Scientific Research Publishing

Open Journal of Statistics, 2026, 16(1), 38-46
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Asymptotic Properties of the Size-Weighted Mean

Eugene P. Canjels

Division of Economic and Risk Analysis, U.S. Securities and Exchange Commission, Washington, D.C., USA
Email: canjelse@sec.gov

How to cite this paper: Canjels, E.P. (2026) Asymptotic Properties of the Size-Weighted Mean *Open Journal of Statistics*, **16**, 38-46.
https://doi.org/10.4236/ojs.2026.161003

Received: January 1, 2026
Accepted: February 21, 2026
Published: February 24, 2026

Abstract

Applied researchers often compute weighted means where the weights reflect the relative size of each observation. Despite their common use, the statistical properties of these size-weighted means appear to be poorly understood, as the existing statistical literature and most software applications typically address three other types of weights—survey, precision, and frequency weights. In this paper, I demonstrate that the size-weighted mean is asymptotically normal, derive an estimator for its variance, and discuss several practical considerations relevant to its application.

Keywords

Size-Weighted Mean, Weighted Mean, Asymptotic Distribution, Hypothesis Testing

1. Introduction

Applied researchers are often interested in an outcome per unit. Examples include crop yield (e.g., tons per hectare), the concentration of a chemical in a solution (e.g., moles per liter), student performance (e.g., the percentage of students reading at grade level), or investment returns (profit per dollar invested). When the observations collected vary in size, that is, in the number of units they represent, the statistic of interest is often an average outcome per unit over the entire sample. This quantity can be computed as a weighted mean, where the weights reflect the relative size of each observation contributing to the total. I refer to this quantity as the *size-weighted mean*.

To illustrate with an example from finance, suppose that one wants to compare the execution cost of shares purchased for two different investment portfolios. The execution cost of a trade is defined as the difference between the prevailing market price at the time the order is placed and the subsequent execution price. These

two prices may differ due to factors such as broker fees, bid-ask spreads, and execution delays, among others. Execution costs can have substantial implications for the long-term performance of an investment portfolio, and minimizing them is a key objective for traders and institutional investors. The execution cost per dollar for an observation i is defined as

$$cost_i = \frac{p_i^e - p_i^o}{p_i^o}$$

where p_i^e is the execution price and p_i^o is the market price at the time the order is placed. To compute the investment-weighted average execution cost over n trades, one calculates

$$\widetilde{cost} = \sum w_i \times cost_i$$

where w_i represents the proportion of the dollar value of trade i relative to the total value of all trades, and the summation runs from $i=1$ to n. To test if two portfolios face the same execution costs, it is necessary to understand the statistical properties of the size-weighted mean.

Introducing notation in the context of crop yields, suppose each observation represents an individual farm. For each farm i, the researcher observes total output, denoted by z_i, and total acreage, denoted by u_i (for "units"), from which crop yield can be computed as $r_i = z_i / u_i$ (for "rate"). The overall average yield can be calculated as total output across all farms divided by total acreage across all farms, or equivalently, as the acreage-weighted average of the individual yields. That is, the size-weighted mean $\tilde{r}$ is given by

$$\tilde{r} = \frac{\sum z_i}{\sum u_i} = \frac{\sum u_i r_i}{\sum u_i} = \sum w_i r_i \tag{1}$$

where $w_i = u_i / \sum u_i$.

The weighted mean has been widely discussed in the statistical literature. However, most existing work focuses on weights that are either precision weights (also known as analytical or inverse-variance weights) or survey weights (including the closely related sampling and probability weights). Precision weights improve the estimate of a regular mean by using information on the variance of each observation. Survey weights may be used for survey data to address unequal selection probabilities, non-responses, or other sampling issues. In addition, statistical software introduces a third category, frequency weights (also known as case weights), which are used when individual observations have been aggregated into groups and the weight represents the number of observations within each group.

However, size weights differ fundamentally from other types of weights and must be analyzed accordingly. The absence of readily accessible literature describing the statistical properties of the size-weighted mean—combined with the lack of clear guidance on the interpretation of weights in statistical software—has led to considerable confusion among applied researchers, as evidenced by questions and discussions on online forums.[1]

[1]See also e.g. https://notstatschat.rbind.io/2020/08/04/weights-in-statistics.

