Paper 1630-2014

Application of Survey Sampling for Quality Control

Yi Du, Freddie Mac

ABSTRACT

Sampling is widely used in different fields for quality control, population monitoring, and modeling. However, the purposes of sampling might be justified by the business scenario, such as legal or compliance needs. This paper uses one probability sampling method, stratified sampling, combined with quality control review business cost to determine an optimized procedure of sampling that satisfies both statistical selection criteria and business needs. The first step is to determine the total number of strata by grouping the strata with a small number of sample units, using box-and-whisker plots outliers as a whole. Then, the cost to review the sample in each stratum is justified by a corresponding business counter-party, which is the human working hour. Lastly, using the determined number of strata and sample review cost, optimal allocation of predetermined total sample population is applied to allocate the sample into different strata.

INTRODUCTION

In an information-intensive society, collecting and analyzing large-volume of data efficiently become critical for people to catch up with fast-moving environment. One of the methods to attain this goal is to applying sampling methodology. Stratified random sampling is one of probability sampling methods to separate the total target population into different strata by certain categorical variable or business criteria so that the population in each stratum is as homogeneous as possible while the population between strata is as heterogeneous as possible. Practically, the determination of stratum could be driven by either categorical variable or business needs to satisfy the interest of research. Furthermore, the stratified probability sampling could increase the efficiency of estimator of overall population parameters by the choice of strata and make the administration operate more easily.

In reality, we may face the problem which contains large number of strata for sampling. For example, the on-line shopping system may need to be monitored to its sales tax implementation to check if the system does reflect the real tax requirement because each state has different sales tax and each state may have different sale taxes levied to different categories of goods. This may require the stratified sampling using state combined with goods category as stratum. The third-party automotive evaluation company may care about the defect rate of brand new cars at the level of make, model even years. Under this scenario, the stratification of total new auto sales into the corresponding strata by make, model and years becomes necessary so that enough detailed information could be estimated. Census statisticians usually use regional, demographic, or socio-economic factors as strata to estimate the overall interest of statistic.

However, in reality, the selection of stratum could lead to the fact that the number of stratum is so large that the benefit from marginal increase in precision (decrease in variability) is offset by the marginal increase in cost by stratification. Furthermore, the distribution of population by stratification could be highly skewed with long tails to make judgment based on the same criterion by the stratified sampling. According to Eric Falk and Wendy Rotz 2003 Joint Statistical Meeting's survey research methods paper, the statistical goal in defining certainty strata is to identify extreme values within a population that heavily influences the estimate and its variance. They proposed that the certainty stratum cut-off value is determined to isolate outliers and is based on statistical relative precision. I extend this idea into a new data-driven stratified sampling framework. The whole population is first stratified using some targeted interest and Box-and-Whisker outliers are used to isolate the population into different parts for sampling. More detailed, the outliers as a whole are treated as an independent stratum and non-outliers are stratified as is. Then the optimal allocation algorithm is employed to allocate the total targeted sample size into each re-defined stratum in consideration the variability and sample cost of each stratum.

STATISTICAL APPROACH

Suppose the total population *N* is divided into *S* strata. Each stratum is denoted with the index *s*, *s* = 1, …, *S* and comprises N_s elements. The strata must not overlap and each element could and only could belong to exactly one stratum. The strata are mutually exclusive each other but jointly form the whole population such that:

1) $\bigcup N_s = N \text{ for } s = 1, \ldots, S$
2) $\cap\, N_s = \phi \text{ for } s = 1, \ldots, S$

The size of each stratum could be derived from the whole population by strata. In reality, this information is not always available so assume the size of each stratum is known as X_s. The following result is easily attained:

3) $\sum N_s * X_s = N \text{ for } s = 1, \ldots, S$

Once the distribution of the whole population is known, I propose to use box-and-whisker plot outliers to re-define the stratum. The main idea is to first rank order the data points from distribution generated by 3) and find out the median (Q_2) of the whole population. In each half, a 'sub-median' is found and denoted as Q_1 and Q_3. The three points equally divide the distribution into four parts and each part takes account 25% of the data points. Q_1 and Q_3 becomes the border of box part of the plot and endpoints become 'whiskers' part of the plot.

