Root N Allocation

Kimberly Fyfe

July 2020

1. Introduction

In business surveys, the first level of stratification is typically based on publication groups such as geography or industrial groupings. Within each of these groups, a second level of stratification makes use of the highly skewed nature of the population to create further groupings of small, medium and large businesses within each of the initial strata. It is within this lowest level of stratification that sample allocation will take place.

The strata containing the large units are often identified by a subject matter expert in order to ensure that the most influential units are placed into this stratum which is typically sampled at a 100% rate. The small and medium sized strata are then created using an algorithm, to which the initial sample is allocated using Root N, Neyman or Proportional allocation.

A small study based on Caribbean data was conducted to compare the results of Neyman, Proportional and Root N allocation. It was found that Neyman and Root N produced very similar results, with the Root N performing slightly better. Additionally, given that Root N is simpler to apply than Neyman and is not dependent on a design variable, it was chosen as a possible method for the Caribbean statistical offices to use when designing business surveys.

Once the initial sample is allocated, it needs to be updated to take into account subject matter knowledge such as expected response rates, importance of the strata to certain estimation domains and predetermined minimum samples sizes to ensure accurate estimation. The expected coefficients of variation (CV) should also be reviewed for the domain-based strata and further refinements should be made to ensure that the survey is expected to produce quality estimates.

2. Defining Root N Allocation

Root N allocation divides the available sample size into predetermined strata proportional to the square root of the total population. In particular:

$$n_h = n * (\sqrt{N_h} / \Sigma\sqrt{N_h})$$

Where:

- n_h is the sample size determined for a given stratum,
- n is the overall sample size to be allocated,
- $\sqrt{N_h}$ is the square root of the population size of a given stratum and
- $\Sigma\sqrt{N_h}$ is the sum of the square root of the population size over all strata involved in the allocation.

As previously mentioned, it seems to provide good results within the Caribbean context while being fairly easy to apply.

3. Implementing Root N Allocation in Excel

The following section describes the various steps involved in performing a sample allocation, along with specific information for conducting the work within an Excel workbook.

a. Preparing the Input Information

To calculate the allocation and to review the expected Coefficients of Variation (CV) the following columns of information are required:

1. Variables identifying the highest level of stratification that are typically based on key publication **domains**. Examples include geography variables and industrial sectors. To retain the correct order within the pivot tables, a counter should be added to these variables. For example:
   1_Crops
   2_Livestock
   3_Manufacturing

2. **Stratum identifiers** that contain descriptive information about the sector and the stratum. Again, to retain the correct order when using pivot tables, the strata should also contain a counter. For example:
   1_Crops_1_TA,
   1_ Crops_2_TS,
   1_ Crops_3_TF
   This type of stratum identifier tells us that there are 3 strata for the Crops sector, and the first stratum is a take-all stratum containing the largest units, the next is a take-some stratum containing the medium-sized units and the final stratum is a take-few stratum containing the smallest units.

   The strata should already be determined. For business surveys, PRASC is recommending the Horgan and Gunning method which has been supplied within another PRASC document. As previously mentioned, the units within the take-all strata should be specified by subject matter specialists as the Horgan and Gunning method does not work well for this.

3. The **design variable** that was used to determine the stratification. For business surveys this is usually a variable that provides an indication of the size of the business such as revenue, assets, wages or number of employees.

The spreadsheet should look similar to the following:

A	B	C
Sector	**Stratum**	**Design variable**
1_Crops	1_Crops_1_TF	1,500.00
1_Crops	1_Crops_1_TF	1,500.00
1_Crops	1_Crops_1_TF	1,600.00
.		
.		
.		
1_Crops	1_Crops_1_TF	4,000.00
1_Crops	1_Crops_1_TF	5,000.00
1_Crops	1_Crops_1_TF	7,200.00
1_Crops	1_Crops_1_TF	8,000.00
1_Crops	1_Crops_1_TF	11,500.00
1_Crops	1_Crops_2_TS	12,500.00
1_Crops	1_Crops_2_TS	12,500.00
1_Crops	1_Crops_2_TS	13,000.00
.		
.		
.		
1_Crops	1_Crops_2_TS	64,000.00
1_Crops	1_Crops_2_TS	78,000.00
1_Crops	1_Crops_2_TS	80,000.00
1_Crops	1_Crops_3_TA	150,000.00
1_Crops	1_Crops_3_TA	220,000.00
1_Crops	1_Crops_3_TA	350,000.00
1_Crops	1_Crops_3_TA	600,000.00

