Lecture 13

Design and Analysis Techniques for Epidemiologic Studies

awilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Population Health

Learning Objectives

1. Define study designs
2. Measures of effects for categorical data
3. Confounders and effects modifications
4. Stratified analysis (Mantel Haenszel statistic, multiple logistic regression)
5. Use of Computer Program for Logistic regression (in Lab)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Study Design

To call in the statistician after the experiment is done may be no more than asking him to perform a postmortem examination:

He may be able to say what the experiment died of.

- Sir Ronald A. Fisher

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Study Design

- Designing a study is possibly the most important role for the statistical expert in a research team
- Design of the study (one need to have a good knowledge about the exposures, disease of interest and study objectives and hypotheses)
  - Sampling design (selection of subjects)
  - Sample size calculation (new study) or power calculation (if study is from existing data)
  - Analysis plan

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Study Design

Most clinical studies can be broadly classified into one of two categories, namely

– Experimental Studies (Clinical Trials): Experimental units are randomly assigned to a specific level of the exposure (intervention).

– Observational Studies : Data are collected in a given situation, without intentional interference (randomization) by the observer.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Experimental Study

- Gold-Standard for the proof of an effect of a treatment (required for registration-FDA)

- Four Phases in Drug research
  - Phase I (How the body copes with the drug , the safe dose range, the side effects, some therapeutic effect, few subjects)
  - Phase II (If the new treatment works well enough to test in phase 3, More about side effects and how to manage them, More about the most effective dose to use, more subjects)
  - Phase III (The new treatment or procedure is compared with the standard treatment, Different doses or ways of giving a standard treatment, sample size is large)
  - Phase IV (More about the side effects and safety of the drug, what the long term risks and benefits are, how well the drug works when it's used more widely than in clinical trials)

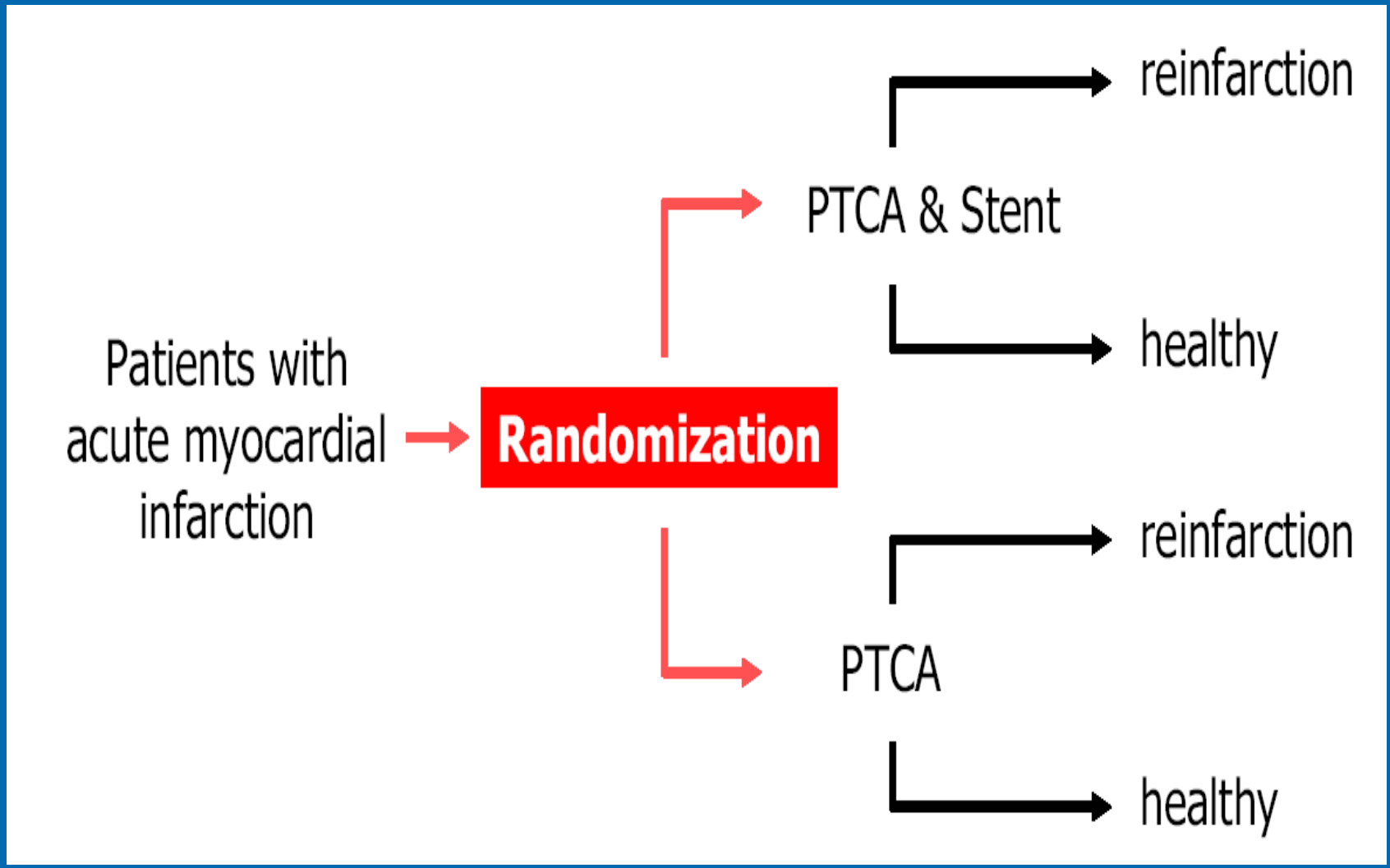
wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Experimental Study

- Randomization protects against bias in assignment to groups.
- Blinding protects against bias in outcome assessment or measurement.
- Control for (major) sources of variability, although not necessarily reflecting real life conditions
- Expensive in terms of time and money

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Experimental Study-Benefit of additional Stent in MI-Therapy



wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Some definitions for the example

- Percutaneous means access to the blood vessel is made through the skin
- Transluminal means the procedure is performed within the blood vessel
- Coronary specifies that the coronary artery is being treated
- Angioplasty means "to reshape" the blood vessel (with balloon inflation)
- A stent is a small, metal coil that helps to keep a “ballooned” artery open.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

