Paper 829-2017

Choose Carefully! An Assessment of Different Sample Designs on Estimates of Official Statistics

Leesha R. Delatie-Budair, Statistical Institute of Jamaica; Jessica J. Campbell, Statistical Institute of Jamaica

ABSTRACT

Designing a survey is a meticulous process involving a number of steps and many complex choices. For most survey researchers, the choice of a probability or non-probability sample is somewhat simple. However, throughout the sample design process, even with probability samples, there are more complex choices—each of which may affect the survey estimates. For example, the sampling statistician must decide whether to stratify the frame. If so, s/he has to decide how many strata, whether to explicitly stratify, and how should a stratum be defined. S/He also has to decide whether to use clusters, and, if so, how to define a cluster and what should be the ideal cluster size.

The factors affecting these choices, along with the impact of different sample designs on survey estimates, are explored in this paper. The SURVEYSELECT procedure in SAS/STAT® 14.1 in SAS Enterprise Guide® Version 7.13, is used to select a number of samples based on different designs using data from Jamaica's 2011 Population and Housing Census. Census results are assumed to be equal to the true population parameter. The estimates from each selected sample are evaluated against this parameter to assess the impact of different sample designs on point estimates. Design-adjusted survey estimates are computed using the SURVEYFREQ procedure in SAS/STAT 14.1. The resultant variances are evaluated to determine the sample design that yields the most precise estimates.

BACKGROUND AND INTRODUCTION

The current Labour Force Survey (LFS) in Jamaica has a sample size of 10,464 dwellings per survey round covering 20,928 dwellings per year. With an estimated 3.1 persons per household[1] this survey covers approximately 65,000 Jamaicans each year or 2.4 per cent of the de jure population[2]. The LFS was first conducted in 1968, and has been conducted on a quarterly basis since the 1970s. Since the move to a quarterly survey, the LFS has maintained the same sample design.

The survey has a paired sample design, with two (2) primary sampling units selected from each sampling region. All dwelling units in Jamaica are assigned to Enumeration Districts (EDs) which are used either alone, or merged with other EDs to form Primary Sampling Units (PSUs). Contiguous or adjoining PSUs are placed into Sampling Regions or strata of similar size. The PSUs are joined in such a way that each sampling region:

a) Is wholly contained within one of Jamaica's 14 parishes

b) Contains approximately the same number of dwellings

c) Is composed of similar dwelling units

d) Contains only urban or only rural PSUs

The number of sampling regions varies from parish to parish because of the unequal distribution of dwellings per parish. The survey is administered with half-rotating panels, with selected dwellings staying in the survey for two (2) quarters then rotated out of the survey until the next year. The sample as a whole is replaced every three to four years in order to manage respondent fatigue and to guard against obsolescence.

The survey is however currently under review, in light of changes in the International Labor Organization's (ILO) recommendations for the measurement of the labour force (19th ICLS). The opportunity is also being

[1] The ratio of households to dwellings is 1:1.03

[2] Persons usually resident in Jamaica

taken to do a comprehensive review of the design of the LFS including the sample and survey design. A fundamental principle of Official Statistics is methodological soundness[3]. This requires periodic technical reviews of statistical methodology to ensure that they are consistent with international standards and best practices.

Official statistics also has to be responsive to the needs of policy makers and other stakeholders. STATIN currently publishes labour force estimates at a highly aggregated level[4]. In recent years, there has been increased demand for reliable estimates of the unemployment rate at lower levels of geographical and other types of disaggregation such as at the parish level. As such, one of the objectives of this review is to assess the efficiency of the existing design, and the precision of the estimates.

DESIGN OBJECTIVES OF THE LFS

The objectives of the LFS sample design is to produce a nationally representative sample that produces reliable estimates at the *95% confidence level* with a low *margin of error (<±3%)*. It is also desirable to have reliable estimates at lower levels of disaggregation. The following minimum levels of disaggregation are required:

1. Urban/ Rural (Parish preferred)
2. Sex
3. Age
4. Occupation – ISCO Major Groups
5. Industry – ISIC Sections
6. Status in Employment

IDENTIFYING ODS OUTPUT NAMES

The ODS TRACE option may be used to identify the name of the tables that are created by SAS® Procedures. This is done using the following code:

```
ods trace on;
   [Insert Proc];
   Run;
ods trace off;
```

This option writes to the SAS log the name of every table and graph produced by the PROC sequentially in the same order as the procedure output.

