Comparative study on the complex samples design features using SPSS Complex Samples, SAS Complex Samples and WesVarPc.

Nur Faezah Jamal, Nor Mariyah Abdul Ghafar, Isma Liana Ismail and Mohd Zaki Awang Chek

To Link this Article: http://dx.doi.org/10.6007/IJARBSS/v8-i4/4238

Received: 27 Feb 2018, Revised: 19 Mar 2018, Accepted: 07 April 2018

Published Online: 21 April 2018

In-Text Citation: (Jamal, Ghafar, Ismail, & Chek, 2018)
To Cite this Article: Jamal, N. F., Ghafar, N. M. A., Ismail, I. L., & Chek, M. Z. A. (2018). Comparative study on the complex samples design features using SPSS Complex Samples, SAS Complex Samples and WesVarPc. International Journal of Academic Research in Business and Social Sciences, 8(4), 1282–1292.

Copyright: © 2018 The Author(s)
Published by Human Resource Management Academic Research Society (www.hrmars.com)
This article is published under the Creative Commons Attribution (CC BY 4.0) license. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this license may be seen at: http://creativecommons.org/licences/by/4.0/legalcode
Comparative study on the complex samples design features using SPSS Complex Samples, SAS Complex Samples and WesVarPc.

Nur Faezah Jamal, Nor Mariyah Abdul Ghafar, Isma Liana Ismail and Mohd Zaki Awang Chek

Faculty of Computer and Mathematical Sciences (FSKM), Universiti Teknologi MARA, Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia

Abstract
Unlike simple random sampling, complex sample designs involve additional considerations such as multistage sampling, stratification, and unequal probability of selection. A basic problem with complex surveys is in variance estimation which requires the use of approximate methods. Generally, such methods are based on either the Taylor series linearization or the replication techniques. Statistical software that use standard packages usually assume that simple random sampling of elements is inadequate for data analysis from complex surveys, especially for purpose of variance estimation. This study compares the complex sample design features produced by three statistical software packages designed to handle complex surveys (SPSS 16.0 Complex Samples, SAS 9.0 Complex Samples, and WesVarPc 5.1). Comparisons among the software are made based on the types of sample design, sampling error estimates, method of variance estimation and cost of software packages. The results of the finding show that WesVarPc can be downloaded for free from Web and offers complete basic of descriptive analyses. Although expensive, SPSS 16.0 Complex Samples and SAS 9.0 Complex Samples have been dominant in the field of data management and data analysis.

Keywords: Simple Random Sampling, Complex Sample Designs, Taylor Series Linearization, replication techniques, SPSS 16.0 Complex Samples, SAS 9.0 Complex Samples, WesVarPC 5.1.

Introduction
In complex sample surveys, these collected data are usually used in secondary data analysis. Secondary data sources are usually more cost effective than collecting primary data because data that collected from secondary data sources are typically easier to obtain than primary data (Hofferth, 2005; Rosenberg, Greenfield and Dimick, 2006).

Unlike simple random sampling, complex sample design surveys require the procedures of statistical analysis used to analyze the data that can account for the various aspects that
contribute to the complexity, such as unequal probability of selection, multiple stages of selection, stratification, and clustering.

Analyzing sample survey data generally cannot be use the standard statistical software packages since they usually assume that simple random sampling of elements. These packages underestimation of standard errors for point estimates and yielded biased point estimates of population parameters. Sampling weight variable using standard packages produce appropriate point estimates of population parameters. Standard version of statistical packages such as SPSS and SAS are inappropriate for complex survey data since based on the assumption of identically and independently distributed observations or simple random sampling with replacement. The estimated standard errors are usually incorrect because the variance estimation procedure typically does not consider the clustering or stratification in the sampling plan.

Variance estimation for estimators depends upon the specific sampling plan used and the complexity of the sampling design requires approximate methods, which generally belong to either the Taylor series linearization or the replication techniques. Variance estimation is important because it exhibit the precision of estimators, which in turn affect the confidence intervals and the subsequent testing of hypotheses of population parameters. Based on complex sample survey data for variance estimators must recognize the following factors:

1. Most estimators are non-linearity
2. Estimators are weighted
3. Sampling plan will be using stratification prior to first-stage sampling (and perhaps also at subsequent sampling stages)
4. Elements in the sample will generally not be statistically independent due to multistage cluster sampling.

