

Survey Methods & Sampling Techniques

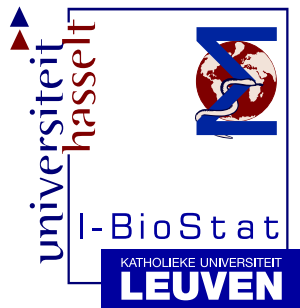
Geert Molenberghs

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

Katholieke Universiteit Leuven & Universiteit Hasselt, Belgium

geert.molenberghs@uhasselt.be

www.censtat.uhasselt.be



Master in Statistics, Universiteit Hasselt

Master in Quantitative Methods, Katholieke Universiteit Brussel

Contents

I	Introduction and Basic Concepts	1
0	Relevant References	2
1	The Belgian Health Interview Survey	7
2	General Concepts of Surveys	30
3	Population and Sampling	44
II	Simple Random Sampling	122
4	General Concepts and Design	123

5	Analysis	129
6	Sample Size Determination	156
III	A First Perspective on Software	180
7	General Considerations Regarding Software	181
8	SAS and The Belgian Health Interview Survey	195
IV	Systematic Sampling	221
9	General Concepts and Design	222
10	Analysis	245
V	Benchmark (Ratio) Estimators	272
11	General Concepts and Design	273
12	Analysis	279

VI	Stratification	321
13	General Concepts and Design	322
14	Analysis	350
15	Sample Size Determination and Allocation	392
VII	Multi-Stage Sampling and Clustering	409
16	General Concepts and Design	410
17	Analysis	437
18	Complex-Model-Based Analysis	461
VIII	Weighting	527
19	Analysis	561
20	Example: The Belgian Health Interview Survey	588

IX	Integrated Analysis of Belgian Health Interview Survey	610
21	Key Perspective Elements	611
22	Means, Proportions, and Frequencies	616
23	Linear Regression	662
24	Logistic Regression	694
25	Selecting a Sample Using SURVEYSELECT	740
26	Some Selected Examples From STATA	762
X	Incompleteness	769
27	General Concepts	770
28	Simplistic Methods	779
29	Direct Likelihood Maximization	785
30	Multiple Imputation	795

31 **Non-Gaussian Data**836

32 **Incompleteness in the Belgian Health Interview Survey** 842

33 **Sensitivity Analysis: A Case Study** 852

Part I

Introduction and Basic Concepts

Chapter 0

Relevant References

- Barnett, V. (2002). *Sample Survey: Principles and Methods (3rd ed.)*. London: Arnold.
- Billiet, J. (1990). *Methoden van Sociaal-Wetenschappelijk Onderzoek: Onderwerp en Dataverzameling*. Leuven: Acco.
- Billiet, J., Loosveldt, G., and Waterplas, L. (1984). *Het Survey-Interview Onderzocht*. Sociologische Studies en Documenten, **19**, Leuven.
- Brinkman, J. (1994). *De Vragenlijst*. Groningen: Wolters-Noordhoff.
- Chambers, R.L. and Skinner, C.J. (2003). *Analysis of Survey Data*. New York:

Wiley.

- Cochran, W.G. (1977). *Sampling Techniques*. New York: Wiley.
- Foreman, E. K. (1991). *Survey Sampling Principles*. New York: Marcel Dekker.
- Fowler, Jr., F.J. (1988). *Survey Research Methods*. Newbury Park, CA: Sage.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R. (2004). *Survey Methodology*. New York: Wiley.
- Heeringa, S.G., West, B.T., and Berglund, P.A. (2010). *Applied Survey Data Analysis*. Boca Raton: Chapman & Hall/CRC.
- Kish, L. (1965). *Survey Sampling*. New York: Wiley.
- Kottnerus, P. (2003). *Sample Survey Theory*. New York: Springer.

- Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: Wiley.
- Lehtonen, R. and Pahkinen, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. Chichester: Wiley.
- Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley.
- Levy, P. and Lemeshow, S. (1999). *Sampling of Populations*. New York: Wiley.
- Little, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, **77**, 237–250.
- Little, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistical Institute*, **15**, 1–15.
- Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data* (2nd

ed.). New York: Wiley.

- Lynn, P. (2009). *Methodology of Longitudinal Surveys*. Chichester: Wiley.
- Molenberghs, G. and Kenward, M.G. (2007). *Missing Data in Clinical Studies*. New York: Wiley.
- Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. New York: Springer.
- Moser, C.A., Kalton, G. (1971). *Survey Methods in Social Investigation*. London: Heinemann.
- Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Scheaffer, R.L., Mendenhall, W., and Ott L. (1990). *Elementary Survey Sampling*. Boston: Duxbury Press.

- Skinner, C.J., Holt, D., and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.
- Som, R.J. (1996). *Practical Sampling Techniques* (3rd ed.). New York: Marcel Dekker.
- Swyngedouw, M. (1993). Transitietabelanalyse en ML-schattingen voor partieel geclassificeerde verkiezingsdata via loglineaire modellen. *Kwantitatieve Methoden*, **43**, 119–149.
- Vehovar, V. (1999). Field substitution and unit nonresponse. *Journal of Official Statistics*, **15**, 335–350.
- Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer.

Chapter 1

The Belgian Health Interview Survey

- ▷ Background
- ▷ Information about the sample
- ▷ Information about the design

1.1 Background

- **Conducted in years:** 1997 – 2001 – 2004
- **Commissioned by:**
 - ▷ Federal government
 - ▷ Flemish Community
 - ▷ French Community
 - ▷ German Community
 - ▷ Walloon Region
 - ▷ Brussels Region

- **Executing partners:**

- ▷ Scientific Institute Public Health–Louis Pasteur
- ▷ National Institute of Statistics
- ▷ Hasselt University (formerly known as Limburgs Universitair Centrum)
- ▷ Website: <http://www.iph.fgov.be/epidemie/epien/index4.htm>

- **Goals:**

- ▷ Subjective health, from the respondent's perspective
- ▷ Identification of health problems
- ▷ Information that cannot be obtained from care givers, such as
 - * Estimation of prevalence and distribution of health indicators
 - * Analysis of social inequality in health and access to health care
 - * Study of possible trends in the health status of the population

- **Domains:**

- ▷ Complaints and symptoms
- ▷ Health status
- ▷ Use of health services
- ▷ Life style
- ▷ Socio-economic variables

1.2 Differences in Categories Covered

Category	1997	2001	2004
Household questionnaire			
Health and society		*	*
Face-to-face interview			
Acute affections		*	
<i>Physical activity</i>	*	*	
Waiting list			*
Contacts with primary health care		*	*
Maternal and infantile health	*		
<i>Traumatism, accidents, violence, dog bites</i>	*	*	
Written questionnaire			
Morbidity	*		
Health complaints		*	*
Knowledge of/and behavior towards HIV/AIDS	*		*
Consumption of other products		*	*
<i>Traumatism, accidents, violence, dog bites</i>			*
Patient satisfaction		*	
Prevention: diabetes		*	*
<i>Physical activity</i>			*

1.3 Overview of Design

- **Regional stratification:** *fixed a priori*
- **Provincial stratification:** *for convenience*
- **Three-stage sampling:**
 - ▷ Primary sampling units (PSU): Municipalities: *proportional to size*
 - ▷ Secondary sampling units (SSU): Households
 - ▷ Tertiary sampling units (TSU): Individuals

- Over-representation of German Community
- Over-representation of 4 (**2**) provinces in 2001 (**2004**):

Limburg	Hainaut
Antwerpen	Luxembourg

- Sampling done in 4 quarters: Q1, Q2, Q3, Q4

1.4 Regional Stratification

Region	1997		2001		2004	
	Goal	Obt'd	Goal	Obt'd	Goal	Obt'd
Flanders	3500	3536	$3500+550=4050$	4100	$3500+450$	
+ elderly					$+450=4400$	4513
Wallonia	3500	3634	$3500+1500=5000$	4711	$3500+900$	
+ elderly					$+450=4850$	4992
Brussels	3000	3051		3000	3000	
+ elderly					$+350=3350$	3440
Belgium	10,000	10,221	$10,000+2050=12,050$	12,111	$10,000+1350$	
+ elderly					$+1250=12,600$	12,945

1.5 Provincial Stratification in 1997

Province	sample #	sample %	pop. %
Antwerpen	945	26.7	27.7
Oost-Vlaanderen	812	23.0	23.0
West-Vlaanderen	733	20.7	19.1
Vlaams-Brabant	593	16.8	17.0
Limburg	453	12.8	13.2
Hainaut	1325	36.5	38.7
Liège	1210	33.3	30.6
Namur	465	12.8	13.2
Brabant-Wallon	356	9.8	10.3
Luxembourg	278	7.6	7.3
Brussels	3051		

1.6 Provincial Stratification in 2001

Province	pop.	% in region	# interviews				actual	# groups	# towns	rate p. 1000
			theor.	round	oversp.	sum				
Antwerpen	1,640,966	27.7	969	950	350	1300	1302	26	19	0.79
Oost-Vlaanderen	1,359,702	22.9	803	850	0	850	874	17	17	0.63
West-Vlaanderen	1,127,091	19.0	665	650	0	650	673	13	13	0.58
Vlaams-Brabant	1,011,588	17.1	598	600	0	600	590	12	12	0.59
Limburg	787,491	13.3	465	450	200	650	661	13	13	0.83
Flanders	5,926,838	100	3500	3500	550	4050	4100	81	74	0.68
Hainaut	1,280,427	39.3	1256	1250	500	1750	1747	35	27	1.37
Liège	947,787	29.0	929	950	0	950	935	19	19	1.00
Namur	441,205	13.5	433	450	0	450	435	9	7	1.02
Brabant Wallon	347,423	10.7	341	300	0	300	291	6	6	0.86
Luxembourg	245,140	7.5	241	250	1000	1250	1303	25	21	5.10
Wallonia	3,261,982	100	3200	3200	1500	4700	4711	94	80	1.44
German comm.	70,472	1.1	300	300	0	300	294	6	6	4.26
Wallonia+German	3,332,454	100	3500	3500	1500	5000	5005	100	86	1.50
Brussels	954,460	100	3000	3000	0	3000	3006	60	18	3.14
Belgium	10,213,752	100	10,000	10,000	2050	12,050	12,111	241	178	1.18

1.7 Provincial Stratification in 2004

Province	Goal	Obtained
Antwerpen	1100	1171
Oost-Vlaanderen	900	944
West-Vlaanderen	750	814
Vlaams-Brabant	650	561
Limburg	1000	1023
Hainaut	1500	1502
Liège	1200	1181
Namur	550	531
Brabant-Wallon	400	446
Luxembourg	1200	1332
Brussels	3350	3440

1.8 Overview of Stratification

- Regions (Flanders, Brussels, Wallonia) within the country
- Provinces within a region
- The corresponding selection probabilities factor into the *weights* of the previous section
- A full account of stratification requires more than just the introduction of weights, but including weights that properly reflect stratification is a first and very important step towards a correct analysis

1.9 Multi-Stage Sampling: Primary Sampling Units



Towns

- Within each province, order communities \propto *size*
- Systematically sample in groups of 50

- Representation with certainty of larger cities.

For 1997:

- ▷ Antwerpen: 6 groups
- ▷ Liège and Charlerloi: 4 groups each
- ▷ Gent: 3 groups
- ▷ Mons and Namur: 2 groups each
- ▷ All towns in Brussels

- Representation ensured of respondents, living in smaller towns

- For 2001, the list of municipalities selected as least once:

Municipality	# times selected			Municipality	# times selected		
	min	max	actual		min	max	actual
Antwerpen	7	8	8	La Louvière	2	3	2
Mechelen	1	2	1	Tournai	1	2	2
Leuven	1	2	1	Mouscron	1	2	1
Gent	2	3	2	Arlon	2	3	3
Hasselt	1	2	1	Marche en Famenne	1	2	2
Brugge	1	2	1	Aubange	1	2	1
Liège	3	4	4	Bastogne	1	2	1
Seraing	1	2	1	Namur	2	3	3
Verviers	1	2	1	Eupen	1	2	2
Charleroi	5	6	5	Brussels	All towns at least once		
Mons	2	3	3				

1.10 Multi-Stage Sampling: Secondary Sampling Units



Households

- List of households, ordered following
 - ▷ statistical sector
 - ▷ age of reference person
 - ▷ size of household
- clusters of 4 households selected
- households within clusters randomized
- twice as many *clusters* as *households* needed, to account for refusal and non-responders

1.11 Multi-stage Sampling: Tertiary Sampling Units

Individual Respondents

- **Households of size ≤ 4 :** all members
- **Households of size ≥ 5 :**
 - ▷ reference person and partner (if applicable)
 - ▷ other households members selected on birthday rule in 1997 or by prior sampling from household members in 2001 and 2004
 - ▷ maximum of 4 interviews per household

1.12 Overview of Multi-Stage Sampling and Clustering

- Due to the three-way sampling method used
- Clustering and multi-stage sampling are **not** the same, even though they often occur together
- (Artificial) examples where they do not occur together:
 - ▷ **Clustering without multi-stage sampling:** select households and then always all members
 - ▷ **Multi-stage sampling without clustering:** select towns, then one household, then one member within a household

- Within this study, there are two sources of clustering:
 - ▷ Households within towns
 - ▷ Individual respondents within households
- Taking clustering into account can be done in several ways:
 - ▷ Ad hoc, using the so-called *design factor*
 - ▷ Using specific survey analysis methods, when the emphasis is not on the clustering itself but it is taken into account as a nuisance factor
 - ▷ Using models for hierarchical (clustered) data, such as linear or generalized linear mixed models, multi-level models, etc.

1.13 Weights

- Region
- Province
- Age of reference person
- Household size
- Quarter
- Selection probability of individual within household
- Taking this into account is relatively easy, even with standard software

1.14 Incomplete Data

- Types of incompleteness in this survey:
 - ▷ **Household level**
 - * Households with which no interview was realized
 - * Households which explicitly refused
 - * Households which could not be contacted
 - ▷ **Individual level**
 - ▷ **Item level**
- In addition, the reason of missingness needs to be considered. For example, is missingness due to illness of the interviewer, or is it related to the income and social class of the potential respondent?

- General missing data concepts as well as survey-specific missing data concepts need to be combined
- The study of incomplete survey data requires some non-trivial statistical skill

1.15 Design → Analysis

- Weights & selection probabilities
- Stratification
- Multi-stage sampling & clustering
- Incomplete data

Chapter 2

General Concepts of Surveys

- ▷ Census *versus* survey
- ▷ Applications of surveys
- ▷ Ingredients of surveys

2.1 The Census

- (*volkstelling, recensement*).
- The oldest form of data collection: the Bible reports on the census, for which everyone had to go back to their native town.
- Original goals: organization of tax payments; political representation.
- Currently: the same, supplemented with collection of a wide variety of relevant information (race, age, constitution of households, quality of life, . . .).
- Censuses are broad: it is hard to go in any depth on a particular topic.

- Census are infrequent: A common periodicity is 10 years (Belgium: 1991, 2001,...).
- Often conducted by the national statistical offices:
 - ▷ Belgium: National Institute of Statistics (NIS/INS).
 - ▷ US: Bureau of Census (federal).

2.2 A Survey Rather Than a Census

- Alternative to census: organization of a well-targetted



survey

with a limited but precise scope.

- ▷ “Which are the major themes in the public opinion?” In view of organizing the election campaigns of political parties.
- ▷ “What are consumers’ demands?” in market research.
- ▷ Research on facts, behavior,... in sociology, psychology.

- While originating from the humanities, they are nowadays broadly applied:
 - ▷ **Health Interview Survey:** subjective health of population (NIS/INS; US National Institutes of Health).
 - ▷ Quality of life in patients with serious illnesses, such as cancer, AIDS, Alzheimer.
 - ▷ For many **mental health** outcomes, surveys/questionnaires may be the only way to collect data: schizophrenia (Positive and Negative Symptoms Scale, PANSS; Brief psychiatric rating scale, BPRS), depression (Hamilton depression scale, HAMD),...
 - ▷ **Unemployment:** Statistics about jobs and the employment market.
 - ▷ **Income and expenses:** Patterns of consumer behavior and expectations are important predictors for trends in the economy.

▷ **Crime research.**

- * Traditionally, police reports were used to compile crime-related statistics.
- * This leads to a distorted (biased) picture: not all crime is being reported, especially not the smaller or very common crimes.
- * The major crimes, where casualties or other victims have to be counted, are relatively well reported.
- * Advantage of surveys: not only the crime itself, but also related large subjective aspects, e.g., feeling (un)safe, can be documented; better coverage.

▷ **Agriculture:** To obtain a good picture of yield, yearly variations, variations on a longer time scale, etc.

▷ **Housing:** Costs, expectations,...

▷ **Job satisfaction.**

2.3 Aspects of Surveys

- ▷ scientific question
 - ▷ selection of instruments
 - ▷ questionnaire design
 - ▷ other design aspects
 - ▷ fieldwork organization
 - ▷ interviewing methods
 - ▷ sample selection
 - ▷ analysis methods
- All aspects have an impact on quality, captured through:
 - ▷ psychometric concepts: reliability, validity,...
 - ▷ statistical concepts: precision, bias,...
 - ▷ general, vaguely defined concepts: accuracy,...

- Surveys almost always result from multi-disciplinary teamwork:
 - ▷ sociology, psychometrics, statistics, mathematics,...
 - ▷ supplemented with substantive sciences (subject matter areas): medicine, political sciences, epidemiology, economy and market research,...
- Surveys are used for a wide variety of measurement processes and methods of data collection.
- We will focus on
 - ▷ Surveys that produce **statistics**: quantitative, numerical descriptions of relevant aspects of a **study population**.
 - ▷ Data generally arise from respondents' answers to questions.
 - ▷ The group of respondents is a small portion of the population: the **sample** (*steekproef, échantillon*).

- This course's focus will be on the quantitative design and analysis aspects.
- It is important to study all options which lead to data collection.
- If the survey option is chosen, then all aspects of design, conduct, and analysis have to be studied and planned very carefully.

2.4 Who Organizes Surveys?

- **Government:** central, regional, and local governments; government-sponsored research institutes: NIH, CDC,...
- **Research institutes:** universities, colleges, other research institutes,...
- **Private initiative:** market research companies,...

2.5 Overview of Survey Ingredients

Choice for a survey. A survey is expensive.

Use it when no other source to obtain the data exists:

- ▷ The variables/items are not available.
- ▷ The variables/items are available, *but not in conjunction with other variables.*

Example: both health information as well as life style information is available, but not jointly so.

- ▷ Otherwise, avoid requesting information that is already available.

Standardized measurements.

- ▷ Measurement instruments which collect data in a standardized fashion.
- ▷ Good psychometric properties:
 - * Are questions designed by experts?
 - * Are literature results available about validity and reliability?
 - * Is the validity and reliability studied for the purpose of this research?
 - * Is question lab being used?
 - * Is a pilot study being undertaken?

Data collection and interviewing.

- ▷ Collect information in the same way for all respondents.
- ▷ Level and type of training for interviewers:
 - * manual
 - * on-line documentation
 - * hotline

- ▷ The interviewer must not influence the response.
- ▷ The interviewer has to ensure that the question is answered with the highest possible accuracy.
- ▷ A good question has to fulfill the following properties:
 - * It has to be possible to ask the question as formulated.
 - * It has to be possible to formulate and answer the question without having to amplify on it.
 - * If amplification is necessary nevertheless, standardized procedures must exist as to how this should take place.

Design. Includes:

- ▷ definition of population
- ▷ sample frame
- ▷ probability sampling method
- ▷ See next chapter

Probability sampling.

Analysis methodology. Choose the analysis methodology in accordance with the design.

Non-response.

Chapter 3

Population and Sampling

- ▷ Non-sampling-based methods
- ▷ Sampling
- ▷ Key definitions
- ▷ Notation
- ▷ Examples
- ▷ Basic quantities

3.1 Non-Sampling-Based Methods

3.1.1 Census

- In a census, the entire population is studied:

sample = population

- Theoretically simple \longleftrightarrow practically complicated and expensive.
- Alternative: a portion of the population.
- How is this portion selected?

3.1.2 Pilot Study

- Sometimes, only a global picture is required:
 - ▷ Press reporters or politicians, feeling the pulse of the public opinion.
 - ▷ Product developers, obtaining a feel for promising products.
- An informal study or **pilot study** is then sufficient.
- Who is then eligible for interviewing?
 - ▷ those immediately available: friends, colleagues, mother-in-law,...
 - ▷ volunteers: those who return a form, etc.
- This is largely an *exception*.
- A pilot study can also be used as a 'preamble' to a full-fledged survey:
 - ▷ To try out the feasibility of the survey, also in terms of fieldwork.
 - ▷ As a specific device to support sample size calculation.

3.2 Sampling

- Sampling allows one to obtain a **representative** picture about the population, *without studying the entire population*.
- Two essential questions:
 - ▷ How is a sample selected?
 - ▷ How are the resulting sample data analyzed, to allow for statements about the population?
- In both cases we need **statistical sampling theory**.

3.3 Definitions

Survey population: The collection of units (individuals) about which the researcher wants to make quantitative statements.

Sample frame: The set of units (individuals) that has non-zero probability of being selected.

Sample: The subset of units that have been selected.

Probability sampling: The family of probabilistic (stochastic) methods by which a subset of the units from the sample frame is selected.

Design properties: The entire collection of methodological aspects that leads to the selection of a sample.

The probability sampling method is the most important design aspect.

Sample size: The number of units in the sample.

Analysis and inference: The collection of statistical techniques by which population estimands are estimated.

Examples: estimation of means, averages, totals, linear regression, ANOVA, logistic regression, loglinear models.

Estimand: The true population quantity (e.g., the average body mass index of the Belgian population).

Estimator: A (stochastic) function of the sample data, with the aim to “come close” to the estimand.

Estimate: A particular realization of the estimator, for the particular sample taken (e.g., 22.37).

We will consider several of these aspects in turn.

3.4 Population

- A population can be physical and/or geographical, but
- does not have to be an entire country or region.
- A population can be a cohort: all males born in Brussels in 1980.
- There can be geographical, temporal, and definition characteristics at the same time: all females living in Brussels, diagnosed with breast cancer between from 1990 until 1999 inclusive.

3.5 Sample Frame

- The sample frame “operationalizes” the population.
 - ▷ **Population:** All females living in Brussels, diagnosed with breast cancer between from 1990 until 1999 inclusive.
 - ▷ **Sample frame:** The National Cancer Register for the given years.
- There are three groups of units:
 - ▷ **1. Belonging to both the population and the sample frame:** This fraction should be as large as possible.
Their probability is ≥ 0 of being selected.
 - ▷ **2. Belonging to the population but not to the sample frame:** Can be damaging if too large and/or too different units.
Their probability of selection is 0.

- * If a selection is based on households, then dormitories, prisons, elderly homes, and homeless people have no chance of being selected.
 - * Driving licenses (US)
 - * Registered voters
 - * House owners
 - * Phone directories: excludes those without phone and those unlisted.
- ▷ 3. **Belonging to the sample frame but not to the population:** May contribute to cost, but is not so harmful otherwise.

For example, a survey on elderly can be conducted as follows:

- * select households from the general population
- * retain those who are “sufficiently old”
- * collect data on this subselected sample
- * But this procedure is clearly inefficient.

If group 1 is sufficiently large, then the sample frame is sufficiently representative.

- It is important to answer such questions as:
 - ▷ What percentage is excluded from selection?
 - ▷ How different are these groups?
 - It is possible to opt for a selection scheme with less than full coverage of the population, if it is sufficiently cheaper.
- Statistical and economic arguments have to be balanced.

3.6 Types of Sample Frame

- It is useful to think of a sample frame as a list.
- A list is a broad concept, there are widely different types.
 - ▷ **Static, exhaustive lists:**
 - * A single list contains all sample frame units
 - * The list exists prior to the start of the study
 - ▷ **Dynamic lists:**
 - * The list is generated together with the sample
 - * For example: all patients visiting a general practitioner during the coming year
 - * There are implications for knowledge about the selection probability

▷ **Multi-stage lists:**

- * The natural companion to multi-stage sampling (see Part **VII**)

● If selection is undertaken based on a list, one has to consider the list's quality:

▷ How has the list been composed?

▷ How does the updating take place?

▷ **Always** report:

- * who cannot be selected?

- * in what way do those who have selection probability equal to zero differ from the others?

- * who did have unknown selection probability

⇒ trustworthy, useful results

3.7 Sampling Methods

- We will study various sampling methods, and their rationale:

Simple random sampling: the standard method; studied to compare other methods with.

Systematic sampling: chosen to increase precision and/or to ensure sampling with certainty for a subgroup of units.

Stratification: performed:

- ▷ to increase precision of population-level estimates
- ▷ to allow for estimation at sub-population level
- ▷ a combination of both

Multi-stage procedures: decrease precision but facilitate fieldwork.

Differential rates: will often result from other sampling methods; the overall precision will decrease.

Benchmark estimation: may introduce some bias but is aimed to increase precision; there is a need for external sources.

- All methods, aimed at increasing precision, may actually decrease it in pathological cases, and vice versa.

3.8 Selection Probability

- The probability of an individual to be selected:
 - ▷ Should be known or estimable (consistently)
 - ▷ Does not have to be constant
 - ▷ The selection probability may not be known a priori, it is sufficient to know or estimate it *by the time of analysis*.

This is natural with dynamic lists.

Example: patients visiting a general practitioner during the coming year, by asking:

“How frequently have you visited the doctor during the last [time frame]?”

- If external factors, such as initiatives by respondents, influence the chance of being included, the integrity of the study is in jeopardy.

So, watch out for

- ▷ people who come to a meeting
- ▷ people who speak up most
- ▷ people who volunteer to respond
- ▷ people who are easy to access

- Procedure:

- ▷ Attach to each member of the sample frame a non-zero probability of being selected
- ▷ use probabilistic techniques to draw the sample

3.9 Sample Units

- A study can have units at several levels simultaneously (multi-stage sampling): towns, households, individuals.
- In such a case, either one or more levels can be of scientific interest:
 - ▷ Possibility 1: interest only in individuals
 - ▷ Possibility 2: interest in households and individuals simultaneously

- Examples of units:

- ▷ lots

- ▷ dwellings within lots

- ▷ apartments within dwellings

- ▷ property

- ▷ individuals

- ▷ children

- ▷ families

- ▷ households

3.10 Notation

- Within sampling theory, it is customary to identify population and sample frame: one speaks about *population*, but it actually should be *sample frame*.
- The notational conventions are slightly different than in other areas of statistics.

- In mathematical statistics, for example, one uses:

- ▷ Population:

$$X \sim N(\mu, \sigma^2)$$

- ▷ Sample (stochastic values):

$$X_i \sim N(\mu, \sigma^2), \quad i = 1, \dots, n$$

- ▷ Sample (realized values):

$$x_i, \quad i = 1, \dots, n$$

- ▷ Average:

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

- Our conventions:

Quantity	Population	Sample
Size	N	n
Unit index	I	i
Value for a unit	X_I	x_i
Average	\bar{X}	\bar{x}
Total	X	x
Total, estimated from sample		\hat{x}

- Estimators will be studied in Part II and later.

3.11 A Small Artificial Population

- Population

$$\mathcal{P} = \{1, 2, 3, 4\}$$

- Listing of Artificial Population:

I	Y_I
1	1
2	2
3	3
4	4

- $I = 1, \dots, 4$
- $N = 4$

3.11.1 Samples from Artificial Population

- Samples of size $n = 1$:

- ▷ Enumeration:

$$\mathcal{S}_1 = \left\{ \begin{array}{l} \{1\}, \\ \{2\}, \\ \{3\}, \\ \{4\} \end{array} \right\}$$

- ▷ $S = 4$

- ▷ $s = 1, 2, 3, 4$

- Samples of size $n = 2$, with ordering taken into account:

▷ Enumeration:

$$\mathcal{S}_2 = \{ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \\ \{2, 1\}, \{3, 1\}, \{4, 1\}, \{3, 2\}, \{4, 2\}, \{4, 3\}, \\ \{1, 1\}, \{2, 2\}, \{3, 3\}, \{4, 4\} \}$$

▷ $S = 16$

▷ $s = 1, \dots, 16$

- Samples of size $n = 2$, with ordering **not** taken into account:

▷ Enumeration:

$$\mathcal{S}_2 = \{ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \\ \{1, 1\}, \{2, 2\}, \{3, 3\}, \{4, 4\} \}$$

▷ $S = 10$

▷ $s = 1, \dots, 10$

- \mathcal{S} is itself a population, a **meta-population** of size S .
- A sampling mechanism assigns, to each member of the collection of samples, a probability of being selected.
- These probabilities are necessary to:
 - ▷ Study the properties of a sampling methods
 - ▷ Conduct estimation and statistical inferences
- The population itself can be studied for characteristics.

3.11.2 Characteristics of the Population

- Population average:

$$\bar{Y} = \frac{1}{4} \sum_{I=1}^4 Y_I = \frac{1 + 2 + 3 + 4}{4} = 2.5$$

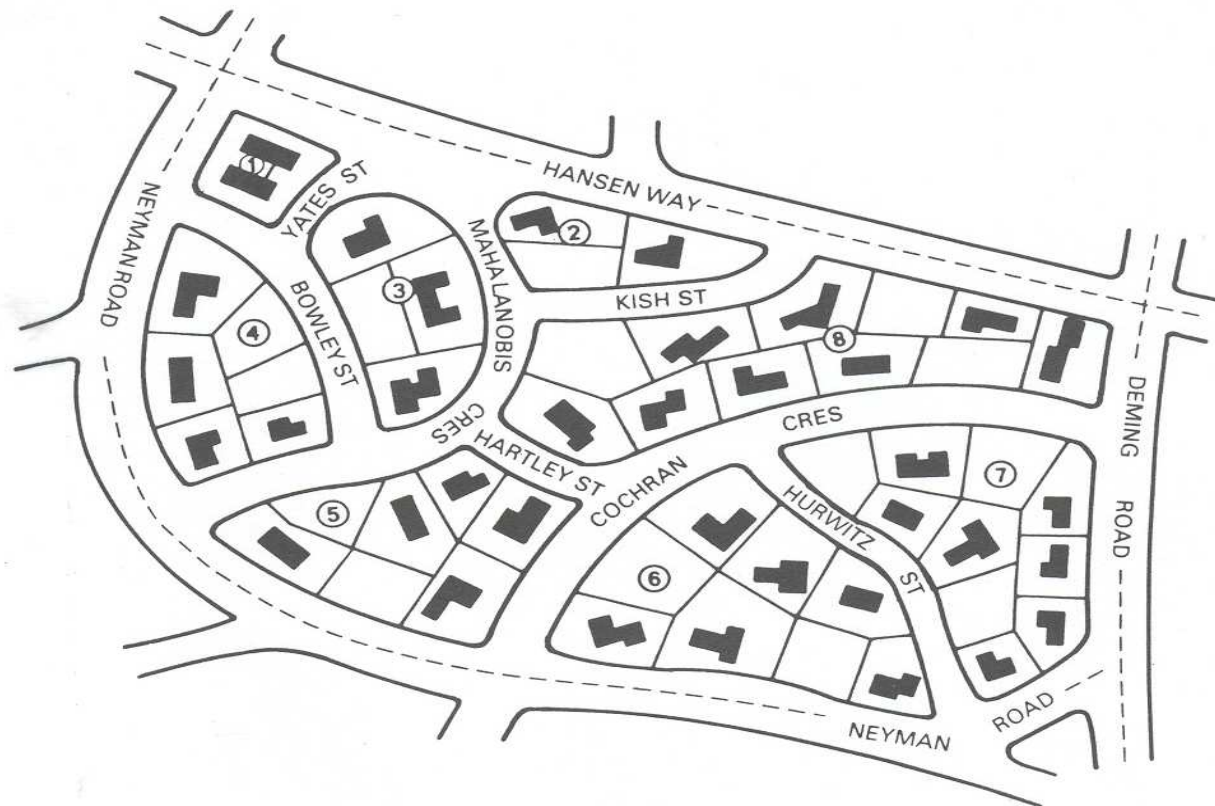
- Population variance:

$$\sigma_Y^2 = \frac{1}{4} \sum_{I=1}^4 (Y_I - \bar{Y})^2 = \frac{(1 - 2.5)^2 + (2 - 2.5)^2 + (3 - 3.5)^2 + (4 - 4.5)^2}{4} = 1.25$$

- Population total:

$$Y = \sum_{I=1}^4 Y_I = 1 + 2 + 3 + 4 = 10$$

3.12 Surveytown



3.12.1 Surveytown

- $N = 8$
- $I = 1, \dots, N = 8$
- Two variables:
 - ▷ X_I : number of building lots in block I
 - ▷ Y_I : number of dwellings (buildings) in block I

- Listing of Surveytown:

I	X_I	Y_I
1	1	1
2	3	2
3	4	3
4	6	4
5	7	5
6	8	6
7	10	7
8	11	8

- Population totals:

$$X = 50$$

$$Y = 36$$

There are 50 lots, 36 with dwellings, hence 14 empty lots.

- Population averages:

$$\bar{X} = 6.25$$

$$\bar{Y} = 4.50$$

- Population variances:

$$\sigma_X^2 = \frac{1}{8} \sum_{I=1}^8 (X_I - 6.25)^2 = 10.4375$$

$$\sigma_Y^2 = \frac{1}{8} \sum_{I=1}^8 (Y_I - 4.50)^2 = 5.25$$

3.12.2 Proportion

- The ratio of the number of dwellings to the number of lots:

$$R = \pi = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}} = 0.72$$

- A proportion can be considered the average of a random variable:
 - ▷ Define the (related, but different) population of all **lots**: $I = 1, \dots, 50$
 - ▷ Let

$$Z_I = \begin{cases} 1 & \text{if lot } I \text{ is built upon} \\ 0 & \text{if lot } I \text{ is empty} \end{cases}$$

Then,

$$Z = \sum_{I=1}^{50} Z_I = 36$$

$$\bar{Z} = \frac{1}{50} \sum_{I=1}^{50} Z_I = 0.72$$

- The population variance:

$$\begin{aligned}\sigma_Z^2 &= \frac{1}{50} \sum_{I=1}^{50} (Z_I - 0.72)^2 \\&= \frac{1}{50} [36(1 - 0.72)^2 + 14(0 - 0.72)^2] \\&= \frac{36}{50} \cdot (1 - 0.72)^2 + \frac{14}{50} \cdot (-0.72)^2 \\&= 0.72 \cdot (1 - 0.72)^2 + (1 - 0.72) \cdot (0.72)^2 \\&= 0.72 \cdot (1 - 0.72) \cdot [(1 - 0.72) + 0.72] \\&= 0.72 \cdot (1 - 0.72) \\&= R(1 - R) = \pi(1 - \pi)\end{aligned}$$

3.13 Population Quantities

- Population average:

$$\bar{Y} = \frac{1}{N} \sum_{I=1}^N Y_I$$

- Population total:

$$Y = \sum_{I=1}^N Y_I$$

- Population variance:

▷ We have calculated before:

$$\sigma_Y^2 = \frac{1}{N} \sum_{I=1}^N (Y_I - \bar{Y})^2$$

but we can also calculate:

$$S_Y^2 = \frac{1}{N-1} \sum_{I=1}^N (Y_I - \bar{Y})^2$$

▷ There is a rationale for each one of them:

- * σ_Y^2 is compatible with the maximum likelihood principle, and hence asymptotically unbiased

- * S_Y^2 is unbiased even in small samples; it follows from the least-squares principle

▷ The square root S_Y (σ_Y) is the standard deviation.

- Population covariance:

$$\sigma_{XY} = \frac{1}{N} \sum_{I=1}^N (X_I - \bar{X})(Y_i - \bar{Y})$$

$$S_{XY} = \frac{1}{N-1} \sum_{I=1}^N (X_I - \bar{X})(Y_i - \bar{Y})$$

- Population correlation:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{S_{XY}}{S_X S_Y}$$

3.14 Sampling Mechanisms

- Recall that a population \mathcal{P} with N members gives rise to a meta-population \mathcal{S} of S samples.
- A sampling mechanism assigns a probability P_s ($s = 1, \dots, S$) to each sample.
- Obviously, to be valid, the P_s must satisfy:
 - ▷ $P_s \geq 0$, for all $s = 1, \dots, S$
 - ▷ $\sum_{s=1}^S P_s = 1$

- For the Artificial Population, with $n = 2$:

s	Sample	Probability
1	$\{1,2\}$	P_1
2	$\{1,3\}$	P_2
3	$\{1,4\}$	P_3
4	$\{2,3\}$	P_4
5	$\{2,4\}$	P_5
6	$\{3,4\}$	P_6
7	$\{1,1\}$	P_7
8	$\{2,2\}$	P_8
9	$\{3,3\}$	P_9
10	$\{4,4\}$	P_{10}

3.14.1 Sampling With Equal Probabilities

- The simplest mechanism is to assign the same selection probability to each individual.
- There are two versions:

Without Replacement: Every individual can enter the sample at most once.

With Replacement: Every individual can enter the sample multiple times; precisely, between 0 and n times.

- Both give rise to **Simple Random Sampling** (see also Part II).
- For the Artificial Population, with $n = 2$:

s	Sample	P_s	
		Without	With
1	{1,2}	1/6	2/16
2	{1,3}	1/6	2/16
3	{1,4}	1/6	2/16
4	{2,3}	1/6	2/16
5	{2,4}	1/6	2/16
6	{3,4}	1/6	2/16
7	{1,1}	0	1/16
8	{2,2}	0	1/16
9	{3,3}	0	1/16
10	{4,4}	0	1/16

- ▷ Selection without replacement sets the selection probability for all samples with replication equal to 0.
- ▷ Under sampling with replacement, the heterogeneous samples are twice as likely to be selected as the homogeneous samples.
- ▷ The reason is that, for example, $\{1,2\}$, can be selected in the orders (1,2) and (2,1).
- ▷ In contrast, $\{1,1\}$ comes into being in only one way.
- ▷ (In general, probability depends on the number of permutations a sequence can have.)
- ▷ The above consideration implies that assigning the same probability of being selected to an individual is not the same as giving every sample the same probability of being selected.

- At any time in the sequence of sample takes, the selection probability of a given individual is $1/N$:

With Replacement: Since at any time there are N “balls in the urn”, the probability is

$$\frac{1}{N}$$

Without Replacement: For an individual to be selected at a given time (take), let us say $t + 1$:

$$\frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdot \frac{N-3}{N-2} \cdot \cdots \cdot \frac{N-t}{N-t+1} \cdot \frac{1}{N-t} = \frac{1}{N}$$

- Sampling without replacement is the norm:
 - ▷ Sampling with replacement has lower precision (see later).
 - ▷ Sampling with replacement is inconvenient for the fieldwork.

3.15 Sample Selection: Remarks

- It is important that samples be taken in a totally random fashion (or the closest approximation to it that one can accomplish in practice).
- Classical, historic models:
 - ▷ Balls drawn from an urn (e.g., lotto games)
 - ▷ Tossing of dies
- Modern, realistic model: computerized pseudo-random generators

- Samples can be taken for various units simultaneously:
 - ▷ Households and individuals within households simultaneously.
 - * Select all individuals within a household.
 The selection probability of an individual within household h :

$$\frac{1}{N_{HH}}$$
 with N_{HH} the number of households.
 - * Select one individual within a household.
 The selection probability of an individual within household h :

$$\frac{1}{N_{HH}} \cdot \frac{1}{M_h}$$
 with N_{HH} the number of households and M_h the number of individuals within household h .
 - ▷ The first probability is constant, the second one depends on the size of the household.
 - ▷ This has implications for the analysis.

3.16 Sample Quantities

- Sample fraction:

$$f = \frac{n}{N}$$

This quantity is relevant only in *finite populations*.

- Carefully distinguish between **three** quantities:

Population quantity: a quantity, computed using all N population units.

Sample quantity: the same quantity, computed using the n units selected into the sample.

Estimate: an “approximation” of the population quantity, using only of the n sample units.

Quantity	Population	Sample	Estimate from sample for population
Average (mean)	$\bar{Y} = \frac{1}{N} \sum_{I=1}^N Y_I$	$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$	
Total (sum)	$Y = \sum_{I=1}^N Y_I$	$y = \sum_{i=1}^n y_i$	$\hat{y} = \frac{N}{n} \cdot \sum_{i=1}^n y_i$

3.16.1 Example: Artificial Population (Without Replacement)

s	Sample	P_s	\bar{y}_s	y_s	\hat{y}_s
1	{1,2}	1/6	1.5	3.0	6.0
2	{1,3}	1/6	2.0	4.0	8.0
3	{1,4}	1/6	2.5	5.0	10.0
4	{2,3}	1/6	2.5	5.0	10.0
5	{2,4}	1/6	3.0	6.0	12.0
6	{3,4}	1/6	3.5	7.0	14.0

3.16.2 Example: Artificial Population (With Replacement)

s	Sample	P_s	\bar{y}_s	y_s	\hat{y}_s
7	{1,1}	1/16	1.0	2.0	4.0
1	{1,2}	2/16	1.5	3.0	6.0
2	{1,3}	2/16	2.0	4.0	8.0
8	{2,2}	1/16	2.0	4.0	8.0
3	{1,4}	2/16	2.5	5.0	10.0
4	{2,3}	2/16	2.5	5.0	10.0
5	{2,4}	2/16	3.0	6.0	12.0
9	{3,3}	1/16	3.0	6.0	12.0
6	{3,4}	2/16	3.5	7.0	14.0
10	{4,4}	1/16	4.0	8.0	16.0

3.16.3 Some Observations

- When sampling with replacement, two estimates can be obtained that cannot be obtained when sampling is done without replacement:
 - ▷ 1.0 and 4.0 for the average
 - ▷ 4.0 and 16.0 for the total
- These happen to be the most extreme values.
- We now have four **estimators**:
 - ▷ The column of all values \bar{y} is the estimator of the mean, obtained with/without replacement.
 - ▷ The column of all values \hat{y} is the estimator of the total, obtained with/without replacement.

- When is an estimator good?
- To answer this question, we study characteristics of the estimators, i.e., the column of estimates.
- The quantities commonly used are:
 - ▷ expectation
 - ▷ variance (precision), leading to the standard error
 - ▷ bias
 - ▷ mean square error

3.17 Expectation and Bias

- **Definitions of expectation.**

- ▷ The **expectation** is the average of all possible estimates.
- ▷ The **expectation** is the average of the estimator.

- The expectation can be considered the population average of population \mathcal{S} .

- Expectation for an estimator \hat{y} :

$$E(\hat{y}) = \sum_{s=1}^S P_s \hat{y}_s$$

- This appears to be the notation for the total only, but it holds for every estimator; for the mean:

$$E(\bar{y}) = \sum_{s=1}^S P_s \bar{y}_s$$

- When all samples are equally likely to be taken, like in simple random sampling without replacement, then

$$P_s = \frac{1}{S}$$

and

$$E(\hat{y}) = \frac{1}{S} \sum_{s=1}^S \hat{y}_s$$

- **Definition of bias.**

- ▷ If $E(\hat{y}) = Y$, i.e., the expected value of the estimator is equal to the population value, then the estimator is termed **unbiased**.
- ▷ The **bias** is $Y - E(\hat{y})$.

3.17.1 Example: Artificial Population

- Expectation for the average under sampling without replacement:

$$\begin{aligned} E(\bar{y}) &= \frac{1}{S} \sum_{s=1}^S \bar{y}_s \\ &= \frac{1.5 + 2.0 + 2.5 + 2.5 + 3.0 + 3.5}{6} \\ &= 2.5 \end{aligned}$$

- Expectation for the total under sampling without replacement:

$$\begin{aligned} E(\hat{y}) &= \frac{1}{S} \sum_{s=1}^S \hat{y}_s \\ &= \frac{6.0 + 8.0 + 10.0 + 10.0 + 12.0 + 14.0}{6} \\ &= 10.0 \end{aligned}$$

- Expectation for the average under sampling with replacement:

$$\begin{aligned} E(\bar{y}) &= \sum_{s=1}^S P_s \bar{y}_s \\ &= \frac{2}{16} \cdot [1.5 + 2.0 + 2.5 + 2.5 + 3.0 + 3.5] + \frac{1}{16} \cdot [1.0 + 2.0 + 3.0 + 4.0] \\ &= \frac{40}{16} = 2.5 \end{aligned}$$

- Expectation for the total under sampling with replacement:

$$\begin{aligned} E(\hat{y}) &= \sum_{s=1}^S P_s \hat{y}_s \\ &= \frac{2}{16} \cdot [6.0 + 8.0 + 10.0 + 10.0 + 12.0 + 14.0] \\ &\quad + \frac{1}{16} \cdot [4.0 + 8.0 + 12.0 + 16.0] = \frac{40}{16} = 10.0 \end{aligned}$$

- Summary:

Quantity	Population value	Expectation of estimator	
		Without	With
Average (mean)	2.5	2.5	2.5
Total (sum)	10.0	10.0	10.0

- The estimators are unbiased, regardless of whether applied with or without replacement.
- The same computations for $n = 1, 3, 4$ will equally well produce unbiased estimators.
- Nevertheless, we feel there is a difference between both: this is where **variance** comes in.

3.18 Variability, Precision, Variance, Standard Error, and Standard Deviation

- Some definitions:

Variability: (informal term) fluctuation in a quantity.

Precision: (informal term) absence of fluctuation in a quantity.

- The above terminology is too informal to be useful; they combine aspects of bias and precision.

- Therefore, we prefer variance and its derived quantities:

Variance: Averaged squared deviation of a random variable around its mean.

Standard deviation: The square root of the variance.

Standard error: In the specific case of an estimator, the standard deviation is termed standard error.

- Thus:
 - ▷ The standard deviation is about population \mathcal{P}
 - ▷ The standard error is about meta-population \mathcal{S}
 - ▷ While \mathcal{P} is given, we can influence \mathcal{S} by selecting a sampling mechanism, a sample size, and opting for either with replacement or without replacement.

- The variance of a sample estimator has general form:

$$\begin{aligned}\sigma_{\hat{y}}^2 &= E(\hat{y} - E\hat{y})^2 \\ &= \sum_{s=1}^S P_s \left(\hat{y}_s - \sum_{s=1}^S P_s \hat{y}_s \right)^2\end{aligned}$$

- When every sample has the same selection probability:

$$\sigma_{\hat{y}}^2 = \frac{1}{S} \sum_{s=1}^S \left(\hat{y}_s - \frac{1}{S} \sum_{s=1}^S \hat{y}_s \right)^2$$

3.18.1 Example: Artificial Population

- Variance of the average under sampling without replacement:

$$\begin{aligned}\sigma_{\bar{y}}^2 &= \frac{1}{S} \sum_{s=1}^S \left(\bar{y}_s - \frac{1}{S} \sum_{s=1}^S \bar{y}_s \right)^2 \\ &= \frac{(1.5-2.5)^2 + (2.0-2.5)^2 + (2.5-2.5)^2 + (2.5-2.5)^2 + (3.0-2.5)^2 + (3.5-2.5)^2}{6} \\ &= \frac{2.5}{6} = 0.4167\end{aligned}$$

- Variance of the total under sampling without replacement:

$$\begin{aligned}\sigma_{\hat{y}}^2 &= \frac{1}{S} \sum_{s=1}^S \left(\hat{y}_s - \frac{1}{S} \sum_{s=1}^S \hat{y}_s \right)^2 \\ &= \frac{(6.0-10)^2 + (8.0-10.0)^2 + (10.0-10.0)^2 + (10.0-10.0)^2 + (12.0-10.0)^2 + (14.0-10.0)^2}{6} \\ &= \frac{40.0}{6} = 6.6667\end{aligned}$$

- Note that the expectation of the population total is 4 times the expectation of the population average,
while the variance of the population total is 16 times the variance of the population average.
- Variance of the average under sampling with replacement:

$$\begin{aligned}
 \sigma_{\bar{y}}^2 &= \sum_{s=1}^S P_s \left(\bar{y}_s - \sum_{s=1}^S P_s \bar{y}_s \right)^2 \\
 &= \frac{2}{16} \cdot [(1.5 - 2.5)^2 + (2.0 - 2.5)^2 + (2.5 - 2.5)^2 + (2.5 - 2.5)^2 + (3.0 - 2.5)^2 + (3.5 - 2.5)^2] \\
 &\quad + \frac{1}{16} \cdot [(1.0 - 2.5)^2 + (2.0 - 2.5)^2 + (3.0 - 2.5)^2 + (4.0 - 2.5)^2] \\
 &= \frac{10.0}{16} = 0.6250
 \end{aligned}$$

- Variance of the total under sampling with replacement:

$$\begin{aligned}\sigma_{\hat{y}}^2 &= \sum_{s=1}^S P_s \left(\hat{y}_s - \sum_{s=1}^S P_s \hat{y}_s \right)^2 \\&= \frac{2}{16} \cdot [(6.0 - 10.0)^2 + (8.0 - 10.0)^2 + (10.0 - 10.0)^2 + (10.0 - 10.0)^2 + (12.0 - 10.0)^2 + (14.0 - 10.0)^2] \\&\quad + \frac{1}{16} \cdot [(4.0 - 10.0)^2 + (8.0 - 10)^2 + (12.0 - 10.0)^2 + (16.0 - 10.0)^2] \\&= \frac{160.0}{16} = 10.0\end{aligned}$$

- Summary:

Expectation			
Quantity	Population value	Expectation of estimator	
		Without	With
Average (mean)	2.5	2.5	2.5
Total (sum)	10.0	10.0	10.0
Variances			
Quantity	Population value	Variance of estimator	
		Without	With
Average (mean)	1.25	0.4167	0.6250
Total (sum)	—	6.6667	10.0000

- The variance at the population level is **not** comparable to the variance of the estimators, except for $n = 1$.
- The variance of the estimator without replacement is smaller than the variance of the estimator with replacement.

3.18.2 Some Concerns

1. The **enumeration** we have conducted is feasible only in small samples only: we would need a computationally more parsimonious method in large populations and/or large samples.

This problem will be tackled now.

2. The calculations seem to need knowledge of the entire population.
In practice, we dispose of a single sample only.

This problem will be tackled in the following part.

3.19 Algebraic Computation Rather Than Tedious Enumeration

- The explicit enumeration to calculate these expectations is only possible for very small populations for which the entire population is known:
 - ▷ **Examples** where it is possible:
 - * the Artificial Population
 - * Surveytown
 - ▷ **Counterexample** where it is not possible:
 - * Belgian Health Interview Survey
- When it is possible, there is actually no point in sampling any longer.
- However, we can derive the expectation through algebraic manipulations, using the expectation (E) operator.

- Let us illustrate this for a total:

$$E(\hat{y}) = E\left(\frac{N}{n} \sum_{i=1}^n y_i\right) = \frac{N}{n} \sum_{i=1}^n E y_i.$$

- We have reduced the operation to the expectation of a single unit.
- Let us assume every unit has the same probability of being selected:

$$E(y_i) = \frac{1}{N} \sum_{I=1}^N Y_I = \frac{1}{N} Y = \bar{Y}$$

- Hence, we obtain

$$E(\hat{y}) = \frac{N}{n} \sum_{i=1}^n \frac{1}{N} Y = Y$$

- Conclusion: every sample, taken such that every unit has the same probability of being selected, is unbiased, regardless of the population and sample sizes, and whether a sample is taken with or without replacement.

- Now assume the unit selection probabilities are unequal:

- ▷ P_I for unit I in the population

- ▷ p_i for unit i in the sample

- Unbiased estimators are then given by

$$\bar{y} = \sum_{i=1}^n \frac{y_i}{np_i}$$

$$\hat{y} = \frac{1}{N} \sum_{i=1}^n \frac{y_i}{np_i}$$

3.20 When Is an Unbiased Estimator Unbiased?

There are a number of (non-quantitative) conditions:

- The existence of every unit in the population is known.
In survey terms, it means that population=sample frame.
This is never true in practice.
- A truly random sample has been taken.
- All variables we need to know (size of household, income,...) can be collected.
- The values that need to be collected, are collected.
- The sample estimates have been obtained by means of correct calculations.

- No other errors occurred.
- The sample values (the values recorded) are equal to their population values.
- Information is obtained in the same fashion for all individuals.

All deviations have an impact on bias (and possible on the variance).

3.21 Bias, Variance, and Mean Squared Error

- A triangular relationship:
 - ▷ **Bias**: the discrepancy between expectation and the true population value.
 - ▷ **Variance (standard error)**: the discrepancy between a sample realization and the expectation.
 - ▷ **What an investigator wants to know**: the discrepancy between a sample realization and the true population value.

- We can place them in a single, triangular relationship:

$$\begin{aligned}E(\hat{y} - Y)^2 &= E(\hat{y} - E\hat{y} + E\hat{y} - Y)^2 \\&= E(\hat{y} - E\hat{y})^2 + (E\hat{y} - Y)^2 \\ \text{MSE}(\hat{y}) &= \sigma_{\hat{y}}^2 + [\text{bias}(\hat{y})]^2\end{aligned}$$

- What an investigator wants to know = **MSE** = **mean square error**.
- Variance receives more attention than bias, since it is easier to study algebraically.
- Practically, when conducting a survey, we have to split resources over:
 - ▷ selecting a sample which is large enough (to reduce variance and hence standard error)
 - ▷ the reduction and avoidance of bias
- But, reducing the standard error is routine (sample sizes formulae abound), while reduction of bias requires insight and the consideration of a lot of aspects, usually outside the control and/or knowledge of the investigator.

3.22 Example: Surveytown

- Let us recall a few facts about Surveytown.
- Enumeration:

I	X_I	Y_I
1	1	1
2	3	2
3	4	3
4	6	4
5	7	5
6	8	6
7	10	7
8	11	8

- Population totals:

$$X = 50$$

$$Y = 36$$

- Population averages:

$$\bar{X} = 6.25$$

$$\bar{Y} = 4.50$$

- Population variances:

$$\sigma_X^2 = \frac{1}{8} \sum_{I=1}^8 (X_I - 6.25)^2 = 10.4375$$

$$\sigma_Y^2 = \frac{1}{8} \sum_{I=1}^8 (Y_I - 4.50)^2 = 5.25$$

- Samples (without replacement) of size $n = 1$:

s	Sample	P_s	y_s	\hat{y}_s	$(\hat{y}_s - E\hat{y}_s)^2$
1	{1}	1/8	1	8	$(8 - 36)^2$
2	{2}	1/8	2	16	$(16 - 36)^2$
3	{3}	1/8	3	24	$(24 - 36)^2$
4	{4}	1/8	4	32	$(32 - 36)^2$
5	{5}	1/8	5	40	$(40 - 36)^2$
6	{6}	1/8	6	48	$(48 - 36)^2$
7	{7}	1/8	7	56	$(56 - 36)^2$
8	{8}	1/8	8	64	$(64 - 36)^2$
Expectation				36	
Variance					336 (s.e. 18.33)

- Samples (without replacement) of size $n = 2$ (Part A):

s	Sample	P_s	y_s	\hat{y}_s	$(\hat{y}_s - E\hat{y}_s)^2$
1	{1,2}	1/28	3	12	$(12 - 36)^2$
2	{1,3}	1/28	4	16	\vdots
3	{1,4}	1/28	5	20	
4	{1,5}	1/28	6	24	
5	{1,6}	1/28	7	28	
6	{1,7}	1/28	8	32	
7	{1,8}	1/28	9	36	
8	{2,3}	1/28	5	20	
9	{2,4}	1/28	6	24	
10	{2,5}	1/28	7	28	
11	{2,6}	1/28	8	32	
12	{2,7}	1/28	9	36	
13	{2,8}	1/28	10	40	
14	{3,4}	1/28	7	28	
15	{3,5}	1/28	8	32	

- Samples (without replacement) of size $n = 2$ (Part B):

s	Sample	P_s	y_s	\hat{y}_s	$(\hat{y}_s - E\hat{y}_s)^2$
16	{3,6}	1/28	9	36	
17	{3,7}	1/28	10	40	
18	{3,8}	1/28	11	44	
19	{4,5}	1/28	9	36	
20	{4,6}	1/28	10	40	
21	{4,7}	1/28	11	44	
22	{4,8}	1/28	12	48	
23	{5,6}	1/28	11	44	
24	{5,7}	1/28	12	48	
25	{5,8}	1/28	13	52	
26	{6,7}	1/28	13	52	
27	{6,8}	1/28	14	56	\vdots
28	{7,8}	1/28	15	60	$(60 - 36)^2$
Expectation				36	
Variance					144 (s.e. 12.00)

- Consider the biased situation where unit $I = 8$ has been omitted.
- Biased samples (without replacement) of size $n = 2$ (Part A):

s	Sample	P_s	y_s	\hat{y}_s	$(\hat{y}_s - E\hat{y}_s)^2$	$(\hat{y}_s - Y)^2$
1	{1,2}	1/21	3	10.5	$(10.5 - 28)^2$	$(10.5 - 36)^2$
2	{1,3}	1/21	4	14.0	\vdots	\vdots
3	{1,4}	1/21	5	17.5		
4	{1,5}	1/21	6	21.0		
5	{1,6}	1/21	7	24.5		
6	{1,7}	1/21	8	28.0		
8	{2,3}	1/21	5	17.5		
9	{2,4}	1/21	6	21.0		
10	{2,5}	1/21	7	24.5		
11	{2,6}	1/21	8	28.0		
12	{2,7}	1/21	9	31.5		
14	{3,4}	1/21	7	24.5		

- Biased samples (without replacement) of size $n = 2$ (Part B):

s	Sample	P_s	y_s	\hat{y}_s	$(\hat{y}_s - E\hat{y}_s)^2$	$(\hat{y}_s - Y)^2$
15	{3,5}	1/21	8	28.0		
16	{3,6}	1/21	9	31.5		
17	{3,7}	1/21	10	35.0		
19	{4,5}	1/21	9	31.5		
20	{4,6}	1/21	10	35.0		
21	{4,7}	1/21	11	38.5		
23	{5,6}	1/21	11	38.5		
24	{5,7}	1/21	12	42.0	\vdots	\vdots
26	{6,7}	1/21	13	45.5	$(45.5 - 28)^2$	$(45.5 - 36)^2$
Expectation				28		
Variance					81.6667	
Bias ²					$+ (28 - 36)^2$	
MSE						$= 145.6667$
					s.e. 9.04	RMSE 12.07

Part II

Simple Random Sampling

Chapter 4

General Concepts and Design

- ▷ Principle of Simple Random Sampling
- ▷ Examples

4.1 Simple Random Sampling

- The most basic form of sampling
- Used as background, to compare other method with
- Recall the two classical model: drawing balls from an urn:
 - ▷ one after the other
 - ▷ independently from one another
 - ▷ choice between with/without replacement
- General principles already discussed in Chapter 3

4.1.1 Quantities

- We need the following information:
 - ▷ Population \mathcal{P}
 - ▷ Population size N
 - ▷ Sample size n
 - ▷ Whether sampling is done with or without replacement
- Recall that N and n produce the **sample fraction**:

$$f = \frac{n}{N}$$

4.1.2 Number of Samples

Data	N	n	S	
			Without	With
General	N	n	$\binom{N}{n}$	N^n
Artificial Population	4	2	6	16
Surveytown	8	2	28	64
Health Interview Survey	10,000,000	10,000	$10^{34,338}$	$10^{70,000}$

- Note that, for sampling with replacement, we have counted permutations separately, like in

$$\mathcal{S}_2 = \{ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \\ \{2, 1\}, \{3, 1\}, \{4, 1\}, \{3, 2\}, \{4, 2\}, \{4, 3\}, \\ \{1, 1\}, \{2, 2\}, \{3, 3\}, \{4, 4\} \}$$

- In case we want a formula for unordered pairs only, like in

$$\mathcal{S}_2 = \{ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \\ \{1, 1\}, \{2, 2\}, \{3, 3\}, \{4, 4\} \}$$

the formula becomes:

$$S = \binom{N - 1 + n}{n} = \frac{(N - 1 + n)!}{n! (N - 1)!}$$

- For the Artificial Population, and $n = 2$:

$$S = \binom{4 - 1 + 2}{2} = \frac{5!}{2! 3!} = 10$$

Chapter 5

Analysis

- ▷ With and without replacement
- ▷ Variance: enumeration, algebraic calculation, and estimation
- ▷ Subgroups
- ▷ Totals within subgroups

5.1 With and Without Replacement

- For the artificial population, we produced the following summary in Chapter 3:

Expectation			
Quantity	Population value	Expectation of estimator	
		Without	With
Average (mean)	2.5	2.5	2.5
Total (sum)	10.0	10.0	10.0
Variances			
Quantity	Population value	Variance of estimator	
		Without	With
Average (mean)	1.25	0.4167	0.6250
Total (sum)	—	6.6667	10.0000

- We derived that, while the expectation is equal to its population value for both sampling with and without replacement, this is not true for the variances:
 - ▷ The variance is smaller without replacement than with replacement.
We will show this is always true.
 - ▷ The sampling variances are different from the population variance.
Notwithstanding this, they are connected.
 - ▷ The variance resulted from (tedious) enumeration.
Algebraic calculations are possible.

5.1.1 General Variance Formulae

- Estimators:

- ▷ For the **average**:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- ▷ For the **total**:

$$\hat{y} = \frac{N}{n} \sum_{i=1}^n y_i$$

- Variances:

For the Average

Quantity	General	Artificial Population
Population variance	$\sigma_Y^2 = \frac{1}{N} \sum_{I=1}^N (Y_I - \bar{Y})^2$	1.2500
With replacement	$\sigma_{\bar{y}}^2 = \frac{1}{n} \sigma_Y^2$	$\frac{1}{2} \cdot 1.2500 = 0.6250$
Population variance	$S_Y^2 = \frac{1}{N-1} \sum_{I=1}^N (Y_I - \bar{Y})^2$	1.6667
Without replacement	$\sigma_{\bar{y}}^2 = \frac{1}{n} (1-f) S_Y^2$	$\frac{1}{2} \cdot \frac{1}{2} \cdot 1.6667 = 0.4167$

For the Total

Quantity	General	Artificial Population
Population variance	$\sigma_Y^2 = \frac{1}{N} \sum_{I=1}^N (Y_I - \bar{Y})^2$	1.2500
With replacement	$\sigma_{\hat{y}}^2 = \frac{N^2}{n} \sigma_Y^2$	$\frac{16}{2} \cdot 1.2500 = 10.0000$
Population variance	$S_Y^2 = \frac{1}{N-1} \sum_{I=1}^N (Y_I - \bar{Y})^2$	1.6667
Without replacement	$\sigma_{\hat{y}}^2 = \frac{N^2}{n} (1-f) S_Y^2$	$\frac{16}{2} \cdot \frac{1}{2} \cdot 1.6667 = 6.6667$

5.1.2 Considerations

- For sampling with replacement, also S_Y^2 can be used.
- The difference between σ_Y^2 and S_Y^2 is irrelevant for moderate to large populations.
- The essential difference between both situations is $1 - f$.
 - ▷ If $f = 1$, then sampling with replacement is equal to the census, and there is no residual uncertainty (provided measurements y_i are equal to their true values Y_I , i.e., there is no measurement error).
 - ▷ $f \simeq 0$ if
 - * N is large or infinite
 - * $n \ll N$: sample size much smaller than population size
- Note that, if $N = \infty$, estimating the total has no meaning.

5.1.3 Example: Surveytown

- Previously, the computations for Surveytown have been carried out, for samples of size $n = 1$ and $n = 2$, by enumeration.
- They can now easily be repeated by computation, using the above formulas:
- Population variance: $S_Y^2 = 6$
 - ▷ Samples of size $n = 1$:
$$\sigma_{\hat{y}}^2 = \frac{8^2}{1} \times \left(1 - \frac{1}{8}\right) \times 6 = \frac{64 \times 7 \times 6}{8} = 336$$
 - ▷ Samples of size $n = 2$:
$$\sigma_{\hat{y}}^2 = \frac{8^2}{2} \times \left(1 - \frac{2}{8}\right) \times 6 = \frac{64 \times 6 \times 6}{2 \times 8} = 144$$
- Let us give the frequencies of the estimators for the number of buildings within Surveytown, based on samples of size 1–8.

