Nordic Summerschool of Cancer Epidemiology

Bendix Carstensen Steno Diabetes Center Copenhagen
Herlev, Denmark
http://BendixCarstensen.com

Esa Läärä University of Oulu
Oulu, Finland

Danish Cancer Society, August 2024 / Janauy 2025

http://BendixCarstensen.com/NSCE/2017

From C:\Bendix\teach\NSCE\2024\slides/slides.tex

Monday 19th August, 2024, 14:59

Introduction

- Starters
- Analysis and statistics
- Uses of statistics in epidemiology
- References

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?

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	94,029	217
Never	240	128,528	187

Results:

- Observed rate ratio RR $= 217/187 = 1.16$
- P-value 0.12
- 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?

Analysis and statistics

By **analysis** we mean **statistical** analysis.

Statistics:

- (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.
- (plural) the numerical facts or data themselves

(Webster's Dictionary)

Use of statistics in epidemiology

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

Use of statistics

Basic approaches and tools:

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

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.

BxC: B. Carstensen (2022).
Epidemiology with R. OUP, Oxford.

Chance

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,August 2024 / Januay 2025

Chance variation

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

Systematic and random variation

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

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

This is **systematic variation**.

Systematic and random variation

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

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

Example: Smoking and lung cancer

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

Example: Breast cancer

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

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

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

Example: Breast cancer

Look at observed numbers of **cases**!

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

Reality of changes over the years?

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

Simple probability model for cancer occurrence

Assume that the population is **homogeneous**

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

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

Simple probability model for cancer occurrence

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

How a probability model works

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

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

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

Component of a probability model

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

Statistics

The opposite of a probability models:

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

Thus **statistical modelling** is how to estimate parameters from observed data. Data → Model

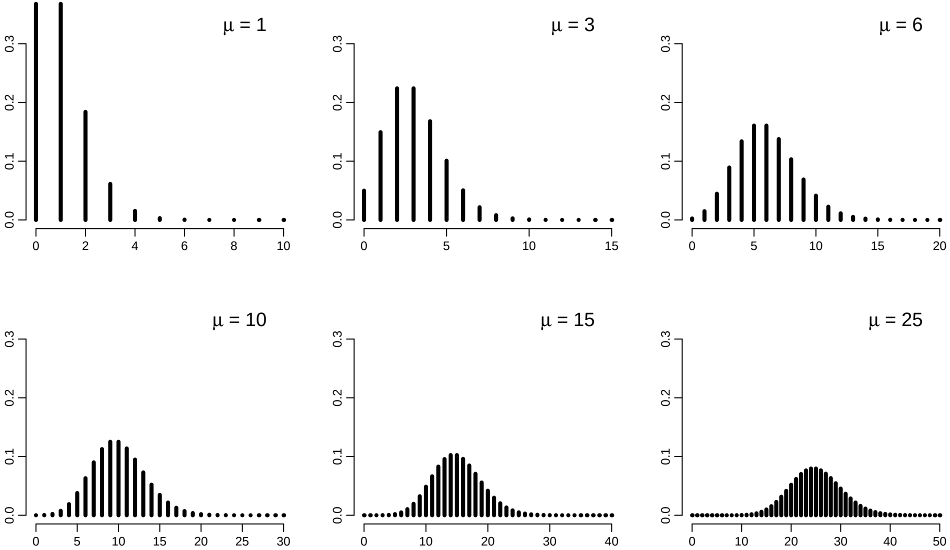
Statistics — the workings

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

Poisson and Gaussian models

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

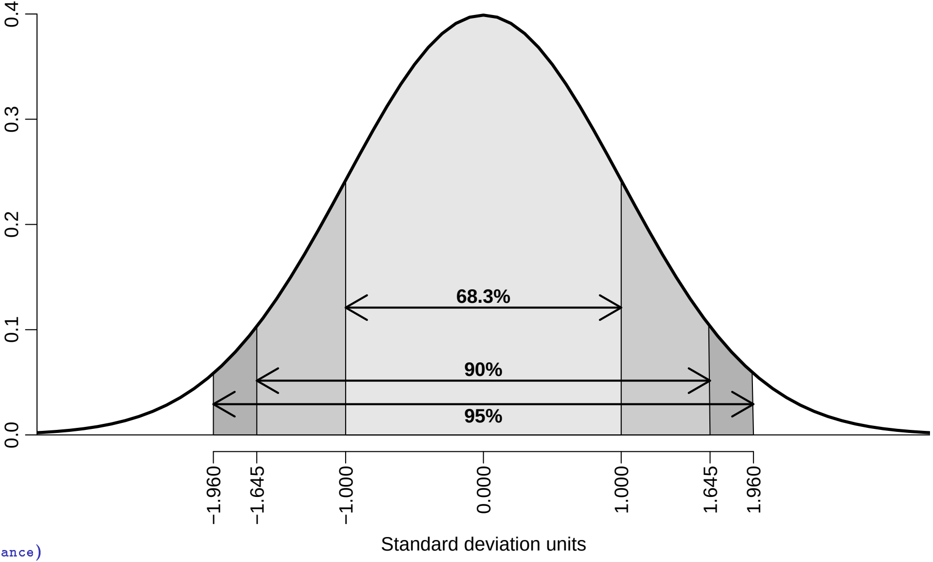
Poisson distribution with different means (μ)



Normal (Gaussian) distribution

- common model for continuous variables
  - symmetric and bell-shaped
  - has two parameters:
    - μ = expectation or mean
    - σ = standard deviation
- Central limit theorem:
  A sum of many small independent quantities will follow a normal distribution
- Conseqence:
  When we compute various functions based on our data we can approximate the distribution with the normal distribution
- . . . so we just need to compute mean and standard deviation — the shape is fixed by the theory

Areas under curve limited by selected quantiles



Example: Observed incidence rate

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

Example: Observed incidence rate

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

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

⇒ Random error depends inversely on the number of cases.

⇒ s.e. of R is proportional to R.

Example: Observed incidence rate

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

⇒ Observed R is with 95% proability in the interval

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

⇒ with 95% probability λ is in the interval

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

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

Chance summary

- Observations vary systematically by **known** factors
- Observations vary randomly by **unknown** factors
- Probability model describes the random variation
- We observe random variables — draws from a probability distribution
- Central limit theorem allows us to quantify the random variation
- . . . and construct confidence interval

Inference

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,August 2024 / January 2025

Models and data

- A probability model can be used to generate data (by simulation) — from **model** to **data**
- Inference is the **inverse**:
- What model generated the data?
- — from data to model
- . . . if we know data we can say something sensible about disease process in the population that generated data

Models and data — model components

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

Statistical concepts

- Probability: parameters → data
- Statistics: data → parameter(estimate)s
- Notation:
  - Parameter denoted by a Greek letter, β
  - Estimator & estimate by the same Greek letter with "hat", $\hat{\beta}$
- Example: Incidence rate:
  - Theoretical rate — the value of the rate in the model that could have generated data: λ
  - Estimator: $\widehat{\lambda} = R = D/Y$, empirical rate.
- . . . but where did the D/Y come from?

