Nordic Summer School in Cancer Epidemiology

14 to 25 August, 2017

Danish Cancer Society, Copenhagen

Measures of disease frequency and effects

Esa Läärä

Unit of Mathematical Sciences, University of Oulu, Finland

esa.laara@oulu.fi http://math.oulu.fi/

& Bendix Carstensen

Steno Diabetes Center, Denmark

& Department of Biostatistics, University of Copenhagen

bendix.carstensen@regionh.dk

www.bendixcarstensen.com

Outline

Key references

IS: dos Santos Silva, I. (1999).
Cancer Epidemiology: Principles and Methods.
International Agency for Research on Cancer, Lyon.

B&D: Breslow, N.E., Day, N.E. (1987).
Statistical Methods in Cancer Research Vol. II – The Design and Analysis of Cohort Studies.
IARC, Lyon.

C&H: Clayton, D., Hills, M. (1993).
Statistical Models in Epidemiology. OUP, Oxford.

Internet resources on cancer statistics

- **NORDCAN:** Incidence, mortality, prevalence and survival statistics from 41 major cancers in the Nordic countries.

  Association of the Nordic Cancer Registries (ANCR), Danish Cancer Society

  `http://www-dep.iarc.fr/nordcan/English/frame.asp`

  Reference: Engholm, G. *et al.* (2010) NORDCAN – a Nordic tool for cancer information, planning, quality control and research. *Acta Oncologica* **49**: 725-736.

- **GLOBOCAN:** Estimates of the incidence of, mortality, prevalence and disability-adjusted life years (DALYs) from major type of cancers, at national level, for 184 countries of the world in 2008.

  International Agency for Research on Cancer (IARC);

  `http://globocan.iarc.fr/`

INTRODUCTION

What is epidemiology?

Some textbook definitions:

- "study of the **distribution** and **determinants** of disease **frequency** in man" (MacMahon and Pugh 1970)
- "study of the distribution and determinants of health related **states** and **events** in specified populations, ..." (Porta (ed.) Dictionary of Epidemiology, 2014)
- "discipline on principles of **occurrence** research in medicine" (Miettinen 1985)

Different epidemiologies

- **descriptive** epidemiology – monitoring & surveillance of diseases for planning of health services
  – a major activity of cancer registries.
- **etiologic** or "analytic" epidemiology – study of cause-effect relationships
- **disease** epidemiologies – *e.g.* of cancer, cardiovascular diseases, infectious diseases, musculoskeletal disorders, mental health, ...
- **determinant-based** epidemiologies – *e.g.* occupational epidemiology, nutritional epidemiology, ...
- **clinical** epidemiology – study of diagnosis, prognosis and effectiveness of therapies in patient populations
  – basis of evidence-based medicine

Frequency (from Webster's Dictionary)

Etymology: $<$ L *frequentia* $=$ assembly, multitude, crowd.

2. rate of occurrence
3. *Physics*. number of . . . regularly occurring events . . . in unit of time,
5. *Statistics*. the number of items occurring in a given category. Cf. **relative frequency**.

These meanings are all relevant in epidemiology.

But what are **rate** and **occurrence**?

Cancer in Norden 1997 (NORDCAN)

Frequency of cancer (all sites excl. non-melanoma skin) in Nordic male populations expressed by different measures.

	New cases	Crude rate	ASR (World)	Cumul. risk	SIR
Denmark	11 787	452	281	27.8	104
Finland	10 058	401	269	26.5	101
Iceland	633	464	**347**	**32.6**	**132**
Norway	10 246	**469**	294	29.4	109
Sweden	**19 908**	455	249	25.4	93

- Where is the frequency truly **highest**, where lowest?
- What do these measures mean?

Questions on frequency & occurrence

How many women in Denmark

- are carriers of breast cancer today at 12? – **prevalence**
- will contract a new breast ca. during 2015? – **incidence**
- die from breast ca. in 2015? – **mortality**
- will be alive after 5 years since diagnosis among those getting breast ca. in 2015? – **survival**
- are cured of breast cancer during 2015? – **cure**

What are the **proportions** or/and **rates** of occurrence of these states and events?

Questions on risk

- How great are the **risks** of these events?
- Is the risk of breast ca. among nulliparous **greater than** among parous women?
- What are the **excess** and **relative risks** for nulliparous compared to parous women?
- What is the **dose-response relationship** between occupational exposure to crystalline silica and the risk of getting lung cancer in terms of level and length of exposure?

Descriptive and causal questions

- **Descriptive:** What is the occurrence of lung cancer workers exposed to silica dust as compared to that in subjects of other occupations?
- **Causal:** What is the risk of lung cancer among silica dust workers *as compared to* . . . what the risk in these same men *would be, had they not been* exposed to silica?

NB. Causal question – **counterfactual conditional**!

Challenge: *How to find a* **comparable** *group of unexposed?*

What is risk?

Phrase "Risk of disease S" may refer to different concepts:

(i) **probability** of *getting* S during a given **risk period**
$\rightarrow$ **incidence** probability,

(ii) **rate** of change of that probability
$\rightarrow$ **hazard** or intensity, or

(iii) **probability** of *carrying* S at a given *time point*
$\rightarrow$ **prevalence** probability.

Most commonly meaning (i) is attached with risk.

NB. "Risk" should not be used in the meaning of **risk factor**.

However, in **risk assessment** literature: "hazard" is often used in that meaning. In statistics, though, hazard refers to notion (ii): change of probability per unit time.

Risks are conditional probabilities

- There are no "absolute risks".
- All risks are conditional on a multitude of factors, like
  - length of risk period (*e.g.* next week or lifetime),
  - age and gender,
  - genetic constitution,
  - health behaviour & environmental exposures.
- In principle each individual has an own quantitative value for the risk of given disease in any defined risk period, depending on his/her own risk factor profile.
- Yet, these individual risks are latent and unmeasurable.
- **Average risks** of disease in large groups sharing common characteristics (like gender, age, smoking status) are estimable from appropriate epidemiologic studies by pertinent **measures of occurrence**.

BASIC MEASURES OF FREQUENCY OR OCCURRENCE

Quantification of the occurence of disease (or any other health-related state or event) requires specification of:

(1) what is meant by a **case**, *i.e.*, an individual in a population who has or gets the disease

(more generally: possesses the state or undergoes the event of interest).

$\Rightarrow$ challenge to accurate diagnosis and classification!

(2) the **population** from which the cases originate.

(3) the **time point** or **period** of observation.

Types of occurrence measures

- Longitudinal – **incidence** measures:
  incidence rate & incidence proportion
- Cross-sectional – **prevalence** measures.

General form of frequency or occurrence measures

$$\frac{\text{numerator}}{\text{denominator}}$$

Numerator: number of cases observed in the population.

Denominator: generally proportional to the size of the population from which the cases emerge.

Numerator and denominator must cover the *same population*, and the *same period* or *same time point*.

Incidence measures

- **Incidence proportion** (Q) over a fixed *risk period*:

$$Q = \frac{\text{number of incident (new) cases during period}}{\text{size of pop'n at risk at start of the period}}$$

  Also called **cumulative incidence** (even "risk"; *e.g.* **IS**).

  NB. "Cumulative incidence" has other meanings, too.

- **Indidence rate** (I) over a defined observation period:

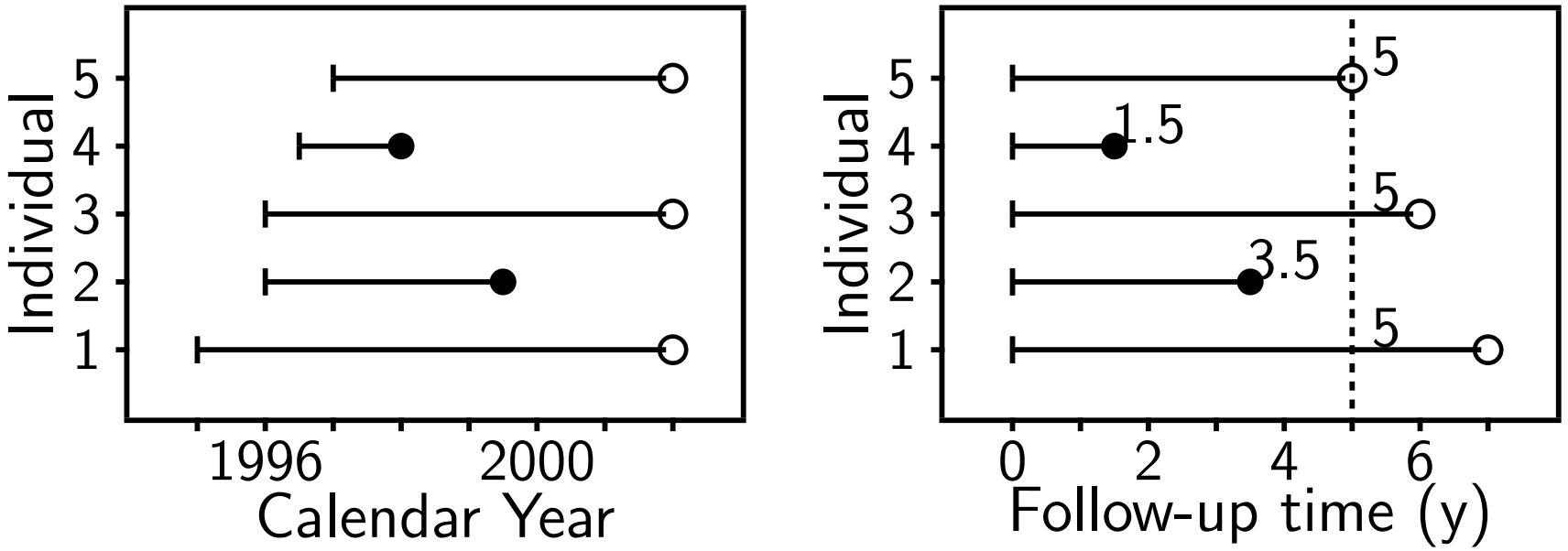
$$I = \frac{\text{number of incident (new) cases during period}}{\text{sum of follow-up times of pop'n at risk}}$$

  Also called **incidence density**.

Example: Follow-up of a small cohort

| = entry, ∘ = exit with censoring; outcome not observed,

• = exit with outcome event (disease onset) observed



Complete follow-up in the 5-year risk period

$\Rightarrow$ can calculate both measures:

$$\text{Inc. rate} = \frac{2 \text{ cases}}{5 + 3.5 + 5 + 1.5 + 5 \text{ years}} = 10 \text{ per } 100 \text{ years},$$

$$\text{Inc. prop.} = 2/5 = 0.4 \text{ or } 40 \text{ per cent.}$$

Properties of incidence proportion

- Dimensionless quantity ranging from 0 to 1 (0% to 100%) = *relative frequency*,
- Estimates the average theoretical **risk** or probability of the outcome occurring during the risk period, in the **population at risk** – *i.e.* among those who are still free from the outcome at the start of the period,
- Simple formula valid when the follow-up time is fixed & equals the risk period, and when there are no **competing events** or **censoring**.
- Competing events & censoring $\Rightarrow$ Calculations need to be corrected using special methods of survival analysis.

Properties of incidence rate

- Like *a frequency* quantity in physics; measurement unit: *e.g.* Hz $= 1/$second, $1/$year, or $1/1000$ y.
- Estimates the average underlying **intensity** or **hazard rate** of the outcome in a population,
- Estimation accurate in the **constant hazard model**,
- Calculation straightforward also with competing events and censored observations.
- Hazard depends on age (& other time variables)
  $\Rightarrow$ rates *specific to age group etc.* needed,
- Incidence proportions can be estimated from rates.
  In the constant hazard model with no competing risks:

$$Q = 1 - \exp(-I \times \Delta) \approx I \times \Delta$$

Competing events and censoring

The outcome event of interest (*e.g.* onset of disease) is not always observed for all subjects during the chosen risk period.

- Some subjects die (from other causes) before the event.

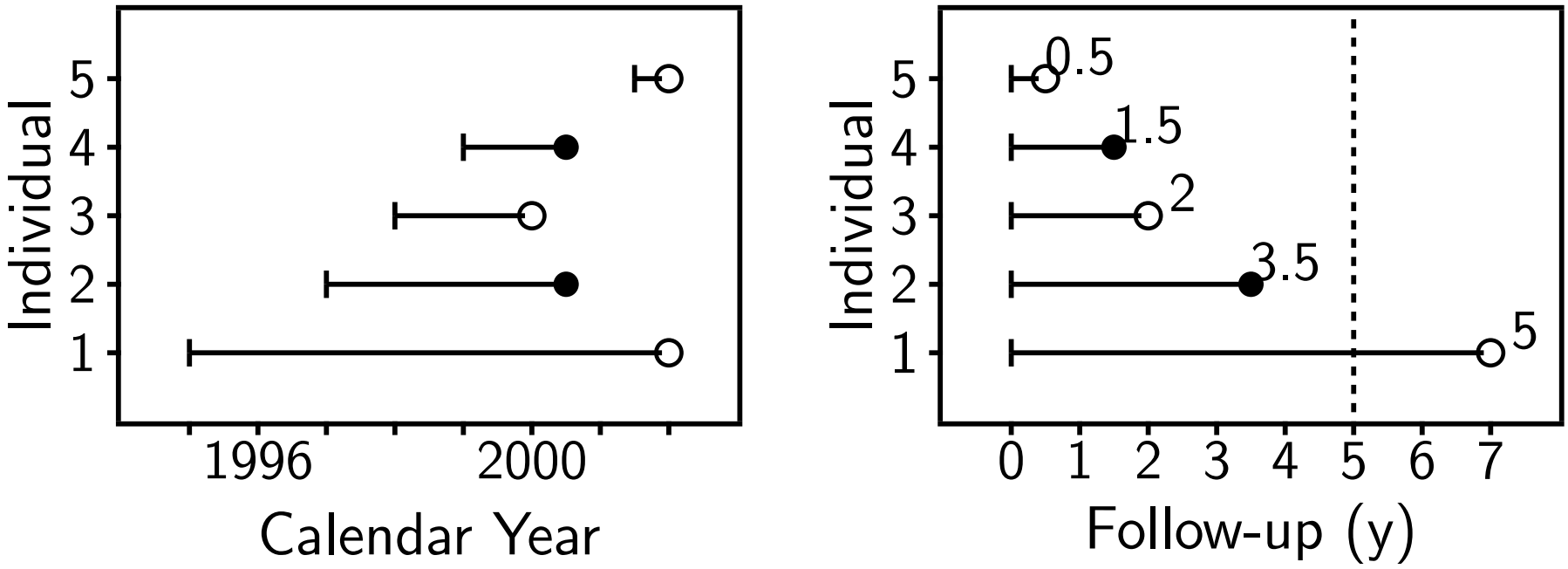
  $\Rightarrow$ Death is a **competing event** after which the outcome cannot occur any more.

- Others emigrate and escape national disease registration, or the whole study is closed "now", which prematurely interrupts the follow-up of some individuals

  $\Rightarrow$ **censoring**, **withdrawal**, or **loss to follow-up**

Competing events and censorings require special statistical treatment in estimation of incidence and risk.

Follow-up of another small cohort



Two censored observations $\Rightarrow$ the rate can be calculated:

$$I = 2/12.5 \text{ y} = 16 \text{ per } 100 \text{ years}$$

but the 5-year incidence proportion **IS NO MORE** $2/5$!

However, under the constant rate model and in the absence of competing risks, the incidence proportion is obtained:

$$Q = 1 - \exp(-5 \times 2/12.5) = 0.55 \text{ (or 55\%)}$$

Person-years in dynamic populations

With dynamic study population individual follow-up times are always variable and impossible to measure accurately.

Common approximation – **mid-population** principle:

(1) Let the population size be N_{t-1} at start and N_t at the end of the observation period t with length u_t years,

(2) Mid-population for the period: $\bar{N}_t = \frac{1}{2} \times (N_{t-1} + N_t)$.

(3) Approximate person-years: $\widetilde{Y}_t = \bar{N}_t \times u_t$.

NB. The actual study population often contains also some already affected, who thus do not belong to the population at risk. With rare outcomes the influence of this is small.

Male person-years in Finland 1991-95

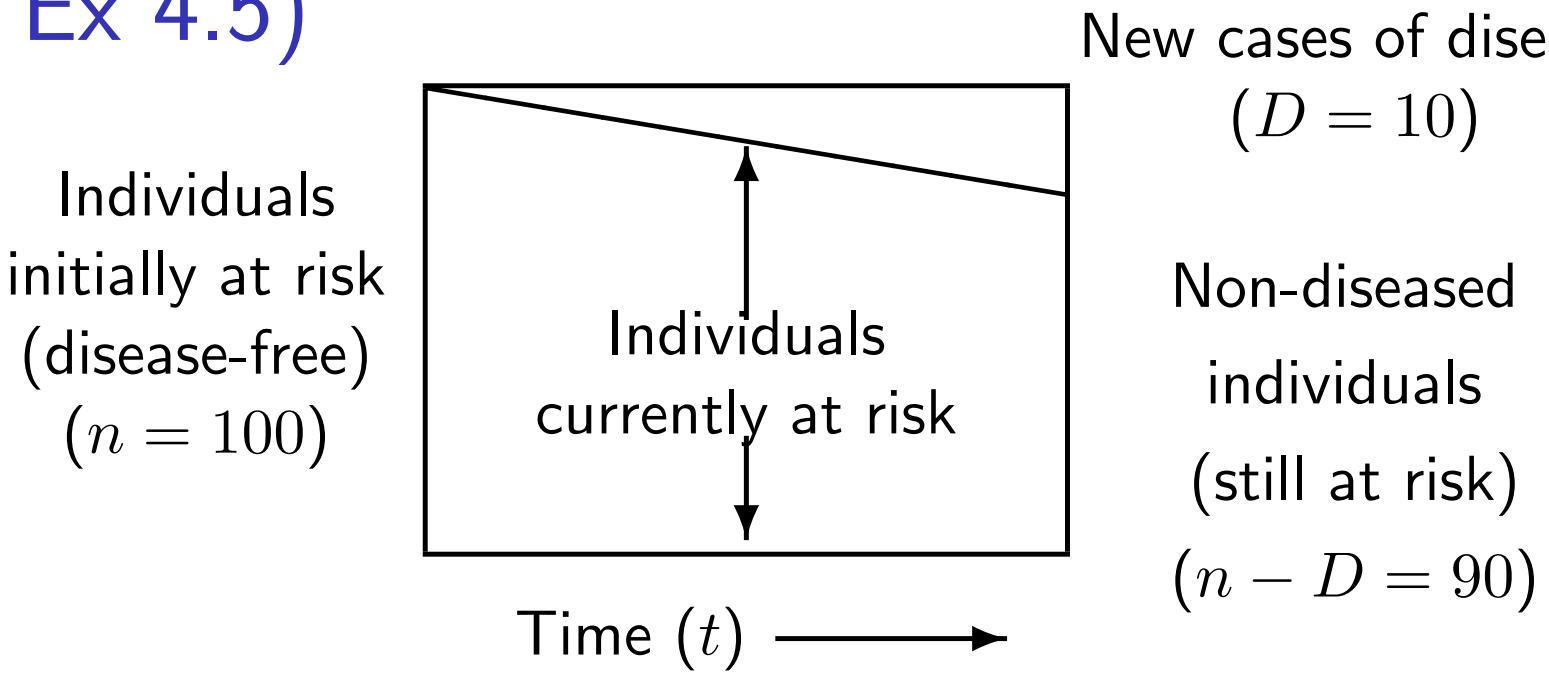
Total male population (1000s) on 31 December by year:

1990	1991	1992	1993	1994	1995
2431	2443	2457	2470	2482	2492

Approximate person-years (1000s) in various periods:

$$\begin{aligned}
\text{1992:} \quad & \tfrac{1}{2} \times (2443 + 2457) \times 1 = && 2450 \\
\text{1993-94:} \quad & \tfrac{1}{2} \times (2457 + 2482) \times 2 = && 4937 \\
\text{1991-95:} \quad & \tfrac{1}{2} \times (2431 + 2492) \times 5 = && 12307.5
\end{aligned}$$

Incidence proportion, rate, and odds (IS, Ex 4.5)



Assuming a risk period of 1 year with complete follow-up:

$$\begin{aligned} \text{Incidence proportion } Q &= 10/100 = 0.10 = 10\% \\ \text{Incidence rate } I &= 10/95 \text{ y} = 10.5 \text{ per 100 y} \\ \textbf{Incidence odds } Q/(1-Q) &= 10/90 = 0.11 = 11 \text{ per } 100 \end{aligned}$$

Approximate relations btw measures

With sufficiently

- "short" length Δ of risk period and
- "low" risk (say $Q < 5\%$)

the incidence proportion Q, rate I and odds are approximately related as follows:

$$\frac{Q}{1-Q} \approx Q \approx I \times \Delta$$

The "**rare disease assumption**".

Mortality

Cause-specific mortality from disease S is described by **mortality rates** defined like I but

- cases are *deaths* from S, and
- follow-up is extended until death or censoring.

Cause-specific **mortality proportions** must be corrected for the incidence of **competing causes of death**

Total mortality:

- cases are deaths from any cause.

Mortality depends on the incidence and the **prognosis** or **case fatality** of the disease, *i.e.* the **survival** of those affected by it.

Prevalence measures

Point prevalence or simply **prevalence** P of a health state C in a population at a given time point t is defined

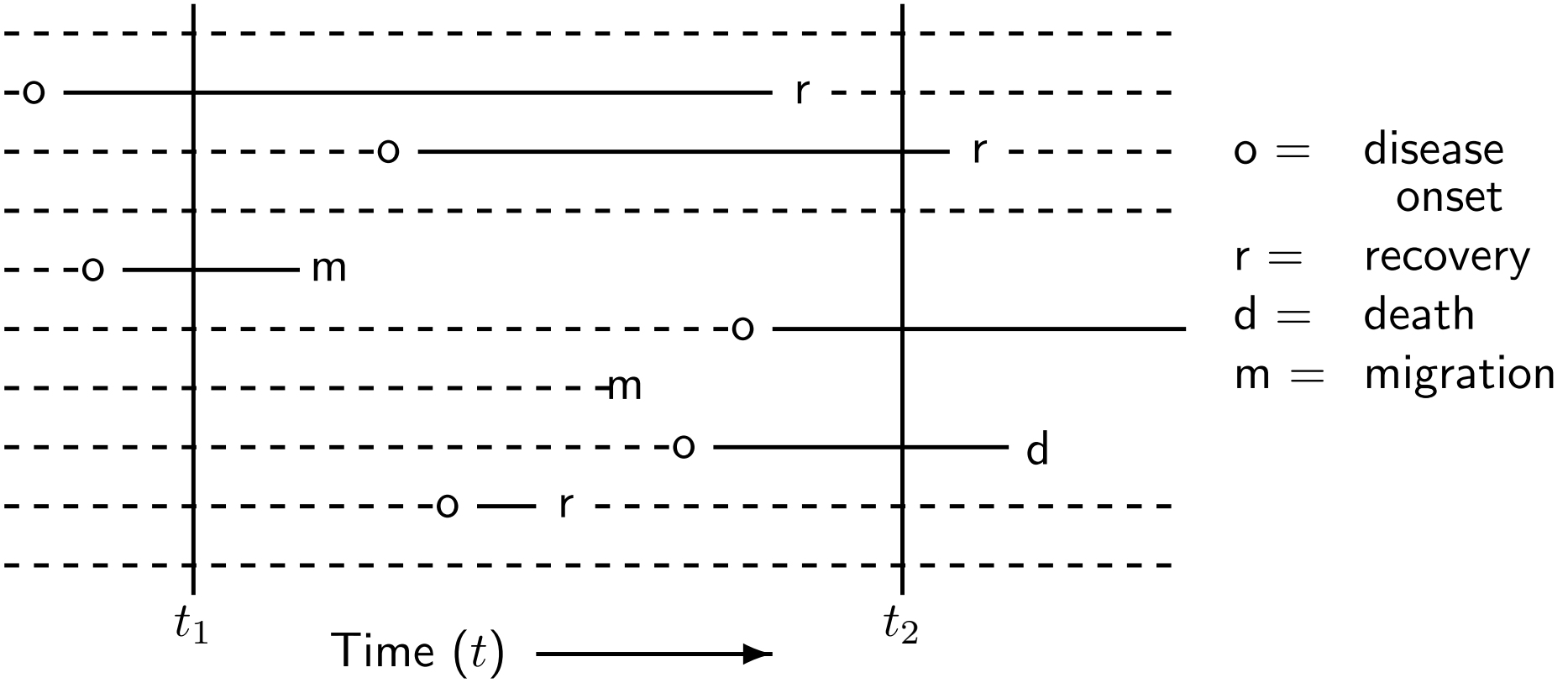
$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

This is calculable from a cross-sectional study base.

Period prevalence for period from t_1 to t_2 is like P but

- numerator refers to all cases prevalent already at t_1 plus new cases occurring during the period, and
- denominator is the population size at t_2.

Example 4.1 (IS: p. 59)



Prevalence at time t_1 : $2/10 = 0.2 = 20\%$

Prevalence at time t_2 : $3/8 = 0.38 = 38\%$

Period prevalence: $5/8 = 0.62 = 62\%$

Prevalence and incidence are related

Point prevalence of S at given time point t depends on the

(a) *incidence* of new cases of S before t, and the

(b) *duration* of S, depending in turn on the probability of *cure* or recovery from S, or *survival* of those affected

typically in a complicated way.

Simple special case: In a **stationary** population, the prevalence (P), incidence (I), and average duration ($\bar{d}$) of S have a simple relationship:

$$P = \frac{I \times \bar{d}}{I \times \bar{d} + 1} \quad \approx \quad I \times \bar{d}$$

The approximation works well, when $P < 0.1$ (10%).

Prevalence of cancer?

- How do we know, whether and when cancer is cured?

  $\Rightarrow$ Existing or prevalent case problematic to define.

- NORDCAN: Prevalence of cancer C at time point t in the target population refers to the

  • number & proportion of population members who

  (a) are alive and resident in the population at t, and
  (b) have a record of an incident cancer C diagnosed before t.

- **Partial prevalence:** Cases limited to those diagnosed during a fixed time in the past; *e.g.* within 1 y (initial treatment period), 3 y (clinical follow-up), or 5 y (cure?).

Ex: Cancers with poor and good prognosis

Age-standardized[a] incidence, mortality, prevalence, and survival for cancers of kidney and thyroid in women of Finland.

	Kidney	Thyroid
Incidence rate in 2011 (per 10^5 y)	12	11
Mortality rate in 2011 (per 10^5 y)	5	1
Prevalence on 31.12.2011 (per 10^5)	92	198
– diagnosed $<$ 1 y ago	*9*	*10*
– diagnosed $<$ 3 y ago	*24*	*29*
– diagnosed $<$ 5 y ago	*35*	*47*
– diagnosed $>$ 5 y ago	*57*	*151*
5-y relative survival; cases 2004–8 (%)	64	90

[a] Standard: Nordic population in 2000

MEASURES OF EFFECT – COMPARATIVE MEASURES

- Quantification of the **association** between a determinant (risk factor) and an outcome (disease) is based on

  comparison of occurrence between the *index* ("exposed") and the *reference* ("unexposed") groups by
  - relative comparative measures (ratio)
  - absolute comparative measures (difference)

- In causal studies these are used to estimate the **causal effect** of the factor on the disease risk.

  $\Rightarrow$ **comparative measure** $\approx$ **effect measure**

- Yet, caution is needed in inferences on causal effects, as often the groups to be compared suffer from **poor comparability** $\Leftrightarrow$ **Confounding**.

Relative comparative measures

Generic name **"relative risk"** (RR) comparing occurrences between exposed (1) and unexposed (0) groups can refer to

- incidence rate ratio I_1/I_0,
- incidence proportion ratio Q_1/Q_0,
- incidence odds ratio $[Q_1/(1-Q_1)]/[Q_0/(1-Q_0)]$,
- prevalence ratio P_1/P_0, or
- prevalence odds ratio $[P_1/(1-P_1)]/[P_0/(1-P_0)]$,

depending on study base and details of its design.

Incidence rate ratio is the most commonly used comparative measure in cancer epidemiology.

Absolute comparative measures

Generic term **"excess risk"** or **"risk difference"** (RD) btw exposed and unexposed can refer to

- incidence rate difference $I_1 - I_0$,
- incidence proportion difference $Q_1 - Q_0$, or
- prevalence difference $P_1 - P_0$.

Use of relative and absolute comparisons

- Ratios – describe the **biological strength** of the exposure
- Differences – inform about its **public health importance**.

Example: (IS, Table 5.2, p.97)

Relative and absolute comparisons between the exposed and the unexposed to risk factor X in two diseases.

	Disease A	Disease B
Incidence rate among exposed[a]	20	80
Incidence rate among unexposed[a]	5	40
Rate ratio	4.0	2.0
Rate difference[a]	15	40

[a] Rates per 100 000 pyrs.

Factor X has a stronger biological potency for disease A, but it has a greater public health importance for disease B.

Ratio measures in "rare diseases"

(IS: Ex 5.13)

	Exposure	
	Yes	No
No. initially at risk	4 000	16 000
No. of cases	30	60
Person-years at risk	7 970	31 940

$$\text{Inc. prop'n ratio} = \frac{30/4\,000}{60/16\,000} = \frac{7.5 \text{ per } 1\,000}{3.75 \text{ per } 1\,000} = \mathbf{2.0000}$$

$$\text{Inc. rate ratio} = \frac{30/7\,970 \text{ y}}{60/31\,940 \text{ y}} = \frac{3.76 \text{ per } 1\,000 \text{ y}}{1.88 \text{ per } 1\,000 \text{ y}} = \mathbf{2.0038}$$

$$\text{Inc. odds ratio} = \frac{30/(4\,000-30)}{60/(16\,000-60)} = \frac{0.00756}{0.00376} = \mathbf{2.0076}$$

With low incidence these ratios are very similar.

Attributable fraction (excess fraction)

- **Measures of potential impact**:
  Combination of absolute and relative comparisons.
- When the incidence is higher in the exposed, the **attributable fraction** (AF) for the exposure or risk factor is defined as:

$$\text{AF} = \frac{I_1 - I_0}{I_1} = \frac{\text{RR} - 1}{\text{RR}}.$$

  Also called **excess fraction** (or even "attributable risk" in old texts).
- This measure estimates the fraction out of all new cases of disease *among those exposed*, which are attributable to (or "caused" by) the exposure itself, and which thus could be avoided if the exposure were absent.

Population attributable fraction

- Suppose we ask instead:

  "How large a fraction of all cases in the population would be prevented, if the exposure were eliminated?"

- The answer to this question depends in addition on

$$p_{\mathsf{E}} = \text{ proportion of exposed in the population.}$$

- **Population excess fraction** (PAF) is now defined:

$$\mathsf{PAF} = \frac{I - I_0}{I} = \frac{p_{\mathsf{E}}(\mathsf{RR} - 1)}{1 + p_{\mathsf{E}}(\mathsf{RR} - 1)}$$

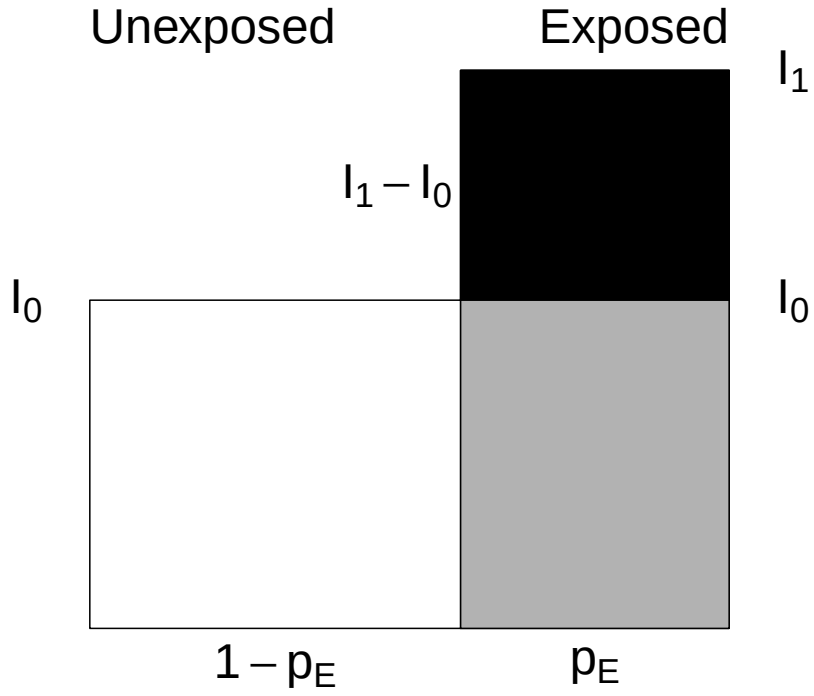
- AF: biological impact of exposure,
- PAF: impact of exposure on the population level.