All three columns of information will be required to calculate the expected CVs while only the first 2 are needed for Root N Allocation.

It is noted that the most efficient sample designs are based on variables that have the highest correlation with the variables to be measured by the survey. For example, if the survey will measure revenue, then the best variable for determining the stratum boundaries and reviewing the expected CVs would be revenue. The less correlated the design variable, the less efficient the design will be.

b. Preparing Stratum Level Information for Allocation

To conduct the allocation, the count of units within each stratum and the overall number of units for the strata involved in the allocation are required. If these values are not already available, they can be calculated from the information mentioned in the previous section. Appendix A provides detailed instructions for using pivot tables to calculate:

Number units (or population count) in each stratum
Number of units overall

Following the instructions in Appendix A should have produced a pivot table containing the stratum identifier in column A with the label "Row Label", and the population count (N_h) for each stratum in column B. The final row is the total population count (N), which needs to be noted before moving to the next step.

Since we already know the allocation for the take-all strata, they need to be removed from the following steps and the total population count (N') should be calculated. This can be done in the pivot table by clicking on the small triangle to the right of "Row Labels" (

) and clicking to remove the take-all strata from the pivot table. To be able to manipulate these values, the columns (stratum identifier and population counts) from the pivot table need to be copied into a new table using "Paste Special" to make a copy of the actual values (Highlight the columns to be copied, then Right-click / select Copy. Move the cursor to cell A1 in a new spreadsheet and Right-click / Paste Special / Values and number formats. **Rename the new sheet to RootN Alloc**).

The square root of the population counts and their total are also needed. In Cell C3 enter the title "SQRT Count" and in Cell C4 enter:

=SQRT(B4)

Then copy the formula to the remaining rows. At the bottom, use the Sum function to calculate the sum of these values (this will be referred to as C$## in the next section).

The only piece of information that is missing in order to perform the calculations is the total sample size to be allocated. This would be calculated as:

n'=n – (N-N')

in order to remove the units which have already been allocated to the take-all strata.

c. Calculating Root N Allocation

With all of the required pieces of information known for the allocation, n' needs to be entered into the spreadsheet. In Cell A1, enter the label "Total Sample size (n'):" then in C1, enter the number of sampling units to be considered in this exercise (n'). Please note that the sample size to be allocated should be lower than the total available sample because the manual adjustments that will need to be made in the next step will cause the sample size to increase. The amount of the increase will depend on the population and the adjustments that are determined necessary in order to achieve the desired level of accuracy for the various subgroups to be published.

In Cell D4 enter the following formula and copy the formula to the other strata to calculate the preliminary sample sizes for each stratum:

=C$1*(C4/C$##)

Cell **C1** contains the sample size to be allocated to the remaining strata (n'). The "$" anchors cell reference to the first row as the formula is copied.
Cell **C4** contains the square root of the number of records in the first stratum.
Cell **C$##** is the sum of the square root of the total number of records in the strata involved in the allocation that was previously calculated. This cell reference will depend on the number of strata involved in the allocation.

With the preliminary sample sizes calculated for each stratum, two quick checks should be done to ensure that the formulae are correct. First, sum all of the allocated values; they should sum to n', in cell C1. Secondly, copy the formula to the Grand Total line; the result should also be n'. If either of these do not hold true, then the allocation needs to be reviewed for errors before proceeding to the next step.

