RECENSEMENT · CENSUS

Using the 2021 Canadian Census Data Quality Indicators for Statistical Inference

February 14, 2023

Alexander Imbrogno
Emily Campling

RECENSEMENT · CENSUS

Outline

- Background on the 2021 Census of Population
- Non-response and the Total Non-response Rate
- Quality indicators per question
- Five-digit data quality numeric codes
- Confidence intervals for long-form estimates
- Using confidence intervals for statistical hypothesis testing

RECENSEMENT • CENSUS

The 2021 Census of Population

The 2021 Census of Population

RECENSEMENT · CENSUS

- Two main components.

Short-form questionnaire (2A):

- Used to enumerate usual residents in 75% of dwellings.
- The questions on the 2A are referred to as “short-form content”.

Long-form questionnaire (2A-L & 2A-R):

- Contains all the questions from the 2A plus additional questions known as “long-form content”.
- 2A-L form was used to enumerate usual residents in 25% of private dwellings.
  - The dwellings were selected using a systematic sampling design.
- 2A-R form was used to enumerate usual residents in 100% of private dwellings located in First Nations communities, Métis Settlements, Inuit regions and other remote areas.

- **Short-form questions** were given to **100%** of dwellings (a **census**).
- **Long-form questions** were given to **25%** of private dwellings (a **sample**).

RECENSEMENT • CENSUS

Non-Response

RECENSEMENT • CENSUS

Non-response (NR)

- There are **two types of non-response** in the 2021 Census of Population.
  - 1) Total non-response (TNR).
  - 2) Partial non-response.

- TNR occurs when:
  - 1) **All questions are unanswered** for a dwelling that received a questionnaire.
  - 2) Or a returned questionnaire does **not meet the minimum amount of content**.

- Partial non-response occurs when a dwelling **answered some** questions but left **others un-answered**.

- TNR & partial non-response can occur in the long-form or short-form content.

RECENSEMENT • CENSUS

Non-response (NR)

- Non-response is a **potential source of bias** in census counts and long-form estimates.

- Bias **occurs** when the characteristics of respondents **differ from those of non-respondents**.
  - Ex. Respondents to telephone surveys tend to be older than non-respondents.
    - age can be related to various characteristics of interest: income, education, work experience etc.

- In general, bias cannot be directly measured as the characteristics of non-respondents are unknown.

Total Non-Response Rate

RECENSEMENT · CENSUS

Total non-response rate

- The TNR rate indicates the risk, and it's potential magnitude, that a **significant bias** may be introduced by **total non-response**.

- Is the **primary quality indicator** that accompanies each **disseminated output** from the 2021 Census.

- **Replaced** the Global Non-Response (GNR) rate used in 2016 and previous Census cycles.
  - GNR rate accounted for both partial & total non-response -> TNR rate accounts for only total non-response.
  - TNR rates have been the primary indicator for many other surveys at Statistics Canada.
    - Replacing GNR with TNR -> standardizing data quality information.

RECENSEMENT · CENSUS

Total non-response rate

- Calculated at the **household level**.

- Two definitions of the TNR rate:
  - **Short-form rate**: computed for the **short-form content** (all households in the population).
  - **Long-form rate**: computed for the **long-form content** (only households in the sample).

- Calculated for a population of interest (POI).
  - POI corresponds to long-form or short-form households in a given geography.

Short-form TNR rate

For a given POI, the short-form (SF) TNR rate is calculated as:

$$\textbf{SF TNR} = \frac{\#\ of\ non\text{-}responding\ households\ in\ the\ POI}{\#\ of\ households\ in\ the\ POI}\ .$$

Total non-response rate

Long-form TNR rate

For a given POI, the long-form (LF) TNR rate is computed as:

$$\textbf{LF TNR} = \frac{\textit{design weighted \# of non-responding households in the POI}}{\textit{design weighted \# of households in the POI}}.$$

RECENSEMENT · CENSUS

Total non-response rate

Census Profile, 2021 Census of Population

More information: Quebec [Province]

Map | Data quality | Geographic hierarchy | Related data

- Excludes census data for one or more incompletely enumerated reserves or settlements.
- Total non-response (TNR) rate, short-form census questionnaire: 2.9%
- Total non-response (TNR) rate, long-form census questionnaire: 3.7%

Census Profile, 2021 Census of Population (statcan.gc.ca)

RECENSEMENT • CENSUS

Recommendation

Short-form data in areas with a short-form TNR above 50% should be used with caution.

Long-form data in areas with a long-form TNR above 50% should be used with caution.

RECENSEMENT • CENSUS

Quality Indicators per Question

RECENSEMENT · CENSUS

Quality indicators per question

- Data quality measures **specific to each question**.
- Quantify two related sources of error in the data: **non-response and imputation**.
- Questions can be at either the **household** or **person** level, depending on the characteristic of interest.
  - Household income (household) vs highest level of education (person).
- Calculated over the set of units in a POI which are **“in scope” to the question.**
  - If the question is not applicable to the person/ household, then they are excluded from the calculation -> “out of scope”.

RECENSEMENT · CENSUS

Quality indicators per question

- Data tables containing quality indicators per question are disseminated for:
  - Most of the questions asked on the short-form and for key income variables (long-form).
  - Many standard geographic areas.

- The set of tables disseminated are for variables and geographies that are likely to be of interest to analysts and users.
  - More complex tables can be produced by custom request.

Data quality tables, 2021 Census of Population (statcan.gc.ca)

RECENSEMENT · CENSUS

Quality indicators per question: short-form rates

- The available short-form rates are:

Non-response rate

$$\text{SF NR} = \frac{\#\ \textit{inscope units in the POI who did not respond to the question}}{\#\ \textit{inscope units in the POI}}.$$

Imputation rate

$$\text{SF IMP} = \frac{\#\ \textit{inscope units in the POI whose response to the question was imputed}}{\#\ \textit{inscope units in the POI}}.$$

Quality indicators per question:

short-form rates

Short-form data quality indicators (24)		Sex at birth - Non-response rate	Sex at birth - Imputation rate	Gender - Non-response rate	Gender - Imputation rate	Age - Non-response rate	Age - Imputation rate
Geography	Short-form total non-response rate						
Canada (map)	3.1	3.8	3.5	4.0	3.9	3.3	3.7
Newfoundland and Labrador (map)	3.1	3.5	3.4	3.7	3.8	3.1	3.6
Division No. 1 (map)	3.0	3.1	3.1	3.4	3.6	2.9	3.3
Division No. 1, Subd. V (map)	10.3	9.1	9.1	9.1	9.1	9.1	9.1
Portugal Cove South (map)	4.3	5.9	5.9	5.9	5.9	5.9	11.8

