PRASC

Project for the Regional Advancement of Statistics in the Caribbean

Projet régional pour l'avancement de la statistique dans les Caraïbes

Funded by the Government of Canada

Canada

Estimation

Project for the Regional Advancement of Statistics in the Caribbean (PRASC)

Gavin Thompson
April 20, 2021, Virtual, from Ottawa, Canada

Delivering insight through data for a better Canada

Overarching Processes

Specify needs	Design	Build	Collect	Process	Analyse	Disseminate	Evaluate
1.1 Identify needs	2.1 Design outputs	3.1 Reuse or build collection instruments	4.1 Create frame and select sample	5.1 Integrate data	6.1 Prepare draft outputs	7.1 Update output systems	8.1 Gather evaluation inputs
1.2 Consult and confirm needs	2.2 Design variable descriptions	3.2 Reuse or build processing and analysis components	4.2 Set up collection	5.2 Classify and code	6.2 Validate outputs	7.2 Produce dissemination products	8.2 Conduct evaluation
1.3 Establish output objectives	2.3 Design collection	3.3 Reuse or build dissemination components	4.3 Run collection	5.3 Review and validate	6.3 Interpret and explain outputs	7.3 Manage release of dissemination products	8.3 Agree an action plan
1.4 Identify concepts	2.4 Design frame and sample	3.4 Configure workflows	4.4 Finalise collection	5.4 Edit and impute	6.4 Apply disclosure control	7.4 Promote dissemination products	
1.5 Check data availability	2.5 Design processing and analysis	3.5 Test production systems		5.5 Derive new variables and units	6.5 Finalise outputs	7.5 Manage user support	
1.6 Prepare and submit business case	2.6 Design production systems and workflow	3.6 Test statistical business process		5.6 Calculate weights			
		3.7 Finalise production systems		5.7 Calculate aggregates			
				5.8 Finalise data files			

Estimation

- Extension of data collected from the sample to the population from which it was drawn

- Consists of:
  - Calculating weights
  - Using weights and sample data to calculate estimates

Weights

- Number of units in the survey population that each sampled unit represents
- Steps to calculate the weights:
  - Calculation of design weights (based upon sample design)
  - Weight adjustment for complete non-response
  - Weight adjustment using auxiliary information (if available)

Design Weight

- Is determined by the sample design (thus, can be calculated before collection).
- Is the inverse of the probability of selection.
- Example: For a Simple Random Sample (SRS), if N=100 and n=25, then the probability of selection is ¼ and the design weight is 4.

Design Weight for Equal Probability Sample Designs

- Design weights are the same for all units in the sample
- This occurs when each unit has the same probability of selection.
- Examples:
  - SRS
  - Systematic Samples
  - Stratified samples, in some situations (e.g. when an SRS is selected in each stratum with an N-proportional allocation)

Example – Design Weights for Equal Probability Sample Designs

Stratified SRS with N-proportional allocation

Stratum	Population Size (N)	Sample Size (n)	Probability of selection (p)	Design weight (1/p)
Men	400	80	80/400 = 0.2	5
Women	600	120	120/600 = 0.2	5
Total	1,000	200	---	---

Design Weight for Unequal Probability Sample Designs

- Design weights are not the same for all units in the sample
- This occurs when the units have unequal probabilities of selection.
- Examples:
  - Stratified samples with disproportional allocation
  - PPS samples

Example – Design Weights for Unequal Probability Sample Designs

Stratified SRS with disproportional allocation

Stratum	Population Size (N)	Sample Size (n)	Probability of selection (p)	Design weight (1/p)
Urban	1,000	200	200/1,000 = 0.2	5
Rural	100	50	50/100 = 0.5	2
Total	1,100	250	---	---

Weight Adjustment for Complete Non-response

- Why do we adjust for complete non-response? Because not compensating for non-responding units leads to the underestimation of totals.
- Based on the assumption that non-respondents are like respondents for the characteristics measured in the survey.
- This assumption is often more reasonable to make within subgroups of the population (e.g., local geographic area) rather than for the population as a whole.

Weight Adjustment for Complete Non-response

- To produce a weight adjusted for non-response, the design weight is multiplied by a non-response adjustment factor.
- This adjustment factor is calculated after collection.

