Microdata User Guide

Labour Force Survey (LFS)
Public Use Microdata File (PUMF)

January 2025

Statistics Canada Statistique Canada

Canada

Table of Contents

1.0 Introduction

The Labour Force Survey (LFS) is a household survey carried out monthly by Statistics Canada. Since its inception in 1945, the objectives of the LFS have been to divide the working-age population into three mutually exclusive categories in relation to the labour market – employed, unemployed, and not in the labour force – and to provide descriptive and explanatory data on each of these groups. Data from the survey provide information on major labour market trends, such as shifts in employment across industrial sectors, hours worked, labour force participation and unemployment rates.

This public use microdata file (PUMF) contains non-aggregated data for a wide variety of variables collected from the LFS. This product is for users who prefer to do their own analysis by focusing on specific subgroups in the population or by cross-classifying variables that are not in LFS catalogued products. The data have been modified to ensure that no individual or business is directly or indirectly identified. Variables most likely to lead to identification of an individual are removed from the microdata file or are collapsed to broader categories.

This guide has been produced to facilitate the use of the PUMF and the interpretation of the results. For more detailed information on the LFS and its methodology, please refer to the Guide to the Labour Force Survey (71-543-G).

Specific inquiries about this product and related statistics or services should be directed to:

Statistics Canada
Centre for Labour Market Information
E-mail: **statcan.labour-travail.statcan@statcan.gc.ca**

2.0 Concepts and Definitions

The concepts and definitions of employment and unemployment adopted by the Labour Force Survey are based on those endorsed by the International Labour Organization (ILO).

The concepts of employment and unemployment are derived from the theory of labour supply as a factor in production. In this context, production refers to the goods and services included in the System of National Accounts. For this reason, unpaid housework and volunteer work are not counted as work for purposes of the LFS, although these activities need not differ from paid work either in purpose or in the nature of the tasks involved.

Employment includes:

1. Persons who, during the LFS reference week, did at least one hour of paid work, either in the context of an employer-employee relationship, or in the context of self-employment;[1] **AND**
2. Persons who, during the LFS reference week, had a job or business but were absent from work.[2]

Unemployment includes:

1. Persons who, during the reference week: did not have a job or business, were available to work, and had looked for work in the four weeks ending with the reference week; **AND**
2. Persons who, during the reference week: did not have a job or business, were available to work, and were on temporary layoff due to business conditions with an expectation of recall; **AND**
3. Persons who, during the reference week, did not have a job or business, were available to work, and had a job to start in the future within four weeks from the reference period:

   (Note that a person in categories 2 and 3 need not have looked for work during the four weeks ending with the reference week)

Not in the Labour Force includes:

1. Persons who were neither employed nor unemployed during the reference week and wanted to work, but did not meet the criteria for unemployment described above; **AND**
2. Persons who were unavailable or unable to work or who did not want to work.

On the LFS PUMF, the variable labour force status (LFSSTAT) indicates whether an individual is employed, unemployed or not in the labour force.

A full list of variables on the LFS PUMF can be found in the LFS PUMF codebook.

[1] Employment also includes persons who did unpaid family work, which is defined as unpaid work contributing directly to the operation of a farm, business or professional practice owned and operated by a related member of the same household.

[2] This excludes persons not at work because they were on layoff or between casual jobs, and those who had a job to start at a future date.

3.0 Survey Methodology

This section provides a brief overview of the methodology of the LFS to support users of the PUMF. For more detailed information on the methodology of the LFS, refer to the Guide to the Labour Force Survey (71-543-G).

3.1 Target population

The Labour Force Survey (LFS) target population includes all persons aged 15 years and older whose usual place of residence is in Canada, including both non-permanent residents (NPRs) —that is, those with a work or study permit, their families, asylum claimants, protected persons and related groups — as well as permanent residents (landed immigrants) and the Canadian-born population.

Populations excluded from the LFS target population include those living on reserves, full-time members of the regular Armed Forces and persons living in institutions (including inmates of penal institutions and patients in hospitals and nursing homes). These groups together represent an exclusion of less than 2% of the Canadian population aged 15 and over.

3.2 Sampling

The LFS sample is drawn from an area frame and is based on a stratified, multi-stage design that uses probability sampling. The monthly LFS sample size is approximately 68,000 households, resulting in the collection of labour market information for approximately 100,000 individuals.

The LFS uses a rotating panel sample design. In the provinces, selected dwellings remain in the LFS sample for six consecutive months. Each month, about one-sixth of the LFS sampled dwellings are in their first month of the survey, one-sixth are in their second month of the survey, and so on. These six independent samples are called rotation groups. On the LFS PUMF, respondent identifiers (REC_NUM) are randomly assigned each month; therefore, it is not possible to calculate labour market flows or track a respondent through their six months in the sample.

3.3 Data collection

The LFS collects information related to the work activities of each household member during the LFS reference week, which is generally the week containing the 15th day of the month. Interviewing for the LFS is carried out each month over the ten days immediately following the reference week.

During the data collection period, selected households are contacted by trained Statistics Canada interviewers either by telephone or in person. Interviews can also be completed online, with members of selected households receiving an invitation to complete their questionnaire on a secure Statistics Canada data collection platform. These interviews are completed without the involvement of an interviewer.

In each dwelling, information about all household members is usually obtained from one knowledgeable household member. This proxy reporting, which accounts for approximately 65% of the information collected, is used to avoid the high cost and extended time requirements that would be involved in repeat calls necessary to obtain information directly from each respondent.

The LFS questionnaire is available online at Questionnaire(s) and Reporting guide(s) - Labour Force Survey questionnaire

3.4 Editing and Imputation

At the end of data collection, a series of verification steps are performed to identify and eliminate potential duplicate records and to exclude non-responding and out-of-scope records. Editing is also performed at this step to identify and correct inconsistent or invalid responses.

