Lecture 1: Introduction to Epidemiology

Dankmar Böhning

Department of Mathematics and Statistics
University of Reading, UK

Summer School in Cesme, May/June 2011

What is Epidemiology?

Epidemiology is the study of the determinants, distribution, and frequency of disease (who gets the disease and why)

-
- epidemiologists study sick people
- epidemiologists study healthy people
- to determine the crucial difference between those who get the disease and those who are spared
-
- epidemiologists study exposed people
- epidemiologists study non-exposed people
- to determine the crucial effect of the exposure

What is Epidemiology? Last's dictionary gives a detailed definition:

The study of the distribution and determinants of health-related states or events in specified populations, and the application of this study to control of health problems.

Uses of Epidemiology

- to determine, describe, and report on the natural course of disease, disability, injury, and death
- to aid in the planning and development of health services and programs
- to provide administrative and planning data

Uses of Epidemiology

- to study the cause (or etiology) of disease(s), or conditions, disorders, disabilities, etc.
- to determine the primary agent responsible or ascertain causative factors
- to determine the characteristics of the agent or causative factors
- to determine the mode of transmission
- to determine contributing factors
- to identify and determine geographic patterns

Purpose of Epidemiology

- to provide a basis for developing disease control and prevention measures for groups at risk
- this translates into developing measures to prevent or control disease

Two Broad Types of Epidemiology:

- descriptive epidemiology: examining the distribution of disease in a population, and observing the basic features of its distribution
- analytic epidemiology: investigating a hypothesis about the cause of disease by studying how exposures relate to disease

descriptive epidemiology is antecedent to analytical epidemiology:

analytical epidemiology studies require information to ...

- know where to look
- know what to control for
- develop viable hypotheses

three essentials characteristics of disease that we look for in descriptive studies are ...

- Person
- Place
- Time

Person

- age, gender, ethnic group
- genetic predisposition
- concurrent disease
- diet, physical activity, smoking
- risk taking behavior
- SES, education, occupation

geographic Place

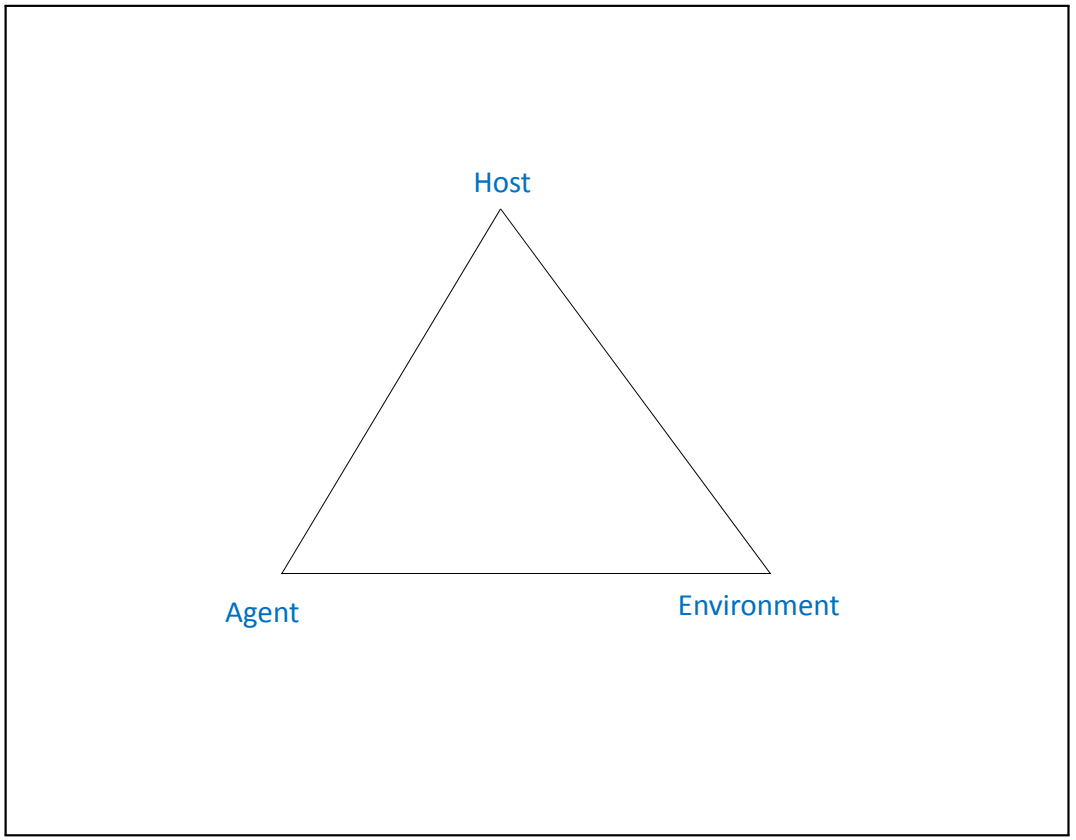
- presence of agents or vectors
- climate
- geology
- population density
- economic development
- nutritional practices
- medical practices

Time

- calendar time
- time since an event
- physiologic cycles
- age (time since birth)
- seasonality
- temporal trends

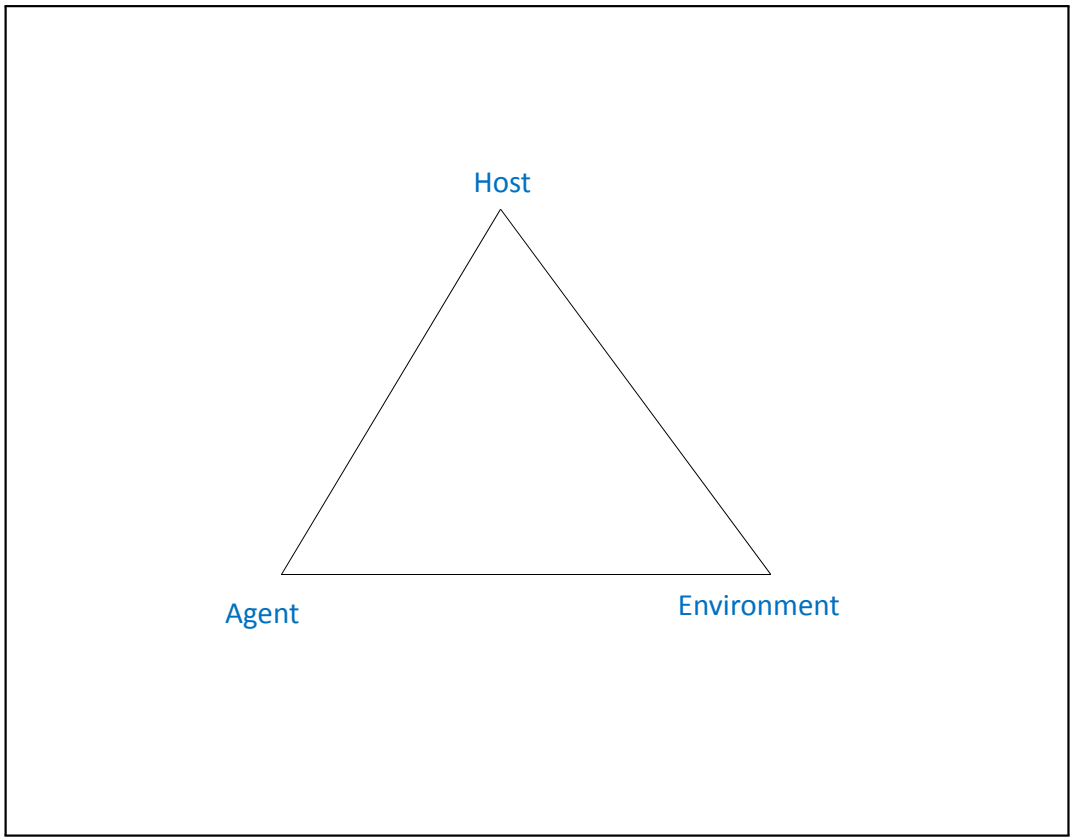
The Epidemiologic Triangle: three characteristics that are examined to study the cause(s) for disease in analytic epidemiology

- host
- agent
- environment



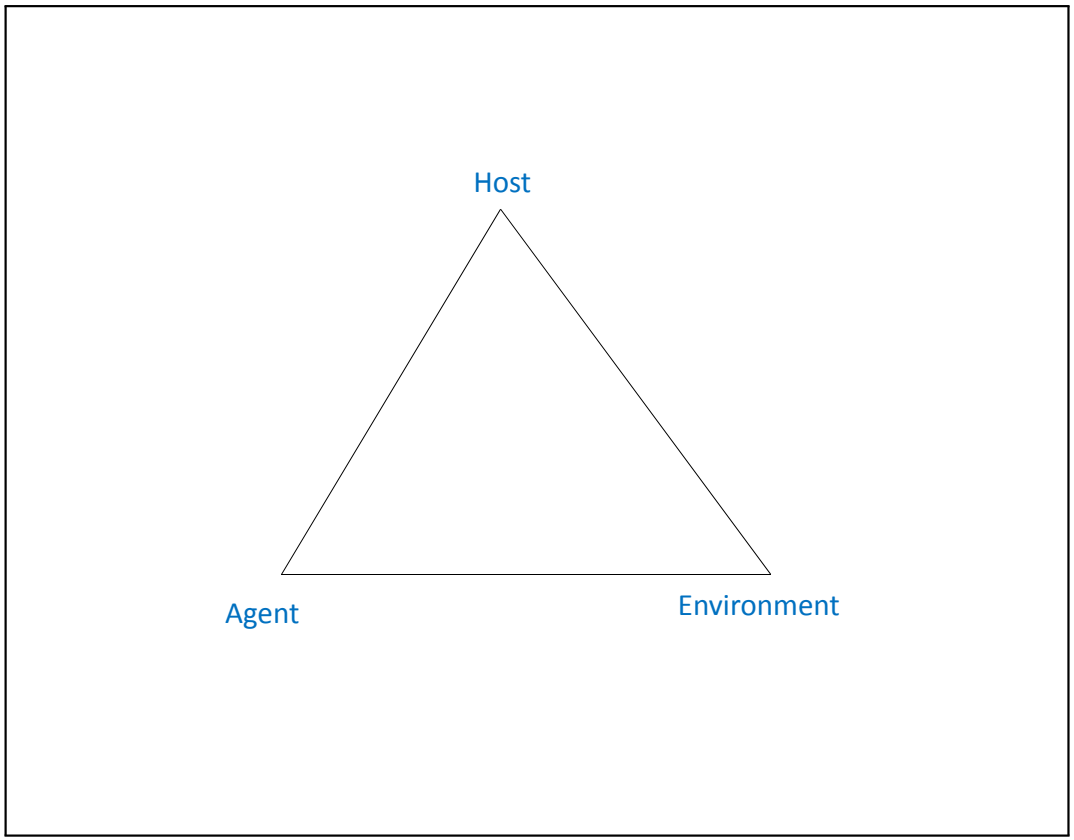
The Epidemiologic Triangle

- **host**
- personal traits
- behaviors
- genetic predisposition
- immunologic factors
- ...



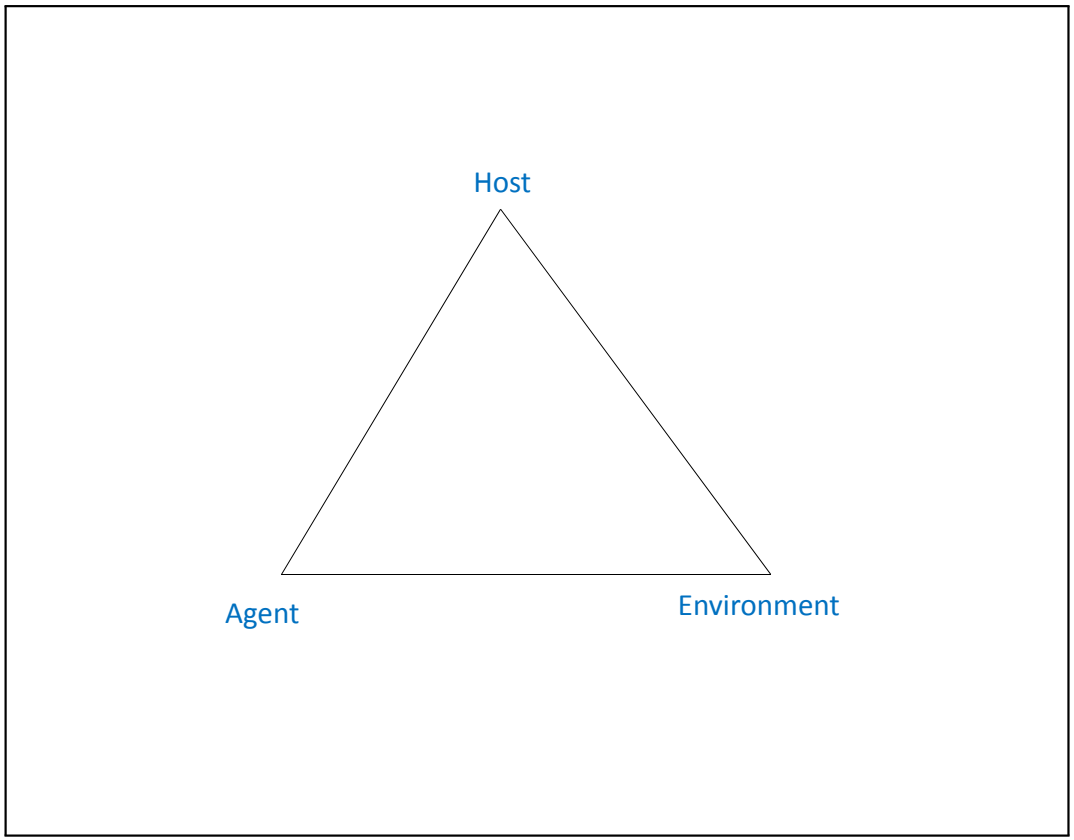
The Epidemiologic Triangle

- **agents**
- biological
- physical
- chemical
- ...
- influence the chance for disease or its severity



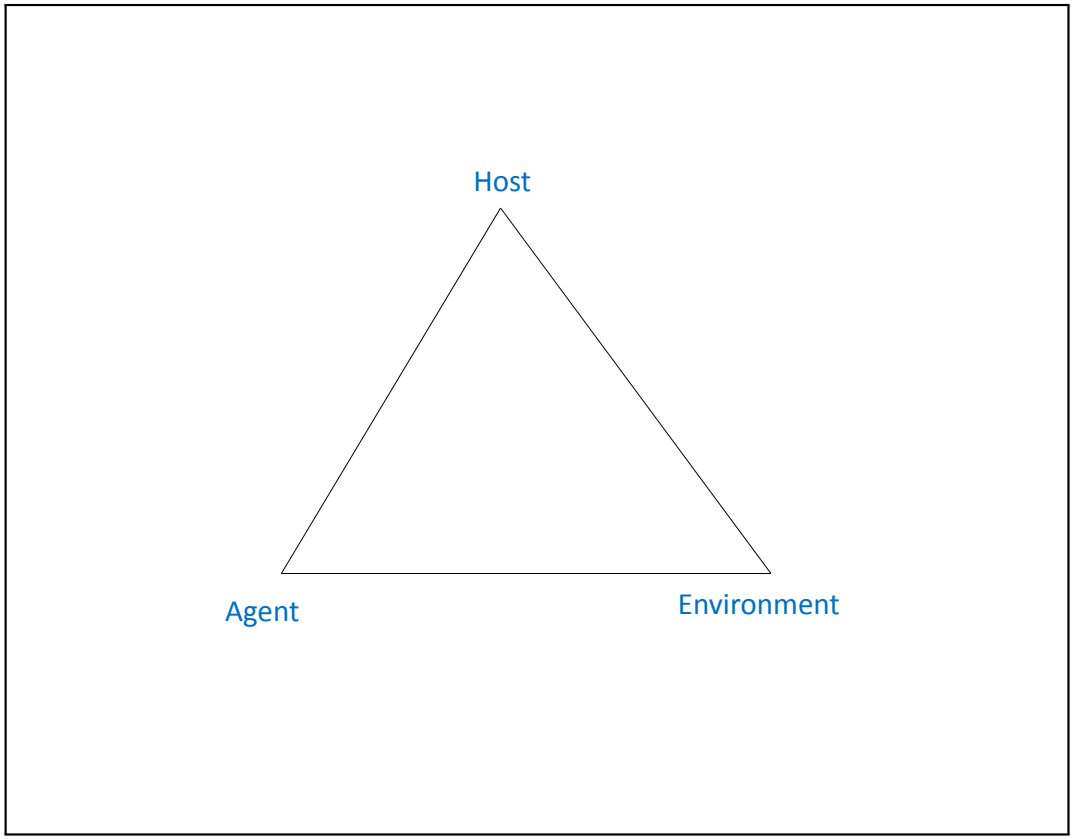
The Epidemiologic Triangle

- **environment**
- external conditions
- physical/biological/social
- ...
- contribute to the disease process



Epidemics occur when ..

- **host, agent and environmental factors are not in balance**
- due to new agent
- due to change in existing agent (infectivity, pathogenicity, virulence)
- due to change in number of susceptibles in the population
- due to environmental changes that affect transmission of the agent of growth of the agent



Epidemiologic Activities

- often concentrate on PPT
- demographic distribution
- geographic distribution
- seasonal patterns and temporal trends
- frequency of disease patterns

Epidemiologic Activities

- are built around the analysis of the relationship between
  - exposures
  - disease occurrence
- are built around the analysis of differences between
  - cases
  - healthy controls

Lecture 2: Measuring Disease Occurrence (Morbidity and Mortality): Prevalence, incidence, incidence density

Dankmar Böhning

Department of Mathematics and Statistics
University of Reading, UK

Summer School in Cesme, May/June 2011

Purpose

The purpose of this material is to provide an overview on the most important measures of disease occurrence:

- prevalence
- incidence (cumulative incidence or risk)
- incidence density

Examples

The concepts will be illustrated with examples and practicals.

Epidemiology and it's Definition

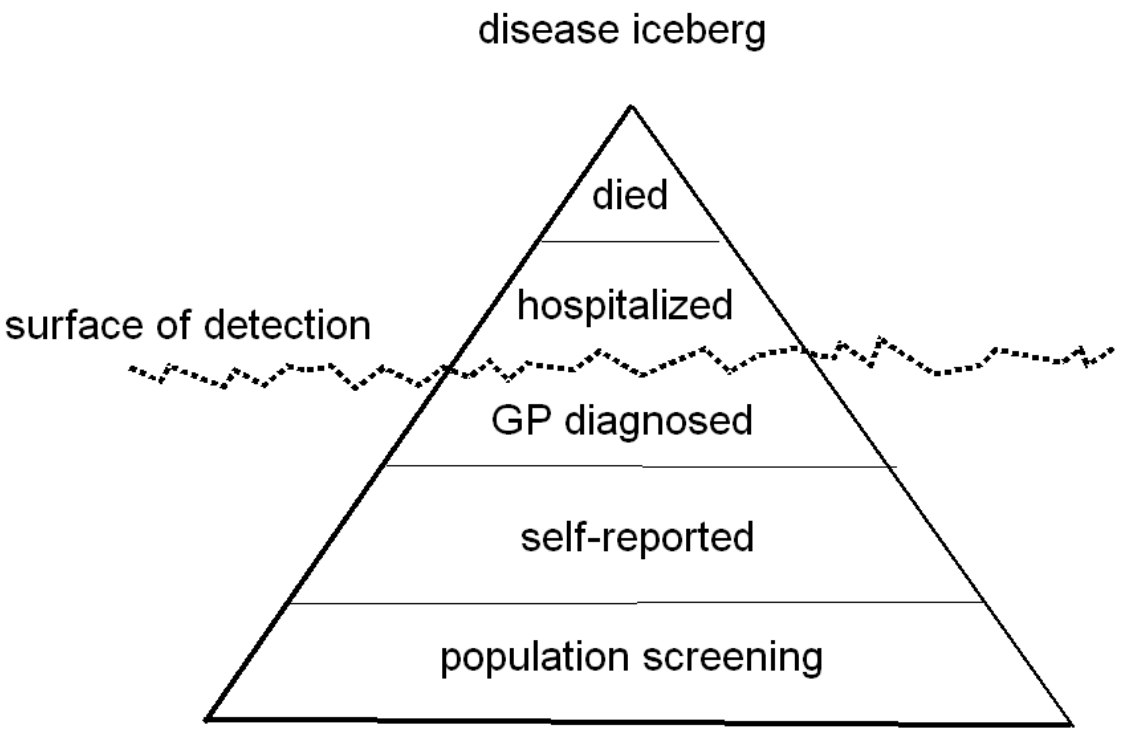
Definition

Epidemiology studies the distribution of diseases in populations and factors related to them.

This definition leads to **two** questions:

1. How can we measure diseases and their distributions?

- morbidity
  - prevalence
  - incidence
- mortality
  - incidence


disease iceberg
died
hospitalized
surface of detection
GP diagnosed
self-reported
population screening

2. How can we measure differences in disease occurrence in different populations?

- epidemiological study types
  - cross-sectional
  - clinical trials
  - cohort studies
  - case-control studies
- epidemiological measures of effect
  - differences in disease risk
  - ratios in disease risk
  - relative differences in disease risk

Measuring Disease Occurrence: Prevalence

Prevalence:

is the **proportion** (denoted as p) of a specific population having a particular disease. p is a number between 0 and 1. If multiplied by 100 it is **percentage**.

Examples

In a population of 1000 there are two cases of malaria:
$p = 2/1000 = 0.002$ or 0.2%.
In a population of 10,000 there are 4 cases of skin cancer:
$p = 4/10,000 = 0.0004$ or 0.04%.

Measuring Disease Occurrence: Prevalence

epidemiological terminology

In epidemiology, disease occurrence is frequently small relative to the population size. Therefore, the proportion figures are multiplied by an appropriate number such as 10,000. In the above second example, we have a prevalence of 4 per 10,000 persons.

Exercise

In a county with 2300 inhabitant there have occurred 2 cases of leukemia. Prevalence?

Quantitative Aspects:

What is Variance and Confidence Interval for the Prevalence!

sample:

sample (population survey) of size n provides for disease status for each unit of the sample:

$$X_i = 1, \text{ disease present}$$

$$X_i = 0, \text{ disease not present}$$

consequently,

$$\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n}$$

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

plausible **estimator of prevalence**.

Computing Variance of Prevalence of X_i:

$$E(X_i) = 1 \times P(X_i = 1) + 0 \times P(X_i = 0)$$
$$= 1 \times p + 0 \times (1 - p) = p$$

$$Var(X_i) = (1 - p)^2 P(X_i = 1) + (0 - p)^2 P(X_i = 0)$$
$$= (1 - p)^2 p + p^2(1 - p) = (1 - p)p[1 - p + p]$$
$$= p(1 - p)$$

Computing Variance of Prevalence of X_i:

consequently,

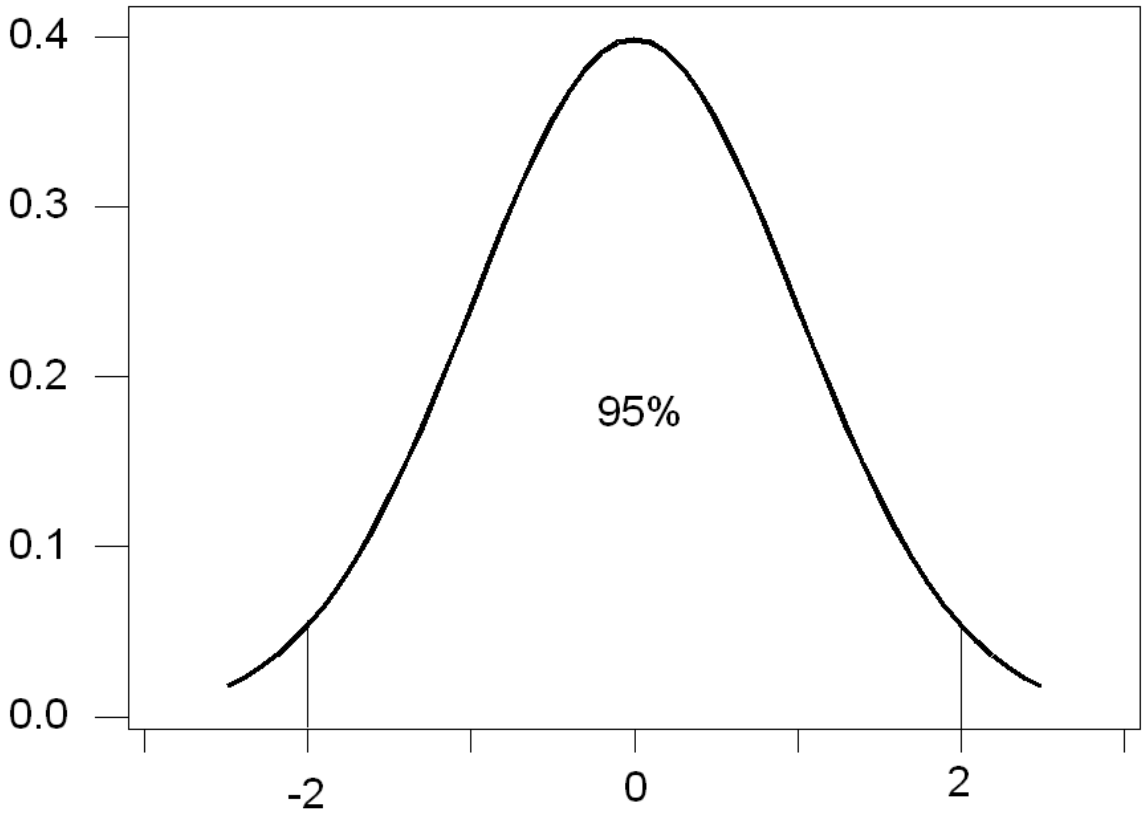
$$Var\left(\hat{p}\right) = Var\left(\frac{\sum_i X_i}{n}\right) = \frac{1}{n^2} Var(\sum_i X_i)$$

$$= \frac{1}{n^2} \sum_i Var(X_i) = \frac{1}{n^2} n \times p(1-p)$$

$$= \frac{p(1-p)}{n}$$

$$SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

$\hat{p}$ is approx. normal



using the normal distribution for $\hat{p}$:

with 95% probability

$$-2 \leq \frac{\hat{p} - p}{SD(\hat{p})} \leq +2$$

$\Leftrightarrow$

$$\hat{p} - 2SD(\hat{p}) \leq p \leq \hat{p} + 2SD(\hat{p})$$

$\Leftrightarrow$

$$95\%CI : \hat{p} \pm 2SD(\hat{p})$$
$$= \hat{p} \pm 2\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}$$

Examples

In a population of 1000 there are two cases of malaria: $p = 2/1000 = 0.002$ or 0.2%.

$$Var(\hat{p}) = 0.002(1 - 0.002)/1000 = (0.00141280)^2,$$

$$SD(\hat{p}) = 0.00141280$$

$$95\%CI : \hat{p} \pm 2\sqrt{\hat{p}(1 - \hat{p})}/\sqrt{n}$$

$$= 0.002 \pm 2 \times 0.0014 = (0 - 0.0048)$$

Exercise

In a county with 2300 inhabitants there have occurred 2 cases of leukemia. Prevalence with CI?