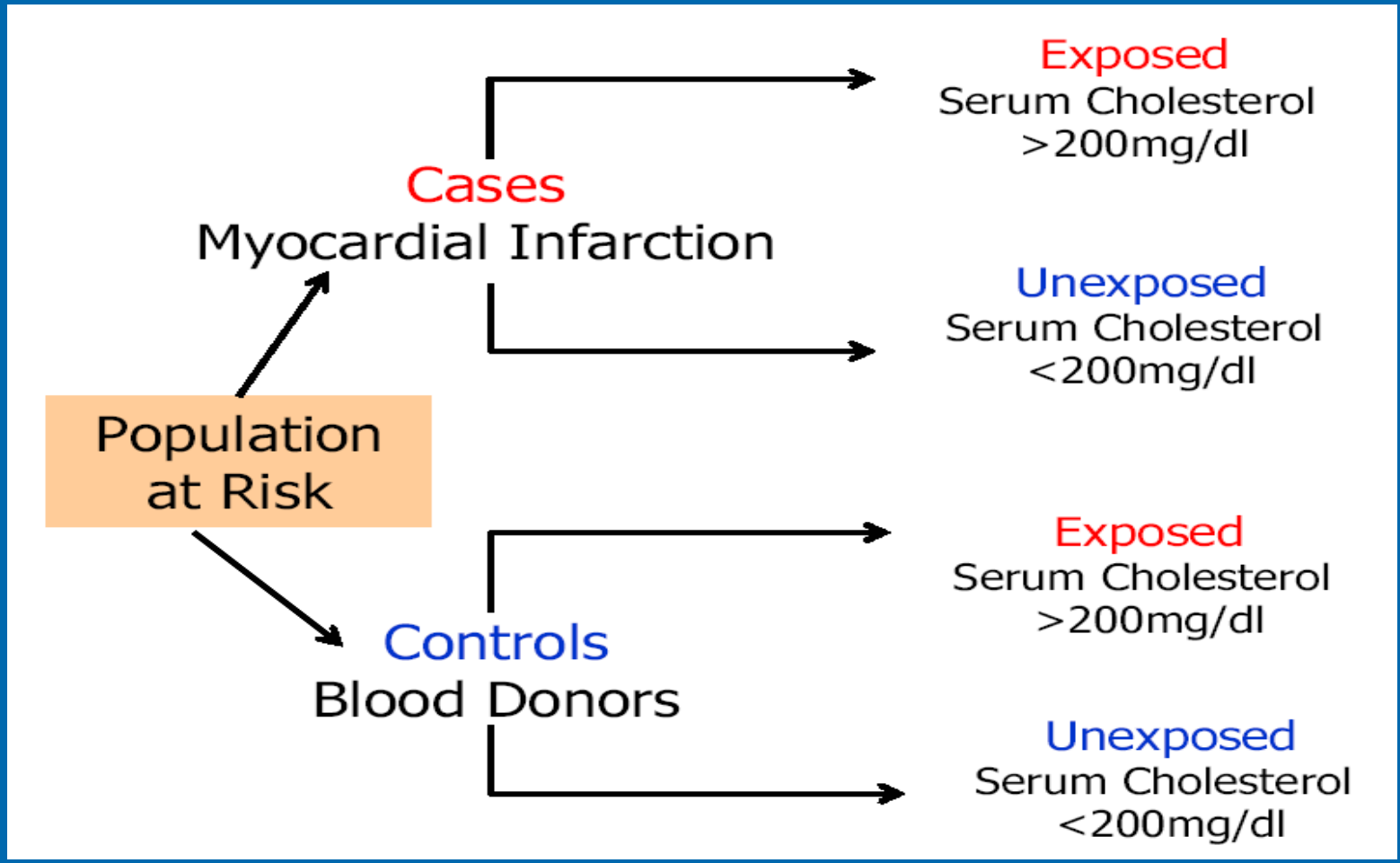
Observational Study

- Survey to characterize a target population with respect to specific parameters
- May include all population members (census)
- Typically includes only a part of the population

  (sample) because of time, cost and other practical constraints

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Observational Study-Risk for MI and High Serum Cholesterol



wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Observational Study most likely used in Epidemiology

- Types of study
  - Cross-sectional study
  - Case-control study (retrospective)
  - Cohort study (Prospective)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

