Measures of disease frequency and effects
Analysis of epidemiological data

Esa Läärä
University of Oulu, Finland
esa.laara@oulu.fi http://stat.oulu.fi/laara

Bendix Carstensen
Steno Diabetes Center, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk http://BendixCarstensen.com

Nordic Summer School in Cancer Epidemiology
August 2011, Danish Cancer Society, Copenhagen
January 2012, Virrat, Finland
http://BendixCarstensen.com/NSCE

Introduction
Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
http://BendixCarstensen.com/NSCE

Key references

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

B&D: Breslow, N.E., Day, N.E. (1987).
Statistical Methods in Cancer Research Volume II – The Design and Analysis of Cohort Studies.
IARC Scientific Publications No. 82, IARC, Lyon.

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

Internet resources on cancer statistics

NORDCAN : Cancer Incidence and Mortality in the Nordic Countries, Version 4.0. Association of Nordic Cancer Registries, Danish Cancer Society, 2002. `http://www-dep.iarc.fr/nordcan.htm`
NORDCAN is a graphical package providing data on the incidence of, and mortality from 40 major cancers for 80 regions of the Nordic countries (Denmark, Finland, Iceland, Norway and Sweden). Using NORDCAN, these data can be presented as a variety of tables and graphs that can be easily exported or printed. NORDCAN allows countries and cancer sites to be grouped and compared as desired.

GLOBOCAN 2008 : Cancer Incidence and Mortality Worldwide in 2008 `http://globocan.iarc.fr/`

Basic Concepts

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

What is Epidemiology?

Some textbook definitions of epidemiology:
Greek: *epi* = upon, *demos* = people

- "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,..." (Last (ed.) Dictionary of Epidemiology, 2000)
- "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: *frequentia* = assembly, multitude, crowd.

1. Also, **frequency**. the state or fact of being frequent; frequent occurrence. We are alarmed by the frequency of fires in the neighborhood.
2. Rate of occurrence:
   The doctor has increased the frequency of his visits.
3. *Physics*: number of periods or ... regularly occurring events ... of any given kind in unit of time, usually in one second.
4. *Math*: the number of times a value recurs in a unit change of the independent variable of a given function.
5. *Statistics*: the number of items occurring in a given category. Cf. **relative frequency**.

Meanings 2 and 5 are both relevant in epidemiology.

But what is "rate" and "occurrence"?

Cancer i 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? — **prevalence**
- will contract a new breast ca. during 2007? — **incidence**
- die from breast ca. in 2007? — **mortality**
- will be alive after 5 years since diagnosis among those getting breast ca. in 2007? — **survival**
- are cured from breast cancer during 2007? — **cure**

Questions on frequency & occurrence

- What is the relative frequency or/and rate of occurrence of these states and events?
- How great are the **risks** of these events?
- Is the frequency/occurrence/risk of breast cancer greater among nulliparous than 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?

What is risk?

What do we mean by "risk of disease S"?

(a) **probability** of *getting* S during a given **risk period**
→ **incidence** probability,

(b) **rate** of change of that probability
→ **hazard** or intensity,

(c) **probability** of *carrying* S at a given *time point*
→ **prevalence** probability.

Most commonly meaning (a) 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 (b): 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 a "personal" 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**.

Types of epidemiologic studies

Can crudely be classified along the following axes:

- *study question*: descriptive $\leftrightarrow$ causal
- *study unit*: individual $\leftrightarrow$ aggregate (ecological study)
- *allocation of exposure*: experimental $\leftrightarrow$ observational
- *population*: closed (cohort) $\leftrightarrow$ open (dynamic)
- *dimensionality*: cross-sectional $\leftrightarrow$ longitudinal
- *timing of observations*: concurrent $\leftrightarrow$ historical ("pro-" vs. "retrospective")
- *sampling of exposure data*: cohort $\leftrightarrow$ case-control

Focus in this course: *observational*, and *longitudinal cohort and case-control* studies.

Descriptive and causal questions

Descriptive: What is the occurrence of outcome C in different population groups.
— Medical demography

Descriptive (II) — groups defined *e.g.* by exposure to a determinant or risk factor X?

Causal (also **etiological** or "analytical"): What is the occurrence of outcome C in a population exposed to risk factor X as compared to ... what the occurrence in the same population *would have been, if not* exposed?

N.B.: Causal question — *counterfactual conditional*!

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

Experimental and observational studies

Allocation of exposure in etiologic studies?

• **Experimental**: Exposure controlled by investigators, its levels being **randomized** among the study subjects.

+ **Comparability** of exposure groups.
+ Feasible in clinical and preventive trials.
– Ethically impossible for hazardous exposures.

• **Observational**: Exposure imposed by the own behaviour of the subjects themselves & and by their environment.

– Possibility of **confounding**: due to other determinants of the outcome, correlated with exposure.
* Challenges: **Valid**: and **efficient** non-randomized design and statistical analysis.

Experimental and observational studies

Allocation of exposure or risk factor in causal studies?

Experimental (Intervention trial): Exposure is controlled by investigators; its levels are allocated among recruited subjects by **randomization**,

⇒ **comparability** of exposure groups.

Observational: Exposure imposed by own behaviour of study subjects and/or by their environment,

⇒ possibility of **confounding** due to other determinants.

Time dimensionality of a study

Cross-sectional: Outcome *status* and its *prevalence* in population at given *time point* are studied, *e.g.*

- number of Danish citizens living with existing cancer on 13 August 2007.

Longitudinal: *Change* in health status, like the *incidence* of new cases over a *time period* is of interest, *e.g.*

- number of Danish citizens getting a new cancer diagnosed during year 2007.

Causal question ⟶ longitudinal study preferred.

Study population & study base

Types of **study population** & its membership defined

- **closed – cohort**: members taken by certain event, *e.g.*
  1. birth cohort, people born during same year,
  2. workers employed by Carlsberg brewery during 1970's, followed up since then, even after retirement
- **open – dynamic**: defined by changeable status, *e.g.*
  1. citizens of Copenhagen, currently resident;
  2. *catchment population* of the Oncological Clinic at Rigshospitalet (CPH),

Study base = study population × its experience in time.

Study base (SB): population experience

Cross-sectional: SB = study population at a *time point*,

Longitudinal: SB comprises **follow-up times** of individuals in the study population over a given *period*.

Cohort: Follow-up time = period from **entry** until a single **exit** at which **outcome** or **censoring** occurs.

Dynamic: Follow-up time consists of possibly several periods of membership since the first entry until the final exit.

- Follow-up calculation complicated.
- Approximation by *mid-population*.

Matematical reminder

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

Logarithms and exponentials

$$10^2 = 10 \times 10$$
$$10^3 = 10 \times 10 \times 10$$
$$10^2 \times 10^3 = 10^5$$
$$10^3/10^2 = 10^1$$
$$(10^3)^2 = 10^6$$

$$10^2/10^2 = 10^0 = 1$$
$$10^2/10^3 = 10^{-1} = 1/10$$
$$10^{1/2} \times 10^{1/2} = 10^1$$
$$10^{1/3} \times 10^{1/3} \times 10^{1/3} = 10^1$$

$$10^{0.3010} = 2$$
$$\log_{10}(2) = 0.3010$$

$$10^{0.4771} = 3$$
$$\log_{10}(3) = 0.4771$$

Multiplication and division

$$2 \times 3 = 6$$

$$10^{0.3010} \times 10^{0.4771} = 10^{0.7781}$$
$$10^{0.7781} = 6$$

$$\log_{10}(2) = 0.3010$$
$$\log_{10}(3) = 0.4771$$
$$0.3010 + 0.4771 = 0.7781$$
$$10^{0.7781} = 6$$

In general: $\log(xy) = \log(x) + \log(y)$

$$\log(x/y) = \log(x) - \log(y)$$
$$\log(x^a) = a \log(x)$$

Natural logarithms $e = 2.7183$

$$\log_e(e) = 1$$

$$e^{0.6931} = 2$$
$$\log_e(2) = 0.6931$$

$$e^{1.0986} = 3$$
$$\log_e(3) = 1.0986$$

$$2 \times 3 = 6$$

$$e^{0.6931} \times e^{1.0986} = e^{1.7918}$$
$$e^{1.7918} = 6$$

In general: $e^x \times e^y = e^{x+y}$

$$e^x/e^y = e^{x-y}$$
$$(e^x)^y = e^{x \times y}$$

Names for the logartithms

Engineers and calculators:

$\log$ is the logarithm to base 10.

$\ln$ is the logarithm to base e, the natural log

Matematicians:

$\log$ is the logarithm to base e, the natural log

$\log_{10}$ is the logarithm to base 10.

We use $\log$ for the natural logarithm, and explicitly $\log_{10}$ when this is needed.

Why natural logarithms?

For small values of x: $e^x \approx 1 + x$

$$e^{-x} \approx 1 - x$$

$$\ln(1 + x) \approx x$$

$$\ln(1 - x) \approx -x$$

For example: $\ln(1.01) = 0.01$

$$\ln(0.99) = -0.01$$

But: $\log_{10}(1.01) = 0.4343 \times 0.01$

$$\log_{10}(0.99) = 0.4343 \times -0.01$$

In general: $\log_{10}(x) = 0.4343 \times \ln(x)$

R and how we use it

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

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.
- Free.
- Runs on any computer.
- Updated every 6 months.

A simple calculator

R lets you enter simple arithmetic and giver you back the ansver straightaway:

```
> 5+8
[1] 13
> sqrt( 1/12 + 1/17 )
[1] 0.3770370
> exp( 1.96 * sqrt( 1/12 + 1/17 ) )
[1] 2.093825
> D0 <- 12
> D1 <- 17
> exp( 1.96 * sqrt( 1/D0 + 1/D1 ) )
[1] 2.093825
```

Handy in daily life too.

A smart calculator

Case-control study of MI:

	Men		Women	
PA index	Case	Cont	Case	Cont
$2500+$ kcals	141	208	49	58
< 2500 kcals	144	112	32	45
Total	285	320	81	103

```
> (141/208)/(144/112)
[1] 0.5272436
> (49/58)/(32/45)
[1] 1.188039
```

A smart calculator

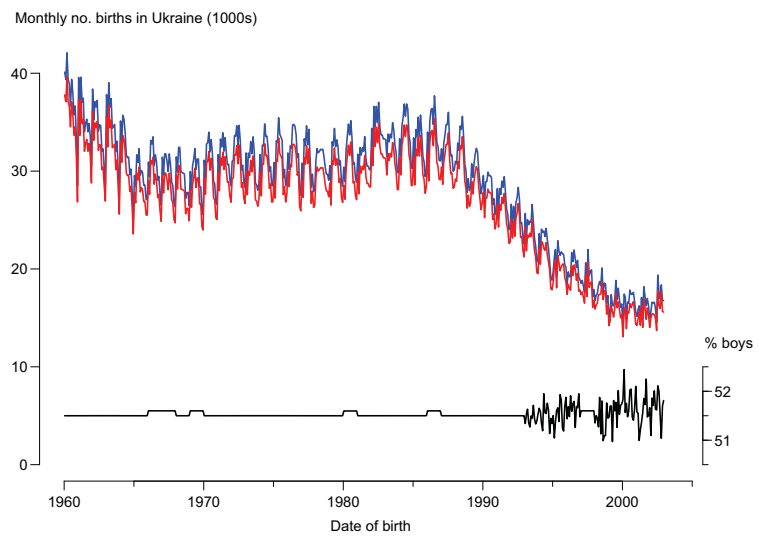
```
> D1 <- c(141, 49)
> D0 <- c(144, 32)
> H1 <- c(208, 58)
> H0 <- c(112, 45)
> OR <- (D1/D0)/(H1/H0)
> OR
[1] 0.5272436 1.1880388
```

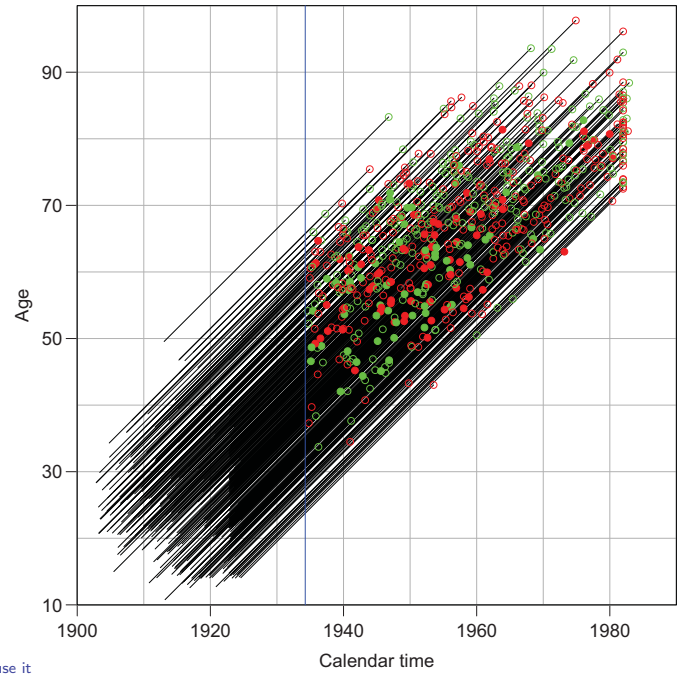
Things done in parallel for the two exposure groups.

