



Statistics Canada

www.statcan.gc.ca

Methodology Q&A

QICSS

February 25 2010

Georgia Roberts

Data Analysis Resource Centre (DARC)

Georgia.Roberts@statcan.gc.ca



Statistics
Canada

Statistique
Canada

Canada 

QUESTION 0

**What is
the Data Analysis Resource Centre
(DARC)?**

ANSWER to QUESTION 0

- **DARC is a small group of methodologists who specialize in analytical methods & tools.**
- **PURPOSE of DARC**
 - To encourage, suggest and provide good analytical methods and tools for use with Statistics Canada data
- **APPROACH of DARC**
 - Consultation and collaboration with researchers and other methodologists

QUESTION 1

What is the position of Statistics Canada on accounting for the impact of the survey design when fitting multi-level models to complex survey data?

QUICK ANSWER TO QUESTION 1

- **Statistics Canada does not have a position or policy on any particular type of analysis, and thus nothing particular about multi-level modeling.**
- **The chapter on Data Analysis and Presentation in the Statistics Canada Quality Guidelines (2009) is also not specific in its recommendations.**

Excerpts from Quality Guidelines

- **“Choose an analytical approach that is appropriate to the question being investigated and the data to be analyzed.”**
- **Assess whether the survey design information can be incorporated into the analysis and if so how this should be done – such as by using a design-based approach.”**

Are there design-based methods that are appropriate for multi-level models?

- **Short answer: It depends on what you are interested in measuring.**
- **I am going to assume that the focus is the parameters of a multi-level model:**
 - the parameters associated with the random effects, and
 - the parameters associated with the fixed part of the model.

Are there design-based methods that are appropriate for multi-level models?

- **Multi-level modeling of survey data by a design-based method is still an area of research.**
- **Most statistical research on this has been restricted to the particular case of the hierarchical structure of the model matching the hierarchical structure of the design (e.g., Pfeffermann et al., 1998).**
- **Several commercial software packages have implemented these research results (e.g., Chantala and Suchindran, 2006).**

Are there any problems with using these software with Statistics Canada survey data?

- **There are not many cases where the hierarchical structure of model and design match.**
- **You need to have weights for each level of the model.**
- **The software currently can only provide design-based variance estimates by a Taylor approach, not by using survey bootstrap weights.**

Are there any cases where these problems can be surmounted?

- **There are some special cases where the model and the design levels do coincide:**
 - **WES: Businesses and then employees within businesses are sampled and modeled**
 - **People and then times within people are sampled and modeled for a two-level growth model**
- **For these special cases, weights do exist for both levels:**
 - **WES: Weights for both levels are provided**
 - **Growth model: Person-level weight is provided and times within a person would have weight of 1, since times were not sampled**

But what about software that can use bootstrap weights?

- **DARC members have written a SAS macro that links with software package HLM to produce variance estimates using survey bootstrap weights (Pierre and Saidi, 2008).**
- **We have heard that MPLUS is working on a version that will be able to use replication weights, based on a papers by Kovacevic et al. (2006) and Stapleton (2008).**

What other options are there?

- **Take a model-based approach for multi-level modeling and try to include any effects of the design that you might suspect directly in the model.**
- **Describe your approach in your paper, including the caveat that you are not taking a design-based approach.**



END OF QUESTION 1

QUESTION 2

**Is it possible to combine cycles
1.1, 2.1, 3.1, 4.1 (2007) and 5.1(2008)
of CCHS?**

**What would be an appropriate way
to do it?**

QUICK ANSWER TO QUESTION 2

It depends on the sorts of parameters that you want to measure,

as well as on the particular data sources that you wish to combine.

Three types of estimates people wish to make when combining

- 1. Descriptive: Estimate characteristics of one or more finite populations, such as:**
 - **(Composite weighted) average rate of the different populations**
 $\alpha p_1 + (1 - \alpha)p_2$
Ex: Average of 2001 and 2003 mammography rates
 - **Rate for combined target populations**
 $(N_1 p_1 + N_2 p_2) / (N_1 + N_2)$
Ex: Rate of mammography for the combined female populations of 2001 & 2003
 - **Difference in rates between 2 of the populations**
 $p_2 - p_1$
Ex: Change between 2001 and 2003 mammography rates

Three types of estimates

2. Assuming a relationship to estimate a descriptive quantity (of a finite population), such as:

- Rate over the combined populations, where each population is assumed to have the same rate

Ex: % of females who had a mammography in 2001, 2003 or 2005 where rate assumed the same in all 3 years

- Rate at a mid-point in time among sampled populations, where some relationship is assumed among this rate and the rates at the sampled time points

