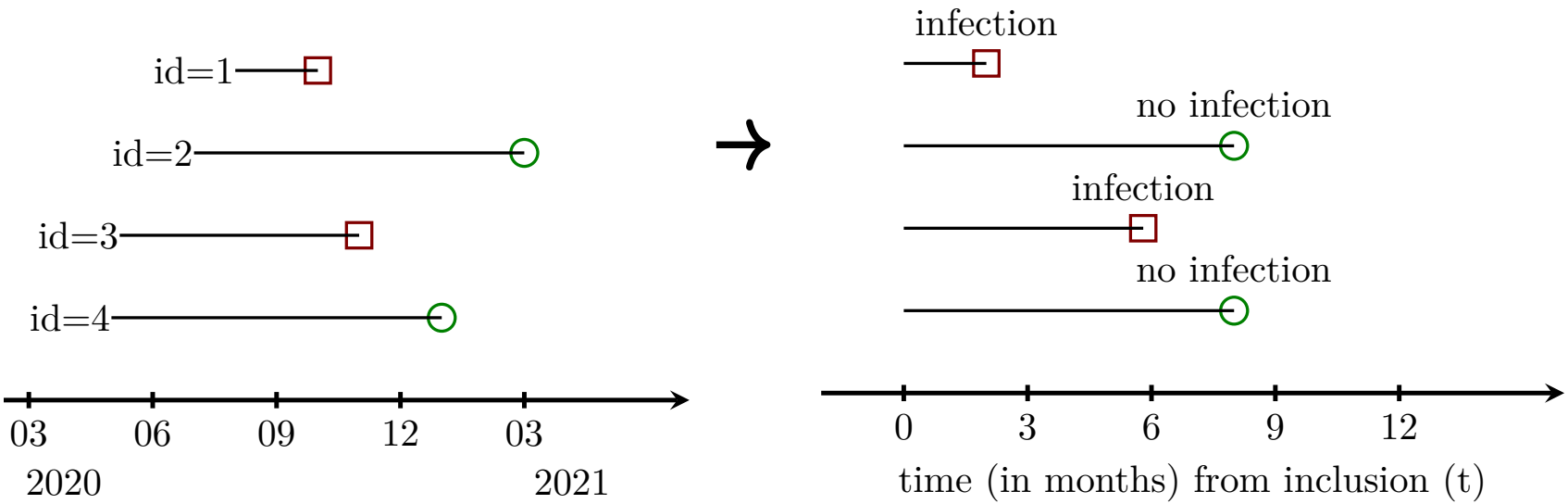
- $\widetilde{T}_1 = 2$ months, $\widetilde{Y}_1 = 1$
- $\widetilde{T}_2 = 8$ months, $\widetilde{Y}_2 = 0$
- $\widetilde{T}_3 = 5.9$ months, $\widetilde{Y}_3 = 1$
- $\widetilde{T}_4 = 8$ months, $\widetilde{Y}_4 = 0$

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

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

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

Incidence rate - example



- $\widetilde{T}_1 = 2$ months, $\widetilde{Y}_1 = 1$
- $\widetilde{T}_2 = 8$ months, $\widetilde{Y}_2 = 0$
- $\widetilde{T}_3 = 5.9$ months, $\widetilde{Y}_3 = 1$
- $\widetilde{T}_4 = 8$ months, $\widetilde{Y}_4 = 0$

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

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

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

Incidence rate - 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 $100 \times (1 - \text{IRR})$, 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.

Incidence rate - example 2

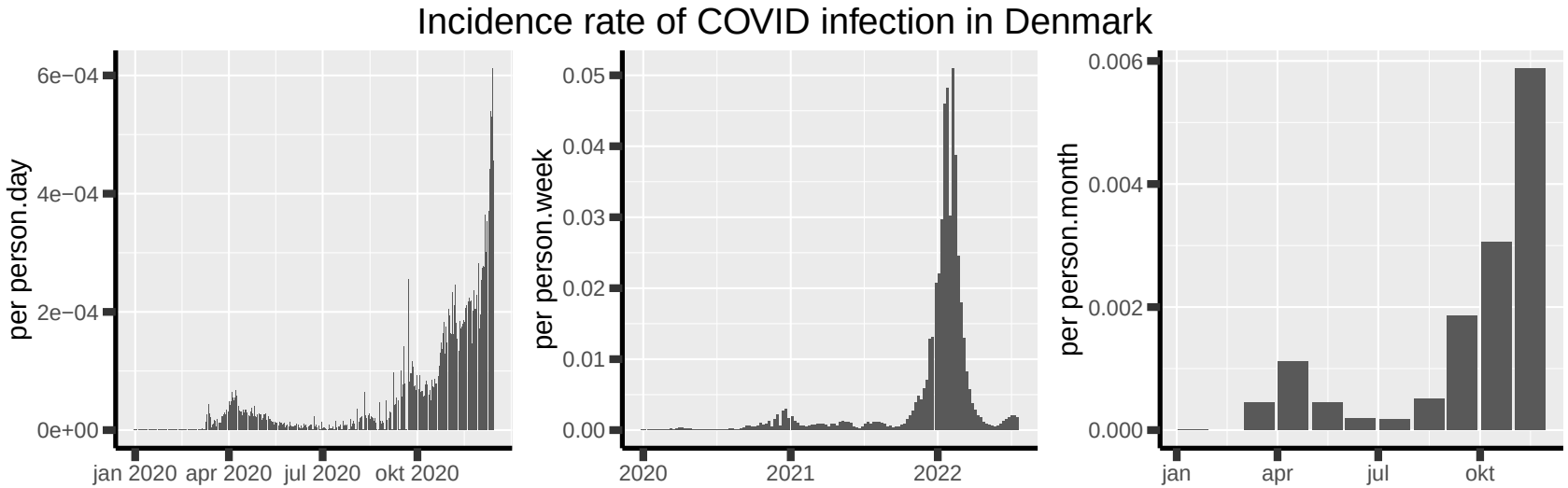
3 datasets:

- daily number of cases (up to end of 2020)
- weekly number of cases (up to end of 2022)
- monthly number of cases based on the daily number