Measure	Sample size n							
	1	2	3	4	5	6	7	8
Mean	36	36	36	36	36	36	36	36
Range								
Minimum	8	12	16	20	24	28	32	36
Maximum	64	60	56	52	48	44	40	36
Variance	336	144	80	48	28.8	16	6.9	0
Standard error	18.3	12	8.9	6.9	5.4	4	2.6	0
Number	8	28	56	70	56	28	8	1

• Observations:

- ▷ All estimators are unbiased.
- ▷ The extremes and the variance reduce with increasing sample size.
- ▷ The variances, calculated from the variance formulae, are in agreement with those based on enumeration, as should be the case.
- ▷ The last column represents the census.

5.1.4 Graphical Representation of Some of the Estimators

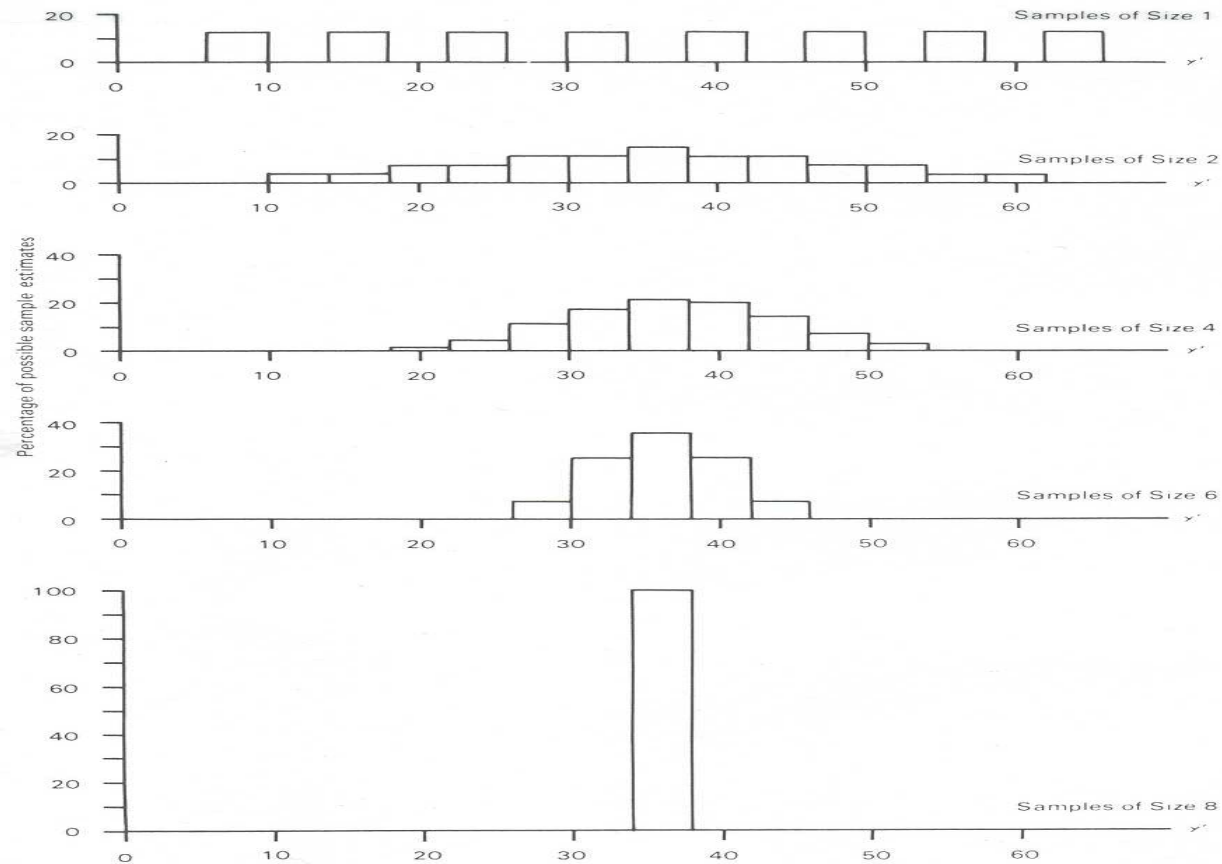


Figure 3.1 Frequency distribution of unbiased estimates of the number of dwellings in Surveytown based on all possible simple random samples of size 1, 2, 4, 6, and 8 blocks, selected without replacement.

5.2 Subgroups

- We have focused on averages and totals of (continuous) quantities.
- Let us shift focus to a proportion (fraction, subgroup).
- Indeed, a subgroup is defined by a variable Z_I taken values

$$Z_I = \begin{cases} 1 & \text{if unit } I \text{ belongs to the subgroup,} \\ 0 & \text{if unit } I \text{ does not belong to the subgroup} \end{cases}$$

- The proportion of units belonging to the subgroup, at population level, is denoted by P or π .
- Often, also the notation $Q = 1 - P$ is used.
- The population proportion is defined as

$$P = \frac{1}{N} \sum_{I=1}^N Z_I$$

and estimated from the sample as

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n z_i$$

- The population variance is given by:

$$\sigma_Z^2 = \frac{N}{N-1} PQ \simeq PQ$$

- The variance for the estimated proportion, without replacement:

$$\sigma_{\hat{p}}^2 = \frac{1}{n} (1-f) \frac{N}{N-1} PQ$$

- For (infinitely) large samples and/or with replacement, we have that:

$$N/(N-1) \simeq 1$$

$$1-f \simeq 1$$

and hence

$$\sigma_{\hat{p}}^2 \simeq \frac{1}{n} PQ$$

5.2.1 Example: Surveytown

- Let us consider the proportion of Surveytown blocks with two or more vacant lots.
- Consider samples of sizes $n = 1, \dots, 8$

Measure	Sample size n							
	1	2	3	4	5	6	7	8
Mean	0.625	0.625	0.625	0.625	0.625	0.625	0.625	0.625
Range								
Minimum	0	0	0.250	0.400	0.500	0.571	0.625	
Maximum	1.000	1.000	1.000	1.000	1.000	0.833	0.714	0.625
Variance	0.234	0.100	0.056	0.034	0.020	0.011	0.005	0
Standard error	0.484	0.317	0.236	0.183	0.142	0.106	0.069	0
Number	8	28	56	70	56	28	8	1

5.2.2 Estimating the Size of a Subgroup

- Consider a population \mathcal{P} of size N .
- Assume that a proportion P belongs to a subgroup (subpopulation, e.g., a region).
- The size of the subgroup is then:

$$N_g = N \cdot P$$

- It can be estimated from a sample of size n by

$$\hat{n}_g = N \cdot \hat{p}$$

with variance

$$\begin{aligned}\sigma_{\hat{n}_g} &= \text{var}(\hat{n}_g) \\ &= \text{var}(N\hat{p}) \\ &= N^2 \text{var}(\hat{p}) \\ &= N^2 \cdot \frac{1}{n} \cdot (1 - f) \cdot \frac{N}{N - 1} \cdot PQ\end{aligned}$$

- The large sample approximation / version for sampling with replacement:

$$\sigma_{\hat{n}_g} = N^2 \cdot \frac{1}{n} PQ$$

5.2.3 Estimating a Quantity for a Subgroup

- Often, we want to estimate quantities (average, sum) for a subpopulation:
 - ▷ The average income of all inhabitants of Flanders
 - ▷ The total income of all inhabitants of Wallonia
- If we would know N_g , then the problem would not differ for the population problem already considered.
- However, we usually have to estimate N_g as well, e.g., by means of \hat{n}_g , studied above.
- The population estimand is

$$Y_g = \sum_{I=1}^{N_g} Y_{gI}$$

- Assume we dispose of a sample of size n :
 - ▷ of which n_g units belong to the subgroup
 - ▷ and for each of which y_{gi} has been recorded
- Then we can construct the estimator:

$$\begin{aligned}
 \widehat{y}_g &= \frac{N}{n} \sum_{i=1}^{n_g} y_{gi} \\
 &= \frac{N}{n} n_g \left(\frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi} \right) \\
 &= N \frac{n_g}{n} \left(\frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi} \right) \\
 &= [N\widehat{p}] \cdot \left(\frac{1}{n_g} \sum_{i=1}^{n_g} y_{gi} \right) \\
 &= \widehat{n}_g \overline{y}_g
 \end{aligned}$$

5.2.4 Variance of a Quantity, Estimated for a Subgroup

- We need the variance of the above product:

$$\begin{aligned}\sigma_{\hat{y}_g}^2 &= \text{var}(\hat{y}_g) \simeq \text{var}(\hat{n}_g \bar{y}_g) \\ &= \bar{Y}_g^2 \sigma_{\hat{n}_g}^2 + N_g^2 \sigma_{\bar{Y}_g}^2\end{aligned}$$

- This formula is different from the one for the size of a subgroup, since we now have two sources of uncertainty:
 - ▷ we do not know the size of the subpopulation
 - ▷ we do not know the value of the average within the subgroup
- The above formula is an approximation, based on the so-called delta method.

5.2.5 Delta Method

- Assume X and Y are random variables
- Variance of the sum:

$$\text{var}(X + Y) = \text{var}(X) + 2\text{cov}(X, Y) + \text{var}(Y)$$

- Variance of the sum under independence:

$$\text{var}(X + Y) = \text{var}(X) + \text{var}(Y)$$

- Variance of the difference:

$$\text{var}(X - Y) = \text{var}(X) - 2\text{cov}(X, Y) + \text{var}(Y)$$

- Variance of the difference under independence:

$$\text{var}(X - Y) = \text{var}(X) + \text{var}(Y)$$

- Note that, under independence, sum and difference have the same variance.
- Variance of the product:

$$\text{var}(X \cdot Y) \simeq Y^2 \text{var}(X) + XY \text{cov}(X, Y) + X^2 \text{var}(Y)$$

or, equivalently

$$\frac{\text{var}(X \cdot Y)}{X^2 Y^2} \simeq \frac{Y^2 \text{var}(X)}{X^2 Y^2} + \frac{XY \text{cov}(X, Y)}{X^2 Y^2} + \frac{X^2 \text{var}(Y)}{X^2 Y^2}$$

$$\frac{\text{var}(X \cdot Y)}{X^2 Y^2} \simeq \frac{\text{var}(X)}{X^2} + \frac{\text{cov}(X, Y)}{XY} + \frac{\text{var}(Y)}{Y^2}$$

$$\text{Rvar}(X \cdot Y) \simeq \text{Rvar}(X) + \text{Rcov}(X, Y) + \text{Rvar}(Y)$$

with

$$\text{Rvar}(X) = \frac{\text{var}(X)}{X^2}$$

$$\text{Rcov}(X, Y) = \frac{\text{cov}(X, Y)}{XY}$$

- Variance of the product under independence:

$$\frac{\text{var}(X \cdot Y)}{X^2 Y^2} \simeq \frac{Y^2 \text{var}(X)}{X^2 Y^2} + \frac{X^2 \text{var}(Y)}{X^2 Y^2}$$

$$\text{Rvar}(X \cdot Y) \simeq \text{Rvar}(X) + \text{Rvar}(Y)$$

- Variance for a general function $Z = f(X, Y)$ of two random variables:

$$\text{var}(Z) \simeq \left(\frac{\partial f(X, Y)}{\partial X}, \frac{\partial f(X, Y)}{\partial Y} \right) \begin{pmatrix} \text{var}(X) & \text{cov}(X, Y) \\ \text{cov}(X, Y) & \text{var}(Y) \end{pmatrix} \begin{pmatrix} \frac{\partial f(X, Y)}{\partial X} \\ \frac{\partial f(X, Y)}{\partial Y} \end{pmatrix}$$

- This method is known as the delta method.

5.3 Estimating a Variance

- We have constructed variance expressions in two ways:
 - ▷ Enumeration
 - ▷ Algebraic computation
- The first one is tedious, since it requires constructing **all** samples.
- While the second is more convenient, more general, and one can derive general insight, **it cannot be used in practice neither**, since it requires knowledge of the population variance, for which **all population units** need to be known.

In practice, a variance can neither be **enumerated** nor **calculated**, but it can be **estimated**.

- In the expression for the variances, the population quantities are replaced by estimates, based on the sample:

Quantity	Calculated	Estimated
Population variance	$S_Y^2 = \frac{1}{N-1} \sum_{I=1}^N (Y_I - \bar{Y})^2$	$\hat{s}_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$
Total	$\sigma_{\bar{y}}^2 = \frac{N^2}{n} (1-f) S_Y^2$	$\hat{\sigma}_{\bar{y}}^2 = \frac{N^2}{n} (1-f) \hat{s}_y^2$
Average	$\sigma_{\bar{y}}^2 = \frac{1}{n} (1-f) S_Y^2$	$\hat{\sigma}_{\bar{y}}^2 = \frac{1}{n} (1-f) \hat{s}_y^2$

5.3.1 Example: Artificial Population

- Consider samples of size $n = 2$, without replacement
- Calculated versus estimated variance:

s	Sample	σ_y^2	s_y^2	$\hat{\sigma}_y^2$
1	{1,2}	6.6667	0.5000	2.0000
2	{1,3}	6.6667	2.0000	8.0000
3	{1,4}	6.6667	4.5000	18.0000
4	{2,3}	6.6667	0.5000	2.0000
5	{2,4}	6.6667	2.0000	8.0000
6	{3,4}	6.6667	0.5000	2.0000
Mean			1.6667	6.6667
			$= S_Y^2$	$= \sigma_y^2$

- The estimated variance constitutes itself a random variable, and apparently is unbiased (which can be proven).

5.4 Covariance

- It is equally possible to construct estimators for covariance and correlation.
- For the covariance, the calculated

$$S_{XY} = \frac{1}{N-1} \sum_{I=1}^N (X_I - \bar{X})(Y_I - \bar{Y})$$

is estimated by:

$$\hat{s}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Other quantities, such as correlations, allow for similar manipulations.

Chapter 6

Sample Size Determination

- ▷ Example with continuous outcomes
- ▷ Example with binary outcomes
- ▷ General expressions

6.1 Example of Sample Size Determination With Continuous Outcome

- Suppose we wish to know the number of failings happening withing a group of $N = 1000$ small retail stores.
- Regarding precision, it is often easier to make relative statements:
 - ▷ We know or assume that the relative population standard deviation, is 1, i.e.,

$$\text{Rvar}(Y) = 1.0^2$$

- ▷ A relative standard error of $10\%=0.1$ is requested, i.e.,

$$\text{Rvar}(\hat{y}) = 0.10^2.$$

This means that we want to estimate the population quantity to within 10% of its value.

- The relative quantities are in the same relationship than the absolute ones:

$$\text{Rvar}(\hat{y}) = \frac{1}{n} \frac{N - n}{N} \text{Rvar}(Y),$$

$$0.10^2 = \frac{1}{n} \left(\frac{1000 - n}{1000} \right) 1.0^2$$

- Solving for n produces the required sample size:

$$n = 91$$

- If we omit the finite population correction and/or consider sampling with replacement:

$$\text{Rvar}(\hat{y}) = \frac{1}{n} \text{Rvar}(Y),$$

$$0.10^2 = \frac{1}{n} 1.0^2$$

- Solving for n produces the required sample size:

$$n = 100$$

- We therefore see, once more, that sampling with replacement is less precise than sampling without replacement. It shows here through the need for a larger sample size.

6.2 Example of Sample Size Determination for a Proportion

- Suppose we wish to know what proportion of shops sells toys.
- A standard error of 5% is requested.
- We assume the proportion which sells toys is about $P = 60\% = 0.6$.
- Using the variance formula for a binary variable Z :

$$\sigma_Z^2 = \frac{N}{N-1}PQ = \frac{1000}{999}(0.6 \times 0.4) = 0.24$$

and including this in the expression for the variance of the estimated proportion:

$$\sigma_{\hat{p}}^2 = \frac{1}{n} \frac{N-n}{N} \sigma_Z^2$$
$$(0.05)^2 = \frac{1}{n} \frac{1000-n}{1000} 0.24$$

- Solving for n produces the required sample size:

$$n = 88$$

- If we omit the finite population correction and/or consider sampling with replacement:

$$\sigma_Z^2 = PQ = 0.6 \times 0.4 = 0.24$$

$$\sigma_{\hat{p}}^2 = \frac{1}{n} \sigma_Z^2$$

$$(0.05)^2 = \frac{1}{n} 0.24$$

- Solving for n produces the required sample size:

$$n = 96$$

6.3 Where Does the Information Come From?

- The information we need in both examples can be divided into two groups:
 - ▷ **Related to what we want to achieve:** the (relative) standard error [or (relative) variance] for the estimator.

This is a completely natural request of information.

- ▷ **Related to the population quantities:**
 - * The (relative) standard deviation [or (relative) variance] of the population quantity in the continuous case.
 - * The proportion itself in the case of a proportion.
 - * Note that, for the proportion, we actually only need the variance too, but for binary data the proportion P produces the variance: $P(1 - P) = PQ$.
- The problem is that the second group of quantities constitutes **circularity**: we need information about what we want to estimate, prior to estimation.

- Therefore, the information has to come from other sources:

Historical information. This refers to studies already conducted about the same or similar variables, in the same or similar populations.

Expert opinion. Watch out with expert opinion!

Pilot study. A small study, conducted to obtain a (rough) idea about the precision of the population quantity, or the proportion we want to estimate. The pilot study can sometimes be integrated into the actual survey that is subsequently set up.

- For all of these reasons, a sample size calculation should be seen as a **rough indication** only of the required sample.
- The most important considerations for choosing a sample size are:
 - ▷ A sample size calculation.
 - ▷ The budget available.
 - ▷ Constraints on the organization of the fieldwork (e.g., number of interviewers available).

6.4 Sample Size Determination: General Expressions

- The above examples may have generated the impression that we have to do algebraic manipulation every time we perform a sample size calculation.
- This is not necessary: general expressions can be derived once and for all.
- We will study in turn:
 - ▷ Total and average
 - ▷ Proportion

6.4.1 Sample Sizes for Total and Average

- Re-consider the case of the total:

$$\text{Rvar}(\hat{y}) = \frac{1}{n} \cdot \frac{N - n}{N} \cdot \text{Rvar}(Y)$$

$$N \cdot n \cdot \text{Rvar}(\hat{y}) = N \cdot \text{Rvar}(Y) - n \cdot \text{Rvar}(Y)$$

$$n[N \cdot \text{Rvar}(\hat{y}) + \text{Rvar}(Y)] = N \cdot \text{Rvar}(Y)$$

$$n = \frac{N \cdot \text{Rvar}(Y)}{\text{Rvar}(Y) + N \cdot \text{Rvar}(\hat{y})}$$

- Furthermore, we can consider an expression like this for the variance, rather than the relative variance:

- Use the facts that:

$$\text{Rvar}(Y) = \frac{\sigma_Y^2}{Y^2}$$

$$\text{Rvar}(\hat{y}) = \frac{\sigma_{\hat{y}}^2}{(NY)^2}$$

- Plugging this in and simplifying, produces:

$$n = \frac{N^2 \cdot \sigma_Y^2}{\sigma_{\hat{y}}^2 + N \cdot \sigma_Y^2}$$

- The same is possible for an average.

- We obtain the following summary:

Situation	Total (\hat{y})	Average (\bar{y})
Without replacement	$n = \frac{N^2 \sigma_Y^2}{\sigma_{\hat{y}}^2 + N \sigma_Y^2}$	$n = \frac{\sigma_Y^2}{\sigma_{\bar{y}}^2 + (1/N) \sigma_Y^2}$
With replacement	$n = \frac{N^2 \sigma_Y^2}{\sigma_{\hat{y}}^2}$	$n = \frac{\sigma_Y^2}{\sigma_{\bar{y}}^2}$
$N \rightarrow +\infty$	—	$n = \frac{\sigma_Y^2}{\sigma_{\bar{y}}^2}$

6.4.2 Sample Sizes for a Proportion

- Using the expressions for $\sigma_{\hat{p}}^2$ and σ_Z^2 , we obtain:

$$\sigma_{\hat{p}}^2 = \frac{1}{n} \cdot \frac{N - n}{N - 1} \cdot PQ$$
$$n = \frac{NPQ}{\sigma_{\hat{p}}^2 \cdot (N - 1) + PQ}$$

- When $N \rightarrow +\infty$ in the above expression, we obtain:

$$n = \frac{PQ}{\sigma_{\hat{p}}^2}$$

- When we start from the original expressions for $\sigma_{\hat{p}}^2$ and σ_Z^2 , but ignoring the correction for sampling without replacement, i.e., turning to sampling with replacement, we find

$$\sigma_{\hat{p}}^2 = \frac{1}{n} \cdot PQ$$

$$n = \frac{PQ}{\sigma_{\hat{p}}^2}$$

- Just like with the average, sampling with replacement is like sampling from an infinite population.

- Let us apply the formula (with replacement for simplicity), for $\sigma_{\hat{p}}^2 = 0.05^2$ like in the example, for a range of P values:

P	Q	n
0.0	1.0	0.0
0.1	0.9	36.0
0.2	0.8	64.0
0.3	0.7	84.0
0.4	0.6	96.0
0.5	0.5	100.0
0.6	0.4	96.0
0.7	0.3	84.0
0.8	0.2	64.0
0.9	0.1	36.0
1.0	0.0	0.0

- A few observations are in place:

- ▷ The sample size is not stable over the range $[0.3; 0.7]$.

- ▷ The sample sizes are symmetric in P and Q .

But is it realistic to need the same sample size for, say, $P = 0.001$ and $P = 0.999$?

- ▷ The sample size is largest for $P = 0.5$ and then decreases. In fact, it is a quadratic function in P :

$$n = \frac{P(1 - P)}{\sigma_{\hat{p}}^2} = \frac{-P^2 + P}{\sigma_{\hat{p}}^2}$$

But wouldn't we expect a proportion of $P = 50\%$ to be the easiest, rather than the most difficult, to estimate precisely?

- The reason for the latter two, rather paradoxical results is that we consider the formula for a **constant** standard error:
 - ▷ We require a standard error of $0.05=5\%$ when $P = 50\%$
 - ▷ We require a standard error of $0.05=5\%$ when $P = 1\%$
- Of course, the latter requirement is easier, since we require a, relatively speaking, less precise result.
- Thus, the formulas derived can be seen as **absolute**: in terms of the absolute standard error.

But since the variance is a function of P , this is less meaningful.

- Alternatively, let us require a standard error proportional to P :

$$\sigma_{\hat{p}}^2 = k^2 P^2$$

where k typically ranges in $[0,1]$.

k is a proportionality constant, describing the required precision in **relative** terms.

- The formula for the sample size can now be rewritten:

$$n = \frac{NPQ}{k^2 P^2 \cdot (N - 1) + PQ}$$

$$n = \frac{NQ}{k^2 P \cdot (N - 1) + Q}$$

- The version for infinite samples and/or sampling with replacement:

$$n = \frac{1}{k^2} \cdot \frac{Q}{P}$$

- Let us again apply this formula (with replacement for simplicity), for $k^2 = 0.05^2$ like in the example, for a range of P values:

P	Q	n
0.0	1.0	$+\infty$
0.0001	0.9999	3,999,600.0
0.001	0.999	399,600.0
0.01	0.99	39,600.0
0.1	0.9	3600.0
0.2	0.8	1600.0
0.3	0.7	933.3
0.4	0.6	600.0
0.5	0.5	400.0
0.6	0.4	266.7
0.7	0.3	171.4
0.8	0.2	100.0
0.9	0.1	44.4
1.0	0.0	0.0

- The observations now become:
 - ▷ The sample size is quite stable over the range $[0.3; 0.7]$, even over $[0.2; 0.8]$.
 - ▷ The sample sizes are asymmetric in P and Q .
 - ▷ The sample size decrease with P ; the largest sample sizes are needed for the smallest P .

These are now in line with intuition.

- We obtain the following summary:

Situation	Absolute (\hat{y})	Relative (\bar{y})
Without replacement	$n = \frac{NPQ}{\sigma_{\hat{p}}^2(N-1) + PQ}$	$n = \frac{NQ}{k^2 P(N-1) + Q}$
With replacement	$n = \frac{PQ}{\sigma_{\hat{p}}^2}$	$n = \frac{1}{k^2} \cdot \frac{Q}{P}$
$N \rightarrow +\infty$	$n = \frac{PQ}{\sigma_{\hat{p}}^2}$	$n = \frac{1}{k^2} \cdot \frac{Q}{P}$

Part III

A First Perspective on Software

Chapter 7

General Considerations Regarding Software

- ▷ Taxonomy
- ▷ Implementations in SAS
- ▷ Other software packages

7.1 Design

- Some software tools are constructed for **design** purposes.
- The **input data base** is then the population or, stated more accurately, the **sample frame**.
- The **output data base** is then a sample selected from the input data base, and taking 0, 1, or more design aspects into account.
- **SAS: PROC SURVEYSELECT**

7.2 Analysis

- Not surprisingly, most software tools are geared towards **analysis**.
- Several views can be taken:

Simple estimators versus model:

Estimating a mean, total, or frequency \longleftrightarrow Regression, ANOVA

Simple cross-sectional data structure versus complex data structure:

Cross-sectional data \longleftrightarrow Multivariate, multi-level, clustered, longitudinal data

To survey or not to survey:

Non-survey data (or SRS) \longleftrightarrow one or more survey-design aspects

7.3 Analysis With SAS for a Continuous Outcome

Model	Data structure	Survey design	Method	SAS procedure
no	simple	no	mean	MEANS
yes	simple	no	linear regression ANOVA	REG ANOVA GLM
no	simple	yes	mean	SURVEYMEANS
yes	simple	yes	linear regression ANOVA	SURVEYREG
yes	complex	no	multivariate regression MANOVA	GLM
yes	complex	somehow	linear mixed model \equiv multi-level model	MIXED

- The word 'somehow' means that some design aspects can be taken into account, even though the procedure is not built for surveys.
 - ▷ In fact, most procedures have a 'weight' statement, allowing to account for sampling with unequal probability and the most important consequences of stratification.
 - ▷ Methods allowing for hierarchies (linear mixed model, multi-level model) also accommodate, to a large extent, clustering and multi-stage sampling.
 - ▷ Methods with a likelihood or Bayesian basis are attractive in the light of incomplete data (see Part X).
- The above table is not exhaustive:
 - ▷ not every analysis possibility is mentioned,
 - ▷ only the most common ones are mentioned, by way of illustration.
- We can compose a similar table for a binary outcome.

7.4 Analysis With SAS for a Binary Outcome

Model	Data structure	Survey design	Method	SAS procedure
no	simple	no	proportion frequency	FREQ
yes	simple	no	logistic regression probit regression	LOGISTIC GENMOD
no	simple	yes	proportion frequency	SURVEYFREQ
yes	simple	yes	logistic regression probit regression	SURVEYLOGISTIC
yes	complex	no	generalized estimating equations	GENMOD
yes	complex	somehow	gen. lin. mixed model non-linear mixed model	GLIMMIX NLMIXED

7.5 Other Software Packages

- Virtually all packages allow to take the survey design **somehow** into account:

weight: most packages have a weight statement \Rightarrow correction for unequal weights and aspects of stratification.

hierarchical data: an increasing number of software packages allow for the analysis of hierarchical data; these features can be usefully used to take the multi-stage and/or clustering nature into account.

Examples: MLwiN, GAUSS, R, SAS, SPlus, Stata

- Note that using these features is not without danger: weights in a non-survey context usually refer to **replication**: if there are 7 records that are exactly equal, they are represented only once with a 'repeat count' 7.
- Some packages have purposefully written survey design and/or analysis tools.

7.5.1 STATA

- STATA has a suite of functions for the analysis of survey data: the **svy*** functions:

svydes: for describing strata and PSU's

svytab: for two-way tables

svymean: for mean estimation

svyprop: for the estimation of a proportion

svyratio: for ratios

svytotal: for totals

svyreg: for linear regression

svyintrg: for censored and interval regression

svylogit: for logistic regression

svymlog: for multinomial logistic regression

svyolog: for probit regression

svyprobt: for probit regression

svyoprob: for ordered probit regression

svypoiss: for Poisson regression

svylc: for estimating linear combinations of parameters

svytest: for hypothesis tests

- Design aspects that can be taken into account:

pweight: sampling weights (in sampling with unequal probabilities)

psu: primary sampling units (in multi-stage sampling)

strata: strata (in stratification)

- A general purpose package

- <http://www.stata.com/>

7.5.2 SPSS

- SPSS has an interface, called **SPSS Complex Samples**.
- It comprises two components:
 - Sampling Plan Wizard:** To draw samples from a database (sample frame), taking the sampling plan (\equiv design) into account.
 - Analysis Preparation Wizard:** Performs statistical analysis, taking the sampling plan (\equiv design) into account.
- The following design types can be used with SPSS Complex Samples:
 - ▷ Stratified sampling
 - ▷ Clustered sampling
 - ▷ Multistage sampling

- A general purpose package
- <http://www.spss.com/>

7.5.3 SUDAAN

- One of the primary aims of SUDAAN is the analysis of survey data:

MULTILOG: Fits multinomial logistic regression models to ordinal and nominal categorical data and computes hypothesis tests for model parameters.
Has GEE (Generalized Estimating Equation) modeling capabilities for correlated (non-)Gaussian data.

REGRESS: Fits linear regression models to continuous outcomes and performs hypothesis tests concerning the model parameters.

LOGISTIC: Fits logistic regression models to binary data and computes hypothesis tests for model parameters.

SURVIVAL: Fits proportional hazards (Cox regression) models to failure time data.

CROSSTAB: Computes frequencies, percentage distributions, odds ratios, relative risks, and their standard errors (or confidence intervals) for user-specified cross-tabulations, as well as chi-square tests of independence and the Cochran-Mantel-Haenszel chi-square test for stratified two-way tables.

DESCRIPT: Computes estimates of means, totals, proportions, percentages, geometric means, quantiles. Also allows for contrasts.

RATIO: Estimates generalized ratios of the form

$$(\text{Summation } y) / (\text{Summation } x),$$

where x and y are observed variables.

- Design aspects that can be taken into account:
 - ▷ stratification (unlimited number of strata)
 - ▷ cluster sampling
 - ▷ multi-stage sampling (unlimited number of stages – this is a powerful and uncommon feature)
 - ▷ unequal selection probabilities
 - ▷ with and without replacement
- Is *not* a general-purpose package.
- Nevertheless, also deals with longitudinal data, clustered data, and incomplete data.
- <http://www.rti.org/sudaan/>

Chapter 8

SAS and The Belgian Health Interview Survey

- ▷ Variables used in this course
- ▷ Three continuous variables
- ▷ A binary variable

8.1 Key Variables Used

Body Mass Index (BMI):

- ▷ Defined as:

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height}^2 (\text{m}^2)} \quad \left[\frac{\text{kg}}{\text{m}^2} \right]$$

- ▷ A continuous measure
- ▷ Frequently analyzed on the log scale: $\ln(\text{BMI})$

General Health Questionnaire–12 (GHQ-12):

- ▷ Comprises 12 questions, yielding a 13 category outcome
- ▷ The focus is on mental health
- ▷ Can be dichotomized as well

“Vragenlijst voor Onderzoek naar de Ervaren Gezondheid” (VOEG):

- ▷ Dutch instrument, leading to a sum score
- ▷ “Questionnaire for Research Regarding Subjective Health Score”
- ▷ translated into French for Belgium
- ▷ to obtain a more symmetric score, the analysis takes place on the log scale:
 $\ln(\text{VOEG} + 1)$

Stable General Practitioner (SGP):

- ▷ “Do you have a steady general practitioner?” (GP)
- ▷ Obviously a binary indicator

8.2 The Belgian Health Interview Survey 1997 Dataset

- Dataset: `bmi_voeg.sas7bdat`
- Brief description of the variables:
 - ▷ Design variables:
 - ID:** Individual identification number
 - HH:** Household indicator
 - WFIN:** Weight, combining all sources taken into account
 - BRU:** Indicator for whether respondent lives in Brussels
 - FLA:** Indicator for whether respondent lives in Flanders
 - WAL:** Indicator for whether respondent lives in Wallonia
 - REGION:** Respondent's region (numerically coded)
 - REGIONCH:** Respondent's region (character coded)
 - PROVINCE:** Respondent's province

▷ Age and sex:

AGE7: Age; categorical variable with 7 categories

AGEGR1–AGEGR7: Binary indicators (dummies) for each of the 7 age categories

SEX: Respondent's sex

▷ Outcome variables:

BMI: body mass index

LNBM: natural logarithm of body mass index

VOEG: VOEG score

LNVOEG: natural logarithm of VOEG score

GHQ12: general health questionnaire – 12 items

GHQBIN: dichotomized version of general health questionnaire – 12 items

SGP: indicator for whether respondent has a stable general practitioner

▷ Socio-economic status:

EDU3: educational level; categorical variable with 3 categories

EDUHIGH: indicator for whether educational level is high school

EDUPRIM: indicator for whether educational level is primary education

EDUSEC: indicator for whether educational level is secondary education

FA3: income level; categorical variable with 3 categories

INCHIG: indicator for whether income category is high

INCLOW: indicator for whether income category is low

INCMED: indicator for whether income category is medium

▷ Life style variable:

TA2: indicator for whether or not a respondent smokes

8.2.1 Coding and Categories for Some of the Variables

mental:		province:		smoke:	
	0 good		1 Antwerpen		1 Non-smoker
	1 bad		2 Vlaams Brabant		2 Smoker
educ:			3 Limburg	sex:	
	1 <=Primary		4 Oost Vlaandaren		1 Male
	2 Secondary		5 West Vlaanderen		2 Female
	3 Higher		6 Brabant Wallon		
income:			7 Hainaut		
	1 <30000		8 Liege		
	2 30000-40000		9 Luxembourg		
	3 40000+		10 Namur		
agegroup:			11 Brussels		
	1 15-24		12 Eupen		
	2 25-34	region:			
	3 35-44		1 Flanders		
	4 45-54		2 Brussels		
	5 55-64		3 Wallonia		
	6 65-74				
	7 75+				

8.3 Some Tables, Created with STATA

```
. tab sex
```

Gender	Freq.	Percent	Cum.
Male	4140	48.34	48.34
Female	4424	51.66	100.00
<hr/>			
Total	8564	100.00	

```
. tab region
```

Region	Freq.	Percent	Cum.
Flanders	2987	34.88	34.88
Brussels	2571	30.02	64.90
Wallonia	3006	35.10	100.00
<hr/>			
Total	8564	100.00	


```
. tab edu3
```

Education	Freq.	Percent	Cum.
<hr/>			
<=Primary	2979	36.29	36.29
Secondary	2425	29.54	65.82
Higher	2806	34.18	100.00
<hr/>			
Total	8210	100.00	

```
. tab fa3
```

Income	Freq.	Percent	Cum.
<hr/>			
<30000	4326	53.03	53.03
30000-40000	2701	33.11	86.14
40000+	1131	13.86	100.00
<hr/>			
Total	8158	100.00	

```
. tab ta2
```

Smoking	Freq.	Percent	Cum.
<hr/>			
Non-smoker	3725	46.20	46.20
Smoker	4338	53.80	100.00
<hr/>			
Total	8063	100.00	

```
. tab age7
```

Age group	Freq.	Percent	Cum.
15-24	1150	13.43	13.43
25-34	1644	19.20	32.62
35-44	1615	18.86	51.48
45-54	1297	15.14	66.63
55-64	1095	12.79	79.41
65-74	1079	12.60	92.01
75+	684	7.99	100.00
Total	8564	100.00	

```
. tab sgp
```

Gen. pract.	Freq.	Percent	Cum.
no	823	9.65	9.65
yes	7709	90.35	100.00
Total	8532	100.00	

8.4 Simple Random Sample Analysis

- We will estimate the means of:
 - ▷ LNBMI
 - ▷ LNVOEG
 - ▷ GHQ12
 - ▷ SGP
- For the geographical entities:
 - ▷ The **country**: Belgium
 - ▷ The **regions**: Brussels, Flanders, Wallonia

- Methods used:

- ▷ Ordinary mean estimation: PROC MEANS

- ▷ Using the survey procedure SURVEYMEANS, under the assumption of SRS and further

- * Infinite population

- * Finite population of $N = 10,000,000$: this is (approximately) the true Belgian population size

- * Finite population of $N = 8564$: this is the actual sample size and thus mimicks the situation of a census

8.4.1 Ordinary Mean Estimation

- The following programs can be used:

```
proc means data=m.bmi_voeg n mean stderr;  
title 'SRS means - for Belgium';  
where (regionch^='');  
var lnbmi lnvoeg ghq12 sgp;  
run;
```

```
proc means data=m.bmi_voeg n mean stderr;  
title 'SRS means - for regions';  
where (regionch^='');  
var lnbmi lnvoeg ghq12 sgp;  
by regionch;  
run;
```

- The options have the following meaning:
 - ▷ **Keywords** n, mean, and stderr: request these statistics to be displayed; there is a variety available.
 - ▷ **WHERE** statement: specifies a condition that needs to be satisfied for an observation to be included.

Here, we omit observations for which **region** is not defined.

- ▷ **VAR** statement: specifies the variables for which the statistics are requested.
- ▷ **BY** statement: requests separate analysis for the groups (here, regions).

- The following output is generated:

SRS means - for Belgium
The MEANS Procedure

Variable	N	Mean	Std Error

LNBMI	8384	3.1872184	0.0018447
LNVOEG	8250	1.7029508	0.0089543
GHQ12	8212	1.6613492	0.0295842
SGP	8532	0.9035396	0.0031963

SRS means - for regions

REGIONCH=Brussels

The MEANS Procedure

Variable	N	Mean	Std Error

LNBMI	2499	3.1758770	0.0033726
LNVOEG	2412	1.8097483	0.0162057
GHQ12	2397	1.8627451	0.0569024
SGP	2557	0.8056316	0.0078271

REGIONCH=Flanders

Variable	N	Mean	Std Error

LNBMI	2933	3.1824771	0.0029933
LNVOEG	2917	1.5163521	0.0152027
GHQ12	2914	1.3853809	0.0462510
SGP	2976	0.9522849	0.0039081

REGIONCH=Walloonia

Variable	N	Mean	Std Error

LNBMI	2952	3.2015302	0.0032165
LNVOEG	2921	1.8011065	0.0145518
GHQ12	2901	1.7721475	0.0510285
SGP	2999	0.9386462	0.0043828

8.4.2 Mean Estimation With Survey Procedure

- It is possible, and advisable, to use the SURVEYMEANS procedure:

```
proc surveymeans data=m.bmi_voeg mean stderr;  
title 'SRS means - infinite population for Belgium and regions';  
where (regionch^='');  
domain regionch;  
var lnBMI lnvoeg ghq12 sgp;  
run;
```

- The options are the same as in the MEANS procedure. Additionally:
 - ▷ **DOMAIN** option: requests separate analyses for each of the domain variable levels (here, regions).

It is similar to the BY statement, except that, at the same time, an analysis for the entire population (here, Belgium) is conducted.

Thus, one SURVEYMEANS call replaces both MEANS calls at the same time.

- The output generated is:

SRS means - infinite population for Belgium and regions

The SURVEYMEANS Procedure

Number of Observations 8564

Statistics

Variable	Mean	Std Error of Mean

LNBMI	3.187218	0.001845
LNVOEG	1.702951	0.008954
GHQ12	1.661349	0.029584
SGP	0.903540	0.003196

Domain Analysis: REGIONCH

REGIONCH	Variable	Mean	Std Error of Mean
Brussels	LNBMI	3.175877	0.003372
	LNVOEG	1.809748	0.016203
	GHQ12	1.862745	0.056894
	SGP	0.805632	0.007826
Flanders	LNBMI	3.182477	0.002993
	LNVOEG	1.516352	0.015201
	GHQ12	1.385381	0.046246
	SGP	0.952285	0.003908
Walloonia	LNBMI	3.201530	0.003216
	LNVOEG	1.801107	0.014550
	GHQ12	1.772148	0.051023
	SGP	0.938646	0.004382

- Note that the results are identical to those obtained with ordinary mean estimation, as it should.

- An important advantage is that also finite sampling corrections can be used:
 - ▷ When we want to take into account the size of the Belgian population, change the first line to:

```
proc surveymeans data=m.bmi_voeg total=10000000 mean stderr;
```

- ▷ The output then changes to:

SRS means - 1st finite population for Belgium and regions

The SURVEYMEANS Procedure

Number of Observations 8564

Statistics

Variable	Mean	Std Error of Mean

LNBMI	3.187218	0.001844
LNVOEG	1.702951	0.008950
GHQ12	1.661349	0.029572
SGP	0.903540	0.003195

Domain Analysis: REGIONCH

REGIONCH	Variable	Mean	Std Error of Mean
Brussels	LNBMI	3.175877	0.003371
	LNVOEG	1.809748	0.016196
	GHQ12	1.862745	0.056870
	SGP	0.805632	0.007823
Flanders	LNBMI	3.182477	0.002992
	LNVOEG	1.516352	0.015194
	GHQ12	1.385381	0.046226
	SGP	0.952285	0.003906
Walloonia	LNBMI	3.201530	0.003215
	LNVOEG	1.801107	0.014544
	GHQ12	1.772148	0.051001
	SGP	0.938646	0.004380

- ▷ As is clear here and in the overview tables to follow, the impact of the population is negligible since, for practical purposes:

$$N = 10,000,000 \simeq +\infty$$

- ▷ For the sake of illustration, suppose we actually conducted a census in a population of $N = n = 8564$.
- ▷ The first line then changes to:

```
proc surveymeans data=m.bmi_voeg total=8564 mean stderr;
```

- ▷ The output becomes:

```
SRS means - census-finite population for Belgium and regions
The SURVEYMEANS Procedure
Number of Observations           8564
```

Statistics		
Variable	Mean	Std Error of Mean

LNBMI	3.187218	0
LNVOEG	1.702951	0
GHQ12	1.661349	0
SGP	0.903540	0

Domain Analysis: REGIONCH

REGIONCH	Variable	Mean	Std Error of Mean
Brussels	LNBMI	3.175877	0
	LNVOEG	1.809748	0
	GHQ12	1.862745	0
	SGP	0.805632	0
Flanders	LNBMI	3.182477	0
	LNVOEG	1.516352	0
	GHQ12	1.385381	0
	SGP	0.952285	0
Walloonia	LNBMI	3.201530	0
	LNVOEG	1.801107	0
	GHQ12	1.772148	0
	SGP	0.938646	0

- ▷ As we have seen before, when $N = n$, it follows that $f = 1$ and hence the standard error vanishes.

8.4.3 Overviews

Logarithm of Body Mass Index					
Analysis	Procedure	Belgium	Brussels	Flanders	Wallonia
SRS	MEANS	3.187218(0.001845)	3.175877(0.003372)	3.182477(0.002993)	3.201530(0.003216)
SRS	SURVEYMEANS	3.187218(0.001845)	3.175877(0.003372)	3.182477(0.002993)	3.201530(0.003216)
SRS ($N = 10^7$)	SURVEYMEANS	3.187218(0.001845)	3.175877(0.003371)	3.182477(0.002992)	3.201530(0.003215)
SRS ($N = 8546$)	SURVEYMEANS	3.187218(0.000000)	3.175877(0.000000)	3.182477(0.000000)	3.201530(0.000000)

Logarithm of VOEG Score					
Analysis	Procedure	Belgium	Brussels	Flanders	Wallonia
SRS	MEANS	1.702951(0.008954)	1.809748(0.016203)	1.516352(0.015201)	1.801107(0.014550)
SRS	SURVEYMEANS	1.702951(0.008954)	1.809748(0.016203)	1.516352(0.015201)	1.801107(0.014550)
SRS ($N = 10^7$)	SURVEYMEANS	1.702951(0.008950)	1.809748(0.016196)	1.516352(0.015194)	1.801107(0.014544)
SRS ($N = 8546$)	SURVEYMEANS	1.702951(0.000000)	1.809748(0.000000)	1.516352(0.000000)	1.801107(0.000000)

General Health Questionnaire – 12					
Analysis	Procedure	Belgium	Brussels	Flanders	Wallonia
SRS	MEANS	1.661349(0.029584)	1.862745(0.056894)	1.385381(0.046246)	1.772148(0.051023)
SRS	SURVEYMEANS	1.661349(0.029584)	1.862745(0.056894)	1.385381(0.046246)	1.772148(0.051023)
SRS ($N = 10^7$)	SURVEYMEANS	1.661349(0.029572)	1.862745(0.056870)	1.385381(0.046226)	1.772148(0.051001)
SRS ($N = 8546$)	SURVEYMEANS	1.661349(0.000000)	1.862745(0.000000)	1.385381(0.000000)	1.772148(0.000000)

Stable General Practitioner (0/1)					
Analysis	Procedure	Belgium	Brussels	Flanders	Wallonia
SRS	MEANS	0.903540(0.003196)	0.805632(0.007826)	0.952285(0.003908)	0.938646(0.004382)
SRS	SURVEYMEANS	0.903540(0.003196)	0.805632(0.007826)	0.952285(0.003908)	0.938646(0.004382)
SRS ($N = 10^7$)	SURVEYMEANS	0.903540(0.003195)	0.805632(0.007823)	0.952285(0.003906)	0.938646(0.004380)
SRS ($N = 8546$)	SURVEYMEANS	0.903540(0.000000)	0.805632(0.000000)	0.952285(0.000000)	0.938646(0.000000)

8.4.4 What Comes Later?

- At the end of most chapters, we will re-estimate the means, accounting for the design feature under consideration.
- In Part IX, we will consider:
 - ▷ All design features combined
 - ▷ Frequency tables
 - ▷ Linear regression
 - ▷ Logistic regression
 - ▷ The use of analysis tools for complex data structures

Part IV

Systematic Sampling

Chapter 9

General Concepts and Design

- ▷ Principle of systematic sampling
- ▷ Examples

9.1 Systematic Sampling

- At first sight, a relatively simple variation to SRS.
- Earlier, SRS was labor-intensive, especially for long lists.
Systematic sampling was an “equivalent” but simpler method.
- It is always done without replacement.

- Essentially done to increase precision:
 - ▷ The units are ordered according to a variable that is related with the survey variable Y ; say from small to large.
 - ▷ By 'jumping' through the list, one ensures that small, medium, and large units are all present.
 - ▷ With SRS, it is possible, purely by chance, to have imbalance.
 - ▷ While this does not create bias, it does make the resulting estimators variable.

9.1.1 Quantities and Procedure

- We need the following information:

- ▷ Population \mathcal{P}
- ▷ Population size N
- ▷ Sample size n
- ▷ A list of the population units

- The sample fraction

$$f = \frac{n}{N}$$

- Write the sample fraction as

$$f = \frac{1}{g}$$

- We then say that 1 in $g = f^{-1}$ units is selected.
- Two quantities describe the procedure:
 - ▷ **The random start:** a random number s , uniformly drawn between 1 and g .
 - ▷ **The jump:** g , which follows by design.

9.1.2 Example

- $N = 8500$
- $n = 100$
- Then,

$$f = \frac{n}{N} = \frac{100}{8500} = \frac{1}{85}$$

and hence $g = 85$, the jump.

- Generate a random start; let us say, $s = 17$.

i	General	Example
1	s	17
2	$s + 1 \times g$	$17 + 1 \times 85 = 102$
3	$s + 2 \times g$	$17 + 2 \times 85 = 187$
4	$s + 3 \times g$	$17 + 3 \times 85 = 272$
\vdots	\vdots	\vdots
i	$s + (i - 1) \cdot g$	$17 + (i - 1) \times 85$
\vdots	\vdots	\vdots
100		$17 + 99 \times 85 = 8432$

9.1.3 Number of Samples

Data	N	n	S	
			SRS	Systematic
General	N	n	$\binom{N}{n}$	$\frac{N}{n} = \frac{1}{f} = g$
Artificial Population	4	2	6	2
Surveytown	8	2	28	4
Health Interview Survey	10,000,000	10,000	$10^{34,338}$	1000

- There obviously is a **huge** difference between the number of SRS's and the number of systematic samples.
- The reason is that there is a relatively small number of samples possible, **given the list**.
- At the same time, the number of possible lists will be huge for large populations (e.g., Belgian population).
- Enumeration formulas for the number of lists are not very elegant, since a lot of different lists will give rise to the same samples.
Neither are they very relevant.

9.1.4 Example: Artificial Population

- Consider the three lists that give rise to different samples:

$$\mathcal{L}_1 = (1\ 2\ 3\ 4)$$

$$\mathcal{L}_2 = (1\ 3\ 2\ 4)$$

$$\mathcal{L}_3 = (1\ 2\ 4\ 3)$$

- All other lists (there are 24 permutations of 4 numbers) produce the same samples as one of the three lists above.

- The sampling mechanism then is:

P_s					
Systematic					
s	Sample	SRS	\mathcal{L}_1	\mathcal{L}_2	\mathcal{L}_3
1	{1,2}	1/6	0	1/2	0
2	{1,3}	1/6	1/2	0	0
3	{1,4}	1/6	0	0	1/2
4	{2,3}	1/6	0	0	1/2
5	{2,4}	1/6	1/2	0	0
6	{3,4}	1/6	0	1/2	0

- Thus, all 6 samples that can be realized with SRS (without replacement), can be realized with systematic sampling too.

- However, for a given list, only 2 samples are possible.
- The expectations for the average:

$$\mathcal{L}_1 : E(\bar{y}) = \frac{1}{2} \cdot [2.0 + 3.0] = 2.5$$

$$\mathcal{L}_2 : E(\bar{y}) = \frac{1}{2} \cdot [1.5 + 3.5] = 2.5$$

$$\mathcal{L}_3 : E(\bar{y}) = \frac{1}{2} \cdot [2.5 + 2.5] = 2.5$$

- Hence, all three lists produce unbiased estimators.

- The variances:

$$\mathcal{L}_1 : \sigma_{\bar{y}}^2 = \frac{(2.0 - 2.5)^2 + (3.0 - 2.5)^2}{2} = \frac{0.5}{2} = 0.25$$

$$\mathcal{L}_2 : \sigma_{\bar{y}}^2 = \frac{(1.5 - 2.5)^2 + (3.5 - 2.5)^2}{2} = \frac{2.0}{2} = 1.00$$

$$\mathcal{L}_3 : \sigma_{\bar{y}}^2 = \frac{(2.5 - 2.5)^2 + (2.5 - 2.5)^2}{2} = \frac{0.0}{2} = 0.00$$

- Recall that the variance under SRS was 0.4167.
- Thus, some lists decrease the variance, while others increase the variance.
- (Note that \mathcal{L}_3 is a somewhat special case, owing to the fact that the list is very small.

- Note that the average of the three variances is:

$$\frac{0.25 + 1.00 + 0.00}{3} = 0.4167$$

- Thus, there are two views possible:
 - ▷ **Conditional view:** The variance under systematic sampling is a function of the list chosen: it is important to choose a **good** list.
 - ▷ **Marginal view:** The variance, averaged (marginalized) over all lists, is the same as under SRS without replacement.
- The second fact sometimes leads to the statement that the computations and procedures under systematic sampling are exactly the same as with SRS: this is true under one view only.

9.2 A Good List in Practice

- A list is **good** if the variable used for ordering is as close to monotonically (increasing or decreasing) related to the survey variable Y as possible.
 - ▷ **Health Interview Survey**: towns ordered from large to small in terms of their population.
 - ▷ **Health Interview Survey**: households ordered in terms of their statistical sector, HH size, and age of reference person.
- A **bad** list shows cyclic behavior in synchrony with the jump:
 - ▷ **The train time table**: if you select every 5th train, in a station with exactly 5 trains an hour.
 - ▷ **Blocks in cities in the Americas**: the regular block patron may play tricks on the survey scientist.

9.3 Example: Surveytown

- Let us add a third variable Z_I to the existing ones X_I and Y_I :
 - ▷ X_I : number of building lots in block I
 - ▷ Z_I : number of newspapers delivered in block I
 - ▷ Y_I : number of dwellings (buildings) in block I

- Listing of Surveytown:

I	X_I	Z_I	Y_I
1	1	8	1
2	3	1	2
3	4	6	3
4	6	10	4
5	7	4	5
6	8	3	6
7	10	7	7
8	11	11	8

- One of our estimands is the population total $Y = 36$

- Construct lists based on X_I and Z_I :

$$\mathcal{L}_X = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)$$

$$\mathcal{L}_Z = (2\ 6\ 5\ 3\ 7\ 1\ 4\ 8)$$

- Consider systematic samples of size $n = 2$:
- Sample fraction and jump:

$$f = \frac{2}{8} = \frac{1}{4}$$

and hence $g = 4$

- This produces the following samples:

$$\mathcal{L}_X = (\textcolor{red}{1} \ 2 \ 3 \ 4 \mid \textcolor{red}{5} \ 6 \ 7 \ 8)$$

$$\mathcal{L}_X = (1 \ \textcolor{red}{2} \ 3 \ 4 \mid 5 \ \textcolor{red}{6} \ 7 \ 8)$$

$$\mathcal{L}_X = (1 \ 2 \ \textcolor{red}{3} \ 4 \mid 5 \ 6 \ \textcolor{red}{7} \ 8)$$

$$\mathcal{L}_X = (1 \ 2 \ 3 \ \textcolor{red}{4} \mid 5 \ 6 \ 7 \ \textcolor{red}{8})$$

and

$$\mathcal{L}_Z = (\textcolor{red}{2} \ 6 \ 5 \ 3 \mid \textcolor{red}{7} \ 1 \ 4 \ 8)$$

$$\mathcal{L}_Z = (2 \ \textcolor{red}{6} \ 5 \ 3 \mid 7 \ \textcolor{red}{1} \ 4 \ 8)$$

$$\mathcal{L}_Z = (2 \ 6 \ \textcolor{red}{5} \ 3 \mid 7 \ 1 \ \textcolor{red}{4} \ 8)$$

$$\mathcal{L}_Z = (2 \ 6 \ 5 \ \textcolor{red}{3} \mid 7 \ 1 \ 4 \ \textcolor{red}{8})$$

- In summary, the samples are:

$$\mathcal{S}_X = \{ \{1, 5\}, \{2, 6\}, \{3, 7\}, \{4, 8\} \}$$

$$\mathcal{S}_Z = \{ \{1, 6\}, \{2, 7\}, \{3, 8\}, \{4, 5\} \}$$

- The following two pages present:

- ▷ sample probabilities P_s

- ▷ estimates \hat{y}_s

for

- ▷ SRS,

- ▷ systematic sampling with list \mathcal{L}_X

- ▷ systematic sampling with list \mathcal{L}_Z

s	Sample	P_s			\hat{y}_s		
		SRS	Systematic		SRS	Systematic	
			\mathcal{L}_X	\mathcal{L}_Z		\mathcal{L}_X	\mathcal{L}_Z
1	{1,2}	1/28	0	0	12		
2	{1,3}	1/28	0	0	16		
3	{1,4}	1/28	0	0	20		
4	{1,5}	1/28	1/4	0	24	24	
5	{1,6}	1/28	0	1/4	28		28
6	{1,7}	1/28	0	0	32		
7	{1,8}	1/28	0	0	36		
8	{2,3}	1/28	0	0	20		
9	{2,4}	1/28	0	0	24		
10	{2,5}	1/28	0	0	28		
11	{2,6}	1/28	1/4	0	32	32	
12	{2,7}	1/28	0	1/4	36		36
13	{2,8}	1/28	0	0	40		
14	{3,4}	1/28	0	0	28		
15	{3,5}	1/28	0	0	32		
16	{3,6}	1/28	0	0	36		

s	Sample	P_s			\hat{y}_s		
		SRS	Systematic		SRS	Systematic	
			\mathcal{L}_X	\mathcal{L}_Z		\mathcal{L}_X	\mathcal{L}_Z
17	{3,7}	1/28	1/4	0	40	40	
18	{3,8}	1/28	0	1/4	44		44
19	{4,5}	1/28	0	1/4	36		36
20	{4,6}	1/28	0	0	40		
21	{4,7}	1/28	0	0	44		
22	{4,8}	1/28	1/4	0	48	48	
23	{5,6}	1/28	0	0	44		
24	{5,7}	1/28	0	0	48		
25	{5,8}	1/28	0	0	52		
26	{6,7}	1/28	0	0	52		
27	{6,8}	1/28	0	0	56		
28	{7,8}	1/28	0	0	60		
Expectation					36	36	36
Variance					144	80	32
Standard error					12.00	8.94	2.83

- The expectations for the total:

$$\mathcal{L}_X : E(\bar{y}) = \frac{1}{4} \cdot [24 + 32 + 40 + 48] = \frac{144}{4} = 36$$

$$\mathcal{L}_Z : E(\bar{y}) = \frac{1}{4} \cdot [28 + 36 + 36 + 44] = \frac{144}{4} = 36$$

- Hence, both lists produce unbiased estimators.

- The variances:

$$\mathcal{L}_X : \sigma_{\bar{y}}^2 = \frac{(24 - 36)^2 + (32 - 36)^2 + (40 - 36)^2 + (48 - 36)^2}{4} = \frac{320}{4} = 80$$

$$\mathcal{L}_Z : \sigma_{\bar{y}}^2 = \frac{(36 - 36)^2 + (28 - 36)^2 + (36 - 36)^2 + (44 - 36)^2}{4} = \frac{128}{4} = 32$$

- Recall that the variance under SRS was 144.
- Both lists increase precision by reducing the variance, but \mathcal{L}_Z more spectacularly so.

Chapter 10

Analysis

- ▷ Estimators
- ▷ Variances
- ▷ The intra-class correlation
- ▷ Sample size determination

10.1 Estimators

Quantity	SRS	SYS
Total	$\hat{y} = \frac{N}{n} \sum_{i=1}^n y_i$	$\hat{y} = \frac{N}{n} \sum_{i=1}^n y_i$
Average	$\hat{y} = \frac{1}{n} \sum_{i=1}^n y_i$	$\hat{y} = \frac{1}{n} \sum_{i=1}^n y_i$
Proportion	$\hat{z} = \frac{1}{n} \sum_{i=1}^n z_i$	$\hat{z} = \frac{1}{n} \sum_{i=1}^n z_i$

- The estimators under SRS and SYS are identical.

10.2 Variances

- Recall the confusion between a **conditional** and **marginal** view.
- Several authors suggest using the same formulas for estimating the variance under SYS as under SRS, acknowledging that the true population variance may be different:
 - ▷ Scheaffer, Mendenhall, and Ott (1990)
- Several authors consider corrections, in terms of intra-class correlation:
 - ▷ Kish (1965)
 - ▷ Lehtonen and Pahkinen (1995)
 - ▷ Kottnerus (2003)

- Some of these corrections are a bit awkward to calculate in practice.
- The availability of modern software tools has made the task a bit easier.
- We will present formulas, based on a combination of the various proposals.
 - ▷ Given a list, there are g samples, equal to the jump.
 - ▷ Each of these g samples can be seen as a **cluster**.
 - ▷ The idea is that, with a **good** list, 'small', 'medium', and 'large' units are represented in all samples (clusters).
 - ▷ This implies that, within a cluster, the units are maximally different.
 - ▷ This implies that, within a cluster, there is negative correlation ρ .
 - ▷ Therefore, a key quantity is the **within-cluster correlation** ρ .
- Overview of the variances:

Quantity	SRS	SYS
Pop. var.	$\hat{s}_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$	$\hat{s}_{y,\text{sys}}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \cdot [1 + (n-1)\rho]$
Total	$\hat{\sigma}_{\bar{y}}^2 = \frac{N^2}{n} (1-f) \hat{s}_y^2$	$\hat{\sigma}_{\bar{y}}^2 = \frac{N^2}{n} (1-f) \hat{s}_{y,\text{sys}}^2$
Average	$\hat{\sigma}_{\bar{y}}^2 = \frac{1}{n} (1-f) \hat{s}_y^2$	$\hat{\sigma}_{\bar{y}}^2 = \frac{1}{n} (1-f) \hat{s}_{y,\text{sys}}^2$
Proportion	$\sigma_{\hat{p}}^2 = \frac{1}{n} \frac{N-n}{N-1} \hat{p}\hat{q}$	$\sigma_{\hat{p}}^2 = \frac{1}{n} \frac{N-n}{N-1} \hat{p}\hat{q} \cdot [1 + (n-1)\rho]$

10.3 The Intra-Cluster Correlation

- The intra-cluster (intraclass) correlation can be derived in several ways:
 - ▷ Using ANOVA sums of squares
 - ▷ Using a hierarchical model
- We will illustrate the latter.

- Assume the model:

$$Y_{IJ} = \mu + b_I + \varepsilon_{IJ}$$

- ▷ Y_{IJ} is the population quantity for subject J in cluster (sample) I
- ▷ μ is the overall mean (population average)
- ▷ $\mu + b_I$ is the cluster-specific average:

$$b_I \sim N(0, \tau^2)$$

- ▷ ε_{IJ} is an individual-level deviation:

$$\varepsilon_{IJ} \sim N(0, \lambda^2)$$

- ▷ The following terminology is commonly used:
 - * μ is a **fixed effect** (fixed intercept).
 - * b_I is a **random effect** (random intercept).
 - * ε_{IJ} is a residual deviation ('error' in samples).

- This is an instance of a **linear mixed model**.

- We can then show that:

$$\text{var}(Y_{IJ}) = \text{var}(b_I + \varepsilon_{IJ}) = \text{var}(b_I) + \text{var}(\varepsilon_{IJ}) = \tau^2 + \lambda^2$$

$$\text{cov}(Y_{IJ}, Y_{IJ'}) = \text{cov}(b_I + \varepsilon_{IJ}, b_I + \varepsilon_{IJ'}) = \text{var}(b_I) = \tau^2$$

and hence

$$\rho = \text{corr}(Y_{IJ}, Y_{IJ'}) = \frac{\tau^2}{\lambda^2 + \tau^2}$$

- Given this, we can also specify the model as:

$$\begin{pmatrix} Y_{I1} \\ Y_{I2} \\ \vdots \\ Y_{In} \end{pmatrix} \sim N \left[\begin{pmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{pmatrix}, \begin{pmatrix} \lambda^2 + \tau^2 & \tau^2 & \dots & \tau^2 \\ \tau^2 & \lambda^2 + \tau^2 & \dots & \tau^2 \\ \vdots & \vdots & \ddots & \vdots \\ \tau^2 & \tau^2 & \dots & \lambda^2 + \tau^2 \end{pmatrix} \right]$$

This is called the **compound-symmetry** model.

- Practically, we can fit such a model in SAS.

10.3.1 Example: Surveytown

- Let us consider both lists \mathcal{L}_X and \mathcal{L}_Z .
- The population is entered into a dataset **by cluster (sample)**.
- A program to display the data:

```
proc print data=m.surveytown01;  
title 'Listing Surveytown - List LX';  
run;
```

with listings

Listing Surveytown - List LX

Obs	sample	y
1	1	1
2	1	5
3	2	2
4	2	6
5	3	3
6	3	7
7	4	4
8	4	8

Listing Surveytown - List LZ

Obs	sample	y
1	1	1
2	1	6
3	2	2
4	2	7
5	3	3
6	3	8
7	4	4
8	4	5

- The linear mixed model can now be fitted as follows:

```
proc mixed data=m.surveytown01 method=ml;  
title 'Intraclass correlation Surveytown - List LX';  
class sample;  
model y = / solution;  
repeated / subject=sample type=cs rcorr;  
run;
```

with a similar program for the second sample.

- A perspective on the statements and options:
 - ▷ **CLASS** statement: states that the variable SAMPLE is an indicator, and not a continuous variable.

