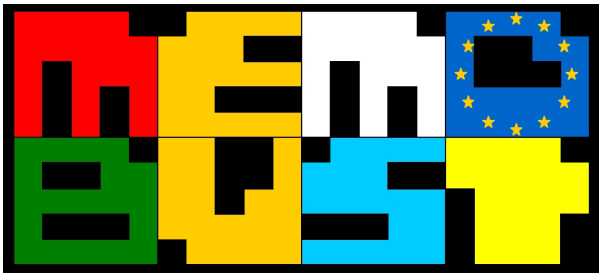


This module is part of the

Memobust Handbook

on Methodology of Modern Business Statistics

26 March 2014

Theme: Sample Selection – Main Module

Contents

General section

1. Summary

Sample selection in business statistics can be challenging because of several reasons. The population is often skewed, new businesses are created or they go out of business, and businesses may be related to each other in different ways. The use of a stratified simple random sampling design can enable researchers to draw inferences about specific subgroups that may be lost in a more generalised random sample, but this requires the selection of relevant stratification variables. An important option here, which is commonly used for business surveys whenever element size varies greatly, is probability proportional to size (pps) sampling, often in combination with cut-off sampling. This method can improve accuracy for a given sample size by concentrating the sample on large elements that have the greatest impact on population estimates. An alternative to stratified simple random sampling is systematic sampling. Cluster or multistage sampling is motivated by the need for practical, economical and sometimes administrative efficiency. The use of fixed panels will produce very efficient estimates of periodic change. In most periodic surveys sample rotation is used in order to reduce response burden.

2. General description

Most samples for business surveys are sampled from lists (list frames) or registers. Developing the associated sample designs can be challenging because business populations can have the following characteristics (Sigman and Monsour, 1995):

- *Skewness*. A small number of businesses account for a large proportion of the population total.
- *Dynamic membership*. Businesses are created, go out of business, change their type or level of activity, or change their identity.
- *Inter-business relationships*. Businesses may be related to each other in such a way that they are owned by the same legal entity, they employ the same accountant, or their activities are combined in a common set of financial records.

2.1 Probability and non-probability sampling

In a **probability sampling** scheme every unit in the population has a chance (greater than zero) of being selected in the sample, and this probability can be accurately determined. The combination of these traits makes it possible to produce unbiased estimates of population totals, by weighting sampled units according to their probability of selection.

Data collected by statistical offices are often not consistent with the accounting rules. This happens, for example, because economic data are frequently collected by different methods, using different sample surveys and different data processing methods and because of estimation error in case of missing data.

Probability sampling includes stratified simple random, probability proportional to size, systematic and balanced sampling. These various ways of probability sampling have two things in common:

- Every element in the population of interest has a known non-zero probability of being sampled.

- They involve random selection at some points.

Non-probability sampling can be divided into two categories. The first category is when we have a situation where some elements of the population have *no* chance of selection (these are sometimes referred to as 'out of coverage', 'undercovered', 'take no'). Cut-off sampling, see Section 2.4, is an example of such a scheme, which is commonly applied in business statistics.

The second category is where the probability of selection cannot be accurately determined. This involves the selection of elements based on assumptions regarding the population of interest, which forms the criteria for selection. Hence, because the selection of elements is non-random, non-probability sampling does not depend on the rationale of probability theory, thus it does not allow the estimation of sampling errors. These conditions give rise to exclusion bias, placing limits on how much information a sample can provide about the population. Information about the relationship between sample and population is limited, making it difficult to extrapolate from the sample to the population. Non-probability also sampling includes accidental (haphazard, convenience) and purposive sampling. These methods are usually not applied in business statistics and will therefore not be discussed further.

In addition to cut-off sampling, a non-probability sampling scheme may also be applied when the goal is to find preliminary estimates (see the module "Sample Selection – Subsampling for Preliminary Estimates"), which merely involves the use of quick respondent units.