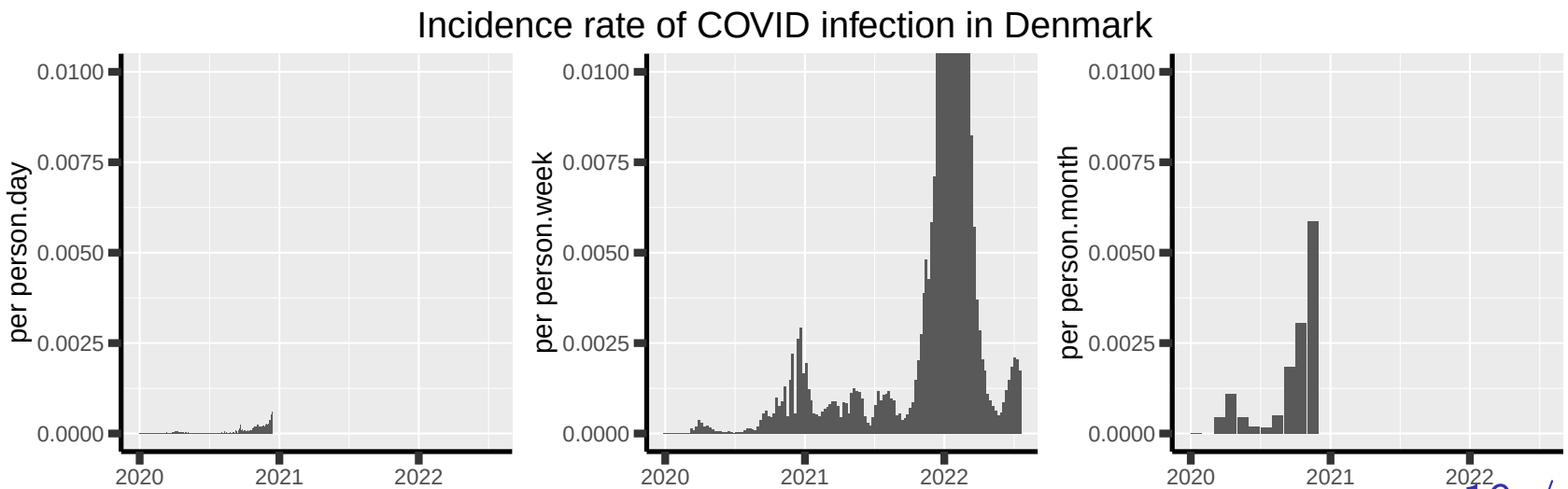
At risk time: unknown

- rough approximation: population size minus cumulative number of cases

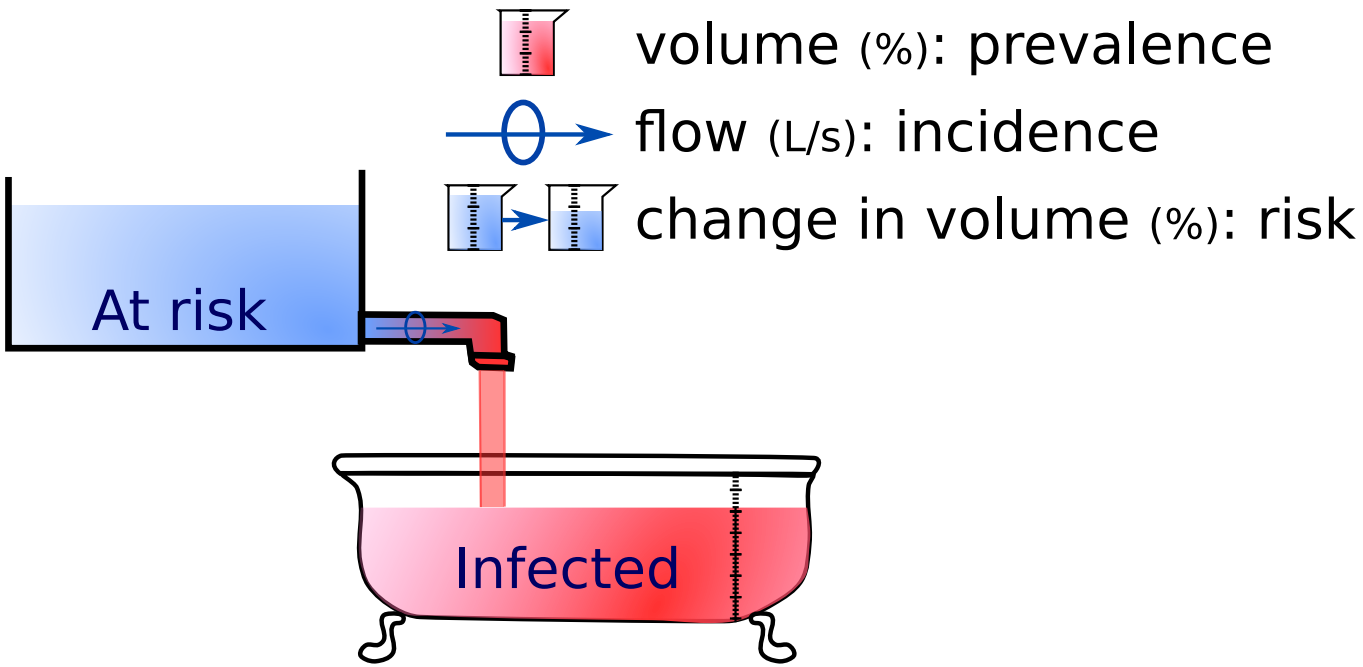
Example 2 - illustration



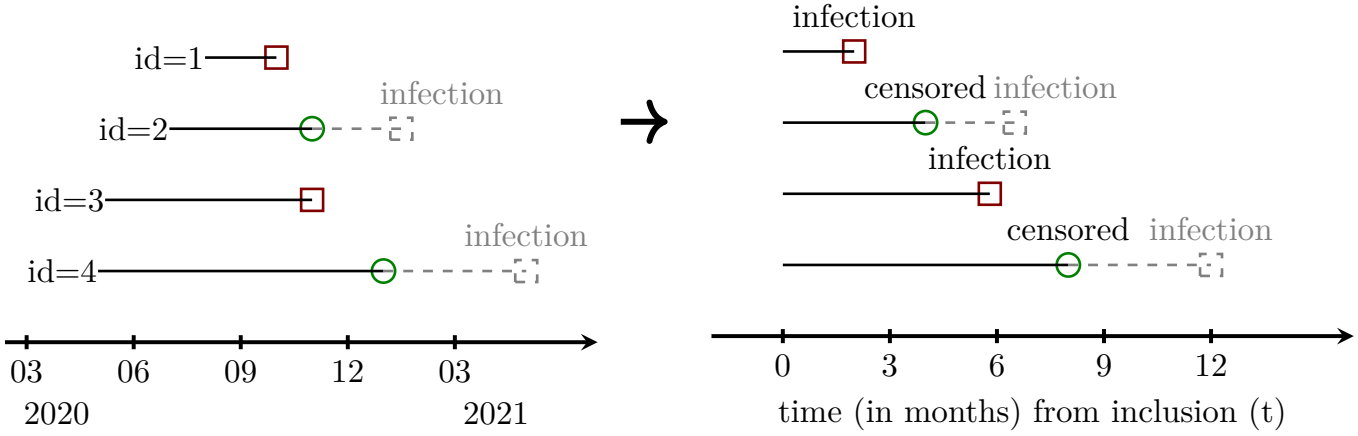
Same but with the same x- and y-scale



Risk-rate relationship



Cohort data: example 1 bis



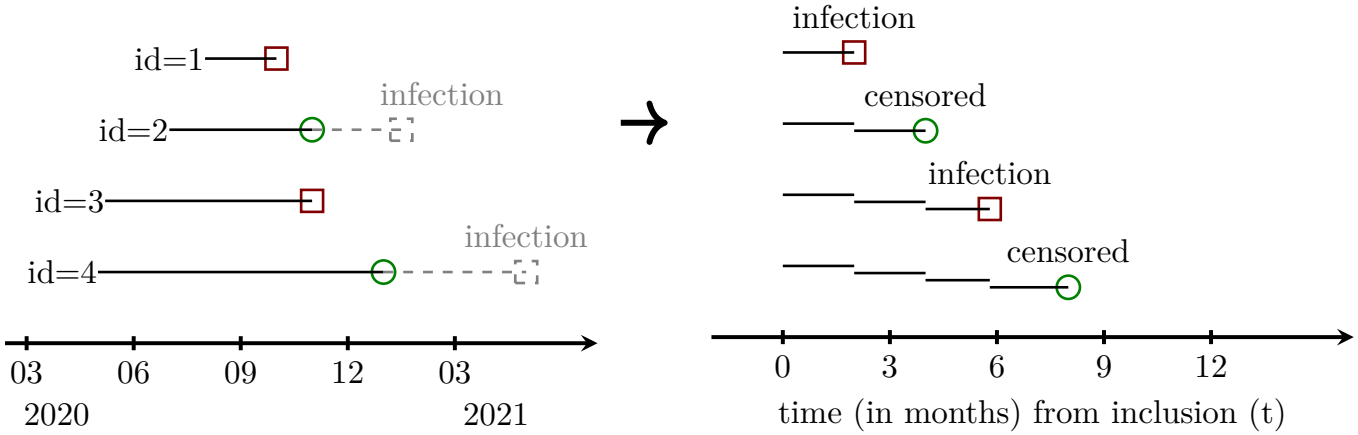
Risk after 8 months:

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

Incidence:

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

Cohort data: example 1 bis



Risk after 8 months:

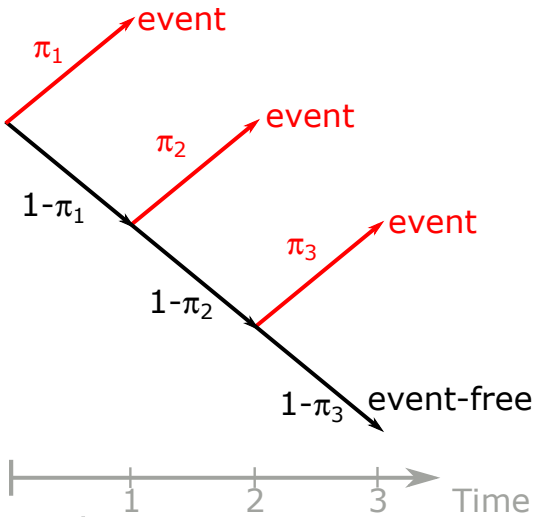
- $\widehat{r}(8) = (2+?)/4 = 0.5$ or 0.75

Incidence:

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

Binary probability models

Assuming piecewise constant hazard:



Survival (probability of not getting the event)

$S(3) =$

$=$

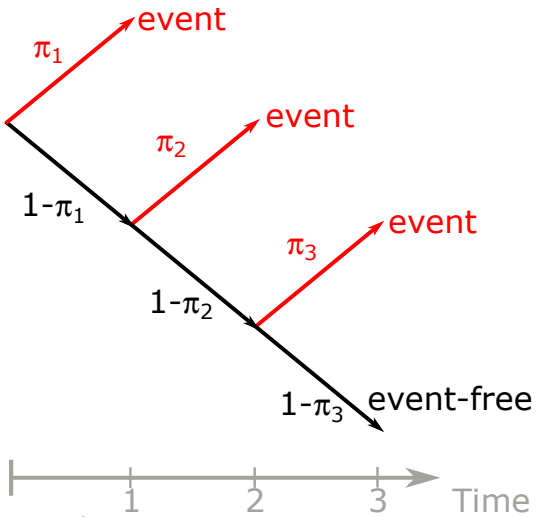
Risk (probability of getting the event)

