ROCKY MOUNTAIN
JOURNAL OF MATHEMATICS
Volume 1, Number 4, Fall 1971

# FIXED LENGTH CONFIDENCE INTERVALS FOR PARAMETERS OF THE NORMAL DISTRIBUTION BASED ON TWO-STAGE SAMPLING PROCEDURES[1]

L.H. KOOPMANS AND CLIFFORD QUALLS

1. **Introduction and summary.** In many industrial situations the statistician is required to estimate a statistical parameter not only with prescribed confidence or reliability but also with prescribed precision. The most natural procedure is to construct a confidence interval for the parameter for which both the confidence coefficient, $1 - \alpha$, and the length of the interval, $2L$, can be specified in advance.

In this paper fixed length confidence intervals based on two-stage sampling procedures are proposed for the variance and coefficient of variation in the case of a single normal distribution and for the difference in means and ratio and difference of variances in the case of two populations.

The usual one-stage sampling methods do not lead to confidence intervals with both prescribed confidence coefficient and length for any of the parameters we consider. In fact, no one-stage confidence interval can be constructed for any of these parameters which satisfy both requirements. (See, e.g. [1], [2].) The reason for this difficulty can be seen, heuristically, by studying the classical confidence interval for the mean $\mu$ of a normal distribution when the variance $\sigma^2$ is also unknown. The endpoints are $\bar{X} \pm t_\alpha s/\sqrt{n}$ where $\bar{X}$ is the mean of a sample of size $n$, $t_\alpha$ is a percentile of the Student's $t$ distribution, and $s^2$ is the unbiased sample variance. Now our ignorance of the magnitude of $\sigma$ and consequently of its estimate $s$ makes it impossible to select, in advance, a sample size $n$ which will guarantee a prescribed bound on the length of this confidence interval.

In a pioneering paper [7] Stein showed how to overcome this problem by employing two stages of sampling. The first sampling stage is used to obtain an estimate of $\sigma$. If the usual $100(1 - \alpha)\%$ confidence interval (above) computed for the first sample is not short enough to meet the length requirement, a second sample size based on the esti-

Received by the editors November 18, 1969 and, in revised form, February 20, 1970.

*AMS* 1970 *subject classifications.* Primary 62F25.

[1]This research supported in part by National Science Foundation Grant GP-12464.

mate of $\sigma$ is completed which guarantees the satisfaction of both the length and confidence requirements.

A fortuitous property of the normal distribution — the independence of $\bar{X}$ and $s^2$ — makes it possible to do two things in Stein's procedure which are seldom possible in two-stage procedures for other distributions or other parameters of the normal distribution. First, if the usual confidence interval for the parameter computed in the first stage is already short enough, a second sample need not be taken. For the procedures we propose a second sample of size at least one must be taken. Secondly, the information from the first sample can be used to estimate the endpoints of the confidence interval as well as determine the second sample size in Stein's procedure. In the procedures we propose, the sole use made of the first sample is to determine the size of the second sample. This, coupled with the fact that every problem proposed depends on the global inequality $P(A \cap B) \geqq P(A) + P(B) - 1$, leads to the expectation that these two-stage procedures are wasteful of data or inefficient in the traditional statistical sense.

Why then does not one use sequential procedures? This would be the best in many cases. However, in many other important situations items are relatively inexpensive to test but the tests are very time consuming. For example in determining the yield of, say, a new strain of rice, each experimental stage will require at least the time for the growth of the plants to maturity — a period of several weeks. The experimental units, either individual plants or small plots, will be relatively inexpensive compared with the possible economic and sociological cost of delaying the use of an improved product. In yet other situations items are relatively inexpensive to test, but the "set up" costs for each stage of experimentation are high. This is often the case when the experiment requires an expensive laboratory. In the above cases, sample size is not the appropriate measure of loss. The number of sampling stages becomes the important component of the loss function. Now, sequential procedures become quite unattractive and, since at least two stages are needed to guarantee both preassigned confidence and length, procedures with exactly two stages are best.

Though sample size is of secondary importance (when compared to the number of stages) in the cost of experimentation, sample sizes are kept within reason as much as possible in the procedures proposed here by taking advantage of the best one-stage procedures in each stage of sampling.

Before the detailed procedures of this paper are given, two comments are in order. Unfortunately, little guidance is available in selecting the first stage sample sizes, since an optimal selection (in terms of the total sample size of both stages) would depend on the unknown parameters of the distribution. For a discussion of this problem for Stein's procedure, see Seelbinder [6] and Moshman [5]. In terms of controlling total sample size it is probably better to take too large a first sample than one which is too small since the second stage sample size is often relatively sensitive to errors of estimation in the first stage. In lieu of no prior information, one might suggest initial sample sizes of from 25 to 50.

Secondly, note that in the two-stage procedures given here the preselected length of the confidence interval, $2L$, is not used until the calculation of the second stage sample sizes. It is therefore possible after the first stage is completed to adjust $L$ to obtain smaller second stage sample sizes, and still preserve the confidence coefficient. Thus an initially "unrealistic" selection of $L$ can be adjusted. Such a "mixed" scheme destroys the fixed length property, but will prove to be a useful (and necessary) technique in many cases.

For the convenience of potential users, the procedures are given in the body of the paper and all proofs are relegated to a later section.

2. **Confidence interval for the parameters of a single normal distribution.** Let $X_{i1}, X_{i2}, \cdots, X_{in_i}, i = 1, 2$, be independent, identically distributed $N(\mu, \sigma^2)$ random variables. We will use subscripts 1 and 2 on sample sizes and estimators throughout to denote the stage of sampling to which that statistic applies. In all cases, $1 - \alpha$ and $2L$ will denote the preassigned confidence coefficient and confidence interval length. $\bar{X}_i$ and $s_i^2$ will denote the sample mean and variance calculated by

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}, \qquad i = 1, 2.$$

$$s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \quad ,$$

For completeness, Stein's procedure is included.