Imputation is the process that replaces invalid or missing information with valid values. The new values are supplied in such a way as to preserve the underlying structure of the data and to ensure that the resulting records will pass all required edits. In other words, the objective is not to reproduce the true microdata values, but rather to establish internally consistent data records that yield good aggregate estimates.

The imputation methods employed in the LFS include carry-forward, deterministic and nearest neighbour donor imputation. In some cases, complete non-response – where all questionnaire data for a household are missing – is resolved by a non-response adjustment, as described in the sub-section entitled Weighting.

3.5 Creation of Derived Variables

Most variables on the LFS PUMF have been derived by combining responses or performing calculations based on responses to the questionnaire. For example, in the LFS questionnaire, respondents are asked to report their wage or salary before taxes and other deductions and include tips and commissions. Then, hourly wages (HRLYEARN) are calculated in conjunction with usual paid work hours per week.

Data users should refer to the LFS PUMF codebook for the code set and universe for each of the derived variables.

3.6 Weighting

LFS data are weighted to enable tabulations of estimates at national, provincial, and sub-provincial levels of aggregation.

The sample design determines a certain number of weighting factors to be used in the calculation of the individual weights. The main component is the inverse of the probability of selection, known as the basic weight. For example, in an area where 2% of the households are sampled, each household would be assigned a basic weight of 1/.02 = 50. The basic weight is then adjusted for any sub-sampling due to growth that may have occurred in the area and for non-response and coverage error.

In the LFS, some survey non-response is resolved by imputation: carry forward, substitution or donor imputation methods. Any remaining non-response is accounted for by adjusting the weights for the responding households in the same area. This non-response adjustment assumes that the characteristics of the responding households are not significantly different from the non-responding households.

The final adjustment to the weight is made to correct for coverage errors and to reduce the sampling variance of the estimates. The subweights are adjusted using composite calibration: this increases efficiency by leveraging the overlap between consecutive monthly samples, and it ensures that survey estimates conform to control totals based on population estimates by age, gender, and geography.

3.7 Revisions

Adjustments are made to LFS data every five years after new population estimates become available following the most recent census. As of January 2025, LFS estimates have been adjusted to reflect population counts from the 2021 Census, with revisions going back to 2011.

Occasionally, LFS data are revised to incorporate updated industry and occupation standards, or enhancements to methodology, data processing and technological systems.

All revisions to the LFS are described in the series "Improvements to the Labour Force Survey (LFS)" (71F0031X).

4.0 Disclosure control

Statistics Canada is prohibited by law from releasing any data which would divulge information obtained under the *Statistics Act* that relates to any identifiable person, business or organization without the prior knowledge or the consent in writing of that person, business or organization. Confidentiality rules are applied to all data that are released or published to prevent the disclosure of any information deemed confidential. If necessary, data are suppressed to prevent direct or residual disclosure of identifiable data.

As such, data in the Public Use Microdata Files (PUMF) may differ from the survey master files held by Statistics Canada. These differences usually are the result of actions taken to protect the anonymity of individual survey respondents. The most common actions are the suppression of data items and grouping values into wider categories.

For example, the LFS master file includes geographic identifiers for detailed sub-provincial areas including census metropolitan areas (CMAs), census agglomerations, economic regions and census subdivisions, whereas geography variables on the PUMF are limited to the provinces and the nine largest CMAs. Furthermore, variables such as age, industry and occupation have been grouped on the PUMF to protect the confidentiality of respondents. In addition, some records on the PUMF have been perturbed as an added layer of security, such that any individual record on the PUMF may not reflect a true record on the LFS master file.

Measures have been taken to ensure that resulting estimates from the PUMF remain close to the official estimates calculated from the master data file; however, there may be some discrepancies compared with estimates published on the Statistics Canada website, particularly for smaller domains. In the event of a discrepancy, estimates in published tables and other data products on the Statistics Canada website should be considered official statistics.

5.0 Tabulation, Analysis and Release

5.1 Estimation and Statistical Analysis

The LFS is based upon a complex sample design, with stratification, multiple stages of selection, and unequal probabilities of selection of respondents. Using data from such complex surveys presents challenges for analysts because the survey design and the selection probabilities affect the estimation and variance calculation procedures that should be used. To ensure accurate results, the provided survey weights (FINALWT) must be used for all estimation, tabulation and statistical analysis using the LFS PUMF.

For small domains, it may be necessary to combine monthly files to produce reliable estimates. For the LFS, it is typically recommended to use a three-month moving average or annual average estimate for small domains such as CMAs or immigrants. To calculate a three-month moving average, users should stack the three monthly PUMFs into one file and divide the survey weight by 3. For example, a three-month moving average estimate for April would be based on PUMFs for February, March and April and the weight used would be WT_3MMA = FINALWT/3. Similarly, to calculate an annual estimate, users should stack the twelve monthly PUMFs and divide the survey weight by 12, i.e., WT_ANNUAL = FINALWT/12. The combined file and adjusted weight can then be used for tabulation or analysis.

Before completing any analysis with the dataset, it is essential to review the LFS PUMF codebook. The codebook describes the code set and universe for each variable on the data file as well as notes on special formatting, inclusions or exclusions.

For example, variables on the PUMF describing hours (UHRSMAIN, AHRSMAIN, UTOTHRS, ATOTHRS, HRSAWAY, PAIDOT, UNPAIDOT, XTRAHRS) and wages (HRLYEARN) are presented as whole numbers with implied decimals. Therefore, data in these variables must be transformed to ensure that they are used correctly. For example, actual total hours worked per week at all jobs has one decimal implied and should be divided by 10 to obtain a value in hours, i.e., a value of 435 for ATOTHRS corresponds to 43.5 hours per week. Average hourly wages has two implied decimals and should be divided by 100 to obtain a value in dollars and cents, i.e., a value of 2345 for HRLYEARN corresponds to $23.45.