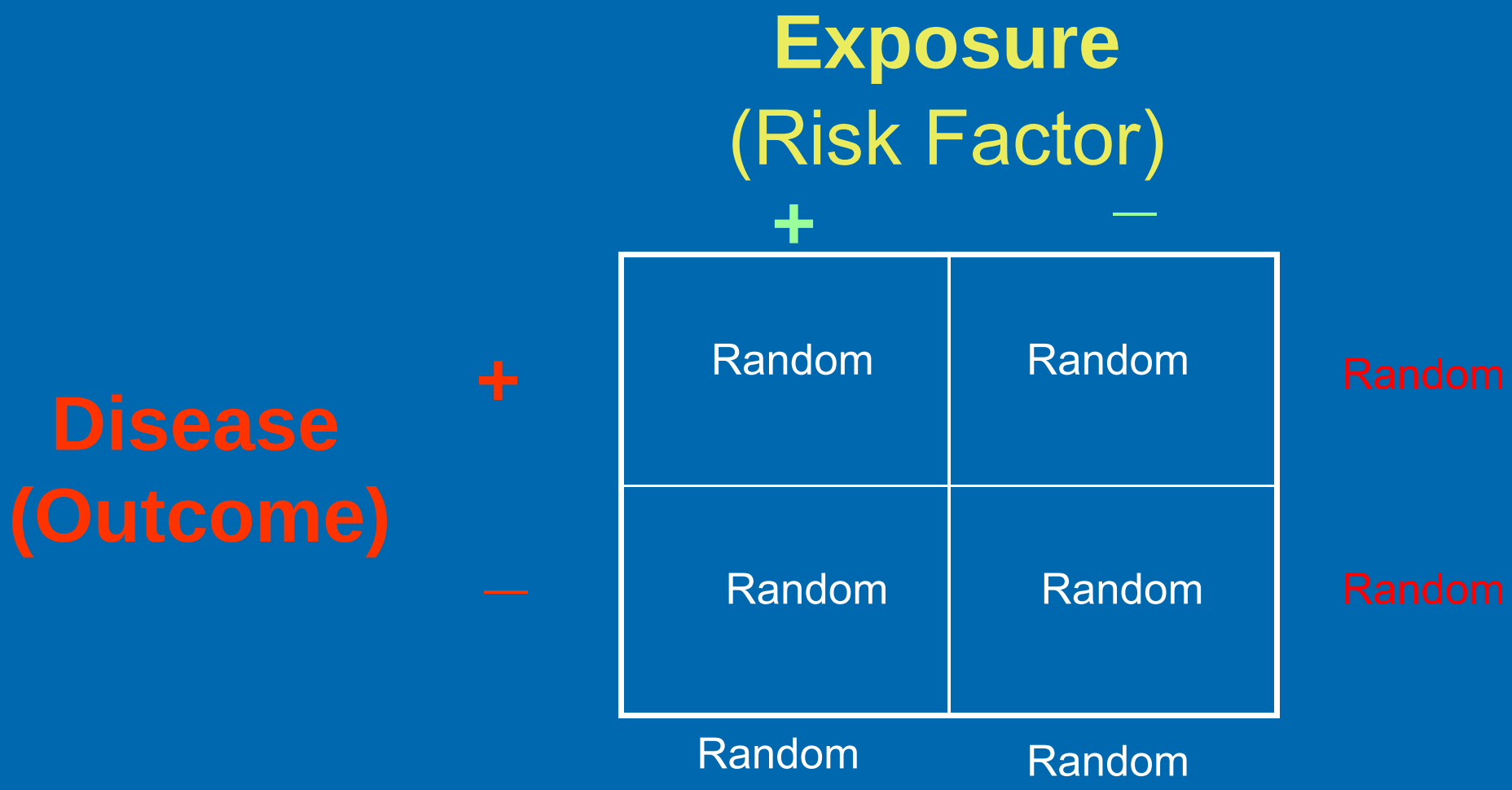
Cross-Sectional Studies

→ Begin with "Cross-sectional" sample

→ Determine Exposure and Disease at same time

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Cross-Sectional Studies



wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

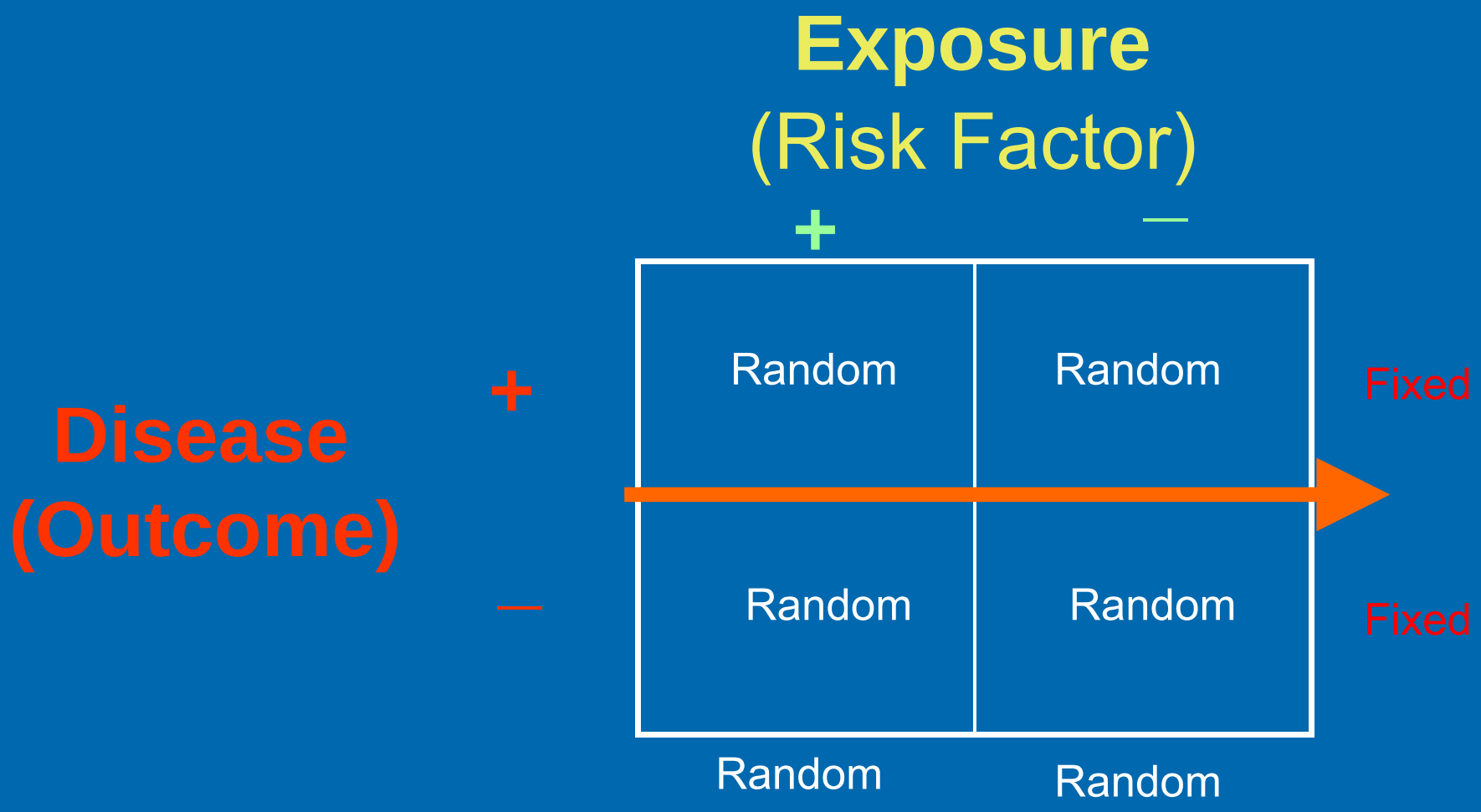
Case-Control Studies

→ Begin with sample of "Cases and Controls"