A. *Confidence interval for* $\mu$ *(Stein's procedure).*

*Stage* 1. A first sample size $n_1$ is selected. (See [5], [6] for guidelines on the selection of the first sample size.) On the basis of the first sample of $n_1$ observations, $\bar{X}_1$ and $s_1^2$ are computed. If $t_\alpha$ denotes the solution of the equation

$$P(t_{n_1-1} \geqq t_\alpha) = \alpha/2,$$

where $t_{n_1-1}$ has Student's $t$ distribution with $n_1 - 1$ degrees of freedom, then the confidence interval

$$\bar{X}_1 - s_1 t_\alpha/\sqrt{n_1} \leqq \mu \leqq \bar{X}_1 + s_1 t_\alpha/\sqrt{n_1}$$

can be used, without further sampling, if

$$s_1 t_\alpha / \sqrt{n_1} \leqq L.$$

If this inequality is not satisfied we must proceed to:

*Stage* 2. A second sample of size $n_2 = n - n_1$ is taken where $n = [s_1^2 t_\alpha^2 / L^2] + 1$. (Here and henceforth, $[x]$ denotes the largest integer strictly smaller than $x$. Thus, for example, $[7.5] = 7$ but $[7] = 6$.)

Now, if $\bar{X}$ denotes the sample mean based on the *total* sample of $n$ observations, the desired $100(1 - \alpha)\%$ confidence interval is

$$\bar{X} - L \leqq \mu \leqq \bar{X} + L.$$

B. *Confidence interval for* $\sigma^2$.

*Stage* 1. For a preselected sample size $n_1$, a $100(1 - \alpha/2)\%$ upper confidence limit for $\sigma^2$ is obtained:

$$\bar{\sigma}_1^2 = (n_1 - 1)s_1^2 / B_{\alpha/2},$$

where $B_{\alpha/2}$ is the solution of

$$P(\chi^2_{n_1 - 1} \leqq B_{\alpha/2}) = \alpha/2,$$

and $\chi^2_{n_1-1}$ has the chi-square distribution with $n_1 - 1$ degrees of freedom.

*Stage* 2. The exact second sample size, $n_2$, is the smallest integer for which

$$P\left( n_2 - \frac{n_2 L}{\bar{\sigma}_1^2} \leqq \chi^2_{n_2 - 1} \leqq n_2 + \frac{n_2 L}{\bar{\sigma}_1^2} \right) \geqq 1 - \frac{\alpha}{2}. \tag{2.1}$$

(Throughout, "exact" is to be interpreted to mean that the prescribed level and confidence coefficient are guaranteed by use of the exact sample size.)

It is shown in [3] that $n_2$ is a well defined random variable and in §4 that the appropriate $100(1 - \alpha)\%$ confidence interval is

$$\frac{n_2 - 1}{n_2} s_2^2 - L \leqq \sigma^2 \leqq \frac{n_2 - 1}{n_2} s_2^2 + L.$$

If $[(n_2 - 1)/n_2] s_2^2 - L < 0$, the lower endpoint of the interval can be replaced by 0.

An approximate but explicit expression for the second sample size $n_2$ based on an application of the central limit theorem is

$$n_2^* = \left[ \frac{2 x_\alpha^2 \bar{\sigma}_1^4}{L^2} \right] + 1, \tag{2.2}$$

where $x_\alpha$ is the solution of the equation

$$\Phi(x) = 1 - \alpha/4$$

and $\Phi$ is the standard normal $N(0, 1)$ distribution function.

The use of the approximate value of $n_2$ may cause the computed confidence interval to have confidence coefficient somewhat smaller than $1 - \alpha$. When this is to be avoided, the explicit expression will lead to a good "first guess" toward computing the exact sample size by means of (2.1).

This procedure can also be used to obtain a $100(1 - \alpha)\%$, length $2\sqrt{L}$ confidence interval for $\sigma$. See reference [3].

C. *Confidence interval for* $\tau = \sigma/\mu$ *(the coefficient of variation) when it is known that* $\mu > c$ *for some constant* $c > 0$. The coefficient of variation, $\tau = \sigma/\mu$, is a useful parameter for measuring the (normalized) variability of a nonnegative random variable when the distribution of such a random variable is approximated by a normal distribution. It is often the case that a positive lower bound, $c$, is known for $\mu$. In such cases, fixed length confidence intervals for $\tau$ can be constructed in two stages of sampling. When no such bound for $\mu$ is known, it can be shown [2] that a purely sequential scheme is required to obtain confidence intervals of prescribed length.

*Stage* 1. On the basis of the first sample of size $n_1$, $100(1 - \alpha/4)\%$ confidence intervals for $\sigma$ and $\mu$ are obtained:

$$0 \leqq \sigma \leqq \bar{\sigma}, \qquad \underline{\mu} \leqq \mu \leqq \bar{\mu},$$

where

$$\bar{\sigma} = \left( \frac{(n_1 - 1)s_1^2}{B_{\alpha/4}} \right)^{1/2},$$

$$\underline{\mu} = \max\left(\bar{X}_1 - \frac{t_{\alpha/4}s_1}{\sqrt{n_1}},\ c\right), \qquad \bar{\mu} = \bar{X}_1 + \frac{t_{\alpha/4}s_1}{\sqrt{n_1}}$$

and where

$$B_{\alpha/4} \text{ satisfies } P(\chi^2_{n_1-1} \leqq B_{\alpha/4}) = \alpha/4$$

and

$$t_{\alpha/4} \text{ satisfies } P(t_{n_1-1} \geqq t_{\alpha/4}) = \alpha/8.$$

*Stage* 2. Let $l_\sigma$ be the unique positive real solution of the cubic equation

$$x^3 + \frac{\bar{\sigma}}{2}x^2 + \frac{\underline{\mu}\bar{\sigma}}{\sqrt{2}}x - \frac{L\underline{\mu}^2\bar{\sigma}}{\sqrt{2}} = 0.$$