Data users should also review the universe for each variable before calculating estimates or performing analysis. For example, the universe for job permanency (PERMTEMP) is “Currently employed, employees.” Therefore, before doing analysis using this variable, the population of interest should be limited to persons who are employed employees, i.e., (LFSSTAT = 1 or LFSSTAT = 2) and (COWMAIN = 1 or COWMAIN = 2).

5.2 Interpretation of Variance

It is important to determine the quality of any estimates used in statistical analysis. While data quality is affected by both sampling and non-sampling errors, the quality guidelines in this user guide address quality assessments determined only on the basis of sampling error. For more information on sampling and non-sampling errors, refer to *Methodology of the Canadian Labour Force Survey* (71-526-x).

There are two main factors which should be considered when determining the quality of an estimate: the number of respondents who contribute to the calculation of the estimate and the sampling variability of the estimate.

To produce an estimate of acceptable quality, at least 5 respondents should contribute to the calculation of the estimate.

For most weighted estimates, an appropriate measure of quality can be determined by calculating the coefficient of variation (CV) of the estimate by dividing the standard error of the estimate by the estimate itself, then following the guidelines in the table below.

Table 5: Quality Guidelines

Quality of Estimate	Guidelines
1) Acceptable	Estimates have a sample size of 5 or more, and coefficients of variation less than 15%. No warning is required.
2) Marginal	Estimates have a sample size of 5 or more, and coefficients of variation of 15% to 35%. Estimates should be accompanied by a warning regarding data quality.
3) Unacceptable	Estimates have a sample size of less than 5, or coefficients of variation in excess of 35%. Statistics Canada recommends not to release estimates of unacceptable quality.

Estimates of marginal or unacceptable quality should be accompanied by a warning to caution subsequent users.

Note that for estimates of small ratios, such as the unemployment rate, the CV may be misleadingly large. For small ratios, differences, or any other type of estimate, a confidence interval may provide a better representation of the data quality. More information on calculating a confidence interval using the PUMF is found in section 6.2.2.

6.0 Poisson Bootstrap Weights for Variance Estimation

The calculation of precise variance estimates requires detailed knowledge of the sample design of the survey. To protect the confidentiality of respondents, this level of detail cannot be included in a PUMF. For LFS PUMF users, the Poisson bootstrap method can be used to approximate the true value of the variance (Beaumont & Patak, 2012). If required, more precise variance estimates for most statistics which account for the complexities of the LFS sample design can be requested on a cost-recovery basis using Rao-Wu Bootstrap weights and the survey master file.

The Poisson Bootstrap is a special case of the Generalized Bootstrap. Its practicality to users, its flexibility, and its robustness makes bootstrapping a widely accepted technique for variance estimation for PUMF users. With the large number of observations on the LFS PUMF and enough bootstrap replicates, assumptions to use the Poisson bootstrap technique are easily met. The technique is a one-way design and assumes that all observations are independent of each other. The proposed method utilizes one of the simplest implementations of the Poisson bootstrap, as described in Beaumont and Patak (2012).

6.1 How to create Poisson bootstrap weights:

For each unit *k* on the PUMF, calculate an adjustment factor as follows:

$$adjustment\ factor_k = 1 + poisson\ factor_k * \sqrt{\frac{(finalwt_k - 1)}{finalwt_k}} \qquad (1)$$

where $poisson\ factor = 1\ or - 1\ with\ 50\%\ probability$ and $finalwt_k$ is the survey weight for unit *k* on the PUMF.

Then, calculate the bootstrap weight as:

$$bootstrap\ weight = finalwt * \ adjustment\ factor \qquad (2)$$

Equation (1) and (2) can be combined to view the calculations as:

$$bootstrap\ weight = finalwt + \left(poisson\ factor * finalwt * \sqrt{\frac{(finalwt_k - 1)}{finalwt_k}} \right)$$

which can also be written as:

$$bootstrap\ weight = finalwt\ \pm finalwt * \sqrt{\frac{(finalwt_k - 1)}{finalwt_k}}$$

where the + or - is determined by the random poisson factor.

Repeat this procedure to create 1,000 bootstrap replicates.

A calibrated version of the Poisson bootstrap can lead to variance estimates that are closer to those estimated by the LFS master file. Calibration can be done by adjusting each of the 1,000 uncalibrated bootstrap weights using:

$$calibrated\ bootstrap\ weight = \frac{sum\ of\ finalwt\ by\ domain}{sum\ of\ bootstrap\ weights\ by\ domain} * bootstrap\ weight \qquad (3)$$

where $bootstrap\ weight$ are the 1,000 weights from (2), and the sum of bootstrap weights are calculated for each bootstrap replicate.

In essence, each bootstrap replicate is calibrated to match estimates made using the FINALWT variable on the PUMF. The most efficient way to do this is to choose domains that are similar to those used in the calibration of the master data file: province, age group and gender. See Appendix B for recommended calibration domains and Appendix C.1 for an example showing how to use the formulae to generate bootstrap weights.

These weights can then be used to calculate the variances in the same way as described in the Rao-Wu-Yue bootstrap, found in Chapter 7 of the *Methodology of the Canadian Labour Force Survey* (Statistics Canada, 2017).