Excess fraction illustrated

- The population divided into exposed and unexposed.
- The rate I_1 among exposed would be I_0, *i.e.* same as in unexposed, if the exposure had no effect.
- The excess $I_1 - I_0$ is caused by the exposure.

- AF $= \dfrac{I_1 - I_0}{I_1}$,

$=$ fraction of black area out of total black + gray area.



PAF illustrated

- Total incidence I in population – weighted average:

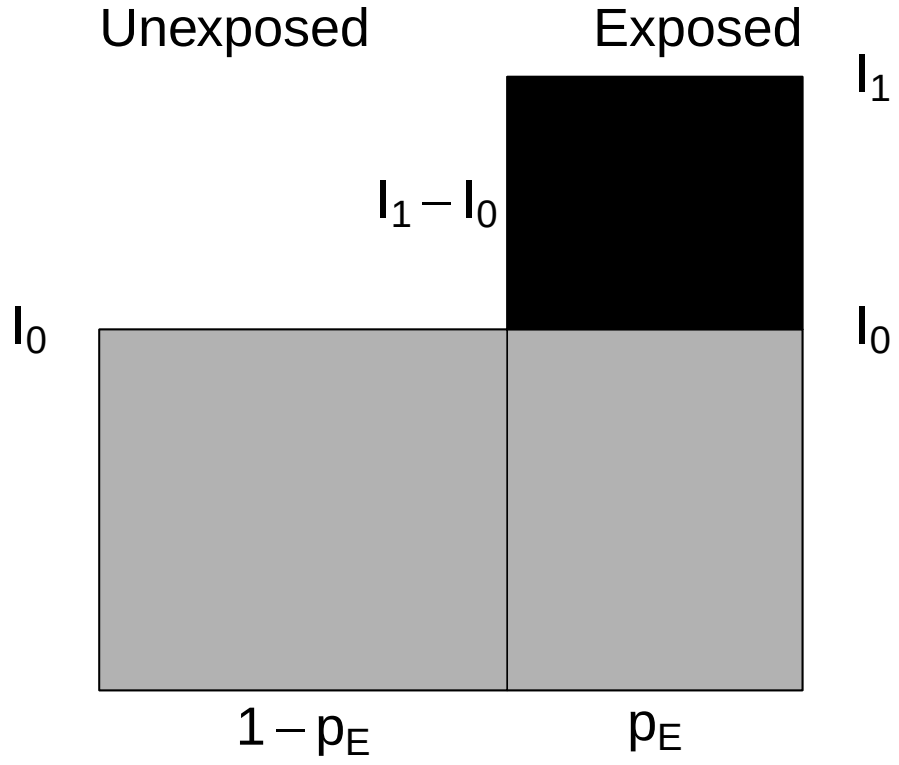
$$I = p_{\mathsf{E}} \times I_1 + (1 - p_{\mathsf{E}}) \times I_0 \quad \text{(total area)}$$

  would equal I_0, if exposure had no effect

- Excess incidence caused by exposure:
  $I - I_0 = p_{\mathsf{E}} \times (I_1 - I_0)$ (black area).

- PAF $= \dfrac{I - I_0}{I}$,

= fraction of black area out of total black + gray area.



Prevented fractions

- When the incidence in exposed is lower, we define the **prevented fraction** for such a preventive factor:

$$\mathrm{PF} = \frac{I_0 - I_1}{I_0} = 1 - \mathrm{RR}$$

  also called **relative risk reduction** = percentage of cases prevented among the exposed due to the exposure.

- Used to evaluate the relative effect of a preventive intervention ("exposure") vs. no intervention.

- **Population prevented fraction** (PPF) combines this with the prevalence of exposure in the population:

$$\mathrm{PPF} = \frac{I_0 - I}{I_0} = p_{\mathrm{E}} \times (1 - \mathrm{RR}),$$

  measuring the relative reduction in caseload attributable to the presence of preventive factor in the population.

Effect of smoking on mortality by cause

(**IS**: Example 5.14, p. 98)

Underlying cause of death	Never smoked regularly Rate[b] (1)	Current cigarette smoker Rate[b] (2)	Rate ratio (2)/(1)	Rate difference[b] (2) − (1)	Excess fraction (%) $\frac{(2)-(1)}{(2)} \times 100$
Cancer					
All sites	305	656	2.2	351	54
Lung	14	209	14.9	195	93
Oesophagus	4	30	7.5	26	87
Bladder	13	30	2.3	17	57
Respiratory diseases (except cancer)	107	313	2.9	206	66
Vascular diseases	1037	1643	1.6	606	37
All causes	1706	3038	1.8	1332	44

[a] Data from Doll *et al.*, 1994a.
[b] Age-adjusted rates per 100 000 pyrs.

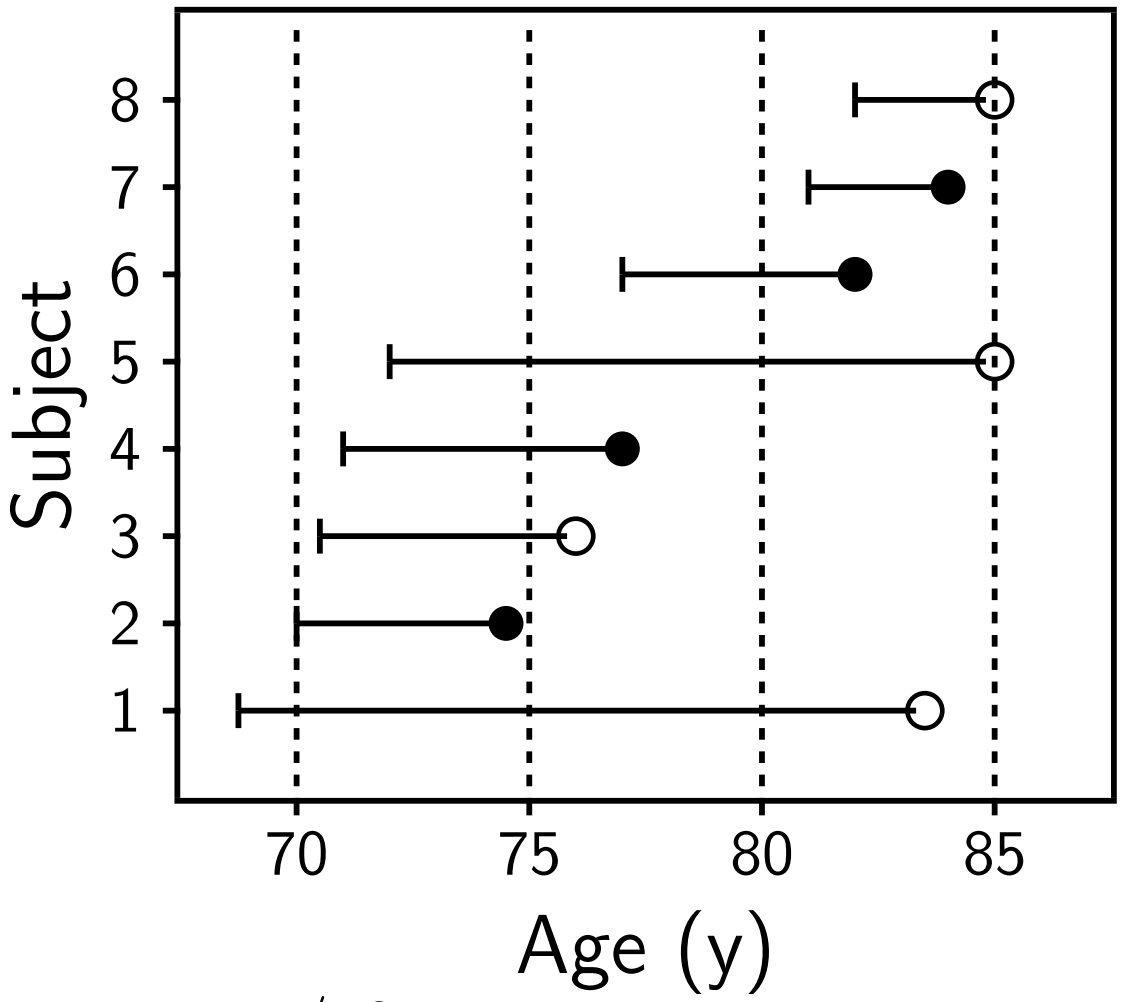
RATES IN MANY TIME SCALES

Incidence can be studied on various distinct time scales, *e.g.*

Time scale	Origin: date of ...
age	birth
exposure time	first exposure
follow-up time	entry to study
duration of disease	diagnosis

- Age is usully the strongest time-dependent determinant of health outcomes.
- Age is also often correlated with duration of "chronic" exposure (*e.g.* years of smoking).

Follow-up of a small geriatric cohort



Overall rate: 4 cases/53.5 person-years = 7.5 per 100 y.

Hides the fact that the "true" rate probably varies by age, being higher among the old.

Splitting follow-up into agebands

- To describe, how incidence varies by age, individual person-years from age of entry to age of exit must first be split or divided into narrower agebands.
- Usually these are based on common 5-year age grouping.
- Numbers of cases are equally divided into same agebands.

- **Age-specific incidence rate** for age group k is

$$I_k = \frac{\text{number of cases observed in ageband}}{\text{person-years contained in ageband}}$$

- Underlying assumption:
  piecewise constant rates model

Person-years and cases in agebands: age-specific rates

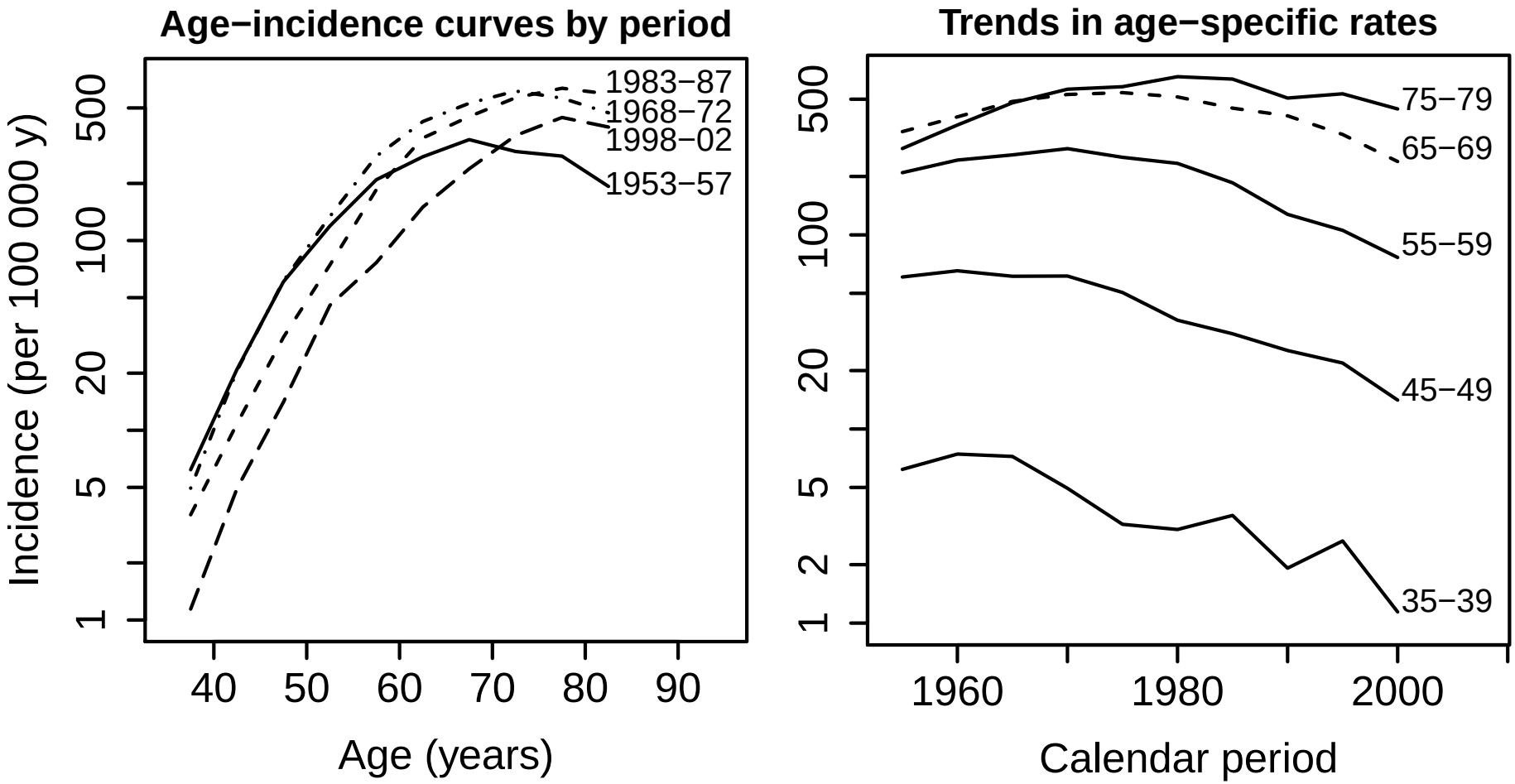
	Ageband			
Subject	70-74	75-79	80-84	Total
1	5.0	5.0	3.5	13.5
2	4.5	-	-	4.5
3	4.5	1.0	-	5.5
4	4.0	2.0	-	6.0
5	3.0	5.0	5.0	13.0
6	-	3.0	2.0	5.0
7	-	-	3.0	3.0
8	-	-	3.0	3.0
Sum of person-years	21.0	16.0	16.5	53.5
Cases	1	1	2	4
Rate (/100 y)	4.8	6.2	12.1	7.5
	Age-specific rates			overall

Ex. Lung cancer incidence in Finland by age and period (compare IS, Table 4.1)

Calendar	Age group (y)									
period	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85+
1953-57	21	61	119	209	276	340	295	279	193	93
1958-62	22	65	135	243	360	405	429	368	265	224
1963-67	24	61	143	258	395	487	509	479	430	280
1968-72	21	61	134	278	424	529	614	563	471	358
1973-77	16	50	134	251	413	541	629	580	490	392
1978-82	13	36	115	234	369	514	621	653	593	442
1983-87	11	31	74	186	347	450	566	635	592	447
1988-92	9	25	57	128	262	411	506	507	471	441
1993-97	7	22	48	106	188	329	467	533	487	367
1998-02	5	14	46	77	150	239	358	445	396	346

- Rows: age-incidence pattern in different calendar periods.
- Columns: Trends of age-specific rates over calendar time.

Lung cancer rates by age and period



- Age-incidence curves: overall level and peak age variable across periods.
- Time trends inconsistent across age groups.

Incidence by age, period & cohort

- **Secular trends** of specific and adjusted rates show, how the "cancer burden" has developed over periods of calendar time.

Birth cohort = people born during the same limited time interval, *e.g.* single calendar year, or 5 years period.

- Analysis of rates by birth cohort reveals, how the level of incidence (or mortality) differs between successive generations – may reflect differences in risk factor levels.
- Often more informative about "true" age-incidence pattern than age-specific incidences of single calendar period.

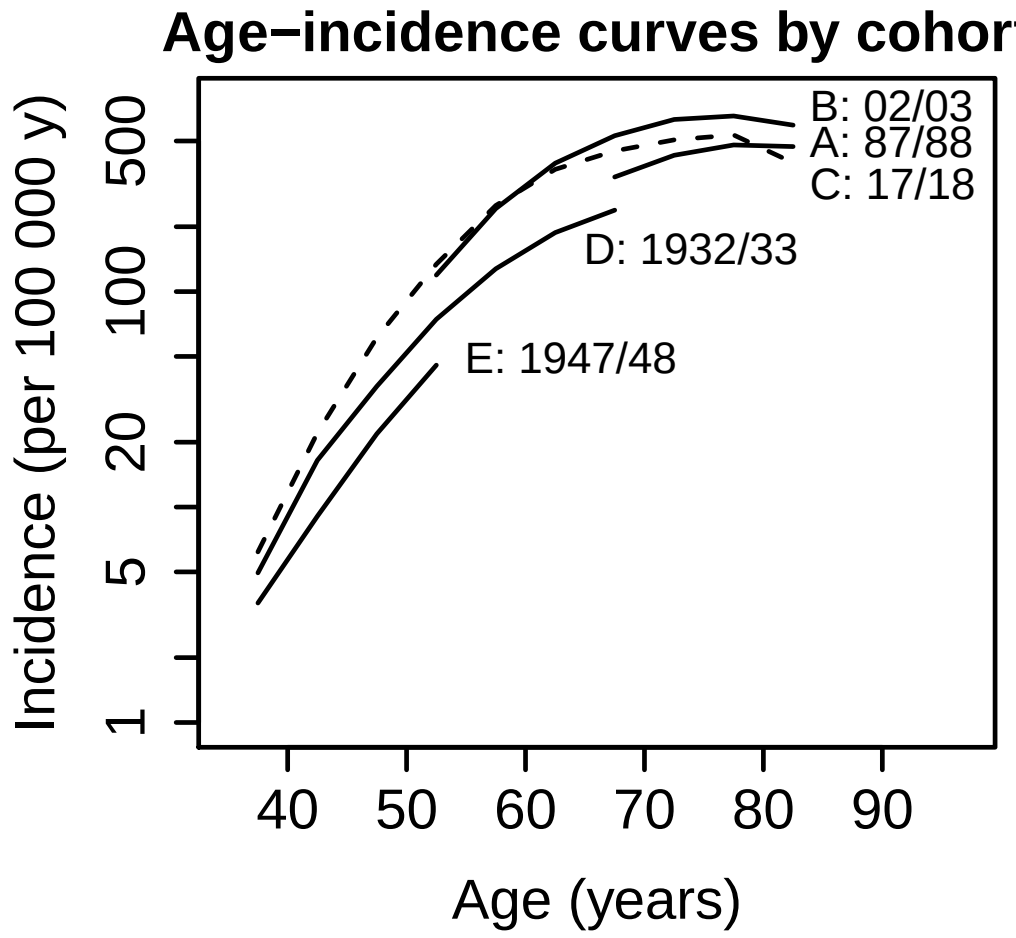
Age-specific rates by birth cohort

Calendar	Age group (y)								
period	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	
1953-57	21	61	[119]	209	276	[340]	295	279	
1958-62	[22]	65	135	[243]	360	405	[429]	368	
1963-67	24	[61]	143	258	[395]	487	509	[479]	A
1968-72	21	61	[134]	278	424	[529]	614	563	
1973-77	[16]	50	134	[251]	413	541	[629]	580	
1978-82	13	[36]	115	234	[369]	514	621	[653]	B
1983-87	11	31	[74]	186	347	[450]	566	635	
1988-92	[9]	25	57	[128]	262	411	[506]	507	
1993-97	7	[22]	48	106	[188]	329	467	[533]	C
1998-02	5	14	[46]	77	150	[239]	358	445	
			E: 1947/48			D: 1932/33			

A = synthetic cohort born around 1887/88, B: 1902/03, C: 1917/18

Diagonals reflect age-incidence pattern in birth cohorts.

Age-incidence curves in 5 birth cohorts



Variable overall levels but fairly consistent form and similar peak age across different birth cohorts.

Split of follow-up by age and period

- Incidence of (or mortality from) disease C in special **cohort of exposed** (*e.g.* occupational group, users of certain medicine)
  $\rightarrow$ often compared to incidence in a **reference** or "general" population.

- For examples, see Laufey's lecture on cohort studies (*e.g.* atomic bomb survivors, rubber workers, and those exposed to dyestaff)

- Adjustment for age and calendar time needed, *e.g.* by comparing **observed** to **expected** cases with SIR (see p. 70-74).

  $\Rightarrow$ Cases and person-years in the study cohort must be split by more than one time scale (age).

Example (C&H, Tables 6.2 & 6.3, p. 54)

Entry and exit dates for a small cohort of four subjects

Subject	Born	Entry	Exit	Age at entry	Outcome
1	1904	1943	1952	39	Migrated
2	1924	1948	1955	24	Disease C
3	1914	1945	1961	31	Study ends
4	1920	1948	1956	28	Unrelated death

Subject 1: Follow-up time spent in each ageband

Age band	Date in	Date out	Time (years)
35–39	1943	1944	1
40–44	1944	1949	5
45–49	1949	1952	3

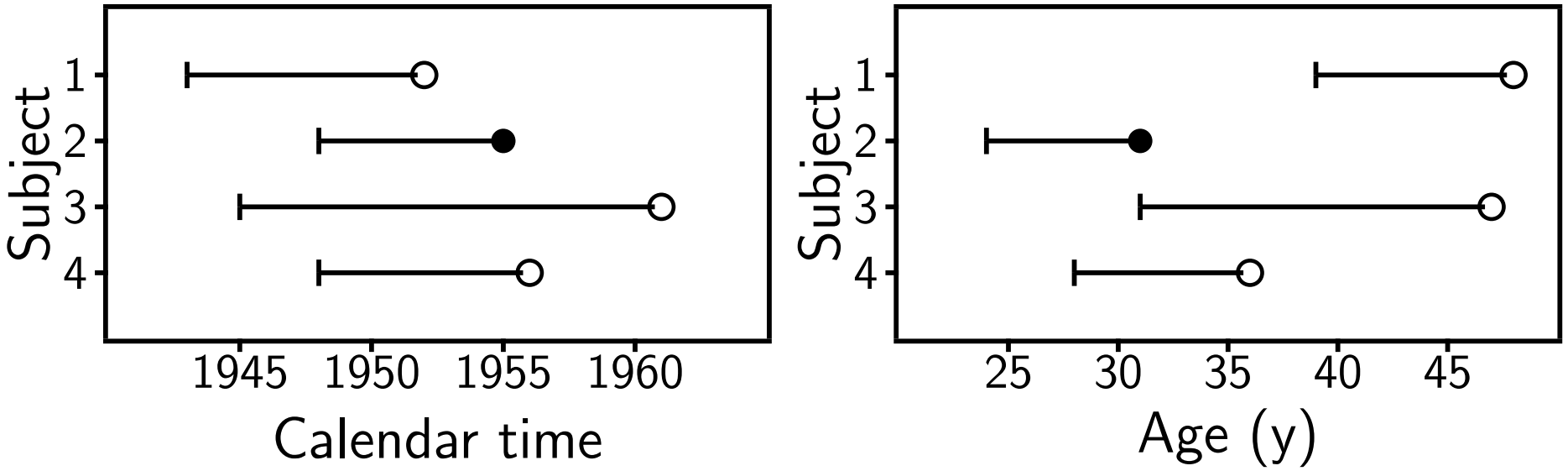
Example: (C&H, Figures 6.1 & 6.2, p. 55)

Follow-up of cohort members by calendar time and age

| entry

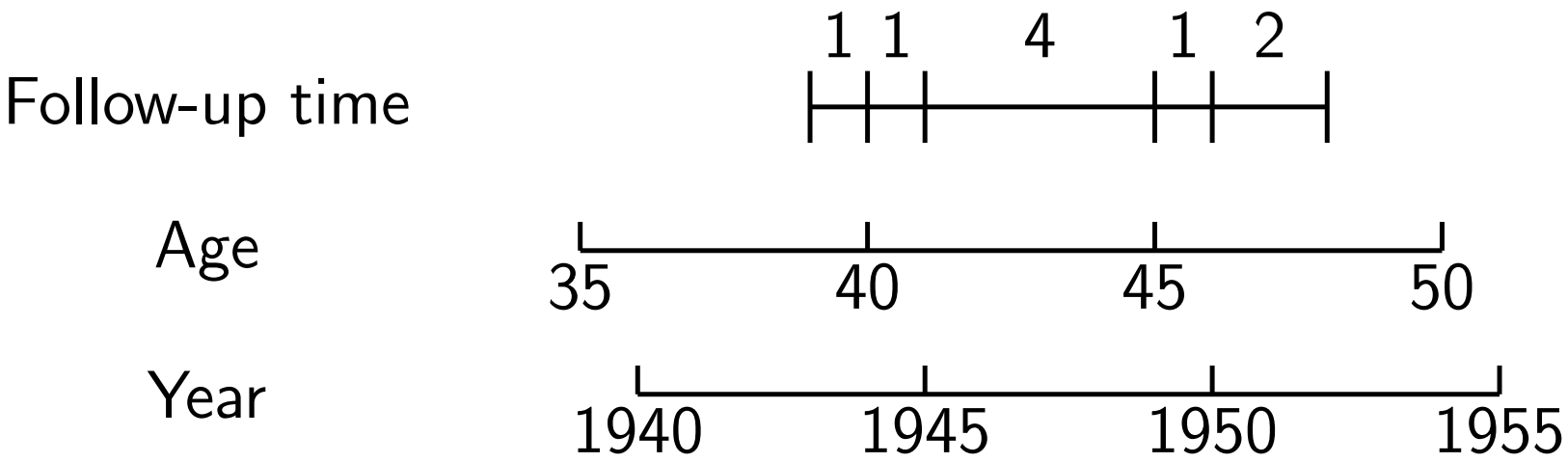
• exit because of disease onset (outcome of interest)

◦ exit due to other reason (censoring)



Person-years by age and period (C&H, Figure 6.4)

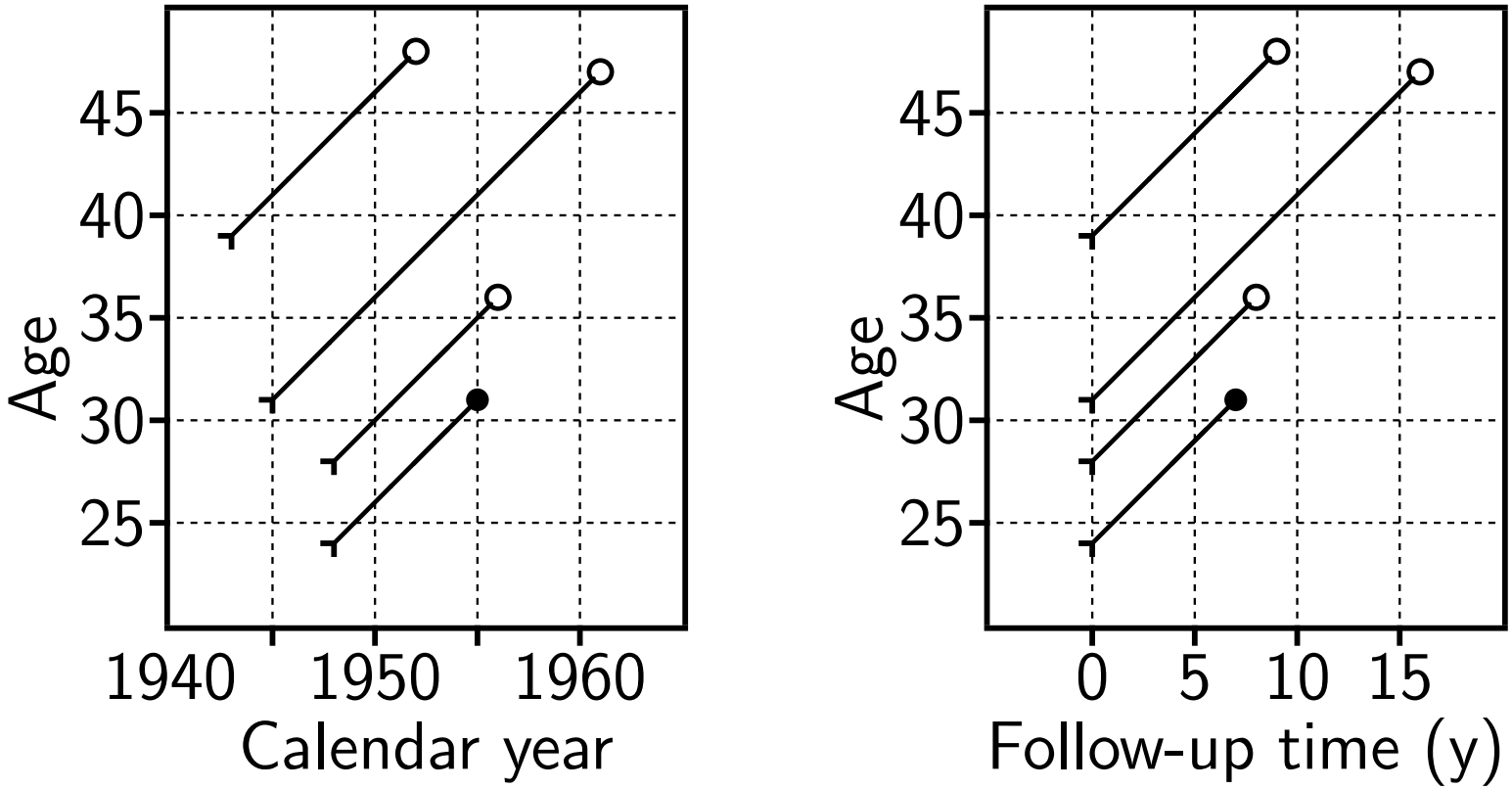
Subject 1: Follow-up jointly split by age and calendar time:



This subject contributes person-time into 5 different cells defined by ageband & calendar period

Follow-up in Lexis-diagrams

(C&H, pp. 58-59)



Follow-up lines run diagonally through different ages and calendar periods.

See also Laufey's lecture on cohort studies, slide 4.

STANDARDIZATION OF RATES

- Incidence of most cancers (and many other diseases) increases strongly by age in all populations.

  $\Rightarrow$ Most of the caseload comes from older age groups.

- **Crude incidence rate** $= \dfrac{\text{total no. of new cases}}{\text{total person-years}}$,
  - numerator = sum of age-specific numbers of cases,
  - denominator = sum of age-specific person-years.

- This is generally a poor summary measure.

- Comparisons of crude incidences between populations can be very misleading, when the age structures differ.

- **Adjustment** or **standardization** for age needed!

Ex. Male stomach cancer in Cali and Birmingham (IS, Table 4.2, p. 71)

	Cali			Birmingham			
Age (y)	Male cases 1982 -86	Male Population 1984 ($\times 10^3$)	Incid. Rate ($/10^5$y) 1982 -86	Male cases 1983 -86	Male Population 1985 ($\times 10^3$))	Incid. Rate ($/10^5$y) 1983 -86	Rate ratio
0–44	39	524.2	**1.5**	79	1 683.6	**1.2**	*1.25*
45-64	266	76.3	**69.7**	1037	581.5	**44.6**	*1.56*
65+	315	22.4	**281.3**	2352	291.1	**202.0**	*1.39*
Total	620	622.9	**19.9**	3468	2 556.2	**33.9**	*0.59*

- In each age group Cali has a higher incidence but the crude incidence is higher in Birmingham.
- **Is there a paradox?**

Comparison of age structures (IS, Tables 4.3,4.4)

	% of male population			
Age (years)	Cali 1984	B'ham 1985	Finland 2011	World Stand.
0–44	84	66	56	74
45–64	12	23	29	19
65+	4	11	15	7
All ages	100	100	100	100

The fraction of old men greater in Birmingham than in Cali.

$\Rightarrow$ Crude rates are **confounded** by age.

$\Rightarrow$ Any summary rate must be **adjusted for age**.