To illustrate the issue in software applications, consider the PROC MEANS procedure in SAS (Version 9.4). This procedure includes a WEIGHT option, but the documentation provides little guidance, stating only that the weight "specifies a numeric variable whose values weight the values of the analysis variables." In fact, the calculation of the standard error of the weighted mean in SAS assumes that the weights are precision weights, and the formula is incorrect when applied to size weights. Consequently, applied researchers who use PROC MEANS with the WEIGHT option may draw misleading conclusions about the precision of their estimates when calculating size-weighted means.

For practitioners, the key result of this paper is presented in Theorem 2. In brief, the size-weighted mean is asymptotically normal, with a standard error that can be approximated by

$$s_{\tilde{r}}^2 = \sum w_i^2 \left(r_i - \tilde{r}\right)^2$$

Surprisingly, this result and its derivation do not appear to have been presented in the literature in a straightforward manner. The most relevant discussions can be found in the sampling literature—for example, [1] [ch. 2.12]—but they are often presented in a rather complicated form. Furthermore, many empirical problems, such as comparing execution costs across two investment portfolios, are not naturally framed as sampling problems. As a result, applied researchers are unlikely to consult sampling textbooks in search of answers.

In addition to the main result, I discuss the implementation of weighted means in statistical software. To date, I am unaware of any software that explicitly addresses size-weighted means or provides procedures to estimate their standard errors. The paper concludes by demonstrating that valid inference for size-weighted means can be obtained by estimating a regression with precision weights while using heteroscedasticity-robust standard errors.

2. Asymptotic Theory

When researchers calculate standard errors for equally weighted means, they typically justify the analysis either by assuming a random sampling scheme in which each observation is independently drawn with equal probability from a known population much larger than the sample, or by invoking an abstract super-population framework. The analysis in this paper adopts a similar assumption. Specifically, I assume that the n pairs $\left(z_1,u_1\right),\cdots,\left(z_n,u_n\right)$ are independently and identically distributed (i.i.d.) with a bivariate probability distribution $g\left(z,u\right)$. I further assume that the number of units u_i is strictly positive, so that $r_i = z_i/u_i$ is always well-defined. The population means of z_i, u_i, and r_i are denoted μ_z, μ_u, and μ_r, respectively, with corresponding population variances σ_z^2, σ_u^2, and σ_r^2, and covariances σ_{zu}, σ_{zr}, and σ_{ur}.[2] Equation (1) above char-

[2]In certain settings, the researcher may encounter observations with $u_i = 0$. In these settings, Theorem 1 can still be used as it depends only on the assumption that $\mu_u \neq 0$. Other results presented below depend on r_i being well-defined for all i.

acterizes the size-weighted mean as either the ratio of total observed output to the total number of observed units or as a linear combination of observed rates. It is also useful to express it as the ratio of sample means:

$$\tilde{r} = \frac{n^{-1}\sum z_i}{n^{-1}\sum u_i} = \frac{\overline{z}}{\overline{u}} \tag{2}$$

From Equation (2) it follows that, under mild regularity conditions, the weighted mean $\tilde{r}$ is a consistent estimator of $\mu_{\tilde{r}}$, where $\mu_{\tilde{r}} = \mu_z / \mu_u$.

The size-weighted and equal-weighted means capture different parameters, both of which may be of interest to the researcher. In the farm example described above, the equal-weighted mean represents the crop yield of a typical farm, whereas the size-weighted mean represents the crop yield of a typical acre. When farm size and yield are correlated, these two measures do not necessarily coincide. A straightforward derivation shows that the difference between them is given by the covariance between u_i and r_i divided by the expectation of u_i

$$\mu_{\tilde{r}} - \mu_r = \frac{\mu_z - \mu_u \mu_r}{\mu_u} = \frac{E(u_i r_i) - \mu_u \mu_r}{\mu_u} = \frac{\sigma_{ur}}{\mu_u}$$