Nowadays, there are several currently available statistical estimation software which make available for clustering and other complex features of stratified multistage samples such as Cenvar, Clusters, Epi Info, Pc Carp, Stata, Sudaan, Vplx, WesVarPc, SPSS Complex Samples, and SAS Complex Samples. However, this list comprises only for documented free-ware statistical packages that are accessible for use by the general survey data analysts.

Thus, objective of this study is to explore the complex samples design features employed by SPSS Complex Samples 16.0, SAS Complex Samples 9.0 and WesVarPc 5.1. Due to cost and time constraint, this study will only focus on three software packages since certain packages such as CLUSTERS, STATA and SUDAAN cannot be downloaded from the website.

Literature Review
A design is not a simple random sample (where each target population units do not have the same opportunity to choose from in this study) is known as complex survey design. Complex survey sampling is widely used for a fraction of the population confined to large while accounting for the size and characteristics of the sample. Based on the characteristics of the subject (e.g., age, race, gender) some individuals are over or under sampled. The individual in the population
has a different probability to be selected into the sample (Natarajan, & Lipsitz, Moore and Gonin, 2008)

Generally, sampling is always multistage. Usually, there are simple random sampling in some levels, but there may be separate sampling of population subgroups called strata. In addition, there is a possibility of group observation, otherwise known as clusters, which may not be sampled independently and may have over sampling or under sampling from a specific group. Combination of different sampling scheme forms what is known as a complex design (Oyeyemi, 2009).

Variance estimates are calculated using only between the number of units the first stage regardless of variance components at each level selection (Lepkowski, J. and Bowles, J., 1996). There are two main methods of variance estimation of employed to design complex sample: Taylor series linearization and repeated replication (Wolter, 1985).

The basis of the Taylor series linearization procedure is the approximation of the estimator of interest by linear estimator. Linear approximation of variance was then estimated using method of standard variance estimation available for linear estimators. The linear approximation is derived by taking the first order Taylor series linearization for the estimator \( \hat{\mu} \) about the parameter \( \mu \). As \( \hat{\mu} \) is in general biased for \( \mu \), the resulting variance estimator actually estimates the mean square error of \( \hat{\mu} \) (Rust, 1985).

For surveys with many strata, but only few sampled primary sampling unit (PSU) per stratum, the random groups method is unsatisfactory. The method of balanced repeated replication (BRR) has been developed for use with such designs and has proved most useful in the common case where two PSUs are selected per stratum (Rust K, 1985).

They show that the Linearization package only can produce standard error statistics for Taylor Series approximation formula was included in the software, while the replication method can be used for almost any statistics. They consider documentation WesVarPc 5.1 lie the easiest to read and the best exit but warned that users who are unfamiliar with replication may get specifications weight replicate of misleading. They note that CENVAR, CLUSTERS, Epilinfo, VPLX, and WesVarPc 5.1 are free, and except for CLUSTERS, can be downloaded off the Internet. Virtually all standard statistical software will handle the weight correctly to estimate the approximate point in most analysis available. A little handle properly weighted for estimates of variance. In addition, only one, Stata has estimated for calculating stratification and multistage selection used in design (Lepkowski and Bowles, 1996).

The National Center for Health Statistics (NCHS) conducts surveys with probability-based complex sample designs. The complex sample designs packages examined in this study using SAS, SPSS, Stata and Sudaan produced identical results using the Taylor series linearization method. Therefore, other factors such as cost, existence of point and click operation, the entire data
management capabilities and an alternative method of variance estimation will affect the choice of the user software.