Adjustment by standardisation

Age-standardised incidence rate (ASR):

$$\text{ASR} = \sum_{k=1}^{K} \text{weight}_k \times \text{rate}_k \ / \ \text{sum of weights}$$

$=$ **Weighted average** of age-specific rates over the age-groups $k = 1, \ldots, K$.

- Weights describe the age distribution of some **standard population**.
- Standard population can be real (*e.g.* one of the populations under comparison, or their average) or fictitious (*e.g.* World Standard Population, WSP)
- Choice of standard population always more or less arbitrary.

Some standard populations:

Age group (years)	African	World	European	Nordic[a]
0–4	10 000	12 000	8 000	5 900
5–9	10 000	10 000	7 000	6 600
10–14	10 000	9 000	7 000	6 200
15–19	10 000	9 000	7 000	5 800
20–24	10 000	8 000	7 000	6 100
25–29	10 000	8 000	7 000	6 800
30–34	10 000	6 000	7 000	7 300
35–39	10 000	6 000	7 000	7 300
40–44	5 000	6 000	7 000	7 000
45–49	5 000	6 000	7 000	6 900
50–54	3 000	5 000	7 000	7 400
55–59	2 000	4 000	6 000	6 100
60–64	2 000	4 000	5 000	4 800
65–69	1 000	3 000	4 000	4 100
70–74	1 000	2 000	3 000	3 900
75–79	500	1 000	2 000	3 500
80–84	300	500	1 000	2 400
85+	200	500	1 000	1 900
Total	100 000	100 000	100 000	100 000

[a] NORDCAN population in 2000.

Stomach cancer in Cali & B'ham

Age-standardized rates by the World Standard Population:

	Cali			Birmingham		
Age	Rate^a	Weight		Rate^a	Weight	
0–44	1.5 ×	0.74 =	1.11	1.2 ×	0.74 =	0.89
45–64	69.7 ×	0.19 =	13.24	44.6 ×	0.19 =	8.47
65+	281.3 ×	0.07 =	19.69	202.0 ×	0.07 =	14.14
Age-standardised rate			**34.04**			**23.50**

- ASR in Cali higher – coherent with the age-specific rates.
- Summary rate ratio estimate: **standardized rate ratio**

$$\text{SRR} = 34.0/23.5 = 1.44.$$

- Known as **comparative mortality figure (CMF)** when the outcome is death (from cause C or all causes).

Cumulative rate and "cumulative risk"

- A neutral alternative to arbitrary standard population for age-adjustment is provided by **cumulative rate**:

$$\text{CumRate} = \sum_{k=1}^{K} \text{width}_k \times \text{rate}_k,$$

- Weights are now widths of the agebands to be included, usually up to 65 or 75 y with 5-y bands.
- NORDCAN & GLOBOCAN use a transformation:

$$\text{CumRisk} = 1 - \exp(-\text{CumRate}),$$

  calling it as the **cumulative risk** of getting the disease by given age, in the absence of competing causes.
- Yet, in reality competing events are present, so the probability interpretation of CumRisk is problematic.

Stomach cancer in Cali & B'ham

From age-specific rates of Table 4.2. the cumulative rates up to 65 years and their ratio are

$$\text{Cali:} \quad 45\ y \times \frac{1.5}{10^5 y} + 20\ y \times \frac{69.7}{10^5 y} = 0.0146 = \mathbf{1.46} \text{ per } 100$$

$$\text{B'ham:} \quad 45\ y \times \frac{1.2}{10^5 y} + 20\ y \times \frac{44.6}{10^5 y} = 0.0095 = \mathbf{0.95} \text{ per } 100$$

$$\text{ratio:} \quad 1.46/0.95 = \mathbf{1.54}$$

"Cumulative risks" & their ratio up to 65 y:

$$\text{Cali:} \quad 1 - \exp(-0.0146) = 0.0145 = \mathbf{1.45}\%$$

$$\text{B'ham:} \quad 1 - \exp(-0.0095) = 0.0094 = \mathbf{0.94}\%$$

$$\text{ratio:} \quad 1.45/0.94 = \mathbf{1.54}$$

NB. For more appropriate estimates of cumulative risks, correction for total mortality (competing event) needed.

Cumulative measures using 5-y groups

(IS, Fig 4.11, p. 77)

Age-group (years)	Incidence rate (per 100 000 pyrs)
0–4, . . . , 15–19	0.0
20–24, 25–29	0.1
30–34	0.9
35–39	3.5
40–44	6.7
45–49	14.5
50–54	26.8
55–59	52.6
60–64	87.2
65–69	141.7
70–74	190.8
Sum	524.9

$$\text{Cum. rate 0-75 y} = 5\text{ y} \times \frac{524.9}{10^5\text{ y}} = 0.0262 = \mathbf{2.6}\text{ per 100}$$

$$\text{“Cum. risk” 0-75 y} = 1 - \exp(-0.0262) = 0.0259 = \mathbf{2.6}\%.$$

Cumulative and life-time risks

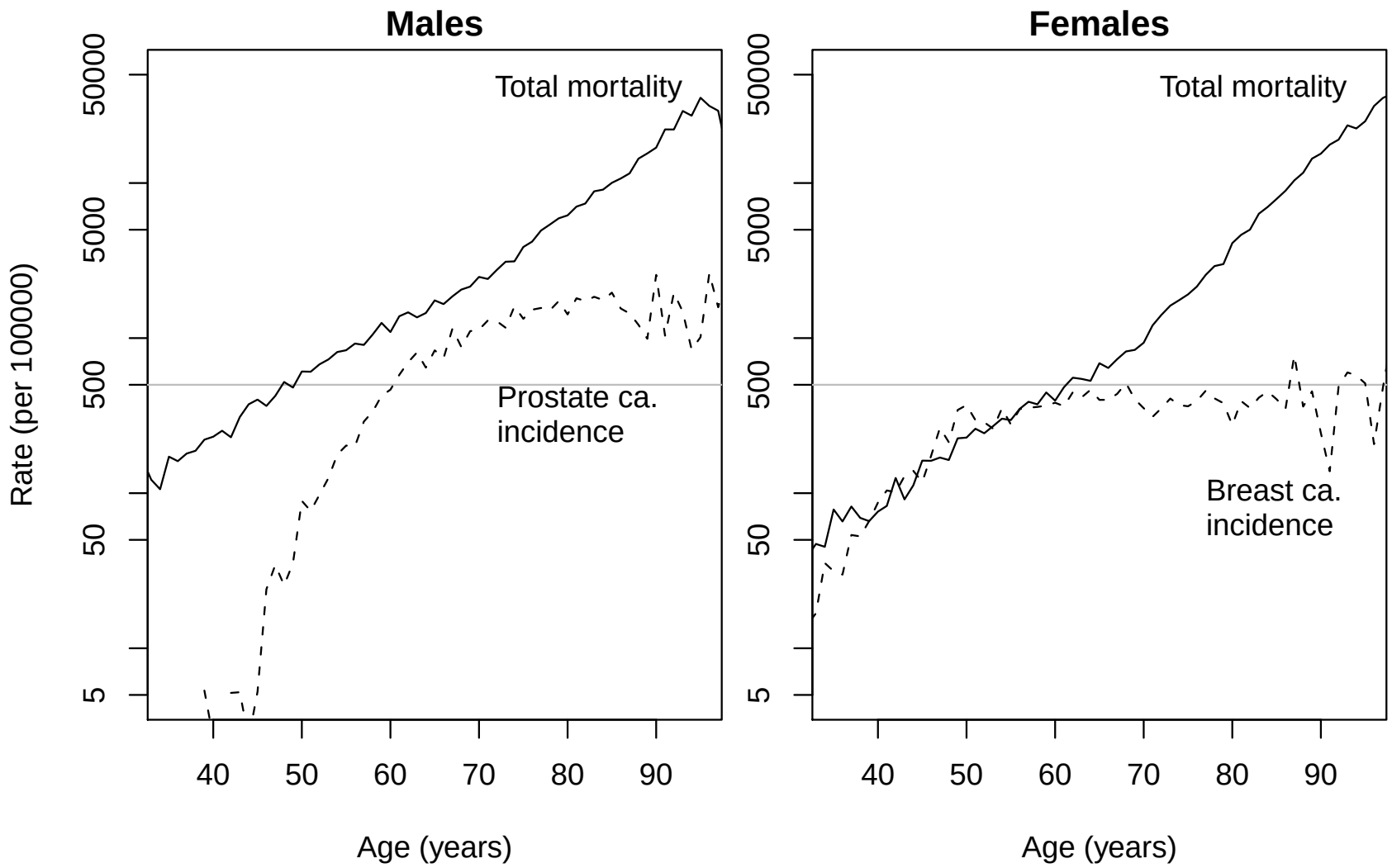
Of course, it is an interesting and relevant question to ask:

> "*What are my chances of getting cancer C, say, in the next 10 years, between ages 50 to 75 years, or during the whole lifetime?*"

However, this is difficult to answer.

- Fully individualized risks are unidentifiable.
- Age-specific and standardized rates are not very informative as such.
- Average cumulative risks are often estimated from cumulative rates using the simple formula above.
- Yet, these naive estimates fictitiously presume that a person would not die from any cause before cancer hits him/her, but could even survive forever!

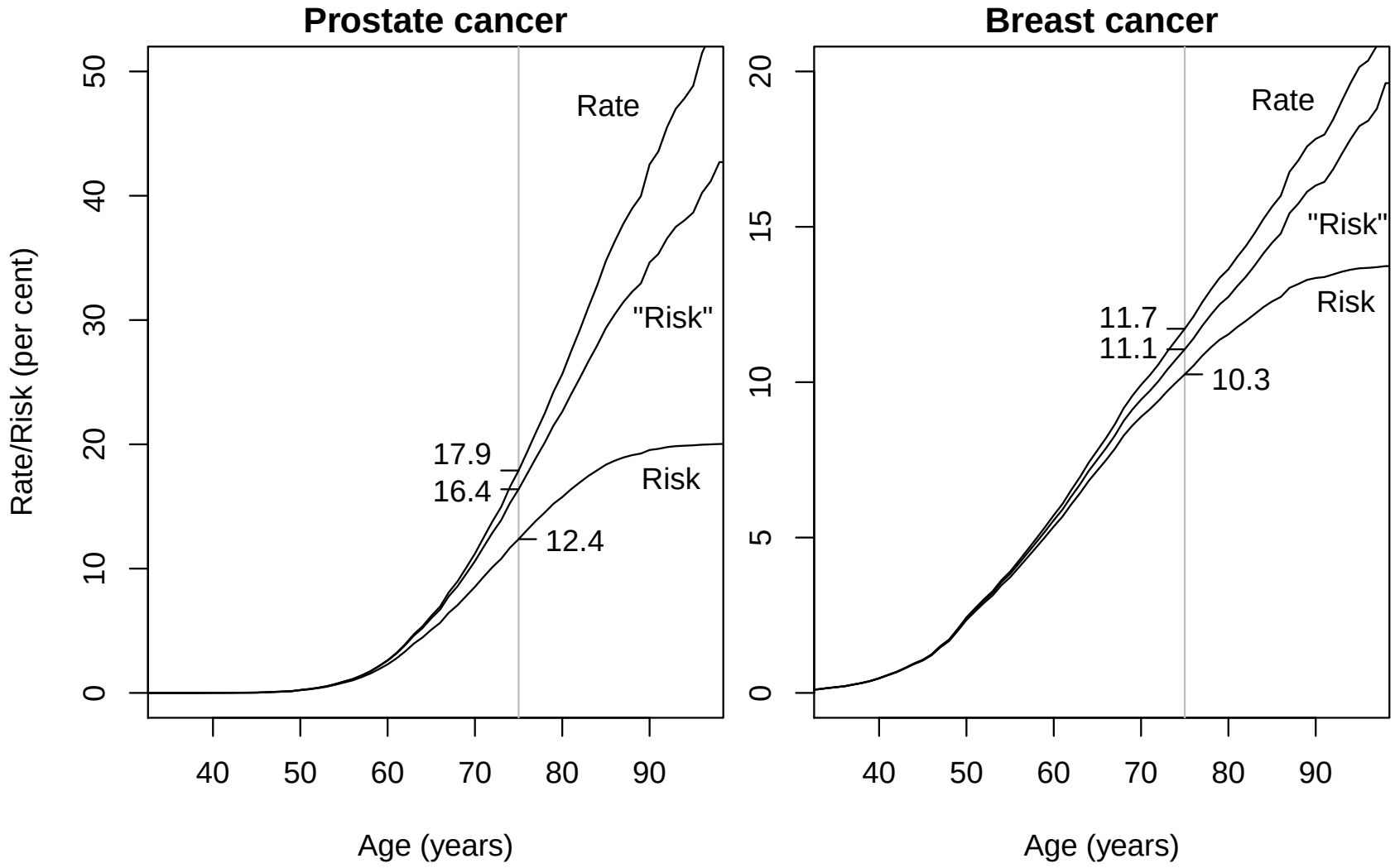
Total mortality and incidence of two common cancers by age, Finland 2005



Estimation of cumulative risks

- The probability of contracting cancer during realistic lifespan or in any age range depends not only on age-specific hazard rates of cancer itself but also of probabilities of overall survival up to relevant ages,
- Hence, the dependence of total mortality by age in the population at risk must be incorporated in the estimation of cumulative risks of cancer.
- When this is properly done, the corrected estimates of cumulative risk will always be lower than the uncorrected "risks".
- The magnitude of bias in the latter grows by age, but is reduced with increased life expectancy.

Cumulative measures, Finland 2005



Greater differences in males reflect shorter life expectancy and relatively high rates of prostate ca. in old ages.

Special cohorts of exposed subjects

- Occupational cohorts, exposed to potentially hazardous agents (*e.g.* rubber workers, see Laufey's lecture on cohort studies)
- Cohorts of patients on chronic medication, which may have harmful long-term side-effects
- No internal comparison group of unexposed subjects.

Question: Do incidence or mortality rates in the *exposed* target cohort differ from those of a roughly comparable **reference** population?

Reference rates obtained from:

- population statistics (mortality rates)
- disease & hospital discharge registers (incidence)

Observed and expected cases – SIR

- Compare rates in a study cohort with a standard set of age–specific rates from the reference population.
- Reference rates normally based on large numbers of cases, so they are assumed to be "known" without error.
- Calculate **expected** number of cases, E, if the standard age-specific rates had applied in our study cohort.
- Compare this with the **observed** number of cases, D, by the **standardized incidence ratio** SIR
  (or st'zed mortality ratio SMR with death as outcome)

$$\text{SIR} = D/E, \qquad \text{SE}(\log[\text{SIR}]) = 1/\sqrt{D}$$

Example: HT and breast ca.

- A cohort of 974 women treated with hormone (replacement) therapy were followed up.
- $D = 15$ incident cases of breast cancer were observed.
- Person-years (Y) and reference rates (λ_a^*, per 100000 y) by age group (a) were:

Age	Y	λ_a^*	E
40–44	975	113	1.10
45–49	1079	162	1.75
50–54	2161	151	3.26
55–59	2793	183	5.11
60–64	3096	179	5.54
$\sum$			16.77

Ex: HT and breast ca. (cont'd)

- "Expected" cases at ages 40–44:

$$975 \times \frac{113}{100\,000} = 1.10$$

- Total "expected" cases is $E = 16.77$
- SIR $= 15/16.77 = 0.89$.
- Error-factor: $\exp(1.96 \times \sqrt{1/15}) = 1.66$
- 95% confidence interval is:

$$0.89 \overset{\times}{\div} 1.66 = (0.54, 1.48)$$

SIR for Cali with B'ham as reference

Total person-years at risk and expected number of cases in Cali 1982-86 based on age-specific rates in Birmingham (**IS**: Fig. 4.9, p. 74)

Age	**Person-years**	**Expected cases in Cali**
0–44	524 220×5= 2 621 100	0.000012×2 621 100= 31.45
45–64	76 304×5= 381 520	0.000446× 381 520=170.15
65+	22 398×5= 111 990	0.002020× 111 990=226.00
All ages	**=3 114 610**	**Total expected (E) 427.82**

Total observed number $O = 620$.

Standardised incidence ratio:

$$\text{SIR} = \frac{O}{E} = \frac{620}{427.8} = 1.45 \quad (\text{or } 145 \text{ per } 100)$$

Crude and adjusted rates compared

(IS: Table 4.6, p. 78, extended)

	Cali, 1982-86	B'ham, 1983-86	Rate ratio
Crude rates ($/10^5$ y)	19.9	33.9	*0.59*
ASR ($/10^5$ y)^B with 3 broad age groups	48.0	33.9	*1.42*
ASR ($/10^5$ y)^C –"–	19.9	14.4	*1.38*
ASR ($/10^5$ y)^W –"–	34.0	23.5	*1.44*
Cum. rate $<$ 65 y (per 1000) –"–	14.6	9.5	*1.54*
ASR ($/10^5$ y)^W with 18 5-year age groups	36.3	21.2	*1.71*
Cum. rate $<$ 75 y (per 1000) –"–	46.0	26.0	*1.77*

Standard population: ^{B} Birmingham 1985, ^{C} Cali 1985, ^{W} World SP

NB: The ratios of age-adjusted rates appear less dependent on the choice of standard weights than on the coarseness of age grouping. 5-year age groups are preferred.

SURVIVAL ANALYSIS

Questions of interest on the **prognosis** of cancer:

- what are the patients' chances to **survive** at least 1 year, or 5 years *etc.*, since diagnosis?

Survival analysis: In principle like incidence analysis but

- population at risk = patients with cancer,
- basic time variable = time since the date of diagnosis, on which the follow-up starts,
- outcome event of interest = death,
- measures and methods used somewhat different from those used in incidence analysis.

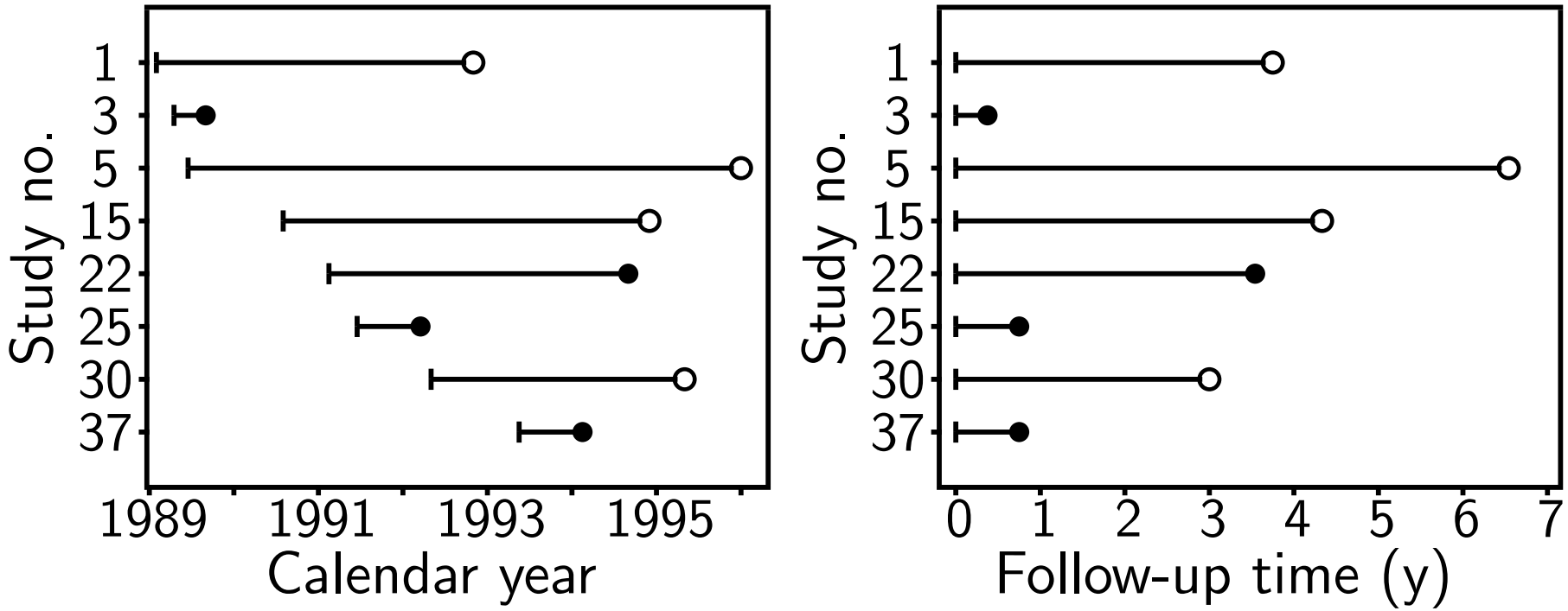
Follow-up of 8 out of 40 breast cancer patients (from IS, table 12.1., p. 264)

No.	Age (y)	Stage[a]	Date of diagnosis	Date at end of follow-up	Vital status at end of follow-up	Cause of death[c]	Full years from diagn's up to end of follow-up	Days from diagn's up to end of follow-up
1	39	1	01/02/89	23/10/92	A	–	3	1360
3	56	2	16/04/89	05/09/89	D	BC	0	142
5	62	2	12/06/89	28/12/95	A	–	6	2390
15	60	2	03/08/90	27/11/94	A	–	4	1577
22	64	2	17/02/91	06/09/94	D	O	3	1297
25	42	2	20/06/91	15/03/92	D	BC	0	269
30	77	1	05/05/92	10/05/95	A	–	3	1100
37	45	1	11/05/93	07/02/94	D	BC	0	272

[a] 1 = absence of regional lymph node involment and metastases
2 = involvment of regional lymph node and/or presence of metastases
[b] A = alive; D = dead; [c] BC = breast cancer; O = other causes

Follow-up of breast ca. pts (cont'd)

| entry = diagnosis; • exit = death; ∘ exit = censoring



(**IS**: Figure 12.1, p. 265)

Life table or actuarial method

Commonly used in population-based survival analysis by cancer registries. (In clinical applications the Kaplan-Meier method is more popular.)

(1) Divide the follow-up time into subintervals $k = 1, \ldots K$; most of these having width of 1 year.

Often the first year is divided into two intervals with widths of 3 mo and 9 mo, respectively.

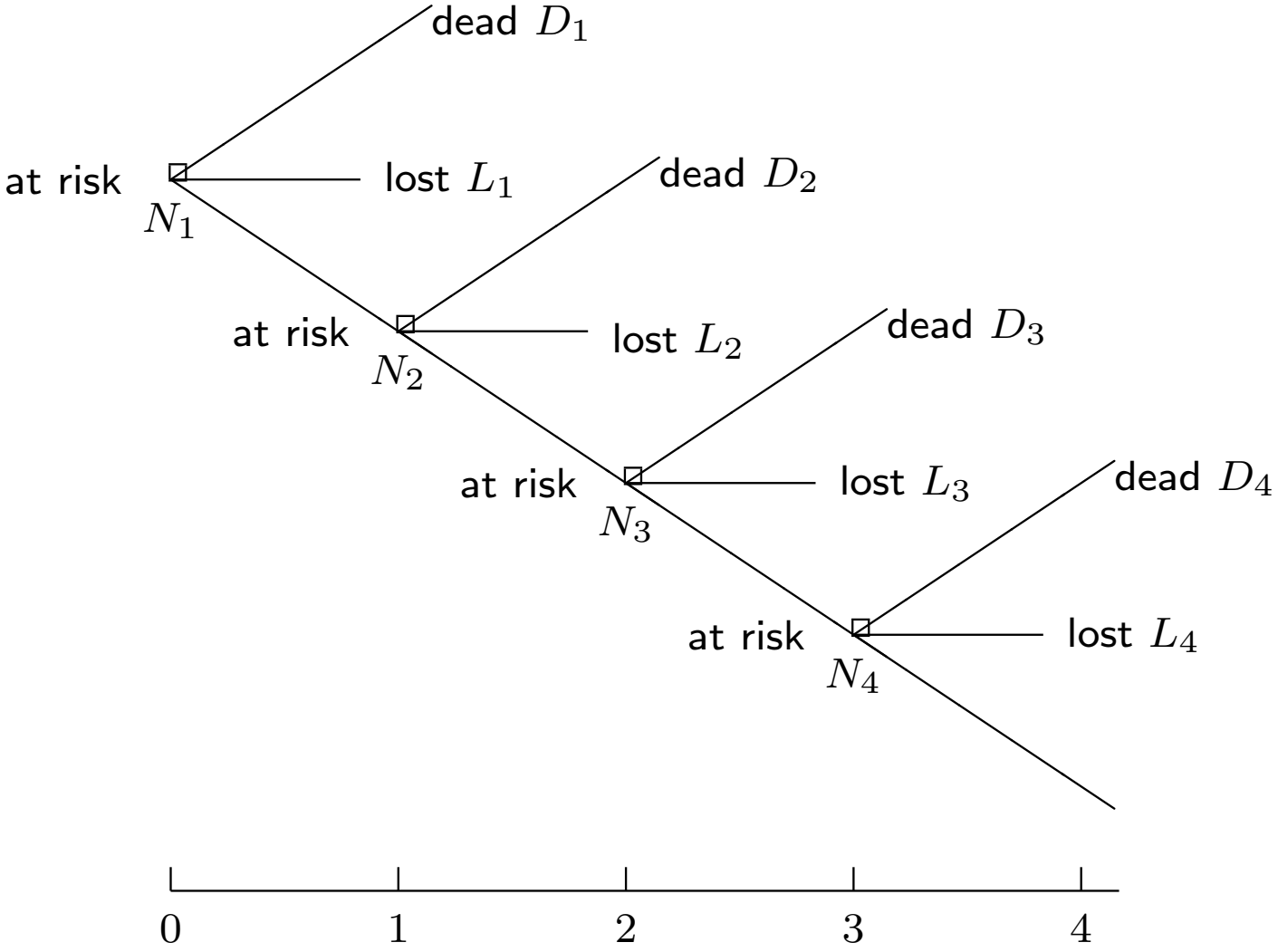
(2) Tabulate from original data for each interval

$N_k =$ size of the **risk set**, *i.e.* the no. of subjects still alive and under follow-up at the start of interval,

$D_k =$ no. of **cases**, *i.e.* deaths observed in the interval,

$L_k =$ no. of **losses**, *i.e.* individuals **censored** during the interval before being observed to die.

Life table items in a tree diagram



Follow-up time (years), divided into 1-y subintervals

N_k = population at risk at the start of the kth subinterval

D_k = no. of deaths, L_k = no. of losses or censorings in interval k

Life table items for breast ca. patients

(**IS**: Table 12.2., p. 273, first 4 columns)

Interval (k)	Years since diagnosis	No. at start of interval (N_k)	No. of deaths (D_k)	No. of losses (L_k)
1	0– < 1	40	7	0
2	1– < 2	33	3	6
3	2– < 3	24	4	3
4	3– < 4	17	4	4
5	4– < 5	9	2	3
6	5– < 6	4	1	2
7	6– < 7	1	0	1
Total			21	19

Life table calculations (cont'd)

(3) Calculate and tabulate for each interval

$N'_k =$ $N_k - L_k/2 =$ corrected size of the risk set, or "effective denominator" at start of the interval,

$q_k =$ $D_k/N'_k =$ estimated conditional probability of dying during the interval given survival up to its start,

$p_k =$ $1 - q_k =$ conditional survival proportion over the int'l,

$S_k =$ $p_1 \times \cdots \times p_k =$ **cumulative survival proportion** from date of diagnosis until the end of the kth interval

$=$ estimate of **survival probability** up to this time point.

Follow-up of breast ca. patients (cont'd)

Actuarial life table completed (IS, table 12.2, p. 273)

Interval (k)	Years since diagnosis	No. at start of interval (N_k)	No. of deaths (D_k)	No. of losses (L_k)	Effective denominator (N'_k)	Cond'l prop'n of deaths during int'l (q_k)	Survival prop'n over int'l (p_k)	Cumul. survival; est'd survival prob'ty (S_k)
1	0– < 1	40	7	0	40.0	0.175	0.825	0.825
2	1– < 2	33	3	6	30.0	0.100	0.900	0.743
3	2– < 3	24	4	3	22.5	0.178	0.822	0.610
4	3– < 4	17	4	4	15.0	0.267	0.733	0.447
5	4– < 5	9	2	3	7.5	0.267	0.733	0.328
6	5– < 6	4	1	2	3.0	0.333	0.667	0.219
7	6– < 7	1	0	1	0.5	0.0	1.0	0.219

1-year survival probability is thus estimated 82.5% and 5-year probability 32.8%.

Comparison to previous methods

- Complement of survival proportion $Q_k = 1 - S_k$
  $=$ incidence proportion of deaths.
  Estimates the cumulative risk of death from the start of follow-up till the end of kth interval.
- Indidence rate in the kth interval is computed as:

$$I_k = \frac{\text{number of cases } (D_k)}{\text{approximate person-time } (\widetilde{Y}_k)}$$

where the approximate person-time is given by

$$\widetilde{Y}_k = \left[N_k - \frac{1}{2}(D_k + L_k) \right] \times \text{width of interval}$$

The dead and censored thus contribute half of the interval width.

Survival curve and other measures

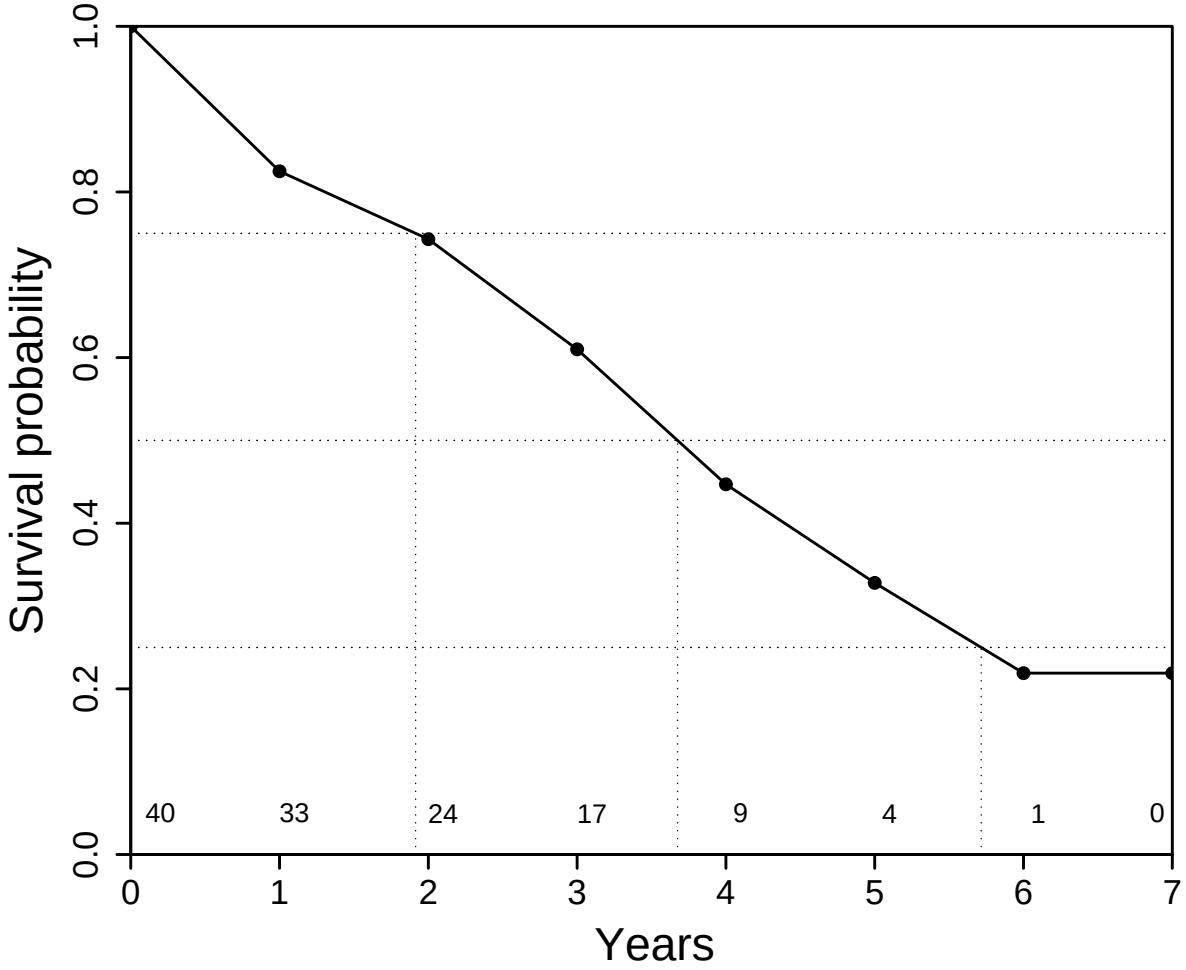
Line diagram of survival proportions through interval endpoints provides graphical estimates of interesting parameters of the survival time distribution, *e.g.*:

- **median** and **quartiles**: time points at which the curve crosses the 50%, 75%, and 25% levels
- **mean residual lifetime**: area under the curve, given that it decreases all the way down to the 0% level.

