

Estimation of the Coefficient of Variation with Minimum Risk: A Sequential Method for Minimizing Sampling Error and Study Cost

Bhargab Chattopadhyay^a and Ken Kelley^b

^aDepartment of Mathematical Sciences, University of Texas at Dallas; ^bDepartment of Information Technology, Analytics, and Operations, University of Notre Dame

ABSTRACT

The coefficient of variation is an effect size measure with many potential uses in psychology and related disciplines. We propose a general theory for a sequential estimation of the population coefficient of variation that considers both the sampling error and the study cost, importantly without specific distributional assumptions. Fixed sample size planning methods, commonly used in psychology and related fields, cannot simultaneously minimize both the sampling error and the study cost. The sequential procedure we develop is the first sequential sampling procedure developed for estimating the coefficient of variation. We first present a method of planning a pilot sample size after the research goals are specified by the researcher. Then, after collecting a sample size as large as the estimated pilot sample size, a check is performed to assess whether the conditions necessary to stop the data collection have been satisfied. If not an additional observation is collected and the check is performed again. This process continues, sequentially, until a stopping rule involving a risk function is satisfied. Our method ensures that the sampling error and the study costs are considered simultaneously so that the cost is not higher than necessary for the tolerable sampling error. We also demonstrate a variety of properties of the distribution of the final sample size for five different distributions under a variety of conditions with a Monte Carlo simulation study. In addition, we provide freely available functions via the MBESS package in R to implement the methods discussed.

KEYWORDS

Coefficient of variation; sample size planning; sequential analysis; research design; sequential point estimation; U-statistics; accuracy; precision; stopping rule; minimum risk



The coefficient of variation is a standardized effect size measure that expresses the degree of variability with respect to central tendency. More specifically, the coefficient of variation for a set of scores is the standard deviation of the scores divided by the mean of the scores. The population coefficient of variation, denoted by κ , is

$$\kappa = \frac{\sigma}{\mu}, \quad (1)$$


where $\sigma = \sqrt{E[(X - \mu)^2]}$ is the population standard deviation and $\mu = E[X]$ is the population mean, with X representing a random variable.

The coefficient of variation is only meaningful when X is a nonnegative random variable, which we assume throughout this article. The coefficient of variation has an unambiguous meaning only when X is measured on a ratio scale, implying a true zero point and equal intervals. For example, Kendall and Stuart (1977) noted that the coefficient of variation suffers “from the disadvantage

of being very much affected by ...the value of the mean measured from some arbitrary origin, and [is] not usually employed unless there is a natural origin of measurement ...” (p. 48; see also Abdi, 2010; Snedecor, 1956). Work by Velleman and Wilkinson (1933) has shown that statistics such as the coefficient of variation can also be meaningfully interpreted for discrete scales involving a true zero point and equal intervals (count data), scales that do not meet Stevens’s (1946) classic definition of ratio scales. Some researchers have ignored these cautions and have used the coefficient of variation on Likert-type scales that have an arbitrary zero point and may not meet the criterion of equal intervals. Allison (1978) showed that the coefficient of variation is generally uninterpretable for such Likert-type scales. In addition, such scales often do not have the property that originally motivated the use of the coefficient of variation: “large things tend to vary much and small things little” (Snedecor, 1956, p. 44). Thus, the coefficient of variation will be clearly interpretable when there is a true zero point and equal

CONTACT Ken Kelley  kkelley@nd.edu  Department of Information Technology, Analytics, and Operations, Mendoza College of Business, University of Notre Dame, Notre Dame, IN 46556.

Correspondence may also be addressed to Bhargab Chattopadhyay, Department of Mathematical Sciences, 800 West Campbell Rd, FO 2.402A, University of Texas at Dallas, Richardson, TX 75080 (email: bhargab@utdallas.edu). Both authors contributed equally and authorship is alphabetical.

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2016 Taylor & Francis Group, LLC

intervals; its interpretation is far more controversial if these conditions are not met.¹

Kelley (2007c) discussed the coefficient of variation in psychology and related fields and proposed a method of sample-size planning from the accuracy in parameter estimation framework in order to accurately estimate the population coefficient of variation. Kelley (2007c) developed a method to plan a (fixed) sample size in order to have a specified degree of assurance that the confidence interval width would be sufficiently narrow. For example, the method Kelley (2007c) developed answers questions such as “for a specified population value of the coefficient of variation, what sample size is necessary in order for the 95% confidence interval to have 99% assurance of being .10 units or less?” We also consider the coefficient of variation here and the accuracy with which the population value has been estimated, but we approach sample-size planning from a different perspective.

The perspective that we take in this article considers sampling error and study cost simultaneously in a *sequential analytic framework*. Unlike Kelley (2007c), who did not consider study cost but rather only considered a sample size determined by a prespecified population coefficient of variation, here we do not make such an assumption of requiring a population value to be specified. The sequential framework does not have an a priori specified sample size to use for the study, as is the case with the traditional power analytic or the traditional accuracy in parameter estimation approach to sample-size planning that are often considered in psychology and related fields that base calculations on an unknown population value. Further, most sample-size planning methods do not consider study cost when planning necessary sample size. However, in this article we explicitly incorporate study cost into our method, which is a very salient issue when implementing a research study. That is to say, the sequential analytic framework used here depends on an a priori specified criterion or criteria with regard to estimation accuracy (in terms of sampling error) and the study cost. The sampling procedure stops once the specified

condition(s) is satisfied, fulfilling a *stopping rule* that is a characteristic of sequential estimation methods. A stopping rule determines whether sampling (i.e., collecting more data) should continue or stop after one (or more) additional observation(s) has been collected. The stopping rule for a traditional research design is reached when the a priori planned sample size from a power analysis or the accuracy in parameter estimation approach is satisfied. As will be shown, our method does not impose an a priori sample size, but rather the sample size ultimately used is unknown a priori and depends on satisfying the criteria specified by the researcher.

To summarize the problem that this article solves, we will develop a method that simultaneously considers the sampling error and the study cost when estimating the coefficient of variation, and we do so in a sequential estimation framework. This general framework is important because, in practice, both the study cost and the sampling error are of concern, yet most sample-size planning methods do not consider study cost and sampling error simultaneously. Cost is generally ignored when designing a study from a statistical perspective, yet cost is a very real consideration for researchers conducting a study. Our primary contribution in this article is a novel approach to a very practical problem in psychology and related disciplines as it relates to the coefficient of variation and the appropriate sample size for its accurate estimation.

To motivate our interest in the coefficient of variation, we first note that the coefficient of variation has a wide variety of potential uses in psychology and related disciplines and we believe it is poised to grow to be a more widely used effect size measure. In experimental psychology, Babkoff, Kelly, and Naitoh (2001) used the coefficient of variation to study reaction time in the context of sleep deprivation for three groups. In neurology reaction time study, Hayashi (2000) examined the coefficient of variation for reaction time when participants were using benzodiazepine (a drug with sedative, hypnotic, anxiolytic, and relaxant properties) in an effort to manipulate their cognitive state (Ornoy, Arnon, Shechtman, Moerman, & Lukashova, 1998). In organizational studies, Harrison, Price, and Bell (1998) used the coefficient of variation as a measure of group inequality (heterogeneity or diversity) with regard to the age of specific group members. In the context of speech disorders, Shriberg, Green, Campbell, McSweeney, and Scheer (2003) used the coefficient of variation to “normalize” the variability in durations of a participant’s speech events (actual speaking), as well as another coefficient of variation for the pause events (pauses during speaking; p. 581). We believe that the coefficient of variation will be of increasing importance due, in part, to the growing interest in simultaneously considering psychological and physiological systems as

¹ When the scale has a true zero point and equal intervals, the coefficient of variation has invariance properties that meet all of the desiderata for a standardized effect size (Kelley & Preacher, 2012). For example, imagine that a researcher wishes to calculate the coefficient of variation of temperatures in a city in the month of February. If temperature is measured on the Kelvin scale (whose units correspond to those of the Celsius temperature scale, but 0 is absolute 0) or the Rankine temperature scale (whose units correspond to those of the Fahrenheit temperature scale, but 0 is absolute 0), the two coefficients of variation will be identical. Such properties facilitate the development of guidelines for defining a sufficiently narrow value of the coefficient of variation. By contrast, if temperature is measured on the Fahrenheit and Celsius scales, which do not have true 0 points, the two coefficients of variation will differ substantially. Comparison of coefficients of variation for different mean temperature values on the same scale (e.g., Celsius) is no longer straightforward (Allison, 1978). In the absence of a true zero point, specification of consistent guidelines that define a sufficiently narrow value of the coefficient of variation becomes a challenging task.

do some of the aforementioned examples. Along those lines, Reed, Lynn, and Meade (2002) explained that, in many laboratories, the variability of chemical assays is summarized by the coefficient of variation. They argued that the main appeal of the coefficient of variation, as opposed to, for example, the standard deviation, is that “[standard deviations] of such assays generally increase or decrease proportionally as the mean increases or decreases, so that division [of the standard deviation] by the mean removes it as a factor in the variability” (p. 1235). It is known in statistical theory that when estimating a parameter of interest from a sample, error in estimation is unavoidable due to sampling. This error is known as sampling error: the random discrepancy between an estimate and the parameter it estimates. A typical approach to reducing the sampling error, holding everything else constant, is to increase sample size. Increasing sample size yields smaller sampling error but also increases study costs, specifically due to the increase in sampling cost. By “sampling cost,” we mean the cost involved in collecting data from the participants, which we regard here as a constant value (i.e., it cost the same to sample the 1st, 2nd, ..., n th observation).

We consider sampling cost to be one of two components of study cost: “structural cost” and “sampling cost.” We posit that for essentially any empirical study that will add to the scientific literature, a certain amount of resources for implementing the study are required. Beyond sampling costs, the financial resources that are required to design, conduct, analyze the data, including but not limited to costs for software licenses, equipment, salary, laboratory fees, and so on, all factor into the structural cost. These (nonsampling) *structural costs* are named as such because they speak to the infrastructure investment that one is willing to pay in order to have a sufficiently small sampling error of the coefficient of variation. The structural costs that are necessary for conducting a study can be considered *the amount one is willing to pay for a sufficient degree of accuracy*. Structural costs are important and pose a real limitation to what can be done in any given investigation. We discuss study cost (= sampling cost + structural cost) more as we develop our method.

Use of a proprietary scale that requires payment for use, scoring an assessment, recruiting an additional participant, participant honorarium, among other things, all affect the sampling cost. Suppose that it is calculated that each participant included in a study costs researchers, monetarily speaking according to all required resources, \$127.50. Having 50 participants in a study would thus entail sampling cost of \$6,375.00, whereas sampling cost would be \$12,750.00 to have 100 participants in a study. Thus, smaller estimation error, holding everything else constant, comes by increasing sample size,

which increases the sampling cost. Of course, if sampling cost were of no concern, the largest sample possible would be best from an accuracy standpoint. However, sampling cost is almost always a concern in empirical studies, and thus there is a practical limit to the size of a sample due to cost.

In this article we solve the general problem of obtaining an accurate estimate while considering the cost of estimating the coefficient of variation (i.e., study cost). We approach this problem from a sequential analysis framework, specifically what is known as the *minimum risk point estimation* problem (e.g., see De & Chattopadhyay, 2015; Sen & Ghosh, 1981). A method of conducting research that simultaneously considers the study cost *and* the sampling error is thus our focus here and offers advantages not usually considered in research design work. Fixed sample size procedures, procedures in which the sample size is fixed in advance before sampling, cannot achieve a trade-off between study cost and sampling error (e.g., see Dantzig, 1940; De & Chattopadhyay, 2015; Sen & Ghosh, 1981). If an approach to sample size planning depends only on the study cost, sampling error is not considered; by extension, statistical power and accuracy in parameter estimation were not considered. On the other hand, if an approach to planning sample size depends only on sampling error, the study cost is not considered. In either case there is a very important aspect of the research that is being ignored. Our work combines both sampling error and study cost into a unified framework for estimating the coefficient of variation.

A purely sequential procedure is proposed that yields a sample size for accurately estimating the unknown population coefficient of variation, taking into account the study cost. By *purely sequential* we mean that after pilot sampling stage, at every stage, one collects a single observation. In the next section we discuss estimating the coefficient of variation. We then discuss the minimum risk point estimation problem followed by the sequential optimization procedure we propose. We follow this with characteristics and properties of the procedure with proofs and justification of the sequential optimization procedure. We then provide an example scenario and include open source and freely available R code via the MBESS package (Kelley, 2007a, 2007b, 2016).

