Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 16

4-6-2020

Using SPSS to Analyze Complex Survey Data: A Primer

Danjie Zou
University of British Columbia, zoudj@mail.ubc.ca

Jennifer E. V. Lloyd
University of British Columbia, jennifer.lloyd@ubc.ca

Jennifer L. Baumbusch
University of British Columbia, jennifer.baumbusch@ubc.ca

 Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation

Zou, D., Lloyd, J. E. V., & Baumbusch, J. L. (2019). Using SPSS to analyze complex survey data: A primer Journal of Modern Applied Statistical Methods, 18(1), eP3253. doi: 10.22237/jmasm/1556670300

Using SPSS to Analyze Complex Survey Data: A Primer

Cover Page Footnote

Thank you to the McCreary Centre Society (https://www.mcs.bc.ca/), who collects and owns the British Columbia Adolescent Health Survey data. Thanks also to Dr. Colleen Poon, Allysha Ram, Dr. Elizabeth Saewyc, and Annie Smith for their guidance as we worked with the data. We also thank the Social Sciences and Humanities Research Council of Canada (SSHRC) for an Insight Development grant awarded to Dr. Baumbusch. Finally, thanks to blind reviewers for their comments that improved the paper. An SPSS syntax file with the commands outlined in this paper is available for download at: http://blogs.ubc.ca/jenniferlloyd/

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP3253.
doi: 10.22237/jmasm/1556670300

בס"ד
Copyright © 2020 JMASM, Inc.
ISSN 1538 − 9472

Using SPSS to Analyze Complex Survey Data: A Primer

Danjie Zou
University of British Columbia
Vancouver, British Columbia

Jennifer E. V. Lloyd
University of British Columbia
Vancouver, British Columbia

Jennifer L. Baumbusch
University of British Columbia
Vancouver, British Columbia

An introduction to using SPSS to analyze complex survey data is given. Key features of complex survey design are described briefly, including stratification, clustering, multiple stages, and weights. Then, annotated SPSS syntax for complex survey data analysis is presented to demonstrate the step-by-step process using real complex samples data.

Keywords: Complex sample design, complex survey data, SPSS, syntax, primer

Introduction

Complex survey data (a.k.a., complex sample data) refer to data collected using a complex survey design. They involve features such as stratification, clustering, multiple stages, and unequal probability of selection (e.g., Muthén & Muthén, 2017; Everitt & Skrondal, 2010). A complex survey design is usually applied in cases where simple random sampling is not convenient, feasible, efficient, or cost-effective (e.g., Lewis, 2017; Heeringa et al., 2017; Lumley, 2010).

Complex sample designs are utilized extensively in education and the social sciences, especially with respect to large scale surveys. Some Canadian examples include the Canadian Labour Force Survey (LFS) (Statistics Canada, 2017), the Canadian Community Health Survey (CCHS) (Statistics Canada, 2018), the National Population Health Survey (NPHS) (Statistics Canada, 2012), and the McCreary Centre Society's British Columbia Adolescent Health Survey (BCAHS) (e.g., Saewyc et al., 2014).

In the United States, large scale complex surveys are common as well, including the High School and Beyond Longitudinal Study (HS&B) (National Center for Education Statistics [NCES], 2017) and the National Postsecondary Student Aid Study (NPSAS) (NCES, n.d.). Some international examples include

doi: 10.22237/jmasm/1556670300 | Accepted: July 15, 2019; Published: April 6, 2020.
Correspondence: Jennifer E. V. Lloyd, jennifer.lloyd@ubc.ca

the Programme for International Student Assessment (PISA) (Westat, 2017) and the World Values Survey (WVS) (World Values Survey, n.d.).

Simple Random Sampling vs. Complex Sampling

One characteristic of simple random sampling (SRS) is that each element in the target population has an equal probability of being chosen for inclusion in the data collection (e.g., Lewis, 2017; Heeringa et al., 2017; Lee & Forthofer, 2005; Neyman, 1934). In contrast, complex sampling (CS) considers composition characteristics of the target population – that is, a group of individual participants a researcher would like to collect information on and make inferences about (Everitt & Skrondal, 2010; Statistics Canada, 2003). Therefore, each element may not have the same probability of being selected.

It is common to find natural strata and clusters of individuals in a population. One simple example is the population of a country – which can be stratified geographically or stratified sociodemographically, according to such characteristics as age, gender, and ethnicity. Clustering is ubiquitous as well -- for example, family members living together in a house or different families with similar socioeconomic status residing in the same neighborhood. A survey designer may therefore plan how to draw a complex sample according to such features, then assign weights to individuals in the sample to reflect the unequal probability of being selected (a process which is described more fully later).

Popular statistical software packages often use the analysis of a simple random sample as its default setting. Fortunately, special modules or procedures in such packages have been developed for the analysis of complex sample data. Such examples include COMPLEX SAMPLES in SPSS (International Business Machines Corporation [IBM], 2017a), SURVEY in SAS (SAS Institute Inc., 2017), SVY in Stata (StataCorp LLC, 2017), TYPE=COMPLEX in the ANALYSIS command in Mplus (Muthén & Muthén, 2017), and the Survey package in R (Lumley, 2019).