Maximum likelihood principle

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

Likelihood

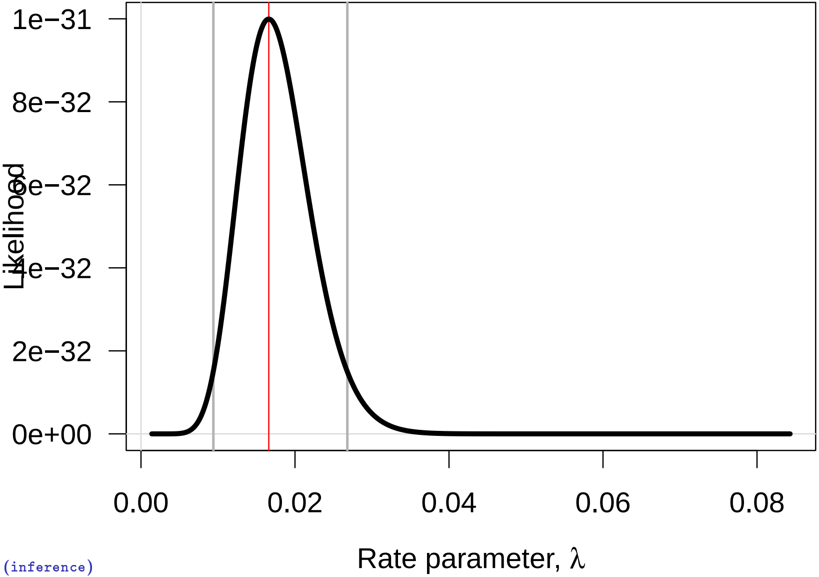
Probability of the data given the parameter:

Assuming the rate (intensity) is constant, λ, the probability of observing 14 deaths in the course of 843.6 person-years:

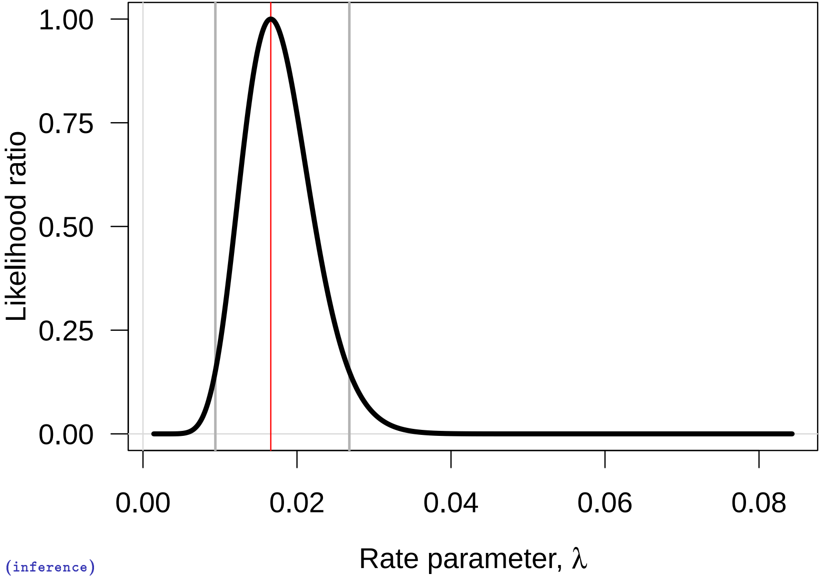
$$\begin{aligned}
\mathrm{P}\left\{D = 14, Y = 843.6|\lambda\right\} &= \lambda^D \mathrm{e}^{\lambda Y} \times K \\
&= \lambda^{14} \mathrm{e}^{\lambda \times 843.6} \times K \\
&= L(\lambda|\text{data})
\end{aligned}$$

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

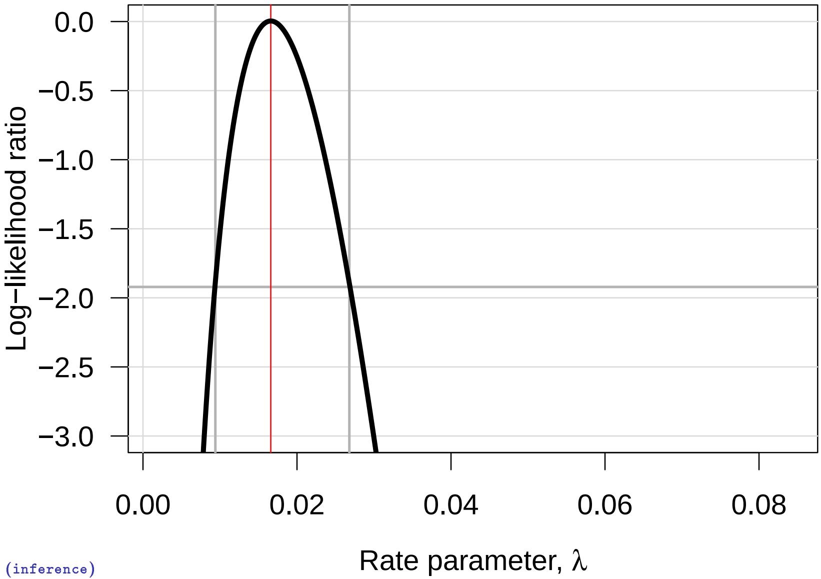
Likelihood function, 14 events, 843.6 PY



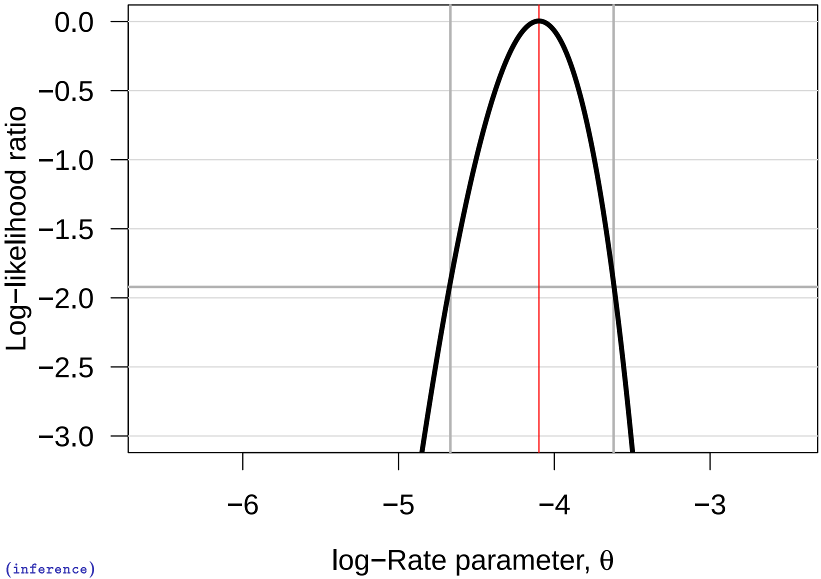
Likelihood function, 14 events, 843.6 PY