The 'interquartile range', abbreviated as 'IQR', is just the width of the box-and-whisker plot. That is, $IQR = Q_3 – Q_1$. Outliers could be defined as any data points too far away from the middle data point. Mathematically, outliers are defined as any data points beyond the range of $[Q_1 – k * IQR, Q_3 + k * IQR]$ and k is a positive constant. Employed with this idea, the stratum is re-defined such that any stratum considered as outliers would be grouped as a whole into one stratum and the rest non-outliers would still keep their own strata.

Once the number of strata is re-defined by the box-and-whisker plot, the whole population is re-stratified however the conditions 1) - 3) still hold except the number of strata would be smaller than S, denoted as T. Then how many samples should be selected in each stratum becomes the next question. In reality, a targeted total sample size may be pre-determined resulting from some possible considerations, such as overall timing issue. And the total sample size would be distributed across each stratum. According to William Cochran (1977), a proportional allocation according to the size of stratum would yield a self-weighting sample, which means the sample size in individual stratum to the total sample size is proportional to the size of stratum to the whole population. Suppose the proposed total sample size is n. Then the sample size in each stratum by proportional allocation is represented as n_p and could be derived by:

4) $n_p = n * N_t / N$

But it is possible that different strata with the same size could have difference in variability. Also, the cost of survey in different strata often varies. Then optimal allocation is introduced to allocate the total sample size into each stratum such that either the variance is minimized subject to a specified cost or the cost is minimized subject to a specified variance. To make the mathematics simple without losing any intuition, a linear function of cost across strata is assumed. The stratum survey cost is denoted as c_t and stratum variance is denoted as V_t. Then the sample size in each stratum by optimal allocation is represented as n_o and could be derived by:

5) $n_o = n * (N_t / N * V_t / \sqrt{c_t}) / \sum(N_t / N * V_t / \sqrt{c_t})$

The learning curve behind 5) is that optimal allocation strategy not only considers the weight of stratum as proportional allocation, but also introduces the stratum variance and cost as factors to allocate total sample size. Intuitively, the more population stratum has and the more variable it is, the more samples would be selected. The more expensive to sample, the few samples would be selected.

A special case of optimal allocation is when the cost in each stratum holds the same, known as Neyman allocation. Then the sample size in each stratum is represented as n_n and could be derived by 5):

6) $\mathrm{n_n} = n * (N_t / N * V_t) / \sum(N_t / N * V_t)$

From the above three allocation methods, the proportional allocation and Neyman allocation could be treated as special cases of optimal allocation. In general, the sample size in each stratum is positive correlated to the size and variability of stratum and negatively correlated to its cost given a pre-determined total sample size. Furthermore, the total sample size could also be determined. Since the optimal allocation is solved depending on a fixed total cost or a specified variance, the total sample size could be solved separately:

7) $n = (C – c_0) \sum(N_t * V_t / \sqrt{c_t}) / \sum(N_t * V_t / \sqrt{c_t})$

8) $n = (\sum N_t * V_t * \sqrt{c_t} / N)\sum N_t * V_t /(N * \sqrt{c_t}) / (V +1 / N * \sum N_t * V_t^2 / N)$

where C is the total cost and c_0 is sunk cost for sampling, V is the targeted total variance.

IMPLEMENTATION WITH MACRO

The above statistical approach uses the box-and-whisker outliers to re-define the stratification and use stratified sampling by optimal allocation.

In order to find out the box-and-whisker outliers to re-define stratum, SAS® procedure PROC UNIVARIATE are used.

PROC UNIVARIATE SYNTAX:

PROC UNIVARIATE <*options*>;

VAR *variables*;

OUTPUT <*OUT=SAS-data-set* < *keyword1=names ...keywordk=names* > < *percentile-options* >> ;

RUN;

Under the scenario here, ***variables*** are the frequency in each stratum and ***output*** statement would evoke the lower quartile (Q_1) and upper quartile (Q_3) with interquartile range (IQR) and an SAS data set would be generated to contain these information.

Data: specify the input data set against which the descriptive statistic is analyzed

Out: specify the output data set containing the descriptive statistic related to the analysis

Q1: lower quartile

Q3: upper quartile

Qrange: interquartile range ($Q_3 – Q_1$)

In order to implement optimal allocation to do stratified sampling, SAS procedure PROC SURVEYSELECT is used.