As it is not possible to select portions of a sampling unit, the allocated values will need to be rounded to the nearest integer. In Cell E4, enter:

=INT(D4+0.5)

And copy this to the other strata, thus completing the initial sample allocation. Next, adjustments will need to be made to ensure that there are enough units to produce a good estimate.

d. Adjustment to the Allocation – Minimum Sample size

In order to have enough units to create an accurate estimate, a minimum number of records need to be selected within each stratum and the sample sizes should be adjusted for nonresponse.

Depending on the size of the strata and the budget, the minimum sample within each stratum typically ranges from 3 or 4 units, up to 10. The higher the sample size, the better the estimates will be. To have Excel apply the minimum sample size, a two-step process is specified below. The first step applies the minimum and the second ensures that it is not larger than the population size. To start, define the minimum by entering the label "Minimum Sample Size:" into Cell E1 and enter the appropriate value in Cell H1. In Cell F4, enter the following formula and copy it to the rest of the rows:

=IF(E4<H$1,H$1,E4)

In Cell G4, enter the following and copy it to all rows:

=IF(F4>B4,B4,F4)

Column G should now contain the sample sizes with the specified minimum applied. By placing the overall sample size to be allocated to the take-some and take-few strata in Cell C1 and the minimum in Cell H1, it is now quick and easy to experiment with various values to find one that works well.

e. Calculating Expected Quality

To ensure that the expected quality of the estimates is at an acceptable level and that a similar level of quality can be expected for all domains (or better for more important domains), the expected coefficient of variation (CV) should be checked. The formula for this is:

CV= (Standard Deviation / Mean)

where

Standard Deviation is the square root of the variance of the design variable. The design variable should be a good proxy for the variables to be estimated by the survey and was likely used to determine the stratum boundaries and

the **Mean** is the average of the design variable.

As described in Wikipedia, the mean and variance of a stratified random sample are given by:

$$\bar{x} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{x_h}$$

$$s_{\bar{x}}^2 = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

where,

L = number of strata

N = the sum of all stratum sizes

N_h = size of stratum h

$\bar{x_h}$ = sample mean of stratum h

n_h = number of observations in stratum h

s_h = sample standard deviation of stratum h

(https://en.wikipedia.org/wiki/Stratified_sampling)

where:

s_h^2 is $(\Sigma(x_{hi}-\mu_h)^2) / (N_h-1)$

x_{hi} is the values of the design variable falling within stratum h,

μ_h is the average for design variable values falling within stratum h,

$(N_h-n_h)/N_h$ is the Finite Population Correction factor (FPC) which is being applied due to the high sampling fractions that will likely occur and because we will be sampling without replacement.

The expected CV should be calculated for each publication domain. For example, if the first level of stratification is industrial sector followed by sub-strata based on size, it will be desirable to review the expected CV for each industrial sector. This is done by summing the Variance (adjusted for stratification) across the sub-strata and calculating the mean for the group, then applying the formula for the expected CV. As the take-all strata will contribute to the domain estimates, it is recommended that a new spreadsheet be created for reviewing the expected CVs so that the contribution of these strata can be accounted for without editing the allocation spreadsheet (in case the allocation needs to be refined). Detailed instructions for calculating the expected CVs in Excel can be found in Appendix B.

f. Refining the Allocation

The expected CVs should be reviewed to ensure that quality estimates are expected from the sample. Appendix B supplies tables used by StatCan to classify survey CVs by various quality levels. For domains of moderate importance, an expected CV of 15% (or less) should be targeted. The more important domains should be targeting at most 10%. Additionally, if the expected CVs were calculated using a proxy variable, it is possible that worse (or maybe better) CVs could be expected, and that should also be taken into account.

If acceptable CVs are being achieved, it may be possible to lower the initial sample size and still achieve acceptable results. It may be worthwhile to rerun the allocation to see if this is possible. On the other hand, if high expected CVs are being observed it will be necessary to increase the initial sample size and rerun the allocation. However, if very high expected CVs are observed for a few strata and increasing the sample size does not have much impact on the expected CVs, these strata should be reviewed again to see if it contains large records that should be promoted to Must-Takes. If a large number of records are moved to the Must-Take stratum, the stratification may need to be rerun in order to achieve an efficient sample design.