Weight Adjustment for Complete Non-response

Non-response adjustment factor

$$= \frac{\text{Sum of design weights in the original sample}}{\text{Sum of the design weights of the responding units}}$$

For an Equal Probability Sample Design, this is equivalent to

$$= \frac{\text{Number of units in the original sample}}{\text{Number of responding units}}$$

Non-response adjusted weight

= Design weight × Non-response adjustment factor

Example: Weight Adjustment for Complete Non-response (for a SRS)

- Number of units in the population = N = 500
- Number of units in the original sample = n = 50

- Probability of selection = p = n/N = 1/10
- Design weight = 1/p = 10
- Suppose that the number of respondent units is 40

- Non-response adjusted weight
  = Design weight × Non-response adjustment factor
  = 10 × (50/40)
  = 12.5

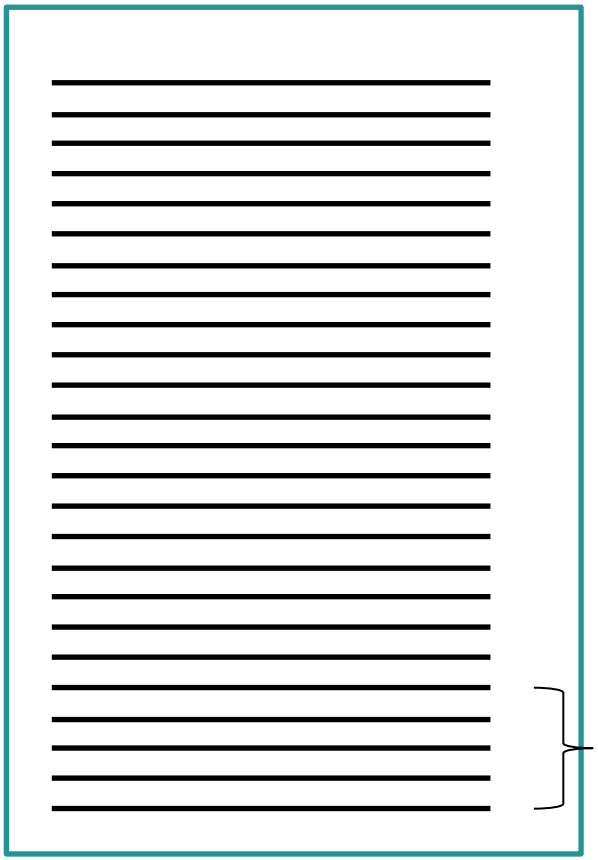
Example: Weight Adjustment for Complete Non-response with unequal design weights

- Sample of n=3 units, in a population of N=28.
- Design weights for the 3 units: 12, 10 and 6.
- Unit 3 refuses to participate.
- Sum of design weights in the original sample = 12 + 10 + 6 = 28
- Sum of design weights of the respondent units = 12 + 10 = 22

Non-response adjusted weights

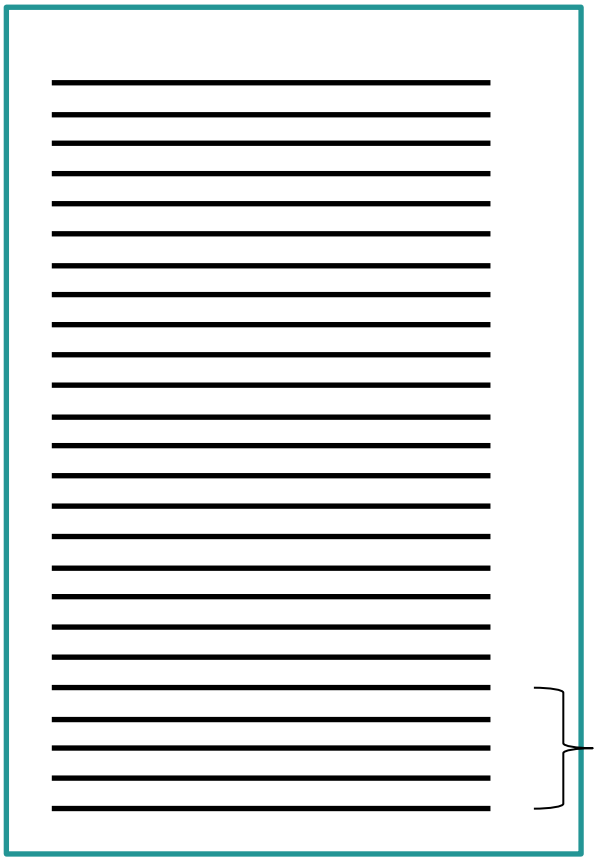
Responding Unit	Design Weight	Non-response adjustment factor	Non-response adjusted weight
1	12	= 28/22 = 1.2727	= 12 × 1.2727 =15.27
2	10	= 28/22 = 1.2727	= 10 × 1.2727 =12.73
Sum	22	---	28

