

MPH specialmodul Epidemiologi og Biostatistik

Bestemmelse af stikprøvestørrelse

Matchning

8. juni 2009

www.biostat.ku.dk/~pka/mpspec09

Per Kragh Andersen

1

Planning of investigations.

How many persons are needed?

For what purpose?

- (1) To obtain a given precision of an estimate.
- (2) To obtain a given *power* of a test (the most common situation).

Slightly different techniques for categorical data and quantitative data.

2

Categorical data: Estimation.

Example (1):

We wish to estimate the risk of pre-eclampsia among pregnant women and we desire a given precision of the estimate: $p \pm a$.

We know that a 95% confidence interval around a relative frequency is $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$ where n is the number of women, i.e.

$a = 1.96\sqrt{\frac{p(1-p)}{n}}$, and thereby $n = \frac{3.84p(1-p)}{a^2}$ (where $3.84 = 1.96^2$).

To obtain this, we need to know the order of magnitude of p in advance.

For example: $p = 0.10$, $a = 0.04$ gives $n = 216$.

3

Categorical data: Testing.

Example (2):

We want to *treat* pregnant women with pre-eclampsia and compare two treatments with respect to the risk of some pregnancy outcome, e.g. stillbirth. We want to be “pretty certain” to detect a treatment difference of Δ between the two risks.

What do we mean by “pretty certain”? We need the statistical concept of the *power* of a test.

If we test using a given level of significance α (i.e. 5%) and if the true treatment difference is Δ then we want to have a large probability of rejecting the null hypothesis: $\Delta = 0$. This probability, the power $1 - \beta$, is, for example, set to 80%.

In general: the larger power we want (and the smaller α we use), the larger needs n to be.

4

The probability β is called the “Type 2 error risk” and α is called the “Type 1 error risk”.

	H_0 correct	H_0 wrong
Accept	Type 2 error β	
Reject	Type 1 error α	power $1 - \beta$

To find n , a good guess of the risk in the control group (p_1) is needed. Letting $p_2 = p_1 - \Delta$,

$$n = \frac{p_1(1 - p_1) + p_2(1 - p_2)}{\Delta^2} \times f(\alpha, \beta)$$

women are needed *in each group*.

Quantitative data: Estimation.

Example (1):

We wish to estimate the level of cholesterol in a population and we desire a given precision of the estimate: $\bar{X} \pm a$.

We know that a 95% confidence interval around a mean is $\bar{X} \pm 1.96 \frac{SD}{\sqrt{n}}$ where n is the number of persons in the sample, i.e. $a = 1.96 \frac{SD}{\sqrt{n}}$, and thereby $n = 3.84 \left(\frac{SD}{a}\right)^2$ (where $3.84 = 1.96^2$).

To obtain this, we need an estimate of SD in advance or, alternatively, we can report the desired precision as a/SD .

For example: $a/SD = 20\%$ gives $n = 3.84/(0.2)^2 = 96$.

Here $f(\alpha, \beta)$ is given by:

	β			
α	0.05	0.10	0.20	0.50
0.10	10.8	8.6	6.2	2.7
0.05	13.0	10.5	7.9	3.8
0.02	15.8	13.0	10.0	5.5
0.01	17.8	14.9	11.7	6.6

Example: $p_1 = 0.15, \Delta = 0.07, \alpha = 0.05, \beta = 0.20$

$$\text{Then: } n = \frac{0.15 \cdot 0.85 + 0.08 \cdot 0.92}{0.07^2} \times 7.9 = 324$$

NB! this is the number in each group.

Example: $p_1 = 0.1, RR = 1.5, \alpha = 0.05, \beta = 0.20$

Then: $p_2 = p_1 \cdot RR = 0.15, \Delta = 0.05$ and

$$n = \frac{0.1 \cdot 0.9 + 0.15 \cdot 0.85}{0.05^2} \times 7.9 = 687$$

Quantitative data: Testing.

Example (2):

We want to *treat* persons with increased level of cholesterol and believe that the level of those treated is Δ lower than that of the untreated.

How many persons do we need in order to be “pretty certain” to detect a treatment difference of Δ between the two mean values.

Let n be the number *in each group*.

Again, we use the statistical concept of the *power* of a test.

To find n , a good guess of SD is needed (or of the ratio Δ/SD).

We then need:

$$n = 2 \left(\frac{SD}{\Delta}\right)^2 \times f(\alpha, \beta)$$

persons *in each group*.