RECENSEMENT · CENSUS

Quality indicators per question: long-form rates

- The available long-form rates are:

Non-response rate

$$\text{LF NR} = \frac{sum\ of\ final\ weights\ of\ in{-}scope\ units\ in\ the\ POI\ who\ did\ not\ respond\ to\ the\ question}{sum\ of\ final\ weights\ of\ in{-}scope\ units\ in\ the\ POI}.$$

Imputation rate

$$\text{LF IMP} = \frac{sum\ of\ final\ weights\ of\ in{-}scope\ units\ in\ the\ POI\ whose\ response\ to\ the\ question\ was\ imputed}{sum\ of\ final\ weights\ of\ in{-}scope\ units\ in\ the\ POI}.$$

Quality indicators per question: long-form rates

Impact of imputation rate

- Only available for income variables.

$$IMPACT = The\ proportion\ of\ the\ total\ of\ y\ values\ which\ have\ been\ imputed$$

$$= \frac{the\ weighted\ total\ of\ y\ which\ has\ been\ imputed\ for\ in-scope\ units\ in\ the\ POI}{the\ weighted\ total\ of\ y\ for\ in-scope\ units\ in\ the\ POI}.$$

Where y is a variable from an income question.

Quality indicators per question: long-form rates

Impact of imputation rate: example

RECENSEMENT · CENSUS

- Let y_i be a variable of interest corresponding to an income question for unit i in the sample (Ex. total income).
  - $y_i = total\ income_i = \ income\ source\ 1_i + \ income\ source\ 2_i$.

$$\boldsymbol{IMPACT = The\ proportion\ of\ the\ total\ of\ y\ values\ which\ has\ been\ imputed.}$$

$$= \frac{the\ weighted\ total\ of y_i * z_i\ for\ in-scope\ units\ in\ the\ POI}{the\ weighted\ total\ of y_i\ for\ in-scope\ units\ in\ the\ POI}$$

- z_i is a unit level variable which takes values between 0 and 1.
  - Represents the proportion of y_i which has been imputed.
  - Ex. If $income\ source\ 1_i$ was imputed and $income\ source\ 2_i$ wasn't, then z_i = 0.5.
    - i.e. Half of z_i was imputed.

RECENSEMENT • CENSUS

Interpretation of Non-Response & Imputation Rates

Interpretation of non-response rates

RECENSEMENT • CENSUS

- The TNR rate and the non-response rates per question **indicate the risk, and it's potential magnitude**, that a **significant bias** may be introduced by **non-response.**

- In general, a **lower non-response rate** indicates a **lower risk of non-response bias** and, therefore, **more reliable figures and estimates**.

- **Both** the TNR rate and applicable non-response rate(s) per question should be consulted as they may offer **different perspectives on data quality**.
  - Ex. a given region has a low TNR rate but a high non-response rate for labour questions.

- When comparing the TNR rate and the non-response rate per question, users should be aware of **differences in their definition**.

RECENSEMENT · CENSUS

Interpretation of imputation rates

- The imputation rate indicates whether the **quantity of imputed values is large relative to the quantity of reported values**.
  - The impact of imputation can be interpreted similarly by replacing the “quantity” by the “sum”.

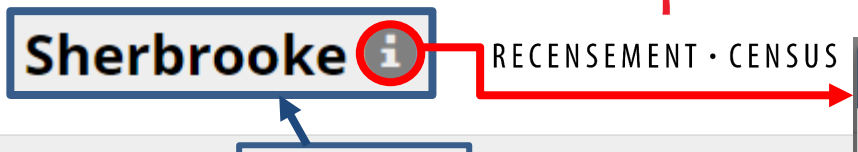
- Generally, the higher these rates, the more reason there is to question the quality of the estimates and the potential for bias.

- However, the rates themselves **do not indicate the level of quality of the imputed data**.
  - Imputation models for the 2021 Census are based on the use of auxiliary information well correlated with the characteristic of interest -> accurate imputed values.

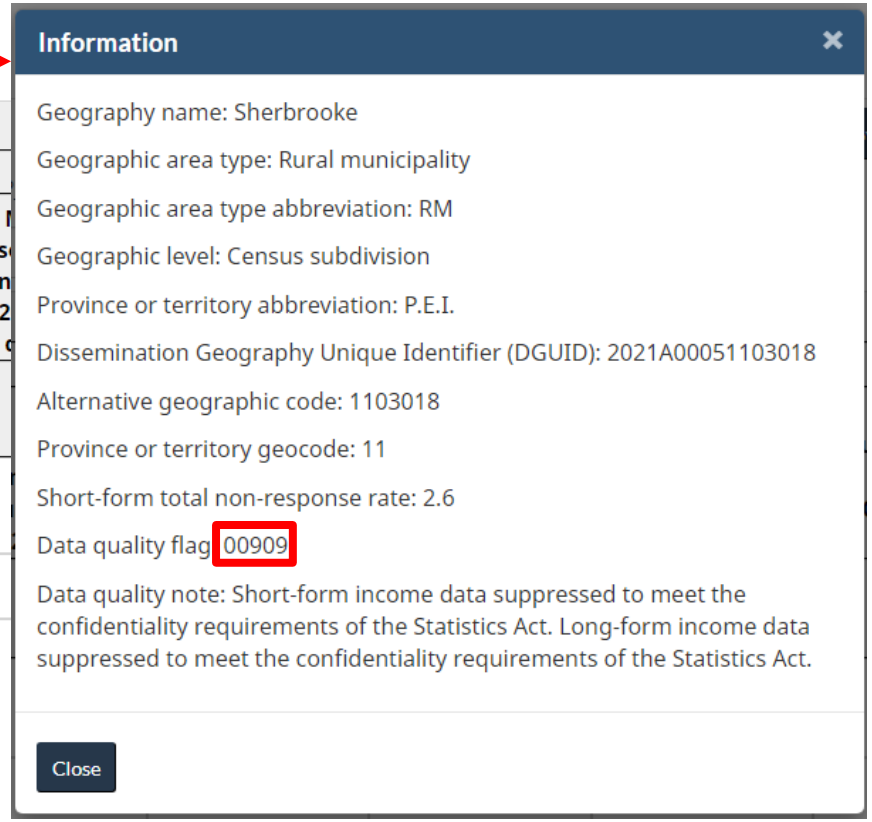
RECENSEMENT • CENSUS

Five-digit data quality flags

Five-digit data quality flags

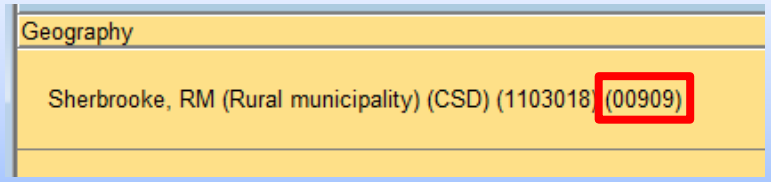


Geography	Sherbrooke (map)					
Household size (7)[2]	Total - Households by household size					
Household income statistics (6)	Number of households (2021)[3, 4, 5]	Number of households (2016)	Median household total income (2020) (2020 constant dollars)	Median household total income (2015) (2020 constant dollars)	Median household after-tax income (2020) (2020 constant dollars)	hous tax in (202
Household type including census family structure (11)						
Total – Household type including census family structure[6]	x	x	x	x	x	
Census-family households	x	x	x	x	x	

