Analysis of Epidemiological Data

Esa Läärä University of Oulu
Oulu, Finland
https://www.oulu.fi/en/researchers/esa-laara
Bendix Carstensen Steno Diabetes Center
Gentofte, Denmark
http://BendixCarstensen.com

Nordic Summer School in Cancer Epidemiology, Copenhagen, 8/2022

http://BendixCarstensen.com/NSCE/2022

From Thursday 18th August, 2022, 14:59

Contents & programme

1 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)

- Association of oral contraceptive (OC) use with the risk of breast cancer.
- *Null hypothesis* H_0:
  OC use does not affect the 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 incidence rate ratio IR $= 217/187 = 1.16$,
(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 inter- pretation 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** & **interaction**
- 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
chance or **random 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?
- Statistical information is in the number of cases.

2.2 Probability models for incidence

- Random variation in incidence rates (or other measures of occurrence and quantitative variables) is analysed by suitable **probability models**.
- A model represents the assumed **probability distribution** of the relevant observable **random variable(s)**
- It contains **parameters**, constants with unknown value, that are of interest.
- More detailed specification of the distribution is based on certain well-defined mathematical functions.

Constant rate model

- In a sufficiently homogenous population we assume constant "true" but unknown theoretical incidence rate – **hazard** or **intensity** – of contracting cancer over short period of time.
- Example 4: Assume that the hazard of breast in Finnish men aged 65-69 y has a constant value over the whole period 1989-91.
- The unknown rate is denoted by Greek λ.
- **NB.** More complex models are needed for a realistic description of, how the hazard depends on time and other factors.

Probability models (cont'd)

- The observable number of cases D and empirical incidence rate $I = D/Y$ in Y person-years in a given population at risk are:

  random variables with beforehand unpredictable values in given observation periods.

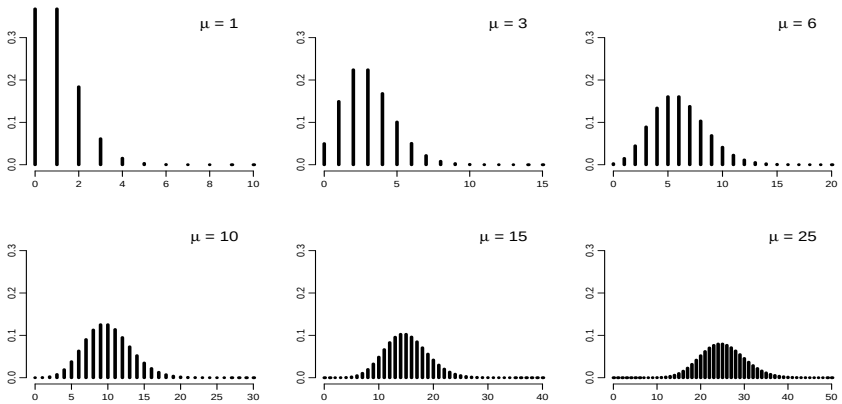
- The *probability distribution* of possible values of the pertinent random variable, D or I, has some known mathematical form.
- In constant rate model, the *parameter* of interest is the unknown true hazard λ.

2.3 Poisson distribution

- Elementary model for the distribution for the number of cases D assuming a constant hazard rate λ.
- Key characteristics of the Poisson distribution for D
  - **expected value** (theoretical mean) $\mu = \lambda Y$, and
  - **standard deviation** $\sigma = \sqrt{\mu}$.
- Example 4: If in 65-69 old males, the hazard or true rate of BC were $22.5/10^6$ y, the expected number of cases would be
  - in 1990: $22.5 \times 89000/10^6 = 2$,
  - 1989-91: $22.5 \times (88 + 89 + 90)/10^3 = 6$.

Poisson distributions with varying μ

Point probabilities of different possible values 0, 1, 2, ..., of D



2.4 Gaussian distribution

When the expected value μ of D is large enough, the Poisson distribution resembles more and more the **Gaussian** or **Normal** distribution, which is

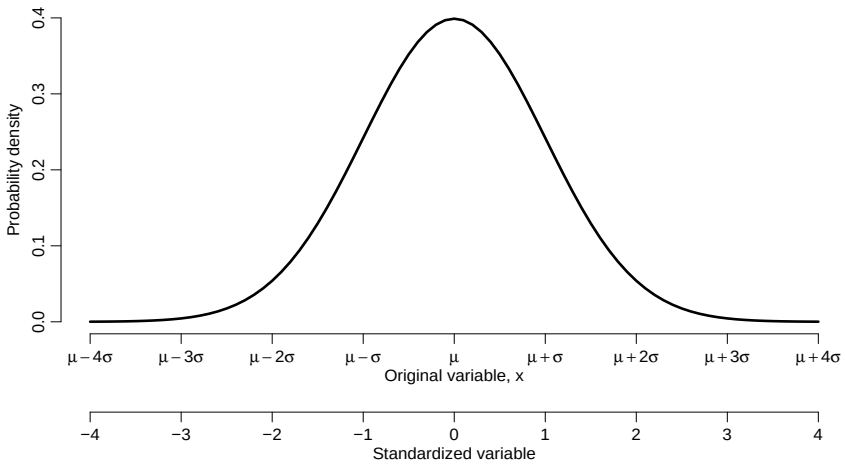
- a common model for continuous variables,
- symmetric and bell-shaped,
- has two parameters:
  μ = expectation, and σ = standard deviation.

