



**Karolinska  
Institutet**

## **4.2 Optimal 2-stage designs**

# Two-stage studies in Lecture 4.1

Not "planned" *a priori* as 2-stage studies

Ectopic pregnancy investigators realised during the study that chlamydia antibody was important risk factor.

**BUT:**

did **not** try to get this data for all subjects  
**nor** for all subjects from then on

**INSTEAD:**

all cases, sample of controls (oversampled black women)

Suggests consideration of:

costs

power (intuition about "informative" subjects)

# Two-stage studies in Lecture 4.1

Not "planned" *a priori* as 2-stage studies

H. Pylori investigators already had a sample of schoolchildren and wanted to gather additional variables (infection in family member) as a new research question.

**BUT:**

did **not** try to get this data for all children

**AGAIN:**

all cases, sample of controls (targeting high prevalence)

Suggests consideration of:

costs

power (intuition about "informative" subjects:

SES, immigrant status)

# Efficient analysis vs. efficient design

In these examples we have seen, the data were already gathered, so best we can do is

*analyse it efficiently*

But what if we are at **planning stage**?

Can we **design** an “efficient” two-stage study?

# If a two-stage study is planned in advance

1. How many should we plan to sample?

2. Which subjects?

**Q1.** is like the usual “sample size” for any study, to achieve a specific power

# Sample size for difference in means

Usually

power = 80%, significance = 5%

We need some information:

- Specify smallest difference of clinical interest
- Provide estimate of variance/SD in two populations

# Quiz

Suppose we plan to randomly assign half of a group of 5-year old children to get extra milk for one year in order to test if they have greater height gain.

At this age, the average height gain is approx **6cm**, and the SD is approx **2cm**. Let us assume that an increase to **6.5cm** is important.

Use [openepi.com](https://openepi.com) to find the sample size necessary for power of **80%** and significance level=**.05**

# Sample size for difference in proportions (or RR or OR)

**Example:** a clinical trial to compare a new drug to an existing treatment that has a 60% success rate. Assume that it would be clinically important to detect a 75% success rate for the new drug.

## Quiz



# Efficient case-control studies: an “old” problem

Optimal sampling has long concerned medical investigators, particularly in case-control studies.

Common designs have equal numbers of cases and controls, overall or within strata (**‘balanced’** design):

This not necessarily optimum when

- different cost for case/control and maybe across strata
- different “information” across strata

# Examples of work on optimal design

## **Cain & Breslow 1988**

### **case-control study, logistic regression**

- propose balanced design as efficient
- no consideration of cost

We saw examples of very efficient balanced designs  
(Malawian monitoring and evaluation of HIV services)

# Designs considering cost

## Cochran (1963) and Nam(1973)

cost per control =  $C_0$

cost per case =  $C_1$

optimum ratio  $\frac{n_0}{n_1} = \sqrt{\frac{C_1}{C_0}}$

## Nam & Fears 1990

### strata-matched C-C study

- cost per case ( $C_1$ ) fixed, cost per control ( $C_0$ ) varies by stratum,
- fixed total cost
- when  $C_0=C_1$ , and exposure rate same in all strata,  $n_0=n_1$  optimal.

# Summary of early “optimal” designs

## Balanced sample:

- simple and intuitive
- Can apply to second-stage sampling from any classic design
- Very good efficiency in many settings
- Does not consider costs.

## Designs that consider costs:

- Allow for differential costs (between case and control)
- only for case-control design (not for other first-stage designs),
- not for two-stage sampling.

cost of case vs. control not the issue in most practical applications  
(exposure is expensive!)

## In weighted logistic regression of 2-stage data

The variance of the estimates depends on:

- proportion of first-stage sample in each of Z,Y strata
- proportion of each stratum selected in the second stage

⇒ *Can choose sampling fractions to minimise standard error!*

Formula for the optimal sampling strategy (won't show!)