PROC SURVEYSELECT SYNTAX:

PROC SURVEYSELECT *options*;

STRATA *variables </ options>*;

RUN;

Under the scenario of stratified sampling, the total sample size needs to be specified. Also, in order to apply different allocation methods, the variance and cost by strata should be specified,

Data: specify the input data set where the sample would be selected

Out: specify the output data set containing sampling information

Seed: specify the random number seed which allows replicating the result

Method: specify the selection method

Sampsize: specify the sample size

Strata: used for stratified sampling

Alloc: specify the method for allocating the total sample size among the strata

Allocmin: specify the minimum sample size to allocate to a stratum

Var: indicates the variance in each stratum as a secondary input data set used for optimal/Neyman allocation method

Cost: indicates the cost in each stratum as a secondary input data set used for optimal allocation method

The following table shows the requirement of macro variables needed to be input into the customized SAS macro to implement the survey sample method proposed above:

Macro Variable	**Valid Value**	**Description**
POP	Valid SAS data set	The survey population is to be sampled under the analysis
STRATA	Valid SAS variable name (separated by space for multiple variables)	The strata defined by the analysis and could be univariate or multivariate
K	Positive number	The number used to calculate outliers based on quartile and inter-range quartile
ONE_SIDED	Y or N	If one-sided equals to N, the outliers would be considered by the strata with small population. Otherwise, the strata with both small population and large population would be considered.

VAR	Valid SAS variable name	The variable used to calculate the variance by strata and could be univariate or multivariate. Could be null if proportional allocation is applied
COST	Valid SAS variable name	The variables used to calculate the cost by strata and should be univariate. Could be null if proportional allocation or Neyman allocation is applied
TOT_SAMP	Positive number	The total sample size needed for survey sample
ALLOCMETH	Optimal, Neyman or Proportional	Three allocation methods used for stratified sampling
ALLOCMIN	Positive number	The minimum sample size for each stratum
OUT_POP	Valid SAS data set	The final sampled population with sampling information

Table 1. Macro Variables Used in Customized SAS Macro for Analysis

There are several places where I give users flexibility when using this customized macro to do stratified sampling under the proposed framework.

First of all, when using box-and-whisker outliers to re-stratify the whole population, users could also adjust the positive parameter *k* to get corresponding outliers. Bot-and-whisker outliers is defined when *k* is equal to *1.5*. Another more common outliers is to set *k* to *3* by the similar formula from first quartile, third quartile and inter-range quartile and it is commonly called ‘extreme value’.

Secondly, the customized SAS macro is meant to re-stratify the whole population based on the outlier’s algorithm according to the size of strata defined originally. However, it is possible to group the strata with large population into the outliers too. In reality, the analysts probably would like to keep strata with large population as is since they alone as an independent group may differentiate from other groups. Therefore, the macro variable ‘*one-sided*’ could be set to ‘*Y*’ to only group the strata with small population into one stratum.

Thirdly, the stratification could be defined by more than one identification and users could use multivariate methods to stratify the population and the strata will be calculated at the combination of multiple variables. Also, the variance by strata could be driven by more than one variable and users could enjoy the freedom to use relevant information to calculate stratum-level variance correspondingly.

APPLICATIONS IN DIFFERENT AREAS

Thus far, we discuss to apply outliers detection technique to re-stratify the population for sampling with optimal allocation strategy into different strata in consideration of differences in variability and sampling cost between strata. For quality control purposes, the information analysts are interested in deriving from survey is usually unknown and actually is the goal of survey sampling. Rather, people would use information of sample to estimate the whole population’s characteristics. For example, manufacture lines usually have quality control process to assure the product is made according to the standard of guideline. Under this scenario, the truth of matter is actually unknown which also makes the variability of the population unknown. However, some signal variables could be used as proxy for estimation purposes, such as weight or volume of product to estimate the variance. In order to apply optimal allocation strategy, the sample cost needs to be justified and review time of each product from former sample or to similar product could be used as proxy.

Moreover, the proposed sampling strategy could not only be used for quality control purposes, but also be applied in other areas as well. For example, people always have the interesting question whether the major could become a key factor to graduates’ career success. A survey could be conducted according to different groups by major. Obviously, the total number of students in each group varies a lot and there are tons of majors offered by different levels of schools. And students in each major may behave differently in different industries and areas, plus the factor that the availability of each individual student could vary a lot. When applying the proposed sampling method, the majors with small graduates could be grouped into one group as a whole and the income or frequency of promotion in the same time of period could be treated as signal variables of career success and the survey method, such as outbound calls versus mails would be used to collect the sample. The re-stratified sampling method with optimal allocation could be applied to design the sample to meet the needs. This method may be also used to estimate the health cost by age group, or TV ratings by channels and audience.