NB. Often the curve ends at higher level than 0%, in which case some measures cannot be calculated.

Survical curve of breast ca. patients

(IS: Fig 12.8)



Numbers above x-axis show the size of population at risk.

Relative survival analysis

- Another interesting and relevant question:

  "*How much worse are the chances of a cancer patient to survive, say, 5 years, as compared with a comparable person without the disease?*"

- An answer is provided by **relative survival proportions**:

$$R_k = S_k^{\text{obs}} / S_k^{\text{exp}}, \qquad \text{where}$$

- S_k^{obs} = **observed** survival proportion in cancer patient group k by age, gender and year of diagnosis,
- S_k^{exp} = **expected** survival proportion based on the age-specific mortality rates of the same gender and calendar time in a reference population (*cf.* SIR!)

\+ No information on causes of death needed.

CONCLUSION

Measuring and comparing disease frequencies

- not a trivial task but
- demands expert skills in epidemiologic methods.

Major challenges:

- obtain the right denominator for each numerator,
- valid calculation of person-years,
- appropriate treatment of time and its various aspects,
- removal of confounding from comparisons.

APPENDIX: Introduction to R

What is R?

- A practical calculator:
  - You can see what you compute
  - . . . and change easily to do similar calculations.
- A statistical program.
- An environment for data analysis and graphics.
- A programming language
- Developed by international community of volunteers.
- Free.
- Runs on any computer.
- Updated every 6 months.

What does R offer for epidemiologists?

- Descriptive tools
  - Versatile tabulation
  - High-quality graphics
- Analytic methods
  - Basic epidemiologic statistics
  - Survival analysis methods
  - Common regression models and their extensions
  - Other...

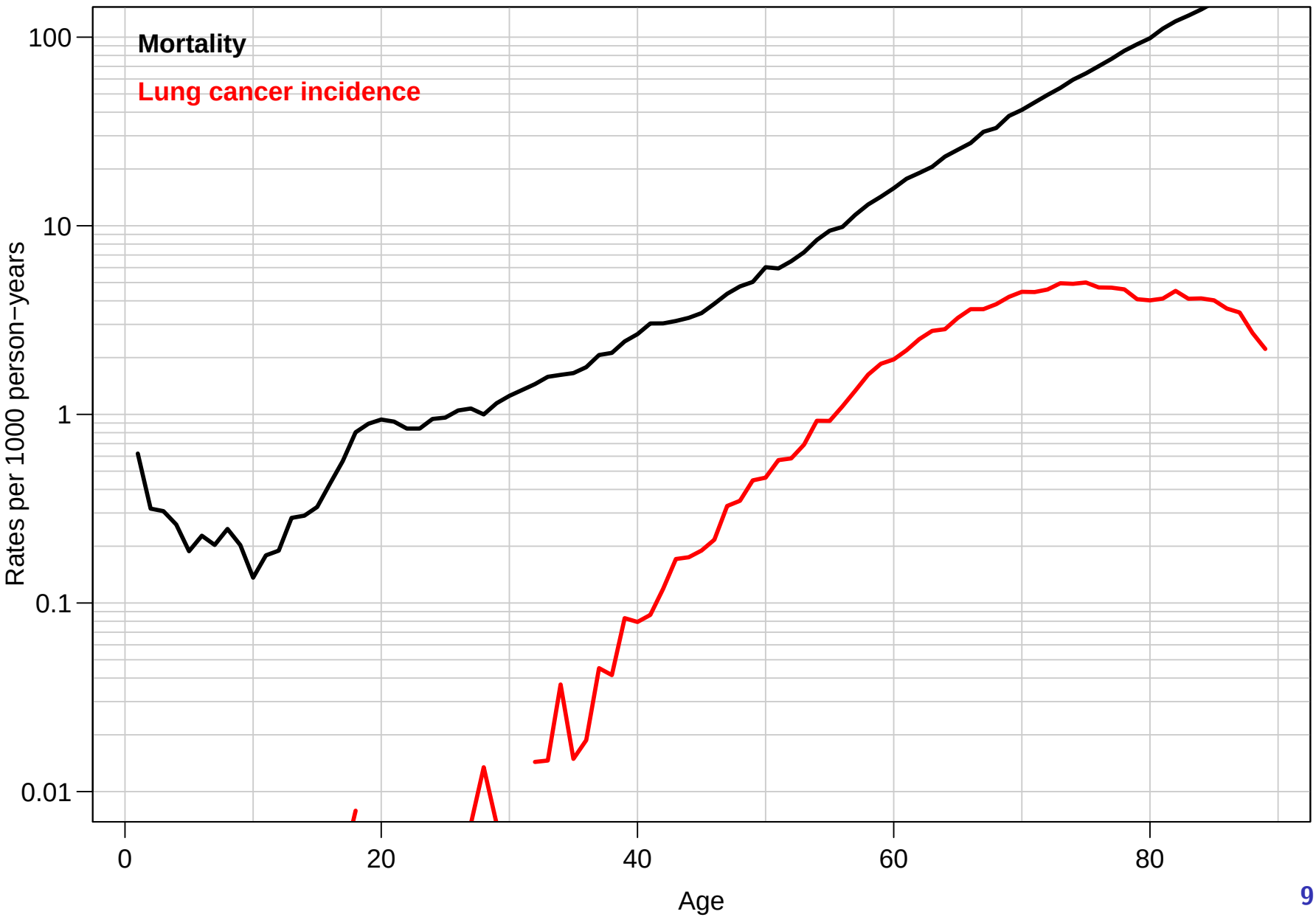
These are provided by *e.g.* SPSS, SAS and Stata, too, so ...?

Many features of R are more appealing in the long run.

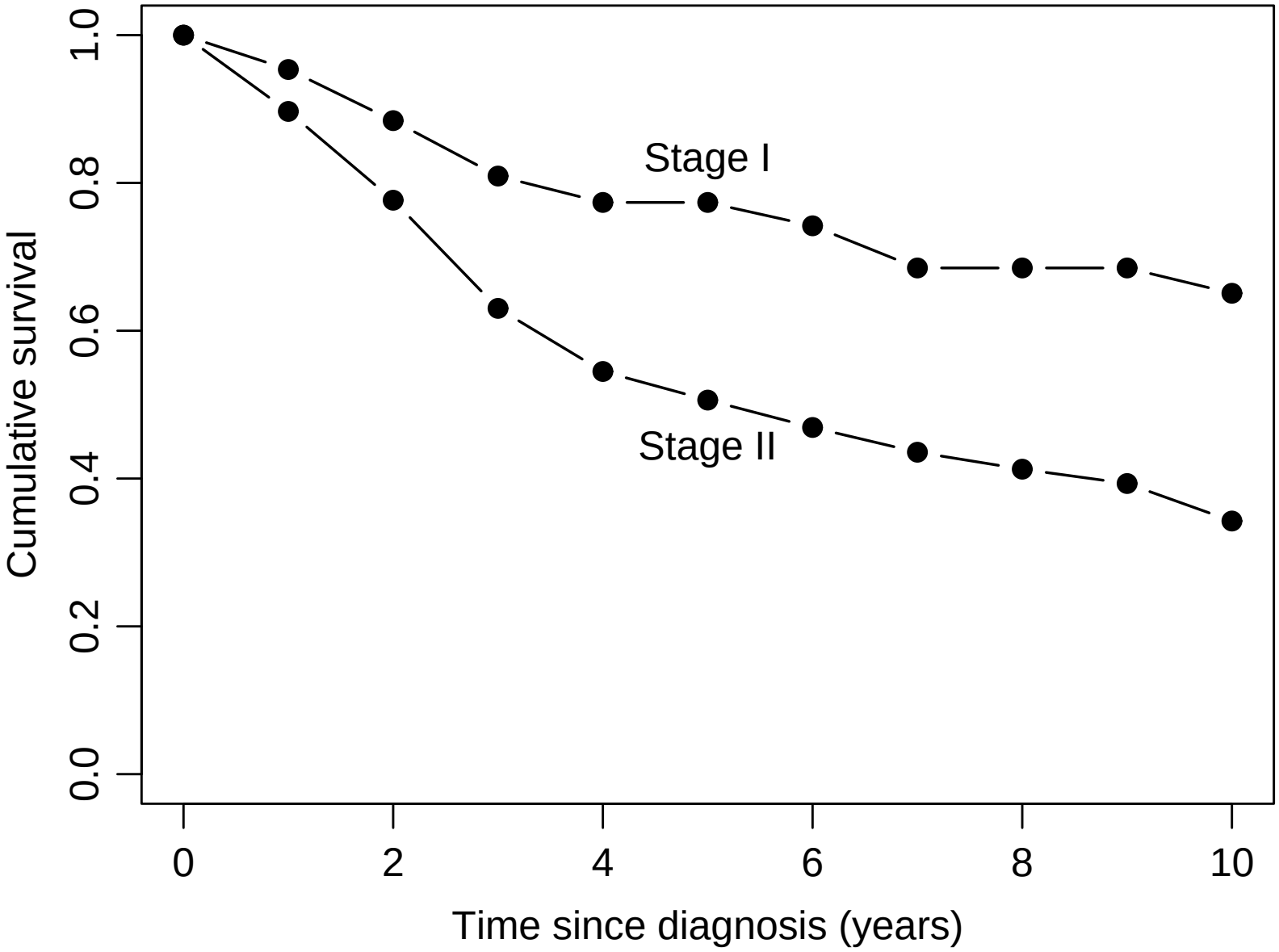
Graphics in R

- Versatile, flexible, high quality, ...
- Easy to add items (points, lines, text, legends ...) to an existing graph.
- Fine tuning of symbols, lines, axes, colours, etc. by *graphical parameters* ($>$ 67 of them!)
- Interactive tools using the mouse
  - Put new things on a graph
  - Identify points

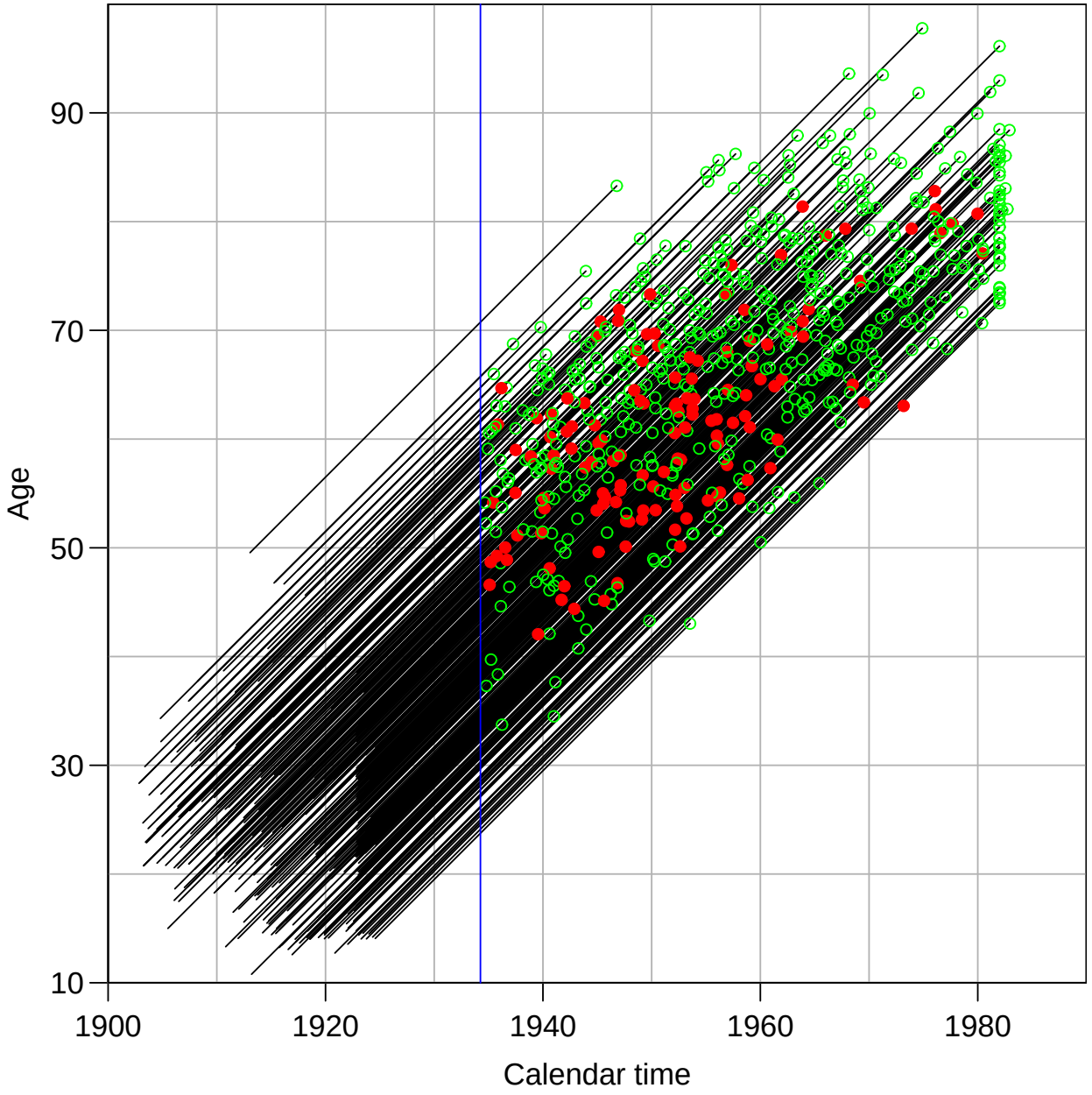
Total mortality and lung ca incidence in DK



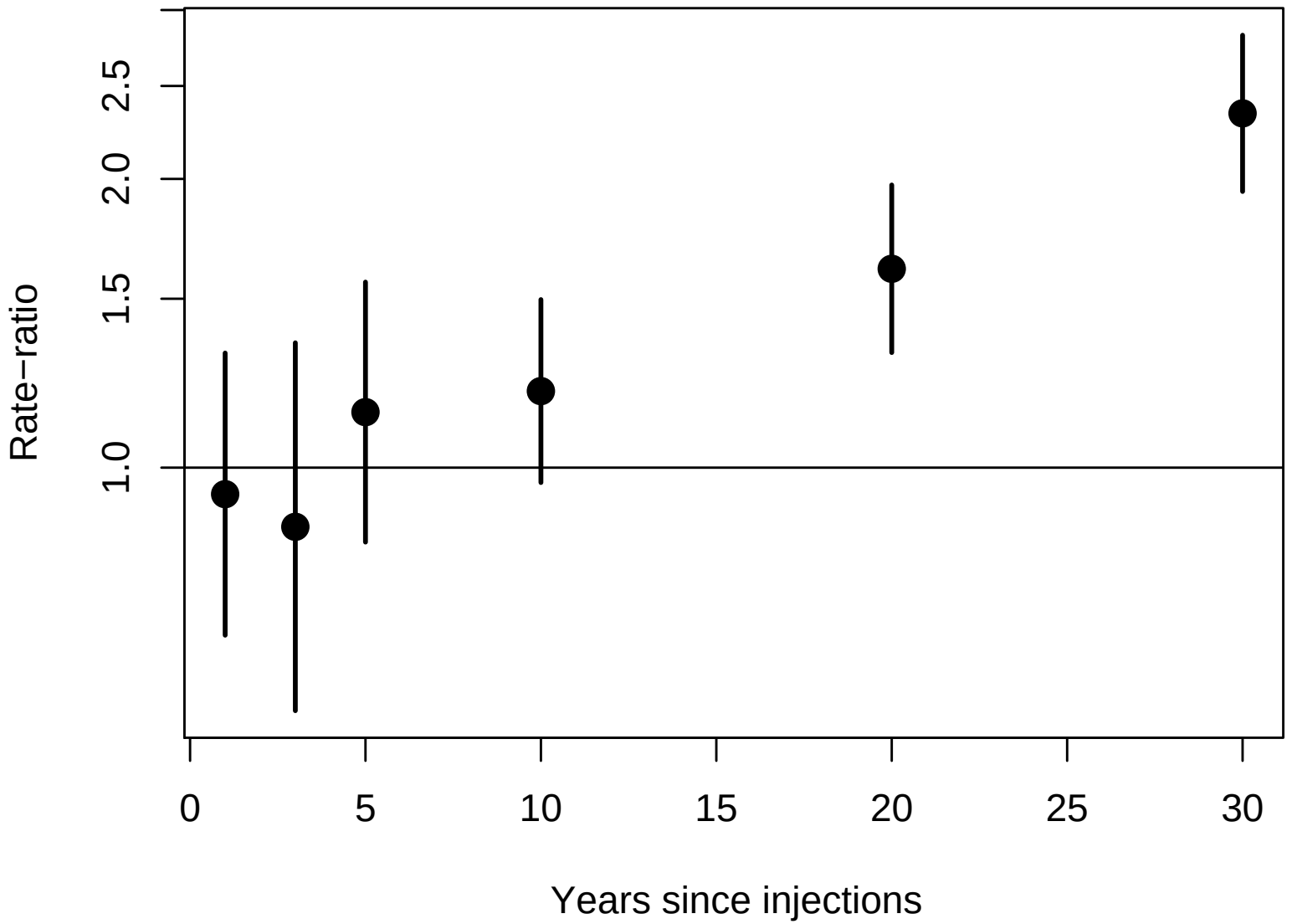
Survival of cervix ca patients (C&H, 34)



Lexis diagram of Welsh nickel cohort



Rate ratios with confidence intervals



Getting your graphs out

- Graphs can be saved to disk in almost any format

  `.eps`, `.pdf`, `.bmp`, `.jpg`, `.png`, ...

- Save graphs from the screen or write directly to a file.
- You can also directly transport an R graph as a metafile into a Word document!

Tools for nearly anything!

- Thousands of add-on packages.
- Several packages for epidemiological analyses:
  - `Epi`: focus on chronic disease epidemiology:
    - Cohort studies, splitting follow-up time
    - Lexis diagram, several timescales
    - Multistate model support
    - Advanced tabulation
    - Informative reporting of estimation results
  - `epicalc`:
  - `epitools`: Mostly infectious diseases.
  - `epiR`: Leaning towards veterinary epidemiology.
- Packages may be installed and updated from within R.

Running R

- Interactive but not mouse-driven!
- Commands typed from keyboard.
- More practical: commands written and saved in a **script file** from which they are run.
- Execution of tasks:
  - evaluation of **expressions** contained in commands,
  - based on calls of **functions**.

Difficult to learn & slow to use?

- Maybe in the beginning.
- Versatility and flexibility rewarding in the long run.

Running R on Windows

- Start by double-clicking the R-icon.
- `R Console`: the **console window**
  - command lines to be typed – or pasted from a script file – after prompt '>',
  - prompt '+' marks continuation of an incomplete command line,
  - output follows a completed command requesting it,
  - arrow key ↑ leads to previous command lines.
- Menu bar for a few useful pull-down menus.
- On-line help in HTML form.

R as a simple calculator

Write the arithmetic expression on the empty line after the prompt and press `Enter`. The result is displayed immediately.

```
> 2+2
[1] 4
> 3*5 - 6/2
[1] 12
> (2+3)^2
[1] 25
> sqrt( 1/12 + 1/17 )
[1] 0.3770370
> exp( 1.96 * sqrt( 1/12 + 1/17 ) )
[1] 2.093825
```

R as a smart calculator

Simple summary of results from a cohort study:

	Exposed	Unexposed
No. of cases/Person-years	20/2000	25/5000

- Numbers of cases and person-years are first assigned & saved into vectors `D` and `Y`;
- Incidence rates in the two groups as well as their ratio and difference are then calculated and printed:

```
> D <- c(20, 25) ; Y <- c(2000, 5000)
> rate <- 1000*D/Y ; rate
[1] 10  5
> ratio <- rate[1]/rate[2] ; diff <- rate[1]-rate[2]
> c(ratio, diff)
[1] 2 5
```

A couple of important things

- Names of **variables** (or any other **objects**)
  - Start with a letter from `A,...,Z` or `a,...,z`; lower case separated from upper case, *e.g.* 'x' $\neq$ 'X'
  - Letters, integers `0, ..., 9`, dots '.', and underlines '_' allowed after 1st letter.
- **Assignment operator** '<-' (consists of '<' and '-')
  - assigns a value to an object, for example

    ```
    > A <- 5+2 ; A
    [1] 7
    ```

    means that a numeric variable 'A' is given $5+2 = 7$ as its value, and is then printed,
  - the equal sign '=' is also allowed as assignment operator.

Vectors and their arithmetics

Vector = ordered set of numbers (or other similar elements)

- Can be assigned values elementwise by function `c()`
- Vector `x` with 4 elements 1, 2, 4, 7 assigned and printed:

```
> x <- c(1,2,4,7)
> x
[1] 1 2 4 7
```

- Arithmetic operations `+`, `-`, `*`, `/`, `^` (power) for vectors of same **length** *i.e.* same number of elements.

$\Rightarrow$ Outcome: a new vector whose elements are results of the operation on the corresponding elements in original vectors.

- Common mathematical functions, like `sqrt()`, `log()`, `exp()` work in the same way for numeric vectors.

R script – commands in a file

R script file is an ASCII file containing a sequence of R commands to be executed.

The **script editor** of R works as follows:

1. In `RGui` open the script editor window: *File - New script*, or when editing an existing script file: *File - Open script*,
2. Write the command lines without prompt '>' or '+'.
3. Save the script file: *File - Save e.g.* as `c:\...\mycmds.R` or with some other file name having extension `.R`

R script (cont'd)

4. Paint the lines to be excecuted and paste them on the console window using the third icon on the toolbar.
5. Edit the file using *Edit* menu, save & continue.

- To run a whole script file, write in console window:
  ```
  > source("c:/.../mycmds.R", echo=TRUE)
  ```
- The script can also be written and edited by any external editor programs (like Notepad).
- Of these, *Tinn-R* provides nice facilities for editing, checking and running R scripts, see `http://www.sciviews.org/Tinn-R/`.
- *R Studio* – very versatile interface; see `http://www.rstudio.com/`.

R in this course

- The main purpose is to inform you about the existence and potential of R, which you might find useful in any future work involving serious epidemiologic data analysis.
- Here, R will be used only as a simple calculator.
- No need for a lot of the more fancy stuff.
- The script editor will help you keep your solutions for future reference.
- After the course, solutions to all exercises will be provided.
- A good workbook introduction to R:
  `http://bendixcarstensen.com/Epi/R-intro.pdf`

Nordic Summerschool of Cancer Epidemiology

Bendix Carstensen Steno Diabetes Center
Gentofte, Denmark
http://BendixCarstensen.com
Esa Läärä University of Oulu
Oulu, Finland

Danish Cancer Society,
August 2017 / Januay 2018

http://BendixCarstensen.com/NSCE/2017

Chance

Bendix Carstensen & Esa Laara

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,
August 2017 / Januay 2018

http://BendixCarstensen.com/NSCE/2017

Chance variation

- Systematic and random variation
- Probability model:
  - random variable — observation — data
  - distribution
  - parameters
- Statistic
- Standard error

Systematic and random variation

Cancer incidence rates vary by known & measured determinants of disease, such as:

- age,
- gender,
- region,
- time,
- specific risk factors.

This is **systematic variation**.

Systematic and random variation

In addition, observed rates are subject to **random** or **chance variation**:
— variation due to unknown sources like

- latent genetic differences,
- unknown concomitant exposures,
- sampling,
- "pure chance" — quantum mechanics

Example: Smoking and lung cancer

- Only a minority of smokers get lung cancer
- . . . and some non-smokers get the disease, too.
- At the **individual** level the outcome is unpredictable.
- When cancer occurs, it can eventually only be explained just by "bad luck".
- Unpredictability of individual outcomes implies largely unpredictable — **random** — variation of disease rates at population level.

Example: Breast cancer

Breast cancer incidence rates in Finland, age group 65-69 years in three successive years.

Year	Males (per 10^6 P-years)	Females (per 10^4 P-years)
1989	46	21
1990	11	20
1991	33	19

- Big annual changes in risk among males?
- Is there steady decline in females?

Example: Breast cancer

Look at observed numbers of cases!

	Males		Females	
Year	Cases	P-years	Cases	P-years
1989	4	88,000	275	131,000
1990	1	89,000	264	132,000
1991	3	90,000	253	133,000

Reality of changes over the years?

The information is in the **number** of cases

Simple probability model for cancer occurrence

Assume that the population is **homogeneous**

- the theoretical incidence rate
- **hazard** or **intensity** — λ
- of contracting cancer
- is **constant** over a short period of time, dt

$$\lambda = \Pr\{\text{Cancer in}(t, t + \mathrm{d}t)\}/\mathrm{d}t$$

Simple probability model for cancer occurrence

- The observations:
  - Number of cases D in
  - Y person-years at risk
  - $\Rightarrow$ empirical incidence rate $R = D/Y$
- are all **random variables** with unpredictable values
- The **probability distribution** of possible values of a random variable has some known mathematical form
- . . . some properties of the probability distribution are determined by the **assumptions**
- . . . other properties are determined by quantities called **parameters**
- — in this case the theoretical rate λ.

How a probability model works

If the hazard of lung cancer, λ, is constant over time, we can **simulate** lung cancer occurrence in a population:

- Start with N persons
- 1st day: $\mathrm{P}\{\text{lung cancer}\} = \lambda \times 1$ day for all N
- 2nd day: $\mathrm{P}\{\text{lung cancer}\} = \lambda \times 1$ day for those left w/o LC
- 3rd day: $\mathrm{P}\{\text{lung cancer}\} = \lambda \times 1$ day for those left w/o LC
- ...

Thus a **probability model** shows how to generate data with known parameters. Model $\rightarrow$ Data

Component of a probability model

- **structure** of the model
  — *a priori* assumptions:
  — constant incidence rate
- `parameters` of the model
  — *size* of the incidence rate:
  — derived from data **conditional** on structure

Statistics

The opposite of a probability models:

- the **data** is known
- want to find **parameters**
- this is called estimation
- . . . mostly using maximum likelihood

Thus **statistical modelling** is how to estimate parameters from observed data. Data $\rightarrow$ Model

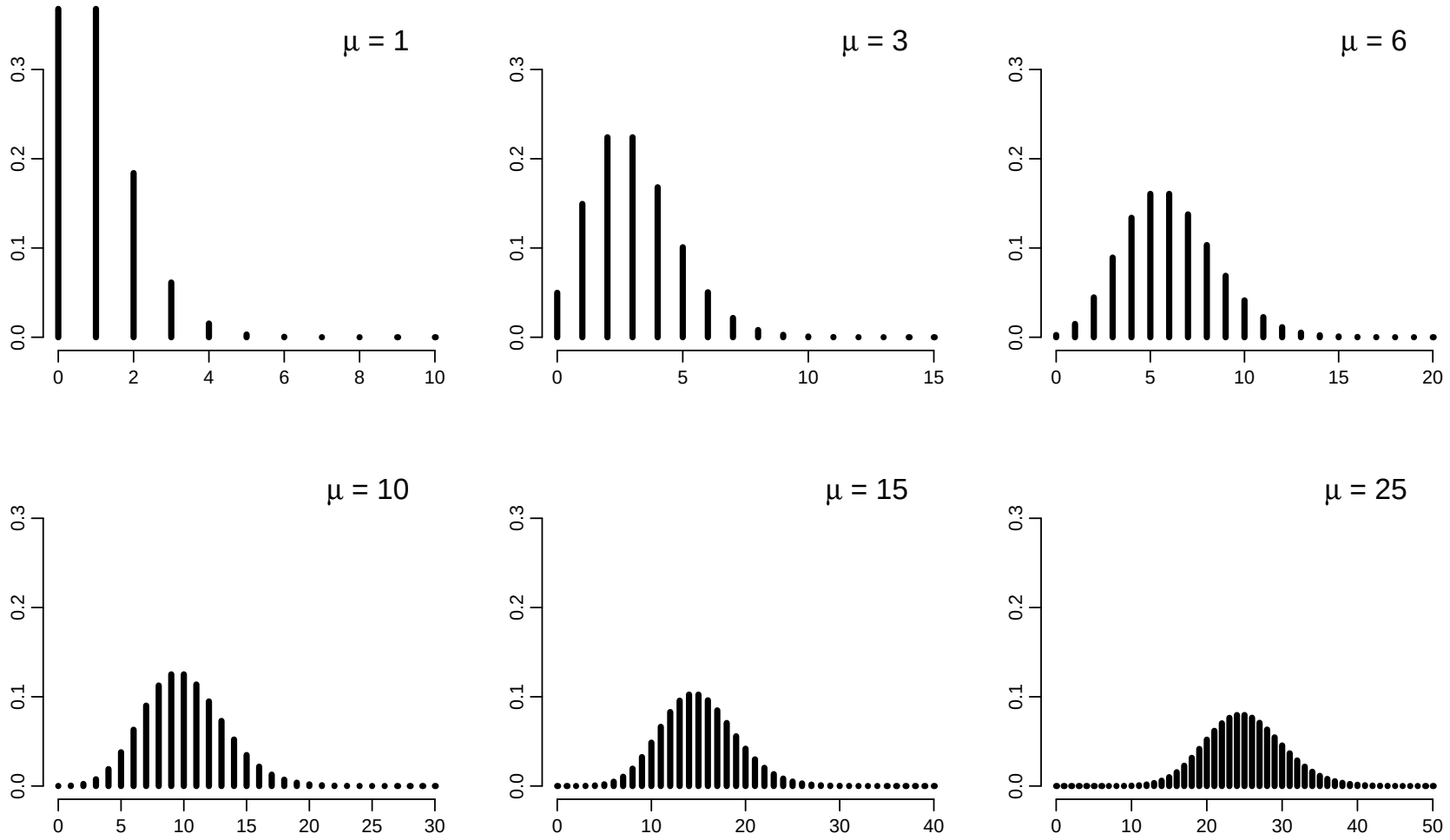
Statistics — the workings

- Fix the **model** (structure)
- For any set of parameters we can generate data
- Find parameters that generates data that look most like the observed data
- Recall the notion of **random variables**:
  - Given model and parameter
  - we know the distribution of **functions of** data
- Essential distributions are **Poisson** and **Normal** (Gaussian) distributions

Poisson and Gaussian models

- **Poisson distribution**: simple probability model for number of cases D (in a fixed follow-up time, Y) with
- **expectation** (theoretical mean) $\mu = \lambda Y$,
- **standard deviation** $\sqrt{\mu}$
- When the expectation μ of D is large enough, the Poisson distribution resembles more and more the **Gaussian** or **Normal** distribution.