→Start with Disease status, then assess and compare Exposures in cases vs. controls.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Case-control Studies



wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

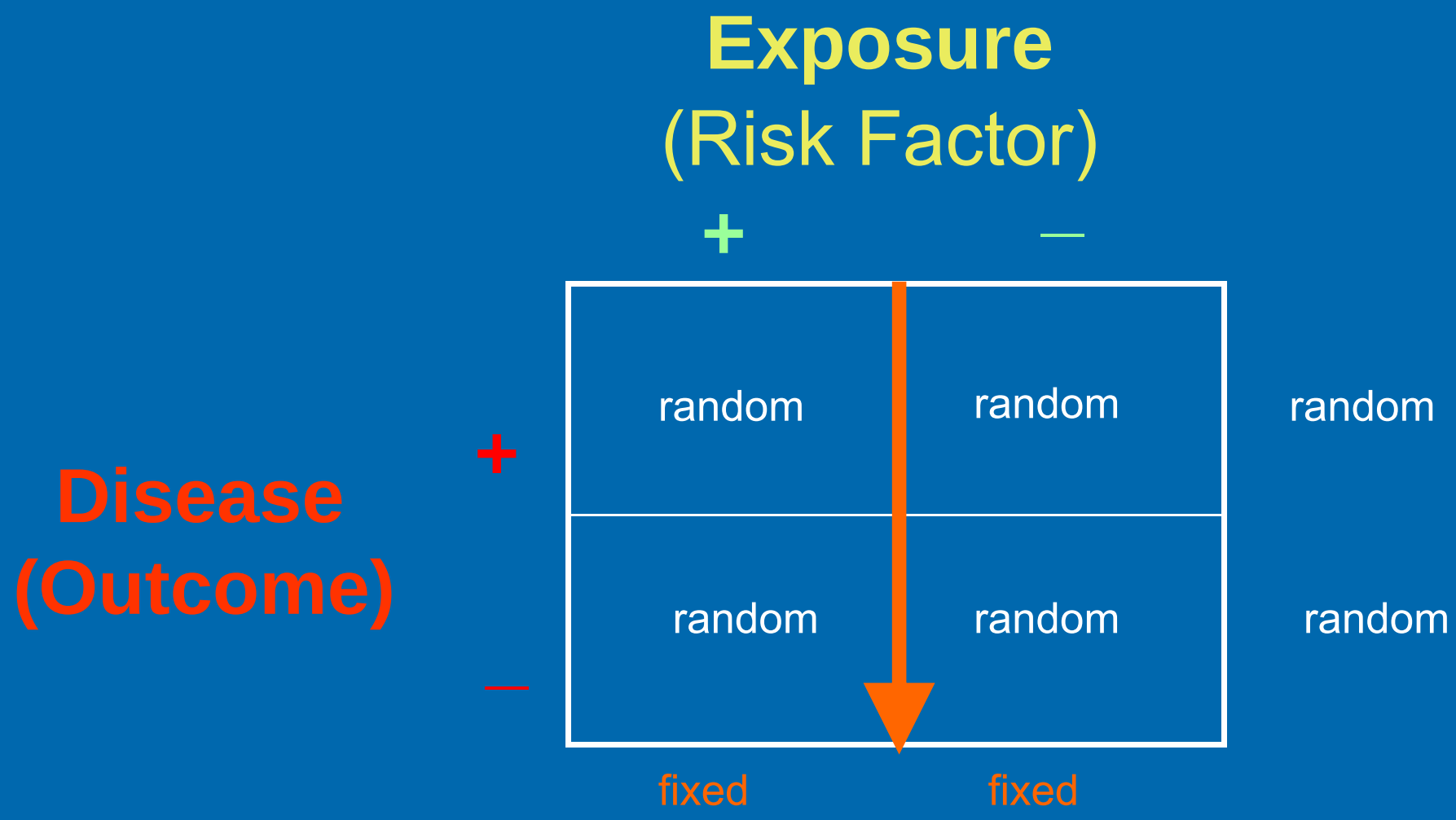
Cohort Studies

→Begin with sample → "Healthy Cohort" (i.e., subjects without the outcome *yet*)

→Start with Exposure status, then compare subsequent disease experience in exposed vs. unexposed.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Cohort Studies



wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Measures of effects for categorical data

- Depends on study design
  - Prospective study: Incidence of disease (risk difference, relative risk, odds ratio of disease)
  - Cross-sectional: Prevalence of disease (risk difference, relative risk, odds ratio of disease)
  - Case-cohort: study of exposure (odds ratio of exposure)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

2X2 tables notations

Exposure
(Risk Factor)

Disease (Outcome)	E	$\bar{E}$	
D	a	c	m_1
$\bar{D}$	b	d	m_2
	n_1	n_2	N

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Risk difference

Only for cross-sectional and cohort studies
Measured the attributable risk due to exposure

$$RD = P(D|E) - P(D|\overline{E})$$

$$\hat{p}_1 = a/n_1 \qquad \hat{p}_2 = c/n_2$$

$$\hat{RD} = \hat{p}_2 - \hat{p}_1$$

$$se(\hat{RD}) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{ab}{n_1^3} + \frac{cd}{n_2^3}}$$

awilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Risk Difference

- A Confidence interval for the true risk difference can be easily constructed

- If the confidence interval contains 0, there is no evidence from data to suggest that the probability of the disease differs for the exposed and the unexposed groups

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Relative Risk

Only for cross-sectional and cohort studies: Ratio of the probability that the outcome characteristic is present for one group, relative to the other

$$RR = \frac{P(D \mid E)}{P(D \mid \overline{E})}$$

The range of RR is [0, ∞). By taking the logarithm, we have (- ∞, +∞) as the range for ln(RR) and a better approximation to normality for the estimated $\ln(\hat{RR})$:

$$\ln(\hat{RR}) = \ln\left(\frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \overline{E})}\right) = \ln\left(\frac{a/n_1}{c/n_2}\right)$$

$$\ln(\hat{RR}) \sim N\left(\ln(p_1/p_2),\ \frac{1-p_1}{p_1 n_1} + \frac{1-p_2}{p_2 n_2}\right)$$

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Relative Risk

	Cold - Y	Cold - N	Total
Vitamin C	17	122	139
Placebo	31	109	140
Total	48	231	279

The estimated relative risk is:

$$\hat{RR} = \frac{\hat{P}(D \mid E)}{\hat{P}(D \mid \overline{E})} = \frac{17/139}{31/140} = 0.55$$

$$\ln(\hat{RR}) \pm Z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{1-\hat{p}_1}{\hat{p}_1 n_1} + \frac{1-\hat{p}_2}{\hat{p}_2 n_2}}$$

We can obtain a confidence interval for the relative risk by first obtaining a confidence interval for the log-RR: and exponentiating the endpoints of the CI.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Odds Ratio