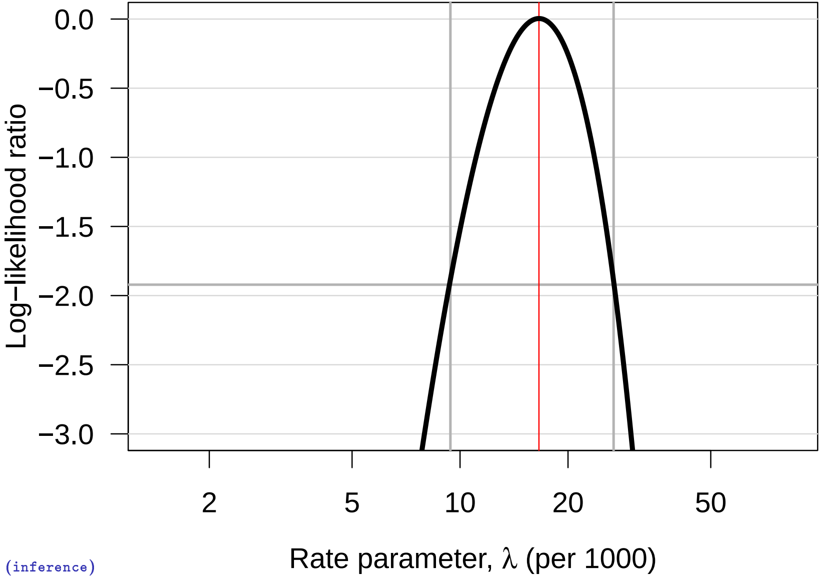
Log-likelihood function 14 events, 843.6 PY



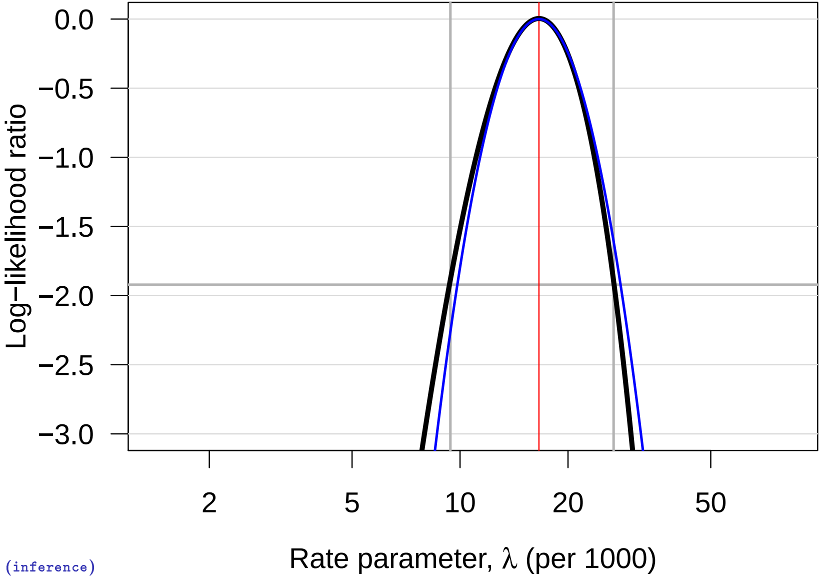
Log-likelihood function 14 events, 843.6 PY



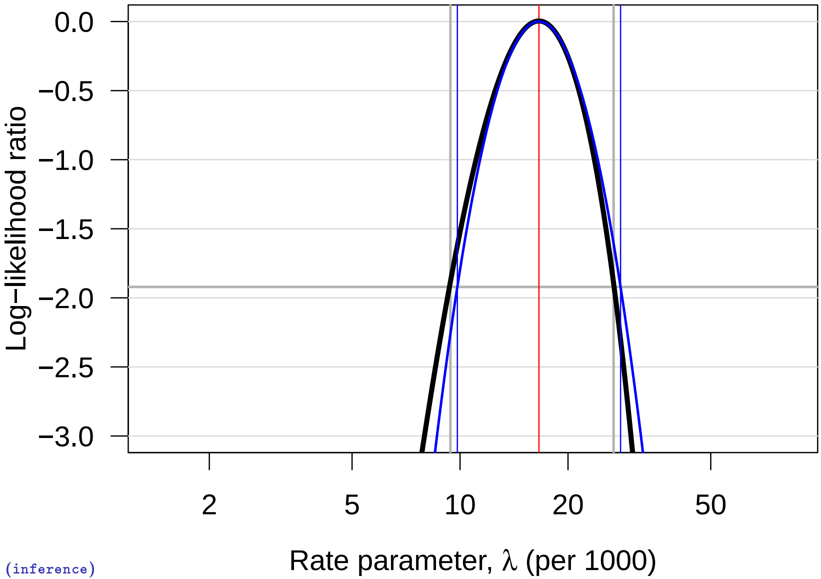
Log-likelihood function 14 events, 843.6 PY



Log-likelihood function 14 events, 843.6 PY



Log-likelihood function 14 events, 843.6 PY



Confidence interval for a rate

- Based on the quadratic approximation to the normal density
- A 95% confidence interval for the log of a rate, θ is:

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

— the 1.96 is from the normal distribution:
± 1.96 is the middle 95% of the normal distribution.

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

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

— the probability the theoretical rate λ is in this interval is 95%.

Example for a single rate

In the example we had 14 deaths during 843.6 years of follow-up.

The rate is computed as:

$$\hat{\lambda} = D/Y = 14/843.6 = 0.0165 \text{ years}^{-1} = 16.5 \text{ per } 1000 \text{ years}$$

The confidence interval is computed as:

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

per 1000 person-years.

Comparing two rates

If we have observations of two rates λ_1 and λ_0,
based on (D_1, Y_1) and (D_0, Y_0)
—from independent samples:

- The variance of the difference of the rates is the **sum** of the variances of each of the rates
- The variance of the difference of the $\log$ of the rates is the **sum** of the variances of the $\log$ of them

. . . this can be used to construct confidence intervals for rate differences and rate ratios.