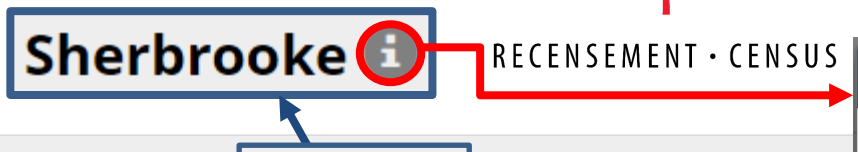


Geography: Sherbrooke, RM (Rural municipality) (CSD)
Data quality flag: **00909**

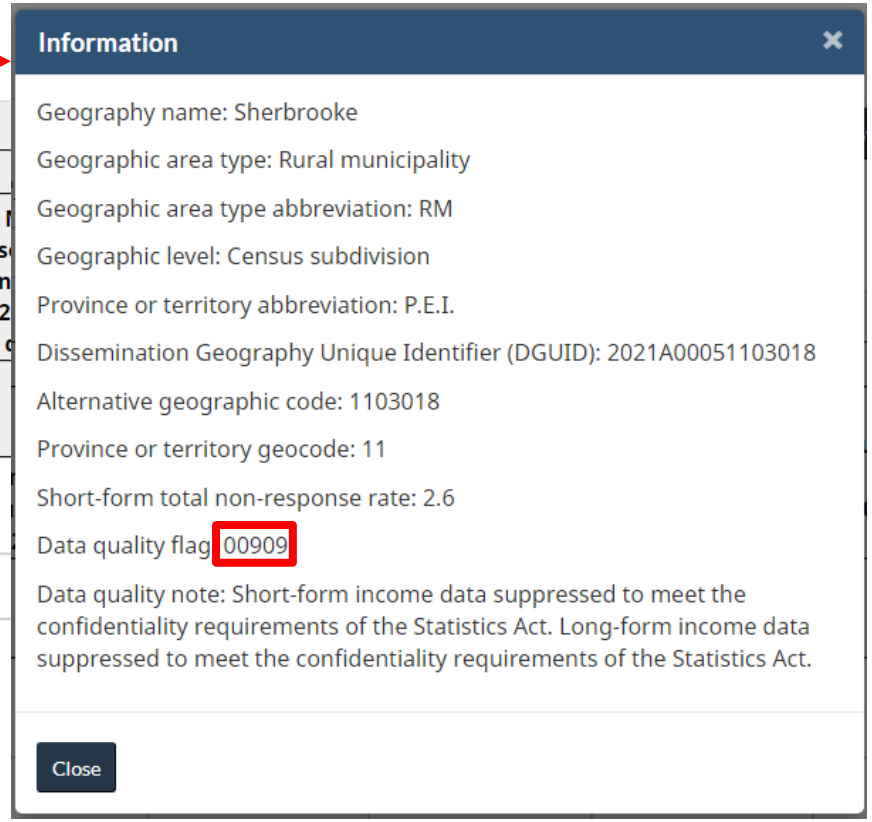
Beyond 20/20



Five-digit data quality flags



Geography	Sherbrooke (map)					
Household size (7)[2]	Total - Households by household size					
Household income statistics (6)	Number of households (2021)[3, 4, 5]	Number of households (2016)	Median household total income (2020) (2020 constant dollars)	Median household total income (2015) (2020 constant dollars)	Median household after-tax income (2020) (2020 constant dollars)	hous tax in (202
Household type including census family structure (11)						
Total – Household type including census family structure[6]	x	x	x	x	x	
Census-family households	x	x	x	x	x	

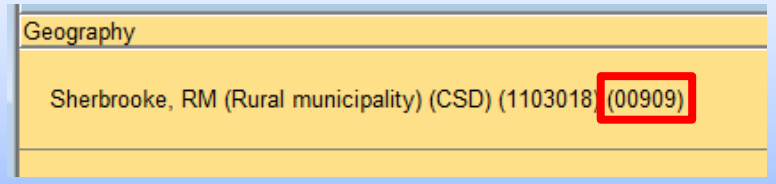


Geography: Sherbrooke, RM (Rural municipality) (CSD)
Data quality flag: **0**0909

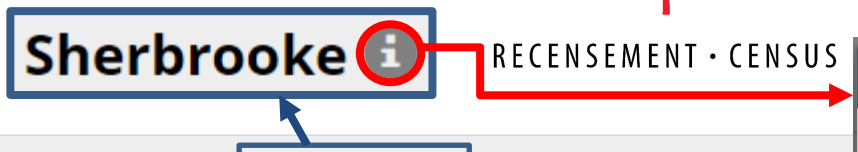
Incomplete enumeration indicator

- **Default value of zero**
- **Area is not an incompletely enumerated reserve or settlement**

