Two-Step Sequential Sampling for Gamma Distributions

Moors, J.J.A.; Strijbosch, L.W.G.

Publication date:
2002

No. 2002-78

TWO-STEP SEQUENTIAL SAMPLING FOR GAMMA DISTRIBUTIONS

By J.J.A. Moors, L.W.G. Strijbosch

August 2002

ISSN 0924-7815

Discussion paper

TWO-STEP SEQUENTIAL SAMPLING FOR GAMMA DISTRIBUTIONS

J.J.A. Moors

L.W.G. Strijbosch

CDP 2002-78

Abstract

Even after a careful choice of the sample size, the estimates obtained often are unsatisfactory; in particular, their accuracy may turn out to be insufficient. If time and resources allow, a rather usual remedy is to make additional observations. The same situation arises whenever a pilot sample is used to determine the final sample size, as many textbooks advise.

This type of two-step sequential sampling results in a random final sample size, leading to several complications; in particular, most standard estimators will be biased. To illustrate and quantify these complications, gamma distributed populations will be considered here.

Two procedures to acquire an additional sample will be dealt with. The first is based on the observed sample variance: if it is considered too high, the original sample size is *doubled.* The second sample extension procedure concerns independent samples from two populations: the ratio of the observed variances determines the number of additional observations from one of the populations. The main conclusion is that sizable biases remain, even for intermediate sample sizes.

JEL-codes: C13, C15, C42, M41

Key-words: estimation bias, extended sampling, sample extension, stochastic sample size, two-step sampling

1 Introduction

Two-step sampling procedures will be considered, where the first step consists of a random sample of fixed size. An additional sample is drawn *only* if the outcomes satisfy

a given criterion C; if so, the second step observations are combined with the original ones. The resulting final sample size therefore is a random variable.

Generally the random variables $\underline{n}$, $\overline{\underline{y}}$ and $\underline{s}^2$ will denote size, mean and variance of the final sample. The indices 1, 2 and 3 indicate step 1, step 2 and the union of both samples, respectively. Consequently,

$$(\underline{n}, \overline{\underline{y}}, \underline{s}^2) = \begin{cases} (n_1,\ \overline{\underline{y}}_1,\ \underline{s}_1^2) \text{ if } C' \\ (\underline{n}_3,\ \overline{\underline{y}}_3,\ \underline{s}_3^2) \text{ if } C \end{cases} \tag{1.1}$$

with the obvious relations

$$\begin{cases} \underline{n}_3 = n_1 + \underline{n}_2, \quad \overline{\underline{y}}_3 = (n_1 \overline{\underline{y}}_1 + \underline{n}_2 \overline{\underline{y}}_2)/\underline{n}_3 \\ (\underline{n}_3 - 1)\underline{s}_3^2 = (n_1 - 1)\underline{s}_1^2 + (\underline{n}_2 - 1)\underline{s}_2^2 + n_1 \underline{n}_2 (\overline{\underline{y}}_1 - \overline{\underline{y}}_2)^2/\underline{n}_3 \end{cases} \tag{1.2}$$

Note that a certain criterion C is used only once: either step 1 is followed by step 2, or not; that explains our title.

The set-up of this paper is very similar to MOORS & STRIJBOSCH (2000), to be abbreviated to M&S in the sequel, the important difference being that now populations will be assumed to have a gamma distribution - instead of the normal distribution $N(\mu, \sigma^2)$. But the same framework is used, with main features being:

- ◦ simple random sampling (with replacement) is used,
- ◦ the final sample mean $\overline{\underline{y}}$ is the estimator of μ,
- ◦ its accuracy is measured by $\underline{s}^2/\underline{n}$.

Besides, the same two sample extension procedures as in M&S are applied.

In Sections 2 and 3, the extension criterion is $C = \{\underline{s}_1^2 > c\}$ with predetermined constant c; if C occurs, a step 2 sample of - again - size n_1 is observed. So, $\underline{n}$ can take either the value n_1 or $2n_1$. This situation ocurs in practice if unsatisfactory accuracy of $\overline{\underline{y}}_1$ brings the investigator to doubling the original sample.

In Section 4, $\underline{s}_1^2$ is compared with the variance, $\underline{t}_1^2$ say, of an independent size n_1 sample from another population. Only if $C = \{\underline{s}_1^2 > \underline{t}_1^2\}$ occurs, a step 2 sample of random size

$$\underline{n}_2 = \text{entier}[n_1(\underline{s}_1^2/\underline{t}_1^2 - 1)] \tag{1.3}$$

is drawn; in case C' occurs, the sample from the second population is increased comparably. This procedure is relevant when the investigator wants to estimate two population means with about equal accuracy.

For both two-step sampling procedures M&S showed the consequences of applying standard statistical procedures in the case of normally distributed populations. For two reasons, however, this latter assumption does not fully illustrate all pitfalls: the normal distribution is symmetric *and* it has the (unique) property that sample mean and variance are independent. Consequently, the estimator $\underline{\overline{y}}$ for μ for example was still unbiased in M&S. Hence, we wanted to repeat our investigations for other distributions. We chose the class of gamma distributions; for variables taking only positive values, this rich class is a logical choice. Besides, for large shape parameter the gamma distribution resembles a normal one; hence, we are able to compare our new results with M&S. Finally, the nice analytical properties of gamma distributions allow some theoretical derivations.

So, we start with some theoretical derivations in Section 2; we restricted ourselves to a few simple parameter values. They were checked by means of simulations, that gave also results for other parameter values.

To be more precise, for both two-step sampling procedures we present:

- the expectation of the estimator $\underline{\overline{y}}$ for μ,
- variance and Mean Squared Error of $\underline{\overline{y}}$ as accuracy measures,
- the loss of accuracy due to the random final sample size,
- the expectation of the estimator $\underline{s}^2$ for σ^2,
- the expectation of the estimator $\underline{s}^2/\underline{n}$ for $MSE(\underline{\overline{y}})$.

