How Many Simulations Should One Run?

NITIS MUKHOPADHYAY
University of Connecticut, Storrs, U.S.A.
GREG CICCONETTI
Muhlenberg College, Allentown,, U.S.A.

1. INTRODUCTION: THE CUSTOMARY ONE-SAMPLE T TEST

We consider a well-known statistical methodology for tests of hypotheses. Suppose that we have a fixed number of independent observations $X_1, ..., X_m$ from a single population with a continuous probability density function (pdf) $f(x)$ and the distribution function (df) $F(x)$. We assume that this population has the finite mean μ and finite variance $\sigma^2(> 0)$. In order to test a null hypothesis $H_0 : \mu = \mu_0$ against a two-sided alternative hypothesis $H_1 : \mu \neq \mu_0$ at a given size or level $\alpha, 0 < \alpha < 1$, a customary two-sided test

$$\text{would reject } H_0 \text{ in favor of } H_1 \text{ if and only if } \left|\frac{\sqrt{m}(\overline{X}-\mu_0)}{S}\right| > t_{m-1,\alpha/2} \tag{1.1}$$

where $\overline{X} \equiv \overline{X}_m = m^{-1}\Sigma_{i=1}^m X_i$ is the sample mean, $S^2 \equiv S_m^2 = (m-1)^{-1}\Sigma_{i=1}^m (X_i - \overline{X}_m)^2$ is the sample variance, and $t_{m-1,\alpha/2}$ is the upper $50\alpha\%$ point of the Student's t distribution with $m-1$ degrees of freedom. If the population happened to be normal, that is if

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, -\infty < x < \infty, \tag{1.2}$$

then the size of the test procedure (1.1) will be exactly α. In statistical literature, this is a widely used and recommended test procedure to decide between H_0 versus H_1 even when f is not given by (1.2). When f is not given by (1.2), that is when the normality of the population distribution can not be justified from practical perspectives, then many standard books and manuals on statistical methodologies recommend that the *sample size* m should preferably be "large", customarily $m \geq 30$. The wisdom behind this comes from the following persuasion:

$$\begin{gathered}\text{Under } H_0, \text{ the test statistic } \frac{\sqrt{m}(\overline{X}_m-\mu_0)}{S_m} \xrightarrow{\mathcal{L}} N(0,1) \text{ as } m \to \infty. \text{ In} \\ \text{other words, the distribution of } \frac{\sqrt{m}(\overline{X}_m-\mu_0)}{S_m} \text{ can be approximated} \\ \text{by the standard normal distribution if the sample size } m \text{ is "large".} \\ \text{Then, the percentage point } t_{m-1,\alpha/2} \text{ may be replaced in (1.1) by} \\ z_{\alpha/2}, \text{ the upper } \tfrac{1}{2}\alpha \text{ point of the standard normal distribution.}\end{gathered} \tag{1.3}$$

Since a population distribution is rarely exactly normal in real applications, practitioners often invoke the idea summarized in (1.3) and then follow-up with the test procedure (1.1). But, then the size or level of that test may not coincide with α.

Now, one may like to address the following pertinent question: If the test procedure (1.1) is used in the case of a *specific* non-normal distribution f with fixed m, how different will the true size of the test (1.1) be from the nominal target α? One may try to reach practically useful guidelines derived from computer *simulation* runs with a few specifically chosen f. An obvious limitation of any computer *simulation*, however, is that one can not include all possible non-normal f in this investigation.

First we formulate the problem, rationalize a loss function, and then suggest a criterion to measure the "goodness" of the proposed estimator of the size associated with the test (1.1) in the case of a normal and three *specific* non-normal population distributions (Section 2). Then, the relevance of an appropriate sequential simulation methodology is described (Section 3). In this problem, we argue that a natural two-stage simulation methodology can also be pursued and its description is provided (Section 4). These two methodologies are followed by numerical examples and analyses (Section 5) where we find clear evidence that the two-stage simulation strategy operationally outperforms its sequential counterpart.

We emphasize that for the specific problem on hand, neither loss function nor these proposed sequential and two-stage simulation methodologies can be explicitly found in the literature. Also, the derivations of some of the associated asymptotics do not appear too routine and hence these are appended at the end of the paper (Section 6).

We end this section with a passing remark. As we began reviewing the literature on "simulation" within the areas of computer science and numerical analyses, we quickly discovered that phrases such as *simulation*, *repetition*, *replication*, *run*, *simulation size*, *batch* are interchangeably used without discretion. Unfortunately, these taken-for-granted terminologies do create a confusing maze from which it becomes quite a challenge to figure out exactly "what" part of an experiment is being "simulated". In one situation, simply generating $m = 10$ observations from a known population itself is called a *simulation*. In yet another situation, generating $m = 10$ observations from a known population a fixed number of times, say $n = 100$, is called a *simulation*. In the second situation, is the "simulation size" 10 or 100? From this point onward, we will try to be careful in our usage of various terminologies.

2. FORMULATION, LOSS FUNCTION, RISK FUNCTION, AND THE OBJECTIVE

Even with the help of a sophisticated computer, it will be physically impossible to investigate the performance of the statistical methodology (1.1) for every single *non-normal* population distribution f. In this investigation, we include one normal and three *specific* non-normal population distributions. Let $\Phi(.)$ denote the standard normal df, that is

$$\Phi(x) = \int_{y=-\infty}^{x} \phi(y)dy, \phi(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}), -\infty < x < \infty, \tag{2.1}$$

and we consider the following fixed choices for F.

1. $F(x) \equiv F_1(x) = \Phi\left(\frac{x-5}{\sqrt{2}}\right)$, that is the population distribution is $N(5, 2)$;
2. $F(x) \equiv F_2(x) = 0.75\Phi\left(\frac{x-5}{1}\right) + 0.25\Phi\left(\frac{x-5}{\sqrt{5}}\right)$, that is the population distribution is a mixture of 75% $N(5, 1)$ and 25% $N(5, 5)$ distributions;

3. $F(x) \equiv F_3(x) = 0.75\Phi\left(\frac{x-5}{\sqrt{4/3}}\right) + 0.25\Phi\left(\frac{x-5}{2}\right)$, that is the population distribution is a mixture of 75% $N(5, \frac{4}{3})$ and 25% $N(5, 4)$ distributions;
4. $F(x) \equiv F_4(x) = 0.75\Phi\left(\frac{x-5}{\sqrt{1/3}}\right) + 0.25\Phi\left(\frac{x-5}{\sqrt{7}}\right)$, that is the population distribution is a mixture of 75% $N(5, \frac{1}{3})$ and 25% $N(5, 7)$ distributions.

(2.2)

Incidentally, the random variables associated with choices #2-#4 are sometimes referred to as Tukey noises or Tukey random variables.

Our goal is to examine some of the key roles that sequential analysis can play in designing "efficient" experiments with computer simulations. We proceed to highlight our ideas by simulating the one-sample test (1.1) when f corresponds to any one of the four choices from (2.2). This is just an example! One may examine the one-sample test (1.1) by considering f that corresponds to any other kind of distribution not included in (2.2). This article emphasizes the importance of sequential and multi-stage sampling on any field that may rely upon computer simulations. Obviously, in principle, similar ideas may be carried over for designing "efficient" computer simulations in general.