Three types of estimates

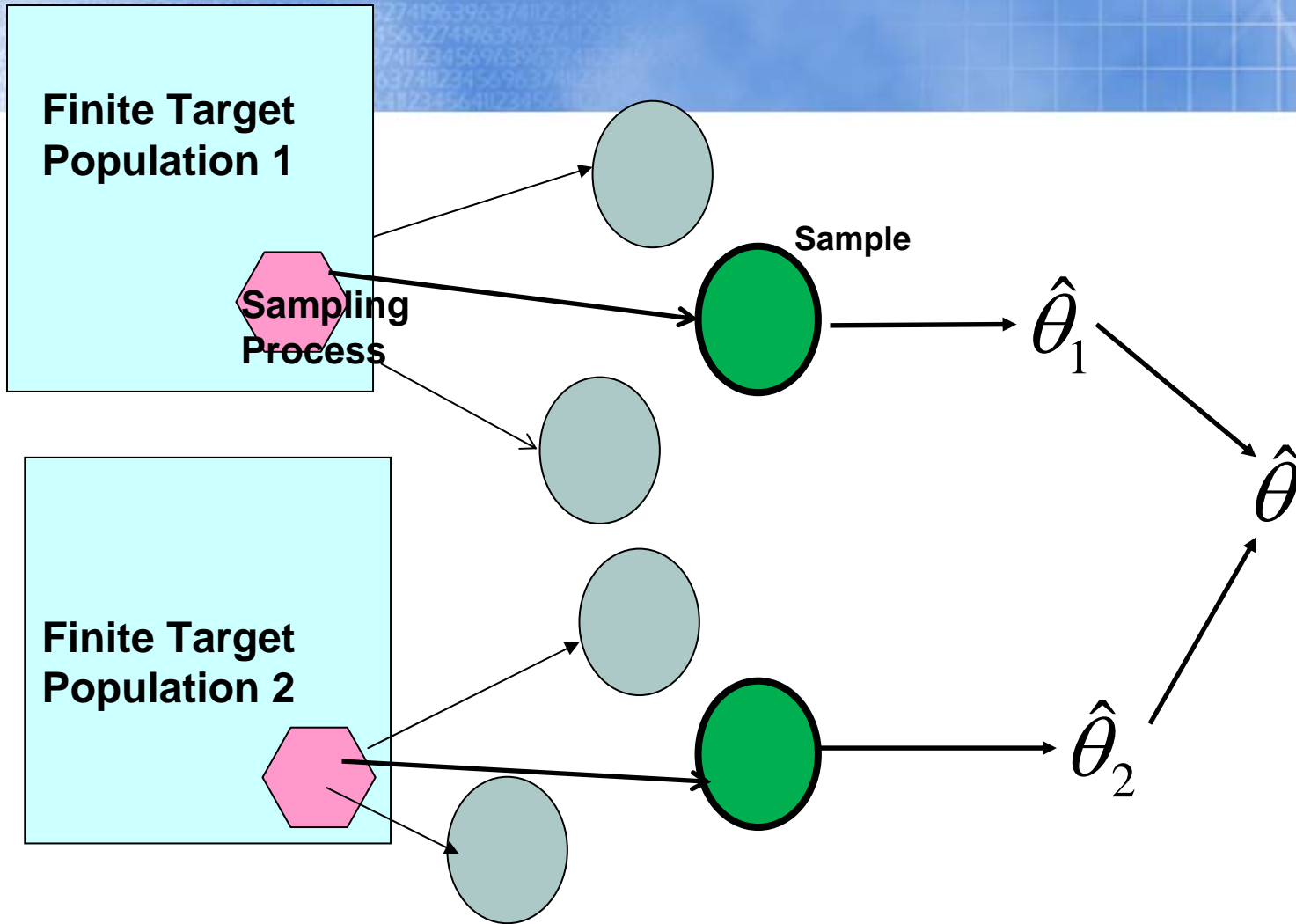
3. Analytic: Estimate relationships thought to hold beyond specific finite populations, often using a model

Ex: Odds of having a mammogram for poor women compared to rich women, controlling for effect of age, education and race, using a logistic model