Here $f(\alpha, \beta)$ is given by:

α	β			
	0.05	0.10	0.20	0.50
0.10	10.8	8.6	6.2	2.7
0.05	13.0	10.5	7.9	3.8
0.02	15.8	13.0	10.0	5.5
0.01	17.8	14.9	11.7	6.6

Example: $SD = 1.0 \text{ mmol/L}, \Delta = 0.5 \text{ mmol/L}, \alpha = 0.05, \beta = 0.20$

Then: $n = 2\left(\frac{1.0}{0.5}\right)^2 \times 7.9 = 63$

NB! this is the number in each group.

Doing it in SAS.

For this particular purpose we will be using SAS ANALYST where, for QUANTITATIVE DATA, the computations may be carried out quite easily.

'STATISTICS→SAMPLE SIZE'
→'TWO-SAMPLE T-TEST'

(It is also possible to do the reversed computations, i.e. find the power $1 - \beta$ from n .) We need to specify:

- Group 1 Mean
- Group 2 Mean
- SD
- α
- an interval (or just a single value) for the power $1 - \beta$

The expressions for n for categorical data and for quantitative data look quite alike.

If we replace

$$p_1(1 - p_1) + p_2(1 - p_2)$$

by

$$2(SD)^2 = 2\bar{p}(1 - \bar{p})$$

where $\bar{p} = (p_1 + p_2)/2$ is the average risk then we may cheat SAS ANALYST to do the computations also for categorical data.

Let Group 1 Mean = p_1 , Group 2 Mean = p_2 , compute $\bar{p} = (p_1 + p_2)/2$ and $SD = \sqrt{\bar{p}(1 - \bar{p})}$.

Finally choose α and the power $1 - \beta$ as before.

Unequal group sizes.

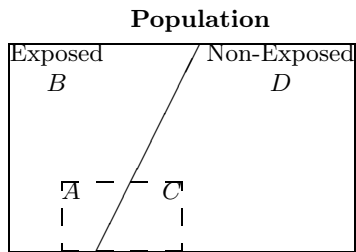
If the two groups do not have the same size:

- first compute the total size $N = 2n$ as if the two groups were equally large,
- then compute $k = n_1/n_2 =$ the ratio between the group sizes
- the total number needed is then $N' = N \frac{(1+k)^2}{4k}$.

Example. If, in the first example above, group 1 is twice as big as group 2:

- $N = 2 \cdot 324 = 648$
- $k = 2$
- $N' = N \frac{(1+k)^2}{4k} = 648 \frac{9}{8} = 729$, i.e. $n_1 = 486, n_2 = 243$.

Cohort and case-control studies.

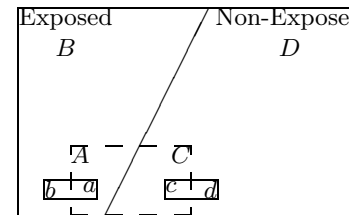


The population consists of $A + B$ exposed and $C + D$ non-exposed individuals.

After some time, (say, 2 years) A out of the exposed and C out of the non-exposed have developed the disease. That is,

$$\text{Relative risk} = \frac{A/(A+B)}{C/(C+D)} \quad \text{Odds ratio} = \frac{A/B}{C/D} = \frac{A \cdot D}{B \cdot C}$$

Population: cohort sample



$$\text{Relative risk} = \frac{A/(A+B)}{C/(C+D)} \quad \text{Odds ratio} = \frac{A/B}{C/D} = \frac{A \cdot D}{B \cdot C}$$

Cohort sample:

$$\begin{aligned} \text{Exposed:} \quad k_1(A+B) &= k_1A + k_1B \\ &\sim a \quad \sim b \end{aligned}$$

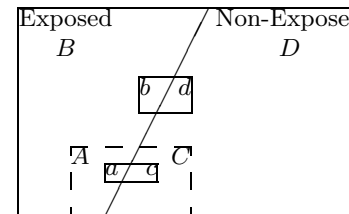
$$\begin{aligned} \text{Non-exposed:} \quad k_2(C+D) &= k_2C + k_2D \\ &\sim c \quad \sim d \end{aligned}$$

$$\begin{aligned} \text{Then:} \quad \frac{a}{a+b} &\sim \frac{k_1A}{k_1A+k_1B} = \frac{A}{A+B} \\ \frac{c}{c+d} &\sim \frac{k_2C}{k_2C+k_2D} = \frac{C}{C+D} \end{aligned} \Rightarrow \text{We can estimate}$$

relative risk

$$\text{AND odds ratio, since } \frac{a \cdot d}{b \cdot c} \sim \frac{k_1A \cdot k_2D}{k_1B \cdot k_2C} = \frac{A \cdot D}{B \cdot C}$$