- ▷ **MODEL** statement: specifies the fixed effects; the intercept comes by default, so there is no reason to specify it.
 - * **'solution'** option: requests outputting of the fixed effects.
- ▷ **'REPEATED'** statement: used to specify the variance-covariance structure.
 - * **'subject='** option: specifies the level of independent replication; samples in our case.
 - * **'type='** option: specifies the covariance structure, compound symmetry (CS) in our case.
 - * **'rcorr'** option: requests outputting of the corresponding correlation matrix.

- A selection of the output for \mathcal{L}_X :

▷ The correlation:

```
Intraclass correlation Surveytown - List LX
The Mixed Procedure
```

```
Estimated R Correlation
Matrix for sample 1
```

Row	Col1	Col2
1	1.0000	-0.5238
2	-0.5238	1.0000

* The correlation is $\rho_{\mathcal{L}_X} = -0.5238$.

* Note the title 'Matrix for sample 1': this is all right, since the matrix is common to all 4 samples.

▷ The fixed effects:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	4.5000	0.5590	3	8.05	0.0040

* The value $\mu = 4.5$ is the proper population average, indeed.

▷ Recall the connection between both variances:

$$\hat{s}_{y,\text{sys}}^2 = \sigma_{\hat{y}}^2 [1 + \rho(n - 1)]$$

- ▷ However, this is assuming there is no correlation in the SRS case, but this is not true, since the corresponding panel for the SRS case is:

Estimated R Correlation

Matrix for sample 1

Row	Col1	Col2
1	1.0000	-0.1429
2	-0.1429	1.0000

- ▷ Hence, the correlation here is $\rho_{\text{SRS}(\text{without})} = -0.1429$.
- ▷ However, the correlation for SRS with replacement is $\rho_{\text{SRS}(\text{with})} = 0$.
- ▷ The reason is that selection without replacement forces sample units to be different, hence the negative correlation.

- Similar output for \mathcal{L}_Z

- ▷ Correlation and mean:

Intraclass correlation Surveytown - List LZ

Estimated R Correlation
Matrix for sample 1

Row	Col1	Col2
1	1.0000	-0.8095
2	-0.8095	1.0000

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	4.5000	0.3536	3	12.73	0.0010

- ▷ The correlation is $\rho_{\mathcal{L}_Z} = -0.8095$, more negative than with \mathcal{L}_Z , underscoring that the variance reduction is more important here.

- Return to the relationship between the variances, and rewrite it as:

$$\frac{\sigma_{\hat{y}, \text{SRS}(\text{with})}^2}{1 + \rho_{\text{SRS}(\text{with})}(n-1)} = \frac{\sigma_{\hat{y}, \text{SRS}(\text{without})}^2}{1 + \rho_{\text{SRS}(\text{without})}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{L}_1}^2}{1 + \rho_{\mathcal{L}_X}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{L}_2}^2}{1 + \rho_{\mathcal{L}_Z}(n-1)}$$

$$\frac{168}{1 + 0.0000 \times (2-1)} = \frac{144}{1 - 0.1429 \times (2-1)} = \frac{80}{1 - 0.5238 \times (2-1)} = \frac{32}{1 - 0.8095 \times (2-1)}$$

10.3.2 Example: Artificial Population

- The intra-cluster correlations for the three lists of the Artificial Population

Method	Variance	ρ	Relationship
SRS (without)	0.4167	-0.33	$\frac{0.4167}{1-0.33 \times (2-1)} = 0.6250$
SRS (with)	0.6250	0.00	$\frac{0.6250}{1+0.00 \times (2-1)} = 0.6250$
SYS(\mathcal{L}_1)	0.2500	-0.60	$\frac{0.2500}{1-0.60 \times (2-1)} = 0.6250$
SYS(\mathcal{L}_2)	1.0000	0.60	$\frac{1.0000}{1+0.60 \times (2-1)} = 0.6250$
SYS(\mathcal{L}_3)	0.0000	-1.00	undetermined

- The variance of SRS with replacement can be seen as a 'norm', which is recovered by all methods, when correction for the within-sample correlations.
- If samples are heterogeneous, we obtain a negative correlation, which is a good thing, since it decreases the variance of the estimator.
- Note that the first and second lists have precisely opposite effects.

10.4 Sample Size Calculation

- Consider the case of an **average**.
- The variance takes the form:

$$\sigma_{\bar{y}}^2 = \frac{1}{n} \left(\frac{N-n}{N} \right) \sigma_Y^2 \cdot [1 + (n-1)\rho]$$

- Algebraic manipulation, and ordering the terms along the powers of n produces:

$$\rho\sigma_Y^2 n^2 + [N\sigma_{\bar{y}}^2 - N\rho\sigma_Y^2 + (1-\rho)\sigma_Y^2]n - N(1-\rho)\sigma_Y^2 = 0$$

which is a quadratic equation.

- It is straightforward to solve such an equation for n .
- Even though a closed form exists, it is not an elegant expression.
- Similar quadratics exist for a total and a proportion.
- Let us consider the case of sampling with replacement and/or sampling.

- **Overview for sampling without replacement:**

- ▷ **Average:**

$$[\rho\sigma_Y^2]n^2 + [N\sigma_{\hat{y}}^2 - N\rho\sigma_Y^2 + (1 - \rho)\sigma_Y^2]n - N(1 - \rho)\sigma_Y^2 = 0$$

- ▷ **Total:**

$$[N^2\rho\sigma_Y^2]n^2 + [N\sigma_{\hat{y}}^2 - N^3\rho\sigma_Y^2 + N^2(1 - \rho)\sigma_Y^2]n - N^3(1 - \rho)\sigma_Y^2 = 0$$

- ▷ **Proportion (absolute):**

$$[\rho PQ]n^2 + [(N - 1)\sigma_{\hat{p}}^2 - N\rho PQ + (1 - \rho)PQ]n - N(1 - \rho)PQ = 0$$

- ▷ **Proportion (relative):**

$$[\rho Q]n^2 + [(N - 1)k^2 P - N\rho Q + (1 - \rho)Q]n - N(1 - \rho)Q = 0$$

- **Overview for Sampling with replacement and/or $N \rightarrow +\infty$:**

Quantity	SRS	SYS
Total	$n = \frac{N^2 \sigma_Y^2}{\sigma_{\hat{y}}^2}$	$n = \frac{N^2 \sigma_Y^2 (1 - \rho)}{\sigma_{\hat{y}}^2 - \rho N^2 \sigma_Y^2}$
Average	$n = \frac{\sigma_Y^2}{\sigma_{\hat{y}}^2}$	$n = \frac{\sigma_Y^2 (1 - \rho)}{\sigma_{\hat{y}}^2 - \rho \sigma_Y^2}$
Proportion (absolute)	$n = \frac{PQ}{\sigma_{\hat{p}}^2}$	$n = \frac{PQ(1 - \rho)}{\sigma_{\hat{p}}^2 - \rho PQ}$
Proportion (relative)	$n = \frac{Q}{k^2 P}$	$n = \frac{Q(1 - \rho)}{k^2 P - \rho Q}$

10.4.1 Illustration of the Correlation's Impact

- Re-consider the example of sample size determination for a proportion.
- $P = 0.6 \Rightarrow Q = 0.4$
- $\sigma_{\hat{p}}^2 = 0.05^2$
- The expression for large sample becomes

$$n = \frac{0.24(1 - \rho)}{0.05^2 - 0.24\rho}$$

- We also solve the corresponding quadratic, assuming $N = 10,000$.

ρ	n		ρ	n	
	With/ $N \rightarrow +\infty$	Without (quadr.)		With/ $N \rightarrow +\infty$	Without (quadr.)
-1.00	1.98	1.98	0.00	96.00	95.99
-0.90	2.09	2.09	0.01	2376.00	770.41
-0.80	2.22	2.22	0.02	-102.26	4844.34
-0.70	2.39	2.39	0.03	-49.53	6545.19
-0.60	2.62	2.62	0.04	-32.45	7404.51
-0.50	2.94	2.94	0.05	-24.00	7921.86
-0.40	3.41	3.41	0.10	-10.05	8959.48
-0.30	4.19	4.19	0.20	-4.22	9479.44
-0.20	5.70	5.70	0.40	-1.54	9739.65
-0.15	7.17	7.17	0.60	-0.68	9826.42
-0.10	9.96	9.96	0.80	-0.25	9869.81
-0.08	11.94	11.94	0.90	-0.11	9884.27
-0.06	15.05	15.05	0.96	-0.04	9891.50
-0.04	20.63	20.62	0.97	-0.03	9892.62
-0.02	33.53	33.50	0.98	-0.02	9893.72
-0.01	49.47	49.35	0.99	-0.01	9894.79
0.00	96.00	95.99	1.00	0.00	9895.84

- The quantities for $\rho = 0$ correspond to SRS.
- $\rho < 0$ produces smaller sample sizes than SRS.
- $\rho > 0$ produces larger sample sizes, but only the quadratic formula makes sense now.

Part V

Benchmark (Ratio) Estimators

Chapter 11

General Concepts and Design

- ▷ Principle of benchmark estimation
- ▷ Connection with estimation of a ratio
- ▷ Examples

11.1 Benchmark Estimation is a Cuckoo's Egg

- SRS, SYS, and later STRAT and CLUST are **sampling methods**.
- Benchmark estimation is an (enhanced) estimation method, in two steps:
 - ▷ **Step 1:** Estimate a population quantity using a conventional method (e.g., SRS).
 - ▷ **Step 2:** Construct a second estimator, using the first estimator and a so-called benchmark as input.

11.1.1 Example

- Suppose a survey of farm yield is conducted.
- Suppose (SRS) estimators are available for two quantities:
 - ▷ X : total planting area for wheat:

$$\hat{x} = 3.75 \text{ million ha}$$

- ▷ Y : total wheat yield

$$\hat{y} = 6.00 \text{ million tonnes}$$

- ▷ $\Rightarrow R = \frac{Y}{X}$: wheat yield per ha

$$\Rightarrow \hat{r} = \frac{\hat{y}}{\hat{x}} = \frac{6.00}{3.75} = 1.60 \text{ tonnes/ha}$$

- Hence, we considered an **estimator of a ratio**.
- Note that both numerator and denominator have random error attached to them.

11.1.2 The General Principle

- Suppose we are confronted with a discrepancy:
 - ▷ From the **survey** we conclude that the planting area is $\hat{x} = 3.75$ million ha.
 - ▷ From a **census** we conclude that the planting area is $\hat{x}_b = 4.00$ million ha.
- It is sensible to assume the census is the gold standard (or at least more accurate).
- The original estimator for Y can now be **corrected**:
- We can then obtain a precise estimate of yield by multiplying the estimated ratio \hat{r} with the census quantity:

$$\hat{y}_b = \hat{r} \cdot \hat{x}_b = \frac{\hat{y}}{\hat{x}} \cdot \hat{x}_b = \frac{6.00}{3.75} \times 4.00 = 6.40 \text{ million tonnes}$$

- ▷ The subscript b refers to benchmark.

- ▷ We use a benchmark \hat{x}_b (in the ideal case, it *is* the true population quantity) to replace the original estimator \hat{y} with a hopefully improved **benchmark estimator** \hat{y}_b .
- ▷ In the literature, the benchmark estimator is traditionally called **ratio estimator**; due to the potential confusion between **estimator of a ratio** and **ratio estimator**, we prefer benchmark estimator.
- Some assumptions need to be verified for the benchmark estimator to be “better”:
 - ▷ (Unbiased) estimators \hat{x} and \hat{y} need to vary around the true population quantities in a proportional fashion: when \hat{x} is *large*, \hat{y} must be too, and vice versa.
 - ▷ The benchmark must not be too variable.
 - ▷ Both of these conditions will be formalized.
 - ▷ They imply that benchmarks can, but not always will, improve precision, or at least MSE.

- It will be shown that the benchmark estimator can be biased and still useful to use.
- A benchmark estimator can be applied to averages and totals alike.
- The technique is easy to apply given the required benchmark information is available.

Chapter 12

Analysis

- ▷ Estimators
- ▷ Variances
- ▷ Extensions
- ▷ Sample size determination

12.1 Estimators

- General expressions
- Application to one sample from Surveytown:
 - ▷ The Y sample is:

$\{1, 2\}$

- ▷ The corresponding X sample is:

$\{1, 3\}$

Quantity	Expression	Estimator	Expression
Total	Y	SRS estimator	$\hat{y} = \frac{N}{n} \sum_{i=1}^n y_i$
Average	\bar{Y}	SRS estimator	$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$
Ratio	$R = \frac{Y}{X} = \frac{\bar{Y}}{\bar{X}}$	Estimator of ratio	$\hat{r} = \frac{\hat{y}}{\hat{x}} = \frac{\bar{y}}{\bar{x}}$
Total	Y	Benchmark estimator	$\hat{y}_b = \hat{r} \cdot \hat{x}_b = \frac{\hat{y}}{\hat{x}} \hat{x}_b$
Average	\bar{Y}	Benchmark estimator	$\bar{y}_b = \frac{1}{N} \hat{y}_b$

Quantity	Expression	Estimator	Expression
Total	$Y = 36$	SRS	$\hat{y} = \frac{8}{2} \times (1 + 2) = \boxed{12}$
Total	$X = 50$	SRS	$\hat{x} = \frac{8}{2}(1 + 3) = 16$
Ratio	$R = \frac{Y}{X} = 0.72$	SRS	$\hat{r} = \frac{\hat{y}}{\hat{x}} = \frac{12}{16} = 0.75$
Total	$Y = 36$	Benchmark	$\hat{y}_b = \hat{r} \cdot \hat{x}_b = \frac{12}{16} \times 50 = \boxed{37.5}$

12.2 Example: Surveytown

- Re-consider both auxiliary variables, as in Section 9.3:
 - ▷ X_I : number of building lots in block I
 - ▷ Z_I : number of newspapers delivered in block I
 - ▷ Y_I : number of dwellings (buildings) in block I
- Recall the listing of Surveytown:

I	X_I	Z_I	Y_I
1	1	8	1
2	3	1	2
3	4	6	3
4	6	10	4
5	7	4	5
6	8	3	6
7	10	7	7
8	11	11	8

- Consider the estimators based on SRS without replacement, of size $n = 2$:
 - ▷ SRS for Y
 - ▷ Benchmark estimator for Y , based on benchmark X
 - ▷ Benchmark estimator for Y , based on benchmark Z

s	Y -sample	\hat{y}	X -sample	\hat{x}	\hat{r}_x	$\hat{y}_{b=X}$	Z -sample	\hat{z}	\hat{r}_z	$\hat{y}_{b=Z}$
1	{1,2}	12	{1,3}	16	0.75	37.50	{8,1}	36	0.33	16.67
2	{1,3}	16	{1,4}	20	0.80	40.00	{8,6}	56	0.29	14.29
3	{1,4}	20	{1,6}	28	0.71	35.71	{8,10}	72	0.28	13.89
4	{1,5}	24	{1,7}	32	0.75	37.50	{8,4}	48	0.50	25.00
5	{1,6}	28	{1,8}	36	0.78	38.89	{8,3}	44	0.56	31.82
6	{1,7}	32	{1,10}	44	0.73	36.36	{8,7}	60	0.53	26.67
7	{1,8}	36	{1,11}	48	0.85	37.50	{8,11}	76	0.47	23.68
8	{2,3}	20	{3,4}	28	0.71	35.71	{1,6}	28	0.71	35.71
9	{2,4}	24	{3,6}	36	0.67	33.33	{1,10}	44	0.55	27.27
10	{2,5}	28	{3,7}	40	0.70	35.00	{1,4}	20	1.40	70.00
11	{2,6}	32	{3,8}	44	0.73	36.36	{1,3}	16	2.00	100.00
12	{2,7}	36	{3,10}	52	0.69	37.50	{1,7}	32	1.13	56.25
13	{2,8}	40	{3,11}	56	0.71	35.71	{1,11}	48	0.83	41.67
14	{3,4}	28	{4,6}	40	0.70	35.00	{6,10}	64	0.44	21.88
15	{3,5}	32	{4,7}	44	0.73	36.36	{6,4}	40	0.80	40.00
16	{3,6}	36	{4,8}	48	0.75	37.50	{6,3}	36	1.00	50.00
17	{3,7}	40	{4,10}	56	0.71	35.71	{6,7}	52	0.77	38.46

s	Y -sample	\hat{y}	X -sample	\hat{x}	\hat{r}_x	$\hat{y}_{b=X}$	Z -sample	\hat{z}	\hat{r}_z	$\hat{y}_{b=Z}$
18	{3,8}	44	{4,11}	60	0.73	36.67	{6,11}	68	0.65	32.35
19	{4,5}	36	{6,7}	52	0.69	34.62	{10,4}	56	0.64	32.14
20	{4,6}	40	{6,8}	56	0.71	35.71	{10,3}	52	0.77	38.46
21	{4,7}	44	{6,10}	64	0.69	34.38	{10,7}	68	0.65	32.35
22	{4,8}	48	{6,11}	68	0.71	35.29	{10,11}	84	0.57	28.57
23	{5,6}	44	{7,8}	60	0.73	36.67	{4,3}	28	1.57	78.57
24	{5,7}	48	{7,10}	68	0.71	35.29	{4,7}	44	1.09	54.55
25	{5,8}	52	{7,11}	72	0.72	36.11	{4,11}	60	0.87	43.33
26	{6,7}	52	{8,10}	72	0.72	36.11	{3,7}	40	1.30	65.00
27	{6,8}	56	{8,11}	76	0.74	36.84	{3,11}	56	1.00	50.00
28	{7,8}	60	{10,11}	84	0.71	35.71	{7,11}	72	0.83	41.67
Expectation		36		50	0.72	36.15		50	0.81	40.37
Variance		144		296.89	$74 \cdot 10^{-5}$	1.85		296.89	$15 \cdot 10^{-2}$	383.84
s.e.		12				1.36				19.59
Bias		0				0.15				4.37
MSE		144				1.87				402.90
RMSE		12				1.37				20.07

- Benchmark X decreases the MSE enormously.
- Benchmark Z dramatically increases the MSE.
- Like with lists in SYS, and with mechanisms to follow: the impact of benchmark estimation, relative to SRS, can be beneficial or detrimental.
- Consider a graphical comparison of both benchmark estimators with SRS:

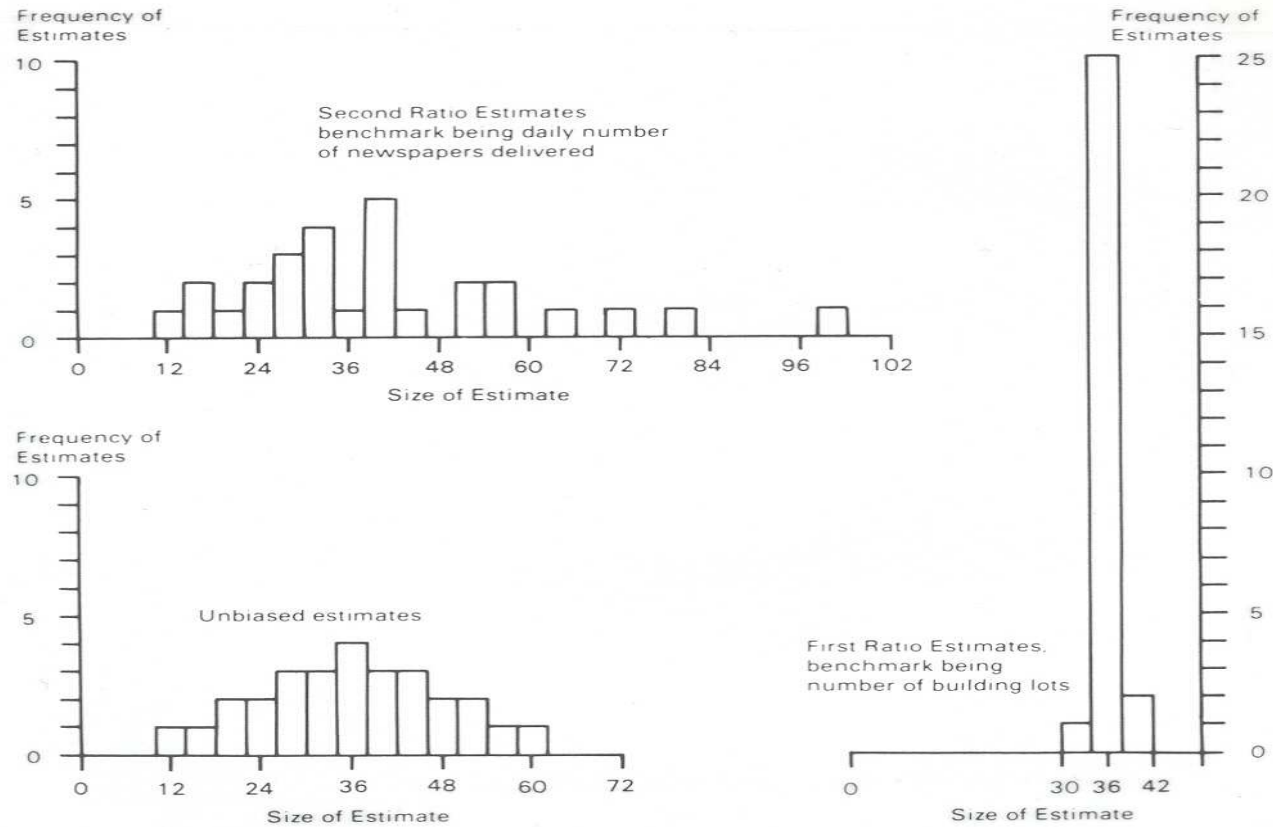
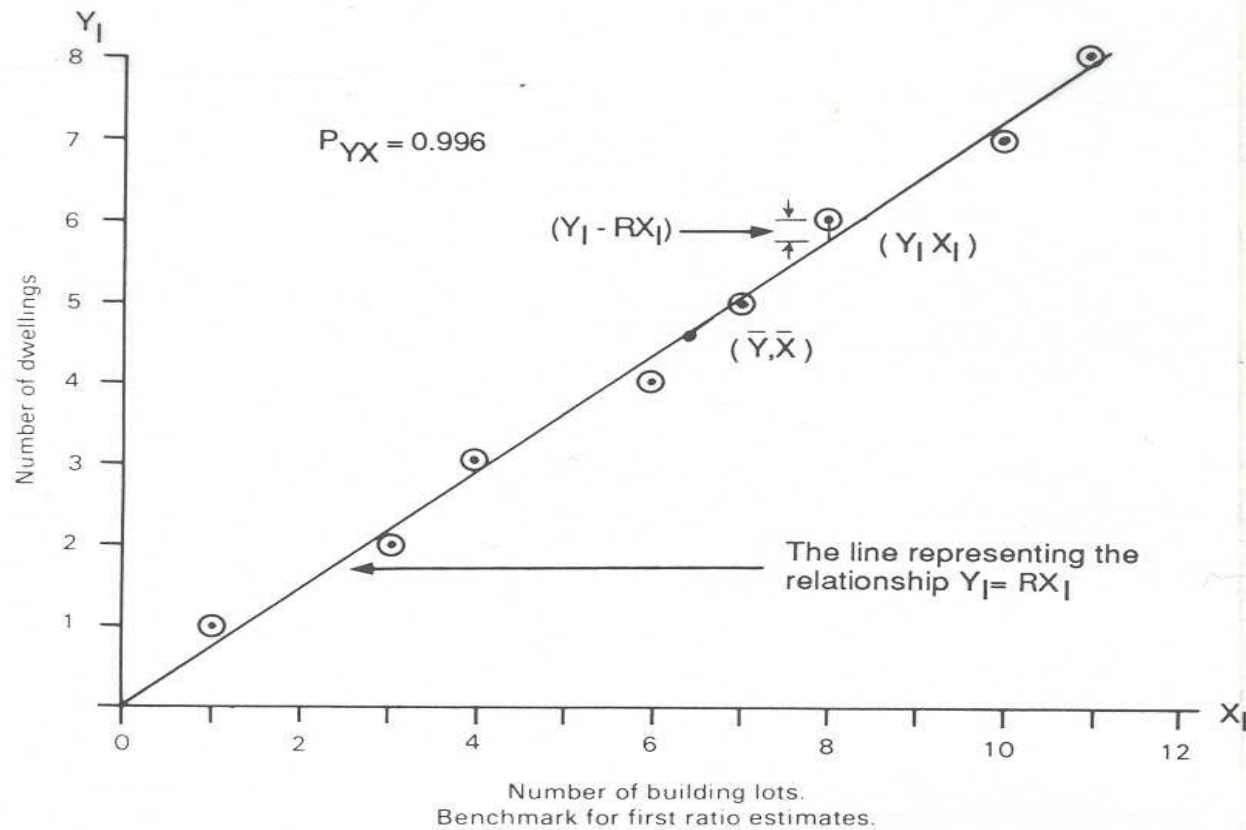


Figure 4.1 Frequency distributions of unbiased and ratio estimates using two different benchmarks derived from all 28 possible simple random samples of two blocks.

- The increased spread of estimates with Z , relative to X , also follows from the regression lines through the origin of Y_I on X_I , on the one hand:



and Z_I on X_I on the other hand:

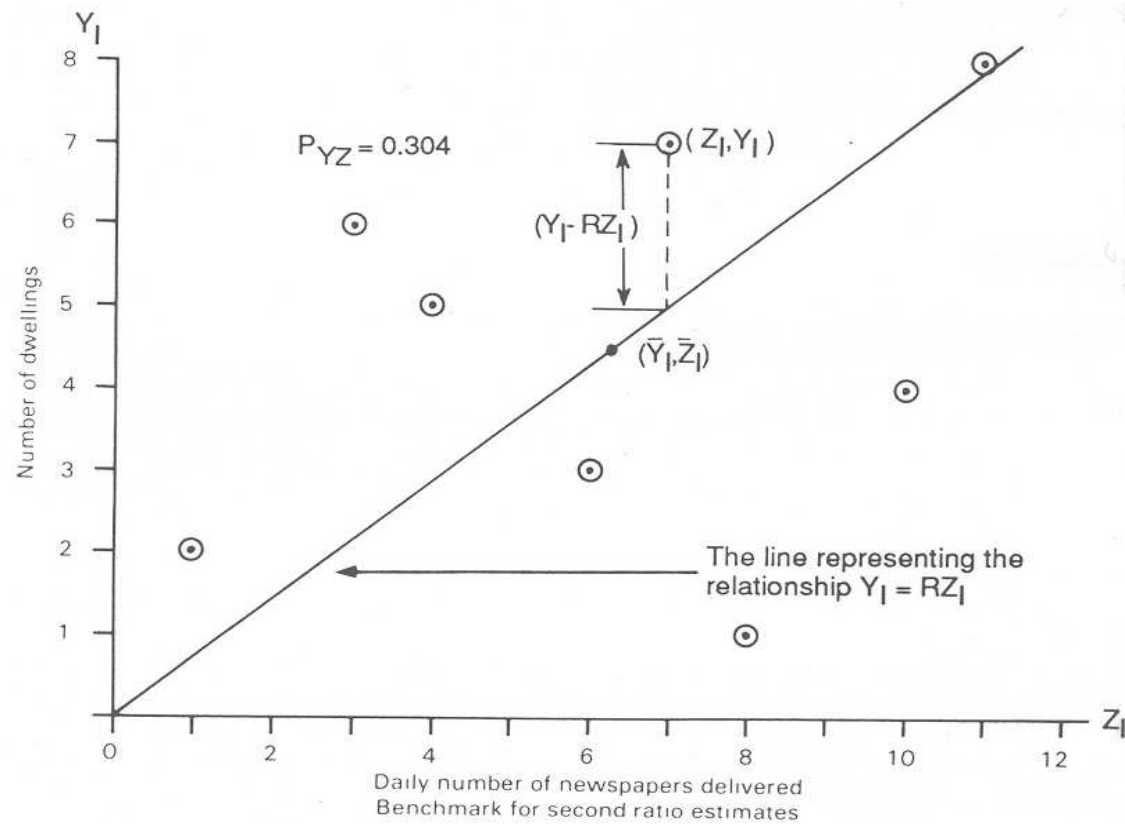


Figure 4.2 Scattergrams of unit (block) values of Y_I and X_I and of Y_I and Z_I relevant to the first and second sets of ratio estimates

- We observe two differences:
 - ▷ In the Z case the deviations are larger than in the X case: **precision**.
 - ▷ The line does not seem as appropriate in the Z case than in the X case: **bias**.
- In what follows, we will more formally study the conditions as to when this occurs.

12.2.1 Other Sample Sizes

- Let us consider benchmark estimators, based on X , for sample sizes $n = 1, 2, 4, 6, 8$:

Measure	Sample size n				
	1	2	4	6	8
Population estimand Y	36	36	36	36	36
Expectation $E(\hat{y}_{b=X})$	37.343	36.150	36.038	36.008	36
Bias	1.343	0.150	0.038	0.008	0
Range: lowest estimate	33.333	33.333	34.615	35.366	36
Range: highest estimate	50.000	40.000	37.500	36.765	36
Variance $\text{Var}(\hat{y}_{b=X})$	25.189	1.850	0.380	0.125	0
Mean square error	26.993	1.874	0.381	0.125	0
Standard error	5.019	1.360	0.616	0.353	0
Root mean square error	5.196	1.369	0.617	0.353	0

- The ratio estimator is biased.
- Both bias and variance decrease with increasing sample size: asymptotically unbiased.
- The variance is considerably smaller than for \hat{y} .

12.3 General Variance Formulae

- Let us display the formulas for two approaches:
 - ▷ Simple random sampling
 - ▷ Benchmark estimation
- and for three quantities:
 - ▷ average
 - ▷ total
 - ▷ ratio
- Note that for a ratio, by definition Y and X are used simultaneously, hence it is only listed in the benchmark column.

With replacement

Quantity	SRS	Benchmark
Pop. var.	$\sigma_Y^2 = \frac{1}{N} \sum_{I=1}^N (Y_I - \bar{Y})^2$	$\sigma^2 = \frac{1}{N} \sum_{I=1}^N (Y_I - RX_I)^2$
\bar{y}	$\sigma_{\bar{y}}^2 = \frac{1}{n} \sigma_Y^2$	$\sigma_{\bar{y}, \text{b.m.}}^2 = \frac{1}{n} \sigma^2$
\hat{y}	$\sigma_{\hat{y}}^2 = \frac{N^2}{n} \sigma_Y^2$	$\sigma_{\hat{y}, \text{b.m.}}^2 = \frac{N^2}{n} \sigma^2$
\hat{r}	—	$\sigma_{\hat{r}}^2 = \frac{1}{\bar{X}^2} \frac{1}{n} \sigma^2$

Without replacement

Quantity	SRS	Benchmark
Pop. var.	$S_Y^2 = \frac{1}{N-1} \sum_{I=1}^N (Y_I - \bar{Y})^2$	$S^2 = \frac{1}{N-1} \sum_{I=1}^N (Y_I - RX_I)^2$
\bar{y}	$\sigma_{\bar{y}}^2 = \frac{1}{n}(1-f)S_Y^2$	$\sigma_{\bar{y}, \text{b.m.}}^2 = \frac{1}{n}(1-f)S^2$
\hat{y}	$\sigma_{\hat{y}}^2 = \frac{N^2}{n}(1-f)S_Y^2$	$\sigma_{\hat{y}, \text{b.m.}}^2 = \frac{N^2}{n}(1-f)S^2$
\hat{r}	—	$\sigma_{\hat{r}}^2 = \frac{1}{\bar{X}^2} \frac{1}{n}(1-f)S^2$

12.3.1 Example: Surveytown

- In Part II, we calculated the variances of SRS estimators, taken without replacement, for $n = 1$ and $n = 2$.
- Let us double these up for benchmark estimation.
- The population variance, necessary for SRS: $S_Y^2 = 6$.
- For benchmark estimation, $\delta_I = Y_I - R X_I$ needs to be calculated:

I	X_I	Y_I	R	$\delta_I = Y_I - RX_I$
1	1	1	0.72	0.28
2	3	2	0.72	-0.16
3	4	3	0.72	0.12
4	6	4	0.72	-0.32
5	7	5	0.72	-0.04
6	8	6	0.72	0.24
7	10	7	0.72	-0.20
8	11	8	0.72	0.08

- The corresponding variance: $S^2 = 0.0466$

▷ Samples of size $n = 1$:

$$\text{SRS: } \sigma_{\hat{y}}^2 = \frac{8^2}{1} \times \left(1 - \frac{1}{8}\right) \times 6 = \frac{64 \times 7 \times 6}{8} = 336$$

$$\text{B.M.: } \sigma_{\hat{y}, \text{b.m.}}^2 = \frac{8^2}{1} \times \left(1 - \frac{1}{8}\right) \times 0.0466 = \frac{64 \times 7 \times 0.0466}{8} = 2.61$$

▷ Samples of size $n = 2$:

$$\text{SRS: } \sigma_{\hat{y}}^2 = \frac{8^2}{2} \times \left(1 - \frac{2}{8}\right) \times 6 = \frac{64 \times 6 \times 6}{2 \times 8} = 144$$

$$\text{B.M.: } \sigma_{\hat{y}, \text{b.m.}}^2 = \frac{8^2}{2} \times \left(1 - \frac{2}{8}\right) \times 0.0466 = \frac{64 \times 6 \times 0.0466}{2 \times 8} = 1.12$$

- We see, once more, there is a large beneficial impact in using X as a benchmark.

12.3.2 Relationship Between Variances

- Using that $\bar{Y} = R\bar{X}$, we can rewrite S^2 :

$$\begin{aligned} S^2 &= \frac{1}{N-1} \sum_I^N [(Y_I - \bar{Y}) - R(X_I - \bar{X})]^2 \\ &= S_Y^2 - 2RS_{YX} + R^2 S_X^2 \end{aligned}$$

- This can be used to rewrite the variances of the estimators:

$$\begin{aligned} \sigma_{\hat{r}}^2 &= \frac{1}{\bar{X}^2} (\sigma_{\hat{y}}^2 - 2R\sigma_{\hat{y}\hat{x}} + R^2\sigma_{\hat{x}}^2) \\ \sigma_{\hat{y}_r}^2 &= \sigma_{\hat{y}}^2 - 2R\sigma_{\hat{y}\hat{x}} + R^2\sigma_{\hat{x}}^2 \end{aligned}$$

where

$$\sigma_{\hat{y}\hat{x}} = E(\hat{y} - E\hat{y})(\hat{x} - E\hat{x}) = \frac{N^2}{n} \frac{N-n}{N} S_{YX}$$

12.4 Bias of a Benchmark Estimator

- We repeatedly used the quantities:

$$\delta_I = Y_I - R X_I$$

as a basis for variance estimation.

- This can be seen as a regression relationship:

$$Y_I = 0 + R X_I + \delta_I$$

- It clearly is a very particular linear regression:

linear regression through the origin

- This is a (sometimes strong) assumption.
- For example, if the true regression relationship is of the general linear type:

$$Y_I = \alpha + \beta X_I + \varepsilon_I$$

- The regression can be displayed graphically:

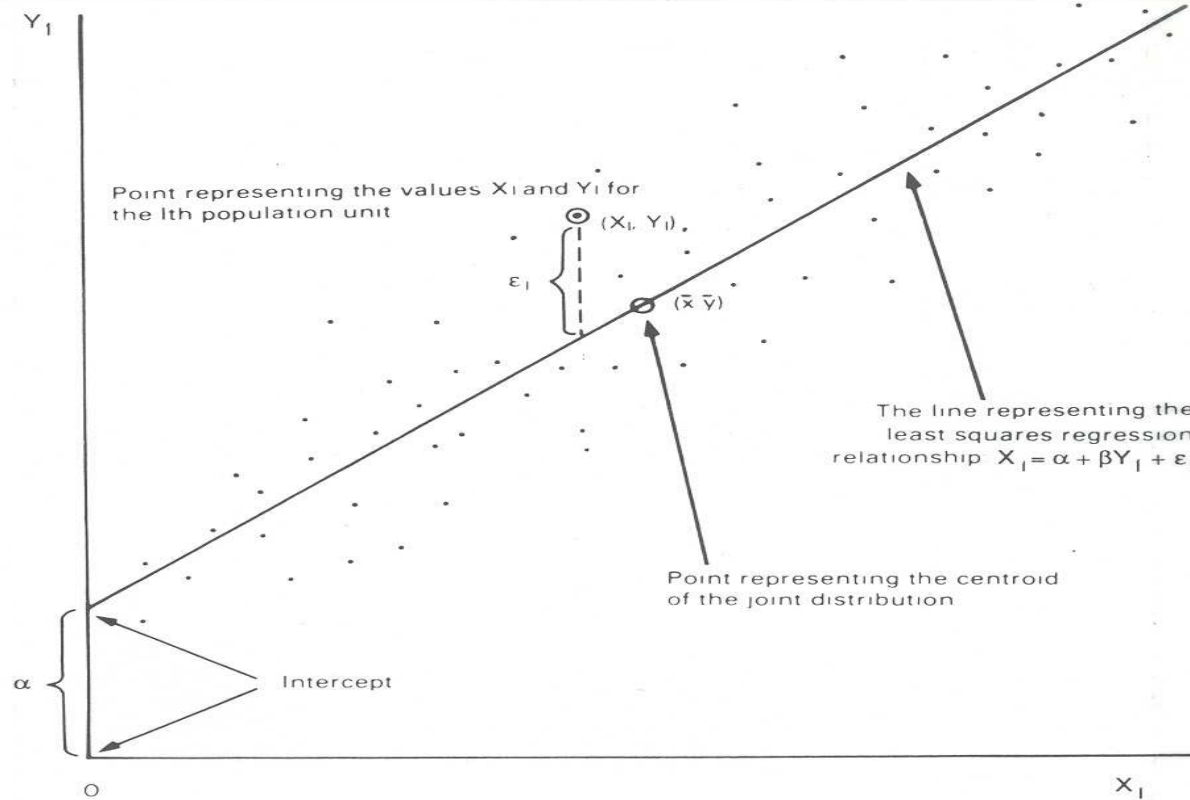


Figure 4.3 Scattergram representing the joint distribution of the pairs of values Y_i and X_i for each population unit, and the least squares regression line representing the relationship between the Y_i and X_i .

- The biases can be expressed as:

$$\text{bias}(\hat{r}) \simeq R(V_{\hat{x}}^2 - V_{\widehat{y\hat{x}}}) \simeq \alpha \cdot \frac{1}{\overline{X}} \cdot \frac{1}{n} \cdot (1 - f) \cdot V_Y^2$$

$$\text{bias}(\widehat{y_{\hat{r}}}) \simeq Y(V_{\hat{x}}^2 - V_{\widehat{y\hat{x}}}) \simeq \alpha \cdot \frac{N}{n} \cdot (1 - f) \cdot V_Y^2$$

- The bias decreases with:

▷ α (and disappears if $\alpha = 0$);

▷ increasing n (and disappears when $f = 1$, i.e., $n = N$).

- This implies that both estimators are **consistent**.

- A good benchmark X should be (roughly) proportional to the survey variable Y .
- In many situations, the fixed cost comes in the way of proportionality, even though linearity would be satisfied.
- In what follows, we will briefly consider appropriate extensions of the benchmark estimator.

12.5 Estimating the Variance

- Like in the SRS case (page151f), we first replace the **calculated** population-level variances by **estimates**:

$$\hat{s}_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y})^2$$

$$\hat{s}_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \hat{x})^2$$

$$\hat{s}_{yx} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y})(x_i - \hat{x})$$

- Next, the **calculated variances of the estimators** are replaced by **estimates variances of the estimators**:

$$\hat{\sigma}_{\hat{r}}^2 = \frac{1}{\overline{X}^2} \cdot \frac{1}{n} \cdot (1 - f) \cdot (\hat{s}_y^2 - 2\hat{r}\hat{s}_{yx} + \hat{r}^2\hat{s}_x^2)$$

$$\hat{\sigma}_{\hat{y}_{b.m.}}^2 = \frac{N^2}{n} \cdot (1 - f) \cdot (\hat{s}_y^2 - 2\hat{r}\hat{s}_{yx} + \hat{r}^2\hat{s}_x^2)$$

$$\hat{\sigma}_{\bar{y}_{b.m.}}^2 = \frac{1}{n} \cdot (1 - f) \cdot (\hat{s}_y^2 - 2\hat{r}\hat{s}_{yx} + \hat{r}^2\hat{s}_x^2)$$

12.6 Asymptotic Relative Efficiency (ARE)

- We have seen:
 - ▷ the benchmark estimator based on X is **more efficient** than SRS;
 - ▷ the benchmark estimator based on Z is **less efficient** than SRS.
- Efficiency is defined as the variance ratio and can be expanded as follows:

$$\begin{aligned}
\text{ARE}^{-1} &= \frac{\sigma_{\hat{y}_{\text{b.m.}}}^2}{\sigma_{\hat{y}}^2} = \frac{\frac{N^2}{n}(1-f)(S_Y^2 - 2RS_{YX} + R^2S_X^2)}{\frac{N^2}{n}(1-f)S_Y^2} \\
&= \frac{(S_Y^2 - 2RS_{YX} + R^2S_X^2)}{S_Y^2} \\
&= \frac{\left(\frac{S_Y^2}{Y^2} - 2\frac{Y}{X}\frac{\rho_{YX}S_XS_Y}{Y^2} + \frac{Y^2}{X^2}\frac{S_X^2}{Y^2}\right)}{\frac{S_Y^2}{Y^2}} \\
&= \frac{(V_Y^2 - 2\rho_{YX}V_XV_Y + V_X^2)}{V_Y^2} \\
&= 1 - 2\rho_{YX}\frac{V_X}{V_Y} + \left(\frac{V_X}{V_Y}\right)^2
\end{aligned}$$

- We then have that

$$\text{ARE}^{-1} \leq 1 \iff -2\rho_{YX}V + V^2 \leq 0$$

$$\iff -2\rho_{YX} + V \leq 0$$

$$\iff \frac{V}{2} \leq \rho_{YX}$$

$$\iff \boxed{\rho_{YX} \geq \frac{1}{2} \frac{V_X}{V_Y}}$$

- Thus, a benchmark is **good** if:
 - ▷ **(Efficiency:)** the survey and benchmark variables are sufficiently highly correlated;
 - ▷ **(Efficiency:)** the benchmark is sufficiently precise, relative to the precision of the survey variable;
 - ▷ **(Bias:)** The regression relationship between survey and benchmark variables passes (approximately) through the origin.

12.7 Extensions of Benchmark Estimators: Regression and Difference Estimators

- The definition of the ratio implies

$$\hat{y} = r\hat{x}$$

- The construction of the benchmark estimator implies

$$\hat{y}_{\text{b.m.}} = rX$$

- These two facts, taken together, allow us to derive the following relationship:

$$\hat{y}_{b.m.} = \hat{y} + \hat{y}_{b.m.} - \hat{y}$$

$$\hat{y}_{b.m.} = \hat{y} + rX - r\hat{x}$$

$$\hat{y}_{b.m.} = \hat{y} + r(X - \hat{x})$$

- **Interpretation:** the ratio r implies a correction of the SRS estimator \hat{y} , using the discrepancy between two quantities:
 - ▷ X , the known population total and
 - ▷ \hat{x} , the unbiased estimate

- The same is true for the mean

$$\bar{y}_{b.m.} = \bar{y} + r(\bar{X} - \bar{x})$$

- Note that this relationship is related to the regression relationship at population level:

$$Y_I = 0 + R X_I + \delta_I$$

- These considerations give rise to a wider class of estimators.

12.7.1 Difference and Regression Estimators

Estimator	Expression	Parameters
Benchmark	$\hat{y}_{\text{b.m.}} = \hat{y} + r(X - \hat{x})$	r : ratio
Difference	$\hat{y}_{\text{diff}} = \hat{y} + d(X - \hat{x}) = \alpha N + dX$	d : arbitrary
Regression	$\hat{y}_{\text{reg}} = \hat{y} + \beta(X - \hat{x}) = \alpha N + \beta X$	α : intercept β : slope

- The latter relationship follows from the fact that

$$\begin{aligned}\widehat{y} - \beta \widehat{x} &= \frac{N}{n} \sum_{i=1}^n y_i - \beta \frac{N}{n} \sum_{i=1}^n x_i \\&= \frac{N}{n} \left(\sum_{i=1}^n y_i - \beta x_i \right) \\&= \frac{N}{n} \left(\sum_{i=1}^n \alpha + \beta x_i + \varepsilon_i - \beta x_i \right) \\&= \frac{N}{n} \left(\sum_{i=1}^n \alpha + \varepsilon_i \right) \\&= \frac{N}{n} (n\alpha + 0) \\&= N \cdot \alpha\end{aligned}$$

- The regression estimator for the mean:

$$\bar{y}_{\text{reg}} = \bar{y} + \beta(\bar{X} - \bar{x}) = \alpha + \beta\bar{X}$$

- Variance computations are rather straightforward in these cases, too.

12.7.2 Some Comments

- Benchmarks are, in many instances, relatively easy to find.
- When a single benchmark is used for a series of estimates, then the corrections from unbiased estimators towards ratio estimators will occur in a consistent, comparable fashion.
- In many settings, fixed costs are involved, implying that then regression estimators may be more desirable than benchmark estimators.
- When relationships are non-linear, further extension is needed.

12.8 Sample Size Determination

- We presented a summary for the SRS case on page 169.
- We now merely have to replace the population variances (e.g., S_Y^2) with the benchmark-estimation version (e.g., S^2).
- It is sensible to use S^2 rather than σ^2 in the formulas without replacement.
- A tabular representation:

Situation	Total ($\widehat{y}_{b.m.}$)	Average ($\overline{y}_{b.m.}$)	Ratio (\hat{r})
Without r.	$n = \frac{N^2 \sigma^2}{\sigma_{\widehat{y}_{b.m.}}^2 + N \sigma^2}$	$n = \frac{\sigma^2}{\sigma_{\overline{y}_{b.m.}}^2 + (1/N) \sigma^2}$	$n = \frac{V^2}{V_{\hat{r}}^2 + (1/N) V^2}$
With r.	$n = \frac{N^2 \sigma^2}{\sigma_{\widehat{y}_{b.m.}}^2}$	$n = \frac{\sigma^2}{\sigma_{\overline{y}_{b.m.}}^2}$	$n = \frac{V^2}{V_{\hat{r}}^2}$
$N \rightarrow +\infty$	—	$n = \frac{\sigma^2}{\sigma_{\overline{y}_{b.m.}}^2}$	$n = \frac{V^2}{V_{\hat{r}}^2}$

Part VI

Stratification

Chapter 13

General Concepts and Design

- ▷ Principles of stratification
- ▷ Post-stratification
- ▷ Examples

13.1 Stratification

- We have seen that SRS is unbiased, but can be rather variable: some samples, and hence some estimates, can be extreme:
 - ▷ containing by chance a undue amount of large or small units
 - ▷ containing by chance an unusual fraction of males and females
 - ▷ containing by chance an unusual fraction of Brussels, Flemish, or Walloon residents

- We have already seen two ways of compensating for this:
 - ▷ **Systematic sampling:** by streamlining the sample frame as a monotonic list, 'small' and 'large' units both occur in roughly the right proportions.
 - ▷ **Benchmark estimation:** by correction an SRS estimator, in a second phase, using a more precise piece of information stemming from a larger survey, a census, a register,...
- The auxiliary variables typically used in the above mechanisms (e.g., X in Surveytown), can also be used in a further correcting mechanism:
 - ▷ **Stratification:** *partition the population* in subgroups according to the levels of an auxiliary variable, so that the survey variable is more homogenous within such a subgroup, or **stratum**, than in the population as a whole.

- The effect of stratification is that 'extreme' samples are assigned probability 0, just like in SYS and BENCH.
- It will be shown that, while stratification is intended for increase in precision, it is technically possible for the reverse effect to occur, like in SYS and BENCH.
- The condition for STRAT to work better than SRS is that the correlation between stratifying variable and survey variable should be positive (see further).
- Clearly, such stratifying variables need to be known prior to the sampling process commences.

- Typical candidates for stratification:

- ▷ age
- ▷ sex
- ▷ geographical information
- ▷ size of units
- ▷ socio-economic status
- ▷ educational level
- ▷ occupational status
- ▷ type of activity/occupation

- The number of stratifying variables and the number of categories per stratifying variable should not be too large.

- Suppose, we use all stratifying variables listed above, with the number of categories in parenthesis:

- ▷ age (5)
- ▷ sex (2)
- ▷ geographical information (12)
- ▷ size of units (5)
- ▷ socio-economic status (4)
- ▷ educational level (4)
- ▷ occupational status (4)
- ▷ type of activity/occupation (5)

- Then, the number of strata is

$$H = 5 \times 2 \times 12 \times 5 \times 4 \times 4 \times 4 \times 5 = 192,000$$

Assuming that an overall sample size of $n = 10,000$ is required, it will be hard to ensure all strata contribute, for example, the same number of units, since we would need

$$n_h = \frac{10,000}{192,000} = 0.0521$$

units per stratum!

- We have clearly **over-stratified**.
- The difference between SRS and stratification diminishes for increasing sample sizes.

13.1.1 Two Reasons for Stratification

Goal 1: to increase precision

- Example: better precision for the Belgian estimator, based upon regional stratification.

Goal 2: to obtain inferences about the strata (as well)

- Example: interest in Brussels, Flemish, and Walloon estimators.
- We will see that these different goals have differential implications for sample size calculations.

13.2 Stratified Samples

13.2.1 Quantities

- As before, we need the following information:
 - ▷ Population \mathcal{P}
 - ▷ Population size N
 - ▷ Sample size n
 - ▷ Whether sampling is done with or without replacement

- In addition, we need:

- ▷ The strata indicators $h = 1, \dots, H$

- ▷ The number of subjects in stratum h : $I = 1, \dots, N_h$, with

$$N = \sum_{h=1}^H N_h$$

- ▷ Y_{hI} is the survey variable value for subject I in stratum h

- ▷ This defines the subpopulations, or **population strata**, \mathcal{P}_h

- ▷ The way the sample of n units is **allocated** to the strata: n_h , with

$$n = \sum_{h=1}^H n_h$$

- ▷ We can calculate the stratum-specific sample fraction:

$$f_h = \frac{n_h}{N_h}$$

- ▷ One sometimes writes the samples sizes as a vector:

$$n = (n_1, n_2, \dots, n_h, \dots, n_H)$$

- * For example, $n = (4, 3)$ implies there are two strata, 4 units are selected from the first stratum, and 3 units are selected from the second stratum.

13.2.2 Number of Samples

- Calculate the number of samples that can be obtained within a stratum: S_h
- The number of stratified samples that can be taken from the entire population then simply is

$$S = \prod_{h=1}^H S_h = S_1 \times \cdots S_H$$

13.2.3 Example of Stratification

- Consider a list of school children.
- Stratify according to:
 - ▷ school district
 - ▷ study year
- Take a sample of 10% out of every stratum h , formed as a school district by study year combination.
- We then have a 10% sample, not only overall, but within every stratum.

13.3 Post-stratification

- Stratification can be done at two levels:
 - ▷ **design stage**: stratify when selecting the sample
 - ▷ **analysis stage**: construct stratified estimators, by:
 - * first: constructing estimators for each stratum
 - * second: combining these in an estimator for the entire population
- Whether or not the method is applied at either one of the stages can be used for characterizing a method:

		At design stage	
		No	Yes
At analysis stage	No	SRS	Problematic
	Yes	Post-stratification	Stratification

- **Post-stratification** is defined as the stratified analysis of a sample that was taken in an un-stratified way.

(Slightly more general: Post-stratification is defined as an analysis that used strictly more stratifying variables than at design stage.)

- The advantage over SRS is, typically, increase of precision, **but not as much as full stratification**.
- The intuitive reasons is that:
 - ▷ **yes:** by constructing stratum-specific estimators that are then combined, important sources of variability are controlled.
 - ▷ **no:** the sample size per stratum is not fixed by design, unlike in full stratification; hence, the variability in the sample size contributes to the overall variability.

- The **problematic** case:
 - ▷ does not take the design into account at analysis stage;
 - ▷ this is problematic for surveys
 - ▷ this is problematic for retrospective (case-control) studies
 - ▷ this is fine for randomized studies

13.4 Example: Artificial Population

- Similar to the illustration in Section 9.1.4, consider two stratifications of the artificial population:

$$\mathcal{P}_{s1} = (1\ 2 \mid 3\ 4)$$

$$\mathcal{P}_{s2} = (1\ 4 \mid 2\ 3)$$

- In both cases, 4 samples of size $n = (1, 1)$ are possible.

- The sampling mechanisms then are:

P_s				
Stratified				
s	Sample	SRS	\mathcal{P}_{s1}	\mathcal{P}_{s2}
1	{1,2}	1/6	0	1/4
2	{1,3}	1/6	1/4	1/4
3	{1,4}	1/6	1/4	0
4	{2,3}	1/6	1/4	0
5	{2,4}	1/6	1/4	1/4
6	{3,4}	1/6	0	1/4

- Stratification \mathcal{P}_{s1} is **good** in the sense that it prohibits the most extreme, outer samples.
- Stratification \mathcal{P}_{s2} is **bad** in the sense that it prohibits the most moderate, middle samples.
- The expectations for the average:

$$\mathcal{P}_{s1} : E(\bar{y}) = \frac{1}{4} \cdot [2.0 + 2.5 + 2.5 + 3.0] = 2.5$$

$$\mathcal{P}_{s2} : E(\bar{y}) = \frac{1}{4} \cdot [1.5 + 2.0 + 3.0 + 3.5] = 2.5$$

- Hence, both stratifications produce unbiased estimators.

- The variances for SRS (without), SRS (with), SYS, and STRAT:

$$\begin{aligned}\text{SRS (without)} : \sigma_y^2 &= \frac{(1.5 - 2.5)^2 + (2.0 - 2.5)^2 + (2.5 - 2.5)^2 + (2.5 - 2.5)^2 + (3.0 - 2.5)^2 + (3.5 - 2.5)^2}{6} \\ &= \frac{2.5}{6} = 0.4167\end{aligned}$$

$$\begin{aligned}\text{SRS (with)} : \frac{2}{16} \cdot [(1.5 - 2.5)^2 + (2.0 - 2.5)^2 + (2.5 - 2.5)^2 + (2.5 - 2.5)^2 + (3.0 - 2.5)^2 + (3.5 - 2.5)^2] \\ + \frac{1}{16} \cdot [(1.0 - 2.5)^2 + (2.0 - 2.5)^2 + (3.0 - 2.5)^2 + (4.0 - 2.5)^2] = \frac{10.0}{16} = 0.6250\end{aligned}$$

$$\mathcal{L}_1 : \sigma_y^2 = \frac{(2.0 - 2.5)^2 + (3.0 - 2.5)^2}{2} = \frac{0.5}{2} = 0.25$$

$$\mathcal{L}_2 : \sigma_y^2 = \frac{(1.5 - 2.5)^2 + (3.5 - 2.5)^2}{2} = \frac{2.0}{2} = 1.00$$

$$\mathcal{L}_3 : \sigma_y^2 = \frac{(2.5 - 2.5)^2 + (2.5 - 2.5)^2}{2} = \frac{0.0}{2} = 0.00$$

$$\mathcal{P}_{s1} : \sigma_y^2 = \frac{(2.0 - 2.5)^2 + (2.5 - 2.5)^2 + (2.5 - 2.5)^2 + (3.0 - 2.5)^2}{4} = \frac{0.5}{4} = 0.125$$

$$\mathcal{P}_{s2} : \sigma_y^2 = \frac{(1.5 - 2.5)^2 + (2.0 - 2.5)^2 + (3.0 - 2.5)^2 + (3.5 - 2.5)^2}{4} = \frac{2.5}{4} = 0.625$$

- Recall: some **lists** decrease the variance, while others increase the variance.
- Equally: some **stratifications** decrease the variance, while others increase the variance.

13.5 Example: Surveytown

- In Section 9.3, two lists were considered:

$$\mathcal{L}_X = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)$$

$$\mathcal{L}_Z = (2\ 6\ 5\ 3\ 7\ 1\ 4\ 8)$$

based on, respectively,

- ▷ X_I : number of building lots in block I
- ▷ Z_I : number of newspapers delivered in block I

- In the same spirit, we can stratify the population in two ways:

$$\mathcal{P}_{sX} = (1 \ 2 \ 3 \ 4 \mid 5 \ 6 \ 7 \ 8)$$

$$\mathcal{P}_{sZ} = (2 \ 6 \ 5 \ 3 \mid 7 \ 1 \ 4 \ 8)$$

- Selecting, as usual, samples of size $n = 2$, implies that we have $4 \times 4 = 16$ possible samples in each case

$$\begin{aligned} \mathcal{S}_{sX} = \{ & \{1, 5\}, \{1, 6\}, \{1, 7\}, \{1, 8\}, \\ & \{2, 5\}, \{2, 6\}, \{2, 7\}, \{2, 8\}, \\ & \{3, 5\}, \{3, 6\}, \{3, 7\}, \{3, 8\}, \\ & \{4, 5\}, \{4, 6\}, \{4, 7\}, \{4, 8\} \} \end{aligned}$$

$$\mathcal{S}_{sZ} = \left\{ \begin{array}{l} \{2, 1\}, \{2, 4\}, \{2, 7\}, \{2, 8\}, \\ \{3, 1\}, \{3, 4\}, \{3, 7\}, \{3, 8\}, \\ \{5, 1\}, \{5, 4\}, \{5, 7\}, \{5, 8\}, \\ \{6, 1\}, \{6, 4\}, \{6, 7\}, \{6, 8\} \end{array} \right\}$$

- Let us enumerate the samples:

s	Sample	P_s					\hat{y}_s				
		SRS	Systematic		Stratified		SRS	Systematic		Stratified	
			\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{sX}	\mathcal{P}_{sZ}		\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{sX}	\mathcal{P}_{sZ}
1	{1,2}	1/28	0	0	0	1/16	12				12
2	{1,3}	1/28	0	0	0	1/16	16				16
3	{1,4}	1/28	0	0	0	0	20				
4	{1,5}	1/28	1/4	0	1/16	1/16	24	24		24	24
5	{1,6}	1/28	0	1/4	1/16	1/16	28		28	28	28
6	{1,7}	1/28	0	0	1/16	0	32			32	
7	{1,8}	1/28	0	0	1/16	0	36			36	
8	{2,3}	1/28	0	0	0	0	20				
9	{2,4}	1/28	0	0	0	1/16	24				24
10	{2,5}	1/28	0	0	1/16	0	28			28	
11	{2,6}	1/28	1/4	0	1/16	0	32	32		32	
12	{2,7}	1/28	0	1/4	1/16	1/16	36		36	36	36
13	{2,8}	1/28	0	0	1/16	1/16	40			40	40
14	{3,4}	1/28	0	0	0	1/16	28				28
15	{3,5}	1/28	0	0	1/16	0	32			32	
16	{3,6}	1/28	0	0	1/16	0	36			36	

<i>s</i>	Sample	P_s					\hat{y}_s				
		SRS	Systematic		Stratified		SRS	Systematic		Stratified	
			\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{sX}	\mathcal{P}_{sZ}		\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{sX}	\mathcal{P}_{sZ}
17	{3,7}	1/28	1/4	0	1/16	1/16	40	40		40	40
18	{3,8}	1/28	0	1/4	1/16	1/16	44		44	44	44
19	{4,5}	1/28	0	1/4	1/16	1/16	36		36	36	36
20	{4,6}	1/28	0	0	1/16	1/16	40			40	40
21	{4,7}	1/28	0	0	1/16	0	44			44	
22	{4,8}	1/28	1/4	0	1/16	0	48	48		48	
23	{5,6}	1/28	0	0	0	0	44				
24	{5,7}	1/28	0	0	0	1/16	48				48
25	{5,8}	1/28	0	0	0	1/16	52				52
26	{6,7}	1/28	0	0	0	1/16	52				52
27	{6,8}	1/28	0	0	0	1/16	56				56
28	{7,8}	1/28	0	0	0	0	60				
Expectation							36	36	36	36	36
Variance							144	80	32	40	160
Standard error							12.00	8.94	2.83	6.32	12.65

- The expectations for the total:

$$\mathcal{P}_{sX} : E(\bar{y}) = \frac{1}{16} \cdot [24 + 28 + \cdots + 44 + 48] = \frac{576}{16} = 36$$

$$\mathcal{P}_{sZ} : E(\bar{y}) = \frac{1}{16} \cdot [12 + 16 + \cdots + 52 + 56] = \frac{576}{16} = 36$$

- Hence, both lists produce unbiased estimators.

- The variances:

$$\mathcal{P}_{cX} : \sigma_{\bar{y}}^2 = \frac{(24 - 36)^2 + (28 - 36)^2 + \cdots + (44 - 36)^2 + (48 - 36)^2}{16} = \frac{640}{16} = 40$$

$$\mathcal{P}_{cZ} : \sigma_{\bar{y}}^2 = \frac{(12 - 36)^2 + (16 - 36)^2 + \cdots + (52 - 36)^2 + (56 - 36)^2}{16} = \frac{2560}{16} = 160$$

- Recall that the variance under SRS was 144.
- \mathcal{P}_{sX} decreases variability dramatically, while \mathcal{P}_{sZ} increases variability, relative to SRS.

- This underscores that homogeneous strata have a beneficial impact, while heterogeneous strata have a detrimental effect.

Chapter 14

Analysis

- ▷ Estimators
- ▷ Variances
- ▷ Examples

14.1 Population Quantities and Estimators

- The general principle for estimation is:
 - ▷ Construct an estimator for each stratum separately.
 - ▷ Combine the stratum-specific estimators to a population-level estimator.
- Let Y take value Y_{hI} for unit I in stratum h .
- Let Y_h be the total within stratum h .
- Let \bar{Y}_h be the average within stratum h .

14.1.1 The Population Total

- The population total simply is:

$$Y = \sum_{h=1}^H Y_h = \sum_{h=1}^H \sum_{I=1}^{N_h} Y_{hI}$$

- It follows as the **unweighted sum** of the stratum-specific totals.
- It follows as the double sum of the population units.

14.1.2 The Population Average

- The average within stratum h :

$$\bar{Y}_h = \frac{1}{N_h} Y_h = \frac{1}{N_h} \sum_{I=1}^{N_h} Y_{hI}$$

- The derivation of the population average needs a bit of algebra:

$$\begin{aligned} \bar{Y} &= \frac{1}{N} Y & \Rightarrow \bar{Y} &= \frac{1}{N} \sum_{h=1}^H \frac{N_h}{N_h} \sum_{I=1}^{N_h} Y_{hI} \\ \Rightarrow \bar{Y} &= \frac{1}{N} \sum_{h=1}^H Y_h & \Rightarrow \bar{Y} &= \sum_{h=1}^H \frac{N_h}{N} \left(\frac{1}{N_h} \sum_{I=1}^{N_h} Y_{hI} \right) \\ \Rightarrow \bar{Y} &= \frac{1}{N} \sum_{h=1}^H \sum_{I=1}^{N_h} Y_{hI} & \Rightarrow \bar{Y} &= \sum_{h=1}^H \frac{N_h}{N} \bar{Y}_h \end{aligned}$$

$$\Rightarrow \boxed{\bar{Y} = \sum_{h=1}^H W_h \bar{Y}_h}$$

- The population average follows as the **weighted sum** of the stratum-specific averages.
- The weights

$$W_h = \frac{N_h}{N}, \quad \sum_{h=1}^H W_h = 1$$

are proportional to the population within a stratum.

- We can rewrite the average as:

$$\bar{Y} = \frac{\sum_{h=1}^H W_h \bar{Y}_h}{\sum_{h=1}^H W_h}$$

14.1.3 Estimators

- The total of the sub-sample within stratum h :

$$y_h = \sum_{i=1}^{n_h} y_{hi}$$

- Estimator for the stratum-specific total:

$$\hat{y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{N_h}{n_h} y_h$$

- Estimator for the population total:

$$\hat{y} = \sum_{h=1}^H \hat{y}_h$$

▷ It is the unweighted average of the stratum-specific totals.

