# Nordic Summerschool of Cancer Epidemiology

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

Danish Cancer Society

August 2015

http://BendixCarstensen.com/NSCE/2015



# Chance

Nordic Summerschool of Cancer Epidemiology
August 2015
Danish Cancer Society
http://BendixCarstensen.com/NSCE/2015

chance

## Chance variation

- Systematic and random variation
- Probability model:
  random variable, distribution, parameters
- Poisson and Gaussian models
- Statistic, sampling distribution and
  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 & random (cont'd)

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 3: 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 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 | 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?



## Probability models

**Simple model for cancer incidence:** In homogenous population we assume:

- the unknown theoretical incidence rate
- *hazard* or *intensity* — $\lambda$
- of contracting cancer
- is **constant** over a short period of time.

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

## Simple model (cont'd)

- Observations:
  - Number of cases $D$ in
  - $Y$ person-years at risk
  - and empirical incidence rate $R = D/Y$
- are **random variables** with unpredictable values in a given observation period.
- The **probability distribution** of possible values of a random variable has some known mathematical form.
- Key properties of the distribution are determined by quantities called **parameters**;
- — in this case the theoretical rate $\lambda$.



## Probability model

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

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

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



## Statistics

The opposite of a probability models:

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

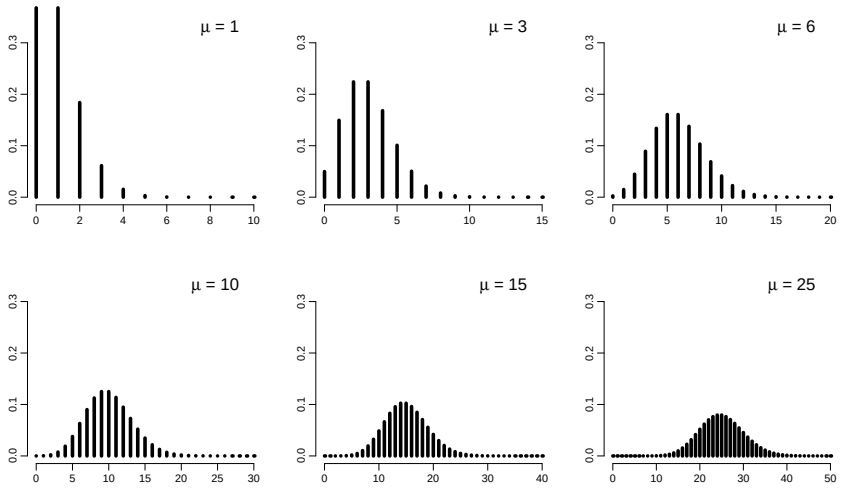
Thus **statistical modelling** is how to estimate parameters from observed data.

## 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 $\mu$ of $D$ is large enough, the Poisson distribution resembles more and more the **Gaussian** or **Normal** distribution.



## Poisson distribution with different means $\mu$:





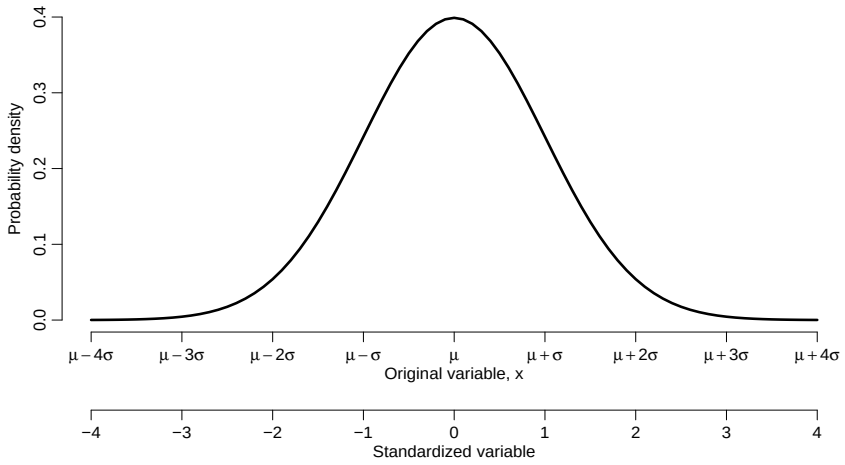

## Gaussian distribution

Gaussian or Normal distribution:

- common model for continuous variables,
  - symmetric and bell-shaped,
  - has two parameters:
    - $\mu$ = expectation or mean,
    - $\sigma$ = standard deviation.
- Most important use of Gaussian model:
- Approximation of **sampling distribution** of empirical measures:
- observed incidence rates
- $\log$(observed incidence rates)

# Gaussian distribution (cont'd)

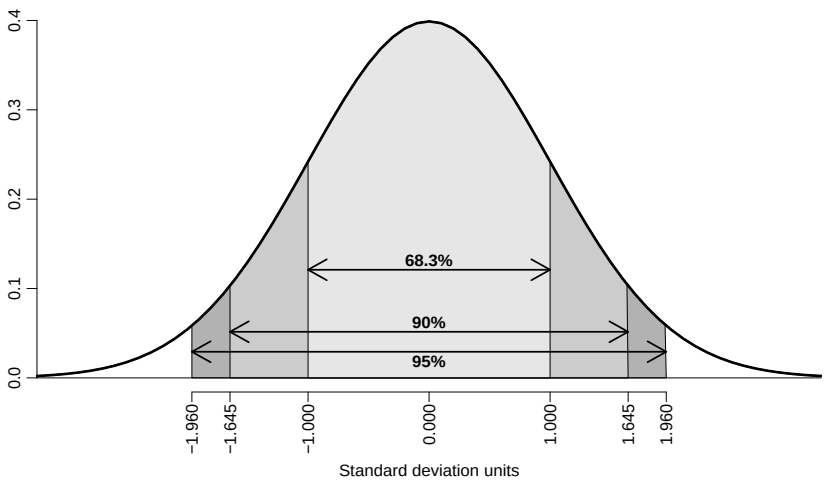
Probability density funtion – the "Bell Curve".






# Gaussian distribution (cont'd)

Areas under curve limited by selected quantiles






# 2.4 Statistic, sampling distribution and standard error

A "**statistic**" is a summary measure calculated from empirical data (a "formula").

- $X$ — a variable having certain distribution in population with mean $\mu$ and standard deviation $\sigma$.
- Take a random sample of $n$ subjects.
- Values of $X$ in the sample: $X_1, X_2, \ldots, X_n$.
- Before sampling these are random variables.

## Sample statistics:

- Sample mean (arithmetic):

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Sample standard deviation:

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



## Sample statistics:

- One-sample $t$-statistic:

$$t = \frac{\bar{X} - \mu_0}{\mathrm{SD}/\sqrt{n}}$$

($\mu_0$ is the hypothesized value of $\mu$).
How far from $\mu_0$ is the observed $X$?



## Sampling distribution

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

## Example 5

- Sampling distribution of the sample mean $\bar{X}$ of variable $X$ with mean $\mu$ and standard deviation $\sigma$ is approximately Gaussian with:
  - expectation $\mu$,
  - standard deviation $\sigma/\sqrt{n}$,
- with sufficiently big sample size, whatever the original distribution of $X$.
- This is the **Central Limit Theorem** (CLT) from probability theory.



## Standard error (s.e.)

Estimated standard deviation of sampling distribution of statistic.

**Example 5 (cont'd):**

- Sample $X_1, \ldots, X_n$ drawn of variable $X$ from population distribution with mean $\mu$ and standard deviation $\sigma$.
  The sample mean is $\bar{X}$ and the sample standard deviation $\mathrm{s.e.}$.
- $\Rightarrow$ *Standard error* of the mean:

$$\mathrm{s.e.}(\bar{X}) = \frac{\mathrm{s.e.}}{\sqrt{n}}$$

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



## Standard error (cont'd)

- Confidence interval (CI) for $\mu$:

$$\bar{X} \pm z \times \mathrm{s.e.}(\bar{X})$$

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

- Used in one-sample $t$-statistic:

$$t = \frac{\bar{X} - \mu_0}{\mathrm{s.e.}(\bar{X})}$$

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

## Example 6: Single incidence rate

Parameter $\lambda$
= true unknown incidence rate in population.