A Quick Note on Out of Scope



- Suppose we have a sample 10,000 from a population of 100,000.
- Each record has a design weight of 10.
- If we are missing 500 records, due to non-response, we are missing 5,000 people in our estimation.
- Our totals would only sum to 95,000.

A Quick Note on Out of Scope



- Now suppose we have the same scenario, but the 5% are Out of Scope.
- We have found 1 in 20 of our sample is not what we expected.
- Having found this, we need to reduce our estimate of the population by 5% or 5,000.
- So not reweighting for Out of Scope gives us the correct, new estimate, of the population at 95,000.

Example: Weight Adjustment for Complete Non-response and Out of Scope (for a SRS)

- Number of units in the population = N = 500
- Number of units in the original sample = n = 50

- Probability of selection = p = n/N = 1/10
- Design weight = 1/p = 10
- Suppose that the number of respondent units is 40
- Out of Scope Units is 5

- Non-response adjusted weight
  = Design weight × Non-response adjustment factor
  = 10 × (45/40)
  = 11.25
- 11.25x40=450…our new estimate of the total population

Weight adjustment using auxiliary information

- It is often important for survey estimates to match known population totals or estimates from another, more reliable source (e.g. Census of population, administrative data, another survey with larger sample size)
- Use auxiliary data to improve the precision of the estimates.
- Requirements:
  - Accurate external sources of information concerning the population
  - Collection of corresponding information for all responding sample units
  - Auxiliary data must be well correlated with the survey variables
- The external data source must pertain to the same target population and must be based on comparable concepts, definitions, reference periods, etc.

Weight adjustment using auxiliary information - Methods commonly used

- Post-Stratification
  - Used to adjust the weights using variables that are suitable for stratification but which could not be used at the design stage (because the data were not available, or more up-to-date information became available after sample selection).
  - Used when the auxiliary data are in the form of counts.
  - Example: age-sex groups by province.

Weight adjustment using auxiliary information - Methods commonly used

- Ratio Estimation
  - When the data are in the form of counts, ratio estimation corresponds to post-stratification.
  - The weights of the records in a classification group are adjusted by a multiplicative factor (this factor is the ratio of the estimate from the auxiliary data to the survey estimate for the same variable, for the classification group).

- More complex methods (e.g. calibration, generalized regression)

Weight adjustment using auxiliary information

Post-stratification adjustment factor =

$$\frac{\text{Number of population units in the post-stratum}}{\text{Estimated number of population units in the post-stratum}}$$

Final weight =

Non-response adjusted weight × Post-stratification adjustment factor

*Note that we need to know what is an *estimate* to be able to calculate the post-stratification adjustment factor.

Overview of calculation of weights

- Design weight = 1 / probability of selection
- Non-response adjusted weight =
  Design weight × Non-response adjustment factor
- Final weight =
  Non-response adjusted weight × Post-stratification adjustment factor

Estimation

Notation:

- w_i = final weight of unit i
- y_i = value of variable of interest for unit i
- i = i^{th} responding unit in the sample

Estimation

Estimate of the total number of units

= sum of w_i over all responding units

Estimate of a total value

= sum of $(w_i \times y_i)$ over all responding units

Estimate of an average value

= estimate of the total value /

estimate of the total number of units

Estimate of the proportion of units

= sum of w_i over the units having a characteristic /

estimate of the total number of units

Estimation

Estimating for specific domains of the population

- Examples of domains: age groups, type of dwelling, income classes.
- The estimate formulas stay the same, except that we sum over responding units in the domain of interest.