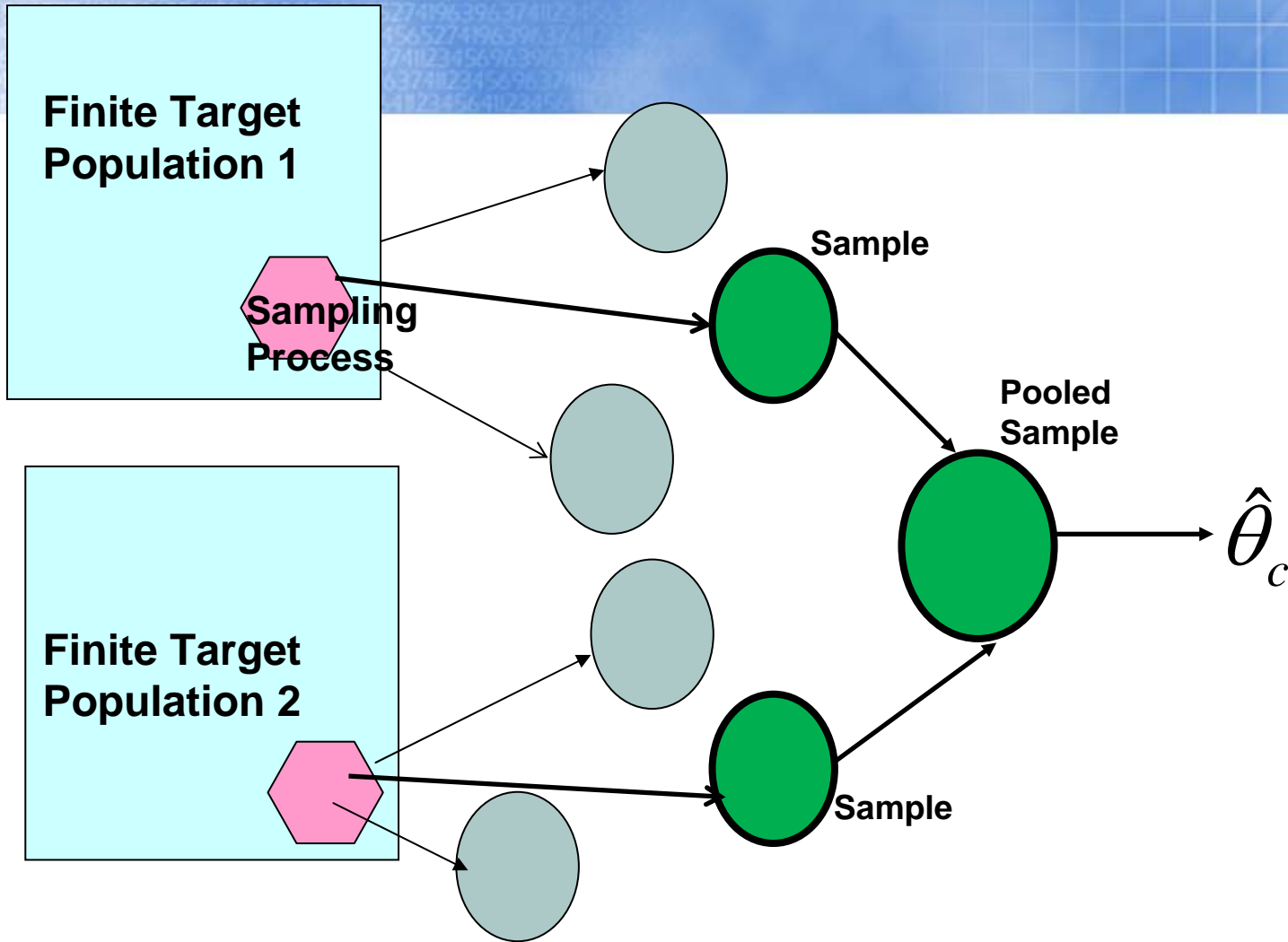
How to decide on appropriate method(s)

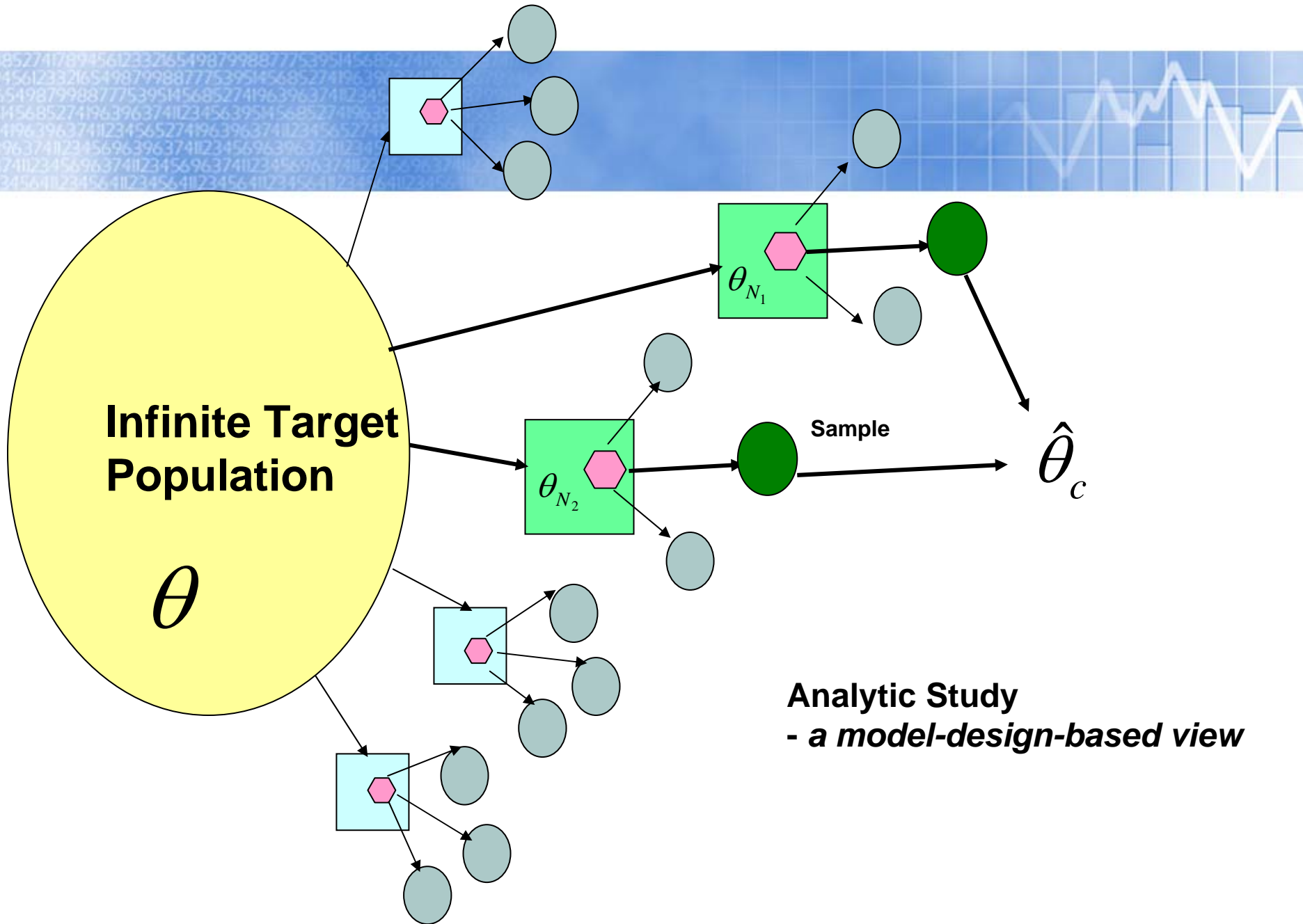
An examination of some different randomization frameworks could give strong guidance about how to estimate quantities of interest and make desired inferences.

