Hands-on Workshop on the use of Health Surveys data (NPHS & CCHS)

François Brisebois, Patrice Mathieu, Mario Bédard

Statistics Canada

Outline of the workshop

- Quick review of the survey(s)
- General context of a statistical analysis
- How to handle missing data
- Overview of some types of cross-sectional analysis
- Comparing populations
  - Age-sex standardisation
  - Warnings about the comparison of cross-sectional estimates over time
- Overview of software programs for data analysis
- General references on aspects covered

Module 1

Quick Review of the Surveys

Outline

- The STC Health Surveys Program
- Terminology
- National Population Health Survey (NPHS)
- Canadian Community Health Survey (CCHS)

Health Survey Program

1994-1999

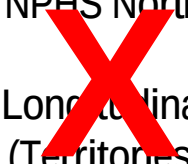
HEALTH SURVEYS

NPHS Household	NPHS North	NPHS Institutions
Longitudinal & X-sectional (Provinces)	Longitudinal & X-sectional (Territories)	Longitudinal & X-sectional (-C3) (Provincial Health Institutions)

Since 2000

HEALTH SURVEYS

NPHS Household	NPHS North	NPHS Institutions	CCHS
Longitudinal (Provinces)	Longitudinal (Territories)	Longitudinal (Prov. Hlth Institutions)	Cross-sectional (Provinces + Territories)



Terminology

Longitudinal versus Cross-sectional

- **Cross-sectional:**
  - **Survey a specific population at a given period of time**
- **Longitudinal:**
  - **Survey a specific population repeatedly over a period of time**

Terminology

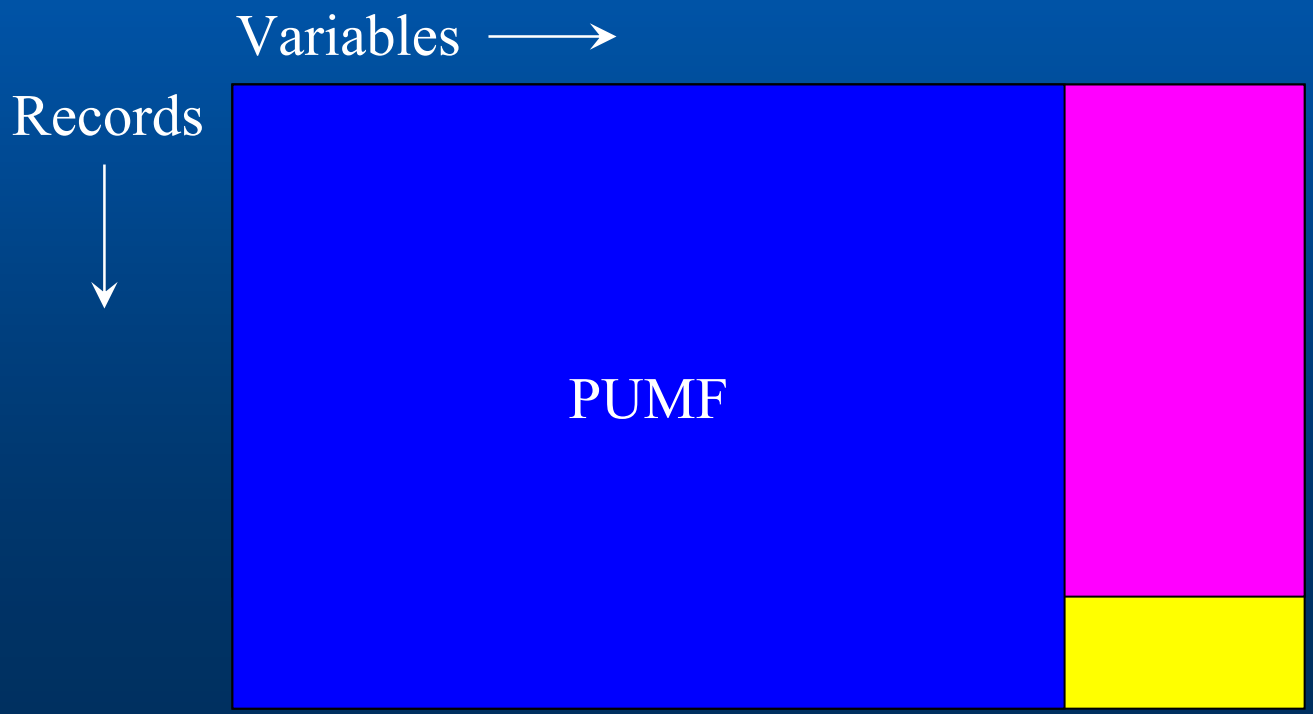
Master, Share, PUMF, Dummy files

- **Master:** File that contains all variables for all respondents
- **Share:** Contains all variables but only for people who accepted to share *(subset of records)*
- **PUMF:** File that contains a subset of variables for all respondents *(subset of variables)*
- **Dummy:** Scrambled version of the master file (for testing only / remote access)

Terminology

Master, Share, PUMF, Dummy files

- Illustration of Master vs. Share vs. PUMF

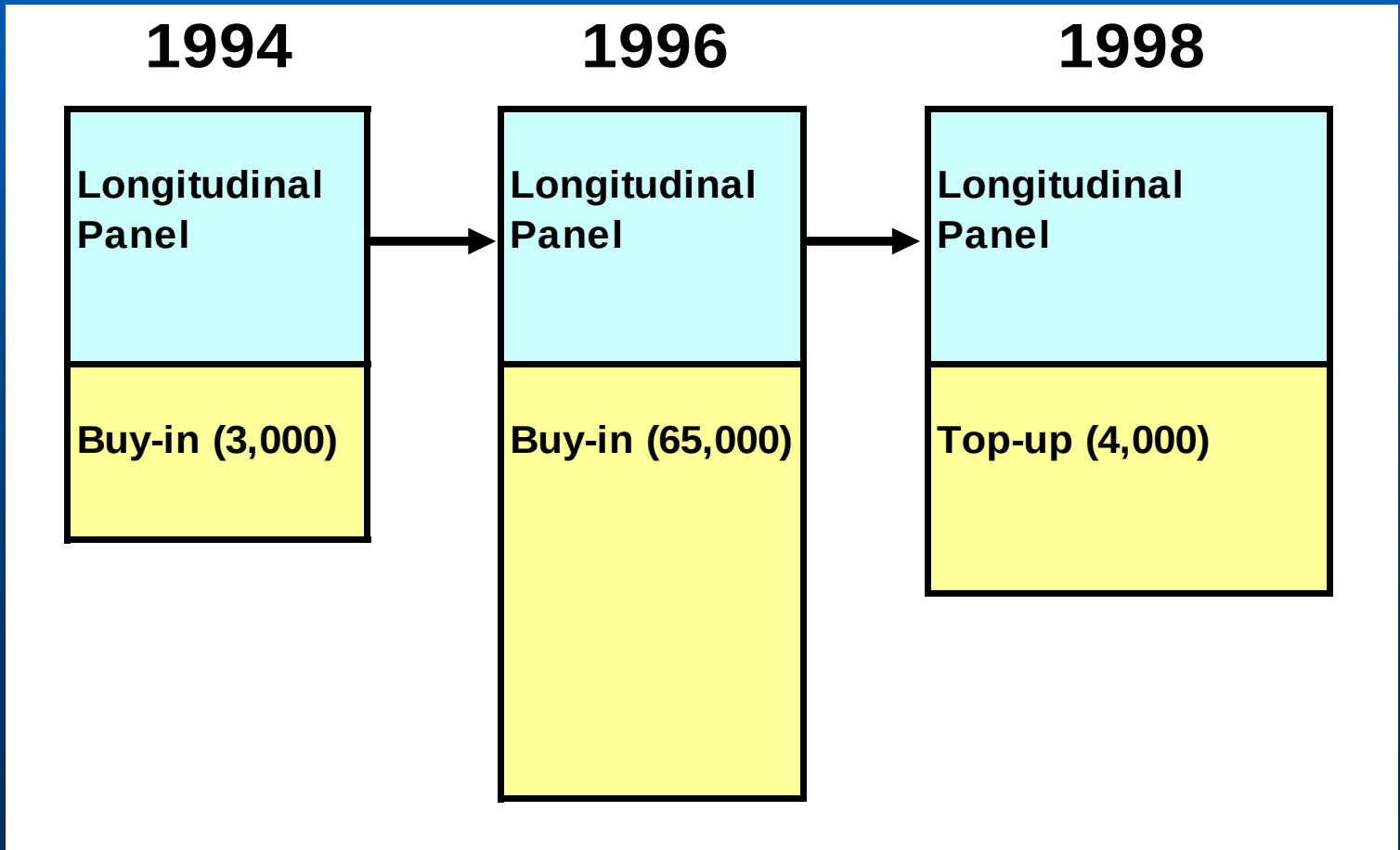


National Population Health Survey (NPHS)

- Longitudinal survey
  (but served both cross-sectional & longitudinal purposes for the 3 first cycles)
- Panel = 17,276 respondents in 1994
- Collection started in 1994-95 (Cycle 1), and panel is contacted every second year for 20 years

National Population Health Survey (NPHS)

- Cross-sectional Samples Composition



National Population Health Survey (NPHS)

