General Psychiatry

Guidance for use of weights: an analysis of different types of weights and their implications when using SAS PROCs

Sabrina Richardson,[1,2] Tuo Lin,[3] Yangyi Li,[4] Xiaohui Niu,[5] Manfei Xu,[6] Valerie Stander,[2] Xin M Tu[2,3]

To cite: Richardson S, Lin T, Li Y, *et al*. Guidance for use of weights: an analysis of different types of weights and their implications when using SAS PROCs. *General Psychiatry* 2019;**32**:e100038. doi:10.1136/gpsych-2018-100038

Received 13 January 2019
Accepted 15 January 2019

[1]The Henry M. Jackson Foundation for the Advancement of Military Medicine, Inc, Bethesda, Maryland, United States
[2]Military Population Health, Naval Health Research Center, San Diego, Ca, United States
[3]Clinical and Translational Research Institute, University of California San Diego, San Diego, California, USA
[4]School of Mathematical and Statistical Sciences, Clemson University, Clemson, South Carolina, USA
[5]College of Informatics, Huazhong Agriculture University, Wuhan, China
[6]Shanghai Mental Health Center, Medical College, Shanghai Jiao Tong University, Shanghai, China

Correspondence to
Dr Xin M Tu; x2tu@ucsd.edu

ABSTRACT

SAS and other popular statistical packages provide support for survey data with sampling weights. For example, PROC MEANS and PROC LOGISTIC in SAS have their counterparts PROC SURVEYMEANS and PROC SURVEYLOGISTIC to facilitate analysis of data from complex survey studies. On the other hand, PROC MEANS and many other classic SAS procedures also provide an option for including weights and yield identical point estimates, but different standard errors (SEs), as their corresponding survey procedures. This paper takes an in-depth look at different types of weights and provides guidance on use of different SAS procedures.

INTRODUCTION

All popular SAS procedures support use of weights, such as the classic PROC MEANS, PROC GLM and PROC LOGISTIC. To facilitate analysis of survey study data, SAS also provides an array of procedures with a 'SURVEY' prefix, such as the PROC SURVEYMEANS, PROC SURVEYREG and PROC SURVEYLOGISTIC. As the latter are designed and developed specifically for survey study data, weights reflecting sampling and non-response are an integral part of these procedures. A common question is which SAS PROC to use when analysing survey study data involving sampling weights. Although they produce identical point estimates, the classic SAS procedures and their SURVEY counterparts generally yield different SEs, which in some cases lead to quite different p-values and conclusions. In this document, we discuss conceptual differences underlying the different types of weights and their implications in statistical methods developed touse the different SAS procedures. Although we focus on SAS in this paper, the discussions also apply to other statistical packages such as R and SPSS. For ease of exposition, we focus on the PROC MEANS and PROC SURVEYMEANS and illustrate our considerations with real and Monte Carlo simulated data. These same considerations will also apply to other SAS procedures for multiple variable analyses such as PROC LOGISTIC versus PROC SURVEYLOGISTIC; however, here we will focus on univariate analyses.

TYPES OF WEIGHTS AND METHODS UNDERLYING SAS PROC MEANS AND PROC SURVEYMEANS

Within these two procedures, the weights have different uses and meaning. Weights used in PROC MEANS are designed to address violations of 'homoscedasticity', a key assumption underlying many statistical methods such as inference for population means with the current context and, more generally, regression analysis. In linear regression models, 'the best linear unbiased estimate' (BLUE) is the most popular estimate. The BLUE estimate has the smallest variance among all competing estimates that are a linear combination of observations. However, if the assumption of homoscedasticity is not met, BLUE will generally be biased. In some cases, weights, a series of known constants for each of the observations, can be used to address such violations, or heteroscedasticity.

Note that regularised estimates, such as the popular 'least absolute shrinkage and selection operator' (LASSO), have become increasingly popular in recent years due to the surge of high-dimensional data arising in biomedical and online social media research. Although these estimates are generally biased, the bias is typically small. Moreover, when the number of independent variables exceeds the sample size in regression models, it is no longer possible to obtain unbiased estimates. Thus, the class of BLUE estimates becomes irrelevant in such applications.

Within the context of inference for the population mean, the sample mean is a BLUE estimate, if all observations have the same population mean and variance in addition to being independent, representing the so-called 'independently and identically distributed' (i.i.d.) sample. If the observations do not have the same variance, that is, a violation of the homoscedasticity assumption, the sample mean is no longer a BLUE estimate, even though it is still an unbiased estimate of the population mean. As the usual sample variance formula and t-statistic are both based on the i.i.d. assumption, the sample variance no longer describes the variability of the sample mean and t-statistic cannot be used to provide valid inference about the population mean. For example, a 95% CI based on the sample variance and t-statistic no longer covers the population mean 95% of the time. In most cases, causes of such heteroscedasticity are unknown and other statistical methods must be used to provide valid inference about the population mean, such as by using estimating equations (see below for details and examples). In some studies, heteroscedasticity is due to aggregating data, in which case weights can be used to 'correct' the type of heteroscedasticity so that the usual sample variance and t-statistic can continue to provide valid inference.