Descriptive Estimation – Separate Approach



Descriptive Estimation – Pooled Approach





What is usually meant by the
“Pooled (or Pooling) Approach” to analysis

- (i) Pool data from different surveys into a single file**
- (ii) Create a weight variable appropriate for the pooled data, your target population , your assumptions, and the quantities of interest**
- (iii) Create design information for variance estimation (e.g. new bootstrap weights if you have modified the weight variable)**
- (iv) Do confirmatory work about any assumptions being made**
- (v) Carry out estimation and inference on the pooled data using techniques that would be appropriate for data from a single sample**

Comments

- **It may be possible to estimate the same quantity by separate and pooled approaches, but the estimates themselves may not be the same.**
- **Using a pooled approach when fitting models seems like the most reasonable way to proceed. Differences between target populations of the different data sources **SHOULD** be considered for inclusion in the model.**

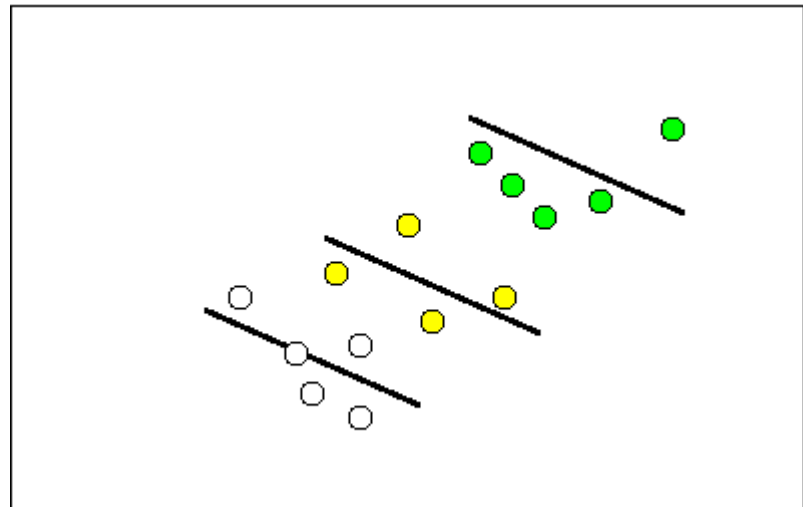
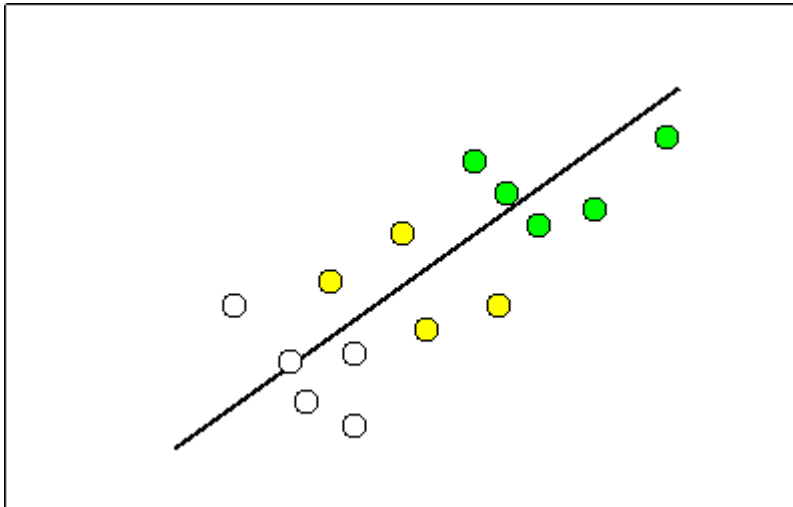
POOLING DATA

Infinite Target Population

Example: LINEAR MODEL RELATING TWO VARIABLES

Left: (Incorrectly) assuming same intercept and slope for all 3 finite populations

Right: (Correctly) allowing for different intercepts but same slope for different finite populations