- **Current situation:**
  - **Cycle 4 (2000-01):**
    - Longitudinal file was released in May 2002
  - **Cycle 5 (2002-03):**
    - Collection completed; data processed
    - New focus content: Sleep, nutrition, residential history, brand of cigarette
    - Dissemination of file planned for Spring 2004
  - **Cycle 6 (2004-05):**
    - Getting ready for collection starting in June 2004

Canadian Community Health Survey (CCHS)

- Cross-sectional survey
- Biennial cycle:
  - 1st year (.1 -- regional component):
    - Health region level -- Sample 130K
    - General content (with optional content)
  - 2nd year (.2 -- provincial component)
    - Province level -- Sample 30K
    - Focus content

Canadian Community Health Survey (CCHS)

- **Current situation:**
  - **Cycle 1.1 (2000-01):**
    - Master / Share files released in May 2002
    - Public-Use Microdata file released in January 2003
  - **Cycle 1.2 (2002):**
    - Mental Health
    - Ontario & Nova Scotia sample buy-ins - > Health Region level representative samples
    - Master & share files released in September 2003
    - Public-Use Microdata file -> Winter 2004

Canadian Community Health Survey (CCHS)

- **Current situation (cont'd):**
  - **Cycle 2.1 (2003):**
    - **Collection until the end of 2003**
    - **Changes in the geography**
    - **Sample buy-ins for 3 HR in Québec (CLSC representative samples)**
  - **Cycle 2.2 (2004):**
    - **Nutrition**

Health Surveys (CCHS +NPHS)

- Summary of cross-sectional files for the household population

Year	Survey	Age	Master & PUMF	Share
1994-95	NPHS – C1	0+	General / Health	General / Health
1996-97	NPHS – C2	2+ *	General / Health	General / Health
1998-99	NPHS – C3	0+	General / Health	General / Health
2000-01	CCHS 1.1	12+	Health	Health
2002	CCHS 1.2	15+	Health	Health

Material Used for the Workshop

Material

- **Dummy data**
  - **CCHS Cycle 1.1 dummy file**
- **Information included**
  - **Data (ASCII) with record layouts**
  - **Data dictionnaries**
  - **Documentation**
  - **Bootstrap**

SPSS / SAS - Read the Data

- ASCII file -> Need record layout
- Very large number of variables
- Use provided peripheral SPSS / SAS files
  - Record layout
  - Variable labels
  - Value labels

SPSS / SAS - Read the Data

- **Computer directories used:**
  - **Network personal directory**
    - **\\ciqss-s2\utilisateurs$\formation\ordinateur xx**
      - **\Data**
      - **\Layout**
      - **\Bootstrp\Data**
      - **\Bootstrp\Layout**

SPSS / SAS - Read the Data

- Read the data in SPSS / SAS
  - Use the template provided to read all records
  - Next, keep only records for the selected region

Module 2

General Context of Statistical Analysis

Outline

- Do I have enough sample?
- Estimation of a statistic
- Estimation of the statistic's precision

Do I have enough sample?

- **Want sample analyzed to be representative**
  - **What does it mean to be representative?**
    - **Can produce estimate unbiasedly, and with a reasonable precision**
      - **Unbiasedly: Proper coverage of population studied**
      - **Precision: Is function of the sample size and of the magnitude of the proportion examined**

Do I have enough sample?

- NPHS was designed to guarantee a good representativity for 10 age-sex groups within each province
  - Age groups: 0-11, 12-24, 25-44, 45-64, 65+
  - Within each health region for provinces that bought more sample in cycles 1 and 2

Do I have enough sample?

- For CCHS,
  - Cycle 1.1:
    - Within each of the 136 Health Regions
    - 10 age-sex groups (12-19, 20-29, 30-44, 45-64, 65+)
  - Cycle 1.2:
    - Provincial; regional for Ontario and Nova Scotia
    - 8 age-sex groups (15-24, 25-44, 45-64, 65+)

Do I have enough sample?

- **For CCHS,**
  - **Cycle 2.1:**
    - **Within each of the 133 Health Regions**
    - **(Possibly) within each CLSC for 3 health regions in Québec**
    - **10 age-sex groups (12-19, 20-29, 30-44, 45-64, 65+)**
  - **Cycle 2.2:**
    - **Within each province (within sub-provincial regions for provinces that bought extra sample)**
    - **15 age-sex groups**
      (<1*, 1-3*, 4-8*, 9-13, 14-18, 19-30, 31-50, 51-70, 71+)

Do I have enough sample?

- **What can I check?**
  - **Does the survey's target population perfectly includes the population studied?**
    - **Refer to documentation for definition of target population**
  - **Is the precision ok ?**
    - **Usually do not know in advance**
    - **Scenarios:**

      Sample size (large or small) VS. Proportion studied (large or small)
    - **CV Tables can be helpful**

Do I have enough sample?

- **Scenarios (related to CCHS 1.1) :**
  - **Smoking rate for a Health Region**
  - **Smoking rate for teenagers in Canada**
  - **Alzheimer in Canada**
  - **Alzheimer for a Health Region**
  - **Diabetes (CCCA_101 = 1) for a Health Region (GEOA_HR4)**
    - **Obtain count and proportion in sample**

Do I have enough sample?

- For analyses from the PUMF, guidelines require a minimum sample size of 30 (see section 10.4 of PUMF Users Guide)
- For Share and Master data, precision rules (must be calculated -- with Bootstrap)
  - Small area estimation techniques

Estimation of a statistic

- Estimation relates sample back to population (inference)
  - Done through the use of sampling weights
- What are sampling weights?
  - Number of people the interviewed person represents in the population
    - Ex.: Weight = 500

Estimation of a statistic

- **What are sampling weights? (cont'd)**
  - **Based on the probability of selection**
    - **A person selected according to a sampling fraction of 1% will have a (initial) weight of 100**
    - **Sampling fractions differ between regions, therefore weights are different from one person to another**
  - **Corrected for total nonresponse, and to match population projections**

Estimation of a statistic

► Name of weight variables:

Survey	Cycle	File	Details	Name (Master/Share)
NPHS	1	General		WT54 / SHRWT5
		Health		WT64 / SHRWT6
NPHS	2	General		WT56 / WT56_S
		Health	Except HPS and Child Hlth Services	WT66 / WT66_S
NPHS	3	General		WT58 / WT58_S
		Health		WT68 / WT68_S
CCHS	1.1	HSI	Total sample	WTSA_M / WTSA_S
			Quarter 4 only	WTSA_Q4M / WTSA_Q4S
			PEI sample buy-in	WTSA_PEM / WTSA_PES
			BC - 16 regions	WTSAM
CCHS	1.2	HSI	Total sample	WTSB_M / WTSB_S

Estimation of a statistic

- How to use the weights in computations:
  - In SPSS:
    - Use “Weight Cases” under “Data” menu
  - In SAS:
    - Use WEIGHT statement within the procedure used to compute the statistic
- Example:
  - Estimate of the number of people with diabetes in Ontario

Estimation of a statistic

- **Exercise:**
  - **Estimate the number of people with diabetes in your region (total number and rate)**
  - **Compare unweighted and weighted rates**

Estimation of the precision

- Precision = Sampling error
- From the fact that results are obtained from a sample and not a census
- Precision is a function of:
  - Sample and population size
  - Sampling design used (design effect)
  - Magnitude of the proportion estimated

Estimation of the precision

- **Measures of precision:**
  - **Variance, Std deviation**
  - **Coefficient of variation (CV)**
    - **CV = $\frac{\text{Std deviation of estimate x 100\%}}{\text{estimate itself}}$**
    - **E.g.: 24% of Cdn population are daily smokers,**
      Std deviation = 0.003
      CV = 0.003 / 0.24 x 100% = 1.25%
    - **CV allows comparison of precision of estimates with different scales**

Estimation of the precision

- Computing an estimate is simple (use weight)
- When using NPHS/CCHS data, the precision of an estimate is more complex to calculate
  - Why?
    - Data is collected from a survey with a complex survey design
    - For complex survey design, no exact formulae for calculation of precision