Ratio of two rates

For two rates λ_1 and λ_0, based on (D_1, Y_1) and (D_0, Y_0); the $\log$ of the ratio (RR) is the difference of the $\log$s of each of the rates: $\log(\mathrm{RR}) = \log(\lambda_1) - \log(\lambda_0)$, and so:

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

As before a 95% c.i. for the RR is then, using the normal distribution:

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

Difference of two rates

For two rates λ_1 and λ_0, based on (D_1, Y_1) and (D_0, Y_0); the variance of the difference of the rates, $\mathrm{RD} = \lambda_1 - \lambda_0$, is:

$$\begin{aligned} \mathrm{var}(\mathrm{RD}) &= \mathrm{var}(\lambda_1 - \lambda_0) \\ &= \mathrm{var}(\lambda_1) + \mathrm{var}(\lambda_0) \\ &= D_1/Y_1^2 + D_0/Y_0^2 \end{aligned}$$

As before a 95% c.i. for the RD is then, using the normal distribution:

$$\mathrm{RD} \pm 1.96 \underbrace{\sqrt{\frac{D_1}{Y_1^2} + \frac{D_0}{Y_0^2}}}_{\text{standard error}}$$

Example: $(14, 843.6\text{py})$ and $(28, 632.3\text{py})$

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

The rate-ratio is computed as:

$$\begin{aligned}\mathrm{RR} &= \hat{\lambda}_1/\hat{\lambda}_0 = (D_1/Y_1)/(D_0/Y_0) \\ &= (28/632.3)/(14/843.6) = 0.0443/0.0165 = 2.669\end{aligned}$$

The 95% confidence interval is computed as:

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

Example: $(14, 843.6\textbf{py})$ and $(28, 632.3\textbf{py})$

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

The rate-difference is computed as:

$$\begin{aligned}\mathrm{RR} &= \hat{\lambda}_1 - \hat{\lambda}_0 = (D_1/Y_1) - (D_0/Y_0) = (28/632.3) - (14/843.6) \\ &= 0.0443 - 0.0165 = 0.0277 = 27.7_{\text{per 1000py}}\end{aligned}$$

The 95% confidence interval is computed as:

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

Estimating a rate using R

Poisson likelihood for one rate, based on 14 events in 843.6 PY:

```
> library( Epi )
> D <- 14 ; Y <- 843.6
> m1 <- glm(D ~ 1, offset = log(Y / 1000), family = poisson)
> ci.exp(m1)

             exp(Est.)     2.5%    97.5%
(Intercept)   16.59554 9.82875 28.02107
```

Conventional description for mortality rates:
"We used Poisson regression with log-person-years as offset..."

But really both D and Y are outcomes (random variables)

Estimating a rate using R

But really both D and Y are outcomes (random variables): use poisreg instead of poisson:

```
> mm <- glm(cbind(D, Y / 1000) ~ 1, family = poisreg)
> ci.exp( mm )
```

```
            exp(Est.)    2.5%    97.5%
(Intercept)  16.59554 9.82875 28.02107
```

...then you write:
"We used multiplicative Poisson regression for events and person-years..."

RR example using R

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

```
> D <- c(14, 28) ; Y <- c(843.6, 632.3) ; gg <- factor(0:1)
> cbind(D, Y, gg)

      D     Y gg
[1,] 14 843.6  1
[2,] 28 632.3  2

> m2 <- glm(cbind(D, Y / 1000) ~ gg, family = poisreg)
> ci.exp(m2)

            exp(Est.)     2.5%     97.5%
(Intercept) 16.595543 9.828750 28.021066
gg1          2.668354 1.404825  5.068325

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

    exp(Est.)      2.5%    97.5%
gg0  16.59554  9.82875 28.02107
gg1  44.28278 30.57545 64.13525
```

RD example using R

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

```
> a2 <- glm(cbind(D, Y / 1000) ~ gg, family = poisreg(link = "identity") )
> ci.exp(a2, Exp=FALSE)

            Estimate     2.5%    97.5%
(Intercept) 16.59554 7.902426 25.28866
gg1         27.68723 9.123703 46.25077

> a3 <- glm(cbind(D, Y / 1000) ~ gg - 1, family = poisreg(link = "identity") )
> ci.exp(a3, Exp = FALSE)

    Estimate      2.5%    97.5%
gg0 16.59554  7.902426 25.28866
gg1 44.28278 27.880508 60.68505
```

You do it (**both** RR and RD):

What is the interpretation of the parameters in `m2`, `m3`, `a2` and `a3`?

Statistical tests

- Are the observed data consistent with a given value of the parameter?
- Such a value is often a **null value**
- Typically a conservative assumption, *e.g.*:
  "no difference in outcome between the groups"
- $\mathrm{RR} = 1$ or $\mathrm{RD} = 0$
- This is called a **null hypothesis**, H_0

Computing a statistical test

$$Z_{\text{obs}} = \frac{\hat{\text{RR}} - 1}{\text{s.e.}(\text{RR})} \approx \mathcal{N}(0,1), \qquad \text{or}$$

$$Z_{\text{obs}} = \frac{\log(\hat{\text{RR}}) - 0}{\text{s.e.}\big(\log(\text{RR})\big)} \approx \mathcal{N}(0,1), \qquad \text{or}$$

$$Z_{\text{obs}} = \frac{\hat{\text{RD}} - 0}{\text{s.e.}(\text{RD})} \approx \mathcal{N}(0,1), \qquad \text{or} \ \ldots$$

- How far are are we from the null in terms of the precision?
- **How far** is quantified by the P-value:
  $P = \mathrm{P}\left\{Z \text{ is more extreme than } Z_{\text{obs}} | H_0 \text{ is true}\right\}$

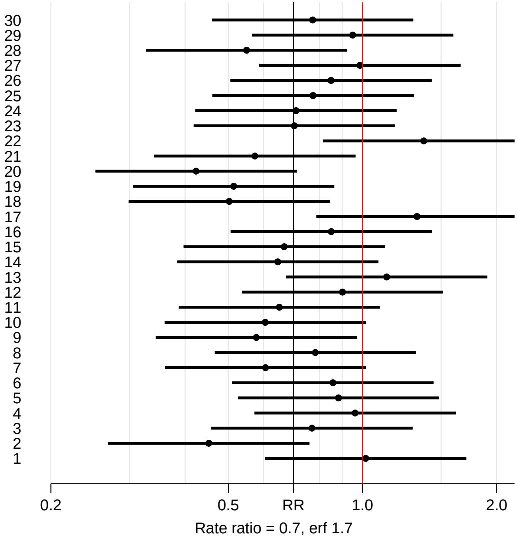
Interpretation of P-values

- Note **it is not** “the probability that H_0 is true” !
- 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 of 5% — **nonsense!**
- ...remember that the 5% is an arbitrary number taken out of thin air.