As before $2L$ is the preselected length of the confidence interval for $\tau$. Let $l_\mu = l_\sigma^2/\sqrt{2}\bar{\sigma}$. Now the exact second sample size $n_2$ can be computed as the larger of the integers $n'$ and $n''$ where $n' = [k^2\bar{\sigma}^2/l_\mu^2] + 1$ and $k$ is the solution of the equation $\Phi(k) = 1 - \alpha/8$, and $n''$ is the smallest integer satisfying the inequality

$$P\left(n'' - \frac{n'' l_\mu^2}{\bar{\sigma}^2} \leqq \chi^2_{n''-1} \leqq n'' + \frac{n'' l_\mu^2}{\bar{\sigma}^2}\right) \geqq 1 - \frac{\alpha}{2}.$$

Since $l_\sigma$ and $l_\mu$ were selected to make $n'$ and an approximate expression for $n''$ equal, the choice $n_2^* = [k^2\bar{\sigma}^2/l_\mu^2] + 1$ is an approximate, explicit solution for $n_2$.

Now, let

$$\mu_* = \max\,[\underline{\mu}, \bar{X}_2 - l_\mu], \qquad \mu^* = \min\,[\bar{\mu}, \bar{X}_2 + l_\mu],$$

and

$$\sigma_* = \max\,[0, ((n_2 - 1)/n_2)s_2 - l_\sigma],$$

$$\sigma^* = \min\,[\bar{\sigma}, ((n_2 - 1)/n_2)s_2 + l_\sigma].$$

If $\mu_* \leqq \mu^*$ and $\sigma_* \leqq \sigma^*$, set $\underline{\tau} = \sigma_*/\mu^*$, $\bar{\tau} = \sigma^*/\mu_*$. Otherwise, set $\underline{\tau} = 0, \bar{\tau} = 2L$.

Then, $\underline{\tau} \leqq \tau \leqq \bar{\tau}$ is a $100(1 - \alpha)\%$ confidence interval for $\tau$ of (maximum) length $2L$.

3. **Confidence intervals for the parameters of two normal distributions.** Let $X_1, X_2, \cdots; Y_1, Y_2, \cdots$ be independent random variables. The $X_i$'s are all assumed to have the $N(\mu_X, \sigma_X^2)$ distribution and the $Y_i$'s the $N(\mu_Y, \sigma_Y^2)$ distribution. First and second stage sample sizes will be denoted by $n_{X1}, n_{Y1}$ and $n_{X2}, n_{Y2}$ respectively.

A. *Confidence interval for* $\mu_X - \mu_Y$.

*Stage* 1. Upper $100\sqrt{1 - \alpha/2}\%$ confidence limits are obtained for $\sigma_X^2$ and $\sigma_Y^2$ on the basis of the first samples of size $n_{X1}$ and $n_{Y1}$:

$$\bar{\sigma}^2_{X1} = \frac{(n_{X1} - 1)s^2_{X1}}{B_X}, \qquad \bar{\sigma}^2_{Y1} = \frac{(n_{Y1} - 1)s^2_{Y1}}{B_Y}$$

where $B_X$ and $B_Y$ satisfy the equations

$$P(\chi^2_{m_1} \leqq B_X) = 1 - \sqrt{1 - \alpha/2}$$

$$P(\chi^2_{m_2} \leqq B_Y) = 1 - \sqrt{1 - \alpha/2}$$

and

$$m_1 = n_{X1} - 1, \qquad m_2 = n_{Y1} - 1.$$

*Stage* 2. Let $k_\alpha$ be the solution of the equation

$$\Phi(k_\alpha) = 1 - \alpha/4.$$

The second sample sizes $n_{X2}$ and $n_{Y2}$ are selected so as to minimize the cost of the second sample as a whole subject to the restriction

$$k_\alpha \left( \frac{\bar{\sigma}_X^2}{n_{X2}} + \frac{\bar{\sigma}_Y^2}{n_{Y2}} \right)^{1/2} \leqq L. \tag{3.1}$$

If $c_X$ and $c_Y$ are the unit sampling costs for the two populations then the cost of the second sample is

$$c_X n_{X2} + c_Y n_{Y2}.$$

An explicit allocation of sample sizes which minimizes this cost subject to (3.1) is

$$n_{X2} = \left[ \frac{k_\alpha^2}{L^2} \frac{\sqrt{c_X}\bar{\sigma}_X + \sqrt{c_Y}\bar{\sigma}_Y}{\sqrt{c_X}} \bar{\sigma}_X \right] + 1,$$

$$n_{Y2} = \left[ \frac{k_\alpha^2}{L^2} \frac{\sqrt{c_X}\bar{\sigma}_X + \sqrt{c_Y}\bar{\sigma}_Y}{\sqrt{c_Y}} \bar{\sigma}_Y \right] + 1.$$

These sample sizes are exact in the sense previously defined but are possibly larger than the implicitly defined solution. In the case of equal unit sampling costs, $c_X = c_Y$, these expressions become

$$n_{X2} = \left[ \frac{k_\alpha^2}{L^2} \bar{\sigma}_X(\bar{\sigma}_X + \bar{\sigma}_Y) \right] + 1,$$

$$n_{Y2} = \left[ \frac{k_\alpha^2}{L^2} \bar{\sigma}_Y(\bar{\sigma}_X + \bar{\sigma}_Y) \right] + 1.$$

