Quality of Life Survey 6 2020/21

QUALITY OF LIFE SURVEY 6 (2020/21)

GUIDE TO WEIGHTED DATA ANALYSIS

NOVEMBER 2021

Author:
Ariane Neethling

GCRO Gauteng City-Region Observatory

Quality of Life 6 (2020/21) survey: Guide to weighted data analysis

Authors: Ariane Neethling[1].

Date: November 2021

Type of output: Technical Report

Research theme: Understanding quality of life

Cover Image: Alexandra Parker

Published by the Gauteng City-Region Observatory (GCRO), a partnership of the University of Johannesburg, the University of the Witwatersrand, Johannesburg, the Gauteng Provincial Government and organised local government in Gauteng (SALGA).

Suggested citation: Neethling, A. (2021). *Quality of Life Survey 6 (2020/21): Guide to weighted data analysis*. Gauteng City-Region Observatory (GCRO). Johannesburg.

[1] Statistical consultant and part-time lecturer, Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa.
Email: ariane_neethling@yahoo.com

TABLE OF CONTENTS

1 INTRODUCTION

Household surveys often have a stratified multistage cluster sampling designs. This is also called complex sampling, and is the approach used for the GCRO Quality of Life surveys. When working with survey data obtained from a complex sample it is crucial to understand the correct application of weights, along with the approporiate statistical techniques.

Standard statistical techniques in statistical software assume that the data were derived from a simple random sample, and therefore do not account for stratification, clustering, unequal-probability sampling, etc. When analysing complex sample data, it is consequently not enough to simply specify or apply the appropriate weights, as this does not take sample design into account. Although point estimates will be accurate when standard techniques are used on weighted data, these techniques will underestimate the standard error. In turn, this results in overestimation of precision (margins of error), and confidence intervals that are too narrow. Consequently, p-values in hypothesis tests and other statistical inference tests will be too small. This means that conclusions drawn, and decisions made on the basis of these analyses, can be wrong and misleading.

For complex sample data, the calculation of the sampling error requires specific techniques that take into account the complexity of the sample design as well as the weights of the respondents. Different textbooks about the analysis of complex survey data are available, such as Chambers & Skinner, 2003; Heeringa et al., 2017; Lehtonen & Pahkinen, 2004; Lohr, 2010; Valliant et al., 2018, but their use requires an advanced knowledge of statistics. Fortunately, all major statistical software packages have special procedures that can be used for the analysis of complex sample data, which can be applied without undue difficulty.

This document provides an overview of how to analyse complex sample survey data, and in particular the data from the GCRO Quality of Life 6 survey (2020/21). The QoL 6 (2020/21) sample was drawn using a multistage stratified cluster sampling strategy, as outlined in the sample design report (Hamann & de Kadt, 2021). The survey data is also weighted, as described the weighting report for the survey (Neethling, 2021).

Two sets of weights are included in the QoL 6 (2020/21) survey dataset – an individual weight for use when results are required per individual, and a household weight when conclusions are required in terms of households. For example, to determine the estimated percentage of *households* without running water inside the house, use the household weight. The use of the individual weight is required for the estimated percentage of *people* without running water inside the house. It is essential to use the weight variable which best suits your desired purposes, because results will vary, even when presented in percentage form, depending on which weight variable is used.

The weighting variable 'DOWNSCALE_MUN_PP_BENCHWGT' provides an individual level weight, scaled to the QoL 6 (2020/21) sample size of 13 616, after it was calculated to sum to the population total. This is the default variable used in GCRO analysis, and is appropriate for use in all individual level analyses. Note that this weight can only be used for percentage (proportions) and mean estimates. If an estimate of actual population size is desired, the original weight (before

it was downscaled) should be used. The weighting variable 'HH_WEIGHT' provides a household level weight. This weight is suitable for use in any household level analyses. This weight has not been downscaled, so the frequency figures are the estimated total number of households in Gauteng province.

In Section 2 of this document, an introduction is given to some of the survey analysis procedures in SAS, SPSS, R and Stata that can be used when analysing complex sample data. Section 3 provides a brief comparison between the results of standard techniques and complex sample techniques.