Practical 1: Prevalence of Caries in Belo Horizonte

The BELCAP Study; background:

- Dental epidemiological study.
- A prospective study of school-children from an urban area of Belo Horizonte, Brazil.
- The Belo Horizonte caries prevention (BELCAP) study.
- The aim of the study was to compare different methods to prevent caries.

- Children selected were all 7 years-old and from a similar socio-economic background.
- Interventions:
  - Control (3),
  - Oral health education (1),
  - Enrichment of the school diet with rice bran (4),
  - Mouthwash (5),
  - Oral hygiene (6),
  - All four methods together (2).
- Interventions were cluster randomised to 6 different schools.
- Response, or outcome variable = DMFT index. (Number of decayed, missing or filled teeth.) DMFT index was calculated at the start of the study and 2 years later. Only the 8 deciduous molars were considered.
- Potential confounders: sex (female 0 male 1), ethnicity.
- Data analysed by Böhning et al. (1999,*Journ. Royal Statist. Soc. A*).

Practical 1: Prevalence of Caries in Belo Horizonte

Questions:

calculate prevalence of caries (DMFT > 0) with 95% CI at **study begin**:

- overall
- stratified by gender
- stratified by school
- stratified by gender and school

Measuring Disease Occurrence: Incidence

Incidence:

is the proportion (denoted as I) of a specific, **disease-free** population **developing** a particular disease **in a specific study period**. I is a number between 0 and 1. If multiplied by 100 it is percentage.

Examples

In a malaria-free population of 1000 there are four new cases of malaria within one year : $I = 4/1000 = 0.004$ or 0.4%.
In a skin-cancer free population of 10,000 there are 11 new cases of skin cancer: $I = 11/10,000 = 0.0011$ or 0.11%.

Measuring Disease Occurrence: Incidence

Exercise

In a rural county with 2000 children within pre-school age there have occurred 15 new cases of leukemia within 10 years. Incidence?

Quantitative Aspects: How to determine Variance and Confidence Interval for the Incidence?

sample (population cohort - longitudinal) of size n, which is **initially disease-free**, provides the disease status for each unit of the sample **at the end of study period**:

$$X_i = 1, \text{ new case}$$

$$X_i = 0, \text{ disease not present}$$

consequently,

$$\hat{I} = \frac{X_1 + X_2 + ... + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

plausible **estimator of incidence**.

Computing Variance of Incidence

Consider any of the X_i:

$$E(X_i) = 1 \times P(X_i = 1) + 0 \times P(X_i = 0)$$

$$= 1 \times I + 0 \times (1 - I) = I$$

$$Var(X_i) = (1 - I)^2 P(X_i = 1) + (0 - I)^2 P(X_i = 0)$$

$$= (1 - I)^2 I + I^2 (1 - I) = (1 - I) I [1 - I + I]$$

$$= I(1 - I)$$

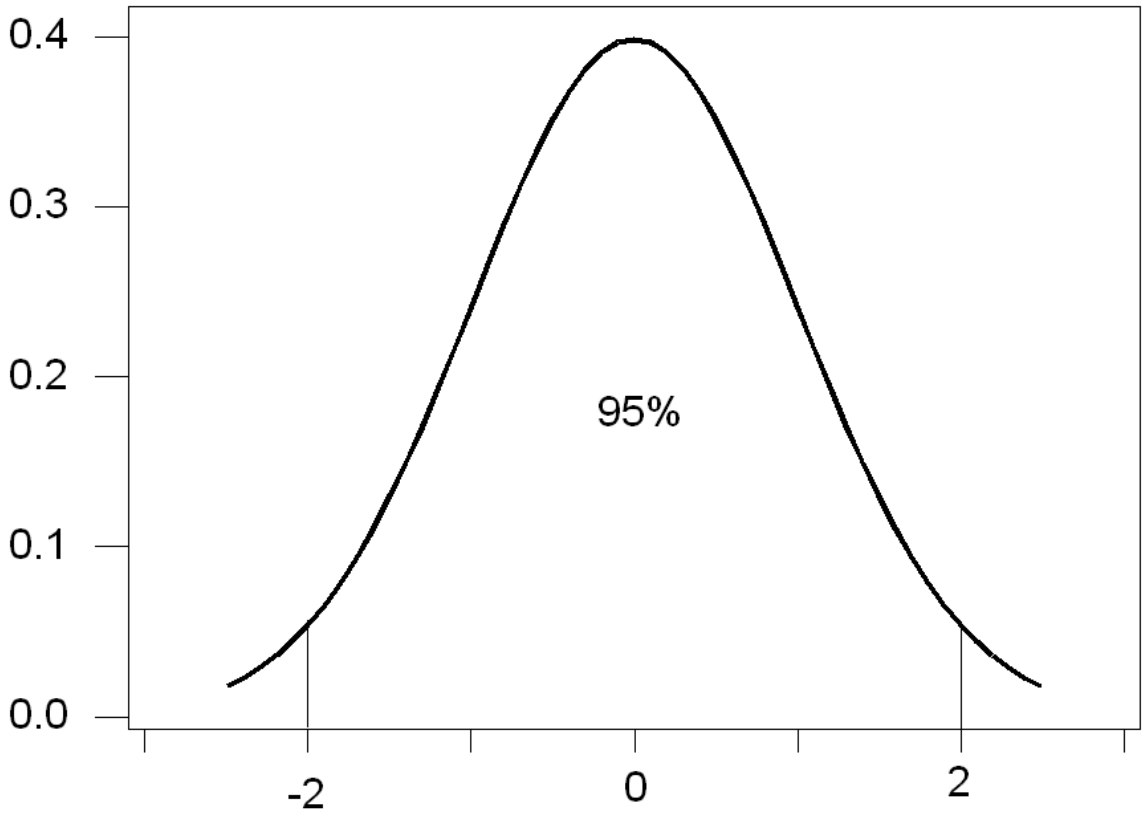
consequently,

$$Var\left(\frac{\sum_i X_i}{n}\right) = \frac{1}{n^2} Var(\sum_i X_i)$$

$$= \frac{1}{n^2} \sum_i Var(X_i) = \frac{1}{n^2} n \times I(1-I) = \frac{I(1-I)}{n}$$

$$SD(\hat{I}) = \sqrt{\frac{I(1-I)}{n}}$$

$\hat{p}$ is approx. normal



95% confidence interval for the incidence density

with 95% probability

$$-2 \leq \frac{\hat{I} - I}{SD(\hat{I})} \leq +2$$

$\Leftrightarrow$

$$\hat{I} - 2SD(\hat{I}) \leq I \leq \hat{I} + 2SD(\hat{I})$$

$\Leftrightarrow$

$$95\%CI : \hat{I} \pm 2SD(\hat{I})$$

$$= \hat{I} \pm 2\sqrt{\hat{I}(1-\hat{I})}/\sqrt{n}$$

Examples

In a malaria-free population of 1000 there are four new cases of malaria within one year : $I = 4/1000 = 0.004$ or .4%.

$$Var(\hat{I}) = 0.004(1 - 0.004)/1000 = (0.001996)^2,$$

$$SD(\hat{I}) = 0.001996$$

$$95\%CI : \hat{I} \pm 2\sqrt{\hat{I}(1 - \hat{I})}/\sqrt{n}$$

$$= 0.004 \pm 2 \times 0.001996 = (0.000008 - 0.0080)$$

Exercise

In a rural county with 2000 children within pre-school age there have occurred 15 new cases of leukemia within 10 years. Incidence with 95% CI?

Practical 1: Prevalence of Caries in Belo Horizonte

Questions:

calculate incidence of caries (DMFT $= 0$ begin of study **and** at DMFT > 0 at the end of study) with 95% CI:

- overall
- stratified by gender
- stratified by school
- stratified by gender and school
- why is it useless here to stratify by age?

Measuring Disease Occurrence: Incidence Density

Incidence Density:

is the rate (denoted as ID) of a specific, **disease-free** population **developing** a particular disease **w. r. t. a specific study period of length** T. ID is a positive number, but not necessarily between 0 and 1.

estimating incidence density

suppose a disease-free population of size n is under risk for a time period T. Then a plausible estimator of ID is given as

$$\widehat{ID} = \frac{\sum_{i=1}^{n} X_i}{n \times T} = \frac{\text{count of events}}{\text{person–time}}$$

where $X_i = 1$ if for person i disease occurs and 0 otherwise.

Examples

A cohort study is conducted to evaluate the relationship between dietary fat intake and the development in prostate cancer in men. In the study, 100 men with high fat diet are compared with 100 men who are on low fat diet. Both groups start at age 65 and are followed for 10 years. During the follow-up period, 10 men in the high fat intake group are diagnosed with prostate cancer and 5 men in the low fat intake group develop prostate cancer.
The incidence density is $\widehat{ID} = 10/(1,000) = 0.01$ in the high fat intake group and $\widehat{ID} = 5/(1,000) = 0.005$ in the low fat intake group.

most useful generalization

occurs if persons are **different times under risk** and hence contributing differently to the person–time–denominator

estimating incidence density with different risk-times

suppose a disease-free population of size n is under risk for a time periods $T_1, T_2, ..., T_n$, respectively. Then a plausible estimator of ID is given as

$$\widehat{ID} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} T_i} = \frac{\text{count of events}}{\text{person–time}}$$

where $X_i = 1$ if for person i disease occurs and 0 otherwise, and T_i represents the person-time of person i in the study period.

Examples

Consider a population of $n = 5$ factory workers with $X_2 = 1$ and all other $X_i = 0$(here the disease incidence might be a lung disease). We have also $T_1 = 12,\ T_2 = 2, T_3 = 6, T_4 = 12, T_5 = 5$, so that

$$\widehat{ID} = \frac{1}{12 + 2 + 6 + 12 + 5} = 1/37.$$

interpretation of incidence density:

In the above example of diet-cancer study: $\widehat{ID} = 0.01$ means what? There is no longer the interpretation of 1 case per 100 men, **but** 1 case per 100 men-years!

The interpretation is now **number of events per person–time**!

Quantitative Aspects for the Incidence Density

sample (population cohort - longitudinal) of size n available:

$$\text{event indicators: } X_1, ..., X_n$$

$$\text{person times: } T_1, ..., T_n$$

estimate of incidence density

$$\widehat{ID} = \frac{X_1 + X_2 + ... + X_n}{T_1 + T_2 + ... + T_n} = \frac{X}{T}$$

a variance estimate can be found as

$$\widehat{Var}(\widehat{ID}) = \frac{\widehat{ID}}{T} = \frac{X}{T^2}$$

Quantitative Aspects for the Incidence Density

variance estimate can be found as

$$\widehat{Var}(\widehat{ID}) = \frac{\widehat{ID}}{T} = \frac{X}{T^2}$$

so that a 95% confidence interval is given as

$$\widehat{ID} \pm 2\sqrt{\frac{\widehat{ID}}{T}}$$

Example

Consider the population of $n = 5$ factory workers with $X_2 = 1$ and all other $X_i = 0$ (here the disease incidence might be a lung disease). We have $X = 1$ and $T = 37$, so that $\widehat{ID} = 1/37 = 0.027$. The variance is $\frac{\widehat{ID}}{T} = 0.0007$ and standard deviation 0.027. This leads to a 95% CI

$$\widehat{ID} \pm 2\sqrt{\frac{\widehat{ID}}{T}} = 0.027 \pm 2 \times 0.027 = (0, 0.081).$$

Exercise

We return to the cohort study mentioned before. It had been conducted to evaluate the relationship between dietary fat intake and the development in prostate cancer in men. In the study, 100 men with high fat diet are compared with 100 men who are on low fat diet. Both groups start at age 65 and are followed for 10 years. During the follow-up period, 10 men in the high fat intake group are diagnosed with prostate cancer and 5 men in the low fat intake group develop prostate cancer.

Compute 95% CI for incidence densities:

$$\text{high fat intake group: } \widehat{ID} = 10/(1,000) = 0.01$$

$$\text{low fat intake group: } \widehat{ID} = 5/(1,000) = 0.005$$

Lecture 3: Direct Standardization of Measures of Disease Occurrence

Dankmar Böhning

Department of Mathematics and Statistics
University of Reading, UK

Summer School in Cesme, May/June 2011

Purpose

The purpose of this material is to provide an introduction to the problems of medical surveillance and associated standardization problems:

- comparing disease (risk factor) occurrence
- standardization methodology
- examples

Medical Surveillance

Example on problems with comparison of rates

The Directly Standardized Rate

How to execute in STATA?

Definition

detection of the occurrence of health-related events or exposures in a target population

Goal

to identify changes in the distributions of diseases in order to prevent or control these diseases within a population

potential specific goals

- identification of pattern of disease occurrence
- detection of disease outbreaks
- development of clues about possible risk factors (ecological study)
- finding of cases for further investigation
- anticipation of health service needs

traditionally

medical surveillance activities were developed to monitor the spread of infectious disease through a population

today

target are all diseases and health related conditions and exposures such as traffic accident morbidity and mortality, smoking, sexual habits, etc

Data Sources

Surveillance of deaths

- mortality statistics

Surveillance of morbidity

- important function of registries such as cancer registries, traffic accident registries, etc.
- legislation on certain transmittable diseases

Surveillance of risk factors

- micro-census
- survey

to detect change

morbidity or mortality needs frequently be compared

- in time (weekly, monthly, yearly, ...)
- in space (county, states, city-areas, ...)

such a comparison - if done without care - can be quite problematic

Comparing Mortality from Lung Cancer in Berlin (West) 1960 and 1989

age-group	deaths 1989	under risk	deaths 1960	under risk
35-39	3	78862	2	44454
40-44	15	74485	5	38932
45-49	49	96516	24	66595
50-54	64	78693	63	83553
55-59	88	48942	145	83353
60-64	83	38789	202	65947
65-69	125	29128	181	50805
70-74	86	19168	160	40282
75-79	126	25109	114	25545
80-84	113	17417	43	12431
85+	54	8821	9	4183
total	**806**	**515930**	**948**	**516080**

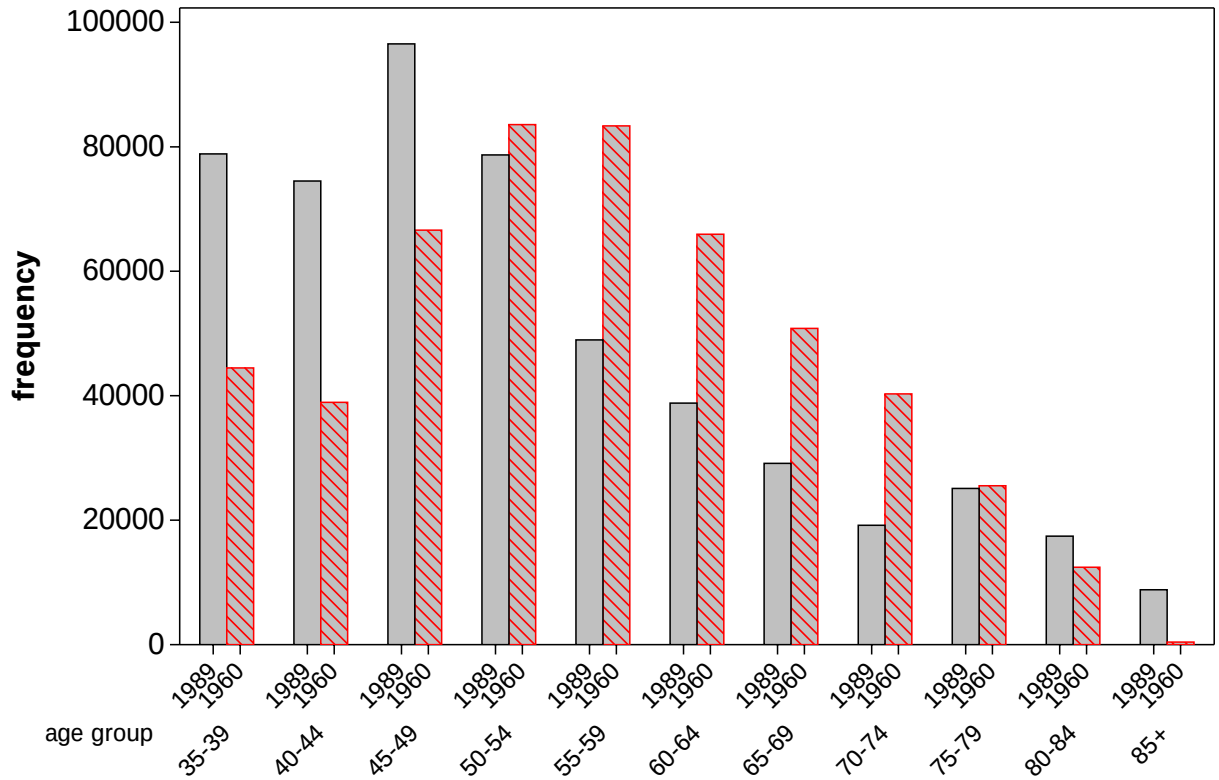
Comparing Mortality from Lung Cancer in Berlin (West) 1960 and 1989

- mortality rate 1960 $= \frac{948}{516080} \times 1000 = 1.84$
- mortality rate 1989 $= \frac{806}{515930} \times 1000 = 1.56$

coming to the **perplexing conclusion** that mortality has **dropped** from 1960 to 1989!

Comparing Mortality Rates from Lung Cancer in Berlin (West) 1960 and 1989

age-group	mortality rate 1989	mortality rate 1960
35-39	0.04	0.04
40-44	0.20	0.13
45-49	0.51	0.36
50-54	0.81	0.75
55-59	1.89	1.74
60-64	2.14	3.06
65-69	4.29	3.56
70-74	4.49	3.97
75-79	5.02	4.46
80-84	6.49	3.46
85+	6.12	2.15
total	**1.56**	**1.84**


frequency
100000
80000
60000
40000
20000
0
1989
1960
age group
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+

Explanation

- age distributions 1960 and 1989 are quite different
- 1989 age distribution puts more weight on younger ages
- 1960 age distribution puts more weight on older ages
- hence **crude rates** are not comparable

Solution

use identical age distribution

- World (Segi's Standard)
- Europe
- national

Two Reference Populations

age-group	World	Europe
...	...	...
35-39	6000	7000
40-44	6000	7000
45-49	6000	7000
50-54	5000	7000
55-59	4000	6000
60-64	4000	5000
65-69	3000	4000
70-74	2000	3000
75-79	1000	2000
80-84	500	1000
85+	500	1000
total	**100000**	**100000**

Construction of Directly Standardized Rate

	study population			reference population
age-group	**deaths**	**at risk**	**rate**	**at risk**
1	d_1	n_1	$p_1 = \frac{d_1}{n_1}$	N_1
2	d_2	n_2	$p_2 = \frac{d_2}{n_2}$	N_2
...	...	...	...	
k	d_k	n_k	$p_k = \frac{d_k}{n_k}$	N_k
total	d	n	$p = \frac{d}{n}$	N

crude rate:

$$p = \sum_{i=1}^{k} \frac{d_i}{n_i} \times \frac{n_i}{n}$$

standardized rate:

$$p_{\mathrm{DS}} = \sum_{i=1}^{k} \frac{d_i}{n_i} \times \frac{N_i}{N}$$

Computing the Standardized Mortality Rate for Lung Cancer in Berlin (West) 1989

age	deaths	under risk	rate	World	Expect.
35-39	3	78862	3/78862=0.00004	6000	0.23
40-44	15	74485	15/74485=0.00020	6000	1.21
45-49	49	96516	49/96516=0.00051	6000	3.05
50-54	64	78693	64/78693=0.00081	5000	4.07
...	...	...	...	...	
85+	54	8821	54/8821=0.00612	500	3.06
total	**806**	**515930**		**38000**	**57.47**

standardized rate (1989):

$$p_{\mathrm{DS}} = \frac{57.47}{38000} \times 1000 = 1.51$$

and, similarly, (1960): $p_{\mathrm{DS}} = \frac{52.08}{38000} \times 1000 = 1.37$

how to execute in STATA?

organization of data

first a data file needs to be constructed containing

- the stratums variable (age)
- the event variable (cases or deaths)
- the population size variable (population)
- the group variable containing information on the groups to be compared (year)

an example is given as follows:

	age	death	population	Year
1.	35-39	3	78862	1989
2.	40-44	15	74485	1989
3.	45-49	49	96516	1989
4.	50-54	64	78693	1989
5.	55-59	88	48942	1989
6.	60-64	83	38789	1989
7.	65-69	125	29128	1989
8.	70-74	86	19168	1989
9.	75-79	126	25109	1989
10.	80-84	113	17417	1989

	age	death	population	Year
11.	85+	54	8821	1989
12.	35-39	2	44454	1960
13.	40-44	5	38932	1960
14.	45-49	24	66595	1960
15.	50-54	63	83553	1960
16.	55-59	145	83353	1960
17.	60-64	202	65947	1960
18.	65-69	181	50805	1960
19.	70-74	160	40282	1960
20.	75-79	114	25545	1960
21.	80-84	43	12431	1960
22.	85+	9	4183	1960

how to execute in STATA?

organization of data

a second data file needs to be constructed containing

- the stratums variable (age) matching with **exactly the same name**
- the population size variable containing the **reference population** carrying the same name as the study population variable

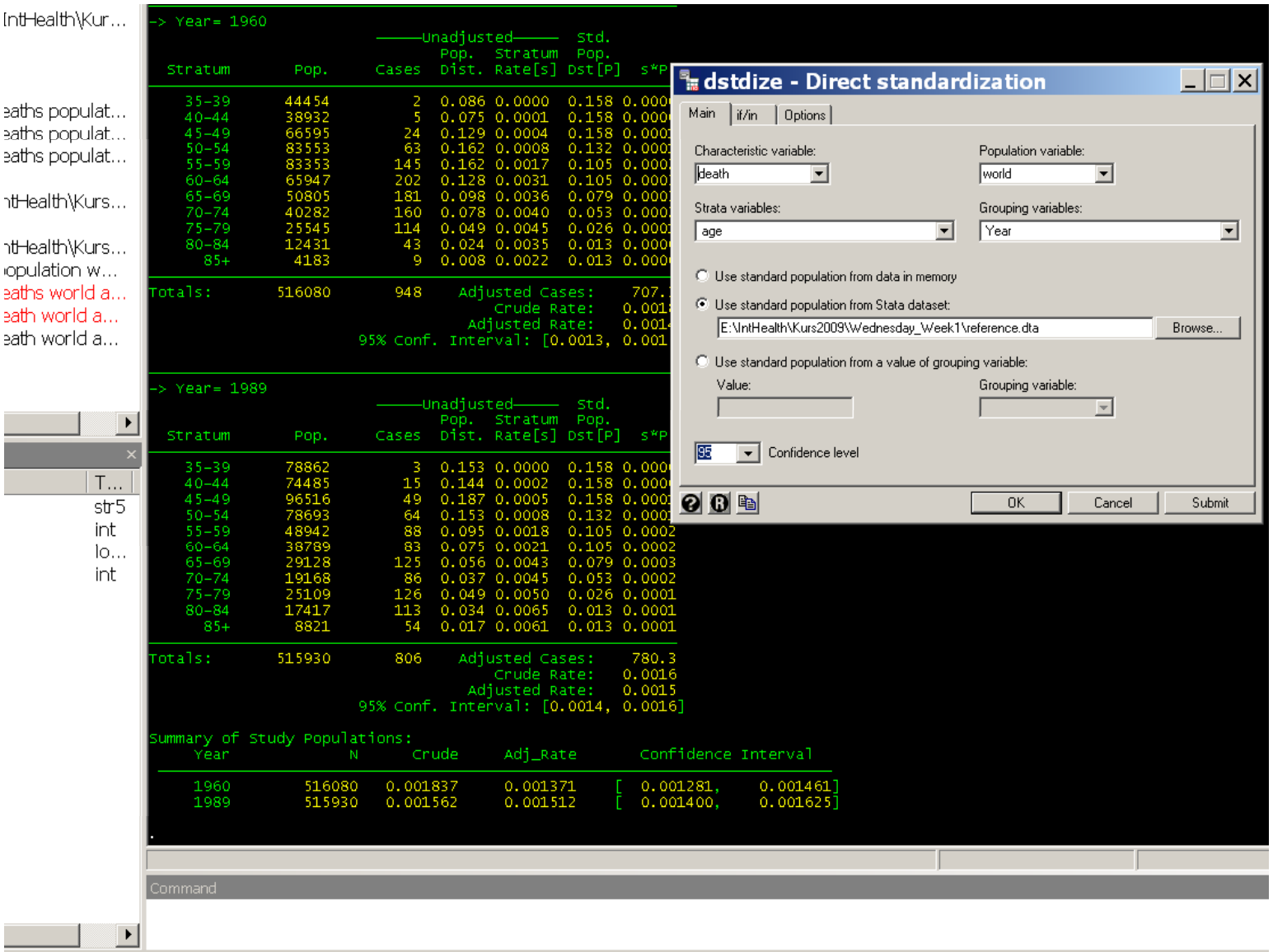
an example is given as follows in which `population` contains now the distribution of the world standard

	age	world	europe
1.	35-39	6000	7000
2.	40-44	6000	7000
3.	45-49	6000	7000
4.	50-54	5000	7000
5.	55-59	4000	6000
6.	60-64	4000	5000
7.	65-69	3000	4000
8.	70-74	2000	3000
9.	75-79	1000	2000
10.	80-84	500	1000
11.	85+	500	1000

how to execute in STATA?

execution of standardization

a very practical way to accomplish this is to choose in the first file the population name as the name of the reference standard, in this example `world`


-> Year= 1960
Stratum Pop. Cases Pop. Dist. Stratum Rate[s] Std. Pop. Dst[P]
35-39 44454 2 0.086 0.0000 0.158
40-44 38932 5 0.075 0.0001 0.158
45-49 66595 24 0.129 0.0004 0.158
50-54 83553 63 0.162 0.0008 0.132
55-59 83353 145 0.162 0.0017 0.105
60-64 65947 202 0.128 0.0031 0.105
65-69 50805 181 0.098 0.0036 0.079
70-74 40282 160 0.078 0.0040 0.053
75-79 25545 114 0.049 0.0045 0.026
80-84 12431 43 0.024 0.0035 0.013
85+ 4183 9 0.008 0.0022 0.013
Totals: 516080 948 Adjusted Cases: 707.
Crude Rate: 0.001
Adjusted Rate: 0.001
95% Conf. Interval: [0.0013, 0.001
-> Year= 1989
Stratum Pop. Cases Pop. Dist. Stratum Rate[s] Std. Pop. Dst[P]
35-39 78862 3 0.153 0.0000 0.158
40-44 74485 15 0.144 0.0002 0.158
45-49 96516 49 0.187 0.0005 0.158
50-54 78693 64 0.153 0.0008 0.132
55-59 48942 88 0.095 0.0018 0.105 0.0002
60-64 38789 83 0.075 0.0021 0.105 0.0002
65-69 29128 125 0.056 0.0043 0.079 0.0003
70-74 19168 86 0.037 0.0045 0.053 0.0002
75-79 25109 126 0.049 0.0050 0.026 0.0001
80-84 17417 113 0.034 0.0065 0.013 0.0001
85+ 8821 54 0.017 0.0061 0.013 0.0001
Totals: 515930 806 Adjusted Cases: 780.3
Crude Rate: 0.0016
Adjusted Rate: 0.0015
95% Conf. Interval: [0.0014, 0.0016]
Summary of Study Populations:
Year N Crude Adj_Rate Confidence Interval
1960 516080 0.001837 0.001371 [0.001281, 0.001461]
1989 515930 0.001562 0.001512 [0.001400, 0.001625]
dstdize - Direct standardization
Main
if/in
Options
Characteristic variable:
death
Population variable:
world
Strata variables:
age
Grouping variables:
Year
Use standard population from data in memory
Use standard population from Stata dataset:
E:\IntHealth\Kurs2009\Wednesday_Week1\reference.dta
Browse...
Use standard population from a value of grouping variable:
Value:
Grouping variable:
95
Confidence level
OK
Cancel
Submit
Command

Lecture 4: Indirect standardization with examples in Stata

Fazil Baksh

Department of Mathematics and Statistics
University of Reading, UK

Summer School - May/June 2011
Çeşme

Indirect standardization

Calculating the rate in STATA

Direct Standardization: age-specific health related event (e.g. disease, death) rates in **study** population are applied to the **reference** population

Indirect Standardization: age-specific rates in **reference** population are applied to the **study** population

Typically used when:

1. Age-specific rates are unavailable for the study population
   - direct standardization is not possible
2. We have a small number of events in the study population and age-specific rates are not stable
   - indirection standardization based on rates from a larger population provides a more precise estimate

Data required:

- Size of the study population in each age group
- Observed total number of events in the study population
- Age-specific event rates in a reference (standard) population

Choosing a reference population:

- the reference population should be similar to the years of available data for the study population.
- For example, to calculate a standardized mortality rate for London in 1989, the reference population could be the 1989 mortality rate of the UK.