$r(3) =$

$=$

Binary probability models

Assuming piecewise constant hazard:



Survival (probability of not getting the event)

$$S(3) = \mathbb{P}[N(1) = 0]\ \mathbb{P}[N(2) = 0|N(1) = 0]\ \mathbb{P}[N(3) = 0|N(2) = 0]$$
$$= (1 - \pi_1)(1 - \pi_2)(1 - \pi_3)$$

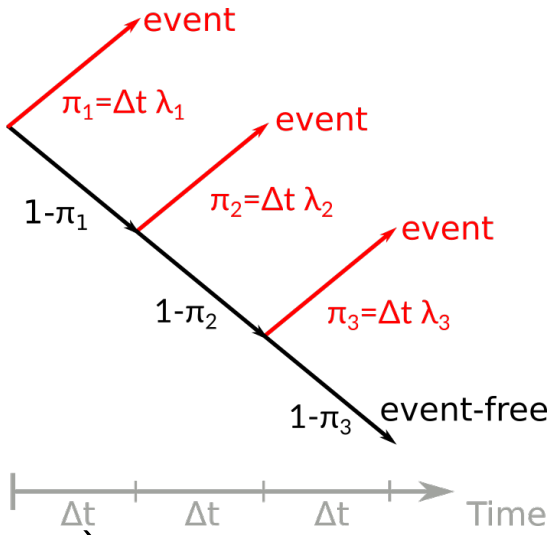
Risk (probability of getting the event)

$$r(3) = 1 - S(3) = 1 - (1 - \pi_1)(1 - \pi_2)(1 - \pi_3)$$
$$=$$

Binary probability models

Assuming piecewise constant hazard:

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



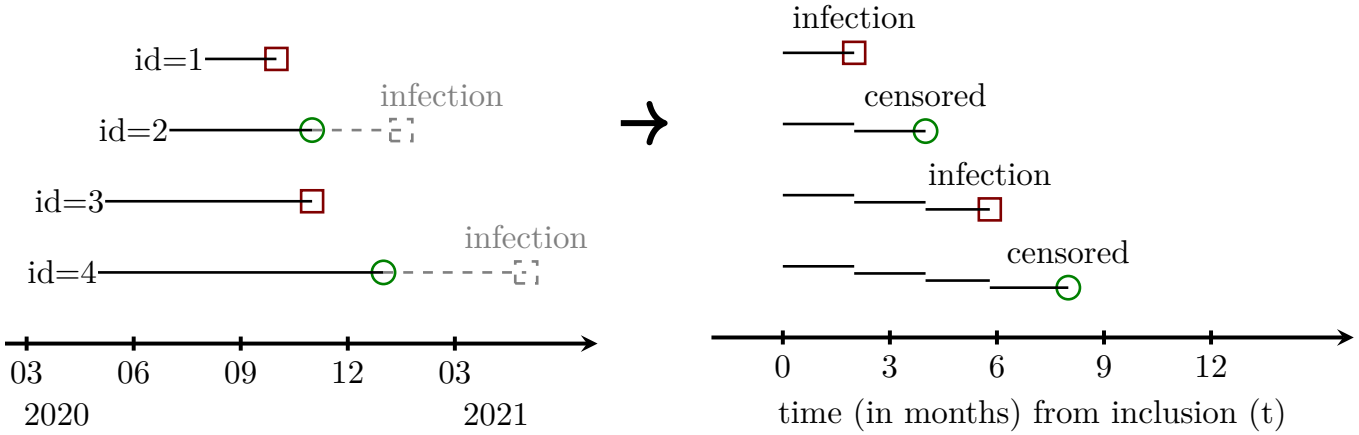
Survival (probability of not getting the event)

$$S(3) = \mathbb{P}[N(1) = 0]\ \mathbb{P}[N(2) = 0|N(1) = 0]\ \mathbb{P}[N(3) = 0|N(2) = 0]$$
$$= (1 - \pi_1)(1 - \pi_2)(1 - \pi_3)$$

Risk (probability of getting the event)

$$r(3) = 1 - S(3) = 1 - (1 - \pi_1)(1 - \pi_2)(1 - \pi_3)$$
$$= 1 - (1 - \Delta t\lambda_1)(1 - \Delta t\lambda_2)(1 - \Delta t\lambda_3)$$

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:

- $\widehat{\lambda}_1 = 1/8$ $\quad t \in [0; 2]$
- $\widehat{\lambda}_2 = 0$ $\quad t \in [2; 4]$
- $\widehat{\lambda}_3 = 1/7.8$ $\quad t \in [4; 5.9]$
- $\widehat{\lambda}_4 = 0$ $\quad t \in [5.9; 8]$