- **Empirical rate** $R = D/Y$,
- **estimator** of $\lambda$.
- $R$ is a statistic, random variable whose:
  - value varies from one study population ("sample") to another in hypothetical repetitions,
  - sampling distribution is (under the Poisson model & other conditions) a transformation of the Poisson distribution,



## Example 6 (cont'd)

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

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

$\Rightarrow$ The amount of random error depends inversely on (the square root of) the number of cases.

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



# Inference

Nordic Summerschool of Cancer Epidemiology
August 2015
Danish Cancer Society
`http://BendixCarstensen.com/NSCE/2015`

`inference`

# 3 STATISTICAL INFERENCE





## Inferential questions

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



## Statistical notation:

- Probability: parameters $\rightarrow$ data
- Statistics: data $\rightarrow$ parameter(estimate)s
- Notation:
  - Parameter denoted by a Greek letter
  - Estimator & estimate by the same Greek letter with "hat".
- Ex: Incidence rate:
  - True unknown rate: $\lambda$
  - Estimator: $\widehat{\lambda} = R = D/Y$, empirical rate.
- Rate ratio:
  - True rate ratio $\rho = \lambda_1/\lambda_0$ between exposed and unexposed,
  - Estimator: $\widehat{\rho} = \text{RR} = R_1/R_0$,
    ratio between the empirical rates.

## 3.2 Statistical testing

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



## Purpose of statistical testing

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

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

- . . . but if we have evidence enough to assert is as **FALSE**



## Test statistic

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

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

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

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

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

Null hypothesis:

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

$O$ = Observed rate difference

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

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

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



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

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

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

- One-tailed $P = 0.06$, two-tailed $P = 0.12$
- What does this mean?
- How do we proceed?



## Questions about the test statistic

- How does the observed value $Z_{\text{obs}}$ locate itself in the sampling distribution of $Z$?
- How common or how rare it is to obtain $Z_{\text{obs}}$ under $H_0$?
- What is the probability of getting $Z$ larger than observed $Z_{\text{obs}}$ if $H_0$ were true?
- The latter probability is the **one-tailed observed significance level** or *$P$-value* against alternative $\rho > 1$.

## Two-tailed $P$ value

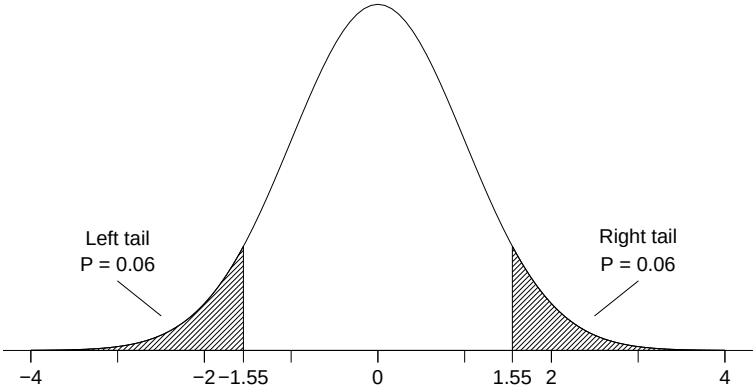
$=$ probability for test statistic $Z$ being more extreme than the absolute value of $Z_{\text{obs}}$.

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



## Distribution of test statistic

— under $H_0$ and graphical derivation of $P$-value.