If it is necessary to lower the expected CV in a few strata, it is possible to manually adjust those sample sizes. This is often the final refinement step, especially to ensure that the most important domains will have quality estimates.

g. Adjusting the Allocation for Nonresponse

The next adjustment will be for expected nonresponse. The previous steps determined the number of observations that are required in order to obtain estimates of a certain level of quality. As it is very rare to achieve a 100% response rate, to obtain the number of observations needed to achieve the expected CVs, it will be necessary to contact more units.

If possible, the non-response rates can be taken from past surveys or results from surveys of similar types of units. As different strata usually have different response rates, the following suggests having column H contain the expected response rates for each stratum represented as a percentage. Column I would then contain the following formula to calculate the final sample sizes, rounded up to the nearest integer:

=INT(G4/H4+0.99)

As very low response rates can greatly increase the final sample size, and can also introduce a bias to estimates, it is strongly recommended that time and effort be invested into developing a collection plan that will maximize response rates. As this investment will permit a smaller sample size for collection, it will reduce collection efforts and costs. This is especially true for repeated surveys.

One of the final steps for collection should be a debriefing meeting with the interviewers to determine aspects of the collection that worked well and where improvements are needed. The interviewers are often a good source for suggestions for improvements. Consultations with international partners can also offer ideas for improving collection.

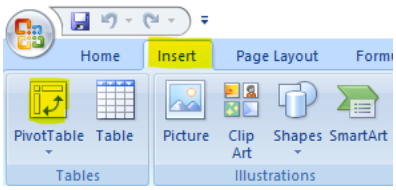
4. Concluding Remark

Sample allocation is an iterative process that often requires the allocation to be run multiple times in order to determine an optimal balance between sample size and quality. Reviewing the expected CVs can even reveal deficiencies in the assignment of the Must-Take units, which could cause the stratification step to be rerun. For these reasons, it is important to allocate sufficient time to this step of the sample design in order to allow the refinements necessary for an efficient sample design.

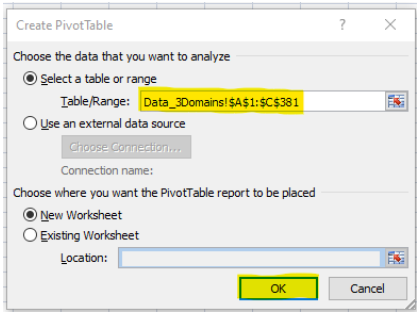
Appendix A
Generating a Pivot Table

Pivot tables within Excel are quick and easy for calculating key items for analysis. To generate the pivot table, click on Cell A1 of the sheet containing the source data, then:

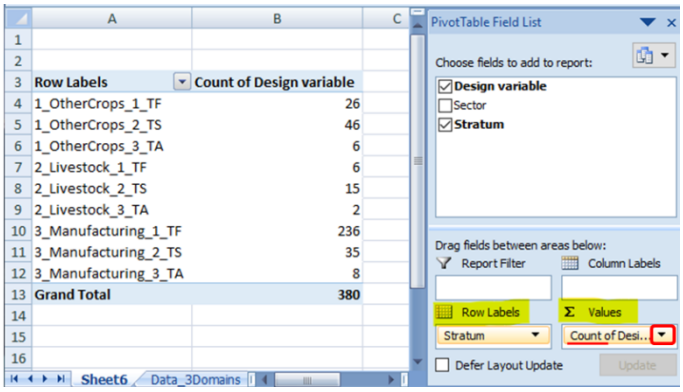
- From the top ribbon menu select *Insert* (highlighted in yellow)
    - This will bring up a new set of choices as shown in the graphic. Click on the PivotTable icon (highlighted in yellow) so that the "Create Pivot Table" window appears.