A reader may feel tempted to place this investigation within some general context of robust hypothesis testing. It should be clear that our intent is quite different. We certainly do not present a full-blown investigation of robustness properties of one- or two-sample t-test. But, it may not be out of place to mention that many relevant robustness issues associated with both one-sample and two-sample t-tests were addressed by Tukey (1960), Tukey and McLaughlin (1963), Tiku and Singh (1981), Posten et al.(1982), and others.

Let us first explain the particular choices of parameter values in (2.2). Suppose that a population pdf is given by

$$f(x) = \tau g(x) + (1-\tau)h(x), x \in R, 0 < \tau < 1, \tag{2.3}$$

where g, h are also both pdf's on R. Let μ_g, μ_h denote the means corresponding to the pdf's g and h respectively. Then, one verifies easily that

$$E_f[X] = \tau\mu_g + (1-\tau)\mu_h \text{ and } E_f[X^2] = \tau E_g[X^2] + (1-\tau)E_h[X^2] \tag{2.4}$$

Hence, it checks out easily that all four selected population distributions are such that each corresponding pdf f is continuous, symmetric about $x = 5$, and we have $\mu \equiv \mu_f = 5, \sigma^2 \equiv \sigma_f^2 = E_f[X^2] - E_f^2[X] = 2$ where $F(x) \equiv F_i(x), i = 1, 2, 3, 4$. The plots of the associated pdf's $f_i(x), i = 1, 2, 3, 4$, are given in Figure 2.1.

While the variance of the four distributions in (2.2) coincide, we look for other measures that might help in understanding the effect of these distributions on the test (1.1). We compare, for example, the tail areas falling beyond the 97.5^{th} percentile of the distribution $F_i, i = 1, 2, 3, 4$.

It should be clear that we have labelled F_2, F_3 and F_4 respectively as "heavier", "lighter", and "heavier" than normal in the sense of "slower", "faster", and "slower" decreasing tails. In this sense, one also notes that F_4 is heavier than F_2. Now, since the t-test (1.1) typically performs worse in the presence of heavy tails and

skewness, we might expect a bit more variability in the simulation results obtained when $F = F_2, F_4$.

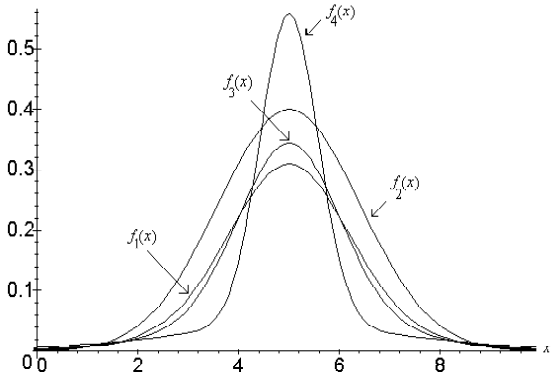

Figure 2.1. Plots of the Four PDF's Corresponding to (2.3).

Table 2.1. Tail Behavior of the Distributions $F_1 - F_4$

Distribution	$P_F\,(X > 7.77198)$	Tail Characteristic
F_1 Normal: $P_{F_1}\,(X > 7.77198) = 0.025$		
F_2	0.028977	Heavier than normal
F_3	0.012386	Lighter than normal
F_4	0.036847	Heavier than normal

We apply the statistical methodology (1.1) to test the null hypothesis $H_0 : \mu = 5$ against a two-sided alternative hypothesis $H_1 : \mu \neq 5$ having a preassigned size $\alpha = 0.05$ on four separate occasions, that is when the underlying population distribution is governed by $F_i(x), i = 1, 2, 3, 4$.

Goal: We wish to estimate the Type I error probability θ_i when the population density f corresponds to $F(x) \equiv F_i(x)$, $i = 1, 2, 3, 4$.

This is accomplished with the help of a computer "simulation" algorithm which is laid out precisely as follows: For the problem on hand, we keep the *sample size* m fixed.

Step #1: First, we fix $m = 5$ and $F(x) \equiv F_1(x)$. We have $t_{4,0.025} = 2.7764$.

Step #2: We generate m observations from this population and obtain the sample mean $\overline{x} = \overline{x}_1$ and standard deviation $s = s_1$ respectively.

Step #3: We evaluate the test statistic $t_1 = \frac{\sqrt{m}(\overline{x}_1 - 5)}{s_1}$ and check whether $\left|\frac{\sqrt{m}(\overline{x}_1 - 5)}{s_1}\right| > 2.7764$. We reject H_0 when $\left|\frac{\sqrt{m}(\overline{x}_1 - 5)}{s_1}\right| > 2.7764$.

Step #4: We generate and save an indicator variable $u \equiv u_{11}$ which takes the value 1 or 0 according as $\left|\frac{\sqrt{m}(\overline{x}_1 - 5)}{s_1}\right| > 2.7764$ or $\left|\frac{\sqrt{m}(\overline{x}_1 - 5)}{s_1}\right| \leq 2.7764$ respectively.

When a cycle through Steps #1-#4 is completed, we refer to this as one *pass* or one *simulation*.

But, naturally from one single observed value u_{11}, which is either zero or one, we can not hope to find a reasonable estimate of the Type I error probability θ_1. We may mention that all standard texts on *simulation* will advocate repeating the previously laid out Steps #1-#4 on n separate occasions. Thus, one would obtain n values $u_{11}, ..., u_{1n}$ for the indicator variable u. The Type I error probability θ_1 is then customarily estimated by $\overline{u}_{1n} = n^{-1}\Sigma_{j=1}^{n} u_{1j}$, that is the observed sample proportion of rejecting H_0.

Next, we successively fix $m = 10, 15, 30, 50$, go back to Steps #1-#4 and finish all n passes or simulations to examine the effect of the choice of m on the respective estimate $\overline{u}_{1n}$ for the Type I error probability θ_1. Then, we can repeat this whole exercise when $m = 5, 10, 15, 30, 50$ and $F(x)$ is successively replaced by $F_i(x)$ in order to estimate θ_i by $\overline{u}_{in}$, the observed sample proportion of rejecting H_0, $i = 2, 3, 4$. But, then what should be chosen as n? Should we fix $n = 100$ or 500 or $1,000$ or $10,000$ or even larger? Customarily, however, one will simply fix $n = 1,000$ or $5,000$ and pass through Steps #1-#4 n times. But, any reasonable way to determine some appropriate n must depend upon how one quantifies the "closeness" between the estimated Type I error probability $\overline{u}_{in}$ and unknown $\theta_i, i = 1, 2, 3, 4$. Designing the simulation experiment itself and appropriately determining the number of simulations, namely n, form the core of this article.

In some textbooks, authors have recommended using appropriate sequential experiments for determining the simulation size n. One may refer to Ross (1999, Chapter 7) or Bratley et al. (1988, pp. 29, 40, 76-78, 92, 101) for basic motivations. We remark in passing that Bratley et al. (1988, p. 92) have mentioned the classic sequential paper of Chow and Robbins (1965) a number of times and proposed that n should be determined sequentially depending upon appropriate quantification of the associated "estimation error". Some Chow-Robbins type sequential simulation strategies and their merits relative to fixed-size (that is, when n is fixed in advance) simulation strategies are discussed sporadically in Ross (1999) and somewhat more systematically in Bratley et al. (1988). Neither source offered readers a clear roadmap involving applications of sequential methods.