The $100(1 - \alpha)\%$ confidence interval is now

$$\bar{X}_2 - \bar{Y}_2 - L \leqq \mu_X - \mu_Y \leqq \bar{X}_2 - \bar{Y}_2 + L.$$

B. *Confidence interval for* $\sigma_X^2/\sigma_Y^2$.

*Stage* 1. A joint $100(1 - \alpha/2)\%$ confidence region for $\sigma_X^2$ and $\sigma_Y^2$ is obtained based on the initial samples of sizes $n_{X1}$, $n_{Y1}$:

$$\underline{\sigma}_X^2 = \frac{(n_{X1} - 1)s_{X1}^2}{B_{\alpha,X}}, \qquad \bar{\sigma}_X^2 = \frac{(n_{X1} - 1)s_{X1}^2}{A_{\alpha,X}},$$

where $A_{\alpha,X}$, $B_{\alpha,X}$ satisfy

$$P(\chi^2_{n_{X1}-1} \leqq A_{\alpha,X}) = \frac{\sqrt{1 - \alpha/2}}{2},$$

$$P(\chi^2_{n_{X1}-1} \leqq B_{\alpha,X}) = \frac{\sqrt{1 - \alpha/2}}{2}.$$

Upper and lower limits of the confidence interval for $\sigma_Y^2$ are obtained from the same expressions with $X$ replaced by $Y$.

*Stage* 2. The (exact) second stage sample sizes $n_{X2}$, $n_{Y2}$ are again selected to minimize the total cost of the second stage of sampling: Let $n_{X2} = m_1 + 1$, $n_{Y2} = m_2 + 1$. Then, $m_1$ and $m_2$ are to be the integers for which $c_X m_1 + c_Y m_1$ is smallest subject to the condition

$$P(1 - L\underline{\sigma}_Y^2/\bar{\sigma}_X^2 \leqq F_{m_1, m_2} \leqq 1 + L\bar{\sigma}_Y^2/\underline{\sigma}_X^2) \geqq 1 - \alpha/2,$$

where $F_{m_1, m_2}$ has the $F$-distribution with $(m_1, m_2)$ degrees of freedom.

Explicit values of $m_1$ and $m_2$ based on an approximation detailed in the proof are

$$m_1^* = \left[\frac{2\bar{\sigma}_X^4 k_\alpha(\sqrt{c_X} + \sqrt{c_Y})}{L^2\underline{\sigma}_Y^2\sqrt{c_X}}\right] + 1,$$

$$m_2^* = \left[\frac{2\bar{\sigma}_X^4 k_\alpha(\sqrt{c_X} + \sqrt{c_Y})}{L^2\underline{\sigma}_Y^2\sqrt{c_Y}}\right] + 1$$

when, as before, $c_X$ and $c_Y$ are the unit sampling costs and $k_\alpha$ is the solution of the equation

$$\Phi(\beta k_\alpha) - \Phi(-k_\alpha) = 1 - \frac{\alpha}{2}, \qquad \beta = \frac{\bar{\sigma}_X^2\bar{\sigma}_Y^2}{\underline{\sigma}_X^2\underline{\sigma}_Y^2}.$$

In the case $c_X = c_Y$,

$$m_1^* = m_2^* = \left[\frac{4\bar{\sigma}_X^4 k_\alpha}{L^2\underline{\sigma}_Y^4}\right] + 1.$$

A simpler equation for $k_\alpha$, which leads to a somewhat larger value of $k_\alpha$ thus to larger sample sizes, is

$$\Phi(k_\alpha) = 1 - \alpha/4.$$

Finally, the desired $100(1 - \alpha)\%$ confidence interval for $\sigma_X^2/\sigma_Y^2$ is

$$s_{X2}^2/s_{Y2}^2 - L \leqq \sigma_X^2/\sigma_Y^2 \leqq s_{X2}^2/s_{Y2}^2 + L.$$

C. *Confidence interval for* $\sigma_X^2 - \sigma_Y^2$. The procedure of 2.B is used to obtain $100\sqrt{1 - \alpha}\%$ confidence intervals for $\sigma_X^2$ and $\sigma_Y^2$ separately.

*Stage* 1. Let

$$\bar{\sigma}_X^2 = \frac{(n_{X1} - 1)s_{X1}^2}{B_X}, \qquad \bar{\sigma}_Y^2 = \frac{(n_{Y1} - 1)s_{Y1}^2}{B_Y}$$

when $B_X$ is the solution of the equation

$$P(\chi^2_{m_1} \leqq B_X) = \frac{1 - \sqrt{1-\alpha}}{2}$$

and $B_Y$ is the solution of

$$P(\chi^2_{m_2} \leqq B_Y) = \frac{1 - \sqrt{1-\alpha}}{2}$$

for $m_1 = n_{X1} - 1$, $m_2 = n_{Y1} - 1$.