Example: Estimation with non-response and post-stratification adjustments

Survey on smoking habits of the employees of a small company

At the design step, no auxiliary information was available that could be used for stratification

Sampling method: SRS

n = 25

N = 78

Number of respondent units =15

Probability of selection (p) = n/N = 25/78 = 0.32

Design weight (1/p) = 1/0.32 = 3.12

Sex of each respondent is collected

Responding Unit	Sex	Smoke	Design Weight (1/p)	Non-response adjustment factor	Non-response adjusted weight
1	F	No	=1/0.32 = 3.12	= 25/15 = 1.67	=3.12 x 1.67 = 5.21
2	F	Yes	3.12	1.67	5.21
3	M	Yes	3.12	1.67	5.21
4	F	Yes	3.12	1.67	5.21
5	M	No	3.12	1.67	5.21
6	F	No	3.12	1.67	5.21
7	F	Yes	3.12	1.67	5.21
8	F	Yes	3.12	1.67	5.21
9	F	No	3.12	1.67	5.21
10	F	No	3.12	1.67	5.21
11	F	Yes	3.12	1.67	5.21
12	M	No	3.12	1.67	5.21
13	F	No	3.12	1.67	5.21
14	F	Yes	3.12	1.67	5.21
15	F	Yes	3.12	1.67	5.21

Example (cont'd): Estimation with non-response and post-stratification adjustments

- Auxiliary information available after survey:
  - 42 men are working in the company
  - 36 women are working in the company

- Create post-strata on sex

Responding Unit	Sex	Smoke	Design Weight (1/p)	Non-response adjustment factor	Non-response adjusted weight	Post-Stratum
1	F	No	3.12	1.67	5.21	2
2	F	Yes	3.12	1.67	5.21	2
3	M	Yes	3.12	1.67	5.21	1
4	F	Yes	3.12	1.67	5.21	2
5	M	No	3.12	1.67	5.21	1
6	F	No	3.12	1.67	5.21	2
7	F	Yes	3.12	1.67	5.21	2
8	F	Yes	3.12	1.67	5.21	2
9	F	No	3.12	1.67	5.21	2
10	F	No	3.12	1.67	5.21	2
11	F	Yes	3.12	1.67	5.21	2
12	M	No	3.12	1.67	5.21	1
13	F	No	3.12	1.67	5.21	2
14	F	Yes	3.12	1.67	5.21	2
15	F	Yes	3.12	1.67	5.21	2

Example (cont'd): Estimation with non-response and post-stratification adjustments

- The information on the n=15 respondents is given in the table below.

Respondent Units	Men (Post-Stratum 1)	Women (Post-Stratum 2)	Total
Smokers	1	7	8
Non-Smokers	2	5	7
Total	3	12	15

Calculate estimates with non-response adjustment:

- Estimated number of *men* in the company

  = sum of non-response adjusted weights over responding men
  = 5.21 + 5.21 + 5.21
  = 5.21 × 3 =15.63 ≈ 16

- Estimated number of *women* in the company = 5.21 + 5.21 + …+ 5.21

  = 5.21 × 12 = 62.52 ≈ 63

Example (cont'd): Estimation with non-response and post-stratification adjustments

- Estimated number of *men who smoke* in the company = 5.21 × 1 = 5.21 ≈ 5
- Estimated number of *women who smoke* in the company = 5.21 × 7 = 36.47 ≈ 36
- Estimated proportion of *men who smoke* in the *company* = 5.21 / 15.63 = 0.33
- Estimated proportion of wo*men who smoke* in the company = 36.47 / 62.52 = 0.58
- The **estimates with non-response adjustment** are summarized in the table below.

	Men (Post-Stratum 1)	Women (Post-Stratum 2)	Total
Estimated Number of Smokers	5.21	36.47	41.68
Estimated Number of Non-Smokers	10.42	26.05	36.47
Estimated Number of Employees	15.63	62.52	78.15
Estimated Proportion of Smokers	*0.33*	*0.58*	*0.53*

Unit	Sex	Smoke	Non-response adjusted weight	Post-Stratum	Post-Stratification adjustment factor	Final Weight
1	F	No	5.21	2	=36/62.52=0.58	=5.21×0.58=3.02
2	F	Yes	5.21	2	0.58	3.02
3	M	Yes	5.21	1	=42/15.63=2.69	=5.21×2.69=14.01
4	F	Yes	5.21	2	0.58	3.02
5	M	No	5.21	1	2.69	14.01
6	F	No	5.21	2	0.58	3.02
7	F	Yes	5.21	2	0.58	3.02
8	F	Yes	5.21	2	0.58	3.02
9	F	No	5.21	2	0.58	3.02
10	F	No	5.21	2	0.58	3.02
11	F	Yes	5.21	2	0.58	3.02
12	M	No	5.21	1	2.69	14.01
13	F	No	5.21	2	0.58	3.02
14	F	Yes	5.21	2	0.58	3.02
15	F	Yes	5.21	2	0.58	3.02

Example (cont'd): Estimation with non-response and post-stratification adjustments

Calculate estimates with non-response and post-stratification adjustments:

- Estimated number of *men* in the company = 14.01 × 3 = 42.03 ≈ 42
- Estimated number of *women* in the company = 3.02 × 12 = 36.24 ≈ 36
- …

- The **final estimates with non-response AND post-stratification adjustments** are summarized in the table below.