Important Issues to be considered first *in all cases*

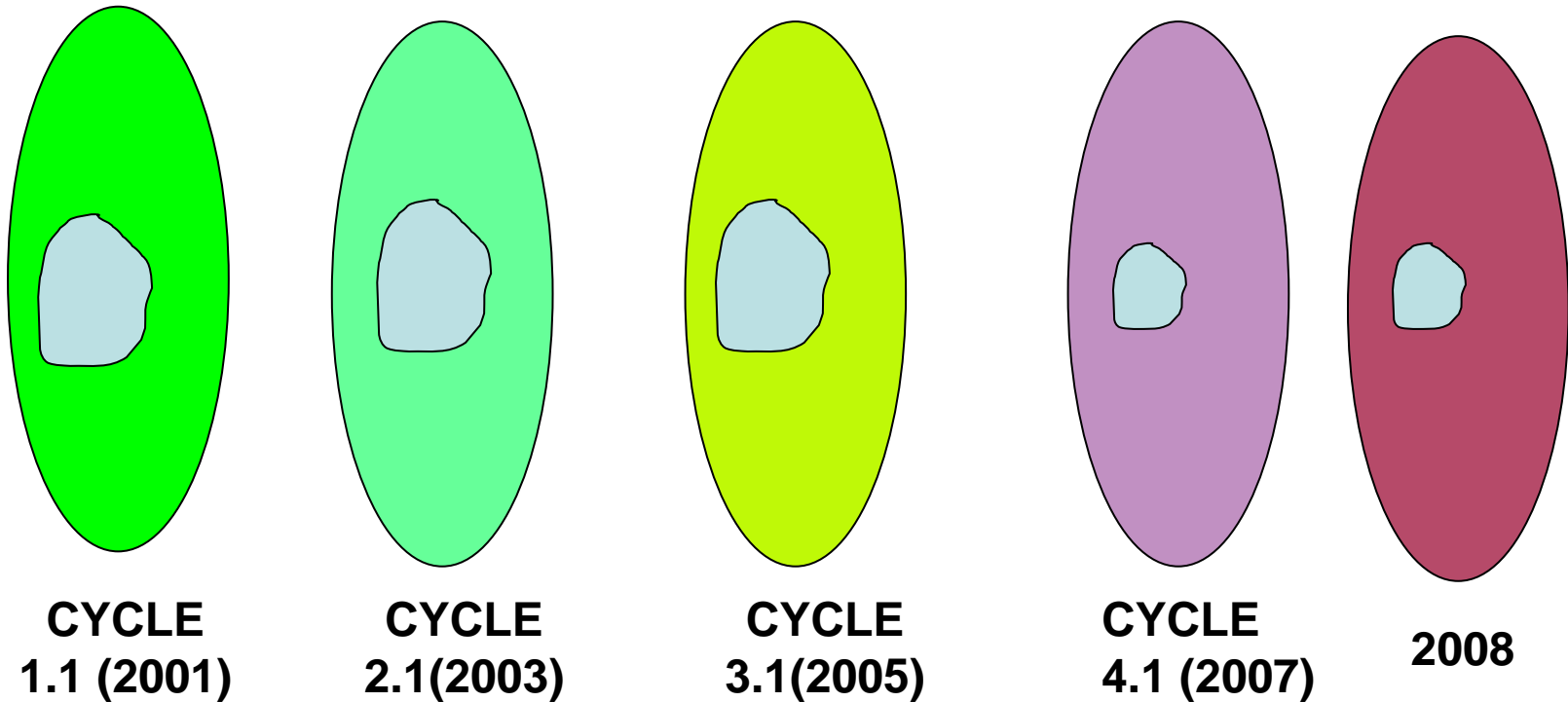
- Are survey designs the same or very similar?
- Are questions and questionnaires the same?
- Is the mode of delivery the same?
- Is the type of respondent the same?
- Are the samples independent?

■ **ALSO IMPORTANT comparability QUESTION**

How do the target populations of the data sources compare?

- same (both target group and time)?
- same target group but different time? (Most common case)
- different target group and time?

The CCHS cycles have the same target group at different times. Each target group is sampled and interviewed once.



While survey designs were fairly constant over Cycles 1.1, 2.1 and 3.1, the data collected for 2007 and 2008 were under a new design with a smaller sample size each year.

Additional Comments

In general, suitable variance estimation may be difficult when combining data from different samples, especially if samples are not independently selected. However, Cycles 1.1, 2.1 and 3.1 of CCHS can be assumed independent. I think that the combined 2007-2008 data can be assumed independent of these 3 cycles.

Health Reports paper by Thomas and Wannell (2009) discuss combining of data from Cycles 1.1, 2.1 and 3.1.