2 USING STATISTICAL SOFTWARE PACKAGES FOR ANALYSING COMPLEX SURVEY DATA

2.1 Complex sample analysis using SAS

SAS has a range of different procedures that should be used when data are obtained from a probability sample design including stratification and/or clustering. All these procedures start with the word 'survey' (SAS Institute Inc., 2017). The procedures 'surveymeans', 'surveyfreq', 'surveyreg', 'surveylogistic', and 'surveyphreg' take the sample design into account, whether it is a single-stage or multistage design, with or without stratification, and with equal or unequal weighting. Box 1 provides an overview of the process in SAS.

Box 1: Complex sample analysis using SAS

Use the command 'proc surveyfreq'

For the QoL 6 (2020/21) question: "*Q7.5 - How satisfied are you with the performance of Gauteng Provincial Government?*"

```
proc surveyfreq data=sd.gcro_qol6;
stratum ward_code;
cluster ea_code;
weight DOWNSCALE_MUN_PP_BENCHWGT;
tables q7_5_pg;
run;
```

Output:

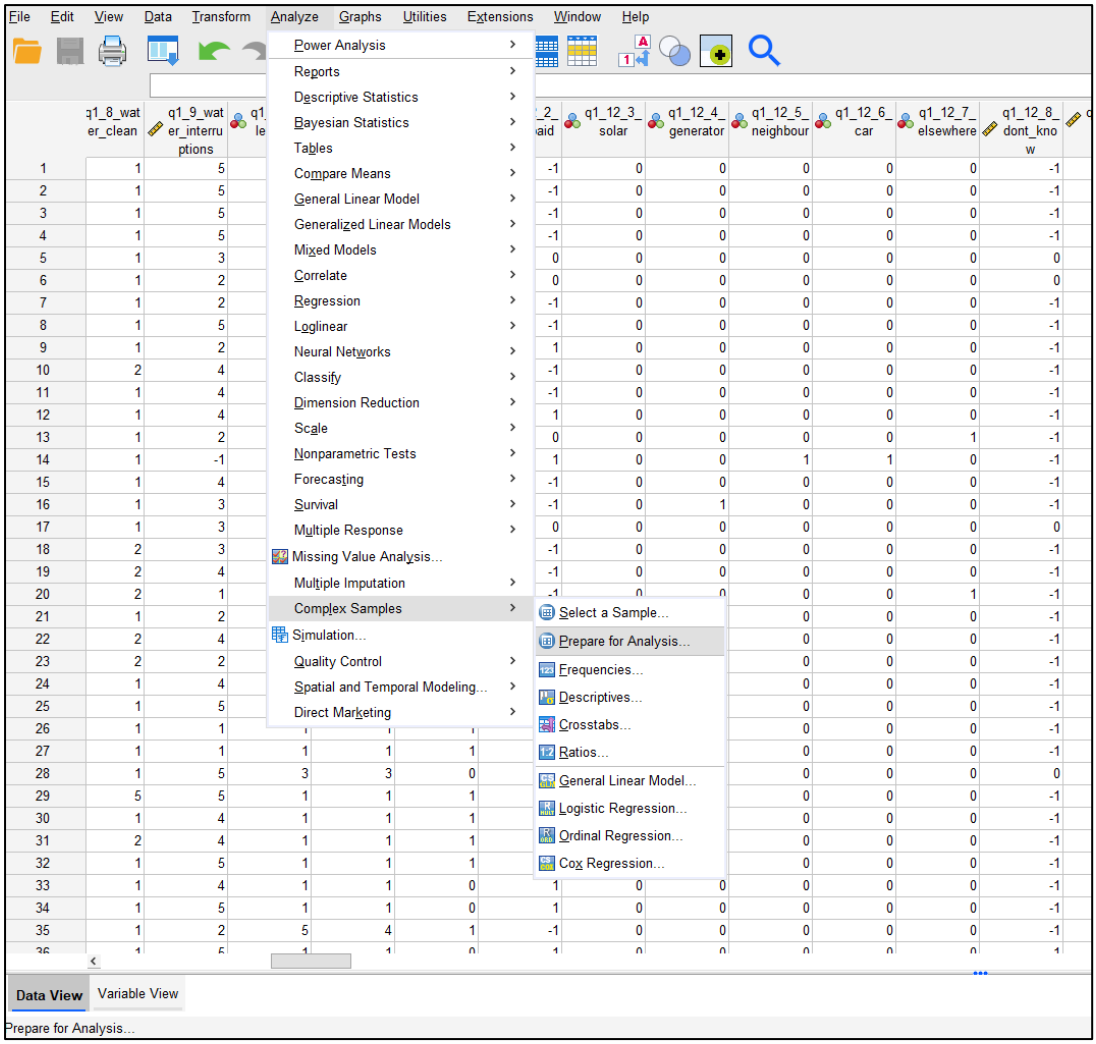
Q7.5 - How satisfied are you with the performance of Gauteng Provincial Government?					
q7_5_pg	Frequency	Weighted Frequency	Std Dev of Wgt Freq	Percent	Std Err of Percent
Very satisfied	455	482.75262	29.02530	3.5455	0.2117
Satisfied	3535	3453	68.36115	25.3629	0.4806
Neither satisfied nor dissatisfied	2408	2496	60.51266	18.3290	0.4294
Dissatisfied	4456	4553	79.52815	33.4375	0.5284
Very dissatisfied	2762	2631	61.88471	19.3251	0.4332
Total	13616	13616	91.84168	100.000	

