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 www.biostat.ku.dk/~bxc

Nordic Summer School in Cancer Epidemiology
August 2011, Danish Cancer Society, Copenhagen
January 2012, Virrat, Finland

Outline

Introduction

1.1 Starters – Example 1

Cohort of male asbestos workers, $N = 17800$.

Observed $D = 24$ cases of lung cancer deaths.
Expected $E = 7$ cases based on age-specific rates in general population.

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

Observed rate ratio > 1:

- true as such?
- biased? by which factors?
- due to play of chance?

Example 2

Nurses Health Study (NHS) on
oral contraceptive (OC) use and breast cancer.

Null hypothesis H_0:
OC use does not affect risk of breast cancer; true rate ratio $= 1$ between ever and never users.

Summary of study outcomes:

OC use	No. of Cases	Person-years	Rate ($/10^5$ y)
Ever	204	94029	217
Never	240	128528	187

Example 2 (cont'd)

Results:

(i) Observed rate ratio RR $= 217/187 = 1.16$,

(ii) P-value 0.12,

(iii) 95% confidence interval $[0.96, 1.40]$

Interpretation?

- true rate ratio $= 1.16$?
- probability that H_0 is true $=$ 12% ?
- probability $=$ 95%, that true rate ratio is between 0.96 and 1.40?
- other? further analysis needed?

1.2 Analysis and statistics

By *analysis* we mean *statistical analysis*.

What is statistics?

1. "(singular) the science that deals with the
   - collection, classification, analysis, and interpretation of numerical facts or data, and that,
   - by use of mathematical theories of probability, imposes order and regularity on aggregates of more or less disparate elements."

2. "(plural) the numerical facts or data themselves." (Webster's Dictionary)

1.3 Uses of statistics in epidemiology

Major tasks:

- assessment of *random variation*
- control of *confounding* and evaluation of *modification*
- guiding study planning:
  choice of design, group sizes,
  length of follow-up, sampling.

Uses of statistics (cont'd)

Basic approaches and tools:

- descriptive summarization of data,
- mathematical models for random variation,
- statistical inference: estimation and testing,
- crude and stratified analysis,
- regression methods.

1.4 References

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

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

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

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 $I = D/Y$ 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.