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



## $P$-value

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

## 3.4 Interpretation of $P$-values

- No mechanical rules of inference
- Rough guidelines
  - "large" value ($p > 0.1$): consistent with $H_0$ but not necessarily supporting it,
  - "small" value ($p < 0.01$): indicates evidence against $H_0$
  - "intermediate" value ($p \approx 0.05$): weak evidence against $H_0$
- Division of $p$-values into "significant" or "non-significant" by cut-off $0.05$ — **To be avoided!**



## 3.5 Confidence interval (CI)

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

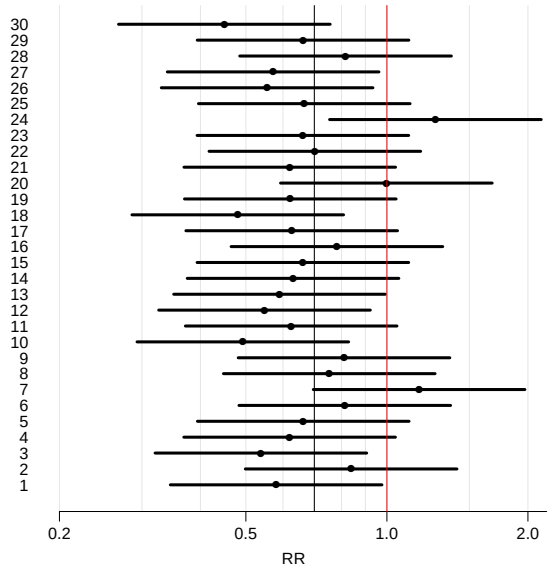


## Confidence interval (cont'd)

- The latter is the **long-term property** of the **procedure** for calculating CI under hypothetical "repeated sampling".
- Yet, the obtained CI from data at hand either covers or does not cover the parameter of interest.
- (N.B. As with $P$ values the accuracy of nominal confidence level depends on lack of bias and on validity of some statistical assumptions.)

# Long-term behaviour of CI

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






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

- Observed rate difference RD $= 30$ per $10^5$ y.
- Standard error SE(RD) $= 19.4$ per $10^5$ y.
- Limits of the 95% approximate CI (per $10^5$ y):
  - lower: $30 - 1.96 \times 19.4 = -8$,
  - upper: $30 + 1.96 \times 19.4 = 68$
- For 90% level, use $1.645$ instead of $1.960$.
  For 99% level, $2.58$ is the multiplier.



# Interpretation of obtained CI

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

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

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

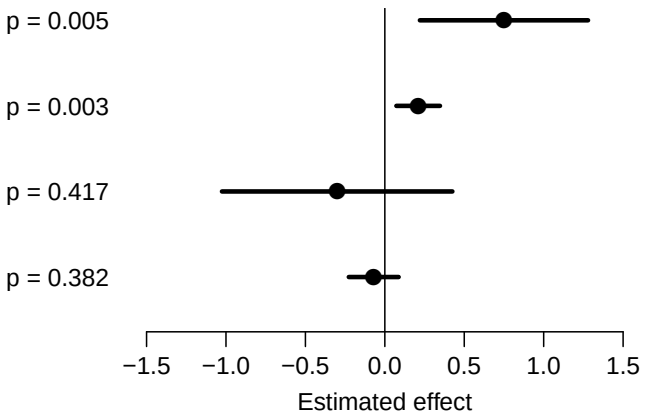
## Interpretation of CI (cont'd)

- CI gives more quantitative information on the parameter and on statistical uncertainty about its value than $P$ value.
- In particular, interpretation of "non-significant" results, *i.e.* large $P$ values:
  - narrow CI about $H_0$ value:
    $\rightarrow$ results give support to $H_0$.
  - wide CI about $H_0$ value:
    $\rightarrow$ results inconclusive.
- The latter is more commonly encountered.



## CI and $P$-value

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




Similar $P$-values but different interpretation!