*Stage* 2. The computation of exact sample sizes $n_{X2}$, $n_{Y2}$ to minimize $c_X n_{X2} + c_Y n_{Y2}$ can be carried out as follows: Fix $\gamma$, $0 < \gamma < 1$, and determine the smallest integers $m_1 = m_1(\gamma)$ and $m_2 = m_2(\gamma)$ such that

$$P\left( m_1 - \frac{m_1 \gamma L}{\bar{\sigma}_X{}^2} \leqq \chi^2_{m_1 - 1} \leqq m_1 + \frac{m_1 \gamma L}{\bar{\sigma}_X{}^2} \right) \geqq \frac{1 + \sqrt{1-\alpha}}{2}$$

and

$$P\left( m_2 - \frac{m_2(1-\gamma)L}{\bar{\sigma}_Y{}^2} \leqq \chi^2_{m_2 - 1} \leqq m_2 + \frac{m_2(1-\gamma)L}{\bar{\sigma}_Y{}^2} \right) \geqq \frac{1 + \sqrt{1-\alpha}}{2}.$$

As $\gamma$ varies between 0 and 1, $m_1(\gamma)$ and $m_2(\gamma)$ will vary discontinuously. For some interval of $\gamma$ values, $c_X m_1(\gamma) + c_Y m_2(\gamma)$ will assume its minimum among the possible values it can assume for $0 < \gamma < 1$. If $\gamma^*$ is any value of $\gamma$ in this interval, set

$$n_{X2} = m_1(\gamma^*), \qquad n_{Y2} = m_2(\gamma^*).$$

Approximate but explicit expressions for $n_{X2}$ and $n_{Y2}$ are

$$n^*_{X2} = \left[ \frac{2x_\alpha{}^2 \bar{\sigma}_X{}^4 (u+v)^2}{L^2 u^2} \right] + 1,$$

$$n^*_{Y2} = \left[ \frac{2x_\alpha{}^2 \bar{\sigma}_Y{}^4 (u+v)^2}{L^2 v^2} \right] + 1,$$

where

$$u = (c_X \bar{\sigma}_X{}^4)^{1/3}, \qquad v = (c_Y \bar{\sigma}_Y{}^4)^{1/3}$$

and $x_\alpha$ is the solution of the equation

$$\Phi(x) = \frac{3 + \sqrt{1-\alpha}}{4}.$$

If $c_X = c_Y$,

$$n^*_{X2} = [\bar{\sigma}_X{}^{8/3} D] + 1, \qquad n^*_{Y2} = [\bar{\sigma}_Y{}^{8/3} D] + 1,$$

where $D = 2x_\alpha{}^2(\bar{\sigma}_X{}^{4/3} + \bar{\sigma}_Y{}^{4/3})/L_2$. Finally, the desired $100(1-\alpha)\%$ confidence interval for $\sigma_X{}^2 - \sigma_Y{}^2$ is

$$\frac{n_{X2}-1}{n_{X2}}s_{X2}^2 - \frac{n_{Y2}-1}{n_{Y2}}s_{Y2}^2 - L \leqq \sigma_X{}^2 - \sigma_Y{}^2$$
$$\leqq \frac{n_{X2}-1}{n_{X2}}s_{X2}^2 - \frac{n_{Y2}-1}{n_{Y2}}s_{Y2}^2 + L\,.$$

4. **Proofs.** The proofs for the various procedures will be given using the notation established in the main part of the paper as much as possible. For clarity, it will be necessary to introduce some new notation in places and to emphasize the underlying probability space $(\Omega, \mathcal{B})$ upon which all the random variables are defined. Our notation will follow that of Loève [4]. For simplicity, we will designate the various schemes by their section and letter indices. Thus, Stein's procedure (which we do not reprove) would be designated as 2.A.

2.B. First, the second sample size, the smallest integer satisfying (2.1), is a well defined random variable. The details of the proof of the measurability and finiteness of $n_2(\omega)$ are given in [3]. The proof in [3] is the prototype of the arguments for the remainder of the implicitly defined sample sizes, and we will omit these discussions hereafter.

Next, we show that the confidence interval given in 2.B has the prescribed length and confidence coefficient. Let

$$A_\sigma = [\sigma^2 \leqq \bar{\sigma}_1{}^2] \quad \text{and} \quad B_\sigma = \left[\, \left|\frac{n_2-1}{n_2}\, s_2^2 - \sigma^2\right| \leqq \frac{\sigma^2}{\bar{\sigma}_1{}^2}\, L \,\right].$$

Recall that $s_2{}^2$ depends on $\omega$ both through the second sample size $n_2(\omega)$ and through the second stage random sample.

Now if $\omega \in A_\sigma \cap B_\sigma$, then

$$\left|\frac{n_2-1}{n_2} s_2{}^2(\omega) - \sigma^2\right| \leqq \frac{\sigma^2}{\bar{\sigma}_1{}^2(\omega)} L \leqq L, \quad \text{since} \; \frac{\sigma^2}{\bar{\sigma}_1{}^2(\omega)} \leqq 1.$$

Thus,

$$\left[\, \left|\frac{n_2-1}{n_2}\, s_2{}^2 - \sigma^2\right| \leqq L \,\right] \supset A_\sigma \cap B_\sigma.$$

If $P_{\mu,\sigma}(A)$ denotes the probability of the event $A$ based on the $N(\mu, \sigma^2)$ distribution, we obtain from the well known inequality

$$P(A \cap B) \geqq P(A) + P(B) - 1,$$

the inequalities

$$P_{\mu,\sigma}\left( \left|\frac{n_2-1}{n_2}\, s_2{}^2 - \sigma^2\right| \leqq L \right) \geqq P_{\mu,\sigma}(A_\sigma \cap B_\sigma)$$
$$\geqq P_{\mu,\sigma}(A_\sigma) + P_{\mu,\sigma}(B_\sigma) - 1.$$

But by the construction of $\bar{\sigma}_1^2$,

$$P_{\mu,\sigma}(A_\alpha) = P(\chi^2_{n_1-1} \geqq B_{\alpha/2}) = 1 - \alpha/2.$$