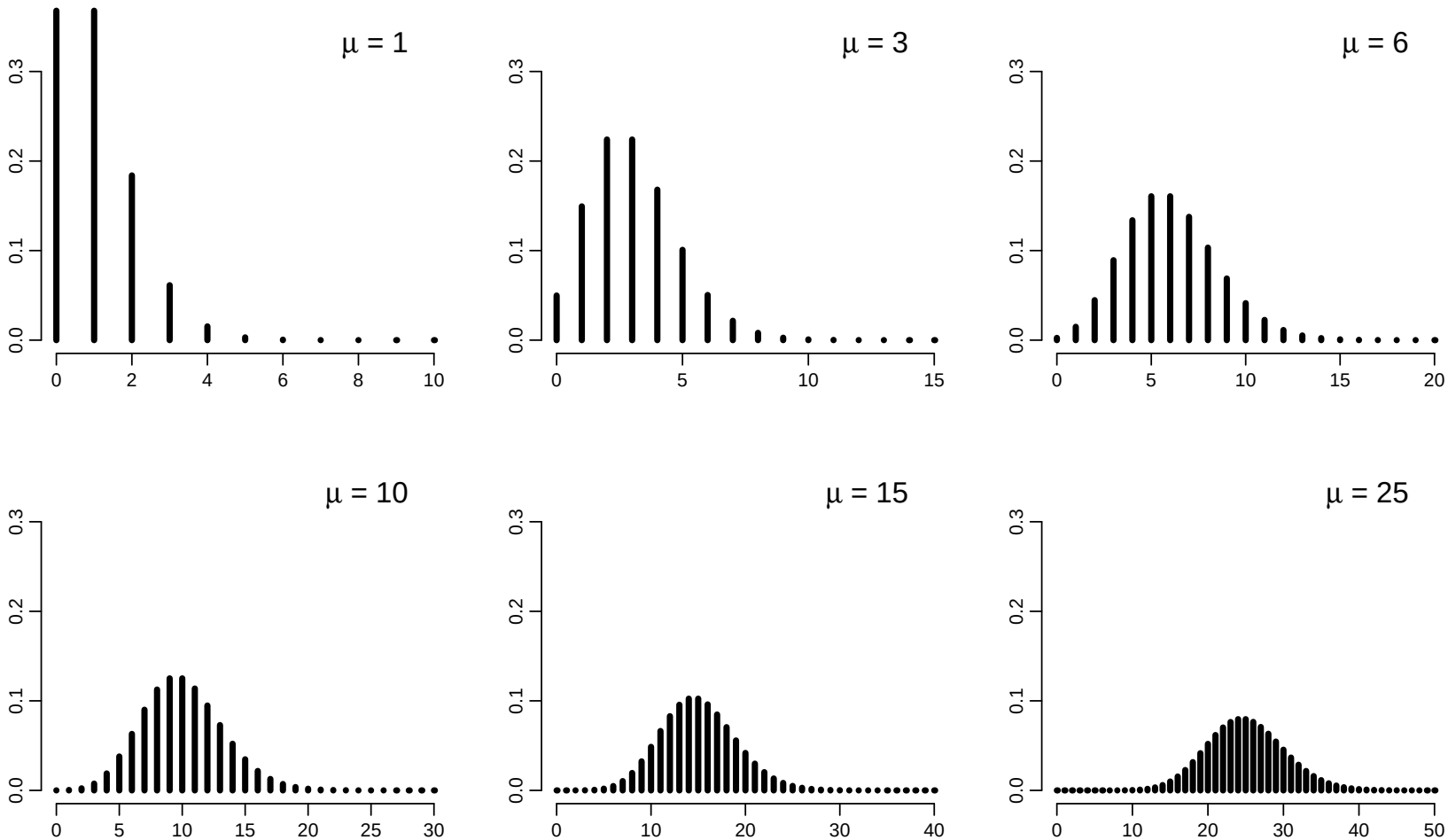
2.3 Poisson and Gaussian models

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:

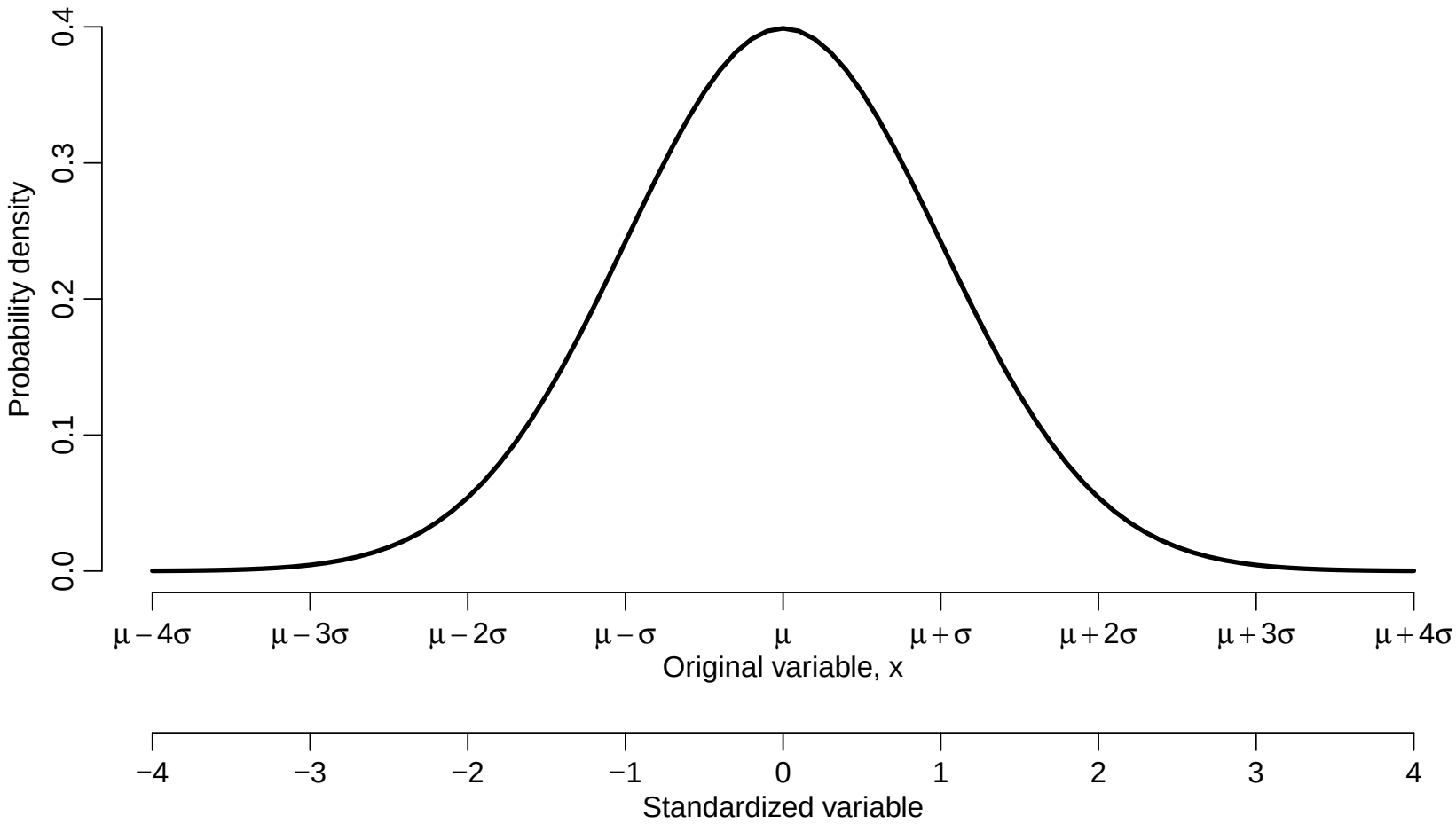
- common model for continuous variables,
- symmetric and bell-shaped,
- has two parameters:
  – $\mu =$ expectation or mean,
  – $\sigma =$ standard deviation.