Although complex sampling and its related analytical techniques have been discussed since the 1930s (e.g., Groves, 2011; Valliant et al., 2018; Lewis, 2017; Heeringa et al., 2017; Lee & Forthofer, 2005; Lumley, 2010; Chambers & Skinner, 2003; Biemer & Lyberg, 2003), much of the often very mathematical content assumes that readers already possess advanced training in psychometrics and statistics. In practice, however, many applied researchers may lack such formal training, so many pieces may be inaccessible and lacking user-friendliness to a broader readership. Although certain resources do exist in the research literature

which are designed to assist users with the analysis of complex survey data in SAS (e.g., Lewis, 2017), Stata (e.g., Heeringa et al., 2017; Kreuter & Valliant, 2007), and R (e.g., Lumley, 2010), resources focusing on the analytical procedures required for SPSS' Complex Samples module, specifically, are few. It should be noted that Heeringa et al. (2017) focus on Stata specifically, but their companion web site provides syntax to replicate their book examples using several software packages, including SPSS (see http://www.isr.umich.edu/src/smp/asda/; a blind reviewer is thanked for this information).

For these reasons, the overarching aim is to provide a step-by-step introduction to the analytic procedures required for SPSS' Complex Samples module. The objectives are two-fold. First, an explanation is provided for the most commonly-discussed procedures and technical terms relevant to complex sampling. The concepts described are introductory and not exhaustive. Second, using the McCreary Centre Society's British Columbia Adolescent Health Survey (Green et al., 2013) as case study data by which to demonstrate the complex samples analytical process, annotated SPSS syntax that may be saved and modified freely in the SPSS syntax editor is provided. The syntax is included because point-and-click sequences performed in graphical user interfaces are often cumbersome and cannot be saved, which is not advisable when undertaking a complicated analytic process.

Complex Survey Design

Process and Key Terms

Target population. A survey designer's first priority should be to define a target population that best meets the specific research question(s). If the size of a target population can be enumerated, then it is a finite population (e.g., Everitt & Skrondal, 2010). For the BCAHS 2013, the target population was specified as six cohorts of students from Grades 7 to 12, who were enrolled in public schools across the province of British Columbia, Canada, during the 2012/2013 school year (N = 260,632) (Saewyc et al., 2014). With the finite target population defined, a survey designer can proceed with defining a sampling frame.

Sampling frame. A sampling frame refers to a list of the target population from which a sample can be drawn, such as geographical listings or membership lists (Biemer & Lyberg, 2003; Everitt & Skrondal, 2010; Dodge et al., 2003). In the case of over-coverage, the sampling frame may also contain ineligible units which

are beyond scope (a blind reviewer is thanked for this reminder). In complex sampling, a sampling frame can be specified by various characteristics. For example, in BCAHS 2013, the sampling frame was comprised of all classrooms in British Columbia, stratified by school district and by grade, from Grades 7 to 12 (Saewyc et al., 2014). The 59 public school districts in British Columbia constitute 16 Health Service Delivery Areas (HSDA), which in turn group into five larger Health Authority (HA) areas (Saewyc et al., 2014). The division of a target population into partitions is called stratification, which is used particularly for the purpose of sampling (Everitt & Skrondal, 2010; Statistics Canada, 2003). A sampling unit refers to an entity that will be sampled from the target population according to the sampling design (Everitt & Skrondal, 2010; Biemer & Lyberg, 2003; Valliant et al., 2018). It can be a group of people with a specified characteristic or a set of hospitals in the same geographical area, as two examples. Here, the term "unit" differs from an "individual" in the sample, the latter of which typically refers to a person.

Cluster sampling. Another common feature of complex sampling is cluster sampling, which refers to sampling natural groups from the sampling frame, such as households in a community or classrooms in a school (Everitt & Skrondal, 2010; Marriott, 1990). With the BCAHS 2013, the drawing unit is the classroom rather than the individual student (Saewyc et al., 2014). If a classroom is drawn, all students in the classroom are included in the sample. In this sense, classrooms represent natural clusters of students.

Sampling stages. Unlike simple random sampling which only has one stage, complex sampling may have multiple sampling stages – meaning the sample is completed in two or more stages (Lewis, 2017; Marriott, 1990). First, a sample of broader units are drawn from the sampling frame. Then a sample of smaller units are drawn, only from those selected broader units in the first stage, and so forth for any additional stages required. Last, the final sample is comprised of all selected individuals in the last stage – for instance, a sample of schools drawn from a sampling frame of school districts (first stage), then a sample of classrooms drawn from the chosen schools (second stage). Assuming the sampling design only involves two stages, then all students in those selected classrooms make up the final sample.

In multiple-stage sampling, units drawn in the first stage are called primary sampling units (PSU), those in the second stage are called secondary sampling units (SSU), and so forth (e.g., Lewis, 2017; Heeringa et al., 2017). It should be clarified,

however, that multiple-stage sampling is not mandatory in complex sampling. For instance, BCAHS 2013 involves only one stage. A random sample of classrooms was drawn from each grade within each school district separately; these samples constituted the entire sample (Saewyc et al., 2014). If a sample of schools had been drawn first, followed by a sample of classrooms from each selected school, then it would have been a two-stage sampling.

Sample weights. Unlike the equal probability of being chosen in simple random sampling, individuals in a complex sample may have different probabilities of selection. These different probabilities of selection are specified by sample weights. A sample weight refers to a number assigned to each element or individual in the sample to reflect a relative importance (Upton & Cook, 2014; Lee & Forthofer, 2005). In a data set, each individual has a corresponding value in a weight variable. Usually, the sample weight can be viewed as the inverse of the selection probability, which is also called the base weight (Lee & Forthofer, 2005; Valliant et al., 2018). Imagine a sample of 100 is randomly drawn from a finite population of 1,000,000 with equal probability. The probability of selection is then 100/1,000,000 (i.e., 0.0001). If so, each person in the sample can be given a weight of 10,000 to represent a group of 10,000 people in the population (the reciprocal of 0.0001). Note in a situation of unequal probability of selection, weights may be unequal. Sample weights are crucial to the complex survey design. All estimations are adjusted by multiplying or adding a weight in a complex survey data analysis (Lee & Forthofer, 2005). Each individual in the sample is assigned a weight to reflect the extent to which he or she would occur in repeated sampling from the population using the given sampling design.