Poisson distribution with different means (μ)

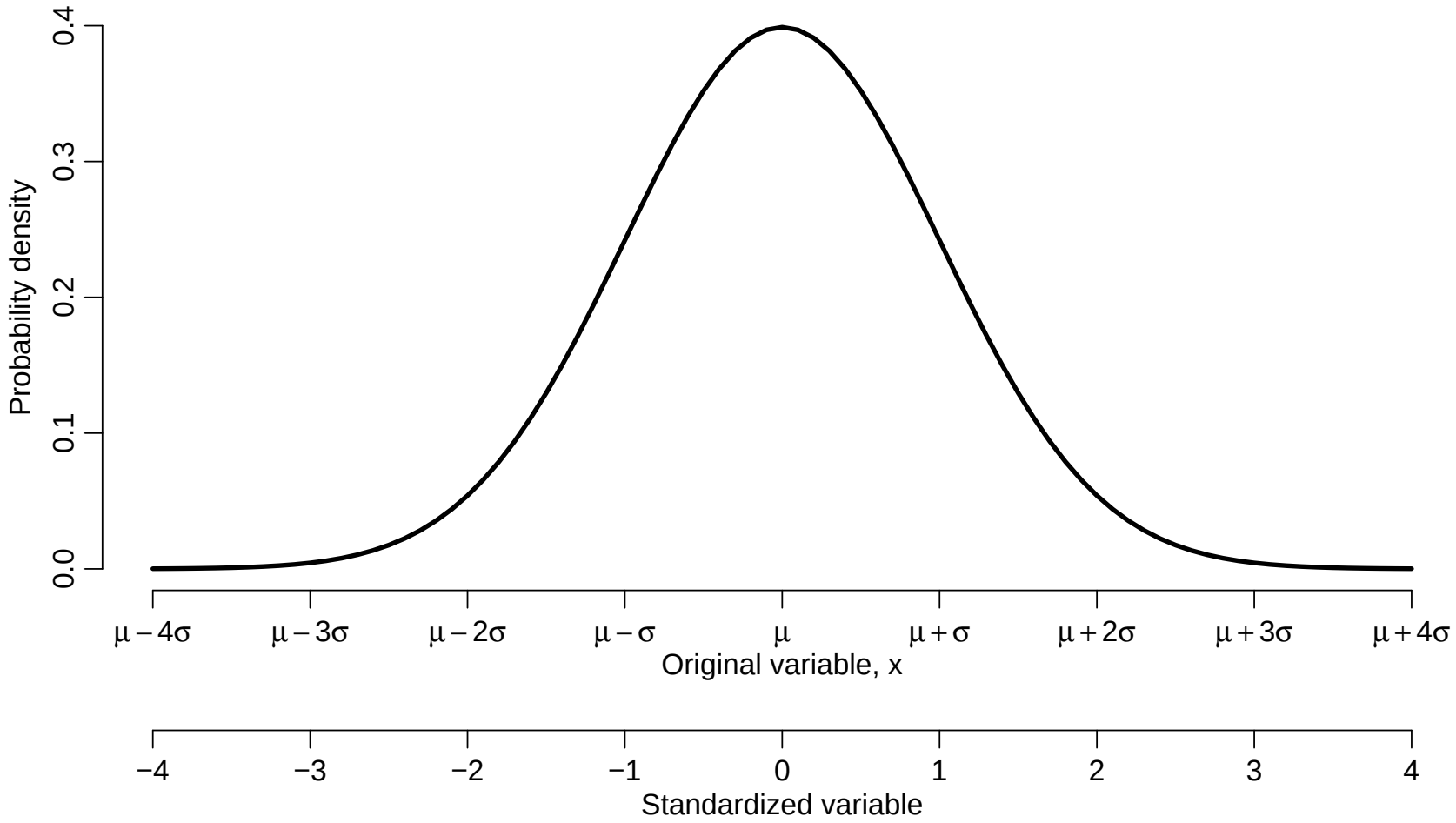


Gaussian distribution

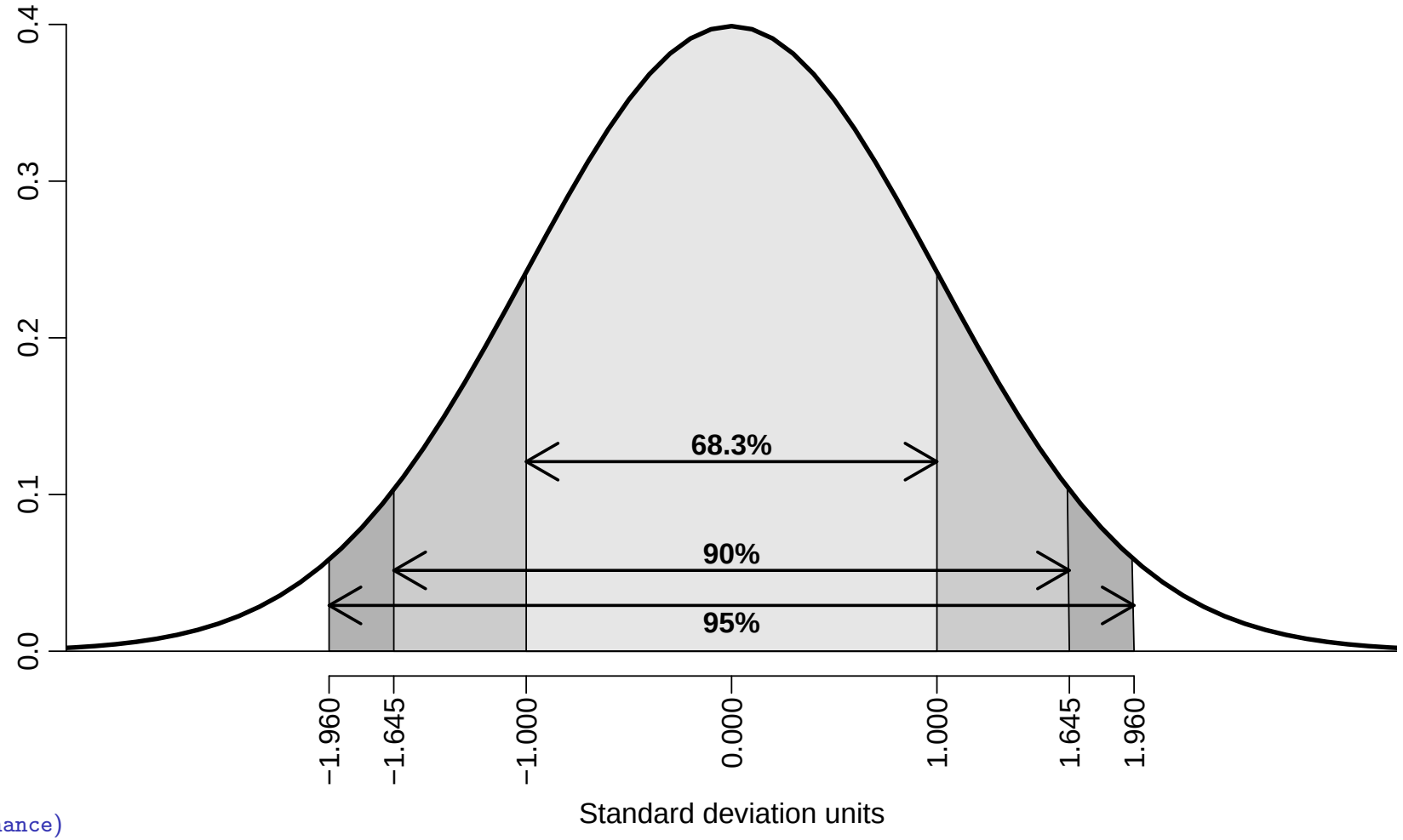
Gaussian or Normal distribution:

- common model for continuous variables,
  - symmetric and bell-shaped,
  - has two parameters:
    - μ = expectation or mean,
    - σ = standard deviation.
- Approximates **sampling distribution** of empirical measures:
  - observed incidence rates
  - $\log$(observed incidence rates)
  - other functions of these

Normal probability density funtion — the “Bell Curve”



Areas under curve limited by selected quantiles



Sampling distribution

- Describes variation of a summary statistic,
- $=$ behaviour of values of the statistic over hypothetical repetitions of taking new random samples of size n.
- Its form depends on:
  - original distribution & parameters,
  - sample size n.
- The larger the sample size $n \rightarrow$ the narrower and more Gaussian-like sampling distribution!

Example: Observed incidence rate

Parameter $\lambda =$ (unknown) incidence rate in population.

- **Model** incidence rate is constant over time
- **Empirical rate** $R = D/Y$,
- **Estimator** of λ, $\hat{\lambda} = R$.
- $\hat{\lambda} = R$ is a statistic, random variable:
  - its value varies from one study population ("sample") to another on hypothetical repetitions
  - its sampling distribution is (under the constant rate model & other conditions) a transformation of the Poisson distribution

Example: Observed incidence rate

- D approximately Poisson, mean λY, sd $\sqrt{\lambda Y}$
- $R = D/Y$ scaled Poisson, mean λ, sd $\sqrt{\lambda Y}/Y = \sqrt{\lambda/Y}$
- Expectation of R is λ,
  standard deviation $\sqrt{\lambda/Y}$.
- Standard error of empirical rate R is estimated by replacing λ with R:

$$\text{s.e.}(R) = \sqrt{\frac{\hat{\lambda}}{Y}} = \sqrt{\frac{R}{Y}} = \frac{\sqrt{D}}{Y} = R \times \frac{1}{\sqrt{D}}$$

$\Rightarrow$ Random error depends inversely on the number of cases.

$\Rightarrow$ s.e. of R is proportional to R.

Example: Observed incidence rate

- Use the central limit theorem:
- $\hat{\lambda} = R \sim \mathcal{N}(\lambda, \lambda/Y) = \mathcal{N}(\lambda, \lambda^2/D)$

$\Rightarrow$ Observed R is with 95% proability in the interval

$$(\lambda - 1.96 \times \lambda/\sqrt{D}; \lambda + 1.96 \times \lambda/\sqrt{D})$$

$\Rightarrow$ with 95% probability λ is in the interval

$$(R - 1.96 \times R/\sqrt{D}; R + 1.96 \times R/\sqrt{D})$$

- . . . a 95% confidence interval for the rate.

Chance summary

- Observations vary systematically by **known** factors
- Observations vary randomly by **unknown** factors
- Probability model describes the random variation
- We observe random variables — draws from a probability distribution
- Central limit theorem allows us to quantify the random variation
- Confidence interval
- . . . but we need a better foundation for the estimators

Inference

Bendix Carstensen & Esa Laara

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,
August 2017 / Januay 2018

`http://BendixCarstensen.com/NSCE/2017`

Inference

- Inferential questions
- Point estimation
- Maximum likelihood
- Statistical testing
- Interpretation of P-values
- Confidence interval
- Recommendations

Inferential questions

- What is the best single-number assesment of the parameter value?
- Is the result consistent or in disagreement with a certain value of the parameter proposed beforehand?
- What is a credible range of parameter values, consistent with our data?

Models and data

- Probability model can be used to **generate** data (by simulation)
- Interest is the **inverse**:
- What model generated the data?

Models and data — model components

- External, *a priori* information on observations
  — structure of the model
- quantitative parameter(s) within model structure
- only the latter is the target for inference

Statistical concepts

- Probability: parameters $\rightarrow$ data
- Statistics: data $\rightarrow$ parameter(estimate)s
- Notation:
  - Parameter denoted by a Greek letter
  - Estimator & estimate by the same Greek letter with "hat".
- Ex: Incidence rate:
  - True unknown rate: λ
  - Estimator: $\widehat{\lambda} = R = D/Y$, empirical rate.
- . . . but where did this come from?

Maximum likelihood principle

- Define your model (*e.g.* constant rate)
- Choose a parameter value
- How likely is it that
  — this model with
  — this parameter
  generated data
- $\mathrm{P}\{\text{data}|\text{parameter}\}$, $\mathrm{P}\{(d, y)|\lambda\}$
- Find the parameter value that gives the maximal probability of data
- Find the interval of parameter values that give probabilities not too far from the maximum.

Likelihood

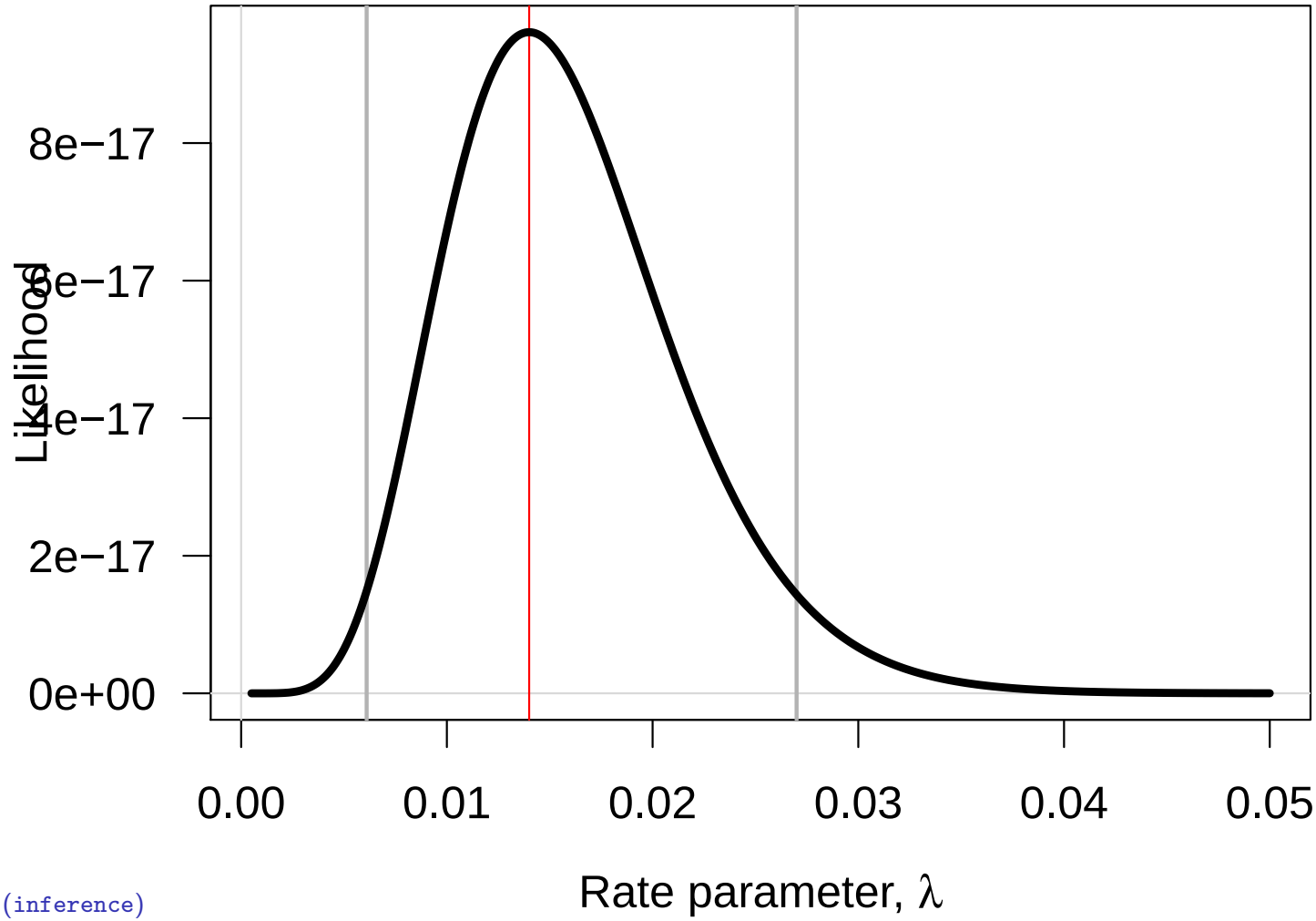
Probability of the data given the parameter:

Assuming the rate (intensity) is constant, λ, the probability of observing 7 deaths in the course of 500 person-years:

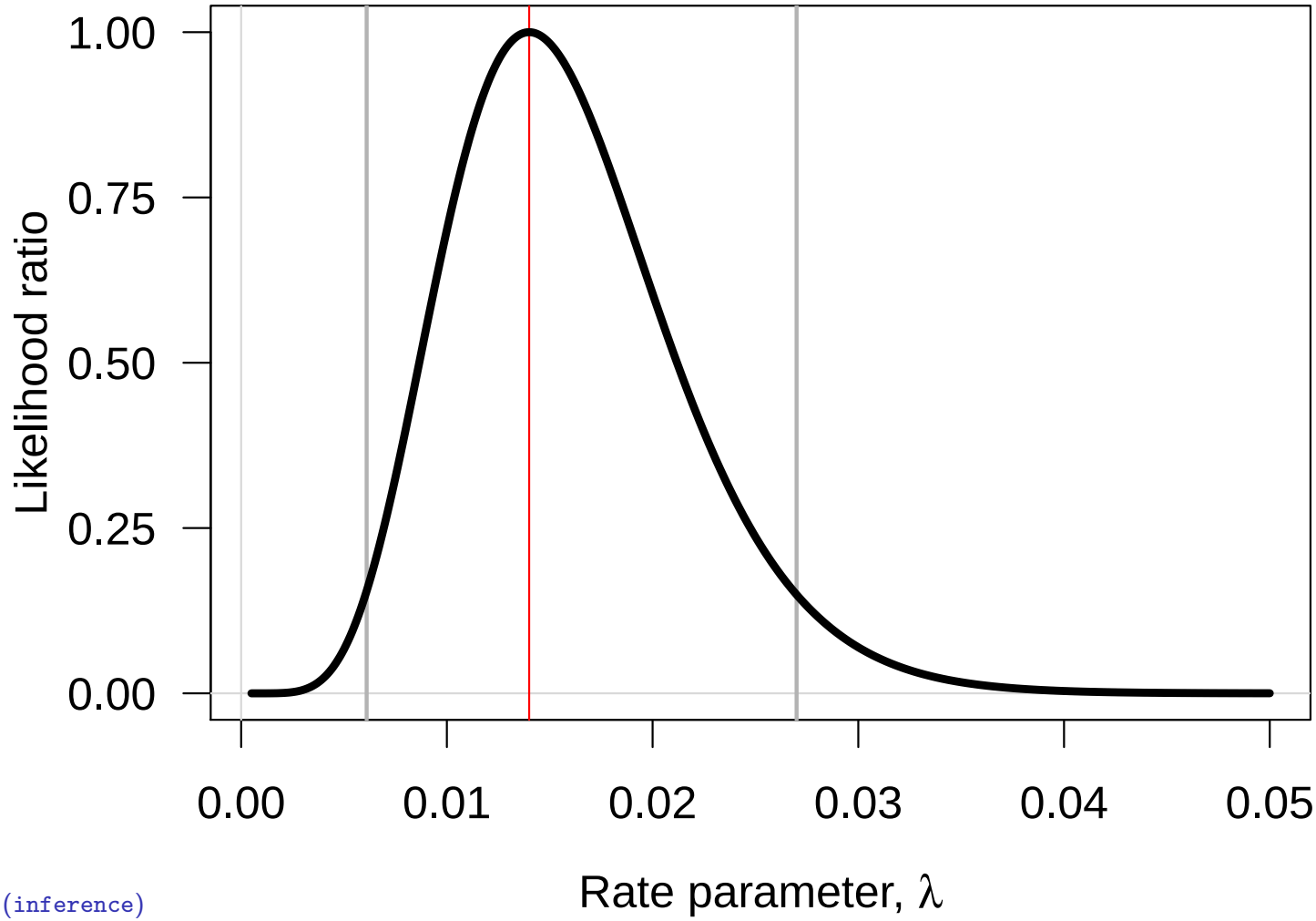
$$\begin{aligned} \mathrm{P}\left\{D = 7, Y = 500 | \lambda\right\} &= \lambda^{D} \mathrm{e}^{\lambda Y} \times K \\ &= \lambda^{7} \mathrm{e}^{\lambda 500} \times K \\ &= L(\lambda | \text{data}) \end{aligned}$$

- Estimate of λ is where this function is as large as possible.
- Confidence interval is where it is not too far from the maximum

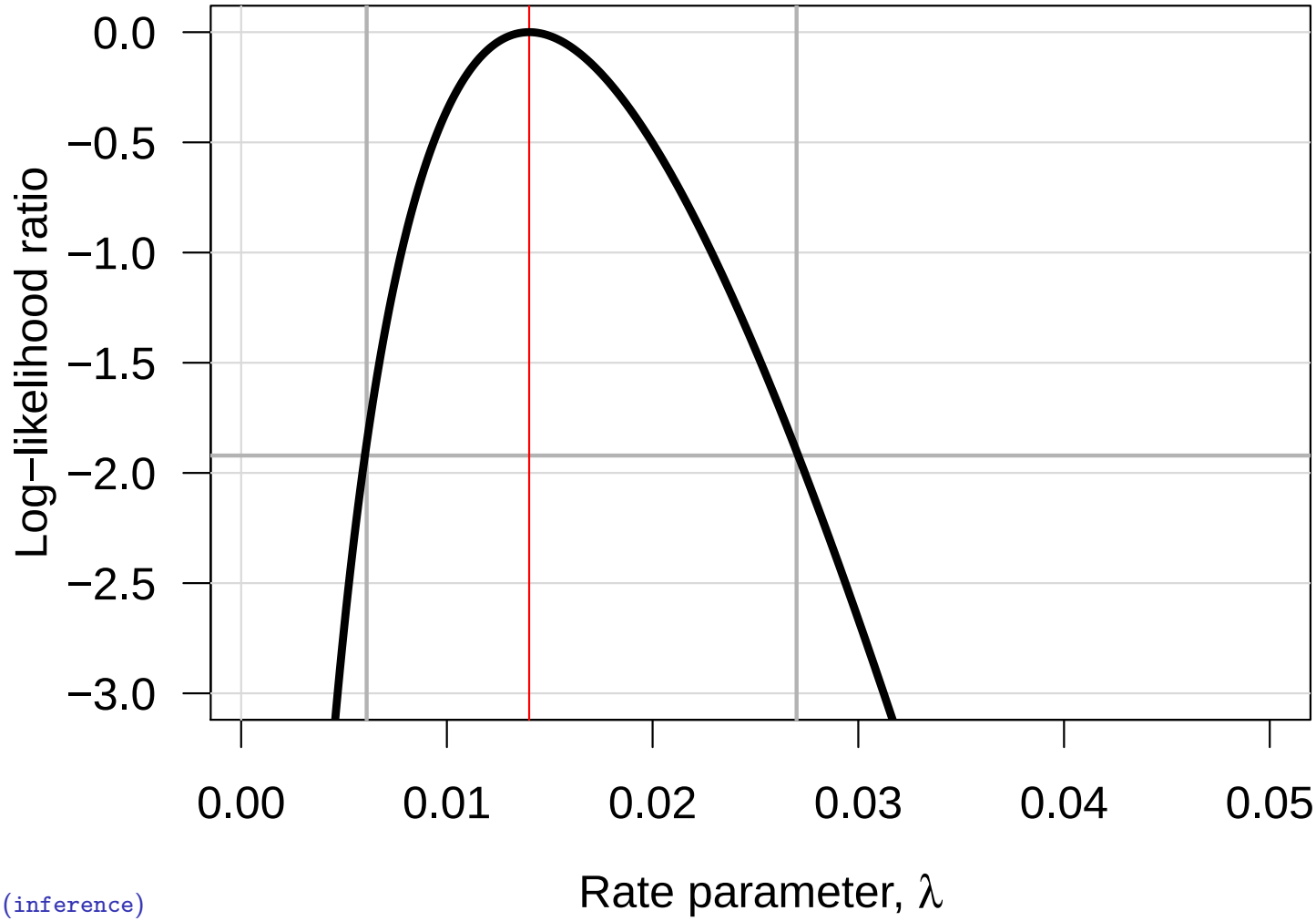
Likelihood function, 7 events, 500 PY



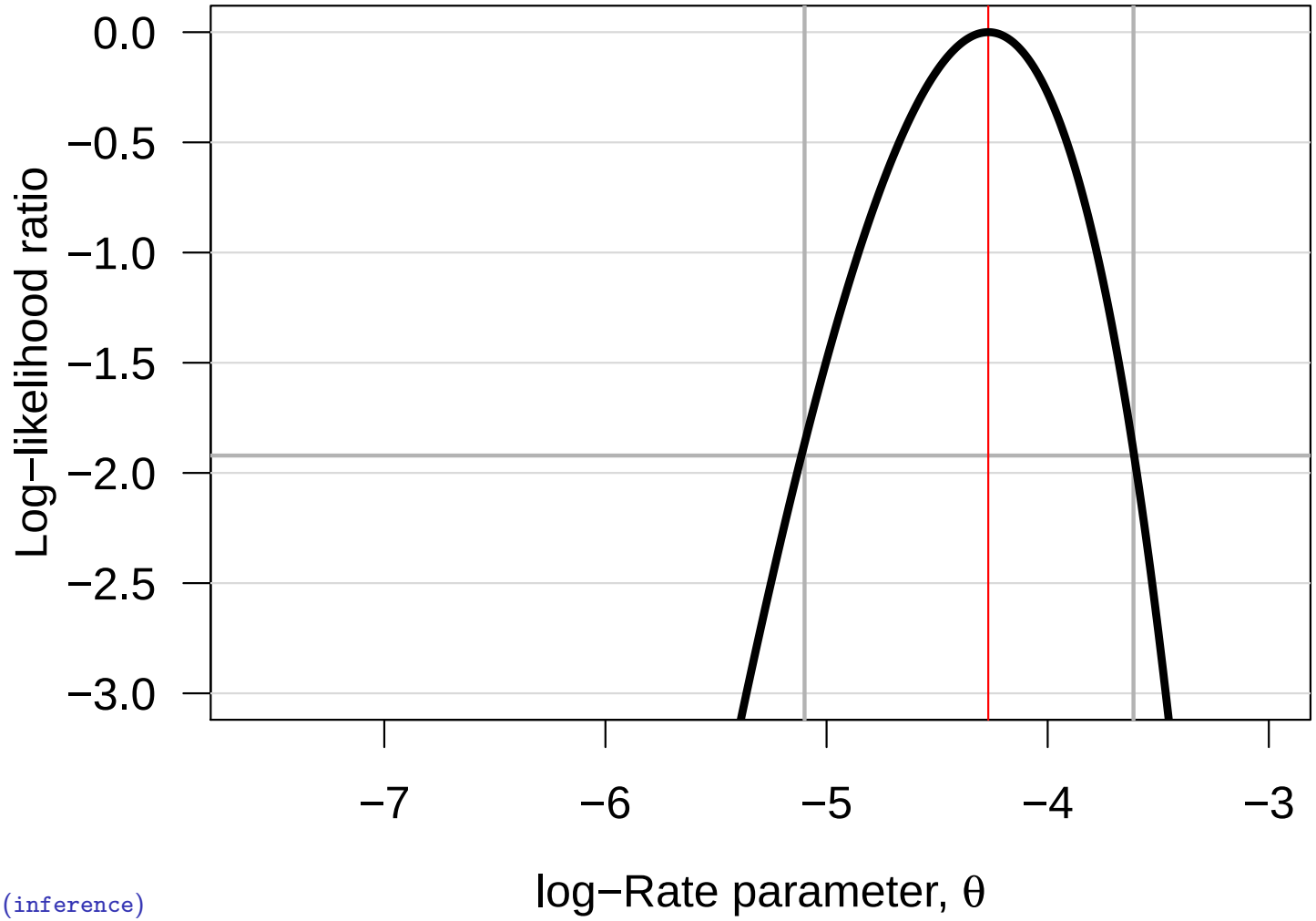
Likelihood function, 7 events, 500 PY



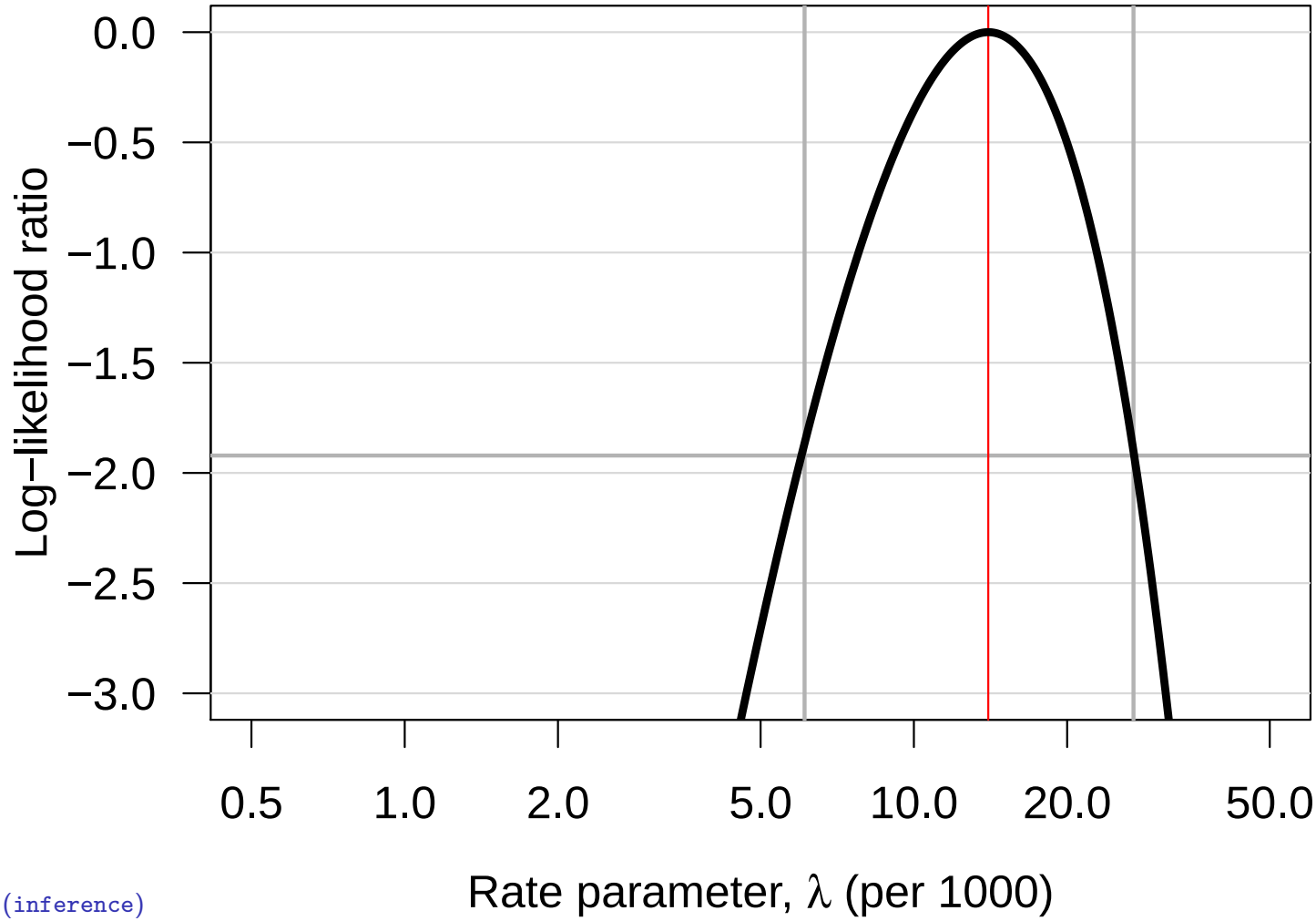
Log-likelihood function, 7 events, 500 PY