2.1. The Formulation

In order to formulate the problem properly, we go back to Step #1. We start with a fixed choice of m and one population distribution function $F(x)$. The true unknown Type I error probability is given by

$$\theta = P_F\left\{\left|\frac{\sqrt{m}(\overline{X}-\mu_0)}{S}\right| > t_{m-1,\alpha/2} \text{ when } \mu = \mu_0\right\}, 0 < \theta < 1, \tag{2.5}$$

and we define an indicator random variable

$$U = \begin{cases} 1 & \text{if } \left|\frac{\sqrt{m}(\overline{X}-\mu_0)}{S}\right| > t_{m-1,\alpha/2}, \text{ that is } H_0 \text{ is rejected} \\ 0 & \text{if } \left|\frac{\sqrt{m}(\overline{X}-\mu_0)}{S}\right| \leq t_{m-1,\alpha/2}, \text{ that is } H_0 \text{ is not rejected,} \end{cases} \tag{2.6}$$

which has a Bernoulli(θ) distribution. After Steps #1-#4 have been repeated n times, one comes up with n independent Bernoulli(θ) random variables $U_1, ..., U_n$ with their realizations $u_1, ..., u_n$. Now, based on n such simulations, the true but unknown rejection probability θ is customarily estimated by $\overline{U}_n = n^{-1}\Sigma_{i=1}^{n} U_i$, the

sample proportion of rejecting H_0 out of n simulations. It is obvious that $\overline{U}_n$ is an unbiased estimator of θ whereas $V_F(\overline{U}_n) = \theta(1-\theta)/n$ for all fixed n.

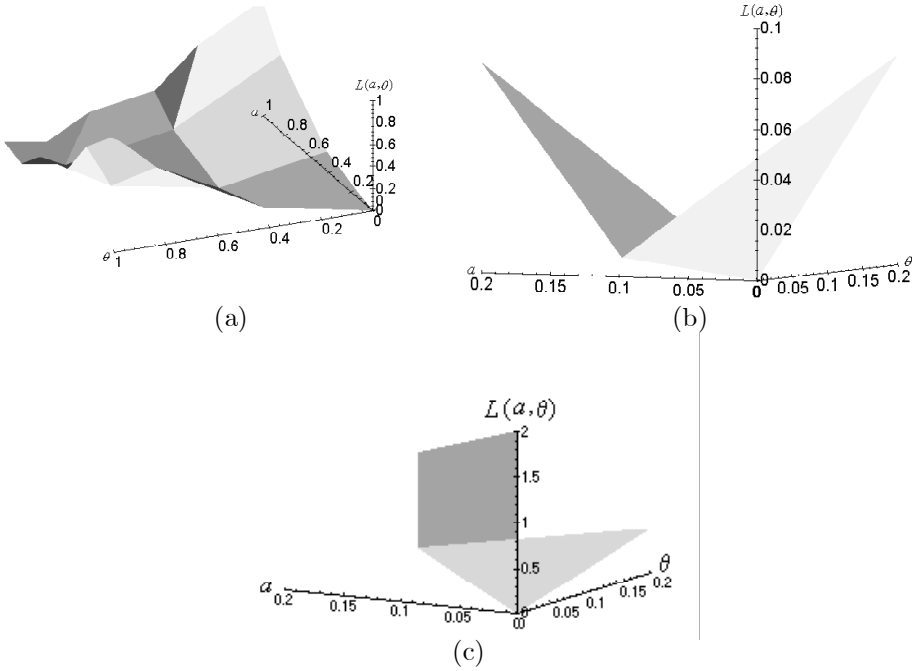

Figure 2.2. The Loss Functions: (a) $L(a,\theta) = (a-\theta)^2$;
(b) $L(a,\theta) = (a-\theta)^2$ near $(0,0)$; (c) $L(a,\theta) = (\frac{a}{\theta} - 1)^2$ near $(0,0)$.

2.2. A Relative Squared Error Loss Function

We realize that θ is supposedly a small number and thus, one may measure the discrepancy between the estimator $\overline{U}_n$ and θ with the following loss function:

$$L_n(\overline{U}_n, \theta) = \left[\frac{\overline{U}_n}{\theta} - 1\right]^2 : \text{“proportionally close” loss.} \tag{2.7}$$

This particular loss function makes good sense because when θ is small, we would look at “large” departure between $\frac{\overline{U}_n}{\theta}$ and 1 very seriously. In a situation like this, a traditional squared error loss function, namely $L_n(\overline{U}_n, \theta) = (\overline{U}_n - \theta)^2$, may not be very sensible. For example if $\overline{U}_n = 0.02, \theta = 0.01$, then the squared error loss is 0.0001 whereas the proportional loss is 1.0. In other words, when θ is assumed small, we may expect a small value of $\overline{U}_n$. Then, a squared error loss function will seem to indicate that $\overline{U}_n$ estimates θ quite accurately but at the same time $\overline{U}_n$ may be several fold larger than θ! The squared error loss is not very sensitive to our specific situation where θ small, but the proportional loss is. A sharp contrast between the squared error loss function and a proportional loss function, particularly near the origin, is clearly visible from the three plots given in Figure 2.2.

In the past, a proportional or a relative squared error loss function has been used by a number of authors in the context of many inference scenarios. Some of the references that come to mind include Nadas (1969), Martinsek (1983,1995), Willson and Folks (1983), Mukhopadhyay and Diaz (1985), and Sriram (1991).

2.3. The Objective: Predetermined Proportional Accuracy

The risk function associated with the loss function L_n, defined in (2.7), is given by

$$E_F[L_n(\overline{U}_n, \theta)] = E_F\left[\frac{\overline{U}_n}{\theta} - 1\right]^2 = \frac{\theta(1-\theta)}{n\theta^2} = \frac{(1-\theta)}{n\theta}. \tag{2.8}$$

Now, one may require that the number of *passes* n be so chosen that the risk of estimation does not exceed a preassigned number $c(> 0)$. In other words, one puts a restriction that

$$\begin{gathered}\frac{(1-\theta)}{n\theta} \leq c, \text{ which holds if and only if } n \text{ is the smallest} \\ \text{integer} \geq c^{-1}\frac{(1-\theta)}{\theta} = n^*(\theta), \text{ say.}\end{gathered} \tag{2.9}$$

In practice, how should one choose the design constant c? It is essential that c is chosen "small" because then we may expect to claim that on the average $\overline{U}_n$ is in a close proximity of θ. The experimenter may decide that it would be fairly adequate to estimate θ within 10% of the magnitude of θ. In this case, the choice $c = (0.1)^2 = 0.01$ may express the experimenter's feeling quite well. On the other hand, the experimenter may wish to estimate θ within 7% of the magnitude of θ instead. In this case, the choice $c = (0.07)^2 = 0.0049$ may express the experimenter's feeling. Which choice of c should one use in practice? A suitable resolution will take into account the required level of accuracy a practical context may demand.

When $\alpha = 0.05$, we may genuinely expect θ to be "around" 0.05. Given this sentiment, an experimenter may then attempt to subjectively quantify the maximum possible error that can be absorbed in practice without causing a significant backlash. Considerations along these lines will guide one to come up with some appropriate number $c(> 0)$ on a case by case basis.

Remark 2.1. Robbins and Siegmund (1974) gave a sequential estimation procedure for a Bernoulli success probability parameter θ with a loss function $\frac{(\overline{U}_n-\theta)^2}{\theta^2(1-\theta)^2}$. This loss function is appropriate when one feels apriori that θ may be close to zero or one. However, this loss function is not very attractive in our situation because we feel that here the parameter θ is close to zero.