See "Find the ODS table names produced by any SAS procedure" by Rick Wicklin (2015) for further information (http://blogs.sas.com/content/iml/2015/09/08/ods-table-names.html)

Design Assumptions

The following assumptions inform the sample size calculation:

- **Design effect** = 2
- **Level of Confidence** = 95% i.e. $\alpha = 0.05$
- **Estimate of the key indicator to be measured by the survey (p)** = 0.5
- **Margin of Error** = ± 3%
- **Response rate =** 90%

For the purposes of this simulation exercise, the following will be held constant across designs where applicable:

- **Final Sample size** [`SAMPSIZE=10464`]: 10,464
- **Ultimate Sampling Unit [**`SAMPLINGUNIT Dwell_ID`**]**: dwellings
- **Number of stages**: 2
- **Number of clusters**: 654 PSUs
- **Number of dwellings per PSU**: 16 dwellings
- **Number of samples selected using each design** [`reps=30`]: 30

Each sampling approach will be repeated thirty (30) times in order to assess the distribution of the results to evaluate which sampling options yields the most accurate and reliable estimates.

[3] UN Fundamental Principles of Official Statistics, 2013
[4] Urban and Rural Areas only

THE POPULATION PARAMETER

For the purposes of this paper, it is assumed that the unemployment rate computed from the Census is the true population parameter. This presents the baseline unemployment rate against which the estimates from different sample designs will be compared. The SURVEYFREQ procedure is used to compute the unemployment rate. Since the Census data (*s.census*) are weighted (*weight*) to compensate for under-coverage, this has to be taken into consideration when computing the estimate. Additionally, the confidence limits and the standard errors are also computed using the following code:

Code 1: One-way Frequency - Unemployment Rate, Census

```
PROC SURVEYFREQ DATA=s.census;
     TABLES LF / ROW CL CLWT CV CVWT;
     ODS OUTPUT OneWay=work.PopParameter;
     WEIGHT weight;
RUN;
```

Using the SURVEYFREQ procedure, a one-way frequency table is created for the LF variable which contains information on the labour force status[5] of individuals. The row percent (ROW), weighted confidence limits (CLWT), and the weighted coefficient of variation (CVWT) are requested along with the table output, to provide information on the level of variability around the point estimate. The resultant output is written to a SAS® dataset called *work.PopParameter* using the SAS ODS OUTPUT option. Table 1 presents the results by Labour Force Status[6].

Table 1: Population Parameter – Labour Force Status, Jamaica 2011

Data Summary	
Number of Observations	2,214,340
Sum of Weights	2,697,990.13

Labour Force Status	*n*	Percent *(Percent)*	Linearized SE *(StdErr)*	95% CI *(LowerCL UpperCL)*
Employed	802,197	86.0	0.0367	(85.94, 86.08)
Unemployed	127,531	**14.0**	0.0367	(13.92, 14.06)
Outside the Labour Force = 1,284,612				

The unemployment rate is also computed for each of the fourteen (14) parishes of Jamaica using domain analysis. This is achieved by a cross-tabulation of the Parish variable and the LF variable. Code 2 was used to perform domain analysis of the unemployment rate by parish.

Code 2: Cross-tabulations - Unemployment Rate by Parish

```
PROC SURVEYFREQ DATA=s.census;
     TABLES Par*LF / ROW CL CLWT CV CVWT DEFF;
     ODS OUTPUT CrossTabs=work.PopParameter_Par;
     WEIGHT weight;
RUN;
```

Presented in Table 2 is the unemployment rate by Parish:

[5] Employed = 1; Unemployed = 2; Outside the Labour Force = missing
[6] The unemployment rate is calculated as a percentage of those in the Labour Force.

Table 2: Population Parameter - Unemployment Rate by Parish

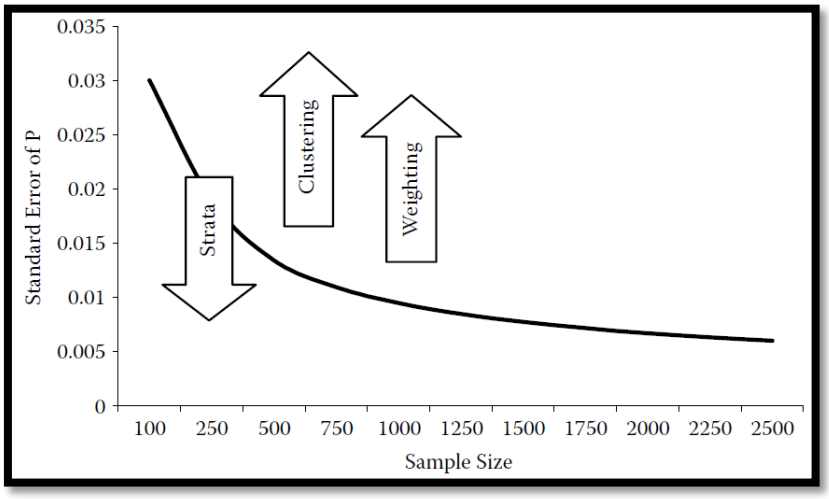
Parish	Unemployment Rate *(RowPercent)*	Linearized SE *(RowStdErr)*	95% CI *(RowLowerCL, RowUpperCL)*
Kingston	15.8	0.2039	(15.4, 16.2)
St Andrew	12.5	0.0781	(12.4, 12.7)
St Thomas	13.6	0.2011	(13.2, 13.9)
Portland	14.8	0.2285	(14.4, 15.3)
St Mary	15.4	0.1902	(15.0, 15.8)
St Ann	15.7	0.1491	(15.4, 16.0)
Trelawny	12.8	0.2037	(12.4, 13.2)
St James	14.1	0.1343	(13.8, 14.3)
Hanover	15.7	0.2184	(15.3, 16.1)
Westmoreland	14.0	0.1489	(13.7, 14.3)
St Elizabeth	12.8	0.1487	(12.5, 13.1)
Manchester	14.7	0.1470	(14.4, 15.0)
Clarendon	15.4	0.1338	(15.1, 15.7)
St Catherine	14.0	0.0814	(13.9, 14.2)