Example 1

Consider taking a random sample of 100 subjects from a population of interest and let y_i denote an outcome of interest from the ith subject. Such a sampling scheme is often called 'simple random sampling'. Let $\mu = E(y_i)$ denote the population mean and $\sigma^2 = Var(y_i)$ denote the population variance, where $E(y_i)$ denotes the mathematical expectation of y_i and $Var(y_i) = E(y_i - E(y_i))^2$.. We are interested in estimating the population mean μ.

Under simple random sampling, the 100 observations y_i form an i.i.d. sample, from which the sample mean $\hat{\mu}$, sample variance $\hat{\sigma}^2$, variance $\hat{\sigma}_\mu^2$ (SE $\hat{\sigma}_\mu$) of the sample mean and t-statistic can be computed:

$$\begin{aligned}
&\hat{\mu} = \frac{1}{100}\left(y_1 + y_2 + \ldots + y_{100}\right) = \frac{1}{100}\sum_{i=1}^{100} y_i,\\
&\hat{\sigma}^2 = \frac{1}{100-1}\left[\left(y_1 - \hat{\mu}\right)^2 + \left(y_2 - \hat{\mu}\right)^2 + \ldots + \left(y_{100} - \hat{\mu}\right)^2\right]\\
&\quad = \frac{1}{100-1}\sum_{i=1}^{100}\left(y_i - \hat{\mu}\right)^2, \qquad (1)\\
&\hat{\sigma}_\mu^2 = \frac{\hat{\sigma}^2}{100},\\
&t = \frac{\hat{\mu}}{\hat{\sigma}_\mu}.
\end{aligned}$$

The above statistics can be used to calculate CIs or test hypotheses of interest. For example, to test if the population mean is 0, we specify the hypothesis as:

$$H_0: \mu = 0 \quad \text{vs} \quad H_a: \mu \neq 0\,.$$

We can readily use the value of the t-statistic $t = \frac{\hat{\mu}}{\hat{\sigma}_\mu}$ to compute the p-value or construct a 95% CI.

Now suppose that we average the first 20 and last 10 observations and replace the first 20 and last 10 observations by their averaged counterparts:

$$\bar{y}_{1-20}, y_{21}, K, y_{90}, \bar{y}_{91-100},$$

where $\bar{y}_{1-20}$ and $\bar{y}_{91-100}$ denote the averaged first 20 and last 10 observations. We can recalculate all the statistics in Equation (1) using the two averaged outcomes plus the remaining 70 observations:

$$\begin{aligned}
&\hat{\mu} = \frac{1}{72}\left(\bar{y}_{1-20} + y_{21} + \ldots + y_{90} + \bar{y}_{91-100}\right),\\
&\hat{\sigma}^2 = \frac{1}{72-1}\left[\left(\bar{y}_{1-20} - \hat{\mu}\right)^2 + \left(y_{21} - \hat{\mu}\right)^2 + \ldots\right.\\
&\qquad \left. + \left(y_{90} - \hat{\mu}\right)^2 + \left(\bar{y}_{91-100} - \hat{\mu}\right)^2\right], \qquad (2)\\
&\hat{\sigma}_\mu^2 = \frac{\hat{\sigma}^2}{72},\\
&t = \frac{\hat{\mu}}{\hat{\sigma}_\mu}.
\end{aligned}$$

Although the sample mean $\hat{\mu}$ in Equation (2) is still an unbiased estimate of μ, the other statistics no longer have the same interpretations as their counterparts in Equation (1); $\hat{\sigma}^2$ is no longer an estimate of σ^2, $\hat{\sigma}_\mu^2$ (or σ_μ) is no longer an estimate of the variability of the sample mean $\hat{\mu}$, and $t = \frac{\hat{\mu}}{\hat{\sigma}_\mu}$ no longer follows the t-distribution. This is because although $\bar{y}_{1-20}$ and $\bar{y}_{91-100}$ still have the same population mean, they no longer have the same variance as the remaining 70 observations:

$$\begin{aligned}
&E\left(\bar{y}_{1-20}\right) = E\left(\bar{y}_{91-100}\right) = E\left(y_i\right) = \mu, \qquad 21 \leq i \leq 90,\\
&Var\left(\bar{y}_{1-20}\right) = \frac{\sigma^2}{20}, \qquad Var\left(\bar{y}_{91-100}\right) = \frac{\sigma^2}{10}, \qquad Var\left(y_i\right) = \sigma^2, \qquad (3)\\
&21 \leq i \leq 90.
\end{aligned}$$

Thus, the reduced 72 observations do not meet the homoscedasticity assumption and all the statistics, except for the sample mean μ, do not have the same meaningful interpretations as their counterparts in Equation (1).