3. A SEQUENTIAL METHODOLOGY

In order to estimate θ, supposedly a small percentage, one will run through the Steps #1-#4 at least some minimal number of times, say k. We can safely postulate that $0 < \theta < \theta_0 (< 1)$ where θ_0 is assumed known. Then, it immediately follows that

$$n^*(\theta) = c^{-1}\frac{(1-\theta)}{\theta} \geq c^{-1}\frac{(1-\theta_0)}{\theta_0}, \tag{3.1}$$

and hence we define the *initial* or *pilot* simulation size

$$\begin{gathered}k \equiv k(c) = \max\left\{k_0, \left\langle c^{-1}\frac{(1-\theta_0)}{\theta_0}\right\rangle + 1\right\} \text{ where } < a > \text{ is the} \\ \text{largest integer} < a \text{ and } k_0 \text{ is known positive integer.}\end{gathered} \tag{3.2}$$

Since θ denotes a small percentage, we may expect only a "small" number of U's to take the value one among 100 or more pilot observations. In practice, if one

can afford to work with larger k_0, then one should. But, since the magnitude of $n^* \equiv n^*(\theta)$ remains unknown, we propose the following sequential simulation strategy by mimicking the classical techniques developed by Chow and Robbins (1965). We first generate k initial Bernoulli observations $U_1, ..., U_k$ where the pilot sample size k is defined by (3.2). Next, we repeat the Steps #1-#4

$$\begin{gathered} N \text{ times with one additionally generated Bernoulli observation} \\ \text{after each pass where } N \equiv N(c) \text{ is the integer } n(\geq k) \text{ for which} \\ \text{we observe } n \geq c^{-1}\tfrac{(1-\overline{U}_n - n^{-1})}{(\overline{U}_n + n^{-1})} \text{ for the first time.} \\ \text{Finally, } \theta \text{ is estimated by } \overline{U}_N = N^{-1}\Sigma_{i=1}^{N} U_i. \end{gathered} \tag{3.3}$$

In the stopping rule (3.3), we first compare the current simulation size k with the current estimate $\widehat{n_k^*(\theta)} = c^{-1}\frac{(1-\overline{U}_k - k^{-1})}{(\overline{U}_k + k^{-1})}$ of $n^*(\theta)$. If $k \geq \widehat{n_k^*(\theta)}$, we stop simulations with the pilot observations $U_1, ..., U_k$ and thus estimate θ by $\overline{U}_k$. If $k < \widehat{n_k^*(\theta)}$, then we go back to Steps #1-#4 to generate a new Bernoulli random variable U_{k+1} and compare the current simulation size $k+1$ with the current estimate $\widehat{n_{k+1}^*(\theta)} = c^{-1}\frac{(1-\overline{U}_{k+1} - (k+1)^{-1})}{(\overline{U}_{k+1} + (k+1)^{-1})}$ of $n^*(\theta)$. If $k+1 \geq \widehat{n_{k+1}^*(\theta)}$, we stop simulations with the observations $U_1, ..., U_k, U_{k+1}$ and thus estimate θ by $\overline{U}_{k+1}$. But, if $k+1 < \widehat{n_{k+1}^*(\theta)}$, then we go back to Steps #1-#4 to generate another Bernoulli random variable U_{k+2} and compare the current simulation size $k+2$ with the current estimate $\widehat{n_{k+2}^*(\theta)} = c^{-1}\frac{(1-\overline{U}_{k+2} - (k+2)^{-1})}{(\overline{U}_{k+2} + (k+2)^{-1})}$ of $n^*(\theta)$. This process continues until the stated boundary condition in the stopping rule (3.3) is satisfied for the very first time.

From the stopping rule (3.3), we note that as long as $\overline{U}_n$ is observed zero, the stopping boundary condition can not be satisfied. Now, is there any guarantee that for all fixed c and θ, the sequential procedure will terminate with probability one? Toward that end, we write

$$\begin{aligned} P_\theta\{N > n\} &\leq P_\theta\left\{n < \tfrac{1}{c}\tfrac{(1-\overline{U}_n - n^{-1})}{(\overline{U}_n + n^{-1})}\right\} = P_\theta\{\overline{U}_n < \tfrac{1}{cn+1}\} \\ &\leq P_\theta\left\{\left|\tfrac{\overline{U}_n}{\theta} - 1\right| > \tfrac{1}{2}\right\}, \end{aligned}$$

since for sufficiently large n, the expression $\frac{1}{\theta(cn+1)} - 1$ can be made smaller than $-\frac{1}{2}$. Thus, for sufficiently large n, using Tchebysheff's inequality, we have

$$P_\theta\{N > n\} \leq 4E_\theta\left[(\tfrac{\overline{U}_n}{\theta} - 1)^2\right] = \tfrac{4\theta(1-\theta)}{n\theta^2} = \tfrac{4(1-\theta)}{n\theta}.$$

In other words, for all fixed $c(>0), 0 < \theta < 1$, we obtain

$$\begin{gathered} 0 \leq P_\theta(N = \infty) = \lim_{n\to\infty} P_\theta\{N > n\} \leq \lim_{n\to\infty} \tfrac{4(1-\theta)}{n\theta} = 0 \\ \Rightarrow P_\theta(N = \infty) = 0 \text{ so that } P_\theta(N < \infty) = 1. \end{gathered}$$

In other words, the sequential procedure (3.3) does terminate with probability one.

4. A TWO-STAGE METHODOLOGY

The motivation is derived from the fact that sampling with larger batches can be observed quickly and hence one may be able to cut down the operational time

significantly compared with sampling sequentially. More explanations are given Section 4.1.

Now, recall the expression of $n^*(\theta)$ from (3.1). Again, we define the *initial* or *pilot* simulation size

$$k \equiv k(c) = \max\left\{k_0, \left\langle c^{-1}\frac{(1-\theta_0)}{\theta_0}\right\rangle + 1\right\} \text{ where } < a > \text{ is the largest integer } < a \text{ and } k_0 \text{ is known positive integer.} \tag{4.1}$$

We note that the pilot simulation size k is going to increase as $n^*(\theta)$ increases and for all practical purposes k is a lower bound of $n^*(\theta)$. Thus, we may proceed to consider a Stein-type (Stein (1945,1949)) two-stage methodology along the line of Mukhopadhyay and Duggan's (1997) adaptation that was shown to enjoy *second-order* properties in the normal case.

We first generate k initial Bernoulli observations $U_1, ..., U_k$ where the pilot sample size k is defined by (4.1). Next, we construct the final simulation size M as follows:

$$M \equiv M(c) = \max\left\{k, \left\langle c^{-1}\frac{(1-\overline{U}_k-k^{-1})}{(\overline{U}_k+k^{-1})}\right\rangle + 1\right\}. \tag{4.2}$$

Note that in (4.2), we have replaced $\overline{U}_k$ with $\overline{U}_k + k^{-1}$ just to make sure that the ratio $\frac{(1-\overline{U}_k-k^{-1})}{(\overline{U}_k+k^{-1})}$ remains well-defined if it happens at all that $\overline{U}_k = 0$. Now, we explain how this simulation strategy is going to be implemented.

(i) If we find that $M = k$, then we do not generate any additional Bernoulli U's so that the simulation will end simply with the k pilot passes, and then $\overline{U}_k$ will be the final estimator of θ.