The final Section 5 discusses our results.

2 Extension based on sample variance; $n_1 = 2$

Only in very simple cases we finished the theoretical derivation of all expectations and variances needed. Hence, in this section we will consider only extension criterion $C = \{\underline{s}_1^2 > c\}$, with $n_2 = n_1 = 2$.

Without loss of generality, the scale parameter λ in the gamma distribution $\Gamma(\lambda, k)$ may be taken equal to 1. Full derivations are presented for $k = 1$ (*i.e.* the exponential distribution), as well as the main results for $k = 2$. Since the latter were obtained quite similar, we expect that solutions for all integer-valued k can be obtained.

So, start with $n_1 = 2$ independent observations $\underline{x}$ and $\underline{z}$ from $\Gamma(1, 1) = Ne(1)$;

$\mu = \sigma^2 = 1$. The joint density p of $(\underline{x}, \underline{z})$ is given by

$$p(x,z) = e^{-x-z},\ x > 0,\ z > 0$$

With the transformation

$$\underline{u} = (\underline{x} + \underline{z})/2,\ \underline{w} = \underline{x} - \underline{z}$$

the density q of $(\underline{u}, \underline{w})$ reads

$$q(u,w) = e^{-2u},\ u > |w|/2$$

The (conditional) densities

$$q_1(u) = 4ue^{-2u},\ u > 0, \quad q_2(w|u) = \frac{1}{4u},\ |w| < 2u \tag{2.1}$$

follow immediately.

Denote the sample variance by $\underline{t}$, so that

$$\underline{t} = (\underline{x} - \underline{z})^2/2 = \underline{w}^2/2$$

and

$$P(\underline{t} \leqslant t|u) = 2P(0 < \underline{w} < \sqrt{2t}|u) = \frac{\sqrt{2t}}{2u},\ 0 < t < 2u^2$$

Differentiation and combination with (2.1) then gives the conditional density $g_2(t|u)$ and related densities:

$$g_2(t|u) = \frac{1}{u\sqrt{8t}},\ 0 < t < 2u^2$$

$$g_2(t) = \frac{1}{\sqrt{2t}} e^{-\sqrt{2t}},\ t > 0$$

$$g_1(u|t) = 2e^{\sqrt{2t}-2u},\ 0 < t < 2u^2 \tag{2.2}$$

Direct results are

$$E(\underline{u}|t) = (1 + \sqrt{2t})/2,\ E(\underline{u}^2|t) = (1 + \sqrt{2t} + t)/2 \tag{2.3}$$

and, with the notation $p = e^{-\sqrt{2c}} (= P(C)$ for $k = 1)$,

$$\begin{cases} E(\underline{u}|C)P(C) = \int_c^\infty E(\underline{u}|t)g_1(t)dt = (1 + \sqrt{c/2})p \\ E(\underline{u}^2|C)P(C) = (3 + 2\sqrt{2c} + c)p/2 \end{cases} \tag{2.4}$$

while $E(\underline{t}|C)P(C) = 1 + \sqrt{2c} + c$.

Note the consequences of (2.3) and (2.4):

$$Var(\underline{u}|t) = 1/4,\ Var(\underline{u}|C) = 1/2 = Var(\underline{u})$$

So, surprisingly, the conditional variance of the sample mean $\underline{u}$, given (a lower bound c for) the sample variance $\underline{t}$, is independent of t and c.

Now, we turn to our general notation:

$$\begin{aligned} E(\overline{\underline{y}}) &= E(\overline{\underline{y}}_1|C')P(C') + E(\overline{\underline{y}}_3|C)P(C) \\ &= E(\overline{\underline{y}}_1) + \tfrac{1}{2}E(\overline{\underline{y}}_2 - \overline{\underline{y}}_1|C)P(C) \\ &= 1 - \tfrac{1}{2}(1 - 1 + \sqrt{2c})p \end{aligned}$$

according to (2.4), so that the bias $B(\overline{\underline{y}})$ equals

$$B(\overline{\underline{y}}) = -\tfrac{p}{4}\sqrt{2c} \tag{2.5}$$

Hence, extension criterion $\{\underline{s}_1^2 > c\}$ now leads to a (negatively) biased estimator $\overline{\underline{y}}$, with minimum bias -0.092 (for $c = 1/2$).

Quite similarly, it follows

$$\begin{aligned} E(\overline{\underline{y}}^2) &= E(\overline{\underline{y}}_1^2) + \tfrac{1}{4}E(\overline{\underline{y}}_2^2 + 2\overline{\underline{y}}_1\overline{\underline{y}}2 - 3\overline{\underline{y}}_1^2|C)P(C) \\ &= \tfrac{3}{2} - (2 + 4\sqrt{2c} + 3c)p/8 \end{aligned}$$

leading to the accuracy measures

$$Var(\overline{\underline{y}}) = \tfrac{1}{2} - p(3c + cp + 2)/8$$

and

$$MSE(\overline{\underline{y}}) = \tfrac{1}{2} - p(3c + 2)/8 \tag{2.6}$$

The latter function is increasing in c from 0.25 to 0.5.

Denote the variance of a sample of fixed size $E(\underline{n}) = 2(1+p)$ by $Var^*(\overline{\underline{y}})$. Then the loss of accuracy due to the random sample size equals

$$MSE(\overline{\underline{y}})/Var^*(\overline{\underline{y}}) - 1 = \frac{p}{4}[2 - 3c - (2 + 3c)p] \tag{2.7}$$

This function is positive only for $c < 0.208$; hence, the accuracy of a sample of fixed size may be *improved* in this case by applying the two-step procedure.