THE SAMPLE DESIGN

COMPLEX SAMPLE DESIGN ELEMENTS

In the world of official statistics, survey samples are complex. Deviations from simple random sampling (SRS) are often necessary in light of real world phenomena and practical considerations. These deviations from SRS impact estimates of variance, and as such have to be accounted for in the estimation process. In official statistics, especially for household surveys, samples are typically selected in multiple stages, employing the use of stratification, clustering and weighting.

Figure 1: Complex sample design effects on standard errors



Source: (Heeringa, West, & Berglund, 2010)

These factors all affect the standard errors of survey estimates in different ways, and as such, careful choices have to be made in each regard. Figure 1 illustrates the impact of different complex sample design features on standard errors. Stratification reduces the standard errors, while clustering and weighting inflate standard errors. The net effect of these design features is dependent on the choices made. Together, these

sample design features produce what is called the design effect, which is the net effect of the complex design on the standard errors of a survey compared to simple random sampling. It is calculated as the ratio of the variance under complex sampling to the variance under simple random sampling of the same sample size. The design effect ($deff$) in practice is often greater than one (1) indicating that standard errors are often larger under a complex design.

Multi-stage Sampling

This is the practice of selecting sampling units in different stages. The first stage includes the selection of groups of sampling elements, with each successive stage involving the selection of subsamples from the groups of elements selected in the previous stage. Selecting samples in stages allows for a more efficient administration of surveys as it relates to costs, and provides for the updating of frame information prior to second stage selection.

For the purposes of this simulation exercise, the second stage design will be held constant.

Stratification

"Strata are non-overlapping, homogeneous groupings of population elements or clusters of elements that are formed by the sample designer prior to the selection of the probability sample." (Heeringa, West, & Berglund, 2010). Strata may be defined at any stage of sampling, and are typically based on geographic areas for which representative estimates are desired.

The choice of allocation method is a choice between the precision of national estimates relative to strata estimates.

With stratification, the sampling statistician is able to decide how many units are selected from each stratum. This allows for the design of samples that provide for the representativeness of important sub-population groups that may otherwise be underrepresented in the sample. The choice of allocation method is an important step in the sample design process. Equal allocation allows for the selection of the same number of sampling units from each stratum, but would require large weights to correct this distortion if stratum sizes are significantly different. On the other hand, proportionate allocation may result in the lack of representativeness for small strata.

There are however, many alternatives that result in distributions that fall between these two extremes. The choice of allocation method is a choice between the precision of national estimates relative to strata estimates. This is termed the allocation parameter, and ranges from 0 to 1, with numbers closer to 0 assigning greater importance to strata estimates. Equal allocation has an allocation parameter = 0, while proportional allocation has an allocation parameter = 1.

The choice of allocation methods is dependent on the desired level of analysis. For example, if the objective of the survey is to inform policy at the strata level only, then an equal allocation approach would be appropriate. On the other hand, if the objectives are to inform policy and decision making at the national level only, then the proportional allocation is appropriate. In the realm of official statistics however, the analytical objectives often require estimates at both the strata and national level. The choice to be made is the relative weight given to strata as opposed to national level estimates. This paper examines allocation approaches giving full weight to national estimates (proportional), equal weight to both national and strata estimates (square-root), and the current design (paired) which falls between these two options.

Clustering

Clusters are groups of sampling elements with similar characteristics and are typically defined based on geographic demarcation. While clusters increase the standard error of estimates, they reduce the cost of data collection and allow for easier survey management and administration. By design, clusters are selected in such a way that they are relatively homogeneous within and heterogeneous without. That is, the sampling elements within clusters tend to be somewhat similar to each other, but different from sampling elements in other clusters (United Nations, 2005).

The survey statistician therefore also has to be mindful of the level of similarity or intra-class correlation within clusters. The design effect from clustering the sample within PSUs depends on two factors: the

subsample size within selected PSUs (b) and the intra-class correlation (ρ). For the purposes of this simulation exercise, the subsample size within selected PSUs will be held constant at sixteen (16) dwellings per PSU to control for the cluster effect.