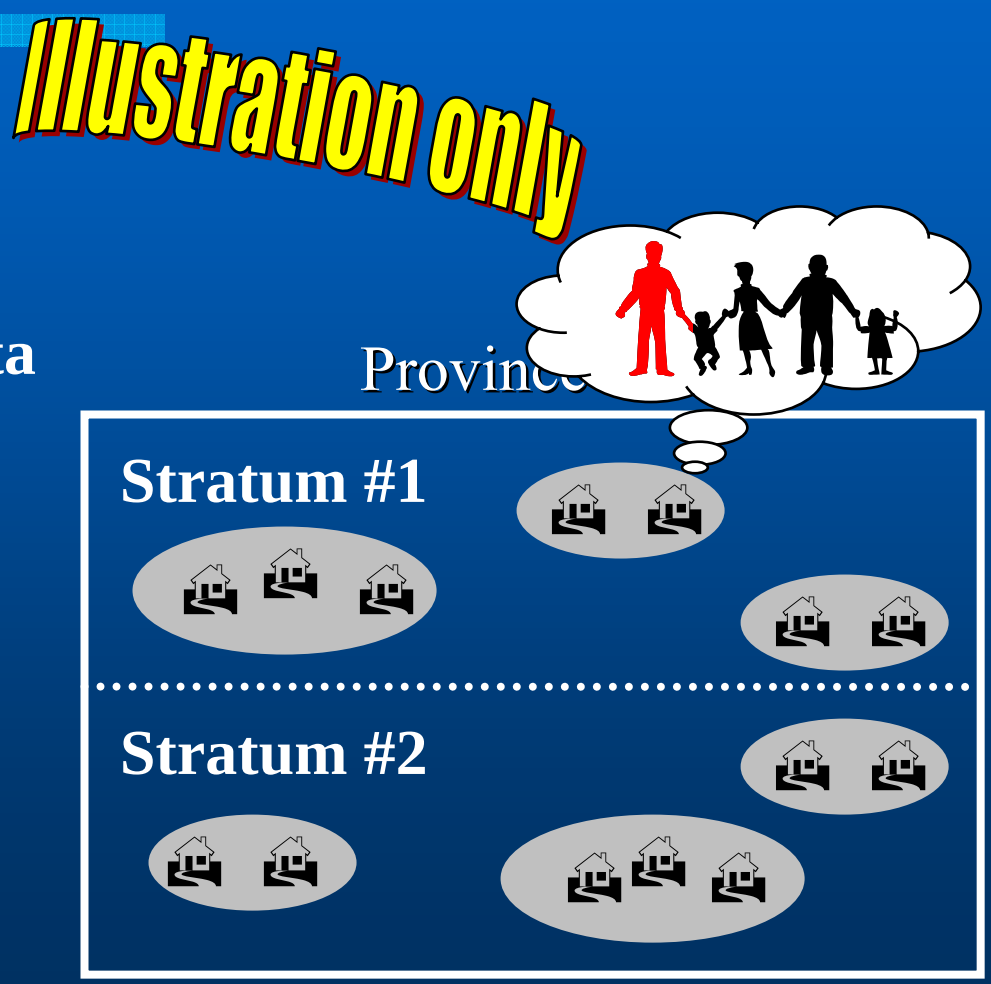
Complex Survey Design

#1: Each province is divided into strata

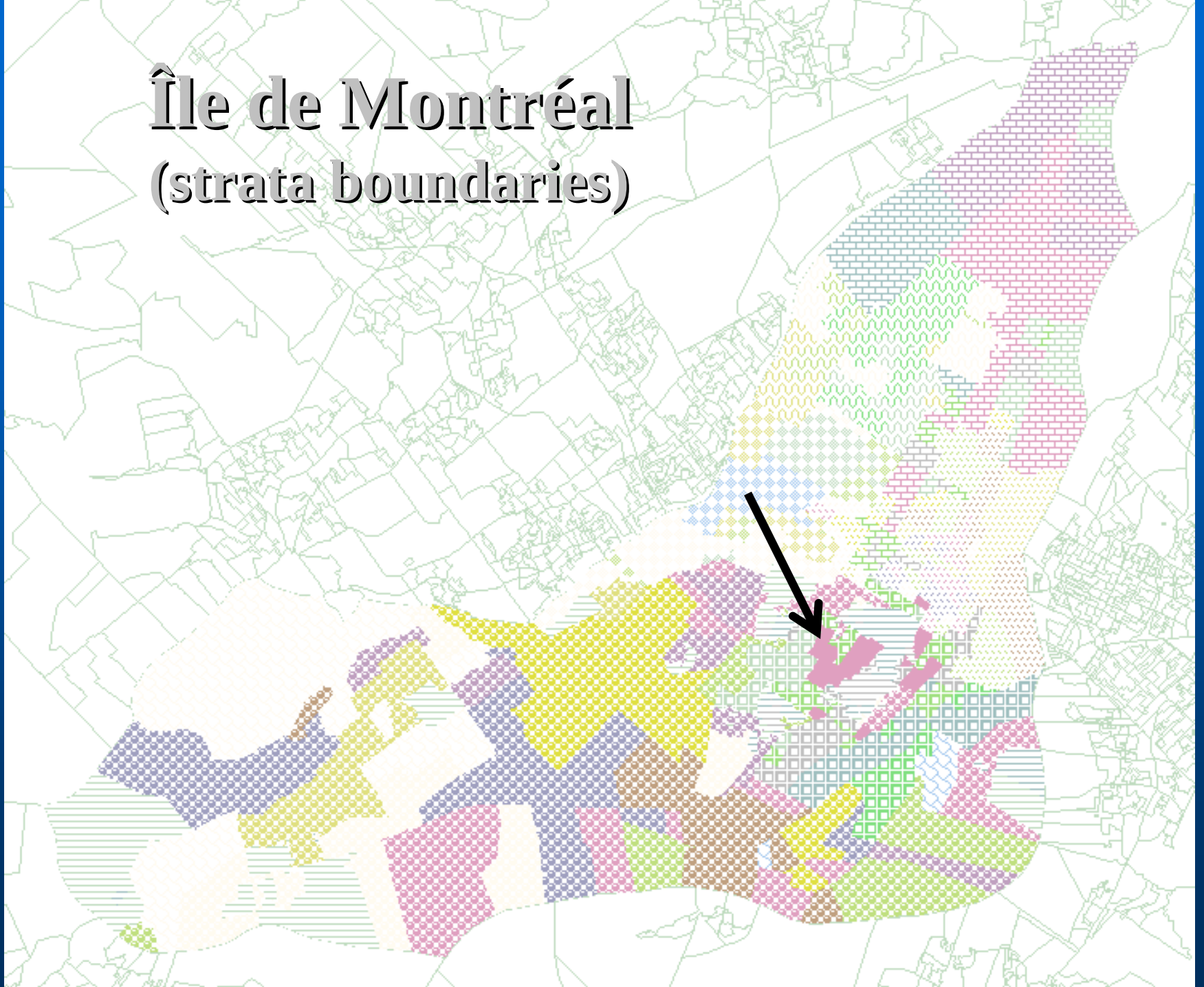
#2: Clusters selected within strata (PPS sampling) (1st stage)

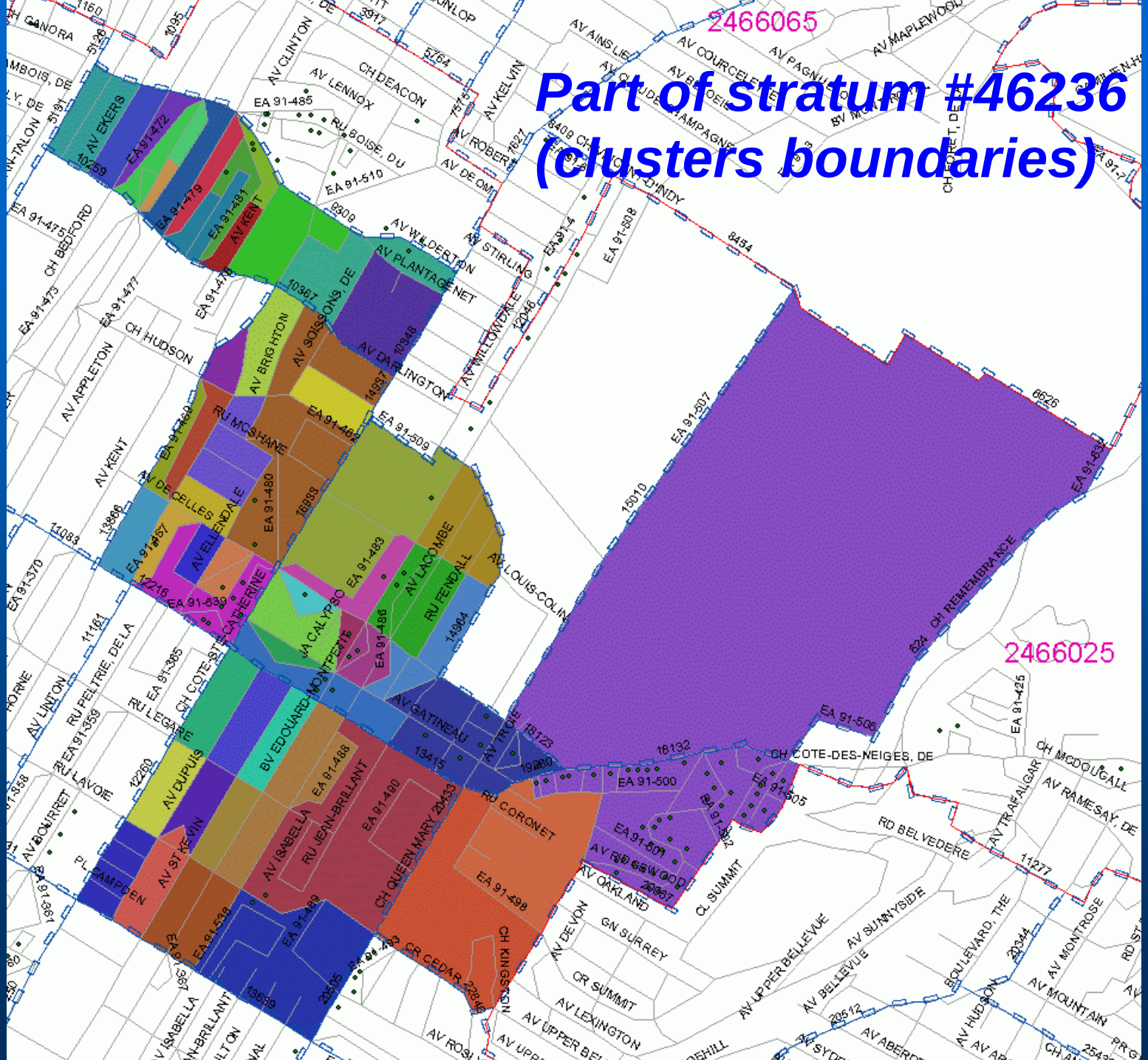
#3: Dwellings selected within clusters (2nd stage)