The standardized mortality ratio (*SMR*):

	study population			reference population		
age-group	**deaths**	**at risk**	**rate**	**deaths**	**at risk**	**rate**
1	d_1	n_1	p_1	D_1	N_1	ρ_1
2	d_2	n_2	p_2	D_2	N_2	ρ_2
...	...	...	...	...	...	...
k	d_k	n_k	p_k	D_k	N_k	ρ_k
total	d	n	p	D	N	ρ

The **expected** number of deaths in the **study** population is:

$$E = \sum_{i=1}^{k} n_i \rho_i = \sum_{i=1}^{k} n_i \frac{D_i}{N_i}$$

$$SMR = \frac{\text{observed number}}{\text{expected number}} = \frac{d}{E}$$

Assuming a Poisson distribution for the observed number of deaths d, the **standard error** is

$$se(SMR) = \frac{\sqrt{d}}{E}$$

- SMR is often multiplied by 100 for presentation purposes
- A value of SMR less than 100 indicate a study population with mortality **less** than the reference, allowing for age differentials.
- Above 100 means a rate above the reference.

If the health related event in **NOT** death, this ratio is called the standardized incidence ratio (SIR).

The **indirect standardized mortality rate** is

$$R_{IDS} = SMR \times \rho = SMR \times \frac{D}{N}$$

Expressed per $1,000$ people, this rate is

$$1000 \times SMR \times \frac{D}{N}$$

With standard error

$$1000 \times \frac{D}{N} \times \frac{\sqrt{d}}{E}$$

Comparing Mortality from Lung Cancer in Berlin (West) 1960 and 1989

age-group	deaths 1989	at risk	deaths 1960	at risk
35-39	3	78862	2	44454
40-44	15	74485	5	38932
45-49	49	96516	24	66595
50-54	64	78693	63	83553
55-59	88	48942	145	83353
60-64	83	38789	202	65947
65-69	125	29128	181	50805
70-74	86	19168	160	40282
75-79	126	25109	114	25545
80-84	113	17417	43	12431
85+	54	8821	9	4183
total	**806**	**515930**	**948**	**516080**

Lung Cancer in Berlin (West) 1960 and 1989

To illustrate the calculation, we use 1960 as reference:

$$E = \sum_{i=1}^{k} n_i \frac{D_i}{N_i} = (78862 \times \frac{2}{44454}) + \ldots + (8821 \times \frac{9}{4183}) = 682.3731$$

So the standardized mortality ratio is

$$SMR = \frac{806}{682.3731} = 1.181$$

with standard error $\frac{\sqrt{806}}{682.3731} = 0.0416$

- Lung cancer mortality in 1989 is thus around 118% that in 1960.

Lung Cancer in Berlin (West) 1960 and 1989

Using the SMR we obtain the indirect standardized rate (per 1000 persons),

$$R_{IDS} = 1000 \times SMR \times \frac{D}{N} = 1000 \times 1.181 \times \frac{948}{516080} = 2.17$$

with standard error

$$1000 \times \frac{948}{516080} \times \frac{\sqrt{806}}{682.3731} = 0.0764$$

- The age adjusted lung cancer mortality rate for 1989 is 2.17 the rate in 1960.

In STATA

Data files needed:

(1) A study population file containing

- the strata variable (age) and the study size for each strata
- the **total** number of events observed
- if necessary, a group variable containing the groups to be compared

(2) A reference population file containing

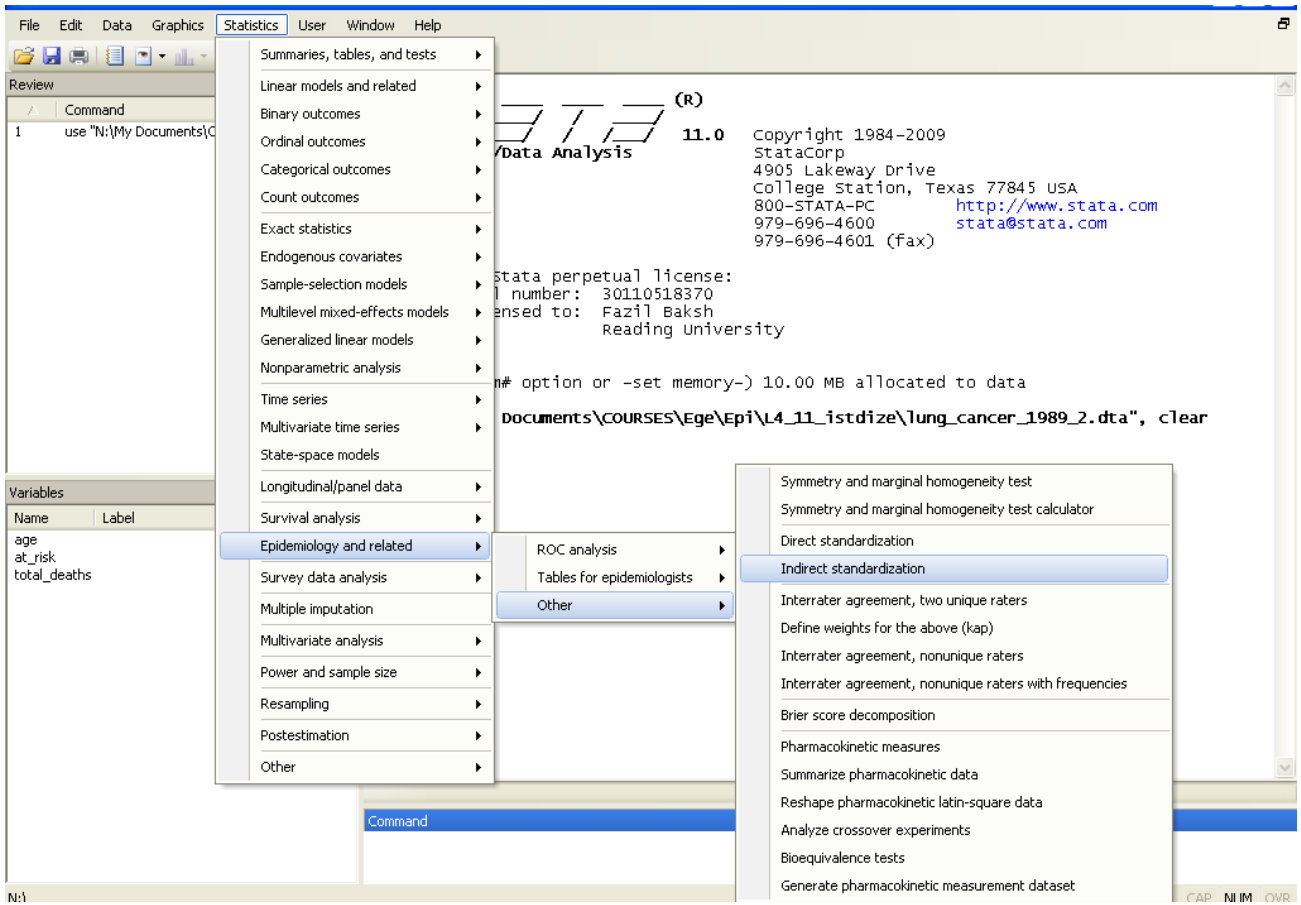
- the strata variable (age) exactly as in study population file
- Age-specific number of events and population size (or age-specific rates)

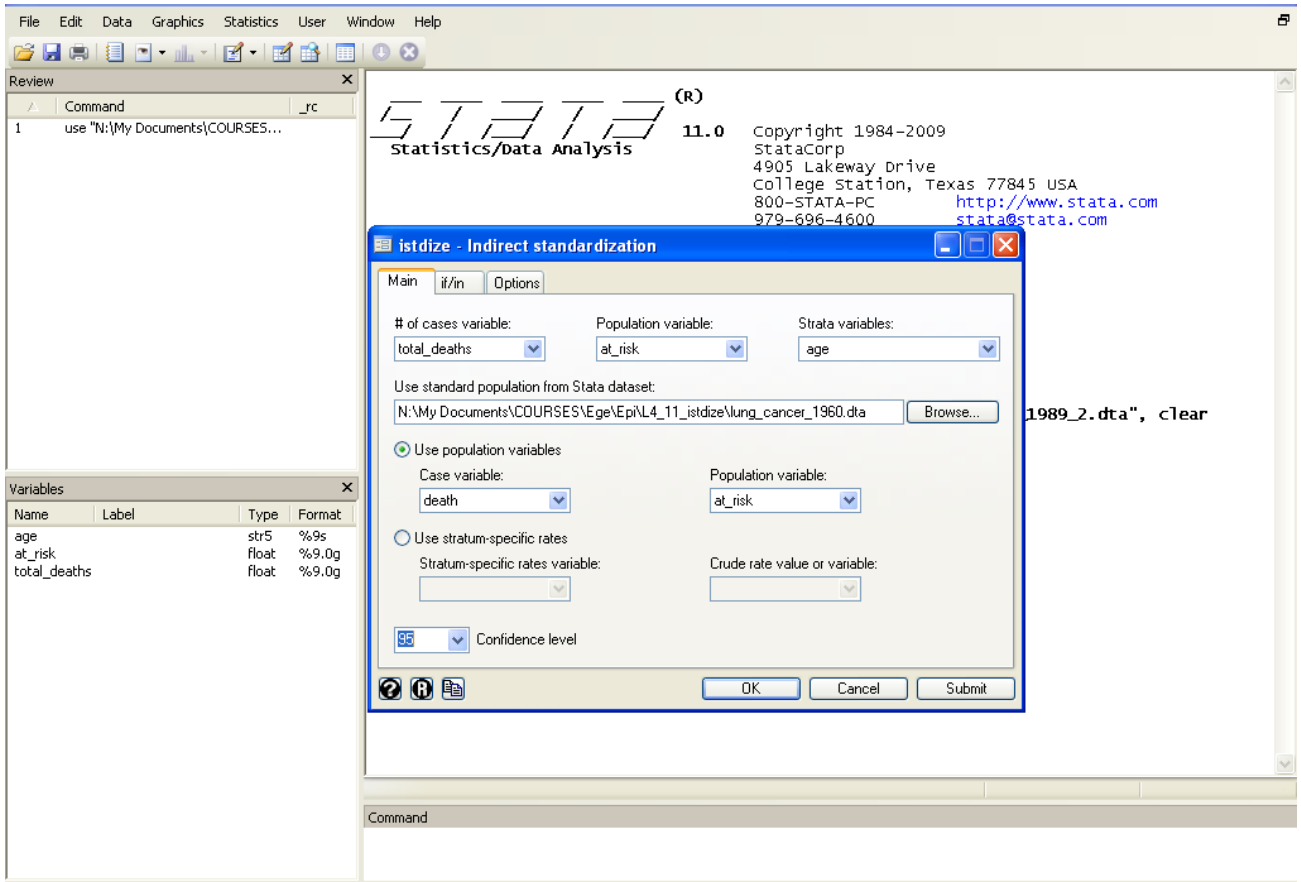
Study population file:

age	at_risk	total_~s
35-39	78862	806
40-44	74485	.
45-49	96516	.
50-54	78693	.
55-59	48942	.
60-64	38789	.
65-69	29128	.
70-74	19168	.
75-79	25109	.
80-84	17417	.
85+	8821	.

Reference population file:

age	death	at_risk
35-39	2	44454
40-44	5	38932
45-49	24	66595
50-54	63	83553
55-59	145	83353
60-64	202	65947
65-69	181	50805
70-74	160	40282
75-79	114	25545
80-84	43	12431
85+	9	4183


File Edit Data Graphics Statistics User Window Help
Review
Command
1 use "N:\My Documents\C
Variables
Name Label
age
at_risk
total_deaths
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Exact statistics
Endogenous covariates
Sample-selection models
Multilevel mixed-effects models
Generalized linear models
Nonparametric analysis
Time series
Multivariate time series
State-space models
Longitudinal/panel data
Survival analysis
Epidemiology and related
Survey data analysis
Multiple imputation
Multivariate analysis
Power and sample size
Resampling
Postestimation
Other
ROC analysis
Tables for epidemiologists
Other
Symmetry and marginal homogeneity test
Symmetry and marginal homogeneity test calculator
Direct standardization
Indirect standardization
Interrater agreement, two unique raters
Define weights for the above (kap)
Interrater agreement, nonunique raters
Interrater agreement, nonunique raters with frequencies
Brier score decomposition
Pharmacokinetic measures
Summarize pharmacokinetic data
Reshape pharmacokinetic latin-square data
Analyze crossover experiments
Bioequivalence tests
Generate pharmacokinetic measurement dataset
(R)
11.0
/Data Analysis
Copyright 1984-2009
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
stata perpetual license:
l number: 30110518370
ensed to: Fazil Baksh
Reading University
n# option or -set memory-) 10.00 MB allocated to data
Documents\COURSES\Ege\Epi\L4_11_istdize\lung_cancer_1989_2.dta", clear
Command
N:\


File Edit Data Graphics Statistics User Window Help
Review
Command _rc
1 use "N:\My Documents\COURSES...
Variables
Name Label Type Format
age str5 %9s
at_risk float %9.0g
total_deaths float %9.0g
Statistics/Data Analysis
11.0
Copyright 1984-2009
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
istdize - Indirect standardization
Main if/in Options
# of cases variable:
total_deaths
Population variable:
at_risk
Strata variables:
age
Use standard population from Stata dataset:
N:\My Documents\COURSES\Ege\Epi\L4_11_istdize\lung_cancer_1960.dta
Browse...
Use population variables
Case variable:
death
Population variable:
at_risk
Use stratum-specific rates
Stratum-specific rates variable:
Crude rate value or variable:
95
Confidence level
OK
Cancel
Submit
1989_2.dta", clear
Command

Stratum	Population Rate	Observed Population	Cases Expected
35-39	0.0000	78862	3.55
40-44	0.0001	74485	9.57
45-49	0.0004	96516	34.78
50-54	0.0008	78693	59.34
55-59	0.0017	48942	85.14
60-64	0.0031	38789	118.81
65-69	0.0036	29128	103.77
70-74	0.0040	19168	76.14
75-79	0.0045	25109	112.05
80-84	0.0035	17417	60.25
85+	0.0022	8821	18.98
Totals:		515930	682.37

```
                                    Observed Cases:      806
                                     SMR (Obs/Exp):     1.18
                  SMR exact 95% Conf. Interval: [1.1010, 1.2656]
                                        Crude Rate:   0.0016
                                     Adjusted Rate:   0.0022
                                95% Conf. Interval: [0.0020, 0.0023]
Summary of Study Populations (Rates):
    Observed     Crude    Adj_Rate  Confidence Interval
------------------------------------------------------------------------------
        806  0.001562   0.002170  [0.002023, 0.002325]

Summary of Study Populations (SMR):
    Observed     Expected    SMR     Confidence Interval
------------------------------------------------------------------------------
        806       682.37   1.181    [1.101024, 1.265611]
```

Lecture 5: Measures of effect I Risk Difference and Attributable Fraction with examples in Stata

Fazil Baksh

Department of Mathematics and Statistics
University of Reading, UK

Summer School - May/June 2011
Çeşme

Measures of differences in disease occurrence

Risk difference

Attributable Fraction

Calculating in STATA

We have seen earlier how to measure diseases and their distributions using prevalence and incidence.

Now we are concerned differences in disease occurrence in different populations.

Common measures are

1. risk difference (RD)
2. relative risk difference or attributable fraction (AF)
3. risk ratio (RR)
4. odds ratio (OR)

In this lecture we will look at the first two.

The risk ratio and odds ratio will be covered in the next lecture.

The **Risk Difference** (RD) is the difference between disease risk in an **exposed** population and risk in an **non-exposed** population.

Let p_1 = disease risk in an **exposed** population

p_0 = disease risk in an **non-exposed** population.

$$RD = p_1 - p_0$$

RD is a number between -1 and 1.

Example 1

In a study of two toothpastes, 10 out of 100 caries-free children using a new toothpaste (exposure) develop caries after 1 year. In another group of 100 caries-free children using a standard toothpaste, 25 develop caries.

$$\widehat{RD} = \frac{10}{100} - \frac{25}{100} = -0.15$$

Example 2

In a group of 1000 persons with heavy sun-exposure, there are 40 cases of skin cancer. In a comparative, equally sized, non-exposed group there are 10 cases of skin cancer.

$$\widehat{RD} = \frac{40}{1000} - \frac{10}{1000} = 0.03$$

Exercise 1

In a cohort study evaluating radiation exposures, 52 tumours developed among 2872 exposed individuals and 6 tumours developed among 5049 unexposed individuals within the observation period.
What is the risk difference?

$$\widehat{RD} = \hat{p}_1 - \hat{p}_0 =$$

Distribution of number of diseased

Suppose that in a cohort study,

Y_1 out of n_1 exposed individuals and

Y_0 out of n_0 non-exposed individuals

developed the disease.

Assume that the probability p_1 of developing the disease is the **same** for everyone in the exposed group

Similarly, assume that the probability p_0 of developing the disease is the **same** for everyone in the non-exposed group

Then $Y_1 \sim B(n_1, p_1)$ distribution

And $Y_0 \sim B(n_0, p_0)$ distribution

Variance of RD

A reasonable estimate for the RD is

$$\widehat{RD} = \hat{p}_1 - \hat{p}_0 = \frac{Y_1}{n_1} - \frac{Y_0}{n_0}$$

From which we get,

$$\begin{aligned} Var(\widehat{RD}) &= Var\left(\frac{Y_1}{n_1} - \frac{Y_0}{n_0}\right) \\ &= Var\left(\frac{Y_1}{n_1}\right) + Var\left(\frac{Y_0}{n_0}\right) \end{aligned}$$

and since both Y_1 and Y_2 follow binomial distributions,

$$Var(\widehat{RD}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}$$

A confidence interval for RD

$$SD(\widehat{RD}) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}}$$

Estimating p_1 and p_0 by $\hat{p}_1 = Y_1/n_1$ and $\hat{p}_0 = Y_0/n_0$

A 95% confidence interval for RD is

$$\widehat{RD} \pm 2SD(\widehat{RD})$$

$$= \widehat{RD} \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}})$$

Example 1 (revisited)

Here we had that 10 children out of 100 using a new toothpaste developed caries while 25 out of 100 using the standard toothpaste developed caries.
The estimated RD was shown to be $\widehat{RD} = \frac{10}{100} - \frac{25}{100} = -0.15$
A 95%CI for RD is $\widehat{RD} \pm 2SD(\widehat{RD})$

$$= \widehat{RD} \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}})$$

$$= -0.15 \pm 2\sqrt{\frac{0.1(1-0.1)}{100} + \frac{.25(1-0.25)}{100}})$$

$$= -0.15 \pm 2\sqrt{0.002775}$$

$$= -0.15 \pm 2 \times 0.0526783 = (-0.255, -0.045)$$

Exercise 1 (revisited)

Here we had a cohort study on radiation exposure where 52 tumours developed among 2872 exposed and 6 tumours developed among 5049 unexposed individuals.
The risk difference was $\widehat{RD} = \hat{p}_1 - \hat{p}_0 =$
A 95% CI for the risk difference is:

$$\widehat{RD} \pm 2\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}})$$

$$=$$

Interpretation:

Attributable Fraction (AF):

The attributable fraction (AF) or **relative risk difference** is a measure that **combines** RD and prevalence

AF due to exposure: Assume that exposure **increases** risk.

That is assume $p_1 > p_0$.

$$AF = \frac{RD}{p_1} = \frac{p_1 - p_0}{p_1}$$

interpretation: Let n be the total number of cases and controls

$$AF = \frac{np_1 - np_0}{np_1}$$

$$= \frac{(\#\text{ cases if everyone exposed}) - (\#\text{ cases if everyone non-exposed})}{\#\text{ cases if everyone exposed}}$$

$$AF = \text{ proportion of cases due to exposure}$$

$$= \text{ proportion of avoidable cases due to exposure}$$

AF is a relative measure:

Effects with similar risks will have similar attributable fractions.

Scenario A): $p_1 = 1/10$, $p_0 = 1/100$

$RD = 0.1 - 0.01 = 0.09 \sim 0.1$

$$AF = 0.09/0.1 = 0.90$$

Scenario B): $p_1 = 1/100$, $p_0 = 1/1000$

$RD = 0.01 - 0.001 = 0.009 \sim 0.01$

$$AF = 0.009/0.01 = 0.90$$

Preventive fraction

If exposure **decreases** risk the preventive fraction is instead calculated:

$$\frac{p_0 - p_1}{p_0}$$

Population attributable fraction (PAF)

This is the proportion of cases occurring in the total population which can be explained by the exposure

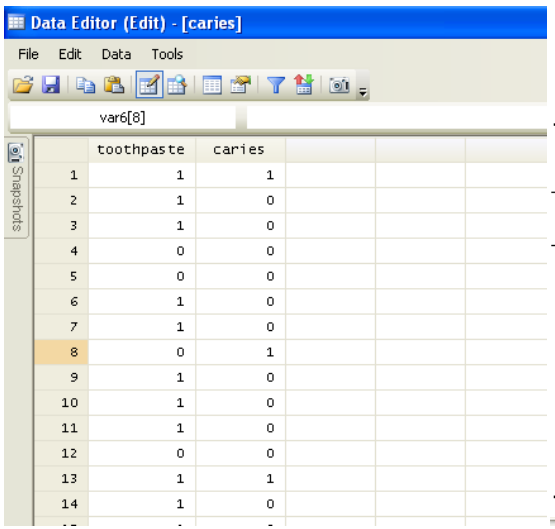
Let the proportion exposed be p

$$PAF = \frac{p(p_1 - p_0)}{pp_1 + (1-p)p_0}$$

In STATA

Example 1: Caries Study

Data in rectangular format:



	toothpaste	caries
1	1	1
2	1	0
3	1	0
4	0	0
5	0	0
6	1	0
7	1	0
8	0	1
9	1	0
10	1	0
11	1	0
12	0	0
13	1	1
14	1	0

```
. cs caries toothpaste

                    |  toothpaste
                    |    Exposed    Unexposed  |      Total
--------------------+-----------------------+-----------
              Cases |         10           25  |         35
           Noncases |         90           75  |        165
--------------------+-----------------------+-----------
              Total |        100          100  |        200
                    |                       |
               Risk |         .1           .25  |       .175
                    |                       |
                    |  Point estimate       |  [95% Conf. Interval]
--------------------+-----------------------+-----------
 Risk difference    |              -.15     |  -.2532475   -.0467525
      Risk ratio    |                .4     |   .2028594    .7887236
 Prev. frac. ex.    |                .6     |   .2112764    .7971406
 Prev. frac. pop    |                .3     |
--------------------+-----------------------+-----------
                              chi2(1) =  7.79  Pr>chi2 = 0.0052
.
```

csi 10 25 90 75

Lecture 6: Measures of effect II Risk Ratio and Odds Ratio with examples in Stata

Fazil Baksh

Department of Mathematics and Statistics
University of Reading, UK

Summer School - May/June 2011
Çeşme

Risk Ratio

Odds Ratio

Calculating in STATA

Risk ratio (RR):

The risk ratio or **relative risk** is the ratio of disease risk in an **exposed** to disease risk in an **non-exposed** population.

$$RR = \frac{p_1}{p_0}$$

where p_1 is disease risk in **exposed** and p_0 is disease risk in **non-exposed** population.

- RR is a number between 0 and ∞.

Interpretation:

For example, RR=2 means that disease occurrence is 2 times more likely in exposure group than in non-exposure group.

RR=1 means **no effect** of exposure.

Example 1

In a study of two toothpastes, 10 out of 100 caries-free children using a new toothpaste (exposure) develop caries after 1 year. In another group of 100 caries-free children using a standard toothpaste, 25 develop caries.

$$\widehat{RR} = \frac{10}{100} / \frac{25}{100} = 0.40$$

Example 2

In a group of 1000 persons with heavy sun-exposure, there are 40 cases of skin cancer. In a comparative, equally sized, non-exposed group there are 10 cases of skin cancer.

$$\widehat{RR} = \frac{40}{1000} / \frac{10}{1000} = 40$$

Exercise 1

In a cohort study evaluating radiation exposures, 52 tumours developed among 2872 exposed individuals and 6 tumours developed among 5049 unexposed individuals within the observation period.
What is the risk ratio?

$$\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0} =$$

Estimator of RR

Suppose that in a cohort study,

Y_1 out of n_1 exposed individuals and

Y_0 out of n_0 non-exposed individuals

developed the disease.