6.2 How to use bootstrap weights to calculate variance

A file of Poisson bootstrap weights can be merged with the corresponding PUMF file and used to calculate the variance of any estimate by following the steps below:

1. Compute the survey estimate using the variable of interest and the final survey weight (FINALWT)
2. Compute a bootstrap estimate for each bootstrap replicate by using the variable of interest and the calibrated bootstrap weights for each replicate. This should result in 1,000 bootstrap estimates

3. Compute the bootstrap variance of the 1,000 bootstrap estimates using the formula:

$$bootstrap\ variance(estimate) = \frac{(bs\ estimate_1 - survey\ estimate)^2 + \cdots + (bs\ estimate_{1000} - survey\ estimate)^2}{1000}$$

where $bs\ estimate_{1-1000}$ are the estimates calculated from the 1,000 bootstrap replicates. If estimates are needed by domain, such as province or age group, the above steps would be implemented separately for each level of the desired domain.

This variance can be used to calculate CVs or confidence intervals as desired. See Appendix C2 for an example showing how to use the bootstrap weights to produce quality indicators.

6.2.1 How to use bootstrap weights to calculate CVs

Once the variance has been calculated as above, the coefficient of variation can be simply computed as follows:

$$CV = \frac{\sqrt{bootstrap\ variance(estimate)}}{estimate}$$

Refer to section 5.2 for more information on how to interpret the CV.

6.2.2 How to use bootstrap weights to calculate confidence intervals

Although coefficients of variation are widely used, a more intuitively meaningful measure of sampling error is the confidence interval of an estimate. A confidence interval constitutes a statement on the level of confidence that the true value for the population lies within a specified range of values. For example, a 95% confidence interval can be described as follows:

> If sampling of the population is repeated indefinitely, each sample leading to a new confidence interval for an estimate, then in 95% of the samples the interval will cover the true population value.

Using the standard error of an estimate, confidence intervals may be obtained under the assumption that under repeated sampling of the population, the various estimates obtained for a population characteristic are normally distributed about the true population value. Under this assumption, the chances are about 68 out of 100 that the difference between a sample estimate and the true population value would be less than one standard error, about 95 out of 100 that the difference would be less than two standard errors, and about 99 out of 100 that the difference would be less than three standard errors. These different degrees of confidence are referred to as the confidence levels.

Confidence intervals for an estimate can be calculated directly by using the following formula to convert to a confidence interval:

$$Confidence\ interval = (estimate - t * \sqrt{variance(estimate)},\ \ estimate + t * \sqrt{variance(estimate)})$$

where t is the approximate endpoint of the normal distribution. Depending on the level of confidence, the following table can be used:

Table 6: Critical values for determining confidence intervals

Critical value (t)	**Confidence Level**
1.0	68% confidence interval
1.6	90% confidence interval
2.0	95% confidence interval
2.6	99% confidence interval

An alternate method to compute confidence intervals using the bootstrap weights is to determine the percentiles of the distribution of estimates corresponding to the desired level of confidence. This is done by first sorting the estimates from the bootstrap replicates in ascending order. Then to obtain the confidence interval, compute which two percentiles needed to evenly bound the desired level of confidence, as shown in this equation:

$$lower\ bound\ percentile = \frac{100\% - desired\ confidence\ level}{2}$$

$$upper\ bound\ percentile = 100\% - \left(\frac{100\% - desired\ confidence\ level}{2}\right)$$

For example, to create a 95% confidence interval, calculate the 2.5th percentile ((100%-95%)/2) and the 97.5th percentile (100%-((100%-95%)/2), determining bounds for the centre 95% of the estimates. Similarly, for a 68% confidence interval, calculate the 16th percentile and the 84th percentile. To then find the values corresponding to those percentiles, multiply the lower and then upper bound percentile by the number of bootstrap estimates and take the corresponding estimate from the sorted estimates. For example, for a 95% confidence interval using 1,000 bootstrap replicates, take the 25th estimate and the 975th estimate when the estimates are sorted in ascending order.

Note: Release guidelines which apply to the estimate also apply to the confidence interval. For example, if the estimate is not releasable, then the confidence interval is also not releasable.

See Appendix C2 for an example showing how to use the bootstrap weights to produce quality indicators such as the CV and confidence intervals.

Appendix A – Formulae

A1 Creation of Poisson Bootstrap Weights

Designate the number of bootstrap replicates, *B*. For each *b* from 1 to *B*, let

$$a_{kb} = 1 + \widetilde{a_{kb}} \sqrt{\frac{(w_k - 1)}{w_k}} \qquad (1)$$

Where: $\widetilde{a_{kb}} = \begin{cases} 1 \text{ with probability } 0.5 \\ -1 \text{ with probability } 0.5 \end{cases}$

w_k is the survey weight for unit *k* on the PUMF.

Then, calculate each bootstrap weight w_{kb}^* as

$$w_{kb}^* = a_{kb} w_k \qquad (2)$$

To calibrate the bootstraps weights, adjust each weight using

$$bw_{kb} = \frac{n_d}{m_{db}} w_{kb}^* \qquad (3)$$

Where: w_{kb}^* is the bootstrap weight *b* for unit *k* from (2)

n_d is the sum of the survey weights w_k in domain *d*

m_{db} is the sum of bootstrap weights w_{kb}^* in domain *d*

A2 Quality indicators

Variance:

$$\widehat{var}(\hat{X}) = \frac{\sum_{b=1}^{1000} (\hat{X}^{*(b)} - \hat{X})^2}{1000}$$

Where: $\hat{X}^{*(b)}$ is the bootstrap estimate for bootstrap replicate *b*

$\hat{X}$ is the survey estimate.