Most important use of Gaussian model:

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

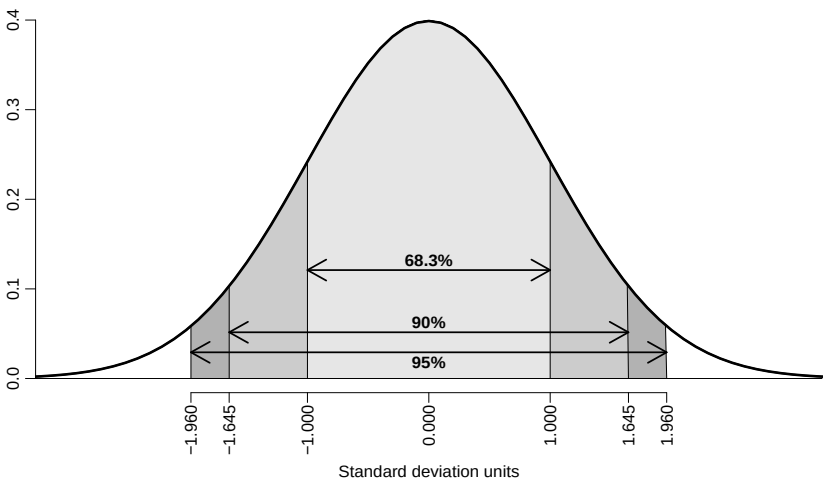
Gaussian distribution (cont'd)

Probability density funtion – the "Bell Curve".



Gaussian distribution (cont'd)

Areas under curve limited by selected quantiles



2.5 Sampling distribution of incidence rate

- Parameter λ = true unknown incidence rate in population.
- Empirical rate $I = D/Y$, **estimator** of λ.
- I is a **statistic**, function of observable quantities. It is a random variable whose:
  - value would vary from one study population – or "sample" – to another in hypothetical repetitions,
  - **sampling distribution** is (under the Poisson model & other conditions) a **scaled** Poisson distribution.

Sampling distribution of incidence rate (cont'd)

- The expected value of I is λ, and its standard deviation is $\sqrt{\lambda/Y}$.
- **Standard error** (SE) of I is the *estimated standard deviation* of the sampling distribution of I

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

$\Rightarrow$ 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's true value is unknown:

  What can we learn about the value?

- Data from empirical study :

  → information on parameter is provided by observed values of relevant statistics,

  → uncertainty on the true value is reduced.

- Still the true value remains unknown.

Inferential questions (cont'd)

- What is the best single-number assesment of the parameter value?
- Is the result compatible or incompatible with a certain value of the parameter proposed beforehand?
- What is a plausible range of values of the parameter, compatible with our observed data?

3.2 Point estimation

- Assessment of 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 at hand.

Point estimation (cont'd)

- Point estimates of parameter are typically obtained by the **method of maximum likelihood**, based on the assumed model and on the observed data,
- **Standard error** (SE) of estimator

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

- SE measures the **(im)precision** of the estimator.

Point estimation and statistical notation:

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

Incidence rate:

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

Example 4: Estimated hazard of BC in 65-69 Finnish men 1989-91:
$\widehat{\lambda} = I = (4 + 1 + 3)/(88 + 89 + 90)/1000 = 30$ per 10^6 y,
when a constant hazard is assumed for the whole period.

Point estimation and statistical notation (cont'd)

Rate ratio:

- true rate ratio $\rho = \lambda_1/\lambda_0$ between exposed and unexposed,
- estimator: $\widehat{\rho} = \text{IR} = I_1/I_0$, **incidence rate ratio**.

Rate difference

- true rate difference $\delta = \lambda_1 - \lambda_0$
- estimator: $\widehat{\delta} = \text{ID} = I_1 - I_0$, **incidence rate difference**.

Example 2: Nurses Health Study, OC and BC

- estimated rate ratio: $\widehat{\rho} = \text{IR} = 217/187 = 1.16$,
- est'd rate difference (per 10^5 y): $\widehat{\delta} = \text{ID} = 217 - 187 = 30$.

3.2 Statistical testing

- *Question*: Are the observed data – summarized by an estimate and its SE – **compatible** with a given value of the parameter?
- Such a given value is often represented in the form of a **null hypothesis**. H_0, which is a statement on the value of the parameter before study.
- In comparative problems H_0 is typically a conservative assumption, *e.g.*
  - "no difference in true rate between exposure groups",
  - "true rate ratio $\rho = 1$".

Purpose of statistical testing

- Evaluation of compatibility or incompatibility of observed data with the null hypothesis H_0
- Checking whether or not the observed difference can reasonably be explained by chance.

NB. These aims are not so ambitious from a quantitative viewpoint.

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 – and given that relevant model assumptions hold – 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

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

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

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

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} when H_0 holds – and assuming that the probability model is sufficiently realistic?
- What is the probability of getting Z larger than observed Z_{obs} if H_0 & assumptions were true.

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

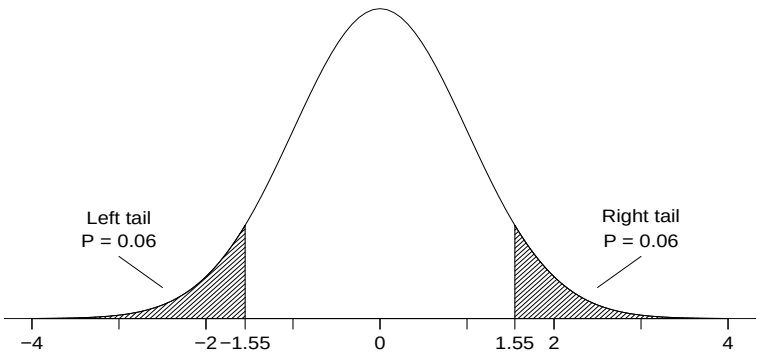
Two-tailed P-value

= probability for test statistic Z being more extreme than the absolute value of Z_{obs},
given the truth of H_0 & model assumptions.