Since only high values of $\underline{s}_1^2$ lead to extension of the original sample, it may be expected that $\underline{s}^2$ is negatively biased. Indeed,

$$E(\underline{s}^2) = E(\underline{s}_1^2) + E(\underline{s}_3^2 - \underline{s}_1^2/C)P(C)$$

and use of (1.2) and (2.4) gives

$$B(\underline{s}^2) = -\frac{p}{6}(3c + 4\sqrt{2c}) \tag{2.8}$$

The minimum bias equals -0.351 for $c = 8/9$.

Finally, with the notation $\underline{v} = \underline{s}^2/n$ for the variance estimator, it follows likewise

$$\begin{aligned} E(\underline{v}) &= E(\underline{s}_1^2/2) + E(\underline{s}_3^2/4 - \underline{s}_1^2/2|C)P(C) \\ &= \tfrac{1}{2} + \tfrac{1}{12}E[\underline{s}_2^2 + (\overline{\underline{y}}_1 - \overline{\underline{y}}_2)^2 - 5\underline{s}_1^2|C]P(C) \\ &= \tfrac{1}{2} - \tfrac{p}{24}(9c + 10\sqrt{2c} + 6) \end{aligned}$$

The bias of the variance estimator with respect to $MSE(\overline{\underline{y}})$ therefore equals

$$B(\underline{v}) = -\frac{5p}{12}\sqrt{2c} \tag{2.9}$$

The minimum is -0.145 for $c = 0.5$.

The analysis for $k = 2$ is more tedious, but quite similar. Two independent variables $\underline{x}$, $\underline{z}$ from $\Gamma(1, 2)$ have joint distribution

$$p(x, z) = xze^{-x-z},\ x > 0,\ z > 0$$

Defining $(\underline{u}, \underline{w})$ as before leads to

$$q(u, w) = (u^2 - w^2/4)e^{-2u},\ |w| < 2u$$

and, consequently,

$$q_2(w|u) = \tfrac{3}{8u}[1 - (\tfrac{w}{2u})^2],\ |w| < 2u$$

Introducing $\underline{t} = \underline{w}^2/2$ again gives

$$g_2(t) = \tfrac{1}{2}(1 + \frac{1}{\sqrt{2t}})e^{-\sqrt{2t}},\ t > 0$$

$$g_1(u|t) = \frac{4u^2 - 2t}{1 + \sqrt{2t}}e^{\sqrt{2t}-2u},\ 0 < t < u^2$$

Hence, the conditional expectations

$$E(\underline{u}|t) = \frac{2t + 3\sqrt{2t} + 3}{2(1 + \sqrt{2t})},\ t > 0$$

$$E(\underline{u}^2|t) = \frac{t\sqrt{2t} + 5t + 6\sqrt{2t} + 6}{2(1 + \sqrt{2t})},\ t > 0$$

are obtained, as well as

$$E(\underline{u}|C)P(C) = \tfrac{1}{4}(2c + 5\sqrt{2c} + 8)p$$

$$E(\underline{u}^2|C)P(C) = \tfrac{1}{4}(c\sqrt{2c} + 8c + 14\sqrt{2c} + 20)p$$

The main formulae for the final estimators are gathered in Table 2.1; to facilitate the comparison the previous results for $k = 1$ are presented as well.

Table 2.1. Theoretical results for gamma distributions ($n_1 = 2$)

$\Gamma(1,k)$	$k=1$	$k=2$
$MSE(\overline{\underline{y}})$	$\frac{1}{2} - \frac{p}{8}(3c+2)$	$1 - \frac{p}{16}(3c\sqrt{2c} + 4\sqrt{2c} + 8)$
$B(\overline{\underline{y}})$	$-\frac{p}{4}\sqrt{2c}$	$-\frac{p}{8}(2c + \sqrt{2c})$
$B(\underline{s}^{2)})$	$-\frac{p}{6}(3c + 4\sqrt{2c})$	$-\frac{p}{12}(3c\sqrt{2c} + 16c + 8\sqrt{2c})$
$B(\underline{v})$	$-\frac{5p}{16}\sqrt{2c}$	$-\frac{5p}{16}(2c + \sqrt{2c})$

The numerical outcomes were checked by means of simulation; details are explained in the next section. Appendix A shows that, overall, the agreement between the simulated and theoretical results is quite good.

Figure 2.1. Theoretical biases for gamma and normal distributions ($n_1 = 2$)