Several methods are available to address heteroscedasticity and its impact on variance estimation and associated p-values. For example, bootstrap and Jackknife resampling methods are commonly implemented in various statistical packages and can be used to provide valid inference in this case. A modern alternative is estimating equations (EE), which do not involve resampling of the observations and provide a more efficient approach. In this example, where the source of heteroscedasticity is known to be caused by averaging some of the observations, weights can be used to 'correct' this special type of heteroscedasticity.

As seen in Equation (3), the variance of the first averaged 20 observation $\bar{y}_{1-20}$ differs from the other 70 observations by a factor of $\frac{1}{20}$ and the variance of the last averaged 10 observation $\bar{y}_{91-100}$ differs from the other 70 observations by a factor of $\frac{1}{10}$. By taking the inverse of these numbers as weights for the two respective averaged observations,

$$w_1 = 20, \qquad w_{72} = 10, \qquad w_i = 1, \qquad 2 \leq i \leq 71,$$

and applying such weights to the sample mean and variance in Equation (2), we obtain a weighted sample mean and sample variance, along with the variance (SE) of the weighted sample mean and t-statistic:

$$
\begin{aligned}
w_g &= \sum_{i=1}^{72} w_i = w_1 + w_2 + \ldots + w_{72} = 20 + 70 + 10 = 100,\\
\hat{\mu} &= \frac{1}{w_g}\left(\sum_{i=1}^{72} w_i y_i\right) = \frac{1}{100}(20\bar{y}_{1-20} + y_{21} + \ldots + y_{90} + 10\bar{y}_{91-100}),\\
\hat{\sigma}^2 &= [\frac{1}{w_g - 1}\sum_{i=1}^{72} w_i (y_i - \hat{\mu})^2]\\
&= \frac{1}{100-1}[20(\bar{y}_{1-20} - \hat{\mu})^2 + (y_{21} - \hat{\mu})^2 + \ldots\\
&\quad + (y_{90} - \hat{\mu})^2 + 10(\bar{y}_{91-100} - \hat{\mu})^2],\\
\hat{\sigma}^2_{\mu} &= \frac{\hat{\sigma}^2}{w_g} = \frac{\hat{\sigma}^2}{w_{100}},\\
t &= \frac{\hat{\mu}}{\hat{\sigma}_{\mu}}.
\end{aligned}
\tag{4}
$$

By comparing Equation (2) and (4), we see that each averaged observation receives more weight than the original observation and the weighted is equal to the number of subjects within the averaged outcome. Also, the mean variance estimates are defined by the sum of the weights, $w_g = 100$, which is the same as the original sample size. Thus, with the weights, the 72 observations in Equation (4) carry the same 'weight' as the original 100 observations. For example, in the sample mean (variance), $\bar{y}_{1-20}$ is weighted 20 times more than each original observation, allowing it to have the same effect as the first 20 observations on the estimated mean (variance). The t-statistic $t = \frac{\hat{\mu}}{\hat{\sigma}_{\mu}}$ in Equation (4) follows the same t-distribution for inference about the population mean.

In Example 1, the heteroscedasticity has a particular form:

$$Var(y_i) = \frac{\sigma_2}{w_i}, \textit{ is a constant}$$

The approach to use weights to correct for heteroscedasticity and construct weighted estimates not only works for inference about the population mean in this example but also for more complex regression models. For example, weighted ordinary least squares (WOLS) uses the same approach to address this type of heteroscedasticity for linear regression. Since the weighted approach here and WOLS in regression setting yield the same point and variance estimate as the maximum likelihood (ML) method (when the outcome is assumed to follow a normal distribution), we will refer to the weighted approach here as the WOLS/ML method throughout the rest of the discussion.

In SAS, the second procedure under consideration, PROC SURVEYMEANS, addresses conceptually different issues arising from complex survey sampling. For the simple random sample as in Example 1, the usual sample mean, sample variance, variance (SE) of the estimate and t-statistic provide valid inference about the population mean. In this case, PROC MEANS and PROC SURVEYMEANS yield identical point estimates and SEs. In most survey studies, more complex sampling strategies are used to more efficiently obtain more reliable estimates. Stratified random sampling is a popular alternative to simple random sampling when sampling heterogeneous populations. Although still yielding the same point estimate, the two SAS PROCs will generally produce different SEs with this type of sampling. We illustrate such difference using data from the Ice Cream Example in the SAS SURVEYMEANS Procedure document, SAS/STAT V.9.2.[1 2]