- 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:

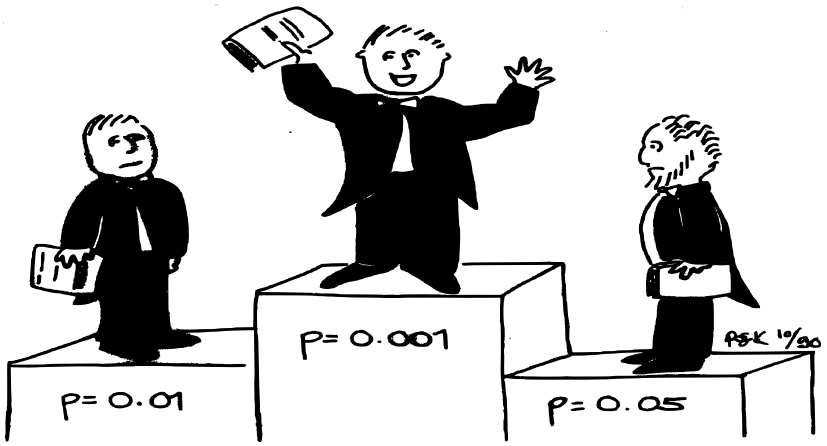
  Under H_0 the probability of getting values of Z larger than the actually observed value 6.43 is < 0.001.

- 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

- Used as a measure of **evidence against** H_0:
  - The smaller is P, the stronger is evidence *against* H_0.
  - Yet, a large P as such **does not** provide supporting evidence *for* H_0.
- Mathematically: Realization of a statistic, *random variable*, whose sampling distribution under H_0 (and given the other assumptions) is uniform in the range $]0, 1[$.
- 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
- Very rough guidelines
  - "large" value $(P > 0.10)$: compatible with H_0 but not necessarily supporting it,
  - "small" value $(P < 0.01)$: indicates evidence against H_0
  - "intermediate" value $(0.01 \leq P \leq 0.10)$: 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 at leas:

- what is a medically relevant deviation of parameter from H_0 (e.g. minimally important elevation of true rate ratio from 1),
- study design: random sampling, randomization or neither,
- possible deviations from model assumptions, like plausible biases due to selection, measurement and/or confounding,
- size of study,
- 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% are sometimes used).
- The limits of CI are statistics, random variables with sampling distribution, such that

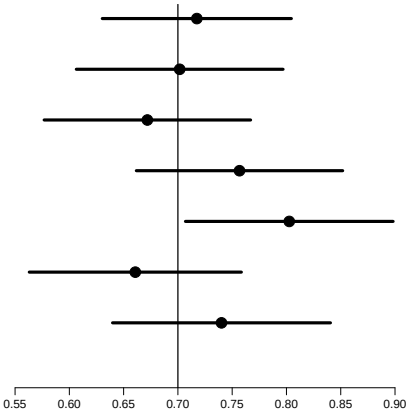
  the probability that the random interval covers the true parameter value equals the confidence level (*e.g.* 95%).

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.
- As with P values, the accuracy of nominal confidence level depends on lack of bias and on validity of model 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 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! – This is 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 compatible, taking into account "probable" random error, and given the model assumptions
  - narrow CI $\rightarrow$ precise estimation $\rightarrow$ small statistical uncertainty about parameter.
  - wide CI $\rightarrow$ imprecise estimation $\rightarrow$ great uncertainty.

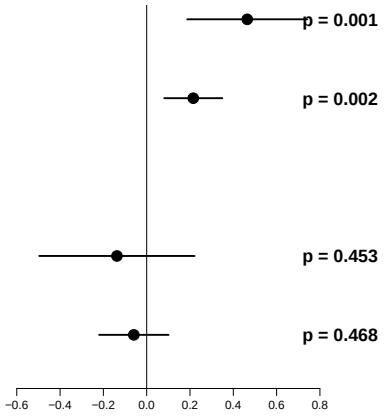
Interpretation of CI (cont'd)

- 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 provide support to H_0.
  - wide CI about H_0 value: $\rightarrow$ results are inconclusive: compatible with H_0, yes, but also with essential deviations from H_0.

  The latter instance 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 – part 1

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

- The description of differences as statistically significant is not acceptable.
- 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.
- 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: λ = true rate in target population
- *Estimator*: $\widehat{\lambda} = I = D/Y$ = cases/person-time,
  = empirical incidence rate in a "representative sample"
- *Model*: D is Poisson-distributed with expectation λY.
- Standard error of empirical rate: $\mathrm{SE}(I) = I/\sqrt{D}$
- Simple approximate 95% CI: $I \pm 1.96 \times \mathrm{SE}(I)$
- Problem: When $D \leq 4$, lower limit ≤ 0!

Single rate (cont'd)

- More accurate approximate CI is based on the log-rate, $\log(I)$, its standard error being $\mathrm{SE}[\log(I)] = 1/\sqrt{D}$.
- From this we get the 95% **error factor** (EF)

$$\mathrm{EF} = \exp\{1.96 \times \mathrm{SE}[\log(I)]\}$$

  where $\exp$ means exponential function or antilog.
- ...and another approximate 95% CI for λ is: $[I/\mathrm{EF},\ I \times \mathrm{EF}]$.
- These limits are always > 0 whenever $D \geq 1$.
- However, when $D = 0$, use the "exact" Poisson limits (or those based on *profile log-likelihood*).

Example 4: BC in Finnish men 65-69 y (cont'd)

- In 1991, the observed rate was 3/90000 y = 33 per 10^6 y.
- Standard error of the rate and the log-rate are

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

- Approximate 95 % CIs for λ in two ways

$$33 \pm 1.96 \times 19 = 33 \pm 37 = [-4, 71] \text{ per } 10^6 \text{ y}$$

$$33 \times \exp(\pm 1.96 \times 0.577) = 33 \stackrel{\times}{\div} 3.1 = [11, 103] \text{ per } 10^6 \text{ y}$$

- Negative lower limit from using $\mathrm{SE}(I)$ is illogical but is avoidable using log-transformation.