Standard Error:

$$\hat{\sigma} = \sqrt{\widehat{var}(\hat{X})}$$

Coefficient of Variation:

$$CV = \frac{\hat{\sigma}}{\hat{X}}$$

Confidence Interval:

$$CI = \hat{X} \pm t * \hat{\sigma}$$

Where: t is the approximate endpoint of the normal distribution for the desired confidence level

Critical value (t)	**Confidence Level**
1.0	68% confidence interval
1.6	90% confidence interval
2.0	95% confidence interval
2.6	99% confidence interval

OR

$$\left(\frac{100\% - p}{2}\right)^{th} percentile, \ \left(100\% - \frac{100\% - p}{2}\right)^{th} percentile$$

Where *p* is the desired confidence level

percentile is calculated from the 1,000 bootstrap estimates

Appendix B – Recommended Calibration Domains

Optimal domains to calibrate bootstrap weights are defined by the cross section of these three variables, giving a total of 220 calibration domains:

Province		Age Group		Gender	
PUMF variable	Description	PUMF variable	Description	PUMF variable	Description
prov = 10	Newfoundland and Labrador	age_6 = 1	15 to 16 years	Gender = 1	Men+
prov = 11	Prince Edward Island	age_6 = 2	17 to 19 years	Gender = 2	Women+
prov = 12	Nova Scotia	age_12 = 2	20 to 24 years		
prov = 13	New Brunswick	age_12 = 3	25 to 29 years		
prov = 24	Quebec	age_12 = 4	30 to 34 years		
prov = 35	Ontario	age_12 = 5 or age_12 = 6	35 to 44 years		
prov = 46	Manitoba	age_12 = 7 or age_12 = 8	45 to 54 years		
prov = 47	Saskatchewan	age_12 = 9	55 to 59 years		
prov = 48	Alberta	age_12 = 10	60 to 64 years		
prov = 59	British Columbia	age_12 = 11	65 to 69 years		
		age_12 = 12	70 and over		

Appendix C – Examples

C1: Generate bootstrap weights

This simple example will use a file that contains 4 sample units.

1. For each unit, create 1,000 replicates with a $poisson\ factor$ equal to 1 or -1 assigned randomly with probability 50%.

Rec_num	FINALWT	Poisson factor$_1$	Poisson factor$_2$	Poisson factor$_3$	…	Poisson factor$_{1000}$
1	500	1	-1	-1		1
2	450	-1	-1	1		-1
3	150	1	-1	1		1
4	250	-1	1	-1		-1

2. Apply formula (1) : $adjustment\ factor_k = 1 + poisson\ factor\ _k * \sqrt{\frac{(finalwt_k - 1)}{finalwt_k}}$

Rec_num	FINALWT	Adjustment factor$_1$	Adjustment factor$_2$	Adjustment factor$_3$	…	Adjustment factor$_{1000}$
1	500	= 1+ 1 * $\sqrt{\frac{500-1}{500}}$ = 1.99900	= 1+ -1* $\sqrt{\frac{500-1}{500}}$ = 0.00100	0.00100		1.99900
2	450	= 1+ -1* $\sqrt{\frac{450-1}{450}}$ = 0.00111	0.00111	1.99889		0.00111
3	150	1.99666	0.00333	1.99666		1.99666
4	250	0.00200	1.99800	0.00200		0.00200

3. Apply formula (2): $bootstrap\ weight = finalwt * adjustment\ factor$

Rec_num	FINALWT	Bootstrap weight$_1$	Bootstrap weight$_2$	Bootstrap weight$_3$	…	Bootstrap weight$_{1000}$
1	500	= 500*1.9990 = 999.500	=500*0.0010 = 0.50025	0.50025		999.500
2	450	= 450*0.0011 = 0.50027	0.50027	899.500		0.50027
3	150	299.499	0.50083	299.499		299.499
4	250	0.50050	499.499	0.50050		0.50050

4. For the next step, simple prov/age/gender data was added to the example dataset. Calculate the sums of FINALWT and all the bootstrap weights grouped by prov/age/gender combinations.

Rec_num	Prov	Age	Gender	FINALWT	Bootstrap weight$_1$	Bootstrap weight$_2$	Bootstrap weight$_3$	…	Bootstrap weight$_{1000}$
1	**10**	**1**	**1**	500	999.500	0.50025	0.50025		999.500
2	**10**	**1**	**1**	450	0.50027	0.50027	899.500		0.50027
3	**59**	**1**	**1**	150	299.499	0.50083	299.499		299.499
4	**59**	**1**	**1**	250	0.50050	499.499	0.50050		0.50050

Prov	Age	Gender	Sum of FINALWT	Sum of bootstrap weight$_1$	Sum of bootstrap weight$_2$	Sum of bootstrap weight$_3$	…	Sum of bootstrap weight$_{1000}$
10	1	1	= 500+450 = 950	=999.500+0.50027 = 1,000.0027	1.00052	900.0003		1,000.00027
59	1	1	= 150+250 = 400	=299.499+0.50050 = 299.9995	499.9998	299.9995		299.9995

5. Apply ratios to bootstrap weights for corresponding domains to create final calibrated bootstrap weights using formula (3):

$$calibrated\ bootstrap\ weight = \frac{sum\ of\ finalwt\ by\ domain}{sum\ of\ bootstrap\ weights\ by\ domain} * bootstrap\ weight$$

Rec_num	Prov	Age	Gender	bw_1	bw_2	bw_3	…	bw_{1000}
1	10	1	1	=(950/1,000.0027)*999.5 = **949.5247**	**474.99054**	**0.528042**		**949.5247**
2	10	1	1	**0.475256**	**475.0095**	**949.472**		**0.475256**
3	59	1	1	= (400/299.9995)*299.499 = **399.3327**	**0.400664**	**399.3327**		**399.3327**
4	59	1	1	**0.667334**	**399.5993**	**0.667334**		**0.667334**

C2: Estimate variance using bootstrap weights

This simple example will estimate the variance of the total number unemployed (LFSStat = 3) using the following example data set:

Rec_num	LFSStat	FINALWT	bw_1	bw_2	bw_3	…	bw_{1000}
1	1	500	704.5136	0.646157	0.596452		1547.675
2	1	450	0.352623	0.646182	1072.481		0.774643
3	3	400	563.5404	1032.688	0.596524		0.774705
4	4	200	281.5933	516.0197	476.3259		0.7752
5	1	150	399.3327	0.400664	399.3327		399.3327
6	3	250	0.667334	399.5993	0.667334		0.667334

Note that in the example calculations, the results will only be calculated from the 4 bootstrap replicates shown. In practice, all 1,000 bootstrap replicates should be used.