END OF QUESTION 2

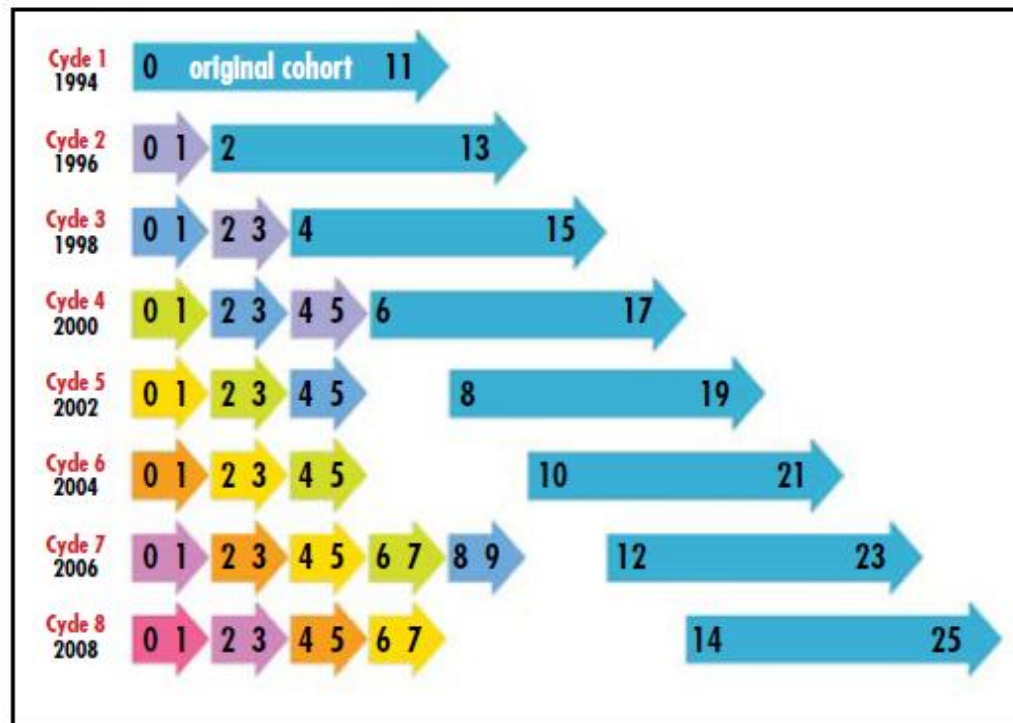
Design of NLSCY – series of cohorts

Different target groups and different times for the cohorts

REFERENCE YEAR					
94/95	96/97	98/99	00/01	02/03	04/05
		COHORT 1			
		Aged 0-11 when selected			
	COHORT 2 Aged 0-1 when selected				
		COHORT 3 Aged 0-1 when selected			
			COHORT 4 Aged 0-1 when selected		
				COHORT 5 Aged 0-1...	
					COHORT 6 Aged 0-1 when .
30					

Another view of NLSCY – showing the age ranges of data collected each cycle, as well as identifying cohorts by different colours

Age of children at each cycle, original cohort versus the early childhood development cohorts



Source: Statistics Canada, National Longitudinal Survey of Children and Youth.

QUESTION 3

- **Consider the following analysis to be done with NLSCY data:**
 - Estimation and comparison of proportions of Canadian children aged 2 to 9 in 1994, 1996, 1998, & 2000 who are hyperactive and who are taking Ritalin
- **What weight would you use for doing this analysis?**
- **How is the dependence in the observations accounted for when doing variance estimation?**

ANSWER TO QUESTION 3

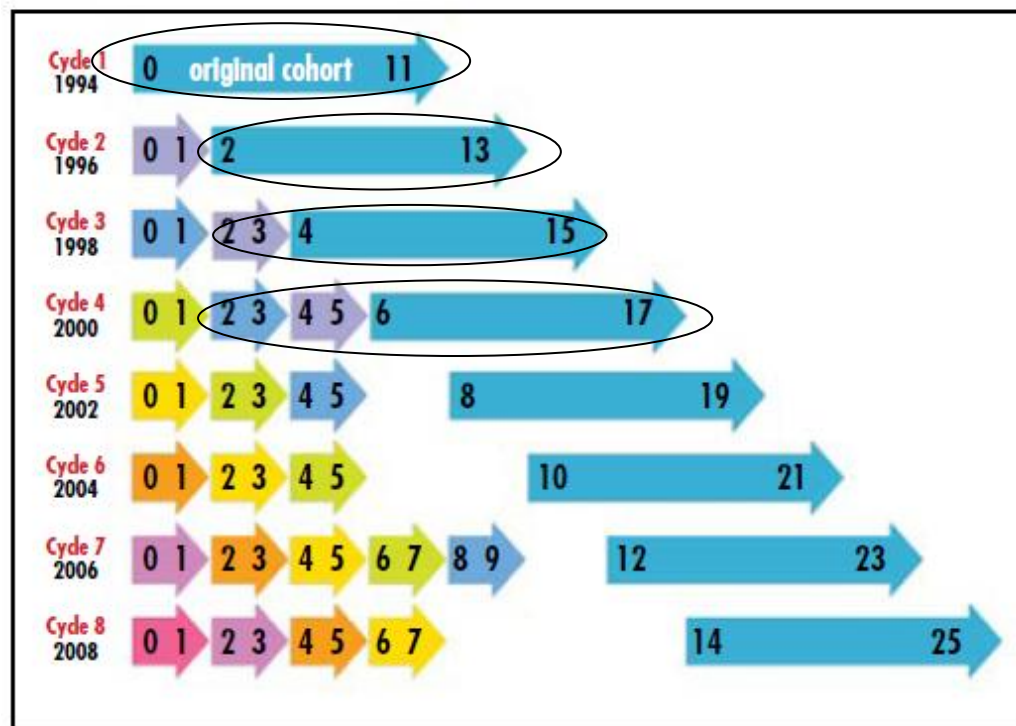
- The analyst is interested in some characteristics of all Canadians aged 2 to 9 in 1994, of all Canadians aged 2 to 9 in 1996, of all Canadians aged 2 to 9 in 1998 and of all Canadians aged 2 to 9 in 2000. He wants to compare across these populations.
- The ideal survey design would have been a set of cross-sectional samples, one drawn from the Canadian population at each of these years.