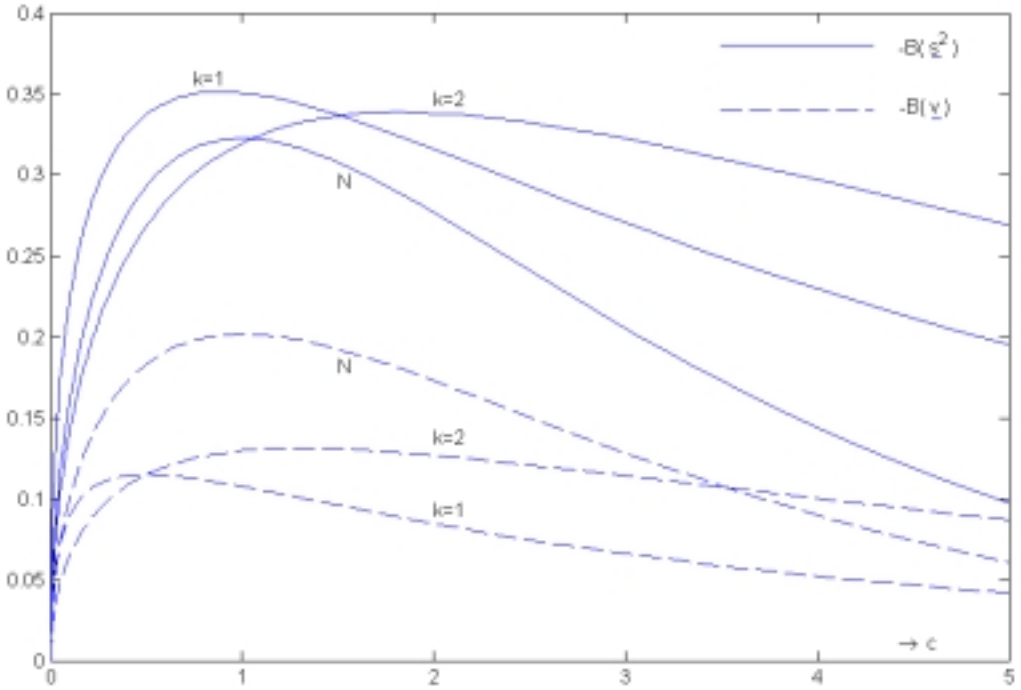


Figure 2.1 shows the biases of $\underline{s}^2$ and $\underline{v}$ from Table 2.1, plus the comparable results of M&S for the normal distribution. To achieve a fair comparison, all three distributions were given variance 1; this implies that the values for $k = 2$ had to be halved. Note that σ^2 and $MSE(\underline{\overline{y}})$ may be underestimated by 35 and 20%, respectively.

3 Extension based on sample variance; general n_1

Since theoretical derivations for $n_1 > 2$ become very tedious, this section is fully based upon simulation results. In principle, the same simulation set-up was used in M&S, which can be summarized as follows. For each value of $k \in \{1, 2, 3, 5, 10, 20\}$ separately, a vector V was drawn consisting of 400,000 random observations from $\Gamma(1, k)$; this vector was used to obtain results for all $n_1 \in \{4, 9, 16, 25\}$. To achieve this, approximately the first half of V was split up into (approximately) 200,000/n_1 initial samples. If a sample satisfied $C = \{s_1^2 > c\}$, then the corresponding second step sample was taken from the remaining part of V. To be precise: in this way 50,000; 22,000; 12,500 and 8,000 final samples were obtained respectively for the four chosen values of n_1. For any pair (k, n_1), $\overline{y}$, s^2 and v were calculated from each final sample: the simulated values of $E(\underline{\overline{y}})$, $Var(\underline{\overline{y}})$, $E(\underline{s}^2)$ and $E(\underline{v})$ followed directly.

To enable a fair comparison of the results for different values of k (and for the normal distribution), we made the population variance σ^2 equal to 1 throughout; furthermore, we wanted to study $c \in \{0.5, 1, 2\}$. That implies that we had to transform our gamma distribution to $\Gamma(\sqrt{k}, k)$. This was achieved by taking $c \in \{k/2,\ k, 2k\}$ in our simulation run, while the final sample mean $\overline{y}$ was divided by $\sqrt{k}$, and both s^2 and v by k.

For the final variance estimator $\underline{v}$, we think the bias relative to the MSE of $\underline{\overline{y}}$ is a more interesting quantity; hence we present the simulated values of the relative bias

$$RB(\underline{v}) = E(\underline{v})/MSE(\underline{\overline{y}}) - 1$$

All corresponding values for the standardnormal distribution can be found in M&S; particularly, their formulae (4.2) and (4.3) give:

$$B(\underline{s}^2) = -\frac{2\nu + 2}{2\nu + 1} c g_\nu(\nu c)$$

$$RB(\underline{v}) = -\frac{6\nu + 4}{2\nu + 1} * \frac{c g_\nu(\nu c)}{1 + G_\nu(\nu c)}$$

Here, $\nu = n_1 - 1$, while g_ν and G_ν denote density and distribution function of the $\mathcal{X}^2_\nu$-distribution. Since $\Gamma(\sqrt{k}, k) - \sqrt{k} \to N(0, 1)$ for $k \to \infty$, this is an extra check

on our results. Figures 3.1-3.3 present the results graphically; the four broken lines in every box correspond to the n_1-values 4 (top) to 25 (bottom). Note that the scale on the horizontal axes is not equidistant! Appendix B shows the calculated biases, but for brevity only for $c = 1$. A discussion of these results is postponed to Section 5.

Figure 3.1. $B(\overline{\underline{y}})$ for distributions $\Gamma(\sqrt{k}, k)$ and $N(0, 1)$

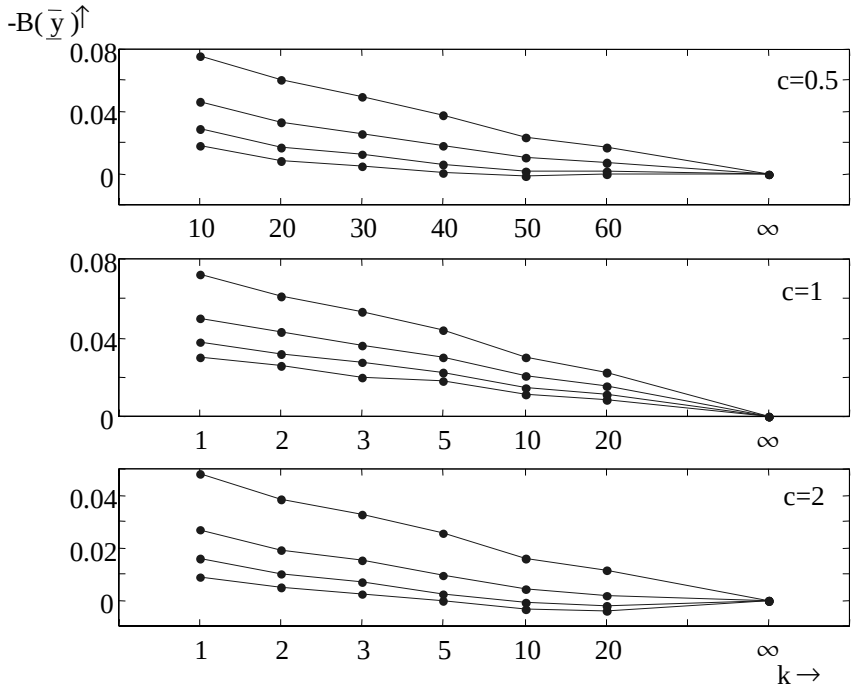


Figure 3.2. $B(\underline{s}^2)$ for distributions $\Gamma(\sqrt{k}, k)$ and $N(0,1)$