R for epidemiology

Versatile graphics:

- Simple graphs easy
- Complicated graphs possible
- You can add things to a graph
- Interactive graphs:
  - Put things on with the mouse
  - Identify points with the mouse





Getting your graphs out

You can save graphs to disk and later fetch them into your documents in almost any format you like:
(`.eps`, `.pdf`, `.emf`, `.bmp`, `.png`).

You can choose to save graphs from the screen or to write directly to a file.

Tools for anything!

- More than 1500 add-on packages.
- Several packages for epidemiology:
  - `Epi`: Mostly chronic disease epidemiology:
    - Cohort studies, split follow-up time
    - Lexis diagram, sevral timescales
    - Multistate model support
    - Advanced tabulation
    - Parameter reporting
  - `epicalc`: For a book by Virasakdi Chongsuvivatwong.
  - `epitools`: Mostly infectious diseases.
  - `epiR`: Leaning towards veterinay epidemiology.
- Install and update packages from within R.

Versatility is paid by steep learning curve

Command line interface:

- You must write commands
- You must know what they are called
- Easy to repeat analyses, because you always have a script of what you did.
- There is a simple editor built into R.
- A good workbook introduction is: `www.mhills.pwp.blueyonder.co.uk/Rwork_book.html`
- Many other introductions to R on the R homepage.

R in this course

- Only use R as a simple calculator.
- No need for for a lot of fancy stuff.
- The script editor (we will show you what that is) will help you keep your solutions for future reference.
- A short recap of exercises tomorrow morning, and tomorrow afternoon.
- After the course, solutions to all exercises will be provided.

Frequency measures

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

Measuring frequency: Cases, population, time

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.
- Cross-sectional – **prevalence** measures.

General form of frequency or occurrence measures

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

Numerator: number of cases observed in the population — at a certain time point or during a specified period.

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

Numerator and denominator must cover the *same population*.

Prevalence

Prevalence:
Point prevalence, is the proportion of existing cases (old and new) in a population at a single point of time.

$$P = \frac{\text{No. of existing cases in a population at one point of time}}{\text{No. of people in the population at the same point of time}}$$

This measure is called point prevalence, because it refers to a single point in time. It is often referred to simply as prevalence.

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** or **cumulative risk** (*e.g.* by **IS**).

Indidence rate (I) over a defined observation period:

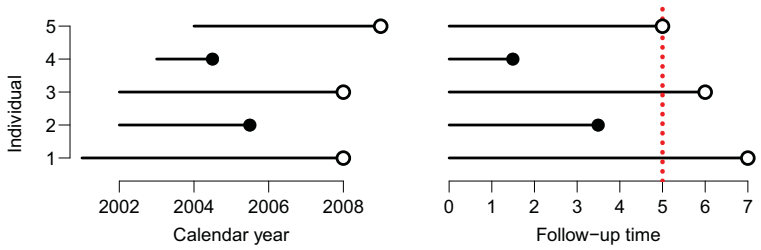
$$I = \frac{\text{number of incident (new) cases during period}}{\text{sum of follow-up times of pop'n at risk}}$$

Also called **incidence density** or **hazard**.

Later we will provide a more precise mathematical definition of the concepts.

Example: Follow-up of a small cohort

- $\circ$ = exit with censoring; outcome not observed,
- $\bullet$ = exit with outcome event (disease onset) observed



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

No censoring in the 5-year risk period $\Rightarrow$ can calculate:

$$\text{Inc. prop.} = 2/5 = 0.4 \ (40 \ \%)$$

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** (see below),
- 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; it is a scaled quantity; it is measured in time^{-1}: cases/1000 Y, say.
- 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:

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

(we shall return to the derivation of this).

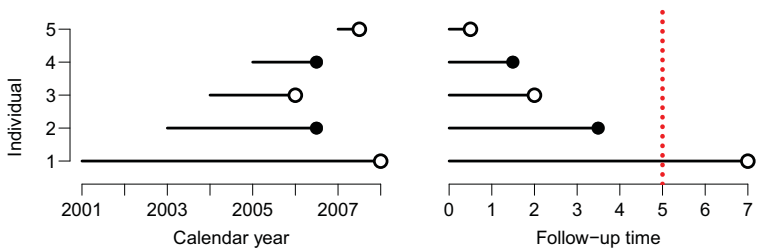
Competing events and censoring

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 incidence and risk calculations.

Follow-up of another small cohort



Two censored observations $\Rightarrow$ can calculate the rate:

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

but the 5-year Q **is no more** $2/5$!
However, under constant rate model

$$Q = 1 - \exp(-5 \times 2/12.5) = 0.55$$

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:

- Let the population size be N_{t-1} at start and N_t at the end of the observation period t with length L_t years,
- Mid-population for the period: $\bar{N}_t = \frac{1}{2} \times (N_{t-1} + N_t)$.
- Approximate person-years: $Y_t \approx \bar{N}_t \times L_t$.

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):

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

Relationships between incidence measures

With constant incidence rate over risk period (length $= \Delta$), incidence proportion Q and rate I are related:

$$\begin{aligned} Q &= 1 - \exp(-I \times \Delta) \approx I \times \Delta \\ I &= -\log(1 - Q)/\Delta \approx Q/\Delta, \end{aligned}$$

The approximations are good when

- the incidence proportion is "small" (under 10 %).
  - incidence rate (I) is small
  - the risk period (Δ) is small

Mortality

Cause-specific mortality from disease C is described by **mortality rate** (and proportion), defined like I (and Q), but

- cases are only *deaths* from cause C, and
- follow-up is extended until death (from *any* acuse) or censoring

The cumulative risk of death from a given cause (cause-specific mortality proportion/risk) requires correction for *competing events*. **Total mortality**: cases are deaths from any cause. Mortality depends on the incidence and the **prognosis** or fatality of the disease, *i.e.* the **survival** of those affected.

Theoretical concepts behind incidences

Analysis of incidences
= analysis of **time to event** or **failure time** or **survival** data.

Mathematical concepts:

$$
\begin{aligned}
T &= \text{time to outcome event – random variable,} \\
S(t) &= P(T > t) = \textbf{ survival} \text{ function of } T, \\
&= \text{probability of avoiding the event up to given time } t, \\
\lambda(t) &= -S'(t)/S(t) = \textbf{ intensity} \text{ or } \textbf{hazard} \text{ function,} \\
\Lambda(t) &= \int_0^t \lambda(u)du = -\log S(t) = \textbf{cumulative hazard}, \\
F(t) &= 1 - S(t) = 1 - \exp\{-\Lambda(t)\} = \textbf{risk} \text{ function} \\
&= \text{probability of the outcome to occur before } t
\end{aligned}
$$

Intensity or hazard function

Can be viewed as *theoretical incidence rate*. Formally:

$$\lambda(t) = \lim_{\Delta \to 0} \frac{P(t < T \leq t + \Delta \mid T > t)}{\Delta}$$

$\approx$ Probability of outcome event occurring in a short risk period $]t, t + \Delta]$, given "survival" or avoidance of the event up to the start t, divided by the period length — "risk per time".

This is equivalent to saying that over a short interval

$$\text{risk} \approx \text{intensity} \times \text{length of interval}$$

or $$P(t < T \leq t + \Delta \mid T > t) \approx \lambda(t) \times \Delta.$$

Exponential survival times (constant hazard)

Simplest probability model for time to event:

Exponential distribution, Exp(λ), in which

$$\text{rate } \lambda(t) = \lambda \text{ (constant)} \;\Rightarrow\; \text{risk over }]0,t] = 1 - \exp(-\lambda t)$$

Analysis of event data of n individuals. For subject i let

$$\begin{aligned} y_i &= \text{time to event or censoring, total: } Y = \sum y_i \\ d_i &= 1/0\text{-indicator for observing event, total: } D = \sum d_i \end{aligned}$$

Exp(λ) model $\Rightarrow$ **Likelihood function** of λ is equivalent to that when number of cases D is *Poisson*-distributed

(Analysis part of the course)

Basic statistical analysis of empirical rates

Asymptotic statistical inference based on likelihood:

- **Maximum likelihood estimator** (MLE) of λ is

$$\widehat{\lambda} = \frac{D}{Y} = \frac{\text{number of cases}}{\text{total person-time}} = I, \text{ empirical incidence rate!}$$

- **Standard error** of the empirical rate is $I/\sqrt{D}$

$\Rightarrow$ The more cases, the greater is **precision** in rate!

- Approximate **confidence interval** for "true" rate λ:

$$\text{estimator } \pm\ 1.96 \times \text{standard error}$$

More about these issues in the analysis lectures.

Prevalence measures

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

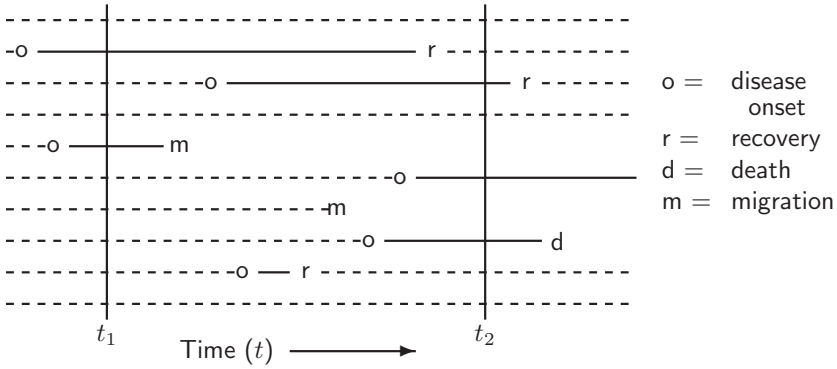
$$P = \frac{\text{number of existing or prevalent cases of } C}{\text{size of the whole population}}$$

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

Relationships between measures

Point prevalence of C at given time point t depends on

- *incidence* of new cases of C before t
- *duration* of C, depending in turn on the probability of **cure** or recovery from C or **survival** of those affected.

Stationary ("stable") population: prevalence (P), incidence (I), and average duration ($\bar{d}$) of C are related:

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

prevalence = incidence $\times$ duration

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

Prevalence of cancer?

Difficult to ascertain, whether and when a cancer is cured.

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

Cancer registry practice: Prevalence of cancer C at time point t in the target population refers to the

number & proportion of population members who

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

Often further classified by years since diagnosis.

Example: Liver and testis cancer

Crude comparison of incidence, mortality and prevalence in the male population of Finland 1999

	Liver	Testis
No. of new cases during 1999	119	103
No. of deaths during 1999	123	8
No. of prevalent cases 1.1.2000	120	1337
– " – diagnosed $<$ 1 y ago	*36*	*97*
– " – diagnosed 1-$<$ 5 y ago	*53*	*291*
– " – diagnosed 5-$<$ 10 y ago	*17*	*304*
– " – diagnosed $>$ 10 y ago	*14*	*642*

Comparative measures

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

Relative and absolute comparisons (IS: Ch 5.2)

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

comparison of occurrence between the *index* ("exposed") and the *reference* ("unexposed") groups or populations by

- relative measures (ratio)
- absolute measures (difference)

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

$\Rightarrow$ **comparative measures** $\approx$ **effect measures**

Relative comparative measures

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

- 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.

Absolute comparative measures

Generic **"excess risk"** btw exposed and unexposed can be

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

Use of relative and absolute comparisons

Ratio – describes the *biological strength* of the exposure

Difference – informs 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
Deaths	30	60
Person-years at risk	7 970	31 940

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

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

$$= \frac{0.00756}{0.00376} = 2.0076$$

Attributable fraction

Combine absolute and relative comparisons.

When incidence is higher for the exposed, we can calculate

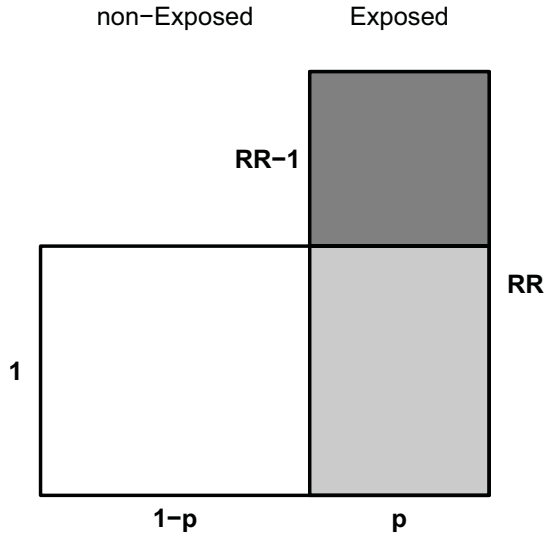
$$\textbf{Excess fraction, EF} = \frac{Q_1 - Q_0}{Q_1} = \frac{\text{RR} - 1}{\text{RR}}$$

also called **attributable fraction**, AF or **attributable risk**.