2.2 Stratified simple random sampling

Whenever the population embraces a number of distinct categories, the frame can be organised by these categories into separate 'strata'. Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected. Questions that need to be answered with this design include:

- How should strata be constructed?
- How should the sample be allocated to strata?

There are several potential benefits to stratified sampling.

First, dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalised random sample. Methods for stratum construction in the case of one characteristic of interest and one continuous stratification variable are described by Dalenius and Hodges (1959), Cochran (1977, pp. 127–133), Godfrey et al. (1984), Kott (1985), Hidiroglou (1986) and Detlefsen and Veum (1991), among others.

Second, utilising a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples). Even if a stratified sampling approach does not lead to increased statistical efficiency, such a tactic will not result in less efficiency than simple random sampling would, provided that each stratum is proportional to the group's size in the population.

Third, since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population.

There are, however, some potential drawbacks to using stratified sampling. First, identifying strata and implementing such an approach can on the one hand increase the cost and complexity of sample selection, on the other hand lead to an increased complexity of population estimates. Second, examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata. Finally, in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than other methods would (although in most cases, the required sample size would be smaller than in case of simple random sampling).

A stratified sampling approach is most effective when the following conditions are met:

- Variability within strata is minimised.
- Variability between strata is maximised.
- The variables upon which the population is stratified are strongly correlated with the desired dependent variable.

Advantages of the approach over other sampling methods are:

- It focuses on important subpopulations and overshadows, possibly ignores irrelevant ones.
- It allows the use of different sampling techniques for different subpopulations.
- It improves the accuracy/efficiency of estimation.
- It permits balancing the statistical power of tests of the various strata by departing from proportional sampling to a greater extent.

Its disadvantages are:

- It requires the selection of relevant stratification variables which can be difficult.
- It is not useful when there are no homogeneous subgroups.
- Its implementation can be expensive.

In some cases the sample designer has an access to an auxiliary variable or the size measure, believed to be correlated with the variable of interest, for each element in the population. It can be used to improve accuracy in sample design. A possible option is to use the auxiliary variable as a basis for stratification, as discussed above. Another option is probability proportional to size sampling (see Section 2.3).

Note that stratification (or *prestratification*) should not be confused with *poststratification*. The latter uses a discrete auxiliary variable to stratify the sample data *after* the sample has been selected. Its purpose is to improve efficiency of an estimator, see "Weighting and Estimation – Main Module".

2.3 Probability proportional to size (pps) sampling

The *pps* approach can improve accuracy for a given sample size by concentrating the sample on large elements that have the greatest impact on population estimates. The *pps* sampling design is commonly used for business surveys whenever element size varies greatly and/or auxiliary information is available – for instance, a survey attempting to measure the number of guest-nights spent in hotels might use each hotel's number of rooms as an auxiliary variable. In other typical examples the

auxiliary information can be the number of employees or turnover as measuring size. In some cases, a former measurement of the variable of interest can be used as an auxiliary variable for attempting to produce more current estimates. Poisson sampling (Hájek, 1960) is a *pps* sampling design with random sample sizes. This tends to be less efficient than *pps* designs with fixed sample sizes, but has the main advantage that it is easy to coordinate the samples, that is, to minimise or maximise overlap between samples selected from the same population. See "Sample Selection – Sample Co-ordination".

2.4 Cut-off sampling

The *pps* approach can be used in combination with *cut-off sampling*. This is often applied to highly skewed populations, such as populations of businesses with a few large units (e.g., defined by the number of employees), and more and more, smaller and smaller values. Therefore, most of the volume for a given variable will be covered by a relatively small number of businesses, hence individual small businesses will have a little impact on population estimates. Together with the fact that the respondent burden on those businesses will be relatively high, we deliberately exclude businesses with size below a certain *cut-off threshold* from the population, i.e., give them a selection probability of zero. A short introduction to cut-off sampling is given by Knaub (2008).