Lepkowski, J. and Bowles, J., (1996) catalog eight packages for the PC that will calculate standard errors for complex survey data using either replication or linearization: CENVAR, CLUSTERS, EpiInfo, PC CARP, Stata, SUDAAN, VPLX, and WesVarPc. The catalog contains useful information about the availability of platforms, cost, installation, purchase, file input and general software capabilities. Although the estimates are not presented in the study, the authors found that the proportion of the estimates, standard error and coefficient of variation “similar” entire package CENVAR, CLUSTERS, EpiInfo, SUDAAN and WesVarPc.

Methodology
In this study, three complex sample design software packages such as SPSS 16.0 Complex Samples, SAS 9.0 Complex Samples and WesVarPc 5.1 were used. In this study, two approaches to the task of estimating the variance of a non-linear estimator be able distinguished.

SPSS 16.0 Complex Samples
SPSS 16.0 Complex Samples is a module of SPSS Statistics. It provides statistical planning tool specialized to work with the sample survey data. It allows users to make more valid statistics inferences by combining the sample design into survey analysis. Besides that, it can work with the results of numerical and categorical design of complex samples using two algorithms for analysis and prediction. The types of sample design information that can be utilized by SPSS 16.0 Complex Samples include:

1. Stratified sampling: This feature allows choosing to sample in subgroups of the population such as the survey of people in certain job categories, or an age group.
2. Clustered sampling: Select clusters, i.e. groups of units sampled for this study. Clusters can include school or geographic area with a sampling unit may be students or people. Clusters are often help make surveys more cost-effective.
3. Multistage sampling: Select a sample of first-stage based on a group of elements in your population; then create a second stage by drawing a sample sub sample from each selected unit in the first-stage sample. By repeating this option, you can select a sample of a higher level. For example, in a survey of face to face, you may feel the individual in blocks of urban environment and households.

To begin your work in SPSS 16.0 Complex samples using the wizard, which asks you to a lot of factors that you must consider. If you create your own samples, using sampling Wizard to specify the sampling scheme. If you are use public-use datasets that have already been tried, as provided by the CDC, use analysis Preparation Wizard to specify how the samples were defined and how to estimate the standard error. Once you create a sample or determine standard errors, you can create plans, analyze your data, and produce results. (see the Figure 3.1 for workflow).
Procedures to estimate the differences regarding the sample design used to select the sample, including the equal probability and probability proportionate to size (PPS) methods, using both with replacement (WR) and without replacement (WOR) sampling procedures.

**SAS 9.0 Complex Samples**

The current SAS software package now provides complex samples design. The SAS 9.0 Complex Samples provides the SURVEYFREQ, SURVEYLOGISTIC, SURVEYMEANS, SURVEYREG and SURVEYSELECT procedures. In analyzing sample survey data, the SURVEYFREQ, SURVEYMEANS, SURVEYREG and SURVEYSELECT procedures, incorporate sample design into the analysis. This procedure can be used for design of multistage or single-stage, with or without stratification, with or without unequal weighting.

The standard procedure of SAS/STAT, like the MEANS and GLM procedures, calculate sample means and estimate the regression relationships. However, in most procedures, statistical inference is assuming that the sample is taken from an infinite population by simple random sampling. These procedures generally do not calculate the estimates and their variances correctly if the sample is selected from a finite population using a complex design.

On the other hand, the procedures in SAS 9.0 Complex Samples can be used to handle data from stratified sampling, cluster sampling and multistage sampling.
For a multistage sample design, the Taylor series method uses only the first stage of the sample design. Inputs required includes only the first-stage stratum identification and first-stage cluster (PSU). You do not need to include design information on any additional stages of sampling.

WesVarPc 5.1

WesVarPc 5.1 is a statistical software system designed by Westat, Inc. for analysis of complex survey data. This program operates in Windows (3.1, 3.11 and 95) environment and are completely menu driven. The main sample design that can be placed is a stratified multistage cluster sample based on the ultimate cluster sampling model.

Users must have a PC compatible microcomputer with 4Mb RAM and at least 10.1 Mb hard disk space, and Windows 3.1, 3.11 or 95. It can be downloaded from the WWW site and may be the easiest installation among all software listed using the software installation available. Documentation is in Adobe format, but the instructions provided to download Adobe reader. Documentation is the easiest to read and best organized among all listed programs.