EF Estimates the fraction out of all new cases among those exposed, which are "caused" by the exposure itself, and which thus could be "avoided" if the exposure were absent

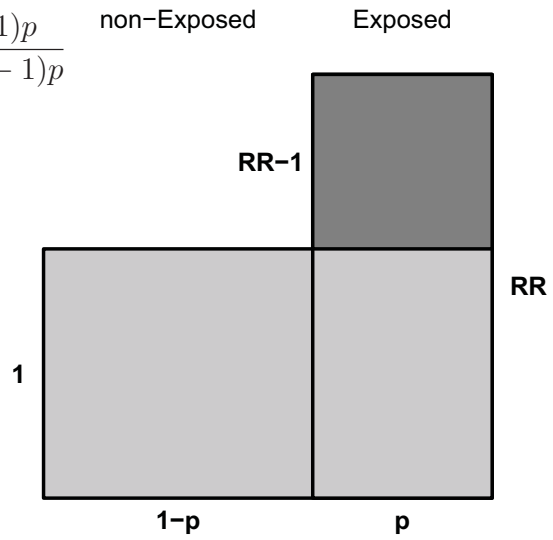
Attributable fraction, AF

$$\text{AF} = \frac{RR - 1}{RR}$$



Population attributable fraction, PAF

$$\text{PAF} = \frac{(RR-1)p}{1+(RR-1)p}$$



Population attributable fraction

If we instead ask:
"How large a fraction of **all** cases would be prevented if exposure was abolished?".

Depends on the fraction of the population which is exposed

$$\text{PAF} = \frac{(\text{RR}-1)p}{1+(\text{RR}-1)p}$$

PAF Estimates the fraction out of all new cases, which are "caused" by the exposure itself, and which thus could be "avoided" if the exposure were absent.

AF is a "biological" measure.
PAF is a "population level" measure.

Measures of potential impact (cont'd)

When the exposed have a lower incidence, we can calculate

$$\textbf{Preventive fraction, PF} = \frac{Q_0 - Q_1}{Q_0} = 1 - \text{RR}$$

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 (exposed) vs. no intervention (unexposed).

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.

Time scales

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

Incidence by age, calendar year, and other time variables

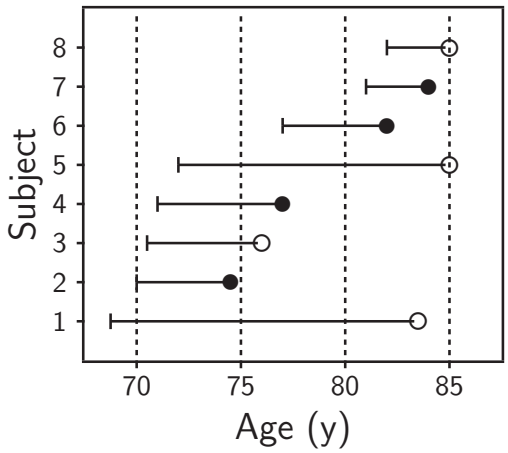
Incidence can be studied on various **time scales**, *e.g.*:

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 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.

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

Lung cancer incidence rates in Finland by age, period and cohort

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.
- Diagonals: age-incidence pattern in birth cohorts.

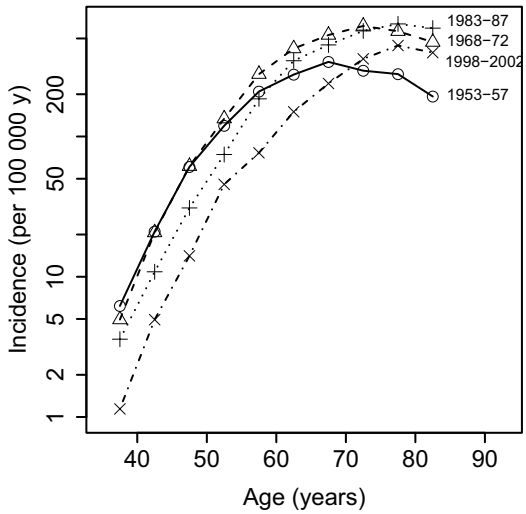
Incidence by age, calendar time & birth cohort

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

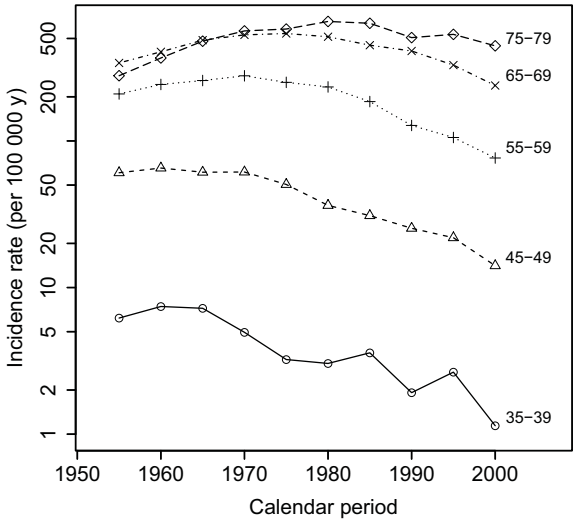
Birth cohort = people born during the same limited time interval, *e.g.* single calendar year, or 5 years period.

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

Age-incidence curves by period (rows)



Time trends by age (columns)

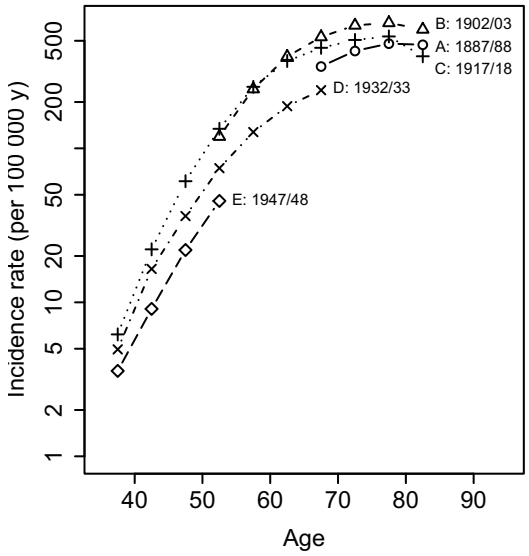


Age-specific rates by birth cohort

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

Age-incidence curves in 5 birth cohorts



Split of follow-up by age and period

Incidence of (or mortality from) disease C in special study cohort (e.g. occupational group, users of certain medicine)

→ often compared to incidence in a *reference* or "general" population

Appropriate adjustment for age and calendar time needed in this, *e.g.* by comparing *observed* to *expected* cases with SIR (see p. 70-71).

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

Example of follow-up

Entry and exit dates for a small cohort of four subjects

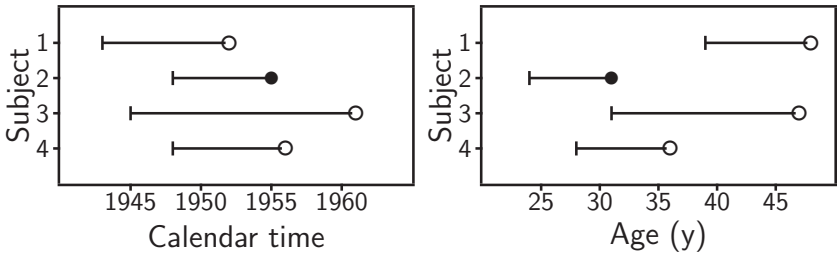
Subject	Born	Entry	Exit	Age at entry	Outcome
1	1904	1943	1952	39	Migrated
2	1924	1948	1955	24	Disease C
3	1914	1945	1961	31	Study ends
4	1920	1948	1956	28	Unrelated death

Subject 1: Follow-up time spent in each ageband

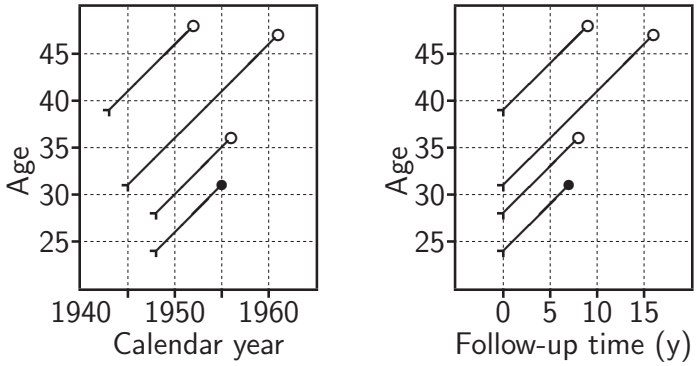
Age band	Date in	Date out	Time (years)
35–39	1943	1944	1
40–44	1944	1949	5
45–49	1949	1952	3

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)



Follow-up in Lexis-diagrams — by age and period



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

Standardization

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

Crude & adjusted 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** is a rate in which:
  - 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.
- Solution: Standardization.

Stomach cancer in Cali and Birmingham (IS, Table 4.2, p. 71)

	Cali			Birmingham			
Age (y)	No. of Male cases 1982 -86	Male Population 1984 (10^3s)	Incidence Rate ($/10^5$ y) 1982 -86	No. of Male cases 1983 -86	Male Population 1985 (10^3s)	Incidence 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 1999	World Stand.
0–44	84	66	61	74
45–64	12	23	27	19
65+	4	11	12	7
All ages	100	100	100	100

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

$\Rightarrow$ The crude rates are **confounded** by age.

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

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 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)

Some standard populations:

Age group (years)	African	World	European	Truncated
0	2 000	2 400	1 600	–
1–4	8 000	9 600	6 400	–
5–9	10 000	10 000	7 000	–
10–14	10 000	9 000	7 000	–
15–19	10 000	9 000	7 000	–
20–24	10 000	8 000	7 000	–
25–29	10 000	8 000	7 000	–
30–34	10 000	6 000	7 000	–
35–39	10 000	6 000	7 000	6 000
40–44	5 000	6 000	7 000	6 000
45–49	5 000	6 000	7 000	6 000
50–54	3 000	5 000	7 000	5 000
55–59	2 000	4 000	6 000	4 000
60–64	2 000	4 000	5 000	4 000
65–69	1 000	3 000	4 000	–
70–74	1 000	2 000	3 000	–
75–79	500	1 000	2 000	–
80–84	300	500	1 000	–
85+	200	500	1 000	–
Total	100 000	100 000	100 000	31 000

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 specific cause C or all causes).

Cumulative rate and cumulative risk

- Choice of standard population weights somewhat arbitrary.
- Alternative and perhaps more "natural" method for age-adjustment is provided by:

$$\textbf{Cumulative rate} = \sum_{k=1}^{K} \text{width}_k \times \text{rate}_k$$

- Weigths are widths of the agebands to be included:

  $\textbf{Cumulative risk} = 1 - \exp(-\text{cumul. rate}) \approx \text{cumul. rate}$
- Usually calculated up to 65 or 75 years with 5-year agebands.
- These estimate the average risk in the population to get the disease by 65 or 75 years given survival until then.
- The competing acuses of exit (death) is **not** taken into account.

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

$$\begin{aligned}
\text{Cali:} \quad & 45\ y \times \tfrac{1.5}{10^5 y} + 20\ y \times \tfrac{69.7}{10^5 y} = 0.0146 = \mathbf{1.46} \text{ per } 100 \\
\text{B'ham:} \quad & 45\ y \times \tfrac{1.2}{10^5 y} + 20\ y \times \tfrac{44.6}{10^5 y} = 0.0095 = \mathbf{0.95} \text{ per } 100 \\
\text{ratio:} \quad & 1.46/0.95 = \mathbf{1.54}
\end{aligned}$$

Cumulative risks (inc. proportions) & their ratio up to 65 y:

$$\begin{aligned}
\text{Cali:} \quad & 1 - \exp(-0.0146) = 0.0145 = \mathbf{1.45}\% \\
\text{B'ham:} \quad & 1 - \exp(-0.0095) = 0.0094 = \mathbf{0.94}\% \\
\text{ratio:} \quad & 1.45/0.94 = \mathbf{1.54}
\end{aligned}$$

Cumulative measures in 5-y groups

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

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

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

Observed and expected cases

- Suppose O cases are **observed** in an **index** population of interest (*e.g.* an occupational cohort) during its follow-up over a lengthy calendar period.
- *Question*: What would be the **expected number of cases** E, if the age- and period-specific rates of a **reference** population for comparison were valid for the index population?
- The ratio "observed/expected" estimates of the "true" rate ratio between the index and the reference populations jointly adjusted for age and period.

Standardized incidence ratio, SIR

Let λ_{kl} = incidence rate in a Lexis-diagram cell defined by ageband k and period l in the reference population. Hence,

$$\text{expected number } (E) = \sum_{k=1}^{K}\sum_{l=1}^{L} \lambda_{kl} \times Y_{kl},$$

where Y_{kl} is the person-years in cell kl of the index population.