In practice, a base weight may be adjusted for practical reasons, beyond simply reflecting the unequal probability of selection. For instance, a weight can be adjusted due to non-response. Non-response is pervasive in survey practice, sometimes because participants may fail to provide some amount of survey information and/or the selected individual is impossible to contact or simply refuses to participate, as some examples (Everitt & Skrondal, 2010; Särndal et al., 1992). Such non-response results in a smaller valid sample size than expected. As a result, their original weights need to be adjusted accordingly. Another common reason a weight may be adjusted is that it may scale statistical estimations to the population rather than to the sample (Everitt & Skrondal, 2010). A weight variable adjusted to the scale of the population is called an expansion weight, while one adjusted to the scale of the sample is called a relative weight (Lee & Forthofer, 2005). Note a data set can have both expansion weight(s) and relative weight(s) simultaneously,

depending on the analyst's research interest. In BCAHS 2013, three aspects were considered when assigning weights to individual participants: the selection probability, non-response adjustment, and population readjustment (Saewyc et al., 2014). In other words, the original weights took the selection probability into account first, then were adjusted for compensation of non-response, and were finally readjusted to represent the corresponding proportion in the population.

Design effect. All of the aforementioned complex sampling features – stratification, clustering, and unequal weighting – influence the estimation of population statistics in different ways (Heeringa et al., 2017; Lewis, 2017). Generally, stratification decreases standard errors, namely the standard deviation of the sampling distribution of a statistic such as a mean. As such, stratification often increases the precision of an estimate (i.e., yields a smaller standard error compared to that in a simple random sampling framework with equal sample size). In contrast, clustering and unequal weighting tend to enlarge standard errors (i.e., yields a larger standard error compared to that in a simple random sampling framework with equal sample size). These factors, then, collectively impact an estimate in a multifaceted manner. A design effect is a measure of the collective influences from these complex sampling features (Lewis, 2017; Kish, 1965). The design effect of an estimate ($\hat{\theta}$), such as a mean, is defined as described in equation (1) (Lewis, 2017, p. 20):

$$Deff = \frac{\mathrm{Var}_{\mathrm{CS}}\left(\hat{\theta}\right)}{\mathrm{Var}_{\mathrm{SRS}}\left(\hat{\theta}\right)}, \tag{1}$$

where *Deff* refers to the design effect, $\mathrm{Var}_{\mathrm{CS}}(\hat{\theta})$ denotes the complex sample design variance of the estimate, and $\mathrm{Var}_{\mathrm{SRS}}(\hat{\theta})$ denotes the variance of the estimate of the same quantity that would be produced by an SRS of the same number of elements that are in the CS.

If *Deff* equals one, then a complex sample has the same design effect as a simple random sample of equal sample size. If *Deff* is larger than one, then $\mathrm{Var}_{\mathrm{CS}}(\hat{\theta})$ is larger than $\mathrm{Var}_{\mathrm{SRS}}(\hat{\theta})$, suggesting $\mathrm{Var}_{\mathrm{CS}}(\hat{\theta})$ is less precise. Theoretically, a design effect may range from 0 to positive infinity. Although a design effect can be smaller than one, it is usually greater than one (Lewis, 2017), particularly for clustered samples. A small design effect (*Deff* < 1) suggests that the complex sample design is more efficient than the corresponding simple random

sampling design. From the perspective of sample size, a design effect of 0.8 would mean that a sample size of 800 elements in a complex sample would produce the same variance as a simple random sample of 1,000. On the other hand, a large design effect (*Deff* > 1) suggests that the complex sample design is less efficient. Correspondingly, if a design effect is 5 in a complex sample of 5,000, then a simple random sample of 1,000 would yield the same variance as the CS.

It is noteworthy that a complex sample design often involves a combination of one or more of the most common features described above. But it may also involve other features not mentioned in the current paper due to space limitations. Because complex sample designs are inherently heterogeneous, and depend heavily on the nature of the research questions and the specific data collected, readers are encouraged to read Valliant et al. (2018), Biemer and Lyberg (2003), Lewis (2017), Heeringa et al. (2017), Lumley (2010), Lee and Forthofer (2005), and Chambers and Skinner (2003) for more in-depth discussion of complex sampling.

Annotated SPSS Syntax

Now that readers have a clearer sense of complex survey design's processes and key terms, attention is now turned to how a researcher may undertake analyzing complex sample data using SPSS (in this case, Version 25 was used; see IBM, 2017a). In this section, annotated syntax is provided so as to serve as a convenient step-by-step guide to performing a complex sample analysis, using BCAHS 2013 as an illustrative case study. For more general information about statistical programming using SPSS syntax, please refer to IBM (2017b).

Because the BCAHS data were already fully anonymized and de-identified for research purposes, ethics approval from the University of British Columbia was not required; however, a Memorandum of Understanding (MOU) with McCreary Centre Society that outlined our proposed uses of the data, including the current manuscript, was signed and approved.

Complex sample plan file specification. A plan file contains the information about complex sample design – including strata (levels of stratification), clusters (nesting or grouping), weights, data collection stages, etc. (IBM, 2017a). In each complex sample command, one command line is required to specify the chosen plan file (described more fully in a later section), so that SPSS recognizes the details of the complex sample design and can, in turn, apply proper estimation methods to the complex survey data. After the features of complex sample design and estimation methods are specified properly in the plan file, the analyst can cite it in

the analytic procedure. Whether a one-stage or multiple-stage design, SPSS can handle it appropriately. No further consideration about the estimation method is needed in the analytical process.