Confidence interval (CI)

- Range of parameter values compatible with the observed data — null values that will give a P-value larger than 5% $(1 -$ confidence level)
- Specified at certain **confidence level**, commonly 95% (also 90% and 99% used)
- The probability that the random interval covers the true parameter value equals the confidence level (*e.g.* 95%).
- The probability that the parameter value is in the interval is confidence level (*e.g.* 95%).

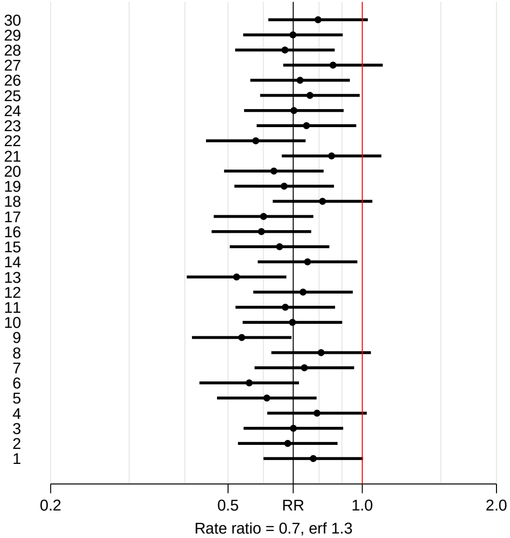
Long-term behaviour of CI



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

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

Long-term behaviour of CI



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

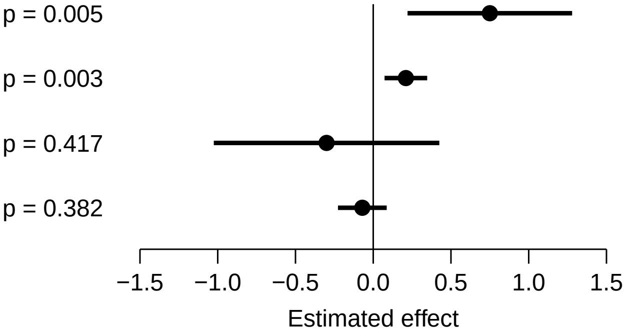
In the long run 95% of these intervals would cover the true value but 5% would not.

Interpretation of CI

- Confidence intervals gives **quantitative** information on the parameter and on statistical uncertainty about its value
- narrow CI about H_0 value → results supports H_0
- narrow CI about non-H_0 value → results supports an alternative
- wide CI about H_0 value → results inconclusive
- wide CI about non-H_0 value → results inconclusive
- **width** of the interval determines the precision
- **location** of the interval determines relevance

Confidence interval and P-value

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



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

Recommendations

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

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

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

Recommendations

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

Analysis

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,August 2024 / Januay 2025

Crude analysis

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

Single incidence rate

- **Data**: Events and risk time (D, Y)
- **Model**: Events occur with constant rate λ.
- **Parameter** of interest:

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

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

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

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

Example using R

Poisson likelihood for one rate, based on 14 events in 843.6 PY:

```
> library( Epi )
> D <- 14 ; Y <- 843.6
> m1 <- glm(D ~ 1, offset = log(Y / 1000), family = poisson)
> ci.exp( m1 )
```

```
            exp(Est.)    2.5%    97.5%
(Intercept)  16.59554 9.82875 28.02107
```

But really both D and Y are outcomes (random variables)

```
> mm <- glm(cbind(D, Y / 1000) ~ 1, family = poisreg)
> ci.exp( mm )
```

```
            exp(Est.)    2.5%    97.5%
(Intercept)  16.59554 9.82875 28.02107
```

Rate ratio in cohort study

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

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

Parameter of interest is ratio of theoretical rates:

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

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

Rate difference in cohort study

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

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

Parameter of interest is difference between theoretical rates:

$$\delta = \lambda_1 - \lambda_0 = \text{rate among exposed} - \text{rate among unexposed}$$

Null hypothesis $H_0 : \delta = 0$: exposure has no effect.

RR example using R

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

```
> D <- c(14,28) ; Y <- c(843.6,632.3) ; gg <- factor(0:1)
> cbind(D, Y, gg)

      D     Y gg
[1,] 14 843.6  1
[2,] 28 632.3  2

> m2 <- glm(cbind(D, Y / 1000) ~ gg, family = poisreg )
> ci.exp(m2)

             exp(Est.)     2.5%     97.5%
(Intercept) 16.595543 9.828750 28.021066
gg1          2.668354 1.404825  5.068325

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

    exp(Est.)     2.5%    97.5%
gg0  16.59554  9.82875 28.02107
gg1  44.28278 30.57545 64.13525
```

RD example using R

Poisson likelihood, one rate and RD **or** two rates:

```
> a2 <- glm(cbind(D, Y / 1000) ~ gg, family = poisreg(link = 'identity') )
> ci.exp(m2, Exp = FALSE )
```

```
             Estimate      2.5%    97.5%
(Intercept) 2.8091342 2.2853118 3.332957
gg1         0.9814617 0.3399129 1.623010
```

```
> a3 <- glm(cbind(D, Y / 1000) ~ gg - 1, family=poisreg(link = 'identity') )
> ci.exp(m3, Exp = FALSE )
```

```
    Estimate     2.5%    97.5%
gg0 2.809134 2.285312 3.332957
gg1 3.790596 3.420197 4.160994
```

You do it (**both** RR and RD):
What is the interpretation of the parameters?

Analysis of proportions

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

Analysis of proportions

Theoretical proportion: probability, π, that a random person becomes a case in a given period.

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

Analysis of proportions

Theoretical **prevalence**: probability, p, that a randomly chosen person in the population is a case (at a given time).

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

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

Analysis of proportions

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

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

- Depends also inversely on $\sqrt{x}$
- ...but not a good approximation to the distribution of $\hat{p} = x/n$

Analysis of proportions

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

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

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

Analysis of proportions

- Use confidence limits based on symmetric (normal) $\log(\text{OR})$:
- Compute error factor: $\text{EF} = \exp\big(1.96/\sqrt{np(1-p)}\big)$
- then use EF to compute confidence interval:

$$p/\big(p + (1-p) \overset{\times}{\div} \text{EF}\big)$$

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

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

Analysis of proportions by `glm`

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

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

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

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

> odds / (odds + 1)

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

Analysis of proportions by `glm`

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

```
> x <- 4 ; n <- 25
> pl <- glm(cbind(x, n - x) ~ 1, family = binomial(link = "log") )
> ci.exp(pl)
            exp(Est.)       2.5%     97.5%
(Intercept)      0.16 0.06517056 0.3928154
> odds / (odds + 1) # (from last slide)
            exp(Est.)       2.5%     97.5%
(Intercept)      0.16 0.06137145 0.3568687
```

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

Rate ratio in case-control study

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

Case-control design:

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

Rate ratio in case-control study

Summarized data on outcome:

Exposure	Cases	Controls
yes	D_1	C_1
no	D_0	C_0

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

Rate ratio in case-control study

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

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

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

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

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

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

Rate ratio from case-control study

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

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

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

Example: mobile phone use and brain cancer

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

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

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

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

Example: mobile phone use and brain cancer

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

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

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

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

OR from binomial model

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

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

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

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

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

Extensions and remarks

- This extends to crude analyses of exposure variables with several categories when each exposure category is separately compared to a reference group
- Evaluation of possible monotone trend in the parameter over increasing levels of exposure: estimation of regression slope
- Crude analysis is insufficient in observational studies:
- control of confounding needed

Short recap

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,August 2024 / Januay 2025

Rates

- dimension time$^-1$
- estimated as $\hat{\lambda} = D/Y$
- confidence interval for λ:
  - multiplicative $\lambda \overset{\times}{\div} \mathrm{erf}$
  - additive $\lambda \pm \mathrm{EM}$

Practical model for rates

```
> library( Epi )
> D <- 14 ; Y <- 843.6/1000 ; D/Y
```

```
[1] 16.59554
```

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

```
            exp(Est.)    2.5%    97.5%
(Intercept)  16.59554 9.82875 28.02107
```

Better way:

```
> mm <- glm( cbind(D,Y) ~ 1, family=poisreg )
> ci.exp( mm )
```

```
            exp(Est.)    2.5%    97.5%
(Intercept)  16.59554 9.82875 28.02107
```

Allows error factor and margin too:

```
> mm <- glm( cbind(D,Y) ~ 1, family=poisreg )
> ci.exp( mm )
            exp(Est.)    2.5%    97.5%
(Intercept)  16.59554 9.82875 28.02107
```

With error margin (conf.int. on rate-scale)

```
> ma <- glm( cbind(D,Y) ~ 1, family=poisreg(link="identity") )
> ci.exp( ma, Exp=FALSE )
            Estimate     2.5%    97.5%
(Intercept) 16.59554 7.902426 25.28866
```

Rate ratio and rate difference

```
> D <- c(14,28) ; Y <- c(843.6,632.3)/1000 ; gg <- factor(0:1)
> mr <- glm( cbind(D,Y) ~ gg, family=poisreg )
> ci.exp( mr )
```

```
            exp(Est.)     2.5%     97.5%
(Intercept) 16.595543 9.828750 28.021066
gg1          2.668354 1.404825  5.068325
```

```
> mR <- glm( cbind(D,Y) ~ gg-1, family=poisreg )
> ci.exp( mR )
```

```
    exp(Est.)     2.5%    97.5%
gg0  16.59554  9.82875 28.02107
gg1  44.28278 30.57545 64.13525
```

Rate ratio and rate difference

```
> ma <- glm( cbind(D,Y) ~ gg, family=poisreg(link="identity") )
> ci.exp( ma, Exp=FALSE )
```

```
            Estimate     2.5%    97.5%
(Intercept) 16.59554 7.902426 25.28866
gg1         27.68723 9.123703 46.25077
```

```
> mA <- glm( cbind(D,Y) ~ gg-1, family=poisreg(link="identity") )
> ci.exp( mA, Exp=FALSE )
```

```
    Estimate      2.5%    97.5%
gg0 16.59554  7.902426 25.28866
gg1 44.28278 27.880508 60.68505
```

Models

- Probability model: Data generator, model to data
- Statistical analysis: From data to model (parameters)
- Maximum likelihood is the basis for parameter estimation
- But only for given model
- Normal approximation provides confidence intervals
- — either for log-rates, rates, RR, RD, OR
- Beware of P-values

Stratified analysis

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,August 2024 / Januay 2025

Stratified analysis

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

Shortcomings of crude analysis

- the rate ratio for the risk factor of interest is not constant, but varies by other determinants of the disease

⇐ heterogeneity of the comparative parameter or **effect modification**

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

⇒ bias in comparison or **confounding**

⇐ exposure varies across other determinants

Models for outcome with effects of

- primary variable (“exposure”)
- secondary variable (“stratum”)
- **effect modification** is the interaction model
  exposure × stratum
  exposure with **different** effects across strata
- **confounding** is the main-effects model
  exposure + stratum exposure with **same** effect across strata

Handling for effect modification and confounding

- **Stratification** of data
  by potentially modifying and/or confounding factor(s)
  & use of **adjusted** estimators
- Conceptually simpler,
  and technically less demanding approach is
  regression modeling
- Regression modeling is feasible because we have computers
- …adjustment estimators are left-overs from teachers taught before the advent of computers (*e.g.* BxC & EL…)

Effect modification

Incidence rates (per 10^5 PY) 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)

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

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

- Rate ratio modified by age
- Rate difference not modified.
- There is no such thing as interaction (effect modification) without reference to the **scale** of the effect (*e.g.* additive or multiplicative)

Handling effect modification

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

Actual example

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

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

CHD and smoking

Both comparative parameters appear heterogeneous:

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

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

Evaluation of modification

- Modification or its absence is an inherent property of the phenomenon:
- cannot be removed or "adjusted" for
- — it depends on the **scale** on which it is measured
- Before looking for effect-modification:
  - what **scale** are we using for description of effects
  - how will we **report** the modified effects (the interaction)
  - . . . do not test for an interaction you have not seen:
    that would be returning to the world of P-values

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.

CHD and smoking example with R I

```
> library( Epi )
> R <- 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 / R # risk time in units of 10^4 PY
> smk <- factor(rep(1:2, each = 5), labels = c("Smoke", "non-Sm") )
> age <- factor(rep(seq(35, 75, 10), 2))
> data.frame(D, Y, age, smk)
```

```
     D         Y age    smk
1   32 5.2459016  35  Smoke
2  104 4.3333333  45  Smoke
3  206 2.8611111  55  Smoke
4  186 1.2653061  65  Smoke
5  102 0.5312500  75  Smoke
6    2 1.8181818  35 non-Sm
7   12 1.0909091  45 non-Sm
8   28 0.5714286  55 non-Sm
9   28 0.2592593  65 non-Sm
10  31 0.1462264  75 non-Sm
```

CHD and smoking example with R II

```
> ma <- glm(cbind(D, Y) ~ age + smk, family = poisreg)
> mi <- update(ma, . ~ . + age:smk) # add the multiplicative 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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> #
> aa <- glm(cbind(D, Y) ~ age + smk, family = poisreg(link = identity))
> ai <- update(ma, . ~ . + age:smk ) # add the additive interaction
> anova(aa, ai, test = "Chisq")
```

CHD and smoking example with R III

```
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
```

Deviance?

