FON: T1 Final Income Statistics - User Guide

Written by: Alexander Hempel

1 Introduction

The T1 Final Statistics provided by the Canada Revenue Agency serve as a snapshot into the lives of Canadians and their behaviour as reported by tax forms. These data allow one to track how certain line items of the tax form are used over time, across provinces and by different income levels. This is the only public source available to find line by line detail of the various tax credits and deductions in Canada. In addition, these are the official counts, unreliant on survey methodology and misreporting.

The downside is that the files are not linked across years and so studying trends in these variables is a challenge. This project has sought to rectify this in two ways. First, the line item names have changed over the years and so a consistent set of names has been created. In some cases, line items are lumped together into one category in some years and separated in others. Rows have been added to allow for consistent comparison over time in these cases.

Second, for the distributional data, only nominal, static income bins are reported in each year (eg. $10,000 to $15,000, $250,000 and over etc.) which makes comparisons over time by income level somewhat challenging due to compositional concerns. If the group of people making more than $250,000 has changed substantially over the years (as it has), then comparing the behaviour of those in these bins across time makes little sense. To solve this, we use an linear interpolation strategy to infer the values at a set of designated percentiles: vingtiles - every 5% and quintiles - every 20%. This lets us compare by a common group of people in the relative income distribution over time; this is particularly useful for studying the top 1% for example.

2 Raw Data

For the tax years 2010 to 2017, data is available publicly on the CRA website here. For previous years (1992-2008), data was provided by the CRA for Milligan and Smart (CJE, 2015) which has been used again in this project.

3 Creating the Dataset

3.1 Crosswalking the Data

The first major piece of work required to create this dataset is to match names across the years. In some cases, the names differ only by some small typo, but in others the name is completely different or is comprised of multiple lines for some years, but is split up in others. The general principle used to match line items was to only match line items that comprised of the exact same box from the tax form. This way, comparisons across time are most clear. What is not differentiated is when policies related to a line item are changed. Discontinuities in the data of a line item associated with policy changes will be left to the user to analyze.

This approach is useful because it means that we are always comparing the same thing over time and the original data can always be recreated. However, one issue is that this can leave some years blank because the raw data are aggregated or disaggregated in some years, but not others. To overcome this, additional line items are added to the dataset with the complete time series for some of these aggregated categories. This allows users to see both the reported components as they are in the original data as well as a more usable complete time series of the aggregated category.

An illustrative example: the Disability Amount prior to 2006 was reported with both the amounts for oneself and the amounts transferred from a dependant as a single item. Afterwards, these two components were split up. To create a consistent series across time, an additional item was created called Disability Amount (Agg.), which captures the complete time series of these combined variables. To ensure transparency and greater detail, the sub-components remain in the dataset. This exercise is repeated for a few other line items as well as broader categories of income and deductions (eg. Self-Employment income, Pension income etc.).

Another issue is that not every line item is reported for every year since 1992. In some cases, this is because the policy was not in place, while in others it was reported as part of an "Other" or "Additional" category. Some effort was made to try identifying which ones fell into which category and this information is available. **Upon consultation about whether missing values are to be included, this section should be amended.**

3.2 Interpolation of Percentiles

The second major exercise to create this dataset involved estimating the data at pre-specified percentiles. The problem we have is that the nominal bins reported in the data represent a different percentile[1] of the income distribution in each year. For example, the lower threshold of the top income bin in Ontario ($250,000 and over) represented the 99.66th percentile in 1996, but the 98.98th percentile in 2014. These are not directly comparable over time. To solve this, we would like to pre-define a set of percentiles to compare over time (eg. vingtiles, the top 1% etc.).

The subsequent challenge is then what values to assign to these pre-defined percentiles. Since we do not have data by percentile, we will have to impute or estimate these values. For this project, we choose to perform a linear interpolation of the cumulative share of the item total using the nearest observed percentiles. We use the cumulative share in part because a cumulative distribution function has well defined behaviour as opposed to a density function of total dollars as an example.