Population: case-control sample



$$\text{Relative risk} = \frac{A/(A+B)}{C/(C+D)} \quad \text{Odds ratio} = \frac{A/B}{C/D} = \frac{A \cdot D}{B \cdot C}$$

Case-control sample:

$$\begin{aligned} \text{Diseased:} \quad k_3(A + C) &= k_3A + k_3C \\ \text{(cases)} \quad &\sim a \quad \sim c \\ \text{Non-diseased:} \quad k_4(B + D) &= k_4B + k_4D \\ \text{(controls)} \quad &\sim b \quad \sim d \end{aligned}$$

$$\text{Then: } \frac{a}{a+b} \sim \frac{k_3A}{k_3A+k_4B}, \quad \frac{c}{c+d} \sim \frac{k_3C}{k_3C+k_4D}$$

⇒ We canNOT estimate relative risk

$$\text{BUT odds ratio, since } \frac{a \cdot d}{b \cdot c} \sim \frac{k_3A \cdot k_4D}{k_4B \cdot k_3C} = \frac{A \cdot D}{B \cdot C}$$

Conclusion:

From the data collected in the case-control study we may:

- estimate odds ratio by $OR = \frac{a \cdot d}{b \cdot c}$
and its confidence limits
- test for no association between exposure and outcome using the chi-square test

PROVIDED THAT

there is no selection bias, that is, when selecting cases and controls, no consideration of exposure must be taken

Exposure odds ratio.

	cases	controls
Exposed	a	b
Non-exposed	c	d

Odds for being exposed among cases = a/c

Odds for being exposed among controls = b/d

⇒ exposure odds ratio = $\frac{a/c}{b/d} = \frac{ad}{bc}$, i.e. the exposure OR estimates the disease OR.

It is the latter that we are interested in.

Matching.

Design method to

- adjust for confounding
- and/or increase efficiency

Example: BCG vaccine and leprosy.

BCG scar	Leprosy cases	Population survey
Present	101	46028
Absent	159	34594

$$OR_{\text{crude}} = 0.477 \quad (0.371, 0.612)$$

Example: only 1000 controls.

BCG scar	Leprosy cases	Population survey
Present	101	554
Absent	159	446

$$OR_{\text{crude}} = 0.511 \quad (0.386, 0.675)$$

Little precision is lost by reducing the number of controls.

Age-stratification.

BCG	Leprosy cases		Healthy population		Odds ratio estimate
	-	+	-	+	
Age					
0-4	1	1	7593	11719	0.65
5-9	11	14	7143	10184	0.89
10-14	28	22	5611	7561	0.58
15-19	16	28	2208	8117	0.48
20-24	20	19	2438	5588	0.41
25-29	36	11	4356	1625	0.82
30-34	47	6	5245	1234	0.54
Total	159	101	34594	46028	0.48

$$OR_{\text{MH}} = 0.587 \quad (0.448, 0.769)$$

Age-stratification in small study.

Age	BCG	Cases		Controls	
		-	+	-	+
0-4		1	1	101	137
5-9		11	14	91	115
10-14		28	22	82	101
15-19		16	28	28	87
20-24		20	19	25	69
25-29		36	11	63	21
30-34		47	6	56	24
		159	101	446	554

$$OR_{\text{MH}} = 0.582 \quad (0.425, 0.796)$$

Here, some of the controls are “wasted”: there are many-many controls per case for the youngest.

Age-matching: age distributions are made identical for cases and controls by choosing 4 controls per case *from the same age group*.

Age	BCG	Cases		Controls	
		-	+	-	+
0-4		1	1	3	5
5-9		11	14	48	52
10-14		28	22	67	133
15-19		16	28	46	130
20-24		20	19	50	106
25-29		36	11	126	62
30-34		47	6	174	38
		101	159	514	526

$$OR_{\text{MH}} = 0.568 \quad (0.420, 0.768)$$

The confidence interval gets (slightly) more narrow, the frequency matching uses the controls more efficiently.

Logistic regression

“Large” study.

Parameter	Estimate	SE
Intercept	-8.880	0.7093
Age 5-9	2.624	0.7340
Age 10-14	3.583	0.7203
Age 15-19	3.824	0.7228
Age 20-24	3.900	0.7244
Age 25-29	4.156	0.7224
Age 30-34	4.158	0.7213
BCG	-0.547	0.1409

(=ln(0.579))

25

Logistic regression

Matched study.