## 3.6 Recommendations

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

Extracts from section *Statistics*:

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

## Recommendations (cont'd)

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

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

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



## Recommendations in BMJ (cont'd)

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

Nordic Summerschool of Cancer Epidemiology
August 2015
Danish Cancer Society
`http://BendixCarstensen.com/NSCE/2015`

`analysis`

# CRUDE ANALYSIS

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



# Single incidence rate

- **Parameter** of interest:

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

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

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

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



# Single rate (cont'd)

- Simple approximate 95% CI:

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

- using 95% **error margin**:

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

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

## Single rate (cont'd)

- Better approximation on log-scale:

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

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

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

where $\exp$ is the exponential function or antilog (inverse of the natural logarithm)

- From these items we get 95% CI for $\lambda$:

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

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



## Example 4 (cont'd)

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

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

- The 95% error margin:

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

Negative lower limit — illogical!



## Example 4 (cont'd)

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

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

- via the 95% error factor:

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

from which the confidence limits (both $> 0$):

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

## Rate estimation in Poisson model

3 male breast cancers in $90,909$ person years, corresponding to a rate of $3/0.090909 = 33$ per $10^6$ PY

```
> library( Epi )
> D <- 3 ; Y <- 90909
> m0 <- glm( D ~ 1, offset=log(Y/10^6), family=poisson )
> ci.exp( m0 )
            exp(Est.)     2.5%    97.5%
(Intercept)  33.00003 10.64322 102.3189
```

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



## 4.2 Rate ratio in cohort study

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

Parameter of interest is true rate ratio:

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

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



## Rate ratio (cont'd)

Summarized data on outcome from cohort study with person-time

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

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

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

## Rate ratio (cont'd)

Point estimator of true rate ratio, $\rho$, is the empirical rate ratio (RR):

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

N.B.: The last form is particularly useful in case-control studies — see next section.



## Rate ratio (cont'd)

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

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

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

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

Note: SE of estimate depends inversely on numbers of cases.



## Example 8: Helsinki Heart Study

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

## Example 8 (cont'd)

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

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

95 % CI for $\rho$ :

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

Two-tailed $P = 0.52$



## Rate ratio in Poisson model

```
> library( Epi )
> D <-c(31,26) ; Y <- c(10000,10000) ; E <- c(1,0)
> cbind( D, Y, E)
      D     Y E
[1,] 31 10000 1
[2,] 26 10000 0
> mr <- glm( D ~ factor(E), offset=log(Y/10^3), family=poisson )
> ci.exp( mr )
            exp(Est.)      2.5%    97.5%
(Intercept)  2.600000 1.7702679 3.818631
factor(E)1   1.192308 0.7079898 2.007935
```

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



## Rates in Poisson model

```
> cbind( D, Y, E)
      D     Y E
[1,] 31 10000 1
[2,] 26 10000 0
> mR <- glm( D ~ factor(E)-1, offset=log(Y/10^3), family=poisson )
> ci.exp( mR )
            exp(Est.)      2.5%    97.5%
factor(E)0        2.6 1.770268 3.818631
factor(E)1        3.1 2.180125 4.408004
```

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

## 4.3 Rate ratio in case-control study

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

Case-control design:

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



## Rate ratio in case-control study

Summarized data on outcome:

| Exposure | Cases | Controls |
|---|---|---|
| yes | $D_1$ | $C_1$ |
| no | $D_0$ | $C_0$ |

- Can we directly estimate the rates $\lambda_0$ and $\lambda_1$ from this?
- — and the ratio of these?
- NO and YES!
- Rates are not directly 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:

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

- $\Rightarrow$ RR estimated by the ratio of the case-control ratios $(D/C)$
- . . . but of course there is a penalty to pay. . .