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

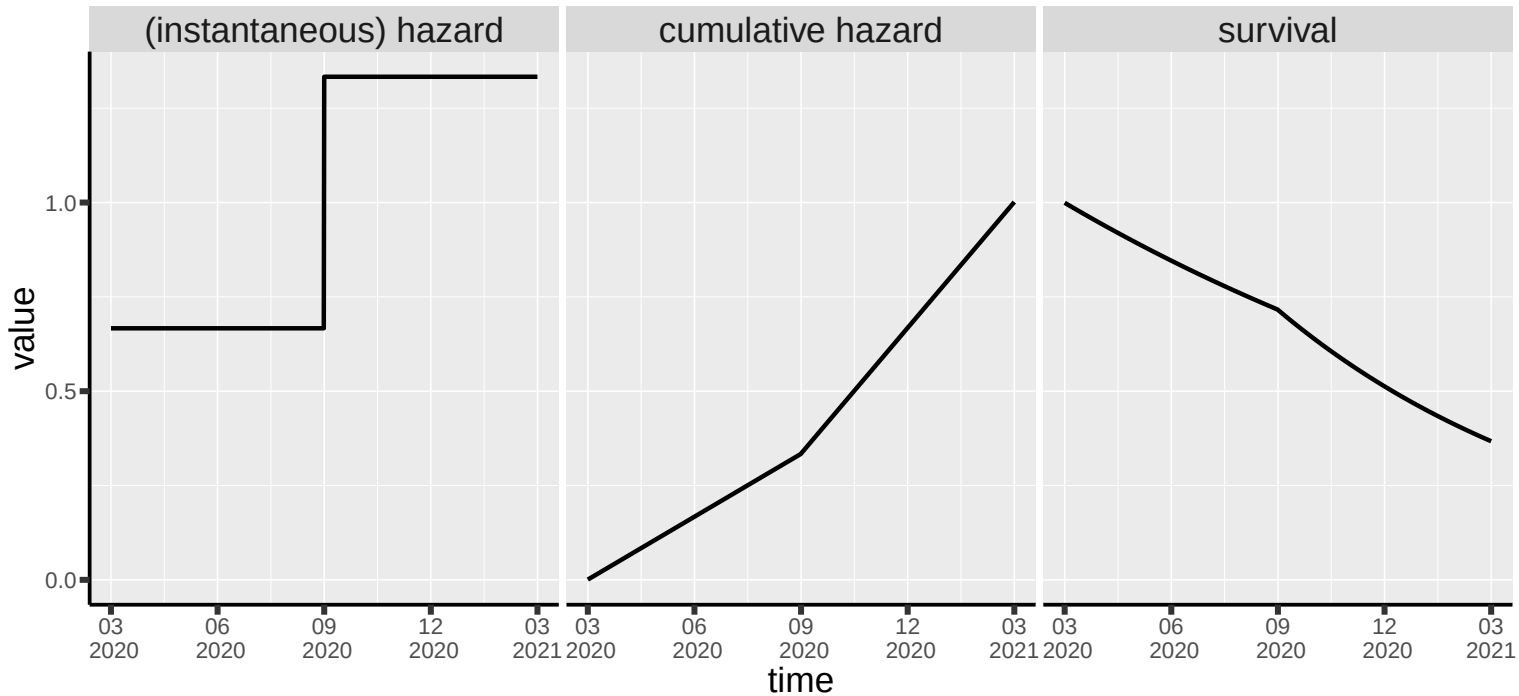
```
1-exp(-sum(covidDK$cases/covidDK$atRisk*1)) # infection
```

```
  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 τ
  ⚠ unit: time^{-1}, e.g. person-year

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

- **Risk**: probability of experiencing the disease before time τ

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

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

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

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**: $RR(\tau) = \frac{r_{SPA}(\tau)}{r_{DEN}(\tau)} = 0.5642$

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

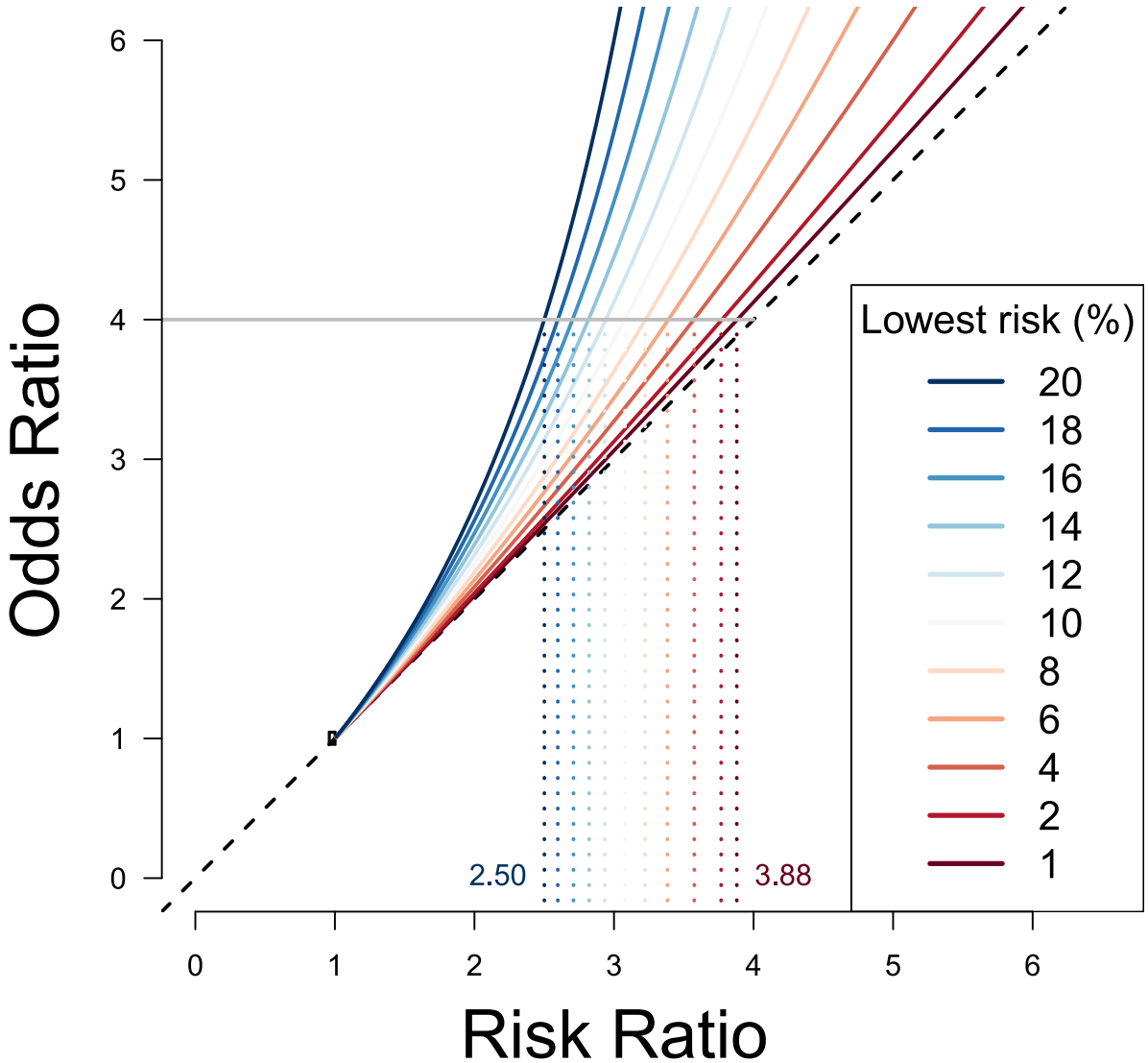
```
risk 0 0.01 0.10 0.25 0.3333333 0.5 0.75  0.99   1
odds  0 0.01 0.11 0.33 0.5000000 1.0 3.00 99.00 Inf
```

- $\Omega \in [0, \infty[$
- 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

Country \ Infection	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[2] a χ^2_1.

[2] chi-square distribution with 1 degree of freedom

Personal opinion

I don't like so much this test.

Consider the following result:

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

What can you conclude?

Personal opinion

I don't like so much this test.

Consider the following result:

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

What can you conclude?

We lack a parameter of interest!

- better use RR or RD with associated confidence intervals

Quantifying uncertainty

Quiz 1 - p-value

Consider comparing two drugs regarding the occurence of a disease.

A low p-value (e.g. below 0.05)

- provides evidence again the null hypothesis, i.e. one drug is better than the other
- cannot tell

A high p-value (e.g. above 0.05)

- provides evidence for the null hypothesis, i.e. the drugs are equivalent
- cannot tell

If two studies report different p-values (e.g. 0.01 vs 0.1)

- the studies disagree
- cannot tell

Quiz 1 - p-value (solution)

A low p-value (e.g. below 0.05)

✔ provides evidence again the null hypothesis, i.e. one drug is better than the other

✘ cannot tell

A high p-value (e.g. above 0.05)

✘ provides evidence for the null hypothesis, i.e. the drugs are equivalent