1. Compute survey estimate using survey weights:

$$\hat{X} = 1 * 400 + 1 * 250 = 650\ unemployed$$

2. Compute bootstrap estimates using bootstrap weights:

$$\hat{X}^{*(1)} = 563.5404 + 0.667334 = 564.2078\ unemployed$$

$$\hat{X}^{*(2)} = 1032.688 + 399.5993 = 1432.28725\ unemployed$$

$$\hat{X}^{*(3)} = 0.596524 + 0.667334 = 1.263858\ unemployed$$

…

$$\hat{X}^{*(1000)} = 0.774705 + 0.667334 = 1.442039\ unemployed$$

3. Compute variance:

$$\widehat{var}(\hat{X}) = [(564.2078 - 650)^2 + (1432.28725 - 650)^2 + (1.263858 - 650)^2 + \cdots + (1.442039 - 650)^2] / 1000$$

$$= 1{,}460.819654$$

Standard deviation:

$$\widehat{std}(\hat{X}) = \sqrt{1{,}460.819654}$$

$$= 38.22067 \ldots$$

Coefficient of variation:

$$CV = 38.22067 / 650$$

$$= 0.0588$$

$$= 5.9\%\ (acceptable\ quality)$$

95% Confidence interval:

$$\mathrm{CI} = 650 \pm 2 * 38.22067$$

$$= 650 \pm 74.9125 \ldots$$

$$= (573.56, 726.44)$$

Using the guidelines in Table 5, the CV for this estimate (5.9%) would fall in the acceptable quality range; however, the actual quality of the estimate would be unacceptable because it is based on a sample size of 2 records.