    - In the Table/ range field, ensure that all data are selected and click *OK*
        - If the correct range does not appear, the cursor was not in Cell A1 when you started or there were blank lines in the data; select *Cancel*, correct the issue and try again.
    - A new spreadsheet should appear with *PivotTable Field List* to the right as in the next image.
    - Double click on the tab name and **rename the sheet to PivotTable**.



- Click the check boxes for the **stratum identifier** and the **design variable.** The two selected variables will appear in the areas below.
- Ensure that the **stratum identifier** is in *Row Labels* and the **design variable** is in the Values field. If they are not in the correct locations, then click and drag them to the correct locations.
- Within the *Values* field, click the small triangle to the right of *Sum of* …. This will reveal a new menu. Select *Value Field Settings* at the bottom.
- From the scroll menu, select ***Count*.**
- The resulting pivot table should look similar to the example.



Within Column B, there should be a population count for each stratum and a grand total which are the required fields for the allocation:

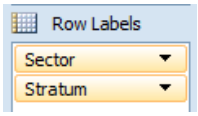
Number units (or population count) in each stratum and
Number of units overall.

With these two key pieces of information, the remaining instructions for "Preparing Stratum Level information for Allocation" can be followed.

Appendix B
Calculating Expected CVs

Returning to the pivot table that was used to obtain the counts required for the allocation, key information required for calculating the expected CVs can be generated.

If the pivot table is still filtering out the take-all strata, it needs to be changed to **include all strata**. Also, to be able to calculate the expected CV for the higher-level stratification variables (geography or industrial sector), they need to be added to the pivot table. This can be done by selecting these variables in the Pivot Table Field list. These variables will be placed in the table areas below. Make sure to drag them all under the *Row Labels* area if not already there, **ensuring that the higher levels of stratification appear before the size strata** within the *Row labels*.



Next, the calculation will require stratum-level **variances**. These can be obtained by returning to the spreadsheet with the PivotTable:

- If the PivotTable Filed list has disappeared, click on the pivot table to have them reappear to the right.
- Within the *Values* box, click the small triangle to the right of *Sum of* .... This will reveal a new menu. Select *Value Field Settings* at the bottom of the list.
- From the scroll menu, go to the bottom and select *Var* (2nd from the bottom).
- The Pivot table will be updated with the variances which will need to be copied to a new spreadsheet:
    - Highlight columns A and B
    - Right-click and select *Copy*
    - Create a new spreadsheet, and **rename it to Expected CVs**
    - Click on Cell A1
    - Right-click and select *Paste Special*; a new menu will appear
    - Select *Values and number formats*, then click *OK.*

Now the **population counts** need to be copied to the new spreadsheet called Expected CVs (N_h):

- Return to the pivot table.
- Within the *Values* field click the small triangle to the right of *Var of* ... to reveal the menu and select *Value Field Settings* again.
- From the scroll menu, select *Counts* (2nd from the top).
- Use *Paste Special* again to copy columns A and B from the pivot table into Columns D and E of the spreadsheet called **Expected CVs**, beside the copy of the variances.

Now to get the **averages**:

- Follow the same steps to calculate the averages and paste them into columns G and H in the spreadsheet called **Expected CVs**. Remember to use *Paste Special*.

Review the row identifiers to make sure that all of the stratum information is properly aligned. Once satisfied, delete the redundant and blank columns between the Variance, counts and averages.