- Odds of an event is the probability that disease occurs divided by the probability it does not occur.
- Can be computed for all study designs
- In cohort studies, we have the odds ratio for disease (fixed # of exposed and non exposed)
- In case-control studies, we have the odds ratio for exposure (fixed # of cases and controls)
- In cross-sectional, we have both the odds ratio for exposure and disease (random margins)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Odds Ratio - Disease

- Odds ratio is the odds of the event for exposed divided by the odds of the event for unexposed

- Sample odds of the outcome for each group:

$$odds_E = \frac{a}{b} \quad and \quad odds_{\bar{E}} = \frac{c}{d}$$

$$OR(disease) = \frac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))} = \frac{odds_E}{odds_{\bar{E}}} = \frac{ad}{bc}$$

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Odds Ratio-Exposure

we fixed the number of cases and controls then ascertained exposure status. The relative risk is therefore not estimable from these data alone. Instead of the relative risk we can estimate the exposure OR which Cornfield (1951) showed equivalent to the disease OR:

$$\frac{P(E|D)/(1-P(E|D))}{P(E|\bar{D})/(1-P(E|\bar{D}))} = \frac{P(D|E)/(1-P(D|E))}{P(D|\bar{E})/(1-P(D|\bar{E}))}$$

In other words, the odds ratio can be estimated regardless of the sampling scheme.

$$OR(disease) = OR(\text{exposure}) = \frac{ad}{bc}$$

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Odds Ratio-Relative risk

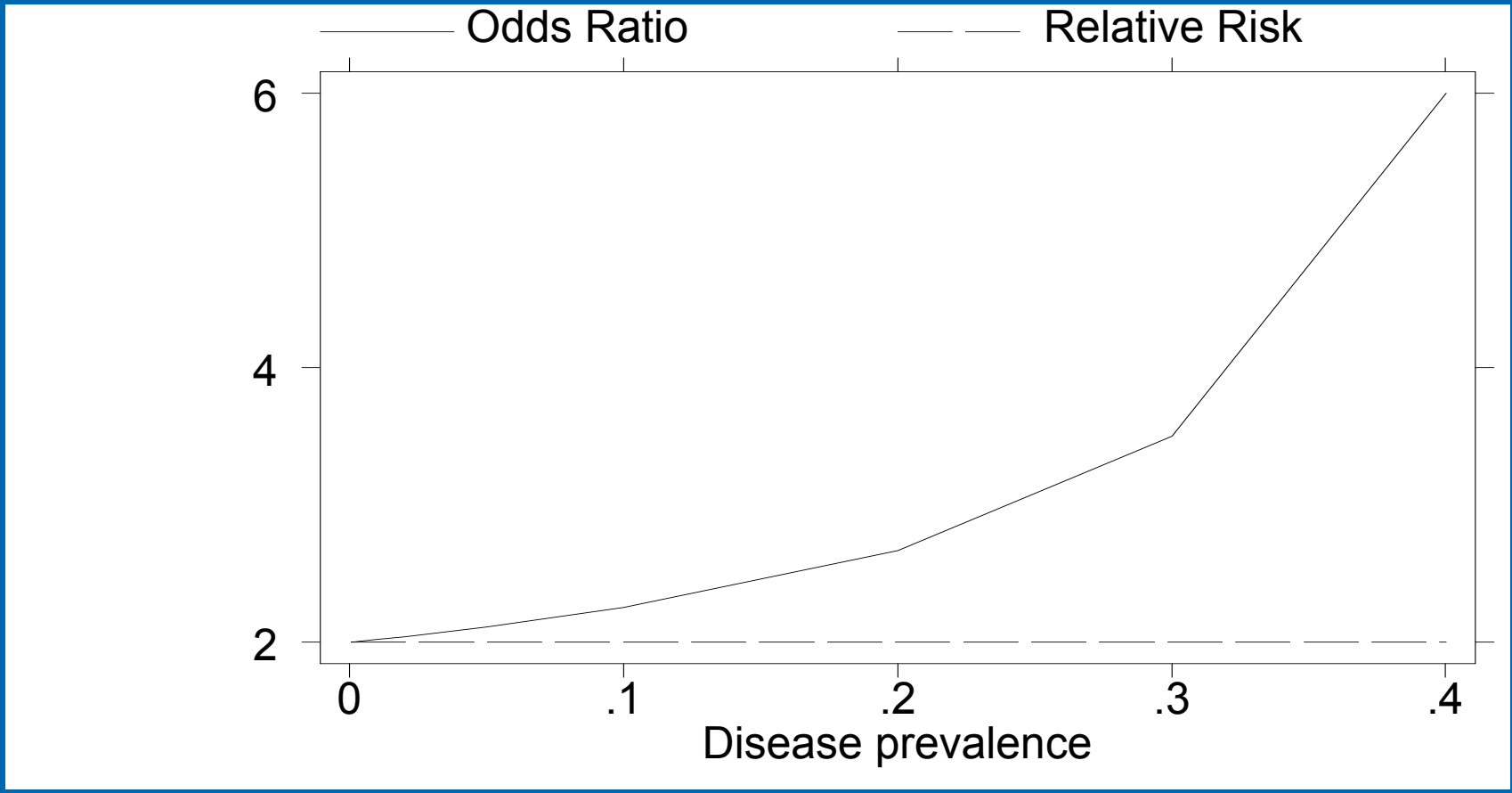
For rare diseases, the **disease odds ratio** approximates the relative risk:

$$\frac{P(D|E)/(1-P(D|E))}{P(D|\overline{E})/(1-P(D|\overline{E}))} \approx \frac{P(D|E)}{P(D|\overline{E})}$$

Since with case-control data we are able to effectively estimate the exposure odds ratio we are then able to equivalently estimate the disease odds ratio which for rare diseases approximates the relative risk.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Odds Ratio-Relative risk



wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Odds Ratio

The odds ratio has [0, ∞) as its range. The **log odds ratio** has (- ∞, +∞) as its range and the normal approximation is better as an approximation to the estimated log odds ratio.

$$\ln\left(\hat{OR}\right) \sim N\left(\ln(OR), \frac{1}{n_1 p_1} + \frac{1}{n_1 q_1} + \frac{1}{n_2 p_2} + \frac{1}{n_2 q_2}\right)$$