Example 2

In the Ice Cream study, researchers are interested in how much students in a junior high school spend weekly on ice cream. The junior high school has a total of 4000 students distributed in grades 7, 8, and 9 as follows:

$$
n_h = \begin{cases} 1824 & \text{if } h = 1 \text{ (Grade 7)} \\ 1025 & \text{if } h = 2 \text{ (Grade 8)}, \\ 1151 & \text{if } h = 3 \text{ (Grade 9)} \end{cases} \quad N = N_1 + N_2 + N_3 = 40,000.
$$

where the three different grades represent three strata indexed by h;, N_h denotes the number of students in the hth stratum and N denotes the population size. To address this question, 40 students are selected from the study population using a stratified random sampling; a random sample of 20, 9 and 11 students is taken from the three strata:

$$
n_h = \begin{cases} 20 & \text{if } h = 1 \text{(Grade 7)} \\ 9 & \text{if } h = 2 \text{(Grade 8)}, \\ 11 & \text{if } h = 3 \text{(Grade 9)} \end{cases} \quad n = n_1 + n_2 + n_3 = 40.
$$

The distribution of the three grades in the sample is

$$
\pi_h^{ST} = \frac{n_h}{n} = \begin{cases} \frac{20}{40}\ (50.0\%) & \text{if } h = 1 \text{ (Grade 7)} \\ \frac{9}{40}\ (22.5\%) & \text{if } h = 2 \text{ (Grade 8)} \\ \frac{11}{40}\ (27.5\%) & \text{if } h = 3 \text{ (Grade 9)} \end{cases}, \tag{5}
$$

If the 40 students were sampled under simple random sampling, the distribution of the grades would be:

$$
\pi_h = \frac{N_h}{N} = \begin{cases} \frac{1824}{4000}\ (45.6\%) & \text{if } h = 1 \text{(Grade 7)} \\ \frac{1025}{4000}\ (25.6\%) & \text{if } h = 2 \text{(Grade 8)} \\ \frac{1151}{4000}\ (28.8\%) & \text{if } h = 3 \text{(Grade 9)} \end{cases}, \tag{6}
$$

By comparing Equation (5) and (6), we see that grade 7 is over-represented while the other two grades are under-represented in the study sample. Thus, the usual sample mean of the whole sample will be biased towards Grade 7. To obtain an estimate of mean weekly spending on ice cream for this junior high school, we must use sampling weights to reduce the over-representation of Grade seven and increase the under-representation of other two grades.

To correctly include a sampling weight, it must be the inverse of the sampling probability that a subject is selected from the population. For simple random sampling, this probability is approximated by the sampling fraction, $f = \frac{n}{N} = \frac{40}{4000}$, which is constant. Similarly, the sampling weight for each randomly sampled subject is $w_i = \frac{1}{f} = \frac{4000}{40}$, is also constant, regardless of grade. Under stratified sampling, the sampling fraction is no

longer constant and both this fraction and the sampling weight depend on the strata:

$$f_h = \frac{n_h}{N_h} = \begin{cases} \frac{20}{1824} & \text{if } h = 1 \text{ (Grade 7)} \\ \frac{9}{1025} & \text{if } h = 2 \text{ (Grade 8)} \\ \frac{11}{1151} & \text{if } h = 3 \text{ (Grade 9)} \end{cases}, \quad w_{hi} = \frac{1}{f_h},\ 1 \le i \le n_h,\ 1 \le h \le 3\ . \tag{7}$$

The sampling weight w_{hi} in this case counterbalances the effect of oversampling or undersampling of a stratum, allowing for unbiased estimation of the population mean.

Let y_{hi} denote the spending on ice cream by a student in stratum i in the sample. We can apply the weighted mean in Equation (4) to estimate the mean spending on ice cream by the students in the junior high school μ (the formula looks slightly different because of the added notation for stratification):

$$w_{gg} = \sum_{h=1}^{3}\sum_{i=1}^{n_h} w_{hi} = \sum_{i=1}^{20}\frac{1824}{20} + \sum_{i=1}^{9}\frac{1025}{9} + \sum_{i=1}^{11}\frac{1151}{11} = 4000,$$
$$\hat{\mu} = \frac{1}{w_{gg}}\left(\sum_{h=1}^{3}\sum_{i=1}^{n_h} w_{hi}y_{hi}\right) = \frac{1}{4000}\left(\frac{1824}{20}\sum_{i=1}^{20} y_{1i} + \frac{1025}{9}\sum_{i=1}^{9} y_{2i} + \frac{1151}{11}\sum_{i=1}^{11} y_{3i}\right), \tag{8}$$

However, the variance $\hat{\sigma}_\mu^2$, or SE $\hat{\sigma}_\mu$, calculated according to Equation (4), no longer estimates the variability of the estimate $\hat{\mu}$ of the mean. Unlike the weights in Example 1, sampling weights in this example are used to address different sampling probabilities between the strata. Even under homoscedasticity, that is, a common variance of y_{hi} across all three strata, we still need to use the weighted mean in Equation (8) to estimate the correct population mean.