Moreover,

$$\begin{aligned} P_{\mu,\sigma}(B_\sigma) &= P_{\mu,\sigma}\left( \left| \frac{n_2 - 1}{n_2} s_2^2 - \sigma^2 \right| \leqq \frac{\sigma^2}{\bar{\sigma}_1^2} L \right) \\ &= P_{\mu,\sigma}\left( \left| \frac{(n_2 - 1)s_2^2}{\sigma^2} - n_2 \right| \leqq \frac{n_2 L}{\bar{\sigma}_1^2} \right) \\ &= \sum_{n=1}^{\infty} \int_0^{\infty} P_{\mu,\sigma}\left( \left| \frac{(n-1)s_2^2(n)}{\sigma^2} - n \right| \leqq \frac{nL}{\xi} \middle| n_2 = n, \right. \\ &\qquad\qquad \left. \bar{\sigma}_1^2 = \xi \right) P_{\mu,\sigma}(n, d\xi), \end{aligned}$$

where $s_2^2(n)$ is the random variable $s_2^2$ based on a second sample of size $n$ and $P_{\mu,\sigma}(n, \xi) = P_{\mu,\sigma}(n_2 = n, \bar{\sigma}_1^2 \leqq \xi)$.

Now, let $F(n, \xi) = P(|\chi^2_{n-1} - n| \leqq nL/\xi)$ and let $C_n$ be the set of values of $\bar{\sigma}_1^2$ which lead to $n_2 = n$ through the definition of $n_2$. It follows that $F(n, \xi) \geqq 1 - \alpha/2$ for every $\xi \in C_n$ and, since $[n_2 = n] = [\bar{\sigma}_1^2 \in C_n]$, $P_{\mu,\sigma}(n_2 = n, \bar{\sigma}_1^2 \in B) = P_{\mu,\sigma}(n_2 = n, \bar{\sigma}_1^2 \in C_n \cap B)$ for every Borel set $B$. Moreover, since the events $[|(n-1)s_2^2(n)/\sigma^2 - n| \leqq nL/\xi]$ depend only on the observations of the second sample, whereas events of the form $[n_2 = n, \bar{\sigma}_1^2 \in B]$ depend only on the observations from the first sample, the independence of the sample observations implies the independence of the two types of events. It follows that the conditioning on $n_2 = n$ and $\bar{\sigma}_1^2 = \xi$ in the last expression for $P_{\mu,\sigma}(B_\sigma)$ can be omitted.

Furthermore, under the condition that $\mu$ and $\sigma^2$ are the true mean and variance of the observations, $(n-1)s_2^2(n)/\sigma^2$ has the $\chi^2_{n-1}$ distribution. Thus,

$$P_{\mu,\sigma}\left( \left| \frac{(n-1)s_2^2(n)}{\sigma^2} - n \right| \leqq \frac{nL}{\xi} \right) = F(n, \xi).$$

Finally, we obtain

$$\begin{aligned} P_{\mu,\sigma}(B_\sigma) &= \sum_{n=1}^{\infty} \int_{C_n} F(n, \xi) P_{\mu,\sigma}(n, d\xi) \\ &\geqq \left(1 - \frac{\alpha}{2}\right) \sum_{n=1}^{\infty} P_{\mu,\sigma}(n, C_n) \\ &\geqq 1 - \frac{\alpha}{2}, \end{aligned}$$

since $\sum_{n=1}^{\infty} P_{\mu,\sigma}(n, C_n) = P_{\mu,\sigma}(1 \leqq n_2 < \infty) = 1$.

We next indicate the basis for the approximate expression for $n_2$. For even moderate values of $n$ the difference between the distribution of $\chi_n^2$ and $\chi_{n-1}^2$ is negligible. Since $E(\chi_n^2) = n$ and $V(\chi_n^2) = 2n$, the Central Limit Theorem implies that

$$P\left( \left|\frac{\chi_n^2 - n}{\sqrt{2n}}\right| \leqq x \right) \cong 2\Phi(x) - 1.$$

Thus, if $x_\alpha$ is the solution of the equation

$$\Phi(x_\alpha) = 1 - \alpha/4,$$

we will have

$$P\left( \left|\frac{\chi_{n-1}^2 - n}{\sqrt{2n}}\right| \leqq x_\alpha \right) \cong 1 - \frac{\alpha}{2}.$$

But $n_2$ is to be the smallest integer for which

$$P(|\chi_{n-1}^2 - n| \leqq nL/\bar{\sigma}_1^2) \geqq 1 - \alpha/2.$$

Thus, $n_2$ will approximately satisfy the equation

$$P\left( \left|\frac{\chi_{n-1}^2 - n}{\sqrt{2n}}\right| \leqq \frac{\sqrt{n}L}{\sqrt{2}\bar{\sigma}_1^2} \right) = 1 - \frac{\alpha}{2}.$$

This suggests the approximation

$$\frac{\sqrt{n_2^*}L}{\sqrt{2}\sigma_1^2} \cong x_\alpha \quad \text{or} \quad n_2^* = \left[\frac{2x_\alpha^2\bar{\sigma}_1^4}{L^2}\right] + 1.$$

2.C. The confidence interval for $\tau$ given here is based on Theorem 3 of [2], and is given in detail in reference [3].

3.A. The statistics $\bar{\sigma}_{X1}^2$ and $\bar{\sigma}_{Y1}^2$ are independent random variables since each depends on a different sample and they are the standard upper endpoints of $100\sqrt{1 - \alpha/2}\%$ confidence intervals for $\sigma_X^2$ and $\sigma_Y^2$ respectively. If we let

$$A = [\sigma_X^2 \leqq \bar{\sigma}_X^2, \sigma_Y^2 \leqq \bar{\sigma}_Y^2]$$

and $\Delta = (\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2)$, it follows that $P_\Delta(A) = 1 - \alpha/2$.