Example 4 (cont'd): Using R as pocket calculator

CI computed from I and SE(I) directly

```
> d <- 3
> y <- 90000
> I <- 10^6*d/y
> SE <- I/sqrt(d)
> CI <- I + c(-1,1)*1.96*SE
> round(c(I, SE, CI), 1)
```

```
[1] 33.3 19.2 -4.4 71.1
```

CI computation based on SE[log(I)] and error factor EF:

```
> SElog <- 1/sqrt(d)
> EF <- exp(1.96*SElog)
> CIlog <- c(I/EF, I*EF)
> round( c(I, SElog, EF, CIlog), 2)
```

```
[1]  33.33   0.58   3.10  10.75 103.35
```


Example 4 (cont'd): Modelling single rate with R

Fit a Poisson model for one rate using **logarithmic link function**

```
> library( Epi )
> m1 <- glm( cbind(d, y/10^6) ~ 1, family=poisreg(link="log") )
> round(ci.exp(m1, Exp=TRUE), 1)  # exp-transformation needed
```

```
            exp(Est.) 2.5% 97.5%
(Intercept)      33.3 10.8 103.4
```

We thus got the CI, which was based on SE[log(I)].
The other approximate CI is obtained by using the **identity link**

```
> a1 <- glm( cbind(d, y/10^6) ~ 1, family=poisreg(link="identity") )
> round(ci.exp(a1, Exp=FALSE), 1) # exp- transformation not needed
```

```
            Estimate 2.5% 97.5%
(Intercept)     33.3 -4.4  71.1
```

NB. Command `ci.exp()` returns the results concisely.

4.2 Rate ratio in cohort study

- *Question*: What is the relative hazard 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* $H_0 : \rho = 1 \Leftrightarrow$ exposure has no effect.

Rate ratio (cont'd)

- Summary of results 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

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

Rate ratio (cont'd)

- Estimator of true rate ratio ρ: incidence rate ratio (IR):

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

- Standard error, 95% error factor, and 95% CI for ρ:

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

- **NB.** Random error depends inversely on numbers of cases!

Example 5: Helsinki Heart Study (HHS)

- In the HHS 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 5: HHS (cont'd)

- Incidence rates 3.1 and 2.6 per 1000.
  Point estimate of the true rate ratio ρ, its SE, and EF:

$$\widehat{\rho} = \text{IR} = 3.1/2.6 = 1.19$$
$$\text{SE}[\log(\text{IR})] = \sqrt{1/31 + 1/26} = 0.2659$$
$$\text{EF} = \exp(1.96 \times 0.2659) = 1.68$$

- 95 % CI for ρ : $\quad [1.19/1.68,\ 1.19 \times 1.68] = [0.7,\ 2.0]$
- $H_0 : \rho = 1$; test statistic $Z = \log(1.19)/0.2659 = 0.654$, two-tailed $P = 0.51$.
- *Interpretation?*

Example 5: Rates and their ratio with R

Poisson & log-link: Estimating the two rates & CIs separately – **NB.** CIs of individual rates are seldom of interest as such

```
> D <- c(26, 31) ; Y <- c(10, 10) ; gemf <- factor(0:1)
> m2 <- glm( cbind(D,Y) ~ gemf - 1, family=poisreg )
> round(ci.exp( m2 ), 2)

      exp(Est.) 2.5% 97.5%
gemf0       2.6 1.77  3.82
gemf1       3.1 2.18  4.41
```

Estimating the rate in unexposed, and the rate ratio & its CI

```
> m2b <- glm( cbind(D, Y) ~ gemf, family=poisreg )
> round(ci.exp(m2b, pval=TRUE), 2)

            exp(Est.) 2.5% 97.5%    P
(Intercept)      2.60 1.77  3.82 0.00
gemf1            1.19 0.71  2.01 0.51
```


4.3 Rate difference in a cohort

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

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

- Same layout for summary data as above for cohort study.
- Point estimator of δ, the empirical rate difference: $\widehat{\delta} = \text{ID}$

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

Rate difference (cont'd)

- Standard error of incidence rate difference, 95% **error margin** (EM) & approximate 95% confidence interval (CI) for δ:

$$\text{SE(ID)} = \sqrt{\frac{I_1^2}{D_1} + \frac{I_0^2}{D_0}}$$
$$\text{EM} = 1.96 \times \text{SE(ID)}$$
$$\text{CI} = [\text{ID} - \text{EM}, \text{ID} + \text{EM}]$$

- Log-transformation is not meaningful here; original scale is used.
- Random error again depends inversely on number of cases.

Example 5: HHS (cont'd)

- Observed rate difference btw exposed and unexposed was
$$\text{RD} = 3.1 - 2.6 = +0.5 \text{ per } 10^3 \text{ y},$$
- Its standard error
$$\text{SE(RD)} = \sqrt{3.1^2/31 + 2.6^2/26} = 0.755 \text{ per } 10^3 \text{ y}$$
and 95% approximate CI:
$$0.5 \pm 1.96 \times 0.755 = 0.5 \pm 1.5 = [-1.0, 2.0] \text{ per } 10^3 \text{ y}.$$
- Ranges from negative to positive values, which is logical, because the rate difference can have either minus or plus sign.
- *Interpretation?*

Example 5 (cont'd): Rate difference using R

Poisson model with identity link: estimating first the two rates separately, and then the rate in unexposed plus the rate difference:

```
> a2 <- glm( cbind(D, Y) ~ gemf - 1, family=poisreg(link='identity') )
> round( ci.exp( a2, Exp=FALSE ), 1)

      Estimate 2.5% 97.5%
gemf0      2.6  1.6   3.6
gemf1      3.1  2.0   4.2

> a2b <- glm( cbind(D,Y) ~ gemf, family=poisreg(link='identity') )
> round( ci.exp( a2b, Exp=FALSE ), 1)

            Estimate 2.5% 97.5%
(Intercept)      2.6  1.6   3.6
gemf1            0.5 -1.0   2.0
```

4.4 Rate ratio in case-control study

- Parameter of interest: $\rho = \lambda_1/\lambda_0$ – same as in cohort study.
- Required case-control design:
  1. **incident cases** occurring during a given period in the source population are collected,
  2. **controls** are obtained by **density sampling** from those at risk in the source.
  3. exposure is ascertained in cases and chosen controls.