Estimation of the coefficient of variation

We now begin to formalize our ideas, beginning with estimating the coefficient of variation. Consider n independent and randomly selected individuals from some population of interest with scores denoted by X_1, X_2, \dots, X_n . The common, yet biased, estimator of the population coefficient of variation from Equation (1) is

$$k_n = \frac{s_n}{\bar{X}_n}, \quad (2)$$

where \bar{X}_n is the sample mean defined as

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \quad (3)$$

and s_n is the sample standard deviation defined as

$$s_n = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1}}. \quad (4)$$

That is, s_n is the square root of the usual unbiased estimator of the population variance. Note that we use the subscript n on the preceding sample estimates to explicitly note the sample size on which the estimator is based. Including the subscript is useful as we are considering the properties of the estimator based on different sample sizes. For an observed coefficient of variation in a particular study, usually only k would be used to represent the sample coefficient of variation actually observed and the n subscript would not be included.²

Suppose we want to estimate the population coefficient of variation, κ , accurately by having minimal sampling error. In other words, we want the estimated coefficient of variation, k_n , to be close to κ . More specifically, we seek to obtain a k_n that is within ϵ units of κ , where ϵ is a positive value (i.e., $\epsilon > 0$). That is, we want to estimate k_n such that it differs from the population value by no more than ϵ , namely, for it to be contained within the interval $(\kappa - \epsilon, \kappa + \epsilon)$. The value of ϵ is defined as the *maximum probable error*, which is the maximum absolute difference between k_n and κ that the researcher wishes to allow. Because we are using only a sample (of size n) to estimate κ , there is a chance that the estimate may fall outside the interval $(\kappa - \epsilon, \kappa + \epsilon)$. However, we seek to balance the trade-off between study cost and the chance that the estimate falls outside of the interval, which is a form of an optimization problem. With regard to the chance that the absolute difference between the population coefficient of variation and the estimated value of the coefficient of variation will exceed ϵ , we rely on Chebyshev's inequality so as to not invoke potentially unrealistic assumptions about the distribution of the data (e.g., Lim & Leek, 2012; Lord, 1953). In particular, the chance (expressed as a percentage) that the k_n will lie outside the interval $(\kappa - \epsilon, \kappa + \epsilon)$ will be

$$P(|k_n - \kappa| \geq \epsilon) \leq \frac{E[(k_n - \kappa)^2]}{\epsilon^2} \times 100\%. \quad (5)$$

² As can be seen from Equation (2), the coefficient of variation is undefined if $\bar{X}_n = 0$. We ignore this special case because we regard $P(\bar{X}_n = 0) \approx 0$ in practical situations.

Thus, we can say that the chance that the absolute difference between the population coefficient of variation, κ , and the estimated value of the coefficient of variation, k_n , exceeds ϵ is at most $E[(k_n - \kappa)^2]/\epsilon^2$, which is the quantity that we seek to minimize while considering study cost. From Abdi (2010), we note that k_n is not an unbiased estimator of κ , that is, $E[k_n] \neq \kappa$. Hence, the term $E[(k_n - \kappa)^2]$ (i.e., the numerator of the quantity we seek to minimize) is the mean square error (MSE) of k_n (not the variance). We define the MSE formally momentarily, but for now, consider that the MSE is a sum of the precision (variance) and squared bias (Rozeboom, 1966). For an estimator that is unbiased, the MSE and the variance are equal. However, due to the bias, the MSE is larger than simply the variance by an amount equal to the squared bias. If the MSE of k_n is very small (i.e., on average, the squared discrepancy between the estimate given by the estimator k_n and the population coefficient of variation κ is very small), then there is a high probability of estimating κ accurately. In other words, the chance that the estimate will lie inside the interval $(\kappa - \epsilon, \kappa + \epsilon)$ may be high.

Suppose that, excluding the sampling costs, a researcher is willing to pay \$100 so that the absolute difference between the point estimate of the coefficient of variation, k_n , and its corresponding population value, κ , will be at most ϵ . In other words, the researcher is willing to invest \$100 in the structural cost of performing a study, again, excluding the sampling costs, so that the difference between the estimate and population value will be sufficiently small. We note that

$$|k_n - \kappa| \leq \epsilon \iff \quad (6)$$

$$(k_n - \kappa)^2 \leq \epsilon^2, \quad (7)$$

where \iff means "if and only if." Therefore, we can say that the researcher is willing to pay \$100 so that the squared difference between the point estimate of the coefficient of variation, k_n , and its corresponding population value, κ , will be at most ϵ^2 . Due to the sampling error, which is unknown because κ is unknown, we must work with the expectation of the squared difference between k_n and κ (i.e., the mean square error). That is, because κ is unknown, the actual amount that is being paid for the expected squared difference (i.e., the mean square error) is given as $AE[(k_n - \kappa)^2]$, where, in this particular example, $A = \$100/\epsilon^2$. Thus, A has a unit of "dollar per square unit of ϵ ." Conceptually, this idea translates into the "price one is willing to pay per squared unit of maximum probable error."

When designing a study, one can choose A directly, by specifying the dollar per square unit of ϵ that one is willing to pay, or indirectly, by specifying its two components, namely, the structural cost one is willing to invest and the

desired ϵ . For example, if one would be willing to pay \$100 for a sufficiently accurate estimate of κ to be within 0.05, the value of A would be \$40,000(= $\frac{100}{.05^2}$). The value that a researcher is willing to pay for a desired level of ϵ is subjective and context specific, as is the desired ϵ itself. The value of A depends on the amount of money (e.g., US dollars) one is willing to pay for a sufficiently small deviation from the parameter (i.e., the maximum absolute difference desired between the population value and its estimate). Smaller values of ϵ will lead to larger values of A , holding constant the structural cost that one is willing to pay.³

When a study's goal is to estimate a parameter accurately, such as the coefficient of variation here or for any effect size more generally, the structural costs and the maximum probable error of the estimate (i.e., ϵ) are combined to form A . When we say "what the researcher is willing to pay," we literally mean the structural cost the researcher is willing to invest in a study in order to estimate the parameter of interest with the desired degree of accuracy. This value is implicitly included (along with anticipated sampling cost) in many grant applications for empirical studies when a certain amount of money is requested to conduct a study (less the sampling cost). Ignoring the overhead cost of many grant applications, consider the total amount of money requested, less whatever funds will be used for sampling costs. The nonsampling costs are the structural costs that a researcher is agreeing to invest in order to obtain the desired outcome, namely, an accurate estimate of the parameter of interest. If a researcher is willing to pay more and/or desires a smaller value of ϵ , A is larger than it would have been. A larger value of A will translate into a more expensive study, holding everything else constant. Notice that A is a fixed value in any investigation and specified a priori, as the researcher specifies A directly or by specifying its two components (structural cost and ϵ) individually (and does not depend on data, as it is specified a priori). However, what is not fixed but rather is evaluated in multiple steps

³ To provide an analogy outside of the research framework for a better conceptual understanding, consider shopping for a car in which a goal is to minimize downtime (e.g., for maintenance, repairs, refueling/charging). There are two types of costs that can be considered: the cost of the car itself and the cost to operate the car per mile. In this scenario, the "cost of the car" is the analog of what we are calling the structural costs, whereas the "cost per mile" of operating the car is the analog of sampling cost per mile. A consumer may be "willing to pay" \$20,000 for a car (structural cost). Separate from the cost of the car itself is the cost of operating the car, which is estimated to be \$.10/mile (sampling cost). Further, consider that the probable downtime (e.g., per week), which maps onto our "probable error," is 5 hours. Thus, we would have $A = \$20,000/5^2$ and we would then add \$.10 for each mile driven to accomplish the goal of minimum downtime. Thus, the total cost involved for a certain amount of usage would be
Total cost = structural cost + mile × .10 = \$20,000 + Miles × \$.10, which is the analog of our study cost (= structural cost + sampling cost).

throughout the process is the sampling cost, and the necessary sample size that will accomplish the study's goal of achieving a sufficiently accurate estimate of the coefficient of variation is unknown. This is the core contribution of this article: minimizing sampling cost, and thereby study cost, by using a sequential procedure that provides a stopping rule once an optimization function is minimized that considers cost and accuracy according to the goals of the researcher. Throughout this article, we regard sampling cost as a constant (fixed) per participant (i.e., the cost for sampling participants is c regardless of the number of participants).

Before moving to the optimization function we discuss accuracy, which statistically is conceptualized as a function of precision and bias (e.g., Rozeboom, 1966). Holding constant bias, improving precision improves accuracy. We are improving precision and, by not increasing bias, we obtain a more accurate estimate. We prefer to use the term *accuracy* instead of *precision* in this context to make clear that we are not focused solely on precision at the expense of bias, but rather that we are concerned with both bias and precision as our procedure improves precision but does not worsen bias. Recalling that we are working in a distribution free environment, we now quantify the MSE. Using Bao (2009), the expression for $E[(k_n - \kappa)^2]$ is

$$E[(k_n - \kappa)^2] = \frac{\xi^2}{n} + \eta, \tag{8}$$

where η is the expected value of the residual term of a Nagar-type expansion (see Nagar, 1959) of k_n and ξ^2 depends on four unknown parameters: (a) population mean (μ), (b) population variance (σ^2), (c) third central moment (μ_3), and (d) fourth central moment (μ_4).⁴ Specifically, ξ^2 is given by

$$\xi^2 = \frac{\mu_4}{4\mu^4} + \frac{\sigma^4}{4\mu^4} - \frac{\mu_3}{\mu^3} + \frac{\sigma^2}{2\mu^2}. \tag{9}$$

The expression of η is given in Bao (2009). For *not too small sample sizes*, η is negligible, and thus ignoring η will have negligible effect on the expression's value for most purposes; we demonstrate this ignorability with a

⁴ The Nagar-type expansion of coefficient of variation, k_n , given in equation (4) in Bao (2009) is

$$k_n = \frac{\sigma}{\mu} \left[1 + \frac{1}{2} \frac{S_n^2 - \sigma^2}{\sigma^2} - \frac{1}{8} \left(\frac{S_n^2 - \sigma^2}{\sigma^2} \right)^2 + \frac{1}{16} \left(\frac{S_n^2 - \sigma^2}{\sigma^2} \right)^3 + o_p(n^{-3/2}) \right] \times \left[1 - \frac{\bar{X}_n - \mu}{\mu} + \left(\frac{\bar{X}_n - \mu}{\mu} \right)^2 - \left(\frac{\bar{X}_n - \mu}{\mu} \right)^3 + o_p(n^{-3/2}) \right].$$

Monte Carlo simulation study in the the supplementary material.⁵

The approximate expression of the MSE of k_n is

$$E[(k_n - \kappa)^2] \approx \frac{\xi^2}{n}. \quad (10)$$

Equations (9) and (10) consist of a mean (μ), variance (σ^2), third-central moment (μ_3), and fourth-central moment (μ_4). To be clear, these four central moments do not define a particular distribution. For example, the log-normal distribution and the perturbed log-normal distribution each have the same mean, variance, third-central moment, and fourth-central moment, yet the distributions are different (in that the shapes differ, e.g., Durrett, 2010, pp. 103–104). Thus, if only the fixed values of μ , σ^2 , μ_3 , and μ_4 are provided, one cannot say for certain that the distribution of the data is normal, exponential, or some other distribution. Equations (9) and (10) are valid for all distributions with finite fourth moment [$E(X^4) < \infty$].

From the approximate expression of the MSE defined in Equation (10), we see that the MSE of k_n depends, in part, on the sample size. To have a higher chance that the estimate of κ will lie within $(\kappa - \epsilon, \kappa + \epsilon)$, a larger sample size is required. Of course, a larger sample size will inflate the study cost, specifically by inflating sampling cost. Thus, the problem we seek to solve is to find the minimum sample size required to estimate κ accurately while taking into consideration the sampling cost, which we solve in a minimum risk point estimation framework.

Minimum risk point estimation problem

Suppose we have n independent observations X_1, \dots, X_n with a common but unknown distribution function, F . We estimate the population coefficient of variation, κ , with the estimator k_n , as defined in Equation (2). As the sample size grows larger and larger, we know, statistically, more and more information about the unknown population coefficient as the MSE (i.e., $E[(k_n - \kappa)^2]$) becomes smaller and smaller (i.e., accuracy improves). However, a larger sample size also leads to a larger sampling cost. Recall that by sampling cost we mean the cost associated with collecting data (and not structural costs). Let c be the known cost

of sampling each observation; for example, the value of c is \$127.50 in the aforementioned example, where it is calculated that every participant that is included in a study costs researchers \$127.50. We hold c constant throughout this article.

To account for both the sampling error and the study cost, drawing on Equation (5) we define the following function, known as a risk function, which provides the expected cost of estimating κ (by using k_n) using a sample of n observations with a maximum probable error ϵ . This risk function is defined as

$$R_n(\kappa) = AE[(k_n - \kappa)^2] + cn, \quad (11)$$

where cn represents the cost of sampling n observations at a cost of c per participant (thus, multiplying n and c yields the sampling cost for n observations). The values of A and c are fixed in any given application and specified by the researcher, but to be clear n is not known a priori but is updated (sample size increased) in the sequential sampling framework we use.