Weighting

There are different types of weights that are used in survey estimation under a complex sample design. These include: design weights, non-response weights and post-stratification weights. These weights are used to adjust the distribution of the sample to match the population distribution to compensate for distortions introduced by the sample design and non-response.

For the purpose of this simulation exercise, only design weights will be applied. Design Weights are the inverse of the probability of selection. In multi-stage selection, this is found by taking the inverse of the product of the probability of selection at each stage. Equation 1 presents the formula used to compute the design weights. The first half of the equation represents the inverse of the probability of selection at stage one (1) while the second half is the inverse of the probability of selection at stage two (2).

Equation 1: Design Weight of the *i*th dwelling in the *j*th PSU in Stratum *d*

$$W_{dji} = \frac{\sum_d H}{P_d \times H_{dj}} \times \frac{H_{dj}}{k}$$

Where:

- P_d is the total number of primary sampling units to be selected
- H_{dj} is the total number of dwellings in the *j*th PSU in strata *d*
- $\sum_d H$ is the total number of dwellings in stratum *d*
- k the number of dwellings selected per PSU

ALTERNATE SAMPLE DESIGNS

SIMPLE RANDOM SAMPLING

Simple random sampling (SRS) is the most basic approach to sample designs. However, as it relates to the design of samples for real world surveys, SRS remains more of a theoretical construct than an actual option. SRS however provides an ideal against which the results of complex samples may be evaluated for efficiency.

Code 3 was used to select thirty (30) samples of 10,464 dwellings each, using SRS [METHOD=srs]:

Code 3: SRS Selection

```
PROC SURVEYSELECT DATA= s.CENSUS
     METHOD= srs SAMPSIZE=10464 REPS=30 SEED=1984
     OUT= s.sample_SRS OUTSIZE STATS;
     SAMPLINGUNIT Dwell_ID;
RUN;
```

A random number seed [SEED=1984] is used to ensure that the results can be reproduced. The results of the sample selection procedure is saved to a new file [OUT=] along with design and sampling frame information and the selection probability and sampling weight.

CURRENT SAMPLE DESIGN – PAIRED SELECTIONS

The LFS uses a two-stage 'paired selection design', with **two** Primary Sampling Units (PSUs) selected from **each** stratum or Sampling Region at the first stage. A total of thirty-two (32) dwellings are selected from each PSU, during the second stage of selection, of which sixteen (16) are surveyed each quarter. The design presented below has been modified to reflect what obtains for one quarter of the survey.

First Stage - Selection of Primary Sampling Units (PSUs), Paired Selections

The current labour force sample has three hundred and twenty-seven (327) strata/ sampling regions and 654 clusters/PSUs. From each sampling region, two PSUs are selected with probability proportionate to size (*pps*), using the number of dwellings as the measure of size.

Code 4 presents the code used for this stage of selection using paired selections. The code has been modified to produce thirty (30) samples of the same design [`REPS=30` *is omitted for routine selection*]:

Code 4: First-stage Selection – Paired Selections (30 Samples)

```
/*Sort data by strata variable*/
PROC SORT DATA=s.census;
      BY sr_id;
RUN;


/*Select 30 First Stage Samples*/
PROC SURVEYSELECT DATA= s.census METHOD=pps SAMPSIZE=2
      REPS=30 SEED=1984 OUT=s.Stage1 OUTSIZE STATS;
      SIZE size;
      SAMPLINGUNIT psu_id;
      STRATA sr_id;
RUN;
```

Careful note should be taken of the variable used to measure size, when selecting groups of observations. "*If you specify a SAMPLINGUNIT statement together with a SIZE statement, the procedure computes a sampling unit's size by summing the size measures of all observations that belong to the sampling unit. Alternatively, if you specify the PPS option in the SAMPLINGUNIT statement and do not specify a SIZE statement, the procedure computes sampling unit size as the number of observations in the sampling unit.*" (SAS Institute Inc, 2015). As such, your measure of size when summed should be equal to the actual size of the sampling unit. For example, if the number of dwellings per PSU is used as the measure of size, then for each PSU, when the size variable is summed, it should be equal to the total number of dwellings for each PSU. That is, the value of the size variable for each observation should be equal to the total number of dwellings in the PSU divided by the number of observations on the file for each PSU.