Assume that the probability p_1 of developing the disease is the **same** for everyone in the exposed group

Similarly, assume that the probability p_0 of developing the disease is the **same** for everyone in the non-exposed group

Then a plausible **estimator of the risk ratio** is

$$\widehat{RR} = \frac{\frac{Y_1}{n_1}}{\frac{Y_0}{n_0}} = \frac{Y_1 n_0}{Y_0 n_1}$$

Variance of RR

Technically it is easier to work with the logarithm of the risk ratio.

$$\log(RR) = \log(p_1) - \log(p_0)$$

Applying the δ **method**, an approximate variance is

$$Var\left(\widehat{\log RR}\right) = \begin{pmatrix} \frac{1}{p_1} & \frac{1}{p_0} \end{pmatrix} \begin{pmatrix} Var(\hat{p}_1) & 0 \\ 0 & Var(\hat{p}_0) \end{pmatrix} \begin{pmatrix} \frac{1}{p_1} \\ \frac{1}{p_0} \end{pmatrix}$$

$$= \frac{1}{p_1^2}\frac{p_1(1-p_1)}{n_1} + \frac{1}{p_0^2}\frac{p_0(1-p_0)}{n_0}$$

Estimating p_1 by Y_1/n_1 and p_0 by Y_0/n_0 and simplifying, we get

$$Var\left(\widehat{\log RR}\right) = \frac{1}{Y_1} - \frac{1}{n_1} + \frac{1}{Y_0} - \frac{1}{n_0}$$

A confidence interval for RR

$$SD(\widehat{\log RR}) = \sqrt{\frac{1}{Y_1} - \frac{1}{n_1} + \frac{1}{Y_0} - \frac{1}{n_0}}$$

Consequently, a 95% confidence interval for the **log relative risk** is

$$\widehat{\log RR} \pm 2SD(\widehat{\log RR})$$

$$= \widehat{\log RR} \pm 2\sqrt{\frac{1}{Y_1} - \frac{1}{n_1} + \frac{1}{Y_0} - \frac{1}{n_0}}$$

and back on the **relative risk scale**, a 95% CI for RR is

$$\exp\left(\widehat{\log RR} \pm 2\sqrt{\frac{1}{Y_1} - \frac{1}{n_1} + \frac{1}{Y_0} - \frac{1}{n_0}}\right)$$

Example 1 (revisited)

Here we had that 10 children out of 100 using a new toothpaste developed caries while 25 out of 100 using the standard toothpaste developed caries.
The estimated RR was shown to be

$$\widehat{RR} = \frac{10}{100} / \frac{25}{100} = 0.4$$

A $95\% CI$ for $\log(RR)$ is

$$\widehat{\log RR} \pm 2\sqrt{\frac{1}{Y_1} - \frac{1}{n_1} + \frac{1}{Y_0} - \frac{1}{n_0}}$$

$$= \log 0.4 \pm 2\sqrt{\frac{1}{10} - \frac{1}{100} + \frac{1}{25} - \frac{1}{100}}$$

$$= -0.92 \pm 2\sqrt{0.12}$$

$$= -0.92 \pm 2 \times 0.3464 = (-1.6128, -0.2272)$$

Hence a 95%CI for the **risk ratio** is

$$(\exp(-1.6128), \exp(-0.2272)) = (0.1993, 0.7968)$$

This shows that the new toothpaste **significantly** reduces the risk of developing caries.

Exercise 1 (revisited)

Here we had a cohort study on radiation exposure where 52 tumours developed among 2872 exposed and 6 tumours developed among 5049 unexposed individuals.
The risk ratio was $\widehat{RR} = \frac{\hat{p}_1}{\hat{p}_0}$

A 95% CI for RR is:

Interpretation:

AF and RR:

Assume that $p_1 > p_0$:

$$\begin{aligned} AF = RD/p_1 &= \frac{p_1 - p_0}{p_1} \\ &= 1 - \frac{p_0}{p_1} \\ &= 1 - \frac{1}{RR} \end{aligned}$$

Hence an **estimate of AF is available if an estimate of RR** is available.

Odds

The odds of an outcome is the number of times the outcome occurs to the number of times it does not.

Suppose that p is the probability of the outcome, then

$$odds = \frac{p}{1-p}$$

It follows that $p = \frac{odds}{odds+1}$

Examples

- $p = 1/2 \Rightarrow odds = 1$
- $p = 1/4 \Rightarrow odds = 1/3$
- $p = 3/4 \Rightarrow odds = 3/1 = 3$

Odds Ratio

$$OR = \frac{odds(\text{ in exposure })}{odds(\text{ in non-exposure })}$$

$$= \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

Properties of Odds Ratio

- $0 < OR < \infty$
- $OR = 1$ if and only if $p_1 = p_0$

Examples

$$\text{risk} = \begin{cases} p_1 = 1/4 \\ p_0 = 1/8 \end{cases} \qquad \text{effect measure} = \begin{cases} OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{1/3}{1/7} = 2.33 \\ RR = \frac{p_1}{p_0} = 2 \end{cases}$$

$$\text{risk} = \begin{cases} p_1 = 1/100 \\ p_0 = 1/1000 \end{cases} \qquad \text{eff. meas.} = \begin{cases} OR = \frac{1/99}{1/999} = 10.09 \\ RR = \frac{p_1}{p_0} = 10 \end{cases}$$

Fundamental Theorem of Epidemiology

$$p_0 \text{ small } \Rightarrow OR \approx RR$$

benefit: OR is interpretable as RR which is easier to deal with

Example: Radiation Exposure and Tumor Development

	cases	non-cases	
E	52	2820	2872
NE	6	5043	5049

odds and *OR*

odds for disease given exposure:

$$\frac{52/2872}{2820/2872} = 52/2820$$

odds for disease given non-exposure:

$$\frac{6/5049}{5043/5049} = 6/5043$$

Example, cont'd

	cases	non-cases	
E	52	2820	2872
NE	6	5043	5049

odds ratio for disease :

$$OR = \frac{52/2820}{6/5043} = \frac{52 \times 5043}{6 \times 2820} = 15.49$$

or, $\log OR = \log 15.49 = 2.74$
for comparison

$$RR = \frac{52/2872}{6/5049} = 15.24$$

	cases	non-cases
E	a	b
NE	c	d

$$OR = \frac{a/b}{c/d} = \frac{ad}{bc}$$

CI for OR: Using

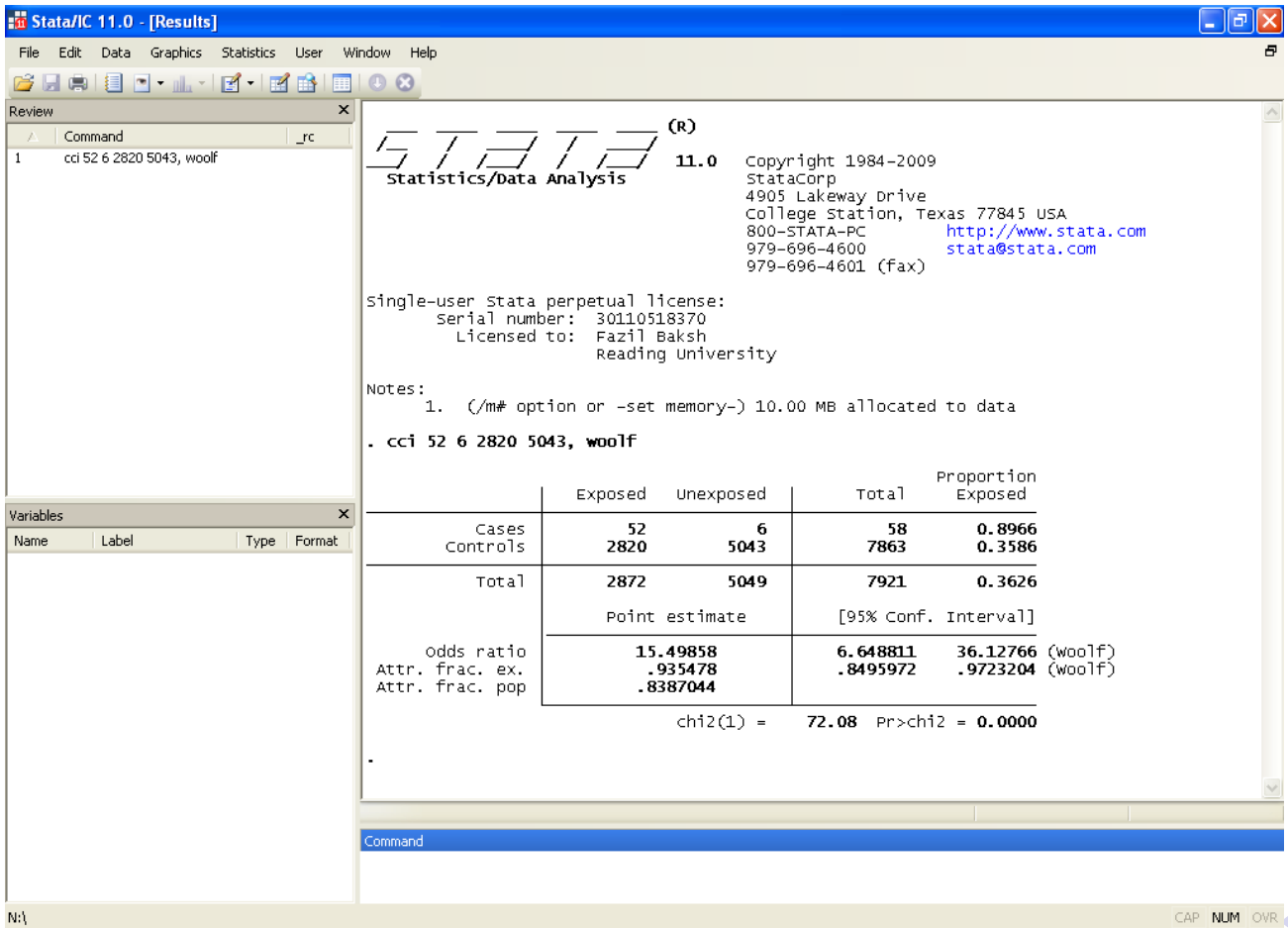
$$Var(\log OR) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

A 95% CI for log OR is $\log OR \pm 2\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

As for RR, the exponent of these limits will provide the CI for OR

In STATA

Example: Radiation Exposure and Tumor Development




```
  ___  ____  ____  ____  ____ (R)
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   11.0   Copyright 1984-2009
  Statistics/Data Analysis             StataCorp
                                       4905 Lakeway Drive
                                       College Station, Texas 77845 USA
                                       800-STATA-PC        http://www.stata.com
                                       979-696-4600        stata@stata.com
                                       979-696-4601 (fax)

Single-user Stata perpetual license:
       Serial number:  30110518370
         Licensed to:  Fazil Baksh
                       Reading University

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data

. cci 52 6 2820 5043, woolf

                                                                  Proportion
                    |  Exposed    Unexposed  |      Total      Exposed
----------------------------------------------------------------------------
              Cases |         52           6  |         58       0.8966
           Controls |       2820        5043  |       7863       0.3586
----------------------------------------------------------------------------
              Total |       2872        5049  |       7921       0.3626
                    |                         |
                    |  Point estimate         |  [95% Conf. Interval]
----------------------------------------------------------------------------
         Odds ratio |      15.49858           |   6.648811   36.12766 (woolf)
 Attr. frac. ex.    |       .935478           |   .8495972   .9723204 (woolf)
 Attr. frac. pop    |       .8387044          |
----------------------------------------------------------------------------
                                        chi2(1) =    72.08  Pr>chi2 = 0.0000

.
```

Confounding and effect modification: Mantel-Haenszel estimation, testing effect homogeneity

Dankmar Böhning

Department of Mathematics and Statistics
University of Reading, UK

Summer School in Cesme, May/June 2011

Overview

1. Cohort Studies with *Similar* Observation Time

2. Cohort Studies with *Individual*, Different Observation Time

3. Case-Control Studies: *Unmatched* Situation

4. Case-Control Studies: *Matched* Situation

1. Cohort Studies with *Similar* Observation Time

Situation in the population:

	Case	Non-Case	
Exposed	p_1	$1-p_1$	
Non-exposed	p_0	$1-p_0$	

interest in: $RR = \frac{p_1}{p_0}$

Situation in the sample:

	Case	Non-Case	At Risk
Exposed	Y_1	$n_1 - Y_1$	n_1
Non-exposed	Y_0	$n_0 - Y_0$	n_0

Interest in estimating $RR = \frac{p_1}{p_0}$:

$$\hat{RR} = \frac{Y_1/n_1}{Y_0/n_0}$$

Example: Radiation Exposure and Cancer Occurrence

	Case	Non-Case	At Risk
Exposed	52	2820	2872
Non-exposed	6	5043	5049

$$\hat{RR} = \frac{52/2872}{6/5049} = \frac{0.0181}{0.0012} = 15.24$$

Tests and Confidence Intervals

Estimated Variance of $\log(\hat{RR})$:

$$\hat{Var}(\log \hat{RR}) = 1/Y_1 - 1/n_1 + 1/Y_0 - 1/n_0$$

Estimated Standard Error of $\log(\hat{RR})$:

$$\hat{SE}(\log \hat{RR}) = \sqrt{1/Y_1 - 1/n_1 + 1/Y_0 - 1/n_0}$$

For the above example:

$$\hat{Var}(\log \hat{RR}) = 1/52 - 1/2872 + 1/6 - 1/5049$$
$$= 0.1854$$

$$\hat{SE}(\log \hat{RR}) = 0.4305$$

Testing

H_0: RR= 1 or log(RR) = 0

H_1: H_0 is false

Statistic used for testing: $Z = \log(\hat{RR}) / \hat{SE}(\log \hat{RR})$

Z is approx. standard normally distributed if H_0 true

Test with Significance level 5%:

reject H_0 if $|Z| > 1.96$
accept H_0 if $|Z| \leq 1.96$

For the example: $Z = \log(15.24)/0.4305 = 6.327$

Confidence Interval

95%-CI covers with 95% confidence the true log (RR):

$$\log(\hat{RR}) \pm 1.96\ \hat{SE}(\log \hat{RR})$$

For the example:

$$\log(15.24) \pm 1.96 \times 0.4305 = (1.8801, 3.5677)$$

and back to the relative risk – scale:

$$(\exp(1.8801), \exp(3.5677)) = (6.55, 35.43)$$

In STATA

	Exposed	Unexposed	Total	
Cases	**52**	**6**	**58**	
Noncases	**2820**	**5043**	**7863**	
Total	**2872**	**5049**	**7921**	
Risk	**.0181058**	**.0011884**	**.0073223**	
	Point estimate		[95% Conf.	Interval]
Risk difference	**.0169175**		**.0119494**	**.0218856**
Risk ratio	**15.23607**		**6.552546**	**35.42713**
Attr. frac. ex.	**.9343663**		**.8473876**	**.971773**
Attr. frac. pop	**.8377077**			
	chi2(1) =	**72.08**	Pr>chi2 =	**0.0000**

Potential Confounding

and Stratification with Respect to the Confounder

Situation:

	Exposed		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	RR
1	50	100	1500	3000	1
2	10	1000	1	100	1
Total	60	1100	1501	3100	0.1585

Explanation?

A more realistic example: *Drinking Coffee and CHD*

	Exposed *(coffee)*		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	RR
Smoker	195	705	21	79	1.03
Non-S	5	95	29	871	1.55
Total	200	800	50	950	4

How to diagnose confounding? Stratify !

Situation:

	Exposed		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	RR
1	$Y_1^{(1)}$	$n_1^{(1)} - Y_1^{(1)}$	$Y_0^{(1)}$	$n_0^{(1)} - Y_0^{(1)}$	$RR^{(1)}$
2	$Y_1^{(2)}$	$n_1^{(2)} - Y_1^{(2)}$	$Y_0^{(2)}$	$n_1^{(2)} - Y_0^{(2)}$	$RR^{(2)}$
…		…		…	
k	$Y_1^{(k)}$	$n_1^{(k)} - Y_1^{(k)}$	$Y_0^{(k)}$	$n_1^{(k)} - Y_0^{(k)}$	$RR^{(k)}$
Total	Y_1	$n_1 - Y_1$	Y_0	$n_1 - Y_0$	RR

How should the RR be estimated?

Use an average of stratum-specific weights:

$$\hat{RR} = w_1\hat{RR}^{(1)} + \ldots + w_k\hat{RR}^{(k)}/(w_1+\ldots+w_k)$$

Which weights?

Mantel-Haenszel Approach

$$\hat{RR}_{MH} = \frac{Y_1^{(1)} n_0^{(1)}/n^{(1)} + \ldots + Y_1^{(k)} n_0^{(k)}/n^{(k)}}{Y_0^{(1)} n_1^{(1)}/n^{(1)} + \ldots + Y_0^{(k)} n_1^{(k)}/n^{(k)}}$$

with $n^{(i)} = n_0^{(i)} + n_1^{(i)}$.

Good Properties!

Mantel-Haenszel Weight: $w_i = Y_0^{(i)} n_1^{(i)}/n^{(i)}$

$$w_1 \hat{RR}^{(1)} + \ldots + w_k \hat{RR}^{(k)}/(w_1 + \ldots + w_k) = \hat{RR}_{MH}$$

Illustration of the MH-weights

	Exposed		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	w_i
1	50	100	1500	3000	1500*150/4650
2	10	1000	1	100	1*1010/1111

In STATA

	Stratum	Case	Exposure	obs
1.	1	1	1	50
2.	1	0	1	100
3.	1	1	0	1500
4.	1	0	0	3000
5.	2	1	1	10
6.	2	0	1	1000
7.	2	1	0	1
8.	2	0	0	100

Stratum	RR	[95% Conf.	Interval]	M-H Weight
1	1	.7944874	1.258673	48.3871
2	1	.1293251	7.732451	.9090909
Crude	.1585495	.123494	.2035559	
M-H combined	1	.7953728	1.257272	

Test of homogeneity (M-H) chi2(1) = 0.000 Pr>chi2 = 1.0000

Illustration: Coffee-CHD-Data

	Case	Exposure	Smoking	freque~y
1.	1	0	1	21
2.	0	0	1	79
3.	1	1	1	195
4.	0	1	1	705
5.	1	0	2	29
6.	0	0	2	871
7.	1	1	2	5
8.	0	1	2	95

Smoking	RR	[95% Conf.	Interval]	M-H Weight
1	1.031746	.6916489	1.539076	18.9
2	1.551724	.6144943	3.918422	2.9
Crude	4	2.971453	5.384571	
M-H combined	1.100917	.7633712	1.587719	

Test of homogeneity (M-H) chi2(1) = 0.629 Pr>chi2 = 0.4279

Inflation, Masking and Effect Modification

Inflation (Confounding): Crude RR is larger (in absolute value) than stratified RR

Masking (Confounding): Crude RR is smaller (in absolute value) than stratified RR

Effect Modification: Crude Rate is in between stratified RR

How can these situations be diagnosed?

Use *heterogeneity or homogeneity* test:

Homogeneity Hypothesis

H_0: $RR^{(1)} = RR^{(2)} = \ldots = RR^{(k)}$

H_1: H_0 is wrong

Teststatistic:

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} (\log \widehat{RR}^{(i)} - \log RR_{MH})^2 / \text{Var}(\log \widehat{RR}^{(i)})$$

Illustration of the Heterogeneity Test for CHD-Coffee

	Exposed		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	χ^2
Smoke	195	705	21	79	0.1011
Non-Smoke	5	95	29	871	0.5274
Total	200	800	50	950	0.6285

Smoking	RR	[95% Conf.	Interval]	M-H Weight
1	**1.031746**	**.6916489**	**1.539076**	**18.9**
2	**1.551724**	**.6144943**	**3.918422**	**2.9**
Crude	**4**	**2.971453**	**5.384571**	
M-H combined	**1.100917**	**.7633712**	**1.587719**	

Test of homogeneity (M-H) chi2(**1**) = **0.629** Pr>chi2 = **0.4279**

.

Cohort Studies with *Individual, different* Observation Time

Situation:

	Event-Risk	Person-Time	At Risk
Exposed	p_1	T_1	n_1
Non-exposed	p_0	T_0	n_0

Definition: Person-Time is the time that n persons spend under risk in the study period

Interest in: $RR = p_1/p_0$

Situation:

	Events	Person-Time	At Risk
Exposed	Y_1	T_1	n_1
Non-exposed	Y_0	T_0	n_0

$$\hat{RR} = \frac{Y_1/T_1}{Y_0/T_0}$$

Y/T is also called the *incidence density* (ID) !

Example: Smoking Exposure and CHD Occurrence

	Events	Person-Time	ID (Events per 10,000 PYs)
Exposed	206	28612	72
Non-exposed	28	5710	49

$$\hat{RR} = \frac{206/28612}{28/5710} = \frac{72}{49} = 1.47$$

Tests and Confidence Intervals

Estimated Variance of $\log(\hat{RR}) = \log(\hat{ID}_1 / \hat{ID}_0)$:

$$\hat{Var}(\log \hat{RR}) = 1/Y_1 + 1/Y_0$$

Estimated Standard Error of $\log(\hat{RR})$:

$$\hat{SE}(\log \hat{RR}) = \sqrt{1/Y_1 + 1/Y_0}$$

For the above example:

$\hat{Var}(\log \hat{RR}) = 1/206 + 1/28 = 0.0405$

$\hat{SE}(\log \hat{RR}) = 0.2013$

Testing

H_0: RR= 1 or log(RR) = 0

H_1: H_0 is false

Statistic used for testing: $Z = \log(\hat{RR}) / \hat{SE}(\log \hat{RR})$

Z is approx. normally distributed if H_0 true:

Test with Significance level 5%:

reject H_0 if $|Z| > 1.96$

accept H_0 if $|Z| \leq 1.96$

For the example: $Z = \log(1.47)/0.2013 = 1.9139$

Confidence Interval

95%-CI covers with 95% confidence the true log (RR):

$$\log(\hat{RR}) \pm 1.96\ \hat{SE}(\log \hat{RR})$$

For the example:

$$\log(1.47) \pm 1.96\ 0.2013 = (-0.0093, 0.7798)$$

and back to the relative risk – scale:

$$(\exp(-0.0093), \exp(0.7798)) = (0.99, 2.18)$$

In STATA

	Exposed	Unexposed	Total		
Cases	**206**	**28**	**234**		
Person-time	**28612**	**5710**	**34322**		
Incidence Rate	**.0071998**	**.0049037**	**.0068178**		
	Point estimate		[95% Conf.	Interval]	
Inc. rate diff.	**.0022961**		**.0002308**	**.0043614**	
Inc. rate ratio	**1.46824**		**.9863624**	**2.264107**	(exact)
Attr. frac. ex.	**.3189125**		**-.0138261**	**.5583247**	(exact)
Attr. frac. pop	**.280752**				
	(midp) Pr(k>=206) =			**0.0243**	(exact)
	(midp) 2*Pr(k>=206) =			**0.0487**	(exact)

.

Stratification with Respect to a Potential Confounder

Example: *energy intake (as surrogate measure for physical inactivity) and Ischaemic Heart Disease*

	Exposed *(<2750 kcal)*		Non-Exposed *(≥2750 kcal)*		
Stratum	Cases	P-Time	Cases	P-Time	RR
40-49	2	311.9	4	607.9	0.97
50-59	12	878.1	5	1272.1	3.48
60-60	14	667.5	8	888.9	2.33
Total	28	1857.5	17	2768.9	2.46

Situation:

	Exposed		Non-Exposed		
Stratum	Cases	P-Time	Cases	P-Time	RR
1	$Y_1^{(1)}$	$T_1^{(1)}$	$Y_0^{(1)}$	$T_0^{(1)}$	$RR^{(1)}$
2	$Y_1^{(2)}$	$T_1^{(2)}$	$Y_0^{(2)}$	$T_0^{(2)}$	$RR^{(2)}$
…		…		…	
k	$Y_1^{(k)}$	$T_1^{(k)}$	$Y_0^{(k)}$	$T_0^{(k)}$	$RR^{(k)}$
Total	Y_1	T_1	Y_0	T_0	RR

How should the RR be estimated?

Use an average of stratum-specific weights:

$$\hat{RR} = w_1{}^{(1)} + \ldots + w_k \hat{RR}^{(k)}/(w_1+\ldots+w_k)$$

Which weights?

Mantel-Haenszel Approach

$$\hat{RR}_{MH} = \frac{Y_1^{(1)}T_0^{(1)}/T^{(1)} + \ldots + Y_1^{(k)}T_0^{(k)}/T^{(k)}}{Y_0^{(1)}T_1^{(1)}/T^{(1)} + \ldots + Y_0^{(k)}T_1^{(k)}/T^{(k)}}$$

with $T^{(i)} = T_0^{(i)} + T_1^{(i)}$.