Those using complex sample data may not always be the designer of the plan file. Instead, a plan file might be provided from the data steward.

In this example, the BCAHS 2013 complex sample data have been stored in a file called `AHS5.sav`, the plan file is called `AHS_plan.csaplan`, and both files have been saved in a data folder on a drive labeled '`E`'. For convenience, the syntax commands are outlined in numbered parentheses, but the parentheses and their contents should, of course, be omitted from the actual program itself.

Part 1: Basic operations

```
(1.1) GET FILE = 'E:\Data\AHS5.sav'.
```

- Opens the SPSS data file located here '`E:\Data\AHS5.sav`'. In SPSS, a command name is usually written in uppercase to distinguish it from other content. This is, however, just a convention, not a requirement. Note an SPSS command may have sub-command(s) and keyword(s), which can be separated with blank space(s) or line break(s). SPSS commands always end with a period.

```
(1.2) DATASET NAME ahs5.
```

- Names the current activated file as '`ahs5`'. The naming is arbitrary but should not conflict with SPSS preset key words or other internal names. A short and meaningful name improves the readability of the syntax. If an open file is not explicitly given a name, SPSS will name it automatically with `DataSet1`, `DataSet2`, and so forth. Once a file is named, the analyst can use the new name to refer to the data set.

```
(1.3) DATASET ACTIVATE ahs5.
```

- Activates data set `ahs5`. SPSS can open several data sets simultaneously. In this case, a data set should be activated first before it can be analyzed. If only one data set is open, then it is automatically considered to be the active file.

```
(1.4) DATASET CLOSE ahs5.
```

- Closes data file `ahs5` without saving changes. This can be done at the end of an analysis session, for example.

Part 2: Plan file

As described earlier, features of complex samples are specified and recorded in a plan file. With the BCAHS 2013, a plan file called `AHS5splt216.csaplan` is used.

The plan file – an XML (eXtensible Markup Language) file – can be opened and read with a text editor, such as Notepad in the Windows operating system. Please note that the contents of the plan file described below were preset by the survey designers/data stewards (Green et al., 2013) and this plan file was required to be used as-is when running analyses. The plan file contents are as follows:

(2.1) `<?xml version="1.0" encoding="utf-8" standalone="no"?>`

- Denotes the plan file is an `.xml` file, with version number 1.0, and is encoded with '`utf-8`'. It is not standalone, implying an `.xml` file parser needs external sources to interpret the content of this `.xml` file (Microsoft, 2011). In this example, the strata variable name, clustering variable name, and weight variable name all come from the data file, rather than from the plan file itself. Note lines in a plan file are not SPSS syntax, so they do not end with a period. Rather, they are organized with angle brackets.

(2.2) `<SPSSComplexSamples version="1.0">`

- Starting line for complex samples specification, showing the CS version of 1.0 (not to be confused with Version 25 of the SPSS software).

(2.3) `<Header copyright="Copyright (c) IBM Corp., 2012. All Rights Reserved."/>`

- Indicates the copyright information.

(2.4) `<AnalysisDesign SRSestimator="wor" numberOfStages="1">`

- Starting line of the complex sample analysis design specification. It also specifies how to estimate variance under the simple random sampling (SRS) assumption. Here, it uses the 'without replacement' (i.e., wor) method, and the number of stages equals one. The SRS estimator is used in the calculation of the design effect.

(2.5) `<AnalysisStage estimationMethod="wr" stageNumber="1">`

- Starting line of the stage specification. It also specifies how to estimate variance for the complex sample. Here, the 'with replacement' (i.e., wr) method is used, and the number of stages equals one.

(2.6) `<StrataVarList numberOfVariables="1">`

- Shows the number of variables specifying the stratification. There is only one variable.

(2.7) `<Variable name="STRspl216"/>`

- Shows the variable specifying the strata. The variable name is '`STRspl216`'.

(2.8) `</StrataVarList>`

- Indicates the end of the stratification specification.

(2.9) `<ClusterVarList numberOfVariables="1">`

- Starting line showing the number of variables specifying the clustering. There is only one variable. In the case of BCAHS 2013, the clusters were classrooms.

(2.10) `<Variable name="PSU2"/>`

- Shows the variable specifying the clustering. The variable name is '`PSU2`'. Note '`PSU`' means primary sampling unit. Because one-stage sampling was conducted in BCAHS 2013, the clustering variable was PSU as well, namely, the variable indexing classrooms.

(2.11) `</ClusterVarList>`

- Indicates the end of the clustering specification.

(2.12) `</AnalysisStage>`

- Indicates the end of the stage specification.

(2.13) `<Weight>`

- Starting line for the weight specification.

(2.14) `<Variable name="WTsplit216"/>`

- Indicates the weight variable. In this case, it is called '`WTsplit216`'. Note a complex sample data set may have more than one weighting variable. But in an analysis, only one weighting variable is specified. The weight used in this example is the adjusted weight discussed earlier.

(2.15) `</Weight>`

- Final line for the weight specification.

(2.16) `</AnalysisDesign>`

- Final line for the complex sample analysis design specification.

(2.17) `</SPSSComplexSamples>`

- Final line for the complex sample specification.

Part 3: Descriptive analyses: Frequencies

Descriptive analyses summarize and tabulate data for exploring the characteristics of data distributions (Everitt & Skrondal, 2010). This section describes how to perform basic frequency analyses in the Complex Samples module. In this example, data for a categorical variable is summarized.

```
(3.1) DATASET ACTIVATE ahs5.
```

- Activates data set `ahs5.sav`, as described above in command 1.3.

```
(3.2) CSTABULATE
```

- Generates a one-way frequency table or a two-way cross-tabulation (IBM, 2017b). CS denotes complex samples. Note all commands in the Complex Samples module begin with '`CS`'.