- . . . is the likelihood-ratio test of a given model **versus** the model with one parameter per record in the data
- In the case of CHD and smoking, stratified by age, the model with one parameter per record is the interaction model so this has 0 deviance
- in general, the deviance *per se* is not meaningful
- . . . but for models fitted to the **same** dataset, the `difference` in deviances between the models is the likelihood ratio test comparing the two models.
- That is what we computed using `anova()`.

Interaction - CHD, age and smoking—your turn!

1. enter data and repeat the analyses as on the slide
2. what is the multiplicative (main) effect of smoking
3. what is the additive (main) effect of smoking
4. use the model “ age / smk” — what does it do?
5. what is the multiplicative (interaction) effect of smoking
6. what is the additive (interaction) effect of smoking
7. try to use `plotEst` to visualize the interactions

CHD and smoking example with R I

```
> library( Epi )
> R <- 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 / R # risk time in units of 10^4 PY
> smk <- factor(rep(1:2, each = 5), labels = c("Smoke", "no-Sm") )
> age <- factor(rep(seq(35, 75, 10), 2))
> data.frame(D, Y, R, D/Y, age, smk)
     D         Y     R   D.Y age   smk
1   32 5.2459016   6.1   6.1  35 Smoke
2  104 4.3333333  24.0  24.0  45 Smoke
3  206 2.8611111  72.0  72.0  55 Smoke
4  186 1.2653061 147.0 147.0  65 Smoke
5  102 0.5312500 192.0 192.0  75 Smoke
6    2 1.8181818   1.1   1.1  35 no-Sm
7   12 1.0909091  11.0  11.0  45 no-Sm
8   28 0.5714286  49.0  49.0  55 no-Sm
9   28 0.2592593 108.0 108.0  65 no-Sm
10  31 0.1462264 212.0 212.0  75 no-Sm
```

CHD and smoking example with R II

```
> # model with log link
> ma <- glm(cbind(D, Y) ~ age + smk, family = poisreg)
> mi <- update(ma, . ~ . + age:smk) # add the multiplicative 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 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> # with identity link
> aa <- glm(cbind(D, Y) ~ age + smk, family = poisreg(link = identity))
> ai <- update(aa, . ~ . + age:smk ) # add the additive interaction
> anova(aa, ai, test = "Chisq")
```

CHD and smoking example with R III

```
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

> # multiplicative interaction
> mI <- update(mi, . ~ -1 + age / smk)
> ci.exp(mI, subset = "smk")

               exp(Est.)       2.5%     97.5%
age35:smkno-Sm 0.1803279 0.04321658 0.7524459
age45:smkno-Sm 0.4583333 0.25215724 0.8330891
age55:smkno-Sm 0.6805556 0.45858247 1.0099729
age65:smkno-Sm 0.7346939 0.49381111 1.0930801
age75:smkno-Sm 1.1041667 0.73868885 1.6504703

> round(1 / ci.exp(mI, subset = "smk"), 2)
```

CHD and smoking example with R IV

```
               exp(Est.)  2.5% 97.5%
age35:smkno-Sm      5.55 23.14  1.33
age45:smkno-Sm      2.18  3.97  1.20
age55:smkno-Sm      1.47  2.18  0.99
age65:smkno-Sm      1.36  2.03  0.91
age75:smkno-Sm      0.91  1.35  0.61
```

```
> # additive interaction
> aI <- update(ai, . ~ -1 + age / smk)
> ci.exp(aI, Exp = FALSE)
```

CHD and smoking example with R V

```
                  Estimate       2.5%      97.5%
age35                  6.1   3.986497   8.213503
age45                 24.0  19.387433  28.612567
age55                 72.0  62.167884  81.832116
age65                147.0 125.874405 168.125595
age75                192.0 154.739452 229.260548
age35:smkno-Sm        -5.0  -7.605951  -2.394049
age45:smkno-Sm       -13.0 -20.746643  -5.253357
age55:smkno-Sm       -23.0 -43.641599  -2.358401
age65:smkno-Sm       -39.0 -84.238620   6.238620
age75:smkno-Sm        20.0 -63.412950 103.412950
```

```
> round(- ci.exp(aI, Exp = FALSE, subset = "smk"), 2)
```

CHD and smoking example with R VI

```
                 Estimate  2.5%   97.5%
age35:smkno-Sm          5  7.61    2.39
age45:smkno-Sm         13 20.75    5.25
age55:smkno-Sm         23 43.64    2.36
age65:smkno-Sm         39 84.24   -6.24
age75:smkno-Sm        -20 63.41 -103.41
```

```
> # forest plots
> par(mfrow = c(1,2))
> plotEst(1 / ci.exp(mI, subset = "smk"),
+         xlog = TRUE, xlab = "smoking RR")
> abline(v = c(1, exp(-coef(ma)[6])))
> plotEst(-ci.exp(aI, Exp = FALSE, subset = "smk"),
+         xlab = "smoking RD")
> abline(v = c(0, -coef(aa)[6]))
> # forest plots again, niceified
> par(mfrow = c(1,2))
> plotEst(1 / ci.exp(mI, subset = "smk"),
+         xlog = TRUE, xlab = "smoking RR", xlim = c(0.5, 20))
```

CHD and smoking example with R VII

```
> abline(v = c(1, exp(-coef(ma)[6])), col = 1:2)
> plotEst(-ci.exp(aI, Exp = FALSE, subset = "smk"),
+         xlab = "smoking RD", xlim = c(-50, 100))
> abline(v = c(0, -coef(aa)[6]), col = 1:2)
```

Confounding - operation example

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

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

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

PN appears more successful than OS?

Operation example

Results stratified by initial diameter size of the stone:

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

OS seems more successful in both subgroups.

Is there a paradox here?

Operation example

- Treatment groups are not comparable w.r.t. initial size.
- Size of the stone (SS) is a **confounder** of the association between operation type and success:

1. 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.

Operation example

- Instance of “confounding by indication”:
  — patient status affects choice of treatment,
  ⇒ bias in comparing treatments.
- This bias is best avoided in planning:
  — randomized allocation of treatment.

Grey hair and cancer incidence

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

Observed crude association nearly vanishes after controlling for age.

Means for control of confounding

Design:

- Randomization
- Restriction
- Matching

Means for control of confounding (cont'd)

Analysis:

- Stratification
- Regression modeling

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

Other methods provide partial removal, but only due to **measured** factors
residual confounding may remain.

Steps of stratified analysis

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

Steps of stratified analysis (cont.)