Now, returning to A specifically in the sequential sampling framework, we formally conceptualize A as *the structural cost that the researcher would be willing to pay per squared unit of ϵ* . The value of ϵ is the desired maximum probable error, $|k_n - \kappa| \leq \epsilon$. The value of A is defined as

$$A = \frac{\text{Structural Cost}}{\epsilon^2}, \quad (12)$$

with Structural Cost being the investment made in the study not due to the cost of sampling. Thus, we are conceptualizing study cost as having two components, the fixed cost that one is willing to pay (for squared unit of ϵ) and the cost of sampling:

$$\text{Study Cost} = \text{Structural Cost} + \text{Sampling Cost}. \quad (13)$$

In our framework, we regard the cost of sampling each observation, as fixed (i.e., for each additional observation, the cost of sampling is the same). Consider A from a very practical perspective, namely, a grant application in which a researcher requested funding to accurately estimate the coefficient of variation. Here, the numerator of A (i.e., the structural cost) would be the funds requested for the grant that do not involve sampling observations. Our objective is to find the sample size for which the expected study cost, defined in the risk function of Equation (11), is minimized. Because A is fixed, for a given c the study cost is minimized by minimizing the necessary sample size while still achieving the specified level of accuracy. We seek to *optimize both the sampling cost and the accuracy of the estimate*. This is known as the *minimum risk point estimation problem*, and $R_n(\kappa)$ is called the risk function of estimating κ with a sample of size n .

⁵ By “not too small sample sizes” here and elsewhere we mean a sample size that is large enough so that the noted properties hold. The exact value of “not too small” is context specific. For example, this is much like the large enough sample size required in order for the sampling distribution of sample means to take on a normal form, which the central limit theorem shows will happen with a large enough sample size. In particular, the central limit theorem says that as sample size gets larger and larger, the sampling distribution of the sample means approaches a normal distribution. Thus, provided sample size is “not too small,” the sampling distribution of sample means will be normal. For very skewed parent distributions, the sampling distribution of the mean can require a larger sample size to become normal than for parent distributions that are themselves close to normal.

The *minimum risk point estimation* problem was developed in the pioneering article of Robbins (1959). He suggested a *purely sequential procedure* for the risk point estimation of the mean of a normal distribution, which we discuss momentarily. We note that a procedure in which, after the pilot sampling stage, one observation is collected at each stage of a sampling process is known as a purely sequential procedure. The minimum risk point estimation problem was generalized by Ghosh and Mukhopadhyay (1979), who introduced a distribution-free scenario and developed a purely sequential procedure for minimum risk point estimation of a population mean. Sen and Ghosh (1981) suggested a purely sequential procedure for the risk point estimation of any parameter using an unbiased estimator based on U-statistics. For estimating the population coefficient of variation, the estimator, k_n , is used. Recall that k_n is the ratio of the sample standard deviation to the sample mean. The sample mean is an unbiased estimator of the population mean, whereas the sample standard deviation is not an unbiased estimator of the population standard deviation. This article considers the minimum risk point estimation of κ in which the estimator is a ratio of two different kinds of estimators, one of which (the standard deviation) is not an unbiased estimator of its parameter.

For not too small sample sizes, combining Equations (8) and (11) and ignoring η from Equation (8), the approximate fixed sample size risk function or the approximate expected study cost for estimation of κ is

$$R_n(\kappa) \approx A \frac{\xi^2}{n} + cn. \tag{14}$$

The risk function—that is, the expected cost of estimating κ , defined in Equation (14)—involves ϵ , sampling error, structural cost, and sampling cost. Again, A and c are fixed in any given application. As the sample size increases, $A\xi^2/n$ decreases while n (and thus cn) increases. This is an optimization problem in which the approximate risk function, defined in Equation (14), needs to be minimized. For not too small samples, if ξ were known, the approximate risk function in Equation (14) is minimized (using derivatives of the right hand side of Equation [14]) at

$$n_c = \sqrt{\frac{A}{c}} \xi, \tag{15}$$

which we call the *theoretically optimal sample size*.

Using a sample of size n_c , which is the theoretically optimal fixed sample size (if the parameter ξ is known) that minimizes both the sampling error and sampling cost, the risk function or the expected cost for estimating κ using the minimum number of observations (using Equation [15], $n_c^2 = A\xi^2/c$; i. e. $cn_c = A\xi^2/n_c$) is denoted

as

$$R_{n_c}^*(\kappa) = A \frac{\xi^2}{n_c} + cn_c = 2cn_c. \tag{16}$$

$R_{n_c}^*(\kappa)$ is called the *minimum asymptotic risk*. In practice, ξ^2 is unknown, and thus an estimator of ξ^2 is desired. Also unknown in practice is n_c . However, as will be proved statistically for not too small samples, our method yields sample sizes with properties that closely approximate n_c in applied situations.

We note that even though the value of ξ^2 depends on the first four central moments of a distribution, it does not depend on a particular distributional assumption. In other words, we are agnostic to the type of distribution that the scores from which the sample coefficient of variation will be calculated follows, as we are working in a distribution-free environment. This is very useful, as the distribution of the scores from the sampled population is generally unknown in practice. Thus, what follows is importantly distribution free. Because ξ^2 is unknown in practice, in the next section, we find an estimator based on U-statistics, which does not rely on distributional assumption. We discuss U-statistics in the next section.

Estimator of the unknown parameter and U-statistics

The estimator of the coefficient of variation in Equation (2) involves a function of the sample mean (Equation (3)) and the sample variance (i.e., the square of Equation (4)), both of which belong to a class of unbiased estimators known as U-statistics, yet the sample standard deviation, Equation (4), is not a U-statistic. Hoeffding (1948) introduced the idea of U-statistics and defined a U-statistic as an unbiased estimator of some parameter, θ , that is associated with an unknown distribution function, F . Suppose that X_1, \dots, X_n are independent and identically distributed (i.i.d.) random variables from a population with a common distribution function F (e.g., F could be a normal distribution, log-normal distribution, gamma distribution, etc.) with an associated parameter θ . More formally, the U-statistic associated with some θ is written as

$$U \equiv U_n^{(r)} = \binom{n}{r}^{-1} \sum_{(n,r)} g^{(r)}(X_{i_1}, \dots, X_{i_r}), \tag{17}$$

where $\sum_{(n,r)}$ denotes the summation over all possible combinations of indices (i_1, \dots, i_r) such that $1 \leq i_1 < i_2 < \dots < i_r \leq n$, and $r < n$. When working with U-statistics, the idea of a kernel is important. A kernel is a generic function of the smallest number of random variables required, which is called the degree, to estimate the

parameter θ unbiasedly. Here, $g^{(r)}(\cdot)$ is a symmetric kernel of degree r , with *symmetric* meaning that changing the arrangement of the r random variables will not affect the value of $g^{(r)}(\cdot)$. For example, $g^{(r)}(X_1, X_2, \dots, X_r) = g^{(r)}(X_2, \dots, X_r, X_1)$ and so on. In addition, $E_F[g^{(r)}(X_1, \dots, X_r)] = \theta$ for all F with r being the minimum sample size required to estimate θ unbiasedly. In this way, we can define unbiased estimators of several parameters using Equation (17). For more details about U-statistics, we suggest readers consult Hollander and Wolfe (1999), Kowalski and Tu (2008), Lee (1990), among others.

We now consider the estimator of the population mean (i.e., μ). Because $E[X_i] = \mu$ for $i = 1, \dots, n$, it is the case that the smallest number of random variables required to estimate μ is 1 (as the expectation does not depend on sample size). The kernel (i.e., the generic function to estimate the parameter unbiasedly) will thus be $g^{(1)}(X_i) = X_i$, which is of degree 1 (i.e., $r = 1$). Applying Equation (17), we can see that

$$U_n^{(1)} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X}_n. \quad (18)$$

Now, suppose we want to estimate σ^2 . Then, $\frac{1}{2}E[(X_{i_1} - X_{i_2})^2] = \sigma^2$. Thus, we need at least two random variables to estimate σ^2 unbiasedly. Hence, for the population variance, the degree is $r = 2$ and the kernel is $g^{(2)}(X_{i_1}, X_{i_2}) = \frac{1}{2}(X_{i_1} - X_{i_2})^2$. If we interchange the position of random variables, $g^{(2)}(X_{i_1}, X_{i_2})$ will remain the same; that is, $g^{(2)}(X_{i_1}, X_{i_2}) = g^{(2)}(X_{i_2}, X_{i_1})$. So, $g^{(2)}(X_{i_1}, X_{i_2})$ is a symmetric kernel of degree 2. Applying Equation (17) for $r = 2$, we can see that

$$U_n^{(2)} = \frac{1}{2} \binom{n}{2}^{-1} \sum_{1 \leq i_1 < i_2 \leq n} (X_{i_1} - X_{i_2})^2 = s_n^2. \quad (19)$$

For technical details about the expression of $U_n^{(2)}$ and the sample variance s_n^2 , we refer the reader to Mukhopadhyay and Chattopadhyay (2012, 2014). Again, suppose we want an estimator based on U-statistics for the population's third-central moment, that is, $\mu_3 = E[(X - \mu)^3]$, and the population's fourth-central moment, $\mu_4 = E[(X - \mu)^4]$. The U-statistics-based unbiased estimators for the third- (μ_3) and the fourth-central moments (μ_4) are

$$\hat{\mu}_{3n} = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n (X_i - \bar{X}_n)^3 \quad (20)$$

and

$$\hat{\mu}_{4n} = \frac{n^2}{(n-1)(n-2)(n-3)} \sum_{i=1}^n (X_i - \bar{X}_n)^4 - \frac{2n-3}{(n-1)(n-2)(n-3)} \sum_{i=1}^n X_i^4$$

$$+ \frac{8n-12}{(n-1)(n-2)(n-3)} \bar{X}_n \sum_{i=1}^n X_i^3 - \frac{6n-9}{n(n-1)(n-2)(n-3)} \left(\sum_{i=1}^n X_i^2 \right)^2, \quad (21)$$

respectively (e.g., Abbasi et al., 2010; Heffernan, 1997). The quantity $\hat{\mu}_{3n}$ is a U-statistic of degree 3 and is an unbiased and consistent estimator of μ_3 , whereas $\hat{\mu}_{4n}$ is a U-statistic of degree 4 and is an unbiased and consistent estimator of μ_4 . Recall that a consistent estimator is an estimator that converges to the population value that it estimates as sample size gets larger. The estimator of ξ^2 that is used to estimate the minimum risk function, defined in Equation (16), is given by

$$V_n^2 = \frac{s_n^4}{4\bar{X}_n^4} + \frac{\hat{\mu}_{4n}}{4\bar{X}_n^4} + \frac{s_n^2}{2\bar{X}_n^2} - \frac{\hat{\mu}_{3n}}{\bar{X}_n^3}, \quad (22)$$

which we find to be a consistent estimator of ξ^2 (using theorem 3.2.1 of Sen, 1981, p. 50). Note that the minimum risk function, defined in Equation (16), contains the theoretically optimal sample size, n_c , which depends on ξ . In practice, ξ is generally unknown and we estimate ξ by V_n , which is the square root of Equation (22).

We have discussed U-statistics here because they are essential to the remainder of the article. In particular, we use U-statistics because, for a large class of probability distributions, the theory of U-statistics allows for a minimum-variance unbiased estimator to be derived from each unbiased estimator of the parameter (e.g., see Cox & Hinkley, 1979). Note that among all unbiased estimators of a parameter, a minimum-variance unbiased estimator is always preferred because (a) it is unbiased and (b) it has the lowest variance (and thus the smallest MSE) among all possible unbiased estimators.

Unless the value of ξ is known, the optimal value of fixed sample size, n_c , cannot be computed. We note that ξ depends on four parameters that would generally be unknown in applied situations. Thus, in an effort to avoid using a potentially poor estimate of ξ , such as that which might be obtained by using supposed population values obtained in some way, which are potentially poor estimates in which to plan a fixed sample size, we develop a new approach. The approach we develop is a sequential sampling procedure that, importantly, does not require that a researcher plug in supposed population values as if they are known. Rather, our method ensures that we are informed by actual data from the population of interest. Correspondingly, our sequential estimation procedure is used to find an estimate of the optimal fixed sample size, n_c , which will provide an accurate estimate of κ with the minimum sampling cost and thereby study cost.

Sequential optimization procedure

In sequential estimation procedures, as opposed to fixed-sample-size estimation methods, the estimation of parameter(s) proceeds in stages. In the first stage of a sequential estimation procedure, a sample (called the pilot sample) is observed to gather preliminary information about the parameter(s) of interest. Then, in successive stages, the researcher collects one (or more) additional observation(s) and then he or she estimates the parameter(s) of interest, which is done again and again until a predefined condition has been satisfied (i.e., the stopping rule is met). That is, after collection of one (or more) additional observation(s), the parameter estimate(s) is (are) recalculated and a check is performed in order to make a decision to either (a) terminate the sampling process or (b) continue with the sampling process. This decision is based on a predefined stopping rule. In a sequential procedure, after the pilot sampling stage, one observation is collected at each stage of a sampling process.