(ii) If we find that $M > k$, then we generate additional $M - k$ Bernoulli random variables $U_{k+1}, ..., U_M$ so that the simulation will end with M passes. Then, $\overline{U}_M = M^{-1}\Sigma_{i=1}^{M} U_i$, that is the sample mean of the combined observations $U_1, ..., U_k, U_{k+1}, ..., U_M$, will be the final estimator of θ.

Combining (i) and (ii), we can say that M is the final simulation size and $\overline{U}_M$ is the final estimator of θ. From the stopping rule (4.2), we note that for all fixed c and θ, the final simulation size M is finite with probability one.

Such a two-stage procedure for a multivariate normal problem was first developed by Mukhopadhyay (1999). Aoshima (2000) and Aoshima and Mukhopadhyay (1999) developed other multivariate two-stage procedures with *second-order* properties.

4.1. The Motivation

One may ask: Why should a practitioner implement the two-stage methodology rather than the earlier sequential methodology from (3.2)-(3.3)? The answer is quite simple indeed. In the sequential strategy (3.3), after every pass through Steps #1-#4, one has to check whether n exceeded $c^{-1}\frac{(1-\overline{U}_n-n^{-1})}{(\overline{U}_n+n^{-1})}$ for the first time, $n \geq k$. We may expect the associated stopping variable N to turn out much larger than k, and then the operational time to check the boundary condition will pile

up fast. The two-stage strategy (4.1)-(4.2) avoids this requirement of checking boundary conditions in its entirety. That is, this latter simulation methodology will terminate faster than the sequential simulation strategy.

In many simulation projects for designing complex systems in computer engineering, one frequently confronts a bigger system that consists of five or ten such *nested* sub-systems. Now, imagine a complex simulation strategy that consists of three nested sub-simulation strategies. In the first sub-system, suppose that checking the boundary condition takes 0.01 seconds each time and it is completed with $20,193$ sequential passes beyond k_1, the pilot simulation size. The estimator $W_{N_1}^{(1)}$ of ν_1 thus obtained is then used in the second sub-system, and suppose that checking the associated boundary condition takes 0.001 seconds each time and it is completed with $73,268$ sequential passes beyond k_2, the pilot simulation size. The estimator $W_{N_2}^{(2)}$ of ν_2 thus obtained is then used in the third sub-system, and suppose that checking the associated boundary condition takes 0.001 seconds each time and it is completed with $28,758$ sequential passes beyond k_3, the pilot simulation size. The estimator $W_{N_3}^{(3)}$ of ν_3 thus obtained is then a final output obtained from the whole system. This is just one simulation pass of the full system. Then, the savings in computer operating time alone will be $20193\times0.01+73268\times0.001+28758\times0.001 = 303.96$ seconds, that is approximately 5.066 minutes, due to implementing two-stage technique (4.1)-(4.2) rather than the sequential technique (3.2)-(3.3). The full system, however, may be simulated 100 or 200 times.

When the full system is simulated 100 times, this gives rise to 100 triplets $(N_{1j}, N_{2j}, N_{3j}), j = 1, ..., 100$. Let us denote $T_{ij} = N_{ij} - k_i + 1, i = 1, 2, 3.$ such that $T_{1\,\min} = 15642, T_{1\,\max} = 24352, T_{2\,\min} = 68274, T_{2\,\max} = 77852$, and $T_{3\,\min} = 21628, T_{3\,\max} = 32291$ with sequential passes beyond the pilot size within subsystems. Now, if all the sequential loops are replaced by the corresponding two-stage simulations within each sub-system, the savings (S) in the computer operating time alone will be between $(0.01 \times 15642 + 0.0001 \times 68274 + 0.001 \times 21628) \times 100$ seconds or 5.14 hours and $(0.01 \times 24352 + 0.0001 \times 77852 + 0.001 \times 32291) \times 100$ seconds or 7.88 hours. In other words, one would have 5.14 hours $\leq S \leq$ 7.88 hours. Often experienced hired hands are paid by the hour to program and test a system, and then this magnitude of savings in computer operation alone will translate into substantial savings in dollars. Therein lies the strength of a two-stage methodology along the line of (4.1).

Remark 4.1. The important message we are sending is that the number of simulations should not be arbitrarily fixed. We have put forth this message forcefully with the help of sound arguments and a specific bounded-risk point estimation problem involving a single parameter θ between 0 and 1 that may be close to zero. We believe that the methodologies proposed here could be extended, in principle, for simulation studies of much more complex problems involving (i) estimation of a high-dimensional parameter vector $\boldsymbol{\theta}$, (ii) infinite-dimensional function estimation, or (iii) Markov Chain Monte Carlo investigations where successive iterations are dependent. These remain important and challenging *open problems* since one must start from ground zero by formulating (i) an appropriate loss function and (ii) the kind of "optimization" one may like to pursue on a case by case basis. It may, however, be too early to ponder about constructing some general theory and methodology embracing all types simulations and optimization techniques.

Remark 4.2. In the determination of the pilot simulation size k, both sequential and two-stage methodologies exploited the assumption of existence of a known number θ_0 such that $0 < \theta < \theta_0$. Practically in any realistic situation, validity of this assumption would almost be taken for granted. But, no matter how we choose θ_0, it may be possible to think of a distribution f, perhaps pathological in nature, for which this assumption would not hold. In such pathological circumstances, different kinds of sequential and two-stage methodologies for appropriately determining the number of simulations can be developed along the lines of Chow and Robbins (1965) and Mukhopadhyay and Diaz (1985) respectively.

5. COMPUTER SIMULATIONS

We have investigated the performance of the sequential methodology (3.2)-(3.3) and the two-stage methodology (4.1)-(4.2) on the hypothesis testing procedure (1.1) under a variety of scenarios. In particular, we worked under the following configuration combinations: $m = 5, 10, 20, 30, 50$, $k_0 = 100$, $c = 0.1, 0.01, 0.0049, 0.001$ and with the population distribution functions given by $F_i(x)$, $i = 1, 2, 3, 4$ described in (2.2). For each of these 240 configurations we are interested in N or M, the stopping simulation size, $\widehat{\alpha}$, the point estimate for the probability of rejecting H_0, and the completion time required by the computer under each configuration. In order to comment on the variability of these recorded values, the entire 240 configurations were repeated 10 times. Since some of the combinations yielded stopping simulation sizes close to $k_0 = 100$, which is supposedly quite small, one complete configuration was run with $k_0 = 200$ to explore how this change might improve estimation of the rejection rate. We will comment on these findings in our concluding remarks.

One may perhaps criticize our decision to stop with only ten complete replications. Once we have a good grasp of how to determine N or M, we can then proceed to come up with the estimate of an appropriate number of complete replications, again sequentially, as the next step in a hierarchical system. Hence, we devote our full energy only on determining N or M with a sequential design first. The variability in N or M and those in the associated estimators of the rejection rates, however, are of secondary interest at this point. We will suffice to use ten complete replications as a means of performing simple initial EDA and speak in some generality of possible variations in these estimates. If we wish to gain point estimators for the measures of variation in the future, the suggested ten complete replications may then be utilized as a part of a "pilot simulation" in an appropriately designed two-stage or sequential follow-up study. Finally, these sentiments may be reminiscent of determining the number of levels in a hierarchical model — but, one would stop at some step in nay hierarchical system!