Rate ratio in case-control study

Summary 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, respectively!

- Rates as such are not directly estimable from these data.

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 incidence 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 log(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 6: Use of mobile phone and brain cancer

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

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

Standard error of log(EOR), and approximate CI for ρ:

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

$$\text{CI} = 0.67 \overset{\times}{\div} \exp\{1.96 \times 0.2266\} = [0.43, 1.05].$$

NB. Model-adjusted estimate: EOR = 0.6 (95% CI 0.3 to 1.0).

Example 6 (cont'd): Crude estimation with R

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

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

> ccmod <- glm( cbind(Ca,Co) ~ Ex, family=binomial ) # Note: a new family
> round( ci.exp( ccmod ), 2)

            exp(Est.) 2.5% 97.5%
(Intercept)      1.02 0.91  1.14
Ex>15            0.67 0.43  1.05
```

- Intercept is meaningless; only estimate for exposure is relevant.
- BTW. The model fitted here is **logistic regression model**.

4.5 Matched case-control study – 1:1 matching

- Suppose each case was matched with 1 individual control subject. Pairwise data w.r.t a binary exposure is tabulated

	Controls		
Cases	Exposed	Unexposed	Total
Exposed	s	t	D_1
Unexposed	u	v	D_0
Total	C_1	C_0	$D = C$

- Analysis ignoring matching would be based on crude EOR $= (D_1/D_0)/(C_1/C_0)$. – Yet, this may be biased.

Matched case-control study (cont'd)

- Proper analysis is based on those case-control pairs which are **discordant** w.r.t exposure, because exposure-**concordant** pairs are non-informative about the exposure effect.
- Valid estimator of rate ratio ρ is $\text{EOR}_M = t/u$.
- Standard error (SE) of $\log(\text{EOR}_M)$ and approximate 95 % confidence interval (CI) for ρ:

$$\text{SE} = \sqrt{1/t + 1/u}, \qquad \text{CI} = \text{EOR}_M \overset{\times}{\div} \exp(1.96 \times \text{SE}).$$

- **NB.** This simple analysis is actually not "crude", as it provides an estimate of ρ that is adjusted for the matching factors.

Example 7: 1:1 matched study

- Cross-tabulation of exposure status in 200 case-control pairs:

	Controls		
Cases	Exposed	Unexposed	Total
Exposed	20	60	80
Unexposed	20	100	120
Total	40	160	200

- Crude EOR $= (80/120)/(40/160) = 2.67$. – Matched analysis:

```
> t <- 60; u <- 20; EOR_M <- t/u;
> SE <- sqrt(1/t+1/u); EF <- exp(1.96*SE)
> CI <- c( EOR_M/EF, EOR_M*EF ); round( c(EOR_M, SE, CI), 2)
[1] 3.00 0.26 1.81 4.98
```

4.6 Analysis of proportions

- Suppose we have cohort data with a fixed **risk period** with complete follow-up for all n subjects (no censoring).
- In this setting the **risk** π of the disease over the risk period is easily estimated by simple **incidence proportion** also known as **cumulative incidence** – or even "risk":

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

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

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

Analysis of proportions (cont'd)

- Proportions 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 $\text{QD} = Q_1 - Q_0$ are treated on the original scale by using the error margin principle: $\text{CI} = \text{QD} \pm 1.96 \times \text{SE(QD)}$.
- Analysis of ratios $\text{QR} = Q_1/Q_0$ leans on SEs of log-proportions & error factor, etc.
- Details of standard error formulas are somewhat different from those of rates.

4.7 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

5 STRATIFIED ANALYSIS

5.1 Shortcomings of crude analysis

- The comparative measure (like rate ratio) for the risk factor of interest is not constant, but varies by other determinants of the disease

$\Leftarrow$ heterogeneity of the comparative parameter:
effect modification or **interaction**

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

$\Rightarrow$ bias in comparison or **confounding**

$\Leftarrow$ exposure varies across other determinants

Need models for outcome including items for

- primary variable ("exposure")
- secondary variable ("stratum")
- **effect modification** represented by a **product term** in an **interaction model** "exposure×stratum", in which exposure effect is assumed **heterogenous** across strata
- **confounding**, which is adjusted for by a **main-effects model** "exposure+stratum", where exposure is assumed to have **same** effect across strata

Handling for effect modification and confounding

- **Stratification** of data
by suspected modifying and/or confounding factor(s) and use of classical **summary estimators**
- Conceptually simpler,
and technically less demanding approach is
regression modeling
- Regression modeling is feasible because we have computers and software
- Classical summary estimators were important to learn for the teachers, who got their initial training before the computer age (BxC & EL . . .)

5.2 Effect modification – Example 8

Incidence rates (per 10^5 y) of lung cancer by occupational asbestos exposure and smoking:

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 **homogeneous**
- Rate difference: **heterogeneous**:
  The value of rate difference is modified by smoking.

Smoking is thus an **effect modifier** of asbestos exposure

- on the absolute scale (rates) but
- **not** on the relative scale (log-rates)

Therefore, it is more accurate to talk about **effect-measure modification**.

Example 9: CHD rate/10^3 y by factor E & age

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

- Rate ratio is modified by age, but rate difference is not.
- Repating the message: There is no such thing as effect modification as such without reference to the **scale** of the effect – additive or multiplicative.

Example 10: Famous real study

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 10: CHD death by smoking (cont'd)

- Both comparative measures appear heterogeneous:
  - ID increases by age (at least up to 75 y)
  - IR decreases by age
- No single-parameter comparison – common rate ratio or rate difference – captures adequately the joint pattern of rates.

NB. In many other real life instances, all comparative measures are more or less heterogeneous across categories of other determinants of disease, i.e. some modification is always to be expected.

Evaluation of modification

- Modification or its absence is an inherent property of the phenomenon.
- It depends on the **scale** on which it is measured
- Cannot be removed or "adjusted" for

When assessing effect-modification, ask yourself

- what is the **scale** that we wish to use for description of effects – multiplicative for ratios or additive for differences?
- how will we **report** the results in the presence of modification?

Evaluation of modification (cont'd)

- Statistical tests for heterogeneity exist. However, with small amounts of data they tend to be insensitive to deviations from homogeneity and thus rarely helpful

$\Rightarrow$ Especially when a "non-significant" P-value for interaction is obtained, it is tempting to assume "no modification":

$+$ simpler analysis and presentation of results,

$-$ misleading if essential modification is present.

Example 10 (cont'd): Analysis with R

Entering the data