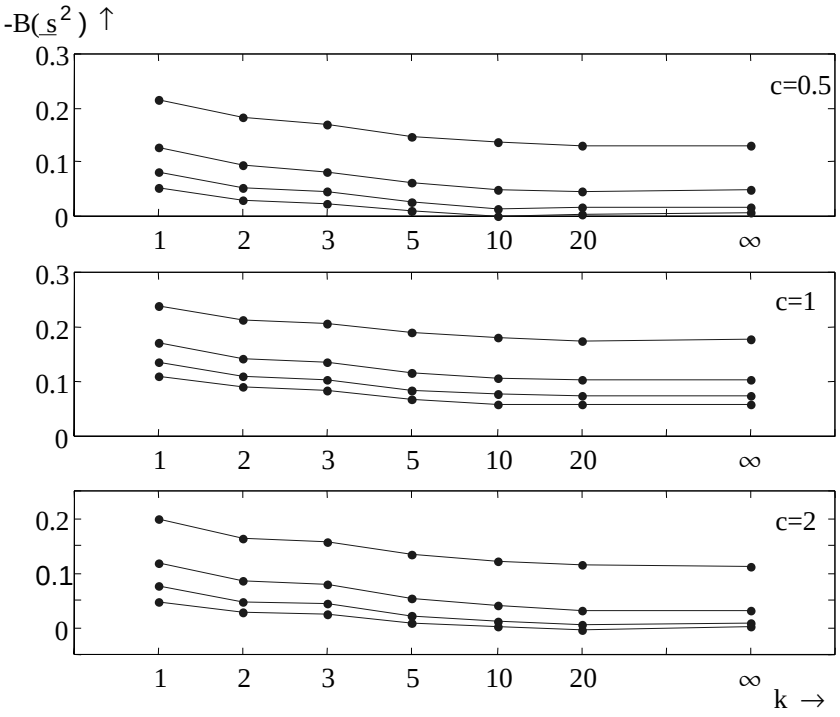
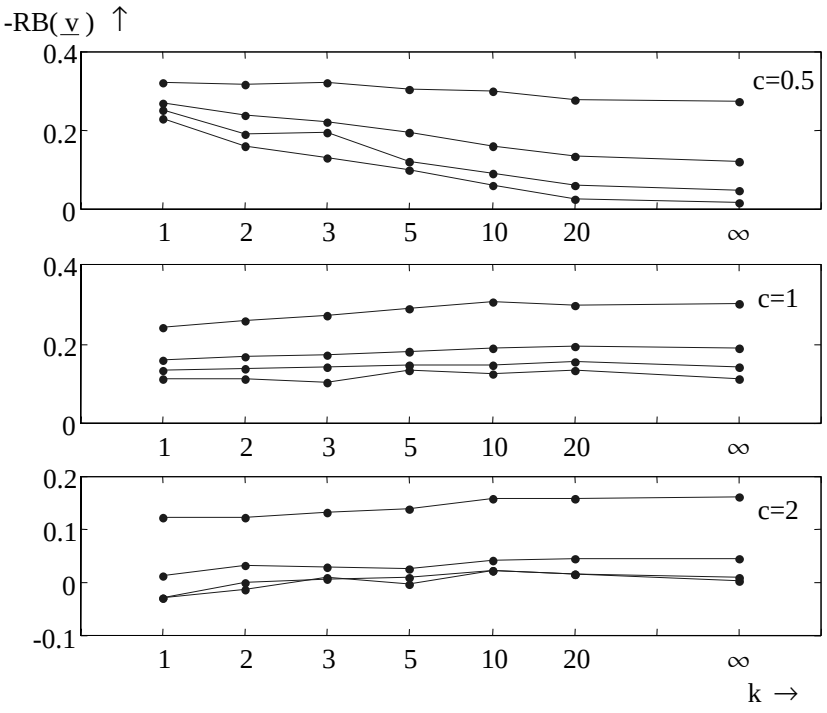


Figure 3.3. $RB(\underline{v})$ for distributions $\Gamma(\sqrt{k}, k)$ and $N(0,1)$



4 Extension based on two sample variances

In this section, two populations are considered, the first having variance k, the second with variance k/τ. Initial size n_1 samples are drawn independently from both populations, leading to sample variances $\underline{s}_1^2$ and $\underline{t}_1^2$, respectively. Since the purpose is to estimate both population means μ_i $(i = 1, 2)$ with about equal accuracy, the extension criterion for the first population now is

$$C = \{\underline{s}_1^2 > \underline{t}_1^2\}.$$

If this event occurs, an additional sample of random size

$$\underline{n}_2 = \text{entier}[n_1(\underline{s}_1^2/\underline{t}_1^2 - 1)]$$

is drawn from population 1. (If C' occurs, the sample from population 2 is extended analogously; we will however concentrate on population 1.)

The final sample of random size

$$\underline{n} = \begin{cases} n_1 & \text{if} \quad C' \\ \text{entier}(n_1\underline{s}_1^2/\underline{t}_1^2) & \text{if} \quad C \end{cases}$$

leads to the estimators $\overline{\underline{y}}$ for μ_1, $\underline{s}^2$ for k and $\underline{v}^2 = \underline{s}^2/\underline{n}$ for $MSE(\overline{\underline{y}})$. Again we are interested in the biases of these standard estimators. For two normal distributions, this problem was studied in M&S, Section 5; here, we consider the case of two gamma-distributions: $\Gamma(1, k)$ and $\Gamma(1, k/\tau)$.