Confidence intervals are based upon:

$$\ln\left(\frac{ad}{bc}\right) \pm Z_{1-\frac{\alpha}{2}} \times \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Therefore, a (1 - α) confidence interval for the odds ratio is given by exponentiating the lower and upper bounds.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Example - NSAIDs and PAIN

- Case-Control Study (Retrospective)
  - Cases: 137 Self-Reporting Patients with back pain reduction
  - Controls: 401 Population-Based Individuals matched to cases wrt demographic factors

	Pain reduction	No pain reduc	Total
NSAID User	32	138	170
NSAID Non-User	105	263	368
Total	137	401	538

Source: Sival-Sears, et al (2004)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Example - NSAIDs and PAIN

$$OR = \frac{32(263)}{138(105)} = \frac{8416}{14490} = 0.58$$

$$\text{var}[\ln(OR)] = \frac{1}{32} + \frac{1}{138} + \frac{1}{105} + \frac{1}{263} = 0.0518$$

$$95\%\ CI: \quad (0.58e^{-1.96\sqrt{0.0518}}, 0.58e^{1.96\sqrt{0.0518}}) \equiv (0.37, 0.91)$$

Interval is entirely below 1, NSAID use appears to be lower among cases than controls

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Summary

RD = $p_1 - p_2$ = risk difference (null: RD = 0)

- also known as **attributable risk** or **excess risk**
- measures **absolute effect** – the proportion of cases among the exposed that can be attributed to exposure

RR = p_1 / p_2 = relative risk (null: RR = 1)

- measures **relative effect** of exposure
- bounded above by $1/p_2$

OR = $[p_1(1-p_2)]/[p_2(1-p_1)]$ = odds ratio (null: OR = 1)

- range is 0 to ∞
- approximates RR for rare events
- invariant of switching rows and cols
- key parameter in logistic regression

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Two main complications of analysis of single exposure effect

(1) Effect modifier- useful information

(2) Confounding factor - bias

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Effect modifier

- Variation in the magnitude of measure of effect across levels of a third variable.

- Effect modification is not a bias but useful information

Happens when RR or OR is different between strata (subgroups of population)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Effect modifier

- To study interaction between risk factors
- To identify a subgroup with a lower or higher risk
- To target public health action

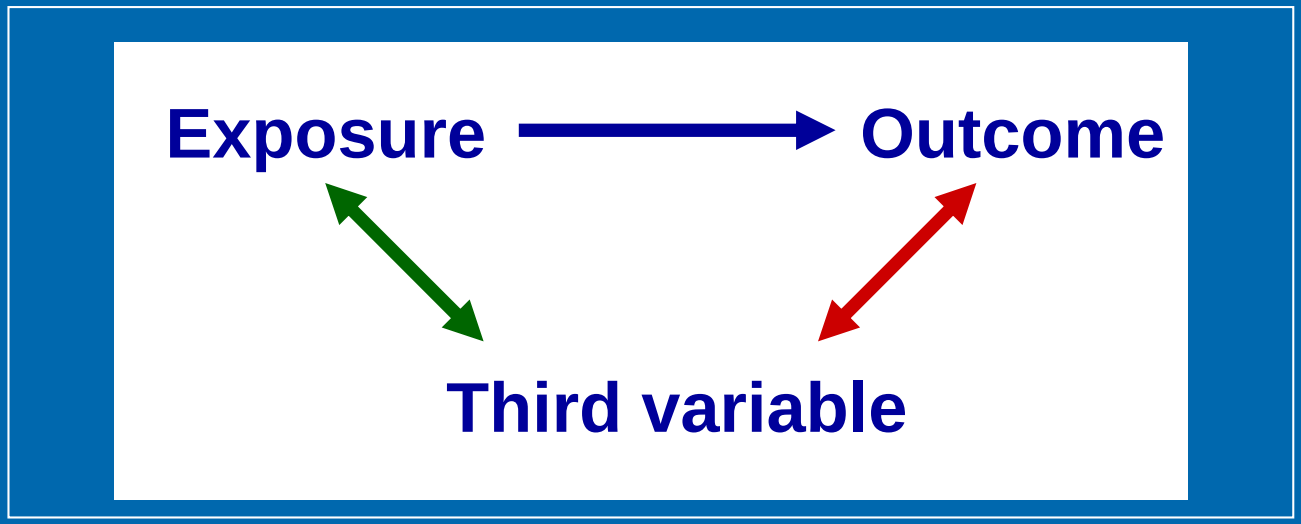
wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Confounding

- Distortion of measure of effect because of a third factor
- Should be prevented or Needs to be controlled for

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Confounding



Be associated with exposure - without being the consequence of exposure

Be associated with outcome - independently of exposure

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Confounding (Simpson's Paradox)

"Condom Use increases the risk of STD"

		STD rate	
Condom Use	Yes	55/95	(61%)
	No	45/105	(43%)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Confounding (Simpson's Paradox)

BUT ...

		STD rate	
# Partners < 5			
Condom Use	Yes	5/15	(33%)
	No	30/82	(37%)
# Partners ≥ 5			
Condom Use	Yes	50/80	(62%)
	No	15/23	(65%)

Explanation: Individuals with more partners are more likely to use condoms. But individuals with more partners are also more likely to get STD.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Confounding - Causal Diagrams

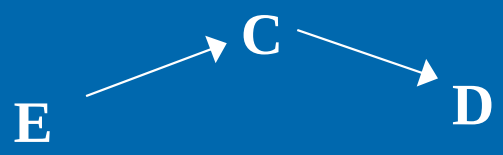
E = Exposure

D = Disease

C = Potential Confounder

An apparent association between E and D is completely explained by C. C is a confounder.

An association between E and D is partly due to variations in C. C is a confounder.



C is in the causal path between E and D, a confounder.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Confounding - Causal Diagrams

C has an independent effect on D.
C is not a confounder.

The effect of C on D is completely contained in E. C is not a confounder

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Example – Genetic Association study

- Idiopathic Pulmonary Fibrosis (IPF) is known to be associated with age and gender (older and male are more likely)
- One study had 174 cases and 225 controls found association of IPF with one gene genotype COX2.8473 (C → T).

Genotype	CC	CT	TT	total
Case	88	72	14	174
Control	84	113	28	225
Total	172	185	42	399

- P-value by Pearson Chi-squares test: p = 0.0241.
- Q: Is this association true?