Most important use of Gaussian model:

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

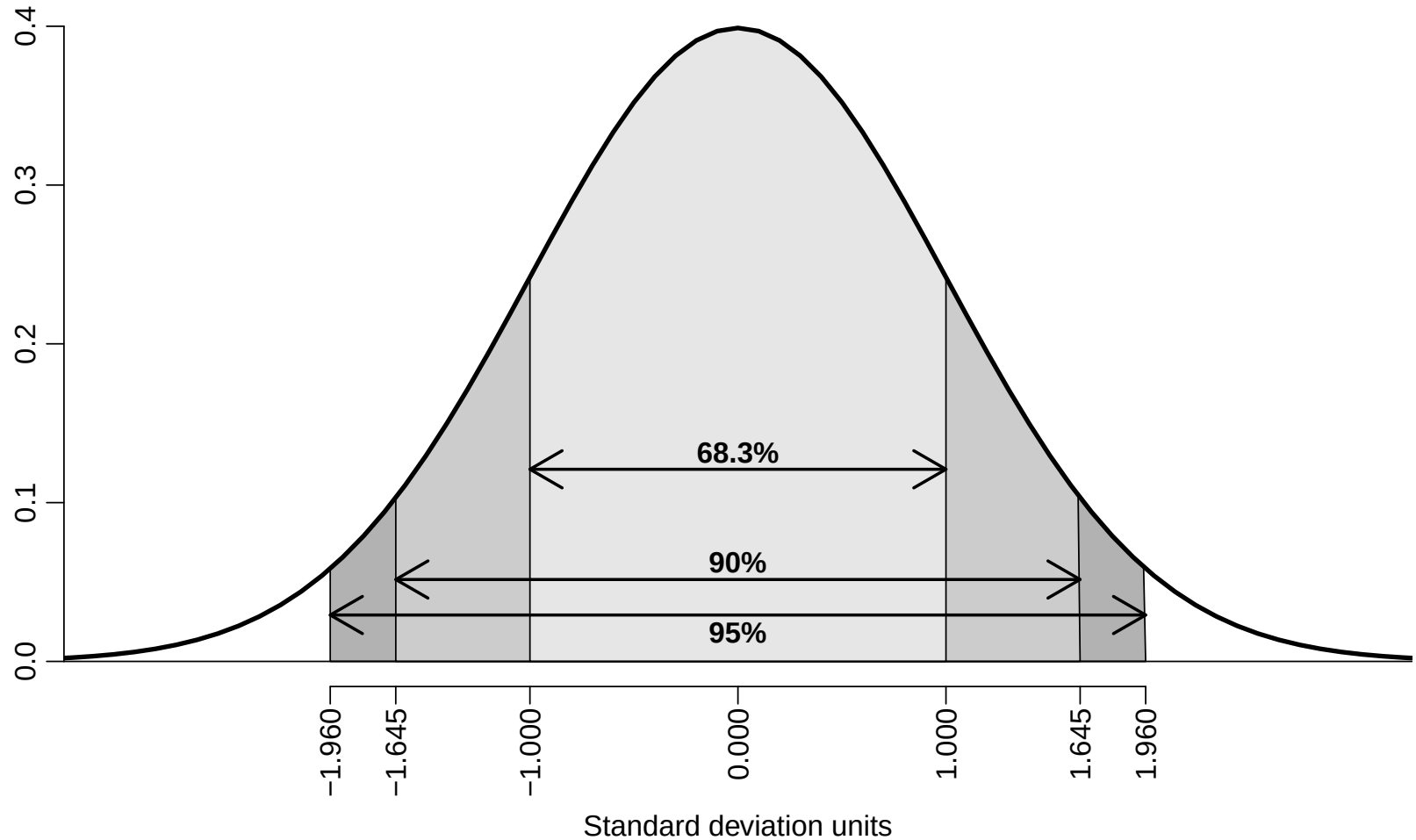
Gaussian distribution (cont'd)