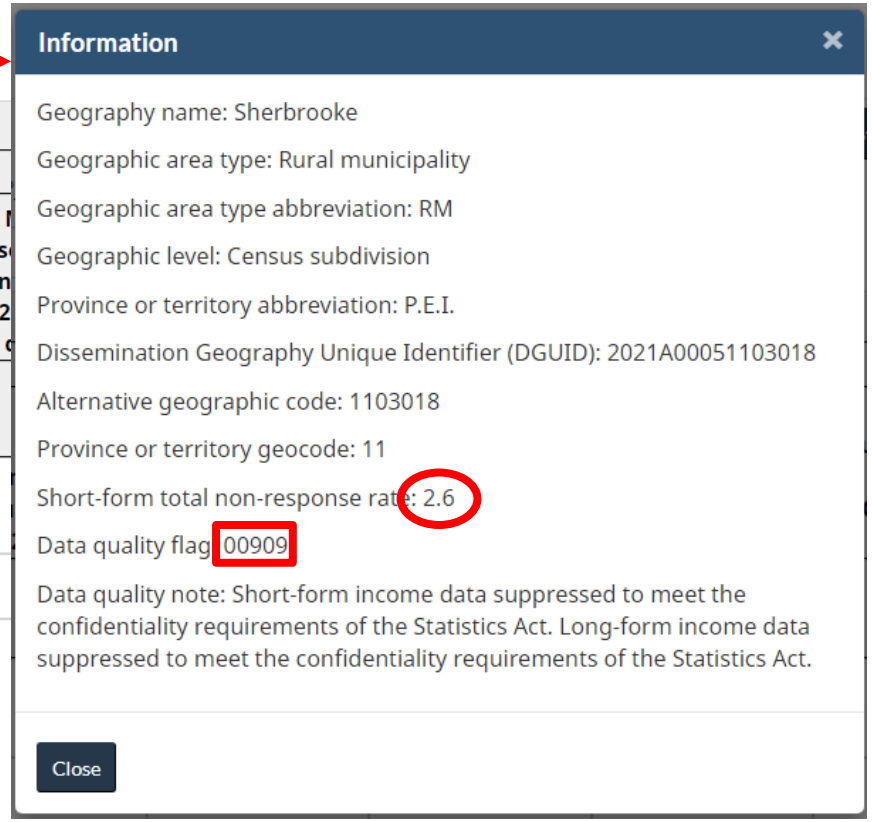
Beyond 20/20



Five-digit data quality flags



Geography	Sherbrooke (map)					
Household size (7)[2]	Total - Households by household size					
Household income statistics (6)	Number of households (2021)[3, 4, 5]	Number of households (2016)	Median household total income (2020) (2020 constant dollars)	Median household total income (2015) (2020 constant dollars)	Median household after-tax income (2020) (2020 constant dollars)	hous tax in (202
Household type including census family structure (11)						
Total – Household type including census family structure[6]	x	x	x	x	x	
Census-family households	x	x	x	x	x	

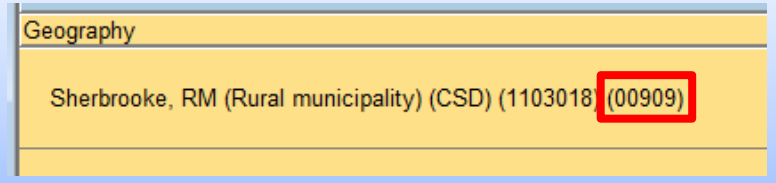


Geography: Sherbrooke, RM (Rural municipality) (CSD)
Data quality flag: **00909**

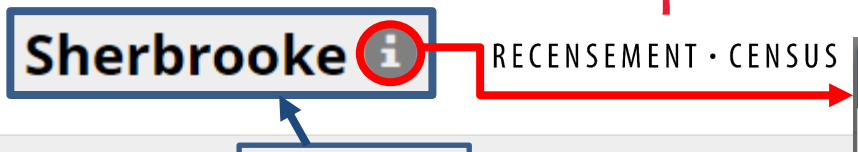
Short-form TNR rate indicator

- **Default value of zero**
- **The short-form TNR rate is less than 10%**

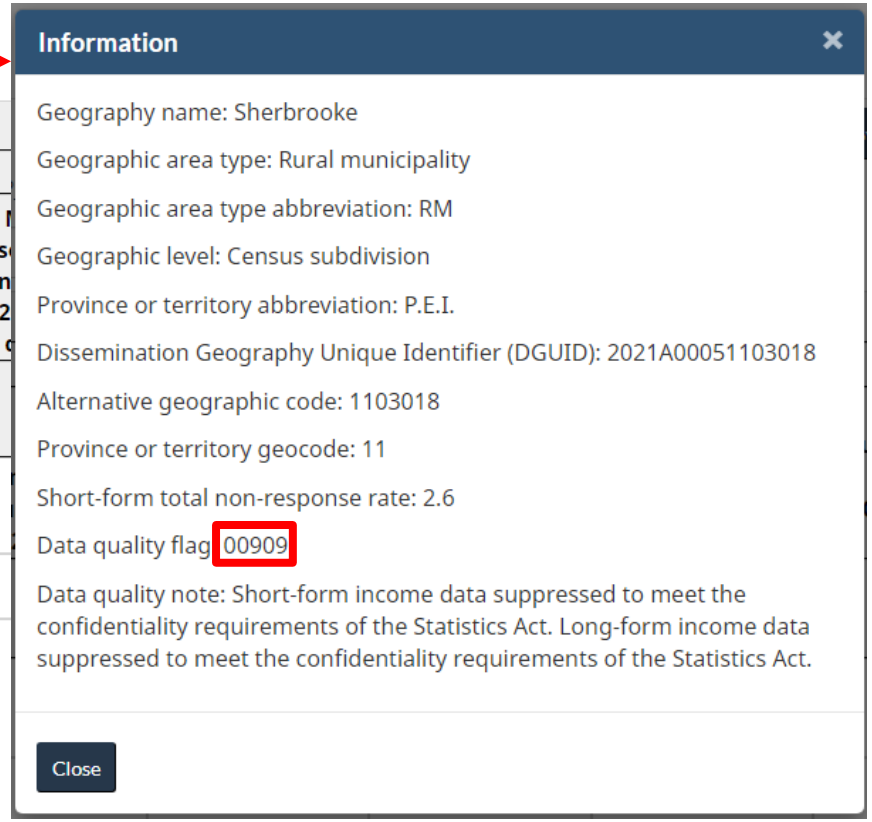
Beyond 20/20



Five-digit data quality flags



Geography	Sherbrooke (map)					
Household size (7)[2]	Total - Households by household size					
Household income statistics (6)	Number of households (2021)[3, 4, 5]	Number of households (2016)	Median household total income (2020) (2020 constant dollars)	Median household total income (2015) (2020 constant dollars)	Median household after-tax income (2020) (2020 constant dollars)	hous tax in (202
Household type including census family structure (11)						
Total – Household type including census family structure[6]	x	x	x	x	x	
Census-family households	x	x	x	x	x	

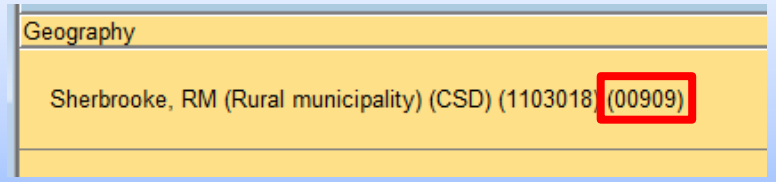


Geography: Sherbrooke, RM (Rural municipality) (CSD)
Data quality flag: **00909**