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Example – Genetic Association study – continued

- Stratify by sex or age

Sex	male	female	total	
Case	108	66	174	
Control	72	153	225	
Total	180	219	399	P < 0.0001

Age	<29	30-49	50-64	65-74	75+	Total	
Case	0	10	42	68	54	174	
Control	104	77	35	7	2	225	
Total	104	87	77	75	56	399	p<0.0001

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Confounding

- **Positive confounding**

  - positively or negatively related to both the disease and exposure

- **Negative confounding**

  - positively related to disease but is negatively related to exposure or the reverse

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

How to prevent/control confounding?

Prevention (Design Stage)

- Restriction to one stratum
- Matching

Control (Analysis Stage)

- Stratified analysis
- Multivariable analysis

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Choosing Confounders for Statistical Adjustment

- One school says choice should be based on **a priori** considerations

  - Confounders selected based on their role as known risk factors for the disease

  - Selection on basis of statistical significance of association with disease can leave residual confounding effect

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Choosing Confounders for Statistical Adjustment

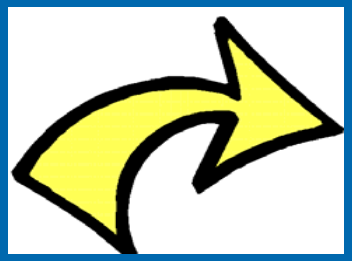
- **Others say choice of confounders should be based on how much they affect RR (OR, RD) when included/ excluded from the model.**

Compare crude measure of effect (RR or OR)

to

adjusted (weighted) measure of effect (Mantel Haenszel RR or OR)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health



To analyse effect modification

To control confounding

- **Solution**

Stratification (stratified analysis) Create strata according to categories inside the range of values taken by the effect modifier or the confounder

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Mantel Haenszel Methods

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Mantel Haenszel Methods-Notations

Assess association between disease and exposure after controlling for one or more confounding variables.

	E	$\overline{E}$	
D	a_i	b_i	$(a_i + b_i)$
$\overline{D}$	c_i	d_i	$(c_i + d_i)$
	$(a_i + c_i)$	$(b_i + d_i)$	n_i

where i = 1,2,…,K is the number of strata.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Cochran Mantel Haenszel Chi-square tests

(1) Correlation Statistic (Mantel-Haenszel statistic) has 1 df and assumes that either exposure or disease are measured on an ordinal (or interval) scale, when you have more than 2 levels.

(2) ANOVA (Row Mean Scores) Statistic has k-1 df and disease lies on an ordinal (or interval) scale when you have more than 2 levels.

(3) General Association Statistic has k-1 df and all scales accepted

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Mantel Haenszel Methods common odds ratio

(1) The Mantel-Haenszel estimate of the odds ratio assumes there is a **common** odds ratio:

$$OR_{pool} = OR_1 = OR_2 = \ldots = OR_K$$

To estimate the common odds ratio we take a weighted average of the stratum-specific odds ratios:

MH estimate:

$$\hat{O}R = \frac{\sum_{i=1}^{K} a_i d_i / n_i}{\sum_{i=1}^{K} b_i c_i / n_i}$$

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Mantel Haenszel Methods

(2) **Test of common odds ratio**

H_o: common OR is 1.0 vs. H_a: common OR $\neq$ 1.0

- A standard error is available for the MH common odds
- Standard CI intervals and test statistics are based on the standard normal distribution.

(3) **Test of effect modification** (heterogeneity, interaction)

H_o: $OR_1 = OR_2 = \ldots = OR_K$

H_a: not all stratum-specific OR's are equal

Breslow-Day (SAS) homogeneity test can be used

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Computing Cochran-Mantel-Haenszel Statistics for a Stratified Table

- The data set Migraine contains hypothetical data for a clinical trial of migraine treatment. Subjects of both genders receive either a new drug therapy or a placebo. Assess the effect of new drug adjusting for gender.

- SAS manual

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Example - Migraine

	Response		
Treatment	Better	Same	Total
Active	28	27	55
Placebo	12	39	51
Total	40	66	106

Pearson Chi-squares test p = 0.0037

But after stratify by sex, it will be different for male vs female.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Example – Migraine

Male	Response		
Treatment	Better	Same	Total
Active	12	16	28
Placebo	7	19	26
Total	19	35	54

p = 0.2205

Female	Response		
Treatment	Better	Same	Total
Active	16	11	27
Placebo	5	20	25
Total	21	31	52

p = 0.0039

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

SAS- codes

```
data Migraine;
    input Gender $ Treatment $ Response $ Count @@;
    datalines;
  female Active  Better 16   female Active  Same 11
  female Placebo Better  5   female Placebo Same 20
  male   Active  Better 12   male   Active  Same 16
  male   Placebo Better  7   male   Placebo Same 19
  ;

   proc freq data=Migraine;
     weight Count;
     tables Gender*Treatment*Response / cmh noprint;
     title1 'Clinical Trial for Treatment of Migraine Headaches';
  run;
********** In SAS, Need to put Exposure BEFORE Disease to
   generate right results for CMH results;
```


wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

SAS Output

The FREQ Procedure

Summary Statistics for Treatment by Response
Controlling for Gender

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	8.3052	0.0040
2	Row Mean Scores Differ	1	8.3052	0.0040
3	General Association	1	8.3052	0.0040

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Limits	
Case-Control (Odds Ratio)	Mantel-Haenszel	3.3132	1.4456	7.5934
	Logit	3.2941	1.4182	7.6515
Cohort (Col1 Risk)	Mantel-Haenszel	2.1636	1.2336	3.7948
	Logit	2.1059	1.1951	3.7108
Cohort (Col2 Risk)	Mantel-Haenszel	0.6420	0.4705	0.8761
	Logit	0.6613	0.4852	0.9013

Breslow-Day Test for
Homogeneity of the Odds Ratios

Chi-Square	1.4929
DF	1
Pr > ChiSq	0.2218

Total Sample Size = 106

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Comments

- The significant *p*-value (0.004) indicates that the association between treatment and response remains strong after adjusting for gender
- The probability of migraine improvement with the new drug is just over two times the probability of improvement with the placebo.
- The large *p*-value for the Breslow-Day test (0.2218) indicates no significant gender difference in the odds ratios.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Limitations of the Stratified Methods

- Can study only one independent variable at a time
- Problematic when there are too many variables to adjust for (too many strata)
- Limited to categorical variables (if continuous, can categorize, which may result in residual confounding)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Multiple Logistic Regression

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Logistic regression

- Often response variables in health studies are binary, eg. disease vs no disease, damage vs no damage, death vs live, etc.