ALTERNATE DESIGN #1 - SQUARE-ROOT ALLOCATION

First Stage - Selection of Primary Sampling Units (PSUs), Square-Root Allocation

This alternative design is stratified by parish, resulting in a total of fourteen (14) strata. A total of 654 PSUs are selected, allocated across the fourteen parishes of Jamaica using the **square-root allocation**[7]. The square-root allocation gives equal weight to both national and strata level estimates with an allocation parameter of 0.5. Equation 2 is the formula used to compute the number of PSUs per strata:

Equation 2: Square-Root Allocation Formula

$$n_d = \frac{n \times \sqrt{Z_d}}{\sum_d \sqrt{Z_d}}$$

Where:

- n_d is the number of PSUs per strata
- n is the total number of PSUs to be selected
- Z_d is the allocation variable, i.e. the number of dwellings per stratum

This allocation method is currently unavailable in SAS and as such it was computed externally and the distribution saved to a SAS file (*work.alloc*). The `ALLOC=` option in the `STRATA` statement was set equal to a SAS® Dataset, with the variable name for the stratum allocation proportions being **_ALLOC_.**

[7] (ILO-IPEC, 2014)

Additionally, the dataset contained all stratum groups, and the strata variable was given the same name and type. Code 5 was used for the first stage selection for using the square-root allocation.

Code 5 presents the code used for this stage of selection using square-root allocation. The code has been modified to produce thirty (30) samples using this design [REPS=30 *is omitted for routine selection*]:

Code 5: First-stage Selection - Square-root Allocation (30 Samples)

```
/*Sort data by strata variable*/
PROC SORT DATA=s.census;
      BY PAR PSU_ID;
RUN;

/*Select 30 First Stage Samples*/
PROC SURVEYSELECT DATA= s.CENSUS       METHOD=pps
      SAMPSIZE=654 REPS=30 SEED=1984
      OUT=s.Stage1 OUTSIZE STATS;
      SIZE size;
      SAMPLINGUNIT psu_id;
      STRATA PAR / ALLOC=work.alloc;
RUN;
```

ALTERNATE DESIGN #2 - PROPORTIONAL ALLOCATION

First Stage - Selection of Primary Sampling Units (PSUs), Proportional Allocation

The alternative design is stratified by parish, resulting in a total of fourteen (14) strata. A total of 654 PSUs are selected, allocated across the fourteen parishes of Jamaica using proportional allocation[8]. Under this allocation method, the allocation proportion is equal to one (1), implying that the design gives no weight to strata level estimates. This is the antithesis of the equal allocation method which gives no weight to national estimates, relative to strata estimates.

Equation 3 represents the formula used to compute the number of PSUs per strata.

Equation 3: Proportional Allocation Formula

$$n_d = \frac{N_d}{N} \times n$$

Where:

- N_d is the total number of dwellings in each strata
- N is the total number of dwellings
- n_d is the number of PSUs per strata
- n is the total number of PSUs to be selected

When allocating samples across strata, SAS® employs a rounding algorithm to ensure that the number of PSUs selected per strata is a whole number and that at least one (1) PSU is selected per stratum.

Code 6 presents the code used for this stage of selection using proportional allocation. The code has been modified to produce thirty (30) samples using this design [REPS=30 *is omitted for routine selection*]:

[8] (ILO-IPEC, 2014)

Code 6: First-stage Selection - Proportional Allocation (30 Samples)

```
/*Sort data by strata variable*/
PROC SORT DATA=s.census;
      BY par psu_id;
RUN;

/*Select 30 First Stage Samples*/
PROC SURVEYSELECT DATA= s.CENSUS       METHOD=pps
      SAMPSIZE=654 REPS=30 SEED=1984
      OUT=s.Stage1 OUTSIZE STATS;
      SIZE size;
      SAMPLINGUNIT psu_id;
      STRATA PAR / ALLOC=prop;
RUN;
```

Presented in Table 3 is the first-stage sample distribution by parish.

Table 3: First-stage Sample Distribution by Parish

Parish	Paired Selections		Proportional allocation		Square-Root allocation	
	# of PSUs	Proportion	# of PSUs	Proportion	# of PSUs	Proportion
Kingston	38	0.058	22	0.034	34	0.052
St Andrew	128	0.196	142	0.216	86	0.132
St Thomas	20	0.031	23	0.036	35	0.054
Portland	36	0.055	21	0.032	33	0.050
St Mary	24	0.037	27	0.042	38	0.058
St Ann	36	0.055	40	0.061	46	0.070
Trelawny	32	0.049	19	0.029	32	0.048
St James	40	0.061	45	0.069	49	0.074
Hanover	32	0.049	18	0.027	31	0.047
Westmoreland	34	0.052	38	0.058	45	0.068
St Elizabeth	34	0.052	37	0.056	44	0.067
Manchester	40	0.061	45	0.069	49	0.074
Clarendon	52	0.080	57	0.087	55	0.084
St Catherine	108	0.165	120	0.184	79	0.121
Total	**654**	**1.000**	**654**	**1.000**	**654**	**1.000**

SECOND STAGE – SELECTION OF DWELLINGS

At the second stage of selection, sixteen (16) dwellings are selected from each PSU systematically with a random start. Prior to selection, the sampling information variables from the first-stage selection are renamed to ensure that they are not overwritten during the second-stage of selection.