To start, we will define the share of income below a certain percentile, $\tilde{y}_j$, for a given line item, z: $H_z(\tilde{y}_j)$. Let:

- y - nominal income level
- $F(y)$ - the CDF of nominal income level
- $\tilde{y} = F(y)$ - a percentile of the income distribution
- $z(\tilde{y}_j)$ - line item dollars as a function of the percentile bin with lower bound $\tilde{y}_j$
- $Z = \sum_j z(\tilde{y}_j)$ - the total amount of line item z

To compute the share of income below percentile $\tilde{y}_j$:

$$H_z(\tilde{y}_j) = \frac{\sum_{i=0}^{j-1} z(\tilde{y}_i)}{Z} \tag{1}$$

Empirically, we do not observe $H_z(\tilde{y}_j)\ \forall\ j \in [0, 100]$. In the CRA data, we observe around 20 income bin - item amount pairs per province per year. We then translate

[1] We define percentile using the total number of returns in each nominal bin as opposed to either the total number of taxable returns or the number of total income assessed

these income bins into percentiles of the distribution using the total number of returns reported in that bin, but none of them exactly correspond to the desired vingtiles. If we define the observed percentiles as $\{\tilde{y}_1, \tilde{y}_j ..., \tilde{y}_n\}_{j=1,...,j,...n}$, then we can compute the cumulative share of income below that bin, $\hat{H}_z(\tilde{y}_j)$ for each observed percentile.

To compute the values associated with a set of consistent percentiles over time, we can perform a linear interpolation. If we define the set of desired vingtiles as v_i for all $i \in \{1, ..., 20\}$, then we can compute $\hat{H}_z(v_i)$ for a desired vingtile using a standard linear interpolation technique:[2]

$$\hat{H}_z(v_i) = \mathbb{1}(\tilde{y}_j < v_i < \tilde{y}_{j+1}) \left[H(\tilde{y}_j) + \frac{v_i - \tilde{y}_j}{\tilde{y}_{j+1} - \tilde{y}_j} (H_z(\tilde{y}_{j+1}) - H_z(\tilde{y}_j)) \right] \tag{2}$$

Once these values are estimated, we can then calculate a number of other relevant variables. First, we can compute the share of income above a given percentile as simply: $1 - \hat{H}_z(v_i)$. Second, we can compute the share in a given vingtile bin as: $\hat{B}_z(v_i) = \hat{H}_z(v_{i+1}) - \hat{H}_z(v_i)$. Third, we can multiply the bin share by the total amount of that line item, to get the implied dollar value for that bin: $\hat{D}_z(v_i) = \hat{B}_z(v_i) * Z$. Lastly, we can deflate $\hat{D}_z(v_i)$ so that we are comparing inflation adjusted values across time rather than nominal values.

3.3 Performance of Interpolation

To verify the success of the interpolation, we have compared the total income item to the Effective Tax Rates data on the Statistics Canada site (Table 11-10-0054-01).[3] These will not be exactly the same because the definitions of income may differ slightly across the two data sources (eg. child tax credits, GST/HST credits, TFSA income etc.). These graphs may also help illustrate how the technique works in practice.

In Figure 1, we plot the CRA data with triangles and Statistics Canada (SC) data using x's for two different years. Here, we compare the lower income thresholds for the various bins (while this is slightly different from the cumulative share, the principle remains the same). We can see that the interpolation does a pretty good

[2]The indicator function here just says that we use the observed CRA percentiles just above and just below the desired vingtile to compute.

[3]Total income is the only item that can be compared because there are no other sources with information on the other items in the CRA data

job of matching up with the known SC data on these income thresholds at specified percentiles. The one issue is the top tail; replacing the top tail linear interpolation with a pareto fit is something that will be corrected moving forward. We also see how the interpolation described above works: we draw a line between the value associated with observed CRA percentiles and take the value on that line at the desired vingtile.