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:

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

- One-sample T-statistic: $T = \frac{\bar{X} - \mu_0}{\text{SD}/\sqrt{n}}$
  (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 $n \longrightarrow$ 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 $I = D/Y$, *estimator* of λ.

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

$$\mathrm{SE}(I) = \sqrt{\frac{I}{Y}} = \frac{\sqrt{D}}{Y} = I \times \frac{1}{\sqrt{D}}$$

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

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

3 STATISTICAL INFERENCE

3.1 Inferential questions

Problem: The parameter value is unknown:

What can we learn about the value?

Data from empirical study :

- $\longrightarrow$ information on parameter is provided by values of some statistics,
- $\longrightarrow$ 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: $\widehat{\lambda} = I = D/Y =$ empirical rate.

Statistical notation (cont'd)

Rate ratio:

- true rate ratio $\rho = \lambda_1/\lambda_0$ between exposed (1) and unexposed (0),
- estimator of true rate ratio: $\widehat{\rho} = \mathsf{IR} = I_1/I_0$ $=$ ratio between the empirical incidence rates in the pertinent exposure groups

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 very 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.

Ex 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

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

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

S = Standard error of ID:

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

Ex 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 $\rho > 1$.

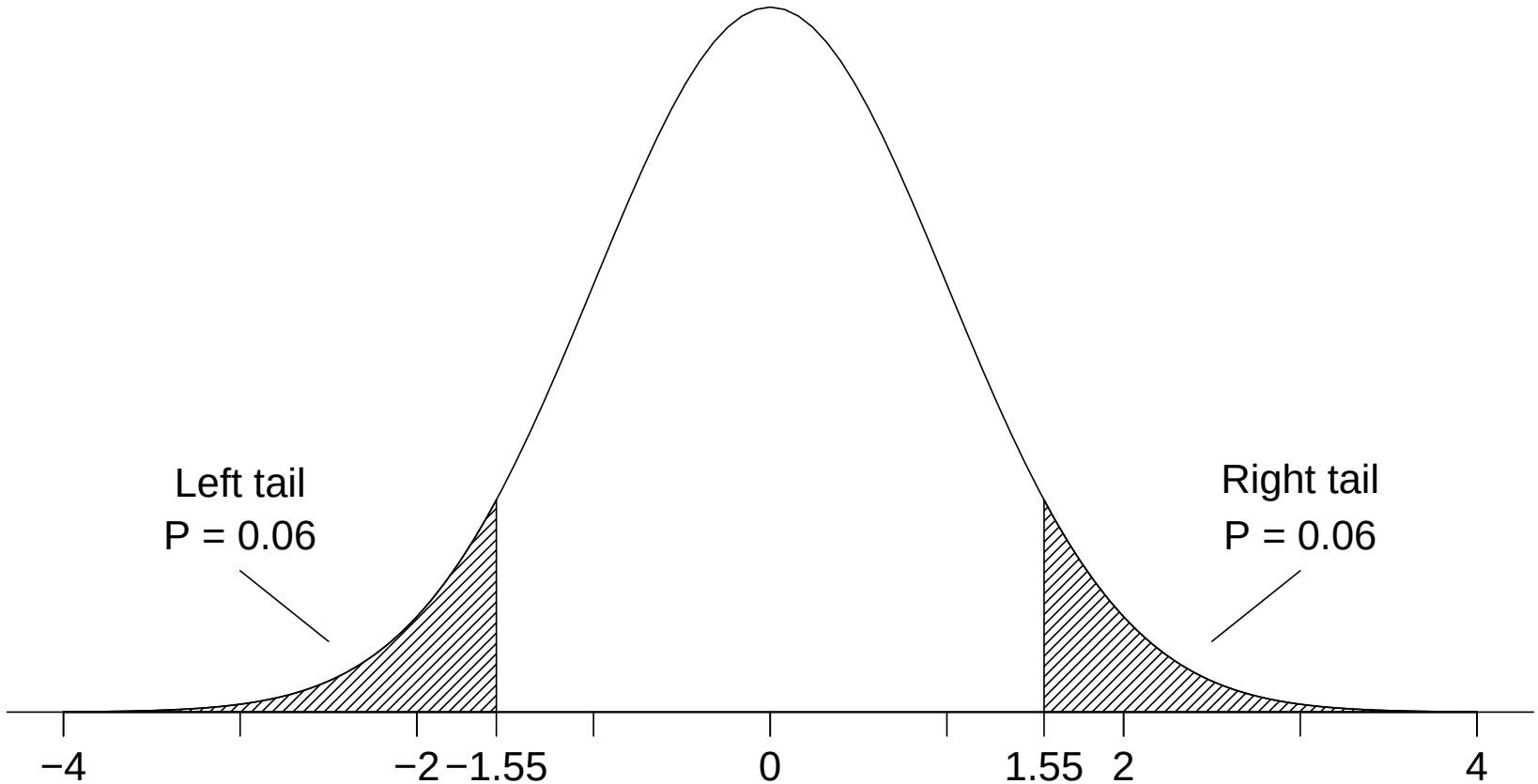
Two-tailed P value

$=$ 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.

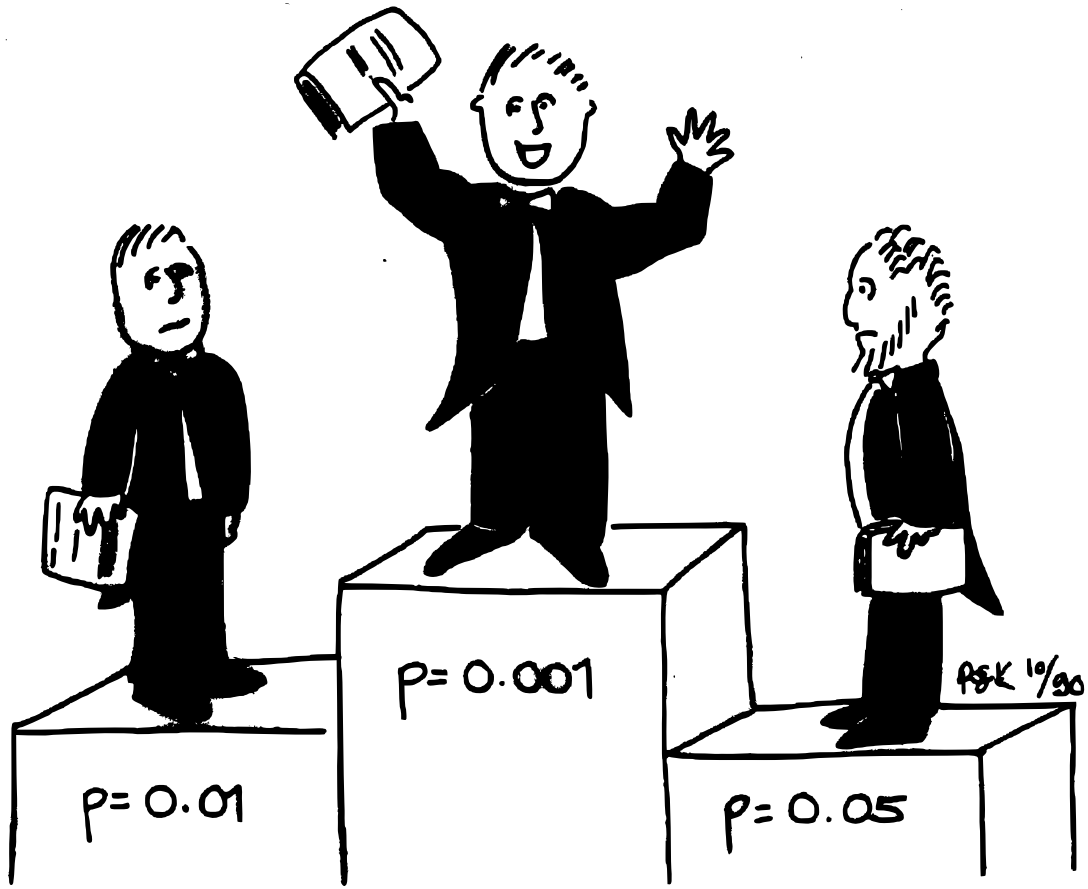
Computer programs show:

> 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

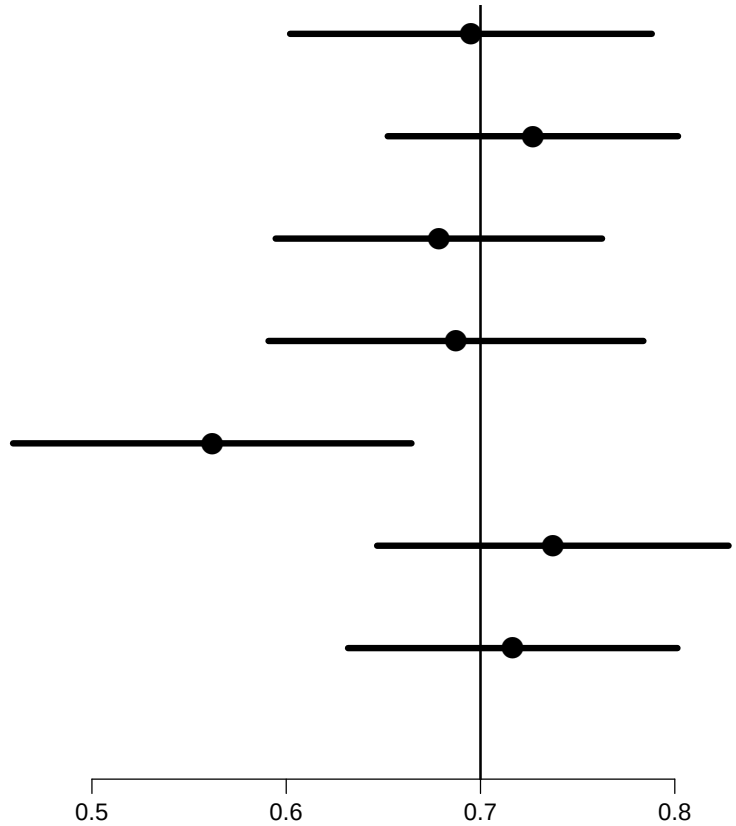
  the probability that the random interval covers the true parameter value equals the confidence level (*e.g.* 95%).

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 $\rho = 0.7$.



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 ID $= 30$ per 10^5 y.
- Standard error SE(ID) $= 19.4$ 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 $\longrightarrow$ precise estimation
  $\longrightarrow$ small statistical uncertainty about parameter.
- wide CI $\longrightarrow$ imprecise estimation
  $\longrightarrow$ 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.