```
> I <- c(6.1, 24, 72,147,192, 1.1,11,49,108,212)
> D <- c( 32,104,206,186,102, 2  ,12,28, 28, 31)
> Y <- D/I # person-times in units of 10^4 years
> smk <- factor( rep(1:2,each=5), labels=c("Smoke","non-Sm") )
> age <- factor( rep(seq(35,75,10),2) )
> data.frame(D,Y=round(Y, 2),age,smk)
```

```
     D    Y age    smk
1   32 5.25  35  Smoke
2  104 4.33  45  Smoke
3  206 2.86  55  Smoke
4  186 1.27  65  Smoke
5  102 0.53  75  Smoke
6    2 1.82  35 non-Sm
7   12 1.09  45 non-Sm
8   28 0.57  55 non-Sm
9   28 0.26  65 non-Sm
10  31 0.15  75 non-Sm
```


Example 10 (cont'd): Modification by age?

Analysing rate ratios both without and with modification, and testing the interaction by **deviance** statistic

```
> options(show.signif.stars=FALSE)
> ma <- glm( cbind(D,Y) ~ age + smk, family=poisreg )
> mi <- update( ma, . ~ . + age:smk ) # add the interaction
> anova( ma, mi, test="Chisq" )
```

```
Analysis of Deviance Table

Model 1: cbind(D, Y) ~ age + smk
Model 2: cbind(D, Y) ~ age + smk + age:smk
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4     11.993
2         0      0.000  4   11.993   0.0174
```

Test result indicates evidence for modification of rate ratio by age.

Example 10 (cont'd): Modification by age?

Analysing rate differences both without and with modification, and testing the interaction.

```
> aa <- glm( cbind(D,Y) ~ age + smk, family=poisreg(link='identity') )
> ai <- update( ma, . ~ . + age:smk ) # add the interaction
> anova( aa, ai, test="Chisq" )

Analysis of Deviance Table

Model 1: cbind(D, Y) ~ age + smk
Model 2: cbind(D, Y) ~ age + smk + age:smk
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4     7.7434
2         0     0.0000  4   7.7434   0.1014
```

Weak evidence for modification of rate difference by age.
However, this does not mean that we have sufficient support for the assumption of homogenous rate difference.

5.3 Confounding - Example 11

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

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

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

PN appears more successful than OS?

Example 11 (cont'd): Stratification

Results stratified by initial diameter size of the stone:

Size	Treatment	Pts	Success	% Success	%-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 11 (cont'd): Confounding

- 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, because it is

(1) a 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.

NB. There is no statistical test for confounding!

Example 10 (cont'd): Confounding by indication

- This is an instance of **confounding by indication**, an important issue in clinical epidemiology:
  – patient status affects choice of treatment,

  $\Rightarrow$ bias in comparing treatments.
- This bias is best avoided in planning:
  – randomized allocation of treatment.

Example 12: Gray hair and cancer incidence

Age	Gray hair	Cases	P-years ×1000	Rate /1000 y	IR
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 seems to vanish when controlling for age.

Means for control of confounding

Design:

- Randomization
- Restriction
- Matching

Analysis:

- Stratification
- Regression modeling

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

- Stratify by levels of the potential confounding/modifying factor(s)
- Compute stratum-specific estimates
- Evaluate similarity of the stratum-specific estimates by "eye-balling", or test of heterogeneity.
- If the comparative measure is judged to be homogeneous enough, calculate an adjusted estimate.
- If effect modification is judged to be present:
  - report stratum-specific estimates with CIs,
  - if desired, calculate an adjusted summary estimate by appropriate standardization.

5.5 Adjusted estimation of rate ratio

- Suppose that the true rate ratio ρ is sufficiently homogeneous across strata (no modification), but confounding is present.

$\Rightarrow$ Crude estimator IR of ρ is biased.

- **Adjusted estimator**, controlling for confounding, must be used.
- These estimators are **weighted** averages of stratum-specific estimators.

Adjusted summary estimators

Different weighting methods:

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

Preferred: Maximum likelihood based on an appropriate model.
– **NB.** These models are actually **regression models** (see ch 6).

Useful methods in some descriptive analyses: CMF & SMR.

Example 12: Gray hair & cancer with R

Data entry

```
> D <- c(6,11,60,19)
> Y <- c(10,20,15,5)
> age <- factor( c("Not old","Not old","Old","Old") )
> hair <- factor( c("Gray","Col","Gray","Col") )
> data.frame( D, Y, age, hair )