Thus, the size-weighted and equal-weighted means coincide only when the covariance between u_i and r_i is zero. At the sample level, the difference between $\tilde{r}$ and the sample mean $\overline{r}$ is given by the sample covariance between the scaled weights $(n \cdot \mathbf{w})$ and the observed rates $\mathbf{r}$:

$$cov(n \cdot \mathbf{w}, \mathbf{r}) = \frac{1}{n}\sum (nw_i - n\overline{w})(r_i - \overline{r}) = \frac{1}{n}\sum (nw_i r_i - n\overline{wr}) = \tilde{r} - \overline{r}$$

Equation (2) provides the foundation for the asymptotic theory used to derive the distribution of the size-weighted mean. By the multivariate central limit theorem,

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{d} N(0, \Omega)$$

where $\overline{X} = [\overline{z} \quad \overline{u}]'$, $\mu = [\mu_z \quad \mu_u]'$ and Ω is the variance-covariance matrix with elements σ_z^2, σ_u^2, and σ_{zu}. Applying the delta method then allows a straightforward derivation of the following theorem:

Theorem 1 *The asymptotic distribution of* $\tilde{r}$ *is given by*:

$$\sqrt{n}(\tilde{r} - \mu_{\tilde{r}}) \sim N(0, \phi^2)$$

where

$$\phi^2 = \left(\frac{1}{\mu_u}\right)^2 \left(\sigma_z^2 + \mu_{\tilde{r}}^2 \sigma_u^2 - 2\mu_{\tilde{r}} \sigma_{zu}\right)$$

Proof. The first-order Taylor expansion of the function $\tilde{r} = \overline{z}/\overline{u}$ is given by

$$\tilde{r} \approx \mu_{\tilde{r}} + \frac{1}{\mu_u}(\overline{z} - \mu_z) - \frac{\mu_z}{\mu_u^2}(\overline{u} - \mu_u) = \mu_{\tilde{r}} + \frac{1}{\mu_u}\left\{(\overline{z} - \mu_z) - \mu_{\tilde{r}}(\overline{u} - \mu_u)\right\}$$

The corresponding formula for the variance of $\tilde{r}$ follows directly.

To estimate the variance of $\tilde{r}$, the population parameters can be replaced with

their consistent sample analogs, yielding a result that can be directly applied by researchers.

Theorem 2 *Under the assumptions above*

$$\tilde{r} \approx N\left(\mu_{\tilde{r}}, s_{\tilde{r}}^2\right)$$

where

$$s_{\tilde{r}}^2 = \sum w_i^2 \left(r_i - \tilde{r}\right)^2$$

Proof. The sample analog of ϕ^2 is given by:

$$f^2 = \left(\frac{1}{\bar{u}}\right)^2 \left(s_z^2 + \tilde{r}^2 s_u^2 - 2\tilde{r}s_{zu}\right)$$

By expanding and rearranging terms, the expression inside the parentheses simplifies as follows:

$$\begin{aligned}
&s_z^2 + \tilde{r}^2 s_u^2 - 2\tilde{r}s_{zu} \\
&= \frac{1}{n}\sum z_i^2 - \bar{z}^2 + \frac{1}{n}\sum \tilde{r}^2 u_i^2 - \tilde{r}^2\bar{u}^2 - \frac{2}{n}\sum \tilde{r}z_i u_i + 2\tilde{r}\bar{z}\bar{u} \\
&= \frac{1}{n}\sum u_i^2 r_i^2 + \frac{1}{n}\sum u_i^2 \tilde{r}^2 - \frac{2}{n}\sum u_i^2 r_i \tilde{r} - \bar{z}^2 - \left(\frac{\bar{z}}{\bar{u}}\right)^2 \bar{u}^2 + 2\left(\frac{\bar{z}}{\bar{u}}\right)\bar{z}\bar{u} \\
&= \frac{1}{n}\sum u_i^2 \left(r_i - \tilde{r}\right)^2
\end{aligned}$$

such that

$$\frac{f^2}{n} = \frac{1}{n}\left(\frac{1}{\bar{u}}\right)^2 \frac{1}{n}\sum u_i^2 \left(r_i - \tilde{r}\right)^2 = \sum w_i^2 \left(r_i - \tilde{r}\right)^2$$

When applying these results, researchers should be aware of several important considerations. First, the size-weighted mean typically requires a larger sample size to achieve approximate normality than the equal-weighted mean. For equal-weighted means a sample size of 30 is often considered sufficient for distributions that are not excessively skewed or fat-tailed. For size-weighted means, this means that if the sample size and distribution of units is such that each observation has a weight smaller than 1/30 the approximation is at least as good as for the equal-weighted mean with a sample size of 30.

