OULUN YLIOPISTO
UNIVERSITY of OULU

Uses of Sampling Methodology in Epidemiologic Research

Esa Läärä

Research Unit of Mathematical Sciences, University of Oulu

esa.laara@oulu.fi

BaNoCoSS, Örebro, 19.6.2019

Outline

- Descriptive and etiological studies
- Study population and study base
- Full cohort design
- Outcome-selective, two-phase sampling
- Nested case-control and case-cohort designs
- Statistical modelling
- Utilization of auxiliary information
- Conclusion

Descriptive or enumerative studies

Questions: Distribution of health traits and related characteristics in a finite target population at a given time?

For instance

> "*What is the prevalence of hormone therapy (HT) among postmenopausal women in Finland, 2019?*"

- **Cross-sectional health survey**: questionnaire, interview, health exams, laboratory tests etc.
- Complex multi-stage sampling often applied.
- Examples: Health 2000 in Finland, NHANES in USA
- Survey business as usual?

Etiological or "analytic" studies

Question: **Causal effect** of exposure to a given factor on the risk of disease among people of certain kind?

Ex. "*What is the 10-year risk of breast cancer in women starting HT at 50 y of age as compared with the risk they would have, had they not started HT?*"

- Involves a **counterfactual conditional**
  – How to find a comparable group of non-users of HT?
- Target population or universe?
  – The whole womankind or a defined domain of it.
- Probability sampling from target – impossible!

End of story?

Case-cohort study on HT & breast cancer

- Study population: Dutch **cohort** of $N \approx 60\,000$ women, 55-69 y, recruited 9/1986. – **Closed** population.
- Questionnaire: reproductive history, health habits, SES, etc.
- Follow-up till 12/1989, mean 3.3 y,
  Total **person-time** $Y \approx 200\,000$ years.
- **Subcohort**, $n = 1800$, simple random sample (3 %).
- $D = 471$ new **cases** in the cohort; 15 in the subcohort.
  Sampling fractions f: cases 100%, others 3%.
- Data for cases and subcohort members analyzed.

Estimated hazard ratio HR 0.99 [95% CI: 0.7 to 1.4] for ever ($D_1 = 58$ cases) *vs.* never ($D_0 = 387$) use of HT.

Schuurman *et al.* (1995) *Cancer Causes and Control*; 6: 416-424,

Nested case-control study: HT & breast ca

- Study population: All women in Finland, 50-62 y, in 1995-2007; $N(t) \approx 450\,000$ at any time, total $N \approx 900\,000$.
  – **Open** or **dynamic** population.

- Follow-up: From variable entry to variable exit times.
  Total person-time $Y \approx 5.85 \times 10^6$ years.

- $D \approx 10\,000$ **cases**. $C \approx 30000$ **controls** were sampled ($f = 3\%$); individually **matched** for age (± 1 mo), alive and cancer-free at diagnosis of the case.

- Data on HT from national reimbursement register for cases and controls analyzed by conditional logistic regression.

$\widehat{\text{HR}}$ 1.36 (95% CI 1.27 to 1.46) for estradiol-progestagen therapy ($D_1 = 1731$) *vs.* no use of HT ($D_0 = 5473$).

Lyytinen *et al.* (2010) *Int J Cancer* 126: 483-489.

Study population and its selection

Study population – or "sample" – in a causal study

- Often a highly selected convenience sample.
- Any available frame population can only cover a very specific subdomain of the whole target.
- Eligibility, feasibility, exclusions, restriction, stratification, participation, etc.
- Aspects of **internal validity** more important than statistical representativeness or generalizability.
- **Generalization?** – Synthesis of independent results obtained from various populations & places.

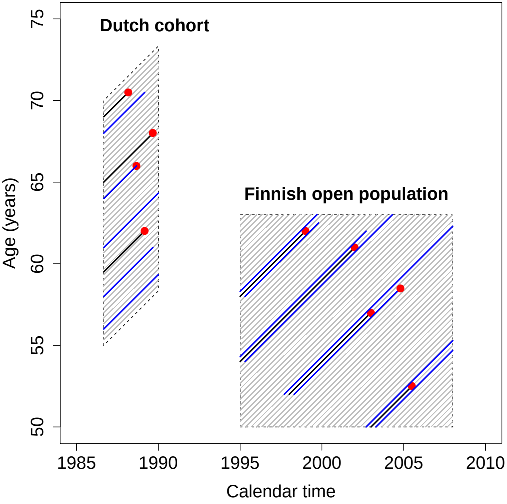
Rothman et al. (2013) Why representativeness should be avoided (with discussion). *Int J Epidemiol* **42**: 1012-1028.