ANSWER TO QUESTION 3 (cont'd)

- We do have a sample of 2-9's drawn from the 1994 Canadian population, and data from them for that time point (Cycle 1 of NLSCY).
- We do have data collected from 2-9's at each of the other time points, although these 2 to 9's are from a variety of cohorts and do not represent all 2-9's in Canada at these times.
- NLSCY has provided a weight variable at each of 1996, 1998 and 2000 so that estimates from the samples may be closer to being cross-sectionally representative.

View of NLSCY – what data might be used for the analyses of Question 3

Age of children at each cycle, original cohort versus the early childhood development cohorts



Source: Statistics Canada, National Longitudinal Survey of Children and Youth.

ANSWER TO QUESTION 3 (cont'd)

- **Use of the cross-sectional weight at each time point should give a reasonable estimate for that time point, and the corresponding cross-sectional bootstrap weights should give good variance estimates for that time-specific point estimate.**
- **Since it was recognized that analysts would wish to compare cross-sectional characteristics for different time points, and since the samples are not independent at different time points, the cross-sectional bootstrap weights were created in a special “coordinated” way.**

ANSWER TO QUESTION 3 (cont'd)

- In the construction of the coordinated bootstrap weights, the sample of psu's from each stratum for each bootstrap weight is selected at the first time point that a cohort could be used for cross-sectional purposes.
- These same samples are then retained in all subsequent cycles for which the cohort is used for cross-sectional purposes.

ANSWER TO QUESTION 3 (cont'd)

- How an analyst should then use the cross-sectional bootstrap weights for estimating variances of quantities involving more than one cross-sectional point estimate:

Be sure to line up bootstrap weights of the same number over the different time points involved.

Example of correctly using coordinated bootstrap weights

p_{94} and p_{96} are estimates of the proportion of children in 1994 and 1996 who are hyperactive, using the cross-sectional weights at each time. We want estimate of variance of $(p_{94} - p_{96})$.

Calculate estimate of difference in proportions using bootstrap weight #1 from each time point, then bootstrap weight #2 from each time point, etc.

Combine these estimates in the usual way to obtain the variance estimate.



END OF QUESTION 3

QUESTION 4

- **Consider the following analysis to be done with NLSCY data:**
 - Fitting of a marginal growth model to children when they are aged 4 to 9, where the children have to be in the initial NLSCY cohort and where the dependent variable for the model was only collected beginning in Cycle 2.
- **How is the dependence in the observations accounted for in the model and when doing variance estimation?**
- **What weight would you use for doing this analysis?**

The analyst states that only Cohort 1 is to be used for this analysis.

REFERENCE YEAR					
94/95	96/97	98/99	00/01	02/03	04/05
		COHORT 1 Aged 0-11 when selected			
	COHORT 2 Aged 0-1 when selected				
		COHORT 3 Aged 0-1 when selected			
			COHORT 4 Aged 0-1 when selected		
				COHORT 5 Aged 0-1...	
					COHORT 6 Aged 0-1 when .
42					

Table shows observations from Cohort 1 (by age and cycle) that satisfy criteria.

94	96	98	00	02	04	06	08
0	2	4	6	8	10	12	14
1	3	5	7	9	11	13	15
2	4	6	8	10	12	14	16
3	5	7	9	11	13	15	17
4	6	8	10	12	14	16	18
5	7	9	11	13	15	17	19
6	8	10	12	14	16	18	20
7	9	11	13	15	17	19	21
8	10	12	14	16	18	20	22
9	11	13	15	17	19	21	23
10	12	14	16	18	20	22	24
11	13	15	17	19	21	23	25

More about the model

- Let i identify the person and j identify the year
- Simple marginal growth model:

$$E(y_{ij}) = \alpha + \beta \text{age}_{ij}$$

- The model could also specify the structure of the dependence among observations
 - Classically, people are considered independent, but some structure is assumed for the dependence (or covariance) among observations from the same person .

More about the model

- **When fitting marginal models to longitudinal survey data, some software does allow you to specify a “working covariance structure” for observations from the same person.**
(e.g., SUDAAN has independence or exchangeable)
- **But, as in the classic case, estimates of model coefficients are approximately unbiased, even if within-person covariance structure is wrongly specified.**
- **Between-person covariance in the estimates is picked up through the design-based variance estimation.**

