

The Sensitivity of Sequential Triple Sampling to Departures from Normality

Alan Kimber¹ and Ali Yousef²

¹Southampton Statistical Sciences Research Institute, University of Southampton, Southampton, UK

²School of Mathematics, University of Southampton, Southampton, UK and University of Kuwait

Abstract: In 1981 Hall introduced the sequential triple sampling procedure for point and interval estimation of the mean from a normal population with unknown variance. As its name implies, it involves up to three stages: a pilot stage, a main study stage and a fine tuning stage and it uses a stopping rule that mimics the fixed sample size rule, had the variance been known. The properties of this procedure are well understood in the case of exact normality of the underlying distribution. However, in practice the exact form of the underlying distribution is rarely known. In this paper we present some asymptotic results for the properties of the procedure when the underlying distribution is unknown. More specifically, we consider underlying t and gamma distributions to investigate the robustness of point and interval estimates derived from the sequential triple sampling procedure. Some practical implications are also discussed briefly.

1. Introduction

Let $X_1, X_2, X_3, \dots$ be a sequence of *i.i.d.* random variables from a continuous distribution function $F(\cdot; \mu, \theta)$, where the mean $\mu \in \mathbb{R}$ and the variance $\theta \in \mathbb{R}^+$ are assumed to be unknown but finite. We also assume that the skewness γ and the kurtosis β are finite. The main interest in this study is sequential estimation of μ treating θ as a nuisance parameter.

Having observed a random sample $X_1, X_2, \dots, X_n$, $\forall n \geq 2$ from the distribution function $F(\cdot; \mu, \theta)$, we use the sample mean and the sample variance as point estimates of μ

and θ respectively: $\bar{X}_n = n^{-1} \sum_{i=1}^n X_i$ and $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$, $\forall n \geq 2$.

It is well-known that these estimators are unbiased for their respective parameters and minimal sufficient statistics when the underlying distribution is normal.

Let n^* be the optimal sample size to meet some appropriate inferential criterion had θ been known. In the literature on sequential sampling for inference about the mean for many underlying distributions n^* takes the form

$$(1) \quad n^* = \lambda g(\theta),$$

where λ depends on some predetermined constants. Further, λ is permitted to tend to infinity if $n^* \rightarrow \infty$. Note that $g(\cdot)$ is an appropriate positive real valued twice continuously differentiable function; see Ghosh *et al.* (1997) for details. We begin by defining the triple sampling scheme of Hall (1981), which is the basis for the results in this paper.

Pilot Phase: Here an initial random sample $X_1, X_2, \dots, X_m$ of size $m \geq 2$ is taken from the distribution $F(\cdot; \mu, \theta)$, from which $\bar{X}_m$ and S_m^2 are calculated as initial estimates of μ and θ respectively.

Main Study Phase: In the main study phase a fraction of n^* is estimated during that phase, say, δn^* , where δ is a fixed number between 0 and 1. The sample size required to complete the main study phase is defined by the following stopping rule as

$$(2) \quad N_1^* = \lceil \delta \lambda g(S_m^2) \rceil + 1, \quad N_1 = \max\{m, N_1^*\}.$$

where $0 < \delta < 1$ and $0 < \lambda < \infty$ are known constants and $\lceil x \rceil$ denotes the integer part of x . Observe that in effect N_1 estimates δn^* in this phase.

If $m \geq N_1^*$, stop sampling at this stage; otherwise continue to observe an extra random sample of size $N_1 - m$ from the distribution function $F(\cdot; \mu, \theta)$, say $X_{m+1}, X_{m+2}, \dots, X_{N_1}$. Hence, we augment the $N_1 - m$ observations by the previous m observations and calculate $\bar{X}_{N_1}$ and $S_{N_1}^2$ as new estimates of μ and θ respectively.

Fine Tuning Phase: This is defined according to the following final stage stopping rule

$$(3) \quad N^* = \lambda g(S_{N_1}^2) + 1, \quad N = \max\{N_1, N^*\}.$$

If $N_1 \geq N^*$, then stop at this stage, otherwise continue to sample $N - N_1$ more observations randomly from the distribution $F(\cdot; \mu, \theta)$, say $X_{N_1+1}, X_{N_1+2}, \dots, X_N$. Whenever sampling is terminated and N is realized, then $\bar{X}_N$ is a natural point estimator for μ , and hence $\bar{X}_N$ is a sequential point estimator of μ . Observe that in effect N estimates n^* in this phase.

The usual approach adopted is to assume that the form of F is known (for example, normal or gamma). In many practical situations it is common to assume that the underlying distribution is normal, perhaps for reasons of simplicity, precedent, a central limit theorem argument, or some combination of these. However, it is of interest to investigate the effect of using the normal-based triple sampling scheme when the underlying distribution is misspecified. We shall concentrate on point and interval estimation in this paper.