- Estimator for the stratum-specific average:

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} = \frac{1}{n_h} y_h$$

- Estimator for the population average:

$$\bar{y} = \frac{1}{N} \hat{y} = \frac{1}{N} \sum_{h=1}^H \hat{y}_h = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h$$

- ▷ The estimator for the population average is a **weighted sum** of the stratum-specific averages.

14.2 Ratios

- A stratum-specific ratio is given by

$$R_h = \frac{Y_h}{X_h} = \frac{\overline{Y_h}}{\overline{X_h}}$$

- The link with the population-level ratio is not immediately straightforward.
- Let us also consider estimators.
 - ▷ The combination with benchmark estimation will be discussed.
 - ▷ Two different options will be considered.

14.2.1 Ratios per Stratum

- The estimators are:

$$\hat{r}_h = \frac{\hat{y}_h}{\hat{x}_h} = \frac{\bar{y}_h}{\bar{x}_h}$$

$$\hat{r} = \frac{\hat{y}}{\hat{x}} = \frac{\bar{y}}{\bar{x}} = \frac{\sum_{h=1}^H \hat{y}_h}{\sum_{h=1}^H \hat{x}_h}$$

14.2.2 Stratum-by-stratum Benchmark Estimator

- Consider the stratum-specific ratios r_h and construct the stratum-specific benchmark estimator for the total:

$$\hat{y}_{\text{b.m.},h} = \hat{r}_h X_h$$

- Combine these to produce the overall benchmark estimator for the total:

$$\hat{y}_{\text{b.m.}} = \sum_{h=1}^H \hat{y}_{\text{b.m.},h} = \sum_{h=1}^H \hat{r}_h X_h$$

14.2.3 Across-stratum Benchmark Estimator

- First, construct the overall ratio \hat{r} .
- Immediately produce the overall benchmark estimator for the total:

$$\hat{y}_{\text{b.m.}} = rX$$

- The stratum-by stratum benchmark estimator \neq the across-stratum benchmark estimator.

14.2.4 Some Comments

- It appears the stratum-specific benchmark estimator uses the information more subtly, and therefore is to be preferred.
- This is not always the case.
- Thus, prefer the across-stratum benchmark estimator when one or both of the following conditions apply:
 - ▷ The stratum-specific sample sizes n_h are very variable and/or very small.
 - ▷ The benchmark X is known at population level but not (or not precise enough) at stratum level (X_h).

14.3 Variance

- We now need three steps:
 - ▷ Derive the population variance per stratum.
 - ▷ Produce a corresponding estimator.
 - ▷ Use these in estimators for the stratum-specific variances.
 - ▷ Combine the results in expressions for the overall population.

Stratum-Level Quantities

Quantity	Calculated	Estimated
Pop. var.	$S_{hY}^2 = \frac{1}{N_h - 1} \sum_{I=1}^{N_h} (Y_{hI} - \bar{Y}_h)^2$	$\hat{s}_{hy}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$
Total	$\sigma_{\hat{y}_h}^2 = \frac{N_h^2}{n_h} (1 - f_h) S_{hY}^2$	$\hat{\sigma}_{\hat{y}_h}^2 = \frac{N_h^2}{n_h} (1 - f_h) \hat{s}_{hy}^2$
Average	$\sigma_{\bar{y}_h}^2 = \frac{1}{n_h} (1 - f_h) S_{hY}^2$	$\hat{\sigma}_{\bar{y}_h}^2 = \frac{1}{n_h} (1 - f_h) \hat{s}_{hy}^2$

Population-Level Quantities

Quantity

Calculated

Estimated

Population variance

$$S_Y^2 = \sum_{h=1}^H S_{hY}^2$$

$$\hat{s}_y^2 = \sum_{h=1}^H \hat{s}_{hy}^2$$

Total

$$\sigma_{\hat{y}}^2 = \sum_{h=1}^H \sigma_{\hat{y}_h}^2$$

$$\hat{\sigma}_{\hat{y}}^2 = \sum_{h=1}^H \hat{\sigma}_{\hat{y}_h}^2$$

Average

$$\sigma_{\hat{y}}^2 = \sum_{h=1}^H w_h^2 \sigma_{\hat{y}_h}^2$$

$$\hat{\sigma}_{\hat{y}}^2 = \sum_{h=1}^H w_h^2 \hat{\sigma}_{\hat{y}_h}^2$$

- For the estimators combining benchmark estimation with stratification, the following expressions need to be used:

Estimator	Calculated	Estimated
Stratum-by -stratum	$S_h^2 = \frac{1}{N_h - 1} \sum_{I=1}^{N_h} (Y_{hI} - R_h X_{hI})^2$ $\simeq S_{hY}^2 - 2R_h S_{hYX} + R_h^2 S_{hX}^2$	$\hat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \hat{r}_h x_{hi})^2$ $\simeq \hat{S}_{hy}^2 - 2\hat{r}_h \hat{S}_{hyx} + \hat{r}_h^2 \hat{S}_{hx}^2$
Across -stratum	$S_h^2 = \frac{1}{N_h - 1} \sum_{I=1}^{N_h} (Y_{hI} - R X_{hI})^2$ $\simeq S_{hY}^2 - 2R S_{hYX} + R^2 S_{hX}^2$	$\hat{S}_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \hat{r} x_{hi})^2$ $\simeq \hat{S}_{hy}^2 - 2\hat{r} \hat{S}_{hyx} + \hat{r}^2 \hat{S}_{hx}^2$

- There is a problem with the latter estimator:
 - ▷ All strata have r in common.
 - ▷ Hence, the strata-specific estimators are not entirely independent of one another.
 - ▷ This results in an (often small) underestimation of the variance (i.e., false precision).

14.4 Example: Artificial Population

- In Section 10.3.2, the intra-cluster (intraclass) correlations were calculated for SRS (without and with replacement), and SYS (lists \mathcal{L}_1 , \mathcal{L}_2 , and \mathcal{L}_3).
- Using similar programs, but now for the stratified sampling mechanisms of Section 13.4, we can expand the table:

Method	Variance	ρ	Relationship
SRS (without)	0.4167	-0.33	$\frac{0.4167}{1-0.33 \times (2-1)} = 0.6250$
SRS (with)	0.6250	0.00	$\frac{0.6250}{1+0.00 \times (2-1)} = 0.6250$
SYS(\mathcal{L}_1)	0.2500	-0.60	$\frac{0.2500}{1-0.60 \times (2-1)} = 0.6250$
SYS(\mathcal{L}_2)	1.0000	0.60	$\frac{1.0000}{1+0.60 \times (2-1)} = 0.6250$
SYS(\mathcal{L}_3)	0.0000	-1.00	undetermined
STRAT(\mathcal{P}_{s1})	0.1250	-0.80	$\frac{0.1250}{1-0.80 \times (2-1)} = 0.6250$
STRAT(\mathcal{P}_{s2})	0.6250	0.00	$\frac{0.6250}{1+0.00 \times (2-1)} = 0.6250$

- Note that the smallest variance is obtained, apart for pathological list \mathcal{L}_3 , for the good stratification.
- Bad stratification annihilates the beneficial effect of sampling without replacement, and effectively returns to the variance of SRS with replacement.

- This underscores that stratification, even though typically used for its beneficial impact on precision, can effectively decrease precision.
- This can be illustrated by partitioning the variance.
- To this effect, consider a classical ANOVA decomposition.
- First, construct a simple dataset, as follows:

Obs	stratum1	stratum2	y
1	1	1	1
2	1	2	2
3	2	2	3
4	2	1	4

- We can now construct ANOVA decompositions using **PROC GLM**:

```
proc glm data=m.artif04;  
title 'GLM - ANOVA table - Non-stratified';  
model y = ;  
run;
```

```
proc glm data=m.artif04;  
title 'GLM - ANOVA table - Good stratification';  
class stratum1;  
model y = stratum1;  
run;
```

- The model for the bad stratification is evidently completely analogous.

- Output for the non-stratified case is:

GLM - ANOVA table - Non-stratified

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	25.00000000	25.00000000	15.00	0.0305
Error	3	5.00000000	1.66666667		
Uncorrected Total	4	30.00000000			

- For the good stratification, we obtain:

GLM - ANOVA table - Good stratification

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	4.00000000	4.00000000	8.00	0.1056
Error	2	1.00000000	0.50000000		
Corrected Total	3	5.00000000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
stratum1	1	4.00000000	4.00000000	8.00	0.1056

- For the bad stratification:

GLM - ANOVA table - Bad stratification

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00000000	0.00000000	0.00	1.0000
Error	2	5.00000000	2.50000000		
Corrected Total	3	5.00000000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
stratum2	1	0	0	0.00	1.0000

- It is intriguing that in the second case, no variability is attributed to the stratification variable, while the reverse is true in the first case.

- We can look at the same problem in a different way, by considering the linear mixed model the systematic sampling part:

$$Y_{IJ} = \mu + b_I + \varepsilon_{IJ}$$

- ▷ Y_{IJ} is the population quantity for subject J in **stratum** I
- ▷ μ is the overall mean (population average)
- ▷ $\mu + b_I$ is the stratum-specific average:

$$b_I \sim N(0, \tau^2)$$

- ▷ ε_{IJ} is an individual-level deviation:

$$\varepsilon_{IJ} \sim N(0, \sigma^2)$$

- The sources of variability in the ANOVA table correspond to τ^2 and σ^2 and can be estimated using **PROC MIXED**.

- ▷ The program for the non-stratified case is, in fact, nothing but a linear regression:

```
proc mixed data=m.artif04 method=ml;  
title 'Artificial Population - Non-stratified';  
model y = / solution;  
run;
```

- ▷ The variance component is:

Cov Parm	Estimate
Residual	1.2500

- ▷ The corresponding program for the first stratification is:

```
proc mixed data=m.artif04 method=ml;  
title 'Artificial Population - Good stratification';  
class stratum1;  
model y = / solution;  
random stratum1;  
run;
```

- ▷ We obtain two variance components:

Cov Parm	Estimate
stratum1	0.7500
Residual	0.5000

- ▷ The sum of the variances is the same as in the non-stratified case, as it should, but a part of the variability is **taken out** by the stratification.

This same phenomenon lead to a negative within-sample correlation, as seen above.

- ▷ The program for the bad stratification is, of course analogous, and produces:

Cov Parm	Estimate
stratum2	0
Residual	1.2500

- ▷ Like in the ANOVA table, we see that apparently no variability is associated to stratification. Yet, the variance actually changed, when the estimator was studied.

In fact, it increased, and this is possible only by assigning a **negative** component of variability to the second stratum.

- ▷ We can allow for this by adding the 'nobound' option to the program:

```
proc mixed data=m.artif04 method=ml nobound;  
title 'Artificial Population - Bad stratification - Nobound';  
class stratum2;  
model y = / solution;  
random stratum2;  
run;
```

- ▷ The result changes to:

Cov Parm	Estimate
stratum2	-1.2500
Residual	2.5000

- ▷ Indeed, while the total variability is still left unchanged, the stratification is now clearly seen to be responsible for an **increase** in error variance, since it adds to the variability, rather than taking away from it.

14.5 Example: Surveytown

- Also for this example, we can calculate the within-cluster (actually now, within-strata) correlation.
- Using the SAS procedure MIXED, the intra-cluster correlation can be calculated, based on the datasets:

Listing Surveytown - Strat. based on X

Obs	sample	y
1	1	1
2	1	5
3	2	1
4	2	6
5	3	1
6	3	7
...		
31	16	4
32	16	8

Listing Surveytown - Strat. based on Z

Obs	sample	y
1	1	2
2	1	1
3	2	2
4	2	4
5	3	2
6	3	7
...		
31	16	6
32	16	8

- The correlations are:

$$\rho_{\mathcal{P}_s X} = -0.7619$$

$$\rho_{\mathcal{P}_s Z} = -0.0476$$

- In Part IV, we obtained relationships between variances, which we can now extend:

$$\begin{aligned} \frac{\sigma_{\hat{y}, \text{SRS}(\text{with})}^2}{1 + \rho_{\text{SRS}(\text{with})}(n-1)} &= \frac{\sigma_{\hat{y}, \text{SRS}(\text{without})}^2}{1 + \rho_{\text{SRS}(\text{without})}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{L}_X}^2}{1 + \rho_{\mathcal{L}_X}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{L}_Z}^2}{1 + \rho_{\mathcal{L}_Z}(n-1)} \\ &= \frac{\sigma_{\hat{y}, \mathcal{P}_s X}^2}{1 + \rho_{\mathcal{P}_s X}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{P}_s Z}^2}{1 + \rho_{\mathcal{P}_s Z}(n-1)} \\ &= \frac{168}{1 + 0.0000 \times (2-1)} = \frac{144}{1 - 0.1429 \times (2-1)} = \frac{80}{1 - 0.5238 \times (2-1)} = \frac{32}{1 - 0.8095 \times (2-1)} \\ &= \frac{40}{1 - 0.7619 \times (2-1)} = \frac{160}{1 - 0.0476 \times (2-1)} \end{aligned}$$

- Additionally, we can calculate the variance decomposition, based on both stratifications.
- In analogy with the Artificial Population, construct the dataset:

Obs	stratumx	stratumz	y
1	1	2	1
2	1	1	2
3	1	1	3
4	1	2	4
5	2	1	5
6	2	1	6
7	2	2	7
8	2	2	8

- Likewise, the following program can be used:

```
proc mixed data=m.surveytown04 method=ml nobound;  
title 'Surveytown - Variance decomposition stratification X';  
class stratumx;  
model y = / solution;  
random stratumx;  
run;
```

- Replace 'X' by 'Z' for the second stratification, and simply omit stratum and the **RANDOM** statement for the unstratified case.

- We obtain the following variance decompositions at population level:

Source	No Stratif.	Stratif. X	Stratif. Z
Stratum		3.5833	-1.4167
Residual	5.2500	1.6667	6.6667
Total	5.2500	5.2500	5.2500
Within-stratum correlation	0.0000	0.6825	-0.2698

- Note that the within-stratum correlation at population level is not the same concept as the within-stratified-samples correlation.

- ▷ The correlation within the population strata is positive for \mathcal{P}_{sX}
 - ⇒ the strata are **homogeneous**
 - ⇒ samples taken have a unit from the first stratum and one from the second stratum and hence are **heterogeneous**
 - ⇒ the correlation between units within a sample **decreases** relative to SRS, with **beneficial** impact on the estimator.

- ▷ The correlation within the population strata is negative for \mathcal{P}_{sZ}
 - ⇒ the strata are **heterogeneous**
 - ⇒ samples taken have a unit from the first stratum and one from the second stratum and hence are **homogeneous**
 - ⇒ the correlation between units within a sample **increases** relative to SRS, with **detrimental** impact on the estimator.

14.5.1 Combining Benchmark Estimation With Stratification

- Let us compare 3 ways of applying benchmark estimation:
 - ▷ Not combined with stratification, as in Section 12.2.
 - ▷ Combined with stratification in the across-stratum fashion.
 - ▷ Combined with stratification in the stratum-by-stratum fashion.

Measure	Stratification		
	No	Across	S-by-s
Population estimand Y	36	36	36
Expectation $E(\hat{y}_{b=X})$	36.150	36.137	38.357
Bias	0.150	0.137	2.357
Range: lowest estimate	33.333	34.375	35.867
Range: highest estimate	40.000	38.889	43.000
Variance $\text{Var}(\hat{y}_{b=X})$	1.850	1.491	5.216
Mean square error	1.874	1.510	10.773
Standard error	1.360	1.221	2.284
Root mean square error	1.369	1.229	3.282

- Note that here the stratifying variable and the benchmark variable are one and the same.
- We should not draw too broad a conclusion from it.
- Nevertheless, stratum-by-stratum benchmark estimation performs **worse** than ordinary benchmark estimation.

14.6 Example: The Belgian Health Interview Survey

- Taking stratification into account, the means are recomputed for

- ▷ LNBMI
- ▷ LNVOEG
- ▷ GHQ12
- ▷ SGP

- The following program can be used:

```
proc surveymeans data=m.bmi_voeg mean stderr;  
title 'stratified means - infinite population for Belgium and regions';  
where (regionch^='');  
domain regionch;  
strata province;  
var lnbmi lnvoeg ghq12 sgp;  
run;
```

- We include the stratification design aspect by way of the **STRATA** statement.
- The output takes the usual form, with now all design aspects listed:

```
stratified means - infinite population for Belgium and regions
The SURVEYMEANS Procedure
```

Data Summary

```
Number of Strata          12
Number of Observations    8560
```

Statistics

Variable	Mean	Std Error of Mean

LNBMI	3.187218	0.001840
LNVOEG	1.702951	0.008801
GHQ12	1.661956	0.029452
SGP	0.903540	0.003116

Domain Analysis: REGIONCH

REGIONCH	Variable	Mean	Std Error of Mean
Brussels	LNBMI	3.175877	0.003373
	LNVOEG	1.809748	0.016206
	GHQ12	1.864301	0.056939
	SGP	0.805632	0.007827
Flanders	LNBMI	3.182477	0.002989
	LNVOEG	1.516352	0.015207
	GHQ12	1.385857	0.046211
	SGP	0.952285	0.003902
Walloonia	LNBMI	3.201530	0.003217
	LNVOEG	1.801107	0.014427
	GHQ12	1.772148	0.050823
	SGP	0.938646	0.004366

- We summarize the results, compare them to SRS (and foreshadow future analyses):

Logarithm of Body Mass Index				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	3.187218(0.001845)	3.175877(0.003372)	3.182477(0.002993)	3.201530(0.003216)
Stratification	3.187218(0.001840)	3.175877(0.003373)	3.182477(0.002989)	3.201530(0.003217)
Clustering	3.187218(0.001999)	3.175877(0.003630)	3.182477(0.003309)	3.201530(0.003429)
Weighting	3.185356(0.002651)	3.171174(0.004578)	3.180865(0.003870)	3.198131(0.004238)
All combined	3.185356(0.003994)	3.171174(0.004844)	3.180865(0.004250)	3.198131(0.004403)

Logarithm of VOG Score				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	1.702951(0.008954)	1.809748(0.016203)	1.516352(0.015201)	1.801107(0.014550)
Stratification	1.702951(0.008801)	1.809748(0.016206)	1.516352(0.015207)	1.801107(0.014427)
Clustering	1.702951(0.010355)	1.809748(0.018073)	1.516352(0.017246)	1.801107(0.016963)
Weighting	1.634690(0.013233)	1.802773(0.021831)	1.511927(0.019155)	1.803178(0.020426)
All combined	1.634690(0.014855)	1.802773(0.023135)	1.511927(0.021409)	1.803178(0.023214)

General Health Questionnaire – 12				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	1.661349(0.029584)	1.862745(0.056894)	1.385381(0.046246)	1.772148(0.051023)
Stratification	1.661956(0.029452)	1.864301(0.056939)	1.385857(0.046211)	1.772148(0.050823)
Clustering	1.661349(0.032824)	1.862745(0.062739)	1.385381(0.052202)	1.772148(0.055780)
Weighting	1.626201(0.044556)	1.924647(0.076313)	1.445957(0.061910)	1.858503(0.078566)
All combined	1.626781(0.048875)	1.924647(0.080508)	1.446286(0.068931)	1.858503(0.084047)

Stable General Practitioner (0/1)				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	0.903540(0.003196)	0.805632(0.007826)	0.952285(0.003908)	0.938646(0.004382)
Stratification	0.903540(0.003116)	0.805632(0.007827)	0.952285(0.003902)	0.938646(0.004366)
Clustering	0.903540(0.003963)	0.805632(0.009766)	0.952285(0.004709)	0.938646(0.005284)
Weighting	0.932702(0.003498)	0.782448(0.011563)	0.954757(0.004722)	0.943191(0.005417)
All combined	0.932702(0.003994)	0.782448(0.013836)	0.954757(0.005379)	0.943191(0.006159)

- We can make the following observations, when comparing stratification to SRS:
 - ▷ The impact on point estimates is minor.
 - ▷ The impact on standard errors is minor, and goes in both directions, with the dominant direction a slight reduction of standard error.

Chapter 15

Sample Size Determination and Allocation

- ▷ General principles
- ▷ Proportional allocation
- ▷ Optimal allocation
- ▷ Cost optimal allocation
- ▷ Compromise allocation

15.1 General Principles

- In a stratified setting, there are two aspects related to sample size:
 - ▷ **Sample size determination:** calculation of the overall sample size n .
 - ▷ **Sample size allocation:** the split of the sample size n over the strata: (n_1, \dots, n_H) .
- Two distinct precision requirements can be put forward:
 - ▷ **Precision at the population level:** the sample sizes n_h are determined so as to reach a certain level of precision for the entire population.
 - ▷ **Precision at stratum level:** the sample sizes n_h are determined to reach a certain level of precision for the strata separately.

▷ These different requirements will produce different results.

▷ **Compromise allocation:** refers to the situation where both the population and the stratum level are of importance.

A compromise between two different allocations is then aimed for.

- A taxonomy of allocations is considered, based on which information is taken into account:

Types of allocation

Stratum-specific

Allocation	size N_h	Var. S_h	cost C_h
Proportional allocation	yes	no	no
Optimal allocation	yes	yes	no
Cost-optimal allocation	yes	yes	yes

- Optimal allocation will differ from proportional allocation when the variability of the survey variable differs a lot between strata.

In practice, for many variables this is not the case.

- Cost-optimal allocation starts from a differential cost between the strata:

$$C = C_0 + \sum_{h=1}^H n_h C_h$$

- ▷ C_0 : fixed costs (overhead)
 - ▷ C_h : average variable cost per unit in stratum h
 - ▷ C : total cost
- Cost-optimal allocation will differ a lot from optimal allocation when the variable cost is different from stratum to stratum.
This may happen, for example, if strata are regions, with some very rural, others very urbanized.

15.2 Sample Allocation

- Let us illustrate the calculations for the case of optimal allocation, when focus is on the entire population.
- Optimal allocation is reached for

$$f_h = \frac{n_h}{N_h} \propto S_h$$

and hence

$$n_h \propto N_h S_h$$

- Requiring that the n_h sum to a pre-fixed n , turns the proportionality result in an equality:

$$n_h = n \frac{N_h S_h}{\sum_h N_h S_h}$$

- These results imply that we have more units
 - ▷ in larger strata
 - ▷ in strata with higher variability

- An overview of all proportionalities:

Proportionalities

Focus on

Allocation	population	strata	compromise
Proportional	$n_h \propto N_h$	$n_h \propto 1$	$n_h \propto N_h^k$
Optimal	$n_h \propto N_h \cdot S_{Yh}$	$n_h \propto S_{Yh}$	$n_h \propto N_h^k \cdot S_{Yh}$
Cost-optim.	$n_h \propto N_h \cdot S_{Yh} \cdot \frac{1}{\sqrt{C_h}}$	$n_h \propto S_{Yh} \cdot \frac{1}{\sqrt{C_h}}$	$n_h \propto N_h^k \cdot S_{Yh} \cdot \frac{1}{\sqrt{C_h}}$

- For compromise allocation, one would typically choose $0 \leq k \leq 1$.
- Some special values deserve attention:
 - ▷ $k = 0$ corresponds with focus on the strata
 - ▷ $k = 0.5$ is a common choice
 - ▷ $k = 1$ corresponds with focus on the population
- The corresponding allocations are:

Allocation

Focus on

Allocation	population	strata	compromise
Proportional	$n_h = n \cdot \frac{N_h}{N}$	$n_h = \frac{n}{H}$	$n_h = n \cdot \frac{N_h^k}{\sum_{h=1}^H N_h^k}$
Optimal	$n_h = n \cdot \frac{N_h S_{Yh}}{\sum_{h=1}^H N_h S_{Yh}}$	$n_h = n \cdot \frac{S_{Yh}}{\sum_{h=1}^H S_{Yh}}$	$n_h = n \cdot \frac{N_h^k S_{Yh}}{\sum_{h=1}^H N_h^k S_{Yh}}$
Cost-opt.	$n_h = n \cdot \frac{N_h S_h \left(\frac{1}{\sqrt{C_h}} \right)}{\sum_{h=1}^H N_h S_h \left(\frac{1}{\sqrt{C_h}} \right)}$	$n_h = n \cdot \frac{S_h \left(\frac{1}{\sqrt{C_h}} \right)}{\sum_{h=1}^H S_h \left(\frac{1}{\sqrt{C_h}} \right)}$	$n_h = n \cdot \frac{N_h^k S_h \left(\frac{1}{\sqrt{C_h}} \right)}{\sum_{h=1}^H N_h^k S_h \left(\frac{1}{\sqrt{C_h}} \right)}$

15.2.1 Example: The Belgian Health Interview Survey

- Let us illustrate this for the Belgian Health Interview Survey:
 - ▷ Consider proportional allocation.
 - ▷ Let $n = 10,000$.
 - ▷ For compromise allocation, set $k = 0.5$.

Allocations for Belgian Health Interview Survey

Region	N_h	Focus on		
		population	strata	compromise
Brussels	1,000,000	1000	$3333.33 \simeq \boxed{3000}$	$1929.93 \simeq 2000$
Flanders	6,000,000	6000	$3333.33 \simeq \boxed{3500}$	$4727.34 \simeq 4750$
Wallonia	3,000,000	3000	$3333.33 \simeq \boxed{3000}$	$3342.73 \simeq 3250$

15.3 Sample Size Determination

- Combined with the sample allocation, an allocation method also yields a specific sample size determination expression.
- Let us present these for the **total**:

Sample size for total

Allocation

n

Proportional allocation

$$n = \frac{N^2 S_Y^2}{\sigma_{\hat{y}}^2 + N S_Y^2}$$

Optimal allocation

$$n = \frac{\left(\sum_{h=1}^H N_h S_{Yh} \right)^2}{\sigma_{\hat{y}}^2 + \sum_{h=1}^H N_h S_{Yh}^2}$$

Cost-optimal allocation

$$n = \frac{\left[\sum_{h=1}^H N_h S_{Yh} \sqrt{C_h} \right] \left[\sum_{h=1}^H N_h S_{Yh} \left(\frac{1}{\sqrt{C_h}} \right) \right]}{\sigma_{\hat{y}}^2 + \sum_{h=1}^H N_h S_{Yh}^2}$$

15.3.1 Impact on Variance

- The various allocation methods have differing impacts on the variance of the estimators.
- Let σ_Y^2 be the variance of the survey variable in the population as a whole.
- The variances for the total can then be expressed as on the following page.
 - ▷ The variance **seems to decrease** with a larger number of effects taken into account.
 - ▷ However, we have illustrated, using the Artificial Population and Surveytown, that the variance **can increase** in some cases.
 - ▷ This is because the additional variance components can be negative, as we have demonstrated numerically using the SAS procedure MIXED.

Variance for total estimator

Allocation

$$\sigma_{\hat{y}}^2$$

Simple random sampling

$$\sigma_{\hat{y}}^2 = \frac{N^2}{n}(1 - f)\sigma_Y^2$$

Proportional allocation

$$\sigma_{\hat{y}}^2 = \frac{N^2}{n}(1 - f)[\sigma_Y^2 - \sigma^2(\bar{Y}_h)]$$

Optimal allocation

$$\sigma_{\hat{y}}^2 = \frac{N^2}{n}(1 - f) \{[\sigma_Y^2 - \sigma^2(\bar{Y}_h)] - \sigma^2(S_{hY})\}$$

Cost-optimal allocation

(more complicated)

15.4 Illustration

- Consider a population subdivided into 5 strata.
- All three quantities vary across population strata: N_h , S_h , and C_h .
- Consider all three allocation methods.

Quantity	Stratum					Total
	1	2	3	4	5	
Population						
N_h	2600	1200	750	300	150	5000
S_h	0.730	1.399	1.722	2.311	2.912	
C_h	10.79	11.63	29.72	45.03	62.89	
Proportional allocation						
n_h	1356	626	391	167	78	2608
Cost	14,631	7280	11,621	7070	4905	45,507
Optimal allocation						
n_h	673	596	458	247	150	2124
Cost	7262	6931	13,612	11,122	9434	48,361
Cost-optimal allocation						
n_h	925	789	380	164	89	2347
Cost	9981	9176	11,294	7385	5597	43,433

Part VII

Multi-Stage Sampling and Clustering

Chapter 16

General Concepts and Design

- ▷ The concepts of multi-stage sampling and clustering
- ▷ Various ways of selecting multi-stage samples
- ▷ Examples

16.1 Multi-Stage Sampling and Clustering

- Informal definition of both concepts:
 - ▷ **Multi-stage sampling:** a hierarchy of units is selected:
 - * starting with **primary sampling units** (PSU),
 - * within with **secondary sampling units** (SSU) are sub-selected,
 - * within which **tertiary sampling units** (TSU) are subselected,
 - * etc.
 - ▷ **Clustering:** refers to the fact that several non-independent units (stemming from a 'cluster') are simultaneously selected.

- Examples of multi-stage sampling:

Unit	Schools	Belgian HIS
PSU	school	town
SSU	class	household
TSU	pupil	individual

- Both concepts go hand in hand, but are **not** the same:
 - ▷ **Multi-stage sampling but not clustering:** select only one household in a town, and only one individual within a household.
 - ▷ **Clustering without multi-stage sampling:** select households from a list of households, and then include all household members. Since there is no sub-selection taking place, this is a one-stage procedure, but there clearly is clustering.

- Some levels are included for sampling convenience only, with no direct scientific interests:
 - ▷ schools and classes
 - ▷ towns in HIS
- At least one level is of direct scientific interest: **target sampling units**:
 - ▷ pupils
 - ▷ individuals in HIS, **but also, to some extent, household**
- The latter situation arises when:
 - ▷ some information exists at household level and is objective: number of rooms in the household's residence,...
 - ▷ some information is personal: political preference, religious beliefs,...

- Multi-stage sampling also goes hand in hand with **weighting**, since primary and secondary units may have different sizes and/or sub-units may be selected with unequal probability (see Part **VIII**).
- The rationales for conducting multi-stage sampling:
 - ▷ Multi-stage 'lists' may be easy to work with: while there is no list of all pupils, there is a list of all schools and every school has got a list of its pupils.
 - ▷ To facilitate the fieldwork: when multi-stage sampling leads to clusters, often geographically close, interviewers will be able to organize their work more efficiently.
- When multi-stage sampling induces clustering and **the within-cluster correlation is positive** (cf. systematic sampling) the precision will go down.

This typically is the situation that happens in practice.

It is aimed for to counter-balance the statistical precision loss by a stronger increase in fieldwork efficiency, so that overall there is a gain.

16.2 Multi-Stage Sampling: the Relative Approach

- Assume a two-stage sample of size n is to be taken out of a population of size N .
- The sample fraction then is

$$f = \frac{n}{N}$$

- This can be done by taking
 - ▷ a fraction f_1 of the PSU
 - ▷ a fraction f_2 of the SSU
 - ▷ so that

$$f = f_1 \cdot f_2$$

- Clearly, given f_1 and f , it follows that $f_2 = f/f_1$.
- In other words, two-stage sampling introduces **one degree of freedom** into the design.
- In general, for K -stage sampling:

$$f = f_1 \cdot f_2 \cdot \dots \cdot f_K$$

introducing $K - 1$ degrees of freedom.

- SRS can be seen as a special case: one-stage sampling, introducing $K - 1 = 0$ degrees of freedom.
- Indeed, SRS is fixed by merely specifying f .

16.2.1 Example

- Goal: sample of students in Flemish schools in the Brussels Region.
- There is no complete list, but there **is** available:
 - ▷ a list of all schools
 - ▷ in each school there is a list of students
- Assume the details are:
 - ▷ **PSU**: schools
 - ▷ **SSU**: $N = 20,000$ students
 - ▷ **required sample size**: $n = 2000$ students
 - ▷ **sample fraction**: $f = 0.1$

Selection probabilities					
	Stage 1: f_1		Stage 2: f_2		Total: f
a.	$\frac{1}{1}$	\times	$\frac{1}{10}$	$=$	$\frac{1}{10}$
b.	$\frac{1}{2}$	\times	$\frac{1}{5}$	$=$	$\frac{1}{10}$
c.	$\frac{1}{5}$	\times	$\frac{1}{2}$	$=$	$\frac{1}{10}$
d.	$\frac{1}{10}$	\times	$\frac{1}{1}$	$=$	$\frac{1}{10}$

- When going towards d, the 'cluster size' (students from the same school) increases, with a detrimental impact on precision, but a beneficial impact on fieldwork.
- When going towards a, the cluster size decreases, with a detrimental impact on the fieldwork, but a beneficial impact on the survey's precision.

- All four example mechanisms produce the required sample fraction and hence sample size.
- Each student has the same selection probability of $1/10$.
- Every school has the same probability of being selected.
- The number of students per school is proportional to the school size.
- The latter property can be inconvenient:
 - ▷ The fieldwork burden in large schools may be too heavy.
 - ▷ Fieldwork hard to organize with unequal PSU sizes.
 - ▷ It is hard to fully control the overall sample size.
- For these reasons, the above **relative** selection is often replaced by an **absolute** one.

16.3 Multi-Stage Sampling: the Absolute Approach

- This is commonly referred to as **area probability sampling**, but it applies more generally, for example also to the school example studied above.
- Suppose we would apply the above mechanism to a city:
 - ▷ **PSU**: There are 400 blocks.
 - ▷ **SSU**: There are $N = 20,000$ houses in the blocks taken together.
 - ▷ **Sample size**: $n = 2000$
 - ▷ **Sample fraction**: $f = 0.1$

- This is the same setting as in the table above, and hence these mechanisms could be used.
- The same burden as described above is bestowed on the fieldwork.
- When the size of the schools, blocks, etc. is available, an alternative, **absolute** approach is possible.

16.3.1 Description of Area Probability Sampling

- Assume N , n , and hence f are prespecified.
- Fix the number of SSU taken per PSU: n_c .
- Construct a cumulative list of the number of SSU per PSU.
- Conduct systematic selection within the cumulative list, with jump

$$g = \frac{1}{f} \cdot n_c$$

- For every hit, select n_c SSU from the corresponding PSU.

16.3.2 Example

- Return to the above example with:
 - ▷ **PSU**: There are 400 blocks.
 - ▷ **SSU**: There are $N = 20,000$ houses in the blocks taken together.
 - ▷ **Sample size**: $n = 2000$
 - ▷ **Sample fraction**: $f = 0.1$
 - ▷ **Cluster size**: $n_c = 10$

- The jump is then:

$$g = \frac{1}{f} \cdot n_c = \frac{1}{0.1} \times 10 = 100$$

- Assume the random start, taken between 1 and 100, is $s = 70$.

- We would then select the blocks where encompassing the cumulative numbers 70, 170, 270, 370,...

block	# houses	cumulative	hits
1	43	43	-
2	87	130	70
3	109	239	170
4	27	266	-
5	15	281	270
⋮	⋮	⋮	⋮

- We selected blocks 2, 3, 5.

- Select 10 houses in each of those blocks.
- If the number of houses within each block were correct, then simple random or systematic sampling could be done and the overall selection probability would be preserved:

block	houses	prob.(1)	prob.(2)	prob.(tot)
2	87	$87/100$	$10/87$	$1/10$
3	109	$109/100$	$10/109$	$1/10$
5	15	$15/100$	$10/15$	$1/10$

- **But:** the number of houses is often reported slightly inaccurately.

- This problem can be solved by determining only the selection **rate**:

$$\frac{\text{cluster size}}{\# \text{ houses}} = \frac{10}{87} = \frac{1}{8.7}$$

- For this particular block, 1 per 8.7 houses is to be selected.
- If a block is larger, then more houses are selected
Otherwise, less houses are selected
- What about “empty areas” ?
⇒ Combine with neighboring areas, to enable selection if the area turns out to be non-empty.

16.4 Cluster Samples

- **Population level:**

- ▷ Population \mathcal{P}

- ▷ **PSU:**

- * M : number of PSU

- ▷ **SSU:**

- * N : number of SSU

- * N_I : number of SSU within PSU I (within **cluster** I)

$$N = \sum_{I=1}^M N_I$$

▷ **Survey variable:**

* Y_{IJ} : value for SSU J within cluster I

* Y_I : sum within cluster I

$$Y_I = \sum_{J=1}^{N_I} Y_{IJ}$$

* Y : overall sum

$$Y = \sum_{I=1}^M Y_I = \sum_{I=1}^M \sum_{J=1}^{N_I} Y_{IJ}$$

- **Sample level:**

- ▷ **PSU:**

- * m : number of selected PSU

- ▷ **SSU:**

- * n : number of SSU

- * N_i : number of SSU within selected PSU i

- * n_i : number of SSU **selected from** the selected PSU i

$$n = \sum_{i=1}^m n_i$$

▷ **Survey variable:**

* y_{ij} : value for selected SSU j within selected cluster i

* y_i : sum within selected cluster i over the selected SSU

$$y_i = \sum_{j=1}^{n_i} y_{ijp}$$

* y : sum over all selected SSU within all selected PSU

$$y = \sum_{i=1}^m y_i = \sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij}$$

▷ **Sample fractions:**

* At the first stage:

$$f_1 = \frac{m}{M}$$

* At the second stage:

$$f_{2i} = \frac{n_i}{N_i}$$

* **Simple cluster sampling:**

$$n_i = N_i \quad \Rightarrow \quad f_{2i} = 1$$

All SSU within a selected PSU are included.

* **Self-weighted sampling:**

$$f_{2i} = \frac{n_i}{N_i} = \frac{\bar{n}}{\bar{N}}$$

The number of SSU selected is proportional to the cluster size and hence the second-stage sample fraction is constant.

16.5 Example: Artificial Population

- Consider three ways of clustering the Artificial Population Units:

$$\mathcal{P}_{c1} = \left(\{1, 3\}, \{2, 4\} \right)$$

$$\mathcal{P}_{c2} = \left(\{1, 2\}, \{3, 4\} \right)$$

$$\mathcal{P}_{c3} = \left(\{1, 4\}, \{2, 3\} \right)$$

- In all three cases, only two samples of size $n = 2$ are possible.
- These samples correspond to the lists \mathcal{L}_1 , \mathcal{L}_2 , and \mathcal{L}_3 , respectively.

- The sampling mechanisms are:

			P_s		
			Systematic / Clustering		
			\mathcal{L}_1	\mathcal{L}_2	\mathcal{L}_3
s	Sample	SRS	\mathcal{P}_1	\mathcal{P}_2	\mathcal{P}_3
1	{1,2}	1/6	0	1/2	0
2	{1,3}	1/6	1/2	0	0
3	{1,4}	1/6	0	0	1/2
4	{2,3}	1/6	0	0	1/2
5	{2,4}	1/6	1/2	0	0
6	{3,4}	1/6	0	1/2	0

- Note that this strong connection between the two mechanisms is a by-product of the artificial population being so small.

- All calculations made for the 3 SYS lists, are also applicable to these three ways of clustering.
- When the emphasis is on lists, there are only 6 possible samples, resulting from 3 essentially different lists (there are other lists, but these will produce the same samples).
- This is not true for clustering, for example:

$$\mathcal{P}_{c4} = (\{1\}, \{2, 3, 4\})$$

is a possible way of defining two clusters, giving rise to 2 possible samples of unequal size (see Part VIII).

- We have stated before that:
 - ▷ Lists **typically** increase precision, although the reverse may happen.
 - ▷ Clustering **typically** decreases precision, although the reverse may happen.
- But now, both mechanisms produce the same 3 situations, how can this be reconciled?
 - ▷ The **natural** list choice is \mathcal{L}_1 : units are ordered monotonically.
 - ▷ The **natural** clustering choice is \mathcal{P}_{c2} : clusters contain units that are more similar.
- Recall that all three lists are unbiased; hence, the same holds for all three ways of clustering.

- The variances for SRS (without), SRS (with), SYS, STRAT, and CLUST:

$$\begin{aligned}\text{SRS (without)} : \sigma_y^2 &= \frac{(1.5 - 2.5)^2 + (2.0 - 2.5)^2 + (2.5 - 2.5)^2 + (2.5 - 2.5)^2 + (3.0 - 2.5)^2 + (3.5 - 2.5)^2}{6} \\ &= \frac{2.5}{6} = 0.4167\end{aligned}$$

$$\begin{aligned}\text{SRS (with)} : \frac{2}{16} \cdot [(1.5 - 2.5)^2 + (2.0 - 2.5)^2 + (2.5 - 2.5)^2 + (2.5 - 2.5)^2 + (3.0 - 2.5)^2 + (3.5 - 2.5)^2] \\ + \frac{1}{16} \cdot [(1.0 - 2.5)^2 + (2.0 - 2.5)^2 + (3.0 - 2.5)^2 + (4.0 - 2.5)^2] = \frac{10.0}{16} = 0.6250\end{aligned}$$

$$\mathcal{L}_1 \equiv \mathcal{P}_{c1} : \sigma_y^2 = \frac{(2.0 - 2.5)^2 + (3.0 - 2.5)^2}{2} = \frac{0.5}{2} = 0.25$$

$$\mathcal{L}_2 \equiv \mathcal{P}_{c2} : \sigma_y^2 = \frac{(1.5 - 2.5)^2 + (3.5 - 2.5)^2}{2} = \frac{2.0}{2} = 1.00$$

$$\mathcal{L}_3 \equiv \mathcal{P}_{c3} : \sigma_y^2 = \frac{(2.5 - 2.5)^2 + (2.5 - 2.5)^2}{2} = \frac{0.0}{2} = 0.00$$

$$\mathcal{P}_{s1} : \sigma_y^2 = \frac{(2.0 - 2.5)^2 + (2.5 - 2.5)^2 + (2.5 - 2.5)^2 + (3.0 - 2.5)^2}{4} = \frac{0.5}{4} = 0.125$$

$$\mathcal{P}_{s2} : \sigma_y^2 = \frac{(1.5 - 2.5)^2 + (2.0 - 2.5)^2 + (3.0 - 2.5)^2 + (3.5 - 2.5)^2}{4} = \frac{2.5}{4} = 0.625$$

Chapter 17

Analysis

- ▷ Estimators
- ▷ Variances
- ▷ Examples
- ▷ Sample size determination

17.1 Estimators

- ▷ We will focus on the two-stage case.
- ▷ Quantities can be estimated at two levels:
 - ▷ Within a PSU
 - ▷ For the entire population
- ▷ The expressions depend on the sample fraction at the SSU level, since this is not constant, with two special cases:
 - ▷ **self-weighting:** f_{2i} is constant
 - ▷ **simple cluster sampling:** f_{2i} is constant and equal to one (entire cluster sampled)
- ▷ We will present expressions for totals.
- ▷ Averages follow simply through dividing by N .

Estimators for Total

Total within SSU

$$\hat{y}_i = \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Population total

General

$$\hat{y} = \frac{M}{m} \sum_{i=1}^m \hat{y}_i = \frac{M}{m} \sum_{i=1}^m \frac{N_i}{n_i} \sum_{j=1}^{n_i} y_{ij}$$

Self-weighted

$$\hat{y} = \frac{M}{m} \frac{\overline{N}}{\bar{n}} \sum_{i=1}^m \sum_{j=1}^{n_i} y_{ij} = \frac{1}{f} \cdot y$$

Simple cluster

$$\hat{y} = \frac{M}{m} \sum_{i=1}^m \sum_{J=1}^{N_i} y_{iJ}$$

17.2 Variances

- We will present expressions for totals.
- Expressions for averages simply follow from dividing the variances for the estimators by $1/N^2$.
- Note that the simple cluster expression is a special case of the self-weighted expression, since for simple cluster sampling $f_2 = 1$ so that the second terms vanish.
- Expressions for non-self-weighted samples exist as well: versions of these will be discussed in Part VIII.

Variances for Total

Quantity

Calculated

Estimated

Population

$$S_{1Y}^2 = \frac{1}{M-1} \sum_{I=1}^M (Y_I - \bar{Y})^2$$

$$\hat{s}_{1y}^2 = \frac{1}{m-1} \sum_{i=1}^m (y_i - \bar{y})^2$$

$$S_{2Y}^2 = \frac{1}{N} \sum_{I=1}^M \frac{N_I}{N_I-1} \sum_{J=1}^{N_I} (Y_{IJ} - \bar{Y}_I)^2$$

$$\hat{s}_{2y}^2 = \frac{1}{n} \sum_{i=1}^m \frac{n_i}{n_i-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

Self-weighted

$$\begin{aligned} \sigma_{\hat{y}}^2 &= \frac{M^2}{m} (1 - f_1) S_{1Y}^2 \\ &\quad + \frac{M^2 \bar{N}^2}{m \bar{n}} (1 - f_2) S_{2Y}^2 \end{aligned}$$

$$\begin{aligned} \hat{\sigma}_{\hat{y}}^2 &= \frac{M^2}{m} (1 - f_1) \hat{s}_{1y}^2 \\ &\quad + \frac{M^2 \bar{N}^2}{m \bar{n}} (1 - f_2) \hat{s}_{2y}^2 \end{aligned}$$

Simple cluster

$$\sigma_{\hat{y}}^2 = \frac{M^2}{m} (1 - f_1) S_{1Y}^2$$

$$\hat{\sigma}_{\hat{y}}^2 = \frac{M^2}{m} (1 - f_1) \hat{s}_{1y}^2$$

17.3 Example: Artificial Population

- In Section 14.4, the intra-cluster (intraclass) correlations were calculated for SRS (without and with replacement), SYS (lists \mathcal{L}_1 , \mathcal{L}_2 , and \mathcal{L}_3), and STRAT (\mathcal{P}_{s1} and \mathcal{P}_{s1}).
- Given the identification between clustering and systematic sampling in this case, we can preserve the table:

Method	Variance	ρ	Relationship
SRS (without)	0.4167	-0.33	$\frac{0.4167}{1-0.33 \times (2-1)} = 0.6250$
SRS (with)	0.6250	0.00	$\frac{0.6250}{1+0.00 \times (2-1)} = 0.6250$
$\text{SYS}(\mathcal{L}_1) \equiv \text{CLUST}(\mathcal{P}_{c1})$	0.2500	-0.60	$\frac{0.2500}{1-0.60 \times (2-1)} = 0.6250$
$\text{SYS}(\mathcal{L}_2) \equiv \text{CLUST}(\mathcal{P}_{c2})$	1.0000	0.60	$\frac{1.0000}{1+0.60 \times (2-1)} = 0.6250$
$\text{SYS}(\mathcal{L}_3) \equiv \text{CLUST}(\mathcal{P}_{c3})$	0.0000	-1.00	undetermined
$\text{STRAT}(\mathcal{P}_{s1})$	0.1250	-0.80	$\frac{0.1250}{1-0.80 \times (2-1)} = 0.6250$
$\text{STRAT}(\mathcal{P}_{s2})$	0.6250	0.00	$\frac{0.6250}{1+0.00 \times (2-1)} = 0.6250$

17.4 Example: Surveytown

- In Section 9.3, two lists were considered:

$$\mathcal{L}_X = (1\ 2\ 3\ 4\ 5\ 6\ 7\ 8)$$

$$\mathcal{L}_Z = (2\ 6\ 5\ 3\ 7\ 1\ 4\ 8)$$

based on, respectively,

- ▷ X_I : number of building lots in block I
- ▷ Z_I : number of newspapers delivered in block I

- In Section 13.5, two stratifications were considered, based on the same information:

$$\mathcal{P}_{sX} = (1 \ 2 \ 3 \ 4 \mid 5 \ 6 \ 7 \ 8)$$

$$\mathcal{P}_{sZ} = (2 \ 6 \ 5 \ 3 \mid 7 \ 1 \ 4 \ 8)$$

- Carrying the idea further, assume we have two ways of defining clusters:

$$\mathcal{P}_{cX} = (\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\})$$

$$\mathcal{P}_{cZ} = (\{1, 7\}, \{2, 6\}, \{3, 5\}, \{4, 8\})$$

- Selecting, as usual, samples of size $n = 2$, implies that every sample reduces to just a single cluster:

$$\mathcal{S}_{cX} = \left\{ \{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\} \right\}$$

$$\mathcal{S}_{cZ} = \left\{ \{1, 7\}, \{2, 6\}, \{3, 5\}, \{4, 8\} \right\}$$

s	Sample	P_s					\hat{y}_s				
		SRS	Systematic		Clustered		SRS	Systematic		Clustered	
			\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{cX}	\mathcal{P}_{cZ}		\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{cX}	\mathcal{P}_{cZ}
1	{1,2}	1/28	0	0	1/4	0	12			12	
2	{1,3}	1/28	0	0	0	0	16				
3	{1,4}	1/28	0	0	0	0	20				
4	{1,5}	1/28	1/4	0	0	0	24	24			
5	{1,6}	1/28	0	1/4	0	0	28			28	
6	{1,7}	1/28	0	0	0	1/4	32				32
7	{1,8}	1/28	0	0	0	0	36				
8	{2,3}	1/28	0	0	0	0	20				
9	{2,4}	1/28	0	0	0	0	24				
10	{2,5}	1/28	0	0	0	0	28				
11	{2,6}	1/28	1/4	0	0	1/4	32	32			32
12	{2,7}	1/28	0	1/4	0	0	36		36		
13	{2,8}	1/28	0	0	1/4	0	40		40		
14	{3,4}	1/28	0	0	0	1/4	28				28
15	{3,5}	1/28	0	0	0	0	32				
16	{3,6}	1/28	0	0	0	0	36				

s	Sample	P_s					\hat{y}_s				
		SRS	Systematic		Clustered		SRS	Systematic		Clustered	
			\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{cX}	\mathcal{P}_{cZ}		\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{cX}	\mathcal{P}_{cZ}
17	{3,7}	1/28	1/4	0	0	0	40	40			
18	{3,8}	1/28	0	1/4	0	0	44		44		
19	{4,5}	1/28	0	1/4	0	0	36		36		
20	{4,6}	1/28	0	0	0	0	40				
21	{4,7}	1/28	0	0	0	0	44				
22	{4,8}	1/28	1/4	0	0	1/4	48	48			48
23	{5,6}	1/28	0	0	1/4	0	44			44	
24	{5,7}	1/28	0	0	0	0	48				
25	{5,8}	1/28	0	0	0	0	52				
26	{6,7}	1/28	0	0	0	0	52				
27	{6,8}	1/28	0	0	0	0	56				
28	{7,8}	1/28	0	0	1/4	0	60			60	
Expectation							36	36	36	36	36
Variance							144	80	32	320	48
Standard error							12.00	8.94	2.83	17.89	6.93

- The expectations for the total:

$$\mathcal{P}_{cX} : E(\bar{y}) = \frac{1}{4} \cdot [12 + 28 + 44 + 60] = \frac{144}{4} = 36$$

$$\mathcal{P}_{cZ} : E(\bar{y}) = \frac{1}{4} \cdot [32 + 32 + 32 + 48] = \frac{144}{4} = 36$$

- Hence, both lists produce unbiased estimators.

- The variances:

$$\mathcal{P}_{cX} : \sigma_{\bar{y}}^2 = \frac{(12 - 36)^2 + (28 - 36)^2 + (44 - 36)^2 + (60 - 36)^2}{4} = \frac{1280}{4} = 320$$

$$\mathcal{P}_{cZ} : \sigma_{\bar{y}}^2 = \frac{(32 - 36)^2 + (32 - 36)^2 + (32 - 36)^2 + (48 - 36)^2}{4} = \frac{192}{4} = 48$$

- Recall that the variance under SRS was 144.
- \mathcal{P}_{cX} increases variability dramatically, while \mathcal{P}_{sZ} decreases variability, relative to SRS.

- Yet, \mathcal{P}_{cX} is the more common, with positive correlation, that we will see in practice.
- Using the SAS procedure MIXED, the intra-cluster correlation can be calculated, based on the datasets:

Surveytown - Clust. based on X

Obs	sample	y
1	1	1
2	1	2
3	2	3
4	2	4
5	3	5
6	3	6
7	4	7
8	4	8

Surveytown - Clust. based on X

Obs	sample	y
1	1	1
2	1	2
3	2	3
4	2	4
5	3	5
6	3	6
7	4	7
8	4	8

- The correlations are:

$$\rho_{\mathcal{P}_cX} = 0.9048$$

$$\rho_{\mathcal{P}_cZ} = -0.7143$$

- In Parts IV and VI, we obtained relationships between variances, which we can now extend:

$$\begin{aligned} \frac{\sigma_{\hat{y}, \text{SRS}(\text{with})}^2}{1 + \rho_{\text{SRS}(\text{with})}(n-1)} &= \frac{\sigma_{\hat{y}, \text{SRS}(\text{without})}^2}{1 + \rho_{\text{SRS}(\text{without})}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{L}_X}^2}{1 + \rho_{\mathcal{L}_X}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{L}_Z}^2}{1 + \rho_{\mathcal{L}_Z}(n-1)} \\ &= \frac{\sigma_{\hat{y}, \mathcal{P}_{sX}}^2}{1 + \rho_{\mathcal{P}_{sX}}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{P}_{sZ}}^2}{1 + \rho_{\mathcal{P}_{sZ}}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{P}_{cX}}^2}{1 + \rho_{\mathcal{P}_{cX}}(n-1)} = \frac{\sigma_{\hat{y}, \mathcal{P}_{cZ}}^2}{1 + \rho_{\mathcal{P}_{cZ}}(n-1)} \\ &= \frac{168}{1 + 0.0000 \times (2-1)} = \frac{144}{1 - 0.1429 \times (2-1)} = \frac{80}{1 - 0.5238 \times (2-1)} = \frac{32}{1 - 0.8095 \times (2-1)} \\ &= \frac{40}{1 - 0.7619 \times (2-1)} = \frac{160}{1 - 0.0476 \times (2-1)} = \frac{320}{1 + 0.9048 \times (2-1)} = \frac{48}{1 - 0.7143 \times (2-1)} \end{aligned}$$

Rank	Method	Variance	ρ
1	SYS (\mathcal{L}_Z)	32	-0.81
2	CLUST (\mathcal{P}_{cZ})	48	-0.71
3	STRAT (\mathcal{P}_{sX})	40	-0.76
4	SYS (\mathcal{L}_X)	80	-0.52
5	SRS (without)	144	-0.14
6	STRAT (\mathcal{P}_{sZ})	160	-0.05
7	SRS (with)	168	0.00
8	CLUST (\mathcal{P}_{cX})	320	+0.90

17.5 Example: The Belgian Health Interview Survey

- Taking stratification into account, the means are recomputed for

- ▷ LNBMI
- ▷ LNVOEG
- ▷ GHQ12
- ▷ SGP

- The following program can be used:

```
proc surveymeans data=m.bmi_voeg mean stderr;  
title 'two-stage (clustered) means - inf. pop. - Belgium and regions';  
where (regionch^='');  
domain regionch;  
cluster hh;  
var lnbmi lnvoeg ghq12 sgp;  
run;
```

- The program includes the **CLUSTER** statement to acknowledge the two-stage nature of the sampling.
- Note that including three or more stages is not possible.
- While it would be possible to include a finite sample correction, as we have seen, the impact is so negligible that it has been omitted.
- The output takes the usual form, with now clustering information listed:

two-stage (clustered) means - infinite population for Belgium and regions

The SURVEYMEANS Procedure

Data Summary

Number of Clusters	4663
Number of Observations	8564

Statistics

Variable	Mean	Std Error of Mean
LNBMI	3.187218	0.001999
LNVOEG	1.702951	0.010335
GHQ12	1.661349	0.032824
SGP	0.903540	0.003963

Domain Analysis: REGIONCH

REGIONCH	Variable	Mean	Std Error of Mean
Brussels	LNBMI	3.175877	0.003630
	LNVOEG	1.809748	0.018073
	GHQ12	1.862745	0.062739
	SGP	0.805632	0.009766
Flanders	LNBMI	3.182477	0.003309
	LNVOEG	1.516352	0.017246
	GHQ12	1.385381	0.052202
	SGP	0.952285	0.004709
Walloonnia	LNBMI	3.201530	0.003429
	LNVOEG	1.801107	0.016963
	GHQ12	1.772148	0.055780
	SGP	0.938646	0.005284

- The summary:

Logarithm of Body Mass Index				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	3.187218(0.001845)	3.175877(0.003372)	3.182477(0.002993)	3.201530(0.003216)
Stratification	3.187218(0.001840)	3.175877(0.003373)	3.182477(0.002989)	3.201530(0.003217)
Clustering	3.187218(0.001999)	3.175877(0.003630)	3.182477(0.003309)	3.201530(0.003429)
Weighting	3.185356(0.002651)	3.171174(0.004578)	3.180865(0.003870)	3.198131(0.004238)
All combined	3.185356(0.003994)	3.171174(0.004844)	3.180865(0.004250)	3.198131(0.004403)

Logarithm of VOG Score				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	1.702951(0.008954)	1.809748(0.016203)	1.516352(0.015201)	1.801107(0.014550)
Stratification	1.702951(0.008801)	1.809748(0.016206)	1.516352(0.015207)	1.801107(0.014427)
Clustering	1.702951(0.010355)	1.809748(0.018073)	1.516352(0.017246)	1.801107(0.016963)
Weighting	1.634690(0.013233)	1.802773(0.021831)	1.511927(0.019155)	1.803178(0.020426)
All combined	1.634690(0.014855)	1.802773(0.023135)	1.511927(0.021409)	1.803178(0.023214)

General Health Questionnaire – 12				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	1.661349(0.029584)	1.862745(0.056894)	1.385381(0.046246)	1.772148(0.051023)
Stratification	1.661956(0.029452)	1.864301(0.056939)	1.385857(0.046211)	1.772148(0.050823)
Clustering	1.661349(0.032824)	1.862745(0.062739)	1.385381(0.052202)	1.772148(0.055780)
Weighting	1.626201(0.044556)	1.924647(0.076313)	1.445957(0.061910)	1.858503(0.078566)
All combined	1.626781(0.048875)	1.924647(0.080508)	1.446286(0.068931)	1.858503(0.084047)

Stable General Practitioner (0/1)				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	0.903540(0.003196)	0.805632(0.007826)	0.952285(0.003908)	0.938646(0.004382)
Stratification	0.903540(0.003116)	0.805632(0.007827)	0.952285(0.003902)	0.938646(0.004366)
Clustering	0.903540(0.003963)	0.805632(0.009766)	0.952285(0.004709)	0.938646(0.005284)
Weighting	0.932702(0.003498)	0.782448(0.011563)	0.954757(0.004722)	0.943191(0.005417)
All combined	0.932702(0.003994)	0.782448(0.013836)	0.954757(0.005379)	0.943191(0.006159)

- We can make the following observations, when comparing clustering to SRS:
 - ▷ The point estimates are invariant; clustering only affects the precision estimates.
 - ▷ The impact on LNBMI is small, a bit higher on LNVOEG, considerable on GHQ-12, and large on SGP.
 - ▷ The reason is that a variable like BMI, while open to genetic and environmental factors, and therefore within-family association, changes a lot between individuals.
- In contrast, whether or not there is a stable GP, a family GP, is virtually a HH-level decision.

17.6 Sample Size Determination

- General expressions are complicated.
- They are similar to SRS for simple cluster sampling (next page).
- Expressions for sampling with equal probability: Levy and Lemeshow (1999, p. 317).

Situation	Total (\hat{y})	Average (\bar{y})
Without replacement	$m = \frac{M^2 \sigma_{1Y}^2}{\sigma_{\hat{y}}^2 + M \sigma_{1Y}^2}$	$m = \frac{\sigma_{1Y}^2}{\sigma_{\bar{y}}^2 + (1/M) \sigma_{1Y}^2}$
With replacement	$m = \frac{M^2 \sigma_{1Y}^2}{\sigma_{\hat{y}}^2}$	$m = \frac{\sigma_{1Y}^2}{\sigma_{\bar{y}}^2}$
$M \rightarrow +\infty$	—	$m = \frac{\sigma_{1Y}^2}{\sigma_{\bar{y}}^2}$

Chapter 18

Complex-Model-Based Analysis

- ▷ General principles
- ▷ Linear mixed models (LMM)
- ▷ Generalized estimating equations (GEE)
- ▷ Generalized linear mixed models (GLMM)
- ▷ Application to the Belgian Health Interview Survey

18.1 Principles

- Analysis methods in Chapter 17 are based on incorporating the multi-stage and/or cluster aspects of the design into simple estimators (mean, total, proportion).
- Modern analysis tools for **hierarchical data** can be used.
- We have to distinguish between methods for continuous and binary data.
- In the binary data case, there are several **non-equivalent** options.

18.2 Linear Mixed Models

- An instance of this model was used in Part IV, where we considered the set of potential systematic samples as **clusters**.
- Virtually the same model can be used for mean (and total) estimation:

$$Y_{ij} = \mu + b_i + \varepsilon_{ij}$$

- ▷ Y_{ij} is the observation for subject j in cluster i
- ▷ μ is the overall, population mean
- ▷ $\mu + b_i$ is the cluster-specific average:

$$b_i \sim N(0, \tau^2)$$

▷ ε_{ij} is an individual-level deviation:

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

▷ We also term b_i the cluster-specific deviation

▷ The following terminology is commonly used:

- * μ is a **fixed effect** (fixed intercept).
- * b_i is a **random effect** (random intercept).
- * ε_{ij} is a residual deviation ('error' in samples).

● This is an instance of a **linear mixed model**.

● Verbeke and Molenberghs (2000)

- Several extensions are possible:
 - ▷ The mean μ can be expanded into a regression function (see Part IX).
 - ▷ The single random effect can be supplemented with more random effects.
 - ▷ The model can be formulated for three and more levels as well.
 - ▷ For example,

$$Y_{ijk} = \mu + b_i + c_{ij} + \varepsilon_{ijk}$$

- * Y_{ijk} is the observation for subject k in household j in town i
- * μ is the overall, population mean
- * b_i is the town-level effect
- * c_{ij} is the household-level effect
- * ε_{ijk} is the individual-level deviation

- ▷ Typical distributional assumptions:

$$b_i \sim N(0, \tau_{\text{town}}^2)$$

$$c_{ij} \sim N(0, \tau_{\text{HH}}^2)$$

$$\varepsilon_{ijk} \sim N(0, \tau_{\text{ind}}^2)$$

- ▷ This is a three-level model.
- ▷ When μ and/or b_i and/or c_{ij} are made functions of covariates, we have a so-called **multi-level approach**.

linear mixed model \equiv **multi-level model**

- **Parameter estimation:**

- ▷ maximum likelihood (ML)
- ▷ restricted maximum likelihood (REML): small-sample correction of ML, to reduce small-sample bias

- **Targets of inference:**

- ▷ fixed effects (e.g., μ)
- ▷ variance components (e.g., τ_{town}^2 , τ_{HH}^2 , and τ_{ind}^2)
- ▷ random effects (e.g., b_i and c_{ij})

- Implementation via **PROC MIXED**

18.2.1 Example: the Belgian Health Interview Survey

- Implementation of the basic, SRS analysis in PROC MIXED, to compute the means for LNBMI, can be done with the following programs (Belgium and regions):

```
proc mixed data=m.bmi_voeg method=reml;  
title 'Survey mean with PROC MIXED, for Belgium';  
title2 'SRS';  
where (regionch^='');  
model lnbmi = / solution;  
run;
```

```
proc mixed data=m.bmi_voeg method=reml;  
title 'Survey mean with PROC MIXED, for regions';  
title2 'SRS';  
where (regionch^='');  
by regionch;  
model lnbmi = / solution;  
run;
```

- This is a special version of the linear mixed model, without random effects, hence ordinary linear regression.
- The following statements and options deserve attention:
 - ▷ The **WHERE** and **BY** statements have their usual meaning.
 - ▷ The **MODEL** statement specifies the mean structure:
 - * The intercept μ is included by default; this is why the right hand side of the equality sign is empty.
 - * The '**solution**' option requests estimates, standard errors, . . . for the fixed effects.