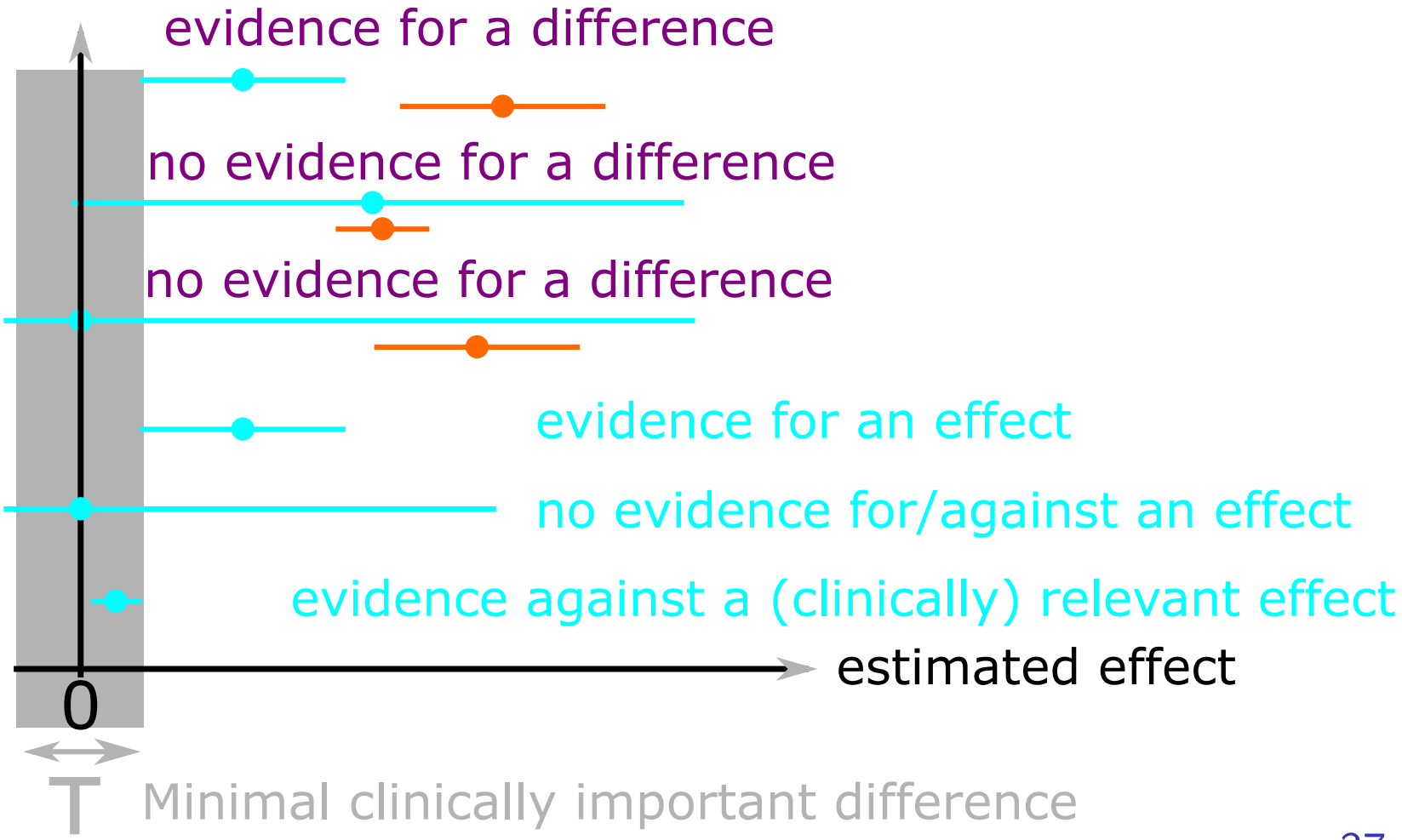
✔ cannot tell, one should look at the CIs

If two studies report different p-values (e.g. 0.01 vs 0.1)

✘ the studies disagree

✔ cannot tell, one should look at the CIs

Comparing confidence intervals



Quiz 2 - 95% confidence interval

For large enough n, the confidence interval $[0.021; 0.336]$:

- contains the true incidence rate with probability 95%.
- contains 95% of the sample data.
- contains incidence rates values compatible with the data

For large enough n, in 95% of the replication studies:

- the (new) $\text{CI}_{\widehat{\lambda}_\tau, 95\%}$ will contain the true incidence rate.
- the (new) estimate will be in the current $\text{CI}_{\widehat{\lambda}_\tau, 95\%}$.

When performing multiple comparisons:

- one should only adjust p-values
- one should adjust both p-values and confidence intervals

Quiz 2 - 95% confidence interval

For large enough n, the confidence interval $[0.021; 0.336]$:

- ✘ contains the true incidence rate with probability 95%.
- ✘ contains 95% of the sample data.
- ✔ contains incidence rates not statistically different with $\widehat{\lambda}_\tau$.

For large enough n, in 95% of the replication studies:

- ✔ the (new) $\text{CI}_{\widehat{\lambda}_\tau, 95\%}$ will contain the true incidence rate.
- ✘ the (new) estimate will be in the current $\text{CI}_{\widehat{\lambda}_\tau, 95\%}$.

When performing multiple comparisons:

- ✘ one should only adjust p-values
- ✔ one should adjust both p-values and confidence intervals

Confidence interval (based on asymptotic results)

95% confidence intervals enable to represent the uncertainty about our estimate, e.g.:

risk: $\text{CI}_{\widehat{r(\tau)},95\%} = \left[\widehat{r}(\tau) - 1.96\sqrt{\frac{r(\tau)(1-r(\tau))}{n}}\,,\, \widehat{r}(\tau) + 1.96\sqrt{\frac{r(\tau)(1-r(\tau))}{n}}\right]$

Incidence rate: $\text{CI}_{\widehat{\lambda_\tau},95\%} = \left[\widehat{\lambda}_\tau \exp\left(-\frac{1.96}{\sqrt{\widetilde{D}}}\right)\,,\, \widehat{\lambda}_\tau \exp\left(\frac{1.96}{\sqrt{\widetilde{D}}}\right)\right]$

Confidence interval (based on asymptotic results)

95% confidence intervals enable to represent the uncertainty about our estimate, e.g.:

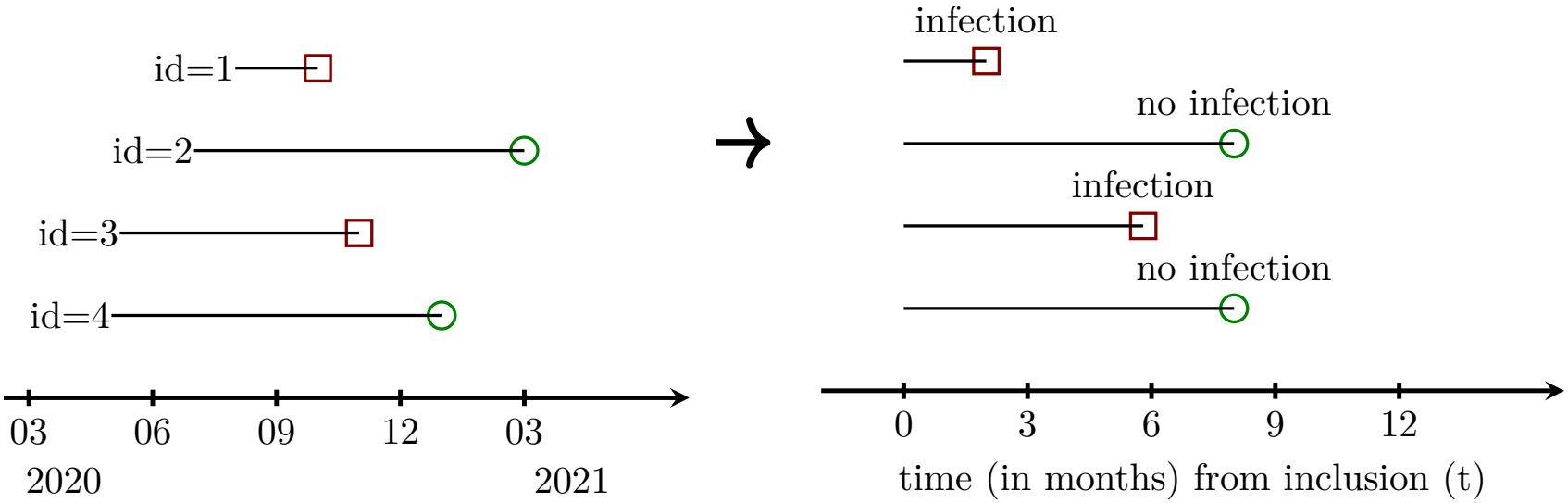