DISCUSSION

In a short, the proposed stratified survey sample framework combined with the idea of outliers is to re-stratify the survey population by grouping the outliers into one stratum and apply the corresponding sample allocation method to sample individual units among strata. In this paper, box-and-whisker outlier is used to identify the outliers. And this method is based on the assumption that the population follows a normal distribution. There are also model-based methods which also base on the normal distribution assumption to identify outliers, such as Peirce’s criterion. There is also non-parametric method to identify outliers without assuming a priori statistical distribution, such as kernel density estimation. There are also many data mining tools to identify outliers which are not relevant to the content of this paper. But more interestingly, these methods could still be applied under this framework to re-stratify the whole population to solve the problem of too many strata leading to the problem of inefficiency of using stratify sampling.

Applying this proposed sampling method in reality, the costs by strata could be measured by different ways. In census survey, the cost in different areas may be determined by how difficult to collect the information. In marketing satisfaction survey, the cost in different groups of customers may be determined by the difference in marketing solicitation channels. While in quality assurance survey of manufacturing industry, the cost of sampling in different parts units may vary because of the time to monitor the process. Or the cost could be a combination of all factors above. However, no matter what factor analysts choose to use to calculate the cost, it should be measurable equivalently and comparable among strata.

CONCLUSION

This paper considers the realistic sampling problem when the stratified sampling method is used while involving with a number of strata. In consideration of advantages of stratified sampling, to increase efficiency and accuracy, too many strata may undermine these benefits. Furthermore, survey cost may vary a lot among strata and the interest of analysis in different strata may behave differently. Combined with these two factors, the optimal allocation method is applied to allocate the total sample size into different strata. Finally, a customized SAS macro is built to implement the proposed sampling method with different SAS procedures.

REFERENCES

Cochran, William, Sampling Techniques Third Edition, New York: John Wiley & Sons.

Falk, Wendy Rotz, 2003, “Stratified Sampling for Sales and Use Tax Highly Skewed Data – Determination of the Certainty Stratum Cut-off Amount”, 2003 Joint Statistical Meetings.

Lohr, Sharon, 2010, Sampling Design and Analysis Second Edition, Boston: Brooks/Cole.

SAS on-line Documentation, SAS/STAT 13.1 User’s Guide, SAS Institute, Cary NC.

CONTACT INFORMATION

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

Name: Yi Du
Organization: Freddie Mac
Address: 1551 Park Run Drive
City, State ZIP: McLean, VA, 22102
Work Phone: 571-382-3816
Email: abraham.du@gmail.com