Furthermore, the derivation of Theorems 1 and 2 relies on a first-order Taylor expansion. This approximation may be poor if the sample mean $\bar{u}$ can approach zero with non-negligible probability. Kish [2] [p. 208] suggests, as a rule of thumb, that a coefficient of variation of $\bar{u}$ below 0.1 is generally sufficient for the approximation to be reliable in empirical applications.

Finally, it is important to note that researchers can make significant errors if they use the formula for precision-weighted means when the size-weighted mean is appropriate. The variance of the precision-weighted mean can be estimated by

$$s_{prec}^2 = \frac{1}{n}\sum w_i \left(r_i - \tilde{r}\right)^2 \tag{3}$$

Except for using Bessel's correction, which replaces n with $n-1$, Equation (3)

is the formula implemented in SAS and Stata. (See also Equation (5) and the corresponding text below.)

To compare the standard errors of the size-weighted and precision-weighted means, it is convenient to write the sample variance of the size-weighted mean as

$$s_{\tilde{r}}^2 = \sum w_i \times \left(w_i \left(r_i - \tilde{r} \right)^2 \right)$$

and the sample variance of the precision-weighted means as

$$s_{prec}^2 = \sum \left(\frac{1}{n} \right) \times \left(w_i \left(r_i - \tilde{r} \right)^2 \right)$$

This highlights that the term w_i in the first equation is replaced by $1/n$ in the second. The estimated standard error for the size-weighted mean is either larger or smaller than that for the precision-weighted mean, depending on whether larger values of w_i tend to be associated with larger or smaller values of $w_i \left(r_i - \tilde{r} \right)^2$. If r_i is homoscedastic and thus not systematically related to w_i, the association tends to be positive, and using the precision-weighted formula will generally underestimate the standard error of the size-weighted mean.

3. Statistical Software

I reviewed several widely used statistical software packages to determine if they include procedures for calculating weighted means and their associated standard errors. Many packages provide built-in procedures for calculating weighted means and for that calculation it does not matter whether the weights are precision, survey, size, or frequency weights. When it comes to estimating the standard error, results vary widely across the packages. Notably, my review identified no package that computes the standard error in accordance with Theorem 2.

Among popular software packages, MATLAB's *mean* and R's *weighted.mean* core functions can calculate weighted means but there appears to be no core function that calculates its standard error. SPSS appears to support only frequency weights. Python NumPy's *average* function can calculate a weighted mean but, like Matlab and R, not its standard error. Python's popular statsmodels package can calculate a weighted mean as well as its standard error, but only for frequency weights.

SAS, in its PROC MEANS command, calculates standard errors for weighted means but only for precision weights. This is particularly concerning because SAS does not explicitly inform users of what type of weights the WEIGHT option assumes. The WEIGHT option doesn't even always reflect the same type of weight. For instance, in PROC SURVEYMEANS the WEIGHT option reflects survey weights instead of precision weights.

Stata provides four types of weights: precision weights (which it calls analytical weights), sampling weights, frequency weights, and weights that it calls "importance" weights. Stata describes its importance weights as follows: "iweights, or importance weights, indicate the relative 'importance' of the observation. They have no formal statistical definition; this is a catch-all category. Any command

that supports iweights will define how they are treated. They are usually intended for use by programmers who want to produce a certain computation."

Thus, Stata's notion of importance weights does not necessarily match the definition of size weights used in this paper. Moreover, Stata's *collapse* command will compute a standard error for the weighted mean only when using precision weights. It is only slightly more reassuring than SAS—which provides no warning at all—that Stata alerts the user when it assumes analytical (*i.e.*, precision) weights if the intended type of weight is not explicitly specified.