Appendix D – Sample code

D1: Example code in SAS

```
/*=========================================================================*/
/* SAMPLE CODE FOR POISSON BOOTSTRAP ON THE LFS PUMF                       */
/* This program creates a file of 1,000 poisson bootstrap weights for use  */
/* with the LFS PUMF, as well as shows a couple examples of how to         */
/* calculate the variance of some estimates                                */
/*-------------------------------------------------------------------------*/
/* Input_data = an LFS PUMF SAS file                                       */
/* seed = number so that you can reproduce the random results              */
/*=========================================================================*/

%let seed = 123;
%let b = 10; *Number of bootstrap replicates. Use 10 for testing and learning
              purposes. Use 1,000 for production of variance estimates;


/*---------CREATE POISSON BOOTSTRAP WEIGHTS--------------------------------*/

* To prepare for the calibration of the bootstrap weights, define the
  calibration age groups;

%macro prep_data(Input_data);
data pumf;
       set &Input_data;

       * age groups as defined in Appendix A of the LFS PUMF user guide;
                  if age_6 = '1'             then age_cal = '1516';
             else if age_6 = '2'             then age_cal = '1719';
             else if age_6 in ('3', '4')     then age_cal = '2024';
             else if age_12 = '03'           then age_cal = '2529';
             else if age_12 = '04'           then age_cal = '3034';
             else if age_12 in ('05', '06')  then age_cal = '3544';
             else if age_12 in ('07', '08')  then age_cal = '4554';
             else if age_12 = '09'           then age_cal = '5559';
             else if age_12 = '10'           then age_cal = '6064';
             else if age_12 = '11'           then age_cal = '6569';
             else if age_12 = '12'           then age_cal = '70+';
run;
%mend;
%prep_data(Input_data);


* Calculate adjustment factors for each bootstrap replicate;
%macro calc_adj_fact(input);

  data adj_fact;
    set &input.;
    array pois_fact{&b.} pois_fact1 - pois_fact&b.;
    array adj_fact{&b.} adj_fact1 - adj_fact&b.;
       * creates &b. adjustment factors;
       do i = 1 to &b.;
           pois_fact[i] = 2 * (ranuni(&seed.) >= 0.5) - 1;
                      * = 1 or -1 with 50% chance probability;
            adj_fact[i] = 1 + pois_fact[i] * sqrt(1 - (1 / finalwt));
       end;
    drop i pois_fact:; * not needed anymore;
   run;

%mend;
%calc_adj_fact(poisson_factors);
```

```
* Calculate the bootstrap weights;
%macro generate_reps(input);
data uncal_bsw;
       set &input;

       array adj_fact{&b.} adj_fact1 - adj_fact&b.;
       * using the adjustment factors to create bootstrap weights;
       array bwun{&b.} bwun1 - bwun&b.;
       * creates &b. bootstrap weights (uncalibrated);
              do i = 1 to &b.;
                     bwun[i] = finalwt * adj_fact[i];
              end;
       drop i adj_fact:; * not needed anymore;
run;
%mend;
%generate_reps(adj_fact);


* Get the sums of the weights by prov, age_cal, and gender for calibration;
%macro calculate_sums(input);
proc summary data = &input;
       var finalwt;
       class prov age_cal gender;
       types prov*age_cal*gender;
       output out = totals_finalwt(drop = _TYPE_ _FREQ_) sum = sum_finalwt;
run;

proc summary data = &input;
       var bwun1 - bwun&b.;
       class prov age_cal gender;
       types prov*age_cal*gender;
       output out = totals_boots(drop = _TYPE_ _FREQ_)
           sum = sum_boot1-sum_boot&b.;
run;
%mend;
%calculate_sums(uncal_bsw);

* Finally, calibrate bwun's to the sums of finalwts;
%macro calibrate_weights();
* merge the tables that contain the uncalibrated bsweights and the sums
  we just calculated;

proc sort data = uncal_bsw; by prov age_cal gender; run;
data to_calibrate;
       merge uncal_bsw totals_finalwt totals_boots;
       by prov age_cal gender;
run;

* multiply each bwun by the sum of the finalwts / sum of the bwuns;

data final_bs;
       set to_calibrate;

        array bw{&b.} bw1 - bw&b.;
              array bwun{&b.} bwun1 - bwun&b.;
              array sum_boot{&b.} sum_boot1 - sum_boot&b.;

           do i = 1 to &b.;
              bw[i] = bwun[i] * (sum_finalwt / sum_boot[i] );
           end;
```

```
             drop i bwun: sum_:; *leave dataset with just final
                                  bootstrap weights;
run;
%mend;
%calibrate_weights();

* Final dataset is "final_bs" which contains all the PUMF variables and the
  calibrated bootstrap weights. You can then use this dataset to calculate
  variance estimates as you wish;

* You can also combine all the steps into one macro and run it all
  at the same time;
%macro generate_bootstrap_weights(Input_data);

%prep_data(Input_data);

*combine steps 1-2;
data uncal_bsw;
set pumf;
    array adj_fact{&b.} adj_fact1 - adj_fact&b.;
    array bwun{&b.} bwun1 - bwun&b.;
           do i = 1 to &b.;
              adj_fact[i] = 2 * (ranuni(&seed.) >= 0.5) - 1;
              bwun[i] = finalwt * (1 + adj_fact[i] * sqrt(1 - (1 / finalwt)));
           end;
       drop i adj_fact1 - adj_fact&b.;
run;

%calculate_sums(uncal_bsw);
%calibrate_weights();

%mend;

%generate_bootstrap_weights(Input_data);

/*---------END OF CREATE POISSON BOOTSTRAP WEIGHTS-------------------------*/


/*---------EXAMPLES OF USING BOOTSRAP WEIGHTS TO CALCULATE VARIANCE--------*/
/* Each example could take at least a couple minutes                      */

/* Variance for totals */
/* Example: total unemployed by province */

proc surveyfreq data=final varmethod=bootstrap; *specify bootstrap;
  tables prov*lfsstat /CLWT varWT CVWT nopercent nototal;
  weight finalwt;
  repweight bw1-bw&b.; *specify name of the bootstrap weights created above;
  where lfsstat = '3'; *only calculate totals for unemployed;
  ods output CrossTabs=Results_totals;
run;


/* Variance of rates/ratios */
/* Example: Unemployment rate by province */

* set up binary indicator variables;
data pumf_unemp;
  set final;
          if lfsstat in ('1','2') then unemployed=0;
          else if lfsstat = '3'   then unemployed=1;
run;
```

```
proc surveyfreq data=pumf_unemp varmethod=bootstrap;
  tables prov*unemployed / nofreq oneway CL CV;
  weight finalwt;
  repweight bw1-bw&b.;
  where lfsstat in ('1','2','3'); *unemployment rate calculated from only
                                   those in labour force;
run;

proc surveyfreq data=pumf_unemp varmethod=bootstrap;
  tables unemployed / nofreq oneway CL CV;
  weight finalwt;
  repweight bw1-bw&b.;
  where lfsstat in ('1','2','3'); *unemployment rate calculated from only
                                   those in labour force;
by prov;
run;


/* Variance of means for numerical variables */
/* Example: Hourly earnings by province */

proc surveymeans data=pumf_results varmethod=bootstrap CV;
  var hrlyearn;
  weight finalwt;
  repweight bw1-bw&b.;
  by prov;
  where cowmain in ('1','2') and lfsstat in ('1','2'); *hrlyearn is only for
                                                        employed employees;
run;
```