Mantel-Haensel Weight: $w_i = Y_0^{(i)}T_1^{(i)}/T^{(i)}$

$$w_1\hat{RR}^{(1)} + \ldots + w_k \hat{RR}^{(k)}/(w_1+\ldots+w_k) = \hat{RR}_{MH}$$

In STATA

	Stratum	Exposure	number~e	Person~e
1.	1	1	2	311.9
2.	1	0	4	607.9
3.	2	1	12	878.1
4.	2	0	5	1272.1
5.	3	1	14	667.5
6.	3	0	8	888.9

Stratum	IRR	[95% Conf.	Interval]	M-H Weight	
1	.9745111	.0881524	6.799694	1.356382	(exact)
2	3.476871	1.14019	12.59783	2.041903	(exact)
3	2.33045	.9123878	6.411597	3.430995	(exact)
Crude	2.455204	1.297757	4.781095		(exact)
M-H combined	2.403914	1.306881	4.421829		

Test of homogeneity (M-H) chi2(2) = 1.57 Pr>chi2 = 0.4555

2. Case-Control Studies: *Unmatched* Situation

Situation:

	Case	Controls
Exposed	q_1	q_0
Non-exposed	$1-q_1$	$1-q_0$

Interest is in: RR = p_1/p_0 which is *not* estimable
not in $RR_e = q_1/q_0$

Illustration with a Hypo-Population:

	Bladder-Ca	Healthy	
Smoking	500	199,500	200,000
Non-smoke	500	799,500	800,000
	1000	999,000	1,000,000

$$RR = p_1/p_0 = 4$$

$$\neq 2.504 = \frac{5/10}{1995/9990} = q_1/q_0 = RR_e$$

However, consider the (disease) **Odds Ratio** defined as

$$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

$Pr(D/E) = p_1$, $Pr(D/NE) = p_0$,

$Pr(E/D) = q_1$, $Pr(E/ND) = q_0$, $p = Pr(D)$

$p_1 = P(D/E)$ *using Bayes Theorem*

$$= \frac{Pr(E/D)Pr(D)}{Pr(E/D)Pr(D)+ Pr(E/ND)Pr(ND)} = \frac{q_1\, p}{q_1 p + q_0\,(1-p)}$$

$p_0 = P(D/NE)$

$$= \frac{Pr(NE/D)Pr(D)}{Pr(NE/D)Pr(D)+ Pr(NE/ND)Pr(ND)} = \frac{(1-q_1)\, p}{(1-q_1)\, p + (1-q_0\,)(1-p)}$$

$p_1/(1-p_1) = q_1p/q_0(1-p)$ und $p_0/(1-p_0) = [(1-q_1)p]/[(1-q_0)(1-p)]$.

it follows that

$$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{q_1/q_0}{(1-q_1)/(1-q_0)} = \frac{q_1/(1-q_1,)}{q_0/(1-q_0)} = OR_e$$

Disease Odds Ratio = Exposure Odds Ratio

Illustration with a Hypo-Population:

	Bladder-Ca	Healthy	
Smoking	500	199,500	200,000
Non-smoke	500	799,500	800,000
	1000	999,000	1,000,000

OR = (500/199,500)/(500/799,500) = (500/500)/(199,500/799,500) = OR_e = 4.007

Also, if disease occurrence is low (low prevalence),

OR ≈ RR

Estimation of OR

Situation:

	Case	Controls
Exposed	X_1	X_0
Non-exposed	m_1-X_1	m_0-X_0
	m_1	m_0

$$\hat{OR} = \frac{\hat{q}_1/(1-\hat{q}_1)}{\hat{q}_0/(1-\hat{q}_0)} = \frac{X_1/(m_1-X_1)}{X_0/(m_0-X_0)} = \frac{X_1(m_0-X_0)}{X_0(m_1-X_1)}$$

Example: Sun Exposure and Lip Cancer Occurrence in Population of 50-69 year old men

	Case	Controls
Exposed	66	14
Non-exposed	27	15
	93	29

$$\hat{OR} = \frac{66 \times 15}{14 \times 27} = 2.619$$

Tests and Confidence Intervals

Estimated Variance of $\log(\hat{OR})$:

$$\hat{Var}(\log \hat{OR}) = \frac{1}{X_1} + \frac{1}{m_1 - X_1} + \frac{1}{X_0} + \frac{1}{m_0 - X_0}$$

Estimated Standard Error of $\log(\hat{OR})$:

$$\hat{SE}(\log \hat{OR}) = \sqrt{\frac{1}{X_1} + \frac{1}{m_1 - X_1} + \frac{1}{X_0} + \frac{1}{m_0 - X_0}}$$

For the above example:

$$\hat{Var}(\log \hat{OR}) = 1/66 + 1/27 + 1/14 + 1/15$$
$$= 0.1903$$

$$\hat{SE}(\log \hat{OR}) = 0.4362$$

Testing

H_0: OR= 1 or log(OR) = 0

H_1: H_0 is false

Statistic used for testing: $Z = \log(\hat{OR}) / \hat{SE}(\log \hat{OR})$

Z is approx. normally distributed if H_0 true:

Test with Significance level 5%:

reject H_0 if $|Z| > 1.96$

accept H_0 if $|Z| \leq 1.96$

For the example: Z = log(2.619)/0.4362 = 2.207

Confidence Interval

95%-CI covers with 95% confidence the true log (RR):

$$\log(\hat{OR}) \pm 1.96\ \hat{SE}(\log \hat{OR})$$

For the example:

$$\log(2.619) \pm 1.96\ 0.4362 = (0.1078, 1.8177)$$

and back to the relative risk – scale:

$$(\exp(0.1078), \exp(1.8177)) = (1.11, 6.16)$$

In STATA

	Exposed	Unexposed	Total	Proportion Exposed	
Cases	**66**	**27**	**93**	**0.7097**	
Controls	**14**	**14**	**28**	**0.5000**	
Total	**80**	**41**	**121**	**0.6612**	
	Point estimate		[95% Conf.	Interval]	
Odds ratio	**2.444444**		**1.028622**	**5.809044**	(Woolf)
Attr. frac. ex.	**.5909091**		**.0278254**	**.8278546**	(Woolf)
Attr. frac. pop	**.4193548**				
	chi2(1) =	**4.22**	Pr>chi2 =	**0.0399**	

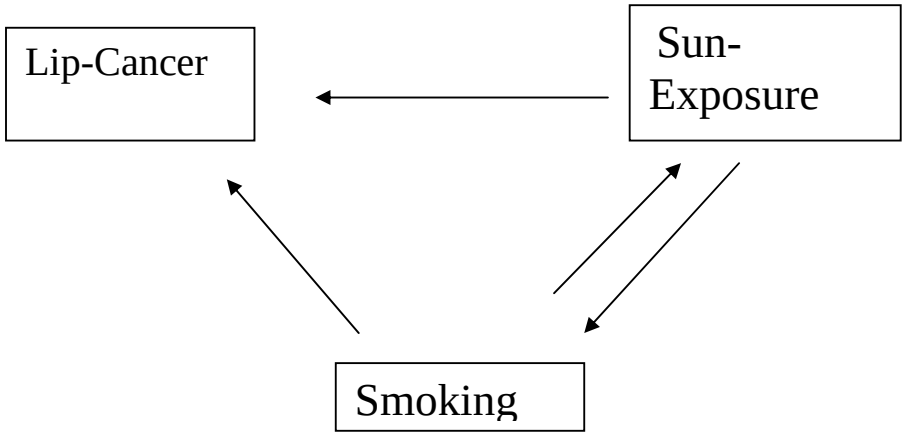
.

Exercise: A case-control study investigates if a keeping a pet bird is a risk factor: Cases: 98 Bird Owners, 141 None, Controls: 101 Bird Owners, 328 None

Potential Confounding

and Stratification with Respect to the Confounder

Situation:



Lip-Cancer and Sun Exposure with Smoking as Potential Confounder

	Cases		Controls		
Stratum	Exposed	Non-Exp.	Exp.	Non-Exp.	OR
Smoke	51	24	6	10	3.54
Non-Smoke	15	3	8	5	3.13
Total	66	27	14	15	2.62

Explanation?

How to diagnose confounding? Stratify !

Situation:

	Cases		Controls		Cases
Stra-tum	Ex-posed	Non-Exp.	Ex-posed	Non-Exp.	OR
1	$X_1^{(1)}$	$m_1^{(1)} - X_1^{(1)}$	$X_0^{(1)}$	$m_0^{(1)} - X_0^{(1)}$	$OR^{(1)}$
2	$X_1^{(2)}$	$m_1^{(2)} - X_1^{(2)}$	$X_0^{(2)}$	$m_1^{(2)} - X_0^{(2)}$	$OR^{(2)}$
…		…		…	
k	$X_1^{(k)}$	$m_1^{(k)} - X_1^{(k)}$	$X_0^{(k)}$	$m_1^{(k)} - X_0^{(k)}$	$OR^{(k)}$
Total	X_1	$m_1 - X_1$	X_0	$m_1 - X_0$	OR

How should the OR *based upon stratification* **be estimated?**

Use an average of stratum-specific weights:

$$\hat{OR} = w_1\hat{OR}^{(1)} + \ldots + w_k\hat{OR}^{(k)}/(w_1+\ldots+w_k)$$

Which weights?

Mantel-Haenszel Weight: $w_i = X_0^{(i)}(m_1^{(i)} - X_1^{(i)})/ m^{(i)}$

Mantel-Haenszel Approach

$$\hat{OR}_{MH} = \frac{X_1^{(1)}(m_0^{(1)} - X_0^{(1)})/m^{(1)} + \ldots + X_1^{(k)}(m_0^{(k)} - X_0^{(k)})/m^{(1)}}{X_0^{(1)}(m_1^{(1)} - X_1^{(1)})/ m^{(1)} + \ldots + X_1^{(1)}(m_0^{(1)} - X_0^{(1)})/ m^{(1)}}$$

with $m^{(i)} = m_0^{(i)} + m_1^{(i)}$.

$$w_1\hat{OR}^{(1)} + \ldots + w_k\hat{OR}^{(k)}/(w_1+\ldots+w_k) = \hat{OR}_{MH}$$

Illustration of the MH-weights

	Cases		Controls		
Stratum	Exposed	Non-Exp.	Exp.	Non-Exp.	w_i
Smoke	51	24	6	10	6*24/91
Non-Smoke	15	3	8	5	8*3/31

In STATA

	Case	Exposure	Smoke	Pop
1.	1	1	0	51
2.	0	1	0	6
3.	1	0	0	24
4.	0	0	0	10
5.	1	1	1	15
6.	0	1	1	8
7.	1	0	1	3
8.	0	0	1	5

. cc Case Control [freq=Pop], by(Smoke)

Smoke	OR	[95% Conf. Interval]	M-H Weight

```
----------------+-------------------------------------------------
           0 |   3.541667    1.011455   13.14962      1.582418 (exact)
           1 |      3.125      .4483337   24.66084     .7741935 (exact)
----------------+-------------------------------------------------
           Crude |   2.619048    1.016247    6.71724            (exact)
M-H combined |   3.404783    1.341535   8.641258
----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.01  Pr>chi2 = 0.9029

                Test that combined OR = 1:
Mantel-Haenszel chi2(1) =      6.96            Pr>chi2 =    0.0083
```

Note that “freq=Pop” is optional, e.g. raw data can be used with this analysis

Inflation, Masking and Effect Modification

Inflation (Confounding): Crude OR is larger (in absolute value) than stratified OR

Masking (Confounding): Crude OR is smaller (in absolute value) than stratified OR

Effect Modification: Crude Rate is in between stratified OR

How can these situations be diagnosed? Use *heterogeneity* or *homogeneity* test:

Homogeneity Hypothesis

H_0: $OR^{(1)} = OR^{(2)} = \ldots = OR^{(k)}$
H_1: H_0 is wrong

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} (\log \widehat{OR}^{(i)} - \log OR_{MH})^2 / \mathrm{Var}(\log \widehat{OR}^{(i)})$$

Illustration of the Heterogeneity Test for Lip Cancer -Sun Exposure

	Cases		Controls		
Stratum	Exposed	Non-Exp.	Exp.	Non-Exp.	χ^2
Smoke	51	24	6	10	0.0043
Non-Smoke	15	3	8	5	0.0101
Total	66	27	14	15	0.0144

	D	E	stratum	freq
1.	0	0	1	10
2.	0	1	2	8
3.	0	1	1	6
4.	1	0	1	24
5.	1	1	1	51
6.	1	0	2	3
7.	0	0	2	5
8.	1	1	2	15

stratum	OR	[95% Conf.	Interval]	M-H Weight	
1	3.541667	1.011455	13.14962	1.582418	(exact)
2	3.125	.4483337	24.66091	.7741935	(exact)
Crude	2.619048	1.016247	6.717228		(exact)
M-H combined	3.404783	1.341535	8.641258		

Test of homogeneity (M-H) chi2(1) = 0.01 Pr>chi2 = 0.9029

Test that combined OR = 1:
Mantel-Haenszel chi2(1) = 6.96
Pr>chi2 = 0.0083

3. Case-Control Studies: *Matched* Situation

Given a *case* is sampled, a *comparable* control is sampled: comparable w.r.t. *matching* criteria

Examples of matching criteria are age, gender, SES, etc.

Matched pairs sampling is more elaborate:
to be effective often a two stage sampling of controls is done:
first stage, controls are sampled as in the unmatched case;
second stage, from the sample of controls.

strata are built according to the matching criteria from which the matched controls are sampled

Result: data consist of *pairs:* (Case,Control)

Because of the design the case-control study the data are *no longer* two independent samples of the diseased and the healthy population, but rather one independent sample of the diseased population, and a stratified sample of the healthy population, stratified by the matching variable as realized for the case

Case 1 (40 ys, man) ⟶ Control 1 (40 ys, man)
Case 2 (33 ys, wom) ⟶ Control 2 (33 ys, wom)
….

Because of the *design* of the matched case-control study, *stratified analysis* is most appropriate with
each pair defining a stratum

What is the principal structure of a pair?

Four Situations

a)

	Case	Control	
exposed	1	1	
non-exposed			
			2

b)

	Case	Control	
exposed	1		
non-exposed		1	
			2

c)

	Case	Control	
exposed		1	
non-exposed	1		
			2

d)

	Case	Control	
exposed			
non-exposed	1	1	
			2

How many pairs of each type?

Four frequencies

a pairs of type a)

	Case	Control	
exposed	1	1	
non-exposed			
			2

b pairs of type b)

	Case	Control	
exposed	1		
non-exposed		1	
			2

c pairs of type c)

	Case	Control	
exposed		1	
non-exposed	1		
			2

d pairs of type d)

	Case	Control	
exposed			
non-exposed	1	1	
			2

$$\hat{OR}_{MH} = \frac{X_1^{(1)}(m_0^{(1)} - X_0^{(1)})/m^{(1)} + \ldots + X_1^{(k)}(m_0^{(k)} - X_0^{(k)})/m^{(1)}}{X_0^{(1)}(m_1^{(1)} - X_1^{(1)})/m^{(1)} + \ldots + X_1^{(1)}(m_0^{(1)} - X_0^{(1)})/m^{(1)}}$$

$$= \frac{a \times 1 \times 0/2 + b \times 1 \times 1/2 + c \times 0 \times 0/2 + d \times 0 \times 1/2}{a \times 0 \times 1/2 + b \times 0 \times 0/2 + c \times 1 \times 1/2 + d \times 1 \times 0/2}$$

$$= b/c$$

$$= \frac{\text{\# pairs with case exposed and control unexposed}}{\text{\# pairs with case unexposed and controlexposed}}$$

In a matched case-control study, the Mantel-Haenszel odds ratio is estimated by the ratio of the frequency of pairs with *case exposed and control unexposed* to the frequency of pairs with *case unexposed and control exposed*:

(typical presentation of paired studies)

		Control		
Case		*exposed*	*unexposed*	
	exposed	a	b	a+b
	unexposed	c	d	c+d
		a+c	b+d	

$$\hat{OR} \text{ (conventional, unadjusted)} = \frac{(a+b)(b+d)}{(a+c)(c+d)}$$

$$\hat{OR}_{MH} = b/c \quad \text{(ratio of } \textit{discordant pairs}\text{)}$$

Example: Reye-Syndrome and Aspirin Intake

		Control		
Case		*exposed*	*unexposed*	
	exposed	132	57	189
	unexposed	5	6	11
		137	63	200

$$\hat{OR} \text{ (conventional, unadjusted)} = \frac{(a+b)(b+d)}{(a+c)(c+d)} = \frac{189 \times 63}{137 \times 11} = 7.90$$

$$\hat{OR}_{MH} = b/c \quad \text{(ratio of } \textit{discordant pairs}\text{)}$$
$$= 57/5 = 11.4$$

Cleary, for the inference only discordant pairs are required! Therefore, *inference is done conditional* upon discordant pairs

What is the probability that a pair is of type (Case exposed, Control unexposed) given it is discordant?

π = Pr (Case E, Control NE | pair is discordant) =

P(Case E, Control NE) / P(pair is discordant) =

P(Case E, Control NE) / P(Case E, Control NE **or** Case NE, Control E)

$$= q_1(1-q_0)/[\, q_1(1-q_0) + (1-q_1)q_0]$$

$$= \frac{q_1(1-q_0)}{(1-q_1)q_0} / \left(\frac{q_1(1-q_0)}{(1-q_1)q_0} + 1 \right) = OR/ (OR+1)$$

How can I estimate π ?

$$\hat{\pi} = \frac{\text{frequency of pairs: Case E; Control NE}}{\text{frequency of all discordant pairs}}$$

$$= b/(b+c)$$

now, $\pi = OR/(OR+1)$ or $OR = \pi/(1-\pi)$

How can I estimate OR?

$$\hat{OR} = \hat{\pi} /(1-\hat{\pi}) = (b/(b+c) / (1- b/(b+c)) = b/c$$

which corresponds to the Mantel-Haenszel-estimate used before!

Testing and CI Estimation

H_0: OR = 1 or π = OR/(OR+1) = ½
H_1: H_0 is false

since $\hat{\pi}$ is a proportion estimator its estimated standard error is:

SE of $\hat{\pi}$: $\sqrt{\pi\,(1-\pi)/m}$ = $_{\text{Null-Hpyothesis}}$ = ½ $\sqrt{1/m}$

where m=b+c *(number of discordant pairs)*

Teststatistic: $Z = (\hat{\pi} - ½)/ (½ \sqrt{1/m})$

$= \sqrt{b+c}\,(2\,b/(b+c) - 1)$

$= (b-c)/\sqrt{b+c}$

and $\chi^2 = Z^2 = (b-c)^2/(b+c)$ is *McNemar's Chi-Square test statistic!*

In the *example*:

$$\chi^2 = (57-5)^2/62 = 43.61$$

Confidence Interval (again using π)

$$\hat{\pi} \pm 1.96\ \hat{SE}\ (\hat{\pi}) = \hat{\pi} \pm 1.96 \sqrt{\hat{\pi}\ (1-\hat{\pi})/m}$$

and, to get Odds Ratios, use transform. OR = $\pi/(1-\pi)$:

$$\frac{\hat{\pi} \pm 1.96 \sqrt{\hat{\pi}\ (1-\hat{\pi})/m}}{1-\ \hat{\pi} \pm 1.96 \sqrt{\hat{\pi}\ (1-\hat{\pi})/m}}$$

to provide a 95% CI for the Odds Ratio!

In the Example,

$$\hat{\pi} = 57/62 = 0.9194,$$

$$\hat{\pi} \pm 1.96 \sqrt{\hat{\pi}\ (1-\hat{\pi})/m} = 0.9194 \pm 1.96 \times 0.0346$$
$$= (0.8516, 0.9871)$$

leading to the 95%-CI for the Odds Ratio:

$$[0.8516/(1-0.8516),\ 0.9871/(1-0.9871)\ ]$$

$$= [5.7375,\ 76.7194\ ]$$

In Stata:

Cases	Controls Exposed	Unexposed	Total
Exposed	**132**	**57**	**189**
Unexposed	**5**	**6**	**11**
Total	**137**	**63**	**200**

McNemar's chi2(1) = **43.61** Prob > chi2 = **0.0000**
Exact McNemar significance probability = **0.0000**

Proportion with factor

		[95% Conf.	Interval]	
Cases	**.945**			
Controls	**.685**			
difference	**.26**	**.1867662**	**.3332338**	
ratio	**1.379562**	**1.253398**	**1.518425**	
rel. diff.	**.8253968**	**.723037**	**.9277566**	
odds ratio	**11.4**	**4.610017**	**36.44671**	(exact)

.

Confounding and effect modification: Mantel-Haenszel estimation, testing effect homogeneity

Dankmar Böhning

Department of Mathematics and Statistics
University of Reading, UK

Summer School in Cesme, May/June 2011

Overview

1. Cohort Studies with *Similar* Observation Time

2. Cohort Studies with *Individual*, Different Observation Time

3. Case-Control Studies: *Unmatched* Situation

4. Case-Control Studies: *Matched* Situation

1. Cohort Studies with *Similar* Observation Time

Situation in the population:

	Case	Non-Case	
Exposed	p_1	$1-p_1$	
Non-exposed	p_0	$1-p_0$	

interest in: $RR = \frac{p_1}{p_0}$

Situation in the sample:

	Case	Non-Case	At Risk
Exposed	Y_1	$n_1 - Y_1$	n_1
Non-exposed	Y_0	$n_0 - Y_0$	n_0

Interest in estimating $RR = \frac{p_1}{p_0}$:

$$\hat{RR} = \frac{Y_1/n_1}{Y_0/n_0}$$

Example: Radiation Exposure and Cancer Occurrence

	Case	Non-Case	At Risk
Exposed	52	2820	2872
Non-exposed	6	5043	5049

$$\hat{RR} = \frac{52/2872}{6/5049} = \frac{0.0181}{0.0012} = 15.24$$

Tests and Confidence Intervals

Estimated Variance of $\log(\hat{RR})$:

$$\hat{Var}(\log \hat{RR}) = 1/Y_1 - 1/n_1 + 1/Y_0 - 1/n_0$$

Estimated Standard Error of $\log(\hat{RR})$:

$$\hat{SE}(\log \hat{RR}) = \sqrt{1/Y_1 - 1/n_1 + 1/Y_0 - 1/n_0}$$

For the above example:

$$\hat{Var}(\log \hat{RR}) = 1/52 - 1/2872 + 1/6 - 1/5049$$
$$= 0.1854$$

$$\hat{SE}(\log \hat{RR}) = 0.4305$$

Testing

H_0: RR= 1 or log(RR) = 0

H_1: H_0 is false

Statistic used for testing: $Z = \log(\hat{RR}) / \hat{SE}(\log \hat{RR})$

Z is approx. standard normally distributed if H_0 true

Test with Significance level 5%:

reject H_0 if $|Z| > 1.96$
accept H_0 if $|Z| \leq 1.96$

For the example: $Z = \log(15.24)/0.4305 = 6.327$

Confidence Interval

95%-CI covers with 95% confidence the true log (RR):

$$\log(\hat{RR}) \pm 1.96\ \hat{SE}(\log \hat{RR})$$

For the example:

$$\log(15.24) \pm 1.96 \times 0.4305 = (1.8801, 3.5677)$$

and back to the relative risk – scale:

$$(\exp(1.8801), \exp(3.5677)) = (6.55, 35.43)$$

In STATA

	Exposed	Unexposed	Total	
Cases	**52**	**6**	**58**	
Noncases	**2820**	**5043**	**7863**	
Total	**2872**	**5049**	**7921**	
Risk	**.0181058**	**.0011884**	**.0073223**	
	Point estimate		[95% Conf.	Interval]
Risk difference	**.0169175**		**.0119494**	**.0218856**
Risk ratio	**15.23607**		**6.552546**	**35.42713**
Attr. frac. ex.	**.9343663**		**.8473876**	**.971773**
Attr. frac. pop	**.8377077**			
	chi2(1) =	**72.08**	Pr>chi2 =	**0.0000**

Potential Confounding

and Stratification with Respect to the Confounder

Situation:

	Exposed		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	RR
1	50	100	1500	3000	1
2	10	1000	1	100	1
Total	60	1100	1501	3100	0.1585

Explanation?

A more realistic example: *Drinking Coffee and CHD*

	Exposed *(coffee)*		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	RR
Smoker	195	705	21	79	1.03
Non-S	5	95	29	871	1.55
Total	200	800	50	950	4

How to diagnose confounding? Stratify !

Situation:

	Exposed		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	RR
1	$Y_1^{(1)}$	$n_1^{(1)} - Y_1^{(1)}$	$Y_0^{(1)}$	$n_0^{(1)} - Y_0^{(1)}$	$RR^{(1)}$
2	$Y_1^{(2)}$	$n_1^{(2)} - Y_1^{(2)}$	$Y_0^{(2)}$	$n_1^{(2)} - Y_0^{(2)}$	$RR^{(2)}$
…		…		…	
k	$Y_1^{(k)}$	$n_1^{(k)} - Y_1^{(k)}$	$Y_0^{(k)}$	$n_1^{(k)} - Y_0^{(k)}$	$RR^{(k)}$
Total	Y_1	$n_1 - Y_1$	Y_0	$n_1 - Y_0$	RR

How should the RR be estimated?

Use an average of stratum-specific weights:

$$\hat{RR} = w_1\hat{RR}^{(1)} + \ldots + w_k\hat{RR}^{(k)}/(w_1+\ldots+w_k)$$

Which weights?

Mantel-Haenszel Approach

$$\hat{RR}_{MH} = \frac{Y_1^{(1)} n_0^{(1)}/n^{(1)} + \ldots + Y_1^{(k)} n_0^{(k)}/n^{(k)}}{Y_0^{(1)} n_1^{(1)}/n^{(1)} + \ldots + Y_0^{(k)} n_1^{(k)}/n^{(k)}}$$

with $n^{(i)} = n_0^{(i)} + n_1^{(i)}$.

Good Properties!

Mantel-Haenszel Weight: $w_i = Y_0^{(i)} n_1^{(i)}/n^{(i)}$

$$w_1 \hat{RR}^{(1)} + \ldots + w_k \hat{RR}^{(k)}/(w_1 + \ldots + w_k) = \hat{RR}_{MH}$$

Illustration of the MH-weights

	Exposed		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	w_i
1	50	100	1500	3000	1500*150/4650
2	10	1000	1	100	1*1010/1111

In STATA

	Stratum	Case	Exposure	obs
1.	1	1	1	50
2.	1	0	1	100
3.	1	1	0	1500
4.	1	0	0	3000
5.	2	1	1	10
6.	2	0	1	1000
7.	2	1	0	1
8.	2	0	0	100

Stratum	RR	[95% Conf.	Interval]	M-H Weight
1	1	.7944874	1.258673	48.3871
2	1	.1293251	7.732451	.9090909
Crude	.1585495	.123494	.2035559	
M-H combined	1	.7953728	1.257272	

Test of homogeneity (M-H) chi2(1) = 0.000 Pr>chi2 = 1.0000

Illustration: Coffee-CHD-Data

	Case	Exposure	Smoking	freque~y
1.	1	0	1	21
2.	0	0	1	79
3.	1	1	1	195
4.	0	1	1	705
5.	1	0	2	29
6.	0	0	2	871
7.	1	1	2	5
8.	0	1	2	95

Smoking	RR	[95% Conf.	Interval]	M-H Weight
1	1.031746	.6916489	1.539076	18.9
2	1.551724	.6144943	3.918422	2.9
Crude	4	2.971453	5.384571	
M-H combined	1.100917	.7633712	1.587719	

Test of homogeneity (M-H) chi2(1) = 0.629 Pr>chi2 = 0.4279

Inflation, Masking and Effect Modification

Inflation (Confounding): Crude RR is larger (in absolute value) than stratified RR

Masking (Confounding): Crude RR is smaller (in absolute value) than stratified RR

Effect Modification: Crude Rate is in between stratified RR

How can these situations be diagnosed?