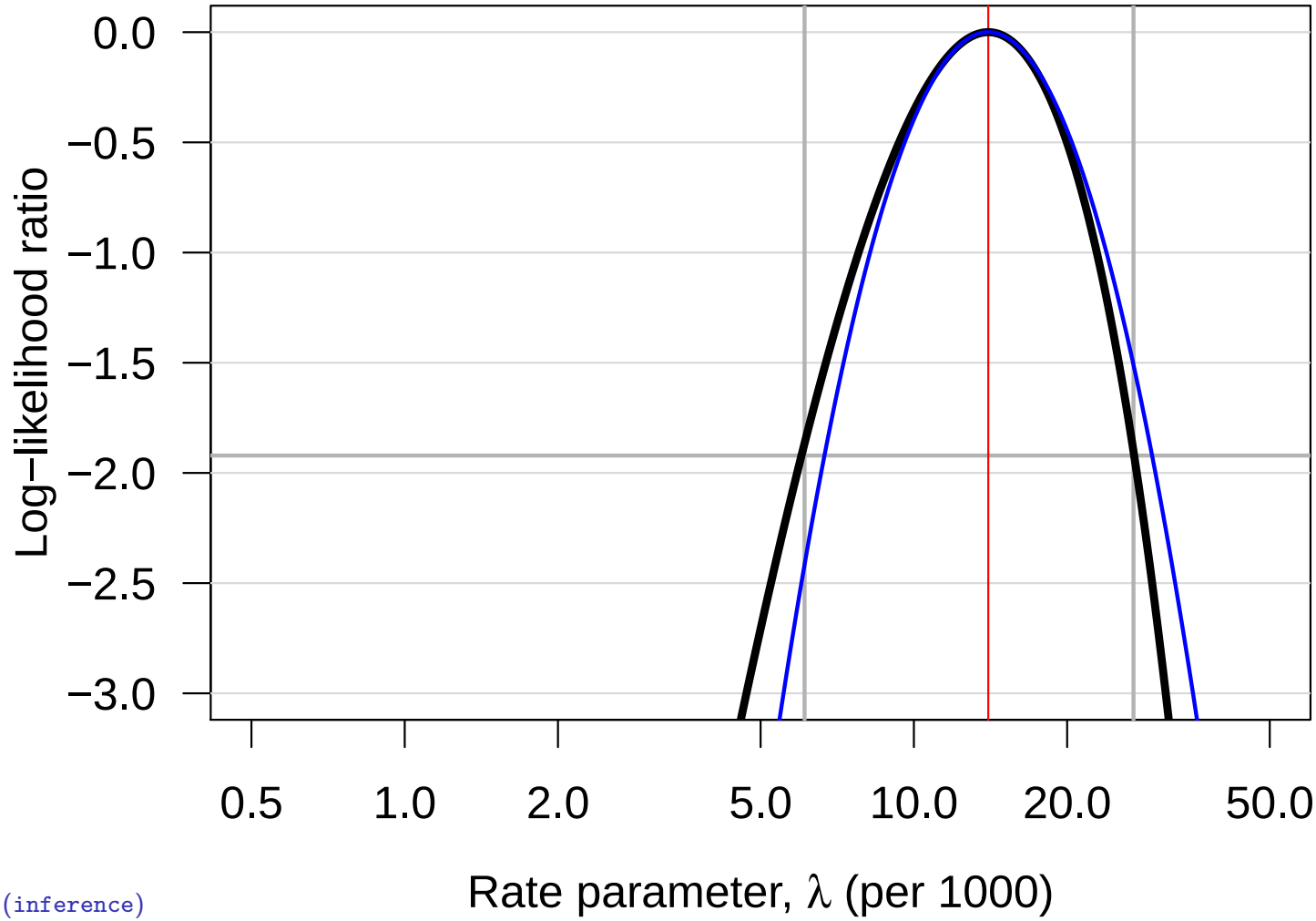
Log-likelihood function, 7 events, 500 PY



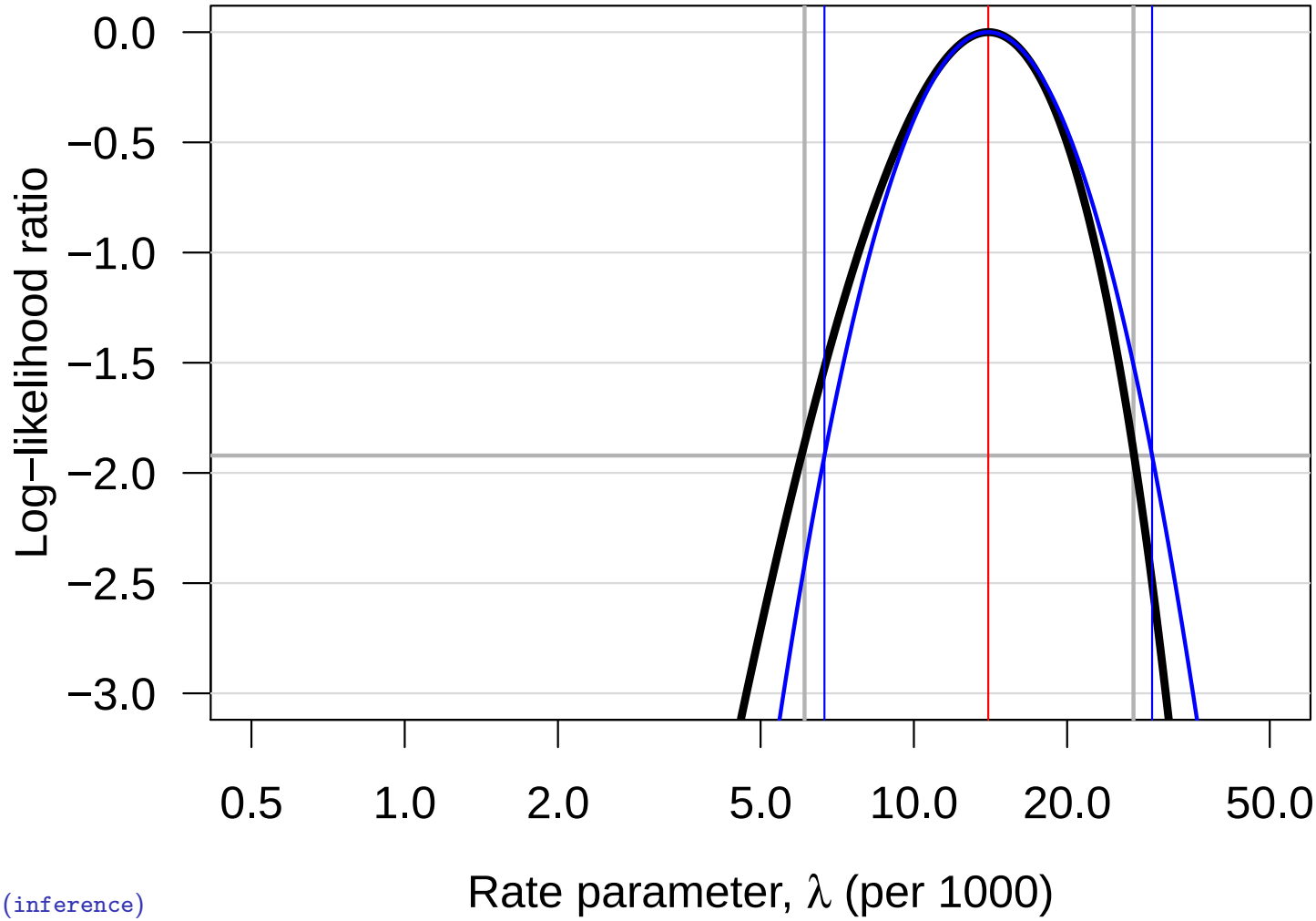
Log-likelihood function, 7 events, 500 PY



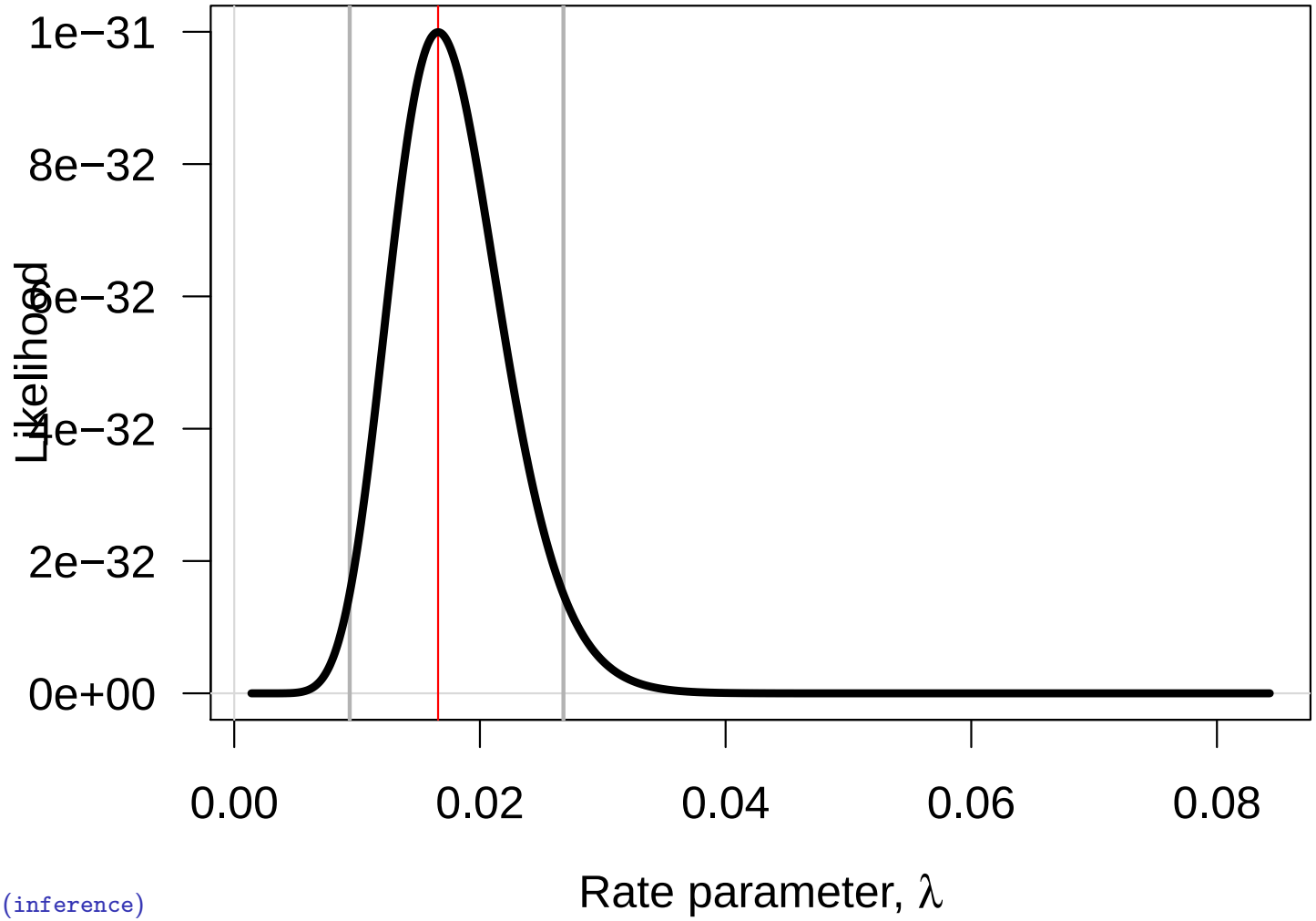
Log-likelihood function, 7 events, 500 PY



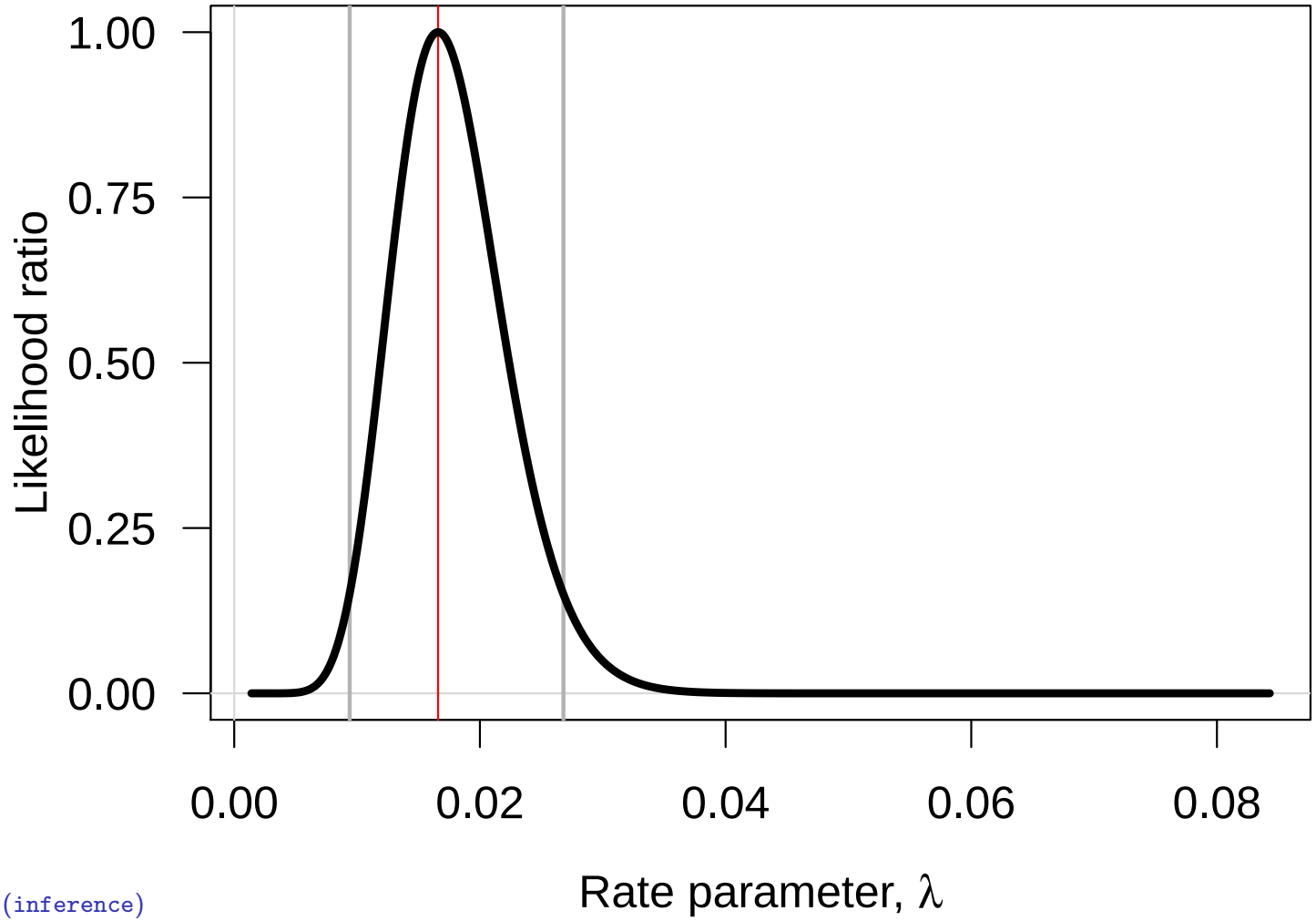
Log-likelihood function, 7 events, 500 PY



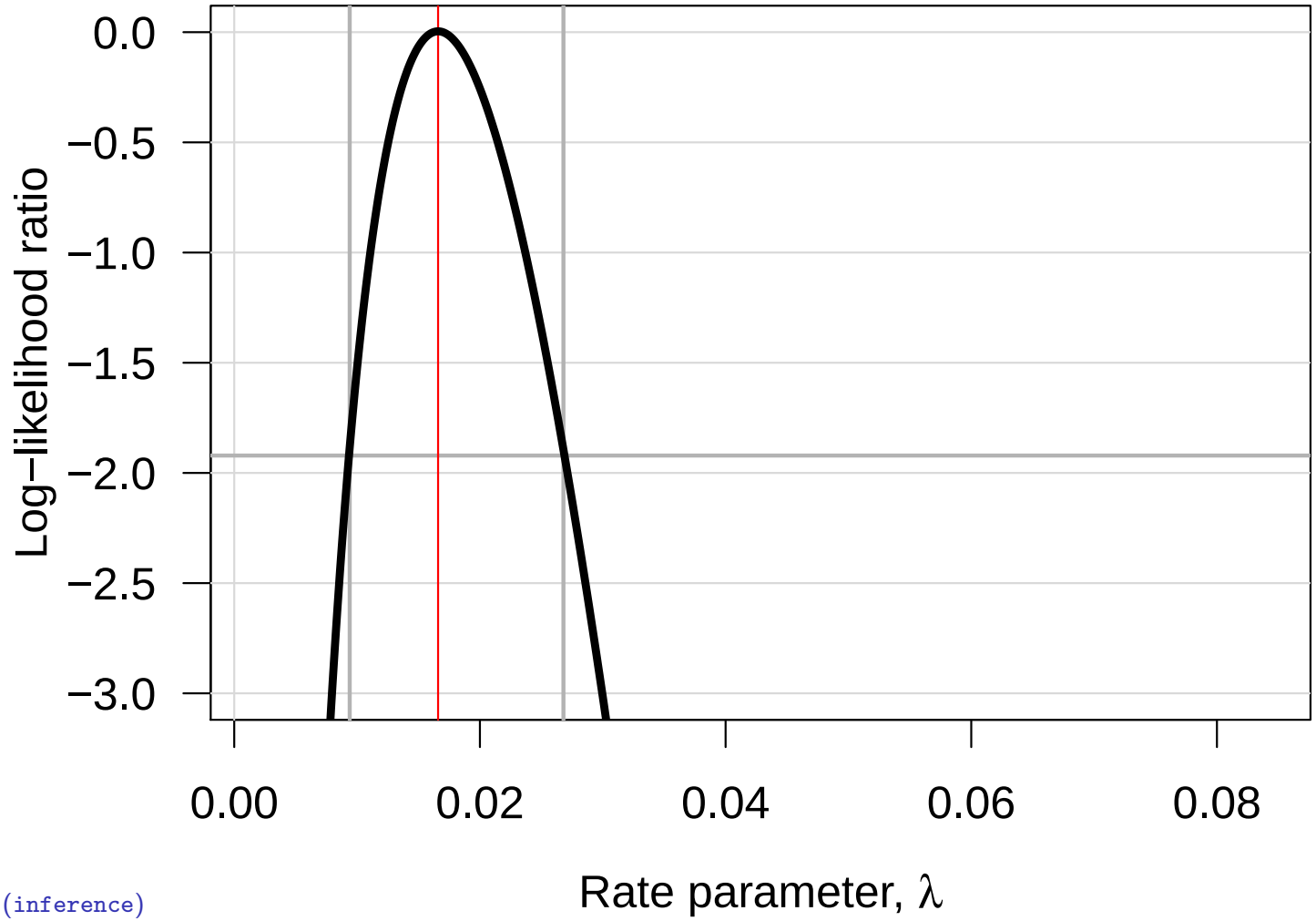
Likelihood function, 14 events, 843.6 PY



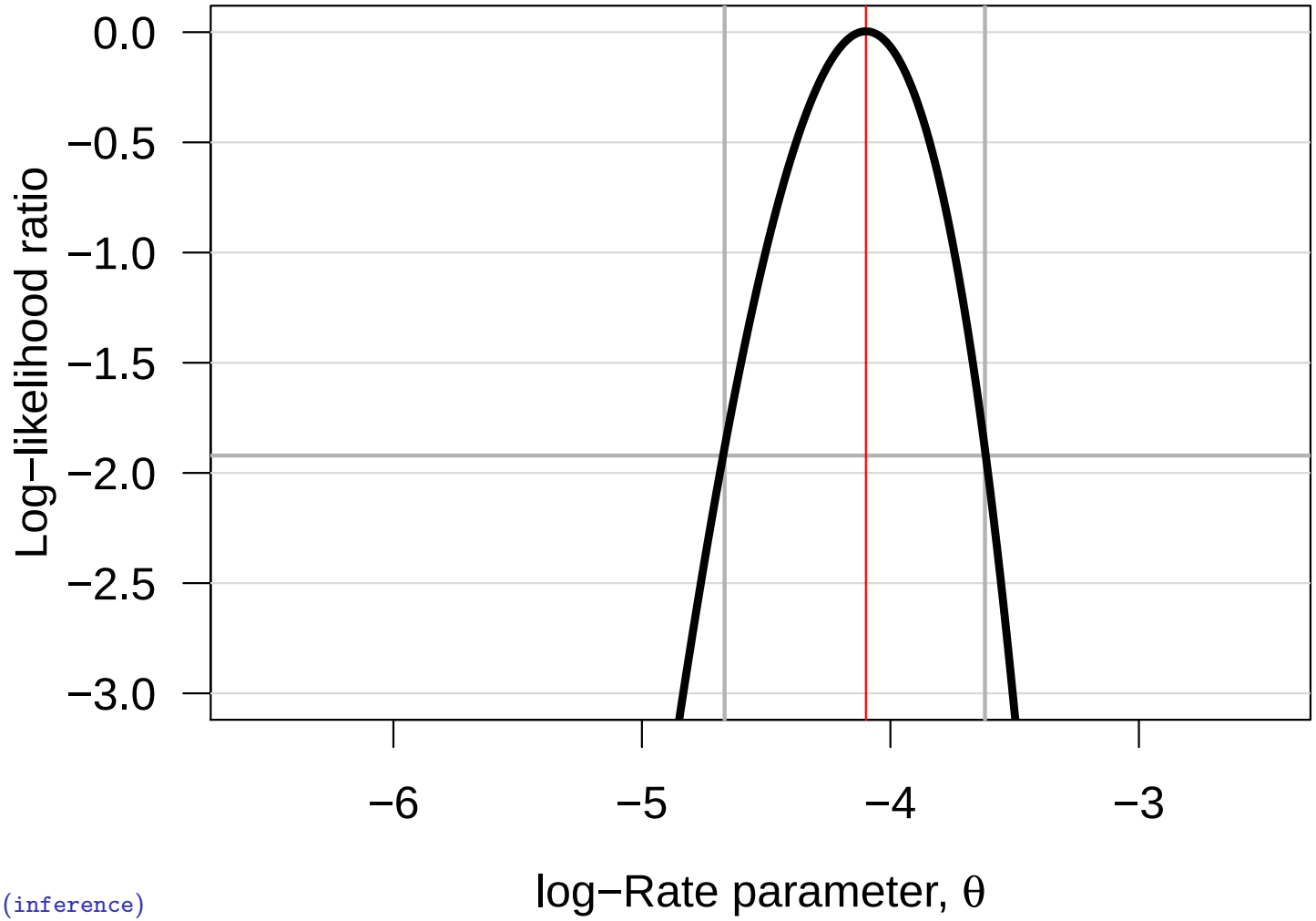
Likelihood function, 14 events, 843.6 PY



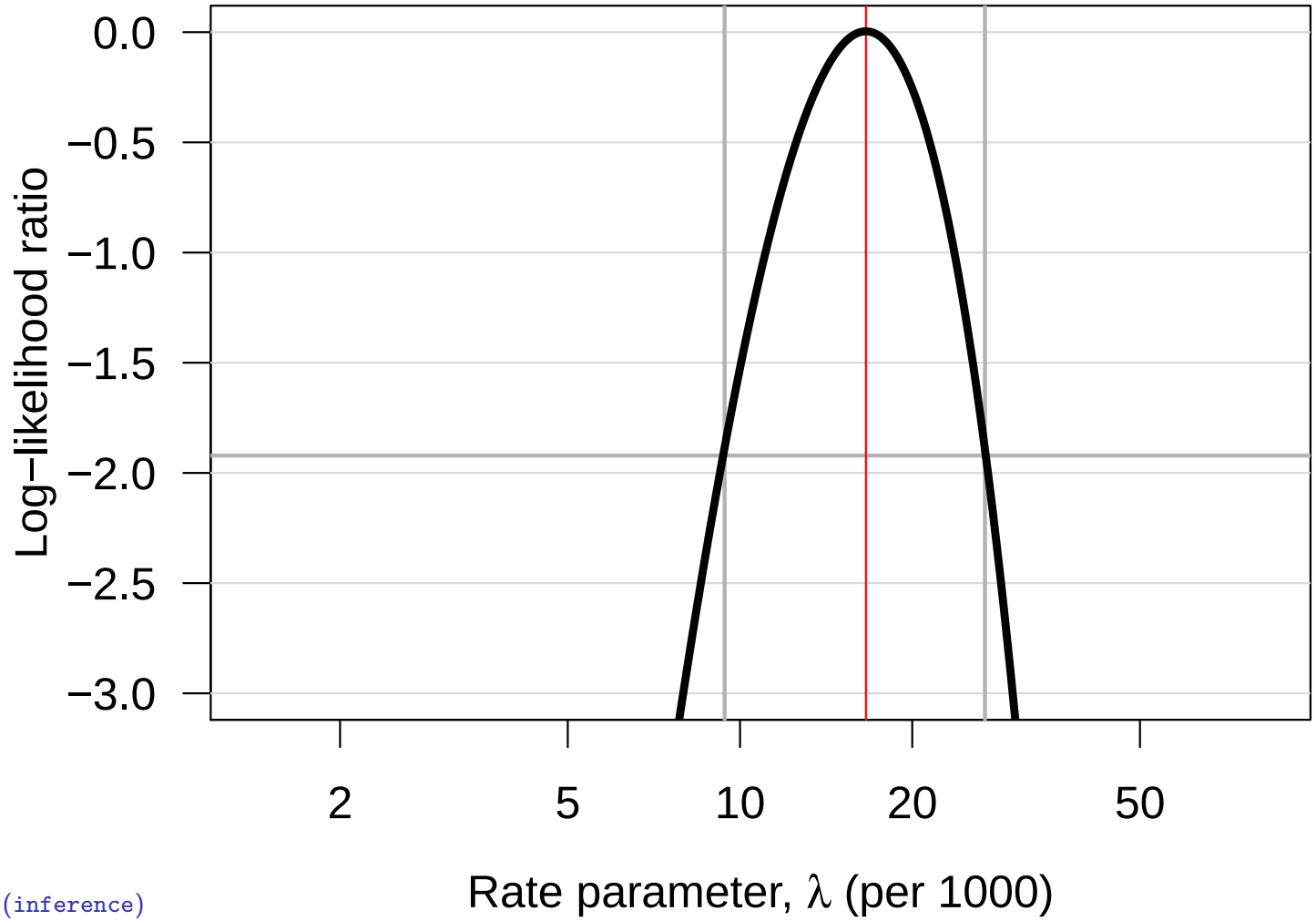
Log-likelihood function 14 events, 843.6 PY



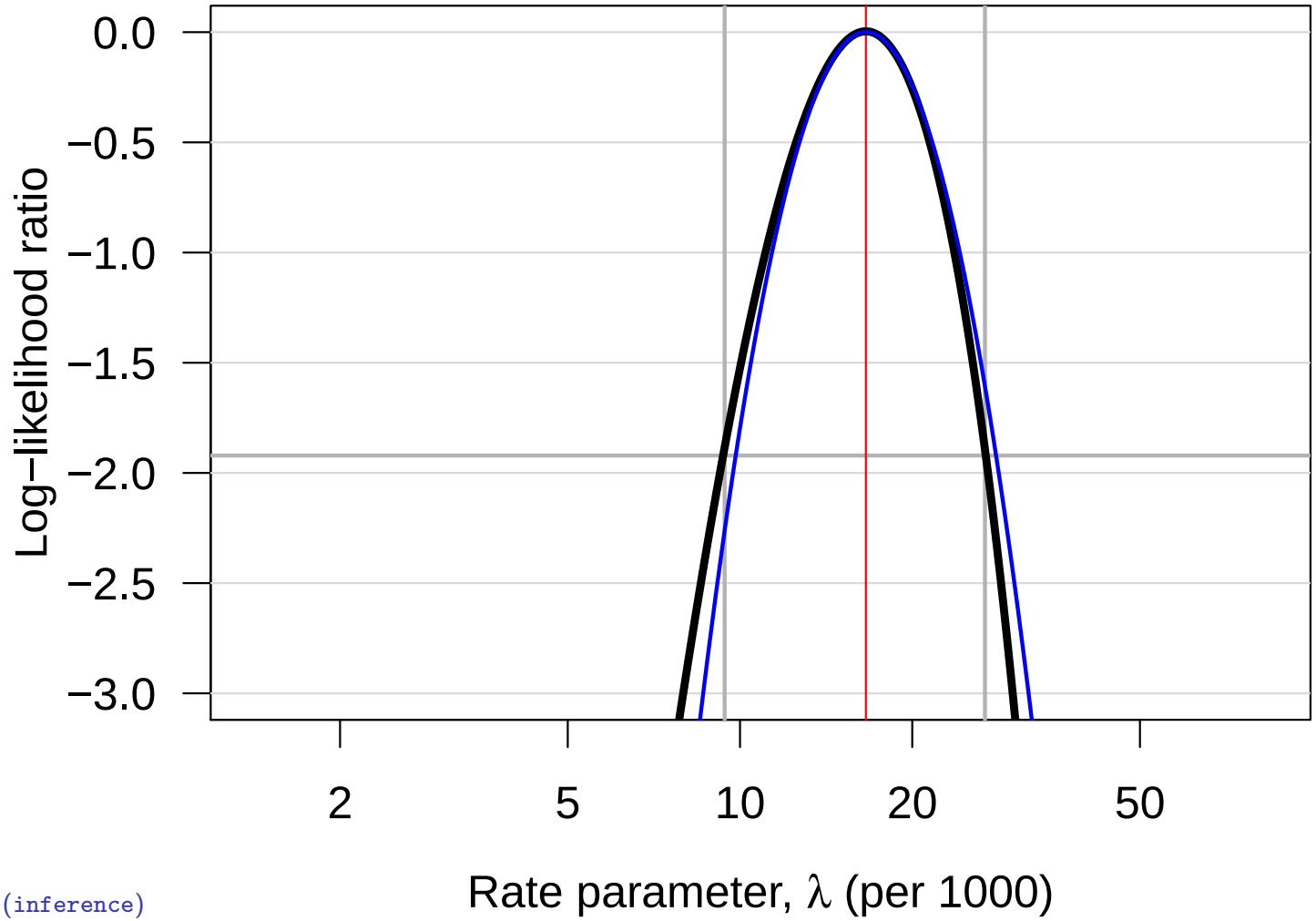
Log-likelihood function 14 events, 843.6 PY



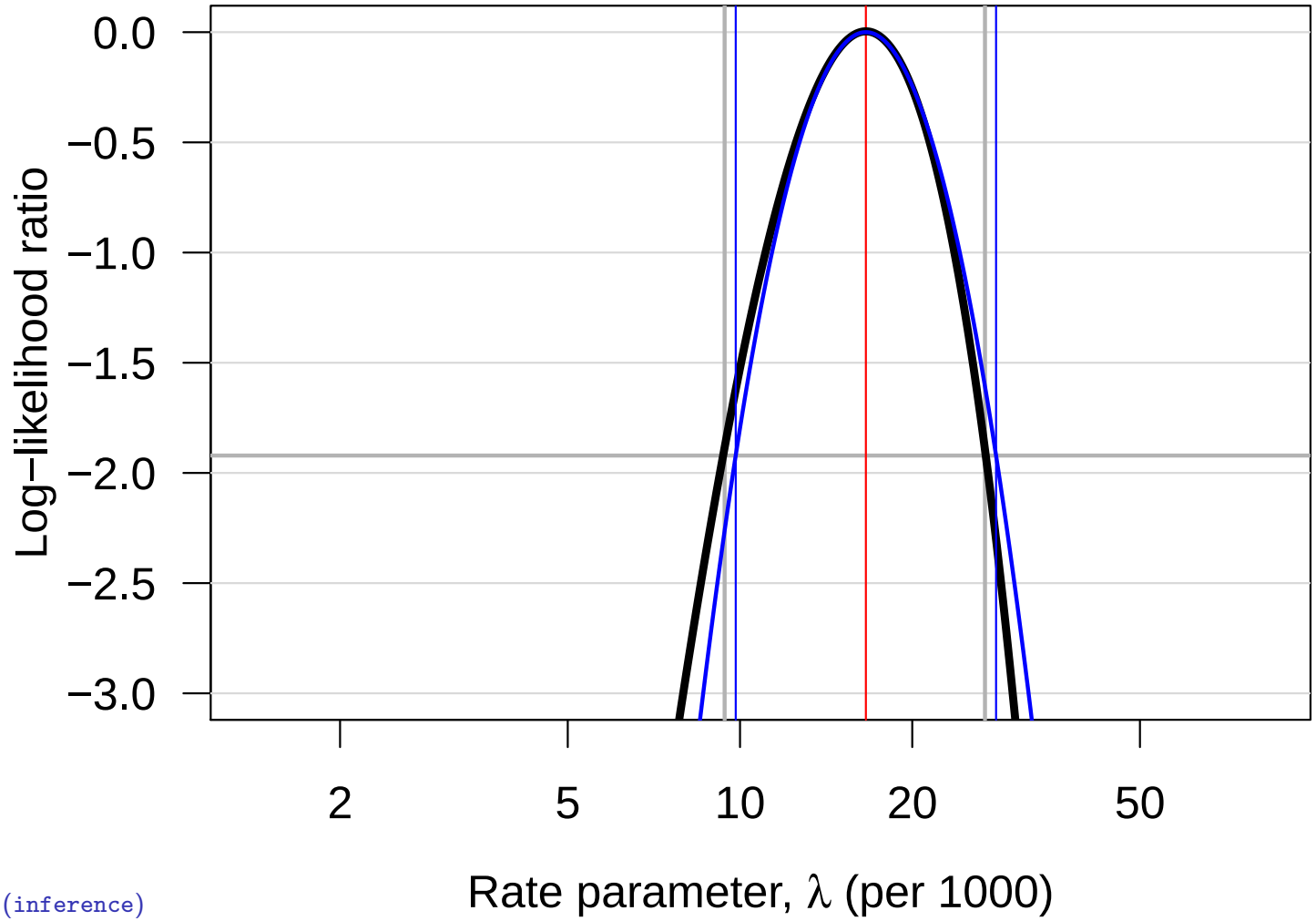
Log-likelihood function 14 events, 843.6 PY



Log-likelihood function 14 events, 843.6 PY



Log-likelihood function 14 events, 843.6 PY



Confidence interval for a rate

- Based on the quadratic approximation:
- A 95% confidence interval for the log of a rate is:

$$\hat{\theta} \pm 1.96/\sqrt{D} = \log(\lambda) \pm 1.96/\sqrt{D}$$

- Take the exponential to get the confidence interval for the rate:

$$\lambda \overset{\times}{\div} \underbrace{\exp(1.96/\sqrt{D})}_{\text{error factor,erf}}$$

Example

Suppose we have 14 deaths during 843.6 years of follow-up.