## Rate ratio from case-control study

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

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

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



## Example 9

Use of mobile phone and brain cancer

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

| Daily use | Cases | Controls |
|---|---|---|
| $\geq 15$ min | 35 | 51 |
| no use | 637 | 625 |

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

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



## Example 9 (cont'd)

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

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

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",
> levels(Ex)
[1] "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
```

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



## Analysis of proportions

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



## Analysis of proportions (cont'd)

Incidence proportion:

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

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

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

## Analysis of proportions (cont'd)

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

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

Depends also inversely on $x$!



## Analysis of proportions (cont'd)

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

```
> 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(4,8:10),ci(3,11:13),ci(2,3:5),ci(1,1:2))
      x  n      p      lo     hi
 [1,] 4  8 0.5000  0.1464 0.8536
 [2,] 4  9 0.4444  0.1132 0.7757
 [3,] 4 10 0.4000  0.0902 0.7098
 [4,] 3 11 0.2727  0.0042 0.5413
 [5,] 3 12 0.2500  0.0000 0.5000
 [6,] 3 13 0.2308 -0.0029 0.4645
 [7,] 2  3 0.6667  0.1223 1.2110
 [8,] 2  4 0.5000  0.0000 1.0000
 [9,] 2  5 0.4000 -0.0382 0.8382
[10,] 1  1 1.0000  1.0000 1.0000
```



## Analysis of proportions (cont'd)

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

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

- then use to compute confidence interval:

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

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

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

## Analysis of proportions by `glm`

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

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

```
> 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
```



## Extensions and remarks

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



# Stratified analysis

Nordic Summerschool of Cancer Epidemiology
August 2015
Danish Cancer Society
`http://BendixCarstensen.com/NSCE/2015`

`strat`

## Stratified analysis

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



## Shortcomings of crude analysis

Crude analysis is misleading, if

- the rate ratio for the risk factor of interest is not constant, but varies by other determinants of the disease
- . . . *i.e.* heterogeneity of the comparative parameter or **effect modification**
- **OR**
- the exposure groups are not comparable w.r.t. other determinants of disease
- . . . *i.e.* bias in comparison or **confounding**



## Remedies

Simple approach for remedy:

- **Stratification** of data by potentially modifying and/or confounding factor(s) & use of **adjusted** estimators
- Conceptually simpler, but technically more demanding approach is **regression modelling**

## Effect modification

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

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

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



## Effect modification (cont'd)

Depends how the effect is measured:

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

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



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

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

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

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

## Effect modification (cont'd)

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



## Example:

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

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



## Example (cont'd)

Both comparative parameters appear heterogenous:

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

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

## Evaluation of modification

- Modification or its absence is an inherent property of the phenomenon:
- cannot be removed or "adjusted" for
- but it depends on the **scale** on which it is measured
- Problems:
  - Stratum-specific numbers have large random error
  - estimates of stratum specific effect parameters variable even if no true modification present,
  - or essential modification may remain undetected



## Evaluation of modification (cont'd)

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



## Confounding - example

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

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

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

PN appears more succesful than OS?

## Example (cont'd)

Results stratified by initial diameter size of the stone:

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

OS seems more succesful in both subgroups.

**Is there a paradox here?**



## Operation example

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



## Example 13 (cont'd)

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

## Grey hair and cancer incidence

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

Observed crude association nearly vanishes after controlling for age.



## Means for control of confounding

Design:

- Randomization
- Restriction
- Matching



## Means for control of confounding (cont'd)

Analysis:

- Stratification
- Regression modelling

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

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

## Steps of stratified analysis

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



## Steps of stratified analysis (cont.)

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



## Estimation of rate ratio

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

## Adjusted summary estimators

Different weighting methods:

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



## Mantel-Haenszel estimators

Cohort study, data summary in each stratum $k$:

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

Stratum-specific rates by exposure group:
$R_{1k} = D_{1k}/Y_{1k}, \quad R_{0k} = D_{0k}/Y_{0k}$
. . . weighted together, by a multiplicative (log-linear) model
$\ln(R_{1k}) = \beta + \ln(R_{0k})$
$\beta$ a common log-$\mathrm{RR}$ across strata.



## Mantel-Haenszel estimator

- Combination of stratum-specific $\mathrm{RR}$s as a proxy for a model estimate of $\beta$
- Formulae devised in times of the hand-calculator — before the advent of computers
- Replaced by statistical models
- Out of date since about mid-1990s

# Gray hair & cancer (cont'd)