Although cut-off sampling is common among practitioners, its theoretical foundations are weak. Elisson and Elvers (2001) performed a univariate analysis that compared cut-off sampling with simple stratified sampling. They conclude that the dimensional variable determining the cut-off threshold has a relevant impact on the result, so they stress that great care must be employed in choosing this variable. Benedetti et al. (2010) proposes a framework that justifies cut-off sampling and provides means for determining cut-off thresholds. They also compute the variance of the resulting estimator and its bias.

2.5 Cluster or multistage sampling

Cluster sampling is an alternative approach for using multiple stratification variables. It is motivated by the need for practical, economical and sometimes administrative efficiency. An important advantage of cluster sampling is that a sampling frame at the element level is not needed. Thus, in multistage sampling, a subsample is drawn from the sampled clusters at each stage except the last. At this stage all the elements from the sampled clusters can be taken in an element level sample, or a subsample of the elements can be drawn (Lehtonen and Pahkinen, 2004, p. 70). For example, in two-stage sampling the first stage can be a frame of geographical areas from which areas (*first-stage units*) are selected, and the second stage a list of businesses (*primary sampling units*) from areas selected in the first stage. Colledge et al. (1987) and Armstrong and St-Jean (1993) give examples on two-stage sampling in business statistics. For mathematical details, see, e.g., Thompson (1997, section 2.6).

A recommended method of a clustering algorithm for stratum construction is given by Jarque (1981).

2.6 Systematic sampling

A popular alternative to simple random sampling is systematic sampling (Cochran 1977, pp. 205–232). Stratified systematic sampling often leads to more efficient estimation than stratified simple random sampling because it can incorporate an additional level of implicit stratification within explicit strata, see example by Garrett and Harter (1995). Systematic sampling is often performed according to

an order based on random numbers. In these cases the notion *systematic* is misleading, because this sampling is random in essentials.

2.7 Balanced sampling

A balanced sampling design (see also "Sample Selection – Balanced Sampling for Multi-Way Stratification") has the property that the estimators of the totals for a set of auxiliary variables are equal to the totals we want to estimate (Tillé, 2006, p. 147). Many types of sampling designs can be interpreted as balanced sampling such as simple random sampling, sampling with fixed size, stratified simple random sampling and unequal probability sampling.

2.8 The use of panels and rotation groups

A *panel* is defined as the collection of all units in the survey for a given period, e.g., a week, a month, a quarter, or whatever. The exclusive use of a fixed panel produces very efficient estimates of periodic change. In most periodic surveys *sample rotation* is used in order to reduce response burden (see "Response – Response Burden"). Sample rotation is closely connected to sample co-ordination (see "Sample Selection – Sample Co-ordination"). A more detailed description of the various forms of rotation sampling is given by Wolter (1979), Sigman and Monsour (1995) and Srinath and Carpenter (1995), among others.

3. Design issues

Not applicable.