4. A Regression Approach

Notwithstanding the lack of canned commands, in some statistical software applications it is possible to obtain standard errors for the size-weighted mean using a regression framework. Specifically, if the application can run a precision-weighted regression while estimating Eicker-Huber-White heteroscedasticity-robust standard errors, then the resulting estimates are identical to those given by Theorem 2, except possibly for small differences due to finite-sample corrections.

To understand why this works, first note that to estimate a regular equal-weighted mean, one can run the regression:

$$r_i = \alpha + \epsilon_i$$

The estimate $\hat{\alpha}$ is equal to $\bar{r}$, and the estimated variance of ϵ_i is $s_\epsilon^2 = n^{-1}\sum\left(r_i - \bar{r}\right)^2$, ignoring Bessel's correction (*i.e.*, replacing n with $n-1$). Consequently, the estimated variance of $\hat{\alpha}$ is $n^{-2}\sum\left(r_i - \bar{r}\right)^2$.

Because heteroscedasticity is only a concern when the explanatory variables are correlated with the variance of the error term, the heteroscedasticity-robust standard errors for $\hat{\alpha}$ are identical to the usual standard errors:

$$\left(X'X\right)^{-1}\left(X'diag\left(\hat{\epsilon}_1^2,\cdots,\hat{\epsilon}_n^2\right)X\right)\left(X'X\right)^{-1} = \frac{1}{n^2}\sum\left(r_i - \bar{r}\right)^2$$

Suppose now that ϵ_i is heteroscedastic where $\sigma_{\epsilon_i}^2$ is equal to $u_i^{-1}\sigma^2$ for some value of σ^2. In that case, precision-weighted means can be estimated using standard OLS as follows:

$$\sqrt{u_i}r_i = \alpha\sqrt{u_i} + e_i \tag{4}$$

because $e_i = \sqrt{u_i}\epsilon_i$ is homoscedastic. Then

$$\hat{\alpha} = \frac{\sum u_i r_i}{\sum u_i} = \tilde{r}$$

with an estimated variance (ignoring Bessel's correction) of

$$\hat{var}\left(\hat{\alpha}\right) = \left(\sum u_i\right)^{-1} s_e^2 = \frac{1}{n}\sum w_i\left(r_i - \tilde{r}\right)^2 \tag{5}$$

where

$$s_e^2 = \frac{1}{n}\sum \hat{e}_i^2 = \frac{1}{n}\sum u_i\left(r_i - \tilde{r}\right)^2$$

Now assume that we are interested in size weights instead of precision weights. Equation (4) can still be estimated using OLS to obtain a size-weighted mean. However, since e_i is now heteroscedastic, it is appropriate to use heteroscedasticity-robust standard errors. The equation for heteroscedascity-robust standard errors reduces to the formula in Theorem 2:

$$\begin{aligned}\hat{v}ar(\hat{\alpha}) &= (X'X)^{-1}\left(X'diag\left(\hat{e}_1^2,\cdots,\hat{e}_n^2\right)X\right)(X'X)^{-1} \\ &= \left(\sum u_i\right)^{-1}\sum u_i^2\left(r_i-\tilde{r}\right)^2\left(\sum u_i\right)^{-1} = \sum w_i^2\left(r_i-\tilde{r}\right)^2\end{aligned} \tag{6}$$

This shows that if one has software capable of estimating precision-weighted regressions with heteroscedasticity-robust standard errors, it can also be used to estimate size-weighted means with the correct standard errors.