Note that if a sample of n is taken from a population of size N, the probability for the first subject sampled is $\frac{1}{N}$, the probability for the second sampled is $\frac{1}{N-1}$ and so on. So, the sampling probabilities for the sampled subjects are not constant and given by:

$$\frac{1}{N}, \frac{1}{N-1}, \frac{1}{N-2}, \cdots \frac{1}{N-n}.$$

In most survey studies, n is much smaller than N, so $N - n$ is well approximated by N. Thus, for all practical purposes, the sampling probability for a random sample of n subjects is the sampling fraction, $f = \frac{n}{N}$. Also, unlike the weights applied in Example 1, sampling weights in Example 2 sum to the total population size. However, a sampling weight is a unit-free quantity and can be multiplied by any number without affecting the statistics. For example, by dividing the weights in Equation (8) by N=4000, they sum to one and the estimate:

$$\hat{\mu} = \left(\sum_{h=1}^{3}\sum_{i=1}^{n_h}\frac{w_{hi}}{N}y_{hi}\right) = \frac{1824}{4000\times 20}\sum_{i=1}^{20} y_{1i} + \frac{1025}{4000\times 9}\sum_{i=1}^{9} y_{2i} + \frac{1151}{4000\times 11}\sum_{i=1}^{11} y_{3i},$$

which is the same as the one in Equation (8).

Shown in table 1 are the weighted means and SEs from the PROC MEANS and PROC SURVEYMEANS for the Ice Cream data. The weighted means are the same, but the SEs are quite different from the two PROCs.

Table 1 Comparison of PROC MEANS and PROC SURVEYMEANS for Ice Cream data in Example 1

PROC MEANS	PROC SURVEYMEANS
Mean: 9.141	Mean: 9.141
SE: 0.858	SE: 0.532

The difference in SEs between the two SAS PROCs above is the result of a different statistical approach used to compute the SE in PROC SURVEYMEANS. Unlike the WOLS/ML variance estimate in the PROC MEANS which is specific to weights selected to address heteroscedasticity, the variance estimate from PROC SURVEYMEANS is derived by estimating equations, or Taylor series expansion, which is valid for any types of weights, including weights computed in PROC MEANS, sampling weights for survey studies, non-response weights and combinations of such weights.

For the stratified sampling in Example 2, the estimating equation variance (SE) of the weighted mean is given by:

$$\hat{\sigma}_\mu^2 = \frac{1}{w_{gg}}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\left[\sum_{i=1}^{n_h} w_{hi}^2\left(y_{hi}-\hat{\mu}\right)^2 - \frac{1}{n_h}\left(\sum_{i=1}^{n_h} w_{hi}\left(y_{hi}-\hat{\mu}\right)\right)^2\right]. \tag{9}$$

If applying the WOLS (ML) variance estimate in Equation (3), the variance of the weighted mean is:

$$\hat{\sigma}_\mu^2 = \frac{1}{w_{gg}}\left[\frac{1}{H}\sum_{h=1}^{H}\frac{1}{n_h-1}\sum_{i=1}^{n_h} w_{hi}\left(y_{hi}-\hat{\mu}\right)^2\right]. \tag{10}$$

The different variance estimates in Equations (9) and (10) can yield quite different SEs as shown by the Ice Cream data in Example 2.

Thus, when analysing survey data involving sampling weights, PROC SURVEYMEANS must be used to provide valid variance (SE) estimates. As noted above, the estimating equation approach also provides valid inference for all other types of weights. The next example shows that this variance estimate also applies when weights are used to address heteroscedasticity.

IMPLICATIONS OF USING THE WRONG PROC WITH A WEIGHT VERSUS A SAMPLING WEIGHT: A MONTE CARLO SIMULATION

In the case of homoscedasticity weights, the choice of PROC MEANS or PROC SURVEYMEANS and associated variance estimates is facilitated as either will give valid and consistent results. We must underscore that this is true only when the weight used is in fact a homoscedasticity (heteroskedasticiy-correction) weight and not sampling weight.

Example 3

In this example, we perform Monte Carlo simulations to show that the estimating equation variance estimate is also valid when used to deal with heteroscedasticity in the data.