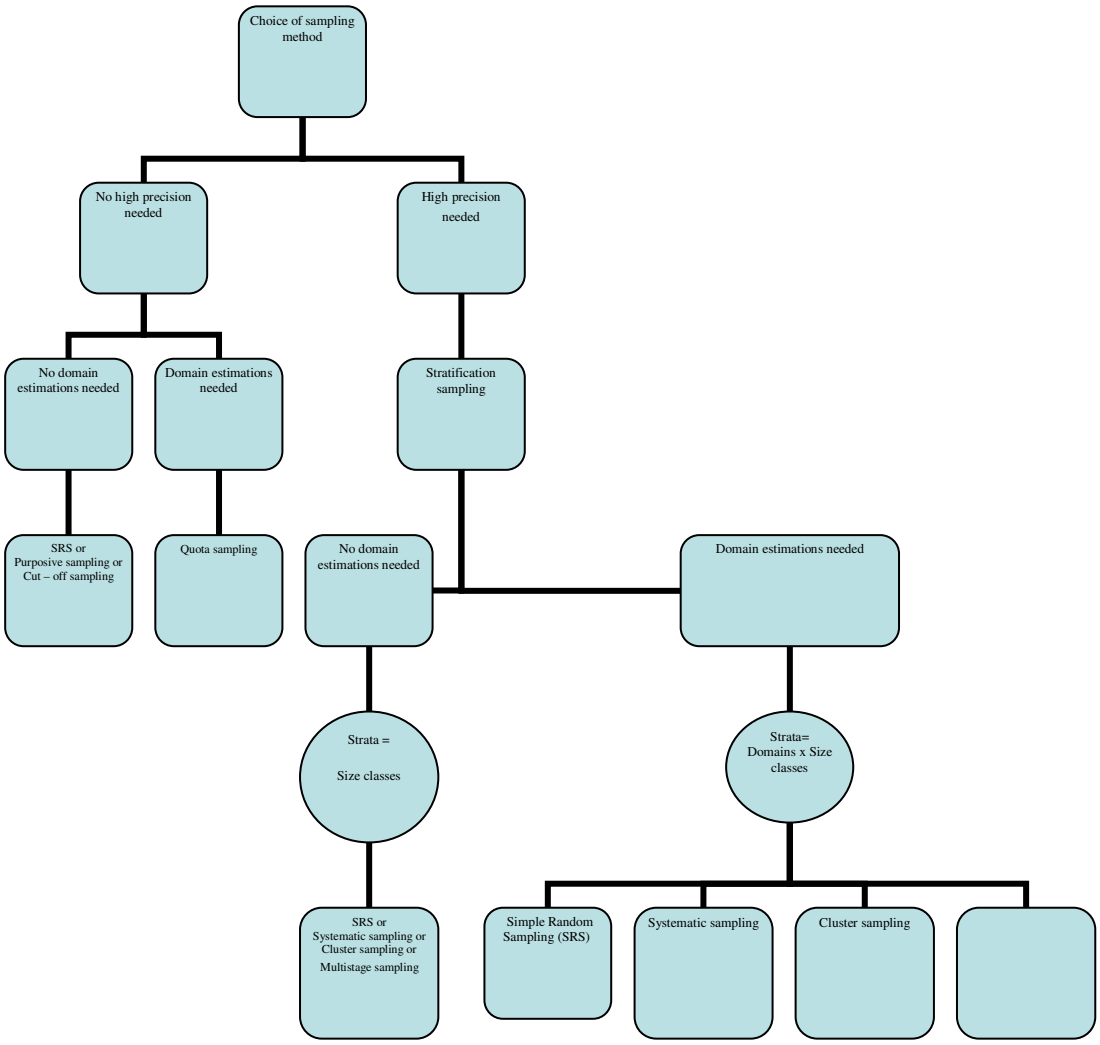
4. Available software tools

Packages for sample designs (http://www.hcp.med.harvard.edu/statistics/survey-soft/):

- **AM Software** from American Institutes for Research.
- **Bascula** from Statistics Netherlands.
- **CENVAR** from U.S. Bureau of the Census.
- **CLUSTERS** from University of Essex.
- **Epi Info** from Centers for Disease Control.
- **Generalized Estimation System (GES)** from Statistics Canada.
- **IVEware** from University of Michigan.
- **PCCARP** from Iowa State University.
- **R survey package** from the R Project.
- **SAS/STAT** from SAS Institute.
- **SPSS Complex Samples** from SPSS Inc.
- **Stata** from Stata Corporation.
- **SUDAAN** from Research Triangle Institute.
- **VPLX** from U.S. Bureau of the Census.

- **WesVar** from Westat, Inc.

5. Decision tree of methods



6. Glossary

For definitions of terms used in this module, please refer to the separate "Glossary" provided as part of the handbook.

7. References

Armstrong, J. and St-Jean, H. (1993), Generalized Regression Estimation for a Two-Phase Sample of Tax Records. *Proceedings of the International Conference on Establishment Surveys*, American Statistical Association, Alexandria, VA, 402–407.

Benedetti, R., Bee, M., and Espa, G. (2010), A Framework for Cut-off Sampling in Business Survey Design. *Journal of Official Statistics* **4**, 651–671.