APPENDIX

```
%MACRO
STRA_SAMP_OUTLIER(POP,STRATA,K,ONE_SIDED,VAR,COST,TOT_SAMP,ALLOCMETH,A
LLOCMIN,OUT_POP);
/******************************************************************/
/* This customized SAS macro is to implement modified stratified */
/* sampling framework by re-defining strata from box-and-whisker */
/* outliers by Yi Du 2014 SAS Global Forum Paper.                */
/* Author: Yi Du                                                 */
/* Macro Variable Definition:                                    */
/* &POP.: the whole population used for sampling                 */
/* &STRATA.: the variable(s) to define stratum                   */
/* &K.: positive number used to calculate outliers               */
/* &ONE_SIDED.: Y/N to indicate to group left-sided outliers or  */
/*              or both-sided outliers                           */
/* &VAR.: the variable(s) used to measure variance by strata     */
/* &COST.: the average value of variable(s) to measure cost      */
/* &TOT_SAMP.: the total sample size requested                   */
/* &ALLOCMETH.: Optimal, Neyman or proportional allocation method*/
/* &ALLOCMIN.: the minimum number of units sampled by strata     */
/* &OUT_POP.: the output SAS data set with sample information    */
/******************************************************************/
%LET i = 1;
  %DO %WHILE(%LENGTH(%SCAN(&strata.,%EVAL(&i.))) > 0);
    %IF %EVAL(&i.) = 1 %THEN %DO;
       %LET newstrata = %SCAN(&strata.,&i.);
     %END;
     %ELSE %DO;
       %LET newstrata = &newstrata.%STR(*)%SCAN(&strata.,%EVAL(&i.));
     %END;
     %LET i = %EVAL(&i.+1);
  %END;

Proc Freq Data = &pop.;
  Tables &newstrata. / Noprint Out = freq_&pop.;
Run;

Proc Univariate Data = freq_&pop.;
  Var count;
  Output Out = quartile_&pop. Q1 = q1 Q3 = q3 Qrange = IQR Min = min
Max = max;
Run;

Proc Sql Noprint;
  Select min,q1,q3,iqr,max Into :min,:q1,:q3,:iqr,:max
  From quartile_&pop.;
Quit;

%IF &min. >= %SYSEVALF(&q1. - &k.*&iqr) %THEN %DO;
  %PUT ERROR: No outliers identified in the population. Please
consider other survey sample methods.;
  %GOTO exit;
```

```
%END;

%ELSE %DO;
Proc Sort Data = &pop. Out = sorted_&pop.;
  By &strata.;
Proc Sort Data = freq_&pop.;
  By &strata.;
Run;

Data new_&pop.;
  Merge sorted_&pop. (in=froms) freq_&pop. (in=fromc);
     By &strata.;
     If froms and fromc Then Output;
Run;

Data restra_&pop.;
  Set new_&pop.;
     %LET j = 1;
     %LET new_strata =;
     %DO %WHILE(%LENGTH(%SCAN(&strata.,%EVAL(&j.))) > 0);
      new_%SCAN(&strata.,&j.) = %SCAN(&strata.,&j.);
       %LET new_strata = &new_strata.%NRSTR( )new_%SCAN(&strata.,&j.);
       %LET j = %EVAL(&j.+1);
     %END;
     %PUT &new_strata.;
     %IF &one_sided. = N %THEN %DO;
       If count < %SYSEVALF(&Q1.- &k.*&IQR.) Or count > %SYSEVALF(&Q3.
+ &k.*&IQR.) Then Do;
         %DO k = 1 %TO %EVAL(&j.-1);
             Format %SCAN(&new_strata.,&k.) $32.;
          %SCAN(&new_strata.,&k.) = 'Combined';
           %END;
       End;
       Else Do;
         %DO k = 1 %TO %EVAL(&j.-1);
             Format %SCAN(&new_strata.,&k.) $32.;
          %SCAN(&new_strata.,&k.) = %SCAN(&strata.,&k.);
           %END;
       End;
     %END;
     %ELSE %DO;
       If count < %SYSEVALF(&Q1.- &k.*&IQR.) Then Do;
         %DO k = 1 %TO %EVAL(&j.-1);
             Format %SCAN(&new_strata.,&k.) $32.;
          %SCAN(&new_strata.,&k.) = 'Combined';
           %END;
       End;
       Else Do;
         %DO k = 1 %TO %EVAL(&j.-1);
             Format %SCAN(&new_strata.,&k.) $32.;
          %SCAN(&new_strata.,&k.) = %SCAN(&strata.,&k.);
           %END;
       End;
      %END;
Run;

%IF &var. ^= %THEN %DO;
Proc Univariate Data = restra_&pop. Noprint;
  Var &var.;
  Class &new_strata.;
  Output Out = var Var=_var_;
Run;
%END;

%IF &cost. ^= %THEN %DO;
Proc Univariate Data = restra_&pop. Noprint;
  Var &cost.;
  Class &new_strata.;
  Output Out = cost Mean = _cost_;
Run;

Proc Sort Data = var;
  By &new_strata.;
Proc Sort Data = cost;
  By &new_strata.;
Run;

Data var;
  Merge var (in=fromv) cost (in=fromc);
    By &new_strata.;
      If fromv and fromc Then Output;
Run;
%END;

Proc Sort Data = restra_&pop.;
  By &new_strata.;
Run;

Proc Surveyselect Data = restra_&pop. Seed = 9234 Method = srs Out =
&out_pop. Sampsize = &tot_samp.;
  Stratum &new_strata./
  %IF %UPCASE(&allocmeth.) = OPTIMAL %THEN %DO;
  alloc = &allocmeth. var = var cost = cost allocmin = &allocmin.;
  %END;
  %ELSE %IF %UPCASE(&allocmeth.) = NEYMAN %THEN %DO;
  alloc = &allocmeth. var = var allocmin = &allocmin.;
  %END;
  %ELSE %DO;
  alloc = &allocmeth. allocmin = &allocmin.;
  %END;
Run;
%END;

%EXIT:
%MEND STRA_SAMP_OUTLIER;
```