In section 2 we give some results for triple sampling where the underlying distribution is correctly specified. The effects of misspecification are considered in sections 3 and 4, paying particular attention to the situations when the true underlying distribution is a t distribution (which is useful for understanding the effect of log-tailedness) and when it is a gamma distribution (which aids understanding of the effect of skewness). We conclude with a brief discussion in section 5.

2. Asymptotic Results for Normal-based Triple Sampling for a Normal Underlying Distribution

We consider two situations here: point estimation with a squared error loss function and fixed width confidence intervals with a given nominal coverage. Details of the necessary regularity conditions may be found in Yousef (2010) but in essence we need n^* to increase at a suitably faster rate than m . For simplicity we shall fix $\delta=0.5$, which is a compromise figure recommended by Hall (1981) and others, though this is not crucial to the general argument.

Without loss of generality we define the squared error loss function for an estimator $\hat{\mu}_n$ of μ based upon a sample of size n as:

$$L(\hat{\mu}_n) = A(\hat{\mu}_n - \mu)^2 + n,$$

where A is a positive constant. Thus, $\lambda = \sqrt{A}$ and $g(\theta) = \sqrt{\theta}$.

Consider the case where the underlying distribution is correctly specified as normal. Then under regularity conditions as $m \rightarrow \infty$ $\bar{X}_N$ and N are asymptotically normal with

$$\begin{aligned} E(\bar{X}_N) &\approx \mu \\ \text{Var}(\bar{X}_N) &\approx (\theta / n^*) \{1 + 2.5 / n^*\} \\ E(N) &\approx n^* - 1 \\ \text{Var}(N) &\approx n^* \end{aligned}$$

An extensive simulation study showed that for m around 15 and for n^* bigger than around 25 the first three of these results are very good approximations but the fourth underestimates the true variability of N unless n^* is very large.

Consider now the confidence interval for μ given by $I_n = \bar{X}_N \pm d$ with nominal coverage $1 - \alpha$. Thus $\lambda = a^2 / d^2$, where a is the upper $100\alpha / 2$ percentage point of the standard normal distribution, and $g(\theta) = \theta$. Then (Hall, 1981) it may be shown that

$$P(\mu \in I_n) \approx (1 - \alpha) - (a / n^*) \phi(a) \{a^2 + 4.5\},$$

where $\phi(\cdot)$ is the standard normal density. Clearly the true coverage approaches the nominal coverage from below as n^* grows large. Again, simulations show that this is a good approximation for a nominal 95% coverage even with modest values of n^* .

3. Point Estimation Results of Normal-based Triple Sampling when the Underlying Distribution is Misspecified

Under similar regularity conditions to those used previously but with the additional condition that the first six moments of the true underlying distribution are finite, and with the squared error loss function specified in the previous section, then as $m \rightarrow \infty$ $\bar{X}_N$ and N are asymptotically normal. General expressions for the expected values and variances have been obtained in Yousef (2010), but in particular where the true underlying distribution is t with r degrees of freedom we have that $\bar{X}_N$ and N are asymptotically normal with

$$\begin{aligned} E(\bar{X}_N) &\approx \mu \\ \text{Var}(\bar{X}_N) &\approx (\theta / n^*) \{1 + (5r - 7) / 2(r - 4)n^*\} \\ E(N) &\approx n^* - 1.5(r - 1) / (r - 4) + 0.5 \\ \text{Var}(N) &\approx n^*(r - 1) / (r - 4) \end{aligned}$$

where in this case $\mu = 0$ and $\theta = r / (r - 2)$. In the classical robustness literature (see, for example, Hampel, 1968) a t -distribution (augmented by location and scale parameters) with modest degrees of freedom is often suggested as a more realistic model for “real” data than the normal distribution.

As can be seen from these results the effect of this type of misspecification is to inflate the variance of the estimator for the mean slightly whilst reducing the variability of the sample size. Again, simulation results show these asymptotic results to be reasonable approximations. Broadly the effect of this type of misspecification is relatively minor in the context of point estimation.

To investigate the effect of skewness we consider the case where the underlying distribution is $\chi^2(r)$. Then we obtain

$$\begin{aligned} E(\bar{X}_N) &\approx \mu - \sqrt{2\theta r^{-1}} / n^* \\ \text{Var}(\bar{X}_N) &\approx (\theta / n^*) \{1 + (2.5 + 3 / r) / n^*\} \\ E(N) &\approx n^* - 1 - 9 / r \\ \text{Var}(N) &\approx n^*(1 + 6 / r) \end{aligned}$$

So the effect of skewness is to give a negatively biased estimator of the mean and to inflate its variance. The variance of the sample size is also inflated.