Study base

Study base = Study population × its experience in time.

- **Cross-sectional base**:
  Study population at a given time point.
  - Perinatal epidemiology: newborn at their dates of birth.
- **Longitudinal base:**
  Comprises **follow-up times** of individuals in the study population over a specified time period.
  - From date of **entry** to date of **outcome** (e.g. breast cancer) – or of **competing event** (e.g. death), or **censoring** (e.g. emigration).
  - Affected by **right censoring** and **late entry** (left truncation), esp. when age is the main time scale.

Study bases in Lexis diagram – simplified



- Outcome cases: black lifelines & red bullets
- Subcohort members: blue lifelines
- Matched controls: blue lifelines
- Subcohort members and controls can become cases.

Obtaining data on risk factors

Main strategies

- **Complete enumeration** of the study base
  – *"Cohort study"* or *full cohort design*

- **Outcome-dependent, 2nd phase sampling**
  – *"Case-control study"*

  Data on exposure to risk factors gathered only for

  (a) **cases:** all (or high % of) D subjects in whom the outcome is observed, and

  (b) **controls:** a random sample of C subjects ($C << N$) of the remaining population at given times.

Full cohort design: binary risk factor X

Simple summary of follow-up data

	$X = 1$	$X = 0$	
	exposed	unexposed	total
No. of cases	D_1	D_0	D
Group size	N_1	N_0	N
Person-time	Y_1	Y_0	Y
Incidence rate	$I_1 = D_1/Y_1$	$I_0 = D_0/Y_0$	$I = D/Y$

The **hazard ratio** (HR) for $X = 1$ vs. $X = 0$ crudely estimated by the empirical **incidence rate ratio** (IR)

$$\text{IR} = \frac{I_1}{I_0} = \frac{D_1/Y_1}{D_0/Y_0} = \frac{D_1/D_0}{Y_1/Y_0}$$

With fixed risk period & complete follow-up, **risk ratios** and **risk odds ratios** are estimable from D_k/N_k, $k = 0, 1$.

Precision in HR estimation

Model-based (Poisson) variance of log(IR) estimated:

$$\widehat{V}_{\text{coh}} = \frac{1}{D_1} + \frac{1}{D_0} = \frac{1}{\text{no. exp'd cases}} + \frac{1}{\text{no. unexp'd cases}}.$$

$\Leftrightarrow$ *the more cases, the better precision!*

- Approximate 95% CI for HR:

$$\text{IR} \times \exp\left\{\pm 1.96 \times \sqrt{\widehat{V}_{\text{coh}}}\right\}$$

- Does not depent on group sizes N_1, N_0 or person-times Y_1, Y_0 as such, even if these were millions.

- Yet, for rare outcomes, large populations are needed to obtain enough cases for adequate precision.

Problems with full cohort design

Obtaining exposure and covariate data

- Slow and expensive in big populations, especially with
  - measurements from biological specimens, like genotyping, antibody assays, *etc.*
  - occupational exposure histories in manual records.
- Easier with questionnaire and register data
  - Yet, analysis of time-dependent exposures can be complicated.
- *Can we obtain equally valid estimates with nearly as good precision by some other strategies*?
- **Yes, we can!**

Crude estimator of HR revisited

- The incidence rate ratio (IR) can be expressed as **exposure odds ratio** (EOR)

$$\mathrm{IR} = \frac{D_1/D_0}{Y_1/Y_0} = \frac{\textit{exp're odds}\text{ in cases}}{\textit{exp're odds}\text{ in study base}} = \mathrm{EOR}$$

- Describes exposure distribution in cases compared to that in the whole study population or study base.
- Implication for more efficient, outcome-selective design:
  - **Numerator**: Collect exposure data on *all cases*.
  - **Denominator**: Estimate the ratio of person-times Y_1/Y_0 by collecting risk factor data from a **random sample** from the whole study population.

Two-phase or case-control designs

General principle: Sampling of subjects from a given study population (SP) – **1st phase sample** – to a **2nd phase sample** is **outcome-selective**.

