

MPH specialmodul Epidemiologi og Biostatistik

Confounding
Stratificeret analyse
16. marts 2009

www.biostat.ku.dk/~pka/mphspec09

Per Kragh Andersen

1

Epidemiology.

Study of *distribution* and *determinants* of disease frequency in human populations.

We need: *measures* of disease frequency.

Typically: disease *outcome* is *binary*

and we may use: *risk, rate, odds, prevalence.*

We want to *compare* these among “*exposed*” and “*non-exposed*” persons and, more generally

to *relate* these to *exposure variables / determinants* and other *explanatory variables.*

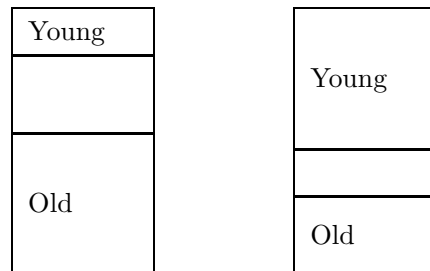
2

Confounding.

Do we always get a fair comparison between exposed and non-exposed?

EXPOSED

NON-EXPOSED



Not necessarily - a randomly selected exposed person tends to be older than a randomly chosen non-exposed. This is a problem if age is a risk factor for the outcome.

3

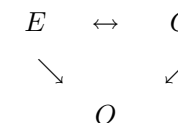
Confounding.

A variable C is a potential confounder for the relation: $E \rightarrow O$

if it is

- 1) related to the exposure: $E \leftrightarrow C$
- 2) an independent risk factor for the outcome: $C \rightarrow O$
- 3) not a consequence of the exposure: $E \rightarrow C \rightarrow O$

That is:



4

Example: 1970 US mortality data (Kahn & Sempos, 1989).

Age	California (a)			Maine (b)		
	Pop.in 1000	No.of deaths	Rate per 1000 ys.	Pop.in 1000	No.of deaths	Rate per 1000 ys.
< 15	5524	8751	1.6	286	535	1.9
15-24	3558	4747	1.3	168	192	1.1
25-34	2677	4036	1.5	110	152	1.4
35-44	2359	6701	2.8	109	313	2.9
45-54	2330	15675	6.7	110	759	6.9
55-64	1704	26276	15.4	94	1622	17.3
65-74	1105	36259	32.8	69	2690	39.0
75+	696	63840	91.7	46	4788	104.1
Total	19953	166285	8.3	992	11051	11.1

“Crude rates”

(a) California: $\frac{166285}{19953000} = 8.3$ per 1000 ys.
 (b) Maine: $\frac{11051}{992000} = 11.1$ per 1000 ys.

$RR = 1.34$

The age distributions in California and Maine differ so the apparent difference may be (partly) ascribed to this.

Confounding

Example: Age is a confounder for this study, since

- 2) age is a risk factor for mortality (obvious)
- 1) the age distributions differ between the two states

Age	California		Maine	
	Pop. in 1000	%	Pop. in 1000	%
< 15	5524	28	286	29
15-24	3558	18	168	17
25-34	2677	13	110	11
35-44	2359	12	109	11
45-54	2330	12	110	11
55-64	1704	9	94	9
65-74	1105	6	69	7
75+	696	3	46	5
Total	19953	100	992	100

Adjustment for confounding using stratification.

Example:

Relationship of age and systolic blood pressure to prevalence of MI in a sample of individuals in the Israeli Ischemic Heart Disease Study (unpublished data - Table 5-3 in Kahn & Sempos (1989)).

	Myocardial infarction		Total
	Present	Absent	
SBP ≥ 140	29	711	740
SBP < 140	27	1244	1271
Total	56	1955	2011

$\ln(OR) = \ln\left(\frac{29 \times 1244}{711 \times 27}\right) = \ln(1.88) = 0.631.$

$$\ln(OR) = \ln\left(\frac{29 \times 1244}{711 \times 27}\right) = \ln(1.88) = 0.631.$$

95% confidence interval:

$$L_1 = 0.631 - 1.96\sqrt{\frac{1}{29} + \frac{1}{27} + \frac{1}{711} + \frac{1}{1244}}$$

$$= 0.631 - 0.532 = 0.099$$

$$L_2 = 0.631 + 0.532 = 1.163.$$

From $\exp(L_1) = 1.10$ to $\exp(L_2) = 3.20$

Age as potential confounder.

2)	Myocardial infarction		Total
	Present	Absent	
Age \geq 60	15	188	203
Age < 60	41	1767	1808
Total	56	1955	2011

$$OR = \frac{15 \times 1767}{188 \times 41} = 3.44$$

1)	Age		Total
	\geq 60	< 60	
SBP \geq 140	124	616	740
SBP < 140	79	1192	1271
Total	203	1808	2011

$$OR = \frac{124 \times 1192}{616 \times 79} = 3.04$$

Separate analyses in strata defined by confounder.