(3.2.1) `/PLAN FILE='E:\Data\AHS_plan.csaplan'`

- Subcommand of '`CSTABULATE`', which specifies the particular plan file for the complex survey data, as described in 'Complex sample plan file specification' section.

(3.2.2) `/TABLES VARIABLES= Q1AgeGroup Q24Grade`

- Another subcommand of '`CSTABULATE`', which specifies the specific variables for which frequency tables are requested. `Q1AgeGroup` is a categorical variable describing age ranges, and `Q24Grade` is a categorical variable denoting grade level. Note this command provides a one-way table. If a two-way table is preferred, one can use the keyword `BY`, such as `/TABLES VARIABLES= Q1AgeGroup BY Q24Grade`.

(3.2.3) `/SUBPOP TABLE=Q4Female DISPLAY=LAYERED`

- Another sub-command of '`CSTABULATE`', which specifies the sub-population(s) variable for which the analysis is performed. Here, frequency tables of values for the aforementioned variables (`Q1AgeGroup` and `Q24Grade`) are to be generated by Q4Female (a binary variable describing gender) for males and females, separately. `DISPLAY=LAYERED` specifies that the results for both sub-populations are shown in one table (i.e., a cross-tabulation table). The alternative is `DISPLAY=SEPARATE`, which specifies that the results for both sub-populations are shown in separate tables. If there is no sub-population, this sub-command can be removed.

(3.2.4) `/CELLS POPSIZE TABLEPCT`

- Selects which estimates will show in table cells. `POPSIZE` represents estimated population size, and `TABLEPCT` represents estimated population percentage. If both are chosen, the results will show both.

(3.2.5) `/STATISTICS SE CV CIN(95) COUNT DEFF DEFFSQRT CUMULATIVE`

- Selects relevant statistics. `SE` denotes the standard error, `CV` denotes the coefficient of variation, `CIN(95)` denotes the confidence interval at a level of 95 percent (a common alternative is 99 percent), `COUNT` denotes the unweighted count of valid cases in the sample, `DEFF` denotes the design effect, `DEFFSQRT` denotes the square root of the design effect, `CUMULATIVE` denotes the cumulative estimate through each value of the variable.

(3.2.6) `/MISSING SCOPE=TABLE CLASSMISSING=EXCLUDE.`

- Shows how to deal with missing values. `SCOPE=TABLE` denotes the table-by-table deletion, which means missing values are determined on the basis

of each variable (IBM, 2017a). Thus, in each table, the number of valid cases may vary from table to table. The alternative choice is `SCOPE=LISTWISE` (listwise deletion). `CLASSMISSING=EXCLUDE` means that the user-defined missing values are treated as invalid. If `CLASSMISSING=INCLUDE` is specified, it treats the user-defined missing values as valid. Note this line ends with a period, which means the end of the `CSTABULATE` command (which began in 3.2).

Part 4: Descriptive analyses: Descriptives

In this section, additional descriptive statistics are specified and estimated.

`(4.1) CSDESCRIPTIVES`

- As the command name suggests, this is a descriptive statistics command for complex samples.

`(4.1.1) /PLAN FILE='E:\Data\AHS_plan.csaplan'`

- The function of this sub-command is the same as that in 3.2.1.

`(4.1.2) /SUMMARY VARIABLES = Q1Age`

- Specifies the variables for summary. Only numeric variables are eligible for this sub-command. Here, `Q1Age` (age in years) is a continuous variable.

`(4.1.3) /MEAN TTEST=10`

- Requests the estimate of the mean and also performs a *t*-test of the null hypothesis that the estimate is equal to 10 (an arbitrary value only for illustration). `TTEST` is optional.

`(4.1.4) /SUM TTEST=100`

- Requests the estimate of the sum and also performs a *t*-test of the null hypothesis that the estimated sum is equal to 100 (once again, an arbitrary value). `TTEST` is optional.

`(4.1.5) /RATIO NUMERATOR=Q1Age DENOMINATOR=Q24Grade TTEST=5`

- Specifies a ratio to be estimated by using keywords `NUMERATOR` and `DENOMINATOR`. This value is the estimation of the ratio in the population. Note both variables should be numeric. Specification of the keyword `TTEST` is similar to that in former sub-command.

`(4.1.6) /STATISTICS SE CV COUNT POPSIZE DEFF DEFFSQRT CIN(95)`

- Similar to that in command 3.2.5.

`(4.1.7) /MISSING SCOPE=ANALYSIS CLASSMISSING=EXCLUDE.`

- Similar to 3.2.6. `SCOPE=ANALYSIS` is the default setting, which means the calculation of statistics is on the basis of all valid cases for analyzed variables (IBM, 2017b).

Part 5: Inferential analyses: General Linear Model (GLM)

Linear regression is a commonly-used analysis in many applied research settings. An analyst may want to explore relationships among variables under the complex survey framework. General Linear Model (GLM) is available in SPSS' Complex Samples module. In addition to linear regression, there are other three regression options: logistic (`CSLOGISTIC`), ordinal (`CSORDINAL`), and Cox proportional hazards (`CSCOXREG`). The syntax of these commands is similar to `CSGLM`. Readers may refer to IBM (2017a, 2017b) for more details.

`(5.1) CSGLM Q29Health BY Q4Female WITH Q1Age`

- Specifies the dependent variables and independent variables. Following the command name `CSGLM`, the dependent variable is specified as `Q29Health` (self-rated health, a continuous variable). The keyword `BY` specifies the categorical independent variables, and the keyword `WITH` defines the continuous variables. Note this command line only specifies the dependent variable and the independent variables with their types. The specification of linear model is done in 5.1.3 by `/MODEL`.