risk: $\text{CI}_{\widehat{r(\tau)},95\%} = \left[\widehat{r}(\tau) - 1.96\sqrt{\frac{r(\tau)(1-r(\tau))}{n}}\,,\, \widehat{r}(\tau) + 1.96\sqrt{\frac{r(\tau)(1-r(\tau))}{n}}\right]$

(original scale: $\text{CI}_{\widehat{\bullet},95\%} = \left[\widehat{\bullet} - 1.96\,\sigma_{\widehat{\bullet}}\,,\, \widehat{\bullet} + 1.96\,\sigma_{\widehat{\bullet}}\right]$)

Incidence rate: $\text{CI}_{\widehat{\lambda_\tau},95\%} = \left[\widehat{\lambda}_\tau \exp\left(-\frac{1.96}{\sqrt{\widetilde{D}}}\right)\,,\, \widehat{\lambda}_\tau \exp\left(\frac{1.96}{\sqrt{\widetilde{D}}}\right)\right]$

(log-scale: $\text{CI}_{\widehat{\bullet},95\%} = \left[\widehat{\bullet}\exp\left(-1.96\,\sigma_{\log\widehat{\bullet}}\right)\,,\, \widehat{\bullet}\exp\left(1.96\,\log\sigma_{\widehat{\bullet}}\right)\right]$)

Confidence interval - example



$$\widehat{\lambda}_\tau = \frac{1+0+1+0}{2+8+5.9+8} = \frac{2}{23.9} \approx 0.084$$

$$\mathsf{CI}_{\widehat{\lambda}_\tau,95\%} = \left[\widehat{\lambda}_\tau \exp\left(-\frac{1.96}{\sqrt{\widetilde{D}}}\right), \widehat{\lambda}_\tau \exp\left(\frac{1.96}{\sqrt{\widetilde{D}}}\right)\right]$$

$$= \left[0.084 \exp\left(-\frac{1.96}{\sqrt{2}}\right), 0.084 \exp\left(\frac{1.96}{\sqrt{2}}\right)\right] \approx [0.021; 0.336]$$

Uncertainty quantification - several approaches

Asymptotic results

✔ fast, easy to describe

✘ not reliable in small samples

Exact tests

✔ very reliable

✘ computer intensive, not always available

Resampling procedures (e.g. boostrap, permutation)

✔ widely applicable - little "math" involved

✘ computer intensive

Confidence intervals - summary

95% confidence intervals:

- represent the uncertainty about our estimate (reasonnable range of values)
- if it does not contain 0, there is evidence for an effect
- if it only contains only "small" values, there is evidence for the absence of a clinically relevant effect

When comparing two estimates

✔ compute the confidence interval of the difference or ratio

✘ do not compare the confidence intervals (unless clear effect)

Likelihood approach - Why?

Systematic approach to:

- estimate parameters
- with their confidence intervals
- and associated significance tests

Especially useful in complex settings, e.g.:

- adjusting on covariates
- handling repeated measurements