The **standardised incidence ratio** (SIR) is defined

$$\text{SIR} = \frac{O}{E}$$

When the outcome is death, this measure is called **standardized mortality ratio**, SMR.

SIR for Cali with Birmingham 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 measures

(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

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

Survival analysis

The **prognosis** of cancer patients:
what is their chance to **survive** 1 year, 5 years etc. after diagnosis?

Survival analysis: In principle like incidence analysis but

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

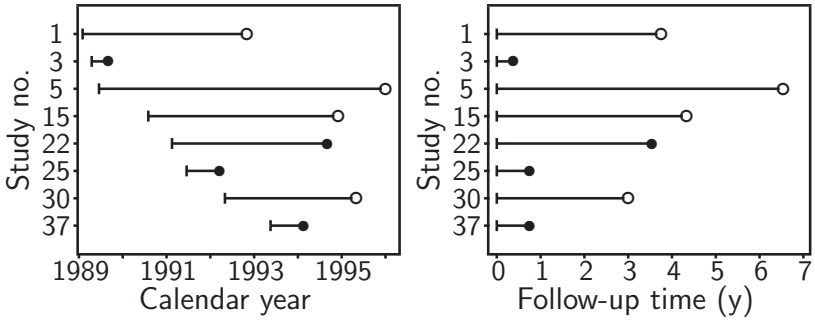
Follow-up of 8 out of 40 breast cancer patients (from IS, table 12.1., p. 264)

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 cancer patients (cont'd)

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



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

Life table or "actuarial" method

(1) Divide the follow-up time into subintervals $k = 1, \ldots K$; usually each with 1 year width.

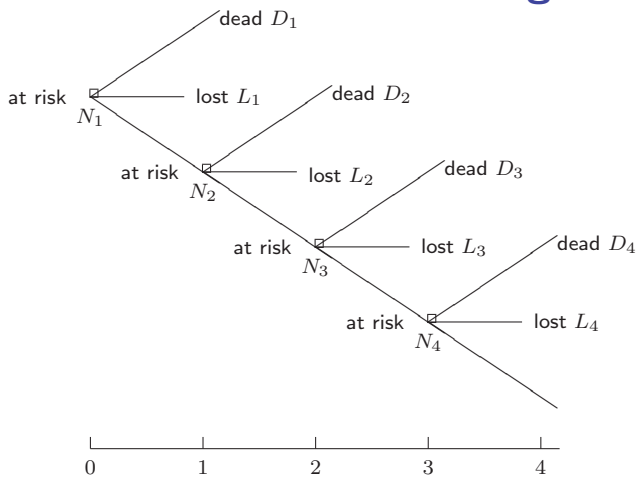
(2) Tabulate from original data for each interval

$N_k =$ 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 measures and methods

Complement of survival proportion $Q_k = 1 - S_k$ is actually incidence proportion of deaths. It estimates cumulative risk of death from start of follow-up till end of kth interval.

"Actuarial" indidence rate in the kth interval:

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

where the person-time is approximated by

$$\left[N_k - \frac{1}{2}(D_k + L_k)\right] \times \text{length of interval}$$

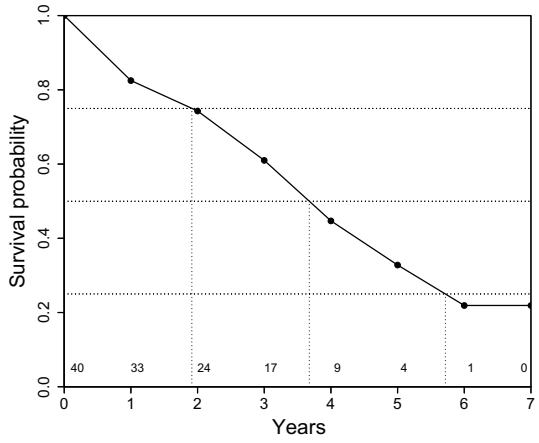
Survival curve and other measures

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.

Cause-specific and relative survival

Cause-specific survival analysis:

- outcome event: death from the disease C itself that defines study population
- deaths from other causes $\rightarrow$ losses.
- problem: ambiguity in cause of death.

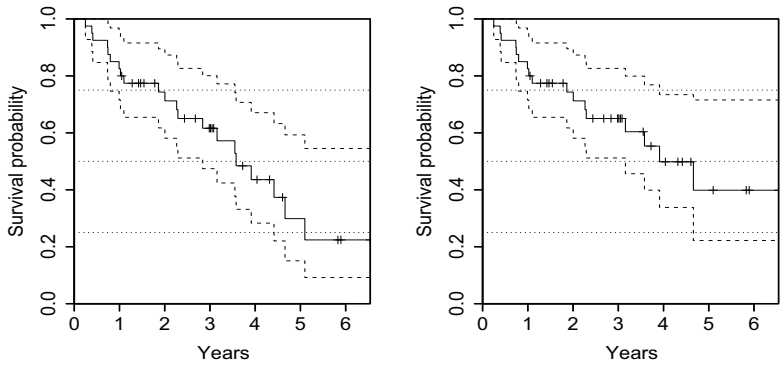
Relative survival: $S_k^{\text{rel}} = S_k^{\text{obs}} / S_k^{\text{exp}}$, i.e. ratio of

- **observed** survival proportion S_k^{obs} in the study population, and
- **expected** survival proportion S_k^{exp} based on age-specific mortalities in the reference (national) population. (See SIR!)

Breast Cancer patients (cont'd)

Overall and cause-specific (death from breast ca.) survival (**IS**: Fig 12.9 & 12.12, p. 271-3)

Kaplan-Meier curves – alternative to "actuarial":



Conclusion

Measures of Disease Occurrence

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

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.

Pracicals

Practicals

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
http://BendixCarstensen.com/NSCE

How to do with practicals

- Read the text
- Find out what you want to do
- Then start using R
- Sequence of practicals:
  1. Monday: 1, 3, 4, 5, 7, 11, 12, 13
  2. Tuesday: 7, 8, 2, 9, 10

Measures of disease frequency and effects
Analysis of epidemiological data

Esa Läärä
University of Oulu, Finland
esa.laara@oulu.fi http://stat.oulu.fi/laara

Bendix Carstensen
Steno Diabetes Center, Denmark
& Department of Biostatistics, University of Copenhagen
bxc@steno.dk http://BendixCarstensen.com

Nordic Summer School in Cancer Epidemiology
August 2011, Danish Cancer Society, Copenhagen
January 2012, Virrat, Finland
http://BendixCarstensen.com/NSCE

Chance

Analysis of Epidemiological Data

Esa Läärä & Bendix Carstensen

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

2 CHANCE VARIATION

2.1 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 & random (cont'd)

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

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

Example 3: Smoking and lung cancer

- Only a minority of smokers get lung cancer. Yet, 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 cause more or less unpredictable – random – variation of disease rates at population level.

Example 4

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?
- Steady decline in females?

Example 4 (cont'd)

Look at observed numbers of cases!

	Males		Females	
Year	Cases	P-years	Cases	P-years
1989	4	88000	275	131000
1990	1	89000	264	132000
1991	3	90000	253	133000

Reality of changes over the years?

2.2 Probability model: random variable, distribution, parameters

Simple model for cancer incidence

In homogenous population we assume

- constant "true" but unknown theoretical incidence rate – *hazard* or *intensity* –
  λ of contracting cancer over short period of time.

Simple model (cont'd)

Number of cases D and empirical incidence rate $R = D/Y$ in Y person-years at risk are:

- *random variables* with unpredictable values in given observation periods.

The *probability distribution* of possible values of a random variable has some known mathematical form.

Key properties of the distribution are determined by quantities called *parameters*; in this case the theoretical rate λ.

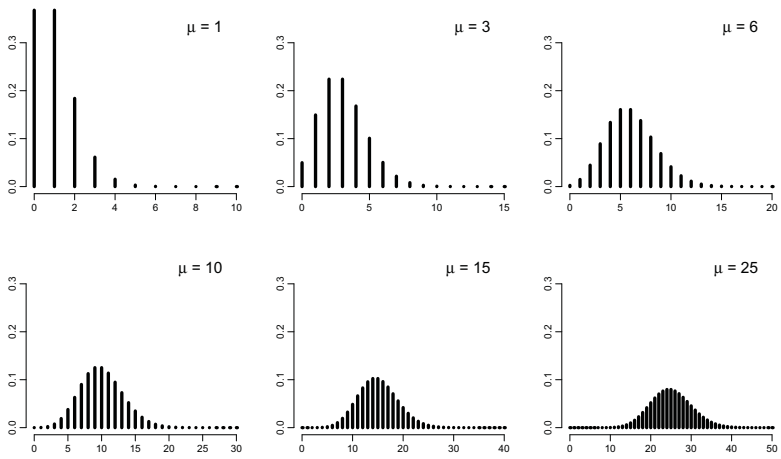
2.3 Poisson and Gaussian models

Poisson distribution: simple probability model for number of cases D 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:

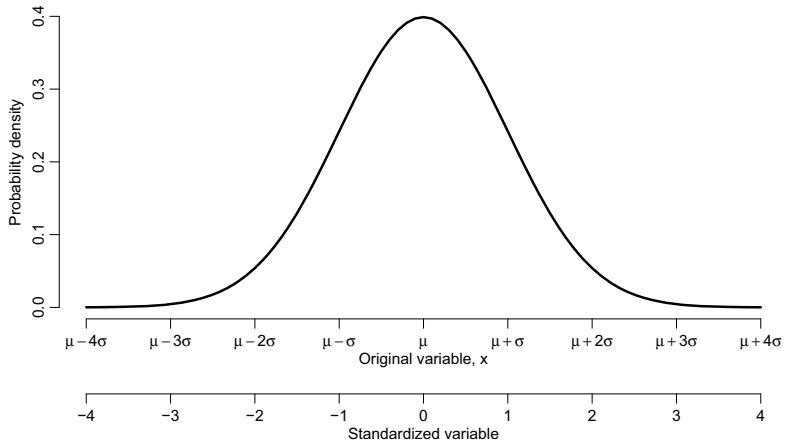
- common model for continuous variables,
- symmetric and bell-shaped,
- has two parameters:
  – μ = expectation or mean,
  – σ = standard deviation.

Most important use of Gaussian model:

> Easy approximation of *sampling distribution* of empirical measures (like observed rates) in certain conditions.

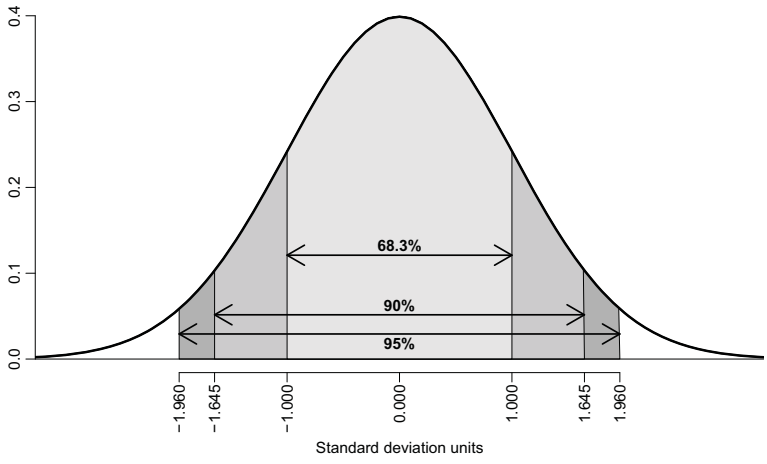
Gaussian distribution (cont'd)

Probability density funtion – the "Bell Curve".



Gaussian distribution (cont'd)

Areas under curve limited by selected quantiles



2.4 Statistic, sampling distribution and standard error

Statistic = summary measure calculated from empirical data (sample).

Let X be a variable having certain distribution in population with mean μ and standard deviation σ.

- Take a random sample of n subjects.
- Values of X in the sample: $X_1, X_2, \ldots, X_n$.
- Before sampling these are random variables.

Statistics (cont'd)

Some statistics derived from this sample:

- Sample mean (arithmetic): $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Sample standard deviation:

$$\mathrm{SD} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

- One-sample T-statistic: $T = \frac{\bar{X} - \mu_0}{\mathrm{SD}/\sqrt{n}}$
  (μ_0 is the hypothesized value of μ).

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 $n \to$ the narrower and more Gaussian-like sampling distribution!

Example 5

Sampling distribution of the sample mean $\bar{X}$ of variable X with mean μ and standard deviation σ is approximately Gaussian with:

- expectation μ,
- standard deviation $\sigma/\sqrt{n}$,

with sufficiently big sample size, whatever the original distribution of X.

This holds by virtue of the *Central Limit Theorem* (CLT) in probability theory.

Standard error (SE)

Estimated standard deviation of sampling distribution of statistic.

Example 5 (cont'd): Sample $X_1, \ldots, X_n$ drawn of variable X from population distribution with mean μ and standard deviation σ. The sample mean is $\bar{X}$ and the sample standard deviation SD.