Additionally, the PSUs selected in the first-stage are treated as strata in the second-stage to ensure that the selection of dwellings is from each PSU. The documentation for the SURVEYSELECT procedure states that "*When you use a SAMPLINGUNIT statement, PROC SURVEYSELECT does not select samples of observations from within the sampling units (clusters). To select independent samples within groups, use the STRATA statement*" (SAS Institute Inc, 2015).

Code 7 is used for second stage selection under each design:

Code 7: Second-stage Selection

```
/*Rename stage 1 sample information*/
```

```
DATA stage1; SET s.stage1;
RENAME      Replicate        =     SampleNum
            UnitSize         =     Stage1_UnitSize
            TotalSize        =     Stage1_TotalSize
            SampleSize       =     Stage1_SampleSize
            Total            =     Stage1_Total
            SelectionProb    =     Stage1_SelectionProb
           SamplingWeight    =     Stage1_SamplingWeight;
RUN;

/*Sort data by cluster variable for each sample*/
PROC SORT DATA=stage1;
      BY SampleNum psu_id;
RUN;

/*Select Second Stage Sample within clusters*/
PROC SURVEYSELECT DATA= stage1 METHOD=sys
      SAMPSIZE=16 SEED=1984
      OUT=s.Stage2 OUTSIZE STATS;
      SAMPLINGUNIT Dwell_ID;
      STRATA SampleNum psu_id;
run;
```

COMPARATIVE ASSESSMENT OF THE UNEMPLOYMENT RATE UNDER ALTERNATE DESIGNS

POINT ESTIMATE – MEAN UNEMPLOYMENT RATE AND MEAN CV

The mean unemployment rate for the thirty (30) samples of each design is computed along with standard errors and coefficient of variation. The results are then compared to the population parameters to assess the accuracy and reliability of the estimates. Both the paired selection design and the square-root allocation introduce a disproportionate distribution of the sample. This implies a compromise between national estimates and strata estimates, with some importance placed on parish level estimates i.e. an allocation parameter less than (1) and greater than zero (0). On the other hand, the proportionate allocation places no relative importance on strata level estimates, which have proven problematic for smaller strata. Table 4 presents the Mean Unemployment Rate over thirty (30) samples for each design along with the mean coefficient of variation (CV) over these samples.

Code 8 is used to compute estimates:

Code 8 is used to compute the design weights for analysis (see Equation 1), labour force status (i.e. the proportions employed and unemployed) and to filter off the unemployment rate for further analysis. The TABULATE procedure is used to compute the mean Unemployment Rate per parish and the Mean CVs obtained for each design.

The results are then compared to the population parameters to assess the accuracy and reliability of the estimates. Both the paired selection design and the square-root allocation introduce a disproportionate distribution of the sample. This implies a compromise between national estimates and strata estimates, with some importance placed on parish level estimates i.e. an allocation parameter less than (1) and greater than zero (0). On the other hand, the proportionate allocation places no relative importance on strata level estimates, which have proven problematic for smaller strata. Table 4 presents the Mean Unemployment Rate over thirty (30) samples for each design along with the mean coefficient of variation (CV) over these samples.

Code 8 is used to compute estimates:

Code 8: Point estimate

```
/* Compute Design Weights*/
DATA s.Stage2; SET s.Stage2;
      prob       = Stage1_SelectionProb * SelectionProb;
      SampleWgt  = 1/prob;
RUN;

/* Compute Unemployment Rate*/
PROC SURVEYFREQ DATA=s.Stage2;
      TABLES Par* SampleNum*LF / ROW CL CLWT CV CVWT DEFF;
      ODS OUTPUT CrossTabs=work.Rates;
      WEIGHT SampleWgt;
RUN;


/* Filter Unemployment Rate and rename variable*/
DATA s.estimates; SET work.Rates;
      WHERE lf=2;
      RENAME RowPercent=UnempRate;
RUN;

/*Tabulate mean Rate and CV by Parish*/
PROC TABULATE
      DATA=S.ESTIMATES_CURRENT;
      VAR UnempRate RowCV;
      CLASS Par /     ORDER=UNFORMATTED MISSING;
      TABLE /* Row Dimension */ Par,
      /* Column Dimension */ UnempRate* Mean
      RowCV* Mean;
            ;
RUN;
```