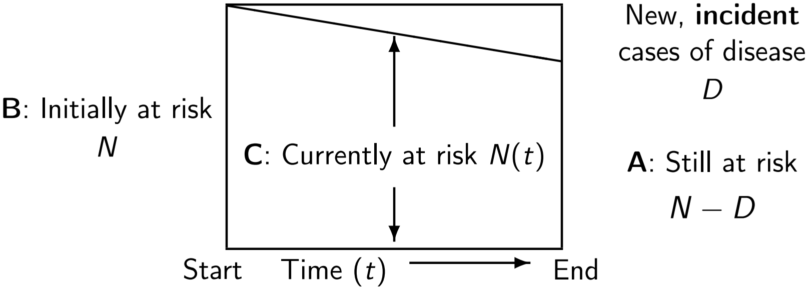
Ideally: SP = subjects who would be included as cases,
if they got the outcome in the study

- *Cohort-based studies*: SP = **cohort** or **closed** population of well-identified subjects under intensive follow-up for outcomes (*e.g.* the Dutch cohort).
- *Register-based studies*: SP = **open** or **dynamic** population in a region covered by a disease register (*e.g.* 50-62 y old women in Finland 1995-2007)
- *Hospital-based* studies: SP = dynamic **catchment** population of cases – may be hard to identify

2nd phase sampling in longitudinal base

Simplified ideal setting – like in outbreak studies:

- Complete follow-up of a cohort of initially healthy subjects with no losses during a fixed risk period.



New, **incident** cases of disease D

B: Initially at risk N

C: Currently at risk $N(t)$

A: Still at risk $N - D$

Start Time (t) End

- Possible sampling frames of controls: **A, B** and **C**

Sampling designs for control selection

A: Case-noncase sampling

- Controls chosen from those $N - D$ subjects still at risk (healthy, non-cases) *at the end* of the follow-up.

B: Case-cohort (CC) sampling

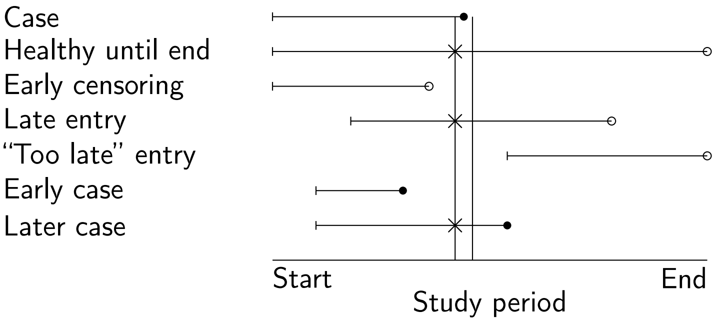
- The control group or **subcohort** is a random sample of the whole cohort (N) *at the start* of the follow-up.

C: Density sampling

- Controls drawn at random times *during the follow-up* from those currently at risk at each of these times.
- **Nested case-control design (NCC)**
  A set of controls is sampled from the **risk set** *at each time t of diagnosis* of a new case.

NCC: Risk-set or time-matched sampling

- Follow-up affected by late entry & right-censoring.
- Sampling frame to select controls for a given case: Other members ($\times$) of the **risk set** $R(t_i)$ at t_i, *i.e.* those at risk at the **time of diagnosis** t_i of case i.



Case

Healthy until end

Early censoring

Late entry

“Too late” entry

Early case

Later case

Start

End

Study period

- **Sampled risk set** $\widetilde{R}(t_i) = \{\text{case}\} \cup \{\text{controls}\}$.
- Controls can be resampled – and may later be cases.

Use of different designs

A: Case-noncase or **epidemic** case-control study

- Works well in studies on acute outbreaks.
- Problems with chronic diseases: variable follow-up, competing events, censoring, left-truncation

B: Case-cohort design

- Good when many outcomes are of interest, and measurements of risk factors from stored material (*e.g.* biological specimens) are relatively stable.

C: Density sampling, esp. nested case-control design

- The most popular in studies of chronic diseases.
- The only viable design in an open population.

Designs* B *and* C *still ignored in many textbooks!