- To explain the variability of the binary variable by other variables, either continuous or categorical, such as age, sex, BMI, marriage status, socio-economic status, etc. use a statistical model to relate the probability of the response event to the explanatory variables.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Logistic regression

- Prob of event labeled as binary outcome
- Event (Y = 1), no event (Y = 0) model the mean:

$$E(Y) = P(Y=1) * 1 + P(Y=0) *0 = P(Y=1)$$

but π = P(Y=1) is between 0 and 1

while $\beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + \ldots$ is a linear combination and may take any value.

A transformation is required.

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Logistic regression

- Logistic regression model:
- logit function.

$$\log [\pi / (1- \pi)] = \beta_0 + x_1\beta_1 + \ldots + x_p\beta_p = \eta$$

equivalent to $p = \exp(\eta) / [\ 1+ \exp(\eta)\ ]$

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Multiple Logistic Regression

Assumptions of Logistic Regression

- The independent variables are linear in the logit which may contain interaction and power terms
- The dependent variable is binary Y=0 or 1
- The independent variables may be binary, categorical, continuous

awilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Population Health

Multiple Logistic Regression-Formulation

$$E(Y|x) = P(Y=1|x) = \pi(x) = \frac{e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\cdots+\beta_pX_p}}$$

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x + \cdots + \beta_p X_p$$

- The relationship between π and x is S shaped
- The *logit* (*log-odds*) transformation (link function)

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Multiple Logistic Regression Assess risk factors

- **Individually H_o: $\beta k = 0$**
- **Globally H_o: $\beta_m = \ldots \beta_{m+t} = 0$**

while controlling for confounders and other important determinants of the event

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Interpretation of the parameters

- If π is the probability of an event and *O* is the odds for that event then

$$Odds = \frac{\pi(x)}{1-\pi(x)} = \frac{probability\ of\ event}{probability\ of\ no\ event}$$

- The link function in logistic regression gives the *log-odds*

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x + \cdots + \beta_p X_p$$

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Interpretation of parameter β in logistic regression

Model : logit (π) = $\beta_0 + x_1\beta_1$

	Y=1	Y=0
X=1	$\pi(1 \mid x=1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$1-\pi(1 \mid x=1) = \frac{1}{1+e^{\beta_0+\beta_1}}$
X=0	$\pi(1 \mid x=0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$	$1-\pi(1 \mid x=0) = \frac{1}{1+e^{\beta_0}}$

$$OR = \frac{\pi(1 \mid x=1) * [1-\pi(1 \mid x=0)]}{[1-\pi(1 \mid x=1)] * \pi(1 \mid x=0)} = e^{\beta_1}$$

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Snoring & Heart Disease: Logistic Example

- An epidemiologic study surveyed 2484 subjects to examine whether snoring was a possible risk factor for heart disease.

	Snoring			
Heart Disease	Never	Occasional	Nearly Every night	Every Night
Yes	24	35	21	30
No	1355	603	192	224
Prop(yes)	.017	.055	.099	.118

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Constructing Indicator variables

- Let Z_1=1 if occasional, 0 otherwise
- Let Z_2=1 if nearly every night, 0 otherwise
- Let Z_3=1 if every night, 0 otherwise

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

SAS Codes

```
data hd;
input hd $ snoring $ count;
Z1=(snoring="occa");
Z2=(snoring="nearly");
Z3=(snoring="every");

cards;
yes never 24
yes occa 35
yes nearly 21
yes every 30

no never 1355
no occa 603
no nearly 192
no every 224
;
run;
```

```
proc logistic data=hd
     descending;
model hd (event="yes") =Z1 Z2
     Z3;
freq  count;
run;
```


wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

SAS OUTPUT

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.0335	0.2059	383.6641	<.0001
Z1	1	1.1869	0.2695	19.3959	<.0001
Z2	1	1.8205	0.3086	34.8027	<.0001
Z3	1	2.0231	0.2832	51.0313	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Z1	3.277	1.932	5.558
Z2	6.175	3.373	11.306
Z3	7.561	4.341	13.172

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Calculating Probabilities

- The fitted logistic regression function is

$\text{Logit}(\pi) = -4.0335 + 1.1869\, Z_1 + 1.8205\, Z_2 + 2.0231\, Z_3$

- So, the probability of heart disease if never snore is exp(-4.0335)/(1+exp(-4.0335))=.0174

- If snore occasionally,

exp(-4.0335+1.1869)/(1+exp(-4.0335 +1.1869)) =.0549

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Calculating Odds Ratios

- If $Z_1=Z_2=Z_3=0$, then odds are exp(-4.0335)
- If $Z_2=Z_3=0$, but $Z_1=1$, then odds are exp(-4.0335+1.1869)
- The ratio of odds is then exp(1.1869)
- What is the odds ratio for comparing those who snore nearly every night with occasional snorers?
- What is the odds ratio for comparing those who snore every night with those who snore nearly every night?

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Old example – Genetic Association study

- Idiopathic Pulmonary Fibrosis (IPF) is known to be associated with age and gender (older and male are more likely)
- One study had 174 cases and 225 controls found association of IPF with one gene genotype COX2.8473 (C → T).

Genotype	CC	CT	TT	total
Case	88	72	14	174
Control	84	113	28	225
Total	172	185	42	399

- P-value by Pearson Chi-squares test: p = 0.0241.
- Q: Is this association true?

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Old example on genetic effect of SNP COX2.8473 on IPF

- Logistic regression model

  logit [Pr(IPF)] = intercept + snp + sex + age

- Results:
- Effect

Effect	DF	Wald Chi-square	P-value
SNP	2	2.7811	0.2489
sex	1	9.1172	0.0025
age	1	100.454	<.0001

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Conclusion: What we have learned

1. Define study designs
2. Measures of effects for categorical data (unstratified analysis)
3. Confounders and effects modifications
4. Stratified analysis (Mantel-Haenszel statistic, multiple logistic regression)
5. Use of SAS Proc FREQ and Proc Logistic

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health

Conclusion: Further readings

Read textbook for

1. Power and sample size calculations

2. Tests in matched pair studies

wilopo@yahoo.com Universitas Gadjah Mada, Faculty of Medicine, Department of Biostatics, Epidemiology and Populayton Health