Table 4: Mean Unemployment Rate and CVs by Parish, Census and Alternate Designs

Parish	Census		Paired Selections		Square-Root Allocation		Proportional Allocation	
	Rate	CV	Rate	CV	Rate	CV	Rate	CV
Kingston	15.8	0.013	15.2	0.090	15.2	0.090	15.9	0.100
St Andrew	12.5	0.006	12.4	0.050	12.0	0.060	12.2	0.050
St Thomas	13.5	0.015	13.1	0.130	13.0	0.100	13.5	0.120
Portland	14.8	0.015	15.1	0.090	14.7	0.100	14.4	0.120
St Mary	15.4	0.012	15.6	0.110	14.7	0.090	14.7	0.100
St Ann	15.7	0.010	14.6	0.090	15.0	0.070	15.0	0.080
Trelawny	12.8	0.016	12.5	0.100	11.8	0.110	12.5	0.130
St James	14.1	0.010	14.2	0.080	14.0	0.080	13.7	0.080
Hanover	15.7	0.014	15.3	0.090	15.1	0.100	15.7	0.120
Westmoreland	14.0	0.011	13.6	0.100	13.7	0.090	13.2	0.090
St Elizabeth	12.8	0.012	12.6	0.100	12.8	0.090	12.4	0.100
Manchester	14.7	0.010	14.8	0.080	14.5	0.080	14.1	0.080
Clarendon	15.4	0.009	15.0	0.080	15.0	0.070	14.8	0.070
St Catherine	14.0	0.006	13.7	0.050	13.8	0.060	13.6	0.050
Jamaica	**14.0**	**0.003**	**13.8**	**0.022**	**13.7**	**0.022**	**14.0**	**0.017**

The results show that on average, the different designs yielded marginally different rates from the population parameter, some of which were found to be statistically significant. The paired selections design yielded an overall rate that was 0.2 percentage points lower than the population parameters. This difference was found to be statistically significant. The results also show that the estimates under this design for seven (7) of the fourteen (14) parishes were statistically significantly lower than the population parameter.

The square-root allocation exhibited a slightly larger downward bias, with rates at an average of 0.4 percentage points lower than the population parameters. Additionally, all estimates for all fourteen parishes and for Jamaica, were lower than the population parameter; these differences were all found to be statistically significant.

The proportional allocation however resulted in an overall unemployment rate that was statistically equal to the population parameter. However, the results show differences that are statistically lower rates for nine (9) of the fourteen parishes.

In terms of accuracy, the paired selections yielded the most accurate results at the parish level, while the proportional allocation yielded the most accurate results at the national level over thirty (30) samples.

In terms of precision, at a threshold of 10% for CVs, on average over thirty (30) samples, the square-root allocation resulted in the most precise estimates. On average, the CVs obtained fell below the 10% threshold, except for one (1) small parish. The average CVs for the estimates from the paired selection fell below the threshold for all but two (2) parishes. The proportional allocation resulted in average CVs falling above the threshold for four (4) of ten (10) parishes.

DISTRIBUTION OF POINT ESTIMATES – UNEMPLOYMENT RATE

Presented in Figure 2 are the distribution of unemployment rates calculated for thirty (30) samples by parish for each sample design. The BOXPLOT option under the GRAPH Task was used to generate the charts. Code 9 was inserted into the system generated code for Axis 1, in order to standardize the axis range to produce comparable graphs:

Code 9: Custom Code to Control y-axis range

```
      order=(6 to 22 by 1)
```

Figure 2 shows the distribution of the estimates over the thirty (30) samples for each of the three (3) sample designs.

The graphs show the distribution of point estimates of the unemployment rate by parish over thirty (30) samples of each design visually. The results show a smaller spread in the distribution of rates under the square-root allocation, relative to the other two designs. This indicates that this design yields more precise estimates that are reproducible over time.