`(5.1.1) /PLAN FILE='E:\Data\AHS_plan.csaplan'`

- The function of this sub-command is the same as that in 3.2.1.

`(5.1.2) /DOMAIN VARIABLE=Q30Allergy(1)`

- Specifies a particular level of categorical variable as the sub-population for which the general regression is to be performed (IBM, 2017b). It is similar to previous `SUBPOP`, specifying which categorical variable is to be used to define the sub-population. For example, `Q30Allergy(1)` in the above syntax denotes youth with allergic conditions, while `Q30Allergy(0)` denotes youth without. The above sub-command means only participants with allergic conditions in the sample are included in the GLM analysis. Note this sub-command is optional.

`(5.1.3) /MODEL Q4Female Q1Age Q4Female*Q1Age`

- Defines the variables included in the GLM. A single variable name represents a main effect (e.g., `Q4Female`, `Q1Age`), and the combination of two or more independent variables with an asterisk denotes their interaction (e.g., `Q4Female*Q1Age`).

`(5.1.4) /INTERCEPT INCLUDE=YES SHOW=YES`

- Specifies the intercept. As the command name suggests, `INCLUDE=YES` indicates an intercept is included, and `SHOW=YES` indicates the estimated intercept is shown in the result.

`(5.1.5) /STATISTICS PARAMETER SE CINTERVAL TTEST DEFF DEFFSQRT`

- Specifies the regression coefficients and relevant statistics. `PARAMETER` denotes the estimated regression coefficients. Other keywords are similar to those in aforementioned syntax.

`(5.1.6) /PRINT COVB CORB SUMMARY VARIABLEINFO SAMPLEINFO`

- Specifies what is displayed in the output. Keyword `COVB` denotes the covariances of estimated parameters. `CORB` denotes the correlations of estimated parameters. `SUMMARY` represents a statistical summary, including value of R^2. `VARIABLEINFO` represents the variable information, and `SAMPLEINFO` represents the sample information.

`(5.1.7) /TEST TYPE=F PADJUST=LSD`

- Specifies what kind of statistical test is used. `TYPE=F` means a Wald *F* test of the null hypothesis that all parameters in the linear model are equal to zero. Alternative options are adjusted Wald *F* (`ADJF`), Wald chi-square (`CHISQUARE`), and adjusted Wald chi-square (`ADJCHISQUARE`). Meanwhile, `PADJUST=LSD` indicates the adjustment method for the significance level is LSD (least significant difference). Alternative options for `PADJUST` are Sidak (`SIDAK`), Sequential Sidak (`SEQSIDAK`), Bonferroni (`BONFERRONI`), and Sequential Bonferroni (`SEQBONFERRONI`) (IBM, 2017b).

`(5.1.8) /MISSING CLASSMISSING=EXCLUDE`

- Similar to 3.2.6.

`(5.1.9) /CRITERIA CILEVEL=95.`

- Controls statistical criteria. Here, the confidence interval of the coefficient estimates is set at a level of 0.95.

Part 6: Macros: DO Loop

An SPSS macro is a user-defined function to automate a variety of SPSS tasks, not just those related to Complex Samples. It is especially useful for simplifying the programming for multiple and/or repetitive commands (IBM, 2017b). Here, a simple macro demonstrating how to generate a frequency table for a series of variables in the BCAHS 2013 data file is presented. Imagine there are four categorical variables (e.g., four special need condition variables, each of which denotes "Yes" or "No"), and an analyst intends to generate a frequency table for each variable. A macro can be defined to do this job automatically – and, in fact, for dozens and even hundreds of similar analyses.

`(6.1) DEFINE !fre (conditions= !CMDEND)`

- A macro starts with `DEFINE`, followed by a user-defined macro name. The macro name usually starts with an exclamation mark to distinguish it from

other variable names (IBM, 2017b). For example, `!fre` denotes a macro name in which `fre` is the abbreviation for frequency. The parentheses contain macro arguments. In the above syntax, a list of the condition variables is specified (which represent a series of special need conditions in BCAHS 2013). `!CMDEND` means the variable list ends at the end of the command (see command 6.7, `condition_1` through `condition_4`). Alternatively, if it is specified as '`conditions= !TOKENS(2)`' in the parentheses, then only the first two variables would be called (i.e., `condition_1` and `condition_2` in command 6.7).

(6.2) `!DO !condition !IN (!conditions)`

- Specifies a `DO` loop in the macro. A loop is a command which runs other commands repetitively. Command `!DO !condition !IN (!conditions)` instructs SPSS to pick a condition variable from the conditions list each time by order and to, in turn, run the analysis specified in the loop (6.3 - 6.4.5). The command sequence 6.3 - 6.4.5 is similar to 3.1 - 3.2.6, except for sub-command 6.4.2, in which `VARIABLES` is specified as the list element `!condition`. When the macro is called, each variable in the list would replace `!condition` once and the command sequence 6.3 - 6.4.5 is iterated four times (the number of variables in the list in command 6.7).

(6.3) `DATASET ACTIVATE ahs5.`

- This command activates the data file `ahs5`.

(6.4) `CSTABULATE`

(6.4.1) `/PLAN FILE='E:\Data\AHS_plan.csaplan'`

(6.4.2) `/TABLES VARIABLES= !condition`

(6.4.3) `/CELLS POPSIZE ROWPCT`

(6.4.4) `/STATISTICS SE CV CIN(99) COUNT DEFF DEFFSQRT`

(6.4.5) `/MISSING SCOPE=TABLE CLASSMISSING=INCLUDE.`

- This line ends with a period, indicating the completion of the `CSTABULATE` command in the `DO` loop.