Let $k_\alpha$, $n_{X2}$ and $n_{Y2}$ be selected as described in the text and take

$$B = \left[ |\bar{X}_2 - \bar{Y}_2 - (\mu_X - \mu_Y)| \leqq k_\alpha \left(\frac{\sigma_X^2}{n_{X2}} + \frac{\sigma_Y^2}{n_{Y2}}\right)^{1/2}\right].$$

Then

$$A \cap B \subset C = [|\bar{X}_2 - \bar{Y}_2 - (\mu_X - \mu_Y)| \leqq L],$$

since, by definition of $n_{X2}$ and $n_{Y2}$, if $\omega \in A \cap B$ then

$$|\bar{X}_2(\omega) - \bar{Y}_2(\omega) - (\mu_X - \mu_Y)| \leqq k_\alpha \left( \frac{\sigma_X^2}{n_{X2}(\omega)} + \frac{\sigma_Y^2}{n_{Y2}(\omega)} \right)^{1/2}$$
$$\leqq k_\alpha \left( \frac{\bar{\sigma}_X^2(\omega)}{n_{X2}(\omega)} + \frac{\bar{\sigma}_Y^2(\omega)}{n_{Y2}(\omega)} \right)^{1/2} \leqq L.$$

But, if $\bar{X}_{2j}$ and $\bar{Y}_{2\ell}$ denote the second stage sample means based on sample sizes of $j$ and $\ell$ respectively, then

$$P_\Delta(B) = \sum_{j=1}^{\infty} \sum_{\ell=1}^{\infty} P_\Delta\Bigg(|\bar{X}_{2j} - \bar{Y}_{2\ell} - (\mu_X - \mu_Y)|$$
$$\leqq k_\alpha \left( \frac{\sigma_X^2}{j} + \frac{\sigma_Y^2}{\ell} \right)^{1/2} \Bigg| n_{X2} = j, n_{Y2} = \ell \Bigg) \tag{3.A.1}$$
$$\times P_\Delta(n_{X2} = j, n_{Y2} = \ell).$$

By the independence of the first and second stage samples the conditioning in the last expression can be omitted as in the proof of 2.B. Furthermore, when $\Delta$ is the "true" vector of parameter values, $\bar{X}_{2j} - \bar{Y}_{2\ell}$ is normally distributed with mean $\mu_X - \mu_Y$ and variance $\sigma_X^2/j + \sigma_Y^2/\ell$. Thus,

$$P_\Delta\left( |\bar{X}_{2j} - \bar{Y}_{2\ell} - (\mu_X - \mu_Y)| \leqq k_\alpha \left( \frac{\sigma_X^2}{j} + \frac{\sigma_Y^2}{\ell} \right)^{1/2} \right) = 2\Phi(k_\alpha) - 1$$
$$= 1 - \frac{\alpha}{2}$$

for all $j, k \geqq 1$. It follows from (3.A.1) that

$$P_\Delta(B) = \sum_{j=1}^{\infty} \sum_{\ell=1}^{\infty} \left( 1 - \frac{\alpha}{2} \right) P_\Delta(n_{X2} = j, n_{Y2} = \ell) = 1 - \frac{\alpha}{2}.$$

Finally, $P_\Delta(C) \geqq P_\Delta(A \cap B) \geqq P_\Delta(A) + P_\Delta(B) - 1 = 2(1 - \alpha/2) - 1 = 1 - \alpha$. Thus, the confidence interval has the desired length and confidence coefficient.

The explicit expressions for the second sample sizes were obtained by replacing the original minimization problem by the continuous version in which $x$ and $y$ are to be found which minimize $c_X x + c_Y y$ subject to the restriction

$$k_\alpha \left( \frac{\bar{\sigma}_X^2}{x} + \frac{\bar{\sigma}_Y^2}{y} \right)^{1/2} = L.$$

The Lagrange multiplier method readily leads to the solution

$$x = \frac{k_\alpha^2}{L^2}\frac{\bar{\sigma}_X}{\sqrt{c_X}}(\sqrt{c_X}\bar{\sigma}_X + \sqrt{c_Y}\bar{\sigma}_Y), \quad y = \frac{k_\alpha^2}{L^2}\frac{\bar{\sigma}_Y}{\sqrt{c_Y}}(\sqrt{c_X}\bar{\sigma}_X + \sqrt{c_Y}\bar{\sigma}_Y).$$

The integer values of $n_{X2}$, $n_{Y2}$ given in the text, which are obtained by taking the smallest integers larger than $x$ and $y$, can be no smaller than the optimal integer solution. Consequently, both length and confidence coefficient specifications are met by the explicit solution.

3.B. The proof that the confidence interval given in the text has the correct confidence coefficient follows closely the pattern of the proofs of 2.B and 3.A with

$$A = [\underline{\sigma}_X^2 \leqq \sigma_X^2 \leqq \bar{\sigma}_X^2, \underline{\sigma}_Y^2 \leqq \sigma_Y^2 \leqq \bar{\sigma}_Y^2],$$

$$B = \left[\frac{\sigma_X^2}{\sigma_Y^2} - L\frac{\sigma_X^2}{\sigma_Y^2}\frac{\underline{\sigma}_Y^2}{\bar{\sigma}_X^2} \leqq \frac{s_{X2}^2}{s_{Y2}^2} \leqq \frac{\sigma_X^2}{\sigma_Y^2} + L\frac{\sigma_X^2}{\sigma_Y^2}\frac{\bar{\sigma}_Y^2}{\underline{\sigma}_X^2}\right],$$

and

$$C = \left[\frac{\sigma_X^2}{\sigma_Y^2} - L \leqq \frac{s_{X2}^2}{s_{Y2}^2} \leqq \frac{\sigma_X^2}{\sigma_Y^2} + L\right] = \left[\frac{s_{X2}^2}{s_{Y2}^2} - L \leqq \frac{\sigma_X^2}{\sigma_Y^2} \leqq \frac{s_{X2}^2}{s_{Y2}^2} + L\right].$$

We omit this proof.