No fixed sample size procedure can provide a solution to the minimum risk point estimation problem (e.g., see Dantzig, 1940; De & Chattopadhyay, 2015), which is why we propose a purely sequential procedure. For details about the general theory of sequential procedures, we refer interested readers to Ghosh and Sen (1991), Mukhopadhyay and Chattopadhyay (2012), and Sen (1981). We also note that the idea of a stopping rule is extensively used to determine the number of interim analyses in clinical trials or in deciding when clinical trials should stop further recruiting. These important problems are discussed not only in the frequentist framework but also in the Bayesian framework. For details about the application of stopping rules and sequential analysis in clinical trials, we refer to Armitage (2014), Ciarleglio, Arendt, Makuch, and Peduzzi (2015), Freedman and Spiegelhalter (1983), Spiegelhalter, Abrams, and Myles (2004), among others.

We use the theory of sequential procedures in this article to develop a method to estimate the coefficient of variation, an effect size of interest in psychology and related fields. In the following subsection we describe how to implement the procedure we have developed for estimating the coefficient of variation.

Implementation of the sequential sampling procedure for the coefficient of variation

As discussed, we will essentially never know all of the population parameters in practice necessary to know n_c , the theoretical sample size. Therefore, to implement a

study that considers the coefficient of variation as we have described, we propose the following method.

Let m be the initial, termed pilot, sample size and let N_c be the final sample size that gives an estimate of the unknown optimal sample size (i.e., N_c estimates n_c). To find an estimate of the desired sample size (i.e., N_c) required to minimize both the approximate sampling error and the sampling cost of estimating the population coefficient of variation, we propose the following purely sequential estimation procedure:

- Stage 1: In the initial stage, obtain a sample, called a pilot sample, of size m . From this pilot sample of size m , obtain an estimate of ξ^2 by finding V_m^2 as given in Equation (22) and check whether $m^2 \geq \frac{A}{c}(V_m^2 + m^{-2\gamma})$. If $m^2 < \frac{A}{c}(V_m^2 + m^{-2\gamma})$, then go to the next step. Otherwise, if $m^2 \geq \frac{A}{c}(V_m^2 + m^{-2\gamma})$, then report that the final sample size is $N_c = m$. We will discuss momentarily the use and choice of γ and a way to obtain the pilot sample size.
- Stage 2: Obtain an additional observation. At this stage there are $(m + 1)$ observations. Update the estimate of ξ^2 by computing V_{m+1}^2 . Now check whether $(m + 1)^2 \geq \frac{A}{c}(V_{m+1}^2 + (m + 1)^{-2\gamma})$. If $(m + 1)^2 \geq \frac{A}{c}(V_{m+1}^2 + (m + 1)^{-2\gamma})$, then stop further sampling and report that the final sample size is $N_c = m + 1$. Otherwise, if $(m + 1)^2 < \frac{A}{c}(V_{m+1}^2 + (m + 1)^{-2\gamma})$, then go to the next step.
- Stage 3: Obtain an additional observation. At this stage there are $(m + 2)$ observations. Update the estimate of ξ^2 by computing $V_{m+2}^2 + (m + 2)^{-2\gamma}$. Now check whether $(m + 2)^2 \geq \frac{A}{c}(V_{m+2}^2 + (m + 2)^{-2\gamma})$. If $(m + 2)^2 \geq \frac{A}{c}(V_{m+2}^2 + (m + 2)^{-2\gamma})$, then stop further sampling and report that the final sample size is $N_c = m + 2$. Otherwise, if $(m + 2)^2 < \frac{A}{c}(V_{m+2}^2 + (m + 2)^{-2\gamma})$, then continue the sampling process and update the sample size until the condition $n^2 \geq \frac{A}{c}(V_n^2 + n^{-2\gamma})$ is met, where $n \geq m$.

This process of collecting one additional observation in each stage after stage 1 is continued until there are N_c observations such that $N_c^2 \geq \frac{A}{c}(V_{N_c}^2 + N_c^{-2\gamma})$. At that stage, we stop further sampling and report that the final sample size is N_c .

For not too small sample sizes, $(V_n^2 + n^{-2\gamma})$ converges to ξ^2 . So the square root of $\frac{A}{c}(V_n^2 + n^{-2\gamma})$ is in fact estimating the optimal sample size, n_c . At each stage in the sequential procedure outlined in the preceding, we are

checking whether the collected sample size is larger than the estimated optimal sample size, or in other words, $n^2 \geq \frac{A}{c}(V_n^2 + n^{-2\gamma})$. From the algorithm just outlined, the stopping rule, N_c can be defined as follows:

$$N_c \text{ is the smallest integer } n(\geq m) \text{ such that} \\ n^2 \geq \frac{A}{c}(V_n^2 + n^{-2\gamma}), \quad (23)$$

where $\gamma \in (0, 1/2)$ with the term $n^{-2\gamma}$ being a correction term that ensures the sampling process does not stop too early (because of the use of the approximate expression) for the estimation of the optimal sample size.⁶ For details about the correction term, refer to De and Chattopadhyay (2015) or Sen and Ghosh (1981). For practical purposes, one can use $\gamma = 0.49$.⁷

If observations are collected using Equation (23), then sampling will stop at some stage with probability one. This is proved in Lemma 1 in the supplementary material, which shows that, under appropriate conditions, $P(N_c < \infty) = 1$. This result is very important as it ensures mathematically that the sampling will be terminated eventually.

To summarize, what we have shown so far is how to find an estimate of the desired sample size in which both the approximate sampling error and the study cost are minimized. This is a useful procedure because it simultaneously considers the sampling error and the study cost when estimating the coefficient of variation. If study cost were of no concern, a larger sample size would always be preferred because the sampling error would be reduced. If sampling error were of no concern, a smaller sample size would be preferred because the study cost of obtaining a sample would be minimal. However, in practice, both the study cost and the sampling error are of concern. Most sample-size planning methods do not consider study cost and sampling error simultaneously. Our primary contribution in this article is thus a novel approach to a very practical problem in psychology and related disciplines as it relates to the coefficient of variation and the appropriate sample size.

Choice of pilot sample size

Recall that in the first stage of a sequential estimation procedure, a sample size m is collected, called the pilot sample. This pilot sample is used to gather preliminary information about the parameter(s) of interest. If the pilot sample size m is too small, the number of sampling stages

in a sequential procedure may be large (e.g., if $m = 5$ yet $N_c = 1,000$, which is 955 additional sampling stages that are necessary). On the other hand, if pilot sample size is very large, we may end up using more samples than we actually need to achieve a certain goal (e.g., if $m > N_c$). A poor choice of the pilot sample size can lead to many sampling stages or inflate the sampling cost (and thereby study cost) by initially collecting more observations than necessary. Clearly, a proper choice of pilot sample size is important. Using the stopping rule defined in Equation (23), the final sample size should always be greater than $(A/c)^{1/(2+2\gamma)}$. Following Mukhopadhyay and De Silva (2009, p. 251), we recommend the use of the pilot sample size m as

$$m = \max \{m_0, \lceil (A/c)^{1/(2+2\gamma)} \rceil\}, \quad (24)$$

where $m_0(\geq 4)$ is the minimum possible sample size required to estimate ξ^2 . Here, $\lceil \cdot \rceil$ is the ceiling function of the quantity, meaning one “rounds up” to the next integer. For example, $\lceil 90.005 \rceil = 91$; $\lceil 90.9995 \rceil = 91$.

Characteristics of our sequential procedure

For a given cost c per observation, the risk function for using the estimator of the coefficient of variation as defined in Equation (2) according to the final sample size N_c is given by

$$R_{N_c}(\kappa) = AE[(k_{N_c} - \kappa)^2] + cE[N_c]. \quad (25)$$

Theorem 1 is defined and proven in the supplementary material and is very important. Theorem 1 is important because, under appropriate conditions, it ensures that, on average, the final sample size, N_c , is close to the optimal sample size, n_c , and that, on average, the risk, $R_{N_c}(\kappa)$, at the final sample size, N_c , is close to the minimized risk, $R_{n_c}^*(\kappa)$, which was defined in Equation (16).

Example

Suppose that a research team seeks to quantify the diversity (which can be conceptualized as inequality or heterogeneity) within schools in a large urban district. Of primary interest is the diversity of age-appropriate books in the homes of third graders.⁸ For purposes of our example, we focus only on a single school.

Diversity can be conceptualized as the coefficient of variation (e.g., Bedeian & Mossholder, 2000; Harrison et al., 1998), which provides a standardized measure of variability relative to the mean. The most appropriate

⁶ We note that incorporating the correction term will not affect the consistency property of $V_n^2 + n^{-2\gamma}$, the estimator of ξ^2 , and ensures that the sampling process does not stop early.

⁷ For not too small sample sizes, $(V_n^2 + n^{-2\gamma})$ converges to ξ^2 . Thus, the convergence rate increases as γ increases. So a higher value of γ , for example $\gamma = 1$, is a good choice. Now, if one uses a value of γ higher than 0.5, then part (ii) of theorem 1 will not be satisfied theoretically.

⁸ Measures such as time spent (a) using computers/tablets, (b) watching television, and (c) playing outside during a typical week might be collected, as well as various demographic, educational attainment, and performance measures.

way to collect the data needed on the number of age-appropriate books for the third graders in the schools of interest is thought to be an in-person survey conducted in each student's home. Although the research team seeks an accurate estimate of the true coefficient of variation of the age-appropriate books of the third graders in the school, there is limited funding to be used on in-home data collection. Of course, the more research funds spent on data collection, the fewer funds available for other research questions or projects. Thus, the research teams seeks a balance between the estimation accuracy (i.e., small sampling error) of the coefficient of variation and the cost of collecting data (i.e., sampling cost). The ideal sample size is not obvious: An accurate estimate is of interest but so too is the minimum sampling cost for the study. Thus, the cost-benefit analysis needs to explicitly consider both of these competing issues. Our method provides a formal way of considering both sampling cost and estimation accuracy, something that may be implicitly done by researchers but has received little attention in the research design literature within psychology and related disciplines.

First, we need to consider the cost of a single in-home visit. This is calculated to be, on average, \$75 per visit (i.e., $c = \$75$). This cost-per-observation includes an honorarium for the participating household, travel expenses for the in-home surveyor, and the cost of the salary of the in-home surveyor who will count age-appropriate books. Estimation of the cost per sample (i.e., the sampling cost for collection of a single observation) can generally be done according to the known or anticipated values of an investigation (e.g., anticipated time data collection will take, anticipated salary of those involved, anticipated honorarium of the surveyor and participants, etc.). The cost per sample is a value generally estimated in, say, a grant application, in that the anticipated sample size multiplied by the cost per observation is a value needed in order to know what amount of money should be invested for data collection.

Second, we need to consider the "maximum probable error" in estimation of the population coefficient of variation (i.e., $|k_n - \kappa| \leq \epsilon$). The probable error is a value not often considered in psychology or related disciplines, but it is important in terms of quantifying the accuracy of an estimate. Suppose that the desire is to have the difference between k_n (the estimate from a sample of size n) and κ (the true value) be 0.05 units or less. Further, suppose the research team is willing to pay \$1,000 for an estimate with such maximum probable error (i.e., $\epsilon = 0.05$), not considering the sampling cost. This \$1,000 translates into the structural cost. Using this, we get $A = \frac{\$1,000}{.05^2} = \frac{\$1,000}{.0025} = \$400,000$, as discussed in Equation (5). For another example to illustrate A , had

the desire been for k_n to be within an interval of 0.1 units around κ and the researcher was willing to pay \$1,000 for the accuracy of such an estimate, A would be \$100,000 ($= \frac{\$1,000}{.1^2} = \frac{\$1,000}{.01}$). Although A is literally the price one is willing to pay per squared unit of ϵ , in and of itself it is not very interpretable as it is a conflation of two values. However, those two values, structural cost (or price one is willing to pay) and ϵ , each are themselves very interpretable.

The information regarding both structural and sampling costs is typically included in grant proposals seeking research funding, as funding agencies require an explicit budget, part of which is the structural cost and another part is the sampling cost (i.e., cost per datum collected). Thus, the information on the cost required for our method is often estimable before the start of an investigation, and our method does not require more advanced knowledge of the study than would be typical in a grant application, other than expectedly considering accuracy.