Works well when we have:

- an iid[3] sample
- a generative model for the sample

[3] independent and identically distributed

Likelihood approach - roadmap (1/3)

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

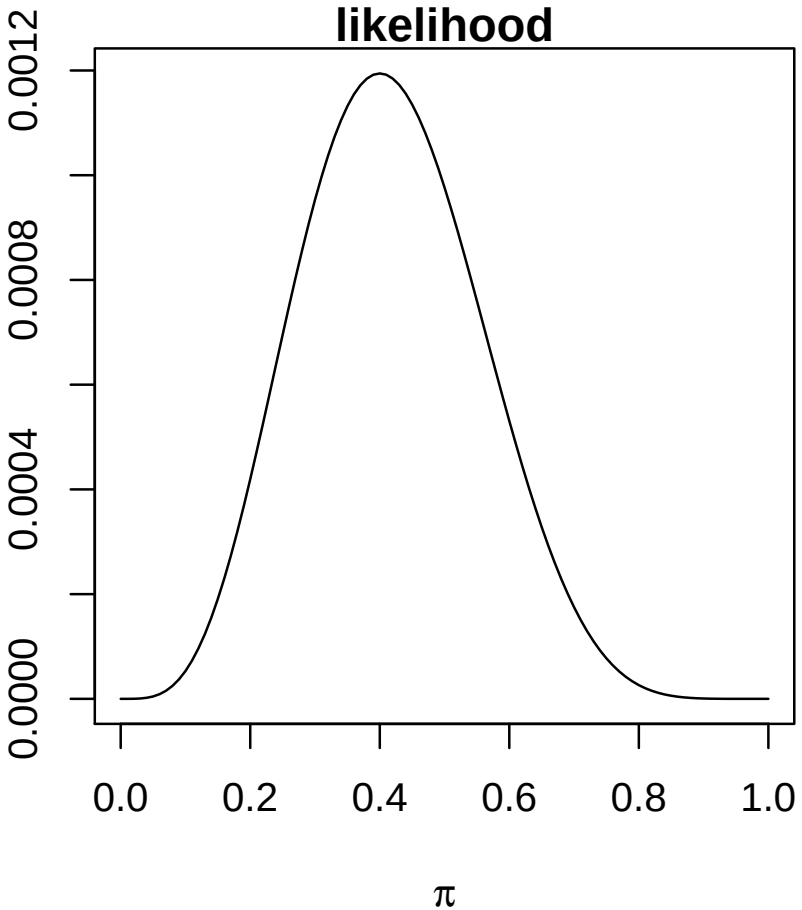
2. express the likelihood
(probability of observing the data given the model)
$\mathcal{L}(\pi) = \prod_{i=1}^{n} \mathbb{P}[Y = Y_i] = \pi^D (1 - \pi)^{n-D}$

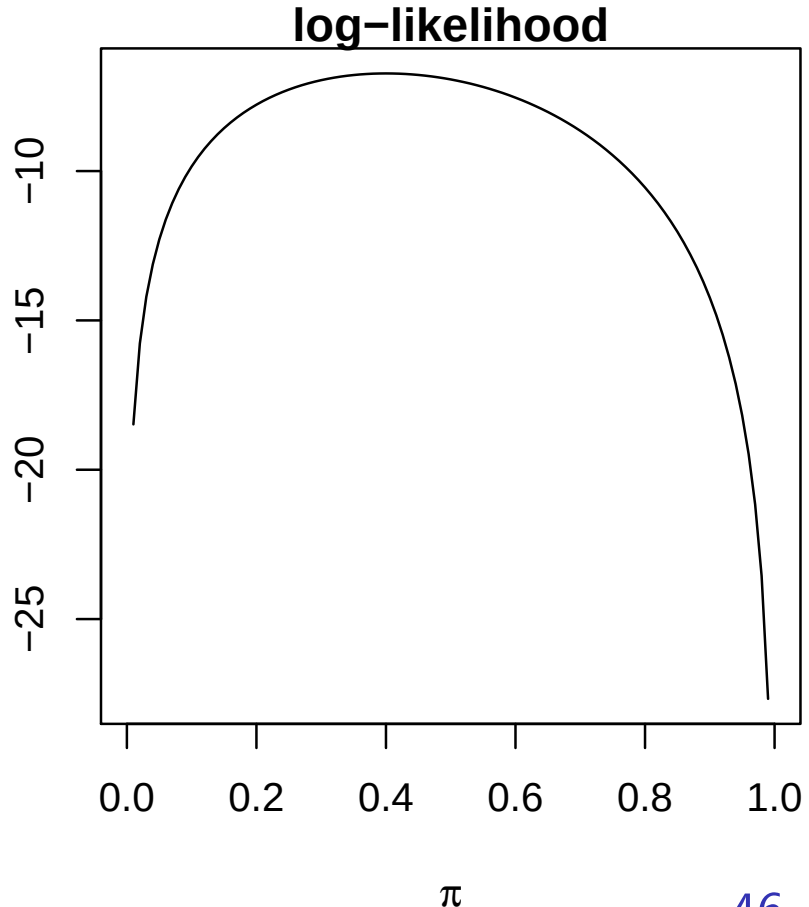
3. express the log-likelihood
$\ell(\pi) = \log(\mathcal{L}(\pi)) = D \log(\pi) + (n - D) \log(1 - \pi)$

Displaying the likelihood

Consider the case where $n = 10$ and $D = 4$

- likelihood: $\mathcal{L}(\pi) = \pi^4(1-\pi)^6$
- log-likelihood $\ell(\pi) = 4\log(\pi) + 6\log(1-\pi)$





Likelihood approach - roadmap (2/3)

log-likelihood: $\ell(\pi) = \log\left(\mathcal{L}(\pi)\right) = D\log(\pi) + (n-D)\log(1-\pi)$

4. find the parameter value maximizing the likelihood (MLE)

i.e. solve[4] $\frac{d\ell(\pi)}{d\pi} = 0$ $\quad \frac{d\ell(\pi)}{d\pi} = \frac{D}{\pi} - \frac{n-D}{1-\pi}$ so $\widehat{\pi} = \frac{D}{n}$

5. quantify the variance of the MLE

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

[4] one should also check that the second derivative of the likelihood is negative

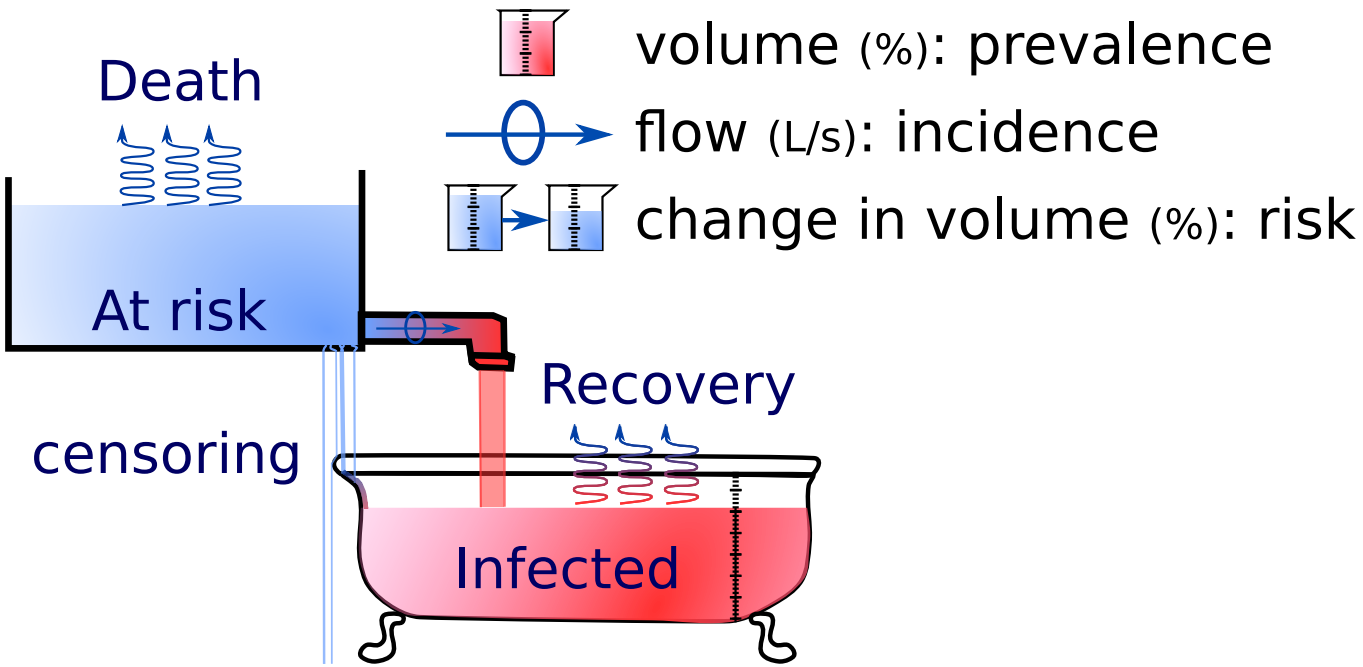
Likelihood approach - roadmap (3/3)

6. The MLE is (asymptotically) unbiased and normally distributed

$$\widehat{\pi} \sim \mathcal{N}\left(\pi, \sigma^2_{\widehat{\pi}}\right)$$

- confidence intervals: $[\widehat{\pi} - 1.96\sigma^2_{\widehat{\pi}}, \widehat{\pi} + 1.96\sigma^2_{\widehat{\pi}}]$
- Wald test $t_W = \frac{\widehat{\pi - 0.5}}{\sigma_{\widehat{\pi}}} \sim \mathcal{N}(0, 1)$ under the null hypothesis of a prevalence of 0.5

Conclusion


volume (%): prevalence
flow (L/s): incidence
change in volume (%): risk
Death
At risk
censoring
Recovery
Infected

What we have seen today

What we have seen today

✔ Introduction:

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

✔ Measures of disease frequency:

- definition and estimation of **prevalence**, odds, **incidence rate**, **risk**,
- unit: per **person.time** for incidence rates
- risk-rate relationship
- estimation of the risk under **right-censoring**

✔ Measures of association