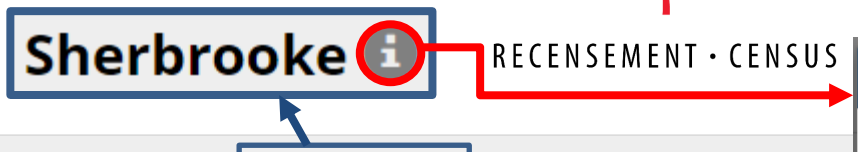
Long-form TNR rate indicator

- **Default value of zero**
- **The long-form TNR rate is less than 10%**

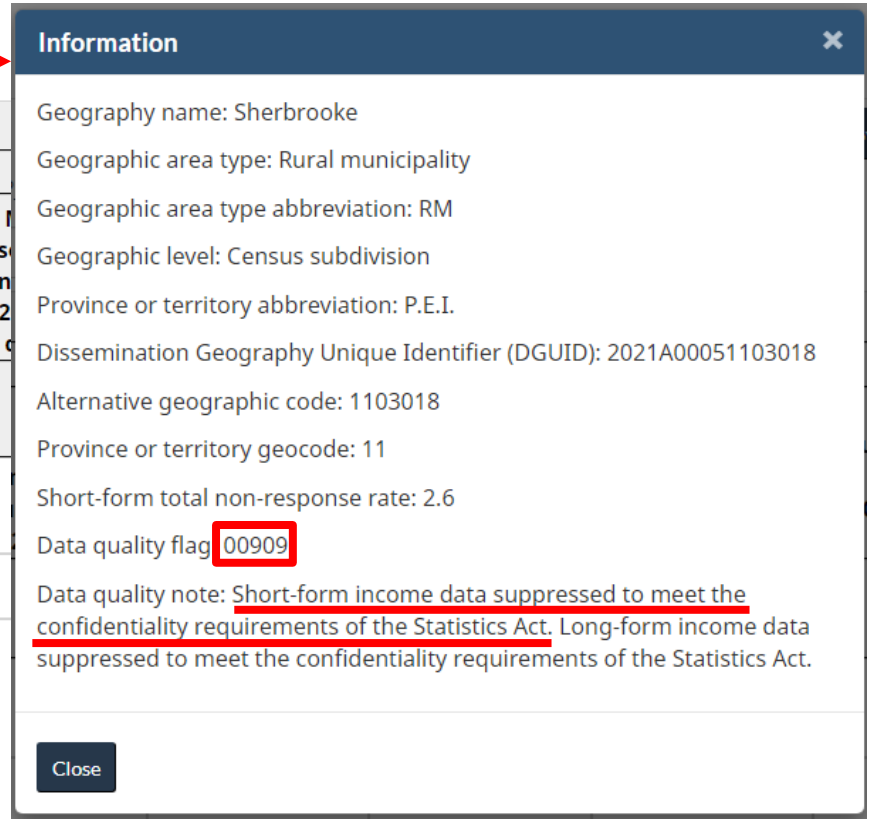
Beyond 20/20



Five-digit data quality flags



Geography	Sherbrooke (map)					
Household size (7)[2]	Total - Households by household size					
Household income statistics (6)	Number of households (2021)[3, 4, 5]	Number of households (2016)	Median household total income (2020) (2020 constant dollars)	Median household total income (2015) (2020 constant dollars)	Median household after-tax income (2020) (2020 constant dollars)	hous tax in (202
Household type including census family structure (11)						
Total – Household type including census family structure[6]	x	x	x	x	x	
Census-family households	x	x	x	x	x	

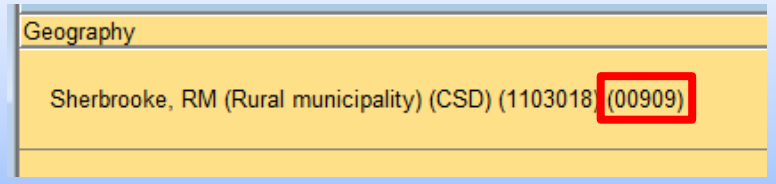


Geography: Sherbrooke, RM (Rural municipality) (CSD)
Data quality flag: **00909**

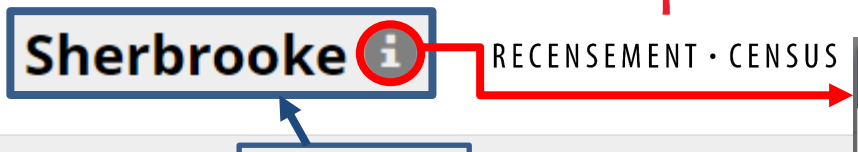
Short-form income suppression indicator

- **Value of 9 indicates short-form income data have been suppressed to meet confidentiality requirements**

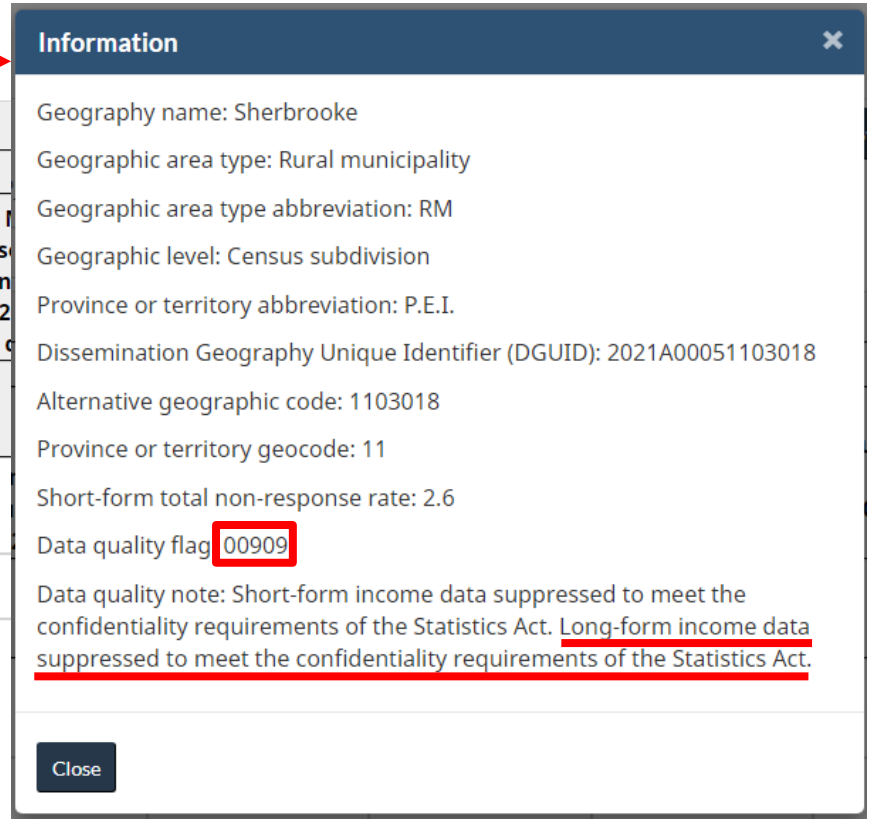
Beyond 20/20



Five-digit data quality flag



Geography	Sherbrooke (map)					
Household size (7)[2]	Total - Households by household size					
Household income statistics (6)	Number of households (2021)[3, 4, 5]	Number of households (2016)	Median household total income (2020) (2020 constant dollars)	Median household total income (2015) (2020 constant dollars)	Median household after-tax income (2020) (2020 constant dollars)	hous tax in (202
Household type including census family structure (11)						
Total – Household type including census family structure[6]	x	x	x	x	x	
Census-family households	x	x	x	x	x	