```
> 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 (cont'd)

Crude and adjusted risk estimate by Poisson model:

```
> library( Epi )
> ci.exp( glm( D ~ hair      , offset=log(Y), family=poisson ) )
            exp(Est.)      2.5%    97.5%
(Intercept)       1.2 0.8390238 1.716280
hairGray          2.2 1.4288764 3.387277
> ci.exp( glm( D ~ hair + age, offset=log(Y), family=poisson ) )
            exp(Est.)       2.5%     97.5%
(Intercept) 3.7782269 2.49962654 5.7108526
hairGray    1.0606186 0.67013527 1.6786339
ageYoung    0.1470116 0.08418635 0.2567211
```



# Case-control study of Alcohol and oesophageal cancer

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

| Exposure $\geq 80$ 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 $\geq 80$ g/d | Cases | Controls | EOR |
|---|---|---|---|---|
| 25-34 | yes | 1 | 9 | $\infty$ |
| | 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 | $\infty$ |
| | no | 8 | 31 | |

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



# Stratified analysis

```
> ca <- c( 1, 0, 4, 5, 2, 2, 4, 3, 1, 3, 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   2  29 >80  45
6   2 138 <80  45
7   4  27 >80  55
8   3 139 <80  55
9   1  18 >80  65
10  3  88 <80  65
11  5   0 >80  75
12  8  31 <80  75
```

## Stratified analysis

The "age:" operator produces a separate alcohol-OR for each age class:

```
> 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        1.114880e+10 0.000    Inf
age55        1.660289e+10 0.000    Inf
age65        2.622503e+10 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 4.759000e+00 0.644 35.177
age55:alc>80 6.864000e+00 1.453 32.427
age65:alc>80 1.630000e+00 0.160 16.570
age75:alc>80 1.755246e+11 0.000    Inf
```

The age-specific ORs are **very** variable...



## 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     0.0000
2         5     8.0552 -5  -8.0552   0.1532
> round( ci.exp( ma, subset="alc" ), 3 )
       exp(Est.)  2.5%  97.5%
alc>80     6.801 3.181 14.539
```

- Stratum-specific ORs somewhat variable.
- Random error in some of them apparently large
- Only weak evidence against homogeneity, so assumption of a common rate ratio seems plausible.



## Oesophageal cancer CC — confounding?

- Is exposure associated with age in the study population?
- Look at variation in the age-specific prevalences of exposure among controls.
- Adjustment for age is generally reasonable.
- There is substatial age-confounding of the `alc`-OR:

```
> ma <- glm( cbind(ca,co) ~ age + alc, family=binomial )
> m0 <- glm( cbind(ca,co) ~       alc, family=binomial )
> round( ci.exp( ma, subset="alc" ), 3 )
       exp(Est.)  2.5%  97.5%
alc>80     6.801 3.181 14.539
> round( ci.exp( m0, subset="alc" ), 3 )
       exp(Est.)  2.5% 97.5%
alc>80     4.946 2.529 9.673
```

**Note:** there exist no **test** for confounding — only test for effect modification (interaction).

# Regression models

Nordic Summerschool of Cancer Epidemiology
August 2015
Danish Cancer Society
http://BendixCarstensen.com/NSCE/2015

regress

## Regression modeling

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



## Limitations of stratified analysis

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

## Limitations (cont'd)

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

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

Key concept: **statistical model**



## Log-linear model for rates

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

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

Equivalent expression, **multiplicative model**:

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



## Log-linear model (cont'd)

Model parameters

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

$\beta = \ln(\rho) =$ slope,
change in $\ln(\lambda)$ for unit change in $X$,
**adjusting for** the effect of $Z$ (& $U, V, \dots$)

$e^\beta = \rho =$ rate ratio for unit change in $X$.

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

## Lung cancer incidence, asbestos exposure and smoking

Dichotomous explanatory variables coded:

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

Log-linear model for theoretical rates

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



## Log-linear model: Variables

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



## Lung cancer, asbestos and smoking

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