Age \geq 60	MI cases	MI negative	Total	
SBP \geq 140	9	115	124	
SBP < 140	6	73	79	
Total	15	188	203	OR=0.95

Age < 60	MI cases	MI negative	Total	
SBP \geq 140	20	596	616	
SBP < 140	21	1171	1192	
Total	41	1767	1808	OR=1.87

Combined analysis over strata (= stratified analysis) (= the Mantel-Haenszel method)

We have a series (here two!) of two by two tables: one from each stratum.

stratum 1	\dots	stratum k
a b	\dots	a b
c d	\dots	c d
n		n

In each stratum, we can estimate odds ratio by

$$\frac{a \cdot d}{b \cdot c} = \frac{a \cdot d/n}{b \cdot c/n}$$

A common odds ratio for all strata may be estimated by the Mantel-Haenszel estimator

$$\frac{\sum \frac{a \cdot d}{n}}{\sum \frac{b \cdot c}{n}} = OR_{MH}$$

(weighted average of separate OR-estimates)

In the example:

$$\frac{\frac{9 \cdot 73}{203} + \frac{20 \cdot 1171}{1808}}{\frac{115 \cdot 6}{203} + \frac{596 \cdot 21}{1808}} = 1.57 = OR_{MH}$$

Interpretation:

OR_{MH} is an estimate of the association between exposure (SBP) and outcome (prevalence of MI), *adjusted for the confounder (age)*.

X_{MH}^2 is a test statistic for no association between exposure and outcome, *adjusted for the confounder*.

The Mantel-Haenszel test.

In each stratum, we can calculate:

$$\begin{aligned} OBServed &= a \\ EXPected &= \frac{(a+b)(a+c)}{n} = E(a) \\ SE &= \sqrt{\frac{(a+b)(c+d)(a+c)(b+d)}{n^2(n-1)}} = SE(a) \end{aligned}$$

The combined Mantel-Haenszel test statistic is

$$\frac{(\sum a - \sum E(a))^2}{\sum (SE(a))^2} = X_{MH}^2 \sim \chi_1^2 \text{ under } H_0$$

In the example: X_{MH}^2

$$\frac{((9 + 20) - (\frac{15 \cdot 124}{203} + \frac{41 \cdot 616}{1808}))^2}{\frac{15 \cdot 124 \cdot 79 \cdot 188}{(203)^2 \cdot 202} + \frac{41 \cdot 616 \cdot 1192 \cdot 1767}{(1808)^2 \cdot 1807}} = \frac{(29 - 23.13)^2}{12.32} = 2.80, \quad P = 0.09$$

Confidence limits for common odds ratio

- 1) calculate $\ln(OR_{MH})$
 $\ln(1.57) = 0.451$
- 2) calculate: $L_1 = \ln(OR_{MH}) - 1.96 \cdot SE$
and $L_2 = \ln(OR_{MH}) + 1.96 \cdot SE$
where $SE = \frac{|\ln(OR_{MH})|}{\sqrt{X_{MH}^2}}$
 $= \frac{0.451}{\sqrt{2.80}} = 0.270$
that is: $L_1 = 0.451 - 1.96 \cdot 0.270$
 $= -0.077$
 $L_2 = 0.451 + 1.96 \cdot 0.270$
 $= 0.979$
- 3) The desired 95% confidence limits are
from $\exp(L_1) = 0.93$
to $\exp(L_2) = 2.66$

Other approaches to confounder adjustment in case-control studies

1. “*Woolf’s method*” (McNeil, pp. 104-105)

Based on using weighted averages of $\ln(OR)$

2. *Regression analysis (logistic regression)*, see McNeil, Chapter 6 and Silva, Chapter 14 (brief).

Using these methods, it is also possible to estimate/test the effect of an exposure on an outcome adjusted for other variables.

Stratified analysis: Silva, Chapter 14; McNeil, Chapter 4.

When is the stratified analysis sensible?

In the stratified analysis, we *average the individual OR’s* from the separate strata.

This makes sense *if the individual OR’s point in the same direction in all strata*, e.g., if the exposure tends to increase the risk of the outcome in all strata

= if there is *no interaction* between exposure and stratification variable on the outcome

= if there is *no effect-modification* of the stratification variable on the relation between exposure and outcome

Tests for no interaction.

Age \geq 60	MI cases	MI negative	Total	
SBP \geq 140	9	115	124	
SBP $<$ 140	6	73	79	
Total	15	188	203	OR=0.95

Age $<$ 60	MI cases	MI negative	Total	
SBP \geq 140	20	596	616	
SBP $<$ 140	21	1171	1192	
Total	41	1767	1808	OR=1.87

Interaction? (= Effect-modification?)

Are the separate *OR’s*, 0.95 and 1.87 different?

Can be tested using Woolf’s method, logistic regression or the “Breslow-Day” test for homogeneity.

Breslow-Day’s test of no interaction.