- Let us discuss selected output:

Survey mean with PROC MIXED, for Belgium
SRS
The Mixed Procedure

Dimensions		Number of Observations	
Covariance Parameters	1	Number of Observations Read	8564
Columns in X	1	Number of Observations Used	8384
Columns in Z	0	Number of Observations Not Used	180
Subjects	1		
Max Obs Per Subject	8564		

- ▷ There is only one covariance parameter, the variance.
- ▷ Columns in X: the number of fixed effects; there is only one, the intercept.
- ▷ Columns in Z: the number of random effects; there are none.
- ▷ The number of subject s is not relevant when there is no hierarchy.

- ▷ The number of observations per subject, since there is no subject specification, is the actual number of measurements.
- ▷ Observations are not used whenever key variables are missing, e.g., when LNBMI is not available.

Covariance Parameter Estimates

Cov Parm	Estimate
Residual	0.02853

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	3.1872	0.001845	8383	1727.76	<.0001

- ▷ The covariance parameter is σ^2 , the estimated population variance.
- ▷ The intercept is the population average μ .

- The output for each of the regions separately takes entirely the same format.
- The version including clustering, i.e., a household-level random effect:

```
proc mixed data=m.bmi_voeg method=reml;  
title 'Survey mean with PROC MIXED, for Belgium';  
title2 'Two-stage (clustered)';  
where (regionch^='');  
model lnbmi = / solution;  
random intercept / subject=hh;  
run;
```

- An additional statement is included:
 - ▷ The **RANDOM** statement specifies the random effect b_i :
 - * The keyword '**intercept**' needs to be used (unlike in the MODEL statement).
 - * The '**subject**' option specifies the level of independent replication.

- The output changes:

Survey mean with PROC MIXED, for Belgium
Two-stage (clustered)
The Mixed Procedure

Dimensions		Number of Observations	
Covariance Parameters	2	Number of Observations Read	8564
Columns in X	1	Number of Observations Used	8384
Columns in Z Per Subject	1	Number of Observations Not Used	180
Subjects	4663		
Max Obs Per Subject	4		

- ▷ There now are two covariance parameters, σ^2 and τ^2 .
- ▷ The 'number of subjects' is **the number of households**.
- ▷ The 'max obs per subject' is **the (maximum) number of individuals within a household**.
- ▷ More observations are not used, since an additional variable in use, **household (hh)**, which can be missing, too.

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	HH	0.004289
Residual		0.02425

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	3.1880	0.001991	4593	1601.34	<.0001

- ▷ There still is one population average estimated, $\hat{\mu} = 3.1880(0.0020)$.
- ▷ Both variance components are present:

$$\hat{\sigma}^2 = 0.0243$$

$$\hat{\tau}^2 = 0.0043$$

$$\hat{\rho} = \frac{\tau^2}{\sigma^2 + \tau^2} = \frac{0.0043}{0.0243 + 0.0043} = 0.15$$

- The correlation ρ is the intra-cluster (intra-household) correlation.
- Note that the intra-household correlation depends on the endpoint; it is different for different variables.

For example, for LNVOEG (details of output not shown), it changes to:

$$\hat{\rho}_{\text{LNVOEG}} = \frac{\tau^2}{\sigma^2 + \tau^2} = \frac{0.1804}{0.4801 + 0.1804} = 0.27$$

- Summary of the various methods for mean estimation on LNBMI:

Logarithm of Body Mass Index					
Analysis	Procedure	Belgium	Brussels	Flanders	Wallonia
SRS	SURVEYMEANS	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
SRS	MIXED	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
Stratification	SURVEYMEANS	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
Clustering	SURVEYMEANS	3.1872(0.0020)	3.1759(0.0036)	3.1825(0.0033)	3.2015(0.0034)
Clustering	MIXED	3.1880(0.0020)	3.1761(0.0036)	3.1840(0.0033)	3.2022(0.0034)
Weighting	SURVEYMEANS	3.1853(0.0027)	3.1712(0.0046)	3.1809(0.0039)	3.1981(0.0042)
Weighting	MIXED	3.1854(0.0018)	3.1712(0.0034)	3.1809(0.0030)	3.1981(0.0032)
All combined	SURVEYMEANS	3.1853(0.0040)	3.1712(0.0048)	3.1809(0.0043)	3.1981(0.0044)
Clust+Wgt	MIXED	3.1865(0.0023)	3.1706(0.0039)	3.1817(0.0036)	3.1994(0.0038)

- **SRS**: Whether the procedure SURVEYMEANS or MIXED is used does not make any difference.
- **Clustering**: There is a small difference between SURVEYMEANS and MIXED for the parameter estimate, but not for the standard error.
This is due to a different handling of incomplete data.

- Note that it is also possible to use the **SURVEYREG** procedure:

```
proc surveyreg data=m.bmi_voeg;  
title 'Mean. Surveyreg, two stage (clustered), for regions';  
by regionch;  
cluster hh;  
model lnbmi = ;  
run;
```

- The statements are self-explanatory, for example:
 - ▷ Removing the **BY** statement produces the results for Belgium.
 - ▷ Removing the **CLUSTER** statement leads to SRS.
 - ▷ There is no right hand side in the model in the **MODEL** statement, since we only want a mean \equiv intercept, which is included by default.

► A selection from the output for Belgium, where clustering is taken into account:

Mean. Surveyreg, two stage (clustered), for Belgium
The SURVEYREG Procedure

Regression Analysis for Dependent Variable LNBMI

Data Summary

Number of Observations	8384
Mean of LNBMI	3.18722
Sum of LNBMI	26721.6

Design Summary

Number of Clusters	4594
--------------------	------

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.18721840	0.00199922	1594.23	<.0001

- * The data summary usefully contains the **mean** and the **total**.
 - * The regression coefficient, which in this case also is the mean, is self-explanatory.
- ▷ Thus, results reported for SURVEYMEANS can also be considered as resulting from SURVEYREG.

18.3 Generalized Estimating Equations

- When an outcome is binary, one can calculate a **proportion** π , which is the **probability** to belong to a group, to have a certain characteristic, etc.
- Alternatively, the logit can be calculated:

$$\beta = \text{logit}(\pi) = \ln \left(\frac{\pi}{1 - \pi} \right),$$

$$\pi = \frac{e^{\beta}}{1 + e^{\beta}}$$

- The model can then be written as:

$$\text{logit}[P(Y_i = 1)] = \beta$$

- Estimation of β typically proceeds through maximum likelihood estimation, which necessitates numerical optimization, since no closed form exists.
- For SRS, this can be implemented the SAS procedures **LOGISTIC** and **GENMOD**
- For the clustered case, the correlation can be incorporated into the model:

$$\text{logit}[P(Y_{ij} = 1)] = \beta,$$

$$\text{Corr}(Y_{ij}, Y_{ik}) = \alpha$$

- Note that we now need the double index again: i for household, j for individual within household.
- β is the logit of the population proportion.
- α is the correlation between the outcome of two individuals within the same

household.

- Full maximum likelihood estimation is tedious.
- Liang and Zeger (Biometrika 1986) have developed a convenient estimation method: **generalized estimating equations (GEE)**.
- A way to think about it is: **correlation-corrected logistic regression**.
- It can also be implemented using the SAS procedure GENMOD.

18.3.1 Example: the Belgian Health Interview Survey

- We will estimate the mean (probability) for SGP:
 - ▷ For Belgium and the regions.
 - ▷ Under SRS and two-stage (cluster) sampling.
 - ▷ Using:
 - * **PROC SURVEYLOGISTIC** for survey-design-based regression.
 - * **PROC GENMOD** for GEE.
- A PROC SURVEYLOGISTIC program for the two-stage case and for the regions is:

```
proc surveylogistic data=m.bmi_voeg;  
title '22. Mean. Surveylogistic, two-stage (clustered), for regions';  
by regionch;  
cluster hh;  
model sgp = ;  
run;
```

- The following statements deserve attention:
 - ▷ The **BY** statement has the same meaning as in PROC MEANS.
 - ▷ Dropping it produces estimates for Belgium.
 - ▷ The **CLUSTER** statements has the same meaning as in PROC SURVEYMEANS.
 - ▷ Dropping it produces SRS estimates.
 - ▷ The **MODEL** specifies the outcome, SGP in our case.
 - ▷ There are no covariates and there is an intercept by default, which is why the right hand side is empty.

- Let us discuss selected output, for SRS and for Belgium:

15. Mean. Surveylogistic, SRS, for Belgium
The SURVEYLOGISTIC Procedure

Model Information

Data Set	M.BMI_VOEG
Response Variable	SGP
Number of Response Levels	2
Model	Binary Logit
Optimization Technique	Fisher's Scoring

Number of Observations Read	8564
Number of Observations Used	8532

Response Profile

Ordered Value	SGP	Total Frequency
1	0	823
2	1	7709

Probability modeled is SGP=0.

NOTE: 32 observations were deleted due to missing values for the response or explanatory variables.

- ▷ The ‘two response levels’ refers to the fact that we have a dichotomous outcome, and we are given the raw frequencies of these, together with information about missingness.
- ▷ The optimization method is Fisher’s scoring, an iterative method: logistic regression and its extensions like survey-design-based logistic regression requires iterative optimization.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.2372	0.0367	3721.3534	<.0001

- ▷ The parameter estimate is a negative value!
- ▷ This is because the logit of the probability of not having a stable GP is modeled:

$$\text{logit}[P(Y_{ij} = 0)] = \beta$$

where $Y_{ij} = 0$ if respondent j in household i does not have a stable GP.

▷ It then follows that

$$\hat{\pi} = \frac{e^{-\hat{\beta}}}{1 + e^{-\hat{\beta}}} = \frac{e^{2.2372}}{1 + e^{2.2372}} = 0.9035$$

which is the same value as obtained with PROC SURVEYMEANS.

▷ The standard error for π follows from the **delta method**:

$$\hat{\sigma}_{\hat{\pi}} = \hat{\pi}[1 - \hat{\pi}]\hat{\sigma}_{\hat{\beta}} = 0.9035 \times 0.0965 \times 0.0367 = 0.0032$$

which is the same value as obtained with PROC SURVEYMEANS.

▷ When clustering is taken into account, we obtain

$$\hat{\beta} = -2.2372(\text{s.e. } 0.0455) \quad \Rightarrow \quad \hat{\pi} = 0.9035(\text{s.e. } 0.0040)$$

This too, coincides with the SURVEYMEANS result.

- **Conclusion:** estimating a proportion (and s.e.) with PROC SURVEYMEANS \equiv estimating the logit of the proportion (and s.e.) with PROC SURVEYLOGISTIC.

This is true for every collection of design aspects taken into account.

- Switching to GEE with PROC GENMOD, for the two-stage case and the regions:

```
proc genmod data=m.bmi_voeg;  
title '30. Mean. GEE logistic regression, for regions';  
title2 'Two-stage (clustered)';  
by regionch;  
class hh;  
model sgp = / dist=b;  
repeated subject = hh / type=cs corrw modelse;  
run;
```

- The following statements deserve attention:
 - ▷ The **BY** statement has the same meaning as before.
 - ▷ Dropping it produces estimates for Belgium.
 - ▷ The **MODEL** specifies the outcome, SGP in our case.

- * Since there are no covariates and since the intercept is included by default, the right hand side is empty.
 - * The **'dist=b'** option specified a Bernoulli distribution, which comes with the logit link as the default link function.
 - * This specification is necessary since the procedure also performs linear regression, Poisson regression, probit regression, etc.
- ▷ Clustering is now accounted for in a different way, through the so-called marginal correlation structure:
- * The **REPEATED** statement ensures we are using GEE.
 - * The **'subject='** option specifies the independent blocks, effectively ensuring a two-stage analysis with HH and individuals.
 - * The **'type='** option specifies the correlation structure, which here is **compound symmetry**, i.e., all correlations within a household are assumed equal.
 - * **Even if this is not true, the resulting estimates and standard errors are still valid!**

This is a main advantage of the method.

- * The 'corr^w' option requests printing of the correlation structure (also named the **working correlation structure**).
- * The 'model^{se}' option requests an alternative set of standard errors, valid only when the correlation structure is correctly specified.

It is advisable to always use the *other* set of standard errors: named the robust, sandwich, or empirically corrected standard errors.

- * The **CLASS** statement is needed, since the subject variable needs to be a class variable.
- ▷ Dropping the REPEATED and CLASS statements produces SRS estimates.

- Let us discuss selected output, for SRS and for Belgium:

25. Mean. GEE logistic regression, for Belgium
SRS

The GENMOD Procedure

Model Information

Data Set	M.BMI_VOEG
Distribution	Binomial
Link Function	Logit
Dependent Variable	SGP

Number of Observations Read	8564
Number of Observations Used	8532
Number of Events	823
Number of Trials	8532
Missing Values	32

Response Profile

Ordered Value	SGP	Total Frequency
1	0	823
2	1	7709

PROC GENMOD is modeling the probability that SGP='0'. One way to change this to model the probability that SGP='1' is to specify the DESCENDING option in the PROC statement.

- ▷ The 'book keeping' information is similar to the one produced by PROC SURVEYLOGISTIC.

Analysis Of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-2.2372	0.0367	-2.3091	-2.1653	3721.79	<.0001

- ▷ The parameter estimate and standard error is exactly the same as with PROC SURVEYLOGISTIC.
- ▷ Hence, also the derived probability and its standard error is the same.
- Let us switch to the output for the clustered case, where genuine GEE is used, through the REPEATED statement.
- The output is more extensive than in the above case, which was in fact merely

ordinary logistic regression.

- The same book keeping information is provided and we do not print it again.
But more information is produced:

29. Mean. GEE logistic regression, for Belgium
Two-stage (clustered)

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	HH (4663 levels)
Number of Clusters	4663
Clusters With Missing Values	30
Correlation Matrix Dimension	4
Maximum Cluster Size	4
Minimum Cluster Size	0

- ▷ This information is geared towards the two-level structure of the model.
- ▷ The maximum cluster size refers, again, to the fact that at most 4 individuals per household are interviewed.

- Three sets of parameter estimates are produced:

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-2.2372	0.0367	-2.3091	-2.1653	3721.79	<.0001

Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits	Z	Pr > Z
Intercept	-2.1504	0.0435	-2.2358 -2.0651	-49.39	<.0001

Analysis Of GEE Parameter Estimates Model-Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits	Z	Pr > Z
Intercept	-2.1504	0.0425	-2.2337 -2.0671	-50.59	<.0001

- ▷ The **initial estimates** are equal to the SRS ones; they are included to start up the iterative GEE estimation process.
They should not be used for inferences.
- ▷ The **model-based estimates** are valid only when the working correlation is correct.
They should not be used for inferences.
- ▷ The **empirically corrected estimates** are the proper GEE estimates.
They are the ones to be used for inferences.
- ▷ In our case, the latter two sets are very similar, indicating that a common within-HH correlation is sensible.
- ▷ The within-HH correlation is estimated and part of the output as well:

Exchangeable Working
Correlation

Correlation 0.4522999388

- Note that now the parameter estimates are **different** from their SURVEYLOGISTIC counterparts. We now have:

$$\hat{\beta} = -2.1504(\text{s.e. } 0.0435) \quad \Rightarrow \quad \hat{\pi} = 0.8957(\text{s.e. } 0.0041)$$

Nevertheless, they are close to each other.

- We can expand the summary table for SGP with our new analyses:

Stable General Practitioner (0/1) — Marginal Models						
Analysis	Procedure	Par.	Belgium	Brussels	Flanders	Wallonia
SRS	SURVEYMEANS	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	SURVEYLOGISTIC.	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	SURVEYLOGISTIC.	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GENMOD	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GENMOD	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
Strat.	SURVEYMEANS	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Strat.	SURVEYLOGISTIC	$-\beta$	2.3272(0.0358)	1.4219(0.0050)	2.9936(0.0859)	2.7278(0.0758)
Strat.	SURVEYLOGISTIC	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Clust.	SURVEYMEANS	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	SURVEYLOGISTIC	$-\beta$	2.2372(0.0455)	1.4219(0.0624)	2.9936(0.1037)	2.7278(0.0918)
Clust.	SURVEYLOGISTIC	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	GENMOD	$-\beta$	2.1504(0.0435)	1.3784(0.0591)	2.9188(0.1019)	2.6470(0.0890)
Clust.	GENMOD	π	0.8957(0.0040)	0.7987(0.0095)	0.9488(0.0050)	0.9338(0.0055)
Wgt.	SURVEYMEANS	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	SURVEYLOGISTIC	$-\beta$	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	SURVEYLOGISTIC	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	GENMOD	$-\beta$	2.6290(0.0642)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
Wgt.	GENMOD	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
All	SURVEYMEANS	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
All	SURVEYLOGISTIC	$-\beta$	2.6290(0.0636)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
All	SURVEYLOGISTIC	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Cl.+Wt.	GENMOD	$-\beta$	2.5233(0.0659)	1.2014(0.0839)	2.9693(0.1284)	2.7251(0.1186)
Cl.+Wt.	GENMOD	π	0.9258(0.0045)	0.7688(0.0149)	0.9512(0.0060)	0.9385(0.0068)

- In summary, we note the following:
 - ▷ SURVEYLOGISTIC consistently produces the same estimates as SURVEYMEANS for the probability, upon transformation.
 - ▷ **SRS**: GEE (GENMOD) produces the same estimates and standard errors as the other methods.
 - ▷ **Clustering**: GEE (GENMOD) produces slightly different estimates and standard errors.
 - ▷ Whatever method chosen, the inferences will be the same.
 - ▷ The advantage of the SURVEYMEANS procedure is that direct estimates are obtained; no need to transform.
 - ▷ The advantage of the modelling procedures is that they allow for more complex models, as we will see in Part **IX**.

18.4 Generalized Linear Mixed Models

- We already considered two models to account for clustering:

▷ The LMM, through random effects:

$$▷ Y_{ij} = \mu + b_i + \varepsilon_{ij}$$

$$▷ b_i \sim N(0, \tau^2)$$

$$▷ \varepsilon_{ij} \sim N(0, \sigma^2)$$

▷ GEE, through marginal correlation:

$$▷ P(Y_{ij} = 1) = \frac{e^\beta}{1+e^\beta}$$

$$▷ \text{Corr}(Y_{ij}, Y_{ik}) = \alpha$$

- Aspects of both can be combined, to produce the **generalized linear mixed model** (GLMM):

$$P(Y_{ij} = 1) = \frac{e^{\beta + b_i}}{1 + e^{\beta + b_i}}$$
$$b_i \sim N(0, \tau^2)$$

- There are a few important differences:
 - ▷ Unlike with the LMM and GEE, it is not straightforward to calculate/obtain the intra-cluster correlation.
 - ▷ ML is an obvious candidate for parameter estimation.

- ▷ **But:** the likelihood contribution for cluster (household) i is:

$$L_i = \int \prod_{j=1}^{n_i} \frac{y_{ij} \cdot e^{\beta+b_i}}{1 + e^{\beta+b_i}} \cdot \varphi(b_i|\tau^2) db_i$$

where $\varphi(b_i|\tau^2)$ is the normal density.

- ▷ There exists no closed-form solution for this integral.
- The stated problem has led to two main approximation approaches:
 - ▷ **Numerical integration:** implemented in the SAS procedure **NLMIXED**.
 - * Allows for high accuracy.
 - * Time consuming.
 - * A bit harder to program.
 - ▷ **Taylor series expansions:** implemented in the SAS procedure **GLIMMIX**.
 - * Bias due to poor approximation.
 - * As easy to use as the MIXED and GENMOD procedures.

18.4.1 Example: the Belgian Health Interview Survey

- We will estimate the mean (probability) for SGP:
 - ▷ For Belgium and the regions.
 - ▷ Under SRS and two-stage (cluster) sampling.
 - ▷ Using **PROC GLIMMIX** for the GLMM.
 - ▷ Using **PROC NLMIXED** for the GLMM.

- A PROC GLIMMIX program for the two-stage case and for the regions is:

```
proc glimmix data=m.bmi_voeg;  
title '42. Mean. GLMM, for regions';  
title2 'with proc glimmix';  
title3 'two-stage (cluster)';  
nloptions maxiter=50;  
by regionch;  
model sgp = / solution dist=b;  
random intercept / subject = hh type=un;  
run;
```

- The following statements deserve attention:
 - ▷ The **MODEL** specifies the outcome, SGP in our case.
 - * Since there are no covariates and since the intercept is included by default, the right hand side is empty.
 - * The '**dist=b**' option specifies a Bernoulli distribution, which comes with the logit link as the default link function.

- * This specification is necessary since the procedure also performs linear regression, Poisson regression, probit regression, etc.
 - ▷ Like in the MIXED procedure, we specify clustering through the **RANDOM** statement:
 - * The '**subject=**' option specifies the independent blocks, effectively ensuring a two-stage analysis with HH and individuals.
 - * The '**type=**' option specifies the correlation structure, which here is **unstructured**.

This actually does not matter here, since there is only one random effect, and then the 'covariance structure' simply is the variance of this single random effect.

 - * Unlike in GENMOD, we do not need the **CLASS** statement, although it is fine to include it for HH: it simply has no impact in this situation.
- ▷ Dropping the RANDOM statement produces SRS estimates.

- Let us discuss selected output, for SRS and for Belgium:

```
37. Mean. GLMM, for Belgium
with proc glimmix
SRS
```

The GLIMMIX Procedure

Model Information

Data Set	M.BMI_VOEG
Response Variable	SGP
Response Distribution	Binomial
Link Function	Logit
Variance Function	Default
Estimation Technique	Maximum Likelihood
Number of Observations Read	8564
Number of Observations Used	8532

Dimensions

Columns in X	1
Columns in Z	0
Subjects (Blocks in V)	1
Max Obs per Subject	8532

- ▷ Similar book keeping information than with the GENMOD and MIXED procedures is provided.

- ▷ The X and Z columns have the same meaning as in the MIXED procedure.

Iteration History

Iteration	Restarts	Evaluations	Objective Function	Change	Max Gradient
0	0	4	2736.7999556	.	227.8632
1	0	3	2706.9270419	29.87291375	20.39023
2	0	3	2706.6515674	0.27547449	0.218305
3	0	3	2706.6515354	0.00003204	0.000026
4	0	8	2706.6515354	-0.00000000	0.000026

Convergence criterion (GCONV=1E-8) satisfied.

- ▷ The iteration panel gives details about the numerical convergence.
- ▷ A similar panel actually is given for GENMOD too, but there it is less relevant.
- ▷ Here, it is best to monitor it, especially since the number of iterations is by default equal to 20.
- ▷ It is therefore better to increase it, as we have done using the **NLOPTIONS** statement.

Parameter Estimates

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	2.2372	0.03667	8531	61.01	<.0001

- ▷ The parameter estimates are the same as with the SURVEYLOGISTIC and GENMOD procedures.
- ▷ This is to be expected with SRS, since in this case everything reduces to ordinary logistic regression.
- ▷ We therefore still find:

$$\hat{\beta} = -2.2372(\text{ s.e. } 0.0367) \Rightarrow \hat{\pi} = 0.9035(\text{ s.e. } 0.0032)$$

- Let us switch to the output for the clustered case:

```
41. Mean. GLMM, for Belgium
with proc glimmix
two-stage (cluster)
The GLIMMIX Procedure
```

Dimensions

G-side Cov. Parameters	1
Columns in X	1
Columns in Z per Subject	1
Subjects (Blocks in V)	4663
Max Obs per Subject	4

Iteration History

Iteration	Restarts	Subiterations	Objective Function	Change	Max Gradient
0	0	4	40394.419563	0.92175167	1.2E-6
1	0	4	41863.211127	0.56258530	0.003163
...					
12	0	1	42683.59837	0.00000019	0.000012
13	0	0	42683.598628	0.00000000	1.407E-6

Convergence criterion (PCONV=1.11022E-8) satisfied.

- ▷ A portion of the book keeping information that has changed is displayed.
- ▷ There now is 1 random effect: 1 column in the Z matrix.
- ▷ The convergence was a little more difficult, necessitating 13 iterations.

Covariance Parameter Estimates

Cov			
Parm	Subject	Estimate	Standard Error
UN(1,1)	HH	1.7506	0.1215

Solutions for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	2.3723	0.04431	4661	53.54	<.0001

- ▷ We obtain the following probability:

$$\hat{\beta} = -2.3723(\text{ s.e. } 0.0443) \Rightarrow \hat{\pi} = 0.9147(\text{ s.e. } 0.0035)$$

- ▷ The estimate for β is supplemented with an estimate for the random effects variance: $\hat{\tau}^2 = 1.75$ with s.e. 0.12.

- ▷ $\hat{\beta}$ and its standard error is not very different from what was obtained with the GENMOD procedure.
 - ▷ The latter is a subtle point, we will return to it after having discussed the NLMIXED program and output.
- We can now consider the NLMIXED program, allowing for clustering and intended for the regions:

```
proc nlmixed data=m.bmi_voeg;  
title '36. Mean. GLMM, for regions';  
title2 'Two-stage (clustered)';  
by regionch;  
theta = beta0 + b;  
exptheta = exp(theta);  
p = exptheta/(1+exptheta);  
model sgp ~ binary(p);  
random b ~ normal(0,tau2) subject=hh;  
estimate 'mean' exp(beta0)/(1+exp(beta0));  
run;
```

- The following statements deserve attention:
 - ▷ Dropping the **BY** statement produces the analysis for Belgium.
 - ▷ The procedure is very different from virtually all other SAS procedures: it is **programming statements based**.
 - ▷ The **MODEL** statement specifies:
 - * the outcome (SGP)
 - * what distribution it follows (binary \equiv Bernoulli in this case)
 - * the parameter ($p = \pi$)
 - * The parameter p itself is modeled through user-defined modeling statements.
 - * 'theta' refers to the linear predictor:
$$\theta = \beta_0 + b_i$$
 - * Then, the logistic transformation is applied to it.
 - * Note that the programming statements are certainly not uniquely defined.

We could make the following replacement:

```
theta = beta0 + b;  
exptheta = exp(theta);  
p = exptheta/(1+exptheta);
```

--> $p = \exp(\beta_0 + b) / (1 + \exp(\beta_0 + b));$

and reach the same result.

▷ the **RANDOM** statement specifies the random-effects structure:

- * The '**subject=**' option specifies the independent blocks, effectively ensuring a two-stage analysis with HH and individuals.
- * The random effect itself is part of the programming statements.
- * It is then declared to follow a distribution, always the normal distribution in this procedure, in the **RANDOM** statement.
- * The mean and variance of this normal distribution are open to programming statements, too.

- ▷ Dropping the RANDOM statement produces SRS estimates.
- ▷ The **ESTIMATE** statement allows for the estimation of additional, perhaps **non-linear**, functions of the fixed effect.

This allows for the direct calculation of the probabilities π from the parameter β .

- Let us discuss selected output, for SRS and for Belgium:

33. Mean. GLMM, for Belgium

SRS

The NLMIXED Procedure

Specifications

Data Set	M.BMI_VOEG
Dependent Variable	SGP
Distribution for Dependent Variable	Binary
Optimization Technique	Dual Quasi-Newton
Integration Method	None

Dimensions

Observations Used	8532
Observations Not Used	32
Total Observations	8564
Parameters	1

Iteration History

Iter	Calls	NegLogLike	Diff	MaxGrad	Slope
1	2	2725.83958	769.9091	158.5903	-21656.3
2	4	2707.66728	18.1723	39.41302	-22.7041
3	5	2706.66115	1.006124	3.77702	-2.24629
4	6	2706.65154	0.009614	0.078502	-0.01883
5	7	2706.65154	4.144E-6	0.000161	-8.3E-6

NOTE: GCONV convergence criterion satisfied.

- ▷ Very similar book keeping information is provided.
- ▷ There is no integration done here, since there are no random effects.

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
beta0	2.2372	0.03667	8532	61.01	<.0001	0.05	2.1653	2.3091

Additional Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
mean	0.9035	0.003196	8532	282.70	<.0001	0.05	0.8973	0.9098

- ▷ The parameter estimate for β is the same as in all previous situations, again in line with expectation.
- ▷ The additional estimate is the value for π we had obtained before: we do not have to calculate it 'by hand' now, nor do we have to apply the delta method ourselves.

- Let us switch to the output for the clustered case:

35. Mean. GLMM, for Belgium
Two-stage (clustered)
The NL MIXED Procedure

	Specifications
Data Set	M.BMI_VOEG
Dependent Variable	SGP
Distribution for Dependent Variable	Binary
Random Effects	b
Distribution for Random Effects	Normal
Subject Variable	HH
Optimization Technique	Dual Quasi-Newton
Integration Method	Adaptive Gaussian Quadrature

	Dimensions
Observations Used	8532
Observations Not Used	32
Total Observations	8564
Subjects	4662
Max Obs Per Subject	4
Parameters	2
Quadrature Points	5

Iter	Calls	Iteration History			
		NegLogLike	Diff	MaxGrad	Slope
1	2	2599.27784	879.9741	207.2481	-15933
2	4	2476.0613	123.2165	52.73273	-69.296
3	6	2408.63443	67.42687	10.91307	-48.9939
4	8	2398.94442	9.690011	7.353038	-7.55216
5	10	2398.55311	0.391314	2.985487	-0.44299
6	12	2398.51796	0.035144	0.95577	-0.03342
7	14	2398.51452	0.003439	0.041302	-0.00531
8	16	2398.51451	9.34E-6	0.000327	-0.00002

NOTE: GCONV convergence criterion satisfied.

- ▷ There is a random effect now, and consequently the so-called ‘adaptive Gaussian quadrature’ method, for numerical integration is used.
The method is efficient but time consuming.
- ▷ The iteration process has been relatively straightforward.

Parameter Estimates								
Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
beta0	4.3770	0.1647	4661	26.57	<.0001	0.05	4.0541	4.6999
tau2	7.8282	0.6424	4661	12.19	<.0001	0.05	6.5688	9.0875

Additional Estimates								
Label	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
mean	0.9876	0.002018	4661	489.31	<.0001	0.05	0.9836	0.9915

- ▷ The model is the same as in the GLIMMIX case, but the estimates are totally different.
- ▷ Let us bring together several estimates for the clustered-data case and for Belgium:

Method	Procedure	Estimate (s.e.)	
		$\hat{\beta}$	$\hat{\pi}$
Marginal approaches			
logistic	SURVEYMEANS	—	0.9035 (0.0040)
logistic	SURVEYLOGISTIC	2.2372 (0.0455)	0.9035 (0.0040)
GEE	GENMOD	2.1504 (0.0435)	0.8957 (0.0040)
Random-effects approaches			
GLMM	GLIMMIX	2.3723 (0.0443)	0.9147 (0.0035)
GLMM	NLMIXED	4.3770 (0.1647)	0.9876 (0.0020)

- ▷ This difference is spectacular and requires careful qualification.
- ▷ Note that the ‘true’ value is the number of people in the dataset with a stable GP divided by the total number of people:

$$\text{pragmatic estimate of } \pi = \frac{7709}{7709 + 823} = 0.9035$$

which, of course, is in agreement with all of the SRS analyses.

▷ Further:

- * The survey-design based procedures are spot on.
- * GEE is a little different, but close.
- * GLIMMIX is a little different, but close, with the deviation going the other way.
- * NLMIXED is spectacularly different.

▷ The strong differences can be explained as follows:

- * Consider our GLMM:

$$Y_{ij}|b_i \sim \text{Bernoulli}(\pi_{ij}), \quad \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \beta_0 + b_i$$

- * The **conditional means** $E(Y_{ij}|b_i)$, are given by

$$E(Y_{ij}|b_i) = \frac{\exp(\beta_0 + b_i)}{1 + \exp(\beta_0 + b_i)}$$

- * The **marginal means** are now obtained from averaging over the random effects:

$$E(Y_{ij}) = E[E(Y_{ij}|b_i)] = E\left[\frac{\exp(\beta_0 + b_i)}{1 + \exp(\beta_0 + b_i)}\right] \neq \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$$

- ▷ Hence, the parameter vector β in the GEE model needs to be interpreted completely differently from the parameter vector β in the GLMM:
 - * GEE: marginal interpretation
 - * GLMM: conditional interpretation, conditionally upon level of random effects
- ▷ In general, the model for the marginal average is not of the same parametric form as the conditional average in the GLMM.

- ▷ For logistic mixed models, with normally distributed random random intercepts, it can be shown that the marginal model can be well approximated by again a logistic model, but with parameters approximately satisfying

$$\frac{\hat{\beta}^{\text{RE}}}{\hat{\beta}^{\text{M}}} = \sqrt{c^2 \tau^2 + 1} > 1, \quad \tau^2 = \text{variance random intercepts}$$

$$c = 16\sqrt{3}/(15\pi)$$

- ▷ For our case:

$$\frac{\hat{\beta}^{\text{RE}}}{\hat{\beta}^{\text{M}}} = \frac{4.3770}{2.1504} = 2.0354$$

$$\sqrt{c^2 \tau^2 + 1} = \sqrt{0.5881^2 \times 7.3232 + 1} = 1.8795$$

- ▷ The relationship is not exact, but sufficiently close.

- ▷ The interpretation of the random-effects-based β is:
The logit of having a stable GP for someone with HH-level effect $b_i = 0$.
 - ▷ The interpretation of the random-effects-based π is:
The probability of having a stable GP for someone with HH-level effect $b_i = 0$.
 - ▷ Thus, the probability corresponding to the average household is different from the probability averaged over all households.
 - ▷ All of these relationships would also hold for the GLIMMIX procedure, if it were not so biased!
-
- We can further expand the summary table for SGP with our new analyses:

Stable General Practitioner (0/1) — Marginal and Random-effects Models						
Analysis	Procedure	Par.	Belgium	Brussels	Flanders	Wallonia
SRS	SURVEYMEANS	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	SURVEYLOGISTIC	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	SURVEYLOGISTIC	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GENMOD	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GENMOD	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GLIMMIX	β	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GLIMMIX	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	NLMIXED	β	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	NLMIXED	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
Strat.	SURVEYMEANS	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Strat.	SURVEYLOGISTIC	$-\beta$	2.3272(0.0358)	1.4219(0.0050)	2.9936(0.0859)	2.7278(0.0758)
Strat.	SURVEYLOGISTIC	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Clust.	SURVEYMEANS	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	SURVEYLOGISTIC	$-\beta$	2.2372(0.0455)	1.4219(0.0624)	2.9936(0.1037)	2.7278(0.0918)
Clust.	SURVEYLOGISTIC	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	GENMOD	$-\beta$	2.1504(0.0435)	1.3784(0.0591)	2.9188(0.1019)	2.6470(0.0890)
Clust.	GENMOD	π	0.8957(0.0040)	0.7987(0.0095)	0.9488(0.0050)	0.9338(0.0055)
Clust.	GLIMMIX	β	2.3723(0.0441)	1.5213(0.0628)	3.1433(0.0988)	—
Clust.	GLIMMIX	π	0.9147(0.0034)	0.8207(0.0092)	0.9586(0.0039)	—
Clust.	NLMIXED	β	4.3770(0.1647)	3.4880(0.3134)	8.4384(1.5434)	6.9047(0.8097)
Clust.	NLMIXED	π	0.9876(0.0020)	0.9703(0.0090)	0.9998(0.0003)	0.9990(0.0008)

Stable General Practitioner (0/1) — Marginal and Random-effects Models						
Analysis	Procedure	Par.	Belgium	Brussels	Flanders	Wallonia
Wgt.	SURVEYMEANS	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	SURVEYLOGISTIC	$-\beta$	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	SURVEYLOGISTIC	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	GENMOD	$-\beta$	2.6290(0.0642)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
Wgt.	GENMOD	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Wgt.	GLIMMIX	β	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	GLIMMIX	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
All	SURVEYMEANS	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
All	SURVEYLOGISTIC	$-\beta$	2.6290(0.0636)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
All	SURVEYLOGISTIC	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Cl.+Wgt.	GENMOD	$-\beta$	2.5233(0.0659)	1.2014(0.0839)	2.9693(0.1284)	2.7251(0.1186)
Cl.+Wgt.	GENMOD	π	0.9258(0.0045)	0.7688(0.0149)	0.9512(0.0060)	0.9385(0.0068)
Cl.+Wgt.	GLIMMIX	β	7.8531(0.1105)	5.1737(0.1906)	9.8501(0.1962)	8.7535(0.1850)
Cl.+Wgt.	GLIMMIX	π	0.9996(0.0000)	0.9944(0.0011)	0.9999(0.0000)	0.9998(0.0000)

- In summary, we note the following:
 - ▷ Compared to the marginal approaches, β and π are not generally interpretable as meaningful population quantities.
 - ▷ In some cases, this has even lead to estimation issues:

- * When parameters are unstable and or diverge, one may need to include the **PARMS** statement into the NLMIXED code. For example,

```
PARMS beta0=3.0 tau2=4.0;
```

- * Nevertheless, the NLMIXED based estimates for π approach the boundary of the $[0, 1]$ interval when clustering is accounted for.
- ▷ It is possible to derive the marginal parameters, but this involves extra numerical integration.
- ▷ Relative to the integration-based NLMIXED estimates, the GLIMMIX estimates are biased downwards.
- ▷ Important uses for the GLMM method:
 - * When estimates are required at more than one level at the same time, e.g., **town** and/or **HH** and/or **individual**.
 - * As a flexible tool for **regression**, rather than for simple population-level estimates (means, totals).

Part VIII

Weighting

18.5 General Concepts and Design

- ▷ The concept of weighting
- ▷ Weighting in the context of stratification
- ▷ Weighting in the context of clustering
- ▷ Selection proportional to size (PPS)
- ▷ Self-weighting
- ▷ Examples

18.6 General Principles

- Weighting arises naturally in a variety of contexts:
 - ▷ **With stratification:** different strata have different selection probabilities.
 - ▷ **With clustering:** weights differ within and between clusters.
 - ▷ **In general:** units are given probabilities of selection, e.g., proportional to their size.
- We will consider the main ones in turn.

- Estimators for averages and total then take the form:

$$\bar{y} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i},$$

$$\hat{y} = N \cdot \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}.$$

- The unweighted expressions result from setting all w_i equal to a constant.
Due to the division by the sum of weights, the actual constant is not important, but sensible choices are 1 or $1/n$.

18.7 Weighting and Stratification

- There are two main reasons why selection probabilities are different between strata:
 - ▷ A subgroup is of interest and not oversampling would lead to too small a sample size.
Example: German Region in the Belgian HIS.
 - ▷ Strata are given equal sample sizes for comparative purposes, but also an estimate for the entire population is required.
Example: Brussels, Flanders, and Wallonia in the Belgian HIS.
 - ▷ Units are then reweighted to ensure proper representativity.

18.7.1 Example

- Suppose a certain subgroup represents 10% of the population.
- With an unweighted scheme (SRS or stratified), this group will also contribute 10% to the sample, on average.
- If we need a sample which includes 100 individuals of the subgroup, then a total sample of 1000 individuals has to be selected.
- Enlarging the subgroup with 50% implies scaling up from 100 to 150, and hence 500 additional interviews for the entire sample are needed.
- It is perfectly possible that 50 extra interviews in the subgroup are essential, but that the other 450 are redundant.

- A solution is to increase the selection probability for the subgroup, relative to the others.

Quantity	Majority	Minority
Population	4500	500
Percentage	90	10
Sample portion	1/10	1/5
Number selected	450	100
Unweighted percentage in sample	81.8	18.2
Weight	1	1/2
Weighted number in sample	450	50
Weighted percentage in sample	90	10

- Unfortunately, it is not always possible to pre-determine whether a respondent belongs to the majority or to the minority.
- This implies that determining the weight is difficult.
- As a surrogate, entire quarters (or other geographical entities) which are known to have large minority populations can be oversampled.
- This procedure works, since the weighting is done at the quarter level, hence producing correct weights, such as in the example above.
- If one calculates the subsample selection probability carefully, then it can be ensured that the sample will contain a sufficient number of minority members.

18.7.2 Example: Artificial Population

- In Section 13.4, stratification was considered

$$\mathcal{P}_{s1} = (1\ 2 \mid 3\ 4)$$

$$\mathcal{P}_{s2} = (1\ 4 \mid 2\ 3)$$

- Samples were selected proportional to the stratum size: 1 out of 2 units in each:
 $n = (1, 1)$.
- Consider a third stratification:

$$\mathcal{P}_{s3} = (1 \mid 2\ 3\ 4)$$

- Retain the sample size $n = (1, 1)$

- The sampling mechanisms then are:

P_s					
Stratified					
s	Sample	SRS	\mathcal{P}_{s1}	\mathcal{P}_{s2}	\mathcal{P}_{s3}
1	{1,2}	1/6	0	1/4	1/3
2	{1,3}	1/6	1/4	1/4	1/3
3	{1,4}	1/6	1/4	0	1/3
4	{2,3}	1/6	1/4	0	0
5	{2,4}	1/6	1/4	1/4	0
6	{3,4}	1/6	0	1/4	0

- The corresponding estimators are:

\widehat{y}					
Stratified					
s	Sample	SRS	\mathcal{P}_{s1}	\mathcal{P}_{s2}	\mathcal{P}_{s3}
1	{1,2}	6		6	7
2	{1,3}	8	8	8	10
3	{1,4}	10	10		13
4	{2,3}	10	10		
5	{2,4}	12	12	12	
6	{3,4}	14		14	

- The expectations for the total:

$$\mathcal{P}_{s1} : E(\bar{y}) = \frac{1}{4} \cdot [8 + 10 + 10 + 12] = 10$$

$$\mathcal{P}_{s2} : E(\bar{y}) = \frac{1}{4} \cdot [6 + 8 + 12 + 14] = 10$$

$$\mathcal{P}_{s2} : E(\bar{y}) = \frac{1}{4} \cdot [7 + 10 + 13] = 10$$

- Hence, also the third stratification produces an unbiased estimator.

- Very important:

The estimates differ depending on the sampling mechanism.

- Indeed, the sample $\{1, 2\}$ produces 6 in the unweighted case and 7 in this weighted case.
- This is because the weighted expression is used. For example:

$$\hat{y} = 4 \cdot \frac{\frac{1}{1/1} + \frac{2}{1/3}}{\frac{1}{1/1} + \frac{1}{1/3}}.$$

- The weights are the inverse of the selection probability.

- The variances for SRS (without), SRS (with), and STRAT:

$$\begin{aligned}\text{SRS (without)} : \sigma_y^2 &= \frac{(6-10)^2 + (8-10)^2 + (10-10)^2 + (10-10)^2 + (12-10)^2 + (14-10)^2}{6} \\ &= \frac{40}{6} = 6.667\end{aligned}$$

$$\begin{aligned}\text{SRS (with)} : \frac{2}{16} \cdot [(6-10)^2 + (8-10)^2 + (10-10)^2 + (10-10)^2 + (12-10)^2 + (14-10)^2] \\ + \frac{1}{16} \cdot [(4-10)^2 + (8-10)^2 + (12-10)^2 + (16-10)^2] = \frac{160}{16} = 10.000\end{aligned}$$

$$\mathcal{P}_{s1} : \sigma_y^2 = \frac{(8-10)^2 + (10-10)^2 + (10-10)^2 + (12-10)^2}{4} = \frac{8}{4} = 2.000$$

$$\mathcal{P}_{s2} : \sigma_y^2 = \frac{(6-10)^2 + (8-10)^2 + (12-10)^2 + (14-10)^2}{4} = \frac{40}{4} = 10.000$$

$$\mathcal{P}_{s3} : \sigma_y^2 = \frac{(7-10)^2 + (10-10)^2 + (13-10)^2}{3} = \frac{18}{3} = 6.000$$

18.8 Weighting and Multi-Stage Sampling / Clustering

- In multi-stage sampling and clustering, subunits may be selected with differential probabilities.

Example: Household members in the Belgian HIS.

- In addition, entire clusters may be selected with variable probabilities.

Example: Towns in the Belgian HIS.

- Just like in the stratified case, this needs to be taken into account via weights.

18.8.1 Example

- Consider a selection of households from a population with two household types:
 - ▷ 1000 2-person households of married couples.
 - ▷ 1000 1-person households of singles.
- Obviously:
 - ▷ 50% of the **households** consist of married couples.
 - ▷ 66.7% of the **people** are married.
- Select a sample of 100 households, and then one person per household.
- We expect, on average, in the sample:
 - ▷ 50 married persons.
 - ▷ 50 unmarried persons.

- If the survey question is: “Are your married?” then a naive estimate would produce: $\hat{z} = 50\%$ are married, which is wrong.
- Weighting the answers by the relative selection probabilities:

$$\hat{z}_1 = \frac{50 \cdot 1 \cdot \frac{1}{1/2} + 50 \cdot 0 \cdot \frac{1}{1/1}}{50 \cdot \frac{1}{1/2} + 50 \cdot \frac{1}{1/1}} = \frac{100}{150} = 0.667$$

- In case we want to assess the proportion of married households, then no weighting is necessary:

$$\hat{z}_2 = \frac{50 \cdot 1 + 50 \cdot 0}{50 + 50} = \frac{50}{100} = 0.5$$

18.8.2 Example: Artificial Population

- In Section 16.5 we considered three ways of clustering:

$$\mathcal{P}_{c1} = (\{1, 3\}, \{2, 4\})$$

$$\mathcal{P}_{c2} = (\{1, 2\}, \{3, 4\})$$

$$\mathcal{P}_{c3} = (\{1, 4\}, \{2, 3\})$$

- Let us add another one:

$$\mathcal{P}_{c4} = (\{1\}, \{2, 3, 4\})$$

- The sampling mechanisms for the original clusterings were:

P_s					
Clustering					
s	Sample	SRS	\mathcal{P}_1	\mathcal{P}_2	\mathcal{P}_3
1	{1,2}	1/6	0	1/2	0
2	{1,3}	1/6	1/2	0	0
3	{1,4}	1/6	0	0	1/2
4	{2,3}	1/6	0	0	1/2
5	{2,4}	1/6	1/2	0	0
6	{3,4}	1/6	0	1/2	0

- We cannot merely add the new samples, since they have a different, and in fact **differing** sample size:

$$\mathcal{S}_{c4} = \{ \{1\}, \{2, 3, 4\} \}$$

- Let us decide to change the selection probabilities so as to comply with **selection proportional to size (PPS)**:

s	Sample	P_s	\hat{y}
1	$\{1\}$	$1/4$	4
2	$\{2,3,4\}$	$3/4$	12

- The expectation of the total:

$$\mathcal{P}_{c4} : E(\bar{y}) = \frac{\frac{1}{4} \times 4 + \frac{3}{4} \times 12}{\frac{1}{4} + \frac{3}{4}} = 10$$

- The variances for SRS (without), SRS (with), SYS, STRAT, and CLUST:

$$\begin{aligned}\text{SRS (without)} : \sigma_y^2 &= \frac{(6-10)^2 + (8-10)^2 + (10-10)^2 + (10-10)^2 + (12-10)^2 + (14-10)^2}{6} \\ &= \frac{40}{6} = 6.667\end{aligned}$$

$$\begin{aligned}\text{SRS (with)} : \frac{2}{16} \cdot [(6-10)^2 + (8-10)^2 + (10-10)^2 + (10-10)^2 + (12-10)^2 + (14-10)^2] \\ + \frac{1}{16} \cdot [(4-10)^2 + (8-10)^2 + (12-10)^2 + (16-10)^2] = \frac{160}{16} = 10.000\end{aligned}$$

$$\mathcal{P}_{c1} : \sigma_y^2 = \frac{(8-10)^2 + (12-10)^2}{2} = \frac{8}{2} = 4.000$$

$$\mathcal{P}_{c2} : \sigma_y^2 = \frac{(6-10)^2 + (14-10)^2}{2} = \frac{32}{2} = 16.000$$

$$\mathcal{P}_{c3} : \sigma_y^2 = \frac{(10-10)^2 + (10-10)^2}{2} = \frac{0.0}{2} = 0.000$$

$$\mathcal{P}_{c4} : \sigma_y^2 = \frac{1}{4}(4-10)^2 + \frac{3}{4}(12-10)^2 = 9 + 3 = 12.000$$

18.8.3 Example: Surveytown

- In Section 17.4, two clusterings were added to the designs already considered prior to that section:

$$\mathcal{P}_{cX} = \left(\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\} \right)$$

$$\mathcal{P}_{cZ} = \left(\{1, 7\}, \{2, 6\}, \{3, 5\}, \{4, 8\} \right)$$

- Samples of size $n = 2$ evidently were composed of a single cluster.
- The list of samples, next to some of the other designs (stratification not shown, but to be found in Sections 13.5 and 14.5):

s	Sample	P_s					\hat{y}_s				
		SRS	Systematic		Clustered		SRS	Systematic		Clustered	
			\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{cX}	\mathcal{P}_{cZ}		\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{cX}	\mathcal{P}_{cZ}
1	{1,2}	1/28	0	0	1/4	0	12			12	
2	{1,3}	1/28	0	0	0	0	16				
3	{1,4}	1/28	0	0	0	0	20				
4	{1,5}	1/28	1/4	0	0	0	24	24			
5	{1,6}	1/28	0	1/4	0	0	28			28	
6	{1,7}	1/28	0	0	0	1/4	32				32
7	{1,8}	1/28	0	0	0	0	36				
8	{2,3}	1/28	0	0	0	0	20				
9	{2,4}	1/28	0	0	0	0	24				
10	{2,5}	1/28	0	0	0	0	28				
11	{2,6}	1/28	1/4	0	0	1/4	32	32			32
12	{2,7}	1/28	0	1/4	0	0	36		36		
13	{2,8}	1/28	0	0	1/4	0	40		40		
14	{3,4}	1/28	0	0	0	1/4	28				28
15	{3,5}	1/28	0	0	0	0	32				
16	{3,6}	1/28	0	0	0	0	36				

<i>s</i>	Sample	P_s					\hat{y}_s				
		SRS	Systematic		Clustered		SRS	Systematic		Clustered	
			\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{cX}	\mathcal{P}_{cZ}		\mathcal{L}_X	\mathcal{L}_Z	\mathcal{P}_{cX}	\mathcal{P}_{cZ}
17	{3,7}	1/28	1/4	0	0	0	40	40			
18	{3,8}	1/28	0	1/4	0	0	44		44		
19	{4,5}	1/28	0	1/4	0	0	36		36		
20	{4,6}	1/28	0	0	0	0	40				
21	{4,7}	1/28	0	0	0	0	44				
22	{4,8}	1/28	1/4	0	0	1/4	48	48			48
23	{5,6}	1/28	0	0	1/4	0	44			44	
24	{5,7}	1/28	0	0	0	0	48				
25	{5,8}	1/28	0	0	0	0	52				
26	{6,7}	1/28	0	0	0	0	52				
27	{6,8}	1/28	0	0	0	0	56				
28	{7,8}	1/28	0	0	1/4	0	60			60	
Expectation							36	36	36	36	36
Variance							144	80	32	320	48
Standard error							12.00	8.94	2.83	17.89	6.93

- These clusterings provided unbiased estimators.

- Variances were:

$$\mathcal{P}_{cX} : \sigma_{\bar{y}}^2 = \frac{(12 - 36)^2 + (28 - 36)^2 + (44 - 36)^2 + (60 - 36)^2}{4} = \frac{1280}{4} = 320$$

$$\mathcal{P}_{cZ} : \sigma_{\bar{y}}^2 = \frac{(32 - 36)^2 + (32 - 36)^2 + (32 - 36)^2 + (48 - 36)^2}{4} = \frac{192}{4} = 48$$

- We noted that \mathcal{P}_{cX} increases variability dramatically, while \mathcal{P}_{sZ} decreases variability, relative to SRS,

But also: that \mathcal{P}_{cX} is the more common choice in practice, with positive correlation, that we will see in practice.

- The relative positions of the methods were:

Rank	Method	Variance	ρ
1	SYS (\mathcal{L}_Z)	32	-0.81
2	CLUST (\mathcal{P}_{cZ})	48	-0.71
3	STRAT (\mathcal{P}_{sX})	40	-0.76
4	SYS (\mathcal{L}_X)	80	-0.52
5	SRS (without)	144	-0.14
6	STRAT (\mathcal{P}_{sZ})	160	-0.05
7	SRS (with)	168	0.00
8	CLUST (\mathcal{P}_{cX})	320	+0.90

- It is possible to reduce variability when using clustering, while using a more relativistic method than switching to not-being-used-in-practice \mathcal{P}_{cZ} .
- This consists of ensuring clusters are:
 - ▷ of variable size (number of blocks)
 - ▷ homogeneous in the survey variable (number of buildings)
- As an example, consider one further clustering:

$$\mathcal{P}_{c3} = \left(\{1, 2, 3\}, \{4\}, \{5\}, \{6\}, \{7a\}, \{7b\}, \{8a\}, \{8b\} \right)$$

Precisely, we:

- ▷ regroup small blocks
- ▷ leave medium sized blocks
- ▷ dissect large blocks

Cluster	Blocks	Y
1	1,2,3	6
2	4	4
3	5	5
4	6	6
5	7a	3
6	7b	4
7	8a	4
8	8b	4

- Let us first take samples of size $n = 1$:

Sample s	Blocks	Y	\hat{y}
1	1,2,3	6	48
2	4	4	32
3	5	5	40
4	6	6	48
5	7a	3	24
6	7b	4	32
7	8a	4	32
8	8b	4	32

- The expectation is:

$$E(\hat{y}) = \frac{1}{8}[48 + 32 + 40 + 48 + 24 + 32 + 32 + 32] = 36$$

- This means we have an unbiased estimator.

- The variance is:

$$\mathcal{P}_{c3} : \sigma_{\hat{y}}^2 = \frac{(48 - 36)^2 + (32 - 36)^2 + \dots + (32 - 36)^2 + (32 - 36)^2}{8} = \frac{512}{8} = 64$$

- The corresponding variance for SRS with $n = 1$ was 336.

- Let us take samples of size $n = 2$:

Sample s	Clusters	Blocks	\hat{y}
1	{1,2}	1,2,3,4	40
2	{1,3}	1,2,3,5	44
\vdots	\vdots	\vdots	\vdots
27	{6,8}	7b,8b	32
28	{7,8}	8a,8b	32

- The list of estimates is

$$\{ 40, 44, 48, 36, 40, 40, 40, 36, 40, 28, 32, 32, 32, 44, \\ 32, 36, 36, 36, 36, 40, 40, 40, 28, 28, 28, 32, 32, 32 \}$$

- The expectation easily follows as

$$E(\hat{y}) = \frac{1}{28}[40 + 44 + \cdots + 32 + 32] = 36$$

- The variance is:

$$\begin{aligned}\mathcal{P}_{c3} : \sigma_{\hat{y}}^2 &= \frac{(40 - 36)^2 + (44 - 36)^2 + \cdots + (32 - 36)^2 + (32 - 36)^2}{8} \\ &= \frac{768}{28} = 27.4286\end{aligned}$$

- The corresponding variance for SRS with $n = 2$ was 144.
- Just as before, we can calculate the within-sample correlations, which now is

$$\rho_{\mathcal{P}_{c3}} = -0.8367$$

- Placing the new estimator among the list of estimators with $n = 2$ produces:

Rank	Method	Variance	ρ
0	CLUST (\mathcal{P}_{c3})	27	-0.84
1	SYS (\mathcal{L}_Z)	32	-0.81
2	CLUST (\mathcal{P}_{cZ})	48	-0.71
3	STRAT (\mathcal{P}_{sX})	40	-0.76
4	SYS (\mathcal{L}_X)	80	-0.52
5	SRS (without)	144	-0.14
6	STRAT (\mathcal{P}_{sZ})	160	-0.05
7	SRS (with)	168	0.00
8	CLUST (\mathcal{P}_{cX})	320	+0.90

- The new estimator is the best one of all!
- Thus, selection proportional to size, as is done here through regrouping the units, can be a very powerful tool to control variability.

Chapter 19

Analysis

- ▷ Selection Proportional to Size
- ▷ Self-weighting
- ▷ Horvitz-Thompson estimator
- ▷ Examples

19.1 Selection Proportional to Size and Self-Weighting

- Define an estimator of the cluster-specific total as:

$$\widehat{y}_i = \frac{1}{f_i} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{f_i} \cdot y_i$$

- Define an estimator for the population total as:

$$\begin{aligned}\widehat{y} &= \sum_{i=1}^m \frac{1}{m} \cdot \frac{1}{\pi_i} \widehat{y}_i \\ &= \sum_{i=1}^m \frac{1}{m} \cdot \frac{1}{\pi_i} \cdot \frac{1}{f_i} \sum_{j=1}^{n_i} y_{ij} \\ &= \sum_{i=1}^m \frac{1}{m} \cdot \frac{1}{\pi_i} \cdot \frac{1}{f_i} \cdot y_i\end{aligned}$$

where

- ▷ f_i is the sample fraction in selected cluster i
- ▷ π_i is the probability to select cluster i
- ▷ y_{ij} is the value of the survey variable for subject j in cluster i

19.1.1 Self-Weighting

- Self-weighting is defined by requiring

$$f = n \cdot \pi_i \cdot f_i$$

to be constant.

- Hence, the estimator for the total reduces to:

$$\begin{aligned}\hat{y} &= \sum_{i=1}^m \frac{1}{m} \cdot \frac{1}{\pi_i} \cdot \frac{1}{f_i} \sum_{j=1}^{n_i} y_{ij} \\ &= \sum_{i=1}^m \frac{1}{f} \sum_{j=1}^{n_i} y_{ij} \\ &= \frac{1}{f} \cdot y\end{aligned}$$

- For the Belgian Health Interview Survey:

$$\pi_i \propto t_i \quad (\text{town size})$$

$$f_i \propto \frac{50}{t_i}$$

$$\Rightarrow n \cdot \pi_i \cdot f_i \propto n \cdot t_i \cdot \frac{50}{t_i} \quad \text{a constant}$$

Hence: the selection of respondents within towns is self-weighting.

19.1.2 Variances for PPS

Quantity	Expression
Pop. var. 1	$S_{1Y}^2 = \sum_{I=1}^M \pi_I \left(\frac{Y_I}{M\pi_I} - \bar{Y} \right)^2 = \frac{1}{M^2} \sum_{I=1}^M \pi_I \left(\frac{Y_I}{\pi_I} - Y \right)^2$
Pop. var. 2	$S_{2Y}^2 = \frac{\bar{N}^2}{\bar{N} - \bar{n}} \cdot \sum_{I=1}^M \frac{N_I}{N} \cdot \frac{N_I - \bar{n}}{N_I} \cdot \frac{1}{N_I - 1} \sum_{J=1}^{N_I} (Y_{IJ} - \bar{Y}_J)^2$
PPS (with)	$\sigma_{\hat{y}}^2 = \frac{M^2}{m} S_{1Y}^2 + \frac{M^2}{m} \cdot \frac{\bar{N}^2}{\bar{n}} \cdot \left(1 - \frac{\bar{n}}{\bar{N}} \right) S_{2Y}^2$
PPS (without)	$\sigma_{\hat{y}}^2 = \frac{M^2}{m} \sum_{I=1}^M \pi_I \left(\frac{1 - n\pi_I}{1 - \pi_I} \right) \cdot \left(\frac{Y_I}{M\pi_I} - \bar{Y} \right)^2 + \frac{M^2}{m} \cdot \frac{\bar{N}}{\bar{n}} \cdot \left(1 - \frac{\bar{n}}{\bar{N}} \right) S_{2Y}^2$

19.2 The Horvitz-Thompson Estimator

- The Horvitz-Thompson (HT) is general and broadly applicable.
- It can be a bit unstable at times.
- Alternatives, such as the Hansen-Hurwitz estimator exist.
- Let
 - ▷ y_i : total for cluster i (which can simply be an individual in the non-clustered case)
 - ▷ π_i : probability of selecting cluster i
 - ▷ v : number of **distinct** clusters sampled

- Note that $v \leq m$, with equality holding when sampling without replacement.
- The Hovitz-Thompson estimator takes the form:

$$\hat{y}_{HT} = \sum_{i=1}^v \frac{y_i}{\pi_i}$$

- The variance:

$$\begin{aligned} \sigma_{\hat{y}_{HT}}^2 &= \sum_{I=1}^M \frac{1 - \pi_I}{\pi_I} Y_I^2 + \sum_{I=1}^M \sum_{J \neq I} \left(\frac{\pi_{IJ} - \pi_I \pi_J}{\pi_I \pi_J} \right) Y_I Y_J \\ &= \sum_{I=1}^M \frac{1 - \pi_I}{\pi_I} Y_I^2 + 2 \sum_{I=1}^{M-1} \sum_{J=I+1}^M \left(\frac{\pi_{IJ} - \pi_I \pi_J}{\pi_I \pi_J} \right) Y_I Y_J \end{aligned}$$

with now in addition

▷ π_{IJ} : probability of **simultaneously** selecting clusters I and J into the sample.

19.3 The Artificial Population and Horvitz-Thompson

- We will consider three situations
 - ▷ SRS without replacement
 - ▷ SRS with replacement
 - ▷ Selection with unequal probabilities
- In all cases, $n = 2$ will be maintained.