   D  Y     age hair
1  6 10 Not old Gray
2 11 20 Not old  Col
3 60 15     Old Gray
4 19  5     Old  Col
```


Example 12: Gray hair & cancer with R (cont'd)

Crude and adjusted estimate of ρ by Poisson model:

```
> library( Epi )
> round(ci.exp( glm( cbind(D,Y) ~ hair      , family=poisreg ) ), 2)

            exp(Est.) 2.5% 97.5%
(Intercept)       1.2 0.84  1.72
hairGray          2.2 1.43  3.39

> round(ci.exp( glm( cbind(D,Y) ~ hair + age, family=poisreg ) ), 2)

            exp(Est.) 2.5% 97.5%
(Intercept)      0.56 0.34  0.92
hairGray         1.06 0.67  1.68
ageOld           6.80 3.90 11.88
```

The adjusted estimate of common rate ratio for the effect of gray hair is in between the two stratum-specific estimates.

Example 13: Case-control study

Alcohol and oesophageal cancer (Tuyns 1977, see B&D)

- 205 incident cases,
- 770 randomly sampled 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	

Recall: EOR estimates the rate ratio ρ!

Example 13 (cont'd): Crude analysis

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

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

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

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

The crude exposure odds-ratio of oesophageal cancer, comparing high vs. low alcohol consumption is 5.64 (95 % CI 4.00 to 7.95).

Example 13 (cont'd): 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! Random sampling of controls: inefficient design
Should have employed stratified sampling by age.

Example 13: Stratification with R

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

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


Example 13: Stratum-specific estimates with R

The "age/alc" term in the model formula produces an EOR for alc in each age class separately

```
> mi <- glm( cbind(ca,co) ~ age/alc, family=binomial )
> round( ci.exp( mi ), 2 )
```

```
                 exp(Est.) 2.5% 97.5%
(Intercept)  0.000000e+00 0.00   Inf
age35        2.345328e+10 0.00   Inf
age45        1.170624e+11 0.00   Inf
age55        1.881661e+11 0.00   Inf
age65        3.147003e+11 0.00   Inf
age75        1.985206e+11 0.00   Inf
age25:alc>80 8.547416e+10 0.00   Inf
age35:alc>80 5.050000e+00 1.27 20.02
age45:alc>80 5.670000e+00 2.80 11.46
age55:alc>80 6.360000e+00 3.45 11.73
age65:alc>80 2.580000e+00 1.22  5.47
age75:alc>80 1.755246e+11 0.00   Inf
```


Example 13: Stratum-specifc estimates (cont'd)

. . . extracting only the relevant estimates:

```
> round( ci.exp( mi, subset="alc" ), 2 )
```

```
                 exp(Est.) 2.5% 97.5%
age25:alc>80 8.547416e+10 0.00   Inf
age35:alc>80 5.050000e+00 1.27 20.02
age45:alc>80 5.670000e+00 2.80 11.46
age55:alc>80 6.360000e+00 3.45 11.73
age65:alc>80 2.580000e+00 1.22  5.47
age75:alc>80 1.755246e+11 0.00   Inf
```

- The age-specific EORs are quite variable.
- Random error in some of them apparently quite large.
- No clear pattern in the possible modification of rate ratio.

Example 13 (cont'd): Modification of rate ratio?

Test of modification using **deviance** statistic

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

Analysis of Deviance Table

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

- Some evidence against homogeneity of rate ratio, but no clear pattern of modification.

Example 13 (cont'd): Common rate ratio ρ?

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

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

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

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

- No clear modification of rate ratio was detected.
- Crude estimate of ρ: 5.64 (95 % CI 4.00 to 7.95)
- Age-adjusted estimate: 5.31 (95 % CI 3.66 to 7.70)

6 REGRESSION MODELLING

6.1 Limitations of stratified analysis

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

Limitations (cont'd)

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

Many of these limitations may be overcome – at least to some extent – by regression modelling.

Key concept – again: **statistical model**

Log-linear model for rates

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

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

Equivalent expression, **multiplicative model**:

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

Log-linear model: Meaning of parameters

Model parameters

α $= \log(\lambda_0) =$ **intercept**, log-baseline rate λ_0
(i.e. rate when $X = Z = \cdots = 0$)

β $= \log(\rho) =$ **slope**,
change in $\log(\lambda)$ for unit change in X,
adjusting for the effect of Z (& $U, V, \ldots$)

e^{β} $= \rho =$ true rate ratio for unit change in X.

The rate ratio for the effect of X is assumed constant in this model, i.e. not modified by Z or any other variable.

Example 8 (cont'd): Lung cancer

Dichotomous explanatory variables coded:

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

Log-linear model for the rates

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

Example 8 (cont'd): Values of variables

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

Note: There will be **4** lines in the dataset, one for each combination of exposure and smoking

Example 8 (cont'd): Analysis with R

Entering the data.

NB. The data here are artificial assuming the amount of person-years among asbestos exposed is $1/4$ of that among non-exposed, and there is no mutual confounding.

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

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


Example 8 (cont'd): Analysis with R

- Regression modelling: Multiplicative Poisson model, requiring logarithmic **link function** (default for Poisson in R).
- Two equivalent approaches
  - response: `D`, offset: `log(Y)` – mostly used in literature,
  - response: `cbind(D,Y)`, and `family=poisreg`
    this latter approach is also useful for fitting **additive** models, like model `ma` below, requiring **identity** link.