#4: People selected within responding dwellings (3rd stage)



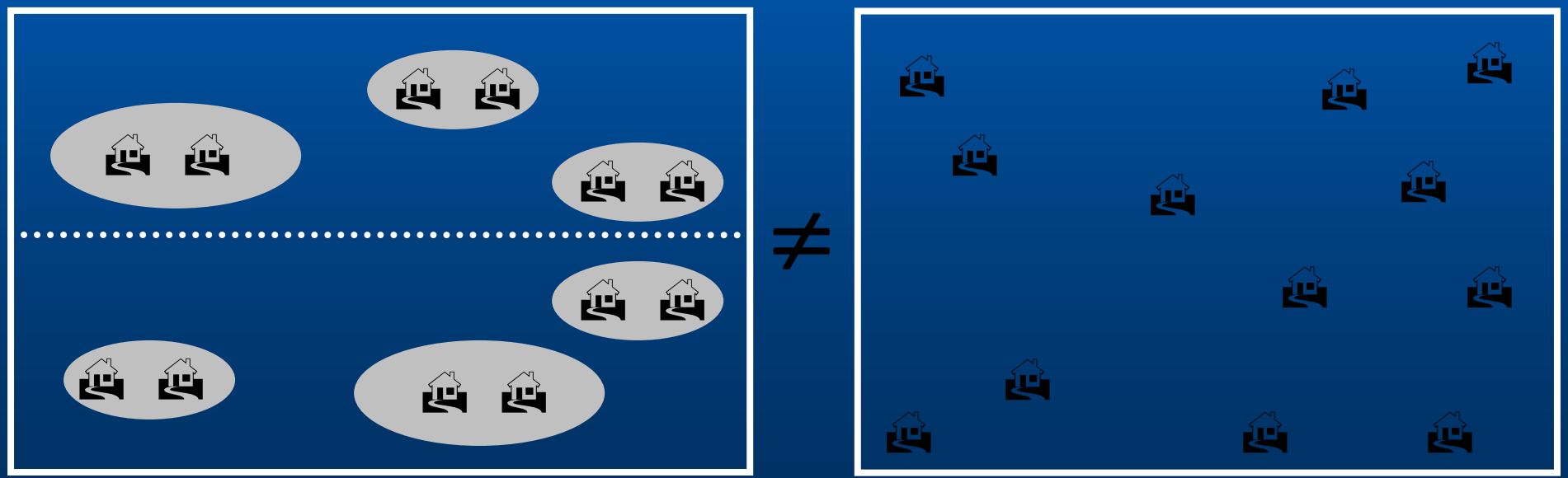
Île de Montréal
(strata boundaries)




Part of stratum #46236
(clusters boundaries)
2466065
2466025
AV KELVIN
AV CLINTON
AV LENNOX
CH DEACON
EA 91-485
RU BOISE, DU
EA 91-510
AV EKERS
EA 91-472
EA 91-479
EA 91-481
AV KENT
AV WILDERTON
AV PLANTAGENET
AV SOISSONS, DE
AV BRIGHTON
AV DARLINGTON
AV STIRLING
AV WILLOWDALE
EA 91-508
EA 91-507
EA 91-509
EA 91-462
RU MCSHANE
AV DE CELLES
AV ELLENDALE
EA 91-480
EA 91-457
AV LACOMBE
AV FENDALL
AV LOUIS-COLIN
CH REMEMBRANCE
EA 91-506
EA 91-500
EA 91-505
EA 91-502
EA 91-501
AV OAKLAND
CL SUMMIT
RD BELVEDERE
AV TRAFALGAR
AV RAMSAY, DE
CH MCDOUGALL
CH COTE-DES-NEIGES, DE
AV GATINEAU
AV TROIE
EA 91-483
EA 91-466
AV CALYPSO
AV MONTPETIT
CH COTE-STE-CATHERINE
AV DUPUIS
BV EDOUARD
EA 91-488
EA 91-490
AV ISABELLA
RU JEAN-BRILLANT
AV ST-KEVIN
CH QUEEN MARY
EA 91-498
RU CORONET
EA 91-499
CR CEDAR
CH KINGSTON
AV DEVON
GN SURREY
CR SUMMIT
AV LEXINGTON
AV UPPER BELLEVUE
AV BELLEVUE
AV SUNNYSIDE
BOULEVARD, THE
AV MONTROSE
AV MOUNTAIN
AV HUDSON
CH HUDSON
AV APPLETON
AV KENT
AV LINTON
RU PELTRIE, DE LA
RU LEGARE
RU LAVOIE
AV BOURRET
PL CAMPDEN
EA 91-475
CH BEDFORD
EA 91-473
EA 91-477
EA 91-370
EA 91-361
EA 91-358
EA 91-359
EA 91-365
EA 91-425
EA 91-538
EA 91-539
EA 91-469
EA 91-470
EA 91-476
EA 91-514
8454
6626
624
15010
18933
14984
14937
13866
13415
18123
19380
18132
20433
22760
22840
20505
20512
20544
20567
12046
10548
10367
9309
8408
7627
7575
5764
3917
1095
1160

Complex Survey Design

Multi-stage design ≠ Simple random design



Complex Survey Design

- **Why such a design?**
  - **Better coverage of the entire region of interest (stratification)**
  - **Efficient for interviewing; less travel, less costly (clustering)**

  Problem: Estimation more complex

Estimation of the precision

- How does the complex sample design affect the precision of estimates?
  - Stratification decreases variability (more precise)
  - Clustering increases variability (less precise)
  - Overall, the multistage design has the effect of increasing variability (less precise than SRS)

Estimation of the precision

- **Variance estimation with complex multistage cluster sample design:**
  - **Exact formula for variance estimation is too complex; use of an approximate approach required**
  - **NOTE: taking account of the design in variance estimation is as crucial as using the sampling weights for the estimation of a statistic**

Estimation of the precision

- Approximate methods for variance estimation:
  - Taylor linearization
  - Replication methods:
    - Balanced Repeated Replication
    - Jackknife
    - Bootstrap

Estimation of the precision

- **Replication methods (Motivation):**
  - You can estimate the variance of an estimated parameter by using a large number of somewhat different subsamples from your original sample
    - **Each subsample, called a replicate, is used to estimate the parameter**
    - **The variability among the resulting estimates is used to estimate the variance of the full-sample estimate**
  - **The replication methods differ in the way the replicates are built**

Estimation of the precision

- **Principle of the replication methods:**
  - **You want to estimate how precise your estimate of the number of smokers in Canada is**
  - **Take a large number of subsamples from your sample**
  - **For each subsample, compute the estimate of the number of smokers**
  - **To estimate the variance, compute the variance among all subsamples estimates**

Estimation of the precision

- **Bootstrap Method:**
  - **Covered in Hands-on Bootstrap Workshop**
- **The result is a file that contains 500 bootstrap weights (representing the 500 bootstrap replicates)**
  - **Used to compute precision (variance, CV)**
  - **Bootvar program**
  - **Confidential info (Master & Share only)**
  - **NPHS Bootstrap weights are coordinated**

▶ Example on diabetes with Bootvar

GEOA_HR4	TYPE	VARIABLE	YHAT	BS_SD	BS_CV	CIL95	CIU95
2401	Total	DIAB	11197.41	1602.82	14.31	8055.88	14338.94
2401	Ratio	DIAB TOTAL	6.42	.92	14.31	4.62	8.22
2402	Total	DIAB	10456.54	3609.92	34.52	3381.10	17531.98
2402	Ratio	DIAB TOTAL	4.33	1.49	34.52	1.40	7.26
2403	Total	DIAB	26320.72	4847.27	18.42	16820.07	35821.37
2403	Ratio	DIAB TOTAL	4.73	.87	18.42	3.02	6.44
2404	Total	DIAB	16485.61	3504.46	21.26	9616.86	23354.36
2404	Ratio	DIAB TOTAL	4.04	.86	21.26	2.35	5.72
2405	Total	DIAB	10249.12	3116.63	30.41	4140.53	16357.71
2405	Ratio	DIAB TOTAL	4.19	1.27	30.41	1.69	6.69
2406	Total	DIAB	78298.47	11005.69	14.06	56727.31	99869.63
2406	Ratio	DIAB TOTAL	4.99	.70	14.06	3.61	6.36
2407	Total	DIAB	11739.11	2477.00	21.10	6884.20	16594.02
2407	Ratio	DIAB TOTAL	4.38	.92	21.10	2.57	6.20
2408	Total	DIAB	5205.24	926.24	17.79	3389.81	7020.67
2408	Ratio	DIAB TOTAL	4.20	.75	17.79	2.74	5.67
2409	Total	DIAB	3756.61	916.56	24.40	1960.16	5553.06
2409	Ratio	DIAB TOTAL	4.83	1.18	24.40	2.52	7.14
2410	Total	DIAB	290.46	111.48	38.38	71.95	508.97
2410	Ratio	DIAB TOTAL	2.02	.78	38.38	.50	3.54
2411	Total	DIAB	2357.23	703.32	29.84	978.73	3735.73
2411	Ratio	DIAB TOTAL	2.76	.82	29.84	1.15	4.38
2412	Total	DIAB	15965.09	2549.22	15.97	10968.62	20961.56
2412	Ratio	DIAB TOTAL	4.82	.77	15.97	3.31	6.33
2413	Total	DIAB	9386.64	2315.15	24.66	4848.95	13924.33
2413	Ratio	DIAB TOTAL	3.16	.78	24.66	1.63	4.69
2414	Total	DIAB	17490.30	2817.56	16.11	11967.89	23012.71
2414	Ratio	DIAB TOTAL	5.29	.85	16.11	3.62	6.96
2415	Total	DIAB	20752.31	4780.71	23.04	11382.11	30122.51
2415	Ratio	DIAB TOTAL	5.26	1.21	23.04	2.88	7.63
2416	Total	DIAB	55656.67	9335.06	16.77	37359.96	73953.38
2416	Ratio	DIAB TOTAL	5.00	.84	16.77	3.35	6.64