19.3.1 SRS Without Replacement

- The clusters in the population are:

$$\mathcal{P} = \left\{ \{1\}, \{2\}, \{3\}, \{4\} \right\}$$

- with samples:

$$\mathcal{S} = \left\{ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\} \right\}$$

- The probability of selecting a '1' (or any other unit) is

$$\pi_I = \frac{3}{6} = \frac{1}{2}$$

- The estimator:

$$\widehat{y}_{HT} = \frac{y_1}{1/2} + \frac{y_2}{1/2} = 2(y_1 + y_2) = 2 \cdot y$$

- The variance:

$$\begin{aligned}\sigma_{\widehat{y}_{HT}}^2 &= \sum_{I=1}^4 \frac{1 - \pi_I}{\pi_I} Y_I^2 + 2 \sum_{I=1}^3 \sum_{J=I+1}^4 \left(\frac{\pi_{IJ} - \pi_I \pi_J}{\pi_I \pi_J} \right) Y_I Y_J \\ &= T_1 + T_2\end{aligned}$$

with

$$\begin{aligned}
 T_1 &= \sum_{I=1}^4 \frac{1 - 1/2}{1/2} Y_I^2 \\
 &= \sum_{I=1}^4 Y_I^2 \\
 &= 1^2 + 2^2 + 3^2 + 4^2 \\
 &= 30
 \end{aligned}$$

$$\begin{aligned}
 \pi_{IJ} &= P(\text{selecting two units simultaneously}) \\
 &= 2 \cdot \frac{1}{4} \cdot \frac{1}{3} \\
 &= \frac{1}{6}
 \end{aligned}$$

$$\frac{\pi_{IJ} - \pi_I \pi_J}{\pi_I \pi_J} = \frac{1/6 - 1/2 \times 1/2}{1/2 \times 1/2}$$

$$= -\frac{1}{3}$$

$$T_2 = -2 \times \frac{1}{3} \times (1 \cdot 2 + 1 \cdot 3 + 1 \cdot 4 + 2 \cdot 3 + 2 \cdot 4 + 3 \cdot 4)$$

$$= \frac{-2 \times 35}{3}$$

Hence,

$$\sigma_{\hat{y}_{HT}}^2 = T_1 + T_2 = 30 - \frac{70}{3} = \frac{20}{3} = 6.667$$

- In Section 3.18.1 we obtained:

$$\begin{aligned}\sigma_{\hat{y}}^2 &= \frac{1}{S} \sum_{s=1}^S \left(\hat{y}_s - \frac{1}{S} \sum_{s=1}^S \hat{y}_s \right)^2 \\&= \frac{(6.0-10)^2 + (8.0-10.0)^2 + (10.0-10.0)^2 + (10.0-10.0)^2 + (12.0-10.0)^2 + (14.0-10.0)^2}{6} \\&= \frac{40.0}{6} = 6.6667\end{aligned}$$

19.3.2 SRS With Replacement

- The clusters in the population are:

$$\mathcal{P} = \left\{ \{1\}, \{2\}, \{3\}, \{4\} \right\}$$

- with samples:

$$\mathcal{S} = \left\{ \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\} \right.$$

$$\left. \{1, 1\} \equiv \{1\}, \{2, 2\} \equiv \{2\}, \{3, 3\} \equiv \{3\}, \{4, 4\} \equiv \{4\} \right\}$$

- The probability of selecting a '1' (or any other unit) is

$$\begin{aligned}\pi_I &= \frac{1}{4} \cdot P(\text{sample with 1 element}) + \frac{1}{2} \cdot P(\text{sample with 2 elements}) \\ &= \frac{1}{4} \cdot \frac{4}{16} + \frac{1}{2} \cdot \frac{12}{16} = \frac{7}{16}\end{aligned}$$

- The estimator:

▷ In a sample with one element:

$$\hat{y}_{HT} = \frac{y_1}{7/16} = \frac{16}{7} \cdot y_1$$

▷ In a sample with two elements:

$$\hat{y}_{HT} = \frac{y_1}{7/16} + \frac{y_2}{7/16} = \frac{16}{7} \cdot (y_1 + y_2)$$

- Enumeration of the estimator:

s	Sample	P_s	\hat{y}	\hat{y}_{HT}
1	{1,2}	2/16	6.0	48/7=6.86
2	{1,3}	2/16	8.0	64/7=9.14
3	{1,4}	2/16	10.0	80/7=11.43
4	{2,3}	2/16	10.0	80/7=11.43
5	{2,4}	2/16	12.0	96/7=13.71
6	{3,4}	2/16	14.0	112/7=16.00
7	{1,1}	1/16	4.0	16/7=2.29
8	{2,2}	1/16	8.0	32/7=4.57
9	{3,3}	1/16	12.0	48/7=6.86
10	{4,4}	1/16	16.0	64/7=9.14

- The expectation of the estimator:

$$\begin{aligned}
 E(\hat{y}_{HT}) &= \frac{1}{16} \left[\frac{48}{7} + \frac{64}{7} + \frac{80}{7} + \frac{80}{7} + \frac{96}{7} + \frac{112}{7} \right] + \frac{2}{16} \left[\frac{16}{7} + \frac{32}{7} + \frac{48}{7} + \frac{64}{7} \right] \\
 &= \frac{70}{7} \\
 &= 10
 \end{aligned}$$

- Thus, the estimator is unbiased, but different from before.
- The variance:

$$\begin{aligned}
 \sigma_{\hat{y}_{HT}}^2 &= \sum_{I=1}^4 \frac{1 - \pi_I}{\pi_I} Y_I^2 + 2 \sum_{I=1}^3 \sum_{J=I+1}^4 \left(\frac{\pi_{IJ} - \pi_I \pi_J}{\pi_I \pi_J} \right) Y_I Y_J \\
 &= T_1 + T_2
 \end{aligned}$$

with

$$\begin{aligned} T_1 &= \sum_{I=1}^4 \frac{1 - 7/16}{7/16} Y_I^2 \\ &= \frac{9}{7}(1^2 + 2^2 + 3^2 + 4^2) \\ &= \frac{270}{7} \end{aligned}$$

$$\begin{aligned} \pi_{IJ} &= P(\text{selecting two units simultaneously}) \\ &= \frac{2}{16} \end{aligned}$$

$$\frac{\pi_{IJ} - \pi_I \pi_J}{\pi_I \pi_J} = \frac{2/16 - 7/16 \times 7/16}{7/16 \times 7/16}$$

$$= -\frac{15}{49}$$

$$T_2 = -2 \times \frac{15}{49} \times (1 \cdot 2 + 1 \cdot 3 + 1 \cdot 4 + 2 \cdot 3 + 2 \cdot 4 + 3 \cdot 4)$$

$$= -\frac{2 \times 15 \times 35}{7}$$

$$= -\frac{150}{7}$$

Hence,

$$\sigma_{\hat{y}_{HT}}^2 = T_1 + T_2 = \frac{270 - 150}{7} = \frac{120}{7} = 17.143$$

- In Section 3.18.1 we obtained:

$$\begin{aligned}\sigma_{\hat{y}}^2 &= \sum_{s=1}^S P_s \left(\hat{y}_s - \sum_{s=1}^S P_s \hat{y}_s \right)^2 \\&= \frac{2}{16} \cdot [(6.0 - 10.0)^2 + (8.0 - 10.0)^2 + (10.0 - 10.0)^2 + (10.0 - 10.0)^2 + (12.0 - 10.0)^2 + (14.0 - 10.0)^2] \\&\quad + \frac{1}{16} \cdot [(4.0 - 10.0)^2 + (8.0 - 10.0)^2 + (12.0 - 10.0)^2 + (16.0 - 10.0)^2] \\&= \frac{160.0}{16} = 10.0\end{aligned}$$

- Hence, the HT estimator is different and less efficient than the ordinary SRS estimator with replacement.

19.3.3 Selection With Unequal Probabilities

- Consider the following set of selection probabilities for the units:

Unit	p_i
1	1/2
2	1/6
3	1/6
4	1/6

- Probability of selecting the various samples:

Sample	p_s	Sample	p_s
$\{1,2\}$	$1/2 \times 1/3 = 1/6$	$\{3,1\}$	$1/6 \times 3/5 = 1/10$
$\{1,3\}$	$1/2 \times 1/3 = 1/6$	$\{3,2\}$	$1/6 \times 1/5 = 1/30$
$\{1,4\}$	$1/2 \times 1/3 = 1/6$	$\{3,4\}$	$1/6 \times 1/5 = 1/30$
$\{2,1\}$	$1/6 \times 3/5 = 1/10$	$\{4,1\}$	$1/6 \times 3/5 = 1/10$
$\{2,3\}$	$1/6 \times 1/5 = 1/30$	$\{4,2\}$	$1/6 \times 1/5 = 1/30$
$\{2,4\}$	$1/6 \times 1/5 = 1/30$	$\{4,3\}$	$1/6 \times 1/5 = 1/30$

- The probabilities of selecting the various units into the samples:

$$\pi_1 = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{10} + \frac{1}{10} + \frac{1}{10} = \frac{4}{5}$$

$$\pi_2 = \pi_3 = \pi_4 = \frac{1}{6} + \frac{1}{10} + \frac{1}{30} + \frac{1}{30} + \frac{1}{30} + \frac{1}{30} = \frac{2}{5}$$

- The estimator:

Sample	\hat{y}_{HT}	π_{IJ}
$\{1,2\}$	$\frac{1}{4/5} + \frac{2}{2/5} = \frac{25}{4}$	$\frac{1}{6} + \frac{1}{10} = \frac{4}{15}$
$\{1,3\}$	$\frac{1}{4/5} + \frac{3}{2/5} = \frac{35}{4}$	$\frac{1}{6} + \frac{1}{10} = \frac{4}{15}$
$\{1,4\}$	$\frac{1}{4/5} + \frac{4}{2/5} = \frac{45}{4}$	$\frac{1}{6} + \frac{1}{10} = \frac{4}{15}$
$\{2,3\}$	$\frac{2}{2/5} + \frac{2}{2/5} = \frac{50}{4}$	$\frac{1}{30} + \frac{1}{30} = \frac{1}{15}$
$\{2,4\}$	$\frac{2}{2/5} + \frac{4}{2/5} = \frac{60}{4}$	$\frac{1}{30} + \frac{1}{30} = \frac{1}{15}$
$\{3,4\}$	$\frac{3}{2/5} + \frac{4}{2/5} = \frac{70}{4}$	$\frac{1}{30} + \frac{1}{30} = \frac{1}{15}$

- The expectation of the estimator:

$$\begin{aligned}
 E(\hat{y}_{HT}) &= \frac{4}{15} \times \left(\frac{25}{4} + \frac{35}{4} + \frac{45}{4} \right) + \frac{1}{15} \times \left(\frac{50}{4} + \frac{60}{4} + \frac{70}{4} \right) \\
 &= \frac{600}{60} \\
 &= 10
 \end{aligned}$$

- The variance:

$$\begin{aligned}
 \sigma_{\hat{y}_{HT}}^2 &= \sum_{I=1}^4 \frac{1 - \pi_I}{\pi_I} Y_I^2 + 2 \sum_{I=1}^3 \sum_{J=I+1}^4 \left(\frac{\pi_{IJ} - \pi_I \pi_J}{\pi_I \pi_J} \right) Y_I Y_J \\
 &= T_1 + T_2
 \end{aligned}$$

with

$$\begin{aligned}
T_1 &= \left(\frac{1 - 4/5}{4/5} \right) \cdot 1^2 + \left(\frac{1 - 2/5}{2/5} \right) \cdot (2^2 + 3^2 + 4^2) \\
&= \frac{175}{4}
\end{aligned}$$

$$\begin{aligned}
T_2 &= 2 \cdot \left(\frac{\pi_{1J} - \pi_1 \pi_J}{\pi_1 \pi_J} \right) \cdot (1 \cdot 2 + 1 \cdot 3 + 1 \cdot 4) + 2 \cdot \left(\frac{\pi_{IJ} - \pi_I \pi_J}{\pi_I \pi_J} \right)_{I,J \geq 2} \cdot (2 \cdot 3 + 2 \cdot 4 + 3 \cdot 4) \\
&= 2 \cdot \left(\frac{4/15 - 4/5 \times 2/5}{4/5 \times 2/5} \right) \cdot (1 \cdot 2 + 1 \cdot 3 + 1 \cdot 4) + 2 \cdot \left(\frac{1/15 - 2/5 \times 2/5}{2/5 \times 2/5} \right) \cdot (2 \cdot 3 + 2 \cdot 4 + 3 \cdot 4) \\
&= 2 \left(-\frac{1}{6} \times 9 - \frac{7}{12} \times 26 \right) = -\frac{100}{3}
\end{aligned}$$

Hence,

$$\sigma_{\hat{y}_{HT}}^2 = T_1 + T_2 = \frac{175}{4} - \frac{100}{3} = \frac{125}{12} = 10.417$$

Chapter 20

Example: The Belgian Health Interview Survey

- ▷ Design-based estimation for LNBMI, LNVOEG, GHQ12, and SGP
- ▷ Regression-based estimation for the continuous LNBMI
- ▷ Logistic regression-based estimation for the binary SGP

20.1 Estimation of Means

- Taking weighting into account, the means are recomputed for

- ▷ LNBMI
- ▷ LNVOEG
- ▷ GHQ12
- ▷ SGP

- The following program can be used:

```
proc surveymeans data=m.bmi_voeg mean stderr;  
title 'weighted means - infinite population for Belgium and regions';  
where (regionch^='');  
domain regionch;  
weight wfin;  
var lnbmi lnvoeg ghq12 sgp;  
run;
```

- The program includes the weights by means of the **WEIGHT** statement.
- While it would be possible to include a finite sample correction, as we have seen, the impact is so negligible that it has been omitted.
- The output takes the usual form, with weighting information listed:

weighted means - infinite population for Belgium and regions
The SURVEYMEANS Procedure

Data Summary

Number of Observations 8564
Sum of Weights 6957597.07

Statistics

Variable	Mean	Std Error of Mean
LNBMI	3.185356	0.002651
LNVOEG	1.634690	0.013233
GHQ12	1.626201	0.044556
SGP	0.932702	0.003498

Domain Analysis: REGIONCH			
REGIONCH	Variable	Mean	Std Error of Mean
Brussels	LNBMI	3.171174	0.004578
	LNVOEG	1.802773	0.021831
	GHQ12	1.924647	0.076313
	SGP	0.782448	0.011563
Flanders	LNBMI	3.180865	0.003870
	LNVOEG	1.511927	0.019155
	GHQ12	1.445957	0.061910
	SGP	0.954757	0.004722
Walloonia	LNBMI	3.198131	0.004238
	LNVOEG	1.803178	0.020426
	GHQ12	1.858503	0.078566
	SGP	0.943191	0.005417

- Note that the weights were chosen so that they recombine the entire population.
- The fact that the sum is not around 10 million is due to empty strata.
- The sum of the weights does not matter for genuine survey procedures, such as the SURVEYMEANS procedure used here.

- It does matter for some of the model-based procedures, as we will see further in this chapter.
- We summarize the results and compare them to SRS (and still foreshadow a bit):

Logarithm of Body Mass Index				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	3.187218(0.001845)	3.175877(0.003372)	3.182477(0.002993)	3.201530(0.003216)
Stratification	3.187218(0.001840)	3.175877(0.003373)	3.182477(0.002989)	3.201530(0.003217)
Clustering	3.187218(0.001999)	3.175877(0.003630)	3.182477(0.003309)	3.201530(0.003429)
Weighting	3.185356(0.002651)	3.171174(0.004578)	3.180865(0.003870)	3.198131(0.004238)
All combined	3.185356(0.003994)	3.171174(0.004844)	3.180865(0.004250)	3.198131(0.004403)

Logarithm of VOG Score				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	1.702951(0.008954)	1.809748(0.016203)	1.516352(0.015201)	1.801107(0.014550)
Stratification	1.702951(0.008801)	1.809748(0.016206)	1.516352(0.015207)	1.801107(0.014427)
Clustering	1.702951(0.010355)	1.809748(0.018073)	1.516352(0.017246)	1.801107(0.016963)
Weighting	1.634690(0.013233)	1.802773(0.021831)	1.511927(0.019155)	1.803178(0.020426)
All combined	1.634690(0.014855)	1.802773(0.023135)	1.511927(0.021409)	1.803178(0.023214)

General Health Questionnaire – 12				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	1.661349(0.029584)	1.862745(0.056894)	1.385381(0.046246)	1.772148(0.051023)
Stratification	1.661956(0.029452)	1.864301(0.056939)	1.385857(0.046211)	1.772148(0.050823)
Clustering	1.661349(0.032824)	1.862745(0.062739)	1.385381(0.052202)	1.772148(0.055780)
Weighting	1.626201(0.044556)	1.924647(0.076313)	1.445957(0.061910)	1.858503(0.078566)
All combined	1.626781(0.048875)	1.924647(0.080508)	1.446286(0.068931)	1.858503(0.084047)

Stable General Practitioner (0/1)				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	0.903540(0.003196)	0.805632(0.007826)	0.952285(0.003908)	0.938646(0.004382)
Stratification	0.903540(0.003116)	0.805632(0.007827)	0.952285(0.003902)	0.938646(0.004366)
Clustering	0.903540(0.003963)	0.805632(0.009766)	0.952285(0.004709)	0.938646(0.005284)
Weighting	0.932702(0.003498)	0.782448(0.011563)	0.954757(0.004722)	0.943191(0.005417)
All combined	0.932702(0.003994)	0.782448(0.013836)	0.954757(0.005379)	0.943191(0.006159)

20.1.1 Discussion

- Unlike with stratification and clustering, the impact is **major** and **differential between outcomes**.
- Recall that an unweighted analysis implicitly assumes the following **incorrect** facts:
 - ▷ the Brussels, Flemish, and Walloon populations are roughly equal
 - ▷ members within a household have roughly the same selection probability
 - ▷ (other components of the weights are relatively unimportant)
- Weighting reduces precision: this is reflected throughout in larger standard errors. *They all increase, roughly, by a factor 1.5.*

- Let us discuss each of the four outcomes:

▷ **LNBMI:**

- * The regional estimates are relatively stable.
- * The Belgian estimate is stable, too.
- * This is a coincidence, as can be seen from the following rounded computations:

$$\text{General: } \hat{\mu}_{\text{Bel}} = w_{\text{Bru}}\hat{\mu}_{\text{Bru}} + w_{\text{Fla}}\hat{\mu}_{\text{Fla}} + w_{\text{Wal}}\hat{\mu}_{\text{Wal}}$$

$$\text{Unweighted: } \hat{\mu}_{\text{Bel}} = \frac{1}{3}3.18 + \frac{1}{3}3.18 + \frac{1}{3}3.20 = 3.1867$$

$$\text{Weighted: } \hat{\mu}_{\text{Bel}} = \frac{1}{10}3.18 + \frac{6}{10}3.18 + \frac{3}{10}3.20 = 3.1860$$

- * Hence, the weights shift a lot between Flanders and Brussels, but these regions have the same average, as a coincidence.

▷ LNVOEG:

* Here, the situation is rather different:

$$\text{General: } \hat{\mu}_{\text{Bel}} = w_{\text{Bru}}\hat{\mu}_{\text{Bru}} + w_{\text{Fla}}\hat{\mu}_{\text{Fla}} + w_{\text{Wal}}\hat{\mu}_{\text{Wal}}$$

$$\text{Unweighted: } \hat{\mu}_{\text{Bel}} = \frac{1}{3}1.8 + \frac{1}{3}1.5 + \frac{1}{3}1.8 = 1.7$$

$$\text{Weighted: } \hat{\mu}_{\text{Bel}} = \frac{1}{10}1.8 + \frac{6}{10}1.5 + \frac{3}{10}1.8 = 1.6$$

- * Since the two smaller regions have a higher average, the unweighted Belgian average is higher than the weighted Belgian average.
- * This also implies there is a larger impact on the standard error for Belgium.
The standard errors for the regions increase with 35, 26, and 40%, while the standard error for Belgium increases with 48%, more than for each of the regions separately.

This is because there are two sources of additional variation: (1) variability in the weights; (2) variability between the regional means.

▷ GHQ-12:

- * The phenomenon is similar to what was observed for LNVOEG.

▷ SGP:

- * The phenomenon is not as extreme, since Brussels and Wallonia are rather different: they do not reinforce each other.
- * But still, weighting downplays the low Brussels estimate and upgrades the high Flemish estimate, producing a higher Belgian average.

20.2 Regression-Based Estimation for LNMBI

- Like before, the procedures SURVEYREG and MIXED can be used to take weighting into account.
- PROC SURVEYREG code is:

```
proc surveyreg data=m.bmi_voeg;  
title '15. Mean. Surveyreg, weighted, for Belgium';  
weight wfin;  
model lnbmi = ;  
run;
```

with straightforward syntax and output (for Belgium):

Estimated Regression Coefficients

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.18535629	0.00265138	1201.39	<.0001

- PROC MIXED code is:

```
proc mixed data=m.bmi_voeg method=reml;  
title '25. Survey mean with PROC MIXED, for Belgium;  
title2 'weighted';  
where (regionch^='');  
weight wfin;  
model lnbmi = / solution;  
run;
```

- There is no need for a RANDOM statement, since no clustering is taken into account.
- The relevant portion of the output for Belgium is:

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	3.1854	0.001836	8383	1734.72	<.0001

- While the estimate is similar, the standard error is considerably smaller.
- An overview of the results:

Logarithm of Body Mass Index					
Analysis	Procedure	Belgium	Brussels	Flanders	Wallonia
SRS	SURVEYMEANS	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
SRS	MIXED	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
Stratification	SURVEYMEANS	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
Clustering	SURVEYMEANS	3.1872(0.0020)	3.1759(0.0036)	3.1825(0.0033)	3.2015(0.0034)
Clustering	MIXED	3.1880(0.0020)	3.1761(0.0036)	3.1840(0.0033)	3.2022(0.0034)
Weighting	SURVEYMEANS	3.1853(0.0027)	3.1712(0.0046)	3.1809(0.0039)	3.1981(0.0042)
Weighting	MIXED	3.1854(0.0018)	3.1712(0.0034)	3.1809(0.0030)	3.1981(0.0032)
All combined	SURVEYMEANS	3.1853(0.0040)	3.1712(0.0048)	3.1809(0.0043)	3.1981(0.0044)
Clust+Wgt	MIXED	3.1865(0.0023)	3.1706(0.0039)	3.1817(0.0036)	3.1994(0.0038)

20.3 Logistic Regression-Based Estimation for SGP

- We will estimate the mean (probability) for SGP:
 - ▷ For Belgium and the regions
 - ▷ Correcting for weighting
 - ▷ Using:
 - * **PROC SURVEYLOGISTIC** for survey-design-based regression.
 - * **PROC GENMOD** for GEE.
 - * **PROC GLIMMIX** for GLMM.
 - * **PROC NLMIXED** for GLMM.

- With straightforward syntax, a PROC SURVEYLOGISTIC program for the weighted mean in Belgium is:

```
proc surveylogistic data=m.bmi_voeg;  
title '17. Mean. Surveylogistic, weighted, for Belgium';  
weight wfin;  
model sgp = ;  
run;
```

- The relevant portion of the output for Belgium:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.6290	0.0557	2225.8554	<.0001

- This too, coincides with the SURVEYMEANS result.

- Switching to GEE with PROC GENMOD, for the weighted means in Belgium:

```
proc genmod data=m.bmi_voeg;
title '27. Mean. GEE logistic regression, for Belgium';
title2 'weighted';
class hh;
weight wfin;
model sgp = / dist=b;
repeated subject = hh / type=ind corrw model;
run;
```

- The use of the **REPEATED** statement is surprising at first sight, since no clustering is taken into account.

Let us study the output to see the reason for this.

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi- Square	Pr > ChiSq
Intercept	1	-2.6290	0.0015	-2.6319	-2.6260	3008181	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

		Standard	95% Confidence			
Parameter	Estimate	Error	Limits		Z	Pr > Z
Intercept	-2.6290	0.0642	-2.7548	-2.5031	-40.95	<.0001

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

		Standard	95% Confidence			
Parameter	Estimate	Error	Limits		Z	Pr > Z
Intercept	-2.6290	0.0015	-2.6319	-2.6260	-1734.4	<.0001

- ▷ The initial parameter, empirically corrected, and model-based estimates are identical.
- ▷ This is not surprising, since the working correlation structure is **independence**: we are assuming no clustering at all.
- ▷ Nevertheless, there is a **huge impact on the standard error**.
- ▷ The initial and model-based standard errors assume the weights are replications!

- ▷ The empirically corrected standard errors adjust the weights (standardizes them) so that they correspond to the proper amount of information available.
 - ▷ In the latter case, we arrive close to the SURVEYLOGISTIC result.
- A similar intervention is needed in the PROC GLIMMIX code:

```
proc glimmix data=m.bmi_voeg empirical;  
title '39a. GLMM, for Belgium';  
title2 'with proc glimmix';  
title3 'weighted - empirical';  
nloptions maxiter=50;  
weight wfin;  
model sgp = / solution dist=b;  
run;
```

- ▷ The '**empirical**' option ensures the empirically corrected standard errors are produced.

- ▷ Output **without** the 'empirical' option:

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	2.6290	0.001516	8531	1734.41	<.0001

- ▷ Output **with** the 'empirical' option:

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	2.6290	0.05572	8531	47.18	<.0001

- ▷ Also here, we see the dramatic impact of neglecting standardization of the weights.
- ▷ The procedure NLMIXED cannot easily take weights into account.
- ▷ We can further expand the summary table for SGP with our new analyses:

Stable General Practitioner (0/1) — Marginal and Random-effects Models						
Analysis	Procedure	Par.	Belgium	Brussels	Flanders	Wallonia
SRS	SURVEYMEANS	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	SURVEYLOGISTIC	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	SURVEYLOGISTIC	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GENMOD	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GENMOD	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GLIMMIX	β	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GLIMMIX	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	NLMIXED	β	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	NLMIXED	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
Strat.	SURVEYMEANS	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Strat.	SURVEYLOGISTIC	$-\beta$	2.3272(0.0358)	1.4219(0.0050)	2.9936(0.0859)	2.7278(0.0758)
Strat.	SURVEYLOGISTIC	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Clust.	SURVEYMEANS	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	SURVEYLOGISTIC	$-\beta$	2.2372(0.0455)	1.4219(0.0624)	2.9936(0.1037)	2.7278(0.0918)
Clust.	SURVEYLOGISTIC	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	GENMOD	$-\beta$	2.1504(0.0435)	1.3784(0.0591)	2.9188(0.1019)	2.6470(0.0890)
Clust.	GENMOD	π	0.8957(0.0040)	0.7987(0.0095)	0.9488(0.0050)	0.9338(0.0055)
Clust.	GLIMMIX	β	2.3723(0.0441)	1.5213(0.0628)	3.1433(0.0988)	—
Clust.	GLIMMIX	π	0.9147(0.0034)	0.8207(0.0092)	0.9586(0.0039)	—
Clust.	NLMIXED	β	4.3770(0.1647)	3.4880(0.3134)	8.4384(1.5434)	6.9047(0.8097)
Clust.	NLMIXED	π	0.9876(0.0020)	0.9703(0.0090)	0.9998(0.0003)	0.9990(0.0008)

Stable General Practitioner (0/1) — Marginal and Random-effects Models						
Analysis	Procedure	Par.	Belgium	Brussels	Flanders	Wallonia
Wgt.	SURVEYMEANS	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	SURVEYLOGISTIC	$-\beta$	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	SURVEYLOGISTIC	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	GENMOD	$-\beta$	2.6290(0.0642)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
Wgt.	GENMOD	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Wgt.	GLIMMIX	β	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	GLIMMIX	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
All	SURVEYMEANS	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
All	SURVEYLOGISTIC	$-\beta$	2.6290(0.0636)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
All	SURVEYLOGISTIC	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Cl.+Wgt.	GENMOD	$-\beta$	2.5233(0.0659)	1.2014(0.0839)	2.9693(0.1284)	2.7251(0.1186)
Cl.+Wgt.	GENMOD	π	0.9258(0.0045)	0.7688(0.0149)	0.9512(0.0060)	0.9385(0.0068)
Cl.+Wgt.	GLIMMIX	β	7.8531(0.1105)	5.1737(0.1906)	9.8501(0.1962)	8.7535(0.1850)
Cl.+Wgt.	GLIMMIX	π	0.9996(0.0000)	0.9944(0.0011)	0.9999(0.0000)	0.9998(0.0000)

- All weighted analyses, properly conducted, produce very similar results.
- The issue of the difference between marginal and random-effects modeling, prominently present in the clustering case, is totally absent here.
- The reason is that now no random effects are included, so **all** analyses are marginal.

Part IX

Integrated Analysis of Belgian Health Interview Survey

Chapter 21

Key Perspective Elements

- ▷ Analysis of continuous data
- ▷ Analysis of binary data
- ▷ Taxonomy

21.1 General Considerations

- Recall that software can be divided into tools for
 - ▷ Design (SAS PROC SURVEYSELECT)
 - ▷ Analysis (various procedures)
 - * Simple estimators *versus* models
 - * Cross-sectional data *versus* complex data
 - * Accounting for survey nature *versus* not accounting for survey nature

21.2 Analysis With SAS for a Continuous Outcome

Model	Data structure	Survey design	Method	SAS procedure
no	simple	no	mean	MEANS
yes	simple	no	linear regression ANOVA	REG ANOVA GLM
no	simple	yes	mean	SURVEYMEANS
yes	simple	yes	linear regression ANOVA	SURVEYREG
yes	complex	no	multivariate regression MANOVA	GLM
yes	complex	somehow	linear mixed model \equiv multi-level model	MIXED

21.3 Analysis With SAS for a Binary Outcome

Model	Data structure	Survey design	Method	SAS procedure
no	simple	no	proportion frequency	FREQ
yes	simple	no	logistic regression probit regression	LOGISTIC GENMOD
no	simple	yes	proportion frequency	SURVEYFREQ
yes	simple	yes	logistic regression probit regression	SURVEYLOGISTIC
yes	complex	no	generalized estimating equations	GENMOD
yes	complex	somehow	gen. lin. mixed model non-linear mixed model	GLIMMIX NLMIXED

- Several of these analysis will be conducted now:
 - ▷ Mean estimation
 - ▷ Frequency tables
 - ▷ Linear regression
 - ▷ Logistic regression

Chapter 22

Means, Proportions, and Frequencies

- ▷ Means using all design aspects
- ▷ Design effects
- ▷ Frequency tables

22.1 Means

22.1.1 Procedures for Means

- The means were calculated for

- ▷ LNBMI
- ▷ LNVOEG
- ▷ GHQ12
- ▷ SGP

assuming

- ▷ SRS: in Part III
- ▷ Stratified sampling: in Part VI
- ▷ Multi-stage sampling (two-stage sampling; clustering): in Part VII

▷ Unequal weights: in Part VIII

- In Parts VII and VIII also modeling procedures were used, each time focusing on one design aspect.
- It is perfectly possible to combine all of these design aspects.
- Using the SURVEYMEANS procedure, the following code can be used:

```
proc surveymeans data=m.bmi_voeg mean stderr;  
title 'weighted/stratified/clustered means';  
title2 'infinite population for Belgium and regions';  
where (regionch^='');  
domain regionch;  
weight wfin;  
strata province;  
cluster hh;  
var lnbmi lnvoeg ghq12 sgp;  
run;
```


- The program merely combines the three design statements: **WEIGHT**, **STRATA**, and **CLUSTER**.
- While it would be possible to include a finite sample correction, as we have seen, the impact is so negligible that it has been omitted.
- The output takes the usual form, with now all design aspects listed in the book keeping part:

The SURVEYMEANS Procedure

Data Summary

Number of Strata	12
Number of Clusters	4663
Number of Observations	8560
Sum of Weights	6954962.18

- The means for Belgium and the regions are:

Statistics		
Variable	Mean	Std Error of Mean
-----	-----	-----
LNBMI	3.185356	0.002867
LNVOEG	1.634690	0.014855
GHQ12	1.626781	0.048875
SGP	0.932702	0.003994
-----	-----	-----

Domain Analysis: REGIONCH			
REGIONCH	Variable	Mean	Std Error of Mean
-----	-----	-----	-----
Brussels	LNBMI	3.171174	0.004844
	LNVOEG	1.802773	0.023135
	GHQ12	1.928896	0.080508
	SGP	0.782448	0.013836
Flanders	LNBMI	3.180865	0.004250
	LNVOEG	1.511927	0.021409
	GHQ12	1.446286	0.068931
	SGP	0.954757	0.005379
Walloonia	LNBMI	3.198131	0.004403
	LNVOEG	1.803178	0.023214
	GHQ12	1.858503	0.084047
	SGP	0.943191	0.006159
-----	-----	-----	-----

- A summary of all analyses is as follows:

Logarithm of Body Mass Index				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	3.187218(0.001845)	3.175877(0.003372)	3.182477(0.002993)	3.201530(0.003216)
Stratification	3.187218(0.001840)	3.175877(0.003373)	3.182477(0.002989)	3.201530(0.003217)
Clustering	3.187218(0.001999)	3.175877(0.003630)	3.182477(0.003309)	3.201530(0.003429)
Weighting	3.185356(0.002651)	3.171174(0.004578)	3.180865(0.003870)	3.198131(0.004238)
All combined	3.185356(0.003994)	3.171174(0.004844)	3.180865(0.004250)	3.198131(0.004403)

Logarithm of VOG Score				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	1.702951(0.008954)	1.809748(0.016203)	1.516352(0.015201)	1.801107(0.014550)
Stratification	1.702951(0.008801)	1.809748(0.016206)	1.516352(0.015207)	1.801107(0.014427)
Clustering	1.702951(0.010355)	1.809748(0.018073)	1.516352(0.017246)	1.801107(0.016963)
Weighting	1.634690(0.013233)	1.802773(0.021831)	1.511927(0.019155)	1.803178(0.020426)
All combined	1.634690(0.014855)	1.802773(0.023135)	1.511927(0.021409)	1.803178(0.023214)

General Health Questionnaire – 12				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	1.661349(0.029584)	1.862745(0.056894)	1.385381(0.046246)	1.772148(0.051023)
Stratification	1.661956(0.029452)	1.864301(0.056939)	1.385857(0.046211)	1.772148(0.050823)
Clustering	1.661349(0.032824)	1.862745(0.062739)	1.385381(0.052202)	1.772148(0.055780)
Weighting	1.626201(0.044556)	1.924647(0.076313)	1.445957(0.061910)	1.858503(0.078566)
All combined	1.626781(0.048875)	1.924647(0.080508)	1.446286(0.068931)	1.858503(0.084047)

Stable General Practitioner (0/1)				
Analysis	Belgium	Brussels	Flanders	Wallonia
SRS	0.903540(0.003196)	0.805632(0.007826)	0.952285(0.003908)	0.938646(0.004382)
Stratification	0.903540(0.003116)	0.805632(0.007827)	0.952285(0.003902)	0.938646(0.004366)
Clustering	0.903540(0.003963)	0.805632(0.009766)	0.952285(0.004709)	0.938646(0.005284)
Weighting	0.932702(0.003498)	0.782448(0.011563)	0.954757(0.004722)	0.943191(0.005417)
All combined	0.932702(0.003994)	0.782448(0.013836)	0.954757(0.005379)	0.943191(0.006159)

- Weighting and clustering each increase the standard error, the combined analysis does more so.
- The point estimate is identical to the weighted one.

22.1.2 Linear Regression Procedures

- Like in Part VII, we can employ the SURVEYREG procedure:

```
proc surveyreg data=m.bmi_voeg;  
title '21. Mean. Surveyreg, all combined, for Belgium';  
strata province;  
cluster hh;  
weight wfin;  
model lnbmi = ;  
run;
```

- A maximal number of design aspects is now taken into account.

- Likewise, it is possible to correct for weighting and clustering simultaneously using the MIXED procedure:

```
proc mixed data=m.bmi_voeg method=reml;  
title '30. Survey mean with PROC MIXED, for Belgium';  
title2 'Weighted + Two-stage (clustered)';  
where (regionch^='');  
weight wfin;  
model lnbmi = / solution;  
random intercept / subject=hh;  
run;
```

- Here and in subsequent procedures, when the regions are of interest, include the statement:

```
by regionch;
```

- A summary of the various methods for mean estimation on LNBMI then becomes:

Logarithm of Body Mass Index					
Analysis	Procedure	Belgium	Brussels	Flanders	Wallonia
SRS	SURVEYMEANS	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
SRS	MIXED	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
Stratification	SURVEYMEANS	3.1872(0.0018)	3.1759(0.0034)	3.1825(0.0030)	3.2015(0.0032)
Clustering	SURVEYMEANS	3.1872(0.0020)	3.1759(0.0036)	3.1825(0.0033)	3.2015(0.0034)
Clustering	MIXED	3.1880(0.0020)	3.1761(0.0036)	3.1840(0.0033)	3.2022(0.0034)
Weighting	SURVEYMEANS	3.1853(0.0027)	3.1712(0.0046)	3.1809(0.0039)	3.1981(0.0042)
Weighting	MIXED	3.1854(0.0018)	3.1712(0.0034)	3.1809(0.0030)	3.1981(0.0032)
All combined	SURVEYMEANS	3.1853(0.0040)	3.1712(0.0048)	3.1809(0.0043)	3.1981(0.0044)
Clust+Wgt	MIXED	3.1865(0.0023)	3.1706(0.0039)	3.1817(0.0036)	3.1994(0.0038)

- Recall that here the results for SURVEYMEANS and SURVEYREG are the same.

22.1.3 Logistic Regression Procedures

- For the binary outcome SGP, we have considered several logistic regression-based procedures.
- A **SURVEYLOGISTIC** call, combining all design aspects:

```
proc surveylogistic data=m.bmi_voeg;  
title '23. Mean. Surveylogistic, weighted,';  
title2 'stratified, two-stage (clustered), for Belgium';  
weight wfin;  
strata province;  
cluster hh;  
model sgp = ;  
run;
```


- By means of GEE, within the **GENMOD** procedure, weighting and clustering can be taken into account:

```
proc genmod data=m.bmi_voeg;  
title '31. Mean. GEE logistic regression, for Belgium';  
title2 'weighted + clustered';  
weight wfin;  
class hh;  
model sgp = / dist=b;  
repeated subject = hh / type=cs corrw modelse;  
run;
```

- The first of two GLMM procedures, the **GLIMMIX** procedure, allows for the inclusion of weighting and clustering:

```
proc glimmix data=m.bmi_voeg empirical;  
title '43a. Mean. GLMM, for Belgium';  
title2 'with proc glimmix maxiter=50';  
title3 'weighted + two-stage (cluster) - empirical';  
nloptions maxiter=50;  
weight wfin;  
model sgp = / solution dist=b;  
random intercept / subject = hh type=un;  
run;
```

- It is important, here and in general, that **empirically corrected standard errors** be used, whenever weights are included, to compensate for not properly calibrated weights in procedures that are not explicitly designed to handle surveys.

- The second procedure, **NLMIXED**, only allows for clustering to be taken into account:

```
proc nlmixed data=m.bmi_voeg;  
title '35. Mean. GLMM, for Belgium';  
title2 'Two-stage (clustered)';  
theta = beta0 + b;  
exptheta = exp(theta);  
p = exptheta/(1+exptheta);  
model sgp ~ binary(p);  
random b ~ normal(0,tau2) subject=hh;  
estimate 'mean' exp(beta0)/(1+exp(beta0));  
run;
```

- Recall that the GLMM based procedures produce a fixed-effects intercept that is **not** the population average, but rather the probability corresponding to someone with random intercept value equal to zero.

Stable General Practitioner (0/1) — Marginal and Random-effects Models						
Analysis	Procedure	Par.	Belgium	Brussels	Flanders	Wallonia
SRS	SURVEYMEANS	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	SURVEYLOGISTIC	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	SURVEYLOGISTIC	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GENMOD	$-\beta$	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GENMOD	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	GLIMMIX	β	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	GLIMMIX	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
SRS	NLMIXED	β	2.2372(0.0367)	1.4219(0.0050)	2.9936(0.0860)	2.7278(0.0761)
SRS	NLMIXED	π	0.9035(0.0032)	0.8056(0.0078)	0.9523(0.0039)	0.9386(0.0044)
Strat.	SURVEYMEANS	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Strat.	SURVEYLOGISTIC	$-\beta$	2.3272(0.0358)	1.4219(0.0050)	2.9936(0.0859)	2.7278(0.0758)
Strat.	SURVEYLOGISTIC	π	0.9035(0.0031)	0.8056(0.0078)	0.9522(0.0039)	0.9386(0.0044)
Clust.	SURVEYMEANS	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	SURVEYLOGISTIC	$-\beta$	2.2372(0.0455)	1.4219(0.0624)	2.9936(0.1037)	2.7278(0.0918)
Clust.	SURVEYLOGISTIC	π	0.9035(0.0040)	0.8056(0.0098)	0.9523(0.0047)	0.9386(0.0053)
Clust.	GENMOD	$-\beta$	2.1504(0.0435)	1.3784(0.0591)	2.9188(0.1019)	2.6470(0.0890)
Clust.	GENMOD	π	0.8957(0.0040)	0.7987(0.0095)	0.9488(0.0050)	0.9338(0.0055)
Clust.	GLIMMIX	β	2.3723(0.0441)	1.5213(0.0628)	3.1433(0.0988)	—
Clust.	GLIMMIX	π	0.9147(0.0034)	0.8207(0.0092)	0.9586(0.0039)	—
Clust.	NLMIXED	β	4.3770(0.1647)	3.4880(0.3134)	8.4384(1.5434)	6.9047(0.8097)
Clust.	NLMIXED	π	0.9876(0.0020)	0.9703(0.0090)	0.9998(0.0003)	0.9990(0.0008)

Stable General Practitioner (0/1) — Marginal and Random-effects Models						
Analysis	Procedure	Par.	Belgium	Brussels	Flanders	Wallonia
Wgt.	SURVEYMEANS	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	SURVEYLOGISTIC	$-\beta$	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	SURVEYLOGISTIC	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
Wgt.	GENMOD	$-\beta$	2.6290(0.0642)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
Wgt.	GENMOD	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Wgt.	GLIMMIX	β	2.6290(0.0557)	1.2800(0.0679)	3.0494(0.1093)	2.8096(0.1011)
Wgt.	GLIMMIX	π	0.9327(0.0035)	0.7824(0.0116)	0.9548(0.0047)	0.9432(0.0054)
All	SURVEYMEANS	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
All	SURVEYLOGISTIC	$-\beta$	2.6290(0.0636)	1.2800(0.0813)	3.0494(0.1245)	2.8096(0.1150)
All	SURVEYLOGISTIC	π	0.9327(0.0040)	0.7824(0.0138)	0.9548(0.0054)	0.9432(0.0062)
Cl.+Wgt.	GENMOD	$-\beta$	2.5233(0.0659)	1.2014(0.0839)	2.9693(0.1284)	2.7251(0.1186)
Cl.+Wgt.	GENMOD	π	0.9258(0.0045)	0.7688(0.0149)	0.9512(0.0060)	0.9385(0.0068)
Cl.+Wgt.	GLIMMIX	β	7.8531(0.1105)	5.1737(0.1906)	9.8501(0.1962)	8.7535(0.1850)
Cl.+Wgt.	GLIMMIX	π	0.9996(0.0000)	0.9944(0.0011)	0.9999(0.0000)	0.9998(0.0000)

22.2 Design Effects

- Most authors define the **design effect** as the ratio of two variances:
 - ▷ the variance of an estimator taking design aspects into account
 - ▷ the variance of the SRS estimator
- Historically, it was used for correction:
 - ▷ compute the SRS estimator and its precision
 - ▷ modify the standard error using the design effect

- This is **not** a good approach:
 - ▷ As we have seen, we have proper design-based and complex model-based estimation methods.
 - ▷ The design effect is **not an invariant** for a method.
- Consider the design effect for clustering.
 - ▷ For example, for LNBMI and Belgium, we find:

$$D_{\text{eff}} = \frac{0.001999}{0.001845} = 1.2$$

- An overview table for clustering and weighting:

Outcome	Belgium	Brussels	Flanders	Wallonia
Design Effects for Clustering				
LNBMI	1.2	1.2	2.1	1.1
LNVOEG	1.3	1.2	1.3	1.4
GHQ-12	2.3	1.8	1.8	2.4
SGP	1.5	1.6	1.5	1.5
Design Effects for Weighting				
LNBMI	2.1	1.8	2.8	1.7
LNVOEG	2.2	1.8	1.6	2.0
GHQ-12	2.3	1.8	1.8	2.4
SGP	1.2	2.2	1.5	1.5

- For clustering, the design effects varies between 1.1 and 2.4.
- For weighting, the design effect varies between 1.2 and 2.8.
- Even within a region and/or within an outcome, there is a lot of variability.
- The differences are a function, not only of the variances, but also the changing point estimates, for example in going from an unweighted to a weighted analysis.
- In conclusion, the design effect gives a numerical summary of the impact of one or several design elements in a particular situation, but should not itself be used as a basis for precision estimation.

22.3 Frequency Tables

- We have calculated means, for all four variables, including SGP, even though it is a binary variable.
- The mean for a binary variable is sensible: it is the proportion to observe a “success” .
- The situation is different for categorical variables with more than 2 categories, in which case frequencies are more advisable.
- In any categorical situation it is sensible to:
 - ▷ calculate **frequencies** for a single variable
 - ▷ construct **contingency** tables for 2 variables or more → 2-way, 3-way, or higher-way contingency tables

- The typical SAS tool is **PROC FREQ**:

```
proc freq data=m.bmi_voeg compress;  
where (regionch^='');  
title '1. proc freq - srs proportions, Belgium';  
table sgp;  
run;
```

```
proc freq data=m.bmi_voeg compress;  
where (regionch^='');  
title '2. proc freq - srs proportions, regions';  
table regionch*sgp;  
run;
```

- ▷ The first program is for the frequencies of having versus not having a stable GP.
- ▷ The second program constructs a 2-way table for region with SGP.
- ▷ The **TABLE** statement is the crucial one, specifying the variable or variables of interest.

- ▷ PROC FREQ produces a large amount of output by default; the 'compress' option reduces this.

- The following output is obtained for the first program:

```
1. proc freq - srs proportions, Belgium
The FREQ Procedure
```

SGP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	823	9.65	823	9.65
1	7709	90.35	8532	100.00

```
Frequency Missing = 32
```

- For the second program, we obtain:

2. proc freq - srs proportions, regions
The FREQ Procedure
Table of REGIONCH by SGP

REGIONCH	SGP		
Frequency			
Percent			
Row Pct			
Col Pct	0	1	Total
----- ----- -----			
Brussels	497	2060	2557
	5.83	24.14	29.97
	19.44	80.56	
	60.39	26.72	
----- ----- -----			
Flanders	142	2834	2976
	1.66	33.22	34.88
	4.77	95.23	
	17.25	36.76	
----- ----- -----			
Walloonnia	184	2815	2999
	2.16	32.99	35.15
	6.14	93.86	
	22.36	36.52	
----- ----- -----			
Total	823	7709	8532
	9.65	90.35	100.00
Frequency Missing = 32			

- Of course, these tables start from the assumption that the sample be representative, as it is, for the population.

In particular, it appears the regional percentages are, roughly 30%, 35%, and 35%.

- We need to take the design into account to rectify this.
- The above programs can be adapted to incorporate weighting, by including:

```
weight wfin;
```

- The output changes to:

```
3. proc freq - weighted proportions, Belgium
The FREQ Procedure
```

SGP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	466652.6	6.73	466652.6	6.73
1	6467487	93.27	6934140	100.00

```
Frequency Missing = 23457.39966
```

4. proc freq - weighted proportions, regions

The FREQ Procedure

Table of REGIONCH by SGP

REGIONCH	SGP		
Frequency			
Percent			
Row Pct			
Col Pct	0	1	Total
----- ----- -----			
Brussels	160892	578665	739558
	2.32	8.35	10.67
	21.76	78.24	
	34.48	8.95	
----- ----- -----			
Flanders	180516	3809385	3989901
	2.60	54.94	57.54
	4.52	95.48	
	38.68	58.90	
----- ----- -----			
Walloonnia	125245	2079437	2204682
	1.81	29.99	31.79
	5.68	94.32	
	26.84	32.15	
----- ----- -----			
Total	466653	6467487	6934140
	6.73	93.27	100.00
Frequency Missing = 23457.39966			

- Note that the region-specific proportions are more in line with reality.
- The frequencies reflect the sum of the weights: PROC FREQ treats them merely as repeat counts, and not the inverse of selection probabilities.
- The procedure **PROC SURVEYFREQ** can be used to properly take the survey design into account:

```
proc surveyfreq data=m.bmi_voeg;  
title '5. proc surveyfreq - srs, infinite proportions, Belgium';  
table sgp;  
run;
```

```
proc surveyfreq data=m.bmi_voeg;  
title '6. proc surveyfreq - srs, infinite proportions, regions';  
table regionch*sgp;  
run;
```


- The procedure is syntactically entirely similar to PROC FREQ, especially when applied to SRS for an infinite population.
- The output is similar to what was obtained for SRS with PROC FREQ:

```
5. proc surveyfreq - srs, infinite proportions, Belgium
The SURVEYFREQ Procedure
```

Data Summary			
Number of Observations		8564	
Table of SGP			
SGP	Frequency	Percent	Std Err of Percent
0	823	9.6460	0.3196
1	7709	90.3540	0.3196
Total	8532	100.000	
Frequency Missing = 32			

and

```
proc surveyfreq - srs, infinite proportions, regions
```

```
The SURVEYFREQ Procedure
```

```
Data Summary
```

```
Number of Observations      8564
```

Table of REGIONCH by SGP

REGIONCH	SGP	Frequency	Percent	Std Err of Percent
Brussels	0	497	5.8251	0.2536
	1	2060	24.1444	0.4633
	Total	2557	29.9695	0.4960
Flanders	0	142	1.6643	0.1385
	1	2834	33.2161	0.5099
	Total	2976	34.8805	0.5160
Walloonia	0	184	2.1566	0.1573
	1	2815	32.9934	0.5091
	Total	2999	35.1500	0.5169
Total	0	823	9.6460	0.3196
	1	7709	90.3540	0.3196
	Total	8532	100.000	

Frequency Missing = 32

- ▷ While displayed a little differently, the numbers coincide with what we obtained from PROC FREQ.
 - ▷ Note that one obtains precision estimates, making the procedure useful *even in a non-survey context*.
 - ▷ The output for the SGP frequencies is exactly a sub-part of the output for the cross-tabulation of region by SGP
⇒ in what follows it will be dropped.
- We can now also correct for finite sampling, changing the PROC SURVEYFREQ statement to:

```
proc surveyfreq data=m.bmi_voeg total=10000000;
```

- The output changes only slightly:

8. proc surveyfreq - srs, finite proportions, regions
The SURVEYFREQ Procedure

Number of Observations 8564

Table of REGIONCH by SGP

REGIONCH	SGP	Frequency	Percent	Std Err of Percent
Brussels	0	497	5.8251	0.2535
	1	2060	24.1444	0.4631
	Total	2557	29.9695	0.4958
Flanders	0	142	1.6643	0.1384
	1	2834	33.2161	0.5097
	Total	2976	34.8805	0.5158
Walloonnia	0	184	2.1566	0.1572
	1	2815	32.9934	0.5088
	Total	2999	35.1500	0.5167
Total	0	823	9.6460	0.3195
	1	7709	90.3540	0.3195
	Total	8532	100.000	

Frequency Missing = 32

- We observe no impact on frequencies and percentages, and a small impact on the standard errors.
- This is in line with observations in the case of mean estimation.
- Setting the **TOTAL** $N = 8564$, the predictable effect is:

10. proc surveyfreq - srs, census-finite proportions, regions
The SURVEYFREQ Procedure

Number of Observations 8564

Table of REGIONCH by SGP

REGIONCH	SGP	Frequency	Percent	Std Err of Percent

Brussels	0	497	5.8251	0.0000
	1	2060	24.1444	0.0000
	Total	2557	29.9695	0.0000

Flanders	0	142	1.6643	0.0000
	1	2834	33.2161	0.0000
	Total	2976	34.8805	0.0000

Walloonia	0	184	2.1566	0.0000
	1	2815	32.9934	0.0000
	Total	2999	35.1500	0.0000

Total	0	823	9.6460	0.0000
	1	7709	90.3540	0.0000
	Total	8532	100.000	

Frequency Missing = 32

- Three further design aspects can be included:

- ▷ **Stratification** by the statement:

```
strata province;
```

- ▷ **Weighting** by the statement:

```
weight wfin;
```

- ▷ **Clustering** by the statement:

```
cluster hh;
```

- The output in the stratified case:

12. proc surveyfreq - stratified proportions, regions

The SURVEYFREQ Procedure

Number of Strata 12

Number of Observations 8560

Table of REGIONCH by SGP

REGIONCH	SGP	Frequency	Percent	Std Err of Percent
Brussels	0	497	5.8251	0.2346
	1	2060	24.1444	0.2346
	Total	2557	29.9695	0.0000
Flanders	0	142	1.6643	0.1361
	1	2834	33.2161	0.1361
	Total	2976	34.8805	0.0000
Walloonia	0	184	2.1566	0.1535
	1	2815	32.9934	0.1535
	Total	2999	35.1500	0.0000
Total	0	823	9.6460	0.3116
	1	7709	90.3540	0.3116
	Total	8532	100.000	

Frequency Missing = 28

- The proportion has not changed, but there is a small impact on the standard error.
- The data summary also included the number of strata.
- The number of available observations has slightly decreased, due to a small number of individuals for which the province has not been recorded in the database.
- The output for weighting:

14. proc surveyfreq - weighted proportions, regions

Data Summary

Number of Observations 8564
Sum of Weights 6957597.07

Table of REGIONCH by SGP

REGIONCH	SGP	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
<hr/>						
Brussels	0	497	160892	9665	2.3203	0.1430
	1	2060	578665	16075	8.3452	0.2624
	Total	2557	739558	18166	10.6655	0.3044
<hr/>						
Flanders	0	142	180516	19170	2.6033	0.2736
	1	2834	3809385	79623	54.9367	0.7635
	Total	2976	3989901	80908	57.5400	0.7462
<hr/>						
Walloonia	0	184	125245	12156	1.8062	0.1755
	1	2815	2079437	49543	29.9884	0.6851
	Total	2999	2204682	50410	31.7946	0.6972
<hr/>						
Total	0	823	466653	24327	6.7298	0.3498
	1	7709	6467487	79980	93.2702	0.3498
	Total	8532	6934140	79253	100.000	
<hr/>						

Frequency Missing = 32

- ▷ The information provided is more extensive, since both frequencies as well as weighted frequencies are given.
 - ▷ The overall percentage of not having a stable GP is smaller, in line with:
 - * the proper **up-weighting of Flanders**, where virtually everyone has a stable GP
 - * the proper **down-weighting of Brussels**, where a large fraction does not have a stable GP
 - ▷ The analysis agrees closely with the weighted analysis within PROC FREQ, but is more informative.
-
- The output for the clustered analysis:

16. proc surveyfreq - two-stage (clustered) proportions, regions

Number of Clusters 4663
 Number of Observations 8564

Table of REGIONCH by SGP

REGIONCH	SGP	Frequency	Percent	Std Err of Percent

Brussels	0	497	5.8251	0.3214
	1	2060	24.1444	0.6543
	Total	2557	29.9695	0.7180

Flanders	0	142	1.6643	0.1673
	1	2834	33.2161	0.7619
	Total	2976	34.8805	0.7760

Walloon	0	184	2.1566	0.1902
	1	2815	32.9934	0.7500
	Total	2999	35.1500	0.7680

Total	0	823	9.6460	0.3963
	1	7709	90.3540	0.3963
	Total	8532	100.000	

Frequency Missing = 32

- ▷ The number of clusters is displayed.
- ▷ There is impact on the standard error.
- A program for all design aspects combined:

```
proc surveyfreq data=m.bmi_voeg;  
title '17. proc surveyfreq - all aspects, proportions, Belgium';  
strata province;  
weight wfin;  
cluster hh;  
table regionch*sgp;  
run;
```

- The output:

```
18. proc surveyfreq - all aspects, proportions, regions
```

```
The SURVEYFREQ Procedure
```

```
      Data Summary
```

Number of Strata	12
Number of Clusters	4663
Number of Observations	8560
Sum of Weights	6954962.18

and

Table of REGIONCH by SGP

REGIONCH	SGP	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Brussels	0	497	160892	10724	2.3203	0.1568
	1	2060	578665	19832	8.3452	0.2952
	Total	2557	739558	20399	10.6655	0.3106
Flanders	0	142	180516	21464	2.6033	0.3090
	1	2834	3809385	91357	54.9367	0.8024
	Total	2976	3989901	90895	57.5400	0.7529
Walloon	0	184	125245	13587	1.8062	0.1962
	1	2815	2079437	57538	29.9884	0.7148
	Total	2999	2204682	57600	31.7946	0.7095
Total	0	823	466653	27574	6.7298	0.3994
	1	7709	6467487	109773	93.2702	0.3994
	Total	8532	6934140	109525	100.000	

Frequency Missing = 28

- Note that the estimated percentages, obtained for Belgium, coincide with the estimated means on pages 630 and 631.
- For the regions, PROC SURVEYFREQ does not provide the marginal percentages, but rather the percentage to belong to a given cell.
- In case the marginal probabilities are required, it is better to change the code to:

```
proc surveyfreq data=m.bmi_voeg;  
title '19. proc surveyfreq - all aspects, proportions, BY regions';  
by regionch;  
strata province;  
weight wfin;  
cluster hh;  
table sgp;  
run;
```


- This produces the following output:

```
19. proc surveyfreq - all aspects, proportions, BY region
The SURVEYFREQ Procedure
```

```
REGIONCH=Brussels
```

Data Summary

```
Number of Strata          1
Number of Clusters       1544
Number of Observations   2568
Sum of Weights           742678.193
```

Table of SGP

SGP	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
0	497	160892	10724	21.7552	1.3836
1	2060	578665	19832	78.2448	1.3836
Total	2557	739558	20399	100.000	

Frequency Missing = 11

and

REGIONCH=Flanders

Data Summary

Number of Strata	5
Number of Clusters	1508
Number of Observations	2986
Sum of Weights	4001968.5

Table of SGP

SGP	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
0	142	180516	21464	4.5243	0.5379
1	2834	3809385	91357	95.4757	0.5379
Total	2976	3989901	90895	100.000	

Frequency Missing = 10

and

REGIONCH=Walloonia

Data Summary

Number of Strata	6
Number of Clusters	1611
Number of Observations	3006
Sum of Weights	2210315.49

Table of SGP

SGP	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
0	184	125245	13587	5.6809	0.6159
1	2815	2079437	57538	94.3191	0.6159
Total	2999	2204682	57600	100.000	

Frequency Missing = 7

Chapter 23

Linear Regression

- ▷ Ordinary linear regression
- ▷ Linear regression for survey data
- ▷ Linear mixed model

23.1 Concept

- In our mean estimation endeavors, we employed procedures for linear regression.
- This implies we can conduct genuine linear regression, using:
 - ▷ **PROC REG, PROC GLM**: Conventional linear regression procedures
 - ▷ **PROC SURVEYREG**: Design-based regression procedure
 - ▷ **PROC MIXED**: Regression procedure for hierarchical data, based on the LMM
- Note that a variety of tools, designed for **generalized linear models** work for:
 - ▷ linear regression
 - ▷ logistic regression
 - ▷ probit regression

▷ Poisson regression

▷ ...

Such procedures can hence be used for linear regression as well.

- Example include PROC GENMOD, PROC GLIMMIX, PROC NLMIXED.
- However, the dedicated linear regression procedures, mentioned earlier, often have more features than the more general purpose tools.

23.2 Model

- Assume we are interested in the effect of **sex** and **age** on **BMI**.
- **Sex** is a binary variable, necessitating a single parameter.
- Define **age** as a 7-point ordinal variable **age7**.
- Construct dummy variables:

$$\text{age7} = \begin{cases} 1 & \iff 15 \leq \text{age} \leq 24 \\ 2 & \iff 25 \leq \text{age} \leq 34 \\ 3 & \iff 35 \leq \text{age} \leq 44 \\ 4 & \iff 45 \leq \text{age} \leq 54 \\ 5 & \iff 55 \leq \text{age} \leq 64 \\ 6 & \iff 65 \leq \text{age} \leq 75 \\ 7 & \iff 75 \leq \text{age} \end{cases}$$

$$A_\ell = \text{agegr}_\ell = \begin{cases} 1 & \iff \text{age7} = \ell \\ 0 & \iff \text{age7} \neq \ell \end{cases}$$

- We now consider the following basic regression model:

$$Y_i = \beta_0 + \beta_1 S_i + \beta_{21} A_{1i} + \beta_{22} A_{2i} + \beta_{23} A_{3i} + \beta_{24} A_{4i} + \beta_{25} A_{5i} + \beta_{26} A_{6i} + \varepsilon_i$$

where

- ▷ Y_i is LNBMI for respondent i
 - ▷ S_i is sex of respondent i (0 for males; 1 for females)
 - ▷ $A_{\ell i}$ is the value age-dummy ℓ takes for respondent i
 - ▷ ε_i is the error term
- In **conventional linear regression**, we assume $\varepsilon_i \sim N(0, \sigma^2)$.
 - In **design-based regression**, the variability will be calculated by properly taking the design-related formulas into account.

- In a **hierarchical model**, e.g., the LMM, our model will change to the two-stage setting:

$$Y_{ij} = \beta_0 + b_i + \beta_1 S_{ij} + \beta_{21} A_{1ij} + \beta_{22} A_{2ij} + \beta_{23} A_{3ij} + \beta_{24} A_{4ij} + \beta_{25} A_{5ij} + \beta_{26} A_{6ij} + \varepsilon_{ij}$$

where now

- ▷ Y_{ij} is LNBMI for individual j in household i
- ▷ S_{ij} is sex of individual j in household i
- ▷ $A_{\ell ij}$ is the value age-dummy ℓ takes for individual j in household i
- ▷ b_i is a **household-level effect on LNBMI**: $b_i \sim N(0, \tau^2)$
- ▷ ε_i is the deviation for individual j in household i : $\varepsilon_{ij} \sim N(0, \sigma^2)$

23.3 Programs

23.3.1 Programs for Ordinary Linear Regression

- Ordinary linear regression can be coded using the dedicated SAS procedures **PROC REG** and **PROC GLM**:

```
proc reg data=m.bmi_voeg;  
title '1. Ordinary linear regression, for Belgium';  
title2 'with PROC REG';  
model lnbmi = sex agegr1 agegr2 agegr3 agegr4 agegr5 agegr6;  
run;
```

```
proc glm data=m.bmi_voeg;  
title '2. Ordinary linear regression, for Belgium';  
title2 'with PROC GLM';  
class age7;  
model lnbmi = sex age7 / solution;  
run;
```

- ▷ PROC REG is more basic and does not allow for dummy variables \Rightarrow the user has to create them.
- ▷ PROC GLM allows for univariate and multivariate regression and contains the **CLASS** statement to automatically create dummies.
- ▷ When there are 7 dummies, PROC GLM removes the last one to ensure estimability, exactly like we have done ourselves with PROC REG.
- ▷ PROC GLM is an “ANOVA-based” procedure: there is more emphasis on ANOVA tables than on parameter estimates; this is why we include the ‘**solution**’ option into the **MODEL** statement.

Note that we have seen the ‘**solution**’ option repeatedly in earlier chapters.

- We can also use the LMM procedure **PROC MIXED**, without the hierarchical features, to fit an ordinary linear regression:

```
proc mixed data=m.bmi_voeg method=reml;  
title '3. Ordinary linear regression, for Belgium';  
title2 'with PROC MIXED - REML estimation';  
class age7;  
model lnbmi = sex age7 / solution;  
run;
```

```
proc mixed data=m.bmi_voeg method=ml;  
title '4. Ordinary linear regression, for Belgium';  
title2 'with PROC MIXED - ML estimation';  
class age7;  
model lnbmi = sex age7 / solution;  
run;
```

- ▷ We can opt for both REML and ML, i.e., restricted maximum likelihood and maximum likelihood.

Recall that the former is a small-sample correction towards ML: since our sample is very large, there will be little or no difference.

- ▷ The syntax of the procedure, used in this way, is very similar to the PROC GLM syntax.
- Finally, we can employ the design-based regression procedure **PROC SURVEYREG**, but confine it to SRS:

```
proc surveyreg data=m.bmi_voeg;  
title '5. Surveyreg, SRS, infinite population';  
class age7;  
model lnbmi = sex age7 / solution;  
run;
```

- ▷ Used in this fashion, the procedure is syntactically similar to PROC GLM and PROC MIXED.

23.3.2 Programs for Design-Based Linear Regression

- Starting from the **PROC SURVEYREG** program on page 671:

```
proc surveyreg data=m.bmi_voeg;  
title '5. Surveyreg, SRS, infinite population';  
class age7;  
model lnbmi = sex age7 / solution;  
run;
```

a number of design features can be built in:

- ▷ **Finite population:** the **PROC SURVEYREG** statement changes to:

```
proc surveyreg data=m.bmi_voeg total=10000000;
```

- ▷ **Census-finite population:** the **PROC SURVEYREG** statement changes to:

```
proc surveyreg data=m.bmi_voeg total=8384;
```

- ▷ **Stratification:** the following statement is added:

```
strata province;
```

- ▷ **Two-stage sampling (clustering):** the following statement is added:

```
cluster hh;
```

- ▷ **Weighting:** the following statement is added:

```
weight wfin;
```

- ▷ **Maximal accommodation for design:** the program becomes:

```
proc surveyreg data=m.bmi_voeg total=10000000;  
title '11. Surveyreg, weighted, stratified,';  
title2 'two-stage (clustered), finite population';  
class age7;  
weight wfin;  
strata province;  
cluster hh;  
model lnBMI = sex age7 / solution;  
run;
```

23.3.3 Programs for the Linear Mixed Model

- The design features that can be accommodated in PROC MIXED are weighting and clustering:

```
proc mixed data=m.bmi_voeg method=reml;  
title '12. Approximate survey regression, for Belgium';  
title2 'with PROC MIXED (weighted + clustered)';  
weight wfin;  
class age7;  
model lnbmi = sex age7 / solution;  
random intercept / subject=hh;  
run;
```

- When there are three or more levels, in a 3-stage or multi-stage design, PROC MIXED can accommodate this through multiple RANDOM statements.