The rate is computed as:

$$\hat{\lambda} = D/Y = 14/843.7 = 0.0165 = 16.5 \text{ per 1000 years}$$

The confidence interval is computed as:

$$\hat{\lambda} \overset{\times}{\div} \text{erf} = 16.5 \overset{\times}{\div} \exp(1.96/\sqrt{14}) = (9.8, 28.0)$$

per 1000 person-years.

Ratio of two rates

If we have observations two rates λ_1 and λ_0, based on (D_1, Y_1) and (D_0, Y_0), the variance of the difference of the log-rates, the $\log(\mathrm{RR})$, is:

$$\begin{aligned} \mathrm{var}(\log(\mathrm{RR})) &= \mathrm{var}(\log(\lambda_1/\lambda_0)) \\ &= \mathrm{var}(\log(\lambda_1)) + \mathrm{var}(\log(\lambda_0)) \\ &= 1/D_1 + 1/D_0 \end{aligned}$$

As before a 95% c.i. for the RR is then:

$$\mathrm{RR} \overset{\times}{\div} \underbrace{\exp\left(1.96\sqrt{\frac{1}{D_1} + \frac{1}{D_0}}\right)}_{\text{error factor}}$$

Example

Suppose we in group 0 have 14 deaths during 843.6 years of follow-up in one group, and in group 1 have 28 deaths during 632.3 years.

The rate-ratio is computed as:

$$\begin{aligned} \mathrm{RR} &= \hat{\lambda}_1/\hat{\lambda}_0 = (D_1/Y_1)/(D_0/Y_0) \\ &= (28/632.3)/(14/843.7) = 0.0443/0.0165 = 2.669 \end{aligned}$$

The 95% confidence interval is computed as:

$$\begin{aligned} \hat{\mathrm{RR}} \overset{\times}{\div} \mathrm{erf} &= 2.669 \overset{\times}{\div} \exp\left(1.96\sqrt{1/14 + 1/28}\right) \\ &= 2.669 \overset{\times}{\div} 1.899 = (1.40, 5.07) \end{aligned}$$

Example using R

Poisson likelihood for one rate, based on 14 events in 843.7 PY:

```
> library( Epi )
> D <- 14 ; Y <- 843.7
> m1 <- glm( D ~ 1, offset=log(Y/1000), family=poisson)
> ci.exp( m1 )

            exp(Est.)     2.5%    97.5%
(Intercept)  16.59358 9.827585 28.01774
```

Poisson likelihood, two rates, or one rate and RR:

```
> D <- c(14,28) ; Y <- c(843.7,632.3) ; gg <- factor(0:1)
> m2 <- glm( D ~ gg, offset=log(Y/1000), family=poisson)
> ci.exp( m2 )

            exp(Est.)     2.5%     97.5%
(Intercept)  16.59358 9.827585 28.017744
gg1           2.66867 1.404992  5.068926
```

Example using R

Poisson likelihood, two rates, or one rate and RR:

```
> D <- c(14,28) ; Y <- c(843.7,632.3) ; gg <- factor(0:1)
> m2 <- glm( D ~ gg, offset=log(Y/1000), family=poisson)
> ci.exp( m2 )

            exp(Est.)     2.5%     97.5%
(Intercept)  16.59358 9.827585 28.017744
gg1           2.66867 1.404992  5.068926

> m3 <- glm( D ~ gg - 1, offset=log(Y/1000), family=poisson)
> ci.exp( m3 )

    exp(Est.)      2.5%    97.5%
gg0  16.59358  9.827585 28.01774
gg1  44.28278 30.575451 64.13525
```

Statistical testing

- Are the observed data
  (possibly summarized by an estimate and its SE)
  consistent with a given value of the parameter?
- Such a value is often represented in the form a *null hypothesis* (H_0), which is a statement about the belief about value of the parameter before study.
- Typically a conservative assumption, *e.g.*:
  "no difference in outcome between the groups"
  "true rate ratio $\rho = 1$".

Purpose of statistical testing

- Evaluation of consistency or disagreement of observed data with H_0.
- Checking whether or not the observed difference can reasonably be explained by chance.
- **Note:** This is not so ambitious.
- The NULL is never true — there is always a difference between two groups

$\Rightarrow$ not testing if H_0 is **TRUE**,

- **if** it were true could we see this kind of data
- . . . not investigating if there were **other** probability models that could have generated the data
- . . . but if we have evidence enough to assert is as **FALSE**

Test statistic

- Function of observed data and null hypothesis value,
- a common form of test statistic is:

$$Z = \frac{O - E}{S}$$

O = some "observed" statistic,
E = "expected value" of O under H_0,
S = SE or standard deviation of O under H_0.

- Evaluates the size of the "signal" $O - E$ against the size of the "noise" S — if numerically large, H_0 unlikely
- Under H_0 the sampling distribution of this statistic is (with sufficient amount of data) close to the standard Gaussian.

Example — rate difference

Null hypothesis:

- OC use has no effect on breast ca. risk
  $\Leftrightarrow$ true rate difference $\delta = \lambda_1 - \lambda_0$ equals 0.

O = Observed rate difference

$$\widehat{\delta} = \mathrm{RD} = (28/632.3) - (14/843.7) = 44.2 - 16.5 = 27.7 \text{ per } 10^3\text{PY}.$$

E = Expected rate difference = 0, if H_0 true.

S = Standard error of RD:

$$\text{SE(RD)} = \sqrt{\frac{28}{632.3^2} + \frac{14}{843.7^2}} = 9.5 \text{ per } 10^3 \text{ y}.$$

Example — rate difference

- Test statistic $Z = (O - E)/S$, its observed value:

$$Z_{\text{obs}} = \frac{27.7 - 0}{9.5} = 2.92$$

- One-tailed $P = 0.0017$:
  probability of more extreme observations in **one** direction
- Two-tailed $P = 0.0034$:
  probability of more extreme observations in **any** direction
- Question of *a priori* assumptions
- Two-tailed is the preferred in most cases

P-value

- Synonym for "observed significance level".
- Measures the **evidence against** H_0:
  - The smaller the p value, the stronger the evidence against H_0.
  - Yet, a large p as such **does not** provide supporting evidence *for* H_0.
- Operationally: the probability of getting a statistic at least as extreme as the observed, **assuming** H_0 is true
- However, **it is not**
  "the probability that H_0 is true"!

Interpretation of P-values

- No mechanical rules of inference
- Rough guidelines
  - "large" value ($p > 0.1$): consistent with H_0 but not necessarily supporting it,
  - "small" value ($p < 0.01$): indicates evidence against H_0
  - "intermediate" value ($p \approx 0.05$): weak evidence against H_0
- Division of p-values into "significant" or "non-significant" by cut-off 0.05 — **To be avoided!**
- . . . remember that the 5% is an arbitrary number taken out pf thin air.

Confidence interval (CI)

- Range of values of the parameter compatible with the observed data
- Specified at certain *confidence level*, commonly 95% (also 90 % and 99% used)
- The limits of a CI are statistics, random variables with sampling distribution, such that
- the probability that the random interval covers the true parameter value equals the confidence level (*e.g.* 95%).

Interpretation of obtained CI

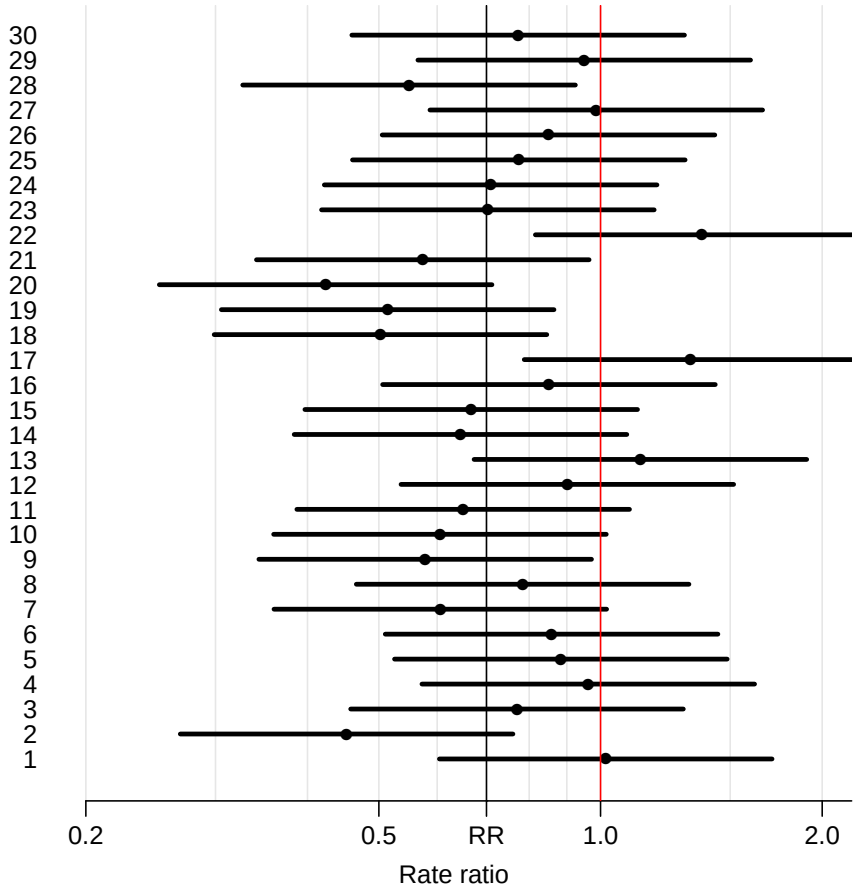
Frequentist school of statistics: no probability interpretation! (In contrast to *Bayesian* school).

Single CI is viewed by frequentists as a range of conceivable values of the unknown parameter with which the observed estimate is fairly consistent, taking into account "probable" random error:

- narrow CI $\rightarrow$ precise estimation
  $\rightarrow$ small statistical uncertainty about parameter.
- wide CI $\rightarrow$ imprecise estimation
  $\rightarrow$ great uncertainty.

Long-term behaviour of CI

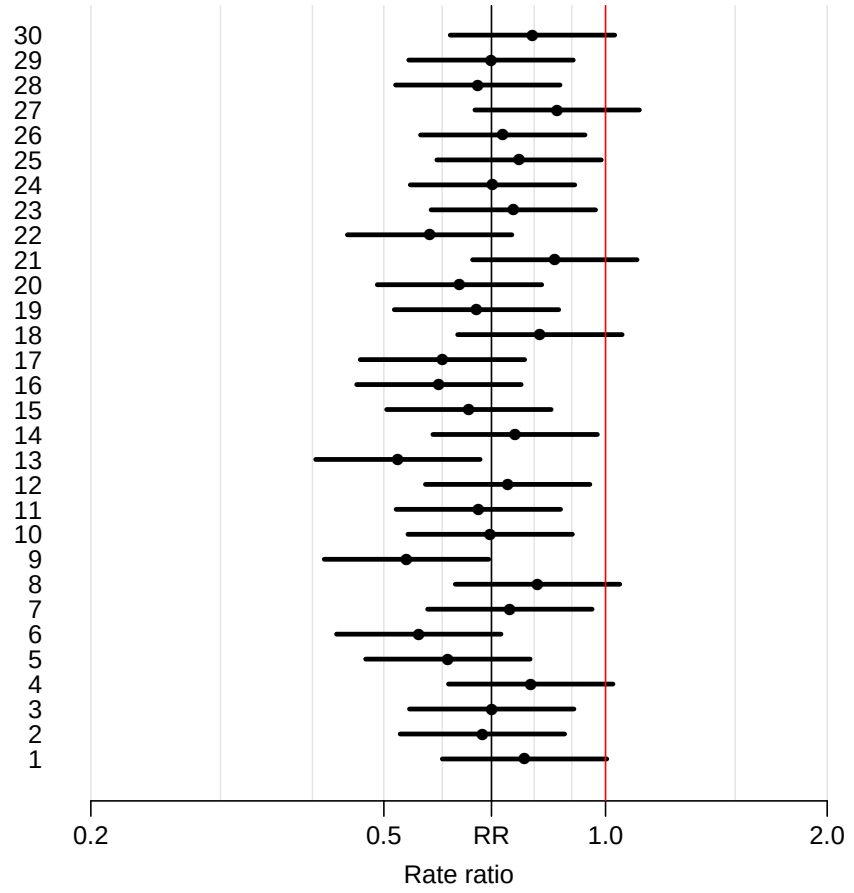
Variability of 95% CI under hypothetical repetitions of similar study, when true rate ratio is RR.



In the long run 95% of these intervals would cover the true value but 5% would not.

Long-term behaviour of CI

Variability of 95% CI under hypothetical repetitions of similar study, when true rate ratio is RR.



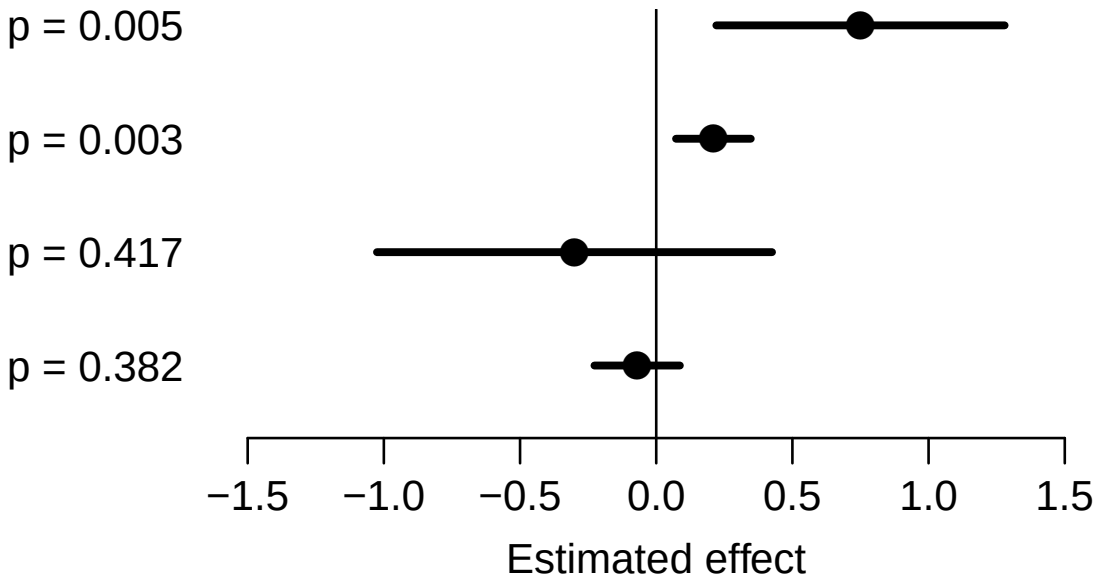
In the long run 95% of these intervals would cover the true value but 5% would not.

Interpretation of CI

- CI gives more quantitative information on the parameter and on statistical uncertainty about its value than P value.
- narrow CI about H_0 value:
  $\rightarrow$ results give support to H_0.
- wide CI about H_0 value:
  $\rightarrow$ results inconclusive.
- The latter is more commonly encountered.

Confidence interval and P-value

95 % CIs of rate difference δ and P values for $H_0 : \delta = 0$ in different studies.



- Which ones are significant?
- Which ones are informative?

Recommendations

ICMJE: Uniform Requirements for Manuscripts submitted to Biomedical Journals. `http://www.icmje.org/`

Extracts from section *Statistics*:

- When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals).
- Avoid relying solely on statistical hypothesis testing, such as the use of p values, which fails to convey important quantitative information.

Recommendations

Sterne and Davey Smith: Sifting the evidence – what's wrong with significance tests? *BMJ* 2001; **322**: 226-231.

"Suggested guidelines for the reporting of results of statistical analyses in medical journals"

1. The description of differences as statistically significant is not acceptable.
2. Confidence intervals (CI) for the main results should always be included, but 90% rather than 95% levels should be used.

Recommendations

3. CIs should not be used as a surrogate means of examining significance at the conventional 5% level.
4. Interpretation of CIs should focus on the implications (clinical importance) of the range of values in the interval.
5. In observational studies it should be remembered that considerations of confounding and bias are at least as important as the issues discussed in this paper.

Analysis

Bendix Carstensen & Esa Laara

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,
August 2017 / Januay 2018

http://BendixCarstensen.com/NSCE/2017

Crude analysis

- Single incidence rate
- Rate ratio in cohort study
- Rate ratio in case-control study
- Rate difference in cohort study
- Analysis of proportions
- Extensions and remarks

Single incidence rate

- **Model**: Events occur with constant rate λ.
- **Parameter** of interest:

$$\lambda = \text{true rate in target population}$$

- **Estimator**: $\widehat{\lambda} = R$, the empirical rate in a "representative sample" from the population:

$$R = \frac{D}{Y} = \frac{\text{no. of cases}}{\text{person-time}}$$

- Standard error of rate: $\text{SE}(R) = R/\sqrt{D}$.

Single rate

- Simple approximate 95% CI:

$$[R - \mathrm{EM}, R + \mathrm{EM}]$$

- using 95% **error margin**:

$$\mathrm{EM} = 1.96 \times \mathrm{SE}(R)$$

- Problem: When $D \leq 4$, lower limit ≤ 0!

Single rate

- Better approximation on log-scale:

$$\mathrm{SE}\big(\log(R)\big) = 1/\sqrt{D}$$

- From this we get the 95% **error factor** (EF)

$$\mathrm{EF} = \exp\Big(1.96 \times \mathrm{SE}\big(\log(R)\big)\Big)$$

  where $\exp$ is the exponential function or antilog (inverse of the natural logarithm)
- From these items we get 95% CI for λ:

$$[R/\mathrm{EF},\ R \times \mathrm{EF}].$$

- These limits are always > 0 whenever $D \geq 1$.

Single rate example

- The observed incidence of breast cancer in Finnish men aged 65-69 y in 1991 was 33 per 10^6 py based on 3 cases.
- Standard error of the rate is:

$$\mathrm{SE}(R) = 33 \times \sqrt{1/3} = 19 \text{ per } 10^6 \text{ y}$$

- The 95% error margin:

$$\begin{aligned} \mathrm{EM} &= 1.96 \times 19 = 37 \text{ per } 10^6 \text{ y} \\ 33 \pm 37 &= [-4, 70] \text{ per } 10^6 \text{ y} \end{aligned}$$

Negative lower limit — illogical!

Single rate example

- A better approximate CI obtained on the log-rate scale:

$$\mathrm{SE}\big(\log(R)\big) = \sqrt{1/3} = 0.577$$

- via the 95% error factor:

$$\mathrm{EF} = \exp(1.96 \times 0.577) = 3.1$$

from which the confidence limits (both > 0):

$$[33/3.1, 33 \times 3.1] = [10.6, 102] \text{ per } 10^6\text{py}$$

Rate estimation in Poisson model

3 male breast cancers in $90,909$ person years:

```
> library( Epi )
> D <- 3 ; Y <- 90909 / 10^6 ; D/Y

[1] 33.00003

> m0 <- glm( D ~ 1, offset=log(Y), family=poisson )
> ci.exp( m0 )

            exp(Est.)     2.5%    97.5%
(Intercept)  33.00003 10.64322 102.3189
```

- Response variable: `D` — no. cases
- Offset variable: `log(Y)` — log-person-years
  note the scaling of Y to the units desired.
- Explanatory variable: "1" — intercept only
- `ci.exp` transforms back to rate scale.

Rate ratio in cohort study

Question: What is the rate ratio of cancer in the exposed as compared to the unexposed group?

Model Cancer incidence rates constant in both groups, values λ_1, λ_0

Parameter of interest is true rate ratio:

$$\rho = \frac{\lambda_1}{\lambda_0} = \frac{\text{rate among exposed}}{\text{rate among unexposed}}$$

Null hypothesis $H_0 : \rho = 1$: exposure has no effect.

Rate ratio

Summarized data on outcome from cohort study with person-time

Exposure to risk factor	Cases	Person-time
Yes	D_1	Y_1
No	D_0	Y_0
Total	D_+	Y_+

Empirical rates by exposure group provide estimates for the true rates:

$$\widehat{\lambda}_1 = R_1 = \frac{D_1}{Y_1}, \qquad \widehat{\lambda}_0 = R_0 = \frac{D_0}{Y_0}$$

Rate ratio

- Point estimate of the true rate ratio, ρ, is the empirical rate ratio (RR):

$$\widehat{\rho} = \mathrm{RR} = \frac{\hat{\lambda}_1}{\hat{\lambda}_0} = \frac{R_1}{R_0} = \frac{D_1/Y_1}{D_0/Y_0} = \frac{D_1/D_0}{Y_1/Y_0}$$

- The last form is particularly useful in case-control studies — see next section.
- Easier to use the $\log$-transformation:

$$\log(\mathrm{RR}) = \log(\hat{\lambda}_1) - \log(\hat{\lambda}_0)$$

Rate ratio

$$\log(\mathrm{RR}) = \log(\hat{\lambda}_1) - \log(\hat{\lambda}_0)$$

$\Rightarrow$ variance of $\log(\mathrm{RR})$ = sum of the variances of the log-rates.

- Standard error of $\log(\mathrm{RR})$, 95% error factor and approximate 95% CI for ρ:

$$\begin{aligned} \mathrm{SE}\big(\log(\mathrm{RR})\big) &= \sqrt{\frac{1}{D_1} + \frac{1}{D_0}} \\ \mathrm{EF} &= \exp\Big(1.96 \times \mathrm{SE}\big(\log(\mathrm{RR})\big)\Big) \\ \mathrm{CI} &= [\mathrm{RR}/\mathrm{EF},\ \mathrm{RR} \times \mathrm{EF}]. \end{aligned}$$

Note: SE (EF) of estimate depends inversely on numbers of cases.

Example: Helsinki Heart Study

- In the study (Frick et al. NEJM 1987) over 4000 men were randomized to daily intake of either:
  - gemfibrozil ("exposed", $N_1 \approx 2000$), or
  - placebo ("unexposed", $N_0 \approx 2000$).
- After mean follow-up of 5 y, the numbers of cases of any cancer in the two groups were:
  $D_1 = 31$ and $D_0 = 26$.
- Rounded person-years were $Y_1 \approx Y_0 \approx 2000 \times 5$ y $= 10000$ y.

Example: Helsinki Heart Study

Incidence rates 3.1 and 2.6 per 1000 y.
Estimate of true rate ratio ρ with SE etc.:

$$\begin{aligned} \widehat{\rho} = \mathrm{RR} &= \frac{3.1/1000\mathrm{y}}{2.6/1000\ \mathrm{y}} = 1.19 \\ \mathrm{SE}[\log(\mathrm{RR})] &= \sqrt{\frac{1}{31} + \frac{1}{26}} = 0.2659 \\ \mathrm{EF} &= \exp(1.96 \times 0.2659) = 1.68 \end{aligned}$$

95 % CI for ρ :

$$[1.19/1.68,\ 1.19 \times 1.68] = [0.7,\ 2.0]$$

Two-tailed $P = 0.52$

Rate ratio in Poisson model

```
> library( Epi )
> D <-c(31,26) ; Y <- c(10000,10000)/10^3 ; E <- c(1,0)
> cbind( D, Y, E)

      D  Y E
[1,] 31 10 1
[2,] 26 10 0

> mr <- glm( D ~ factor(E), offset=log(Y), family=poisson )
> ci.exp( mr )

            exp(Est.)      2.5%    97.5%
(Intercept)  2.600000 1.7702679 3.818631
factor(E)1   1.192308 0.7079898 2.007935
```

- Response variable: `D` — no. cases in each group
- Offset variable: `log(Y)` — log-person-years
  note the scaling to units desired for intercept (the rate)
- Explanatory variable: `factor(E)`

```
> mR <- glm( D ~ factor(E)-1, offset=log(Y), family=poisson )
> ci.exp( mR )

           exp(Est.)     2.5%    97.5%
factor(E)0       2.6 1.770268 3.818631
factor(E)1       3.1 2.180125 4.408004
```

- Response variable: `D` — no. cases in each group
- Offset variable: `log(Y)` — log-person-years
  note scaling to units desired for intercept
- Explanatory variable: `factor(E) - 1`
  omit intercept: rates separately for each group.
- `ci.exp` transforms back to rate scale.

```
> mR <- glm( D/Y ~ factor(E)-1, weight=Y, family=poisson )
> ci.exp( mR )

           exp(Est.)     2.5%    97.5%
factor(E)0       2.6 1.770268 3.818631
factor(E)1       3.1 2.180125 4.408004
```

- Response variable: `D/Y` — rate in each group
- Weight variable: `Y` — person-years, inversely proportional to variance of the rate
- Explanatory variable: `factor(E) - 1`
  omit intercept: rates separately for each group.
- `ci.exp` transforms back to rate scale.

Rate difference in Poisson model

```
> mD <- glm( D/Y ~ factor(E)-1, weight=Y, family=poisson(link="identity") )
> ci.exp( mD, Exp=FALSE )

           Estimate     2.5%    97.5%
factor(E)0      2.6 1.600611 3.599389
factor(E)1      3.1 2.008738 4.191262
```

- Response variable: `D/Y` — rate in each group
- Weight variable: `Y` — person-years, inversely proportional to variance of the rate
- Explanatory variable: `factor(E) - 1`
  omit intercept: rates separately for each group.
- `ci.exp` with `Exp=FALSE` keeps estimate on the rate scale.

Rate difference in Poisson model

```
> md <- glm( D/Y ~ factor(E), weight=Y, family=poisson(link="identity") )
> ci.exp( md, Exp=FALSE )

             Estimate       2.5%    97.5%
(Intercept)       2.6  1.6006105 3.599389
factor(E)1        0.5 -0.9797404 1.979740
```

- Response variable: `D/Y` — rate in each group
- Weight variable: `Y` — person-years, inversely proportional to variance of the rate
- Explanatory variable: `factor(E)`
  rate in reference group and rate difference.
- `ci.exp` with `Exp=FALSE` keep estimate on the rate scale.

Analysis of proportions

- Suppose we have cohort data with a **fixed risk period**, i.e. all subjects are followed over the same period and therfore has the same length, as well as no losses to follow-up (no censoring).
- In this setting the **risk**, π, of the disease over the risk period is estimated by simple
- **incidence proportion** (often called "**cumulative incidence**" or even "**cumulative risk**")

Analysis of proportions

Incidence proportion:

$$\begin{aligned} \widehat{\pi} &= p = \frac{x}{n} \\ &= \frac{\text{number of new cases during period}}{\text{size of population-at-risk at start}} \end{aligned}$$

Analogously, empirical **prevalence** (proportion) p at a certain point of time t

$$p = \frac{\text{no. of prevalent cases at } t}{\text{total population size at } t} = \frac{x}{n}$$

Analysis of proportions

- Proportions (unlike rates) are dimensionless quantities ranging from 0 to 1
- Analysis of proportions based on **binomial distribution**
- Standard error for an estimated proportion:

$$\mathrm{SE}(p) = \sqrt{\frac{p(1-p)}{n}} = p \times \sqrt{\frac{(1-p)}{x}}$$

- Depends also inversely on x!
- . . . but not a good approximation...

Analysis of proportions

- CI : $p \pm 2 \times \mathrm{SE}(p)$ are within $[0;1]$ if $x > 4/(1+4/n)$
- This is always true if $x > 3$ (if $x > 2$ for $n < 12$)
- — but the approximation is not good for $x < 10$

```
> ci <- function(x,n) round(cbind( x, n, p=p<-x/n, lo=p-2*sqrt(p*(1-p)/n),
+                                                  hi=p+2*sqrt(p*(1-p)/n) ),4)
> rbind(ci(3,11:13),ci(2,3:5),ci(1,1:2))
```

```
     x  n      p      lo     hi
[1,] 3 11 0.2727  0.0042 0.5413
[2,] 3 12 0.2500  0.0000 0.5000
[3,] 3 13 0.2308 -0.0029 0.4645
[4,] 2  3 0.6667  0.1223 1.2110
[5,] 2  4 0.5000  0.0000 1.0000
[6,] 2  5 0.4000 -0.0382 0.8382
[7,] 1  1 1.0000  1.0000 1.0000
[8,] 1  2 0.5000 -0.2071 1.2071
```

Analysis of proportions

- Use confidence limits based on symmetric (normal) $\log(\mathrm{OR})$:
- Compute error factor:

$$\mathrm{EF} = \exp\left(1.96/\sqrt{np(1-p)}\right)$$

- then use to compute confidence interval:

$$p/\left(p + (1-p) \overset{\times}{\div} \mathrm{EF}\right)$$

- Observed $x = 4$ out of $n = 25$: $\hat{p} = 4/25 = 0.16$
- Naive CI: $0.16 \pm 1.96 \times \sqrt{0.16 \times 0.84/25} = [0.016; 0.304]$
- Better: $\mathrm{EF} = \exp(1.96/\sqrt{25 \times 0.16 \times 0.84}) = 2.913$

$$\mathrm{CI} : 0.16/\left(0.16 + (0.84 \overset{\times}{\div} 2.913)\right) = [0.061; 0.357]$$

Analysis of proportions by `glm`

- Default is to model $\text{logit}(p) = \log(p/(1-p))$, log-odds
- Using `ci.exp` gives odds (ω):

$$\omega = p/(1-p) \quad \Leftrightarrow \quad p = \omega/(1+\omega)$$

```
> x <- 4 ; n <- 25
> p0 <- glm( cbind( x, n-x ) ~ 1, family=binomial )
> ( odds <- ci.exp( p0 ) )

            exp(Est.)       2.5%     97.5%
(Intercept) 0.1904762 0.06538417 0.5548924

> odds/(odds+1)

            exp(Est.)       2.5%     97.5%
(Intercept)      0.16 0.06137145 0.3568687
```

Analysis of proportions by `glm`

- Default is to model $\text{logit}(p) = \log(p/(1-p))$, log-odds
- Using `ci.exp` gives odds (ω):

$$\omega = p/(1-p) \quad \Leftrightarrow \quad p = \omega/(1+\omega)$$

```
> x <- 4 ; n <- 25
> p0 <- glm( cbind( x, n-x ) ~ 1, family=binomial )
> ( odds <- ci.exp( p0 ) )

            exp(Est.)       2.5%     97.5%
(Intercept) 0.1904762 0.06538417 0.5548924

> odds/(odds+1)

            exp(Est.)       2.5%     97.5%
(Intercept)      0.16 0.06137145 0.3568687
```

Analysis of proportions by `glm`

Also possible to model $\log(p)$, log-probability, by changing the link function:

```
> x <- 4 ; n <- 25
> pl <- glm( cbind( x, n-x ) ~ 1, family=binomial(link="log") )
> ci.exp( pl )

            exp(Est.)       2.5%     97.5%
(Intercept)      0.16 0.06517056 0.3928154
```

We see that the estimated probability is the same but the confidence limits are slightly different.