```
> D <- c( 150, 15, 120,  12 )       # cases
> Y <- c(  25, 25, 100, 100 ) / 100 # PY (100,000s)
> A <- c( 1, 1, 0, 0 ) # Asbestos exposure
> S <- c( 1, 0, 1, 0 ) # Smoking
> cbind( D, Y, A, S )
       D    Y A S
[1,] 150 0.25 1 1
[2,]  15 0.25 1 0
[3,] 120 1.00 0 1
[4,]  12 1.00 0 0
```

# Lung cancer, asbestos and smoking

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

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



# Lung cancer, asbestos and smoking

Summary and extraction of parameters:

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

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

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

(Dispersion parameter for poisson family taken to be 1)

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



# Summary and extraction of parameters

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

Parameters are the same for the two modelling approaches.

## Interpretation of parameters

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

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

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



## 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 | $\alpha+\gamma$ | $\alpha$ |
| Rate ratio | 5 | 5 | $\ln(\beta)$ | $\ln(\beta)$ |
| Rate difference | 480 | 48 | $\beta$ | $\beta$ |



## Log-linear model (cont'd)

Model with effect modification (two regressors only)

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

equivalently

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

where $\alpha$ is as before, but

- $\beta$ = log-rate ratio $\rho$ for a unit change in $X$ when $Z = 0$,
- $\gamma$ = log-rate ratio $\tau$ for a unit change in $Z$ when $X = 0$

## Interaction parameter

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

For binary $X$ and $Z$ we have

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

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



## Log-linear model: Estimated rates

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



## Lung cancer, asbestos and smoking

```
> mi <- glm( D/Y ~ A + S + I(A*S), weight=Y, family=poisson )
> round( ci.exp( mm ), 3 ) ; round( ci.exp( mi ), 3 )
            exp(Est.)  2.5%  97.5%
(Intercept)        12 8.060 17.867
A                   5 3.977  6.286
S                  10 6.733 14.853
            exp(Est.)  2.5%  97.5%
(Intercept)        12 6.815 21.130
A                   5 2.340 10.682
S                  10 5.524 18.101
I(A * S)            1 0.451  2.217
```

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

## Additive model for rates

General form with two regressors

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

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



## Additive model (cont'd)

- $\delta$ = 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
- $\beta$ = rate difference for unit change in $X$ for all values of $Z$
- $\gamma$ = rate difference for unit change in $Z$ for all values of $X$,



## Example: Additive model

```
> mai <- glm( D/Y ~ A + S + A*S, weight=Y, family=poisson(link=identity) )
> ci.exp( mai, Exp=FALSE )
            Estimate       2.5%     97.5%
(Intercept)       12   5.210486  18.78951
A                 48  16.886536  79.11346
S                108  85.481728 130.51827
A:S              432 328.808315 535.19168
```

A very clear interaction (effect modification)

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

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



## Model fitting (cont'd)

Output from computer packages will give:

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

May be difficult to interpret!

Model checking & diagnostics:

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



## Problems in modelling

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

## Modeling

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



# Conclusion

Nordic Summerschool of Cancer Epidemiology
August 2015
Danish Cancer Society
http://BendixCarstensen.com/NSCE/2015

concl-analysis

## Concluding remarks

Epidemiologic study is a

**Measurement excercise**

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

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

Result: **Estimate** of the parameter.

## Estimation and its errors

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

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



## Sources of bias

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



## Sources of random error

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

## Random sampling

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



## Randomization

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



## Controlled randomness

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

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

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

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

## Controlled randomness (cont'd)

*Question*: How often controlled randomness actually employed in epidemiology?

*Answer*: Rarely!

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



## Implications

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

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

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



## Recommendations

Possible remedies for these problems:

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

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

# Conclusion

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

Motto (Campbell & Machin 1983):

**STATISTICS is about**
**COMMON SENSE and**
**GOOD DESIGN!**