- Example:

```
random intercept / subject=town;  
random intercept / subject=hh;
```


23.4 Parameter Estimates

23.4.1 Selected Output

- **PROC REG** for ordinary linear regression:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	3.25992	0.00822	396.64	<.0001
SEX	1	-0.04508	0.00342	-13.17	<.0001
AGEGR1	1	-0.12354	0.00772	-15.99	<.0001
AGEGR2	1	-0.04495	0.00729	-6.17	<.0001
AGEGR3	1	-0.00303	0.00731	-0.41	0.6784
AGEGR4	1	0.03796	0.00757	5.02	<.0001
AGEGR5	1	0.06126	0.00779	7.86	<.0001
AGEGR6	1	0.06156	0.00783	7.86	<.0001

- **PROC GLM** for ordinary linear regression:

Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		3.259921896 B	0.00821889	396.64	<.0001
SEX		-0.045076290	0.00342154	-13.17	<.0001
AGE7	1	-0.123542829 B	0.00772450	-15.99	<.0001
AGE7	2	-0.044953329 B	0.00728596	-6.17	<.0001
AGE7	3	-0.003032890 B	0.00731326	-0.41	0.6784
AGE7	4	0.037962219 B	0.00756814	5.02	<.0001
AGE7	5	0.061264578 B	0.00778986	7.86	<.0001
AGE7	6	0.061560303 B	0.00782798	7.86	<.0001
AGE7	7	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

- Note that there is a warning about non-uniqueness.

This is not an issue, and merely indicates one dummy has to be removed, as stated earlier.

A different choice will lead to differently coded **but equivalent** parameterizations.

- **PROC MIXED** with **REML** and **ML** for ordinary linear regression:

Effect	AGE7	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		3.2599	0.008219	8376	396.64	<.0001
SEX		-0.04508	0.003422	8376	-13.17	<.0001
AGE7	1	-0.1235	0.007725	8376	-15.99	<.0001
AGE7	2	-0.04495	0.007286	8376	-6.17	<.0001
AGE7	3	-0.00303	0.007313	8376	-0.41	0.6784
AGE7	4	0.03796	0.007568	8376	5.02	<.0001
AGE7	5	0.06126	0.007790	8376	7.86	<.0001
AGE7	6	0.06156	0.007828	8376	7.86	<.0001
AGE7	7	0

Effect	AGE7	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		3.2599	0.008215	8376	396.83	<.0001
SEX		-0.04508	0.003420	8376	-13.18	<.0001
AGE7	1	-0.1235	0.007721	8376	-16.00	<.0001
AGE7	2	-0.04495	0.007282	8376	-6.17	<.0001
AGE7	3	-0.00303	0.007310	8376	-0.41	0.6782
AGE7	4	0.03796	0.007565	8376	5.02	<.0001
AGE7	5	0.06126	0.007786	8376	7.87	<.0001
AGE7	6	0.06156	0.007824	8376	7.87	<.0001
AGE7	7	0

- Note that the impact of the ML versus REML choice is not noticeable up to the 4th decimal place of the standard errors.
- **PROC SURVEYREG** for ordinary linear regression:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.2599219	0.00819723	397.69	<.0001
SEX	-0.0450763	0.00340116	-13.25	<.0001
AGE7 1	-0.1235428	0.00774817	-15.94	<.0001
AGE7 2	-0.0449533	0.00760124	-5.91	<.0001
AGE7 3	-0.0030329	0.00769816	-0.39	0.6936
AGE7 4	0.0379622	0.00789804	4.81	<.0001
AGE7 5	0.0612646	0.00810799	7.56	<.0001
AGE7 6	0.0615603	0.00831177	7.41	<.0001
AGE7 7	0.0000000	0.00000000	.	.

- PROC SURVEYREG for a finite and a census-finite population:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.2599219	0.00819380	397.85	<.0001
SEX	-0.0450763	0.00339973	-13.26	<.0001
AGE7 1	-0.1235428	0.00774492	-15.95	<.0001
AGE7 2	-0.0449533	0.00759806	-5.92	<.0001
AGE7 3	-0.0030329	0.00769493	-0.39	0.6935
AGE7 4	0.0379622	0.00789472	4.81	<.0001
AGE7 5	0.0612646	0.00810459	7.56	<.0001
AGE7 6	0.0615603	0.00830828	7.41	<.0001
AGE7 7	0.0000000	0.00000000	.	.

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.2599219	0	Infty	<.0001
SEX	-0.0450763	0	-Infty	<.0001
AGE7 1	-0.1235428	0	-Infty	<.0001
AGE7 2	-0.0449533	0	-Infty	<.0001
AGE7 3	-0.0030329	0	-Infty	<.0001
AGE7 4	0.0379622	0	Infty	<.0001
AGE7 5	0.0612646	0	Infty	<.0001
AGE7 6	0.0615603	0	Infty	<.0001
AGE7 7	0.0000000	0	.	.

- PROC SURVEYREG for all design aspects combined:

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	3.2843384	0.01248738	263.01	<.0001
SEX	-0.0497695	0.00452629	-11.00	<.0001
AGE7 1	-0.1364198	0.01194590	-11.42	<.0001
AGE7 2	-0.0613612	0.01162671	-5.28	<.0001
AGE7 3	-0.0160375	0.01189654	-1.35	0.1777
AGE7 4	0.0231462	0.01215398	1.90	0.0569
AGE7 5	0.0570522	0.01362803	4.19	<.0001
AGE7 6	0.0355099	0.01514205	2.35	0.0191
AGE7 7	0.0000000	0.00000000	.	.

- **PROC MIXED** for weighting and clustering:

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	HH	0.007719
Residual		12.3942

▷ We can now also calculate the intra-class correlation.

▷ Recall the computations from page 474:

$$\widehat{\sigma}^2 = 0.0243$$

$$\widehat{\tau}^2 = 0.0043$$

$$\widehat{\rho}_{\text{LNBI}} = \frac{\tau^2}{\sigma^2 + \tau^2} = \frac{0.0043}{0.0243 + 0.0043} = 0.15$$

- ▷ This now changes to:

$$\widehat{\sigma^2} = 12.3992$$

$$\widehat{\tau^2} = 0.007719$$

$$\widehat{\rho}_{\text{LNBMI}|\text{sex,age}} = \frac{\tau^2}{\sigma^2 + \tau^2} = \frac{0.007719}{12.3992 + 0.007719} = 0.00062$$

- ▷ The total variability is much larger: impact of the weights, which sums to, roughly, the population total.
- ▷ This does not change the relative magnitudes of σ^2 and τ^2 .
- ▷ The resulting intra-cluster correlation, after correcting for sex and age, is much smaller.

Sex and age have the power to explain a large amount of within-household correlation.

- The fixed effects:

Effect	AGE7	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		3.2863	0.009221	4594	356.40	<.0001
SEX		-0.04882	0.002908	3782	-16.79	<.0001
AGE7	1	-0.1467	0.008992	3782	-16.32	<.0001
AGE7	2	-0.06982	0.008837	3782	-7.90	<.0001
AGE7	3	-0.01486	0.008841	3782	-1.68	0.0928
AGE7	4	0.01884	0.008915	3782	2.11	0.0346
AGE7	5	0.05052	0.009453	3782	5.34	<.0001
AGE7	6	0.02971	0.009462	3782	3.14	0.0017
AGE7	7	0

23.4.2 Overview Table

Logarithm of Body Mass Index (Belgium)									
Analysis	Procedure	Parameter estimates (s.e.) $\times 10^4$							
		$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\beta}_{23}$	$\hat{\beta}_{24}$	$\hat{\beta}_{25}$	$\hat{\beta}_{26}$
Ordinary linear regression									
1.–5. SRS	several*	32,599(82)	-451(34)	-1235(77)	-450(73)	-30(83)	380(76)	613(78)	616(78)
Design-based linear regression									
5. SRS, ∞	SURVEYREG	32,599(82)	-451(34)	-1235(77)	-450(73)	-30(73)	380(76)	613(78)	616(78)
6. SRS, 10^7	SURVEYREG	32,599(82)	-451(34)	-1235(77)	-450(73)	-30(73)	380(76)	613(78)	616(78)
7. SRS, 8384	SURVEYREG	32,599(0)	-451(0)	-1235(0)	-450(0)	-30(0)	380(0)	613(0)	616(0)
8. weighted	SURVEYREG	32,843(127)	-498(49)	-1364(118)	-614(115)	-160(120)	231(119)	571(134)	355(143)
9. stratified	SURVEYREG	32,600(82)	-451(34)	-1235(77)	-450(76)	-30(77)	380(79)	613(81)	616(83)
10. clustered	SURVEYREG	32,600(80)	-451(32)	-1235(79)	-450(77)	-30(78)	380(80)	613(82)	616(83)
11. all	SURVEYREG	32,843(125)	-498(45)	-1364(119)	-614(116)	-160(119)	231(122)	571(136)	355(151)
Hierarchical linear regression									
12. wt, clust	MIXED	32,863(92)	-488(29)	-1467(90)	-698(88)	-149(88)	188(89)	505(94)	297(95)

*: REG, GLM, MIXED (REML), MIXED (ML), SURVEYREG (SRS)

- As stated earlier, all **ordinary linear regression** implementations produce exactly the same results, as it should.
- Some analyses (**SRS with finite-population correction** and **stratified analyses**) are only slightly different.
- In this case, there is little clustering left (we derived a small inter-cluster correlation), hence the **clustered analysis** is similar, too.
- Not surprisingly, the largest impact is seen on the **weighted analysis**, with the direction in which the coefficients move hard to predict.
- Due to the different nature of the correction, the **linear mixed model analysis** is different, though not spectacular.

23.4.3 Hypothesis Testing

- Especially in a regression context, we might be interested in testing hypotheses, such as:

$H_{0,1}$: Sex has no effect on LNBMI.

$H_{0,2}$: Age has no effect on LNBMI.

- In formulas:

$$H_{0,1} : \beta_1 = 0$$

$$H_{0,2} : \beta_{21} = \beta_{22} = \beta_{23} = \beta_{24} = \beta_{25} = \beta_{26} = 0$$

- $H_{0,1}$ involves 1 parameter: $d_1 = 1$ (numerator) degrees of freedom (ndf).
- $H_{0,2}$ involves 6 parameters: $d_2 = 6$ (numerator) degrees of freedom.
- One typically, but not exclusively, uses the F_{d_1, d_2} test, where d_2 represents the denominator degrees of freedom (ddf).
- ddf refers to the amount of information available for the test.
- ddf is directly related to the sample size, but in complex designs and/or hierarchical models, calculation is more subtle.
- For the LMM, there are various methods, but the most recommended ones are Satterthwaite and Kenward-Roger.

- Using a high-quality ddf method is essential when the dataset is small (small number of first-level units).
- Since we have a large number of HH, there is little problem here, but when we would start from the town level, differences might become noticeable.

23.4.4 Selected Output

- The output takes various forms.
- **PROC REG** does not foresee such tests by default, even though they can be obtained.
- **PROC GLM** produces:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
SEX	1	4.24172530	4.24172530	173.55	<.0001
AGE7	6	30.21321192	5.03553532	206.03	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
SEX	1	4.24198562	4.24198562	173.56	<.0001
AGE7	6	30.21321192	5.03553532	206.03	<.0001

- ▷ Type I tests focus on an effect, marginal over the others.
- ▷ Type III tests focus on an effect, given the others.
- ▷ Both are similar here: sex and age seem to have relatively independent effects.

- **PROC SURVEYREG** produces:

Tests of Model Effects

Effect	Num DF	F Value	Pr > F
Model	7	229.17	<.0001
Intercept	1	393731	<.0001
SEX	1	175.65	<.0001
AGE7	6	231.17	<.0001

- ▷ Apart from the sex and age effects, the overall **model effect**, referring to all covariates (sex and age here) simultaneously.
- ▷ The intercept effect refers to the null hypothesis that the intercept be zero; usually a less relevant hypothesis.

- PROC MIXED produces:

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
SEX	1	8376	173.56	<.0001
AGE7	6	8376	206.03	<.0001

- ▷ Type III tests are produced.
- ▷ In simple settings, the same results as with PROC GLM are obtained, but not always, since different estimation algorithms and approximations are used.

23.4.5 Overview Table

Logarithm of Body Mass Index (Belgium)					
Analysis	Procedure	sex		age	
		<i>F</i>	<i>p-value</i>	<i>F</i>	<i>p-value</i>
Ordinary linear regression					
2. SRS	GLM	173.56	<0.0001	206.03	<0.0001
3. SRS	MIXED (REML)	173.56	<0.0001	206.03	<0.0001
4. SRS	MIXED (ML)	173.73	<0.0001	206.23	<0.0001
5. SRS, ∞	SURVEYREG	175.65	<0.0001	231.17	<0.0001
Design-based linear regression					
5. SRS, ∞	SURVEYREG	175.65	<0.0001	231.17	<0.0001
6. SRS, 10^7	SURVEYREG	175.80	<0.0001	231.36	<0.0001
7. SRS, 8384	SURVEYREG	0	1.0000	0	1.0000
8. weighted	SURVEYREG	104.58	<0.0001	115.56	<0.0001
9. stratified	SURVEYREG	175.54	<0.0001	231.20	<0.0001
10. clustered	SURVEYREG	195.36	<0.0001	231.94	<0.0001
11. all	SURVEYREG	120.90	<0.0001	113.56	<0.0001
Hierarchical linear regression					
12. wt, clust	MIXED (REML, default)	281.94	<0.0001	262.74	<0.0001
12. wt, clust	MIXED (ML, Kenward-Roger)	281.89	<0.0001	262.66	<0.0001

- We can see the impact of design choices on the tests:
 - ▷ **Stratification** has little impact.
 - ▷ **Weighting** reduces efficiency.
 - ▷ **Clustering** properly partitions the variability and increases efficiency.
 - ▷ **All**: the net result is a smaller test statistic.

Hence, failing to accommodate the survey design might declare effects significant that, in fact, are not.
- The difference between **Kenward-Roger** and the default in the MIXED procedure is small since there is a large number of households.

Chapter 24

Logistic Regression

- ▷ Ordinary logistic regression
- ▷ Logistic regression for survey data
- ▷ Generalized estimating equations
- ▷ Generalized linear mixed model
- ▷ Mean estimation with GEE and GLMM

24.1 Concept

- In our mean estimation endeavors, we employed procedures for logistic regression.
- This implies we can conduct genuine logistic regression, using:
 - ▷ **PROC LOGISTIC, PROC GENMOD**: Conventional logistic regression procedures
 - ▷ **PROC SURVEYLOGISTIC**: Design-based logistic regression procedure
 - ▷ **PROC GENMOD with REPEATED statement**: Marginal logistic regression tool for hierarchical data: GEE
 - ▷ **PROC GLIMMIX, PROC NLMIXED**: Mixed-model based logistic regression procedure for hierarchical data, based on the GLMM
- Several procedures will work for non-binary data, such as ordinal, nominal, and count data, as well.

24.2 Model

- Assume we are interested in the effect of **sex** and **age** on **SGP**.
- **Sex** is a binary variable, necessitating a single parameter.
- As before, define **age** as a 7-point ordinal variable **age7**, together with its dummies $A_\ell = \text{agegr}_\ell$.

- We now consider the following basic logistic regression model:

$$\theta_i = \gamma_0 + \gamma_1 S_i + \gamma_{21} A_{1i} + \gamma_{22} A_{2i} + \gamma_{23} A_{3i} + \gamma_{24} A_{4i} + \gamma_{25} A_{5i} + \gamma_{26} A_{6i}$$

$$P[Z_i = 1 | S_i, A_{1i}, \dots, A_{6i}] = \frac{e^{\theta_i}}{1 + e^{\theta_i}}$$

where

- ▷ Z_i is SGP for respondent i
- ▷ S_i still is sex of respondent i (0 for males; 1 for females)
- ▷ $A_{\ell i}$ still is the value age-dummy ℓ takes for respondent i

- With GEE, the above model changes to

$$\theta_{ij} = \gamma_0 + \gamma_1 S_{ij} + \gamma_{21} A_{1ij} + \gamma_{22} A_{2ij} + \gamma_{23} A_{3ij} + \gamma_{24} A_{4ij} + \gamma_{25} A_{5ij} + \gamma_{26} A_{6ij}$$

$$P[Z_{ij} = 1 | S_{ij}, A_{1ij}, \dots, A_{6ij}] = \frac{e^{\theta_{ij}}}{1 + e^{\theta_{ij}}}$$

$$\text{Corr}(Z_{ij}, Z_{ik}) = \alpha$$

where now

- ▷ Z_{ij} is SGP for individual j in household i
- ▷ S_{ij} is sex of individual j in household i
- ▷ $A_{\ell ij}$ is the value age-dummy ℓ takes for individual j in household i

- With GLMM, the model becomes

$$\begin{aligned}\theta_{ij} = & \gamma_0 + g_i + \gamma_1 S_{ij} \\ & + \gamma_{21} A_{1ij} + \gamma_{22} A_{2ij} + \gamma_{23} A_{3ij} + \gamma_{24} A_{4ij} + \gamma_{25} A_{5ij} + \gamma_{26} A_{6ij}\end{aligned}$$

$$P[Z_{ij} = 1 | S_{ij}, A_{1ij}, \dots, A_{6ij}] = \frac{e^{\theta_{ij}}}{1 + e^{\theta_{ij}}}$$

where now, in addition,

▷ g_i is a household-level effect on LNBMI: $g_i \sim N(0, \tau^2)$

24.3 Programs

24.3.1 Programs for Ordinary Linear Regression

- Ordinary logistic regression can be coded using the dedicated SAS procedures **PROC LOGISTIC** and **PROC GENMOD**.
- Let us first consider **PROC LOGISTIC**:

```
proc logistic data=m.bmi_voeg;  
title '1. Ordinary logistic regression, for Belgium';  
title2 'with PROC LOGISTIC';  
class age7 / param=ref;  
model sgp = sex age7;  
contrast 'sex' sex 1;  
contrast 'age7' age7 1 0 0 0 0 0 -1,  
                age7 0 1 0 0 0 0 -1,  
                age7 0 0 1 0 0 0 -1,  
                age7 0 0 0 1 0 0 -1,  
                age7 0 0 0 0 1 0 -1,  
                age7 0 0 0 0 0 1 -1;  
  
run;
```

```
proc logistic data=m.bmi_voeg;  
title '1a. Ordinary logistic regression, for Belgium';  
title2 'with PROC LOGISTIC - with effect coding';  
class age7;  
model sgp = sex age7;  
run;
```

- ▷ We have used PROC GENMOD before.
- ▷ PROC LOGISTIC was historically the first procedure to fit logistic (and probit) regression.
- ▷ Hence, there is no need to specify the distribution and the default link function is the logit link.
- ▷ The default coding for dummy variables is so-called **effect coding**: every dummy parameter is a comparison between a particular category and the last category. To change this to the **reference coding**, where simply the last (seventh in our case) parameter is set equal to zero, the '**param=ref**' option is included in the **CLASS** statement.

- ▷ We will illustrate the difference by comparing both versions.
- ▷ The **CONTRAST** statement is included since the LOGISTIC procedure does not automatically provide tests for the null hypothesis of no effect in case two or more dummy variables are used.
- ▷ Thus, here, the two instances of the CONTRAST statement refer to, respectively:

$H_{0,1}$: Sex has no effect on SGP.

$H_{0,2}$: Age has no effect on SGP.

- ▷ Equivalently:

$$H_{0,1} : \gamma_1 = 0$$

$$H_{0,2} : \gamma_{21} = \gamma_{22} = \gamma_{23} = \gamma_{24} = \gamma_{25} = \gamma_{26} = 0$$

▷ Indeed, more than one CONTRAST statement is allowed.

- The equivalent PROC GENMOD code is:

```
proc genmod data=m.bmi_voeg;
title '2. Ordinary logistic regression, for Belgium';
title2 'with PROC GENMOD';
class age7;
model sgp = sex age7 / dist=b;
contrast 'sex' sex 1;
contrast 'age7' age7 1 0 0 0 0 0 -1,
                    age7 0 1 0 0 0 0 -1,
                    age7 0 0 1 0 0 0 -1,
                    age7 0 0 0 1 0 0 -1,
                    age7 0 0 0 0 1 0 -1,
                    age7 0 0 0 0 0 1 -1;

run;
```

▷ We have used PROC GENMOD before.

▷ Also here, CONTRAST statements are used.

- We can also use the GLM procedures **PROC GLIMMIX** and **PROC NLMIXED**, without the hierarchical features, to fit an ordinary logistic regression model:

```
proc glimmix data=m.bmi_voeg;  
title '3. Ordinary logistic regression, for Belgium';  
title2 'with proc glimmix';  
nloptions maxiter=50;  
class age7;  
model sgp = sex age7 / solution dist=b;  
run;
```

```
proc nlmixed data=m.bmi_voeg;  
title '4. Ordinary logistic regression, for Belgium';  
title2 'with PROC NLMIXED - ML estimation';  
theta = beta0 + beta1*sex + beta21*agegr1 + beta22*agegr2  
        + beta23*agegr3 + beta24*agegr4 + beta25*agegr5  
        + beta26*agegr6;  
exptheta = exp(theta);  
p = exptheta/(1+exptheta);  
model sgp ~ binary(p);  
run;
```

```
proc nlmixed data=m.bmi_voeg;  
title '4a. Ordinary logistic regression, for Belgium';  
title2 'with PROC NLMIXED - ML estimation';  
title3 'for lik ratio test';  
theta = beta0 + beta1*sex;  
exptheta = exp(theta);  
p = exptheta/(1+exptheta);  
model sgp ~ binary(p);  
run;
```

- ▷ The syntax of both procedures, used here, is a straightforward extension of the versions used for mean estimation.
- ▷ Note, in particular, the linear predictor θ_i has to be spelled out in every exact detail in the NLMIXED procedure.
- ▷ Since NLMIXED is essentially a non-linear procedure, there is **no CONTRAST** statement, which is confined to linear combinations of parameters.

- ▷ An easy solution is by using the **likelihood ratio test**, through fitting a model with and without the **age** parameters.

The difference between both likelihoods at maximum follows, asymptotically, a χ^2_6 distribution.

- ▷ A similar undertaking for the **sex** effect is not necessary, since it is a 1-parameter effect, and a test follows from the parameter estimates table.

- Finally, we can employ the design-based regression procedure **PROC SURVEYLOGISTIC**, but confine it to SRS:

```
proc surveylogistic data=m.bmi_voeg;  
title '5a. Surveylogistic, SRS, infinite population';  
class age7 / param=ref;  
model sgp = sex age7;  
run;
```

- ▷ Used in this fashion, the procedure is syntactically similar to PROC LOGISTIC, PROC GENMOD, and PROC GLIMMIX.

- ▷ Note, also here, the need to change the default **effect coding** to **reference coding**, in line with the LOGISTIC procedure.
- ▷ The same was not true for the linear regression procedures, where the **reference coding** is the default.

24.3.2 Programs for Design-Based Linear Regression

- Starting from the **PROC SURVEYLOGISTIC** program on page 706:

```
proc surveylogistic data=m.bmi_voeg;  
title '5a. Surveylogistic, SRS, infinite population';  
class age7 / param=ref;  
model sgp = sex age7;  
run;
```

a number of design features can be built in:

- ▷ **Finite population:** the **PROC SURVEYLOGISTIC** statement changes to:

```
proc surveylogistic data=m.bmi_voeg total=10000000;
```

- ▷ **Census-finite population:** the **PROC SURVEYLOGISTIC** statement changes to:

```
proc surveylogistic data=m.bmi_voeg total=8532;
```

- ▷ **Stratification:** the following statement is added:

```
strata province;
```

- ▷ **Two-stage sampling (clustering):** the following statement is added:

```
cluster hh;
```

- ▷ **Weighting:** the following statement is added:

```
weight wfin;
```

- ▷ **Maximal accommodation for design:** the program becomes:

```
proc surveylogistic data=m.bmi_voeg total=10000000;  
title '11. Surveylogistic, weighted, stratified, two-stage (clustered),';  
title2 'finite population';  
class age7 / param=ref;  
weight wfin;  
strata province;  
cluster hh;  
model sgp = sex age7;  
run;
```

24.3.3 Programs for Generalized Estimating Equations

- The design features that can be accommodated in PROC MIXED are weighting and clustering:

```
proc genmod data=m.bmi_voeg;
title '12. GEE logistic regression, for Belgium';
title2 'weighted + clustered';
weight wfin;
class age7 hh;
model sgp = sex age7 / dist=b;
repeated subject = hh / type=cs corrw model;
contrast 'sex' sex 1;
contrast 'age7' age7 1 0 0 0 0 0 -1,
                  age7 0 1 0 0 0 0 -1,
                  age7 0 0 1 0 0 0 -1,
                  age7 0 0 0 1 0 0 -1,
                  age7 0 0 0 0 1 0 -1,
                  age7 0 0 0 0 0 1 -1;

run;
```

- Also here, CONTRAST statements are needed for the test statistics.

24.3.4 Programs for the Generalized Linear Mixed Model

- The design features that can be accommodated in PROC GLIMMIX are weighting and clustering:

```
proc glimmix data=m.bmi_voeg empirical;  
title '13. GLMM, for Belgium';  
title2 'with proc glimmix';  
title3 'weighted + two-stage (cluster)';  
nloptions maxiter=50;  
weight wfin;  
class age7;  
model sgp = sex age7 / solution dist=b;  
random intercept / subject = hh type=un;  
run;
```

- The **NLMIXED** procedure only accommodates the clustering feature:

```
proc nlmixed data=m.bmi_voeg;  
title '14. GLMM, for Belgium';  
title2 'with PROC NLMIXED';  
title3 'two-stage (cluster)';  
theta = beta0 + b + beta1*sex + beta21*agegr1 + beta22*agegr2  
        + beta23*agegr3 + beta24*agegr4 + beta25*agegr5  
        + beta26*agegr6;  
exptheta = exp(theta);  
p = exptheta/(1+exptheta);  
model sgp ~ binary(p);  
random b ~ normal(0,tau2) subject=hh;  
run;
```

- Like the MIXED procedure, GLIMMIX allows for multiple **RANDOM** statement, while NLMIXED allows for only one.
- As before, a second copy of the NLMIXED program is needed to conduct a likelihood ratio test.

In the second program the age dummies are omitted.

24.4 Parameter Estimates

24.4.1 Selected Output

- Consider **PROC LOGISTIC** for ordinary logistic regression.
 - ▷ The program version with the '**param=ref**' option produces the following class level information:

Class Level Information

Class	Value	Design Variables					
AGE7	1	1	0	0	0	0	0
	2	0	1	0	0	0	0
	3	0	0	1	0	0	0
	4	0	0	0	1	0	0
	5	0	0	0	0	1	0
	6	0	0	0	0	0	1
	7	0	0	0	0	0	0

whereas the default is:

Class Level Information

Class	Value	Design Variables					
AGE7	1	1	0	0	0	0	0
	2	0	1	0	0	0	0
	3	0	0	1	0	0	0
	4	0	0	0	1	0	0
	5	0	0	0	0	1	0
	6	0	0	0	0	0	1
	7	-1	-1	-1	-1	-1	-1

- ▷ In the first case, the intercept corresponds to the seventh and last dummy category, whereas in the second case the intercept has the meaning of an average over all categories.
- ▷ Parameter estimates for the first and second versions, respectively:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.4678	0.2236	121.7901	<.0001
SEX	1	-0.4398	0.0749	34.4897	<.0001
AGE7 1	1	1.0383	0.2150	23.3229	<.0001
AGE7 2	1	1.2748	0.2062	38.2370	<.0001
AGE7 3	1	1.0939	0.2082	27.6049	<.0001
AGE7 4	1	0.7088	0.2180	10.5766	0.0011
AGE7 5	1	0.6776	0.2230	9.2319	0.0024
AGE7 6	1	0.2433	0.2364	1.0593	0.3034

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.7482	0.1164	225.6591	<.0001
SEX	1	-0.4398	0.0749	34.4897	<.0001
AGE7 1	1	0.3188	0.0915	12.1433	0.0005
AGE7 2	1	0.5553	0.0756	54.0044	<.0001
AGE7 3	1	0.3743	0.0794	22.2475	<.0001
AGE7 4	1	-0.0107	0.0962	0.0124	0.9114
AGE7 5	1	-0.0419	0.1043	0.1616	0.6877
AGE7 6	1	-0.4762	0.1236	14.8484	0.0001

- ▷ While the estimates are different (except for the **sex** effect), one set transforms linearly into the other set.

- **PROC GENMOD** for ordinary logistic regression:

Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-2.4684	0.2237	-2.9067	-2.0300	121.80	<.0001
SEX	1	-0.4398	0.0749	-0.5865	-0.2930	34.49	<.0001
AGE7	1	1.0389	0.2150	0.6174	1.4604	23.34	<.0001
AGE7	2	1.2754	0.2062	0.8713	1.6796	38.25	<.0001
AGE7	3	1.0945	0.2082	0.6863	1.5026	27.62	<.0001
AGE7	4	0.7094	0.2180	0.2822	1.1367	10.59	0.0011
AGE7	5	0.6782	0.2231	0.2410	1.1154	9.24	0.0024
AGE7	6	0.2438	0.2364	-0.2195	0.7072	1.06	0.3024
AGE7	7	0.0000	0.0000	0.0000	0.0000	.	.
Scale	0	1.0000	0.0000	1.0000	1.0000		

- The estimates are the same,

- as is the case for the PROC GLIMMIX version:

Effect	AGE7	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		2.4684	0.2237	8524	11.04	<.0001
SEX		0.4398	0.07488	8524	5.87	<.0001
AGE7	1	-1.0389	0.2150	8524	-4.83	<.0001
AGE7	2	-1.2754	0.2062	8524	-6.19	<.0001
AGE7	3	-1.0945	0.2082	8524	-5.26	<.0001
AGE7	4	-0.7094	0.2180	8524	-3.25	0.0011
AGE7	5	-0.6782	0.2231	8524	-3.04	0.0024
AGE7	6	-0.2438	0.2364	8524	-1.03	0.3024
AGE7	7	0

the PROC NLMIXED version:

Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
beta0	2.4683	0.2237	8532	11.04	<.0001	0.05	2.0298	2.9067
beta1	0.4398	0.07488	8532	5.87	<.0001	0.05	0.2930	0.5866
beta21	-1.0388	0.2150	8532	-4.83	<.0001	0.05	-1.4604	-0.6173
beta22	-1.2754	0.2062	8532	-6.18	<.0001	0.05	-1.6796	-0.8712
beta23	-1.0944	0.2082	8532	-5.26	<.0001	0.05	-1.5026	-0.6862
beta24	-0.7094	0.2180	8532	-3.25	0.0011	0.05	-1.1367	-0.2821
beta25	-0.6782	0.2231	8532	-3.04	0.0024	0.05	-1.1154	-0.2409
beta26	-0.2437	0.2364	8532	-1.03	0.3026	0.05	-0.7072	0.2197

and the PROC SURVEYLOGISTIC version:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.4678	0.2261	119.1579	<.0001
SEX	1	-0.4398	0.0750	34.4267	<.0001
AGE7	1	1.0383	0.2152	23.2771	<.0001
AGE7	2	1.2748	0.2060	38.3096	<.0001
AGE7	3	1.0939	0.2081	27.6175	<.0001
AGE7	4	0.7088	0.2177	10.5967	0.0011
AGE7	5	0.6776	0.2235	9.1894	0.0024
AGE7	6	0.2433	0.2362	1.0605	0.3031

- PROC SURVEYLOGISTIC for a finite and a census-finite population:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.4678	0.2260	119.2596	<.0001
SEX	1	-0.4398	0.0749	34.4561	<.0001
AGE7 1	1	1.0383	0.2151	23.2969	<.0001
AGE7 2	1	1.2748	0.2059	38.3423	<.0001
AGE7 3	1	1.0939	0.2081	27.6411	<.0001
AGE7 4	1	0.7088	0.2177	10.6058	0.0011
AGE7 5	1	0.6776	0.2234	9.1973	0.0024
AGE7 6	1	0.2433	0.2361	1.0615	0.3029

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.4678	0	.	.
SEX	1	-0.4398	0	.	.
AGE7 1	1	1.0383	0	.	.
AGE7 2	1	1.2748	0	.	.
AGE7 3	1	1.0939	0	.	.
AGE7 4	1	0.7088	0	.	.
AGE7 5	1	0.6776	0	.	.
AGE7 6	1	0.2433	0	.	.

- PROC SURVEYLOGISTIC for all design aspects combined:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9833	0.4645	41.2560	<.0001
SEX	1	-0.3217	0.1026	9.8407	0.0017
AGE7	1	1.0020	0.4582	4.7827	0.0287
AGE7	2	1.1202	0.4504	6.1852	0.0129
AGE7	3	1.0737	0.4600	5.4485	0.0196
AGE7	4	0.5692	0.4614	1.5215	0.2174
AGE7	5	0.2009	0.4683	0.1841	0.6679
AGE7	6	0.3574	0.5028	0.5054	0.4771

- **PROC GENMOD** with **REPEATED** for GEE, accommodating weighting and clustering.
 - ▷ The working correlation is considerable, underscoring the strong correlation in SGP within a household:

Exchangeable Working
Correlation

Correlation 0.3943526021

- ▷ Recall that the working correlation structure does not need to be correctly specified and hence should not be overinterpreted.
Nevertheless, we obtain a good indication about the average correlation between HH members in terms of SGP.

▷ The parameter estimates:

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept		-2.9657	0.3765	-3.7037	-2.2277	-7.88	<.0001
SEX		-0.2620	0.0823	-0.4233	-0.1008	-3.19	0.0014
AGE7	1	1.0632	0.3766	0.3252	1.8013	2.82	0.0047
AGE7	2	1.0754	0.3719	0.3464	1.8043	2.89	0.0038
AGE7	3	1.0436	0.3814	0.2962	1.7911	2.74	0.0062
AGE7	4	0.6295	0.3735	-0.1026	1.3616	1.69	0.0919
AGE7	5	0.2568	0.3820	-0.4919	1.0054	0.67	0.5015
AGE7	6	0.3822	0.4020	-0.4057	1.1701	0.95	0.3417
AGE7	7	0.0000	0.0000	0.0000	0.0000	.	.

- ▷ Recall that it is imperative to use the empirically corrected standard errors, since the purely model based ones do not properly deal with the weights:

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept		-2.9657	0.0105	-2.9862	-2.9451	-282.95	<.0001
SEX		-0.2620	0.0024	-0.2668	-0.2573	-108.11	<.0001
AGE7	1	1.0632	0.0102	1.0432	1.0833	104.07	<.0001
AGE7	2	1.0754	0.0102	1.0554	1.0953	105.88	<.0001
AGE7	3	1.0436	0.0102	1.0237	1.0635	102.77	<.0001
AGE7	4	0.6295	0.0103	0.6092	0.6498	60.89	<.0001
AGE7	5	0.2568	0.0111	0.2350	0.2785	23.16	<.0001
AGE7	6	0.3822	0.0110	0.3606	0.4038	34.69	<.0001
AGE7	7	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1.0000

- The **GLIMMIX** procedure for the GLMM:

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Standard Error
UN(1,1)	HH	40.1435	0.9785

Effect	AGE7	Estimate	Standard Error	DF	t Value	Pr > t
Intercept		7.8965	1.1015	4661	7.17	<.0001
SEX		0.7908	0.3995	3863	1.98	0.0478
AGE7	1	-1.8937	1.3218	3863	-1.43	0.1520
AGE7	2	-1.6106	1.4072	3863	-1.14	0.2525
AGE7	3	-1.3059	1.2926	3863	-1.01	0.3124
AGE7	4	-0.7893	1.4074	3863	-0.56	0.5750
AGE7	5	-0.1224	1.4225	3863	-0.09	0.9315
AGE7	6	-1.5910	1.2234	3863	-1.30	0.1935
AGE7	7	0

- The **NLMIXED** procedure for the GLMM:

Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha	Lower	Upper
beta0	4.7993	0.3553	4661	13.51	<.0001	0.05	4.1027	5.4958
beta1	0.6810	0.1112	4661	6.12	<.0001	0.05	0.4629	0.8991
beta21	-1.8426	0.3295	4661	-5.59	<.0001	0.05	-2.4886	-1.1967
beta22	-2.0590	0.3169	4661	-6.50	<.0001	0.05	-2.6803	-1.4377
beta23	-1.7769	0.3177	4661	-5.59	<.0001	0.05	-2.3996	-1.1541
beta24	-1.2543	0.3282	4661	-3.82	0.0001	0.05	-1.8978	-0.6108
beta25	-0.9829	0.3364	4661	-2.92	0.0035	0.05	-1.6424	-0.3233
beta26	-0.4115	0.3400	4661	-1.21	0.2262	0.05	-1.0780	0.2550
tau2	8.1683	0.6827	4661	11.97	<.0001	0.05	6.8300	9.5067

- ▷ Note that, as before, the GLMM parameters are much larger in absolute values than their marginal counterparts, for reasons studied before.
- ▷ Usually, the GLIMMIX estimates are biased downwards relative to the gold-standard NLMIXED ones.
- ▷ However, here, a direct comparison is difficult since the GLIMMIX parameters come from a model correcting for weighting and clustering, whereas in the NLMIXED syntax only clustering is taken into account.

24.4.2 Overview Table

Stable General Practitioner (Belgium)									
Analysis	Procedure	Parameter estimates (s.e.) $\times 10^2$							
		$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_{21}$	$\hat{\gamma}_{22}$	$\hat{\gamma}_{23}$	$\hat{\gamma}_{24}$	$\hat{\gamma}_{25}$	$\hat{\gamma}_{26}$
Ordinary logistic regression									
1.–5. SRS	several*	-247(22)	-44(7)	104(22)	127(21)	109(21)	71(22)	68(22)	24(24)
Design-based logistic regression									
5. SRS, ∞	SURVEYLOGISTIC	-247(22)	-44(7)	104(22)	127(21)	109(21)	71(22)	68(22)	24(24)
6. SRS, 10^7	SURVEYLOGISTIC	-247(23)	-44(7)	104(22)	127(21)	109(21)	71(22)	68(22)	24(24)
7. SRS, 8384	SURVEYLOGISTIC	-247(0)	-44(0)	104(0)	127(0)	109(0)	71(0)	68(0)	24(0)
8. weighted	SURVEYLOGISTIC	-298(38)	-32(11)	100(37)	112(36)	107(37)	57(37)	20(37)	36(42)
9. stratified	SURVEYLOGISTIC	-247(23)	-44(8)	104(22)	127(21)	109(21)	71(22)	68(22)	24(24)
10. clustered	SURVEYLOGISTIC	-247(23)	-44(6)	104(23)	127(22)	109(22)	71(23)	68(24)	24(25)
11. all	SURVEYLOGISTIC	-298(46)	-32(10)	100(46)	112(45)	107(46)	57(46)	20(47)	36(50)
Hierarchical logistic regression									
12. wt, clust	GEMNOD	-297(38)	-26(8)	106(38)	108(37)	104(38)	63(37)	25(38)	38(40)
13. wt, clust	GLIMMIX	-790(110)	-79(40)	189(132)	161(141)	131(129)	79(141)	12(142)	159(122)
14. clust	NLMIXED	-480(36)	-68(11)	184(33)	206(32)	177(32)	125(33)	98(34)	41(34)

*: LOGISTIC, GENMOD, GLIMMIX, NLMIXED, SURVEYLOGISTIC (SRS)

- As for LNBMI, the largest impact is seen for the weighted analyses.
- Recall the relationship between the marginal (GEE) and random-effects parameters (GLMM):

$$\frac{\hat{\beta}^{\text{RE}}}{\hat{\beta}^{\text{M}}} = \sqrt{c^2\tau^2 + 1} > 1, \quad \tau^2 = \text{variance random intercepts}$$

$$c = 16\sqrt{3}/(15\pi)$$

In our case, this becomes:

$$\sqrt{c^2\tau^2 + 1} = \sqrt{0.5881^2 \times 8.17 + 1} = 1.96$$

It is hard to verify the relationship pragmatically:

- ▷ NLMIXED (GLMM) does not correct for weighting, while GENMOD (GEE) does.
- ▷ The GLIMMIX parameter estimates are hard to trust, given the severe bias inherent to this approximate method.
- Recall, once more, that the GLMM parameters have a **different, HH-specific** interpretation, and hence cannot be compared directly to the other analyses.

24.4.3 Hypothesis Testing

- As stated before, we are interested in:

$H_{0,1}$: Sex has no effect on SGP.

$H_{0,2}$: Age has no effect on SGP.

- Mathematically translated:

$$H_{0,1} : \gamma_1 = 0$$

$$H_{0,2} : \gamma_{21} = \gamma_{22} = \gamma_{23} = \gamma_{24} = \gamma_{25} = \gamma_{26} = 0$$

- In the linear context, F tests are rather prominent.
- The situation is less unambiguous with non-Gaussian, e.g., binary, data.
- Some procedures, like GLIMMIX, implement approximate F tests.
- Note that this corresponds to a squared t test for a single parameter:

$$F_{1,d_2} \equiv t_{d_1}^2$$

- The asymptotic versions for $d_2 \rightarrow \infty$ is a Wald test, with then

$$t_{d_2} \rightarrow Z \sim N(0, 1)$$
$$F_{d_1,d_2} \rightarrow X_{d_2}^2 \sim \chi_{d_2}^2$$

- A Wald test essentially compares the difference between a parameter (set of parameters) and its null values on the one hand with its variance (variance-covariance matrix) on the other hand.
- Alternatively, a **likelihood ratio test** can be constructed by fitting a model with and without a set of d_1 parameters, then calculating the double difference between the log-likelihoods at maximum, and referring it to a $\chi^2_{d_1}$:

$$2(\hat{\ell}_1 - \hat{\ell}_0) \sim \chi^2_{d_1}$$

with ℓ_1 (ℓ_0) the log-likelihood under the alternative (null) hypothesis.

- Finally, a **score test** can be considered, which compares the score function (first derivative of the log-likelihood) of the alternative model, evaluated in the null model parameter estimate, to its precision.

- The score test statistics asymptotically follows a $\chi^2_{d_1}$.
- Asymptotically under the null, likelihood ratio (LR), Wald (W), and score (S) tests are equivalent.

24.4.4 Selected Output

- The output takes various forms.
- **PROC LOGISTIC** produces, by default and as a result of the **CONTRAST** statement:

Type 3 Analysis of Effects Wald

Effect	DF	Chi-Square	Pr > ChiSq
SEX	1	34.4897	<.0001
AGE7	6	87.2642	<.0001

Contrast Test Results Wald

Contrast	DF	Chi-Square	Pr > ChiSq
sex	1	34.4897	<.0001
age7	6	87.2642	<.0001

- PROC GENMOD consider the LR test rather than the W test:

Contrast Results				
Chi-				
Contrast	DF	Square	Pr > ChiSq	Type
sex	1	35.01	<.0001	LR
age7	6	97.69	<.0001	LR

- PROC GLIMMIX produces

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
SEX	1	8524	34.49	<.0001
AGE7	6	8524	14.55	<.0001

While the result for **sex** is similar, it is not at all for **age**, due to the relatively poor approximations used.

- PROC NLMIXED:

- ▷ For **sex**, being a single parameter, we can use the appropriate line in the parameter estimates panel:

Parameter	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
beta1	0.4398	0.07488	8532	5.87	<.0001	0.05

- ▷ Here, the p -value follows directly from the t test.
- ▷ In case one is interested in the F statistic:

$$F = t^2 = 5.87^2 = 34.46$$

- ▷ For the age effect, compare **minus twice the log-likelihood** from the model with and without the age effects:

-2 Log Likelihood	5276.0
-2 Log Likelihood	5373.7

producing $X^2 = 97.70$.

- PROC SURVEYLOGISTIC produces

Type 3 Analysis of Effects

Effect	DF	Wald	Pr > ChiSq
		Chi-Square	
SEX	1	34.4267	<.0001
AGE7	6	87.3918	<.0001

24.4.5 Overview Table

Stable General Practitioner (Belgium)						
Analysis	Procedure	Test	sex		age	
			<i>F</i>	<i>p</i> -value	<i>F</i>	<i>p</i> -value
Ordinary logistic regression						
1. SRS	LOGISTIC	Wald	34.49	<0.0001	87.26	<0.0001
2. SRS	GENMOD	χ^2	35.01	<0.0001	97.69	<0.0001
3. SRS	GLIMMIX	<i>F</i>	34.49	<0.0001	14.55	<0.0001
4. SRS	NLMIXED	<i>t</i>	5.87	<0.0001		
4. SRS	NLMIXED	<i>F</i>	34.46	<0.0001		
4. SRS	NLMIXED	LR			97.70	<0.0001
5. SRS, ∞	SURVEYLOGISTIC	Wald	34.43	<0.0001	87.39	<0.0001
Design-based logistic regression						
5. SRS, ∞	SURVEYLOGISTIC	Wald	34.43	<0.0001	87.39	<0.0001
6. SRS, 10^7	SURVEYLOGISTIC	Wald	34.46	<0.0001	87.47	<0.0001
7. SRS, 8384	SURVEYLOGISTIC	Wald	0.00	1.0000	0.00	1.0000
8. weighted	SURVEYLOGISTIC	Wald	8.13	0.0044	45.63	<0.0001
9. stratified	SURVEYLOGISTIC	Wald	34.41	<0.0001	87.44	<0.0001
10. clustered	SURVEYLOGISTIC	Wald	48.26	<0.0001	72.31	<0.0001
11. all	SURVEYLOGISTIC	Wald	9.84	0.0017	37.19	<0.0001
Hierarchical logistic regression						
12. wt, clust	GEMNOD	score	8.82	0.0030	41.00	<0.0001
13. wt, clust	GLIMMIX	<i>F</i>	3.92	0.0478	1.08	0.3706
14. clust	NLMIXED	<i>t</i>	6.12	<0.0001		
14. clust	NLMIXED	<i>F</i>	37.45	<0.0001		
14. clust	NLMIXED	LR			86.80	<0.0001

- We can see the impact of design choices on the tests:
 - ▷ **Stratification** has little impact.
 - ▷ **Weighting** reduces efficiency.
 - ▷ **Clustering** properly partitions the variability and increases efficiency.
 - ▷ **All**: the net result is a smaller test statistic.

Again, failing to accommodate the survey design might declare effects significant that, in fact, are not.
- The GLIMMIX results are, due to the poverty of the approximation, not trustworthy and have been italicized for this reason.
- The F tests in the NLMIXED procedures are simply the squares of the t tests.
- Recall that the 6-df test for **age** in these cases are conducted differently than the 1-df tests for **sex**.

Chapter 25

Selecting a Sample Using SURVEYSELECT

- ▷ General concept
- ▷ Example code for Surveytown
- ▷ Output for Surveytown

25.1 General Concept

- Assume the sample frame is given as a dataset.
- It is then possible to select a sample from it, using **PROC SURVEYSELECT**.
- The sampling methods allowed for are:
 - ▷ **SRS**: simple random sampling
 - ▷ **URS**: sampling with replacement (unrestricted random sampling)
 - ▷ **SYS**: systematic sampling
 - ▷ **SEQ**: sequential sampling: (a way of looping through a stratum, similar in spirit but different from systematic sampling)
 - ▷ **PPS**: sampling with probability proportional to size

- All of these methods can be combined with **STRATIFICATION**.
- The PPS method features several versions, essentially allowing for combination with the other methods (SRS, URS, SYS, and SEQ).
- A versatile collection of sampling methods results.

25.2 Example: Surveytown

- Let us assume Surveytown consists of the following information:

Surveytown sample frame

Obs	block	stratum	y	inhabitants
1	1	1	1	10
2	2	1	2	20
3	3	1	3	30
4	4	1	4	40
5	5	2	5	50
6	6	2	6	60
7	7	2	7	70
8	8	2	8	80

- ▷ The variables block, stratum, and Y (the number of inhabited lots) are in line with their earlier uses.
- ▷ The number of inhabitants is introduced as an example of a **size** variable for a block, to be used in what follows.

- Program for **SRS**:

```
title '1. surveyselect - Surveytown - SRS';  
proc surveyselect data=m.surveytown03 out=m.surveytown_srs  
method=srs n=4 rep=5 seed=498388;  
id block stratum y;  
run;
```

▷ The **SURVEYSELECT** procedure contains all of the essential information:

- * The input and output datasets.
- * The output dataset contains the sample(s) taken.
- * '**method=srs**' option specifies the choice for SRS; which here means SRS without replacement!
- * The '**n=4**' option specifies the size of a sample taken.
- * The '**rep=5**' option requests 5 executions of the sampling.

This is useful to study (asymptotic) properties, or just to study how a method behaves.

* The 'seed=' option initiates the random number generator. This is useful when we want to redo the same analysis.

▷ The ID statement specifies which variables are to be included in the output dataset.

- The output is as follows:

```
1. surveyselect - Surveytown - SRS
```

The SURVEYSELECT Procedure

Selection Method Simple Random Sampling

Input Data Set	SURVEYTOWN03
Random Number Seed	498388
Sample Size	4
Selection Probability	0.5
Sampling Weight	2
Number of Replicates	5
Total Sample Size	20
Output Data Set	SURVEYTOWN_SRS

- This is essentially book keeping information about the sampling method and its application to the set of data at hand.
- A print of the resulting output dataset displays the 5 samples taken, where
 - ▷ The REPLICATE variable is automatically added, to indicate the rank number of the particular sample taken.

1. surveyselect - Surveytown - SRS

Obs	Replicate	block	stratum	y
1	1	1	1	1
2	1	2	1	2
3	1	4	1	4
4	1	5	2	5
5	2	2	1	2
6	2	3	1	3
7	2	4	1	4
8	2	7	2	7
9	3	4	1	4
10	3	6	2	6
11	3	7	2	7
12	3	8	2	8
13	4	2	1	2
14	4	3	1	3
15	4	6	2	6
16	4	8	2	8
17	5	3	1	3
18	5	4	1	4
19	5	6	2	6
20	5	7	2	7

- Switching to **SYS**, we merely have to change one option:

```
method=sys
```

- This produces exactly the same book keeping information.
- The output dataset is:

2. surveyselect - Surveytown - SYS

Obs	Replicate	block	stratum	y
1	1	1	1	1
2	1	3	1	3
3	1	5	2	5
4	1	7	2	7
5	2	2	1	2
6	2	4	1	4
7	2	6	2	6
8	2	8	2	8
9	3	2	1	2
10	3	4	1	4
11	3	6	2	6
12	3	8	2	8
13	4	1	1	1
14	4	3	1	3
15	4	5	2	5
16	4	7	2	7
17	5	2	1	2
18	5	4	1	4
19	5	6	2	6
20	5	8	2	8

▷ We clearly see the impact of the method: only two possible samples arise:

* {1, 3, 5, 7}

* {2, 4, 6, 8}

- For **SRS with replacement (URS)**, the option changes to:

```
method=urs
```

- This produces a slightly updated book keeping panel:

```
3. surveyselect - Surveytown - SRS & replacement
The SURVEYSELECT Procedure
```

Selection Method	Unrestricted Random Sampling
Input Data Set	SURVEYTOWN03
Random Number Seed	498388
Sample Size	4
Expected Number of Hits	0.5
Sampling Weight	2
Number of Replicates	5
Total Sample Size	20
Output Data Set	SURVEYTOWN_SYS

- ▷ The expected number of hits is the probability that an unit will be selected, it is not different from the SRS and SYS selection probability, as we have seen before.

- The output dataset:

3. surveyselect - Surveytown - SRS & replacement

Obs	Replicate	block	stratum	y	Number Hits
1	1	2	1	2	2
2	1	5	2	5	1
3	1	6	2	6	1
4	2	2	1	2	1
5	2	3	1	3	1
6	2	6	2	6	1
7	2	7	2	7	1
8	3	5	2	5	1
9	3	6	2	6	1
10	3	7	2	7	1
11	3	8	2	8	1

12	4	2	1	2	1
13	4	4	1	4	1
14	4	8	2	8	2
15	5	4	1	4	1
16	5	6	2	6	2
17	5	8	2	8	1

- ▷ We clearly see that some units are selected more than once.
 - ▷ This is indicated by the variable 'Number Hits'.
 - ▷ For example, the first sample consists of blocks 2, 2, 5, and 6.
- Switching to **stratification**, this is coded by combining '**method=srs**' with the **STRATA** statement:

```

title '4. surveyselect - Surveytown - stratified';
proc surveyselect data=m.surveytown03 out=m.surveytown_strat
    method=srs n=(2 2) rep=5 seed=498388;
strata stratum;
id block stratum y;
run;

```

- ▷ Note that we use the `'n=(2 2)'` to indicate that our sample should consist of 2 units from the first and two from the second stratum.
- Before printing the output dataset, it is useful to order it by replicate, rather than the default, which is by stratum:

```
proc sort data=m.surveytown_strat;  
by replicate;  
run;
```

- The book keeping information now is:

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling
Strata Variable	stratum

Input Data Set	SURVEYTOWN03
Random Number Seed	498388
Number of Strata	2
Number of Replicates	5
Total Sample Size	20
Output Data Set	SURVEYTOWN_STRAT

- The 5 samples look like:

4. surveyselect - Surveytown - stratified

Obs	stratum	Replicate	block	y	Selection Prob	Sampling Weight
1	1	1	1	1	0.5	2
2	1	1	3	3	0.5	2
3	2	1	6	6	0.5	2
4	2	1	8	8	0.5	2
5	1	2	1	1	0.5	2
6	1	2	2	2	0.5	2
7	2	2	6	6	0.5	2
8	2	2	8	8	0.5	2
9	1	3	1	1	0.5	2
10	1	3	2	2	0.5	2
11	2	3	5	5	0.5	2
12	2	3	8	8	0.5	2
13	1	4	1	1	0.5	2
14	1	4	4	4	0.5	2
15	2	4	6	6	0.5	2
16	2	4	7	7	0.5	2
17	1	5	3	3	0.5	2
18	1	5	4	4	0.5	2
19	2	5	7	7	0.5	2
20	2	5	8	8	0.5	2

- ▷ Every sample nicely has 2 units from each stratum, as requested.
- ▷ The selection probabilities are all equal, and hence the sampling weight.
- This last observation is not always true: assume we change the subsample sizes by changing to `'n=(1,3)'`.
- We then obtain:

```
5. surveyselect - Surveytown - stratified/unequal prob
```

Obs	stratum	Replicate	block	y	Selection Prob	Sampling Weight
1	1	1	1	1	0.25	4.00000
2	2	1	6	6	0.75	1.33333
3	2	1	7	7	0.75	1.33333
4	2	1	8	8	0.75	1.33333

5	1	2	3	3	0.25	4.00000
6	2	2	5	5	0.75	1.33333
7	2	2	7	7	0.75	1.33333
8	2	2	8	8	0.75	1.33333
9	1	3	3	3	0.25	4.00000
10	2	3	5	5	0.75	1.33333
11	2	3	6	6	0.75	1.33333
12	2	3	7	7	0.75	1.33333
13	1	4	1	1	0.25	4.00000
14	2	4	5	5	0.75	1.33333
15	2	4	6	6	0.75	1.33333
16	2	4	8	8	0.75	1.33333
17	1	5	3	3	0.25	4.00000
18	2	5	5	5	0.75	1.33333
19	2	5	6	6	0.75	1.33333
20	2	5	7	7	0.75	1.33333

▷ Now, there is always only 1 unit from the first stratum, while there are 3 from the second.

- ▷ To compensate for this, the sampling weights are inversely proportional to the selection probability, so that proper weighted estimators can be used.
- Assume we want to sample **proportional to size**, and assume the size is given by the number of inhabitants.
- The following program can be used:

```
title '6. surveyselect - Surveytown - prop. to size';  
proc surveyselect data=m.surveytown03  
                  out=m.surveytown_pps  
                  method=pps  
  
n=4  
rep=5  
  
                  seed=498388;  
size inhabitants;  
run;
```

- ▷ The **SIZE** statement is needed to specify which variable will be used as a measure for a block's size.

- The book keeping output is as follows:

```
6. surveyselect - Surveytown - prop. to size
```

The SURVEYSELECT Procedure

Selection Method	PPS, Without Replacement
Size Measure	inhabitants

Input Data Set	SURVEYTOWN03
Random Number Seed	498388
Sample Size	4
Number of Replicates	5
Total Sample Size	20
Output Data Set	SURVEYTOWN_PPS

- The samples taken:

6. surveyselect - Surveytown - prop. to size

Obs	Replicate	block	stratum	y	inhabitants	Selection Prob	Sampling Weight
1	1	4	1	4	40	0.44444	2.25000
2	1	6	2	6	60	0.66667	1.50000
3	1	7	2	7	70	0.77778	1.28571
4	1	8	2	8	80	0.88889	1.12500
5	2	2	1	2	20	0.22222	4.50000
6	2	5	2	5	50	0.55556	1.80000
7	2	6	2	6	60	0.66667	1.50000
8	2	8	2	8	80	0.88889	1.12500
9	3	4	1	4	40	0.44444	2.25000
10	3	6	2	6	60	0.66667	1.50000
11	3	7	2	7	70	0.77778	1.28571
12	3	8	2	8	80	0.88889	1.12500
13	4	3	1	3	30	0.33333	3.00000
14	4	5	2	5	50	0.55556	1.80000
15	4	6	2	6	60	0.66667	1.50000
16	4	8	2	8	80	0.88889	1.12500
17	5	5	2	5	50	0.55556	1.80000
18	5	6	2	6	60	0.66667	1.50000
19	5	7	2	7	70	0.77778	1.28571
20	5	8	2	8	80	0.88889	1.12500

- ▷ Note that the selection probability proportionally increases with the number of inhabitants.
- ▷ As a result, the sampling weight inversely decreases with it.

Chapter 26

Some Selected Examples From STATA

- ▷ Selected programs
- ▷ Selected output

26.1 Programs

```
use "bmi_voeg.dta", clear
log using bmi_voeg.log, replace
label list
svymean bmi voeg lnbmi lnvoeg
      [pw=wfin], by(region) strata(province) psu(hh) obs ci
svyset,clear
svyprop sgp
      [pw=wfin],by(region) strata(province) psu(hh)
svyset,clear
svyreg lnbmi wal fla sex agegr2 agegr3 agegr4 agegr5 agegr6 agegr7
      eduprim edusec inclow incmed ta2
      [pw=wfin], strata(province) psu(hh)
svyset,clear
svylogit sgp wal fla sex agegr2 agegr3 agegr4 agegr5 agegr6 agegr7
      eduprim edusec inclow incmed ta2
      [pw=wfin],or strata(province) psu(hh)
svyset,clear
log close
clear
```

26.2 Selected Output

- Survey means for BMI (LNBMI), VOEG (LNVOEG)
- Survey proportions for SGP
- Survey regression for LNBMI
- Survey regression for SGP

```
. svymean bmi voeg lnbmi lnvoeg [pw=wfin],by(region) strata(province) psu(hh) o
> bs ci
```

Survey mean estimation

```
pweight:  wfin          Number of obs(*) =      8560
Strata:   province      Number of strata =       12
PSU:      hh           Number of PSUs   =      4663
                        Population size = 6954962.2
```

Mean	Subpop.	Estimate	Std. Err.	[95% Conf. Interval]		Obs
-----+-----						
bmi	Flanders	24.40122	.1087409	24.18804	24.61441	2933
	Brussels	24.18994	.1252331	23.94443	24.43546	2499
	Wallonia	24.86484	.113913	24.64152	25.08817	2952
-----+-----						
voeg	Flanders	5.060524	.1112748	4.842372	5.278676	2917
	Brussels	6.892918	.1519949	6.594935	7.190901	2412
	Wallonia	6.807946	.1387637	6.535902	7.07999	2921
-----+-----						
lnbmi	Flanders	3.180865	.0042499	3.172533	3.189197	2933
	Brussels	3.171174	.004844	3.161677	3.18067	2499
	Wallonia	3.198131	.0044034	3.189499	3.206764	2952
-----+-----						
lnvoeg	Flanders	1.511927	.0214095	1.469954	1.5539	2917
	Brussels	1.802773	.0231351	1.757417	1.848129	2412
	Wallonia	1.803178	.0232138	1.757668	1.848689	2921

(*) Some variables contain missing values.

```
. svyset,clear
```

```
. svyprop sgp [pw=wfin],by(region) strata(province) psu(hh)
```

pweight:	wfin	Number of obs	= 8532
Strata:	province	Number of strata	= 12
PSU:	hh	Number of PSUs	= 4662
		Population size	= 6934139.7

Survey proportions estimation

```
-> region=Flanders
```

sgp	_Obs	_EstProp	_StdErr
no	142	0.045243	0.005379
yes	2834	0.954757	0.005379

```
-> region=Brussels
```

sgp	_Obs	_EstProp	_StdErr
no	497	0.217552	0.013836
yes	2060	0.782448	0.013836

```
-> region=Wallonia
```

sgp	_Obs	_EstProp	_StdErr
no	184	0.056809	0.006159
yes	2815	0.943191	0.006159

```
. svyset,clear
```

```
. svyreg lnmbmi wal fla sex agegr2 agegr3 agegr4 agegr5 agegr6 agegr7 eduprim
> edusec inclow incmed ta2 [pw=wfin], strata(province) psu(hh)
```

Survey linear regression

```
pweight:  wfin      Number of obs   =      7272
Strata:    province  Number of strata =       12
PSU:       hh        Number of PSUs   =     4135
                        Population size = 6005749.7
                        F( 14, 4110)    =      62.76
                        Prob > F       =      0.0000
                        R-squared       =      0.1812
```

lnmbmi	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wal	.0200879	.0066976	2.999	0.003	.0069571	.0332188
fla	.0018571	.0064851	0.286	0.775	-.0108572	.0145714
sex	-.0472085	.0048974	-9.639	0.000	-.0568101	-.0376069
agegr2	.0849605	.0077336	10.986	0.000	.0697984	.1001226
agegr3	.1310856	.0078827	16.630	0.000	.1156312	.1465399
agegr4	.1621346	.0084205	19.255	0.000	.1456259	.1786433
agegr5	.1936704	.0111365	17.391	0.000	.1718369	.2155039
agegr6	.1717149	.0134455	12.771	0.000	.1453544	.1980754
agegr7	.1244203	.0125904	9.882	0.000	.0997362	.1491043
eduprim	.0547676	.0081827	6.693	0.000	.0387252	.0708101
edusec	.0389084	.0069964	5.561	0.000	.0251916	.0526251
inclow	.0054668	.0094271	0.580	0.562	-.0130154	.0239489
incmed	.009757	.0086923	1.122	0.262	-.0072845	.0267986
ta2	-.0069546	.0051119	-1.360	0.174	-.0169768	.0030676
_cons	3.108181	.0157486	197.362	0.000	3.077305	3.139057

```
. svylogit sgp wal fla sex agegr2 agegr3 agegr4 agegr5 agegr6 agegr7 eduprim
> edusec inclow incmed ta2 [pw=wfin],or strata(province) psu(hh)
```

Survey logistic regression

```
pweight:  wfin      Number of obs   =      7371
Strata:   province  Number of strata =       12
PSU:      hh        Number of PSUs   =     4185
                        Population size = 6068632.8
                        F( 14, 4160)   =      20.34
                        Prob > F      =      0.0000
```

sgp	Odds Ratio	Std. Err.	t	P> t	[95% Conf. Interval]	
wal	4.416586	.6604814	9.933	0.000	3.294248	5.921301
fla	5.669468	.8964809	10.973	0.000	4.158221	7.729956
sex	1.335315	.1508814	2.559	0.011	1.069981	1.666447
agegr2	.8740069	.1736995	-0.678	0.498	.5919685	1.29042
agegr3	1.012399	.2161323	0.058	0.954	.6661619	1.538591
agegr4	1.547812	.3401891	1.988	0.047	1.005961	2.381528
agegr5	2.273991	.5529406	3.379	0.001	1.41173	3.662906
agegr6	2.062806	.6736768	2.217	0.027	1.087402	3.913152
agegr7	4.203339	2.084346	2.896	0.004	1.589935	11.11244
eduprim	1.502789	.286093	2.140	0.032	1.034675	2.182691
edusec	2.173208	.4043337	4.172	0.000	1.508989	3.1298
inclow	.9598621	.2149789	-0.183	0.855	.6187443	1.48904
incmed	1.343702	.3012734	1.318	0.188	.8657622	2.085487
ta2	.7752901	.1021982	-1.931	0.054	.5987243	1.003926

```
. svyset,clear
```

Part X

Incompleteness

Chapter 27

General Concepts

- ▷ Notation
- ▷ Taxonomies
- ▷ Example

27.1 Notation

- Subject i provides $j = 1, \dots, p$ measurements
- **Measurement** Y_{ij}
- **Missingness indicator** $R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed,} \\ 0 & \text{otherwise.} \end{cases}$
- Group Y_{ij} into a vector $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{ip})' = (\mathbf{Y}_i^o, \mathbf{Y}_i^m)$
 $\begin{cases} \mathbf{Y}_i^o & \text{contains } Y_{ij} \text{ for which } R_{ij} = 1, \\ \mathbf{Y}_i^m & \text{contains } Y_{ij} \text{ for which } R_{ij} = 0. \end{cases}$
- Group R_{ij} into a vector $\mathbf{R}_i = (R_{i1}, \dots, R_{ip})'$

27.2 Framework

$$f(\mathbf{Y}_i, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

Selection Models: $f(\mathbf{Y}_i | \boldsymbol{\theta}) \boxed{f(\mathbf{R}_i | \mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi})}$

$\boxed{\text{MCAR}}$

\longrightarrow

$\boxed{\text{MAR}}$

\longrightarrow

$\boxed{\text{MNAR}}$

$$f(\mathbf{R}_i | \boldsymbol{\psi})$$

$$f(\mathbf{R}_i | \mathbf{Y}_i^o, \boldsymbol{\psi})$$

$$f(\mathbf{R}_i | \mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi})$$

Pattern-Mixture Models: $f(\mathbf{Y}_i | \mathbf{R}_i, \boldsymbol{\theta}) f(\mathbf{R}_i | \boldsymbol{\psi})$

Shared-Parameter Models: $f(\mathbf{Y}_i | \mathbf{b}_i, \boldsymbol{\theta}) f(\mathbf{R}_i | \mathbf{b}_i, \boldsymbol{\psi})$

$$f(\mathbf{Y}_i, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

Selection Models: $f(\mathbf{Y}_i | \boldsymbol{\theta}) \boxed{f(\mathbf{R}_i | \mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi})}$

MCAR



MAR



MNAR

CC?