Compare *OB*Served counts with *EX*Pected counts assuming a constant *OR* over strata.

a	b	$a + b$
c	d	$c + d$
$a + c$	$b + d$	$n = a + b + c + d$

Compute expected $E_{BD}(a)$ such that *OR* in each table is OR_{MH} , i.e.:

$E_{BD}(a)$	$a + b - E_{BD}(a)$	$a + b$
$a + c - E_{BD}(a)$	$d - a + E_{BD}(a)$	$c + d$
$a + c$	$b + d$	$n = a + b + c + d$

such that

$$OR_{MH} = \frac{E_{BD}(a)(d - a + E_{BD}(a))}{(a + c - E_{BD}(a))(a + b - E_{BD}(a))}$$

Quadratic equation! Breslow-Day test:

$$\sum_{\text{strata}} \left(\frac{a - E_{BD}(a)}{SE} \right)^2 \approx \chi_{k-1}^2$$

(where $SE = \dots$).

The Framingham study

Planned as a *20 year cohort study* of residents aged 30-59 in Framingham town, Massachusetts, in 1948.

Aim: 6000 persons (\Rightarrow 2000 CHDs)

Sampling: List of families stratified by size and district. For every 3 families, 2 were selected and all members from relevant age groups were invited.

Hope: 90% accepted

Result: 69% accepted = 4469 persons

Addition: volunteers: 740 persons

Here: Combined data

age \geq 45

CHOL at exam 1

1406 persons, 10 exams (\sim 18 years of follow-up)

13 variables selected:

Baseline: sex, age, FRW, SBP, DBP, CHOL, CIG (CHD)

Follow-up: SBP10, CHD, YRS_CHD, DEATH, YRS_DTH, CAUSE

Coding of variables.

- **sex** 1 for males, 2 for females
- **age** age (years) at baseline (45-62)
- **frw** "Framingham relative weight" (pct.) at baseline (52-222; 11 persons have missing values)
- **sbp** systolic blood pressure at baseline (*mmHg*) (90-300)
- **dbp** diastolic blood pressure at baseline (*mmHg*) 50-160)
- **chol** cholesterol at baseline (*mg/100ml*) (96-430)
- **cig** cigarettes per day at baseline (0-60; 1 person has missing value)
- **chd** 0 if no "coronary heart disease" during follow-up, 1 if "coronary heart disease" at baseline (prevalent cases), x=2-10 if "coronary heart disease" was diagnosed at follow-up no. x

- **sbp10** systolisk blodtryk ved follow-up 10 (*mmHg*) (94-264; 635 missing)
- **yrs_chd** person years at risk of developing "coronary heart disease" (0-18; 43 have missing values)
- **death** 0 if alive at follow-up no. 10, x=2-10 if dead between follow-up x-1 and x
- **yrs_dth** person years at risk of death (1-18)
- **cause** cause of death (0, 1, 2, 3, 4, 5, 6; 19 have missing values):
 - 0 alive at follow-up no. 10
 - 1 sudden CHD, 2 non-sudden CHD
 - 3 stroke, 4 other cardiovascular cause
 - 5 cancer
 - 6 other causes

SAS-øvelser.

De første 3 spørgsmål drejer sig alle om Framingham studiet, og der henvises til variabellisten i forelæsningsnoterne. SAS programmet `framing.sas` læser data fra filen `t:\framing.txt` og konstruerer en ny variabel `chdny`, som er

- 1 hvis `CHD=2, 3, ..., 10`
- 0 hvis `CHD=0`
- “missing” (‘.’) hvis `CHD=1`.
- **1.** Importer SAS-programmet `framing.sas` i program editor vinduet, udfør det, og sammenlign dermed 18-års CHD risikoen for mænd og kvinder.

25

- **2.** Se på aldersfordelingen i materialet ved hjælp af PROC UNIVARIATE og definer en ny variabel, hvor alderen opdeles i passende grupper. Sammenlign ved brug af PROC FREQ 18-års CHD risikoen for mænd og kvinder justeret for alder. Fortolk Breslow-Day testet for ingen interaktion.
- **3.** Sammenlign på tilsvarende vis 18-års CHD risikoen for mænd og kvinder justeret for systolisk blodtryk.

26

- **4.** SAS programmet `israeli.sas` svarer til eksemplet fra forelæsningsnoterne vedr. blodtryk og prevalens af MI. Kør programmet og rekonstruer derved resultaterne, som blev gennemgået, dvs.
 - **a.** separate *OR*'r i strata med 95% sikkerhedsintervaller
 - **b.** Mantel-Haenszel teststørrelsen
 - **c.** Mantel-Haenszel estimatoren for den justerede *OR* med sikkerhedsinterval og fortolk Breslow-Day testet for ingen interaktion.
 - **d.** Modificer programmet til også at analysere de tre første 2×2 tabeller fra eksemplet, dvs. blodtryk vs. mi, alder vs. mi, og alder vs. blodtryk.

27