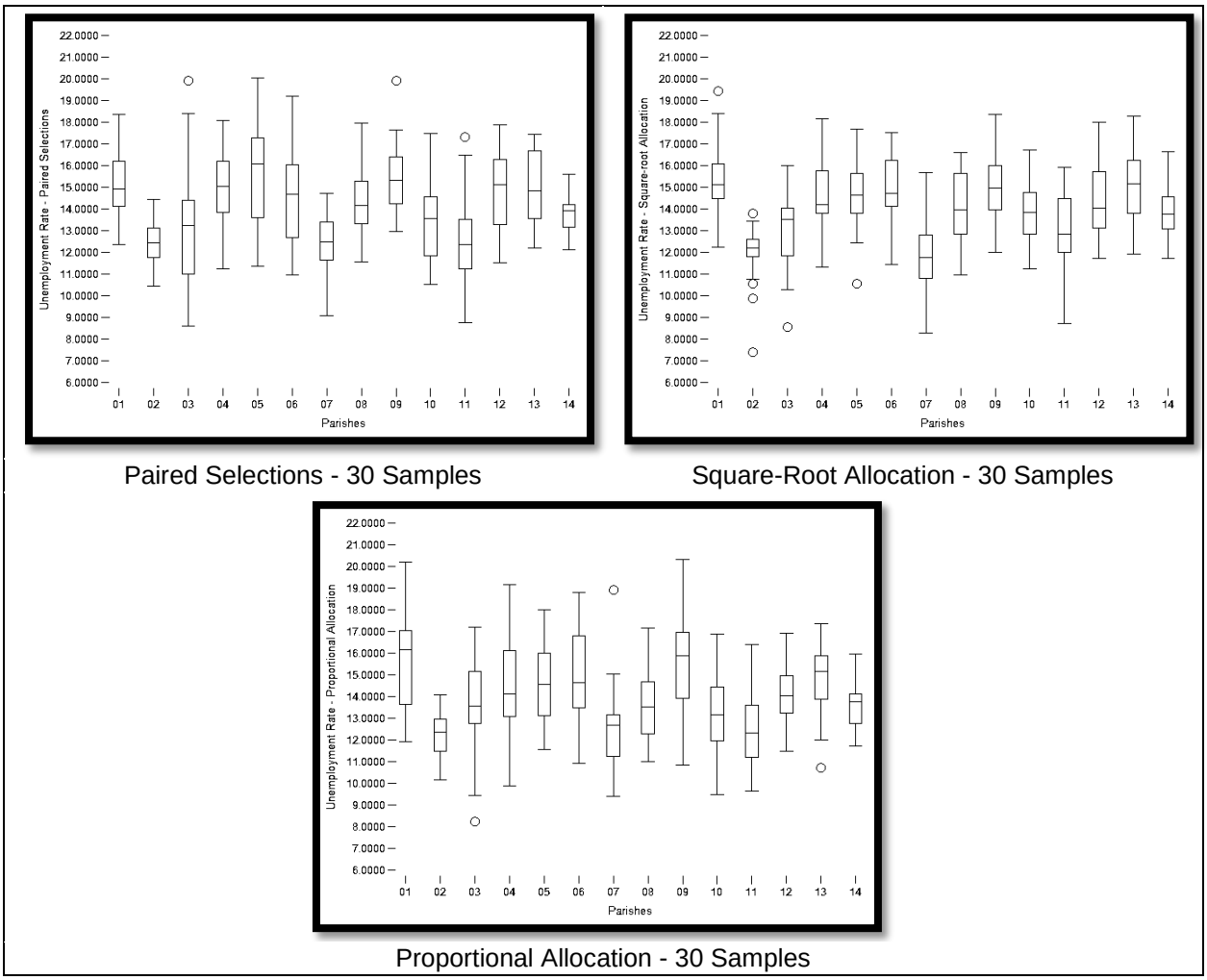


Figure 2: Distribution of Unemployment Rates by Parish and Sample Design

CONCLUSION

This simulation exercise shows that the choices made by the sampling statistician, not only affects the accuracy of the point estimate, but also the reliability of estimates over time with repeated samples.

On average, the proportional allocation yielded the most accurate estimates at the national level, while the paired selections yielded the most accurate estimates at the strata/ parish level. On the other hand however, the square-root selection yielded the most precise estimates at both the parish and national level. It should be noted that none of the three sample designs examined in this simulation exercise produced accurate estimates at the strata level for all parishes.

When designing a sample, the choices are many. These choices however should only be made after deliberation and careful thought about the analytical requirements of the survey results. The design of the sample should take into account the need for national as opposed to strata level estimates and the relative precision of the estimates. Depending on the use of the survey information, the sampling statistician can choose to accept a lower level of precision or accuracy.

In this case, more accurate information is desired at the strata level that prove reliable over time. Further choices therefore have to be made in order to arrive at a design that yields accurate and reliable estimates for all parishes and at the national level.

REFERENCES

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied Survey Data Analysis.* Boca Raton, FL 33487-2742: CRC Press, Taylor and Francis Group, LLC.

ILO-IPEC. (2014). *ILO-IPEC Interactive Sampling Tools No. 2 – Allocation of Sample Size among domains or strata.* Geneva: International Labour Office, International Programme on the Elimination of Child Labour (IPEC).

Kish, L. (1965). *Survey Sampling.* New York: John Wiley and Sons.

SAS Institute Inc. (2015). The SURVEYSELECT Procedure. In S. I. Inc, *SAS/STAT® 14.1 User's Guide* (pp. 9331-9410). Cary, NC, USA.

SAS Institute Inc. (2015). The SURVEYFREQ Procedure. In S. I. Inc., *SAS/STAT® 14.1 User's Guide* (pp. 8823-8922). Cary, NC, USA.

United Nations. (2005). *Household Sample Surveys in Developing and Transition Countries.* New York: Department of Economic and Social Affairs, Statistics Division, United Nations.

Wicklin, R. (2015, September 8). *Find the ODS table names produced by any SAS procedure*. Retrieved from SAS BLOGS: http://blogs.sas.com/content/iml/2015/09/08/ods-table-names.html

RECOMMENDED READING

- *SAS/STAT® 14.1 User's Guide*

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Leesha Delatie-Budair
Statistical Institute of Jamaica
1-876-630-1600
ldelatie-budair@statinja.gov.jm