Available in Stata **optimal** package

# First scenario (no cost consideration)

Data already gathered on  $n$  subjects, and now a forgotten/new covariate to be measured on a sub-sample of size  $n_2$

e.g. database exists for a cohort or case-control study, and new/renewed interest in biochemical or genetic marker, so decide to test 100 stored specimens

**Q: which 100?**

Can compute sampling fractions to minimize SE of biomarker,

**but as with sample size calculations, need some info:**

- estimates of proportions in first-stage strata
- pilot data (few observations in each stratum)

## Other Scenarios (including cost)

Cost per observation:

First stage:  $C_1$

Second stage:  $C_2$

### Fixed budget

study hasn't been designed and we have fixed budget, how many and “who” to sample?

### Fixed precision

study hasn't been designed, we wish to achieve a specified SE while minimising total study cost

# Example of fixed budget

If study was planned in advance to investigate effect of chlamydia antibody status on risk of ectopic pregnancy (adjusting for age, etc....)

Budget €50,000

Cost per observation:

€5 at Stage 1,

€50 for CT antibody (Stage 2)

**Q. How many patients to study overall?**

**How many (and who) should have CT antibody measured?**



# Optimal sampling for fixed Budget

## Objective:

given first stage cost  $C_1$  per unit and second stage cost  $C_2$  per unit, minimise  $SE(\beta)$  for fixed total budget

$$B = n^* C_1 + n_2^* C_2$$

Can find expression for  $n$  (first stage sample size)

and then second-stage sampling fractions use

$$n_2 = \frac{B - nC_1}{C_2} \text{ "affordable" second-stage sample size}$$

# Optimal sampling: fixed precision for minimum cost

**Objective:** given first stage cost  $C_1$  and second stage cost  $C_2$  per unit, to achieve a specified variance for coefficient,  $V(\beta) = \delta$  for minimum total cost  $B = n \cdot C_1 + n_2 \cdot C_2$

Again, there is an expression for  $n$

and sampling fractions at second-stage (total  $n_2$ )

**Note:**

Each of sampling fractions =  $\sqrt{\frac{C_1}{C_2}}$  (...)

# Illustration: ectopic pregnancy example

case-control study: ectopic pregnancy and STDs

Total sample size = 979 (264 cases, 715 controls)

1<sup>st</sup> stage : gonnorrhoea, contraceptive use, sex partners (n=979).

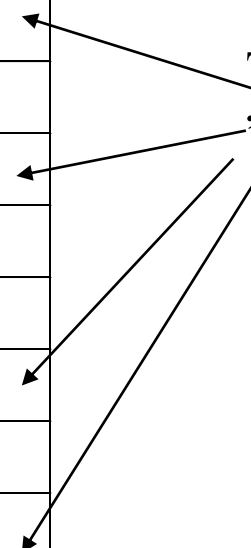
2<sup>nd</sup> stage chlamydia antibody (n=327)

| ectopic | gonn | contra | First Stage sample | Second stage sample |
|---------|------|--------|--------------------|---------------------|
| N       | N    | N      | 175                | 46                  |
| N       | N    | Y      | 490                | 129                 |
| N       | Y    | N      | 19                 | 8                   |
| N       | Y    | Y      | 31                 | 9                   |
| Y       | N    | N      | 186                | 90                  |
| Y       | N    | Y      | 44                 | 25                  |
| Y       | Y    | N      | 29                 | 18                  |
| Y       | Y    | Y      | 5                  | 2                   |

How could 327 be “optimally” chosen  
(to maximise precision of effect of chlamydia)

| <b>ectopic</b> | <b>gonn</b> | <b>contra</b> | <b>Optimal<br/>sampling<br/>fraction</b> | <b>Actual<br/>Sampling<br/>fraction</b> |
|----------------|-------------|---------------|--|---|
| N              | N           | N             | .62                                      | .26                                     |
| N              | N           | Y             | .11                                      | .26                                     |
| N              | Y           | N             | .72                                      | .42                                     |
| N              | Y           | Y             | .15                                      | .29                                     |
| Y              | N           | N             | .47                                      | .48                                     |
| Y              | N           | Y             | .96                                      | .57                                     |
| Y              | Y           | N             | .53                                      | .62                                     |
| Y              | Y           | Y             | 1  | .40                                     |