Given values for c and A from the preceding, we can obtain a pilot sample size (Step 1 of our procedure) using the *minimum risk for the coefficient of variation function*, namely `mr.cv()`, in the MBESS R package (Version 4.0.0 or greater, Kelley, 2007a, 2007b, 2016). The `mr.cv()` is submitted as follows:

```
mr.cv(pilot=TRUE, A=400000,
      sampling.cost=75, gamma=.49)
```

where, after submitting the code, the function returns

```
Pilot.SS
18.
```

Thus, under this scenario, the pilot sample size is $m = 18$. An alternative way to specify the preceding would be to use the structural cost and epsilon directly (rather than specifying A ; note that with the structural cost set to 1,000 and ϵ set to .05, $A = 1000/.05^2 = 400,000$, as used in the preceding):

```
mr.cv(pilot=TRUE, structural.cost=
      1000, epsilon=.05, sampling.cost=
      75, gamma=.49)
```

which again returns

```
Pilot.SS
18
```

After using `mr.cv()` with `pilot=TRUE` specified and the pilot data collected, the function can then be used to check whether the pilot sample size meets the convergence criterion of the procedure. If the convergence criterion is met, the sampling procedure stops; if the convergence criterion is not met, the procedure continues. We now illustrate this procedure.

After the researcher collects the 18 observations (i.e., the pilot sample), the `mr.cv()` function can be used again, but this time using the data to evaluate the stopping rule. After data are collected, we suggest that a vector of scores be assigned to an object, which we call `Data` in this example, and then the `mr.cv()` function is evaluated as follows:

```
Data <- c(36, 53, 19, 11, 10, 24,
          14, 65, 18, 48, 25, 35, 13, 18, 3,
          41, 5, 3)
mr.cv(data=Data, A=400000, sampling.
      cost=75, gamma=.49)
```

at which point the function returns

```
      Risk      N      cv Is.Satisfied?
[1,] 5964.345  18 0.7391157  FALSE
```

The function provides the value of the risk function, the sample size, the sample value of the coefficient of variation, as well as a check to assess whether the criterion is satisfied (i.e., does `Is.Satisfied?` equal `TRUE` or `FALSE`). In the preceding example, the criterion of our procedure is not satisfied with the collected data (notice the final column of the output).

At this point, due to the criterion not being satisfied, another observation is collected. Although one could collect more than a single observation, the procedure that we describe is based on a single additional datum. The observed value, here a value of 44, is appended onto the existing data (shown below), and the function is submitted again (on the updated data set):

```
Data <- c(Data, 44)
mr.cv(data=Data, A=400000, sampling.
      cost=75, gamma=.49)
```

which returns

```
      Risk      N      cv Is.Satisfied?
[1,] 6224.861  19 0.7113385  FALSE
```

Now, sequentially, another observation is collected and added to the data vector before submitting the `mr.cv()` function. This process continues until the output for `Is.Satisfied` returns `TRUE`, signifying that the optimization criteria has been met. Additional observations are collected one at a time and evaluated with the `mr.cv()` function until the function shows that the criterion is satisfied. For our example, after the 35th observation is collected, the function shows the first instance of the criterion being satisfied (note we type in all of the data here for demonstration purposes):

```
Data <- c(36, 53, 19, 11, 10, 24,
          14, 65, 18, 48, 25, 35, 13, 18, 3, 41,
          5, 3, 44, 26, 13, 39, 2, 3, 26, 22, 8,
          15, 12, 22, 5, 21, 23, 40, 18)
mr.cv(data=Data, A=400000, sampling.
      cost=75, gamma=.49),
```

which returns

```
      Risk      N      cv Is.Satisfied?
[1,] 4891.284  35 0.7013904  TRUE
```

At this point, after the 35th observation is collected, the criterion is satisfied, which can be seen with the last column of the output (specifically where `Is.Satisfied?` is shown to be `TRUE`). We now have formal justification via the stopping rule for the sequential procedure, that with the input specifications chosen and the data that were observed, sampling can stop.

Thus, the total sampling cost for conducting this study in the situation outlined here is $35 \times \$75 = \$2,625$. Recalling the structural cost investment of \$1,000, the study cost were thus \$3,625. This study cost, in which the sample size and thus study cost was unknown a priori, was based on our minimum risk optimization procedure, in which the accuracy of the estimated coefficient of variation and study cost were simultaneously considered. Our approach does not consider just accuracy nor does it consider just cost. By combining these two important aspects of study design in the risk function (i.e., Equation [11]), we sought to minimize the risk function, which, when minimized leads to our stopping rule to be satisfied and thus informs the researcher to stop sampling additional participants.

Characteristics of the final sample size: An empirical demonstration

The procedure we developed for minimizing study cost and sampling error simultaneously for the coefficient of variation has been justified mathematically (see the supplementary material) for large sample sizes. However, an interesting outcome that has no known way to be analytically derived is the distribution of the final sample size, which we will demonstrate under a variety of scenarios. In particular, it is interesting to consider the behavior of the final sample size under different distributions and under small to large sample-size scenarios.

To implement the sequential procedure in this Monte Carlo demonstration, we consider two scenarios. In the first scenario, we assume that the researcher is willing to pay \$200,000 so that the absolute difference between the point estimate of the coefficient of variation, k_n , and the true value, κ , will be, at most, $\epsilon = 0.2$, so $A = \$(200,000/0.2^2)$, and we fix the cost of sampling each unit (e.g., person) to be $c = \$10$. In the other scenario, we assume that the researcher is willing to pay \$500,000 so that the absolute difference between the point estimate of the coefficient of variation, k_n , and the true value, κ , will be, at most, $\epsilon = 0.2$, so $A = \$(500,000/0.2^2)$, and we fix the cost of sampling each unit (e.g., person) in the population to be $c = \$100$. Our example values are meant to show the

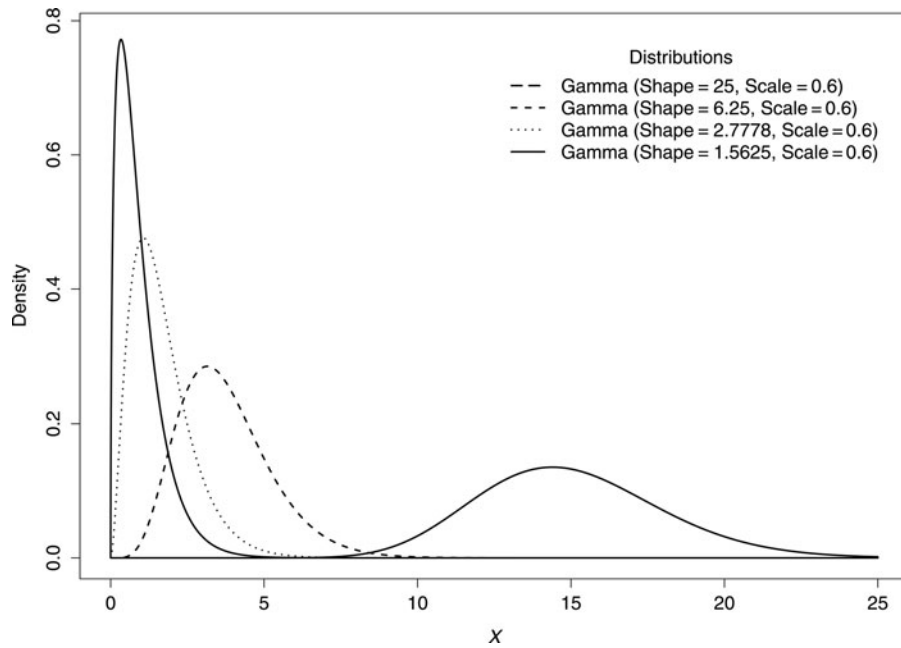


Figure 1. Probability density function of the four gamma distributions used in the simulation study.

flexibility of the method and are scalable to larger/smaller values of structural cost and values of epsilon and the sampling cost; there is nothing special about the values used here other than to illustrate the method in a variety of conditions.

We use $\gamma = 0.49$, as suggested in the previous section, for both scenarios. We compute the pilot sample size by using the pilot sample size formula given in the algorithm from the previous section: $m = \max \{4, \lceil (A/c)^{1/(2+2 \times 0.49)} \rceil\}$. The results are based on random

samples from five different distributions: gamma, log-normal, folded-normal, normal, and Weibull. In all cases, the number of replications used is 5,000. To show the variety of distributions used in our simulation study, we show plots of the gamma distributions, log-normal distributions, folded-normal distributions, normal distributions, and Weibull distributions, respectively, in Figures 1–5.

Tables 1 and 2 present the mean final sample size \bar{N} (estimates $E[N_c]$) from 5,000 replications and the mean risk \bar{r}_N (which estimates $R_{N_c}(\kappa)$) obtained from the

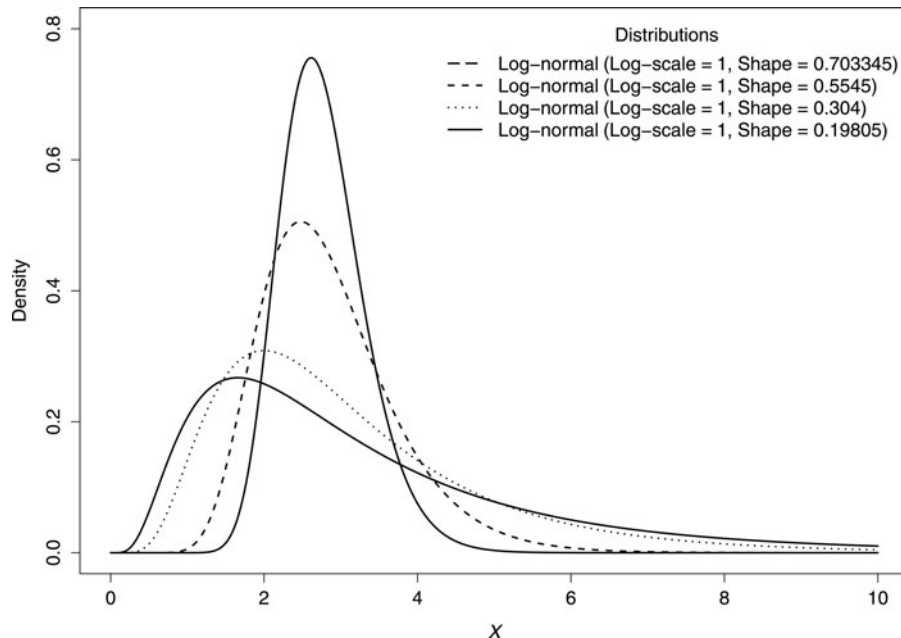


Figure 2. Probability density function of the four log-normal distributions used in the simulation study.

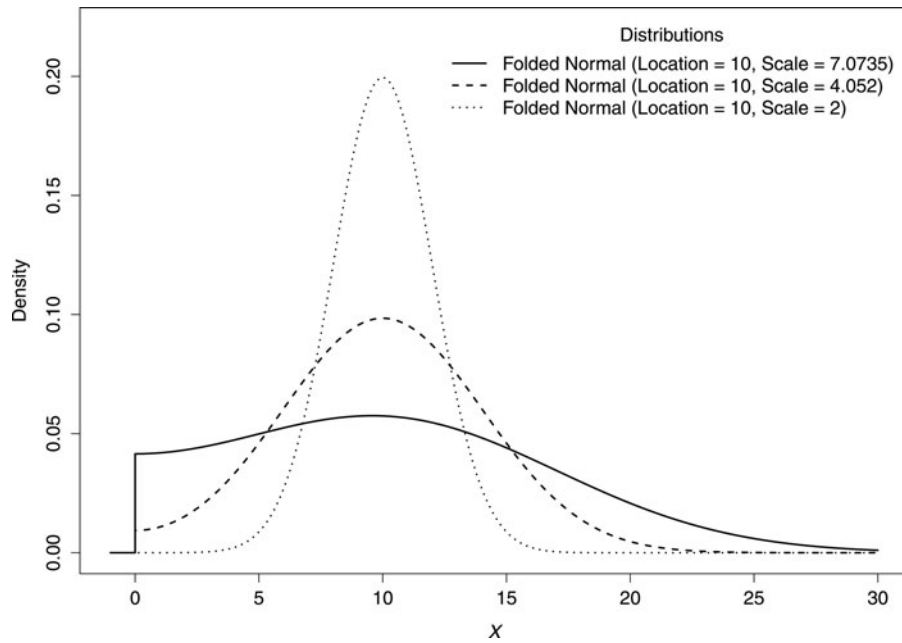


Figure 3. Probability density function of the three folded-normal distributions used in the simulation study.

sample of size N . Moreover, $s(\bar{N})$ and $s(\bar{r}_N)$ represent the standard errors of \bar{N} and \bar{r}_N , respectively. From the fifth column of Table 2, we find that, except for extremely skewed distributions such as the log-normal distribution with parameters 1 and 0.703345 and the log-normal distribution with parameters 1 and 0.5545, the ratio of the average final sample size, \bar{N} , to the optimal sample size, n_c , is close to 1. In all cases, we find that the ratio approaches 1 as sample sizes grow larger. The last column suggests that the ratio of the risk of estimating the coefficient of

variation, using the purely sequential procedure, \bar{r}_N , to the optimal sample size risk, $R_{n_c}^*$, is close to 1. Thus, for not so skewed distributions, our sequential procedure works remarkably well. In fact, except in the extremely skewed cases, the relative cost discrepancy is, at most, about 5%. This implies that the expected cost incurred by our method is almost the same as the optimal sample size risk, $R_{n_c}^*$, defined in Equation (16).