Cross-sectional study base

- Study base = population at a given time point t.
- Cases are **prevalent**: they have the outcome at t.
- Common *e.g.* in studies of birth defects & in genetic epidemiology of "chronic" phenotypes (*e.g.* T2D).
- Alternative sampling designs:

  A: **Case-noncase** sampling: Controls are a random sample from the healthy; free from outcome at t.

  Direct estimability of **prevalence odds ratio**.

  B: **Case-cohort** sampling: Control group = subcohort, i.e. random sample of the whole population at t.

  Direct estimability of **prevalence ratio**.

Study base and sampling strategies

	Type of study base & population		
	Study base longitudinal		Study base
Sampling strategy	Open pop'n	Closed pop'n	cross-sectional
Complete enumeration	Incidence statistics	Classical cohort study	Health survey
A: Case-noncase	–	epidemic case-control s.	prevalence case-control s.
B: Case-cohort	–	case-cohort study	prevalence case-cohort s.
C: Density sampling	density case-control study	density case-control study	–

Dashes denote designs that are not possible

What comparative parameter is estimated?

- Longitudinal base: Simple summary of 2nd phase data

	exposed	unexposed	total
cases	D_1	D_0	D
controls/subcohort	C_1	C_0	C

- Depending on study base & sampling strategy, the empirical **exposure odds ratio** (EOR)

$$\text{EOR} = \frac{D_1/D_0}{C_1/C_0} = \frac{\text{cases: exposed / unexposed}}{\text{controls: exposed / unexposed}}$$

  is a consistent estimator of

  (A) risk odds ratio, (B) risk ratio, (C) hazard ratio,

- **NB.** In case-cohort studies with variable follow-up times C_1/C_0 is substituted by $\widehat{Y}_1/\widehat{Y}_0$, from estimated p-years.

Exposure odds ratio in density sampling

- Simply put: Exposure odds C_1/C_0 among controls estimates consistently exp. odds Y_1/Y_0 of p-times.
  – *An instance of PPS-sampling!*

$\Rightarrow$ Crude EOR $= (D_1/D_0)/(C_1/C_0)$ is a consistent estimator of the incidence rate ratio IR in the whole population = *target of inference at 2nd phase*

$\Rightarrow$ EOR is a consistent estimator of the hazard ratio HR.

- Assumes stability of exposure distribution over time.
  – May be unrealistic with a closed study population.
- Solution: **Time-matched** sampling of controls from **risk sets**, *i.e.* NCC & matched analysis.

Prentice & Breslow (1978, *Biometrika*)

Statistical precision and efficiency

With case-noncase (**A**) or density (**C**) sampling of controls (unmatched), estimated variance of crude log(EOR):

$$\widehat{V}_{\text{caco}} = \frac{1}{D_1} + \frac{1}{D_0} + \frac{1}{C_1} + \frac{1}{C_0} \approx \left(\frac{1}{D_1} + \frac{1}{D_0}\right)\left(1 + \frac{D}{C}\right)$$

= full cohort variance + variance from control sampling

- Determined essentially by the numbers of cases.
- With $C/D \geq 4$, $\widehat{V}_{\text{caco}}$ not much bigger than the full cohort variance $\widehat{V}_{\text{coh}}$ with same numbers of cases.

$\Rightarrow$ ***Small sampling fraction, high cost-efficiency!***

Further matching in NCC studies

- For each case choose 1 or more (rarely > 4) controls with same age, sex, region, exposure time, *etc.*
  – *Maximal stratification?*
- Improves efficiency, if matching factors are strong determinants of outcome.
- Matching for *storage time, freeze-thaw cycle & analytic batch* improves **comparability of measurements** from biospecimens.
- **Overmatching** may induce bias or reduce efficiency.
- **Counter-matching**: choose controls who are **different** from cases w.r.t. a surrogate of the main risk factor
  – Can improve efficiency.

Statistical analysis: full-cohort design

- Model-based, assumes sampling from a superpopulation
- Binary outcome: binary regression models.
- Time-to-event outcomes: **Cox model** for hazard rates

$$h_i(t;\beta) = h_0(t)\exp\{x_{i1}\beta_1 + \cdots + x_{ip}\beta_p\}, \quad i = 1, \ldots, N$$

- e^{β_j} = hazard ratio (HR) for unit change in X_j.
- **Partial log-likelihood**, estimating equations,

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \delta_i \left[\mathbf{x}_i - \sum_{k \in R(t_i)} e_k \mathbf{x}_k \Big/ \sum_{k \in R(t_i)} e_k\right] = 0,$$

$R(t_i)$ = risk set at event time t_i, δ_i = event/censoring indicator for subject i at t_i, and $e_k = \exp\{\mathbf{x}_k^\top \boldsymbol{\beta}\}$.