Rate ratio in case-control study

Parameter of interest: $\rho = \lambda_1/\lambda_0$
— same as in cohort study.

Case-control design:

- **incident cases** occurring during a given period in the source population are collected,
- **controls** are obtained by *incidence density sampling* from those at risk in the source.
- **exposure** is ascertained in cases and chosen controls.

Rate ratio in case-control study

Summarized data on outcome:

Exposure	Cases	Controls
yes	D_1	C_1
no	D_0	C_0

- Can we directly estimate the rates λ_0 and λ_1 from this?
- — and the ratio of these?
- NO and YES (respectively)
- Rates are not estimable from a case-control design

Rate ratio in case-control study

- If controls are representative of the person- years in the population, their division into exposure groups estimates the exposure distribution of the person-years:

$$C_1/C_0 \approx Y_1/Y_0$$

- Hence, we can estimate the RR by the OR:

$$\widehat{\mathrm{RR}} = \mathrm{OR} = \frac{D_1/Y_1}{D_0/Y_0} = \frac{D_1/D_0}{Y_1/Y_0} \approx \frac{D_1/D_0}{C_1/C_0} = \frac{D_1/C_1}{D_0/C_0}$$

$\Rightarrow$ RR estimated by the ratio of the case-control ratios (D/C)

- . . . but of course there is a penalty to pay. . .

Rate ratio from case-control study

Standard error for $\log(\mathrm{OR})$, 95% error factor and approximate CI for ρ:

$$\begin{aligned} \mathrm{SE}\big(\log(\mathrm{OR})\big) &= \sqrt{\frac{1}{D_1}+\frac{1}{D_0}+\frac{1}{C_1}+\frac{1}{C_0}} \\ \mathrm{EF} &= \exp\Big(1.96 \times \mathrm{SE}\big(\log(\mathrm{OR})\big)\Big) \\ \mathrm{CI} &= [\mathrm{OR}/\mathrm{EF}, \mathrm{OR} \times \mathrm{EF}] \end{aligned}$$

NB. Random error again depends inversely on numbers of cases **and** controls — the penalty, in the two exposure groups.

Example: mobile phone use and brain cancer

(Inskip et al. NEJM 2001; 344: 79-86).

Daily use	Cases	Controls
≥ 15 min	35	51
no use	637	625

The RR associated with use of mobole phone longer than 15 min (vs. none) is estimated by the OR:

$$\mathrm{OR} = \frac{35/51}{637/625} = 0.67$$

Example: mobile phone use and brain cancer

SE for $\log(\mathrm{OR})$, 95% error factor and approximate CI for ρ:

$$\begin{aligned}
\mathrm{SE}\big(\log(\mathrm{OR})\big) &= \sqrt{\frac{1}{35}+\frac{1}{637}+\frac{1}{51}+\frac{1}{625}} = 0.2266 \\
\mathrm{EF} &= \exp()1.96 \times 0.2266) = 1.45 \\
\mathrm{CI} &= [0.67/1.45, 0.67 \times 1.45] = [0.43, 1.05]
\end{aligned}$$

N.B. model-adjusted estimate (with 95% CI):

$$\mathrm{OR} = 0.6[0.3, 1.0]$$

OR from binomial model

```
> Ca <- c(638,35); Co <- c(625,51); Ex <- factor(c("None",">15"),levels=c("None",'
> data.frame( Ca, Co, Ex )

   Ca  Co   Ex
1 638 625 None
2  35  51  >15

> mf <- glm( cbind(Ca,Co) ~ Ex, family=binomial )
> ci.exp( mf )

            exp(Est.)      2.5%    97.5%
(Intercept) 1.0208000 0.9141876 1.139845
Ex>15       0.6722909 0.4311979 1.048185
```

- Intercept is meaningless; only exposure estimate is relevant
- The parameter in the model is $\log(\mathrm{OR})$, so using `ci.exp` gives us the estimated OR — same as in the hand-calculation above.
- This is called **logistic regression**

Extensions and remarks

- All these methods extend to crude analyses of exposure variables with several categories when each exposure category is separately compared to a reference group.
- Evaluation of possible monotone trend in the parameter over increasing levels of exposure: estimation of regression slope.
- CI calculations here are based on simple approximate formulas (**Wald statistics**):
  - accurate when numbers of cases are large
  - for small numbers, other methods may be preferred (e.g. "exact" or likelihood ratio-based as shown by `glm`).
- Crude analysis is insufficient in observational studies: control of confounding needed.

Stratified analysis

Bendix Carstensen & Esa Laara

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,
August 2017 / Januay 2018

`http://BendixCarstensen.com/NSCE/2017`

Stratified analysis

- Shortcomings of crude analysis
- Effect modification
- Confounding
- Steps of stratified analysis
- Estimation of rate ratio
- Mantel-Haenszel estimators
- Matched case-control study

Shortcomings of crude analysis

Crude analysis is misleading, if

- the rate ratio for the risk factor of interest is not constant, but varies by other determinants of the disease
  - . . . *i.e.* heterogeneity of the comparative parameter or **effect modification**
- the exposure groups are not comparable w.r.t. other determinants of disease
  - . . . *i.e.* bias in comparison or **confounding**
- Different cases of a model with effects of
  - primary variable ("exposure")
  - secondary variable ("stratum")
  - **effect modification** is the interaction model
  - **confounding** is the main-effects model

Remedies

Simple approach for remedy:

- **Stratification** of data
  by potentially modifying and/or confounding factor(s)
  & use of **adjusted** estimators
- Conceptually simpler,
  and technically less demanding approach is
  regression modelling
- Regression modeling is feasible because we have computers.

Effect modification

Example: True incidence rates (per 10^5 y) of lung cancer by occupational asbestos exposure and smoking in a certain population:

Asbestos	Smokers	Non-smokers
exposed	600	60
unexposed	120	12
Rate ratio	5	5
Rate difference	480	48

Is the effect of asbestos exposure the same or different in smokers than in non-smokers?

Effect modification (cont'd)

Depends how the effect is measured:

- Rate ratio: constant or **homogenous**
- Rate difference: **heterogenous**:
  The value of rate difference is modified by smoking.

Smoking is thus an **effect modifier** of asbestos exposure on the absolute scale but not on the relative scale of comparison.

Example: Incidence of CHD (per 10^3 y) by risk factor E and age:

Factor E	Young	Old
exposed	4	9
unexposed	1	6
rate ratio	4	1.5
rate difference	3	3

- Rate ratio modified by age
- Rate difference not modified.

There is no such thing as interaction without reference to the **effect scale** (*e.g.* additive or multiplicative)

Effect modification (cont'd)

- Usually comparative parameters are more or less heterogenous across categories of other determinants of disease
- This is termed **interaction** or **effect modification**
- The effect of X depend on the level of Z
- The effect of X cannot be described by a single number,
- . . . it is a function of Z

Example:

Age-specific CHD mortality rates (per 10^4 y) and numbers of cases (D) among British male doctors by cigarette smoking, rate differences (RD) and rate ratios (RR) (Doll and Hill, 1966).

	Smokers		Non-smokers			
Age (y)	rate	D	rate	D	RD	RR
35-44	6.1	32	1.1	2	5	5.7
45-54	24	104	11	12	13	2.1
55-64	72	206	49	28	23	1.5
65-74	147	186	108	28	39	1.4
75-84	192	102	212	31	-20	0.9
Total	44	630	26	101	18	1.7

Example (cont'd)

Both comparative parameters appear heterogenous:

- RD increases by age (at least up to 75 y)
- RR decreases by age

No single-parameter (common rate ratio or rate difference) comparison captures adequately the joint pattern of rates.

Evaluation of modification

- Modification or its absence is an inherent property of the phenomenon:
- cannot be removed or "adjusted" for
- but it depends on the **scale** on which it is measured
- Before looking for effect-modification:
  - what scale are we using for desciption of effects
  - how will we report the modified effects (the interaction)

Evaluation of modification (cont'd)

- statistical tests for heterogeneity insensitive and rarely helpful
- $\Rightarrow$ tempting to assume "no essential modification":
  - $+$ simpler analysis and result presentation,
  - $-$ misleading if essential modification present.

Confounding - example

Observational clinical study with comparison of success of treatment between two types of operation for treating renal calculi:

- OS: open surgery (invasive)
- PN: percutaneous nephrolithotomy (non-invasive)

Treatment	Pts	Op. OK	% OK	%-diff.
OS	350	273	**78**	
PN	350	290	**83**	+5

PN appears more succesful than OS?

Example (cont'd)

Results stratified by initial diameter size of the stone:

Size	Treatment	Pts	Op. OK	% OK	%-diff.
< 2 cm:	OS	87	81	**93**	
	PN	270	235	**87**	-6
≥ 2 cm:	OS	263	192	**73**	
	PN	80	55	**69**	-4

OS seems more succesful in both subgroups.

Is there a paradox here?

Operation example

- Treatment groups are not comparable w.r.t. initial size.
- Size of the stone (SS) is a **confounder** of the association between operation type and success:
  1. an independent determinant of outcome (success), based on external knowledge,
  2. statistically associated with operation type in the study population,
  3. not causally affected by operation type.

Example 13 (cont'd)

- Instance of "confounding by indication":
  — patient status affects choice of treatment,
  $\Rightarrow$ bias in comparing treatments.
- This bias is best avoided in planning:
  — randomized allocation of treatment.

Grey hair and cancer incidence

Age	Gray hair	Cases	P-years ×1000	Rate /1000 y	RR
Total	yes	66	25	2.64	2.2
	no	30	25	1.20	
Young	yes	6	10	0.60	1.09
	no	11	20	0.55	
Old	yes	60	15	4.0	1.05
	no	19	5	3.8	

Observed crude association nearly vanishes after controlling for age.

Means for control of confounding

Design:

- Randomization
- Restriction
- Matching

Means for control of confounding (cont'd)

Analysis:

- Stratification
- Regression modelling

Only randomization can remove confounding due to **unmeasured** factors.

Other methods provide partial removal, but **residual** confounding may remain.

Steps of stratified analysis

- Stratify by levels of the potential confounding/modifying factor(s)
- Compute stratum-specific estimates of the effect parameter (*e.g.* RR or RD)
- Evaluate similarity of the stratum-specific estimates by "eye-balling" or test of heterogeneity.

Steps of stratified analysis (cont.)

- If the parameter is judged to be homogenous enough, calculate an adjusted summary estimate.
- If effect modification is judged to be present:
  - report stratum-specific estimates with CIs,
  - if desired, calculate an adjusted summmary estimate by appropriate standardization — (formally meaningless).

Estimation of rate ratio

- Suppose that true rate ratio ρ is sufficiently homogenous across strata (no modification), but confounding is present.
- Crude RR estimator is biased.
- **Adjusted summary estimator**, controlling for confounding, must be used.
- These estimators are **weighted** averages of stratum-specific estimators.

Adjusted summary estimators

Different weighting methods:

- maximum likelihood (ML)
- weighted least squares (WLS)
- Mantel-Haenszel (MH) weights
- (direct) standardization by external standard population (CMF)
- standardized morbidity ratio (SMR)

Mantel-Haenszel estimators

Cohort study, data summary in each stratum k:

Exposure	Cases	Person-time
yes	D_{1k}	Y_{1k}
no	D_{0k}	Y_{0k}
Total	D_{+k}	Y_{+k}

Compaute stratum-specific rates by exposure group:
$R_{1k} = D_{1k}/\,Y_{1k}, \quad R_{0k} = D_{0k}/\,Y_{0k}$
. . . weighted together to give a common log-RR across strata.

Mantel-Haenszel estimator

- Combination of stratum-specific RRs as a proxy for a model estimate of a common parameter
- Formulae devised in times of the hand-calculator — before the advent of computers
- Replaced by statistical models
- Out of date since about mid-1990s
- . . . but you will still see it occasionally

Gray hair & cancer

```
> D <- c(6,11,60,19)
> Y <- c(10,20,15,5)
> age <- factor( c("Young","Young","Old","Old") )
> hair <- factor( c("Gray","Col","Gray","Col") )
> data.frame( D, Y, age, hair )

   D  Y   age hair
1  6 10 Young Gray
2 11 20 Young  Col
3 60 15   Old Gray
4 19  5   Old  Col
```

Gray hair & cancer

Crude and adjusted risk estimate by Poisson model:

```
> library( Epi )
> ci.exp( glm( D ~ hair      , offset=log(Y), family=poisson ) )

            exp(Est.)      2.5%    97.5%
(Intercept)       1.2 0.8390238 1.716280
hairGray          2.2 1.4288764 3.387277

> ci.exp( glm( D ~ hair + age, offset=log(Y), family=poisson ) )

            exp(Est.)       2.5%     97.5%
(Intercept) 3.7782269 2.49962654 5.7108526
hairGray    1.0606186 0.67013527 1.6786339
ageYoung    0.1470116 0.08418635 0.2567211
```

Case-control study of Alcohol and oesophageal cancer

- Tuyns et al 1977, see Breslow & Day 1980,
- 205 incident cases,
- 770 unmatched population controls,
- Risk factor: daily consumption of alcohol.
- Crude summary:

Exposure ≥ 80 g/d	Cases	Controls	OR
yes	96	109	5.64
no	104	666	

Crude analysis of CC-data

```
> Ca <- c( 96,104)
> Co <- c(109,666)
> Ex <- factor(c(">80","<80"))
> data.frame( Ca, Co, Ex )

   Ca  Co  Ex
1  96 109 >80
2 104 666 <80

> m0 <- glm( cbind(Ca,Co) ~ Ex, family=binomial )
> round( ci.exp( m0 ), 2 )

            exp(Est.) 2.5% 97.5%
(Intercept)      0.16 0.13  0.19
Ex>80            5.64 4.00  7.95
```

The odds-ratio of oesophageal cancer, comparing high vs. low alcohol consumption is $5.64[4.00; 7.95]$

Stratification by age

Age	Exposure ≥ 80 g/d	Cases	Controls	EOR
25-34	yes	1	9	∞
	no	0	106	
35-44	yes	4	26	5.05
	no	5	164	
45-54	yes	25	29	5.67
	no	21	138	
55-64	yes	42	27	6.36
	no	34	139	
65-74	yes	19	18	2.58
	no	36	88	
75-84	yes	5	0	∞
	no	8	31	

NB! Selection of controls: inefficient study

Should have employed stratified sampling by age.

Stratified analysis

```
> ca <- c( 1, 0, 4, 5, 25, 21, 42, 34, 19, 36, 5, 8 )
> co <- c(9, 106, 26, 164, 29, 138, 27, 139, 18, 88, 0, 31)
> alc <- rep( c(">80","<80"), 6 )
> age <- factor( rep( seq(25,75,10), each=2 ) )
> data.frame( ca, co, alc, age )
```

```
   ca  co alc age
1   1   9 >80  25
2   0 106 <80  25
3   4  26 >80  35
4   5 164 <80  35
5  25  29 >80  45
6  21 138 <80  45
7  42  27 >80  55
8  34 139 <80  55
9  19  18 >80  65
10 36  88 <80  65
11  5   0 >80  75
12  8  31 <80  75
```

Stratified analysis

The "age:" operator produces a separate alc-OR for each age class (in the absence of a main effect of alc):

```
> mi <- glm( cbind(ca,co) ~ age + age:alc, family=binomial )
> round( ci.exp( mi ), 3 )
```

```
                  exp(Est.)  2.5%  97.5%
(Intercept)  0.000000e+00 0.000    Inf
age35        2.345328e+10 0.000    Inf
age45        1.170624e+11 0.000    Inf
age55        1.881661e+11 0.000    Inf
age65        3.147003e+11 0.000    Inf
age75        1.985206e+11 0.000    Inf
age25:alc>80 8.547416e+10 0.000    Inf
age35:alc>80 5.046000e+00 1.272 20.025
age45:alc>80 5.665000e+00 2.799 11.464
age55:alc>80 6.359000e+00 3.449 11.726
age65:alc>80 2.580000e+00 1.216  5.475
age75:alc>80 1.755246e+11 0.000    Inf
```

Stratified analysis

...only the relevant parameters:

```
> round( ci.exp( mi, subset="alc" ), 3 )

                exp(Est.)  2.5%  97.5%
age25:alc>80 8.547416e+10 0.000    Inf
age35:alc>80 5.046000e+00 1.272 20.025
age45:alc>80 5.665000e+00 2.799 11.464
age55:alc>80 6.359000e+00 3.449 11.726
age65:alc>80 2.580000e+00 1.216  5.475
age75:alc>80 1.755246e+11 0.000    Inf
```

- The age-specific ORs are quite variable
- Random error in some of them apparently large
- No clear pattern in the interaction

Oesophageal cancer CC — effect modification?

```
> ma <- glm( cbind(ca,co) ~ age + alc, family=binomial )
> anova( mi, ma, test="Chisq" )

Analysis of Deviance Table

Model 1: cbind(ca, co) ~ age + age:alc
Model 2: cbind(ca, co) ~ age + alc
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         0      0.000
2         5     11.041 -5  -11.041  0.05057
```

- Some evidence against homogeneity,
  but no clear pattern in the interaction (effect mdodification)
- Extract a common effect from the reduced model

Oesophageal cancer CC — linear effect modification

```
> ml <- glm( cbind(ca,co) ~ age + alc*as.integer(age), family=binomial )
> round( ci.exp( ml, subset="alc" ), 3 )

                       exp(Est.)  2.5%  97.5%
alc>80                     8.584 1.961 37.579
alc>80:as.integer(age)     0.883 0.609  1.279

> ma <- glm( cbind(ca,co) ~ age + alc, family=binomial )
> anova( mi, ml, ma, test="Chisq" )[1:3,1:5]

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         0      0.000
2         4     10.609 -4 -10.6093  0.03132
3         5     11.041 -1  -0.4319  0.51107
```

Evidence against linear interaction (OR decreasing by age)

Oesophageal cancer CC — effect modification?

```
> mn <- glm( cbind(ca,co) ~ alc      , family=binomial )
> round( ci.exp( mn, subset="alc" ), 2 )

       exp(Est.) 2.5% 97.5%
alc>80      5.64    4  7.95

> ma <- glm( cbind(ca,co) ~ age + alc, family=binomial )
> round( ci.exp( ma, subset="alc" ), 2 )

       exp(Est.) 2.5% 97.5%
alc>80      5.31 3.66   7.7
```

- No clear interaction (effect modification) detected
- Crude OR: $5.64(4.00; 7.95)$
- Adjusted OR: $5.31(3.66; 7.70)$
- **Note:** No test for confounding exists.

Regression models

Bendix Carstensen & Esa Laara

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,
August 2017 / Januay 2018

`http://BendixCarstensen.com/NSCE/2017`

Regression modeling

- Limitations of stratified analysis
- Log-linear model for rates
- Additive model for rates
- Model fitting
- Problems in modelling

Limitations of stratified analysis

- Multiple stratification:
  - many strata with sparse data
  - loss of precision
- Continous risk factors must be categorized
  - loss of precision
  - arbitrary (unreasonable) assumptions about effect shape
- More than 2 exposure categories:
  - Pairwise comparisons give inconsistent results
  - (non)Linear trends not easily estimated

Limitations

- Joint effects of several risk factors difficult to quantify
- Matched case-control studies:
  difficult to allow for confounders & modifiers not matched on.

These limitations may be overcome to some extent by regression modelling.

Key concept: **statistical model**

Log-linear model for rates

Assume that the theoretical rate λ depends on **explanatory variables** or **regressors** X, Z (& U, V, $\dots$) according to a **log-linear** model

$$\log\bigl(\lambda(X, Z, \dots)\bigr) = \alpha + \beta X + \gamma Z + \dots$$

Equivalent expression, **multiplicative model**:

$$\begin{aligned} \lambda(X, Z, \dots) &= \exp(\alpha + \beta X + \gamma Z + \dots) \\ &= \lambda_0 \, \rho^X \tau^Z \cdots \end{aligned}$$

Log-linear model

Model parameters

- $\alpha = \log(\lambda_0) =$ intercept, log-baseline rate λ_0
  (i.e. rate when $X = Z = \cdots = 0$)
- $\beta = \log(\rho) =$ slope,
  change in $\log(\lambda)$ for unit change in X,
  adjusting for the effect of Z (& $U, V, \ldots$)
- $e^{\beta} = \rho =$ rate ratio for unit change in X.

No effect modification w.r.t. rate ratios assumed in this model.

Lung cancer incidence, asbestos exposure and smoking

Dichotomous explanatory variables coded:

- $X =$ asbestos: 1: exposed, 0: unexposed,
- $Z =$ smoking: 1: smoker, 0: non-smoker

Log-linear model for theoretical rates

$$\log\big(\lambda(X, Z)\big) = 2.485 + 1.609X + 2.303Z$$

Log-linear model: Variables

	Rates		Variables			
			X		Z	
Asbestos	Smoke	Non-sm	Smoke	Non-sm	Smoke	Non-sm
exposed	600	60	1	1	1	0
unexposed	120	12	0	0	1	0

Lung cancer, asbestos and smoking

Entering the data:

— note that the data are artificial assuming the no. of PY among asbestos exposed is $1/4$ of that among non-exposed

```
> D <- c( 150, 15, 120,  12 )        # cases
> Y <- c(  25, 25, 100, 100 ) / 100 # PY (100,000s)
> A <- c( 1, 1, 0, 0 ) # Asbestos exposure
> S <- c( 1, 0, 1, 0 ) # Smoking
> cbind( D, Y, A, S )

      D    Y A S
[1,] 150 0.25 1 1
[2,]  15 0.25 1 0
[3,] 120 1.00 0 1
[4,]  12 1.00 0 0
```

Lung cancer, asbestos and smoking

- Regression modelling
- Multiplicative (default) Poisson model
- 2 equivalent approaches
  - D response, $\log(Y)$ offset
  - D/Y response, Y weight
    (warning can be ignored)
  - the latter approach also useful for **additive** models

```
> mo <- glm( D   ~ A + S, offset=log(Y), family=poisson )
> mm <- glm( D/Y ~ A + S, weight=Y, family=poisson )
> ma <- glm( D/Y ~ A + S, weight=Y, family=poisson(link=identity) )
```

Lung cancer, asbestos and smoking

Summary and extraction of parameters:

```
> summary( mo )

Call:
glm(formula = D ~ A + S, family = poisson, offset = log(Y))

Deviance Residuals:
         1           2           3           4
 0.000e+00   0.000e+00  -1.032e-07   0.000e+00

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.4849     0.2031   12.23   <2e-16
A             1.6094     0.1168   13.78   <2e-16
S             2.3026     0.2018   11.41   <2e-16

(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  4.1274e+02  on 3  degrees of freedom
Residual deviance: -1.5987e-14  on 1  degrees of freedom
AIC: 28.37
```

Summary and extraction of parameters

```
> ci.exp( mo )

            exp(Est.)     2.5%     97.5%
(Intercept)        12 8.059539 17.867026
A                   5 3.977142  6.285921
S                  10 6.732721 14.852836

> ci.exp( mo, Exp=F )

            Estimate     2.5%    97.5%
(Intercept) 2.484907 2.086856 2.882957
A           1.609438 1.380563 1.838312
S           2.302585 1.906979 2.698191

> ci.exp( mm, Exp=F )

            Estimate     2.5%    97.5%
(Intercept) 2.484907 2.086856 2.882957
A           1.609438 1.380563 1.838312
S           2.302585 1.906979 2.698191
```

Parameters are the same for the two modelling approaches.

Interpretation of parameters

```
> round( cbind( ci.exp( mm, Exp=F ),
+               ci.exp( mm        ) ), 3 )

            Estimate  2.5% 97.5% exp(Est.)  2.5%  97.5%
(Intercept)    2.485 2.087 2.883        12 8.060 17.867
A              1.609 1.381 1.838         5 3.977  6.286
S              2.303 1.907 2.698        10 6.733 14.853
```

- $\alpha = 2.485 = \log(12)$, log of baseline rate,
- $\beta = 1.609 = \log(5)$, log of rate ratio $\rho = 5$ between exposed and unexposed for asbestos
- $\gamma = 2.303 = \log(10)$, log of rate ratio $\tau = 10$ between smokers and non-smokers.

Rates for all 4 asbestos/smoking combinations can be recovered from the above formula.

Log-linear model: Estimated rates

	Rates		Parameters	
Asbestos	Smokers	Non-smokers	Smokers	Non-smokers
exposed	600	60	$\alpha + \gamma + \beta$	$\alpha + \beta$
unexposed	120	12	$\alpha + \gamma$	α
Rate ratio	5	5	$\log(\beta)$	$\log(\beta)$
Rate difference	480	48	β	β

Log-linear model

Model with effect modification (two regressors only)

$$\log\big(\lambda(X,Z)\big) = \alpha + \beta X + \gamma Z + \delta XZ,$$

equivalently

$$\lambda(X,Z) = \exp\big(\alpha + \beta X + \gamma Z + \delta XZ\big) = \lambda_0 \rho^X \tau^Z \theta^{XZ}$$

where α is as before, but

β = log-rate ratio ρ for a unit change in X when $Z = 0$,

γ = log-rate ratio τ for a unit change in Z when $X = 0$

Interaction parameter

$\delta = \log(\theta)$, interaction parameter, describing effect modification

For binary X and Z we have

$$\theta = e^{\delta} = \frac{\lambda(1,1)/\lambda(0,1)}{\lambda(1,0)/\lambda(0,0)},$$

i.e. the ratio of relative risks associated with X between the two categories of Z.

Log-linear model: Estimated rates

	Rates		Parameters	
Asbestos	Smokers	Non-smokers	Smokers	Non-smoker
exposed	600	60	$\alpha + \gamma + \beta + \delta$	$\alpha + \beta$
unexposed	120	12	$\alpha + \gamma$	α
Rate ratio	5	5	$\log(\beta + \delta)$	$\log(\beta)$
Rate difference	480	48	$\beta + \delta$	β

Lung cancer, asbestos and smoking

```
> mi <- glm( D/Y ~ A + S + I(A*S), weight=Y, family=poisson )
> round( ci.exp( mm ), 3 ) ; round( ci.exp( mi ), 3 )

            exp(Est.)  2.5%  97.5%
(Intercept)        12 8.060 17.867
A                   5 3.977  6.286
S                  10 6.733 14.853

            exp(Est.)  2.5%  97.5%
(Intercept)        12 6.815 21.130
A                   5 2.340 10.682
S                  10 5.524 18.101
I(A * S)            1 0.451  2.217
```

- There is no interaction on the multiplicative scale:
- interaction parameter is 1,
- asbestos and smoking parameters are the same,
- but SEs are larger because they refer to RRs for levels $X = 0$ and $Z = 0$ respectively and not both levels **jointly**

Additive model for rates

General form with two regressors

$$\lambda(X, Z) = \alpha + \beta X + \gamma Z + \delta XZ$$

- $\alpha = \lambda(0, 0)$ is the baseline rate,
- $\beta = \lambda(x + 1, 0) - \lambda(x, 0)$, rate difference for unit change in X when $Z = 0$
- $\gamma = \lambda(0, z + 1) - \lambda(0, z)$, rate difference for unit change in Z when $X = 0$,

Additive model

δ = interaction parameter.

- For binary X, Z:

$$\delta = [\lambda(1,1) - \lambda(1,0)] - [\lambda(0,1) - \lambda(0,0)]$$

- If no effect modification present, $\delta = 0$, and

β = rate difference for unit change in X for all values of Z

γ = rate difference for unit change in Z for all values of X,

Example: Additive model

```
> mai <- glm( D/Y ~ A + S + A*S, weight=Y, family=poisson(link=identity) )
> ci.exp( mai, Exp=FALSE )

            Estimate       2.5%     97.5%
(Intercept)       12   5.210486  18.78951
A                 48  16.886536  79.11346
S                108  85.481728 130.51827
A:S              432 328.808315 535.19168
```

A very clear interaction (effect modification)

$$\lambda(X, Z) = \alpha + \beta X + \gamma Z + \delta XZ = 12 + 48X + 108Z + 432XZ$$

- α $= 12$, baseline rate, i.e. that among non-smokers unexposed to asbestos (reference group),
- β $= 48$ $(60 - 12)$, rate difference between asbestos exposed and unexposed among non-smokers only,
- γ $= 108$ $(= 120 - 12)$, rate difference between smokers and non-smokers among only those unexposed to asbestos
- δ $=$ excess of rate difference between smokers and non-smokers among those exposed to asbestos:
  $\delta = (600 - 120) - (60 - 12) = 432$

Model fitting

Output from computer packages will give:

- parameter estimates and SEs,
- goodness-of-fit statistics,
- fitted values,
- residuals,...

May be difficult to interpret!

Model checking & diagnostics:

- assessment whether model assumptions seem reasonable and consistent with data
- involves fitting and comparing different models

Problems in modelling

- Simple model chosen may be far from the "truth".
- possible bias in effect estimation, — underestimation of SEs.
- Multitude of models fit well to the same data
  which model to choose?
- Software easy to use:
- . . . easy to fit models blindly
- . . . possibility of unreasonable results

Modeling

- Modelling should not substitute but complement crude analyses:
- Crude analyses should be seen as initial modeling steps
- Final model for reporting developed mainly from subject matter knowledge
- Adequate training and experience required.
- Ask help from professional statistician!
- **Collaboration** is the keyword.

Conclusion

Bendix Carstensen & Esa Laara

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,
August 2017 / Januay 2018

`http://BendixCarstensen.com/NSCE/2017`

Concluding remarks

Epidemiologic study is a

Measurement excercise

Target is a **parameter** of interest, like

- incidence rate
- rate ratio
- relative risk
- difference in prevalences

Result: **Estimate** of the parameter.

Estimation and its errors

Like errors in measurement, estimation of parameter is prone to error:

$$\begin{aligned} \text{estimate} \;=\; & \text{true parameter value} \\ & + \text{systematic error (bias)} \\ & + \text{random error} \end{aligned}$$

Sources of bias

- confounding, non-comparability,
- measurement error, misclassification,
- non-response, loss to follow-up,
- sampling, selection

Sources of random error

- biological variation between and within individuals in population
- measurement variation
- sampling (random or not)
- allocation of exposure (randomized or not)

Random sampling

- relevant in **descriptive** studies
- estimation of parameters of occurrence of given health outcomes in a target population
- target population well-defined, finite, restricted by time and space
- representativeness of study population (sample) important

Randomization

- relevant in **causal** studies
- estimation of comparative parameters of **effect** of an exposure factor on given health outcomes
- abstract (infinite) target population
- **comparability** of exposure groups important
- study population usually a convenience sample from available source population

Recommendations

Possible remedies for these problems:

- de-emphasize inferential statistics in favor of pure data decriptors: graphs and tables
- adopt statistical techniques based on realistic probability models
- subject the results of these to influence and sensitivity analysis.

(from Greenland 1990) Interpretation of obtained values of inferential statistics
– not mechanical reporting!

Conclusion

"In presenting and discussing the results of an observational study the greatest emphasis should be placed on bias and confounding." (Brennan and Croft 1994)

Motto (Campbell & Machin 1983):

STATISTICS is about
COMMON SENSE and
GOOD DESIGN!