Input data can be in ASCII format, or DBF, SPSS for Windows, SAS Transport, or PC SAS for DOS format. Inputs from any of these formats are easy in the menu-driven environment. WesVarPc 5.1 requires a new version of the data set made in the format of special WesVarPc 5.1. This requires specification replicates and whether post stratification is to be included in the estimates of variance and replicate weights.

Based on the Figure 3.2, this screen is used as the starting point for WesVarPc 5.1. There are four major paths you can take, selected by clicking on one of the red or blue button (WesVar User’s guide).

1. **New WesVar Data File.** If you do not have WesVar data files created from some other file formats (eg, SAS Files or SPSS), use this button first. WesVar will ask you to identify the weight variables, analysis variables, and any ID variables. If you do not have replicate weights to your input file, WesVar will ask you to provide sample design information so that it can create and save the replicate weights on the WesVarformatted version of your data set.

2. **Open WesVar Data File.** Once the WesVar file has been created, use this button to open the WesVar data file for modifications (recoding, labeling, subsetting, etc.) or to start working with the file, creating one or more workbooks with requests.

3. **New WesVar Workbook.** Tables and models are defined in the WesVar workbook. Use this button to create a new WesVar workbook that is linked to a WesVar data file and to determine tables and models.

4. **Open WesVar Workbook.** After the workbook has been created and linked to WesVar’s data file, it can be reopened to generate additional tables and models or modify the previously defined request.
Analysis and Result
This section explores three software packages which are SPSS 16.0 Complex Samples, SAS 9.0 Complex Samples and WesVarPc 5.1 for the complex sample design features. These three software packages that include many, but not all the options possible for sample survey variance estimation.

SPSS 16.0 Complex Samples provides procedures to analyze data from sample survey data such as Complex Samples Descriptive (CSDESCRIPTIVES) are used to estimate means, sums and ratios, and computes standard errors, design effects, confidence intervals hypothesis tests for samples drawn by complex methods. SPSS 16.0 Complex Samples provide highly cost to purchase. SPSS 16.0 Complex Samples use Taylor series approximation for variance estimation.

SAS 9.0 Complex Samples uses PROC SURVEYFREQ and PROC SURVEYMEANS to generate estimates of population means, totals and proportions, standard errors, confidence limits, hypothesis tests, domain analysis and ratio estimates. SAS 9.0 Complex Samples also provide highly cost to purchase. In the survey analysis SAS 9.0 Complex Samples procedures provide a Taylor series linearization method of variance estimation methods for complex survey designs. The selection of variance estimation methods depends on the sample design used, the sample design information available, and the parameters to be estimated.

WesVarPc 5.1 uses repeated replication for variance estimation, including jackknife, balanced half sample, and the Fay modification to the balanced half sample method. WesVarPc
5.1 has facilities at present for contingency Table analysis, regression, and logistic regression. There is an extensive menu-driven system for creating new variables covering a wide range of statistics that WesVarPc 5.1 can be used. This software provides descriptive parameters such as totals, means, ratios and differences of rations and proportions. The design effect estimates also provide for this software.

Table 4.1 summarizes the three software packages on various features, including sampling plans, variance estimation methods and types of analysis.