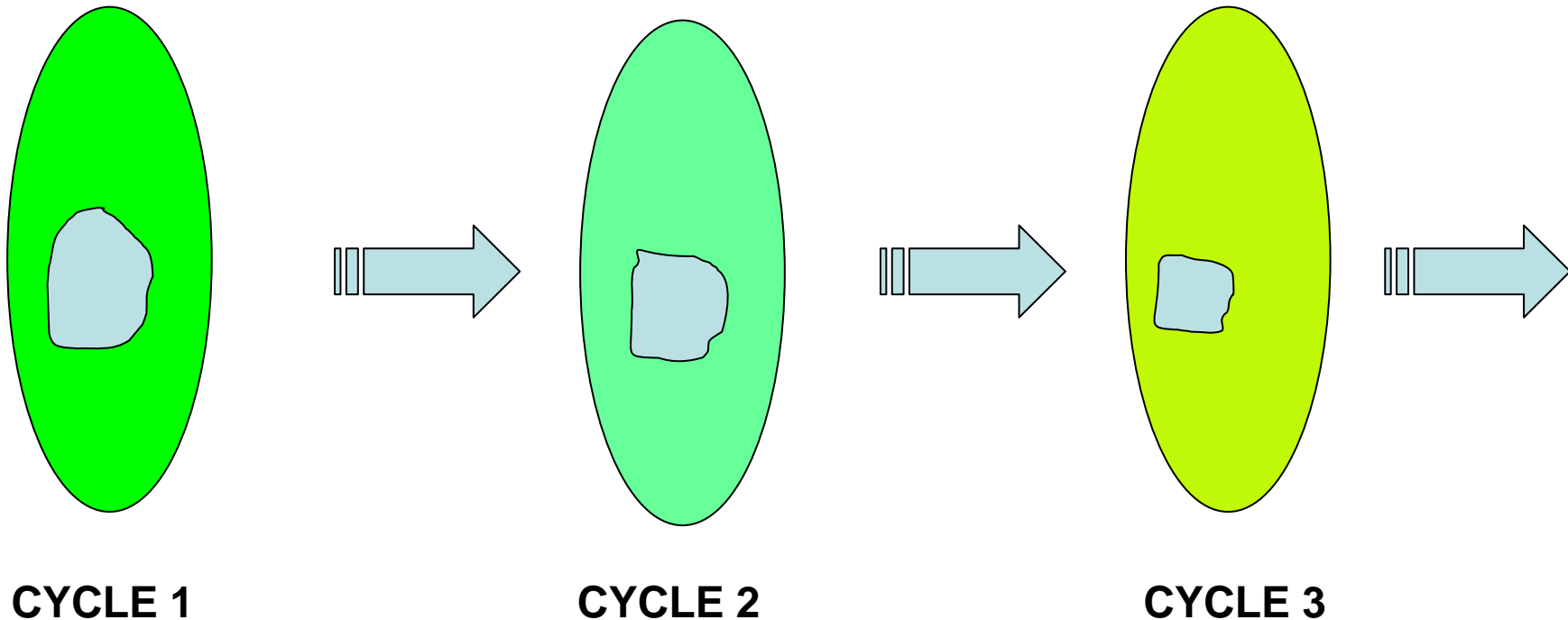
PART 2 of QUESTION 4

What weight variable would make sense when doing this analysis with this longitudinal cohort?

RECALL WHAT HAPPENS IN A LONGITUDINAL SURVEY AND WHAT WEIGHTS TRY TO DO

SAME POPULATION (BUT AGING) OVER TIME

Different subsample representing that population each time
(due to nonresponding sample units)



WHAT IS THE CONSEQUENCE OF USING DIFFERENT LONGITUDINAL WEIGHT VARIABLES?

(a) If you should use the weight of the final cycle from which you use data, you could have observations for individuals for which the weight variable chosen has the value 0.

Impact 1: Perhaps less efficient estimates than if you could use all the data

Impact 2: This final weight could produce biased results if what you are studying is related to dropout.

WHAT IS THE CONSEQUENCE OF USING DIFFERENT LONGITUDINAL WEIGHT VARIABLES?

(b) If you should use the weight of the first cycle, some data may be missing (due to unit nonresponse at one or more cycles) for some people.

Impact: Perhaps more bias in the estimates than if you had all the data for all individuals.

Thus, often a tradeoff as to which weight variable to use.

ARE THERE OTHER ALTERNATIVE WAYS TO PROCEED?

Sometimes the quantity to be estimated can be broken into pieces, where each piece involves a different cycle. Then a different weight variable can be used to estimate each piece.

e.g., repeated measures in marginal models



END OF QUESTION 4

REFERENCES

- Chantala, K. and Suchindran, C. (2006). Adjusting for Unequal Selection Probability in Multilevel Models: A Comparison of Software Packages. Proceedings of the Survey Research Methods Section, JSM, Seattle, Washington.
- Kovacevic, M., Huang, R., and You, Y. (2006). Bootstrapping for Variance Estimation in Multi-Level Models Fitted to Survey Data. Proceedings of the Survey Research Methods Section, JSM, Seattle, Washington.
- Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H. and Rasbash, J. (1998) Weighting for unequal section probabilities in multilevel models. *J. Roy. Statist. Soc. B* 60, 23-56.
- Pierre, F. and Saidi, A. (2008). Implementing Resampling Methods for Design-Based Variance Estimation in Multilevel Models: Using HLM6 and SAS Together. Proceedings of the Survey Research Methods Section, JSM, Denver, Colorado.
- Stapleton, L.M. (2008). Variance estimation using replication methods in structural equation modeling with complex sample data. *Structural Equation Modeling*, 15, 183-210.
- Statistics Canada Quality Guidelines (2009) <http://www.statcan.gc.ca/cgi-bin/af-fdr.cgi?l=eng&loc=http://www.statcan.gc.ca/pub/12-539-x/12-539-x2009001-eng.pdf&t=Statistics%20Canada%20Quality%20Guidelines>
- Thomas, S. and Wannell, B. (2009). Combining cycles of the Canadian Community Health Survey . Health Reports. <http://www.statcan.gc.ca/pub/82-003-x/2009001/article/10795-eng.htm>

Note that papers in JSM Proceedings can be accessed at <http://www.amstat.org/sections/SRMS/Proceedings/>

- For more information, please contact:

- Pour plus d'information, veuillez contacter :

Georgia Roberts

Data Analysis Resource Centre

Methodology Branch, Statistics Canada

Georgia.Roberts@statcan.gc.ca