These strata  
"informative"



# What about Cost....

If cost=€5 at first-stage, €50 at second stage,

Total cost=  $979*5 + 327*50 = €21,245$

**Q1**

For this budget, what would be the most cost-effective study?

**Q2**

What if  $c_1=c_2=5$ ?

| ectopic | gonn | contra | Optimal sampling fraction<br>n2=327 | Optimal sampling fraction<br>c1=5,c2=50 | Optimal sampling fraction<br>C1=5,C2=5 |
|---------|------|--------|-------------------------------------|---|--|
| N       | N    | N      | .62                                 | 1                                       | 1                                      |
| N       | N    | Y      | .11                                 | .19                                     | .33                                    |
| N       | Y    | N      | .72                                 | 1                                       | 1                                      |
| N       | Y    | Y      | .15                                 | .26                                     | .45                                    |
| Y       | N    | N      | .47                                 | .81                                     | 1                                      |
| Y       | N    | Y      | .96                                 | 1                                       | 1                                      |
| Y       | Y    | N      | .53                                 | .91                                     | 1                                      |
| Y       | Y    | Y      | 1                                   | 1                                       | 1                                      |
|         |      |        |                                     | Total= 677                              | Total=2588                             |

**For cheaper data, best design samples more subjects (obvious!) and performs lab tests on all in 6 of the strata (less obvious).**

# Illustration: H.Pylori study

Recall: Cross-sectional study of schoolchildren

1<sup>st</sup> stage : immigrant background, SES (n=664)

2<sup>nd</sup> stage *Hp* status of mother, father, sibs, +.... (n=174)

Assume *mothers Hp status* of primary interest

What is optimal way to sample 200 families?

|                             | <u>Stage 1</u> | <u>Stage 2<br/>Actual<br/>Design</u> | <u>Stage 2<br/>Optimal<br/>Design</u> |
|-----------------------------|----------------|--------------------------------------|---------------------------------------|
| <b><u>Cases (total)</u></b> | 104            | Sample<br>All                        |                                       |
| low prev, low SES           | 6              |                                      | 6                                     |
| high SES                    | 4              |                                      | 4                                     |
| high prev, low SES          | 16             |                                      | 13 (81%)                              |
| High SES                    | 78             |                                      | 59 (76%)                              |
| <b><u>Controls</u></b>      |                |                                      |                                       |
| low prev, low SES           | 241            | 28 (12%)                             | 17 (7%)                               |
| high SES                    | 179            | 37 (21%)                             | 32 (18%)                              |
| high prev, low SES          | 55             | 13 (24%)                             | 13 (24%)                              |
| High SES                    | 84             | 27 (32%)                             | 46 (55%)                              |

these children informative!





## Recent developments:

*Mclsaac and Cook (Stat in Med, April 2015)*

**adaptive** two-phase design

Phase-II sampling divided into subphases

**IIa:** pilot data (no info available yet to optimize)

**IIb:** optimally sampled

**Objective:** allocate the  $n_2$  into  $n_{2a}$  and  $n_{2b}$

> 2 phases: like group sequential monitoring in clinical trials

extreme case: selection probabilities updated after each individual observation!

**Practical recommendation:**

50:50  $n_{2a}$  (balanced) and  $n_{2b}$  (optimized)

# Summary

The optimal designs are for binary outcome (logistic regression)  
(For time-to-event data: next lecture 4.3)

Functions are available in the package **optimal** in Stata  
not (yet!) in R

Power calculations available in **PowerIIPhase()** function in R