$\Rightarrow$ *Standard error* of the mean:

$$\mathrm{SE}(\bar{X}) = \frac{\mathrm{SD}}{\sqrt{n}}$$

Describes *precision* in estimation of μ by $\bar{X}$.

Standard error (cont'd)

- Used in one-sample T-statistic:

$$T = \frac{\bar{X} - \mu_0}{\mathrm{SE}(\bar{X})}$$

to test null hypothesis $H_0 : \mu = \mu_0$.
(How far from μ_0 is $\bar{X}$, in SE units)

- Confidence interval (CI) for μ:

$$\bar{X} \pm z \times \mathrm{SE}(\bar{X})$$

where z is an approriate quantile of the t- or Normal distribution (in Normal dist'n $z = 1.960$ for 95% CI).

Example 6: Single incidence rate

Parameter λ
= true unknown incidence rate in population.

Empirical rate $R = D/Y$, *estimator* of λ.

R is a statistic, random variable whose:

- value varies from one study population ("sample") to another in hypothetical repetitions,
- sampling distribution is (under the Poisson model & other conditions) transformation of the Poisson distribution,

Example 6 (cont'd)

- Expectation of empirical rate R is λ,
  standard deviation in the sampling distribution for R is $\sqrt{\lambda/Y}$.
- Standard error of empirical rate R:

$$\mathsf{SE}(R) = \sqrt{\frac{R}{Y}} = \frac{\sqrt{D}}{Y} = R \times \frac{1}{\sqrt{D}}$$

$\Rightarrow$ The amount of random error depends inversely on the number of cases.

$\Rightarrow$ SE of R is proportional to R.

Inference

Analysis of Epidemiological Data

Esa Läärä & Bendix Carstensen

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

3 STATISTICAL INFERENCE

3.1 Inferential questions

Problem: The parameter value is unknown:

What can we learn about the value?

Data from empirical study :

→ information on parameter is provided by values of some statistics,

→ uncertainty on it is reduced.

Still the true value remains unknown.

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 plausible range of values of the parameter consistent with our data?

3.2 Point estimation

Point estimation

= assessing the value of the unknown parameter by a single number obtained from data.

Estimator (point estimator) of parameter

= statistic to be calculated from observable data (sample), whose sampling distribution is concentrated about the true value of the parameter.

Estimate (point estimate) of parameter

= realized value of the estimator in the data.

Point estimation (cont'd)

Standard error (SE) of estimate

= estimated standard deviation of the sampling distribution of an estimator.

Measures the *(im)precision* of the estimator.

Statistical notation:

- Parameter denoted by a Greek letter
- Estimator & estimate by the same Greek letter with "hat".

Incidence rate:

- true unknown rate: λ
- estimator: $\widehat{\lambda} = R = D/Y$, empirical rate.

Statistical notation (cont'd)

Rate ratio:

- true rate ratio $\rho = \lambda_1/\lambda_0$ between exposed and unexposed,
- estimator: $\widehat{\rho} = \text{RR} = R_1/R_0$, ratio between the empirical rates.

Mean of any variable X

- true mean: μ, expectation
- estimator: $\widehat{\mu} = \bar{X}$, sample mean.

3.2 Statistical testing

Question: Are the observed data
– summarized by an estimate and its SE –
consistent with a given value of the parameter?

Such a given value is often represented in the form a *null hypothesis* (H_0), which is a statement on the true value of the parameter before study.

In comparative problems typically a conservative assumption, *e.g.*

- "no difference in outcome btw 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.

NB. These aims are not so ambitious.

Test statistic

- Function of observed data and null hypothesis value,
- Sampling distribution of it under H_0 is known, at least approximately.

Common form of test statistic:

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

in which ...

Test statistic (cont'd)

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.
- Under H_0 the sampling distribution of this statistic is (with sufficient amount of data) close to the standard Gaussian.

Example 2: OC & breast ca. (cont'd)

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

$$\widehat{\delta} = \text{RD} = 217 - 187 = 30 \text{ per } 10^5 \text{ y.}$$

E = Expected rate difference = 0, if H_0 true.
S = Standard error of RD:

$$\text{SE(RD)} = \sqrt{\frac{217^2}{204} + \frac{187^2}{240}} = 19.4 \text{ per } 10^5 \text{ y.}$$

Example 2: OC & breast ca. (cont'd)

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

$$Z_{\text{obs}} = \frac{30 - 0}{19.4} = 1.55$$

What does this mean?

How do we proceed?

Questions about the test statistic

- How does the observed value Z_{obs} locate itself in the sampling distribution of Z?
- How common or how rare it is to obtain Z_{obs} under H_0?
- What is the probability of getting Z larger than observed Z_{obs} if H_0 were true.

The latter probability is the *one-tailed observed significance level* or P*-value* against alternative $\rho > 1$.

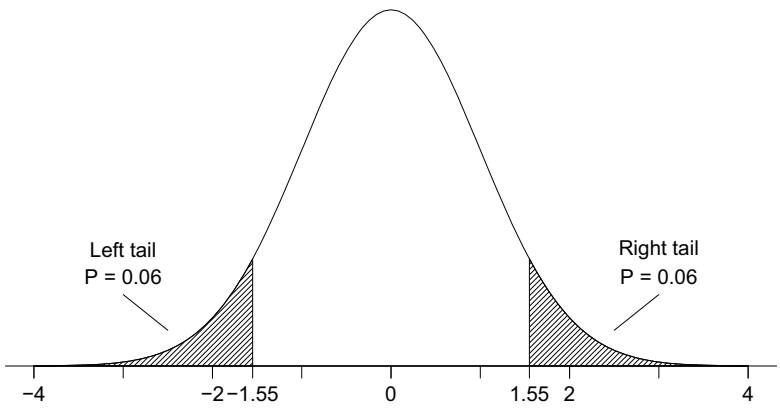
Two-tailed P value

$=$ probability for test statistic Z being more extreme than the absolute value of Z_{obs}.

- Considers deviations from H_0 in either direction.
- Is usually preferred to one-tailed P.

Example 2 (cont'd)

Distribution of test statistic under H_0 and graphical derivation of P-value



One-tailed $P = 0.06$, two-tailed $P = 0.12$

Ex. 1: Lung ca. & asbestos (cont'd)

H_0: Mortality from lung cancer is not elevated in asbestos workers, *i.e.* true rate ratio $\rho = \lambda_1/\lambda_0$ equals 1.

Results:

$O =$ 24 observed cases of lung ca. deaths.

$E =$ 7 expected cases based on age-specific rates in general population.

$$\text{SMR} = \frac{D}{E} = \frac{24}{7} = 3.4$$

Ex. 1: Lung ca. and asbestos (cont'd)

Observed value of test statistic Z:

$$Z_{\text{obs}} = \frac{24 - 7}{\sqrt{7}} = 6.43$$

Under H_0 the sampling distribution of Z is again approximately standard Gaussian.

What is the P-value?

Ex. 1: Lung ca. and asbestos (cont'd)

Tables of standard Gaussian distribution give:

> Under H_0 the probability of getting values of Z larger than the actually observed value 6.43 is < 0.001.

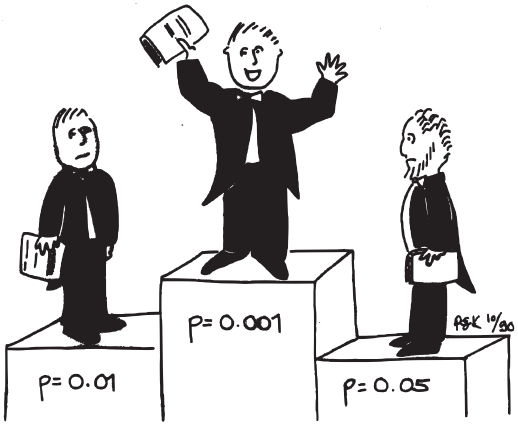
Computer programs show:

> This upper tail P-value is actually 6.4×10^{-11} – extremely small!

Two-tailed $P = 1.28 \times 10^{-10}$ ($2 \times$ one-tailed)

What does this mean?

Great!?



So what?

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, *given that* H_0 is true
- However, **it is not** "the probability that H_0 is true"!

3.4 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!**

Interpretation of P-values (cont'd)

In judging the results, take also into account:

- size of study,
- study design: random sampling, randomization or neither,
- what is a medically relevant deviation of parameter from the H_0 value (e.g. minimally important elevation of true rate ratio from 1),
- Consistency with independent empirical studies and other relevant information & knowledge.

Never base conclusions on a P-value only!

3.5 Confidence interval (CI)

- Range of conceivable values of parameter between lower and upper *confidence limits*.
- Specified at certain *confidence level*, commonly 95% (also 90 % and 99% used).
- The limits of 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%).

Confidence interval (cont'd)

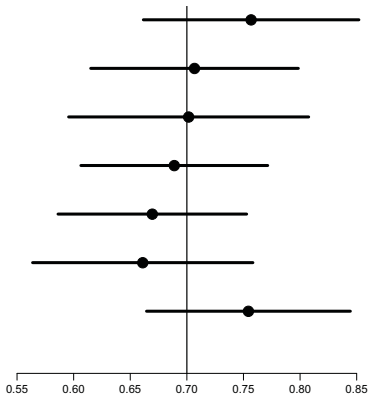
The latter is the *long-term property* of the *procedure* for calculating CI under hypothetical "repeated sampling".

Yet, the obtained CI from data at hand either covers or does not cover the parameter of interest.

(N.B. As with P values the accuracy of nominal confidence level depends on lack of bias and on validity of some statistical assumptions.)

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.

Example 2: OC & breast ca (cont'd)

Observed rate difference RD $= 30$ per 10^5 y.

Standard error SE(RD) $= 19.4$ per 10^5 y.

Limits of the 95% approximate CI (per 10^5 y):

- lower: $30 - 1.96 \times 19.4 = -8$,
- upper: $30 + 1.96 \times 19.4 = 68$

For 90% level, use 1.645 instead of 1.960.
For 99% level, 2.58 is the multiplier.

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.

Interpretation of CI (cont'd)

CI gives more quantitative information on the parameter and on statistical uncertainty about its value than P value.

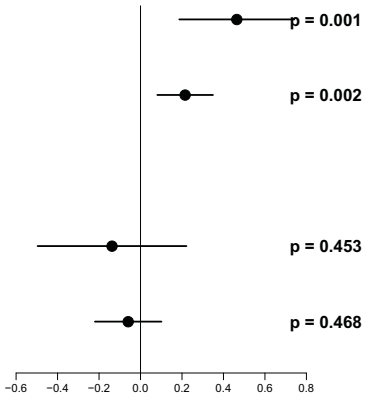
In particular, interpretation of "non-significant" results, *i.e.* large P values:

- 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.

CI and P-value

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



Similar P-values but different interpretation!

3.6 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 (cont'd)

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 in BMJ (cont'd)

CIs should not be used as a surrogate means of examining significance at the conventional 5% level.

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

Analysis of Epidemiological Data

Esa Läärä & Bendix Carstensen

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

4 CRUDE ANALYSIS

4.1 Single incidence rate

Parameter of interest:

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

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

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

Model: D is Poisson with expectation λY.

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

Single rate (cont'd)

Simple approximate 95% CI:

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

where

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

is the 95% *error margin*.

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

Single rate (cont'd)

More accurate approximation of CI by using the log-rate $\ln(R)$, where ln = natural logarithm.
Standard error for log-rate:

$$\mathrm{SE}[\ln(R)] = \frac{1}{\sqrt{D}}$$

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

$$\mathrm{EF} = \exp\{1.96 \times \mathrm{SE}[\ln(R)]\}$$

where exp means exponential function or antilog (inverse of the natural logarithm).

Single rate (cont'd)

From these items we get 95% CI for λ:

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

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

(When $D = 0$, use the "exact" Poisson limits)

N.B.: If the 90% level is desired, then 1.960 substituted by 1.645. For the 99% level the multiplier is 2.576.

Example 4 (cont'd)

The observed incidence of breast cancer in Finnish men aged 65-69 y in 1991 was 33 per 10^6 y based on 3 cases.

Standard error of the rate and the log-rate are

$$\begin{aligned} \mathrm{SE}(R) &= 33 \times \sqrt{1/3} = 19 \text{ per } 10^6 \text{ y} \\ \mathrm{SE}[\ln(R)] &= \sqrt{1/3} = 0.577 \end{aligned}$$

The 95% error margin:

$$\mathrm{EM} = 1.96 \times 19 = 37 \text{ per } 10^6 \text{ y}.$$

Example 4 (cont'd)

For the true rate λ an approximate 95% CI on the original scale:

$$33 \pm 37 = [-4, 70] \text{ per } 10^6 \text{ y.}$$

Negative lower limit – illogical!

A better approximate CI obtained on the log-rate scale via the 95% error factor

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

from which the confidence limits (both > 0):

$$[33/3.1, 33 \times 3.1] = [11, 102] \text{ per } 10^6 \text{ y.}$$

4.2 Rate ratio in cohort study

Question: What is the relative risk of cancer in the exposed as compared to the unexposed?