AC?

imputation?

⋮

direct likelihood!

expectation-maximization (EM).

multiple imputation (MI).

(weighted) GEE!

joint model!?

sensitivity analysis?!

27.3 Ignorability

- Let us decide to use likelihood based estimation.
- The full data likelihood contribution for subject i :

$$L^*(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{Y}_i, \mathbf{R}_i) \propto f(\mathbf{Y}_i, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi}).$$

- Base inference on the observed data:

$$L(\boldsymbol{\theta}, \boldsymbol{\psi} | \mathbf{Y}_i, \mathbf{R}_i) \propto f(\mathbf{Y}_i^o, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi})$$

with

$$\begin{aligned} f(\mathbf{Y}_i^o, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\mathbf{Y}_i, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) d\mathbf{Y}_i^m \\ &= \int f(\mathbf{Y}_i^o, \mathbf{Y}_i^m | \boldsymbol{\theta}) f(\mathbf{R}_i | \mathbf{Y}_i^o, \mathbf{Y}_i^m, \boldsymbol{\psi}) d\mathbf{Y}_i^m. \end{aligned}$$

- Under a MAR process:

$$\begin{aligned}f(\mathbf{Y}_i^o, \mathbf{R}_i | \boldsymbol{\theta}, \boldsymbol{\psi}) &= \int f(\mathbf{Y}_i^o, \mathbf{Y}_i^m | \boldsymbol{\theta}) f(\mathbf{R}_i | \mathbf{Y}_i^o, \boldsymbol{\psi}) d\mathbf{Y}_i^m \\ &= f(\mathbf{Y}_i^o | \boldsymbol{\theta}) f(\mathbf{R}_i | \mathbf{Y}_i^o, \boldsymbol{\psi}),\end{aligned}$$

- The likelihood factorizes into two components.

27.3.1 Ignorability: Summary

Likelihood/Bayesian + MAR

&

Frequentist + MCAR

27.4 Example: Surveytown

- Consider all three variables for Surveytown:
 - ▷ X_I : number of building lots in block I
 - ▷ Z_I : number of newspapers delivered in block I
 - ▷ Y_I : number of dwellings (buildings) in block I
- Assume blocks 7 and 8 miss their values on Y .

- Listing of Surveytown:

I	X_I	Z_I	Y_I
1	1	8	1
2	3	1	2
3	4	6	3
4	6	10	4
5	7	4	5
6	8	3	6
7	10	7	7
8	11	11	8

Chapter 28

Simplistic Methods

- ▷ Complete case analysis
- ▷ Available case analysis
- ▷ Simple imputation
- ▷ Example

28.1 CC, AC, and Simple Imputation

MCAR

Complete case analysis:

⇒ **delete** incomplete subjects

- Standard statistical software
- Loss of information
- Impact on precision and power
- Missingness \neq MCAR \Rightarrow bias
- (Case-wise deletion)

Available case analysis:

⇒ **delete** incomplete subjects per variable(s) studied

- \pm Standard statistical software
- Loss of information
- Impact on precision and power
- Missingness \neq MCAR \Rightarrow bias
- (List-wise deletion)

Simple imputation:

⇒ **impute** missing values

- Standard statistical software
- Increase of information
- Often unrealistic assumptions
- Usually bias

28.2 Example: Surveytown

- Consider four analyses:
 - ▷ Analysis of the original, complete data
 - ▷ **Complete case analysis:** only the 6 blocks with all three variables observed
 - ▷ **Available case analysis:** all 8 blocks for X and Z and the 6 remaining blocks for Y
 - ▷ **Simple mean imputation:** replace the missing values in Y with the average of the remaining ones: 3.5
- The datasets for these analyses are:

Original data					Complete case analysis				
Obs	block	x	z	y	Obs	block	x	z	y
1	1	1	8	1	1	1	1	8	1
2	2	3	1	2	2	2	3	1	2
3	3	4	6	3	3	3	4	6	3
4	4	6	10	4	4	4	6	10	4
5	5	7	4	5	5	5	7	4	5
6	6	8	3	6	6	6	8	3	6
7	7	10	7	7					
8	8	11	11	8					
Available case analysis					Mean imputation				
Obs	block	x	z	y	Obs	block	x	z	y
1	1	1	8	1	1	1	1	8	1.0
2	2	3	1	2	2	2	3	1	2.0
3	3	4	6	3	3	3	4	6	3.0
4	4	6	10	4	4	4	6	10	4.0
5	5	7	4	5	5	5	7	4	5.0
6	6	8	3	6	6	6	8	3	6.0
7	7	10	7	.	7	7	10	7	3.5
8	8	11	11	.	8	8	11	11	3.5

- In each of the four cases, the means of the three variables can simply be calculated with a program like:

```
proc means data=m.surveytown06a n mean stderr;
title 'means for surveytown - original data';
var x z y;
run;
```

- Means and standard errors, assuming this is a simple random sample from an infinite population, for illustration's sake:

Method	\bar{x}	\bar{z}	\bar{y}
Original data	6.25(1.22)	6.25(1.22)	4.50(0.87)
Complete cases	4.83(1.08)	5.33(1.36)	3.50(0.76)
Available cases	6.25(1.22)	6.25(1.22)	3.50(0.76)
Mean imputation	6.25(1.22)	6.25(1.22)	3.50(0.56)

- All simple incomplete data methods produce a downward bias in the point estimate, in this case.
- Mean imputation further artificially (hence incorrectly) reduces the standard error.
- CC further distorts the point estimates for variables, like X and Z , that are actually incomplete.
- We can do better!

Chapter 29

Direct Likelihood Maximization

- ▷ Concept
- ▷ Implications for software use
- ▷ Example

29.1 Concept

$$\boxed{\text{MAR}} : f(\mathbf{Y}_i^o | \boldsymbol{\theta}) \cancel{f(\mathbf{R}_i | \mathbf{Y}_i^o, \boldsymbol{\psi})}$$

Mechanism is MAR
 $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ distinct
Interest in $\boldsymbol{\theta}$
Use observed information matrix

\Rightarrow Likelihood inference is valid

Outcome type	Modeling strategy	Software
Gaussian	Linear mixed model	MIXED
Non-Gaussian	Generalized linear mixed model	GLIMMIX, NL MIXED

29.2 Example: Surveytown

- The key concept of direct likelihood is an analysis based on **all variables, also auxiliary ones**.
- Therefore, consider **Model 1**:

$$\begin{pmatrix} x_i \\ z_i \\ y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_x \\ \mu_z \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xz} & \sigma_{xy} \\ \sigma_{zx} & \sigma_{zz} & \sigma_{zy} \\ \sigma_{yx} & \sigma_{yz} & \sigma_{yy} \end{pmatrix} \right]$$

- Several variations to this model can be considered.

- ▷ Considering a simplified covariance structure, a diagonal one being the most extreme choice: **Model 2**:

$$\begin{pmatrix} x_i \\ z_i \\ y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_x \\ \mu_z \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & 0 & 0 \\ 0 & \sigma_{zz} & 0 \\ 0 & 0 & \sigma_{yy} \end{pmatrix} \right]$$

- ▷ Using X only as auxiliary variable: **Model 3**:

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{pmatrix} \right]$$

- ▷ Using Z only as auxiliary variable: **Model 4**:

$$\begin{pmatrix} z_i \\ y_i \end{pmatrix} \sim N \left[\begin{pmatrix} \mu_z \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_{zz} & \sigma_{zy} \\ \sigma_{yz} & \sigma_{yy} \end{pmatrix} \right]$$

- To fit the model in SAS, first the dataset needs to be transformed.

```

data m.surveytown06e;
set m.surveytown06b;
array w (3) x z y;
do j=1 to 3;
    response=w(j);
    outcome=j;
    output;
end;
run;

```

Obs	block	x	z	y	j	response	outcome
1	1	1	8	1	1	1	1
2	1	1	8	1	2	8	2
3	1	1	8	1	3	1	3
4	2	3	1	2	1	3	1
5	2	3	1	2	2	1	2
6	2	3	1	2	3	2	3
7	3	4	6	3	1	4	1
8	3	4	6	3	2	6	2
9	3	4	6	3	3	3	3
10	4	6	10	4	1	6	1
11	4	6	10	4	2	10	2
12	4	6	10	4	3	4	3
13	5	7	4	5	1	7	1
14	5	7	4	5	2	4	2
15	5	7	4	5	3	5	3
16	6	8	3	6	1	8	1
17	6	8	3	6	2	3	2
18	6	8	3	6	3	6	3
19	7	10	7	.	1	10	1
20	7	10	7	.	2	7	2
21	7	10	7	.	3	.	3
22	8	11	11	.	1	11	1
23	8	11	11	.	2	11	2
24	8	11	11	.	3	.	3

- A program for fitting Model 1 is:

```
proc mixed data=m.surveytown06e method=reml;  
title 'mixed model - x and z as auxiliary - type=un';  
class outcome;  
model response = outcome / noint solution;  
repeated outcome / subject=block type=un rcorr;  
run;
```

- ▷ The three variables are stacked onto each other, with three lines per subject.
- ▷ The '**noint**' option ensures that the three mean parameters follow directly.
- ▷ The unstructured '**type=un**' covariance structure ensure maximal freedom on the covariance model.

This is essential for the model to allow X and Z to predict Y when the latter is unobserved.

▷ The estimated correlation matrix is

Estimated R Correlation
Matrix for Subject 1

Row	Col1	Col2	Col3
1	1.0000	0.3054	0.9954
2	0.3054	1.0000	0.2893
3	0.9954	0.2893	1.0000

establishing a high correlation between X and Y , but a weak one between Z and Y , as we known very well by now.

- ▷ The estimates and standard errors for the mean:

Solution for Fixed Effects

Effect	outcome	Estimate	Standard Error	DF	t Value	Pr > t
outcome	1	6.2500	1.2211	8	5.12	0.0009
outcome	2	6.2500	1.2211	8	5.12	0.0009
outcome	3	4.4825	0.8568	8	5.23	0.0008

- ▷ Thus, the correct means follow for X and Z , which is not surprising since they are completely observed.
- ▷ The mean for Y is corrected a long way towards the true mean, thanks to the correlation with X .

- The table can be updated:

Method	\bar{x}	\bar{z}	\bar{y}
Original data	6.25(1.22)	6.25(1.22)	4.50(0.87)
Complete cases	4.83(1.08)	5.33(1.36)	3.50(0.76)
Available cases	6.25(1.22)	6.25(1.22)	3.50(0.76)
Mean imputation	6.25(1.22)	6.25(1.22)	3.50(0.56)
Model 1 (X , Z , unstr.)	6.25(1.22)	6.25(1.22)	4.48(0.86)
Model 2 (X , Z , indep.)	6.25(1.10)	6.25(1.10)	3.50(1.27)
Model 3 (X, unstr.)	6.25(1.22)	—	4.4964(0.86)
Model 4 (Z , unstr.)	—	6.25(1.22)	3.40(0.76)

- Using the highly predictive X only has the best predictive power.
- This behavior is reminiscent of benchmark estimation.

29.2.1 Why Does It Work?

- R completers $\leftrightarrow N - R$ “incompleters”

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \end{pmatrix} \sim N \left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ & \sigma_{22} \end{pmatrix} \right)$$

- Conditional density

$$Y_{i2}|y_{i1} \sim N(\beta_0 + \beta_1 y_{i1}, \sigma_{22.1})$$

μ_1	freq. & lik.	$\widehat{\mu}_1 = \frac{1}{N} \sum_{i=1}^N y_{i1}$
μ_2	frequentist	$\widetilde{\mu}_2 = \frac{1}{R} \sum_{i=1}^R y_{i2}$
μ_2	likelihood	$\widehat{\mu}_2 = \frac{1}{N} \left\{ \sum_{i=1}^R y_{i2} + \sum_{i=R+1}^N [\overline{y}_2 + \widehat{\beta}_1 (y_{i1} - \overline{y}_1)] \right\}$

Chapter 30

Multiple Imputation

- ▷ General idea
- ▷ Estimation
- ▷ Hypothesis testing
- ▷ Use of MI in practice
- ▷ Example

30.1 General Principles

- Valid under MAR
- Useful next to direct likelihood
- Three steps:
 1. The missing values are filled in M times $\implies M$ complete data sets
 2. The M complete data sets are analyzed by using standard procedures
 3. The results from the M analyses are combined into a single inference
- Rubin (1987), Rubin and Schenker (1986), Little and Rubin (1987)

30.1.1 The Algorithm

1. Draw θ^* from its posterior distribution
2. Draw \mathbf{Y}_i^{m*} from $f(\mathbf{y}_i^m | \mathbf{y}_i^o, \theta^*)$.
3. To estimate β , then calculate the estimate of the parameter of interest, and its estimated variance, using the completed data, $(\mathbf{Y}^o, \mathbf{Y}^{m*})$:

$$\hat{\beta} = \hat{\beta}(\mathbf{Y}) = \hat{\beta}(\mathbf{Y}^o, \mathbf{Y}^{m*})$$

The *within* imputation variance is

$$U = \widehat{\text{Var}}(\hat{\beta})$$

4. Repeat steps 1, 2 and 3 a number of M times

$$\Rightarrow \hat{\beta}^m \quad \& \quad U^m \quad (m = 1, \dots, M)$$

30.1.2 Pooling Information

- With M imputations, the estimate of β is

$$\hat{\beta}^* = \frac{\sum_{m=1}^M \hat{\beta}^m}{M}$$

- Further, one can make normally based inferences for β with

$$(\beta - \hat{\beta}^*) \sim N(\mathbf{0}, V)$$

where

total:
$$V = W + \left(\frac{M+1}{M} \right) B$$

within:
$$W = \frac{\sum_{m=1}^M \mathbf{U}^m}{M}$$

between:
$$B = \frac{\sum_{m=1}^M (\hat{\beta}^m - \hat{\beta}^*)(\hat{\beta}^m - \hat{\beta}^*)'}{M-1}$$

30.1.3 Hypothesis Testing

- Two “sample sizes”:
 - ▷ N : The sample size of the data set
 - ▷ M : The number of imputations
- Both play a role in the asymptotic distribution (Li, Raghunathan, and Rubin 1991)

$$H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$$

↓

$$p = P(F_{k,w} > F)$$

where

k : length of the parameter vector $\boldsymbol{\theta}$

$$F_{k,w} \sim F$$

$$F = \frac{(\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)' W^{-1} (\boldsymbol{\theta}^* - \boldsymbol{\theta}_0)}{k(1 + r)}$$

$$w = 4 + (\tau - 4) \left[1 + \frac{(1 - 2\tau^{-1})}{r} \right]^2$$

$$r = \frac{1}{k} \left(1 + \frac{1}{M} \right) \text{tr}(B W^{-1})$$

$$\tau = k(M - 1)$$

- Limiting behavior:

$$F \xrightarrow{M \rightarrow \infty} F_{k,\infty} = \chi^2/k$$

30.2 Use of MI in Practice

- Many analyses of the same incomplete set of data
- A combination of missing outcomes and missing covariates
- MI can be combined with classical GEE
- MI in SAS:

Imputation Task:

PROC MI



Analysis Task:

PROC "MYFAVORITE"



Inference Task:

PROC MIANALYZE

30.3 Example: Surveytown

- Consider multiple imputation for the incomplete version of the Surveytown data.
- The variables X and Z will be taken along as auxiliary information.
- An advantage of multiple imputation is that, once conducted, several modes of analysis can be considered.
- We will consider:
 - ▷ **SURVEYMEANS**: ordinary mean estimation, but taking the finite population of $N = 8$ into account.
 - ▷ **MIXED**: trivariate normal **Model 1**, as considered in the direct likelihood setting.

30.3.1 The Imputation Task

- The following simple code can be used, to produce multiple imputations:

```
proc mi data=m.surveytown06b seed=486378 simple out=m.surveytown07a
    nimpute=10 round=0.01;
title 'Multiple imputation in Surveytown';
var x z y;
run;
```

- ▷ The '**seed**' option ensures that, every time we run this program, we get exactly the same imputations (for diagnostic purposes).
- ▷ The number of imputations is '**nimpute=10**'.
- ▷ The imputations are generated to two decimal places, due to '**round=0.1**'.

- A portion of the multiply imputed datasets, all organized into one large set of data:

Multiply imputed Surveytown data

Obs	_Imputation_	block	x	z	y
1	1	1	1	8	1.00
2	1	2	3	1	2.00
3	1	3	4	6	3.00
4	1	4	6	10	4.00
5	1	5	7	4	5.00
6	1	6	8	3	6.00
7	1	7	10	7	6.95
8	1	8	11	11	8.20
...					
9	2	1	1	8	1.00
10	2	2	3	1	2.00
11	2	3	4	6	3.00
12	2	4	6	10	4.00
13	2	5	7	4	5.00
14	2	6	8	3	6.00
15	2	7	10	7	7.46
16	2	8	11	11	7.72
...					

Multiply imputed Surveytown data

Obs	_Imputation_	block	x	z	y
...					
23	3	7	10	7	7.09
24	3	8	11	11	8.10
...					
31	4	7	10	7	6.93
32	4	8	11	11	8.11
...					
73	10	1	1	8	1.00
74	10	2	3	1	2.00
75	10	3	4	6	3.00
76	10	4	6	10	4.00
77	10	5	7	4	5.00
78	10	6	8	3	6.00
79	10	7	10	7	7.34
80	10	8	11	11	7.95

- Due to the 'simple' option, a simple analysis, based on a multivariate model, is already produced at this stage.
- Let us present key parts of the output.
 - ▷ Some book keeping information:

Multiple imputation in Surveytown
The MI Procedure

Model Information

Data Set	M.SURVEYTOWN06B
Method	MCMC
Multiple Imputation Chain	Single Chain
Initial Estimates for MCMC	EM Posterior Mode
Start	Starting Value
Prior	Jeffreys
Number of Imputations	10
Number of Burn-in Iterations	200
Number of Iterations	100
Seed for random number generator	486378

- ▷ A relevant overview of the missing data patterns and corresponding statistics:

Missing Data Patterns						-----Group Means-----		
Group	x	z	y	Freq	Percent	x	z	y
1	X	X	X	6	75.00	4.833333	5.333333	3.500000
2	X	X	.	2	25.00	10.500000	9.000000	.

Univariate Statistics						--Missing Values--	
Variable	N	Mean	Std Dev	Minimum	Maximum	Count	Percent
x	8	6.25000	3.45378	1.00000	11.00000	0	0.00
z	8	6.25000	3.45378	1.00000	11.00000	0	0.00
y	6	3.50000	1.87083	1.00000	6.00000	2	25.00

- ▷ The correlations between the variables reveals, again, the tight relationship between Y and X on the one hand, and the loose and negative relationship between Y and Z on the other hand:

Pairwise Correlations			
	x	z	y
x	1.000000000	0.305389222	0.992314968
z	0.305389222	1.000000000	-0.192814109
y	0.992314968	-0.192814109	1.000000000

- ▷ Note that the correlations are different from what was obtained with Model 1 in the direct likelihood method, since here the correlations are based on the completers only.
- ▷ Parameter estimates and the covariance matrix of the outcomes, now properly accounting for missingness, are also obtained:

EM (Posterior Mode) Estimates				
TYPE	_NAME_	x	z	y
MEAN		6.250000	6.250000	4.482408
COV	x	6.958333	2.125000	4.852466
COV	z	2.125000	6.958333	1.410318
COV	y	4.852466	1.410318	3.410916

- ▷ Between and within variability information is displayed:

Multiple Imputation Variance Information				
Variable	-----Variance-----			DF
	Between	Within	Total	
y	0.004691	0.751038	0.756198	5.5616

Multiple Imputation Variance Information			
Variable	Relative	Fraction	Relative
	Increase	Missing	
	in Variance	Information	Efficiency
y	0.006871	0.006834	0.999317

- ▷ It is clear that the between-variance is small relative to the within-variance.

- ▷ Parameter estimates and standard errors for variables with incomplete information is given:

Multiple Imputation Parameter Estimates						
Variable	Mean	Std Error	95% Confidence Limits		DF	Minimum
y	4.500875	0.869597	2.331725	6.670025	5.5616	4.366250

t for H0:				
Variable	Maximum	Mu0	Mean=Mu0	Pr > t
y	4.606250	0	5.18	0.0026

- ▷ Note also here the closeness of the mean estimator for Y to the true value, in spite of missingness.

30.3.2 The Model Task With PROC SURVEYMEANS

- To estimate the means for each of the 10 imputations, use the following program:

```
proc surveymeans data=m.surveytown07a total=8.00000001;  
title 'SURVEYMEANS analysis after multiple imputation';  
title2 'with finite population correction';  
by _imputation_;  
var x z y;  
ods output Statistics = m.surveytown07b;  
run;
```

The syntax is virtually the same than our earlier uses of the **SURVEYMEANS** procedure, except:

- ▷ The 'BY' statement with the variable `_imputation_`, created by **PROC MI**, is mandatory to ensure separate analyses are done for each of the (10) imputations.

- ▷ The 'ODS' (output delivery system) statement is necessary to store the 10 parameter estimates and 10 standard errors, so that they can be passed on to **PROC MIANALYZE**.
- ▷ 'Statistics' is a reserved word for a specific table: the main table outputted by the procedure.
- ▷ The small increment in the '**total=**' option avoids boundary problems in **MIANALYZE**.

This only applies when $N = n$, i.e., with a census.

- PROC SURVEYMEANS produces the following output:

The SURVEYMEANS Procedure

SURVEYMEANS analysis after multiple imputation
with finite population correction

Imputation Number=1

Statistics					
Variable	N	Mean	Std Error of Mean	95% CL for Mean	
x	8	6.250000	0.000043172	6.24989791	6.25010209
z	8	6.250000	0.000043172	6.24989791	6.25010209
y	8	4.518750	0.000031049	4.51867658	4.51882342

Imputation Number=2

Statistics					
Variable	N	Mean	Std Error of Mean	95% CL for Mean	
x	8	6.250000	0.000043172	6.24989791	6.25010209
z	8	6.250000	0.000043172	6.24989791	6.25010209
y	8	4.522500	0.000030846	4.52242706	4.52257294

...

Imputation Number=10

Statistics					
Variable	N	Mean	Std Error of Mean	95% CL for Mean	
x	8	6.250000	0.000043172	6.24989791	6.25010209
z	8	6.250000	0.000043172	6.24989791	6.25010209
y	8	4.536250	0.000031145	4.53617635	4.53632365

- ▷ The means for X and Z do not change, since there is no missingness in these variables.
- ▷ The means for Y change, due to the two missing observations, which are 10 times randomly filled.
- ▷ The standard errors are all within-imputation standard errors, so each one of them underestimates the true variability, until the analysis task (**PROC MIANALYZE**) is performed.
- ▷ The standard errors would be exactly equal to zero if '**total=8**' were used.
- The **ODS** statement, placing the results from the 'Statistics' tables above into a dataset, produces:

Estimates and standard errors from SURVEYMEANS

Obs	_Imputation_	Var Name	N	Mean	StdErr	Lower CLMean	Upper CLMean
1	1	x	8	6.250000	0.000043172	6.24989791	6.25010209
2	1	z	8	6.250000	0.000043172	6.24989791	6.25010209
3	1	y	8	4.518750	0.000031049	4.51867658	4.51882342
4	2	x	8	6.250000	0.000043172	6.24989791	6.25010209
5	2	z	8	6.250000	0.000043172	6.24989791	6.25010209
6	2	y	8	4.522500	0.000030846	4.52242706	4.52257294
7	3	x	8	6.250000	0.000043172	6.24989791	6.25010209
8	3	z	8	6.250000	0.000043172	6.24989791	6.25010209
9	3	y	8	4.523750	0.000031040	4.52367660	4.52382340
...							
19	7	x	8	6.250000	0.000043172	6.24989791	6.25010209
20	7	z	8	6.250000	0.000043172	6.24989791	6.25010209
21	7	y	8	4.547500	0.000031608	4.54742526	4.54757474
22	8	x	8	6.250000	0.000043172	6.24989791	6.25010209
23	8	z	8	6.250000	0.000043172	6.24989791	6.25010209
24	8	y	8	4.366250	0.000028303	4.36618307	4.36631693
25	9	x	8	6.250000	0.000043172	6.24989791	6.25010209
26	9	z	8	6.250000	0.000043172	6.24989791	6.25010209
27	9	y	8	4.465000	0.000029976	4.46492912	4.46507088
28	10	x	8	6.250000	0.000043172	6.24989791	6.25010209
29	10	z	8	6.250000	0.000043172	6.24989791	6.25010209
30	10	y	8	4.536250	0.000031145	4.53617635	4.53632365

- **PROC MIANALYZE** can work with a variety of input forms, but the above dataset is not suitable without re-organization, even though it contains all information.
- One way to organize the the required input for **PROC MIANALYZE** is:
 - ▷ One column per point estimate (there are three in our case).
 - ▷ One column per standard error (there are three in our case).

- Code for this reorganization:

```
data m.helpx;  
set m.surveytown07b;  
meanx=mean;  
stdex=stderr;  
if varname='x' then output;  
run;
```

```
data m.helpz;  
set m.surveytown07b;  
meanz=mean;  
stdez=stderr;  
if varname='z' then output;  
run;
```

```
data m.helpy;  
set m.surveytown07b;  
meany=mean;  
stdey=stderr;  
if varname='y' then output;  
run;
```

```
data m.surveytown07c;  
merge m.helpx m.helpz m.helpy;  
by _imputation_;  
drop varname stderr mean  
      lowerclmean upperclmean;  
run;
```

- The re-organized information looks as follows:

Reorganized estimates and standard errors from SURVEYMEANS

Obs	_Imputation_	N	meanx	stdex	meanz	stdez	meany	stdey
1	1	8	6.25	.000043172	6.25	.000043172	4.51875	.000031049
2	2	8	6.25	.000043172	6.25	.000043172	4.52250	.000030846
3	3	8	6.25	.000043172	6.25	.000043172	4.52375	.000031040
4	4	8	6.25	.000043172	6.25	.000043172	4.50500	.000030777
5	5	8	6.25	.000043172	6.25	.000043172	4.41750	.000029094
6	6	8	6.25	.000043172	6.25	.000043172	4.60625	.000032350
7	7	8	6.25	.000043172	6.25	.000043172	4.54750	.000031608
8	8	8	6.25	.000043172	6.25	.000043172	4.36625	.000028303
9	9	8	6.25	.000043172	6.25	.000043172	4.46500	.000029976
10	10	8	6.25	.000043172	6.25	.000043172	4.53625	.000031145

- We are now in a position to start the analysis task.

30.3.3 The Analysis Task After PROC SURVEYMEANS

- A program for the analysis task takes the following form:

```
proc mianalyze data=m.surveytown07c;  
title 'MIANALYZE of SURVEYMEANS results';  
modeleffects meanx meanz many;  
stderr stdex stdez stdey;  
run;
```

- Key statements and options are:
 - ▷ '**data=**' specifies the input dataset.
 - ▷ In our case, it contains parameter estimates and standard errors for all three means of X , Z , and Y .
 - ▷ The dataset is not of any special form, as such recognized by the procedure.
 - ▷ This implies we must specify:
 - * The estimates through the **MODELEFFECTS** statement.
 - * The standard errors through the **STDERR** statement.

- The output takes the following form:

MIANALYZE of SURVEYMEANS results
The MIANALYZE Procedure

Multiple Imputation Variance Information

Parameter	-----Variance-----			DF
	Between	Within	Total	
meanx	0	1.8638394E-9	1.8638394E-9	.
meanz	0	1.8638394E-9	1.8638394E-9	.
many	0.004691	9.387976E-10	0.005160	9

Multiple Imputation Variance Information

Parameter	Relative	Fraction	Relative
	Increase	Missing	
	in Variance	Information	Efficiency
meanx	0	.	.
meanz	0	.	.
many	5496487	1.000000	0.909091

Multiple Imputation Parameter Estimates

Parameter	Estimate	Std Error	95% Confidence Limits		DF
meanx	6.250000	1.8638394E-9	.	.	.
meanz	6.250000	1.8638394E-9	.	.	.
meany	4.500875	0.071834	4.338376	4.663374	9

Multiple Imputation Parameter Estimates

Parameter	Minimum	Maximum	Theta0	t for H0:	
				Parameter=Theta0	Pr > t
meanx	6.250000	6.250000	0	.	.
meanz	6.250000	6.250000	0	.	.
meany	4.366250	4.606250	0	62.66	<.0001

- The output is rather straightforward.
- The increase in variance is extreme in this case:
 - ▷ If there would have been no missingness, there would have been **zero variance** since $N = n = 8$.

- ▷ Due to missingness, there is some variability (uncertainty) introduced.
 - ▷ This produces an infinite variance increase here.
 - ▷ However, since we set '**total=8.00000001**', the excess is still finite.
 - ▷ Not all information is provided for X and Z since here the reverse happens: there is no missingness so the variance increase is zero.
- Note, once again, the correcting power of primarily X on the mean estimation for Y : even though the raw mean in the available data is 3.5, multiple imputation, like direct likelihood, corrects strongly towards the true mean of 4.5.

30.3.4 The Model Task With PROC MIXED

- One of the appealing features of multiple imputation is that several analyses can be done, based on a single multiple-imputation exercise.
- For example, we can complement the above **SURVEYMEANS** analysis with **MIXED** Model 1.
- Exactly like in the direct-likelihood case, the data need to be organized differently to enable use of **PROC MIXED**:

```
data m.surveytown07e;  
set m.surveytown07a;  
array w (3) x z y;  
do j=1 to 3;  
    response=w(j);  
    outcome=j;  
    output;  
end;  
run;
```

- The re-organized data look like:

Multiply imputed data reorganized to allow for MIXED analysis

Obs	_Imputation_	block	x	z	y	j	response	outcome
1	1	1	1	8	1.00	1	1.00	1
2	1	1	1	8	1.00	2	8.00	2
3	1	1	1	8	1.00	3	1.00	3
4	1	2	3	1	2.00	1	3.00	1
5	1	2	3	1	2.00	2	1.00	2
6	1	2	3	1	2.00	3	2.00	3
...								
22	1	8	11	11	8.20	1	11.00	1
23	1	8	11	11	8.20	2	11.00	2
24	1	8	11	11	8.20	3	8.20	3
...								
217	10	1	1	8	1.00	1	1.00	1
218	10	1	1	8	1.00	2	8.00	2
219	10	1	1	8	1.00	3	1.00	3
220	10	2	3	1	2.00	1	3.00	1
221	10	2	3	1	2.00	2	1.00	2
222	10	2	3	1	2.00	3	2.00	3
223	10	3	4	6	3.00	1	4.00	1
...								
238	10	8	11	11	7.95	1	11.00	1
239	10	8	11	11	7.95	2	11.00	2
240	10	8	11	11	7.95	3	7.95	3

- We are now in a position to apply **PROC MIXED**:

```
proc mixed data=m.surveytown07e method=reml;  
title 'MIXED analysis after multiple imputation';  
title2 'x and z as auxiliary - type=un';  
by _imputation_;  
class outcome;  
model response = outcome / noint solution covb;  
repeated outcome / subject=block type=un rcorr;  
ods output solutionF = m.surveytown07f covb = m.surveytown07g;  
run;
```

- The program is the same as before, with a few additions:
 - ▷ The 'BY' statement with the variable `_imputation_`, created by **PROC MI**, is mandatory to ensure separate analyses are done for each of the (10) imputations.

- ▷ The 'ODS' (output delivery system) statement is necessary to store information that needs to be passed to PROC MIANALYZE:
 - * 'solutionF': the 10 sets of parameter estimates
 - * 'covb': the 10 variance-covariance matrices of the parameter estimates
- ▷ For these to take effect, two options in the MODEL statement are necessary:
 - * For 'solutionF': the 'solution' option
 - * For 'covb': the 'covb' option
- Exactly like in the SURVEYMEANS case, there are 10 distinct analyses, each with their output.
- Since we have seen such output before, we present a small fraction:

Effect	outcome	Estimate	Standard Error	DF	t Value	Pr > t
Imputation Number=1						
outcome	1	6.2500	1.2211	8	5.12	0.0009
outcome	2	6.2500	1.2211	8	5.12	0.0009
outcome	3	4.5187	0.8782	8	5.15	0.0009
Imputation Number=2						
outcome	1	6.2500	1.2211	8	5.12	0.0009
outcome	2	6.2500	1.2211	8	5.12	0.0009
outcome	3	4.5225	0.8725	8	5.18	0.0008
...						
Imputation Number=10						
outcome	1	6.2500	1.2211	8	5.12	0.0009
outcome	2	6.2500	1.2211	8	5.12	0.0009
outcome	3	4.5362	0.8809	8	5.15	0.0009

- The dataset with the parameter estimates:

Parameter estimates from the MIXED model

Obs	_Imputation_	Effect	outcome	Estimate	StdErr	DF	tValue	Probt
1	1	outcome	1	6.2500	1.2211	8	5.12	0.0009
2	1	outcome	2	6.2500	1.2211	8	5.12	0.0009
3	1	outcome	3	4.5187	0.8782	8	5.15	0.0009
4	2	outcome	1	6.2500	1.2211	8	5.12	0.0009
5	2	outcome	2	6.2500	1.2211	8	5.12	0.0009
6	2	outcome	3	4.5225	0.8725	8	5.18	0.0008
...								
28	10	outcome	1	6.2500	1.2211	8	5.12	0.0009
29	10	outcome	2	6.2500	1.2211	8	5.12	0.0009
30	10	outcome	3	4.5362	0.8809	8	5.15	0.0009

- The dataset with the variance-covariance parameters:

Covariance matrices of estimates from the MIXED model

Obs	_Imputation_	Row	Effect	outcome	Col1	Col2	Col3
1	1	1	outcome	1	1.4911	0.4554	1.0672
2	1	2	outcome	2	0.4554	1.4911	0.3377
3	1	3	outcome	3	1.0672	0.3377	0.7712
4	2	1	outcome	1	1.4911	0.4554	1.0606
5	2	2	outcome	2	0.4554	1.4911	0.3038
6	2	3	outcome	3	1.0606	0.3038	0.7612
...							
28	10	1	outcome	1	1.4911	0.4554	1.0721
29	10	2	outcome	2	0.4554	1.4911	0.3217
30	10	3	outcome	3	1.0721	0.3217	0.7760

- We are now in a position to complete the analysis task.

30.3.5 The Analysis Task After PROC MIXED

- **PROC MIANALYZE** can be invoked to process the **PROC MIXED** output:

```
proc mianalyze parms=m.surveytown07f covb=m.surveytown07g;  
title 'MIANALYZE of MIXED results';  
class outcome;  
modeleffects outcome;  
run;
```

- Note that the information is now passed on using **two** options:
 - ▷ **'parms'**: the parameter estimates
 - ▷ **'covb'**: the variance-covariance matrix of the parameter estimates
- Since the information is passed on in a structured way, only the **MODELEFFECTS** is needed.

- Specifying 'outcome' as the **MODELEFFECTS** variable, implies the column labeled 'outcome' is defining; **not the column labeled 'Effect'** which is not used at all.
- Defining 'outcome' as a **CLASS** variable states that every one of the three levels corresponds to a different parameter (X , Z , and Y , respectively).
- The results take a form, equal in layout as the previous use:

MIANALYZE of MIXED results

Multiple Imputation Variance Information

Parameter	outcome	-----Variance-----			DF
		Between	Within	Total	
outcome	1.000000	0	1.491071	1.491071	.
outcome	2.000000	0	1.491071	1.491071	.
outcome	3.000000	0.004691	0.751038	0.756198	193286

Multiple Imputation Variance Information

Parameter	outcome	Relative Increase in Variance	Fraction Missing Information	Relative Efficiency
outcome	1.000000	0	.	.
outcome	2.000000	0	.	.
outcome	3.000000	0.006871	0.006834	0.999317

Multiple Imputation Parameter Estimates

Parameter	outcome	Estimate	Std Error	95% Confidence Limits		DF
outcome	1.000000	6.250000	1.491071	.	.	.
outcome	2.000000	6.250000	1.491071	.	.	.
outcome	3.000000	4.500875	0.869597	2.796486	6.205264	193286

- Now, since we are in a 'large population' context, there is both within- and between-imputation variability.

- The fraction of missing information is so small, since the X values compensate for the missing information on Y .
- If Z only, or no auxiliary variables at all would be used, the fraction would go up, and bias would appear:
 - ▷ Given X , missingness in Y is MAR or even MCAR.
 - ▷ Without X , the mechanism is MNAR.

30.3.6 Summary of the Results

Method	\bar{x}	\bar{z}	\bar{y}
Original data	6.25(1.22)	6.25(1.22)	4.50(0.87)
Simplistic Methods			
Complete cases	4.83(1.08)	5.33(1.36)	3.50(0.76)
Available cases	6.25(1.22)	6.25(1.22)	3.50(0.76)
Mean imputation	6.25(1.22)	6.25(1.22)	3.50(0.56)
Direct Likelihood			
Model 1 (X , Z , unstr.)	6.25(1.22)	6.25(1.22)	4.48(0.86)
Model 2 (X , Z , indep.)	6.25(1.10)	6.25(1.10)	3.50(1.27)
Model 3 (X , unstr.)	6.25(1.22)	—	4.4964(0.86)
Model 4 (Z , unstr.)	—	6.25(1.22)	3.40(0.76)
Multiple Imputation			
MI (posterior mode)	6.25(—)	6.25(—)	4.482408(—)
MI (model based)	—	—	4.500875(0.87)
SURVEYMEANS	6.25(0.00)	6.25(0.00)	4.500875(0.071834)
Model 1 (X , Z , unstr.)	6.25(1.49)	6.25(1.49)	4.500875(0.87)

- All direct likelihood and MI methods provide acceptable results.
- It is important to use X as an auxiliary variable.
- The **posterior mode** analysis is a byproduct of generating the imputations by means of Monte-Carlo Markov Chain (MCMC) estimation.
- The **model based** analysis in MI considers an unstructured mean vector and an unstructured covariance matrix.

These are also the ingredients of Model 1, hence the similarity.

- The MI standard errors are a bit larger, owing to the uncertainty stemming from drawing random imputations.

It typically diminishes when the number of imputations increases.

Chapter 31

Non-Gaussian Data

- ▷ Non-Gaussian data
- ▷ Likelihood-based methods
- ▷ Weighted generalized estimating equations
- ▷ Multiple imputation combined with generalized estimating equations

31.1 Non-Gaussian Data

- We have considered two main families of methods:
 - ▷ **Likelihood-based methods:** generalized linear mixed models
 - ▷ **Non-likelihood methods:** GEE
- They differ in nature:
 - ▷ GLMM: random-effects (hierarchical, multi-level)
 - ▷ GEE: marginal
- This implies that one may have to choose a family based on scientific reasons.
- Thus, it is necessary what to do when data are incomplete.

31.2 Likelihood-Based Methods

- The GLMM is typically fitted using maximum likelihood or approximations thereof.
- Thus: the GLMM produces ignorability under **MAR**.
- In other words: the GLMM is valid under **MAR**.
- Practically:
 - ▷ **PROC NL MIXED**: a bit involved, but accurate.
 - ▷ **PROC GLIMMIX**: the approximation is poor, and even worse with incomplete data.
- Our analyses, conducted with the GLMM, are widely valid.

31.3 Generalized Estimating Equations

- When a marginal model is needed, GEE is a recommendable method.
- But: it is not likelihood based.
- GEE is valid only:
 - ▷ When the mechanism is MCAR.
 - ▷ When the mechanism is MAR and the working correlation matrix is correctly specified.
 - ▷ When the mechanism is MAR and weighted GEE (W-GEE) are used.
 - ▷ When the mechanism is MAR and multiple imputation is used in conjunction with GEE.

31.3.1 Weighted Generalized Estimating Equations

- The principle is: to weigh a unit (respondent) by the inverse of its probability to drop out.
- It is very natural to use with longitudinal data (panel studies).
- Less easy to use with multivariate (survey) data, full of intermittent missingness.
- Very related to inverse probability weighting such as in the Horvitz-Thompson estimator.
- But: a model needs to be specified for the weights, unlike purely design-based uses of the weighting method.
- Example code: `www.uhasselt.be/censtat`

31.3.2 Multiple Imputation Combined with Generalized Estimating Equations

- The concept of GEE can be combined with multiple imputation.
- In the imputation task, a full model needs to be specified.
- This can be done very flexibly:
 - ▷ A general loglinear model.
 - ▷ A general transition model.
 - ▷ ...
- The method is then valid under MAR, and proceeds exactly like in the examples given in the continuous case.

Chapter 32

Incompleteness in the Belgian Health Interview Survey

- ▷ Taxonomy
- ▷ Household-level non-response
- ▷ Individual-level non-response
- ▷ Item-level missingness

32.1 Incomplete Data

- **Household level**

- ▷ Households with which no interview was realized
- ▷ Households which explicitly refused
- ▷ Households which could not be contacted

- **Individual level**

- ▷ Individual refuses to participate, in spite of HH agreement

- **Item level**

- ▷ A participating respondent leaves some questions unanswered

32.2 Design Measures Towards Missing Data

- Increased number of sampled households (HHs)
- Replacement scheme for drop-outs
 - ▷ HHs sampled in clusters of 4
 - ▷ Oversampling of clusters
- Proxy interviews
- Invitation letter
- Multiple attempts to contact a HH
- Coding of the reasons for drop-outs

32.3 Missing Data: HH-Level

- 35,023 HHs sampled
- 11,568 HHs attempted to contact
- Different reasons for a HH non-interview:

Type	Description	#	%
NP: Non-Participation	no interview regardless reason	6904	59.7%
NA: Non-Availability	no interview due to difficulty in contacting	3546	30.7%
NR: Non-Response	no interview due to explicit HH refusal	3358	29.0

32.4 Individual-Level Missingness

- 10,339 HH members selected for interview.
- Similar reasons for missingness at this level:

Type	Description	#	%	Proxy
NP: Non-Participation	no personal interview	785	7.6%	671
NA: Non-Availability	difficulty in contacting	408	3.9%	408
NR: Non-Response	explicit refusal	210	2.0%	96

32.5 Item-Level Missingness

- Only non-response
- More than 1000 variables obtained for the interviewed individuals.
- Frequency of NR depending on the item (question):
 - ▷ BMI: 2.1%
 - ▷ VOEG: 3.7%
 - ▷ Maximum observed: 11%
- May be substantial when several variables are considered jointly.

32.5.1 Factors Influencing Item-Level Missingness

- Different across regions.
- Missingness increases with HH size.
- Effect of the age of the reference person.
- Effect of nationality of reference person.
- Effect of gender of reference person.

32.5.2 Multiple Imputation for LNBMI

Effect	Level	AC (7272 obs.)	MI (8564 obs.)
Region	Brussels	—	—
	Flanders	0.007 (0.006)	0.009 (0.006)
	Wallonia	0.023 (0.007)	0.027 (0.006)
Gender	Male	—	—
	Female	-0.050 (0.004)	-0.054 (0.003)
Education	Primary	—	—
	Secondary	-0.011 (0.005)	-0.013 (0.004)
	Higher	-0.046 (0.005)	-0.045 (0.005)
Income level	< 40,000	—	—
	40,000–60,000	0.008 (0.004)	0.006 (0.004)
	> 60,000	0.003 (0.006)	-0.001 (0.006)
Smoking	Non-smoker	—	—
	Smoker	0.003 (0.004)	0.004 (0.004)
Age	Age-group	0.030 (0.001)	0.001 (0.001)

32.5.3 Multiple Imputation for LNVOEG

Effect	Level	AC (7389 obs.)	MI (8564 obs.)
Region	Brussels	—	—
	Flanders	-0.264 (0.032)	-0.268 (0.031)
	Wallonia	0.015 (0.033)	0.002 (0.033)
Gender	Male	—	—
	Female	0.296 (0.019)	0.284 (0.018)
Education	Primary	—	—
	Secondary	-0.072 (0.023)	-0.069 (0.023)
	Higher	-0.099 (0.025)	-0.088 (0.025)
Income level	< 40, 000	—	—
	40,000–60,000	-0.049 (0.021)	-0.039 (0.021)
	> 60, 000	-0.107 (0.030)	-0.094 (0.034)
Smoking	Non-smoker	—	—
	Smoker	0.238 (0.019)	0.220 (0.019)
Age	Age-group	0.051 (0.006)	0.050 (0.005)

- While the AC analyses are based on a different number of cases for different variable, multiple imputation allows for a common base of inference.
- Differences are not extremely large, but they are noticeable.

Chapter 33

Sensitivity Analysis: A Case Study

- ▷ The Slovenian Public Opinion Survey
- ▷ MAR and MNAR analyses
- ▷ Informal sensitivity analysis
- ▷ Interval of ignorance & interval of uncertainty

33.1 The Slovenian Plebiscite

- Rubin, Stern, and Vehovar (1995)
- Slovenian Public Opinion (SPO) Survey
- Four weeks prior to decisive plebiscite
- Three questions:
 1. Are you in favor of Slovenian independence ?
 2. Are you in favor of Slovenia's secession from Yugoslavia ?
 3. Will you attend the plebiscite ?
- Political decision: **ABSENCE** \equiv **NO**
- Primary Estimand: θ : **Proportion in favor of independence**

- Slovenian Public Opinion Survey Data:

		Independence		
Secession	Attendance	Yes	No	*
Yes	Yes	1191	8	21
	No	8	0	4
	*	107	3	9
No	Yes	158	68	29
	No	7	14	3
	*	18	43	31
*	Yes	90	2	109
	No	1	2	25
	*	19	8	96

33.2 Slovenian Public Opinion: 1st Analysis

- **Pessimistic:** All who *can* say NO *will* say NO

$$\hat{\theta} = \frac{1439}{2074} = 0.694$$

- **Optimistic:** All who *can* say YES *will* say YES

$$\hat{\theta} = \frac{1439 + 159 + 144 + 136}{2074} = \frac{1878}{2076} = 0.904$$

- **Resulting Interval:**

$$\theta \in [0.694; 0.904]$$

- **Resulting Interval:**

$$\theta \in [0.694; 0.904]$$

- **Complete cases:** All who answered on 3 questions

$$\hat{\theta} = \frac{1191 + 158}{1454} = 0.928 \text{ ?}$$

- **Available cases:** All who answered on both questions

$$\hat{\theta} = \frac{1191 + 158 + 90}{1549} = 0.929 \text{ ?}$$

33.3 Slovenian Public Opinion: 2nd Analysis

- **Missing at Random:**

Non-response is allowed to depend on observed, but not on unobserved outcomes:

▷ Based on two questions:

$$\hat{\theta} = 0.892$$

▷ Based on three questions:

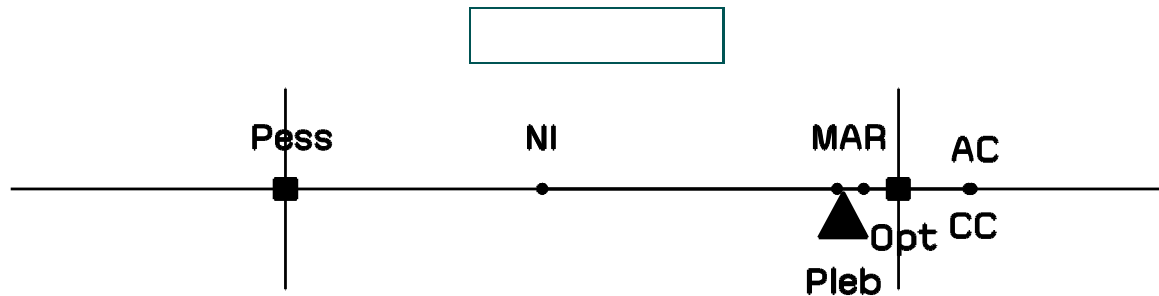
$$\hat{\theta} = 0.883$$

- **Missing Not at Random (NI):**

Non-response is allowed to depend on unobserved measurements:

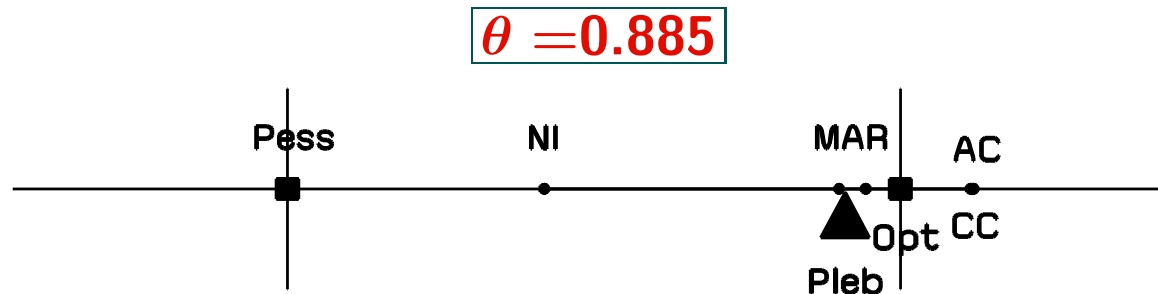
$$\hat{\theta} = 0.782$$

33.4 Slovenian Public Opinion Survey



Estimator	$\hat{\theta}$
Pessimistic bound	0.694
Optimistic bound	0.904
Complete cases	0.928 ?
Available cases	0.929 ?
MAR (2 questions)	0.892
MAR (3 questions)	0.883
MNAR	0.782

33.5 Slovenian Plebiscite: The Truth ?



Estimator	$\hat{\theta}$
Pessimistic bound	0.694
Optimistic bound	0.904
Complete cases	0.928 ?
Available cases	0.929 ?
MAR (2 questions)	0.892
MAR (3 questions)	0.883
MNAR	0.782

33.6 Did “the” MNAR model behave badly ?

Consider a family of MNAR models

- Baker, Rosenberger, and DerSimonian (1992)
- Counts $Y_{r_1 r_2 j k}$
- $j, k = 1, 2$ indicates YES/NO
- $r_1, r_2 = 0, 1$ indicates MISSING/OBSERVED

33.6.1 Model Formulation

$$E(Y_{11jk}) = m_{jk},$$

$$E(Y_{10jk}) = m_{jk}\beta_{jk},$$

$$E(Y_{01jk}) = m_{jk}\alpha_{jk},$$

$$E(Y_{00jk}) = m_{jk}\alpha_{jk}\beta_{jk}\gamma_{jk},$$

Interpretation:

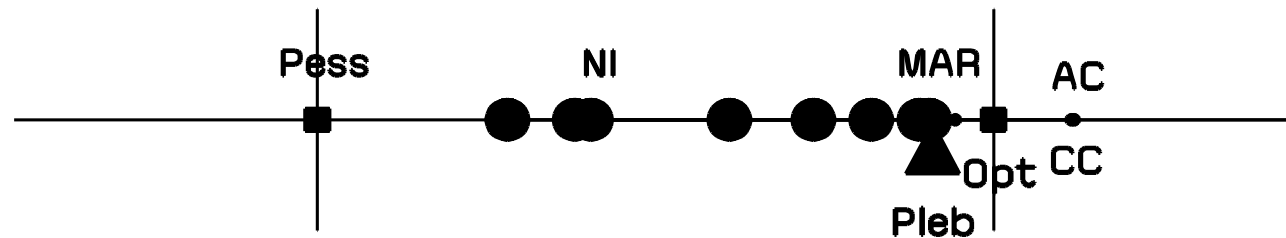
- α_{jk} : models non-response on independence question
- β_{jk} : models non-response on attendance question
- γ_{jk} : interaction between both non-response indicators (cannot depend on j or k)

33.6.2 Identifiable Models

Model	Structure	d.f.	loglik	θ	C.I.
BRD1	(α, β)	6	-2495.29	0.892	[0.878;0.906]
BRD2	(α, β_j)	7	-2467.43	0.884	[0.869;0.900]
BRD3	(α_k, β)	7	-2463.10	0.881	[0.866;0.897]
BRD4	(α, β_k)	7	-2467.43	0.765	[0.674;0.856]
BRD5	(α_j, β)	7	-2463.10	0.844	[0.806;0.882]
BRD6	(α_j, β_j)	8	-2431.06	0.819	[0.788;0.849]
BRD7	(α_k, β_k)	8	-2431.06	0.764	[0.697;0.832]
BRD8	(α_j, β_k)	8	-2431.06	0.741	[0.657;0.826]
BRD9	(α_k, β_j)	8	-2431.06	0.867	[0.851;0.884]

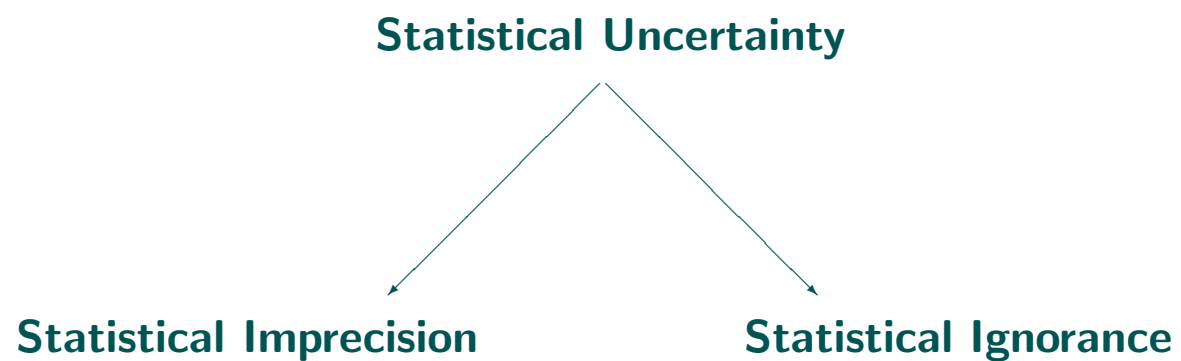
33.6.3 An “Interval” of MNAR Estimates

$$\theta = 0.885$$



Estimator	$\hat{\theta}$
[Pessimistic; optimistic]	[0.694;0.904]
Complete cases	0.928
Available cases	0.929
MAR (2 questions)	0.892
MAR (3 questions)	0.883
MNAR	0.782
MNAR “interval”	[0.741;0.892]

33.7 A More Formal Look



Statistical Imprecision: *Due to finite sampling*

- Fundamental concept of mathematical statistics
- Consistency, efficiency, precision, testing, . . .
- Disappears as sample size increases

Statistical Ignorance: *Due to incomplete observations*

- Received less attention
- Can invalidate conclusions
- Does not disappear with increasing sample size

Kenward, Goetghebeur, and Molenberghs (StatMod 2001)

33.7.1 Monotone Patterns

$$R = 1$$

$Y_{1,11}$	$Y_{1,12}$
$Y_{1,21}$	$Y_{1,22}$



$$R = 0$$

$Y_{0,1}$
$Y_{0,2}$



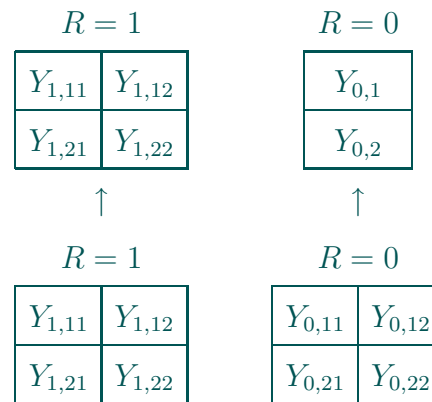
$$R = 1$$

$Y_{1,11}$	$Y_{1,12}$
$Y_{1,21}$	$Y_{1,22}$

$$R = 0$$

$Y_{0,11}$	$Y_{0,12}$
$Y_{0,21}$	$Y_{0,22}$

33.7.2 Models for Monotone Patterns



$$\mu_{r,ij} = p_{ij}q_{r|ij}, \quad (i,j=1,2;r=0,1)$$

Model	$q_{r ij}$	# Par.	Observed d.f.	Complete d.f.
1. MCAR	q_r	4	Non-saturated	Non-saturated
2. MAR	$q_{r i}$	5	Saturated	Non-saturated
3. MNAR(0)	$q_{r j}$	5	Saturated	Non-saturated
4. MNAR(1)	$\text{logit}(q_{r ij}) = \alpha + \beta_i + \gamma_j$	6	Overspecified	Non-saturated
5. MNAR(2)	$q_{r ij}$	7	Overspecified	Saturated

33.7.3 Sensitivity Parameter Method

Sensitivity Parameter: A minimal set η

Estimable Parameter: μ , estimable, given η

Procedure:

- ▷ Given η , calculate parameter and C.I. for μ
- ▷ Set of parameter estimates: **region of ignorance**
- ▷ Set of interval estimates: **region of uncertainty**
- ▷ Single parameter case: 'region' becomes 'interval'

33.8 Slovenian Public Opinion: 3rd Analysis

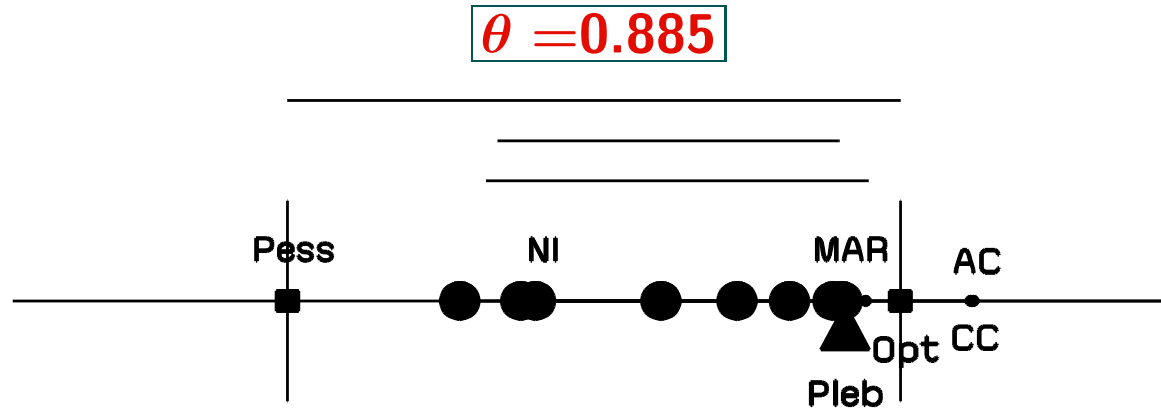
Model	Structure	d.f.	loglik	θ	C.I.
BRD1	(α, β)	6	-2495.29	0.892	[0.878;0.906]
BRD2	(α, β_j)	7	-2467.43	0.884	[0.869;0.900]
BRD3	(α_k, β)	7	-2463.10	0.881	[0.866;0.897]
BRD4	(α, β_k)	7	-2467.43	0.765	[0.674;0.856]
BRD5	(α_j, β)	7	-2463.10	0.844	[0.806;0.882]
BRD6	(α_j, β_j)	8	-2431.06	0.819	[0.788;0.849]
BRD7	(α_k, β_k)	8	-2431.06	0.764	[0.697;0.832]
BRD8	(α_j, β_k)	8	-2431.06	0.741	[0.657;0.826]
BRD9	(α_k, β_j)	8	-2431.06	0.867	[0.851;0.884]
Model 10	(α_k, β_{jk})	9	-2431.06	[0.762;0.893]	[0.744;0.907]
Model 11	(α_{jk}, β_j)	9	-2431.06	[0.766;0.883]	[0.715;0.920]
Model 12	$(\alpha_{jk}, \beta_{jk})$	10	-2431.06	[0.694;0.904]	

33.9 Every MNAR Model Has Got a MAR Bodyguard

- Fit an MNAR model to a set of incomplete data.
- Change the conditional distribution of the unobserved outcomes, given the observed ones, to comply with MAR.
- The resulting new model will have exactly the same fit as the original MNAR model.
- The missing data mechanism has changed.
- This implies that definitively testing for MAR *versus* MNAR is not possible.

33.10 Slovenian Public Opinion: 4rd Analysis

Model	Structure	d.f.	loglik	$\hat{\theta}$	C.I.	$\hat{\theta}_{\text{MAR}}$
BRD1	(α, β)	6	-2495.29	0.892	[0.878;0.906]	0.8920
BRD2	(α, β_j)	7	-2467.43	0.884	[0.869;0.900]	0.8915
BRD3	(α_k, β)	7	-2463.10	0.881	[0.866;0.897]	0.8915
BRD4	(α, β_k)	7	-2467.43	0.765	[0.674;0.856]	0.8915
BRD5	(α_j, β)	7	-2463.10	0.844	[0.806;0.882]	0.8915
BRD6	(α_j, β_j)	8	-2431.06	0.819	[0.788;0.849]	0.8919
BRD7	(α_k, β_k)	8	-2431.06	0.764	[0.697;0.832]	0.8919
BRD8	(α_j, β_k)	8	-2431.06	0.741	[0.657;0.826]	0.8919
BRD9	(α_k, β_j)	8	-2431.06	0.867	[0.851;0.884]	0.8919
Model 10	(α_k, β_{jk})	9	-2431.06	[0.762;0.893]	[0.744;0.907]	0.8919
Model 11	(α_{jk}, β_j)	9	-2431.06	[0.766;0.883]	[0.715;0.920]	0.8919
Model 12	$(\alpha_{jk}, \beta_{jk})$	10	-2431.06	[0.694;0.904]		0.8919



Estimator	$\hat{\theta}$
[Pessimistic; optimistic]	[0.694;0.904]
MAR (3 questions)	0.883
MNAR	0.782
MNAR "interval"	[0.753;0.891]
Model 10	[0.762;0.893]
Model 11	[0.766;0.883]
Model 12	[0.694;0.904]

33.11 Concluding Remarks

MCAR/simple	CC LOCF	biased inefficient not simpler than MAR methods
MAR	direct likelihood weighted GEE	easy to conduct Gaussian & non-Gaussian
MNAR	variety of methods	strong, untestable assumptions most useful in sensitivity analysis