To simulate a sample with heteroscedasticity, we consider a normal distribution consists of five

subpopulations: all with the same (population) mean μ=1, but different variances:

$$y_{hi} \sim N\left(\mu, \frac{\sigma^2}{w_{hi}}\right), \qquad \mu = \sigma^2 = 1, \qquad w_{hi} = \frac{1}{h+2}, \qquad 1 \le h \le 5. \tag{11}$$

When sampled from this five-component mixture, y_{hi}, the variance of the observation varies depending on the subpopulation sampled:

$$\mathrm{Var}(y_{hi}) = \frac{\sigma^2}{w_{hi}} = \frac{\sigma^2}{\frac{1}{h+2}} = (h+2)\ \sigma^2 \begin{cases} 3\sigma^2 & \text{if } h=1 \\ 4\sigma^2 & \text{if } h=2 \\ 5\sigma^2 & \text{if } h=3 \\ 6\sigma^2 & \text{if } h=4 \\ 7\sigma^2 & \text{if } h=5 \end{cases}$$

By using the weights, $w_{hi} = \frac{1}{h+2}$, we can use the weighted mean, WOLS/ML variance (SE) of the weighted mean and t-statistic in Equation (2) and (4) used in PROC MEANS to make inference about the mean μ. Alternatively, we can also apply the estimating equation (EE) variance estimate in Equation (9) to compute variance (SE) of the weighted mean using PROC SURVEYMEANS. Although the two variance estimates look quite different, they are both consistent estimates of the variance of the weighted mean $\hat{\mu}$.

To demonstrate this using Monte Carlo simulations, we perform the following steps:

a. Simulate a sample of 25 000 y_{hi}'s from Equation (11), with 5000 y_{hi}'s from each subpopulation;
b. Compute the estimate $\hat{\mu}$ and two variance estimates of $\hat{\mu}$ according to Equation (4) and (9):

$$\text{Point Estimate}: \quad \hat{\mu} = \frac{1}{w_{gg}} \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi} y_{hi} = \frac{1}{w_{gg}} \sum_{h=1}^{5} \sum_{i=1}^{500} w_{hi}\, y_{hi},$$
$$w_{gg} = \left(\tfrac{1}{3} + \tfrac{1}{4} + \tfrac{1}{5} + \tfrac{1}{6} + \tfrac{1}{7}\right) \times 500 = 546.43,$$

$$\textit{WOLS/ML Variance of } \hat{\mu}: \ \hat{\sigma}^2_{WOLS/ML} = \frac{1}{w_{gg}} \left[\frac{1}{H} \sum_{h=1}^{H} \frac{1}{n_h - 1} \sum_{i=1}^{n_h} w_{hi} \left(y_{hi} - \hat{\mu}\right)^2\right]$$
$$= \frac{1}{546.43} \left[\frac{1}{5} \sum_{h=1}^{5} \frac{1}{500-1} \sum_{i=1}^{500} w_{hi} \left(y_{hi} - \hat{\mu}\right)^2\right],$$

$$\text{EE Variance of}: \ \hat{\sigma}^2_{EE} = \frac{1}{w^2_{gg}} \sum_{h=1}^{H} \frac{n_h}{n_h - 1} \left[\sum_{i=1}^{n_h} w_{hi}(y_{hi} - \hat{\mu})^2 - \frac{1}{n_h} \sum_{i=1}^{n_h} w_{hi}(y_{hi} - \hat{\mu}))^2\right] \tag{12}$$
$$= \frac{1}{(546.43)^2} \sum_{h=1}^{5} \frac{500}{500-1} \left[\sum_{i=1}^{500} w^2_{hi}(y_{hi} - \hat{\mu})^2 - \frac{1}{500} \left(\sum_{i=1}^{n_k} w_{hi}(y_{hi} - \hat{\mu})\right)^2\right].$$

(c) Repeat Step (a) and (b) M=2000 times;

(d) For each Monte Carlo iteration, let $\hat{\mu}^{(m)}$ denote the estimate and $\hat{\sigma}^{2(m)}_{WOLS/ML}$ and $\hat{\sigma}^{2(m)}_{EE}$ denote the two variance estimates;

(e) Compute the Monte Carlo mean $\hat{\mu}_{emp}$ and empirical variance $\hat{\sigma}^2_{emp}$ of the estimate $\hat{\mu}_{emp}$:

$$\hat{\mu}_{emp} = \frac{1}{M} \sum_{m=1}^{M} \hat{\mu}^{(m)}, \qquad \hat{\sigma}^2_{emp} = \frac{1}{M} \sum_{m=1}^{M} \left(\hat{\mu}^{(m)} - \hat{\mu}_{emp}\right)^2. \tag{13}$$

(f) Compute the two variance estimates averaged over the 2000 Monte Carlo simulations:

Table 2 Comparison of PROC MEANS and PROC SURVEYMEANS for simulation study

PROC MEANS	SAS PROC SURVEYMEANS
Mean: 1.000	Mean: 1.000
SE (WOLS/ML): 0.000183 SE (empirical): 0.000184	SE (EE): 0.000183

EE, estimating equation; ML, maximum likelihood; WOLS, weighted ordinary least squares.

$$\hat{\sigma}^2_{WOLS/ML} = \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}^{2(m)}_{WOLS/ML} = \frac{1}{2000} \sum_{m=1}^{2000} \hat{\sigma}^{2(m)}_{WOLS/ML},$$
$$\hat{\sigma}^2_{EE} = \frac{1}{M} \sum_{m=1}^{M} \hat{\sigma}^{2(m)}_{EE} = \frac{1}{2000} \sum_{m=1}^{2000} \hat{\sigma}^{2(m)}_{EE}. \tag{14}$$

The Monte Carlo mean and variance above provide a benchmark to assess and compare performances of estimates. The Monte Carlo sample variance is also known as the empirical variance of the estimate, since it measures the variability of the estimate and is a consistent estimate of the variance of the asymptotic distribution of the estimate.