D2: Example code in R

```r
#===========================================================================
# SAMPLE CODE FOR POISSON BOOTSTRAP ON THE LFS PUMF
# This program creates a file of 1,000 poisson bootstrap weights for use
# with the LFS PUMF, as well as shows a couple examples of how to
# calculate the variance of some estimates

# Input_data = an LFS PUMF file already loaded into environment
# seed = number so that you can reproduce the random results

# Required packages:
# dplyr
# tidyr
#===========================================================================

# load required packages
library(dplyr)
library(tidyr)

# Set seed and number of replicates here
seed = 1234
reps = 10 # Number of bootstrap replicates. Use 10 for testing and learning purposes. Use 1000 
for production of variance estimates.

# done in a series of functions, run the example under each one to follow    along

# As a pre-step for the calibration of bootstrap weights, define the
# calibration age groups as defined in Appendix A of the LFS PUMF user guide
## Temporarily map levels 1 and 2 of age_6 into levels 0 and 1 of age_12
## (splitting age_12 15-19 age bracket into 15-16 and 17-19),
## then convert to factor, merging levels into 10-year brackets for 35-44 and 45-54.

prep_data <- function(pumf) {
  pumf$age_cal <- factor(
    ifelse(
      pumf$age_6 %in% 1:2,
      as.numeric(pumf$age_6)-1,
      as.numeric(pumf$age_12)
    ),
    levels = 0:12,
    labels = c(
      "15-16", "17-19", "20-24", "25-29", "30-34",
      rep("35-44", 2), rep("45-54", 2),
      "55-59", "60-64", "65-69", "70+"
    )
  )
  pumf
}

pumf <- prep_data(Input_data)


# function to create poisson factors for one replicate
sample_poisson_factors <- function(input) {
  sample(c(-1, 1), length(input), replace = TRUE) # random and independent
}

# example to view poisson factors
# list of 1 or -1 for each record and one replicate
poisson_factors <- sample_poisson_factors(pumf$finalwt)
```

```
# function to calculate the uncalibrated bootstrap weights
# (using the sample_poisson_factors function above)
generate_replicates <- function(final_weight, n_reps, seed_value) {
  adjustment_factors <- final_weight * sqrt((final_weight - 1) / final_weight) # equation (1)
  set.seed(seed_value)

  replicate(
    n_reps,
    final_weight + sample_poisson_factors(final_weight) * adjustment_factors,   # equation (1)
+ (2)
    simplify = "array"
  ) |> as.matrix()
}

# example
uncal_bsw <- generate_replicates(pumf$finalwt, 10, seed)

# function to calibrate each bootstrap replicate to the sums of finalwts by  domain
calibrate_weights <- function(uncalibrated_weights, final_weight, domains) {
  domain_indices <- split(seq_len(nrow(uncalibrated_weights)), domains)  #location on the pumf
file where the groups are

  domain_fw_totals <- domain_indices |>   # pull out all the ones in a group and sum the final
wt
    sapply(function(x) sum(final_weight[x]))

  domain_bs_totals <- domain_indices |>          # pull out all the ones in a group and sum th
e bootstrap weights, for each replicate
    sapply(function(x) colSums(uncalibrated_weights[x, ]))

  domain_scaling_factors <- domain_fw_totals / t(domain_bs_totals)  # matrix transpose

  uncalibrated_weights * domain_scaling_factors[domains, ]
}

# example
cal_bsw <- calibrate_weights(uncal_bsw, pumf$finalwt, interaction(pumf$prov, pumf$gender, pumf
$age_cal))


# Final function that puts it all together
generate_bootstrap_weights <- function(d, n_reps, seed_value) {
  uncalibrated_weights <- generate_replicates(d$finalwt, n_reps, seed_value)
  domains <- interaction(d$prov, d$gender, d$age_cal)
  calibrate_weights(uncalibrated_weights, d$finalwt, domains)
}

# example
final_bs <- pumf |>
  dplyr::mutate(bswt = generate_bootstrap_weights(d=pumf, n_reps=reps, seed_value=seed))


##############################################################

#Example of using bootstrap weights to calculate variance

# define indicators
final_bs$employed <- final_bs$lfsstat %in% 1:2
final_bs$unemployed <- final_bs$lfsstat %in% 3
```

```r
final_bs$nilf <- final_bs$lfsstat %in% 4

# Estimates of totals function
bs_total <- function(bootstrap_weights, final_weights) {
  est_fw <- sum(final_weights)
  est_bs <- colSums(bootstrap_weights)
  bs_var <- mean((est_bs - est_fw)^2)
  bs_sd <- sqrt(bs_var)
  bs_cv <- ifelse(est_fw != 0, abs(bs_sd / est_fw), 0)
  data.frame(
    est = est_fw,
    var = bs_var,
    sd = bs_sd,
    cv = bs_cv * 100,
    lb = est_fw - qnorm(0.975) * bs_sd,
    ub = est_fw + qnorm(0.975) * bs_sd,
    lbq = quantile(est_bs, 0.025),
    ubq = quantile(est_bs, 0.975),
    lbq2 = quantile(est_bs, 0.025, type=2),
    ubq2 = quantile(est_bs, 0.975, type=2)

  )
}

# Example of calculating total of unemployed by province using above function
unemp_by_prov <- function(bw_file) {
  res <- bw_file |>
    dplyr::filter(unemployed) |>
    dplyr::group_by(prov) |>
    dplyr::summarize(
      est = bs_total(bswt, finalwt),
      .groups = "drop"
    ) |>
    tidyr::unpack(cols = est, names_sep = "_")
  res
}

results <- unemp_by_prov(final_bs)

# Estimates of ratios / proportions
bs_ratio <- function(bootstrap_weights, final_weights, num) {
  final_weights_num = case_when(num==TRUE ~ final_weights, TRUE ~ 0)
  est_fw_num <- sum(final_weights_num)
  est_fw_den <- sum(final_weights)
  est_fw <- est_fw_num / est_fw_den
  bootstrap_weights_num = case_when(num==TRUE ~ bootstrap_weights, TRUE ~ 0)
  est_bs_num <- colSums(bootstrap_weights_num)
  est_bs_den <- colSums(bootstrap_weights)
  est_bs <- est_bs_num/est_bs_den
  bs_var <- mean((est_bs - est_fw)^2)
  bs_sd <- sqrt(bs_var)
  bs_cv <- ifelse(est_fw != 0, abs(bs_sd / est_fw), 0)
  data.frame(
    est = est_fw,
    var = bs_var,
    sd = bs_sd,
    cv = bs_cv * 100,
    lb = est_fw - qnorm(0.975) * bs_sd,
    ub = est_fw + qnorm(0.975) * bs_sd,
    lbq = quantile(est_bs, 0.025),
    ubq = quantile(est_bs, 0.975),
    lbq2 = quantile(est_bs, 0.025, type=2),
    ubq2 = quantile(est_bs, 0.975, type=2)

  )
}

# Example of unemployment rate by province using above function
unemprate_by_prov <- function(bw_file) {
  res <- bw_file |>
    dplyr::filter(!nilf) |>  #calculate unemployment rate on those in labour force
    dplyr::group_by(prov) |>
    dplyr::summarize(
      est = bs_ratio(bswt, finalwt, unemployed),
      .groups = "drop"
    ) |>
    tidyr::unpack(cols = est, names_sep = "_")
  res
}

results_ratio <- unemprate_by_prov(final_bs)
```

References

Beaumont, J.-F., & Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80(1), 127-148.

Statistics Canada. (2017, 12 21). *Methodology of the Canadian Labour Force Survey.* Retrieved from Statistics Canada: https://www150.statcan.gc.ca/n1/pub/71-526-x/71-526-x2017001-eng.htm