Cochran, W.G. (1977), *Sampling Techniques*. Wiley, New York.

Colledge, M., Estavao, V., and Foy, P. (1987), Experience in Coding and Sampling Administrative Data. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 529–534.

Dalenius, T. and Hodges, J. L. (1959), Minimum Variance Stratification. *Journal of the American Statistical Association* **54**, 88–101.

Detlefsen, R. E. and Veum, C. S. (1991), Design Issues for the Retail Trade Sample Surveys of the U.S. Bureau of the Census. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 592–596.

Elisson, H. and Elvers, E. (2001), Cut-off Sampling and Estimation. *Proceedings of Statistics Canada Symposium*.

Garrett, J. K. and Harter, R. M. (1995), Sample Design Using Peano Key Sequencing in Market Research. In: *Business Survey Methods* (eds. B. G. Cox et al.), Wiley, New York, 205–217.

Godfrey, J., Roshwalb, A., and Wright, R. (1984), Model-Based Stratification in Inventory Cost Estimation. *Journal of Business and Economic Statistics* **2**, 1–9.

Hájek, J. (1960), Limiting Distributions in Simple Random Sampling from a Finite Population. *Publications of the Mathematical Institute of the Hungarian Academy of Science* **5**, 361–374.

Hidiroglou, M. A. (1986), The Construction of a Self-Representing Stratum of Large Units in Survey Design. *The American Statistician* **40**, 27–31.

Jarque, C. M. (1981), A Solution to the Problem of Optimum Stratification in Multivariate Sampling. *Applied Statistics* **30**, 163–169.

Knaub Jr., J. R. (2008), Cutoff Sampling. In: *Encyclopedia of Survey Research Methods* (ed. P. J. Lavrakas), Sage, London.

Kott, P. S. (1985), A Note on Model-Based Stratification. *Journal of Business and Economic Statistics* **3**, 284–288.

Lehtonen, R. and Pahkinen, E. (2004), *Practical Methods for Design and Analysis of Complex Surveys*. Wiley, Chichester.

Sigman, R. S. and Monsour, N. J. (1995), Selecting Samples from List Frames of Businesses. In: *Business Survey Methods* (eds. B.G. Cox et al.), Wiley, New York, 133–152.

Srintah, K. P. and Carpenter, R. M. (1995), Sampling Methods for Repeated Business Surveys. In: *Business Survey Methods* (eds. B.G. Cox et al.), Wiley, New York, 171–184.

Tillé, Y. (2006), *Sampling Algorithms*. Springer, New York.

Thompson, M. E. (1997), *Theory of Sample Surveys*. Chapman and Hall, London.

Wolter, K. M. (1979), Composite Estimation in Finite Populations. *Journal of the American Statistical Association* **74**, 604–613.

Interconnections with other modules

8. Related themes described in other modules

1. Sample Selection – Sample Co-ordination
2. Weighting and Estimation – Main Module
3. Response – Response Burden

9. Methods explicitly referred to in this module

1. Sample Selection – Balanced Sampling for Multi-Way Stratification
2. Sample Selection – Subsampling for Preliminary Estimates

10. Mathematical techniques explicitly referred to in this module

1.

11. GSBPM phases explicitly referred to in this module

1.

12. Tools explicitly referred to in this module

1.

13. Process steps explicitly referred to in this module

1.

Administrative section

14. Module code

Sample Selection-T-Main Module

15. Version history

Version	Date	Description of changes	Author	Institute
0.1	18-02-2013	first version	Magnar Lillegård	SSB
0.2	02-04-2013	second version	Magnar Lillegård	SSB
0.2.1	06-09-2013	preliminary release		
0.2.2	09-09-2013	page numbering adjusted		
1.0	26-03-2014	final version within the Memobust project		

16. Template version and print date

Template version used	1.0 p 4 d.d. 22-11-2012
Print date	21-3-2014 17:41