Tables 3 and 4 present the different measures of location (namely, the 0.5 and 99.5 percentiles, the mean, and

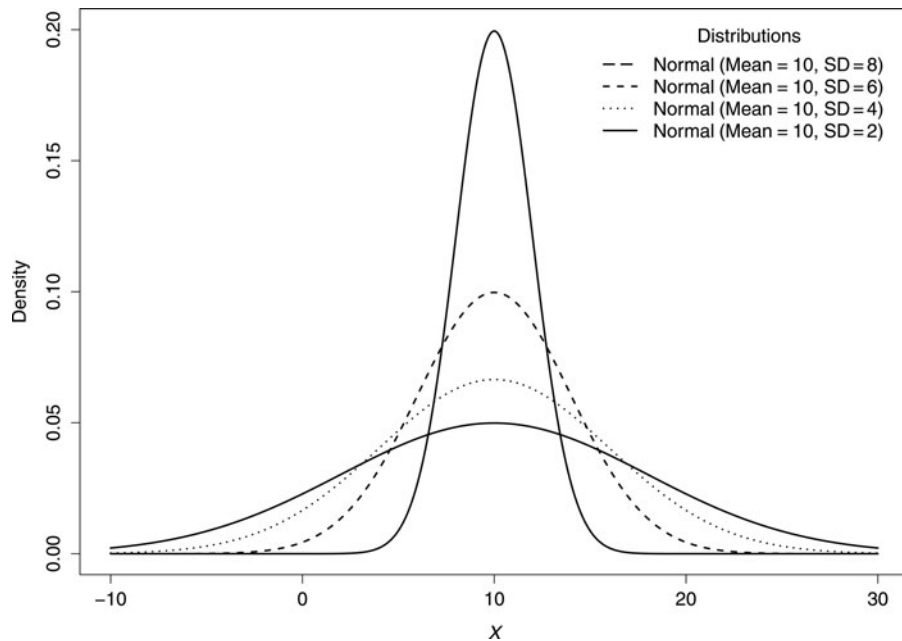


Figure 4. Probability density function of the four normal distributions used in the simulation study.

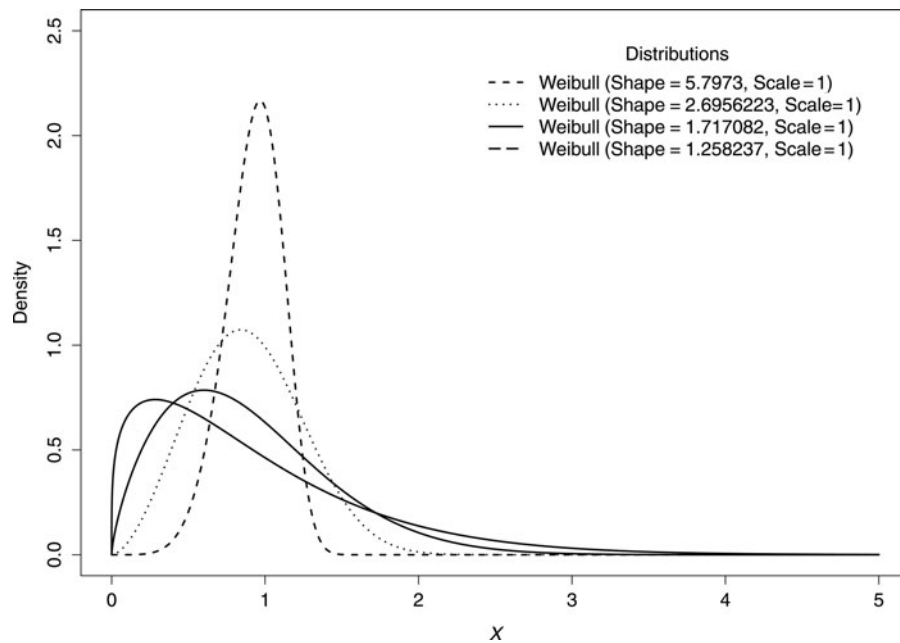


Figure 5. Probability density function of the four Weibull distributions used in the simulation study.

the three quartiles) and the standard deviation of the estimated final sample size. The tables clearly indicate the presence of outliers in the empirical distribution of the estimated final sample size, with the exception of the normal distribution scenarios.

Our procedure is performing very effectively for gamma distributions, normal and folded-normal distributions, Weibull distributions, and, to some extent, log-normal distributions. However, larger values of the coefficients of variation led to wildly different final sample sizes (e.g., in the low 90s for $k = .80$ at the .5 percentile and in the 790s at the 99.5 percentile).⁹ From [Tables 1 and 2](#), we find that our purely sequential procedure works remarkably well as far as the sample size and the cost of estimation are concerned, except in the extremely skewed cases. Statistically, it can be argued that, for extremely skewed distributions, the sample mean and all the estimates of second-, third-, and fourth-central moments are affected by the presence of relatively more extreme observations.

The properties of our sequential method have already been justified mathematically for large samples in a distribution-free environment (i.e., not tied to any particular assumed distribution). [Tables 3 and 4](#) show the

distribution of the final sample size under various scenarios. For some of the log-normal distributions it is apparent that the variability in the final sample size is very large. The stopping rule depends on the estimator of the asymptotic MSE, V_n^2 (i.e., Equation [22]), which further depends on the estimators of coefficient of variation ($= S_n/\bar{X}_n$), skewness ($= \mu_{3n}/\bar{X}_n$), and kurtosis ($= \mu_{4n}/\bar{X}_n$). For highly skewed distributions, large variability in estimates happens with other statistics too, such as with skewness and kurtosis, whose sampling distributions are affected greatly by distributional form. An and Ahmed (2008) studied several kurtosis measures that are widely used in different statistical software (e.g., R, SAS, Stata). They found that for highly skewed or heavily tailed distributions, all widely used estimators of the population kurtosis were substantially underestimated. This is not to say that skewness or kurtosis measures do not generally work well, but rather that they are susceptible to the underlying distribution from which data are sampled. To be clear, our method does not assume any underlying distribution of the data, and thus we do not know whether the distribution is normal or log-normal or gamma or any other distribution. Under that scenario, information about the population coefficient of variation can only be made from data and not based on any presumed distribution characteristics. Correspondingly, the distribution of the final sample size will have different properties under different situations, with some distributions offering a smaller variability in the final sample size than others.

To summarize the performance of our method, for every replication of the simulation for all distributions but the log-normal distribution, the absolute difference

⁹ To help understand why the results, such as the variability of the final sample size, depend on the distribution in a distribution-free environment, we developed an analogy based on a question raised by Will Beasley, who provided an open review of our manuscript. Consider a highway in which there is no speed limit. Some cars are able to go extremely fast, whereas other cars are not. Safety is also a concern. A driver then has to balance his or her safety with an open speed limit. Two drivers starting on the same route would not be expected to complete the drive in the same amount of time (e.g., due to properties of the car, such as top speed, as well as concern for safety). This example has an analog to our distribution free method, in that the properties of our method will depend, even though it is distribution free, on the properties of the population from which observations are sampled.

Table 1. Properties of the distribution of the final sample size when $A = \$(200,000/0.2^2)$ and $c = \$10$.

Distribution	κ	\bar{N} $s(\bar{N})$	n_c	\bar{N}/n_c	\bar{r}_N $s(\bar{r}_N)$	$\frac{\bar{r}_N}{R_{N_c}^* n_c}$
Gamma (shape = 25, scale = 0.6)	0.2	118.3114 0.0660	97	1.2197	1,958.3647 1.7513	1.0095
Gamma (shape = 6.25, scale = 0.6)	0.4	182.3666 0.1598	174	1.0481	3,463.5291 3.5284	0.9953
Gamma (shape = 2.7778, scale = 0.6)	0.6	248.1058 0.2592	246	1.0086	4,854.5673 5.3544	0.98670
Gamma (shape = 1.5625, scale = 0.6)	0.8	374.0366 0.9041	390	0.9591	7,420.1655 18.1623	0.95130
Log-normal (log-scale = 1, shape = 0.19805)	0.2	115.5362 0.0609	93	1.2423	1,883.8360 1.6465	1.0128
Log-normal (log-scale = 1, shape = 0.38525)	0.4	158.2654 0.2056	148	1.0694	2,918.4173 4.8579	0.9860
Log-normal (log-scale = 1, shape = 0.5545)	0.6	220.9894 0.9633	259	0.85324	4,269.8373 19.7387	0.8243
Log-normal (log-scale = 1, shape = 0.703345)	0.8	462.9344 3.8669	740	0.6256	9,188.2414 77.5683	0.6208
Folded-normal (location = 10, scale = 2)	0.2	123.7236 0.0876	104	1.1897	2,098.9917 2.2496	1.0091
Folded-normal (location = 10, scale = 4.052)	0.4	228.2884 0.2033	224	1.0191	4,444.3910 4.2559	0.9921
Folded-normal (location = 10, scale = 7.0735)	0.6	322.2476 0.1935	320	1.0070	6,377.1050 3.9340	0.9964
Normal (Mean = 10, SD = 2)	0.2	123.8528 0.0870	104	1.1908	2,102.2587 2.2357	1.0107
Normal (Mean = 10, SD = 4)	0.4	233.4136 0.2601	230	1.0148	4,550.6635 5.4298	0.9893
Normal (Mean = 10, SD = 6)	0.6	393.9186 0.4454	394	0.9998	7,827.4222 8.9903	0.9933
Normal (Mean = 10, SD = 8)	0.8	602.8416 0.6609	604	0.9981	12,025.4410 13.2456	0.9955
Weibull (shape = 5.7973, scale = 1)	0.2	128.4278 0.1048	111	1.1570	2,217.9524 2.6372	0.9991
Weibull (shape = 2.6956223, scale = 1)	0.4	214.002 0.1630	208	1.0289	4,144.4820 3.4444	0.9963
Weibull (shape = 1.717082, scale = 1)	0.6	277.1446 0.1943	274	1.0115	5,455.0866 3.9952	0.9955
Weibull (shape = 1.258237, scale = 1)	0.8	371.9118 0.6229	378	0.9839	7,380.0198 12.5282	0.9762

Note. κ is the population coefficient of variation; \bar{N} is the mean final sample size; \bar{r}_N is the mean risk; $R_{N_c}(\kappa)$ is the true risk using population parameters; $s(\bar{N})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); n_c is the theoretical sample size if the procedure is used with the population parameters; \bar{r}_N estimates $R_{N_c}(\kappa)$; $s(\bar{r}_N)$ is the standard deviation of the mean estimated risk (i.e., standard error of the risk at the final sample size); tabled values are based on 5,000 replications of a Monte Carlo simulation study.

between the estimated coefficient of variation and the population coefficient of variation was within the fixed value of the specified ϵ . In the case of the log-normal distributions that we considered in Tables 3 and 4, we found that in over 99% of the replications of the simulation, the absolute difference between the estimated coefficient of variation and the population value was within the fixed value of epsilon. We have provided simulation results for illustration purposes only as the method is based on mathematical justification rather than empirical simulation. Nevertheless it is useful to see the effectiveness of our procedure and the properties of various outcomes

(e.g., standard deviation of final sample size) in a variety of situations.

Discussion

The coefficient of variation is a standardized measure of variability defined as the standard deviation divided by the mean. For any given population, the accuracy of the estimated coefficient of variation increases as sampling error decreases. Holding everything else constant, sampling error decreases as the sample size increases. Of course, increasing sample size, in turn, increases the study

Table 2. Properties of the distribution of the final sample size when $A = \$(500,000/0.2^2)$ and $c = \$100$.