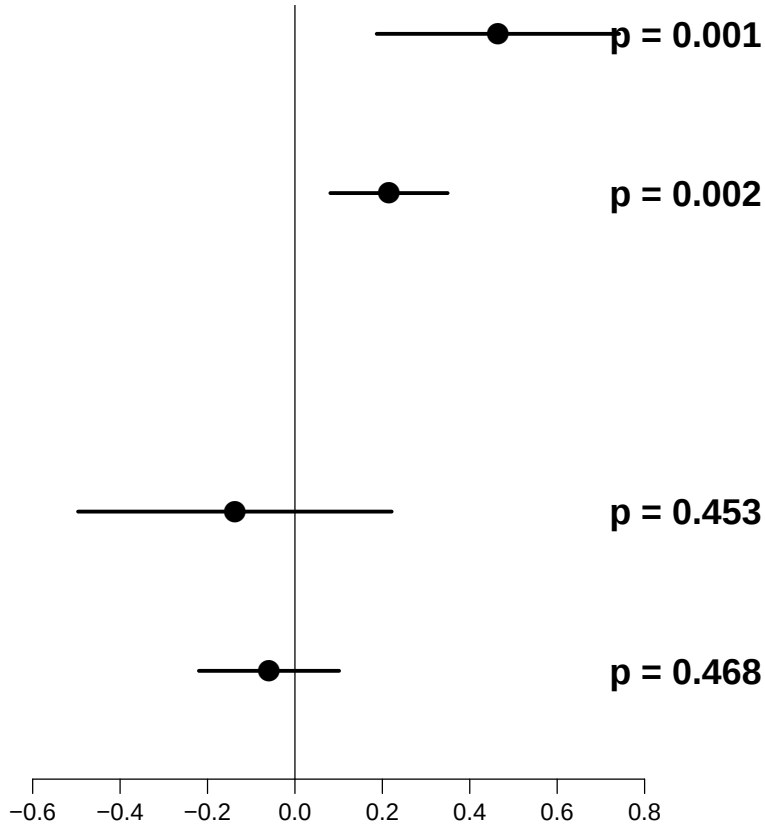
In particular, interpretation of "non-significant" results, *i.e.* large P values:

- narrow CI about H_0 value:
  $\longrightarrow$ results give support to H_0.
- wide CI about H_0 value:
  $\longrightarrow$ 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.

4 CRUDE ANALYSIS

4.1 Single incidence rate

- *Parameter* of interest:

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

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

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

- *Model*: D is Poisson with expectation λY.
- Standard error of rate: $\text{SE}(I) = I/\sqrt{D}$.

Single rate (cont'd)

- Simple approximate 95% CI:

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

where

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

is the 95% *error margin*.

- Problem: When $D < 4$, lower limit < 0!

Single rate (cont'd)

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

$$\text{SE}[\ln(I)] = \frac{1}{\sqrt{D}}$$

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

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

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:

$$[I/\text{EF},\ I \times \text{EF}].$$

- These limits are always > 0 whenever $D \geq 1$.
- When $D = 0$, use the "exact" Poisson limits)
- NB.: 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

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

- The 95% error margin:

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

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 hazard ("relative risk") of cancer in the exposed as compared to the unexposed?
- *Parameter* of interest: true hazard rate ratio

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

- *Null hypothesis* $H_0 : \rho = 1 \Leftrightarrow$ 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 incidence rates by exposure group provide estimates for the true rates:

$$\widehat{\lambda}_1 = I_1 = \frac{D_1}{Y_1}, \qquad \widehat{\lambda}_0 = I_0 = \frac{D_0}{Y_0}$$

Rate ratio (cont'd)

Point estimator of true rate ratio ρ:
empirical *incidence rate ratio* (IR):

$$\widehat{\rho} = \mathsf{IR} = \frac{\hat{\lambda}_1}{\hat{\lambda}_0} = \frac{I_1}{I_0} = \frac{D_1/Y_1}{D_0/Y_0} = \frac{D_1/D_0}{Y_1/Y_0}$$

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

Rate ratio (cont'd)

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

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

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

$$D_1 = 31 \text{ and } D_0 = 26.$$

- Rounded person-years were

$$Y_1 \approx Y_0 \approx 2000 \times 5 \text{ y} = 10000 \text{ y}.$$

Example 8 (cont'd)

Incidence rates 3.1 and 2.6 per 1000 y.