2.2 Complex sample analysis using SPSS

The special module in SPSS that can be used for complex sample data is called 'Complex Samples' (International Business Machines Corporation [IBM], 2017a).

Before any analyses under the Complex Samples module can be executed, the complex sample plan has to be set up. Under Analyze > Complex Samples > Prepare for Analysis... (Figure 1), the stratification variable, clusters, weights, etc are defined and saved for future use of any of the methods specified in the dropdown menu under the Complex Samples module.

Figure 1: Prepare for analysis under the Complex Samples module

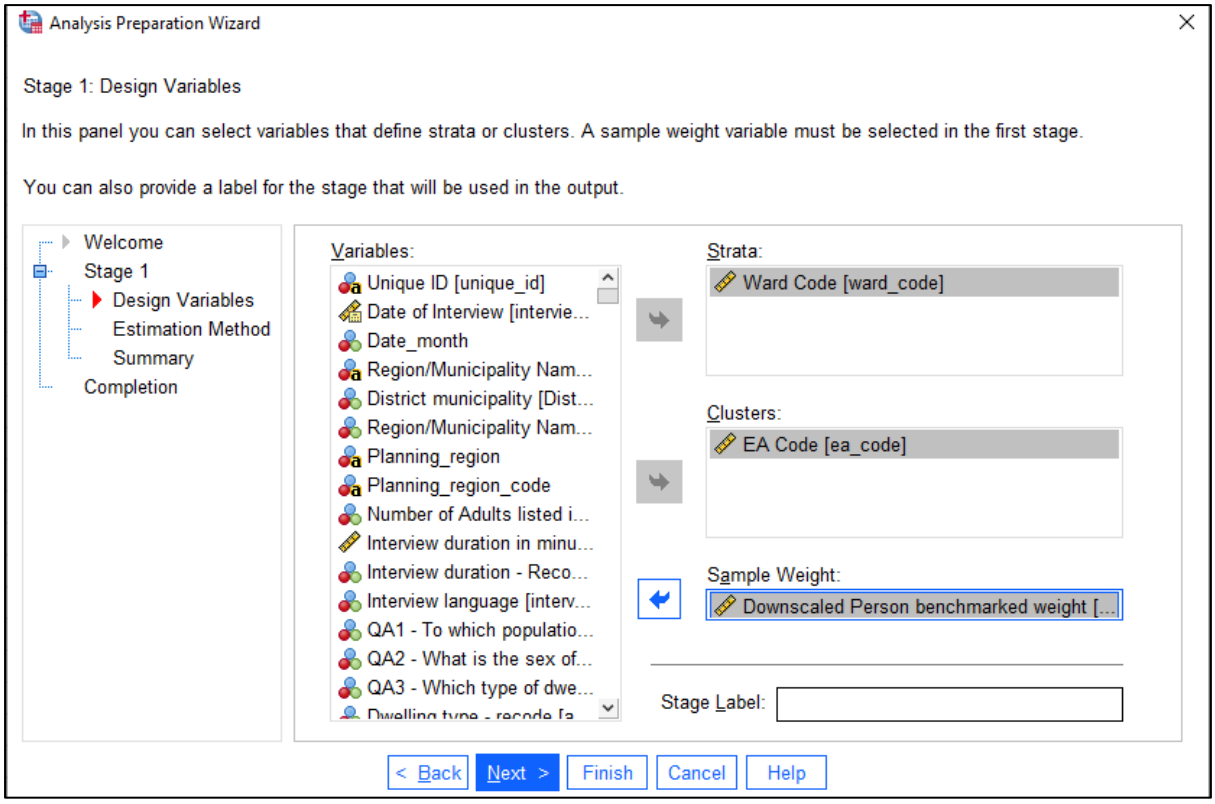


To prepare the complex sample plan for the GCRO QoL 6 survey, select 'Create a plan' on the first window and complete the next steps as follows:

- Use the 'Browse' button to select the location to save the plan that will be created and give it a name. It will be saved with a '.csaplan' file extension. Select 'Next'.
- In the following window, assign variable "Ward Code [ward_code]" to the 'Strata' box, "EA Code [ea_code]" to the 'Clusters' box and, the weight you would like to use to the 'Sample Weight' box. This is illustrated in Figure 2 below.
- Click the 'Finish' button.

The sampling plan is now saved and can be used for any future analysis under the Complex Samples module. Create a sampling plan for each weight of interest – hence, separate plans for the household weight and the individual weight.

Figure 2: Stipulate design variables in the Complex Samples module



After preparing the complex sample plan, any of the methods specified in the dropdown menu under the Complex Samples module in SPSS can be used for the analysis. On selecting any of these methods, first specify which complex sample plan to use in the analysis, whereafter the analysis can be run using that plan. For example, 'Frequencies' can be chosen to obtain univariate tabular statistics, by taking the defined sample design and weights into account. Figure 3, below, shows the result from SPSS when doing a 'Frequencies' for question Q7.5 of QoL 6 (2020/21) – *"How satisfied are you with the performance of Gauteng Provincial Government?"*.

Figure 3: Complex sample analysis using SPSS

Q7.5 - How satisfied are you with the performance of Gauteng Provincial Government?

		Estimate	Standard Error	95% Confidence Interval Lower	95% Confidence Interval Upper
% of Total	Very satisfied	3.5%	0.2%	3.2%	4.0%
	Satisfied	25.4%	0.5%	24.4%	26.3%
	Neither satisfied nor dissatisfied	18.3%	0.4%	17.5%	19.2%
	Dissatisfied	33.4%	0.5%	32.4%	34.5%
	Very dissatisfied	19.3%	0.4%	18.5%	20.2%
	Total	100.0%	0.0%	100.0%	100.0%

2.3 Complex sample analysis using R

In R, several different packages for complex sample analysis are available, and different ways of programming are also possible. This document provides an example of how to use the package *'survey'* in R (see Lumley, 2010 & 2019).

Similar to SAS and SPSS, it is important to define the complex sample features regarding stratification, clusters, weights, etc. In the *'survey'* package, this has to be done with 'svydesign' and is illustrated with the QoL 6 (2020/21) data in Box 2, below.

Box 2: Complex sample analysis using R

For the QoL 6 (2020/21) question: "*Q7.5 - How satisfied are you with the performance of Gauteng Provincial Government?*"

```
# Create survey design object
library("survey")
design <- svydesign( data = gcrodat, ids = ~ea_code, strata = ~ward_code,
 weights = ~ DOWNSCALE_MUN_PP_BENCHWGT, nest = TRUE)


#one-way table - only Est percentage
tb1 = svytable(~q7_5_pg,design=design)
tb2 = prop.table(svytable(~q7_5_pg,design=design))*100
tb2
```

Output:

Dissatisfied	Neither satisfied nor dissatisfied	Satisfied
33.43752	18.32902	25.36290
Very dissatisfied	Very satisfied	
19.32508	3.54548	

The output in the example can be coded further to obtain a better format of choice. An alternative option is to use 'svymean'. The standard error (SE), confidence interval (CI) and/or other estimates could also be calculated. For example, based on the same results from Q7.5 in Box 2, the output in Table 1 could also be generated.

Table 1: Complex sample analysis using R

Response options	Percent	SE	CI 2.5%	CI 97.5%
Very satisfied	3.54548	0.2117442	3.130275	3.960685
Satisfied	25.36290	0.4806008	24.420496	26.305297
Neither satisfied nor dissatisfied	18.32902	0.4294353	17.486948	19.171090
Dissatisfied	33.43752	0.5284071	32.401380	34.473666
Very dissatisfied	19.32508	0.4331570	18.475712	20.174450

2.4 Complex sample analysis using Stata

Stata can also be used to analyse complex sample data. All the survey commands have the prefix 'svy'. Similar to the other software package, the survey design has to be declared. In Stata, this has to be done by using the 'svyset' command. By using 'svytabulate' a one-way frequency table can be generated as shown for the other software tools. Detailed information on the different survey commands and syntax are summarised in StataCorp LLC (2017).