Parameter	Estimate	SE
Intercept	-1.0670	0.800
Age 5-9	-0.0421	0.827
Age 10-14	0.0119	0.812
Age 15-19	0.0713	0.814
Age 20-24	0.0244	0.816
Age 25-29	-0.1628	0.814
Age 30-34	-0.2380	0.813
BCG	-0.5721	0.155

(=ln(0.564))

“Strange” age effects!

This is because when we have matched for age then we *cannot estimate the age effect* (age distributions among cases and controls are identical).

26

Then: why not leave age entirely out of the analysis?

Stratum	Cases		Controls		Odds ratio
	+	-	+	-	
1	89	11	80	20	2.0
2	67	33	50	50	2.0
3	33	67	20	80	2.0
Total	189	111	150	150	1.7

ALWAYS ADJUST FOR AGE IN THE CASE-CONTROL ANALYSIS WHEN YOU HAVE AN AGE-MATCHED DESIGN.

The bias from ignoring matching will always be towards 1.

27

Individual matching.

Each control (or set of controls) is matched to one particular case.
e.g., neighbourhoods, siblings, “time”.

Simplest case: 1:1 matching = matched pairs

Example (McNeil, p. 238):

Cases: 223 women giving pre-term birth

Controls: 223 women giving full-term birth

at the same hospital in Thailand during the same period (1992-93) and matched on age and parity.

Exposure: predominant work during first trimester
standing vs. non-standing

28

Four kinds of pairs:

Case	Control standing	Control non-standing
Standing	177	31
Non-standing	14	1

NB: table of 223 pairs, not 446 women.

Concordant pairs (strata)

	Case	Control	Case	Control
+Exp	1	1	0	0
-Exp	0	0	1	1
	1	1	1	1

Discordant pairs (strata)

	Case	Control	Case	Control
+Exp	1	0	0	1
-Exp	0	1	1	0
	1	1	1	1

Mantel-Haenszel analysis.

Only discordant pairs contribute:

$$OR_{MH} = \frac{31}{14} = 6.42,$$

The MH-test is:

$$X_{MH}^2 = \frac{(31 - 14)^2}{31 + 14} = 6.42, \quad P = 0.011$$

This is known as *McNemar's test*.

Logistic regression

Ordinary logistic regression does not work for individually matched case-control studies.

Instead, one has to use so-called *conditional* logistic regression.

Matching: pros and cons.

- **Pro**
 - intuitively simple
 - efficiency may be increased from within-group (within-pair) comparisons (cf. paired *t*-test)
- **Con**
 - inability to estimate effect of matching variables
 - complicates analysis - matching variables *must* be accounted for
 - risk over “over-matching” if exposure is strongly related to matching variable (e.g., matching on household in nutritional epidemiology)

Opgaver

- **1.** I en kohorte undersøgelse forventes 5-års risikoen for hjerteinfarkt at være 0.05 for personer i erhverv (=de ueksponerede). Hvor mange personer i erhverv og uden for erhverv (=de eksponerede) skal indgå i en 5-årig kohorte undersøgelse for at have en styrke på $1 - \beta = 0.9$ svarende til en relativ risiko på 2. Signifikansniveauet er det sædvanlige $\alpha = 0.05$.
- **2.** Samme spørgsmål, hvis den ueksponerede gruppe er dobbelt så stor som den eksponerede.
- **3.** I en case-control undersøgelse af erhvervsstatus og risiko for hjerteinfarkt regner man med, at 20% af cases er eksponerede (=uden for erhverv). Hvad er styrken for et test på 5% niveau med 300 cases og 300 kontroller, hvis $OR = 2$? (NB: Husk at OR for eksposition er den samme som OR for sygdom, og betragt situationen, at I skal sammenligne ekspositions risikoen blandt cases og kontroller.)

- **4.** Indlæs sas-programmet `bcg.sas`, som drejer sig om eksemplet vedr. bcg vaccination og spedalskhed. Udfør analyserne og genfind resultaterne fra planche 25.
- **5.** Lav om på antallet af kontroller i `cards` kommandoen, så de svarer til det lille datasæt på planche 23 og udfør den logistiske regressionsanalyse.
- **6.** Lav igen om på antallet af kontroller i `cards` kommandoen, så de nu svarer til det matchede datasæt. Genfind resultaterne fra planche 26 og sammenlign med **5.** ovenfor..