If the method for estimating the mean is correct, the Monte Carlo mean $\hat{\mu}_{emp}$ should be close to the population mean μ=1. Likewise, if a variance estimate is consistent, its Monte Carlo average in Equation (13) will be close to the empirical version $\hat{\sigma}^2_{emp}$ and vice versa.

Shown in table 2 are the weighted mean and SEs from the WOLS/ML, EE and empirical variance estimates. The WOLS/ML and EE SEs are virtually identical (difference is 2.67×10^{-8} 2.67×10^{-8}) and both are extremely close to the empirical SE.

Therefore, for non-sampling weights such as weights selected to address heteroscedasticity, the EE variance estimate still describes the sampling variability of the estimate, as illustrated by the simulation study above. The EE variance estimate is more general, as it also provides valid inference for more complex weights such as those used for sampling and non-response bias, while WOLS/ML based variance formulas cannot be applied to all types of weights.

We would like to point out that the EE can also be used to address heteroscedasticity when a correction weight is not available. In many studies, the cause of heteroscedasticity is unknown and weights cannot be computed. In this case, the WOLS/ML approach no longer applies. But, even without a known heteroscedasticity weight, the EE still provides valid variance estimates. For example, when applied to the simulated data in Example 3 without weights, the estimated population mean is 0.9996, which is quite close to 1. The Monte Carlo average of the EE SE, 0.0002, is also quite close to the empirical error, 0.000185. Both the EE and empirical SEs in this case are a bit larger than their weighted counterparts, which is consistent with the property that the weighted mean is the BLUE, that is, the estimate

with the smallest standard error among all estimates that are a linear combination of the observations.

In the next Example, we use simulated data to show thatwhen using survey weights, *the WOLS/ML variance estimate can have severe bias* and the EE variance estimate must be used to provide valid inference about the population mean.

Example 4

We use the Ice Cream Example as the setting to simulate the outcome (spending) for each student. First, we compute sample means and sample variances for each of the three strata (grade). Next, we construct the population distribution as a three-strata normal mixture using these sample means as the population means for the three strata and an averaged sample variance as the common population variance for all the strata. We then estimate the population mean using a weighted mean and compare the two variance estimates.

Specifically, let h denote the index strata and y_{hi} denote spending of ith student sampled from the hth stratum. Let μ_h denote the population mean of the hth stratum:

$$\mu_h = \begin{cases} \mu_1 = 5.0 & \text{if } h = 1 \text{ for Grade } 7 \\ \mu_2 = 15.4 & \text{if } h = 2 \text{ for Grade } 8 \\ \mu_3 = 10.1 & \text{if } h = 3 \text{ for Grade } 9 \end{cases},$$

where 5, 15.4 and 10.1 are the sample means of the corresponding strata in the Ice Cream Example. Let $\sigma^2 = 28.9$ denote the common variance of y_{hi} across all strata (average of three strata variances). Let N_h denote the population size and n_h denote the sample size of the h^{th} stratum. We assume that y_{hi} follows a three-component mixture with the mean μ_h and variance σ^2:

$$y_{hi} \sim N\left(\mu_h, \sigma^2\right), \quad 1 \le i \le N_h, \quad 1 \le h \le 3,$$

$$N_h = \begin{cases} 1824 & \text{if } h = 1 \text{ for Grade } 7 \\ 1025 & \text{if } h = 2 \text{ for Grade } 8 \\ 1151 & \text{if } h = 3 \text{ for Grade } 9 \end{cases},$$

As in the Ice Cream Example, a sample of n=40 is taken from the population with strata sample size following the following distribution:

$$n_h = \begin{cases} 20 & \text{if } h = 1 \text{ for Grade } 7 \\ 9 & \text{if } h = 2 \text{ for Grade } 8 \\ 11 & \text{if } h = 3 \text{ for Grade } 9 \end{cases},$$

The total population size N and overall population mean μ are given by:

$$N = N_1 + N_2 + N_3 = 1824 + 1025 + 1151 = 4000,$$

$$\mu = \pi_1\mu_1 + \pi_2\mu_2 + \pi_3\mu_3$$

$$= \frac{1824}{4000} \times 5 + \frac{1025}{4000} \times 15.4 + \frac{1151}{4000} \times 10.1 = 9.\ 1325,$$

where π_h is the proportion of the hth stratum size to the total population size. Under stratified random sampling, the sampling weights are used to estimate the overall population mean μ:

$$w_{hi} = \begin{cases} \frac{1824}{20} & \text{if } h = 1 \text{ for Grade } 7 \\ \frac{1025}{9} & \text{if } h = 2 \text{ for Grade } 8 \\ \frac{1151}{11} & \text{if } h = 3 \text{ for Grade } 9 \end{cases},$$