```
> library( Epi )
> mo <- glm(       D    ~ asb + smk, family=poisson, offset=log(Y) )
> mm <- glm( cbind(D,Y) ~ asb + smk, family=poisreg )
> ma <- glm( cbind(D,Y) ~ asb + smk, family=poisreg(link=identity) )
```


Example 8 (cont'd): Log-linear model summary

```
> summary( mo )

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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

Number of Fisher Scoring iterations: 3
```

Example 8 (cont'd): Extracting estimates

```
> round( ci.exp( mo, Exp=FALSE ), 3)

            Estimate  2.5% 97.5%
(Intercept)    2.485 2.087 2.883
asb            1.609 1.381 1.838
smk            2.303 1.907 2.698

> round( ci.exp( mm, Exp=FALSE ), 3)

            Estimate  2.5% 97.5%
(Intercept)    2.485 2.087 2.883
asb            1.609 1.381 1.838
smk            2.303 1.907 2.698
```

Estimates of model parameters are the same for the two modelling approaches.

Example 8 (cont'd): Estimating rate ratios

Estimates of rate ratios are obtained by exponential transformation of the estimates of model parameters

```
> cbind( round(ci.exp(mm,Exp=FALSE),3), round(ci.exp(mm),2) )

            Estimate  2.5% 97.5% exp(Est.) 2.5% 97.5%
(Intercept)    2.485 2.087 2.883        12 8.06 17.87
asb            1.609 1.381 1.838         5 3.98  6.29
smk            2.303 1.907 2.698        10 6.73 14.85
```

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

Example 8 (cont'd): Rates from parameters

Fitted rates for all 4 asbestos/smoking combinations can be recovered from the model formula.

	Rates		Obtained from parameters	
Asbestos	Smok	Non-sm	Smok	Non-sm
exposed	600	60	$\exp(\alpha+\gamma+\beta)$	$\exp(\alpha+\beta)$
unexposed	120	12	$\exp(\alpha+\gamma)$	$\exp(\alpha)$
Rate ratio	5	5	$\exp(\beta)$	$\exp(\beta)$

Log-linear model with interaction

Model for describing effect modification (two regressors only)

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

equivalently

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

where α is as before, but

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

Interaction parameter

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

For binary X and Z we have

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

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

Interaction model: Rates from parameters

	Rates		Obtained from parameters	
Asbestos	Smok	Non-sm	Smok	Non-sm
exposed	600	60	$\exp(\alpha+\gamma+\beta+\delta)$	$\exp(\alpha+\beta)$
unexposed	120	12	$\exp(\alpha+\gamma)$	$\exp(\alpha)$
Rate ratio	5	5	$\exp(\beta+\delta)$	$\exp(\beta)$

Example 8 (cont'd): Lung cancer

Fitting a log-linear interaction model `mi` and comparing it with previously fitted main-effects model `mm`

```
> mi <- glm( cbind(D,Y) ~ asb + smk + I(asb*smk), family=poisreg )
> round( cbind(ci.exp( mi ), rbind(ci.exp( mm ),NA) ), 3 )
```

```
             exp(Est.)  2.5%  97.5% exp(Est.)  2.5%  97.5%
(Intercept)         12 6.815 21.130        12 8.060 17.867
asb                  5 2.340 10.682         5 3.977  6.286
smk                 10 5.524 18.101        10 6.733 14.853
I(asb * smk)         1 0.451  2.217        NA    NA     NA
```

- No interaction on the multiplicative scale: Interaction parameter estimated as 1,
- Asbestos and smoking effects remain unchanged,
- Yet, SEs are larger, as they refer to estimated rate ratios for levels $X = 0$ and $Z = 0$, resp.; not both levels **jointly**

Additive model with interaction for rates

General form with two regressors

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

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 $X = 0$.

Additive model (cont'd)

δ $=$ interaction parameter.

- For binary X and 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,

Ex. 8: Additive model with & w/o interaction

```
> mai <- glm( cbind(D,Y) ~ asb + smk + asb*smk, family=poisreg(link=identity) )
> ma <- glm( cbind(D,Y) ~ asb + smk, family=poisreg(link=identity) )
> round( cbind( ci.exp( mai, Exp=FALSE), rbind(ci.exp(ma, Exp=FALSE),NA )), 1 )

            Estimate  2.5% 97.5% Estimate  2.5% 97.5%
(Intercept)       12   5.2  18.8     10.2   3.9  16.4
asb               48  16.9  79.1    202.9 156.8 249.0
smk              108  85.5 130.5    136.1 112.4 159.8
asb:smk          432 328.8 535.2       NA    NA    NA

> anova(ma, mai)

Analysis of Deviance Table

Model 1: cbind(D, Y) ~ asb + smk
Model 2: cbind(D, Y) ~ asb + smk + asb * smk
  Resid. Df Resid. Dev Df Deviance
1         1     80.853
2         0      0.000  1   80.853
```

Very strong evidence for modification of rate difference.

Example 8 (cont'd): Additive model with interaction

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

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

Model fitting

Output from computer packages (like **R**, Stata, etc.) will give:

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

May be difficult to interpret!

Model checking & diagnostics:

- assessment whether model assumptions seem reasonable and sufficiently compatible with observed data
- involves fitting and comparing different models

Problems in modelling

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

$\Rightarrow$ possible bias in effect estimation & underestimation of SEs.

- Multitude of models fit well to the same data
  which model to choose?
- Software easy to use:

  . . . too easy to fit models blindly

  . . . possibility of unreasonable results

Modelling

- Modeling should not substitute, but complement crude and stratified descriptive analyses:
- Crude analyses should be seen as initial modeling steps:
  one or two effects in the model
- Final model for used for reporting developed mainly from subject matter knowledge
- Adequate training and experience required.
- Ask help from a professional statistician!
- **Collaboration** is the keyword.

7 CONCLUDING REMARKS

Epidemiologic study is a

Measurement excercise

Target of measurement: some **parameter** of interest, like

- incidence rate
- rate ratio
- 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 *causal 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, part 2

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!