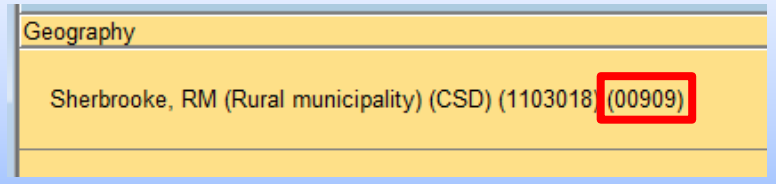


Geography: Sherbrooke, RM (Rural municipality) (CSD)
Data quality flag: **00909**

Long-form income suppression indicator

- **Value of 9 indicates long-form income data have been suppressed to meet confidentiality requirements**

Beyond 20/20



RECENSEMENT · CENSUS

Five-digit data quality flags

- Tables describing the interpretation of values for each of the five digits are available in the 2021 Census Data Quality Guidelines (statcan.gc.ca)

To keep in mind:

- A zero in any of the five digits is the default value.
- A value of 9 for the TNR rate indicators (2nd and 4th digit) or for the income suppression indicators (3rd and 5th digit) indicates data suppression to meet the confidentiality requirements of the *Statistics Act*.
- A value of 5 for the TNR rate indicators indicates a short-form or long-form TNR rate of over 50% -> **Data should be used with caution.**
- Unlike in previous census cycles, there is no longer suppression based on data quality.

RECENSEMENT • CENSUS

Confidence intervals

RECENSEMENT • CENSUS

Variance-based quality indicators for long-form estimates

Variance:

- Measure of uncertainty of an estimate produced from a sample
- Variance is estimated for long-form sample estimates using a replication method
- Meaning of long-form variance estimates depends on the area:
  - In 25% sampled areas, variance estimates measure variance due to sampling and total non-response.
  - In 100% sampled areas, variance estimates measure variance due to total non-response.
- Variance estimates are difficult to interpret. Therefore they are usually used to construct other quality indicators, e.g. standard errors, coefficients of variation, or confidence intervals.

RECENSEMENT · CENSUS

Confidence intervals

- **New** for the 2021 Census: **confidence intervals** are available as variance-based quality indicator for long-form estimates
- The advantage of confidence intervals over other variance-based quality indicators (e.g. coefficients of variation or standard error) is that they allow data users to easily make correct statistical inferences
- Confidence intervals are usually available in data tables with long-form estimates accessible through the Statistics Canada website.
- For the following types of tables, due to technical limitations, they are only available by custom request:
  - disaggregated data tables with a very large number of cells
  - multi-cycle tables
  - tables including both long-form and short-form estimates.

RECENSEMENT · CENSUS

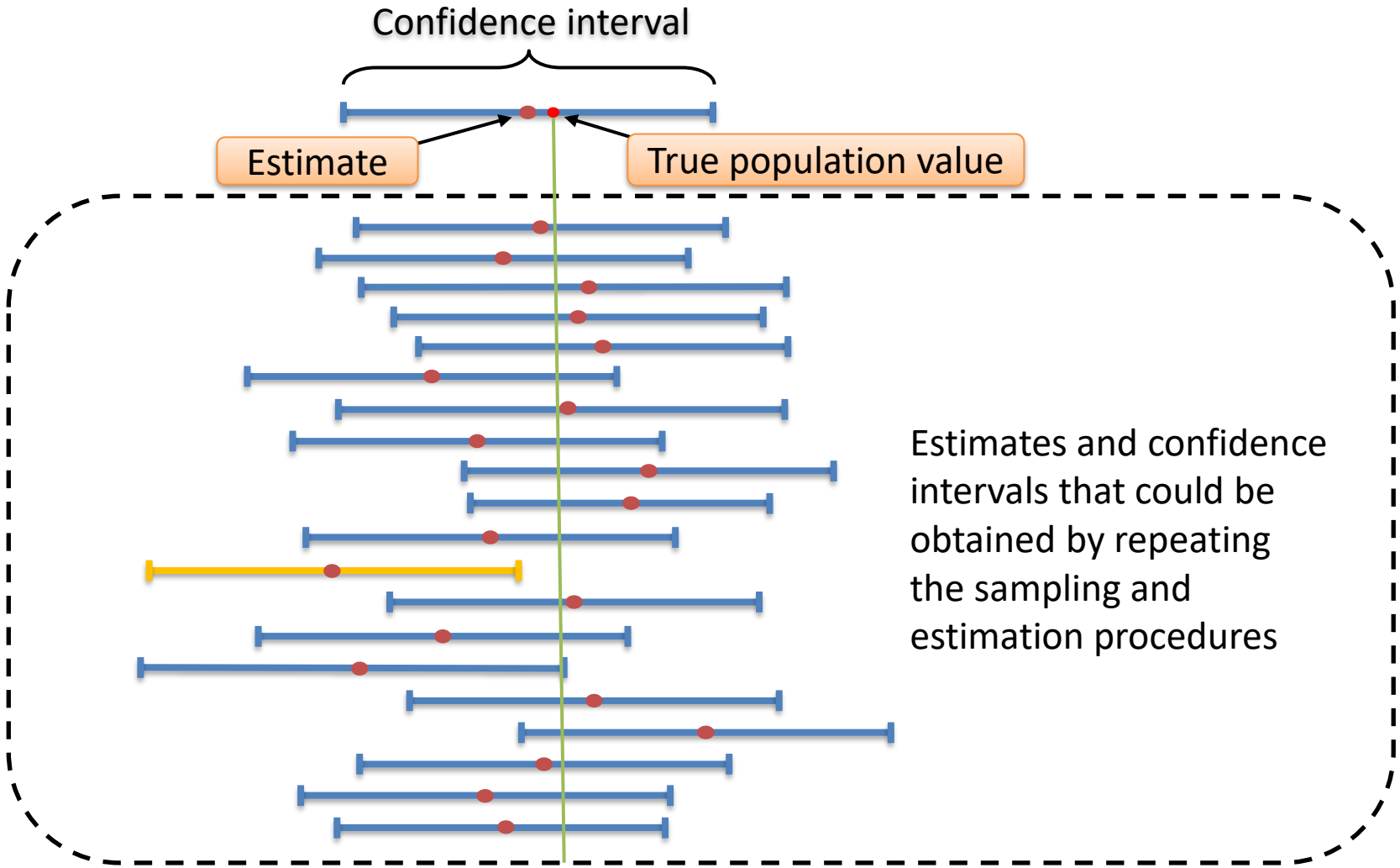
Confidence intervals: Basic definitions and concepts