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

Estimation of rate ratio

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

Adjusted summary estimators

Different weighting methods:

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

Preferred method in analysis: ML Useful method in simple descriptive: CMF / SMR

Gray hair & cancer

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

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

Gray hair & cancer

Crude and adjusted risk estimate by Poisson model:

```
> library(Epi)
> ci.exp(glm(cbind(D, Y) ~ hair      , family = poisreg))
            exp(Est.)      2.5%    97.5%
(Intercept)       1.2 0.8390232 1.716281
hairGray          2.2 1.4288756 3.387279
> ci.exp(glm(cbind(D, Y) ~ hair + age, family = poisreg))
            exp(Est.)       2.5%     97.5%
(Intercept) 3.7782269 2.49962653 5.7108526
hairGray    1.0606186 0.67013527 1.6786339
ageYoung    0.1470116 0.08418635 0.2567211
```

Case-control study of Alcohol and oesophageal cancer

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

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

Crude analysis of CC-data

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

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

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

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

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

Stratification by age

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

NB! Selection of controls: inefficient study
Should have employed stratified sampling by age.

Stratified analysis

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

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

Stratified analysis

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

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

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

Stratified analysis

. . . only the relevant parameters:

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

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

```
> round(pmin(ci.exp(mi, subset = "alc"), 50), 2)

             exp(Est.) 2.5% 97.5%
age25:alc>80     50.00 0.00 50.00
age35:alc>80      5.05 1.27 20.02
age45:alc>80      5.67 2.80 11.46
age55:alc>80      6.36 3.45 11.73
age65:alc>80      2.58 1.22  5.47
age75:alc>80     50.00 0.00 50.00
```

Oesophageal cancer CC — effect modification?

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

Model 1: cbind(ca, co) ~ age + age:alc
Model 2: cbind(ca, co) ~ age + alc
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         0      0.000
2         5     11.041 -5  -11.041  0.05057 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Oesophageal cancer CC — linear effect modification

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

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

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

```
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         0      0.000
2         4     10.609 -4 -10.6093  0.03132 *
3         5     11.041 -1  -0.4319  0.51107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Evidence against linear interaction (OR decreasing by age)

Oesophageal cancer CC — effect modification?

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

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

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

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

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

Regression models

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,August 2024 / Januay 2025

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

Regression modeling

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

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
  - (non)Linear trends not easily estimated

Limitations

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

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

Key concept: **statistical model**

Log-linear model for rates

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

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

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

Model parameters

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

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

Lung cancer incidence, asbestos exposure and smoking

Dichotomous explanatory variables coded:

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

Log-linear model for theoretical rates

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

Log-linear model: Variables

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

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

Lung cancer, asbestos and smoking

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

```
> D <- c(150, 15, 120,  12)        # cases
> Y <- c( 25, 25, 100, 100) / 100 # PY (100,000s)
> 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
```

Lung cancer, asbestos and smoking

- Regression modeling
- Multiplicative (default) Poisson model
- 2 equivalent approaches
  - `D` response, `log(Y)` offset (mostly used in the literature)
  - `cbind(D,Y)` response, `family=poisreg`
  - ...the latter approach also useful for **additive** models

```
> 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))
```

Lung cancer, asbestos and smoking

Summary and extraction of parameters:

```
> summary(mo)

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

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 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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
```

Summary and extraction of parameters I

```
> ci.exp(mo)
            exp(Est.)     2.5%     97.5%
(Intercept)        12 8.059539 17.867026
asb                 5 3.977142  6.285921
smk                10 6.732721 14.852836
> ci.exp(mo, Exp = FALSE)
            Estimate     2.5%    97.5%
(Intercept) 2.484907 2.086856 2.882957
asb         1.609438 1.380563 1.838312
smk         2.302585 1.906979 2.698191
> ci.exp(mm, Exp = FALSE)
            Estimate     2.5%    97.5%
(Intercept) 2.484907 2.086856 2.882957
asb         1.609438 1.380563 1.838312
smk         2.302585 1.906979 2.698191
```

Summary and extraction of parameters II

Parameters are the same for the two modeling approaches.

Interpretation of parameters I

```
> round( cbind( ci.exp( mm, Exp=F ),
+               ci.exp( mm          ) ), 3 )
            Estimate  2.5% 97.5% exp(Est.)  2.5%  97.5%
(Intercept)    2.485 2.087 2.883        12 8.060 17.867
asb            1.609 1.381 1.838         5 3.977  6.286
smk            2.303 1.907 2.698        10 6.733 14.853
```

α $= 2.485 = \log(12)$, log of baseline rate,

β $= 1.609 = \log(5)$, log of rate ratio $\rho = 5$ between exposed and unexposed for asbestos

γ $= 2.303 = \log(10)$, log of rate ratio $\tau = 10$ between smokers and non-smokers.

Interpretation of parameters II

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

Log-linear model: Estimated rates

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

Log-linear model

Model with effect modification (two regressors only)

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

equivalently

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

where α is as before, but

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

Interaction parameter

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 risks associated with X between the two categories of Z.

Log-linear model: Estimated rates

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

Lung cancer, asbestos and smoking

```
> mi <- glm(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 is 1,
- asbestos and smoking effects are the unchanged,
- but SEs are larger because they refer to RRs for levels $X = 0$ and $Z = 0$ respectively and not both levels **jointly**

Additive model for rates

General form with two regressors

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

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

δ = interaction parameter.

- For binary X, Z:

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

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

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

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

Example: Additive model

```
> mai <- glm( cbind(D,Y) ~ asb + smk + asb*smk, family=poisreg(link=identity) )
> round( ci.exp( mai, Exp=FALSE, pval=TRUE ), 4 )
            Estimate     2.5%    97.5%      P
(Intercept)       12   5.2105  18.7895 0.0005
asb               48  16.8865  79.1135 0.0025
smk              108  85.4817 130.5183 0.0000
asb:smk          432 328.8083 535.1917 0.0000
```

A very clear interaction (effect modification)

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

α = 12, baseline rate, i.e. that among non-smokers unexposed to asbestos (reference group),

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

γ = $108\ (= 120 - 12)$, rate difference between smokers and non-smokers among only those unexposed to asbestos

δ = excess of rate difference between smokers and non-smokers among those exposed to asbestos:
$\delta = (600 - 120) - (60 - 12) = 432$

Model fitting

Output from computer packages will give:

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

May be difficult to interpret!

Model checking & diagnostics:

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

Problems in modeling

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

Modeling

- Modeling should not substitute, but complement crude analyses:
- Crude analyses can 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, not data-driven
- Adequate training and experience required.
- Ask help from a professional statistician!
- **Collaboration** is the keyword.

Conclusion

Bendix Carstensen & Esa Läärä

Nordic Summerschool of Cancer Epidemiology
Danish Cancer Society,August 2024 / January 2025

Concluding remarks

Epidemiologic study is a

Measurement excercise

Target is a **parameter**(s) of interest, like

- incidence rate
- rate ratio
- rate difference
- 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:

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

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

Recommendations for analysis and reporting

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

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!