	Men (Post-Stratum 1)	Women (Post-Stratum 2)	Total
Estimated Number of Smokers	14	21	35
Estimated Number of Non-Smokers	28	15	43
Estimated Number of Employees	42	36	78
Estimated Proportion of Smokers	0.33	0.58	0.45

Unadjusted Estimates

	Men (Post-Stratum 1)	Women (Post-Stratum 2)	Total
Estimated Number of Smokers	5.21	36.47	41.68
Estimated Number of Non-Smokers	10.42	26.05	36.47
Estimated Number of Employees	15.63	62.52	78.15
Estimated Proportion of Smokers	*0.33*	*0.58*	*0.53*

Adjusted Estimates

	Men (Post-Stratum 1)	Women (Post-Stratum 2)	Total
Estimated Number of Smokers	14	21	35
Estimated Number of Non-Smokers	28	15	43
Estimated Number of Employees	42	36	78
Estimated Proportion of Smokers	0.33	0.58	0.45

Outliers

- An observation or subset of observations which appears to be inconsistent with the remainder of the dataset.
- Can be detected by doing a graphic of the observations and/or measuring the distance between each observation and the center of the data
- What to do with outliers?

Outliers

- An observation or subset of observations which appears to be inconsistent with the remainder of the dataset.
- Can be detected by doing a graphic of the observations and/or measuring the distance between each observation and the center of the data
- What to do with outliers?
  - Leave them in: loss in precision.
  - Take them out: can bias the results.
  - Mitigation: If possible, use auxiliary information and post-stratification to ensure that they do not contribute an unreasonably large amount to the estimates.

Outlier Example

- Suppose we have a small town of 1,000 hhlds
- Average hhld income last year $30,000
- One family moves out, hhld income $30,000
- What effect does them moving out have on the average?

Outlier Example

- Suppose we have a small town of 1,000 hhlds
- Average hhld income last year $30,000
- One family moves out, hhld income $30,000
- What effect does them moving out have on the average?

The family moving out has no effect, since they make the average income.

Outlier Example – continued

- A new family moves in: hhld income $10M

- What is there effect on the average? (remember pop 1,000 hhlds)

Outlier Example – continued

- A new family moves in: hhld income $10M

- What is there effect on the average? (remember pop 1,000 hhlds)

Average hhld income goes up $10,000 ($10M/1,000 hhlds) to $40,000

(Analysts get excited because they found a community where the average income rose 33% in one year!)

Outlier Example - continued

- How does this hhld arriving affect the median income (assuming a bunch of hhlds make $30,000)?

Outlier Example - continued

- How does this hhld arriving affect the median income (assuming a bunch of hhlds make $40,000)?

It likely won't. If we move one family higher on the ordered list of hhld incomes it will likely be $30,000 (or very close)

Outlier Example – continued

- Average is a non-robust estimator – 1 outlier affects it.
- Median is a robust estimator – at a minimum half of your observations, plus 1, need to be an outlier before the median is affected.
- ($0, $20k, $30k, $30k, $30k, $50k, $250k)
- ($-10k, $20k, $30k, $30k, $100k, $2.5M, $10M)

Outlier Example - continued

- So which is true for the Analysts, did the community get much richer (as a whole) or did everything stay the same?
- Other estimators exist:
- Trimming: taking top (and bottom) # or % and removing
  - For example 1% of highest and lowest hhld incomes

Outlier Example - continued

- Winsorization: “capping” outlier.
  - Analysts, experts, economists etc might conclude that a hhld income above $1.5M has no differing effect on local economy.
  - So, $10M gets set to $1.5M
  - Median $30k
  - Winsorized average: $31.5k

Questions?

You can contact the PRASC team at:

statcan.prasc-prasc.statcan@canada.ca