- A **confidence interval** for an estimate is an interval constructed around the estimate which reflects the estimate's uncertainty.
- It is expressed by two numbers, the lower and upper bounds of the interval
- A confidence interval is associated with a **confidence level**, which is expressed as a percentage.
- The confidence level for the confidence intervals provided with long-form estimates is **95%**. Intuitively, this is the **degree to which we can be confident that the interval contains the true population value**.

Statistics (3)	Count	95% confidence interval lower bound, Count	95% confidence interval upper bound, Count
Mother tongue (9)[2]	Total - Mother tongue[6]	Total - Mother tongue[6]	Total - Mother tongue[6]
Age (19)[3]	Total - Age	Total - Age	Total - Age
Gender (3)[4, 5]	Total - Gender	Total - Gender	Total - Gender
Province or territory of residence 1 year ago (14)	Alberta	Alberta	Alberta
Current - Province or territory of residence			
Canada (map)	57,205	55,926	58,513
Newfoundland and Labrador (map)	1,545	1,337	1,786
Prince Edward Island (map)	440	341	567
Nova Scotia (map)	2,900	2,574	3,268
New Brunswick (map)	1,855	1,662	2,070
Quebec (map)	2,670	2,381	2,994
Ontario (map)	12,340	11,667	13,052
Manitoba (map)	2,220	1,962	2,513
Saskatchewan (map)	4,760	4,407	5,141

RECENSEMENT · CENSUS

Confidence intervals: Interpretation



- Hypothetically repeating the sampling and estimation procedures would result in different estimates and confidence intervals.
- If the process were repeated many times, approximately 95% of the intervals would contain the true population value.

Constructing confidence intervals

Most basic method: Wald confidence interval

The lower bound (LB) and the upper bound (UB) of a 95% Wald confidence interval for a population parameter of interest θ are given by:

$$LB = \hat{\theta} - z \times \widehat{SE}(\hat{\theta}), \qquad UB = \hat{\theta} + z \times \widehat{SE}(\hat{\theta}),$$

where

- $\hat{\theta}$ is the estimate of θ
- z is the 97.5th percentile of the standard normal distribution (approximately 1.96)
- $\widehat{SE}(\hat{\theta})$ is the standard error of $\hat{\theta}$.

Assumes the sampling distribution of the estimator is a normal distribution.

Confidence interval methods for the 2021 Census

RECENSEMENT · CENSUS

- The normality assumption of the Wald interval is often violated, particularly for
  - Small sample sizes
  - Proportion and count statistics.

  Therefore the census uses more elaborate confidence interval methods.
- The method used to construct confidence intervals depends on the type of statistic:
  - All statistics except proportions and counts: **Student's confidence interval**
  - Proportions: **Modified Wilson confidence interval for proportions** (Kott and Carr, 1997; Neusy and Mantel, 2016)
  - Counts: **Modified Wilson confidence interval for counts** (Neusy, Savard, Hidiroglou and Martin, 2021)
- Research and simulations have been done to ensure that confidence intervals for long-form estimates are constructed using methods which achieve coverage close to the stated confidence level in most scenarios.

Student's confidence interval

The lower bound (LB) and the upper bound (UB) of a 95% Student's confidence interval for a population parameter of interest θ are given by:

$$LB = \hat{\theta} - t \times \widehat{SE}(\hat{\theta}), \qquad UB = \hat{\theta} + t \times \widehat{SE}(\hat{\theta}),$$

where

- $\hat{\theta}$ is the estimate of θ
- t is the 97.5th percentile of the Student's t-distribution with R degrees of freedom
- R is the number of replicates used in variance estimation ($R = 32$ for disseminated data products)
- $\widehat{SE}(\hat{\theta})$ is the standard error of $\hat{\theta}$.

- For small sample sizes, the Student's confidence interval has better coverage than the Wald interval.
- If the number of degrees of freedom is large, the Student's confidence interval is very similar to the Wald interval.

RECENSEMENT · CENSUS

Modified Wilson confidence interval for proportions

The lower bound (LB) and the upper bound (UB) of a 95% modified Wilson confidence interval for a proportion-type statistic p are given by:

$$LB = \frac{\hat{p}+t^2/2n_e}{1+t^2/n_e} - \frac{t\sqrt{\hat{p}(1-\hat{p})+t^2/4n_e}}{\sqrt{n_e}(1+t^2/n_e)}, \qquad UB = \frac{\hat{p}+t^2/2n_e}{1+t^2/n_e} + \frac{t\sqrt{\hat{p}(1-\hat{p})+t^2/4n_e}}{\sqrt{n_e}(1+t^2/n_e)},$$

where

- $\hat{p}$ is the estimate of p
- t is the 97.5th percentile of the Student's t-distribution with R degrees of freedom
- $n_e = \min(n/\mathrm{deff}(\hat{p}), n)$ is the effective sample size
- $\mathrm{deff}(\hat{p}) = \frac{\hat{V}(\hat{p})}{\hat{p}(1-\hat{p})/n}$ is the estimated design effect
- n is the in-scope sample size
- $\hat{V}(\hat{p})$ is the estimated variance of $\hat{p}$.

- **Asymmetric interval**
- Has **better coverage** than Wald and Student intervals for small sample sizes and when the population proportion is near zero or one.

Modified Wilson confidence interval for counts

RECENSEMENT · CENSUS

- Method developed by Statistics Canada researchers for the 2021 Census.