Estimation of the precision

- Use of sampling and bootstrap weights

NPHS Estimates for Diabetes - Canada STANDARD DEVIATIONS		
	% People	
	Estimate	Std Dev.
Unweighted	4.1	0.162
Weighted	3.5	0.151
Bootstrap weights	3.5	0.177
Source: 1998 Master Health file		

Estimation of the precision

- Alternative to Bootstrapping: CV tables
- What are they?
  - Approximate sampling variability tables
  - Produced for Canada (total & age groups), for each province, and Health Region (CCHS)
  - Useful only for totals & ratios of categorical variables

Canadian Community Health Survey - 2000/2001

Approximate Sampling Variability Tables for Région de Montréal-Centre (24906)

NUMERATOR OF PERCENTAGE ('000)	ESTIMATED PERCENTAGE												
	0.1%	1.0%	2.0%	5.0%	10.0%	15.0%	20.0%	25.0%	30.0%	35.0%	40.0%	50.0%	70.0%
1	91.1	90.7	90.2	88.8	86.5	84.0	81.5	78.9	76.2	73.5	70.6	64.4	49.9
2	********	64.1	63.8	62.8	61.1	59.4	57.6	55.8	53.9	52.0	49.9	45.6	35.3
3	********	52.4	52.1	51.3	49.9	48.5	47.1	45.6	44.0	42.4	40.8	37.2	28.8
4	********	45.3	45.1	44.4	43.2	42.0	40.8	39.5	38.1	36.7	35.3	32.2	25.0
5	********	40.6	40.3	39.7	38.7	37.6	36.5	35.3	34.1	32.9	31.6	28.8	22.3
. . . . .													
10	********	28.7	28.5	28.1	27.3	26.6	25.8	25.0	24.1	23.2	22.3	20.4	15.8
. . . . .													
15	********	23.4	23.3	22.9	22.3	21.7	21.0	20.4	19.7	19.0	18.2	16.6	12.9
. . . . .													
20	**********	20.2	19.9	19.3	18.8	18.2	17.6	17.0	16.4	15.8	14.4	11.2	
21	**********	19.7	19.4	18.9	18.3	17.8	17.2	16.6	16.0	15.4	14.1	10.9	
22	**********	19.2	18.9	18.4	17.9	17.4	16.8	16.3	15.7	15.1	13.7	10.6	
23	**********	18.8	18.5	18.0	17.5	17.0	16.5	15.9	15.3	14.7	13.4	10.4	
24	**********	18.4	18.1	17.6	17.2	16.6	16.1	15.6	15.0	14.4	13.2	10.2	
25	**********	18.0	17.8	17.3	16.8	16.3	15.8	15.2	14.7	14.1	12.9	10.0	
30	**********	16.5	16.2	15.8	15.3	14.9	14.4	13.9	13.4	12.9	11.8	9.1	
35	**********	15.0	14.6	14.2	13.8	13.3	12.9	12.4	11.9	10.9	8.4		
40	**********	14.0	13.7	13.3	12.9	12.5	12.1	11.6	11.2	10.2	7.9		
45	**********	13.2	12.9	12.5	12.2	11.8	11.4	11.0	10.5	9.6	7.4		
50	**********	12.6	12.2	11.9	11.5	11.2	10.8	10.4	10.0	9.1	7.1		
55	**********	12.0	11.7	11.3	11.0	10.6	10.3	9.9	9.5	8.7	6.7		
60	**********	11.5	11.2	10.8	10.5	10.2	9.8	9.5	9.1	8.3	6.4		
65	**********	11.0	10.7	10.4	10.1	9.8	9.5	9.1	8.8	8.0	6.2		
70	**********	10.6	10.3	10.0	9.7	9.4	9.1	8.8	8.4	7.7	6.0		
75	**********	10.3	10.0	9.7	9.4	9.1	8.8	8.5	8.2	7.4	5.8		
80	**********	9.7	9.4	9.1	8.8	8.5	8.2	7.9	7.2	5.6			
85	**********	9.4	9.1	8.8	8.6	8.3	8.0	7.7	7.0	5.4			
90	**********	9.1	8.9	8.6	8.3	8.0	7.7	7.4	6.8	5.3			
95	**********	8.9	8.6	8.4	8.1	7.8	7.5	7.2	6.6	5.1			
100	**********	8.6	8.4	8.2	7.9	7.6	7.3	7.1	6.4	5.0			
125	**********	7.7	7.5	7.3	7.1	6.8	6.6	6.4	5.8	4.5			

Estimation of the precision

▸ Sampling Variability Guidelines

Type of estimate	*CV*	*Guidelines*
Acceptable	0.0-16.5	General unrestricted release
Marginal	16.6-33.3	General unrestricted release but with warning cautioning users of the high sampling variablitity. Should be identified by letter M.
Unacceptable	> 33.3	No release. Should be flagged with letter U.

Estimation of the precision

- **Exercise: Estimate the CV for the proportion of people with diabetes**
  - **CV Table can be used to check if enough sample size**

Module 3

How to Handle Missing Data

Outline

- What is missing data
- Types of missing data in Statistics Canada surveys
- How to handle missing data
- Missing data with statistical software

Missing Data

- **Missing data is due to nonresponse to some or all questions**
  - **Examples of nonresponse:**
    - **Refusal**
    - **Don’t know**
    - **Skip patterns in the questionnaire**
    - **Death of a longitudinal respondent**

Nonresponse

- 2 types of nonresponse:
  - Total nonresponse (unit nonresponse)
  - Partial nonresponse (item nonresponse)

Nonresponse

- **Total nonresponse:**
  - **All the variables for the respondent are missing**
    - **Examples:**
      - **Complete refusal**
      - **Unable to contact the respondent**
      - **Respondent absent for the duration of the survey (no proxy allowed or available)**
      - **language barrier**

Nonresponse

- **Total nonresponse:**
  - **Accounted for by adjusting the sampling wgts**
    - **Done by methodologists at StatCan**
      - **Using adjustment classes to eliminate possible nonresponse bias**
    - **Users do not need to do anything to the data**

Nonresponse

- **Partial nonresponse:**
  - **Some variables for the respondent are missing**
    - **Examples:**
      - **Respondent refuses to answer specific questions**
      - **Respondent does not know all the answers**
      - **Data unavailable**
  - **There are different approaches on how to handle partial nonresponse**

Types of missing data in Statistics Canada surveys

- There are no “holes” in Statistics Canada data files. A specific value identifies the missing data.

Type of missing data	Value
Not applicable	6, 96, 996
Don’t know	7, 97, 997
Refusal	8, 98, 998
Not stated	9, 99, 999

  - Note: for NPHS (longitudinal) the deads are coded as “not stated”

How to handle missing data

- Depends on the type of missing data, the type of analysis and the number of missing data
  - 1: Keep the missing data, report them separately
  - 2: If missing at random: remove the records with missing data (reweight for the estimation of totals)
  - 3: If not missing at random: remove the records with missing data and reweight
  - 4: Imputation: replace each missing value by a replacement value)

How to handle missing data

- **Examples of imputation:**
  - **deterministic**
  - **by donor**
  - **by the mean**
  - **historical (when longitudinal)**

- **Analysis should be done to check the validity of the imputation (eg: imputation rate, etc.)**
- **Imputation needs to be reported with the results**

How to handle missing data

- **Examples: Number of nights as patient**

Respondent	Overnight patient	# of nights	# of nights
1	Yes	2	2
2	No	996	0
3	Yes	15	15
4	No	996	0
5	No	996	0
6	Yes	10	10

How to handle missing data

- **Examples:**

Respondent	Smoker	Sex	Weight
1	Yes	F	12,5
2	No	F	12,5
3	8	F	
4	No	F	12,5
5	No	F	12,5
6	8	M	
7	Yes	M	25
8	8	M	
9	8	M	
10	No	M	25

Possible problem:

Conclusions:

Nonresponse not at random

ales tend to refuse more than

ation:

males)

Possible solution:

smokers

non-smokers

Adjust the weights

DK or refuse to answer

female smokers: 25%

male smokers: 50%

In the population:

the population:

37.5% smokers

lation:

20% smokers

62.5% non-smokers

mokers

40% non-smokers

40% DK or refuse to answer

on-smokers

nonresponse was ignored)

33.3% smokers

66.6% non-smokers

(item nonresponse was ignored)

Missing data with statistical software

- If nothing is done, the software considers the missing values (6, 7, 8, 9, etc…) as real values
  - SAS: Users must recode the variables (with ‘ ’ or .)
  - SPSS: Users can recode the variables (with ‘ ’ or .) or use the “MISSING VALUES” statement
    - Note: For some procedures (eg: regression), the entire record is rejected when at least one variable of interest is missing.