3 A COMPARISON BETWEEN STANDARD AND COMPLEX SURVEY TECHNIQUES

Most statistical software packages often assume a simple random sample as its default formulae. The package offers the option to add a weight for each observation (respondent). For calculating the percentages and the average of variables (point estimates), the standard technique options of statistical software packages, with the weights specified, will give the same values as the complex survey techniques. For other estimates and statistical inferences, the values and outcomes differ, sometimes significantly. This is because the standard techniques do not account for the stratification and cluster effect of the design used.

Many researchers are under the false impression that if the weights that were calculated to generalize the sample to the population, and thus sum to the population size, are downscaled to the sample size, the standard (default) statistical techniques could be applied in place of complex sample methods. The following example will illustrate that this is not correct, and that using standard statistical techniques for complex sample data should be considered a pitfall to be avoided.

By using SPSS, the following represents a comparison between the different techniques by using a few variables from the QoL 6 (2020/21) survey. The downscaled weight, DOWNSCALE_MUN_PP_BENCHWGT, is used throughout.

In order to illustrate this, compare the opinion of different age groups (18-34, 35-49, 50-64, 65+) for the following variables of the QoL 6 (2020/21) survey:

- ***Q7.5*** *– How satisfied are you with the performance of Gauteng Provincial Government?*

- ***Q7.9*** *– In general, do you think most government officials are doing their best to service the people according to the principles of Batho Pele?*
- ***Q8.4*** *– The country is going in the wrong direction*
- ***Q11.5*** *– How safe do you feel at home?*

Table 2: Comparing the p-values between standard and complex sample analysis.

Method	Q7.5 by Age group	Q7.9 by Age group	Q8.4 by Age group	Q11.5 by Age group
Standard (X wrong)	<0.001	0.002	0.021	<0.001
Complex sample (√ correct)	0.034	0.054	0.281	0.074

Based on the p-values in Table 2, and by using a significant level of 5%, the standard technique indicates significant differences among the age groups for all four variables. However, the complex sample test indicates that there are significant differences among the age groups only for Q7.5 (with a much higher p-value (0.034) compared to for the standard method with a p-value of <0.001). Thus, by using the standard statistical test, incorrect conclusions would have been drawn in three of the four cases.

The same percentages and p-values will be found for the complex sample method when the weights that sum to the population size (not downscaled) are used. This set of weights is the only set that can be used to find the estimated population counts per category or for any analysis done on population counts.

4 CONCLUSION

This document aims to make the users of statistical software aware of the incorrect use of the default standard statistical techniques under complex samples, or any sample with stratification and especially when clusters are present. The standard statistical packages underestimate the true variability under a complex sample design which can lead to too small standard errors, too narrow confidence intervals, incorrect p-values and statistical inference estimates. This can result in erroneous conclusions and decision making. It is recommended that a statistician with the necessary knowledge of complex survey techniques should be consulted, where necessary.

Furthermore, in this document, an introduction of using complex survey techniques with the statistical software packages SAS, SPSS, R and Stata is briefly summarized. The most accurate information regarding the use of the different modules of the software packages must be obtained from the manuals.

5 REFERENCES

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

Hamann, C. & de Kadt, J. (2021). *GCRO Quality of Life Survey 6 (2020/21): Sample design*. Johannesburg: Gauteng City-Region Observatory. Available at https://gcro.ac.za/research/project/detail/quality-life-survey-vi-202021/

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: https://public.dhe.ibm.com/software/analytics/spss/documentation/statistics/25.0/en/client/Manuals/IBM_SPSS_Complex_Samples.pdf

Lehtonen, R. & Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. (2nd edition). John Wiley & Sons.

Lohr, S.L. (2010). *Sampling: Design and Analysis*. (2nd edition). Brooks/Cole, Cengage Learning.

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

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

Neethling, A. (2021). *GCRO Quality of Life Survey 6 (2020/21): Weighting report.* Johannesburg: Gauteng City-Region Observatory. Available at https://gcro.ac.za/research/project/detail/quality-life-survey-vi-202021/

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

Valliant, R., Dever, J. A. & Kreuter, F. (2018). *Practical tools for designing and weighting survey samples* (2nd edition). New York: Springer.