Parameter of interest: 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 (cont'd)

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 (cont'd)

Point estimator of true rate ratio ρ:
empirical rate ratio (RR):

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

N.B.: The last form is particularly useful
(see next section on case-control studies).

Rate ratio (cont'd)

Standard error of ln(RR), 95% error factor and
approximate 95% CI for ρ:

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

NB. Random error depends inversely on numbers of cases.

Example 8: 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 8 (cont'd)

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

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

95 % CI for ρ :

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

Two-tailed $P = 0.52$.

Interpretation?

4.3 Rate ratio in case-control study

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

Required 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 these?
- What about their ratio?

NO and YES!

- Rates as such are not directly estimable.

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, the *exposure odds ratio*

$$\text{EOR} = \frac{D_1/D_0}{C_1/C_0}$$

estimates the same quantity than the rate ratio from a full cohort study

$$\text{RR} = \frac{D_1/Y_1}{D_0/Y_0} = \frac{D_1/D_0}{Y_1/Y_0}$$

Rate ratio in case-control study

Standard error for ln(EOR), 95% error factor and approximate CI for ρ:

$$\begin{aligned} \text{SE}[\ln(\text{EOR})] &= \sqrt{\frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{C_1} + \frac{1}{C_0}} \\ \text{EF} &= \exp\{1.96 \times \text{SE}[\ln(\text{EOR})]\} \\ \text{CI} &= [\text{EOR}/\text{EF}, \text{EOR} \times \text{EF}] \end{aligned}$$

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

Example 9

Use of mobile phone and brain cancer (Inskip et al. NEJM 2001; 344: 79-86).

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

$$\text{EOR} = \frac{35/637}{51/625} = 0.67$$

Example 9 (cont'd)

Standard error for ln(EOR), 95% error factor and approximate CI for ρ:

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

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

$$\text{EOR} = 0.6, \quad [0.3, 1.0].$$

4.4 Rate difference in a cohort

Parameter of interest: true *rate difference* or *"excess rate"*

$$\delta = \lambda_1 - \lambda_0$$

Same data layout as above for cohort study.

Point estimator of δ, the empirical rate difference: $\hat{\delta} = \text{RD}$

$$\text{RD} = R_1 - R_0 = \frac{D_1}{Y_1} - \frac{D_0}{Y_0}$$

Log-transformation is unapplicable here; original scale is used.

Rate difference (cont'd)

Standard error of RD, 95% error margin & approximate 95% CI for δ:

$$\begin{aligned} \text{SE(RD)} &= \sqrt{\frac{R_1^2}{D_1}+\frac{R_0^2}{D_0}} = \sqrt{\frac{R_1}{Y_1}+\frac{R_0}{Y_0}} \\ \text{EM} &= 1.96 \times \text{SE(RD)} \\ \text{CI} &= [\text{RD} - \text{EM}, \text{RD} + \text{EM}] \end{aligned}$$

Random error again depends inversely on number of cases.

Example 8 (cont'd)

In the Helsinki Heart Study the observed rate difference between the exposed and the unexposed groups was

$$\text{RD} = 3.1 - 2.6 = +0.5 \text{ per } 10^3 \text{ y},$$

Its standard error

$$\text{SE(RD)} = \sqrt{\frac{3.1^2}{31} + \frac{2.6^2}{26}} = 0.755 \text{ per } 10^3 \text{ y}$$

giving an 95% error margin

$$\text{EM} = 1.96 \times 0.755 = 1.5 \text{ per } 1000 \text{ y}.$$

Example 8 (cont'd)

95% approximate CI:

$$0.5 \pm 1.5 = [-1.0, 2.0] \text{ per } 10^3 \text{ y}.$$

Ranges from negative to positive values.

Logical here, because the rate difference can have either minus or plus sign.

Interpretation?

4.5 Analysis of proportions

Suppose we have cohort data with a *fixed risk period*, i.e. the follow-up time for all subjects has the same length. Also, no losses to follow-up (no censoring).

In this setting the *risk* π of the disease over the risk period is easily estimated by simple

incidence proportion
(often called "*cumulative incidence*" or even "risk"):

Analysis of proportions (cont'd)

Incidence proportion:

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

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

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

Analysis of proportions (cont'd)

- Proportions (unlike rates) are dimensionless quantities ranging from 0 to 1.
- Statistical analysis of proportions based on *Binomial distribution*.
- Standard error for single incidence proportion (similarly for prevalence):

$$\text{SE}(Q) = \sqrt{\frac{Q(1-Q)}{n}} = Q \times \sqrt{\frac{(1-Q)}{D}}$$

Depends also inversely on D!

Analysis of proportions (cont'd)

The formulae to analyse and compare incidence proportions or prevalences broadly analogous to those for rates.

- differences of proportions treated on original scale by error margin.
- analysis of ratios based on log-proportions & error factors.
- details of standard error formulas different from those of rates.

4.6 Extensions and remarks

1. All these methods are directly extended to crude analyses of polychotomous exposure variables when each exposure category is separately compared to unexposed.
2. Evaluation of possible monotonic trend in the parameter over increasing levels of exposure: estimation of regression slope.
3. Theoretical rates and risks estimated by standardized or cumulative rates or by life-table methods (*e.g.* Kaplan-Meier):
   $\rightarrow$ use appropriate standard errors of these estimators

Extensions (cont'd)

4. 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)
5. Crude analysis insufficient in observational studies: control of confounding needed. (more of this in next chapter)

Stratified analysis

Analysis of Epidemiological Data

Esa Läärä & Bendix Carstensen

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

5 STRATIFIED ANALYSIS

5.1 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

= heterogeneity of comparative parameter or *effect modification*

- the exposure groups are not comparable w.r.t. other determinants of disease

= bias in comparison or *confounding*

Remedies

Simple approach for remedy:

- *Stratification* of data by potentially modifying and/or confounding factor(s) & use of *adjusted* estimators

Conceptually simpler but technically more demanding approach:

- *Regression modelling*

5.2 Effect modification

Example 10: 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.

Effect modification (cont'd)

Example 11: 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.

Effect modification (cont'd)

- Perfect homogeneity is rare
- Usually both comparative parameters are more or less heterogenous across categories of other determinants of disease.
- Implications to analysis and presentation?

Example 12

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 12 (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

- inherent property of the phenomenon; cannot be removed or "adjusted" for,
- needs careful evaluation.

Problems: Stratum-specific numbers have a large random error

- estimates of effect parameters variable even if no "true" modification present,
- essential modification may remain undetected.

Evaluation of modification (cont'd)

- statistical tests for heterogeneity insensitive and rarely helpful

Tempting to assume:
"no essential modification",

+ simpler analysis and result presentation,
− misleading if essential modification present.

5.3 Confounding

Example 13: 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)

		Operation succesful		
Treatment	Pts	Cases	%	%-diff.
OS	350	273	*78*	
PN	350	290	*83*	+5

PN appears more succesful than OS?

Example 13 (cont'd)

Results stratified by initial diameter size of the stone:

Size	Treatment	Pts	Operation succesful: Cases	Operation succesful: %	%-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?

Example 13 (cont'd)

Solution to the paradox:

- 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 $\Rightarrow$ SS is
  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.

Example 14

Association between grey hair and cancer incidence in a cohort study.

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.

5.4 Steps of stratified analysis

1. Stratify by levels of the potential confounding/modifying factor(s)
2. Compute stratum-specific estimates of the effect parameter (e.g. rate ratio)
3. Evaluate similarity of the stratum-specific estimates by "eyeballing" or test of heterogeneity.

Steps of stratified analysis (cont'd)

4. If the parameter is judged to be homogenous enough, calculate an adjusted summary estimate.
5. If effect modification is judged to be present:
   - report stratum-specific estimates & their CIs,
   - if desired, calculate an adjusted summmary estimate by appropriate standardization (e.g. SMR).

5.5 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
- standardization by external standard population (CMF)
- standardized morbidity ratio (SMR)

5.6 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}

Stratum-specific rates by exposure group:

$$R_{1k} = \frac{D_{1k}}{Y_{1k}}, \qquad R_{0k} = \frac{D_{0k}}{Y_{0k}}$$

Mantel-Haenszel estimators (cont'd)

MH-estimator of the common rate ratio ρ:

$$\text{RR}_{\text{MH}} = \frac{\sum_{k=1}^{K} D_{1k}Y_{0k}/Y_{+k}}{\sum_{k=1}^{K} D_{0k}Y_{1k}/Y_{+k}} = \frac{\sum_{k=1}^{K} w_k R_{1k}}{\sum_{k=1}^{K} w_k R_{0k}}$$

i.e. the ratio of weighted rates between the two groups with weights w_k:

$$w_k = \frac{Y_{1k}Y_{0k}}{Y_{+k}} = \frac{1}{\frac{1}{Y_{1k}} + \frac{1}{Y_{0k}}}$$

Mantel-Haenszel estimators (cont'd)

MH-estimator is thus based on standardised rates in which the MH-weights define the "standard population".

Standard error for $\ln(\text{RR}_{\text{MH}})$

$$\text{SE}[\ln(\text{RR}_{\text{MH}})] = \sqrt{\frac{\sum_{k=1}^{K} \frac{D_{+k}Y_{1k}Y_{0k}}{Y_{+k}^2}}{\left(\sum_{k=1}^{K} \frac{D_{1k}Y_{0k}}{Y_{+k}}\right)\left(\sum_{k=1}^{K} \frac{D_{0k}Y_{1k}}{Y_{+k}}\right)}}$$

Mantel-Haenszel estimators (cont'd)

Error factor (95%) as before:

$$\text{EF} = \exp\{1.96 \times \text{SE}[\ln(\text{RR}_{\text{MH}})]\}$$

95 % approximate CI for ρ:

$$[\text{RR}_{\text{MH}}/\text{EF},\ \text{RR}_{\text{MH}} \times \text{EF}].$$

Example 14 (cont'd)

Grey hair and cancer – $K = 2$ strata.

Point estimate:

$$\begin{aligned} \text{RR}_{\text{MH}} &= \frac{6 \times 20/(10+20) + 60 \times 5/(15+5)}{11 \times 10/(10+20) + 19 \times 15/(15+5)} \\ &= 1.06 \end{aligned}$$

Standard error: $\text{SE}[\ln(\text{RR})] = 0.2337$

95% error factor:
$\text{EF} = \exp(1.96 \times 0.2337) = 1.58$

95% CI: $[1.06/1.58,\ 1.06 \times 1.58] = [0.67,\ 1.68]$

Mantel-Haenszel estimators (cont'd)

Case-control study:
data summary from each stratum k:

Exposure	Cases	Controls	Total
yes	D_{1k}	C_{1k}	T_{1k}
no	D_{0k}	C_{0k}	T_{0k}
Total	D_{+k}	C_{+k}	T_{+k}

Stratum-specific *exposure odds ratio*:

$$\mathsf{EOR}_k = \frac{D_{1k}/D_{0k}}{C_{1k}/C_{0k}} = \frac{D_{1k}C_{0k}}{D_{0k}C_{1k}}$$

Undefined $(+\infty)$ when $D_{0k} = 0$ or $C_{1k} = 0$.

Mantel-Haenszel estimators (cont'd)

Estimator of common rate ratio ρ: Mantel-Haenszel summary odds ratio:

$$\mathsf{EOR}_{\mathsf{MH}} = \sum_{k=1}^{K} D_{1k}C_{0k}/T_{+k} \Bigg/ \sum_{k=1}^{K} D_{0k}C_{1k}/T_{+k}$$

- Can also be expressed as a weighted average of stratum-specific EORs.
- Can be calculated even with zero counts in some strata if in one stratum both $D_{0k} > 0$ and $C_{1k} > 0$.
- Statistically rather efficient also with sparse data.

Mantel-Haenszel estimators (cont'd)

- Standard error $\sqrt{V} = \mathsf{SE}[\ln(\mathsf{EOR}_{\mathsf{MH}})]$ is based on somewhat complicated formula for the estimated variance:

$$V = \frac{\sum A_k P_k}{2\left(\sum P_k\right)^2} + \frac{\sum (A_k Q_k + B_k P_k)}{2\left(\sum P_k\right)\left(\sum Q_k\right)} + \frac{\sum B_k Q_k}{2\left(\sum Q_k\right)^2}$$

$$\begin{aligned} \text{where:} \quad A_k &= (D_{1k} + C_{0k})/T_{+k}, \\ B_k &= (D_{0k} + C_{1k})/T_{+k}, \\ P_k &= D_{1k}C_{0k}/T_{+k}, \\ Q_k &= D_{0k}C_{1k}/T_{+k} \end{aligned}$$

for each stratum $k = 1, \ldots, K$.

Example 15: 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	EOR
yes	96	109	5.64
no	104	666	

Example 15: 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. Should have employed stratified sampling by age.

Example 15 (cont'd)

Effect modification?