Example

- **Objective: Create a categorical variable for the body mass index (HWTADBMI) and recode the missing values:**
  - **The categories are:**
    - **0-20 = 1 - low**
    - **20-30 = 2 - middle**
    - **30 et + = 3 - high**
  - **The missing values are:**
    - **999.6 = Not applicable** (Universe: Respondents aged 20 to 64 who answered MAMA_037 <> 1)
    - **999.9 = Not stated**

Exercise

- **Objective: Only for your health region, estimate the ratio and the number of people with a total household income less than $20,000...**
  - **1: … keeping the data as is (report the missing values)**
  - **2: … recoding 7, 8 and 9 to missing values (ignoring the missing values)**

    - **Variable: INCA_3A:**

      Total household income - < $20,000 or >= $20,000

      1- Less than $20,000
      2- $20,000 or more
      3- No income
      7- Don't know
      8- Refusal
      9- Not stated

Exercise

- For Quebec, keeping the missing values

N	Valid	623110
	Missing	0

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1054358	16.9	16.9	16.9
	2	4838040	77.6	77.6	94.6
	3	21365	.3	.3	94.9
	7	102719	1.6	1.6	96.6
	8	116589	1.9	1.9	98.4
	9	98035	1.6	1.6	100.0
Total		6231106	100.0	100.0	

Exercise

- For Quebec, recoding the missing values

N	Valid	5913762
	Missing	317342

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	1075723	17.3	18.2	18.2
	2	4838040	77.6	81.8	100.0
	Total	5913763	94.9	100.0	
	Missing	317343	5.1		
	Total	6231106	100.0		

Exercise

- **Ratio:**
  - **Reporting the missing values:**

    17.3% of the population have a hhld income < $20,000

    77.6% of the population have a hhld income >= $20,000

    5.1% of the population don't know or refuse to answer

  - **Ignoring the missing values**

    18.2% of the population have a hhld income < $20,000

    81.8% of the population have a hhld income >= $20,000

    Item nonresponse was ignored (assuming random nonresponse)

Exercise

- **Totals:**
  - **Reporting the missing values:**

    In the population, 1,075,723 people have a hhld income < $20,000

    In the population, 4,838,040 people have a hhld income >= $20,000

    In the population, 317,342 people don't know or refuse to answer

  - **Ignoring the missing values**

    6,231,106 **X 18.2% = 1,134,061**

    6,231,106 **X 81.8% = 5,097,044**

    In the population, 1,134,061 people have a hhld income < $20,000

    In the population, 5,097,044 people have a hhld income >= $20,000

    Item nonresponse was ignored (assuming random nonresponse)

Module 4

Overview of Some Types of Cross-Sectional Analysis

Outline

- Prevalence (estimation & testing of differences)
- Linear and logistic regression
- Age-sex standardisation (for population comparison)
- Warnings about the comparison of cross-sectional estimates over time

Cross-sectional analysis

- **Possible analysis:**
  - **Descriptive statistics (totals, ratios, means):**
    - **Prevalence: The number of instances of a given disease or other condition in a given population at a designated time**
  - **Study the relationship between two or more variables**
    - **linear regression models**
    - **logistic regression models**
    - **etc...**
    - **The type of analysis depends on the type of the variables of interest (qualitative vs quantitative)**

Simple analysis

- **Descriptive statistics: prevalence (total, mean, ratio)**

  Step 1: Identify the variable(s)

  Step 2: Recode the missing values (if desired)

  Step 3: Use the weights

  Step 4: Calculate the estimate(s)

  Step 5: Calculate the variance(s)

Simple analysis: Example

- **Step 1: Identify the variable**
  - **CCCA_91B: Has emphysema or chronic obstructive pulmonary disease (COPD)**
    - **(Applicable to the 30 and older only)**

Content	Code	Sample	Population
YES	1	118	57,603
NO	2	8,298	4,464,320
NOT APPL.	6	2,832	1,703,636
DON'T KNOW	7	2	2,285
NOT STATED	9	5	3,261
	TOTAL	11,255	6,231,106

- **Source: CCHS cycle 1.1 - dummy file**

Simple analysis: Example

- **Step 2: Recode the missing values**

	Sample
YES (1)	118
NO (2)	8,298
NOT APPL. (6)	2,832
DON'T KNOW (7)	2
NOT STATED (9)	5
TOTAL	11,255

Sample
118
8,298
.
.
.
8,416

Simple analysis: Example

- **Step 3: Use the weight**

	Sample
YES (1)	118
NO (2)	8,298
NOT APPL. (6)	2,832
DON'T KNOW (7)	2
NOT STATED (9)	5
TOTAL	11,255

Sample	Population
118	57,603
8,298	4,464,320
.	
.	
.	
8,416	4,521,923

Simple analysis: Example

- 269,841 people (30 and over) have emphysema or COPD (ignoring nonresponse) (assumption: 1.4% of the nonrespondents have emphysema or COPD)

▸ Step 4: Calculate the estimate(s)

➢ examples:

	Population
OUI (1)	57,603
NON (2)	4,464,320
SANS OBJET (6)	
NE SAIT PAS (7)	2,285
NON DÉCLARÉ (9)	3,261
TOTAL	4,521,923

1,3% of those 30 and over have emphysema or COPD (ignoring nonresponse) **(57,603 ÷ 4,521,923) X 100**

57,675 people (30 and over) have emphysema or COPD (ignoring nonresponse) (assumption: 1.3% of the nonrespondents have emphysema or COPD)

Simple analysis: Example

- **Step 5: Calculate the variance(s)**
  - **Obtain standard deviations, CVs and confidence intervals with the Bootvar program (provided with the data)**
  - **Exercise: Estimate the variance using the program:**

\\ciqss-s2\utilisateurs$\formation\ordinateurXX\Bootvarf_BPCO.sas

where XX is the computer number

Difference of ratios

- **Comparison of populations:**
  - **Interested in testing if the rate of males with emphysema or COPD is significantly different from the rate of females**

	MALES	FEMALES
YES	36,152	21,451
NO	2,152,818	2,311,502

- **Males: 1.65%** of those 30 and over have emphysema or COPD
- **Females: 0.92 %** of those 30 and over have emphysema or COPD

Difference of ratios

- **Statistical test: T-test**

Rate_m = estimated proportion of males having diabetes = 1.65%
Rate_f = estimated proportion of females having diabetes = 0.92%

Hypothesis test: H_0: Rate_m = Rate_f
H_1: Rate_m ≠ Rate_f

Statistic: $$Z = \frac{(\text{Rate\_m} - \text{Rate\_f})}{sd(\text{Rate\_m} - \text{Rate\_f})}$$

Test: At level ✂ = 0,05, we accept H_0 if | z | <= 1.96
We reject H_0 otherwise.

Results: $$Z = \frac{(1.65\% - 0.92\%)}{sd(\text{Rate\_m} - \text{Rate\_f})} = \frac{(1.65\% - 0.92\%)}{0.40} = 1.83$$

(can be obtained with Bootvar)

More complex analysis

- **Regression models**
  - **Regression analysis is a statistical method that utilizes the relationship between two or more variables so that one variable can be predicted from the other(s).**

    (Applied Linear Statistical Models, Neter & al.)
  - **Linear regression analysis consists of a collection of techniques used to explore relationships between variables.**

    (Applied Linear Regression, Weisberg)

More complex analysis

- **Linear regression model:**

$$Y = \text{intercept} + \alpha_1 * X_1 + \alpha_2 * X_2 + \ldots + \alpha_j * X_j + \varepsilon$$

  - Y is quantitative
  - X_i can be quantitative or qualitative (categorical) (Categorical variables need to be «dichotomised»)

More complex analysis

- **Example of a linear regression model:**
  - **BMI (body mass index) VS sex and age**