The lower bound (LB) and the upper bound (UB) of a 95% modified Wilson confidence interval for a count Y are given by:

$$LB = \hat{Y} + t^2\frac{1}{2}\frac{\hat{\mathrm{V}}(\hat{Y})}{\hat{Y}} - \sqrt{t^2\hat{\mathrm{V}}(\hat{Y}) + \left(t^2\frac{1}{2}\frac{\hat{\mathrm{V}}(\hat{Y})}{\hat{Y}}\right)^2}, \qquad UB = \hat{Y} + t^2\frac{1}{2}\frac{\hat{\mathrm{V}}(\hat{Y})}{\hat{Y}} + \sqrt{t^2\hat{\mathrm{V}}(\hat{Y}) + \left(t^2\frac{1}{2}\frac{\hat{\mathrm{V}}(\hat{Y})}{\hat{Y}}\right)^2},$$

where

- $\hat{Y}$ is the estimate of Y
- t is the 97.5th percentile of the Student's t-distribution with R degrees of freedom
- $\hat{V}(\hat{Y})$ is the estimated variance of $\hat{Y}$.

- **Asymmetric** interval
- Has **better coverage** than Wald and Student intervals for small sample sizes and when the population count is near zero or the population size.

RECENSEMENT • CENSUS

Confidence intervals and statistical hypothesis testing

Confidence intervals and statistical hypothesis testing

RECENSEMENT · CENSUS

Confidence intervals are closely related to statistical hypothesis testing.

Statistical hypothesis testing

- A method of statistical inference used to decide if data gathered from a sample support a hypothesis about a population parameter.

Setup

H_0: The **null hypothesis** -> the "status quo"

H_A: The **alternative hypothesis** -> corresponds to scientific/analytical discovery

Usually analysts want to show that the data support rejecting the null hypothesis.

RECENSEMENT · CENSUS

Confidence intervals and statistical hypothesis testing

Example:

H_0: In a given geographic area, the proportion of people in 2021 whose highest level of education was a bachelor's degree was equal to the proportion in 2016, i.e. $p_{2021} = p_{2016}$.

H_A: In the geographic area, these two proportions are not equal, i.e. $p_{2021} \neq p_{2016}$.

Rejection criterion: A criterion involving quantities computed from the sample data that is used to decide whether or not to reject H_0, e.g. reject H_0 in the example if $|\hat{p}_{2021} - \hat{p}_{2016}|$ is "large enough".

Significance level: The probability of rejecting H_0 when H_0 is actually true. This quantity is fixed by the analyst, often set to 0.05.

Confidence intervals and statistical hypothesis testing

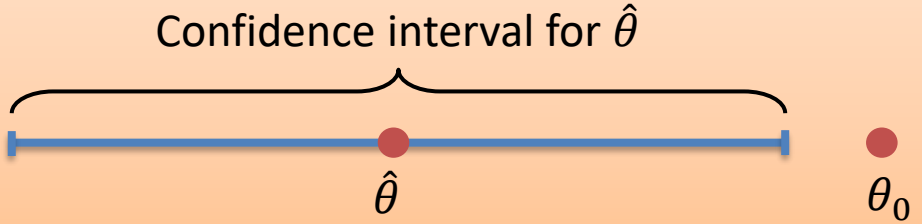
RECENSEMENT · CENSUS

For the following type of test

$$\mathrm{H}_0: \theta = \theta_0$$
$$\mathrm{H}_\mathrm{A}: \theta \neq \theta_0$$

where θ is a population parameter and θ_0 is a fixed value, **a confidence interval for $\hat{\theta}$ provides a rejection criterion:**

H_0 can be rejected at level 0.05 if and only if θ_0 is outside the 95% confidence interval for $\hat{\theta}$.



i.e. Reject H_0 if and only if $\hat{\theta}$ is "far" from θ_0.

Confidence intervals and statistical hypothesis testing

RECENSEMENT · CENSUS

Example: *Performing a test of significant difference for proportions of people in 2016 and 2021 whose highest level of education was a bachelor's degree.*

The hypotheses are

$$\mathrm{H}_0: p_{2021} - p_{2016} = 0 \quad \textit{(the proportions are equal)}$$
$$\mathrm{H}_\mathrm{A}: p_{2021} - p_{2016} \neq 0 \quad \textit{(the proportions are not equal)}$$

- We are given estimates of p_{2021} and p_{2016}: $\hat{p}_{2021} = 0.230$, $\hat{p}_{2016} = 0.195$.
- These can be used to compute an estimate of $p_{2021} - p_{2016}$:

$$\hat{p}_{2021} - \hat{p}_{2016} = 0.230 - 0.195 = 0.035$$

- **A 95% confidence interval for the difference $\hat{p}_{2021} - \hat{p}_{2016}$ provides a criterion for rejecting H_0.**
- If the 95% confidence interval for $\hat{p}_{2021} - \hat{p}_{2016}$ is

$$\mathrm{LB} = 0.025, \quad \mathrm{UB} = 0.045,$$

then the null hypothesis would be rejected at level 0.05 because the confidence interval does not contain 0.

RECENSEMENT · CENSUS

Conclusion

- Several data quality indicators are provided with 2021 Census data products.
- Many of these quality indicators are new for the 2021 Census.
- Overall, census data quality is very good, but it is recommended that users consult the entire suite of quality indicators in assessing the relevance of census data for their specific use cases.
- In particular, confidence intervals for long-form estimates can be used for statistical inference.
- For more information on data quality indicators for the 2021 Census: 2021 Census Data Quality Guidelines (statcan.gc.ca)

RECENSEMENT • CENSUS

References

Kott, P.S. and Carr, D.A. (1997). "Developing an Estimation Strategy for a Pesticide Data Program." Journal of Official Statistics, Vol. 13, No. 4, 367-383.

Neusy E., and Mantel H. (2016). "Confidence Intervals for Proportions Estimated from Complex Survey Data." Proceedings of the Survey Methods Section. SSC Annual Meeting, June 2016.

Neusy E., Savard S.-A., Hidiroglou M., Martin V. (2021). "Modified Wilson Confidence Intervals for Estimated Counts with Application to Census 2021 Long Form Estimation." Statistics Canada internal document.

Statistics Canada (2022). *2021 Census Data Quality Guidelines*. Catalogue no. 98-26-0006. *covers most of the material in this presentation*

Statistics Canada (to be released in August 2023). *Sampling and Weighting Technical Report, Census of Population 2021*. Catalogue no. 98-500-X, issue 2021005.

RECENSEMENT · CENSUS

About the Data Service Centre...

The Data Service Centre (DSC) offers to data users a complete range of services. The DSC assists data users:

- With simple and free data requests, technical and methodological questions;
- With complex requests requiring research, extraction or customization of data from multiple sources;
- By producing customized reports, analyses and maps;
- By offering standard and customized workshops for data users of all levels of expertise;
- Through various outreach activities such as webinars, newsletters, presentations, information sessions and open houses.

For any questions, contact us:

- 1-800-263-1136
- STATCAN.infostats-infostats.STATCAN@canada.ca