Use *heterogeneity or homogeneity* test:

Homogeneity Hypothesis

H_0: $RR^{(1)} = RR^{(2)} = \ldots = RR^{(k)}$

H_1: H_0 is wrong

Teststatistic:

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} (\log \widehat{RR}^{(i)} - \log RR_{MH})^2 / \text{Var}(\log \widehat{RR}^{(i)})$$

Illustration of the Heterogeneity Test for CHD-Coffee

	Exposed		Non-Exposed		
Stratum	Case	Non-Case	Case	Non-Case	χ^2
Smoke	195	705	21	79	0.1011
Non-Smoke	5	95	29	871	0.5274
Total	200	800	50	950	0.6285

Smoking	RR	[95% Conf.	Interval]	M-H Weight
1	**1.031746**	**.6916489**	**1.539076**	**18.9**
2	**1.551724**	**.6144943**	**3.918422**	**2.9**
Crude	**4**	**2.971453**	**5.384571**	
M-H combined	**1.100917**	**.7633712**	**1.587719**	

Test of homogeneity (M-H) chi2(**1**) = **0.629** Pr>chi2 = **0.4279**

.

Cohort Studies with *Individual, different* Observation Time

Situation:

	Event-Risk	Person-Time	At Risk
Exposed	p_1	T_1	n_1
Non-exposed	p_0	T_0	n_0

Definition: Person-Time is the time that n persons spend under risk in the study period

Interest in: $RR = p_1/p_0$

Situation:

	Events	Person-Time	At Risk
Exposed	Y_1	T_1	n_1
Non-exposed	Y_0	T_0	n_0

$$\hat{RR} = \frac{Y_1/T_1}{Y_0/T_0}$$

Y/T is also called the *incidence density* (ID) !

Example: Smoking Exposure and CHD Occurrence

	Events	Person-Time	ID (Events per 10,000 PYs)
Exposed	206	28612	72
Non-exposed	28	5710	49

$$\hat{RR} = \frac{206/28612}{28/5710} = \frac{72}{49} = 1.47$$

Tests and Confidence Intervals

Estimated Variance of $\log(\hat{RR}) = \log(\hat{ID}_1 / \hat{ID}_0)$:

$$\hat{Var}(\log \hat{RR}) = 1/Y_1 + 1/Y_0$$

Estimated Standard Error of $\log(\hat{RR})$:

$$\hat{SE}(\log \hat{RR}) = \sqrt{1/Y_1 + 1/Y_0}$$

For the above example:

$\hat{Var}(\log \hat{RR}) = 1/206 + 1/28 = 0.0405$

$\hat{SE}(\log \hat{RR}) = 0.2013$

Testing

H_0: RR= 1 or log(RR) = 0

H_1: H_0 is false

Statistic used for testing: $Z = \log(\hat{RR}) / \hat{SE}(\log \hat{RR})$

Z is approx. normally distributed if H_0 true:

Test with Significance level 5%:

reject H_0 if $|Z| > 1.96$

accept H_0 if $|Z| \leq 1.96$

For the example: $Z = \log(1.47)/0.2013 = 1.9139$

Confidence Interval

95%-CI covers with 95% confidence the true log (RR):

$$\log(\hat{RR}) \pm 1.96\ \hat{SE}(\log \hat{RR})$$

For the example:

$$\log(1.47) \pm 1.96\ 0.2013 = (-0.0093, 0.7798)$$

and back to the relative risk – scale:

$$(\exp(-0.0093), \exp(0.7798)) = (0.99, 2.18)$$

In STATA

	Exposed	Unexposed	Total		
Cases	**206**	**28**	**234**		
Person-time	**28612**	**5710**	**34322**		
Incidence Rate	**.0071998**	**.0049037**	**.0068178**		
	Point estimate		[95% Conf.	Interval]	
Inc. rate diff.	**.0022961**		**.0002308**	**.0043614**	
Inc. rate ratio	**1.46824**		**.9863624**	**2.264107**	(exact)
Attr. frac. ex.	**.3189125**		**-.0138261**	**.5583247**	(exact)
Attr. frac. pop	**.280752**				
	(midp) Pr(k>=206) =			**0.0243**	(exact)
	(midp) 2*Pr(k>=206) =			**0.0487**	(exact)

.

Stratification with Respect to a Potential Confounder

Example: *energy intake (as surrogate measure for physical inactivity) and Ischaemic Heart Disease*

	Exposed *(<2750 kcal)*		Non-Exposed *(≥2750 kcal)*		
Stratum	Cases	P-Time	Cases	P-Time	RR
40-49	2	311.9	4	607.9	0.97
50-59	12	878.1	5	1272.1	3.48
60-60	14	667.5	8	888.9	2.33
Total	28	1857.5	17	2768.9	2.46

Situation:

	Exposed		Non-Exposed		
Stratum	Cases	P-Time	Cases	P-Time	RR
1	$Y_1^{(1)}$	$T_1^{(1)}$	$Y_0^{(1)}$	$T_0^{(1)}$	$RR^{(1)}$
2	$Y_1^{(2)}$	$T_1^{(2)}$	$Y_0^{(2)}$	$T_0^{(2)}$	$RR^{(2)}$
…		…		…	
k	$Y_1^{(k)}$	$T_1^{(k)}$	$Y_0^{(k)}$	$T_0^{(k)}$	$RR^{(k)}$
Total	Y_1	T_1	Y_0	T_0	RR

How should the RR be estimated?

Use an average of stratum-specific weights:

$$\hat{RR} = w_1{}^{(1)} + \ldots + w_k \hat{RR}^{(k)}/(w_1+\ldots+w_k)$$

Which weights?

Mantel-Haenszel Approach

$$\hat{RR}_{MH} = \frac{Y_1^{(1)}T_0^{(1)}/T^{(1)} + \ldots + Y_1^{(k)}T_0^{(k)}/T^{(k)}}{Y_0^{(1)}T_1^{(1)}/T^{(1)} + \ldots + Y_0^{(k)}T_1^{(k)}/T^{(k)}}$$

with $T^{(i)} = T_0^{(i)} + T_1^{(i)}$.

Mantel-Haensel Weight: $w_i = Y_0^{(i)}T_1^{(i)}/T^{(i)}$

$$w_1\hat{RR}^{(1)} + \ldots + w_k \hat{RR}^{(k)}/(w_1+\ldots+w_k) = \hat{RR}_{MH}$$

In STATA

	Stratum	Exposure	number~e	Person~e
1.	1	1	2	311.9
2.	1	0	4	607.9
3.	2	1	12	878.1
4.	2	0	5	1272.1
5.	3	1	14	667.5
6.	3	0	8	888.9

Stratum	IRR	[95% Conf.	Interval]	M-H Weight	
1	.9745111	.0881524	6.799694	1.356382	(exact)
2	3.476871	1.14019	12.59783	2.041903	(exact)
3	2.33045	.9123878	6.411597	3.430995	(exact)
Crude	2.455204	1.297757	4.781095		(exact)
M-H combined	2.403914	1.306881	4.421829		

Test of homogeneity (M-H) chi2(2) = 1.57 Pr>chi2 = 0.4555

2. Case-Control Studies: *Unmatched* Situation

Situation:

	Case	Controls
Exposed	q_1	q_0
Non-exposed	$1-q_1$	$1-q_0$

Interest is in: RR = p_1/p_0 which is *not* estimable
not in $RR_e = q_1/q_0$

Illustration with a Hypo-Population:

	Bladder-Ca	Healthy	
Smoking	500	199,500	200,000
Non-smoke	500	799,500	800,000
	1000	999,000	1,000,000

$$RR = p_1/p_0 = 4$$

$$\neq 2.504 = \frac{5/10}{1995/9990} = q_1/q_0 = RR_e$$

However, consider the (disease) **Odds Ratio** defined as

$$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

$Pr(D/E) = p_1$, $Pr(D/NE) = p_0$,

$Pr(E/D) = q_1$, $Pr(E/ND) = q_0$, $p = Pr(D)$

$p_1 = P(D/E)$ *using Bayes Theorem*

$$= \frac{Pr(E/D)Pr(D)}{Pr(E/D)Pr(D)+ Pr(E/ND)Pr(ND)} = \frac{q_1\, p}{q_1 p + q_0\,(1-p)}$$

$p_0 = P(D/NE)$

$$= \frac{Pr(NE/D)Pr(D)}{Pr(NE/D)Pr(D)+ Pr(NE/ND)Pr(ND)} = \frac{(1-q_1)\, p}{(1-q_1)\, p + (1-q_0\,)(1-p)}$$

$p_1/(1-p_1) = q_1p/q_0(1-p)$ und $p_0/(1-p_0) = [(1-q_1)p]/[(1-q_0)(1-p)]$.

it follows that

$$OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)} = \frac{q_1/q_0}{(1-q_1)/(1-q_0)} = \frac{q_1/(1-q_1,)}{q_0/(1-q_0)} = OR_e$$

Disease Odds Ratio = Exposure Odds Ratio

Illustration with a Hypo-Population:

	Bladder-Ca	Healthy	
Smoking	500	199,500	200,000
Non-smoke	500	799,500	800,000
	1000	999,000	1,000,000

OR = (500/199,500)/(500/799,500) = (500/500)/(199,500/799,500) = OR_e = 4.007

Also, if disease occurrence is low (low prevalence),

OR ≈ RR

Estimation of OR

Situation:

	Case	Controls
Exposed	X_1	X_0
Non-exposed	m_1-X_1	m_0-X_0
	m_1	m_0

$$\hat{OR} = \frac{\hat{q}_1/(1-\hat{q}_1)}{\hat{q}_0/(1-\hat{q}_0)} = \frac{X_1/(m_1-X_1)}{X_0/(m_0-X_0)} = \frac{X_1(m_0-X_0)}{X_0(m_1-X_1)}$$

Example: Sun Exposure and Lip Cancer Occurrence in Population of 50-69 year old men

	Case	Controls
Exposed	66	14
Non-exposed	27	15
	93	29

$$\hat{OR} = \frac{66 \times 15}{14 \times 27} = 2.619$$

Tests and Confidence Intervals

Estimated Variance of $\log(\hat{OR})$:

$$\hat{Var}(\log \hat{OR}) = \frac{1}{X_1} + \frac{1}{m_1 - X_1} + \frac{1}{X_0} + \frac{1}{m_0 - X_0}$$

Estimated Standard Error of $\log(\hat{OR})$:

$$\hat{SE}(\log \hat{OR}) = \sqrt{\frac{1}{X_1} + \frac{1}{m_1 - X_1} + \frac{1}{X_0} + \frac{1}{m_0 - X_0}}$$

For the above example:

$$\hat{Var}(\log \hat{OR}) = 1/66 + 1/27 + 1/14 + 1/15$$
$$= 0.1903$$

$$\hat{SE}(\log \hat{OR}) = 0.4362$$

Testing

H_0: OR= 1 or log(OR) = 0

H_1: H_0 is false

Statistic used for testing: $Z = \log(\hat{OR}) / \hat{SE}(\log \hat{OR})$

Z is approx. normally distributed if H_0 true:

Test with Significance level 5%:

reject H_0 if $|Z| > 1.96$

accept H_0 if $|Z| \leq 1.96$

For the example: Z = log(2.619)/0.4362 = 2.207

Confidence Interval

95%-CI covers with 95% confidence the true log (RR):

$$\log(\hat{OR}) \pm 1.96\ \hat{SE}(\log \hat{OR})$$

For the example:

$$\log(2.619) \pm 1.96\ 0.4362 = (0.1078, 1.8177)$$

and back to the relative risk – scale:

$$(\exp(0.1078), \exp(1.8177)) = (1.11, 6.16)$$

In STATA

	Exposed	Unexposed	Total	Proportion Exposed	
Cases	**66**	**27**	**93**	**0.7097**	
Controls	**14**	**14**	**28**	**0.5000**	
Total	**80**	**41**	**121**	**0.6612**	
	Point estimate		[95% Conf.	Interval]	
Odds ratio	**2.444444**		**1.028622**	**5.809044**	(Woolf)
Attr. frac. ex.	**.5909091**		**.0278254**	**.8278546**	(Woolf)
Attr. frac. pop	**.4193548**				
	chi2(1) =	**4.22**	Pr>chi2 =	**0.0399**	

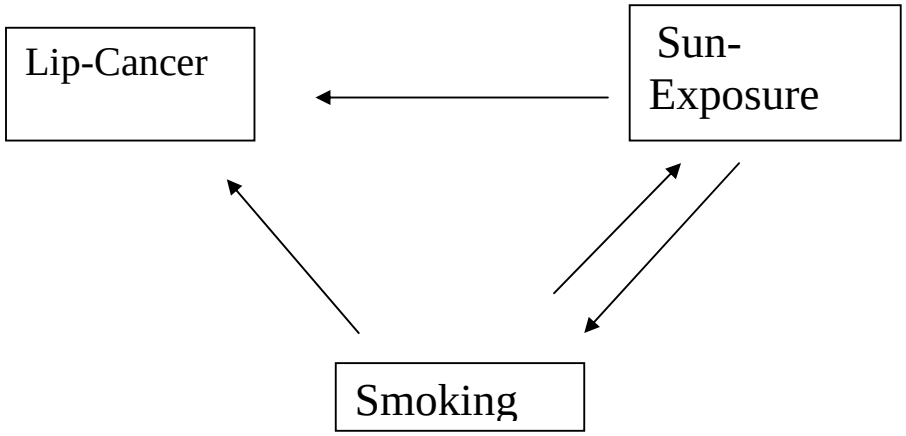
.

Exercise: A case-control study investigates if a keeping a pet bird is a risk factor: Cases: 98 Bird Owners, 141 None, Controls: 101 Bird Owners, 328 None

Potential Confounding

and Stratification with Respect to the Confounder

Situation:



Lip-Cancer and Sun Exposure with Smoking as Potential Confounder

	Cases		Controls		
Stratum	Exposed	Non-Exp.	Exp.	Non-Exp.	OR
Smoke	51	24	6	10	3.54
Non-Smoke	15	3	8	5	3.13
Total	66	27	14	15	2.62

Explanation?

How to diagnose confounding? Stratify !

Situation:

	Cases		Controls		Cases
Stra-tum	Ex-posed	Non-Exp.	Ex-posed	Non-Exp.	OR
1	$X_1^{(1)}$	$m_1^{(1)} - X_1^{(1)}$	$X_0^{(1)}$	$m_0^{(1)} - X_0^{(1)}$	$OR^{(1)}$
2	$X_1^{(2)}$	$m_1^{(2)} - X_1^{(2)}$	$X_0^{(2)}$	$m_1^{(2)} - X_0^{(2)}$	$OR^{(2)}$
…		…		…	
k	$X_1^{(k)}$	$m_1^{(k)} - X_1^{(k)}$	$X_0^{(k)}$	$m_1^{(k)} - X_0^{(k)}$	$OR^{(k)}$
Total	X_1	$m_1 - X_1$	X_0	$m_1 - X_0$	OR

How should the OR *based upon stratification* **be estimated?**

Use an average of stratum-specific weights:

$$\hat{OR} = w_1\hat{OR}^{(1)} + \ldots + w_k\hat{OR}^{(k)}/(w_1+\ldots+w_k)$$

Which weights?

Mantel-Haenszel Weight: $w_i = X_0^{(i)}(m_1^{(i)} - X_1^{(i)})/ m^{(i)}$

Mantel-Haenszel Approach

$$\hat{OR}_{MH} = \frac{X_1^{(1)}(m_0^{(1)} - X_0^{(1)})/m^{(1)} + \ldots + X_1^{(k)}(m_0^{(k)} - X_0^{(k)})/m^{(1)}}{X_0^{(1)}(m_1^{(1)} - X_1^{(1)})/ m^{(1)} + \ldots + X_1^{(1)}(m_0^{(1)} - X_0^{(1)})/ m^{(1)}}$$

with $m^{(i)} = m_0^{(i)} + m_1^{(i)}$.

$$w_1\hat{OR}^{(1)} + \ldots + w_k\hat{OR}^{(k)}/(w_1+\ldots+w_k) = \hat{OR}_{MH}$$

Illustration of the MH-weights

	Cases		Controls		
Stratum	Exposed	Non-Exp.	Exp.	Non-Exp.	w_i
Smoke	51	24	6	10	6*24/91
Non-Smoke	15	3	8	5	8*3/31

In STATA

	Case	Exposure	Smoke	Pop
1.	1	1	0	51
2.	0	1	0	6
3.	1	0	0	24
4.	0	0	0	10
5.	1	1	1	15
6.	0	1	1	8
7.	1	0	1	3
8.	0	0	1	5

. cc Case Control [freq=Pop], by(Smoke)

Smoke	OR	[95% Conf. Interval]	M-H Weight

```
----------------+-------------------------------------------------
           0 |   3.541667    1.011455   13.14962      1.582418 (exact)
           1 |      3.125      .4483337   24.66084     .7741935 (exact)
----------------+-------------------------------------------------
           Crude |   2.619048    1.016247    6.71724            (exact)
M-H combined |   3.404783    1.341535   8.641258
----------------------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =     0.01  Pr>chi2 = 0.9029

                Test that combined OR = 1:
Mantel-Haenszel chi2(1) =      6.96            Pr>chi2 =    0.0083
```

Note that “freq=Pop” is optional, e.g. raw data can be used with this analysis

Inflation, Masking and Effect Modification

Inflation (Confounding): Crude OR is larger (in absolute value) than stratified OR

Masking (Confounding): Crude OR is smaller (in absolute value) than stratified OR

Effect Modification: Crude Rate is in between stratified OR

How can these situations be diagnosed? Use *heterogeneity* or *homogeneity* test:

Homogeneity Hypothesis

H_0: $OR^{(1)} = OR^{(2)} = \ldots = OR^{(k)}$
H_1: H_0 is wrong

$$\chi^2_{(k-1)} = \sum_{i=1}^{k} (\log \widehat{OR}^{(i)} - \log OR_{MH})^2 / \mathrm{Var}(\log \widehat{OR}^{(i)})$$

Illustration of the Heterogeneity Test for Lip Cancer -Sun Exposure

	Cases		Controls		
Stratum	Exposed	Non-Exp.	Exp.	Non-Exp.	χ^2
Smoke	51	24	6	10	0.0043
Non-Smoke	15	3	8	5	0.0101
Total	66	27	14	15	0.0144

	D	E	stratum	freq
1.	0	0	1	10
2.	0	1	2	8
3.	0	1	1	6
4.	1	0	1	24
5.	1	1	1	51
6.	1	0	2	3
7.	0	0	2	5
8.	1	1	2	15

stratum	OR	[95% Conf.	Interval]	M-H Weight	
1	3.541667	1.011455	13.14962	1.582418	(exact)
2	3.125	.4483337	24.66091	.7741935	(exact)
Crude	2.619048	1.016247	6.717228		(exact)
M-H combined	3.404783	1.341535	8.641258		

Test of homogeneity (M-H) chi2(1) = 0.01 Pr>chi2 = 0.9029

Test that combined OR = 1:
Mantel-Haenszel chi2(1) = 6.96
Pr>chi2 = 0.0083

3. Case-Control Studies: *Matched* Situation

Given a *case* is sampled, a *comparable* control is sampled: comparable w.r.t. *matching* criteria

Examples of matching criteria are age, gender, SES, etc.

Matched pairs sampling is more elaborate:
to be effective often a two stage sampling of controls is done:
first stage, controls are sampled as in the unmatched case;
second stage, from the sample of controls.

strata are built according to the matching criteria from which the matched controls are sampled

Result: data consist of *pairs:* (Case,Control)

Because of the design the case-control study the data are *no longer* two independent samples of the diseased and the healthy population, but rather one independent sample of the diseased population, and a stratified sample of the healthy population, stratified by the matching variable as realized for the case

Case 1 (40 ys, man) ⟶ Control 1 (40 ys, man)
Case 2 (33 ys, wom) ⟶ Control 2 (33 ys, wom)
….

Because of the *design* of the matched case-control study, *stratified analysis* is most appropriate with
each pair defining a stratum

What is the principal structure of a pair?

Four Situations

a)

	Case	Control	
exposed	1	1	
non-exposed			
			2

b)

	Case	Control	
exposed	1		
non-exposed		1	
			2

c)

	Case	Control	
exposed		1	
non-exposed	1		
			2

d)

	Case	Control	
exposed			
non-exposed	1	1	
			2

How many pairs of each type?

Four frequencies

a pairs of type a)

	Case	Control	
exposed	1	1	
non-exposed			
			2

b pairs of type b)

	Case	Control	
exposed	1		
non-exposed		1	
			2

c pairs of type c)

	Case	Control	
exposed		1	
non-exposed	1		
			2

d pairs of type d)

	Case	Control	
exposed			
non-exposed	1	1	
			2

$$\hat{OR}_{MH} = \frac{X_1^{(1)}(m_0^{(1)} - X_0^{(1)})/m^{(1)} + \ldots + X_1^{(k)}(m_0^{(k)} - X_0^{(k)})/m^{(1)}}{X_0^{(1)}(m_1^{(1)} - X_1^{(1)})/m^{(1)} + \ldots + X_1^{(1)}(m_0^{(1)} - X_0^{(1)})/m^{(1)}}$$

$$= \frac{a \times 1 \times 0/2 + b \times 1 \times 1/2 + c \times 0 \times 0/2 + d \times 0 \times 1/2}{a \times 0 \times 1/2 + b \times 0 \times 0/2 + c \times 1 \times 1/2 + d \times 1 \times 0/2}$$

$$= b/c$$

$$= \frac{\text{\# pairs with case exposed and control unexposed}}{\text{\# pairs with case unexposed and controlexposed}}$$

In a matched case-control study, the Mantel-Haenszel odds ratio is estimated by the ratio of the frequency of pairs with *case exposed and control unexposed* to the frequency of pairs with *case unexposed and control exposed*:

(typical presentation of paired studies)

		Control		
Case		*exposed*	*unexposed*	
	exposed	a	b	a+b
	unexposed	c	d	c+d
		a+c	b+d	

$$\hat{OR} \text{ (conventional, unadjusted)} = \frac{(a+b)(b+d)}{(a+c)(c+d)}$$

$$\hat{OR}_{MH} = b/c \quad \text{(ratio of } \textit{discordant pairs}\text{)}$$

Example: Reye-Syndrome and Aspirin Intake

		Control		
Case		*exposed*	*unexposed*	
	exposed	132	57	189
	unexposed	5	6	11
		137	63	200

$$\hat{OR} \text{ (conventional, unadjusted)} = \frac{(a+b)(b+d)}{(a+c)(c+d)} = \frac{189 \times 63}{137 \times 11} = 7.90$$

$$\hat{OR}_{MH} = b/c \quad \text{(ratio of } \textit{discordant pairs}\text{)}$$
$$= 57/5 = 11.4$$

Cleary, for the inference only discordant pairs are required! Therefore, *inference is done conditional* upon discordant pairs

What is the probability that a pair is of type (Case exposed, Control unexposed) given it is discordant?

π = Pr (Case E, Control NE | pair is discordant) =

P(Case E, Control NE) / P(pair is discordant) =

P(Case E, Control NE) / P(Case E, Control NE **or** Case NE, Control E)

$$= q_1(1-q_0)/[\, q_1(1-q_0) + (1-q_1)q_0]$$

$$= \frac{q_1(1-q_0)}{(1-q_1)q_0} / \left(\frac{q_1(1-q_0)}{(1-q_1)q_0} + 1 \right) = OR/ (OR+1)$$

How can I estimate π ?

$$\hat{\pi} = \frac{\text{frequency of pairs: Case E; Control NE}}{\text{frequency of all discordant pairs}}$$

$$= b/(b+c)$$

now, $\pi = OR/(OR+1)$ or $OR = \pi/(1-\pi)$

How can I estimate OR?

$$\hat{OR} = \hat{\pi} /(1-\hat{\pi}) = (b/(b+c) / (1- b/(b+c)) = b/c$$

which corresponds to the Mantel-Haenszel-estimate used before!

Testing and CI Estimation

H_0: OR = 1 or π = OR/(OR+1) = ½
H_1: H_0 is false

since $\hat{\pi}$ is a proportion estimator its estimated standard error is:

SE of $\hat{\pi}$: $\sqrt{\pi\,(1-\pi)/m}$ = $_{\text{Null-Hpyothesis}}$ = ½ $\sqrt{1/m}$

where m=b+c *(number of discordant pairs)*

Teststatistic: $Z = (\hat{\pi} - ½)/ (½ \sqrt{1/m})$

$= \sqrt{b+c}\,(2\,b/(b+c) - 1)$

$= (b-c)/\sqrt{b+c}$

and $\chi^2 = Z^2 = (b-c)^2/(b+c)$ is *McNemar's Chi-Square test statistic!*

In the *example*:

$$\chi^2 = (57-5)^2/62 = 43.61$$

Confidence Interval (again using π)

$$\hat{\pi} \pm 1.96\ \hat{SE}\ (\hat{\pi}) = \hat{\pi} \pm 1.96 \sqrt{\hat{\pi}\ (1-\hat{\pi})/m}$$

and, to get Odds Ratios, use transform. OR = $\pi/(1-\pi)$:

$$\frac{\hat{\pi} \pm 1.96 \sqrt{\hat{\pi}\ (1-\hat{\pi})/m}}{1-\ \hat{\pi} \pm 1.96 \sqrt{\hat{\pi}\ (1-\hat{\pi})/m}}$$

to provide a 95% CI for the Odds Ratio!

In the Example,

$$\hat{\pi} = 57/62 = 0.9194,$$

$$\hat{\pi} \pm 1.96 \sqrt{\hat{\pi}\ (1-\hat{\pi})/m} = 0.9194 \pm 1.96 \times 0.0346$$
$$= (0.8516, 0.9871)$$

leading to the 95%-CI for the Odds Ratio:

$$[0.8516/(1-0.8516),\ 0.9871/(1-0.9871)\ ]$$

$$= [5.7375,\ 76.7194\ ]$$

In Stata:

Cases	Controls Exposed	Unexposed	Total
Exposed	**132**	**57**	**189**
Unexposed	**5**	**6**	**11**
Total	**137**	**63**	**200**

McNemar's chi2(1) = **43.61** Prob > chi2 = **0.0000**
Exact McNemar significance probability = **0.0000**

Proportion with factor

		[95% Conf.	Interval]	
Cases	**.945**			
Controls	**.685**			
difference	**.26**	**.1867662**	**.3332338**	
ratio	**1.379562**	**1.253398**	**1.518425**	
rel. diff.	**.8253968**	**.723037**	**.9277566**	
odds ratio	**11.4**	**4.610017**	**36.44671**	(exact)

.

Lecture 8

Modelling with Covariates: Introduction to General Regression

James Gallagher

Director, Statistical Services Centre

University of Reading

Reading

UK

May 2011

Contents

Introduction to Modelling

Example: Does increased sugar consumption lead to dental caries?