Distribution	κ	\bar{N} $s(\bar{N})$	n_c	\bar{N}/n_c	\bar{r}_N $s(\bar{r}_N)$	$\frac{\bar{r}_N}{R_{N_c}^*(\kappa)}$
Gamma (shape = 25, scale = 0.6)	0.2	66.4732 0.0373	49	1.3566	10,072.4981 10.8313	1.0278
Gamma (shape = 6.25, scale = 0.6)	0.4	94.7196 0.10420	87	1.0887	17,233.6881 24.7668	0.9904
Gamma (shape = 2.7778, scale = 0.6)	0.6	125.7062 0.1635	123	1.0220	24,092.2870 35.0619	0.9794
Gamma (shape = 1.5625, scale = 0.6)	0.8	182.9942 0.4796	195	0.9384	35,983.6403 96.9022	0.9227
Log-normal (log-scale = 1, shape = 0.19805)	0.2	65.3924 0.0343	47	1.3913	9,755.6317 10.0950	1.0378
Log-normal (log-scale = 1, shape = 0.38525)	0.4	83.6818 0.1274	74	1.1308	14,497.8467 34.3092	0.9796
Log-normal (log-scale = 1, shape = 0.5545)	0.6	109.1488 0.5081	130	0.8396	20,298.2970 106.6647	0.7807
Log-normal (log-scale = 1, shape = 0.703345)	0.8	201.2408 1.7792	370	0.5439	39,443.1583 358.6241	0.5330
Folded-normal (location = 10, scale = 2)	0.2	68.8436 0.0510	52	1.32392	10,741.5437 14.2890	1.0328
Folded-normal (location = 10, scale = 4.052)	0.4	116.6172 0.1361	112	1.0412	22,158.6498 29.6362	0.9892
Folded-normal (location = 10, scale = 7.0735)	0.6	162.224 0.1363	160	1.0139	31,779.8994 28.1735	0.9931
Normal (Mean = 10, SD = 2)	0.2	68.7838 0.0494	52	1.3228	10,729.6259 13.8814	1.0317
Normal (Mean = 10, SD = 4)	0.4	119.1586 0.1708	115	1.0362	22,700.5080 36.9918	0.9870
Normal (Mean = 10, SD = 6)	0.6	196.4490 0.3070	197	0.9972	38,778.3099 62.5312	0.9842
Normal (Mean = 10, SD = 8)	0.8	300.4782 0.4648	302	0.9950	59,779.6073 93.3771	0.9897
Weibull (shape = 5.7973, scale = 1)	0.2	70.8946 0.0621	56	1.2660	11,307.8055 17.0530	1.0096
Weibull (shape = 2.6956223, scale = 1)	0.4	109.9188 0.1099	104	1.0569	20,697.3890 24.3046	0.9951
Weibull (shape = 1.717082, scale = 1)	0.6	140.1402 0.1359	137	1.0229	27,171.0415 28.6526	0.9916
Weibull (shape = 1.258237, scale = 1)	0.8	184.0418 0.3680	189	0.9738	36,219.3341 74.3809	0.9582

Note. κ is the population coefficient of variation; \bar{N} is the mean final sample size; \bar{r}_N is the mean risk; $R_{N_c}(\kappa)$ is the true risk using population parameters; $s(\bar{N})$ is the standard deviation of the mean final sample size (i.e., standard error of the final sample size); n_c is the theoretical sample size if the procedure is used with the population parameters; \bar{r}_N estimates $R_{N_c}(\kappa)$; $s(\bar{r}_N)$ is the standard deviation of the mean estimated risk (i.e., standard error of the risk at the final sample size); tabled values are based on 5,000 replications of a Monte Carlo simulation study.

cost due to the cost of sampling additional observations (i.e., sampling cost). A fixed-sample-size procedure cannot minimize a function that simultaneously considers both the sampling error and study cost. We have worked in this article to solve this problem for the coefficient of variation.

Unlike “fixed n ” procedures, with sequential methods it is not clear at the beginning of a study what the final sample size will be. This limitation is the consequence of not having to specify one or more generally unknowable population parameters ahead of time. Of course, conducting a study using sequential methods is fundamentally different from saying in a proposal that

“the sample size to be used will be 500.” However, such definitive statements about sample size are based on their own assumptions, such as the distribution shape (usually normality is assumed) and supposed or known population parameters. In general, if the population parameters are not known, they must be supposed in order to obtain the theoretical fixed sample size so that the goals of sample-size planning be met exactly. The problem is that if the supposed values are wrong, then the procedure implied (based on the supposed values) and the theoretically optimal sample size could be very different.

There are five limitations of our method because the method does not directly consider (a) the analysis cost, (b)

Table 3. Summary of locations of the final sample size and standard deviation with $A = \$(200000/0.2^2)$ and $c = \$10$.

Distribution	κ	0.5 Percentile	First quartile	Second quartile	Mean SD	Third quartile	99.5 Percentile
Gamma (shape = 25, scale = 0.6)	0.2	107	115	118	118.3 4.6664	121	131
Gamma (shape = 6.25, scale = 0.6)	0.4	148	175	183	182.4 11.2984	190	210
Gamma (shape = 2.7778, scale = 0.6)	0.6	206	237	247	248.1 18.3272	257	323
Gamma (shape = 1.5625, scale = 0.6)	0.8	291	334	357	374 63.9272	395	650
Log-normal (log-scale = 1, shape = 0.19805)	0.2	104	113	116	115.5 4.3055	118	127
Log-normal (log-scale = 1, shape = 0.38525)	0.4	109	151	160	158.3 14.5397	168	190
Log-normal (log-scale = 1, shape = 0.5545)	0.6	148	189	204	221 68.1180	225	586
Log-normal (log-scale = 1, shape = 0.703345)	0.8	211	289	383	462.9 273.4311	548	1,701
Folded-normal (location = 10, scale = 2)	0.2	110	119	123	123.7 6.1975	128	141
Folded-normal (location = 10, scale = 4.052)	0.4	191	219	229	228.3 14.3741	238	263
Folded-normal (location = 10, scale = 7.0735)	0.6	286	313	323	322.2 13.6797	332	355
Normal (Mean = 10, SD = 2)	0.2	105	120	124	123.9 6.1514	128	141
Normal (Mean = 10, SD = 4)	0.4	167	221	233	233.4 18.3931	246	282
Normal (Mean = 10, SD = 6)	0.6	267	372	393	393.9 31.4961	415	475
Normal (Mean = 10, SD = 8)	0.8	446	571	602	602.8 46.7311	634	726.01
Weibull (shape = 5.7973, scale = 1)	0.2	111	123	128	128.4 7.4124	133	148
Weibull (shape = 2.6956223, scale = 1)	0.4	183	206	214	214 11.5231	222	242
Weibull (shape = 1.717082, scale = 1)	0.6	240	268	278	277.1 13.7408	287	311
Weibull (shape = 1.258237, scale = 1)	0.8	305.9950	344	361	371.9 44.0452	389	564

Note. κ is the population coefficient of variation and SD is the standard deviation; tabled values are based on 5,000 replications of a Monte Carlo simulation study.

economies of scale, (c) the potentially difficult nature of specifying A (or the structural cost and desired ϵ), (d) not knowing the final sample size at the start of the study, and (e) large variability in final sample size for highly skewed distributions. By the *analysis cost* we literally mean the cost incurred for actually performing the analysis at each step of the sequential procedure. However, if one needed to pay for an analysis at each step, our method could be generalized slightly by incorporating analysis cost into c because the cost of collecting an additional observation plus the cost of performing the analysis could be represented in our equations by replacing c with c^* , where c^* is sampling cost plus analysis cost at each step. Therefore, if it costs $c = \$10$ to collect additional data but one is charged $\$50$ for running the analysis (e.g., by the analyst), then the functional cost of adding a single observation is $c^* = 10$

+ 50 = 60 because this is the cost incurred for adding additional data. Thus, this first limitation has a simple solution.

The second limitation we have identified is that there is a fixed cost for sampling regardless of the number of participants. That is, there is no economies-of-scale consideration in our method. By *economies of scale* we mean that we have used a constant c throughout for each participant and thus that there is no reduction (or increase) in cost for sampling larger numbers of participants. Suppose that in some situations the larger the sample size collected the cheaper each observation collected becomes. For example, while partnering with a data collection firm, the first 20 participants might cost $c_1 = \$25$, but thereafter cost $c_2 = \$15$. Or consider that use of a particular assessment cost $c_1 = \$25$ for the first 20 participants and

Table 4. Summary of locations of the final sample size and standard deviation with $A = \$(500000/0.2^2)$ and $c = \$100$.

Distribution	κ	0.5 Percentile	First quartile	Second quartile	Mean SD	Third quartile	99.5 Percentile
Gamma (shape = 25, scale = 0.6)	0.2	60	65	66	66.47 2.6404	68	74
Gamma (shape = 6.25, scale = 0.6)	0.4	72	90	95	94.72 7.3682	100	111
Gamma (shape = 2.7778, scale = 0.6)	0.6	94	119	126	125.7 11.5647	132	168
Gamma (shape = 1.5625, scale = 0.6)	0.8	137	162	173	183 33.9094	194	336
Log-normal (log-scale = 1, shape = 0.19805)	0.2	59	64	65	65.39 2.4237	67	72
Log-normal (log-scale = 1, shape = 0.304)	0.4	52	79	85	83.68 9.0110	90	101
Log-normal (log-scale = 1, shape = 0.5545)	0.6	58	95	103	109.1 35.9304	112	298
Log-normal (log-scale = 1, shape = 0.703345)	0.8	93	128	157	201.2 125.8070	228	791
Folded-normal (location = 10, scale = 2)	0.2	61	66	69	68.84 3.6036	71	79
Folded-normal (location = 10, scale = 4.052)	0.4	92	110	117	116.6 9.6220	123	140
Folded-normal (location = 10, scale = 7.0735)	0.6	126	156	163	162.2 9.6400	169	186
Normal (Mean = 10, SD = 2)	0.2	61	66	69	68.78 3.4927	71	79
Normal (Mean = 10, SD = 4)	0.4	91	111	119	119.2 12.0806	127	152
Normal (Mean = 10, SD = 6)	0.6	141	182	196	196.4 21.7054	211	254
Normal (Mean = 10, SD = 8)	0.8	219	278	301	300.5 32.8679	322	389
Weibull (shape = 5.7973, scale = 1)	0.2	61	68	71	70.89 4.3907	74	84
Weibull (shape = 2.6956223, scale = 1)	0.4	90	104	110	109.9 7.7740	115	129
Weibull (shape = 1.717082, scale = 1)	0.6	114	134	141	140.2 9.6092	147	162
Weibull (shape = 1.258237, scale = 1)	0.8	145	168	178	184 26.01909	192	309

Note. κ is the population coefficient of variation and SD is the standard deviation; tabled values are based on 5,000 replications of a Monte Carlo simulation study.

$c_2 = \$15$ for any additional participants. Our method is not equipped at this time to incorporate fluctuating sampling cost as we have regarded c as a fixed value throughout. One possibility is to approach the problem from the perspective of an expected or average cost per participant, but in so doing one is essentially assuming the final sample size is known. For example, with the current example of $c_1 = \$25$ and $c_2 = \$15$, if we knew that the sample size would be, say, 50, then we could find the average sampling cost, $\bar{c} = (20 \times \$25 + 30 \times \$15)/50 = \$19$. Thus, in such a situation the mean sampling cost, \bar{c} , could replace c in our method. We emphasize, however, that this is likely not a feasible approach because in the sequential framework the final sample size will generally not be known.

The third limitation of specifying A is that there is not much guidance in the literature on how to specify

the structural cost of a research study or the price one is willing to pay for the desired accuracy, that is, achieving the maximum probable error, ϵ . Our approach requires that a researcher specify the structural cost and ϵ (so as to yield A) or specify A directly. Because the idea of A will be new to many researchers, more work on the economics of study design would be useful. The difficulty in determining A is in contrast to ignoring cost, which many studies that consider sample size planning do without an explicit consideration of cost. Because cost is an issue that researchers usually have to contend with, we designed our method with cost as a core part. When one is able to specify the total cost, less sampling, that they are willing to invest in a study, or what can be interpreted as “the price one is willing to pay,” in order to have an estimate with the desired degree of accuracy, then use of

our method (e.g., via the MBESS R package) is straightforward. Not considering the sampling cost, all of the other costs that are needed to implement the study are the structural costs. Although difficult, when budgeting for grants, these costs are considered, as a certain amount of funding will be specified in order to conduct the proposed research. Thus, in this respect, we believe that specifying the structural costs may not be as difficult as it might initially seem, but certainly more discussion of cost considerations in the research design literature would be beneficial.

The fourth limitation of not knowing the final sample size at the start of the study can be mitigated to some extent by using a sensitivity analysis in the sequential framework with supposed distributions and parameters to obtain information about the necessary final sample size the procedure implies. The idea is to use a variety of input parameters and study the distribution of the final sample size. For example, one could suppose that the distribution of scores is normal with a particular value for the population mean and a particular value for the population standard deviation. Then, the `mr.cv()` function could be embedded in a sensitivity analysis in order to obtain a distribution of the final sample size under a variety of plausible scenarios. One would then have low, high, and typical values for what might be an appropriate final sample size (under the variety of conditions specified). One may perform a sensitivity analysis by specifying one or more plausible values for the population coefficient of variation for one or more parent distributions and carry out a large number of replications (e.g., 10,000) to find the typical (or specified percentile) sample size at which the convergence criterion is met. The general framework of sensitivity analyses offers users a great deal of flexibility and provides a great deal of information about how sensitive final sample sizes are in a variety of conditions.