Analysis of two-phase designs: unweighted

- Binary regression or Cox model as appropriate.
- NCC: **Stratified partial likelihood** ⇔ **conditional logistic regression** (Thomas 1977, *JRSS A*)
- CC: **Pseudo-likelihood** (Prentice 1986 *Biometrika*).
- In both instances, estimating equations:

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \delta_i \left[\mathbf{x}_i - \sum_{k \in \widetilde{R}(t_i)} e_k \mathbf{x}_k \Big/ \sum_{k \in \widetilde{R}(t_i)} e_k \right] = 0,$$

$$\begin{aligned} \widetilde{R}(t_i) &= \text{sampled risk set at } t_i \\ &= \{\text{case}\} \cup \{\text{time-matched controls of case}\}, \text{ or} \\ &= \{\text{case}\} \cup \{\text{subcohort members at risk}\}. \end{aligned}$$

- CC: estimation of $\text{cov}(\widehat{\boldsymbol{\beta}})$ requires some extra work.

Analysis of two-phase designs: weighting

- Provides gains in efficiency in certain circumstances.
- Basic idea: **HT-estimation** with weight $d_i = 1/\pi_i$; π_i = inclusion probability to the 2nd phase sample.
- Cases: $\pi_i = 1 \Rightarrow$ weight $= 1$.
- Non-cases: π_i to be sampled …

NCC: … as control – depends on length of follow-up time; estimable in many ways,

CC: … into the subcohort – simply estimated $\widehat{\pi}_i = n_{\text{non-cases}}/N_{\text{non-cases}}$.

- **Weighted partial likelihood**: Sampled risk set $\widetilde{R}(t_i)$ includes in addition future cases who are at risk at t_i

$$\mathbf{U}(\beta) = \sum_{i=1}^{N} \delta_i \left[\mathbf{x}_i - \sum_{k \in \widetilde{R}(t_i)} \frac{1}{\widetilde{\pi}_k} e_k \mathbf{x}_k \Big/ \sum_{k \in \widetilde{R}(t_i)} \frac{1}{\widetilde{\pi}_k} e_k \right] = 0.$$

Utilization of auxiliary variables

- The 1st phase sample (whole cohort) may contain data that are informative of risk factors only obtainable from the 2nd phase sample (cases and controls/subcohort).
- Can be used to increase efficiency via
  - post-stratification,
  - multiple imputation,
  - calibration of weights.
- Ideas recently adopted from sampling theory literature.
- Calibration: Use of delta-beta influence functions coupled with multiple imputation.

see *e.g.* Lumley (2010) and R package `survey`

Conclusion

- Various cost-efficient outcome-selective sampling designs are widely used in epidemiology.
- Plenty of other refinements – not covered here.
- Up to the late 1980s, methods were mostly developed without reference to sampling literature.
- Since 1990s, many ideas learned from sampling theory.
- This has led to further improvements in design and analysis of epidemiologic studies for better efficiency.
- Intensive methodological research continues with more active and fruitful exchange between statisticians working in different realms.

Selected references: classics and newer ones

Borgan, Ø., Samuelsen, S-O. (2003). A review of cohort sampling designs . . . *Norsk Epidemiologi* **13**: 239-248.

Borgan, Ø., Breslow, N.E., . . . , Scott, A., Wild, C.J. (eds.) (2018). *Handbook of Statistical Methods for Case-Control Studies*. Chapman and Hall/CRC.

Keogh, R., Cox, D.R. (2014). *Case-Control Studies*. CUP.

Lumley, T. (2010). *Complex Surveys: . . .* Wiley.

Miettinen, O.S. (1982). Design options in epidemiologic studies . . . *Scand J Work Env Health* **8(Suppl 1)**:1-7.

Prentice, R.L. (1986). A case-cohort design . . . *Biometrika* **73**: 1-11.

Prentice, R.L., Breslow, N.E. (1978). Retrospective studies and failure time models. *Biometrika* **65**: 153-158.

Thomas, D.C. (1977). Addendum. *JRSS A*: **140**: 483-485.