5.1. The Sequential Methodology (3.2)-(3.3)

Throughout the various configurations, general patterns are quite evident in the sequential procedure (3.2)-(3.3) and support intuition. Figures 5.1-5.4 reflect the impact of the distribution, and magnitude of the initial estimate θ_0 on the stopping simulation size, N and the rejection rate θ when $c = 0.1, 0.001$ for each of the four distributions under study $F_i(x), i = 1, 2, 3, 4$ described in (2.2). In the Figures 5.1-5.4, each symbol represents the average obtained from the ten replications for a particular value of the sample size $m = 5, 10, 20, 30, 50$ and the chosen value θ_0. The overall features, trends, and conclusions were very similar when we had fixed

$c = 0.01, 0.0049$ and hence detailed findings are summarized only in the cases when $c = 0.1, 0.001$. Throughout the Figures, we use circles, astersiks and crosses to refer to the situations when $\theta_0 = 0.06, 0.10, 0.30$ respectively. For example, in Figure 5.1, the circle located in the lower left hand corner corresponds to the average of stopping simulation size when $F = F_1, \theta_0 = 0.06, c = 0.1$ and $m = 50$.

This investigation revealed no detectable common pattern for different choices of m and hence we decided to suppress the information concerning the sample size m. In other words, the specific sample size m used in the t-test (1.1) had no appreciable impact on either the final simulation size N or the rejection rate θ.

From Figures 5.1 and 5.2 we make the following assertions: The initial choice of θ_0 does not seem to impact the final simulation size or its variability. But, reviewing the methodologies (3.2)-(3.3) and (4.1)-(4.2), we are reminded that θ_0 plays a role in determining the pilot simulation size, and hence we were relieved to note that we were not penalized in a harsh manner if we decided to use a conservative choice of θ_0. We note that the choice of $\theta_0 = 0.06$, was perhaps a bit too risky in light of our initial ignorance of the behavior of the rejection rate θ.

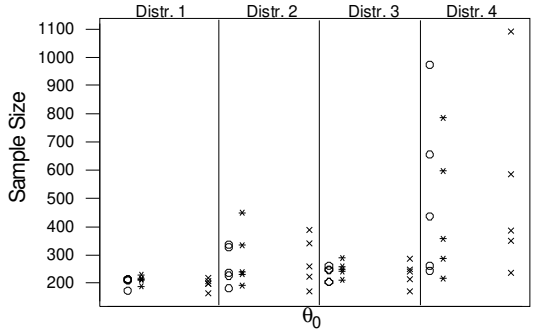

Figure 5.1. Variation in Sequential Simulation (or Sample) Size N by θ_0 for Distributions $F_i, i = 1, 2, 3, 4$, with Risk-Bound $c = 0.1$: Circles, Asteriks and Crosses Correspond to $\theta_0 = 0.06, 0.10, 0.30$ Respectively.

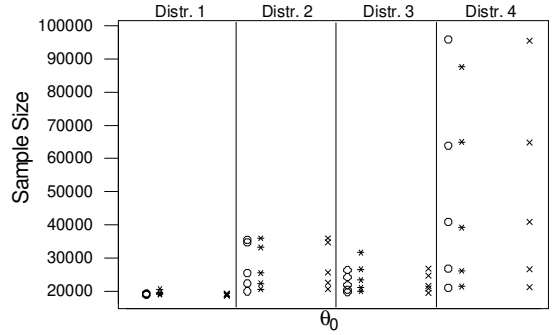

Figure 5.2. Variation in Sequential Simulation (or Sample) Size N by θ_0 for Distributions $F_i, i = 1, 2, 3, 4$, with Risk-Bound $c = 0.001$: Circles, Asteriks and Crosses Correspond to $\theta_0 = 0.06, 0.10, 0.30$ Respectively.

The risk-bound c plays its expected role, that is as the risk decreased, a natural increase in the stopping simulation size could be seen. This established, it is not

surprising to witness an increase in the spread of N. The choice of the population distribution F also plays a prominent role in the determination of the stopping simulation size. Under the ideal normality conditions of F_1 we found the least variation in the stopping simulation size, followed closely by a similar pattern under F_3. Recall that F_3 was found to have lighter than normal tails. Variability increased when we considered F_2 and more so with F_4 as expected since each population distribution F_2 and F_4 had a heavier than normal tail (Table 2.1).

Figures 5.3-5.4 are concerned with the performance of the rejection rate θ. The comments concerning the stopping simulation size N seem to have a direct impact on the rejection rates for the four distributions. Again, we note that the sample size m, did not lead to general patterns and so this information is suppressed in the figures. The dotted line at 0.05 marks the target rejection rate under F_1's normality. The figures suggest that the variation inherent in N is passed to the variation of the rejection rate. We note further that when the risk-bound is $c = 0.1$, the two distributions F_1, F_3 perform quite comparably, while the rejection rates in the case of the distributions F_2, F_4 have much larger variation. With decreased risk-bound

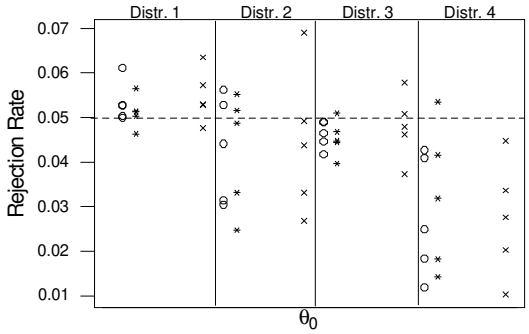

Figure 5.3. Variation in Rejection Rate of Sequential Procedure Due to Simulation Size N by θ_0 for Distributions $F_i, i = 1, 2, 3, 4$, with Risk-Bound $c = 0.1$: Circles, Asteriks and Crosses Correspond to $\theta_0 = 0.06, 0.10, 0.30$ Respectively.

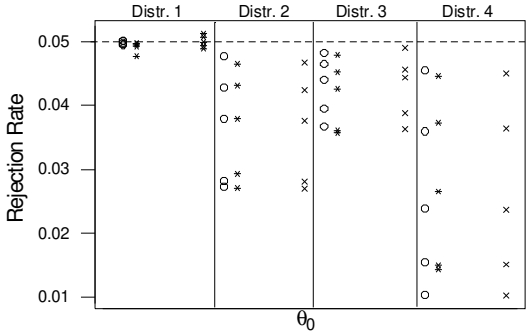

Figure 5.4. Variation in Rejection Rate of Sequential Procedure Due to Simulation Size N by θ_0 for Distributions $F_i, i = 1, 2, 3, 4$, with Risk-Bound $c = 0.001$: Circles, Asteriks and Crosses Correspond to $\theta_0 = 0.06, 0.10, 0.30$ Respectively.

c, and therefore with increased stopping simulation size N, we see a much better picture concerning the rejection rates. The nature of the distributions F_2, F_3, F_4 affects the t-test which seems to perform in a much more conservative fashion. Moreover, large discrepancies among stopping simulation sizes under $F_i, i = 2, 3, 4$ compared with F_1, indicate that establishing a common rejection rate for all values of m is perhaps unattainable.

We should add that heavy tails associated with F_3, F_4 have affected the rejection rates in the same fashion as the stopping simulation sizes.

5.2. The Two-Stage Methodology (4.1)-(4.2)

The two-stage procedure (4.1)-(4.2) mimics the sequential procedure with respect to the behavior of N. We include Figures 5.5-5.6 for comparison purposes.

Comments concerning the rejection rates based on the two-stage procedure are analogous to the sequential procedure. We include Figures 5.7-5.8 for comparison purposes.