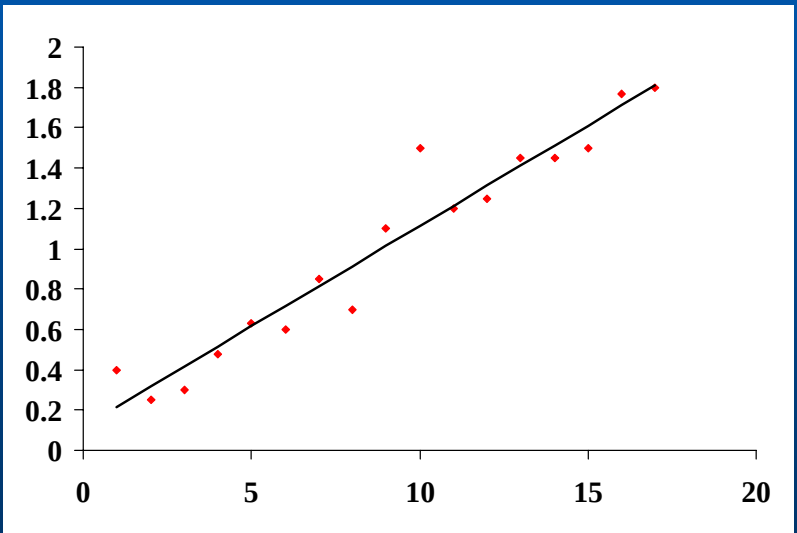
    BMI = intercept + β_1*FEMALE + β_2*AGE + ε

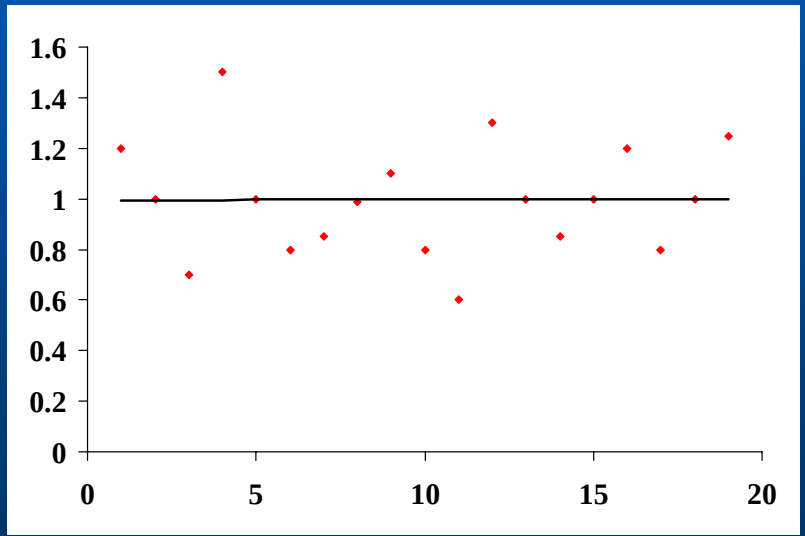
    - **Categorical variables need to be «dichotomised» (If SEX=female then FEMALE=1; else FEMALE=0)**
    - **AGE is a quantitative variable**

More complex analysis

- **Illustration of a simple linear regression model:** (1 predictor variable)

Model: $Y = \text{intercept} + \beta_1 * X_1 + \varepsilon$





The goal is to estimate if the slope ($\beta 1$) and the intercept are statistically significant

More complex analysis

- **Building a regression model:**
  - **1: Explore the data**
  - **2: Select predictor variables**
    - use prior knowledge and expertise
    - parsimony
    - study the interaction between the variables
  - **3: Preliminary models**
    - Verify that all underlying assumptions are respected
    - Diagnostics (transformation, outliers, etc.)
  - **4: Model refinement and selection of the final model**

More complex analysis: Example

- **Example: Fruit and vegetable consumption**

  (From Statistics Canada publication *Health reports, Vol 13 no.3*)

  - **Objective:**

    Analyse associations between the frequency of fruit and vegetable consumption and other health-related behaviours

  - **Data source:**

    First half of CCHS cycle 1.1 (collected from Sept. 2000 to Feb. 2001)

  - **Analytical technique:**

    Multivariate linear regression

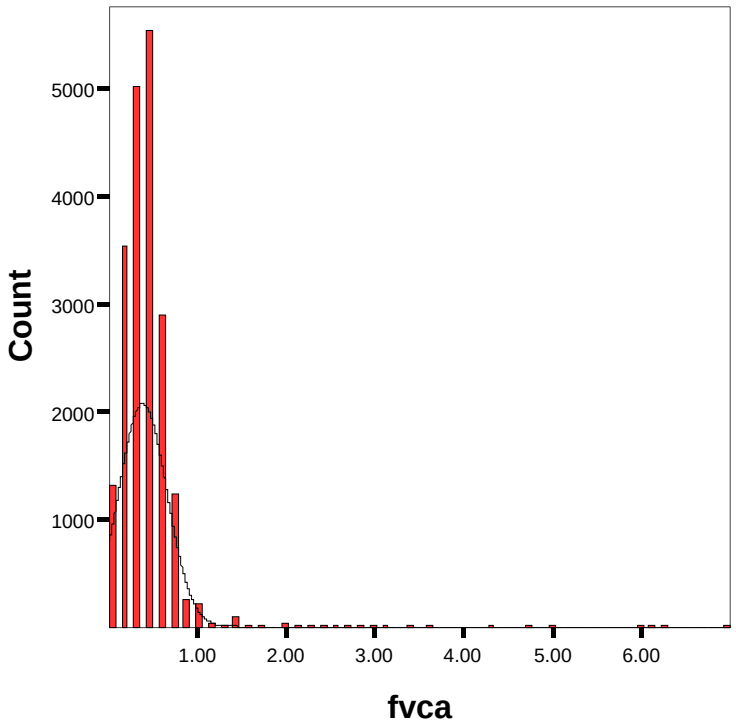
More complex analysis: Example

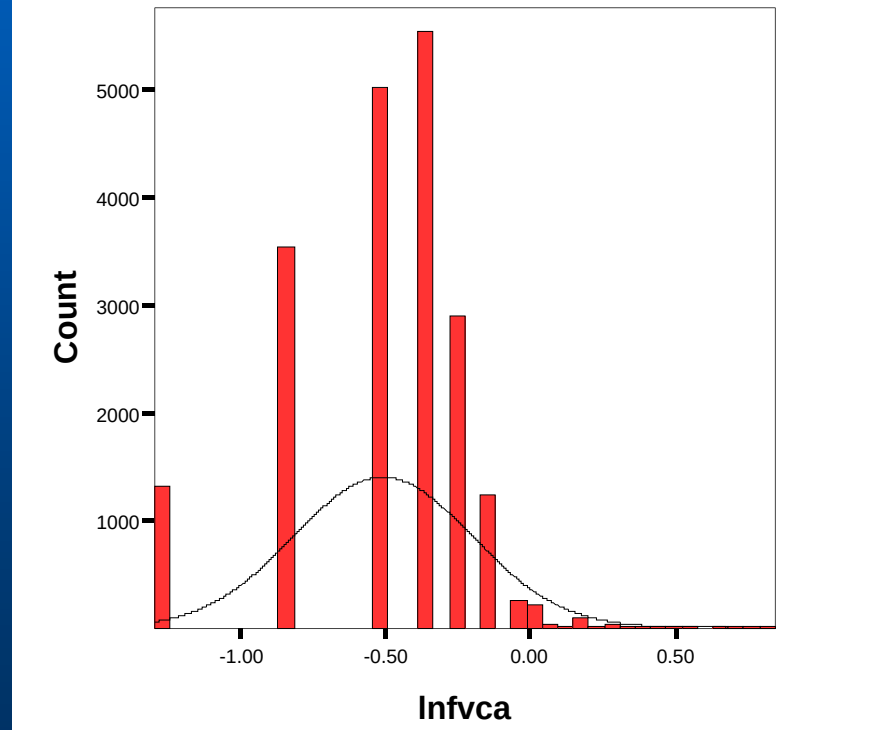
- **Preliminary**
  - **Response rate for first half of CCHS: 80%, total nonresponse adjusted by Statcan (weighting)**
  - **The missing values were not included in the analysis (ie: records with at least one missing value to the variables used in the model were excluded from the analysis)**
  - **Imputation rate: 7.6% (proxy interviews)**
  - **To account for the complex sample design of the survey, the bootstrap technique was used to estimate the variance for the regression coefficients**

More complex analysis: Example

- **Dependent variable:**
  - **Daily frequency of F&V consumption**
    - **Derived from the questions asking how many times specific F&V were consumed**
    - **The model was fitted to a log transformation of the dependent variable because the data were skewed to the right**

More complex analysis: Example





More complex analysis: Example

- **Predictor variables (independent variables):**
  - **Selected health-related behaviours**
    - Physical activity
    - Smoking status
    - Body mass index
    - Alcohol dependence
    - Chronic conditions
    - Disability
    - Self-perceived health
    - Age group
    - Household income
    - Education
    - Marital status

More complex analysis: Example

- **Multivariate linear regression results:**

Characteristics	Beta
Physical activity	0.05 *
Smoking status	
▪ Non-smoker	0.18 *
▪ Occasional	0.20 *
▪ Daily *	--
Body mass index	- 0.005 *
Alcohol dependent	
▪ No	0.08
▪ Yes *	--
No chronic conditions	-0.09 *

Characteristics	Beta
No disability	- 0.06
Self-perceived health	0.04 *
Age	0.01 *
Household income	0.02 *
Education	0.02 *
Marital status	
▪ Married/C-L	0.13 *
▪ Single	0.11 *
▪ Divorced *	--
Intercept	0.62

*** = p-value < 0.05**

More complex analysis

- **Logistic regression model:**
  - **This model is not linear, but can be linearized by a logarithmic transformation. The model becomes:**

$$\log(f(Y)) = \text{intercept} + \beta_1 * X_1 + \beta_2 * X_2 + \ldots + \beta_j * X_j + \varepsilon$$