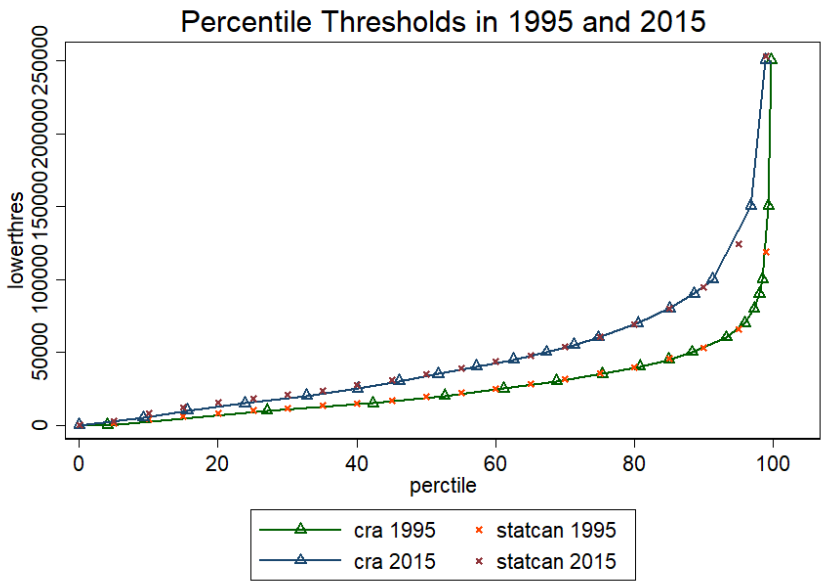

Figure 1: CRA threshold data and SC threshold data are plotted together to illustrate how linear interpolation can do a decent job of estimating the vingtile values for total income.

In Figure 2, we compare the CDF we get from interpolating the CRA total income data with the CDF from the SC data. We can see that their shapes and magnitudes are quite similar. In addition, the trend of income accruing to the Top 1% follows a similar pattern in both the SC data and the interpolated CRA data.

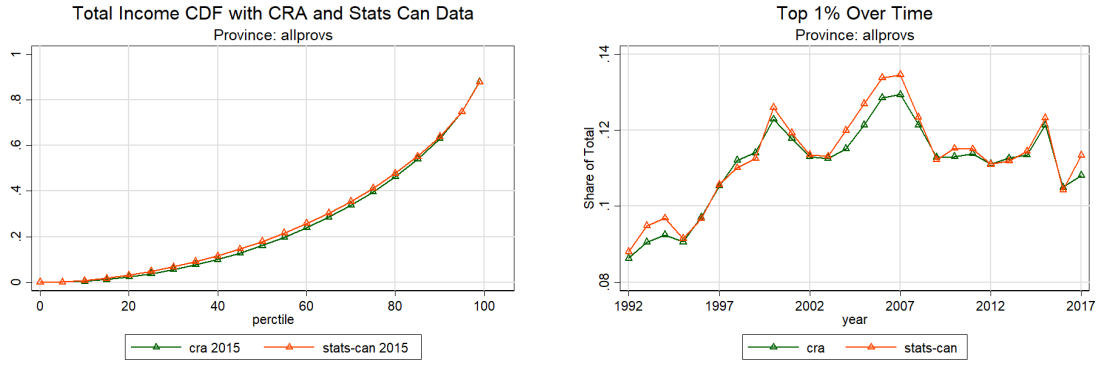

Figure 2: The CDF estimated using the linear interpolation technique is compared to the CDF from the SC data. We can also see that the trend in income going to the Top 1% also follows a similar pattern.

4 Variable Definitions

The variables seen in the public datasets are described below:

Variable	**Description**
provname	Province name
date	Year
pce	Percentile of the income distribution
itemid	A unique item identifier tied to an item name
direct	Relation to percentile (above, below, 5% bin above)
measure	Metric of data (share of total, dollars, dollars per capita)
item	The item name consistent across years
categorylab	A broad category describing the data
provcode	Province code (Statistics Canada)
provabb	Province abbreviation
val	The value associated with this row