Estimate of true rate ratio ρ with SE, etc.

$$\widehat{\rho} = \text{IR} = \frac{3.1/1000 \text{ y}}{2.6/1000 \text{ y}} = 1.19$$

$$\text{SE}[\ln(\text{IR})] = \sqrt{\frac{1}{31} + \frac{1}{26}} = 0.2659$$

$$\text{EF} = \exp(1.96 \times 0.2659) = 1.68$$

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 study population are collected,
- *controls* are obtained by **incidence density sampling** from those at risk in the study population.
- 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

- However, 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 $\rho = \lambda_1/\lambda_0$ than the rate ratio IR from a full cohort study

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

Rate ratio in case-control study

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

$$\mathrm{SE}[\ln(\mathrm{EOR})] = \sqrt{\frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{C_1} + \frac{1}{C_0}}$$

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

$$\mathrm{CI} = [\mathrm{EOR}/\mathrm{EF}, \mathrm{EOR} \times \mathrm{EF}]$$

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:

$$\text{SE}[\ln(\text{EOR})] = \sqrt{\frac{1}{35}+\frac{1}{637}+\frac{1}{51}+\frac{1}{625}} = 0.2266$$

$$\begin{aligned} \text{EF} &= \exp\{1.96 \times 0.2266\} = 1.45 \\ \text{CI} &= [0.67/1.45, 0.67 \times 1.45] = [0.43, 1.05] \end{aligned}$$

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 incidence rate difference: $\hat{\delta} = \mathsf{ID}$

$$\mathsf{ID} = I_1 - I_0 = \frac{D_1}{Y_1} - \frac{D_0}{Y_0}$$

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

Rate difference (cont'd)

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

$$\mathsf{SE(ID)} = \sqrt{\frac{I_1^2}{D_1} + \frac{I_0^2}{D_0}} = \sqrt{\frac{I_1}{Y_1} + \frac{I_0}{Y_0}}$$

$$\mathsf{EM} = 1.96 \times \mathsf{SE(ID)}$$

$$\mathsf{CI} = [\mathsf{ID} - \mathsf{EM}, \mathsf{ID} + \mathsf{EM}]$$

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

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

Its standard error

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

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):

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

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):
   $\longrightarrow$ 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 the next section.

5 STRATIFIED ANALYSIS

5.1 Shortcomings of crude analysis

Crude analysis is misleading, if

(1) 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**

(2) 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 (ID) and rate ratios (IR) (Doll and Hill, 1966).

	Smokers		Non-smokers			
Age (y)	rate	(D)	rate	(D)	ID	IR
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:

- ID increases by age (at least up to 75 y),
- IR 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 is 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:

			Operation succesful		
Size	Treatment	Pts	Cases	%	%-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

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:

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

Mantel-Haenszel estimators (cont'd)

MH-estimator of the common rate ratio ρ:

$$\mathsf{IR}_{\mathsf{MH}} = \frac{\sum_{k=1}^{K} D_{1k} Y_{0k} / Y_{+k}}{\sum_{k=1}^{K} D_{0k} Y_{1k} / Y_{+k}} = \frac{\sum_{k=1}^{K} w_k I_{1k}}{\sum_{k=1}^{K} w_k I_{0k}}$$

i.e. the ratio of weighted incidence 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 $\ln(\text{IR}_{\text{MH}})$

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

Mantel-Haenszel estimators (cont'd)

Error factor (95%) as before:

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

95 % approximate CI for true hazard ratio ρ:

$$[\mathsf{IR}_{\mathsf{MH}}/\mathsf{EF},\ \mathsf{IR}_{\mathsf{MH}} \times \mathsf{EF}].$$

Example 14 (cont'd)

- Grey hair and cancer: $K = 2$ strata.
- Point estimate:

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

- Standard error: SE[ln(IR)] = 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*:

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

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:

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

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

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

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

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 of hazard ratio ρ:

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

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, $\dots$) according to a **log-linear** model

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

Equivalent expression, **multiplicative model**:

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

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

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

$\delta = \ln(\theta)$, 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

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

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

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

- $\alpha = 12$, baseline rate, i.e. that among those both unexposed to asbestos and non-smokers,
- $\beta = 48$ $(60 - 12)$, rate difference between asbestos exposed and unexposed among non-smokers only,

Example 10 (cont'd)

$\gamma = 108\ (= 120 - 12)$, 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!*

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 or selection (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 ...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)

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!