The fifth limitation is that when the distribution from which observations are sampled is highly skewed, like some of the log-normal distributions we considered in the empirical demonstration simulation, the distribution of the final sample size can have high variability due to extreme observations. Thus, even for skewed distributions, in some situations the sequential procedure may yield a mean final sample size from the simulation that is considerably smaller than the theoretically optimal sample size (had parameter values been known), which is one approach to measuring the success of the method. However, under a wide variety of scenarios, our methods produced a mean final sample size that well approximated the theoretically optimal sample size. Although we work in a distribution free environment for the development of the methods, the characteristics of the final sample size will depend on the characteristics of the

population from which data are sampled. This distribution of the final sample size was studied in our empirical investigation, and we found that the distribution can be highly variable for different distributions. Although there were a few highly skewed conditions for the log-normal distribution in which the mean final sample size was considerably smaller than the theoretically optimal sample size, every condition of the simulation showed that in more than 99% of the replications k was within ϵ of κ . That is, the desired degree of accuracy was nearly always satisfied, which, along with considering study cost by minimizing sampling cost, was the goal of the method. Thus, our procedure was shown to work exceptionally well at finding an appropriate sample size such that the parameter of interest was estimated within the desired maximum probable error.

The purely sequential procedure developed here ensures that the ratio of the average final sample size and the theoretically optimal sample size is approximately 1 (meaning that our method recovers approximately the average optimal sample size). In addition, the ratio of the risk function for estimating the coefficient of variation based on the final sample size N_c and the approximate risk function for estimating the coefficient of variation based on the optimal sample size n_c is approximately 1.

We have assumed throughout that the observations are independent and identically distributed (but for arbitrary distributions). Although analytic methods exist for finding confidence intervals for the population coefficient of variation, all such analytic methods assume specific distributions or are approximations. We do not discuss any of these confidence interval methods. We have worked in a distribution-free environment, one in which our methods hold in general and not for a particular type or types of distributions. In our case, after the sampling stops, we recommend a confidence interval be formed. The confidence interval could be one that depends on a particular distribution, or if one wishes to continue in a distribution free environment, a bootstrap procedure could be used to form a confidence interval for the population coefficient of variation.

Ghosh and Sen (1991) argue that sequential procedures are economical in the sense of finding a sample size that reduces study cost while also considering sampling error. The basic theory of sequential analysis is based on the idea of “learn as you go.” Instead of fixing the sample size in advance, the observations are analyzed as they are collected. Fixed-sample-size planning procedures generally depend on values of parameters that are supposed to be true or that are of “minimal importance” (e.g., see Maxwell, Kelley, & Rausch, 2008, for a review). Because of economic consideration and the limitations of many sample size planning procedures with regard to assumed

knowledge of population parameters, there is a natural use and benefit that sequential procedures can have in psychology and related fields.

In this article, we have developed a purely sequential procedure that provides an estimate of the theoretically optimal sample size required to minimize the function containing both the sampling error and study cost without assuming any specific distribution for the data. We focused on the sampling cost by developing a method that, once the structural cost (willingness to pay) and the desired accuracy (ϵ) are specified, the sampling cost could be minimized and done so without assuming any population values for the coefficient of variation. In addition, we developed our work in a distribution-free environment. The lack of any specific distributional assumption is a key piece of our contribution as there is no reason to believe, in many situations, that the distribution of the scores for which the coefficient of variation will be calculated is normal or some other specified distribution. Further, via the MBESS R package, we have provided easy-to-use open source and freely available software to implement the procedures we developed. Given the increased interest in estimating effect sizes in psychology and related fields, we believe that planning studies with the goal of obtaining accurate estimates will continue to increase in importance, which is why our work here on developing a procedure that considers both sampling cost and sampling error for the coefficient of variation seems likely to also grow in importance in psychology and related fields. We hope this article helps to move forward research design in the field by developing a sequential procedure, specifically for the coefficient of variation, but whereupon the ideas can be applied more generally to other effect sizes.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

References

- Abbasi, N., Hemati, S., & Jafarei, A. (2010). Simple Proof of the Theorem: Tending U-Statistics to Central Moments of Sample. *International Journal of Contemporary Mathematical Sciences*, 5(37), 1807–1811. Retrieved from <http://www.m-hikari.com/ijcms-2010/37-40-2010/abbasiIJCMS37-40-2010.pdf>
- Abdi, H. (2010). Coefficient of variation. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 169–171). Thousand Oaks, CA: SAGE Publications, Inc.. doi:10.4135/9781412961288.n56
- Allison, P. D. (1978). Measures of Inequality. *American Sociological Review*, 43, 865–880. Retrieved from <http://www.jstor.org/stable/2094626>
- An, L., & Ahmed, S. E. (2008). Improving the performance of kurtosis estimator. *Computational Statistics & Data Analysis*, 52(5), 2669–2681. doi:10.1016/j.csda.2007.09.024
- Armitage, P. (2014). The evolution of ways of deciding when clinical trials should stop recruiting. *Journal of the Royal Society of Medicine*, 107(1), 34–39. doi:10.1177/0141076813514681
- Bao, Y. (2009). Notes and Problems Finite-Sample Moments of the Coefficient of Variation. *Econometric Theory*, 25, 291–297. doi:10.1017/S0266466608090555
- Bedeian, A. G., & Mossholder, K. W. (2000). On the use of the coefficient of variation as a measure of diversity. *Organizational Research Methods*, 3, 285–297. doi:10.1177/109442810033005
- Ciarleglio, M. M., Arendt, C. D., Makuch, R. W., & Peduzzi, P. N. (2015). Selection of the treatment effect for sample size determination in a superiority clinical trial using a hybrid classical and bayesian procedure. *Contemporary Clinical Trials*, 41, 160–171. doi:10.1016/j.cct.2015.01.002
- Cox, D. R., & Hinkley, D. V. (1979). *Theoretical statistics*. Cambridge, England: CRC Press.
- Dantzig, G. B. (1940). On the Non-Existence of Tests of “Student’s” Hypothesis Having Power Functions Independent of σ . *The Annals of Mathematical Statistics*, 11(2), 186–192. Retrieved from <http://www.jstor.org/stable/2235875>
- De, S. K., & Chattopadhyay, B. (2015). *Minimum Risk Point Estimation of Gini Index*. Retrieved from <http://arxiv.org/abs/1503.08148>
- Durrett, R. (2010). *Probability: Theory and examples*. Cambridge England: Cambridge University Press.
- Freedman, L., & Spiegelhalter, D. (1983). The assessment of the subjective opinion and its use in relation to

- stopping rules for clinical trials. *The Statistician*, 153–160. doi:<http://doi.org/10.2307/2987606>
- Ghosh, B. K., & Sen, P. K. (1991). *Handbook of sequential analysis*. New York NY: CRC Press.
- Ghosh, M., & Mukhopadhyay, N. (1979). Sequential point estimation of the mean when the distribution is unspecified. *Communications in Statistics—Theory and Methods*, 8(7), 637–652. doi: [10.1080/03610927908827789](https://doi.org/10.1080/03610927908827789)
- Harrison, D. A., Price, K. H., & Bell, M. P. (1998). Beyond relational demography: Time and the effects of surface and deep-level diversity on work group cohesion. *Academy of Management*, 41, 96–107. doi:[10.2307/256901](https://doi.org/10.2307/256901)
- Hayashi, R. (2000). Correlation between coefficient of variation of choice reaction time and components of event-related potentials (P300): Effect of benzodiazepine. *Journal of the Neurological Sciences*, 178, 52–56. doi: [10.1016/S0022-510X\(00\)00362-2](https://doi.org/10.1016/S0022-510X(00)00362-2)
- Heffernan, P. M. (1997). Unbiased Estimation of Central Moments by using U-statistics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(4), 861–863. doi:[10.1111/1467-9868.00102](https://doi.org/10.1111/1467-9868.00102)
- Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3), 293–325. Retrieved from <http://www.jstor.org/stable/2235637>
- Hollander, M., & Wolfe, D. A. (1999). *Nonparametric statistical methods*. New York, NY: John Wiley & Sons.
- Kelley, K. (2007a). Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. *Journal of Statistical Software*, 20(8), 1–24. doi:[10.18637/jss.v020.i08](https://doi.org/10.18637/jss.v020.i08)
- Kelley, K. (2007b). Methods for the behavioral, educational, and educational sciences: An R package. *Behavior Research Methods*, 39, 979–984. doi:[10.3758/BF03192993](https://doi.org/10.3758/BF03192993)
- Kelley, K. (2007c). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods*, 39, 755–766. doi: [10.3758/BF03192966](https://doi.org/10.3758/BF03192966)
- Kelley, K. (2016). MBESS 4.0.0 (or greater). [Computer software and manual]. Retrieved from <http://www.cran.r-project.org/>.
- Kelley, K., & Preacher, K. J. (2012). On effect size On effect size. *Psychological Methods*, 17, 137–152. doi: [10.1037/a0028086](https://doi.org/10.1037/a0028086)
- Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed.). London: Charles Griffin & Co.
- Kowalski, J., & Tu, X. M. (2008). *Modern Applied U-statistics*. Hoboken, NJ, John Wiley & Sons.
- Lee, A. J. (1990). *U-statistics: Theory and Practice*. New York, NY: CRC Press.
- Lim, I. S., & Leek, E. C. (2012). Curvature and the visual perception of shape: Theory on information along object boundaries and the minima rule revisited. *Psychological Review*, 119(3), 668. doi:[10.1037/a0025962](https://doi.org/10.1037/a0025962)
- Lord, F. M. (1953). On the Statistical Treatment of Football Numbers. *American Psychologist*. doi: [10.1037/h0063675](https://doi.org/10.1037/h0063675)
- Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, 59, 537–563. doi:[10.1146/annurev.psych.59.103006.093735](https://doi.org/10.1146/annurev.psych.59.103006.093735)
- Mukhopadhyay, N., & Chattopadhyay, B. (2012). A tribute to Frank Anscombe and random central limit theorem from 1952. *Sequential Analysis*, 31(3), 265–277. doi:[10.1080/07474946.2012.694344](https://doi.org/10.1080/07474946.2012.694344)
- Mukhopadhyay, N., & Chattopadhyay, B. (2013). On a new interpretation of the sample variance. *Statistical Papers*, 827–837. doi: [10.1007/s00362-012-0465-y](https://doi.org/10.1007/s00362-012-0465-y)
- Mukhopadhyay, N., & Chattopadhyay, B. (2014). A Note on the Construction of a Sample Variance A note on the construction of a sample variance. *Sri Lankan Journal of Applied Statistics*, 15(1), 71–80. doi:[10.4038/sljastats.v15i1.6795](https://doi.org/10.4038/sljastats.v15i1.6795)
- Mukhopadhyay, N., & De Silva, B. M. (2009). *Sequential methods and their applications*. Boca Raton, FL: CRC Press.
- Nagar, A. (1959). The bias and moment matrix of the general k -class estimators of the parameters in simultaneous equations. *Econometrica: Journal of the Econometric Society*, 575–595. doi:[10.2307/1909352](https://doi.org/10.2307/1909352)
- Ornoy, A., Arnon, J., Shechtman, S., Moerman, L., & Lukashova, I. (1998). Is benzodiazepine use during pregnancy really teratogenic? *Reproductive Toxicology*, 12, 511–515. doi:[10.1016/S0890-6238\(98\)00035-5](https://doi.org/10.1016/S0890-6238(98)00035-5)
- Reed, G. F., Lynn, F., & Meade, B. D. (2002). Use of coefficient of variation in assessing variability of quantitative assays. *Clinical and Diagnostic Laboratory Immunology*, 9(6), 1235–1239. Retrieved from <http://www.jstor.org/stable/25050282>
- Robbins, H. (1959). Sequential estimation of the mean of a normal population. In U. Grenander (Ed.), *In Probability and Statistics (Harold Cramer Volume)* (pp. 235–245). Uppsala, Sweden: Almqvist and Wiksell.
- Rozeboom, W. W. (1966). *Foundations of the theory of prediction*. Homewood, IL: Dorsey Press.
- Sen, P. K. (1981). *Sequential nonparametrics: Invariance principles and statistical inference*. New York, NY: John Wiley & Sons.
- Sen, P. K., & Ghosh, M. (1981). Sequential point estimation of estimable parameters based on U-statistics. *Sankhyā: The Indian Journal of Statistics, Series A*, 331–344. Retrieved from <http://www.jstor.org/stable/25050282>
- Shriberg, L. D., Green, J. R., Campbell, T. F., McSweeney, J. L., & Scheer, A. R. (2003). A diagnostic marker for childhood apraxia of speech: The coefficient of variation ratio. *Clinical Linguistics & Phonetics*, 17, 575–595. doi:[10.1080/0269920031000138141](https://doi.org/10.1080/0269920031000138141)
- Snedecor, G. W. (1956). *Statistical Methods* (5th ed.). Ames, IA: The Iowa State College Press.
- Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. Chichester, England: John Wiley & Sons.
- Stevens, S. S. (1946). On the Theory of Scales of Measurement On the theory of scales of measurement. *Science*, 103(2684), 677–680. Retrieved from <http://www.jstor.org/stable/1671815>
- Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician*, 74, 65–72. doi: [10.1080/00031305.1993.10475938](https://doi.org/10.1080/00031305.1993.10475938)