  - **Y has to be qualitative (categorical) (often dichotomous)**
  - **X_i can be quantitative or qualitative variables**

More complex analysis

- **Example of a logistic regression model:**
  - **Diabetes vs sex and age**

DIAB = intercept + β_1*FEMALE + β_2*AGE + η

  - **DIAB is a dichotomous variable (0 or 1)**
  - **Categorical variables need to be «dichotomised» (If SEX=female then FEMALE=1; else FEMALE=0)**

Example of Bootvar ’s output

prov	beta	bhat	Odds	Wald	pvalue	bs_var	bs_sd	bs_cv	cil95	ciu95
10	Intercept	-4.00	0.01	255.38	0.0000	0.06	0.25	6.26	0.01	0.02
10	female	0.85	2.35	11.81	0.0005	0.06	0.24	29.09	1.44	3.84
10	age	0.46	1.59	2.85	0.0908	0.07	0.27	59.14	0.92	2.73

- Odds ratio: The ratio of the risk of a disease or the death among the exposed to the risk among the unexposed

More complex analysis: Example

- **Example: Food insecurity in Canadian households**

  (From Statistics Canada publication *Health reports, Vol 12 no.4*)

  - **Objective:**

    Examine the prevalence of food insecurity in Canada and the characteristics of people most likely to live in households lacking sufficient funds for food

    Data source:

    Cross-sectional hhld component of the 1998-99 NPHS

    Analytical technique:

    Multivariate logistic regression

More complex analysis: Example

- **Preliminary**
  - **Response rate for cycle 3 of NPHS: 88.2%, total nonresponse adjusted by Statcan (weighting)**
  - **The missing values were not included in the analysis (ie: records with at least one missing value to the variables used in the model were excluded from the analysis)**
  - **To account for the complex sample design of the survey, the bootstrap technique was used to estimate the variance for the odds ratio**

More complex analysis: Example

- Dependent variable: Any food insecurity (yes/no)
- Predictor variables (independent variables):
  - Sex
  - Age group
  - Household income
  - Major source of income
  - Household type
  - Home ownership
  - Marital status
  - Immigration status
  - Aboriginal status

More complex analysis: Example

Characteristics	Odds Ratio
Sex	
Males	1.06
Females *	1.00
Age group	
0 - 17	4.82 *
18 - 44	4.22 *
45 - 64	2.71 *
65 + *	1.00
Household income	
Low	7.96 *
Middle	4.31 *
Upper *	1.00
Major source of income	
Wage & salaries *	1.00
Worker's compensation	1.71 *
Social ass.	3.06 *
CPP, QPP, OAS	0.93
Other	1.02
Home ownership	
Owner*	1.00
Tenant	2.01 *

Characteristics	Odds Ratio
Household type	
Couple with child*	1.00
Couple without child	0.98
Lone mother	1.41 *
Lone father	1.02
Single	0.95
Other	0.99
Marital stauts	
Married *	1.00
With partner	1.06
Single	0.79 *
Widowed	1.04
Divorced	1.45 *
Immigration status	
Canadian born *	1.00
Immigrated 0-9 years	0.66 *
Immigrated 10+ years	1.05
Aboriginal status	
yes	1.48 *
no *	1.00

Standardisation

- **What?**
  - **Standardization is a technique used to compare sub-populations on a given variable, while controlling for the difference in the profile of each sub-population with respect to another variable**
- **Example:**
  - **Comparing asthma rates between different two regions that do not have the same age-sex profile (distribution)**

Standardisation

- **Solution?**
  - **Apply the same standard population distribution to all sub-populations compared**

Standardisation

- **Example:**

	REGION A		Std Pop	REGION B	
Age Grp	**Pop%**	**Rate%**	**Pop%**	**Pop%**	**Rate%**
<60	~~80%~~	40%	50%	~~20%~~	40%
>=60	~~20%~~	60%	50%	~~80%~~	60%
Overall rate:		~~44%~~			~~60%~~
		50%			50%

Standardisation

- **What is needed:**
  - **The (age-sex) distribution for a standard population (D1, D2, …, D10)**
  - **Rates computed for each (age-sex) group within each sub-population compared (R1, R2, …, R10)**
- **The math:**
  - **Std rate = (D1*R1)+(D2*R2)+….+(D10*R10)**

Standardisation

- Direct method (vs. indirect method)
- The actual rates produced do not mean anything on their own, but are directly comparable
- Variance estimation for these std rates would require modifications to the bootstrap weights

Warnings about the comparison of cross-sectional estimates over time

- **Aspects that may explain differences in the estimates obtained from two different cross-sectional surveys**
  - **Methodological aspects:**
    - **Target population**
    - **Questionnaire**
    - **Collection (period, response rate, proxy allowed, etc…)**
    - **Weighting**
    - **Imputation**
    - **Method to calculate the variance**
    - **Sample variability**
  - **Contextual aspects:**
    - **Changes in health standards**
    - **True changes in the population**

Module 6

Overview of Software Programs for Data Analysis

Outline

- Elements to check when using a software program for survey data analysis
- Variance estimation
- Remote access / Support

Elements to check

- Treatment of missing values (multivariate models)
- Proper use of survey weights (for estimation)
- Account for complex design in variance estimation

Variance estimation

- Not all software programs support complex survey designs methods
  - Methods supported vary from one software program to another
  - Bootstrap not yet “directly” supported (use Bootvar)

Complex Survey Design Variance Estimation for some Software Programs

	SAS	Stata	SUDAAN	WesVar	Bootvar
Approximative Methods Supported	Taylor	Taylor	Jackknife, BRR, Taylor, Bootstrap*	Jackkinfe, BRR, Bootstrap*	Bootstrap
Descriptive					
means	*surveymeans*	*svymean*	*descript*	yes	yes
totals	*surveymeans*	*svytotal*	*descript*	yes	yes
proportions	no	*svyprop*	*descript*	yes	yes
ratios	no	*svyratio*	*ratio*	yes	yes
crosstabulations	no	*svytab*	*crosstab*	yes	no
quantiles	no	no	*descript*	yes	no
Modelling					
linear regression	*surveyreg*	*svyreg*	*regress*	yes	yes
instrumental variable regression	no	*svyireg*	no	no	no
interval regression	no	*svyintrg*	no	no	no
logistic regression	no	*svylogit*	*logistic or rlogist*	yes	yes
probit regression	no	*svyprobt*	no	no	no
mutinomial logistic regression	no	*svymlog*	*multilog*	yes	no
ordered logistic regression	no	*svyolog*	*multilog*	no	no
ordered probit regression	no	*svyoprob*	no	no	no
poisson and log-linear regression	no	*svypois*	*loglink*	no	no
hierarchical linear models	no	no	no	no	no
proportional hazards models	no	no	*survival*	no	no

Remote access / Support

- Remote access for NPHS/CCHS supports mainly SAS & SPSS

Module 7

References

References

- Health Report
  - Available through STC web site
    - Our products and services
    - For sale
    - Health
    - Health Reports
- Litterature (books, papers)
  - See handout

References

- **Courses - Statistics Canada**
  - **Statistical Methods for the Analysis of Data — Intermediate Level (0428)**
    - **Standard probability distributions and tables**
    - **Point estimation**
    - **Confidence intervals**
    - **Tests of hypothesis**
    - **Simple linear regression**

References

- **Courses - Statistics Canada**
  - **Survival Analysis (0409)**
    - Accelerated failure time models
    - Weibull and Gompetz models
    - Types of censor data
    - Proportional risk models
    - Time-dependent variables
    - Analysis of discrete data
    - Analysis of sensitivity to censor data
    - Repetitive events
    - Left censor data and left truncation

References

- **Courses - Statistics Canada**
  - **Treatment of Nonresponse in Surveys and Censuses (0424)**
    - Definitions, levels of nonresponse, examples
    - Causes of nonresponse
    - Methods designed to reduce nonresponse
    - Reweighting
    - Imputation
    - Variance estimation in presence of imputation
    - Missing data analysis
    - Evaluation of nonresponse

References

- **Courses - Others**
  - **Summer Programme in Data Analysis (SPIDA, York University)**
    - two-week series of intensive courses designed to train social researchers to analyze complex survey data and especially longitudinal and multi-level data
    - www.math.yorku.ca/SCS/spida/home.html
    - Programme (2003):
      - Linear Models and Model Building Strategies; Logistic Regression; Generalized Linear Models; Nonparametric Regression Models and Generalized Additive Models; Advanced Modeling Topics; Intro to Mixed Models & Models for Hierarchical Data; Mixed Models for Longitudinal Data

References

- **Courses - Others**
  - **CIED / CÉETUM / CIQSS SUMMER DATA TRAINING SCHOOL**
    - http://www.ciqss.umontreal.ca/activities_training.htm
    - **Intensive Course on Longitudinal Data Analysis**
      - **Statistics Canada Longitudinal Surveys**
      - **Introduction to the management of data (Stata)**
      - **Longitudinal data description. Fundamentals of survival analysis**
      - **Introduction to multivariate modelling**
      - **Longitudinal analysis (event history analysis)**
      - **Complex plans, weighting, and robust variances**