The explicit expressions for the second stage sample sizes were obtained by the following argument. The given minimization problem is equivalent to minimizing $m_2 + Km_1$ subject to the constraint

$$P(1 - a \leqq F_{m_1, m_2} \leqq 1 + b) \geqq 1 - \alpha/2$$

where $K = c_X/c_Y$, $a = L(\underline{\sigma}_Y^2/\bar{\sigma}_X^2)$ and $b = L(\bar{\sigma}_Y^2/\underline{\sigma}_X^2)$. If we allow $L$ to approach 0, necessarily $m_1$ and $m_2$ tend to infinity. For large $m_1$ and $m_2$, the distribution of $F_{m_1, m_2}$ is very nearly the same as the distribution of the ratio of independent normal random variables by the Central Limit Theorem:

$$P(1 - a \leqq F_{m_1, m_2} \leqq 1 + b) \cong P\left(1 - a \leqq \frac{\sqrt{2m_1}W_1 + m_1}{\sqrt{2m_2}W_2 + m_2} \leqq 1 + b\right)$$

where $W_1$ and $W_2$ are independent $N(0, 1)$. By a linear change of variables from $W_1$, $W_2$ to $U_1$, $U_2$ this last probability can be written in the form

$$P(U_1 \leqq c, U_2 \leqq d)$$

where $U_1$ and $U_2$ are bivariate $N(0, 1)$ with correlation coefficient

$$\rho = -\{(1-a)^2 + r\}^{-1/2}\{(1+b)^2 + r\}^{-1/2}\{(1-a)(1+b) + r\},$$

and $r = m_2/m_1$. The quantities $c$ and $d$ are $c = a/\sqrt{v_1}$, $d = b/\sqrt{v_2}$ where

$$v_1 = 2((1 - a)^2 m_2^{-1} + m_1^{-1}), \qquad v_2 = 2((1 + b)^2 m_2^{-1} + m_1^{-1}).$$

We restrict attention to sequences of $m_1$ and $m_2$ values which tend to infinity with $L$ in such a manner that $L^2 m_1$ and $L^2 m_2$ have limits (possibly 0 or $\infty$) as $L \to 0$. It follows that $c$ and $d$ have limits $c_0$ and $d_0$ (possibly 0 or $\infty$) and $\rho \to -1$ as $L \to 0$. Thus, the joint distribution of $U_1$ and $U_2$ becomes singular along the line $y = -x$ so that for small $L$

$$\begin{aligned} P(U_1 \leqq c, \ U_2 \leqq d) &\cong P(-c_0 \leqq U_2 \leqq d_0) \\ &= \Phi(d_0) - \Phi(-c_0). \end{aligned} \tag{3.B.1}$$

It is easily seen from the above expressions for $c$ and $d$ that $d_0 = \beta c_0$ where $\beta$ is the quantity given in the text.

We now replace the original minimization problem by the following: Minimize $m_2 + Km_1$ subject to the constraint

$$\Phi(\beta c) - \Phi(-c) = 1 - \alpha/2, \tag{3.B.2}$$

where $c$ is the function of $m_1$ and $m_2$ given above (which is close to $c_0$ for small $L$). Now let $c_\alpha$ be the solution of (3.B.2). Then, ignoring the fact that $m_1$ and $m_2$ are integers, the new constraint is equivalent to the equation

$$m_2^{-1} + m_1^{-1} = a^2/(2c_\alpha^2). \tag{3.B.3}$$

The minimum value of $m_2 + Km_1$ subject to this condition is achieved for $m_2 = \sqrt{K} m_1$. The explicit values of $m_1$ and $m_2$ are now obtained by substituting this expression into (3.B.3). The integer versions of the solutions are those given in the text.

Note that expression (3.B.2) implies that $c_0$ and $d_0$ are necessarily finite and positive. The alternate choice of $c_\alpha$ is justified by the inequality

$$\Phi(\beta c) - \Phi(-c) > 2\Phi(c) - 1$$

which is valid since $\beta > 1$.

3.C. We also refer the interested reader to [3] for the proof of this procedure.

**Acknowledgement.** We have profited from discussions of these procedures with Judah Rosenblatt.

## References

1. J. R. Blum and Judah Rosenblatt, *On multistage estimation*, Ann. Math. Statist. **34** (1963), 1452–1458. MR **27** #5321.

2. L. H. Koopmans, D. B. Owen and J. I. Rosenblatt, *Confidence intervals for the coefficient of variation for the normal and log normal distributions*, Biometrika **51** (1964), 25–32. MR **30** #2599.

3. L. H. Koopmans and Clifford Qualls, *Fixed length confidence intervals for parameters of the normal distribution based on two-stage sampling procedures*, University of New Mexico Technical Report No. 177, Albuquerque, N. M., 1969.

4. Michel Loève, *Probability theory. Foundations. Random sequences*, 3rd ed., Van Nostrand, Princeton, N. J., 1963. MR **34** #3596.

5. Jack Moshman, *A method for selecting the size of the initial sample in Stein's two sample procedure*, Ann. Math. Statist. **29** (1958), 1271–1275. MR **20** #7361.

6. B. M. Seelbinder, *On Stein's two-stage sampling scheme*, Ann. Math. Statist. **24** (1953), 640–649. MR **15**, 637.

7. Charles Stein, *A two-sample test for a linear hypothesis whose power is independent of the variance*, Ann. Math. Statist. **16** (1945), 243–258. MR **7**, 213.

THE UNIVERSITY OF NEW MEXICO, ALBUQUERQUE, NEW MEXICO 87106