Furthermore, one can test the equality of the size-weighted means of two populations using a regression with robust standard errors. It is well-known that to test the equality of the equal-weighted means of two groups, A and B, one can run the regression

$$r_i = \beta_0 + \beta_1 D_i + \epsilon_i$$

where D_i indicates membership of observation i in group B. The analogous weighted regression

$$\sqrt{u_i}r_i = \beta_0\sqrt{u_i} + \beta_1\sqrt{u_i}D_i + \sqrt{u_i}\epsilon_i \tag{7}$$

provides the following intermediate results:

$$X'X = \begin{bmatrix}\sum_A u_i + \sum_B u_i & \sum_B u_i \\ \sum_B u_i & \sum_B u_i\end{bmatrix}$$

$$X'Y = \begin{bmatrix}\sum_A u_i r_i + \sum_B u_i r_i \\ \sum_B u_i r_i\end{bmatrix}$$

and

$$X'diag\left(\hat{\epsilon}_1^2,\cdots,\hat{\epsilon}_n^2\right)X = \begin{bmatrix}\left(\sum_A u_i^2\hat{\epsilon}_i^2 + \sum_B u_i^2\hat{\epsilon}_i^2\right) & \sum_B u_i^2\hat{\epsilon}_i^2 \\ \sum_B u_i^2\hat{\epsilon}_i^2 & \sum_B u_i^2\hat{\epsilon}_i^2\end{bmatrix}$$

where $\sum_A$ and $\sum_B$ indicate that the sums are taken over all observations in group A and group B respectively.

Straightforward calculations show that the estimated coefficients are given by:

$$\hat{\beta} = \begin{bmatrix}\hat{\beta}_0 \\ \hat{\beta}_1\end{bmatrix} = (X'X)^{-1}X'Y = \begin{bmatrix}\tilde{r}_A \\ \tilde{r}_B - \tilde{r}_A\end{bmatrix}$$

and the estimated variances are given by:

$$\begin{aligned}\hat{v}ar\left(\hat{\beta}\right) &= (X'X)^{-1}\left(X'diag\left(\hat{e}_1^2,\cdots,\hat{e}_n^2\right)X\right)(X'X)^{-1} \\ &= \begin{bmatrix}\sum_A w_i^2\left(r_i-\tilde{r}_A\right)^2 & -\sum_A w_i^2\left(r_i-\tilde{r}_A\right)^2 \\ -\sum_A w_i^2\left(r_i-\tilde{r}_A\right)^2 & \left(\sum_A w_i^2\left(r_i-\tilde{r}_A\right)^2 + \sum_B w_i^2\left(r_i-\tilde{r}_B\right)^2\right)\end{bmatrix}\end{aligned}$$

This shows that $\hat{\beta}_1$ in Equation (7) is the difference between the size-weighted

means of group B and group A. Its variance is estimated as $\sum_A w_i^2 \left(r_i - \tilde{r}_A\right)^2 + \sum_B w_i^2 \left(r_i - \tilde{r}_A\right)^2$, identical to the variance obtained by using Theorem 2 directly. The t-statistic for $\hat{\beta}_1$ directly tests the equality of the weighted means for groups A and B.

5. Conclusions and Discussion

Applied researchers frequently calculate and discuss size-weighted means. The statistical literature has often overlooked the interest of applied researchers in size-weighted statistics. This paper shows that under standard regularity conditions, the size-weighted mean for i.i.d. random variables is consistent and asymptotically normal. Its standard error can be estimated using a remarkably simple formula, which in some software can be obtained directly from the output of a weighted regression estimated with heteroscedasticity-robust standard errors. Further research in statistical methodology would be valuable to extend the current analysis to more complex settings.

Statistical software packages have also largely overlooked the interest of researchers in size-weighted statistics and provide limited guidance on the interpretation of weights. Researchers therefore risk substantially misreporting the precision of size-weighted means. Statistical software applications should be improved to avoid unintentional misuse. When a statistical software application provides procedures to calculate weighted means, it should clearly describe what the weights represent and how standard errors are calculated.

Acknowledgements

I thank Erin Smith, Chyhe Becker, and my colleagues in the Office of Litigation Economics at the U.S. Securities and Exchange Commission for valuable comments on drafts of the manuscript. I also like to thank an anonymous reviewer for his/her helpful comments and suggestions. The Securities and Exchange Commission disclaims responsibility for any private publication or statement of any SEC employee or Commissioner. This article expresses the author's views and does not necessarily reflect those of the Commission, the Commissioners, or other members of the staff.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Cochran, W.G. (1977) Sampling Techniques. 3rd Edition, Wiley.

[2] Kish, L. (1995) Survey Sampling (Wiley Classics Library). Wiley.