| PACKAGES            | OVERVIEW                                                                 | METHOD OF VARIANCE ESTIMATION          |
|---------------------|--------------------------------------------------------------------------|----------------------------------------|
| 1. SPSS 16.0        | i. When working with sample survey data need specialized planning tools and statistics. |
| Complex Samples     | ii. Numerical value and categorical outcomes in complex sample designs.   | Taylor series approximation            |
|                     | iii. Types of sample design:                                            |                                        |
|                     | - Stratified sampling                                                   |                                        |
|                     | - Cluster sampling                                                      |                                        |
|                     | - Multistage sampling                                                   |                                        |
|                     | iv. Cost: High cost                                                     |                                        |
|                     | v. Sampling error estimates:                                            |                                        |
|                     | - Means                                                                  |                                        |
|                     | - Totals                                                                |                                        |
|                     | - Ratios                                                                |                                        |
|                     | - Proportions                                                           |                                        |
|                     | - Computes standard errors                                              |                                        |
|                     | - Design effects and others                                             |                                        |
|                     | - Confidence intervals                                                 |                                        |
| 2. SAS 9.0          | i. Provides the SURVEYFREQ, SURVEYMEANS, SURVEYREG and SURVEYSELECT produces. |
| Complex Samples     | ii. Types of sample design:                                            |                                        |
|                     | - Stratified sampling                                                   |                                        |
|                     | - Cluster sampling                                                      |                                        |
|                     | - Multistage sampling                                                   |                                        |
|                     | iii. Cost: High cost                                                    |                                        |
|                     | iv. Sampling error estimates:                                            |                                        |
|                     | - Means                                                                  |                                        |
|                     | - Proportions                                                           |                                        |
|                     | - Totals                                                                |                                        |
|                     | - Computes standard errors                                              |                                        |
### WesVarPc 5.1

|   |   |
|---|---|
| i. | Designed by Westat, Inc |
| ii. | Types of sample design: |
|     | - Stratified sampling |
|     | - Cluster sampling |
|     | - Multistage sampling |
| iii. | Cost: No cost and easily downloaded from the WWW site. |
| iv. | Sampling error estimates: |
|     | - Means |
|     | - Proportions |
|     | - Totals |
|     | - Ratios |
|     | - Design effects and others |

### Repeated replication

- Balanced replicated replication
- Jackknife
- Fay Modification

As a result, WesVarPc 5.1 offers complete the basic of descriptive analyses and can be downloaded from the Web for free. Besides that, this software can be appealing feature for analysts with a limited or no cost for software purchases. Although expensive, SPSS 16.0 Complex Samples and SAS 9.0 Complex Samples were selected because of their dominant in the field of data management and analysis arena. Furthermore, the comparatively new Complex Samples (CS) and PROCs for SPSS Complex Sample and SAS 9.0 Complex Samples respectively represent the more recent developments in data analysis of complex survey.

### Conclusion and Recommendation

Based on the finding of this study, it can be concluded that these three software packages offer the same types of sample design and sampling error estimates. For method of variance estimation, SPSS 16.0 Complex Samples and SAS 9.0 Complex Samples use Taylor series approximation but WESVARPC 5.1 use repeated replication. Based on cost to download the software packages, WESVARPC 5.1 can easily download from website at no cost and offer complete basic descriptive analysis. Although SPSS 16.0 Complex Samples and SAS 9.0 Complex Samples are highly cost, but these packages dominate in the data management and analysis arena. This study can be extended by exploring others software packages such as CENVAR, CLUSTERS, Epi Info and VPLX for complex sample design.

### Corresponding Author

Nur Faezah Jamal  
Faculty of Computer and Mathematical Sciences (FSKM), Universiti Teknologi MARA, Perak Branch, Tapah Campus, 35400 Tapah Road, Perak, Malaysia  
Email: nurfa210@perak.uitm.edu.my
References
Hofferth, S. (2005). Secondary data analysis in family research. *Journal of Marriage and Family*, 67, 891-907.

Lepkowski, J. and Bowles, J. (1996). Sampling Error Software for Personal Computers. *Survey Statistician*, 35, 10-17.

Natarajan, S., S.R. Lipsitz, C.G. Moore and R. Gonin. (2008). Variance estimation in complex survey sampling for generalized linear models. *Applied Statistics*, 57, 75-87.

Oyeyemi, G. A. (2009). *Multivariate Regression in Complex Survey Design*. Indian: ICASTR.

Rosenberg, A.L., Greenfield, M.V.H., and Dimick, J.B. (2006). Secondary data analysis: Using existing data to answer clinical questions. In J.T. Wei (Ed.). In A. G. Rosenberg, *Clinical Research Methods for the Surgeons*. Totowa, NJ: Humana Press.

Rust K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, 1(4), 381-397.

Rust, K.F., and J.N.K. Rao. (1996). Variance Estimation for Complex surveys using replication techniques. *Statistical Methods in Medical Research*, 5, 283-310.

Wolter, K. M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.