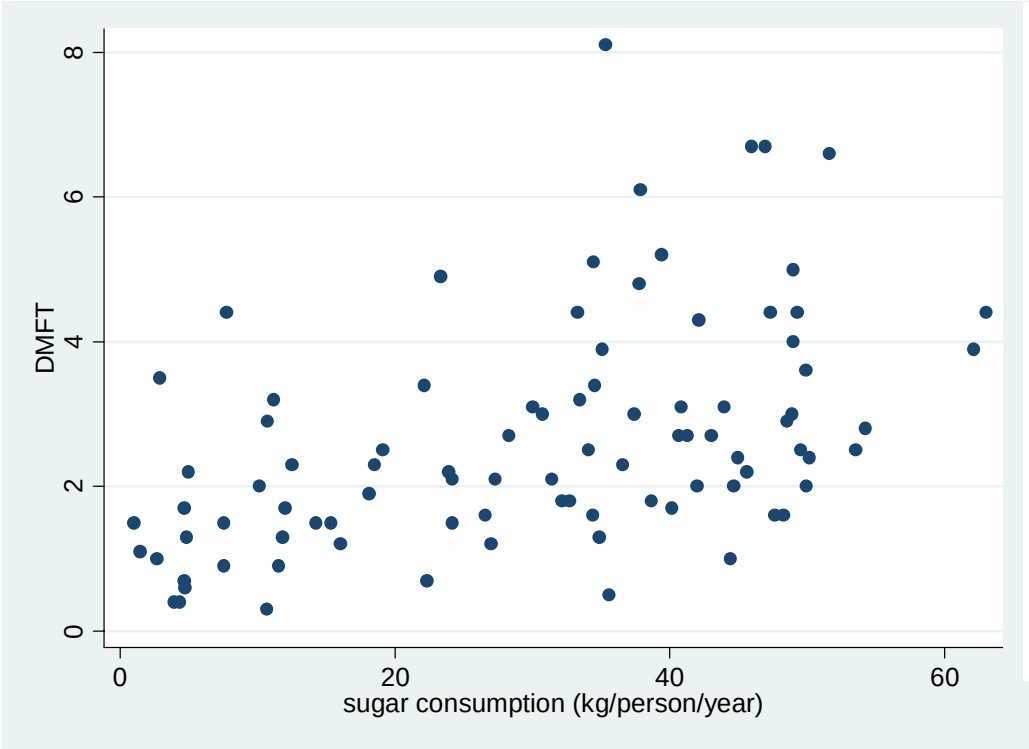
Data on sugar consumption and dental caries in 90 countries.

- Response, or outcome = mean number of decayed, missing or filled teeth (DMFT) at age 12 years-old
    - DMFT score: a continuous response, or outcome
- Exposure = average sugar consumption (kg/head of population/year)
    - A continuous exposure variable
- Data from national surveys between 1979 and 1990, via the WHO Oral Disease Data Bank made available to Woodward and Walker (1994). See Appendix

Exploratory Data Analysis

Graphics: plot of DMFT score against sugar.

[**Stata**: Graphics → Twoway graph (scatter, line, etc.)]



Comments

- DMFT score increases with increasing sugar consumption
- Rough linear association
- Large amount of random variability about the linear trend

A Statistical Model

The simplest summary for the association between 2 continuous variables is a straight line model:

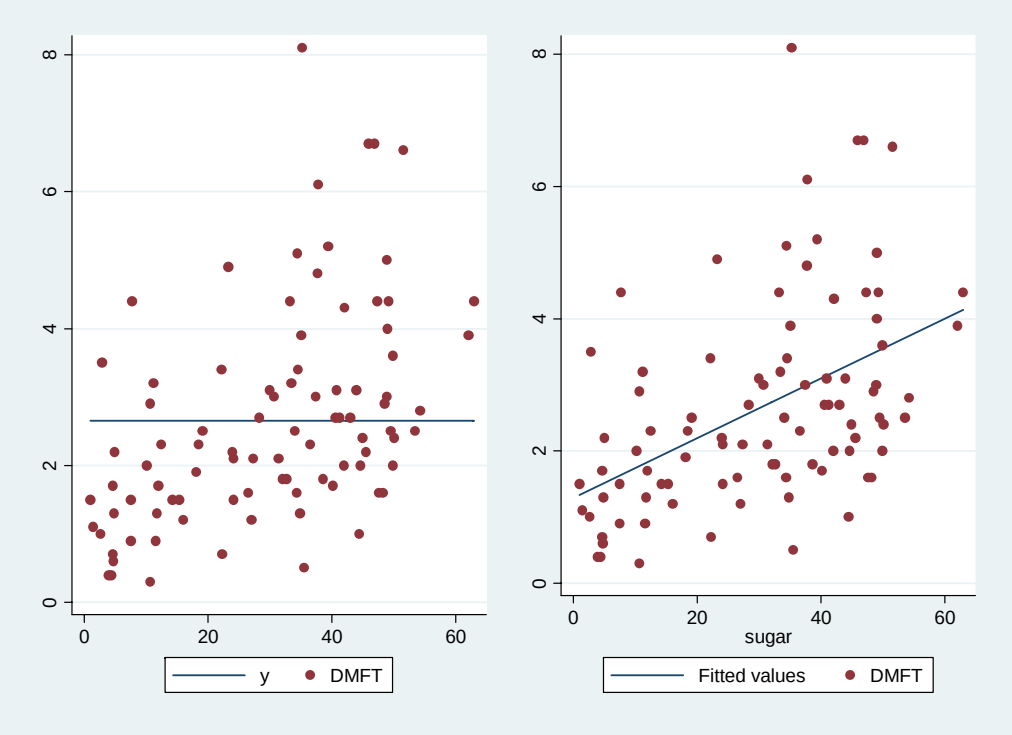
Data = mean (trend) + random error

$y = \alpha + \beta x + \varepsilon$

where y = DMFT score
x = average sugar consumption
ε = independent $N(0,\sigma^2)$ errors

In the literature this regression model is often called a **simple linear regression** model, and is a special case of a **general linear model**.

Competing (nested) models:



mean y = α
DMFT score is not associated with sugar consumption

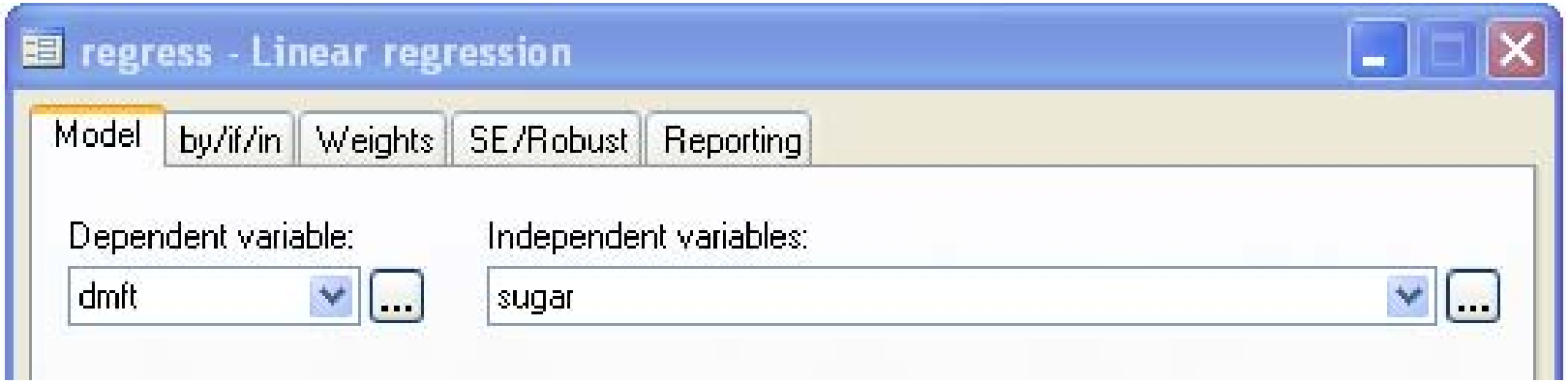
mean y = $\alpha + \beta x$
DMFT score is associated with sugar consumption

If there is truly no association between DMFT score and sugar consumption then $\beta = 0$.

β represents the effect measure in this situation. It is the rate of change in mean y per unit increase in x.

Regression Modelling in Stata

Fit the model in Stata (v.11) to estimate effect of sugar consumption.
[**Stata**: Statistics→Linear models and related→Linear regression]



Stata output:

```
. regress dmft sugar
  Source |       SS       df       MS         Number of obs =      90
---------+------------------------------      F(  1,    88) =   25.60
   Model |  49.8358297     1  49.8358297      Prob > F      =  0.0000
Residual |  171.326395    88  1.94689085      R-squared     =  0.2253
---------+------------------------------      Adj R-squared =  0.2165
   Total |  221.162225    89  2.48496882      Root MSE      =  1.3953

------------------------------------------------------------------------------
    dmft |      Coef.   Std. Err.    t    P>|t|  [95% Conf. Interval]
---------+--------------------------------------------------------------------
   sugar |   .0450854   .0089112   5.06   0.000  .0273763    .0627946
   _cons |   1.296561   .3062384   4.23   0.000  .6879762    1.905145
------------------------------------------------------------------------------
```

$\hat{\beta} = 0.045$.

For a 1 unit increase in sugar consumption, the estimated change in mean DMFT score is an increase of 0.045 units.

95% CI = 0.027 to 0.063, i.e. 0.045 ± 0.018.

$\hat{\alpha} = 1.30$. Estimated mean DMFT score at 0 sugar consumption.

Hypothesis Testing: Model Comparisons

If there is truly no effect of sugar consumption, then $\beta = 0$. This leads to testing:

H_0: $\beta = 0$ (No sugar effect)
against
H_1: $\beta \neq 0$ (There is an effect of sugar)

The F-test. From Stata

```
F(  1,   88)  =   25.60
Prob > F      =  0.0000
```

p-value = <0.001. Hence, there is a statistically significant sugar consumption effect. The higher the sugar consumption, the higher the mean DMFT score.

Notes

- The table of parameter estimates gives an equivalent t-test

dmft	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sugar	.0450854	.0089112	**5.06**	**0.000**	.0273763	.0627946

- Remember the previous F-test (or t-test) is comparing the fit of two models to the data:
  - (1) $y = \alpha + \varepsilon$
  - (2) $y = \alpha + \beta x + \varepsilon$

R^2: Coefficient of Determination

A crude summary measure of the goodness-of-fit of the fitted model.

```
  Source |       SS       df       MS
---------+------------------------------
   Model |  49.8358297     1  49.8358297
Residual |  171.326395    88  1.94689085      R-squared     =  0.2253
---------+------------------------------
   Total |  221.162225    89  2.48496882
```

R^2 = Model SS / Total SS = 0.225 or 22.5%.

22.5% of the variation in the DMFT scores is explained by the fitted the model.

This "low" R^2 indicates that there is a lot of unexplained variability.

The remaining 77.5% could be attributed to many other factors.

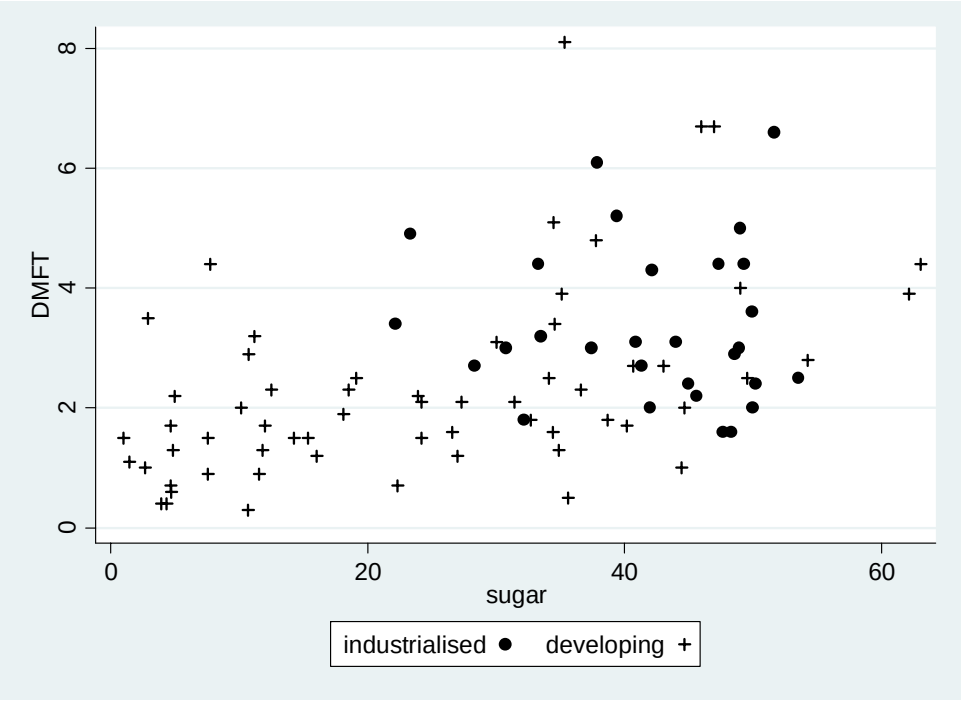
Confounding

- 29 countries were classified as “industrialised” and the remaining 61 as “developing”.
- Consider type of country as a potential confounding factor
  - A categorical variable (2 levels)

How does DMFT score depend upon sugar consumption adjusted for type of country?

What about effect modification? Is there an interaction between sugar consumption and type of country?

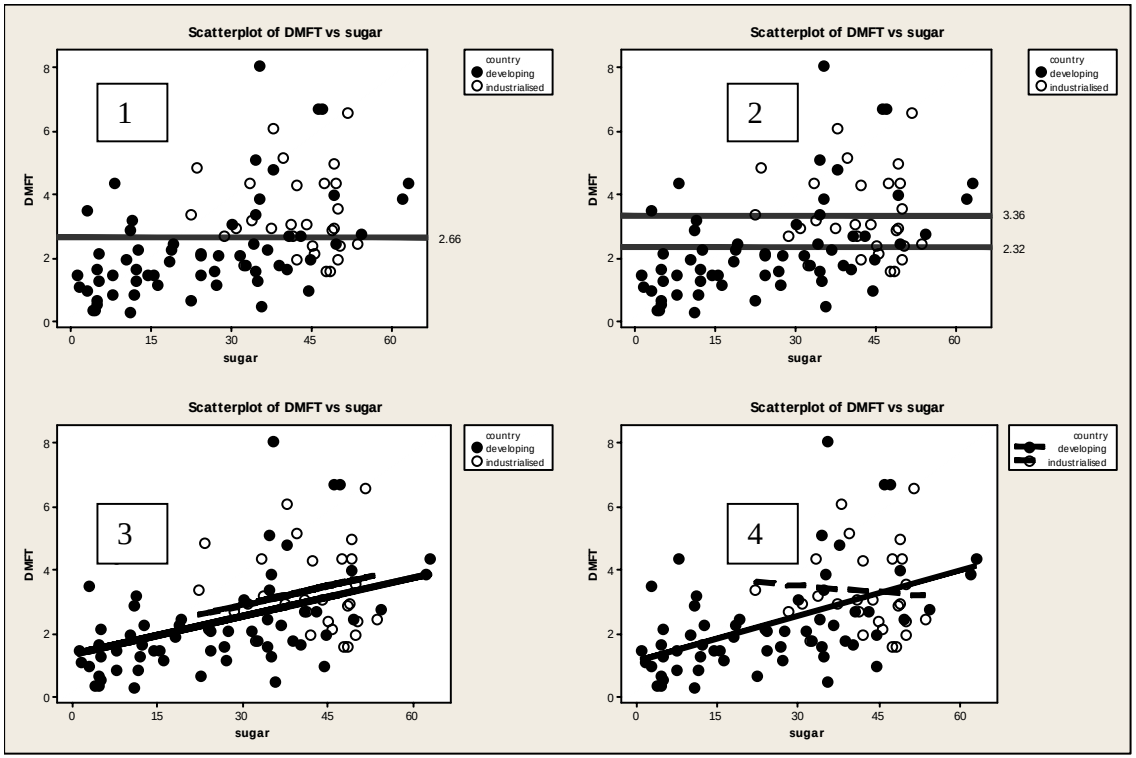
Exploratory Data Analysis



Comments

- Rough linear associations, more clear in the developing countries
- The effect of sugar consumption may be modified by the type of country

Some competing (nested) models:



- Model 1: No effect of sugar or type.
- Model 2: No sugar effect adjusting /allowing for type.
- Model 3: Sugar effect, allowing for type. [Assuming no modification.]
- Model 4: Sugar effect with modification.

No Effect Modification [Model 3]

Data = mean (trend) + random error

$y = \alpha + country_i + \beta x + \varepsilon$

where y = DMFT score

$country_i$ = (main) effect of country, i = 0,1 corresponding to industrialised and developing resepectively

x = average sugar consumption

Constraints

- The model is over parameterised.
- Impose a constraint, say $country_0 = 0$

Note the pattern in the mean trend:

Type = 0, industrialised

$y = \alpha + \text{country}_0 + \beta x = \alpha + \beta x$

Type = 1, developing

$y = \alpha + \text{country}_1 + \beta x = (\alpha + \text{country}_1) + \beta x$

Comments

- Two parallel lines
- β is the rate of change for a fixed country
  - For a 1 unit increase in sugar consumption, the estimated change in mean DMFT score, adjusted for type of country, is an increase of β units

i.e. β represents the (linear) sugar effect **adjusted** for country

Fitting the model in Stata…

[**Stata**: Statistics→Linear models and related → Linear regression]

```
. regress dmft i.country  sugar
------------------------------------------------------------------------------
        dmft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.country |  -.3479401   .3607644    -0.96   0.337    -1.064998    .3691182
       sugar |   .0402757   .0102148     3.94   0.000     .0199726    .0605788
       _cons |   1.677366   .4997554     3.36   0.001     .6840476    2.670684
------------------------------------------------------------------------------
```

- t test: statistically significant sugar effect after adjusting for type of country (p-value = 0.0002)
- $\hat{\beta} = 0.040$, 95% CI = (0.020,0.061)
- For a 1 unit increase in sugar consumption, the estimated change in mean DMFT score, adjusted for type of country, is an increase of 0.040 units

Interaction - Effect Modification

Use **Model 4** to investigate effect modification:

Data = mean (trend) + random error

$y = \alpha + country_i + \beta x + \beta_i x + \varepsilon$

where y = DMFT score

$country_i$ = (main) effect of country, i = 0,1 corresponding to industrialised and developing resepectively

x = average sugar consumption

Constraints

- $country_0 = 0$
- $\beta_0 = 0$

Note the pattern in the mean trend:

Type = 0, industrialised

$$y = \alpha + \text{country}_0 + \beta x + \beta_0 x = \alpha + \beta x$$

Type = 1, developing

$$y = \alpha + \text{country}_1 + \beta x + \beta_1 x = (\alpha + \text{country}_1) + (\beta + \beta_1) x$$

Comments

- Two 'separate' lines
- Effect of increasing sugar depends upon the type of country
  - β_1 represents the interaction effect, or effect modification

Fitting the model in Stata…

```
. regress dmft i.country  sugar i.country#c.sugar
------------------------------------------------------------------------------
        dmft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   1.country |   -2.74389   1.324808    -2.07   0.041    -5.377522   -.1102589
       sugar |   -.013065   .0301432    -0.43   0.666    -.0729876    .0468576
             |
     country#|
     c.sugar |
           1 |   .0600413   .0319804     1.88   0.064    -.0035337    .1236163
             |
       _cons |   3.908571   1.286499     3.04   0.003     1.351096    6.466045
------------------------------------------------------------------------------
```

Type = 0, industrialised	Type = 1, developing
$\hat{y} = \hat{\alpha} + \hat{\beta}x = 3.91 - 0.013x$	$\hat{y} = (\hat{\alpha} + \widehat{\text{country}_1}) + (\hat{\beta} + \hat{\beta}_1)x$
Estimated slope:	$= (3.91 - 2.74) + (-0.013 + 0.060)x$
–0.013, 95% CI = (–0.073, 0.047)	$= 1.17 + 0.047x$
	Estimated slope:
	0.047, 95% CI = (0.026, 0.068)

From the t test for the interaction term: p-value = 0.064. Weak evidence for effect modification.

Conclusions

- No evidence for association between dental status and sugar consumption in industrialised countries
- But there is in developing countries
- A possible epidemiological explanation?
  - Greater use of fluoride toothpastes, and other dental hygiene products in industrialised countries
  - Wider access to dental care in industrialised countries

Extensions

- The modelling framework naturally extends to more complex situations
  - E.g. Adjusting for several potential confounders
- Provides a very flexible framework for statistical analysis

Appendix 1
Sugar Consumption and Dental Caries Data

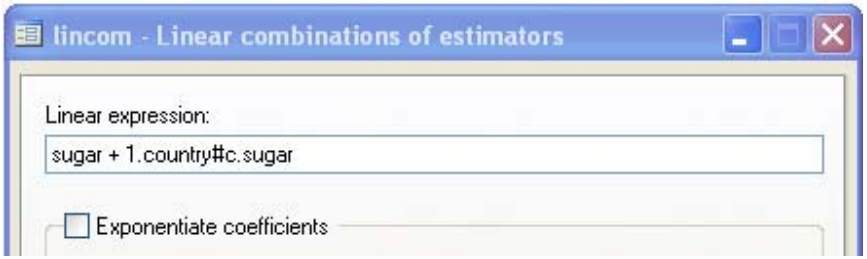
Mean number of decayed , missing or filled teeth (DMFT) at age 12 years old and mean sugar consumption (kg/head of population/year) in 61 developing countries and 29 industrialised countries. Codes for country are 0= industrialised, 1=developing. [Source: Woodward and Walker (1994).]

country	sugar	DMFT	country	Sugar	DMFT	country	sugar	DMFT
0	22.16	3.4	1	54.24	2.8	1	36.6	2.3
0	49.96	2	1	26.56	1.6	1	12	1.7
0	47.32	4.4	1	4.36	0.4	1	34.56	3.4
0	40.86	3.1	1	35.3	8.1	1	34.4	1.6
0	48.92	3	1	40.65	2.7	1	34.86	1.3
0	42.12	4.3	1	11.17	3.2	1	2.88	3.5
0	49.92	3.6	1	24.18	1.5	1	63.02	4.4
0	48.28	1.6	1	12.5	2.3	1	49.02	4
0	41.96	2	1	43	2.7	1	35.6	0.5
0	37.4	3	1	10.74	2.9	1	46.98	6.7
0	39.42	5.2	1	45.98	6.7	1	7.56	1.5
0	33.3	4.4	1	44.44	1	1	4.66	0.7
0	48.98	5	1	11.56	0.9	1	37.76	4.8
0	51.62	6.6	1	44.63	2	1	62.14	3.9
0	48.56	2.9	1	7.76	4.4	1	34.1	2.5
0	30.74	3	1	7.56	0.9	1	34.44	5.1
0	47.62	1.6	1	35.1	3.9	1	3.92	0.4
0	53.54	2.5	1	31.43	2.1	1	11.82	1.3
0	50.16	2.4	1	5	2.2	1	18.1	1.9
0	41.28	2.7	1	32.68	1.8	1	24.16	2.1
0	49.28	4.4	1	1.44	1.1	1	40.18	1.7
0	33.48	3.2	1	4.68	1.7	1	4.72	0.6
0	45.6	2.2	1	10.15	2	1	15.34	1.5
0	44.98	2.4	1	16.02	1.2	1	10.7	0.3
0	28.32	2.7	1	23.93	2.2	1	27.3	2.1
0	43.95	3.1	1	38.66	1.8	1	0.97	1.5
0	32.14	1.8	1	14.26	1.5	1	19.1	2.5
0	37.86	6.1	1	4.84	1.3	1	30	3.1
0	23.32	4.9	1	49.56	2.5	1	22.33	0.7
			1	27	1.2	1	2.66	1
						1	18.53	2.3

Appendix 11
Estimating the Slope for Developing Countries

From Model 4, allowing for effect modification, the estimated slope for developing countries is 0.047, but how do we obtain a corresponding confidence interval? One way is to use a post-estimation command. Having fitted the model including the interaction effect, ask Stata to explicitly estimate the relevant slope. (To do this we need to specify the slope in terms of the sum of two model parameters, $\hat{\beta} + \hat{\beta}_1$)

- Select **Statistics → Postestimation → Linear combinations of estimates**.
- Make the specifications below, which correspond to $\hat{\beta} + \hat{\beta}_1$. **Click Submit**.



Output:

```
. lincom sugar + 1.country#c.sugar

 ( 1)  sugar + 1.country#c.sugar = 0

------------------------------------------------------------------------------
        dmft |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .0469762   .0106835     4.40   0.000     .0257381    .0682144
------------------------------------------------------------------------------
```

References

Woodward, M. and Walker, A.R.P. (1994) Sugar Consumption and Dental Caries: Evidence from 90 Countries. *British Dent. Journal*, **176**, 297-302.

Lecture 9: Logistic Regression Disease Modelling with Covariates

Fazil Baksh

Department of Mathematics and Statistics
University of Reading, UK

Summer School - May/June 2011
Çeşme

This lecture presents an overview of **Logistic Regression** as a tool for evaluating **several exposure** or **confounder** effects.