- Stratum-specific EORs somewhat variable.
- Random error in some of them apparently great (especially in the youngest and the oldest age groups)
- Only weak evidence against homogeneity, so assumption of a common rate ratio seems plausible.

Example 15 (cont'd)

Confounding?

- Is exposure associated with age in the study population?
- Look at variation in the age-specific prevalences of exposure among controls.
- Adjustment for age is generally reasonable.

Summary estimator:

$$\begin{aligned} \text{EOR}_{\text{MH}} &= \frac{1 \times 106/116 + \cdots + 5 \times 31/39}{0 \times 9/116 + \cdots + 8 \times 0/39} \\ &= 5.16 \ \ [3.56,\ 7.47] \end{aligned}$$

Regression models

Analysis of Epidemiological Data

Esa Läärä & Bendix Carstensen

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

6 REGRESSION MODELLING

6.1 Limitations of stratified analysis

- Multiple stratification
  $\Rightarrow$ many strata with sparse data
  $\Rightarrow$ loss of precision
- Continous risk factors must be categorized
  $\Rightarrow$ loss of precision
- More than 2 exposure categories:
  - Pairwise comparisons give inconsistent results
  - Linear trend not easily estimated

Limitations (cont'd)

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

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

The key concept here is the *statistical model*.

6.2 Log-linear model for rates

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

$$\ln\{\lambda(X, Z, \ldots)\} = \alpha + \beta X + \gamma Z + \ldots$$

Equivalent expression, *multiplicative model*:

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

Log-linear model (cont'd)

Model parameters

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

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

Example 10 (cont'd)

Lung cancer incidence by 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

$$\ln\{\lambda(X, Z)\} = 2.485 + 1.609X + 2.303Z$$

Example 10 (cont'd)

Parameters

- α $= 2.485 = \ln(12)$, log of baseline rate,
- β $= 1.609 = \ln(5)$, log of rate ratio $\rho = 5$ between exposed and unexposed for asbestos
- γ $= 2.303 = \ln(10)$, 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.

No extra parameters for effect modification needed.

Log-linear model (cont'd)

Model with effect modification (two regressors only)

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

equivalently

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

where α is as before, but

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

Interaction parameter

- δ = $\ln(\theta)$, 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.

6.3 Additive model for rates

General form with two regressors

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

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

6.3 Additive model (cont'd)

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 10 (cont'd) Additive model

$$\lambda(X, Z) = 12 + 48X + 108Z + 432XZ$$

where

α = 12, baseline rate, i.e. that among those both unexposed to asbestos and non-smokers,

β = 48 $(60 - 12)$, rate difference between asbestos exposed and unexposed among non-smokers only,

Example 10 (cont'd)

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:

$$\delta = [600 - 120] - [60 - 12] = 432.$$

6.4 Model fitting

In real life model parameters unknown.
$\Rightarrow$ Must be estimated from data.

General method for model fitting:
– *maximum likelihood* (ML)

Performed by suitable computer software:
like `R, Stata, S-Plus, SAS`.

Model fitting (cont'd)

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.

6.5 Problems in modelling

Simple model chosen may be far from the "truth".

- possible bias in effect estimation, at least 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.

Modelling

Modelling should not substitute but complement crude & stratified analyses.

Adequate training and experience required.

Ask help from professional statistician!

Conclusion

Analysis of Epidemiological Data

Esa Läärä & Bendix Carstensen

Nordic Summerschool of Cancer Epidemiology
15–26 August 2011
Copenhagen
`http://BendixCarstensen.com/NSCE`

7 CONCLUDING REMARKS

Epidemiologic study is a

Measurement excercise

Object: some **parameter** of interest, like

- incidence rate
- 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
- other

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

Controlled randomness

If *controlled randomness* (random sampling or randomization) is employed as appropriate

$\Rightarrow$ parameter estimate has a well defined *sampling distribution*

This forms the basic tool used in *statistical inference* concerning the value of the parameter

- point estimation
- statistical testing, P-value
- confidence interval

Controlled randomness (cont'd)

Question: How often controlled randomness actually employed in epidemiology?

Answer: Rarely!

"In most epidemiologic studies, randomization and random sampling play little or no role in the assembly of study cohorts."
(Greenland S. Epidemiology 1990; 1: 421-9)

Implications

"... probabilistic interpretations of conventional statistics are rarely justified ... such interpretations may encourage misinterpretation of nonrandomized studies."

"... the continuing application of tests of significance to such non-randomized investigations is inappropriate" (Greenland 1990)

"Confidence intervals should be relegated to a small part of both the results and discussion section as an indication, but no more, of the possible influence of chance imbalance on the result." (Brennan & Croft. *BMJ* 1994; **309**: 727-30)

Recommendations

Possible remedies for these problems

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

(Greenland 1990)

Interpretation of obtained values of inferential statistics – not mechanical!

Recommendations (cont'd)

- "The ability to judge the potential role of chance without the aid of complicated statistics is valuable.
- . . . when confronted with the results from small numbers, and experienced researcher should be able quickly to judge whether statistics are worth calculating at all.
- . . . judgment, that the sample size is sufficient and the observed result so great that chance may be dismissed, can and should be made when one is "confident" that the decision is obvious." (*Jolley, Lancet* 1993; **342**: 27-29)

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!

Cancer in diabetes patients: Basing a wrong conclusion on a wrong or on a correct analyses.

Bendix Carstensen

Nordic Summerschool of Cancer Epidemiology
3–5 February 2012
Virrat, Finland
`http://BendixCarstensen.com/NSCE`

Diabetes and Cancer

Persons with diabetes have long been known to have increased incidence rates and mortality rates from cancer [**?**, **?**, **?**]:

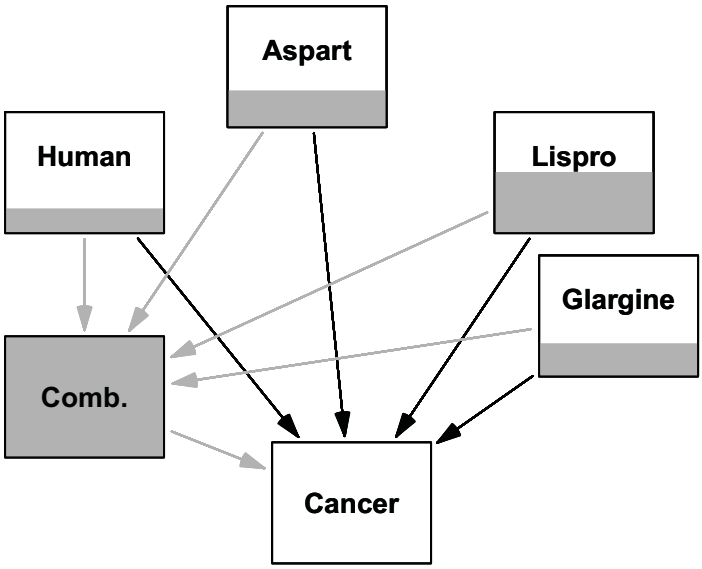
- Pancreas
- Liver
- Colon / Rectum
- Corpus uteri
- Lung
- Kidney
- . . .

Diabetologia, September 2009:

- **Risk of malignancies in patients with diabetes treated with human insulin or insulin analogues: a cohort study.** L. G. Hemkens, U. Grouven, R. Bender, C. Günster, S. Gutschmidt, G. W. Selke, and P. T. Sawicki, Diabetologia, 52:1732–1744, Sep 2009.
- **Insulin glargine use and short-term incidence of malignancies-a population-based follow-up study in Sweden.** J. M. Jonasson, R. Ljung, M. Talbäck, B. Haglund, S. Gudbjörnsdottir, and G. Steineck, Diabetologia, 52:1745–1754, Sep 2009.
- **Use of insulin glargine and cancer incidence in Scotland: a study from the Scottish Diabetes Research Network Epidemiology Group.** H. M. Colhoun and the SDRN Epidemiology Group, Diabetologia, 52:1755–1765, Sep 2009.
- **The influence of glucose-lowering therapies on cancer risk in type 2 diabetes.** C. J. Currie, C. D. Poole, and E. A. Gale, Diabetologia, 52:1766–1777, Sep 2009.
- **Does diabetes therapy influence the risk of cancer?** U. Smith and E. A. Gale, Diabetologia, 52:1699–1708, Sep 2009.

Hemkens et al. [?]

- Data: Insurance database from Germany
- Entry: Newly started treatment for DM
- Exposure:
  Monotherapy (4 classes) throughout follow-up
  - Initial dose
  - Cumulative dose over the entire follow-up
- Outcome: All cancers
- Model: Cox (time since treatment start?)

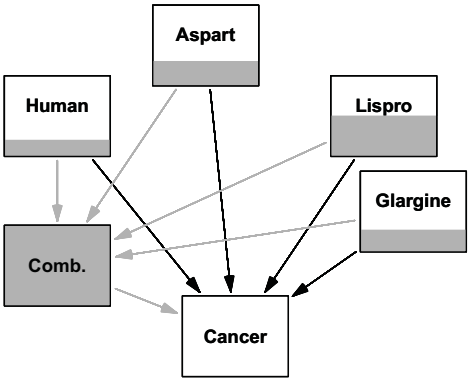


Problems (Hemkens et al.

- Assumes that those who go on to combination therapy are irrelevant, *i.e.* all effects are *instantaneous*.
- The time on monotherapy *before* combination therapy is discarded:

  > *We defined four study groups according to the treatment received: human insulin, aspart, lispro and glargine. Eligible participants were those exposed to only one of these agents during follow-up.*

- . . . thus all cancer rates are too small
- . . . and not necessarily with the same amount
- Conditioning on the future



The gray part of the follow-up time is discarded based on knowledge of the future exit from the groups.

Currie et al. [?]

- Data: THIN database
  (clinical records from GPs)
- "Cohort" of OAD initiators.
- Time-varying exposure,
  i.e. follow-up classified by *current* (maximal?) treatment:
  - Metformin
  - SU
  - Met+SU
  - Insulins: Human basal / Human biphasic / Glargine / other Analog

Currie et al. [?]

- Model: Cox (time since treatment start)
- Persons censored at therapy change:

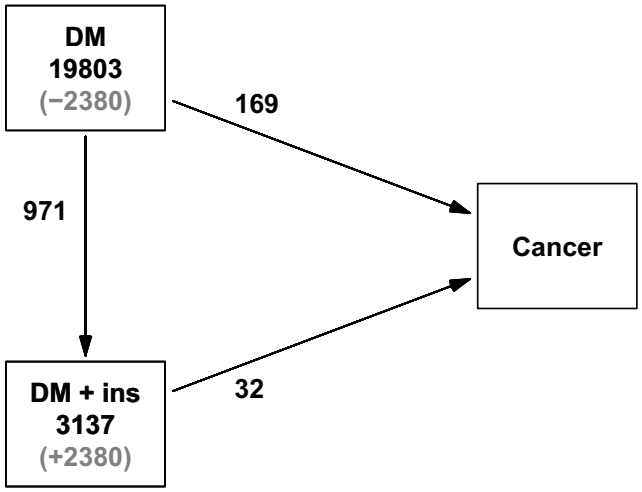
  > *Cohort membership was terminated by progression to a record of the primary or secondary outcomes of interest, right censoring at the final observation of the database, transfer out of the practice, or treatment switching.*

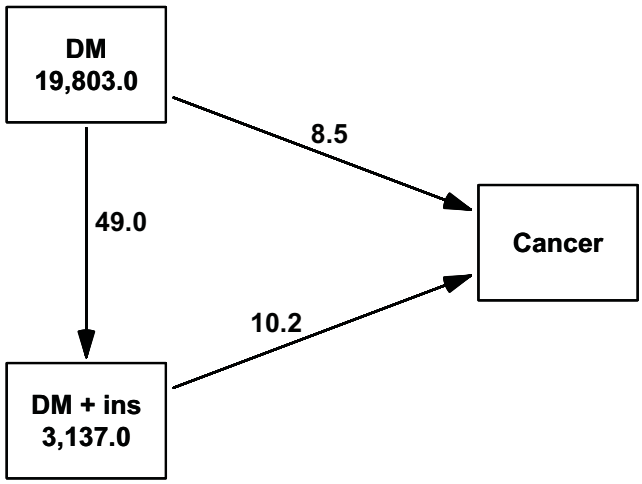
- Censoring is **not** independent of the disease outcome

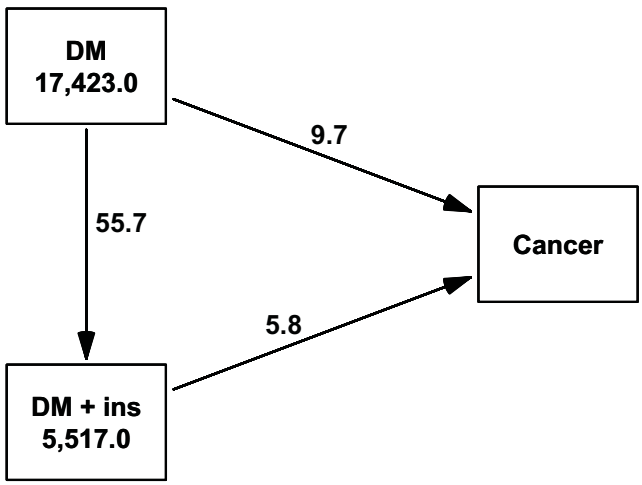
Yang et.al [?]