Since theoretical derivations become rather complicated, we restricted ourselves to simulations. For the same values of k and n_1 as in the previous case, final samples from population 1 were generated, according to the above two-step sampling procedure, leading to sample sizes 80,000; 40,000; 25,000 and 20,000 for $n_1 = 4$, 9, 16 and 25, respectively. For practical reasons however, an upperbound of 1000 was used for $\underline{n}$. This hardly influenced our outcomes, since the upperbound was very rarely exceeded for the chosen values of $\tau \in \{0.5, 1, 2\}$.

Figures 4.1-4.3 show the simulated biases of $\overline{\underline{y}}$ and $\underline{s}^2$ as well as the relative bias of $\underline{v}$. For comparison, under the heading ∞ the values for two normal distributions, derived from M&S, are added. The lay-out of the figures in the same as in the previous section: again, the four broken lines represent the results for $n_1 = 4$ (top), 9, 16 and 25 (bottom); note that the horizontal axes are not equidistant. The precise values for $\tau = 1$ are given in Appendix C.

Figure 4.1. $B(\underline{\overline{y}})$ for gamma (and normal) distributions

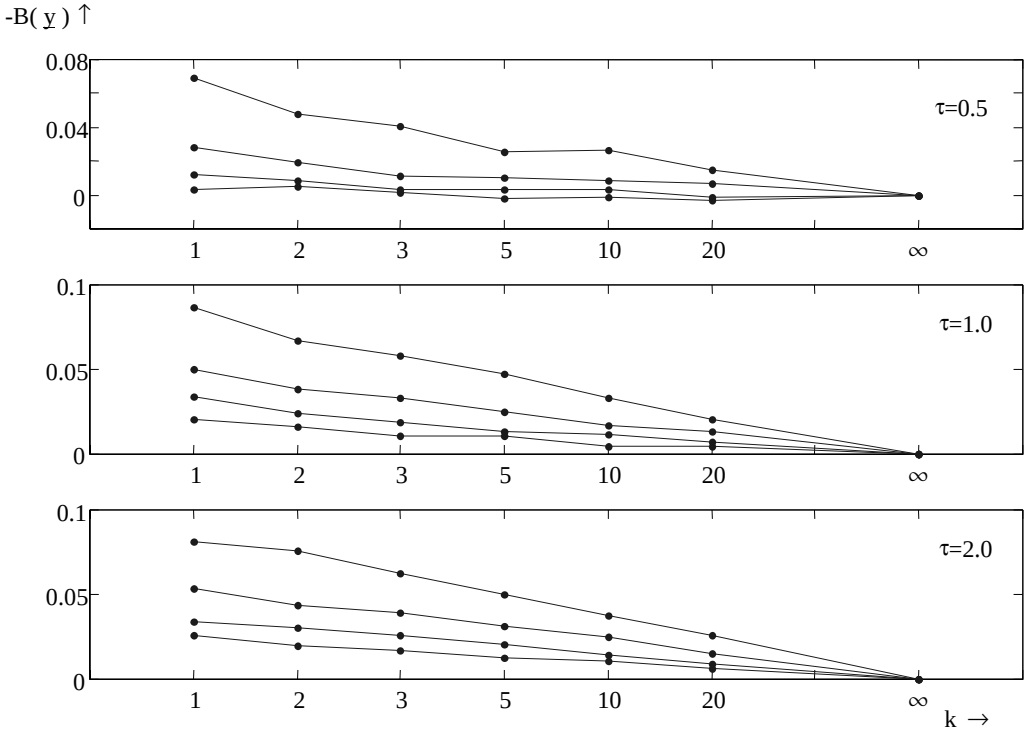


Figure 4.2. $B(\underline{s}^2)$ for gamma (and normal) distributions

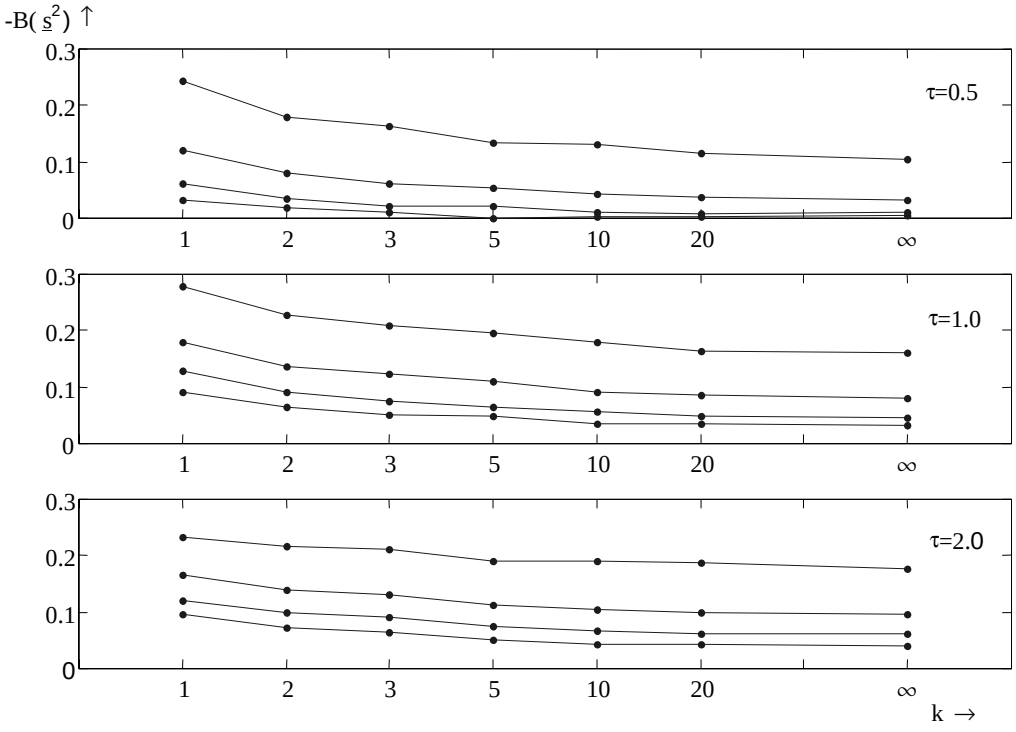
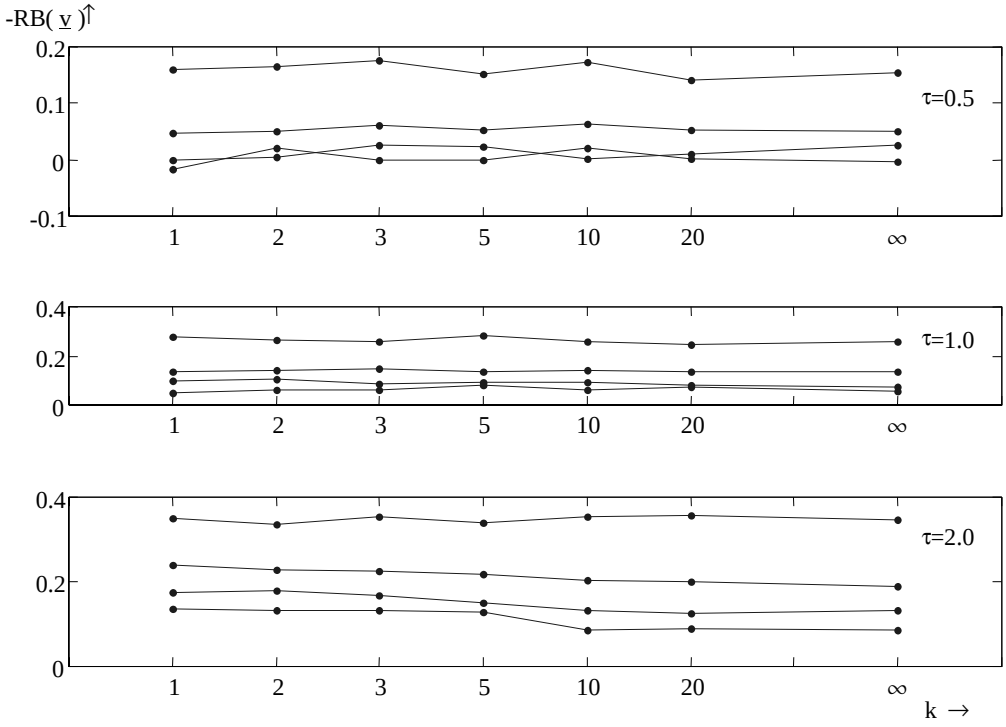


Figure 4.3. $RB(\underline{v})$ for gamma (and normal) distributions



5 Discussion

Stochastic sample sizes are a recurring phenomenon in statistical practice. The relevant literature generally clearly indicates the disadvantages, as the following two quotations from SÄRNDAL *et al.* (1992) illustrate:

> 'BE (Bernoulli) sampling is often considerably less precise that simple random sampling (...) explained by the variablility of the size of the BE sample' (p.55), '(...) there is a nonnegligible loss of precision caused by the lack of control of domain sample size' (p. 397)

Sometimes however, no mention is made of the ensuing problems; e.g. KISH (1965, p. 52) advises:

> 'First, collect a basic sample of reasonable minimum size that might meet the demands. Then compute the results and, if the demands are not met, collect a supplementary sample of desired size. This procedure can be used to obtain either a desired variance or sample size'