5.3. Comparing the Two Methodologies and Conclusions

Our remarks from the previous subsections indicate little difference in the performance of the sequential and the two-stage methodologies as measured by their stopping simulation sizes and rejection rates. On account of such extreme similarities across simulations, our recommendation of one procedure over the other relies heavily on the determination of which procedure runs more efficiently.

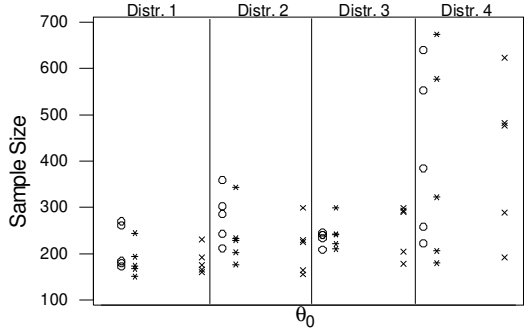

Figure 5.5. Variation in Two-Stage Simulation or Sample Size M by θ_0 for Distributions $F_i, i = 1, 2, 3, 4$, with Risk $c = 0.1$: Circles, Asteriks and Crosses Correspond to $\theta_0 = 0.06, 0.10, 0.30$ Respectively.

Two of the ten configurations ran on a relatively older computer, a Dell x86 Family6 Model 5 Stepping 2 with CPU 448 mHz and 655, 360 KB of RAM, and only 264 MB free space on 3.92 GB hard drive. The remaining eight configurations ran on a Dell Workstation PWS530 with CPU 1700 mHz 1, 047, 564 KB RAM and 33.6 GB free space out of a total hard drive space of 37.2 GB. Simulations were run with the help of S-Plus Version 6.0. Overall time spent for implementing the sequential methodology was 60.14 hours, while 56.46 hours were required for the two-stage study. These times were tallied by having S-Plus call upon the computer's clock

time at the beginning and again at the conclusion of each configuration under study. It is worthy to note that in 55.5% of the simulations, the two-stage methodology terminated with a stopping simulation size that was smaller than that required by the corresponding sequential methodology.

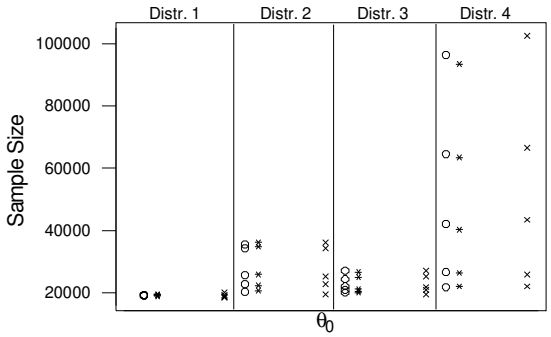

Figure 5.6. Variation in Two-Stage Simulation or Sample Size M by θ_0 for Distributions $F_i, i = 1, 2, 3, 4$, with Risk $c = 0.001$: Circles, Asteriks and Crosses Correspond to $\theta_0 = 0.06, 0.10, 0.30$ Respectively.

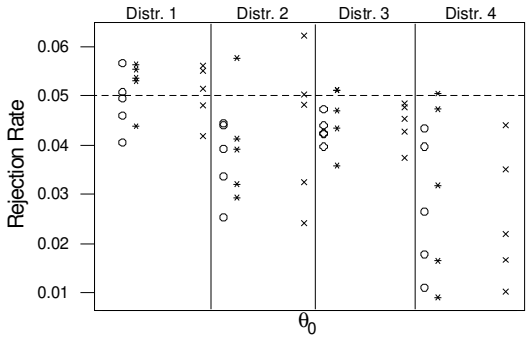

Figure 5.7. Variation in Rejection Rate of Two-Stage Procedure Due to Simulation Size M by θ_0 for Distributions $F_i, i = 1, 2, 3, 4$, with Risk-Bound $c = 0.1$: Circles, Asteriks and Crosses Correspond to $\theta_0 = 0.06, 0.10, 0.30$ Respectively.

However, in 78.875% of the configurations we found that the two-stage methodology completed its task in less time than the corresponding sequential methodology. Based on the arguments made earlier in the last part of Section 4.1, we fully recommend the use of the two-stage methodology (4.1)-(4.2) developed in this paper over the sequential procedure (3.2)-(3.3) for estimating a proportion that is likely to be near zero (or one) via simulation.

Throughout the ten repetitions of our 240 configurations, the purely sequential methodology (3.2)-(3.3) had terminated simulations at the initial value $k_0 = 100$ in 13 out of the 2400 (that is, 0.5417%) configurations, whereas the two-stage methodology (4.1)-(4.2) terminated simulations analogously in 10 out of the 2400 (that is,

0.4167%) configurations. In all such cases we found that the risk was set at the level $c = 0.1$. This prompted us to increase k_0 to 200 and observe any changes. For brevity, however, we ran one repetition of the 240 configurations holding $k_0 = 200$ fixed and found that 7.9% (that is, 19 out of the 240 configurations) of the sequential methodologies and 10.8% (that is, 26 out of the 240 configurations) of the two-stage methodologies terminated with the initial simulation size, $k_0 = 200$. All of these cases occurred when the risk was fixed at the level $c = 0.1$.

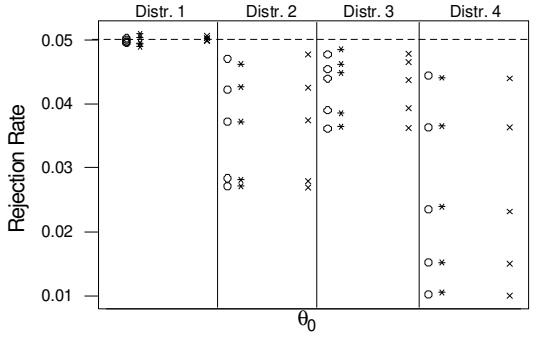

Figure 5.8. Variation in Rejection Rate of Two-Stage Procedure Due to Simulation Size M by θ_0 for Distributions $F_i, i = 1, 2, 3, 4$, with Risk-Bound $c = 0.001$: Circles, Asteriks and Crosses Correspond to $\theta_0 = 0.06, 0.10, 0.30$ Respectively.

One may like to extrapolate what we have just said and consider increasing k_0 even more when $c = 0.1$. But, if that is done, then our proposed stopping rules would terminate at the initial stage with k_0 simulations even more frequently and $\widehat{\theta}$ may not be that great afterall! On the other hand, one must realize that $c = 0.1$ happens to be the largest value of c that we had chosen. For this choice of c and $k_0 = 100$ or 200, one can see that $\left|\widehat{\theta} - \theta\right|$ may be "approximated" by $\sqrt{c}\theta = 0.316\theta$. At this point, rather than worrying too much about the choice of k_0, the experimenter should perhaps do some soul searching and think whether the error bound $c = 0.1$ is "good enough" to live with or not. Perhaps, c should be considerably lowered! Statistically designing an experiment is a great idea, but statisticians just cannot take care of every aspect of an experiment in every field of investigation. We would hate to play "God" by pretending to provide "solutions" for all conceivable "statistical conflicts". The fact is that whatever c, θ_0 and k_0 one may decide to fix, very early stopping may occasionally arise. A suitable "resolution" or "balance" may be reached only through teamwork. For example, one may like to try other choices of c, θ_0 and k_0 and critically "examine" differences that may occur in the performance of computer simulations. That ought to guide many experimenters to come close to a practical solution for this type of problem on a case by case basis.