(6.5) `!DOEND.`

- Indicates the end of `DO` loop in the macro. Note it ends with a period, because the `DO` loop is now complete.

(6.6) `!ENDDEFINE.`

- Indicates the end of entire macro. Note it ends with a period, because the macro is now complete.

(6.7) `!fre conditions = condition_1 condition_2 condition_3 condition_4.`

- Calls the macro with its name `!fre`, and specifies a list of condition variables. Note the argument name here does not start with an exclamation point, which may cause some confusion but it nonetheless correct (IBM, 2017b). According to the number of variables in the list, the macro will call condition variables one by one and produce four frequency tables, one for each variable in the list of conditions.

Part 7: Macros: Nested DO Loop

A more complex example of a macro is introduced in this section, in which two series of variables are used in nested `DO` loops (a `DO` loop inserted into another `DO` loop), generating multiple cross-tabulation tables automatically.

```
(7.1) DEFINE !crosstab (conditions= !CHAREND("/") / vars= !CMDEND)
```

- The macro starts with `DEFINE` and specifies a name `!crosstab`. It then specifies two lists of variables in the parentheses: `conditions` and `vars`. There are two keyword arguments, separated by a slash symbol (the second one). The first argument, `conditions= !CHAREND("/")`, defines the name as 'conditions' and calls variables until reaching a slash symbol; the second argument, `vars= !CMDEND`, defines the name as '`vars`' (i.e., variables) and calls variables until the end of command (see command 7.8).

```
(7.2) !DO !var !IN (!vars)
```

- Starts the first `DO` loop.

```
(7.3) !DO !condition !IN (!conditions)
```

- Starts the second (nested) `DO` loop.

```
(7.4) CSTABULATE
```

- Syntax steps 7.4 - 7.4.5 are similar to 6.3 - 6.4.5, except for sub-command 7.4.2, in which `!condition` is replaced by one element in variable list `!conditions` and `!var` is replaced by one element in variable list `!vars`, respectively.

```
(7.4.1) /PLAN FILE='E:\Data\AHS_plan.csaplan'
(7.4.2) /TABLES VARIABLES= !condition BY !var
(7.4.3) /CELLS POPSIZE ROWPCT
(7.4.4) /STATISTICS SE CV CIN(99) COUNT DEFF DEFFSQRT
(7.4.5) /MISSING SCOPE=TABLE CLASSMISSING=INCLUDE.
(7.5) !DOEND.
```

- Ends the `DO` loop 7.3.

```
(7.6) !DOEND.
```

- Ends the `DO` loop 7.2.

```
(7.7) !ENDDEFINE.
```

- This is the end of entire macro.

```
(7.8) !crosstab conditions = condition_1 condition_2 condition_3
condition_4 / vars= var_1 var_2 var_3.
```

- Calls the macro with its name `!crosstab` and specifies two lists of variables (i.e., `conditions` and `vars`). The first list starts from `condition_1` and ends at the slash symbol. The second list starts from `var_1` and ends when it reaches the period (the end of the command). Depending on the number of variables in the lists, the macro will run the nested loop and produce twelve (3×4) cross-tabulation tables. Macros can do repetitive work much more efficiently, which facilitates an analyst's job. Without macros, an analyst has to repeat numerous lines of syntax to do similar tasks.

Discussion

Complex sampling is widely used in education and the social sciences, especially in large scale survey projects. A sample design is complex due to numerous factors such as stratification, clustering, multiple stages, and weights. Typically, complex sample data should be analyzed under the framework of a complex sample design, rather than that of a simple random sample. If complex sample features are not properly taken into account in analyses, relevant statistical estimations may be biased. Thankfully, many mainstream statistical software packages such as SPSS, SAS, Stata, and MPlus support the analysis of complex sample data.

As shown in the example of BCAHS 2013, using SPSS syntax to analyze complex survey data has several advantages compared to point-and-click methods of data analysis. First, the meaning of SPSS' syntax is generally straightforward, in the sense that commands, sub-commands, and keywords are often readily comprehensible to users. Second, SPSS' syntax is succinct. Several lines of commands can run a complicated analysis, and complex statistical methods can be specified with pre-defined keywords. Third, complex sample features are incorporated in a plan file and therefore are separate from analysis procedures. Therefore, analysts are not always required to know how the original sampling procedure is designed. Rather, they simply need to specify the plan file when running the analysis, based on the instructions of the data steward. That said, analysts are always encouraged to read relevant documents about survey design for a better understanding of any analyses undertaken. Finally, SPSS' syntax can be readily organized and saved, allowing users to edit syntax and to reproduce

analytical results afterward. In summary, the readability, simplicity, and reproducibility of SPSS' syntax render it superior to point-and-click methods, especially in complicated analyses.

There is a convenient means to transfer from the point-and-click method to the syntax method. Regular syntax can be generated by pressing the "Paste" button in the dialogue box when using the point-and-click method. In addition, syntax can be set to be shown in the SPSS output window by ticking the "Display commands in the log" box in the "Options" window. Therefore, beginners may generate and read the SPSS syntax first, then edit it in the syntax editor in accordance with their needs. Some procedures, however, cannot be achieved by the point-and-click method, such as macros. In this situation, users are required to write command lines directly in the syntax editor.

Conclusion

Complex survey design is found increasingly in education and the social sciences, and a wealth of exciting complex sample data are available for analysis, such as the BCAHS 2013. Compared to other popular statistical packages, few resources show how to analyze complex sample data in SPSS, specifically. Filling this gap in the research literature was the motivation of the current primer. It is hoped that this piece facilitates the analyses of applied researchers across a variety of academic disciplines.