Associations of Hyperglycemia and Insulin Usage With the Risk of Cancer in Type 2 Diabetes: The Hong Kong Diabetes Registry.
Yang *et al:* Diabetes, vol. 59, May 2010, pp. 1254 ff.

- Data: DM register of Hong Kong
- Cohort based on any exposure in entire follow-period.
- Additional matching of insulin users to non-users.
- Insulin vs. non-insulin: $\mathrm{RR} = 0.18$!
- Strong bias because of mis-allocation and exclusion of risk time.
- Immortal time bias


DM
19803
(−2380)
169
971
Cancer
DM + ins
3137
(+2380)
32


DM
19,803.0
8.5
49.0
Cancer
10.2
DM + ins
3,137.0

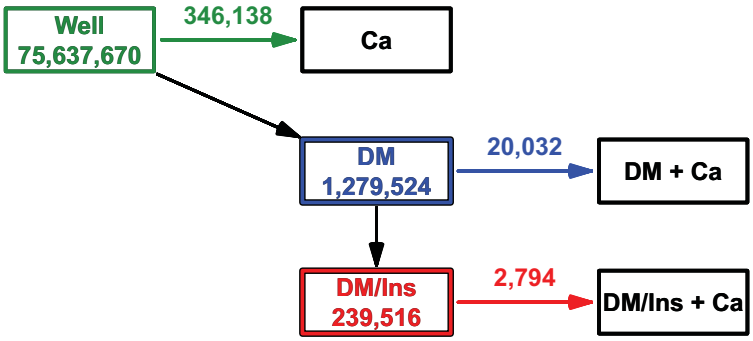

DM
17,423.0
9.7
55.7
Cancer
5.8
DM + ins
5,517.0

Danish study [?]

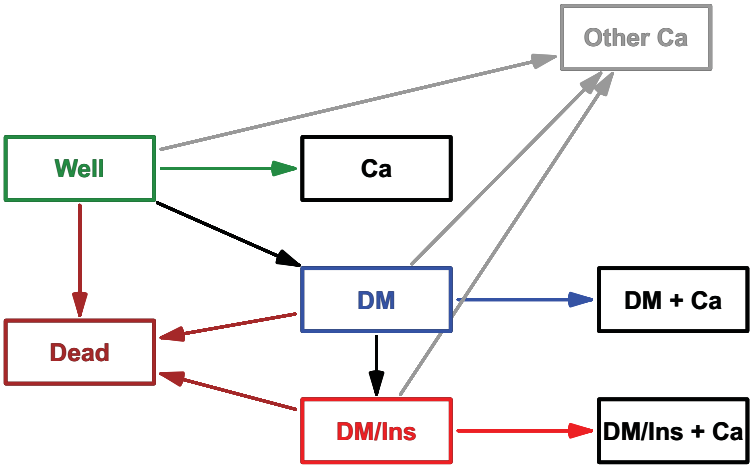
Cancer occurrence in Danish diabetic patients: duration and insulin effects. B. Carstensen, D. R. Witte, and S. Friis. Diabetologia, e-pub ahead of print, Nov 2011.

- Describe cancer incidence rates among diabetes patients in Denmark.
- and how rates vary relative to the non-DM population with:
  - duration of diabetes
  - duration of insulin use
- for all types of cancer
- and for specific sites of cancer

Follow-up of the Danish population



Follow-up of the Danish population



Follow-up in the population

Persons are followed 1 Jan 1995 to:

event: first primary cancer of a given type

censoring:
- diagnosis of any other primary cancer
- death
- 31 Dec 2009

Tabulation & analysis

Follow-up time (person-years) and events (cancer diagnosis) were classified by:

- sex
- current age in 1-year classes
- current date in 1-year classes
- date of birth in 1-year classes
- state of follow-up: Well / DM / DM/Ins
- duration of DM in 6 month classes
- duration of insulin use in 6 month classes

Poisson analysis using class midpoints as continuous variables.

How the data looks — events

	Diabetes duration			Insulin duration		
	Well	DM	DM/Ins	Well	DM	DM/Ins
0	319088	4331	255	319088	17927	781
1	0	2703	196	0	0	407
2	0	2322	222	0	0	329
3	0	1917	238	0	0	248
4	0	1714	210	0	0	181
5	0	1356	211	0	0	133
6	0	1023	216	0	0	132
7	0	828	231	0	0	85
8	0	633	169	0	0	61
9	0	479	180	0	0	46
10	0	297	131	0	0	22
11	0	194	120	0	0	17
12	0	100	62	0	0	11
13	0	30	15	0	0	3
Sum	319088	17927	2456	319088	17927	2456

Model for cancer incidence rates

$$\begin{aligned}\text{rate} = & f(\text{age}) \times g(\text{date of FU}) \times h(\text{date of birth}) \\ & \times t(\text{DM-duration}) \\ & \times s(\text{Ins-duration})\end{aligned}$$

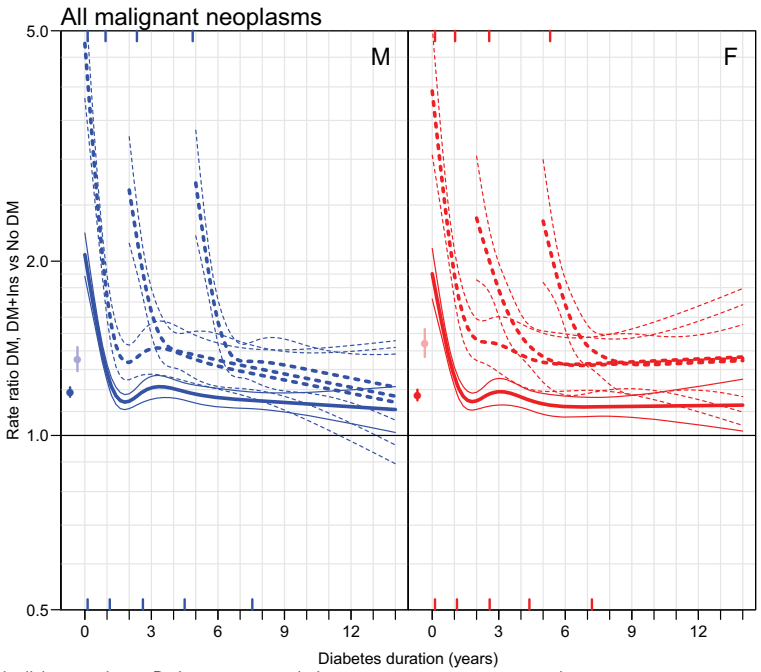
Functions t and s give the **combined** effects of:

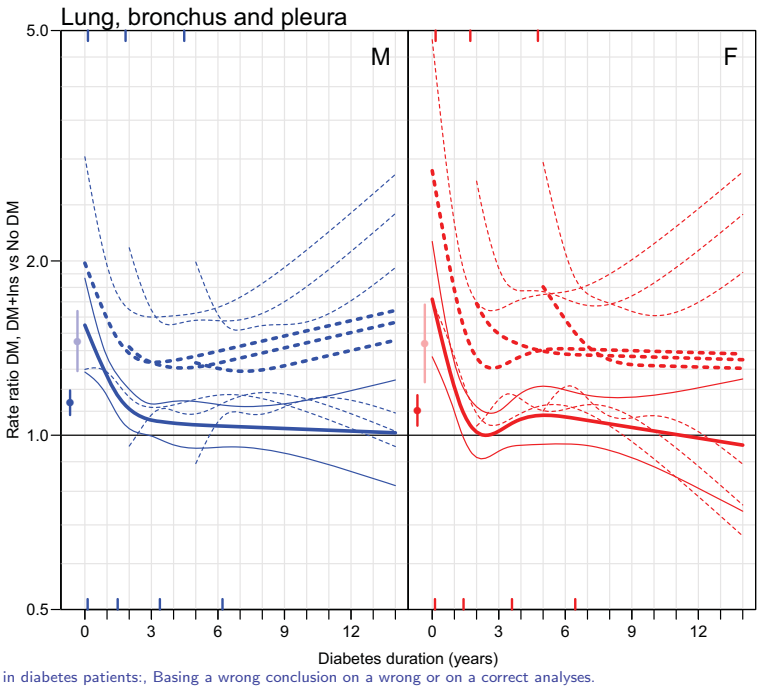
- duration / cumulative dose
  (slowly increasing/decreasing from time 0)
- allocation (jump at time 0) & common risk factors
  (confounding by indication)

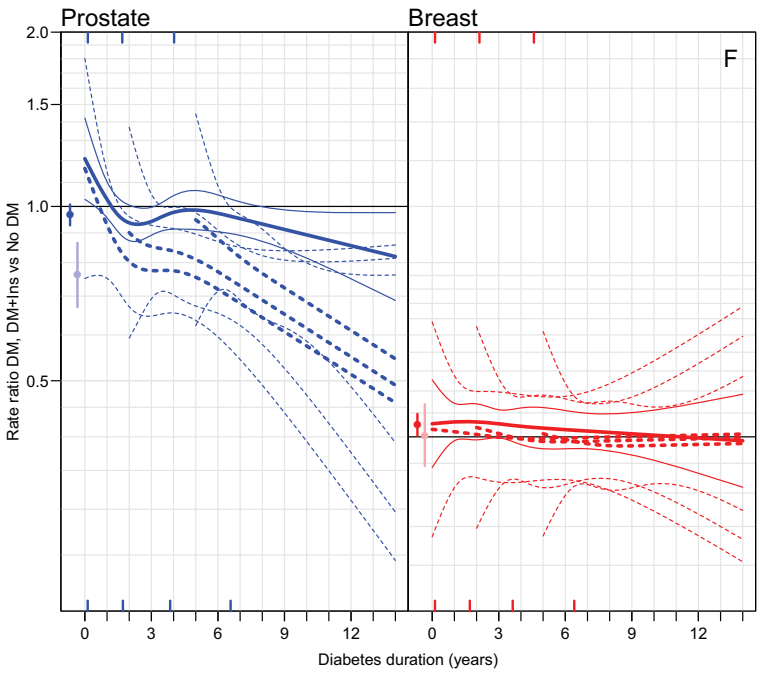
There is **no way** to separate these two effects.

Modelling in R

```
m1 <- glm( D ~ Ns(ax,knots=a.kn) +
               detrend( Ns(px,knots=p.kn), px ) +
               Ns(cx,knots=c.kn) +
               state +
               Ns( DMDur,knots=d.kn) +
               Ns(InsDur,knots=d.kn) +
               offset( log(y) ),
               family = poisson,
               data = subset(data,sex==sx) )
```





Interpretation

Findings are consistent with:

- Common risk factors for DM and cancer
  (obesity, lack of physical exc., eating habits . . .)
- More intense surveillance for cancer following DM diagnosis
- Reverse causation: Undiagnosed cancers lead to DM diagnosis
- Effect of insulin in either direction:
  A cumulative effect of insulin increasing cancer risk cannot be excluded even if RR decrease by insulin duration for most cancer sites — there is a strong mortality selection.

Methodological points for FU-studies

- Follow all persons till death or exit from study
  — never censor persons due to status change, model effect of the status change.
- Only classify follow-up (risk time, events) by currently known features:
  Do not condition on the future.
- Multiple time scales necessary (age, calendar time, duration)

Morale:

- Always draw *all* your boxes.
- Define what they mean.
- When do persons enter.
- When do they exit:
  - as events
  - as censorings (is this independent of the event process?)
- What is counted as events; what is not.

References

The epidemic of matching

Bendix Carstensen

Nordic Summerschool of Cancer Epidemiology
3–5 February 2012
Virrat, Finland
`http://BendixCarstensen.com/NSCE`

Avoid confounding

Confounding of the

- exposure effect on
- the outcome

arises when:

- the confounder is associated with the exposure
- the confounder is associated with the outcome

Sometimes the former can be fixed, but rarely the latter

Avoid confounding

How do you fix the association between a confounder, such as

- age at diagnosis, exposure, ...
- sex

and the exposure, such as:

- IUD
- congenital malformation
- childhood cancer

... you make sure that the confounder distribution is the same among exposed and non-exposed!

$\Rightarrow$ Match your cohort study.

Avoid confounding

What if you cannot fix the confounder distribution?

- Control for the confounder
- Include it in a model

which will allow you to

- Model the exposure effect
- Test for interaction
- . . .

Avoid the "clinical trial" thinking

When you match the control group it is no more representative for the un-exposed.

Analyses based only on the control group are meaningless, such as a Kaplan-Meier curve. . .

. . . only comparisons are relevant.

The precision of the estimates from the control group is smaller that it would have been if you had taken the entire group

Don't think it's a clinical trial

Instead of

Match, Waste, Compare

you should

Use all, Analyze, Report!