4. Fixed Width Confidence Interval Results for Normal-based Triple Sampling when the Underlying Distribution is Misspecified

The approach adopted in Yousef (2010) to approximate the coverage probability when the distribution is misspecified is to use an Edgeworth expansion for the underlying distribution. Such an approximation will only be valid for modest values of skewness and rules out distributions kurtosis above 5.35 (Draper and Tierney, 1972). However, the approximation is satisfactory for t -distributions with modest degrees of freedom and gamma distributions with modest shape parameters. The expressions for approximate coverage probabilities for a general underlying distribution are long and complicated (see Yousef, 2010), so in the interests of transparency we shall concentrate on intervals with nominal 95% coverage and will compare the cases where the underlying distribution is correctly specified (normal case) and misspecified (t with 10 degrees of freedom, chi-squared with 5 and 10 degrees of freedom). Table 1 shows the approximate coverage values of normal-based triple sampling fixed width confidence intervals for the mean with nominal 95% coverage for selected underlying distributions for large n^* .

Underlying distribution	Approximate coverage
Normal	$0.950 - 0.956 / n^*$
$t(10)$	$0.942 - 1.047 / n^*$
$\chi^2(5)$	$0.978 - 1.486 / n^*$
$\chi^2(10)$	$0.962 - 1.319 / n^*$

Table 1 Approximate coverages of fixed width confidence intervals with nominal 95% coverage and n^* large for some underlying distributions.

These results suggest that, even in the case when the underlying distribution is correctly specified, the true coverage only attains the nominal value from below. This was noted by Hall (1981) who suggested taking additional observations after the triple sampling scheme was completed to improve the coverage properties. The effect of long-tailedness reduces the coverage further, whilst the effect of skewness is less clear cut unless n^* is very large. Once again, simulations indicate that for values on m around 15 or 20, which seem reasonable for practical purpose, these approximations are satisfactory.

5. Discussion

In this paper attention has focussed on triple sampling methods for estimating the mean of a supposedly normal population with unknown variance. Although such a set up is one of the simplest possible met by applied statisticians, it does at least cover a variety of situations that are relevant in practice, such as before and after studies, matched pairs and so on.

There have been many interesting theoretical contributions to the literature on triple sampling, including Hall (1981, 1983), Hamdy (1988), Mukhopadhyay (1990) where the

underlying distribution of the observations is assumed to be correctly specified. In practice, the true underlying distribution is usually at best known only approximately. The work of Yousef (2010), which has been summarised only briefly in the present paper, using a combination of rigorous and heuristic approximations and simulation experiments, has shown that triple sampling procedures that assume a normal underlying distribution can be rather sensitive to departures from normality. This to a large extent mirrors the fixed sample case in which standard normal-based inferential procedures (estimation using sample means and sample variances, t-based confidence intervals etc.) are non-robust to outliers and more general departures from normality (see, for example, Barnett and Lewis, 1994, Hampel *et al.*, 1986, and Huber, 1986). This suggests that if triple sampling is to make the leap from being an approach that presents theoreticians with interesting mathematical problems to a being a useful practical tool for the applied statistician, work is needed to produce robust versions of triple sampling that are little affected by departures from the assumed underlying distribution.

References

- Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data* (3rd edn.). Chichester: Wiley.
- Draper, N.R. and Tierney, D.E. (1972) Regions of positive and unimodal series expansion of the Edgeworth and Gram-Charlier approximations. *Biometrika*, **59**, 463-465.
- Ghosh, M., Mukhopadhyay, N, and Sen, P. (1997) *Sequential Estimation*. New York: Wiley.
- Hall, P. (1981) Asymptotic theory and triple sampling of sequential estimation of a mean. *Ann. Statist.*, **9**, 1229-1238.
- Hall, P. (1983) Sequential estimation saving sampling operations. *J. Roy. Statist. Soc.*, **B45**, 1229-1238.
- Hamdy, H.I. (1988) Remarks on the asymptotic theory of triple stage estimation of the normal mean. *Scand. J. Statist.*, **15**, 303-310.
- Hampel, F.R. (1968) *Contributions to the theory of robust estimators*. PhD thesis, Univ. California, Berkeley.
- Hampel, F.R., Rousseuw, P.J., Ronchetti, E. and Stahel, W. (1986) *Robust Statistics – The Approach Based on Influence Functions*. New York: Wiley.
- Huber, P.J. (1986) *Robust Statistics*. New York: Wiley.
- Mukhopadhyay, N. (1990) Some properties of a three-stage procedure with applications in sequential analysis. *Ind. J. Statist.*, **A52**, 218-231.
- Yousef, A.S.A. (2010) Robustness of triple sampling inference procedures to underlying distributions. PhD thesis, Univ. Southampton.