To reduce sampling variability in Monte Carlo estimates, we set Monte Carlo replication size to M=10 000. For each Monte Carlo iteration, let $\hat{\mu}^{(m)}$ denote the estimate and $\hat{\sigma}^{2(m)}_{WOLS/ML}$ and $\hat{\sigma}^{2(m)}_{EE}$ denote the WOLS/ML and EE variance estimates from the mth simulated data. We compute the Monte Carlo sample mean $\hat{\mu}^{(emp)}$,, empirical variance $\hat{\sigma}^2_{emp}$ and averaged variance estimates $\hat{\sigma}^2_{WOLS/ML}$ and $\hat{\sigma}^2_{EE}$ from the two methods the same way as in (13) and (14).

Shown in table 3 are the Monte Carlo mean, (empirical) SE and SEs from the two variance estimates along with the empirical SE. As expected, the Monte Carlo mean is nearly identical to the population mean μ=9.1325 and the averaged EE SE $\hat{\sigma}_{EE}$ is identical to the empirical version $\hat{\sigma}_{emp}$.

Table 3 Monte Carlo mean, empirical, WOLS/ML and EE SE

WOLS/ML	EE
Mean 9.14	9.14
SE (WOLS/ML): 1.36 SE (empirical): 0.726	SE (EE): 0.726

EE, estimating equation; ML, maximum likelihood; WOLS, weighted ordinary least squares.

DISCUSSION

In this paper, we focused on sampling and homoscedasticity weights, discussed the conceptual difference between the two and illustrated the implications of the conceptual difference in SEs of estimated population means though analytic expressions and Monte Carlo simulations. We have demonstrated that homoscedasticity weights have very specific applications. Our experiences with SAS and other popular packages indicate that if weights are available as an option in a procedure such as SAS PROC MEANS, they are typically of the homoscedasticity type. Such procedures should not be used for any other types of weights such as sampling weights. Sampling weights must only be used in survey specific procedures such as SAS PROC SURVEYMEANS, as PROC MEANS will not compute the correct SE and will show substantial bias even in large samples. In contrast, procedures such as SAS PROC SURVEYMEANS are more general and will compute the correct SE when using both sampling and homoscedasticity weights. In the case of heteroscedasticity, estimating equation methods for calculating the SE can even compute the correct variance estimate if a researcher does not have access to a known homoscedasticity weight, correcting for potential distortions in the SE

that can result from this violation. We recommend that researchers identify the type of weight they are using and understand the implications of using the weight within common analytic programs such as SAS, as incorrect application of weights can have important consequences for research analyses.

Contributors SR, XN, VS, MX and XMT had extensive discussions of the statistical issues with different types of weights and their implementations in some popular statistical packages and worked together to structure this report. TL, YL and XN worked together to find the formulas for the estimate and SEs of the estimate of the population mean as implemented in SAS and developed the computer codes to perform data simulations and inference for the examples.

Funding The project described was partially supported by the National Institutes of Health, Grant UL1TR001442 of CTSA funding beginning 13 August 2015 and beyond and by the Navy Bureau of Medicine and Surgery under work unit no. N1240. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Disclaimer I am an employee of the US Government. This work was prepared as part of my official duties. Title 17, U.S.C. §105 provides that copyright protection under this title is not available for any work of the US Government. Title 17, U.S.C. §101 defines a US Government work as work prepared by a military service member or employee of the US Government as part of that person's official duties. This work was supported by the Navy Bureau of Medicine and Surgery under work unit no. N1240. The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, nor the US Government.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

Data sharing statement No additional data are available.

REFERENCES

1 SAS/STAT® 9.2 User's Guide. SAS Institute Inc. *SAS/STAT® 9.2 User's Guide*. Cary, NC: SAS Institute Inc, 2008.
2 Tang W, He H, Tu XM. *Applied categorical and count data analysis*. Florida: Chapman & Hall/CRC, 2012.

Sabrina Richardson received her PhD in developmental psychology from the University of California, Riverside, and is currently a Leidos research psychologist at the Naval Health Research Center working on the Millennium Cohort Family Study, a longitudinal study investigating the well-being of service members and their families. Her research broadly seeks to understand the multifaceted processes of adaptation supporting better than expected outcomes when encountering risk (i.e., resilience), particularly focused on child and family development. She has authored works on children's resilience, sibling relationships, child maltreatment, transition aged foster youth, and military family readiness. She is also interested in quantitative methods such as structural equation modeling, moderation/mediation analyses, and survey methodology.