Acknowledgements

Thank you to the McCreary Centre Society (https://www.mcs.bc.ca/), who collects and owns the British Columbia Adolescent Health Survey data. Thanks also to Dr. Colleen Poon, Allysha Ram, Dr. Elizabeth Saewyc, and Annie Smith for their guidance as we worked with the data. Finally, thanks to blind reviewers for their comments that improved the paper. An SPSS syntax file with the commands outlined in this paper is available for download at: http://blogs.ubc.ca/jenniferlloyd/

References

Biemer, P. P., & Lyberg, L. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley. doi: 10.1002/0471458740

Chambers, R. L., & Skinner, C. J. (2003). *Analysis of survey data*. Hoboken: Wiley. doi: 10.1002/0470867205

Dodge, Y., Marriott, F. H. C., International Statistical Institute. (2003). *The Oxford dictionary of statistical terms* (6th edition). Oxford, UK: Oxford University Press.

Everitt, B., & Skrondal, A. (2010). *The Cambridge dictionary of statistics* (4th edition). New York: Cambridge University Press.

Green, R., Saewyc, E., & Stewart, D. (2013). Plan file for the 2013 British Columbia adolescent health survey [Data file and code book]. Vancouver, British Columbia: McCreary Centre Society.

Groves, R. M. (2011). Three eras of survey research. *The Public Opinion Quarterly, 75*(5), 861-871. doi: 10.1093/poq/nfr057

Heeringa, S., West, B. T., & Berglund. (2017). *Applied survey data analysis* (2nd edition). Boca Raton, FL: Chapman & Hall/CRC.

International Business Machines Corporation. (2017a). *IBM SPSS complex samples 25*. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/25.0/en/client/Manuals/IBM_SPSS_Complex_Samples.pdf

International Business Machines Corporation. (2017b). *IBM SPSS statistics 25 command syntax reference*. Retrieved from ftp://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/25.0/en/client/Manuals/IBM_SPSS_Statistics_Command_Syntax_Reference.pdf

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Kreuter, F., & Valliant, R. (2007). A survey on survey statistics: What is done and can be done in Stata. *The Stata Journal, 7*(1), 1-21. doi: 10.1177/1536867x0700700101

Lee, E. S., & Forthofer, R. N. (2005). *Analyzing complex survey data* (2nd edition). Thousand Oaks, CA: Sage Publications. doi: 10.4135/9781412983341

Lewis, T. H. (2017). *Complex survey data analysis with SAS*. Boca Raton, FL: CRC Press.

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Hoboken, N.J: John Wiley. doi: 10.1002/9780470580066

Lumley, T. (2019). survey: Analysis of complex survey samples [R software package]. Retrieved from: https://cran.r-project.org/package=survey

Marriott, F. H. C. (1990). *A dictionary of statistical terms*. London, UK: Longman Scientific & Technical.

Microsoft. (2011, February 21). XML declaration. Retrieved from https://docs.microsoft.com/en-us/previous-versions/dotnet/netframework-4.0/ms256177(v=vs.100)

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th edition). Los Angeles, CA: Muthén & Muthén. Retrieved from https://www.statmodel.com/ugexcerpts.shtml

National Center for Education Statistics. (2017). High school and beyond (HS&B) longitudinal study. Retrieved from https://nces.ed.gov/statprog/handbook/pdf/hsb.pdf

National Center for Education Statistics. (n.d.). National postsecondary student aid study (NPSAS). Retrieved from: https://nces.ed.gov/surveys/npsas/about.asp

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society, 97*(4), 558-625. doi: 10.2307/2342192

Saewyc, E., Stewart, D., & Green, R. (2014). *Methodology for the 2013 BC adolescent health survey*. Vancouver, British Columbia: McCreary Centre Society. Retrieved from https://www.mcs.bc.ca/pdf/AHSV_methodology.pdf

Särndal, C., Swensson, B., & Wretman, J. H. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.

SAS Institute Inc. (2017). Introduction to survey sampling and analysis procedures. In *SAS/STAT® 14.2 user's guide* (pp. 237-249). Cary, NC: SAS Institute Inc. Retrieved from https://support.sas.com/documentation/onlinedoc/stat/142/introsamp.pdf

StataCorp LLC. (2017). *Stata survey data reference manual* (Release 15). College Station, TX: Stata Press. Retrieved from https://www.stata.com/manuals/svy.pdf

Statistics Canada. (2003). *Statistics Canada quality guidelines* (4th edition). Ottawa, Ontario: Minister of Industry. Retrieved from https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2003001-eng.pdf

Statistics Canada. (2012). National population health survey: Household component, longitudinal (NPHS). Retrieved from https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3225&lang=en&db=imdb&adm=8&dis=2

Statistics Canada. (2017). *Methodology of the Canadian labour force survey*. Ottawa, Ontario: Minister of Industry. Retrieved from: https://www150.statcan.gc.ca/n1/en/catalogue/71-526-X

Statistics Canada. (2018). Canadian community health survey - Annual component (CCHS). Retrieved from: https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3226

Upton, G. J. G., & Cook, I. (2014). *A dictionary of statistics* (3rd edition). Oxford, UK: Oxford University Press.

Valliant, R., Dever, J. A., & Kreuter, F. (2018). *Practical tools for designing and weighting survey samples* (2nd edition). New York: Springer. doi: 10.1007/978-1-4614-6449-5

Westat. (2017, February). *Main survey school sampling preparation manual: Overview*. Paris: Organisation for Economic Co-operation and Development. Retrieved from https://www.oecd.org/pisa/pisaproducts/MAIN-SURVEY-SCHOOL-SAMPLING-PREPARATION-MANUAL.pdf

World Values Survey. (n.d.). Fieldwork and sampling. Retrieved from https://www.worldvaluessurvey.org/WVSContents.jsp?CMSID=FieldworkSampling