An early discussion of specific double sampling procedures was given by STEIN (1945). He described how to obtain a confidence interval of fixed width and confidence level for a normal mean. COX (1952) presented large sample results for general distributions.

To make practitioners aware (once more) of the consequences of stochastic (final) sample sizes, we considered two specific sample extension situations in detail: a one sample case with a fixed second step sample size, and a two sample situation leading to an additional sample of stochastic size. In our previous paper M&S, normal distributions were assumed, for which exact theoretical results could be derived. Since two-step procedures are applied as well in case of unknown population distributions, we felt the need to investigate other distributions too. We chose gamma distributions, allowing theoretical derivations for $n_1 = 2$, at least for the one sample case.

These theoretical results were presented in Section 2. Among them was the curious feature that accuracy may be improved by using the two-step procedure. Note that this phenomenon was found as well in M&S: the accuracy 'loss' in Figure 3.2 can take negative values. The simulation results for general n_1 in Section 3 show that for gamma distributions $\Gamma(\sqrt{k}, k)$ the final sample mean $\underline{\bar{y}}$ is no longer unbiased; biases decrease with increasing k. The bias of the final sample variances $\underline{s}^2$ is more slowly decreasing in k, while the bias of the estimate $\underline{v}$ for MSE $(\underline{\bar{y}})$ is even less dependent of k. As was to be expected, all biases are decreasing in n_1.

For the two sample problem in Section 4 only simulation results were obtained, summarized in Figures 4.1-4.3. The behaviour of the biases here is similar to that of the previous case. Throughout, the results are in accordance with the outcomes for the normal distribution in M&S.

The main numerical conclusion is that for both extension procedures considered here, the relative bias of the estimator $\underline{v}$ for the accuracy of the final estimator $\underline{\bar{y}}$ ranges from about 10% ($n_1 = 25$) up to 30% ($n_1 = 4$). The accuracy of both methods therefore is seriously overestimated by the usual techniques. In a testing situation this may easily lead to incorrect conclusions.

6 References

COX, D.R. (1952), Estimation by double sampling, *Biometrika* **39**, 217-227

KISH, L. (1965), *Survey sampling*, John Wiley and Sons, New York.