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

✔ Estimation and quantification of the uncertainty

- interpretation of **p-values**
- interpretation and calculation of **confidence intervals** (CIs)
- BONUS: a glimpse at the likelihood theory

Take home messages

Statistical softwares can help you with estimation and quantification of the uncertainty . . . but not with defining the parameter(s) of interest:

- prevalence (static) vs. incidence/risk (dynamic)
- e.g. (registry study) average 5-year risk difference between treatment A and B in the danish population .

Time often plays a big role:

- effects may not be constant over time, especially treatment effects.

For the practical, document `L2-summary.pdf` contains

- formula (estimation, CIs)
- useful R functions

Reference I

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

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}\left[t < T \leq t + dt \middle| T > t\right]}{dt}$$

$$= \lim_{dt \to 0} \frac{\frac{\mathbb{P}[t < T \leq t + dt]}{dt}}{\mathbb{P}\left[T > t\right]} = \lim_{dt \to 0} \frac{\frac{\mathbb{P}[T \leq t + dt] - \mathbb{P}[T \leq t]}{dt}}{\mathbb{P}\left[T > t\right]}$$

$$= \lim_{dt \to 0} \frac{\frac{(1 - S(t + dt)) - (1 - S(t))}{dt}}{S(t)} = \frac{-\frac{\partial S(t)}{\partial t}}{S(t)}$$

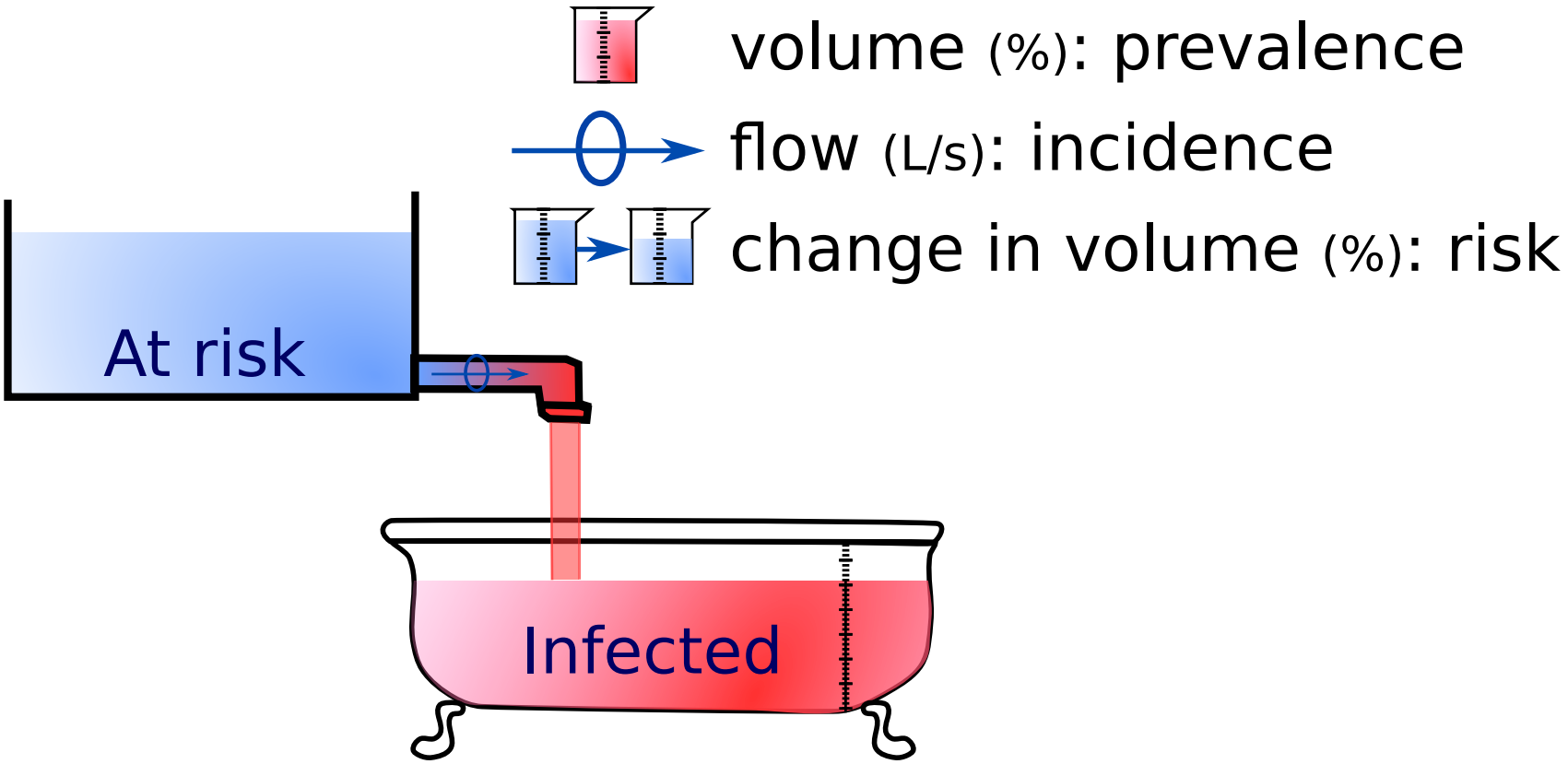
$$\lambda(t) = -\frac{\partial \log S(t)}{\partial t}$$

$$\Lambda(\tau) = \int_0^\tau \lambda(t) dt = -\log S(\tau)$$

$$S(\tau) = \exp(-\Lambda(\tau))$$

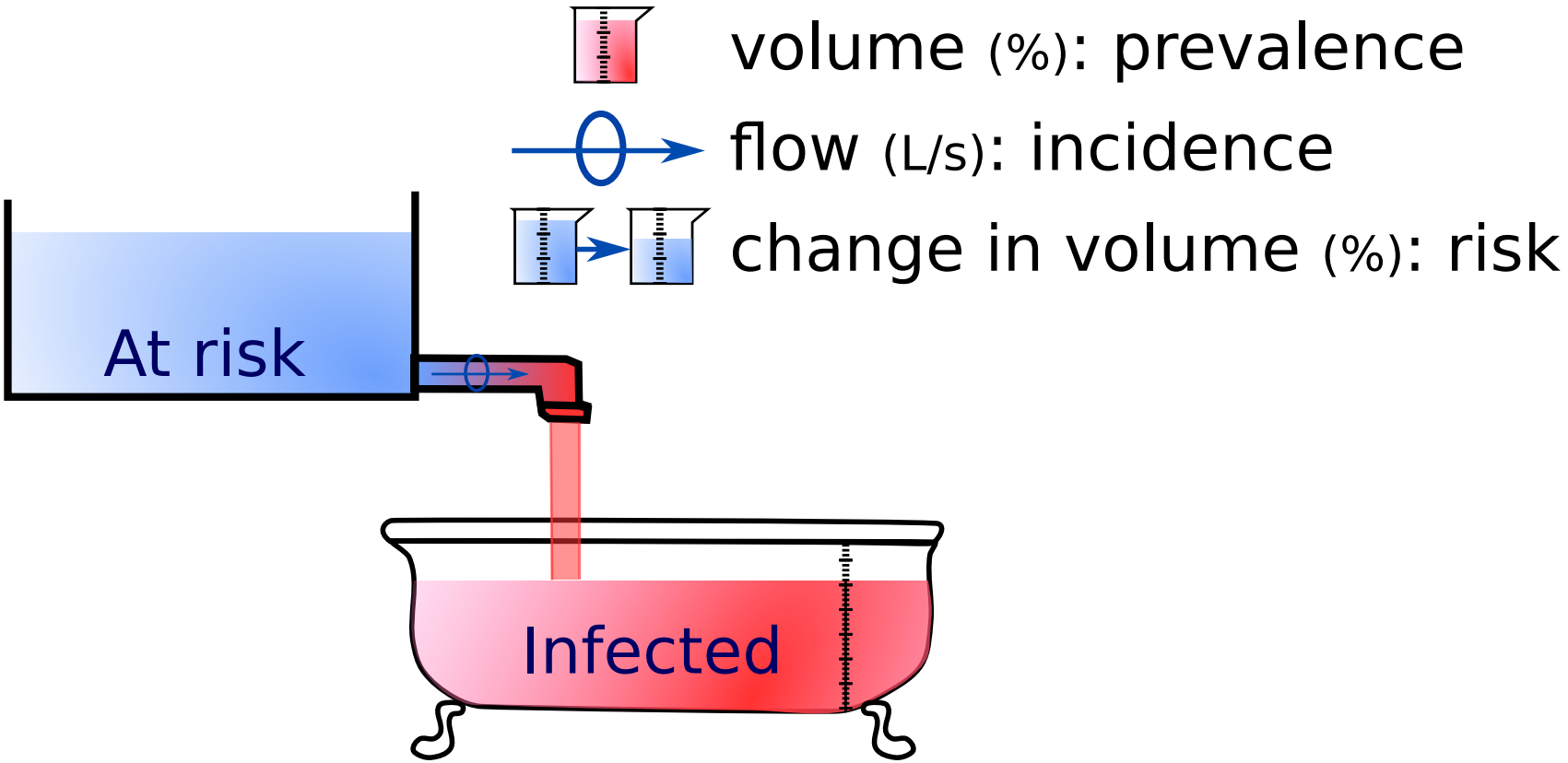
$$r(\tau) = 1 - \exp(-\Lambda(\tau))$$

The epidemiologist's bathtub



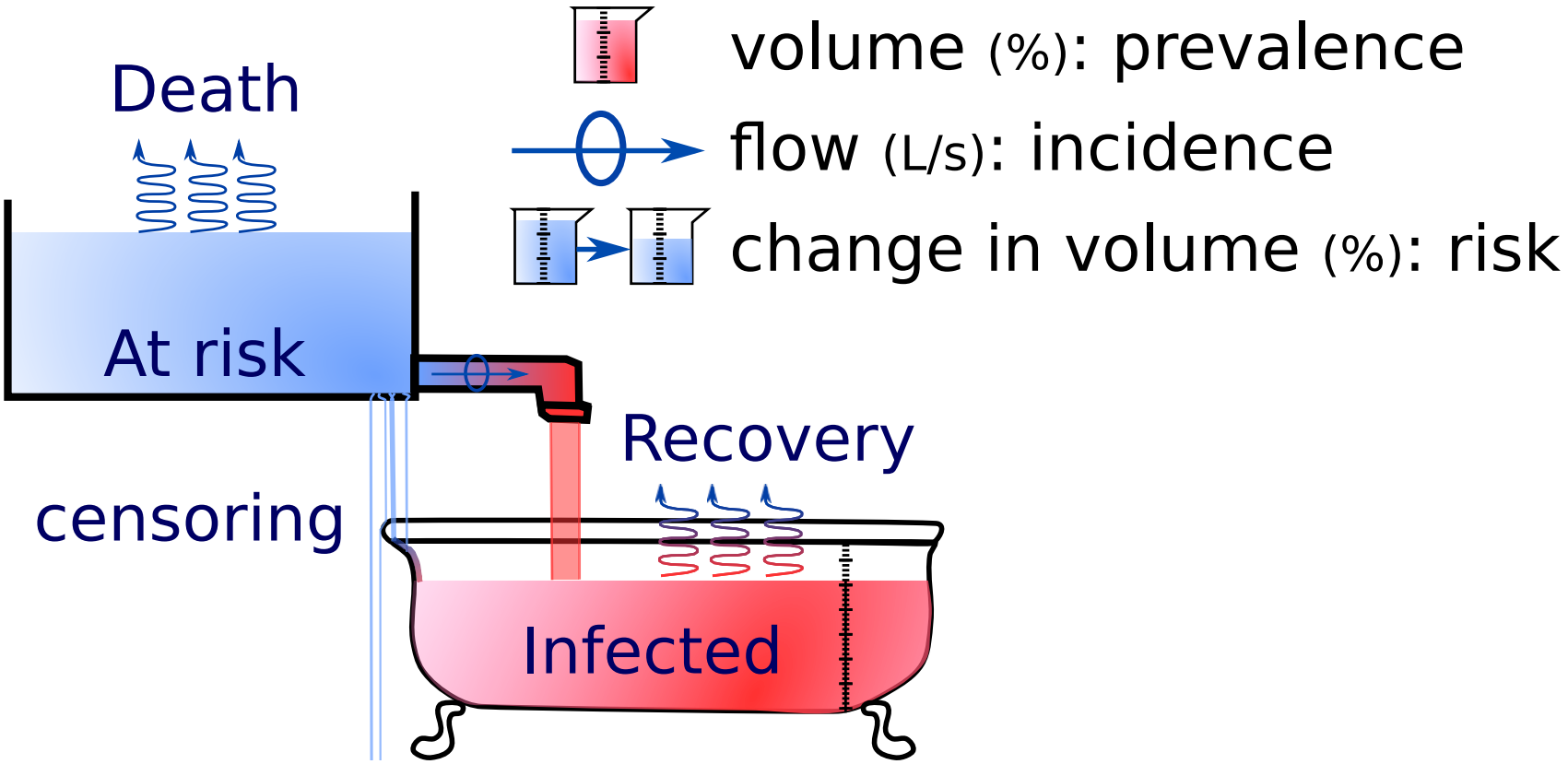
- Prevalence: static
- Incidence rate/rate: dynamic
- risk: dynamic

The epidemiologist's bathtub



- Prevalence: static = incidence x duration
- Incidence rate/rate: dynamic
- risk: dynamic

The epidemiologist's bathtub



- Prevalence: static = ~~incidence x duration~~
- Incidence rate/rate: dynamic
- risk: dynamic

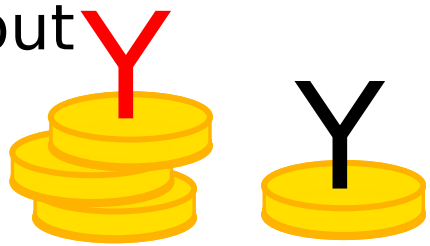
Gambling at 1:3

Expected realisation

Y Y Y Y

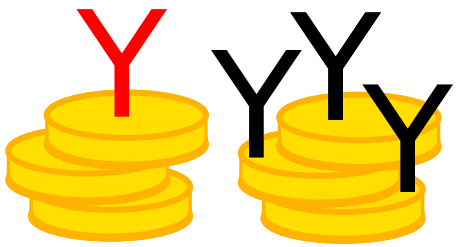
$\Omega =$ 1 / 0

Pay-out



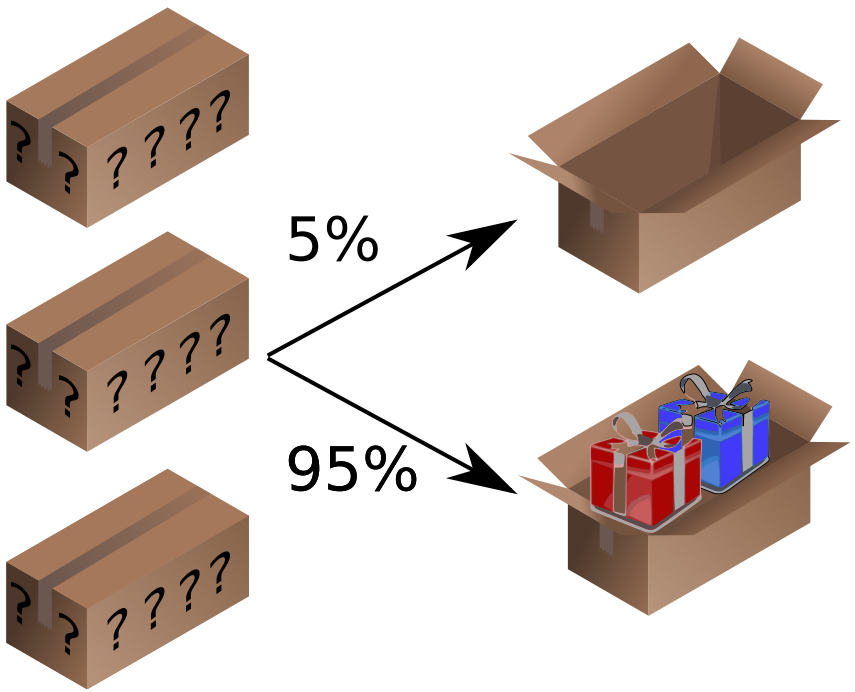
$r =$ 1 / 1 0

Expected gain



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