The stratum-level variances now need to be adjusted and summed in order to obtain the sector-level variances. To do this, the allocated sample sizes need to be entered into the table. Once done, the table should look like Columns A to E in the following graphic (lines for the sector totals and Grand Total have been highlighted):

	A	B	C	D	E	F	G
1							
2					Sample	Adjusted	Expected
3	Row Labels	Var of Design variable	Count of Design variable	Average of Design variable	Size	Variance	CV
4	1_Other Crops	6042818690	78	37285.99987			0.073479
5	1_OtherCrops_1_TF	6725679.385	26	4359.230769	10	279788262	
6	1_OtherCrops_2_TS	274080055	46	28701.71391	10	45387657107	
7	1_OtherCrops_3_TA	32078955412	6	245781.525	6	0	
8	2_Livestock	2818598046	23	31755.99			0.047098
9	2_Livestock_1_TF	3713866.667	6	3053.333333	6	0	
10	2_Livestock_2_TS	157780077.2	15	24137.55467	10	1183350579	
11	2_Livestock_3_TA	12929035570	2	175002.225	2	0	
12	3_Manufacturing - Other	4.18471E+11	279	137807.9095			0.392962
13	3_Manufacturing_1_TF	517189181	236	27327.89864	12	2278390738608	
14	3_Manufacturing_2_TS	2.58281E+12	35	668380.0574	10	225996039379336	
15	3_Manufacturing_3_TA	1.23243E+12	8	1075715.083	8	0	
16	Grand Total	3.10372E+11	380	110755.5329			

To simplify the calculation, N^2 in the adjusted variance will be removed, then added back in for calculating the expected CVs.

$$s_{\bar{x}}^2 = \sum_{h=1}^{L} \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

Based on this, the formula for the adjusted variance for the first stratum will be:

=(C5^2)*((C5-E5)/C5)*(B5/E5)

And can be copied to all of the other strata.

Given the adjusted formula to simplify the calculation, the Expected CV for each domain is:

=(SQRT(SUM(F5:F7))/C4/D4)

Where cell C4 (in the red box) is the N^2 that was removed from the first formula (which is now just N because of the square root).

Note:
For the domains with **fewer than 3 strata**, a blank line will need to be added to the spreadsheet or the formula will need to be adjusted to properly sum over the strata.

The rows highlighted with yellow are the expected CVs for key domains which should be reviewed. To give the user an idea of the type of quality to target, the following table that was being used by Statistics Canada in 2009 to classify estimates gives an idea of what is considered a good CV, when expressed as a percentage:

Quality level of estimate	Guidelines
1) Acceptable	Estimates have a sample size of 30 or more, and low coefficients of variation in the range of **0.0% to 16.5%**. No warning is required.
2) Marginal	Estimates have a sample size of 30 or more, and high coefficients of variation in the range of **16.6% to 33.3%**. Estimates should be flagged with the letter M (or some similar identifier). They should be accompanied by a warning to caution subsequent users about the high levels of error, associated with the estimates.
3) Unacceptable	Estimates have a sample size of less than 30, or very high coefficients of variation in excess of **33.3%**. Statistics Canada recommends not to release estimates of unacceptable quality. However, if the user chooses to do so then estimates should be flagged with the letter U (or some similar identifier) and the following warning should accompany the estimates: "Please be warned that these estimates do not meet Statistics Canada's quality standards. Conclusions based on these data will be unreliable, and most likely invalid."

(Website address: https://www150.statcan.gc.ca/n1/pub/13f0026m/2007001/table/tab5p1-eng.htm)

Another table commonly used by Statistics Canada for classifying CVs offers more refined categories:

Rating	CV Range	Code
excellent	0.00 to 4.99	A
very good	5.00 to 9.99	B
good	10.00 to 14.99	C
acceptable	15.00 to 24.99	D
use with caution	25.00 to 34.99	E
too unreliable to be published	≥ 35.00	F

(Website address:https://www150.statcan.gc.ca/n1/pub/21f0008x/2012001/t091-eng.htm)

Additionally, since response rates can also influence the quality of an estimate, consideration is being given to including this as a factor within the quality rating. For example:

Sampling CV	Response Rates			
	100% to 90%	90% to 67%	67%-40%	<40%
0% to 5%	A	B	C	E
5% to 10%	B	C	D	E
10% to 15%	C	D	E	E
15% to 25%	D	E	E	F
25% to 50%	E	E	E	F
> 50%	F	F	F	F