MOORS, J.J.A. and L.W.G. STRIJBOSCH (2000), *Two step sequential sampling*, CentER Discussion Paper 2000-39, Tilburg University; to appear in *Statistical Neerlandica*

56

SÄRNDAL, C.-E., B. SWENSSON and J. WRETMAN (1992), *Model assisted survey sampling*, Springer-Verlag, New York.

STEIN, C. (1945), A two-sample test for a linear hypothesis whose power is independent of the variance, *Annals of Mathematical Statistics* **16**, 243-258.

Appendix A

For $n_1 = 2$, the simulations described at the beginning of Section 3 were carried out with $N = 150,000$; besides, they were repeated once, with different seeds. The two sets of simulated values of $E(\underline{\overline{y}})$, $E(\underline{s}_2)$ and $E(\underline{v})$ are presented in Tables A1-A3, respectively. The theoretical values are given as well.

Table A1. Simulated and theoretical values of $E(\underline{\overline{y}})$

	$k=1$			$k=2$		
c/k	true	simulated		true	simulated	
1/2	0.9080	0.9082	0.9080	1.8962	1.8970	1.8981
1	0.9140	0.9140	0.9139	1.8985	1.8991	1.8985
2	0.9323	0.9305	0.9323	1.9200	1.9207	1.9215

Table A2. Simulated and theoretical values of $E(s^2)$

	$k=1$			$k=2$		
c/k	true	simulated		true	simulated	
1/2	0.6628	0.6585	0.6631	1.3607	1.3568	1.3486
1	0.6492	0.6445	0.6517	1.3233	1.3187	1.3154
2	0.6842	0.6757	0.6873	1.4061	1.4028	1.4005

Table A3. Simulated and theoretical values of $E(\underline{v})$

	$k=1$			$k=2$		
c/k	true	simulated		true	simulated	
1/2	0.1858	0.1846	0.1859	0.3822	0.3814	0.3790
1	0.2048	0.2036	0.2054	0.4248	0.4241	0.4225
2	0.2519	0.2494	0.2526	0.5366	0.5372	0.5352

Appendix B

The more extensive simulations for $n_1 > 2$ and general k are reported here, but only for $c/k = 1$. The resulting simulated values of (minus) the biases of $\underline{\overline{y}}$ and $\underline{s}^2$ and (minus) the relative bias of $\underline{v} = \underline{s}^2/\underline{n}$ are presented in the Tables B1-B3. The last columns contain the exact values for the standard normal distribution.

Table B1. Values of $-B(\underline{\overline{y}})$: simulated from $\Gamma(\sqrt{k}, k)$ and exact for $N(0,1)$

k	1	2	3	5	10	20	∞
n_1							
4	0.0719	0.0614	0.0531	0.0436	0.0304	0.0222	0
9	0.0501	0.0428	0.0363	0.0301	0.0202	0.0155	0
16	0.0381	0.0317	0.0277	0.0220	0.0149	0.0115	0
25	0.0298	0.0257	0.0201	0.0177	0.0113	0.0088	0

Table B2. Values of $-B(\underline{s}^2)$: simulated from $\Gamma(\sqrt{k}, k)$ and exact for $N(0,1)$

k	1	2	3	5	10	20	∞
n_1							
4	0.2398	0.2127	0.2060	0.1891	0.1808	0.1742	0.1762
9	0.1698	0.1434	0.1353	0.1169	0.1065	0.1038	0.1034
16	0.1343	0.1110	0.1047	0.0848	0.0765	0.0738	0.0744
25	0.1096	0.0911	0.0832	0.0671	0.0589	0.0580	0.0584

Table B3. Values of $-RB(\underline{v})$: simulated from $\Gamma(\sqrt{k}, k)$ and exact for $N(0,1)$

k	1	2	3	5	10	20	∞
n_1							
4	0.2171	0.2426	0.2594	0.2805	0.3012	0.2952	0.3013
9	0.1329	0.1490	0.1583	0.1704	0.1850	0.1914	0.1907
16	0.1026	0.1169	0.1259	0.1386	0.1444	0.1532	0.1410
25	0.0821	0.0897	0.0894	0.1257	0.1198	0.1316	0.1123

Appendix C

For the extension problem based on two independent sample variances only simulation results are available. They are reported here for the case $\tau = 1$ (equal population variances). The Tables C1-C3 have exactly the same lay-out as those in Appendix B.

Table C1. Values of $-B(\underline{y})$: simulated from $\Gamma(\sqrt{k}, k)$

k	1	2	3	5	10	20	∞
n_1							
4	0.0863	0.0670	0.0581	0.0469	0.0333	0.0209	0
9	0.0498	0.0383	0.0330	0.0248	0.0171	0.0132	0
16	0.0342	0.0245	0.0191	0.0137	0.0112	0.0068	0
25	0.0208	0.0163	0.0105	0.0110	0.0046	0.0043	0

Table C2. Values of $-B(\underline{s}^2)$: simulated from $\Gamma(\sqrt{k}, k)$ and $N(0, 1)$

k	1	2	3	5	10	20	∞
n_1							
4	0.2784	0.2268	0.2092	0.1967	0.1789	0.1646	0.1602
9	0.1807	0.1356	0.1240	0.1089	0.0912	0.0849	0.0793
16	0.1278	0.0924	0.0751	0.0638	0.0562	0.0477	0.0465
25	0.0917	0.0646	0.0508	0.0489	0.0358	0.0338	0.0311

Table C3. Values of $-RB(v)$: simulated from $\Gamma(\sqrt{k}, k)$ and $N(0, 1)$

k	1	2	3	5	10	20	∞
n_1							
4	0.2757	0.2618	0.2572	0.2807	0.2570	0.2467	0.2566
9	0.1376	0.1416	0.1507	0.1370	0.1370	0.1355	0.1331
16	0.0984	0.1062	0.0874	0.0940	0.0940	0.0774	0.0725
25	0.0521	0.0591	0.0615	0.0814	0.0596	0.0740	0.0573