Contents

1. Introduction to logistic regression
2. Confounding
3. Effect modification
4. Comparing different logistic regression models

Introduction to Logistic Regression

Simple logistic regression model

$$\text{Let } Y = \begin{cases} 1, & \text{Person diseased} \\ 0, & \text{Person healthy} \end{cases}$$

$$\text{and let } x = \begin{cases} 1, & \text{if exposure present} \\ 0, & \text{if exposure not present} \end{cases}$$

The simple model is

$$\text{logit}(p_x) = \log \frac{p_x}{1 - p_x} = \alpha + \beta x$$

where

$$p_x = Pr(Y = 1|x)$$

Interpretation of parameters α and β

$$\log \frac{p_x}{1-p_x} = \alpha + \beta x$$

$$x = 0: \quad \text{logit}(p_0) = \log \frac{p_0}{1-p_0} = \alpha \tag{1}$$

$$x = 1: \quad \text{logit}(p_1) = \log \frac{p_1}{1-p_1} = \alpha + \beta \tag{2}$$

now

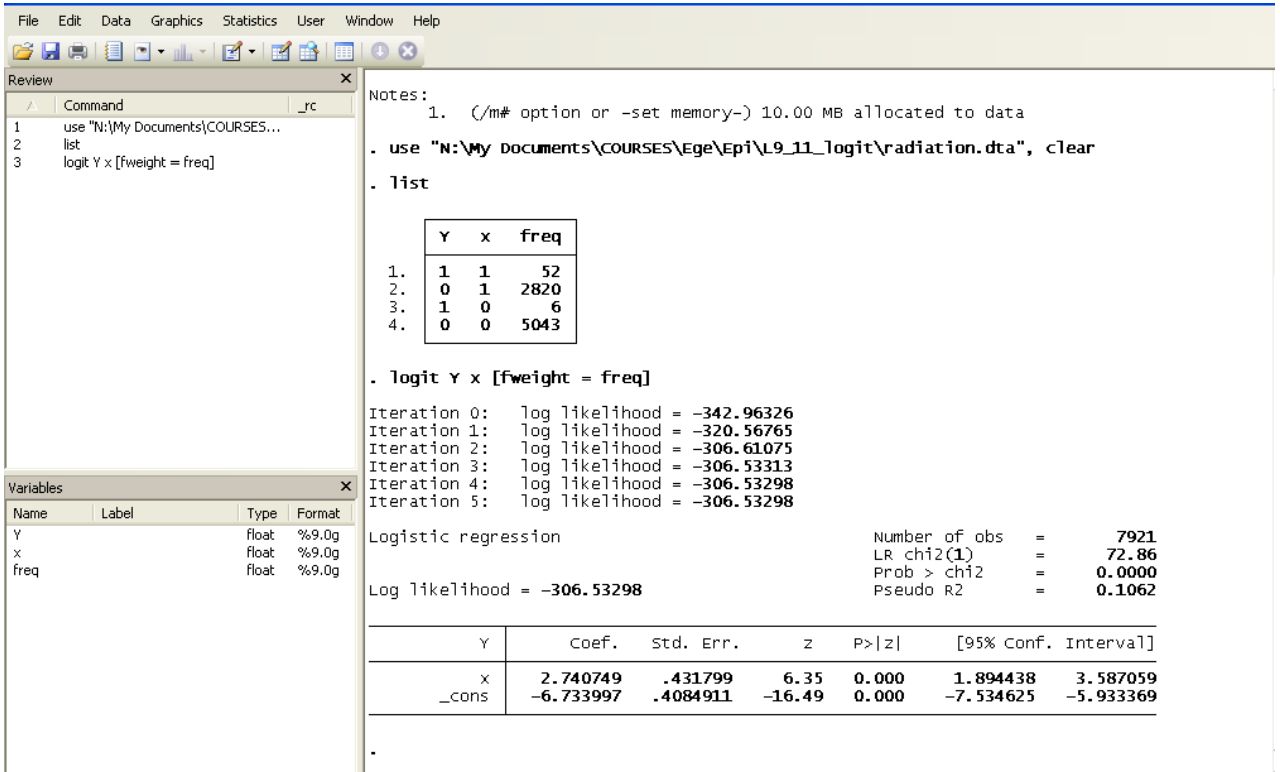
$$(2) - (1) = \underbrace{\log \frac{p_1}{1-p_1} - \log \frac{p_0}{1-p_0}}_{\log \frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}} = \log OR} = \alpha + \beta - \alpha = \beta$$

$$\log OR = \beta \Leftrightarrow OR = e^{\beta}$$

Example: Radiation Exposure and Tumor Development

	cases	non-cases	
E	52	2820	2872
NE	6	5043	5049

Analysis in stata:



Confounding:

Consider the following illustrative example:

	cases	non-cases	
E	60	1100	1160
NE	1501	3100	4601

OR

odds ratio:

$$OR = \frac{60 \times 3100}{1501 \times 1100} = 0.1126$$

This suggests that exposure has a protective effect on disease

However, suppose the data was actually from two strata.

Stratified Data:

Stratum 1:

	cases	non-cases	
E	50	100	150
NE	1500	3000	4500

$$OR = \frac{50 \times 3000}{100 \times 1500} = 1$$

Stratum 2:

	cases	non-cases	
E	10	1000	1010
NE	1	100	101

$$OR = \frac{10 \times 100}{1000 \times 1} = 1$$

	Y	E	S	freq
1.	1	1	0	50
2.	0	1	0	100
3.	1	0	0	1500
4.	0	0	0	3000
5.	1	1	1	10
6.	0	1	1	1000
7.	1	0	1	1
8.	0	0	1	100

The logistic regression model for simple confounding

$$\log \frac{p_{\mathbf{x}}}{1 - p_{\mathbf{x}}} = \alpha + \beta E + \gamma S$$

where

$$\mathbf{x} = (E, S)$$

is the covariate combination of exposure E and stratum S

Interpretation of model parameters

Stratum 1:

$$\log \frac{p_x}{1-p_x} = \alpha + \beta E + \gamma S$$

$$E = 0, S = 0 : \log \frac{p_{0,0}}{1-p_{0,0}} = \alpha \quad (3)$$

$$E = 1, S = 0 : \log \frac{p_{1,0}}{1-p_{1,0}} = \alpha + \beta \quad (4)$$

now

$$(4) - (3) = \log OR_1 = \alpha + \beta - \alpha = \beta$$

$$\log OR = \beta \Leftrightarrow OR = e^{\beta}$$

the log-odds ratio in the first stratum is β

Interpretation of model parameters

Stratum 2:

$$\log \frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}} = \alpha + \beta E + \gamma S$$

$$E = 0, S = 1 : \log \frac{p_{0,1}}{1-p_{0,1}} = \alpha + \gamma \quad (5)$$

$$E = 1, S = 1 : \log \frac{p_{1,1}}{1-p_{1,1}} = \alpha + \beta + \gamma \quad (6)$$

now:

$$(6) - (5) = \log OR_2 = \alpha + \beta + \gamma - \alpha - \gamma = \beta$$

the log-odds ratio in the second stratum is also β

The confounding model assumes **identical exposure effects** in each stratum

```
(crude analysis) Logistic regression
Log likelihood = -3141.5658
------------------------------------------------------------------------------
           Y | Odds Ratio   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------------
           E |   .1126522   .0153479       .0862522    .1471326
------------------------------------------------------------------------------

(adjusted for confounder) Logistic regression

Log likelihood =-3021.5026
------------------------------------------------------------------------------
           Y | Odds Ratio   Std. Err.      [95% Conf. Interval]
-------------+----------------------------------------------------------------
           E |          1   .1736619       .7115062    1.405469
           S |        .02   .0068109       .0102603    .0389853
------------------------------------------------------------------------------
```

Effect modification

Consider the following data on passive smoking and lung cancer:

	cases	non-cases	
E	52	121	173
NE	54	150	204

odds ratio:

$$OR = \frac{52 \times 150}{54 \times 121} = 1.19$$

However, suppose the above is actually **combined** data for males and females

Stratified analysis:

Stratum 1 (females):

	cases	non-cases	
E	41	102	143
NE	26	71	97

$$OR = \frac{41 \times 71}{26 \times 102} = 1.10$$

Stratum 2 (males):

	cases	non-cases	
E	11	19	30
NE	28	79	107

$$OR = \frac{11 \times 79}{19 \times 28} = 1.63$$

interpretation:

The effect is different for males and females

The logistic regression model for effect modification

$$\log \frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}} = \alpha + \beta E + \gamma S + \underbrace{(\beta\gamma)}_{\text{effect modif. par.}} \qquad E \times S$$

where

$$\mathbf{x} = (E, S)$$

is the covariate combination of exposure E and stratum S

Interpretation of model parameters

Stratum 1:

$$\log \frac{p_x}{1-p_x} = \alpha + \beta E + \gamma S + (\beta\gamma) E \times S$$

$$E = 0, S = 0 : \log \frac{p_{0,0}}{1-p_{0,0}} = \alpha \quad (7)$$

$$E = 1, S = 0 : \log \frac{p_{1,0}}{1-p_{1,0}} = \alpha + \beta \quad (8)$$

now

$$(8) - (7) = \log OR_1 = \alpha + \beta - \alpha = \beta$$

$$\log OR = \beta \Leftrightarrow OR = e^{\beta}$$

the log-odds ratio in the first stratum is β

Interpretation of model parameters

Stratum 2:

$$\log \frac{p_{\mathbf{x}}}{1-p_{\mathbf{x}}} = \alpha + \beta E + \gamma S + (\beta\gamma) E \times S$$

$$E = 0, S = 1 : \log \frac{p_{0,1}}{1-p_{0,1}} = \alpha + \gamma \quad (9)$$

$$E = 1, S = 1 : \log \frac{p_{1,1}}{1-p_{1,1}} = \alpha + \beta + \gamma + (\beta\gamma) \quad (10)$$

now:

$$(10) - (9) = \log OR_2 = \alpha + \beta + \gamma + (\beta\gamma) - \alpha - \gamma = \beta + (\beta\gamma)$$

$$\log OR = \beta \Leftrightarrow OR = e^{\beta + (\beta\gamma)}$$

the log-odds ratio in the second stratum is $\beta + (\beta\gamma)$

The effect modification model allows for **different effects** in the strata

Data from passive smoking and LC example are as follows:

	Y	E	S	ES	freq
1.	1	1	0	0	41
2.	0	1	0	0	102
3.	1	0	0	0	26
4.	0	0	0	0	71
5.	1	1	1	1	11
6.	0	1	1	1	19
7.	1	0	1	0	28
8.	0	0	1	0	79

CRUDE EFFECT MODEL

Logistic regression

Log likelihood = -223.66016

Y	Coef.	Std. Err.	z	P>|z|
E	.1771044	.2295221	0.77	0.440
_cons	-1.021651	.1586984	-6.44	0.000

CONFOUNDING MODEL

Logistic regression

Log likelihood = -223.56934

Y	Coef.	Std. Err.	z	P>|z|
E	.2158667	.2472221	0.87	0.383
S	.1093603	.2563249	0.43	0.670
_cons	-1.079714	.2101705	-5.14	0.000

EFFECT MODIFICATION MODEL

Logistic regression

Log likelihood = -223.2886

Y	Coef.	Std. Err.	z	P>|z|
E	.0931826	.2945169	0.32	0.752
S	-.03266	.3176768	-0.10	0.918
ES	.397517	.5278763	0.75	0.451
_cons	-1.004583	.2292292	-4.38	0.000

interpretation of crude effects model:

$$\log OR = 0.1771 \Leftrightarrow OR = e^{0.1771} = 1.19$$

interpretation of confounding model:

$$\log OR = 0.2159 \Leftrightarrow OR = e^{0.2159} = 1.24$$

interpretation of effect modification model:

$$\text{Females:} \quad \log OR_1 = 0.0932 \Leftrightarrow OR_1 = e^{0.0932} = 1.10$$

$$\text{Males:} \quad \log OR_2 = 0.0932+0.3975 \Leftrightarrow OR_2 = e^{0.0932+0.3975} = 1.63$$

Model evaluation:

The likelihood approach:

$$L = \prod_{i=1}^{n} p_{x_i}^{y_i} (1 - p_{x_i})^{1-y_i}$$

is called the **likelihood** for models

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \begin{cases} \alpha + \beta E_i + \gamma S_i + (\beta\gamma) E_i \times S_i, \ (M_1) \\ \alpha + \beta E_i + \gamma S_i, \ (M_0) \end{cases}$$

where M_1 is the effect modification model and

M_0 is the confounding model

Model evaluation using the likelihood ratio:

Let

$$L(M_1) \text{ and } L(M_0)$$

be the **likelihood** for models M_1 and M_0

Then

$$LRT = 2 \log L(M_1) - 2 \log L(M_0) = 2 \log \frac{L(M_1)}{L(M_0)}$$

is called the **likelihood ratio** for models M_1 and M_0

LRT has a **chi-square distribution with 1** df under M_0

Example: passive smoking and LC:

model	log-likelihood	LRT
crude	-223.66016	-
homogeneity	-223.56934	0.1816
effect modification	-223.2886	0.5615

note:

for valid comparison on chi-square scale: models must be **nested**

Model evaluation in general:

Consider the likelihood

$$L = \prod_{i=1}^{n} p_{x_i}^{y_i}(1 - p_{x_i})^{1-y_i}$$

for a general model with k **covariates**:

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} \ (M_0)$$

and for the model with an **additional** p **covariates**:

$$\log \frac{p_{x_i}}{1 - p_{x_i}} = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$$

$$+\beta_{k+1} x_{i,k+1} + ... + \beta_{k+p} x_{i,k+p} \ (M_1)$$

Again let

$$L(M_1) \text{ and } L(M_0)$$

be the **likelihood** for models M_1 and M_0

Then the **likelihood ratio**

$$LRT = 2 \log L(M_1) - 2 \log L(M_0) = 2 \log \frac{L(M_1)}{L(M_0)}$$

has a **chi-square distribution with p** *df* under M_0

Meta-Analysis:

Investigating the results from several independent studies with the purpose of an integrative analysis

Example: BCG vaccine against tuberculosis, Colditz et al. 1974, JAMA

The data consists of 13 studies with each study containing

- TB cases for BCG intervention
- number at risk for BCG intervention
- TB cases for control
- number at risk for control

Also two covariates are given: *year of study* and *latitude expressed in degrees from the equator*

Data analysis

This data can be analyzed by taking

- *TB case* as disease occurrence response
- *intervention* as exposure
- *study* as confounder

			intervention		control	
study	year	latitude	TB cases	total	TB cases	total
1	1933	55	6	306	29	303
2	1935	52	4	123	11	139
3	1935	52	180	1541	372	1451
4	1937	42	17	1716	65	1665
5	1941	42	3	231	11	220
6	1947	33	5	2498	3	2341
7	1949	18	186	50634	141	27338
8	1950	53	62	13598	248	12867
9	1950	13	33	5069	47	5808
10	1950	33	27	16913	29	17854
11	1965	18	8	2545	10	629
12	1965	27	29	7499	45	7277
13	1968	13	505	88391	499	88391

Study	RR	[95% Conf.	Interval]	M-H Weight
1	.2048682	.0862974	.4863523	14.57143
2	.4109387	.1343016	1.257398	5.164122
3	.4556111	.3871323	.536203	191.5949
4	.2537655	.1494209	.4309765	32.99024
5	.2597403	.0734426	.9186087	5.634146
6	1.561916	.3736891	6.528374	1.548667
7	.7122268	.5725137	.8860348	91.56356
8	.2365605	.1792809	.3121408	127.4251
9	.8044895	.5162931	1.253558	21.90337
10	.9828351	.5821375	1.659341	14.10754
11	.197721	.0783566	.4989192	8.018273
12	.6253663	.3925763	.9961964	22.83805
13	1.012024	.894572	1.144897	249.5
Crude	.6138209	.5676759	.6637168	
M-H combined	.6352672	.5881287	.6861838	

BUT:

Test of homogeneity (M-H chi2(12) = 152.568 Pr>chi2 = 0.0000

Conclusions from meta-analysis of BCG and TB

- most studies show preventive effect
- crude and MH-adjusted estimates are rather close
- **but:** homogeneity test is significant

what are the reasons for this heterogeneity in RR?

need to look at

- year effect
- latitude effect

This can be done using logistic regression

Lecture 10
Poisson Regression

James Gallagher
Director, Statistical Services Centre
University of Reading
Reading
UK

May 2011

Contents

The Poisson Distribution

- *Count* data may follow such a distribution, at least approximately

  Examples: Number of

  - Deaths, diseased cases, hospital admissions and so on ….

- Poisson distribution: $Y \sim Poi(\mu)$

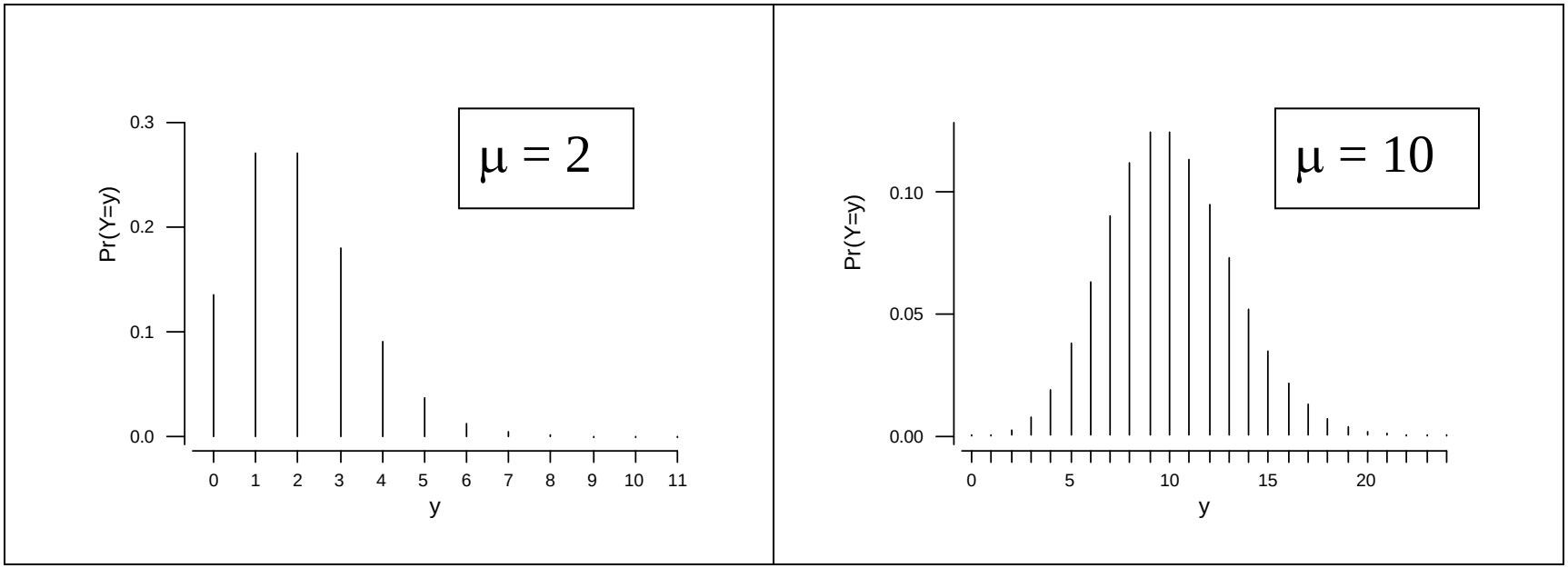
  Y has density function:

$$\Pr(Y = y) = \begin{cases} \dfrac{\mu^y \exp(-\mu)}{y!} & \text{for } y = 0, 1, 2, ..., +\infty \\ 0 \quad \text{otherwise} \end{cases}$$

  where $\mu > 0$.

Properties of the Poisson Distribution

- E(Y) = Var(Y) = μ
- Shape
  - Skewed for small μ
  - Approximately *normal* for large μ



Introduction to Poisson Regression

Example: BELCAP dental epidemiological study

- A prospective study of school-children from an urban area of Belo Horizonte, Brazil
  - The Belo Horizonte caries prevention (BELCAP) study
- The aim of the study was to compare different methods to prevent caries
- Response (outcome) variable=DMFT index. (No. of decayed, missing or filled teeth.)
  - DMFT index was calculated at the start of the study and 2 years later
- Potential confounders: sex, ethnicity, baseline dental score

For simplicity consider only

y = DMFT2, post-intervention DMFT index

and

two interventions: control (i=0) and oral hygiene (i=1)

Poisson regression model:

(1) $y \sim Poi(\mu)$

(2) $\log(\mu) = \alpha + \text{intervent}_i$; $\text{intervent}_0 = 0$

Notes

- Other functions of μ can be modelled but $\log(\mu)$ will always result in $\hat{\mu} > 0$.
- $\alpha + \text{intervent}_i$ is known generically as the **linear predictor**.
- The model is also called a **log-linear model**.

But why can't we use a linear regression model (general linear model)? There are problems:

(a) For a Poisson random variable E(Y)=Var(Y). This violates the constancy of variance assumption.

(b) A linear regression model assumes we are dealing with normal distributions – the Poisson may not look very normal!

(c) Linear regression may give negative predicted means.

Continuing with the Poisson regression model…

Interpretation of the Poisson Regression Model

For children in the **control** group the model says:

$$\log(\mu) = \alpha + \text{intervent}_0 = \alpha$$

$$\mu = \exp(\alpha)$$

For children in the **oral hygiene** group the model says:

$$\log(\mu) = \alpha + \text{intervent}_1$$

$$\mu = \exp(\alpha + \text{intervent}_1)$$

Hence, $$\frac{\mu|_{\text{oral}}}{\mu|_{\text{control}}} = \exp(\text{intervent}_1)$$

$\exp(\text{intervent}_1)$=ratio of true means(oral hygiene/control)=effect measure

Note the interpretation:

$\exp(\text{intervent}_1) < 1$: intervention effect, oral hygiene doing better

$\exp(\text{intervent}_1) = 1$: no intervention effect

$\exp(\text{intervent}_1) > 1$: intervention effect, oral hygiene doing worse

Stata refers to $\exp(\text{intervent}_1)$ as an incidence rate ratio, so intervent_1 is a log incidence rate ratio.

Stata fits the model using the method of maximum likelihood.
[**Stata**: Statistics→Count outcomes→Poisson regression]

```
. poisson dmft2 i.intervent

Iteration 0:   log likelihood = -505.90325
Iteration 1:   log likelihood = -505.90325

Poisson regression                                Number of obs   =        259
                                                  LR chi2(1)      =       9.11
                                                  Prob > chi2     =     0.0025
Log likelihood = -505.90325                       Pseudo R2       =     0.0089
------------------------------------------------------------------------------
       dmft2 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 1.intervent |  -.2620432   .0874031    -3.00   0.003    -.4333501   -.0907363
       _cons |   .8525362   .0559893    15.23   0.000     .7427993    .9622731
------------------------------------------------------------------------------
```

$$\widehat{\text{intervent}}_1 = -0.262,\ \hat{\alpha} = 0.853$$

$$\exp(\widehat{\text{intervent}}_1) = \exp(-0.262) = 0.77$$

Mean DMFT index for the oral hygiene method is estimated to 77% of that for the control.

Confidence Intervals

An approximate [Wald type] 95% confidence interval for the ratio of true means may be calculated using the Stata output.

Stage 1

From the output, an approximate 95% CI for β is

$$-0.433 \text{ to } -0.0907$$

Stage 2

An approximate 95% CI for $\exp(\beta)$ is then

$$\exp(-0.433) \text{ to } \exp(-0.0907)$$

$$\text{i.e. } 0.65 \text{ to } 0.91$$

Hypothesis Testing: Model Comparisons

If there is truly no intervention effect then $\beta = 0$, i.e. $\exp(\beta)=1$. This leads to the hypotheses:

H_0: $\beta = 0$ (No intervention effect)

vs.

H_1: $\beta \neq 0$ (There is an intervention effect)

Stata gives an approximate likelihood ratio test for this:

```
LR chi2(1)      =    9.11
Prob > chi2     =  0.0025
```

Likelihood ratio, statistic X^2= 9.11 (1 df), p-value = 0.0025. Hence, there is evidence for an intervention effect. Oral hygiene improves dental status.

Notes

- The previous likelihood ratio test is comparing the fit of two nested models to the data:
  - (1) $\log(\mu) = \alpha$
  - (2) $\log(\mu) = \alpha + \text{intervent}_i$

Model	$\log \hat{L}$
(1) $\log(\mu) = \alpha$	–510.456
(2) $\log(\mu) = \alpha + \text{intervent}_i$	–505.903
$X^2 = 2[\log \hat{L}(2) - \log \hat{L}(1)] = 9.11$ (1 df)	

Confounding and Effect Modification

- Ignoring the pre-intervention (baseline) DMFT index is clearly not a good idea
- How can the intervention effect be adjusted for baseline?
- Let DMFT1 = Pre-intervention DMFT index
- Böhning et al. (1999) uses log(DMFT1+0.5) as a linear effect…

Poisson regression model:

(1) y, DMFT2~Poi(μ)

(2) $\log(\mu) = \alpha + \beta \times \log(\text{DMFT1}) + \text{intervent}_i$; $\text{intervent}_0 = 0$

Hence, the intervention effect, adjusted for baseline DMFT is

$$\frac{\mu|_{\text{oral}}}{\mu|_{\text{control}}} = \exp(\text{intervent}_1)$$

- Perform statistical analysis as before
- Similarly, effect modification may be assessed by introducing an interaction term into the above model

Effect of Adjusting for Pre-intervention Dental Status

Analysis	Intervention effect (Ratio of means)	95% CI	p-value (LRT)
Unadjusted	0.77	0.65 to 0.91	0.0025
Adjusted	0.93	0.78 to 1.10	0.40

Ignoring pre-intervention dental status gives a misleading result.

Further, there is no evidence for effect modification.

Extensions

- The models discussed naturally extend, to allow the inclusion of other factors
  - E.g. the potential confounders sex and ethnicity
- Interactions (effect modifications) may also be assessed
- Poisson regression may also be used to model rates and ratios. See Practical 3

Appendix

The BELCAP Study

Background

- Dental epidemiological study
- A prospective study of school-children from an urban area of Belo Horizonte, Brazil
    - The Belo Horizonte caries prevention (BELCAP) study
- The aim of the study was to compare different methods to prevent caries

Details

- Children were all 7 years-old and from a similar socio-economic background
    - See Mendonça and Böhning (1994) and Mendonça (1995)

- Interventions:
  - Control,
  - Oral health education,
  - School diet enriched with rice bran,
  - Mouthwash,
  - Oral hygiene,
  - All four methods together
- Response (outcome) variable=DMFT index. (No. of decayed, missing or filled teeth.)
  - DMFT index was calculated at the start of the study and 2 years later
  - Only the 8 deciduous molars were considered
- Potential confounders: sex, ethnicity
- Data on 797 children analysed by Böhning et al. (1999)

- Lesions of the tooth were also included in the index. Graded as:
  - 0 = healthy,
    1 = light chalky spot,
    2 = thin brown-black line,
    3 = damage, not larger than 2mm wide,
    4 = damage, wider than 2mm
  - The $D_{1\text{-}4}MFT$ index. Pilz (1985)
- In the BELCAP study a lesion graded 1-4 contributed 1 to the DMFT index

References

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L. and Kirchner, U. (1999) The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. *Journal of the Royal Statistical Society (Series A)*, **162**, 195-209.

Breslow, N.E. and Day, N.E. (1987). *Statistical Methods in Cancer Research. Volume II - The Design and Analysis of Cohort Studies.* International Agency for Research in Cancer, Lyon.

Mendonça, L. (1995). Longitudinalstudie zu kariespräventiven Methoden, durchgeführt bei 7- bis 10-jährigen urbanen Kindern in Belo Horizonte (Brasilien). *Dissertation*. Free University of Berlin, Berlin.

Mendonça, L. and Böhning, D. (1994). Die Auswirkung von Gesundheitsunterricht und Mundspülung mit Na-Fluorid auf die Prävention von Zahnkaries: eine Kohortenstudie mit urbanen Kindern in Brasilien. *39th A. Conf. Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie, Dresden, September 18th-25th*.

Pilz, M.E.W. (1985). *Praxis der Zahnerhaltung und Oralen Prävention.* Munich: Hanser.