From a practical perspective, we would not advocate estimating rejection rates or proportions near zero with a risk-bound as large as 0.1. Hence, we venture to say that our findings indicate very minimal influence of the choice of k_0 on the methodologies provided the risk-bound c is fixed at a realistically small level.

Since there is some chance that N or M may not go past the pilot phase, an experimenter would need to settle down with a reasonable choice of k_0. The choice of k_0 could be made differently depending upon the substantive field of application.

For example, while using computer simulations in biology or electrical engineering, the choice of k_0 would very likely be different. Statistical considerations alone can not settle such issues. But, generally speaking, within a simulation exercise, we recommend that one should really go by the smallest possible k_0 that one can live with. In many situations, k_0 would probably not exceed 100 or 200. Should more observations be needed, one would let the stopping rule (3.3) or (4.2) determine the appropriate final simulation size. We take solace knowing that estimates have been found under the bounded risk criterion. Such a statement can not be made when simulations are carried out by haphazardly implementing a fixed number of runs.

There is truly no substitute of a well-designed simulation study.

6. LARGE-SAMPLE PROPERTIES

From (3.3), one can claim easily that $N(c_1) \geq N(c_2)$ with probability one if $0 < c_1 < c_2$. That is, with a smaller choice of c, the resulting simulation size can not go down. Also, from the stopping rules (3.3) and (4.2), we note immediately that $\min\{N(c), M(c)\} \geq k(c)$ and thus $\lim_{n^*(\theta)\to\infty} \min\{N(c), M(c)\} \geq \lim_{n^*(\theta)\to\infty} k(c) = \infty$ with probability one. So, we can easily claim that $\lim_{n^*(\theta)\to\infty} N(c) = \infty$ and $\lim_{n^*(\theta)\to\infty} M(c) = \infty$ with probability one. Hence, both randomly stopped estimators $\overline{U}_N, \overline{U}_M$ of θ converge to θ with probability one (and hence, in probability) as $n^*(\theta) \to \infty$. One may refer to Hoeffding (1961) or Sproule (1974).

Now, we are in a position to summarize the main large-sample results associated with both the methodologies (3.2)-(3.3) and (4.1)-(4.2).

Theorem 6.1. *Let $T \equiv T(c)$ be a generic stopping variable where T stands for N (or M) defined in (3.3) (or (4.2)). Recall the expression of $n^*(\theta)$ from (2.9). Then, as $n^*(\theta) \to \infty$, we conclude*:

(i) $T/n^*(\theta) \to 1$ *in probability* (P_θ);

(ii) $E_\theta[T/n^*(\theta)] \to 1$;

(iii) $E_\theta[L_T/c] \to 1$.

for all fixed $0 < \theta < 1$.

Its proof in the case of the sequential and two-stage methodologies are sketched in Sections 6.1 and 6.2 respectively. The properties (i)-(ii) reinforce the feeling that we can expect the final simulation size N or M to be very close to the unknown fixed sample size $n^*(\theta)$. The property (iii) states that the sequential risk $E[L_N]$ or the two-stage risk $E[L_M]$ for estimating θ by $\overline{U}_N$ or $\overline{U}_M$ respectively may be expected to be close to the set target c.

6.1. The Sequential Methodology (3.2)-(3.3): Proof of Theorem 6.1

From (3.3), observe the basic inequality

$$c^{-1}\frac{(1-\overline{U}_N-N^{-1})}{(\overline{U}_N+N^{-1})} \leq N < c^{-1}\frac{(1-\overline{U}_{N-1}-(N-1)^{-1})}{(\overline{U}_{N-1}+(N-1)^{-1})} + (k-1)I(N=k)$$

$$\Rightarrow \frac{(1-\overline{U}_N-N^{-1})}{(\overline{U}_N+N^{-1})}\frac{\theta}{1-\theta} \leq \frac{N}{n^*(\theta)} < \frac{(1-\overline{U}_{N-1}-(N-1)^{-1})}{(\overline{U}_{N-1}+(N-1)^{-1})}\frac{\theta}{1-\theta} + \frac{k-1}{n^*(\theta)} I(N=k). \quad (6.1)$$

We noted earlier that $\lim_{n^*(\theta)\to\infty} N = \infty$ with probability one. So, the randomly stopped estimator $\overline{U}_N$ of θ converges to θ with probability one (and hence, in probability) as $n^*(\theta) \to \infty$. One may refer to Hoeffding (1961), Khan (1969) or Sproule (1974). Also, one notes that $\frac{k-1}{n^*(\theta)}$ is $O(1)$. Thus, part (i) will follow from (6.1) if we verify that

$$I(N=k) \xrightarrow{P} 0 \text{ as } n^*(\theta) \to \infty. \quad (6.2)$$

Now, let us pretend that $n^*(\theta)$ goes to infinity in such a way that $c^{-1}\frac{1-\theta_0}{\theta_0}$ remains a positive integer and that $n^*(\theta)$ is large enough so that $k = c^{-1}\frac{1-\theta_0}{\theta_0}$. Thus, for all fixed $0 < \theta < 1$, we can write

$$P_\theta\{N=k\} \leq P_\theta\left\{k \geq c^{-1}\frac{(1-\overline{U}_k-k^{-1})}{(\overline{U}_k+k^{-1})}\right\} = P_\theta\{\overline{U}_k \geq \theta_0\}$$

and hence for any fixed $p(>0)$, we obtain

$$\begin{aligned} P_\theta\{N=k\} \leq P_\theta[\left|\overline{U}_k-\theta\right| \geq \theta-\theta_0] &\leq (\theta-\theta_0)^{-2p}E_\theta[\left|\overline{U}_k-\theta\right|^{2p}] \\ &= O(k^{-p}) = O([n^*(\theta)]^{-p}), \end{aligned} \quad (6.3)$$

with the help of the Markov inequality. Next, with an arbitrary but fixed $\varepsilon > 0$, we can apply the Markov inequality one more time and express

$$\begin{aligned} P_\theta\{I(N=k) > \varepsilon\} \leq \varepsilon^{-1}E_\theta\{I(N=k)\} = \varepsilon^{-1}P_\theta\{N=k\} \to 0 \\ \text{as } n^*(\theta) \to \infty, \text{ in view of (6.3).} \end{aligned} \quad (6.4)$$

The last equation verifies (6.2) and hence part (i) holds.

Next, we turn to part (ii). For this, we closely follow Woodroofe (1982, Chapter 4) and Ghosh et al. (1997, Section 2.9) to rewrite (3.3) as follows:

$$\begin{aligned} N \equiv N(c) \text{ is the integer } n(\geq k) \text{ for which we observe} \\ ng(\overline{U}_n+n^{-1}) \geq c^{-1} \text{ for the first time where } g(x) = \tfrac{x}{1-x}. \end{aligned} \quad (6.5)$$

Let us denote $S_n = \Sigma_{i=1}^n \left\{\frac{\theta}{1-\theta} + \frac{1}{(1-\theta)^2}(U_i-\theta)\right\}, n \geq 1$. But, note that $g'(x) = (1-x)^{-2}$ and $g''(x) = (1-x)^{-3}$ so that we have

$$\begin{aligned} ng(\overline{U}_n+n^{-1}) &= n\left(\tfrac{\theta}{1-\theta}+n^{-1}\right) + n\tfrac{1}{(1-\theta)^2}(\overline{U}_n+n^{-1}-\theta) + \tfrac{n(\overline{U}_n+n^{-1}-\theta)^2}{(1-\nu_n)^3} \\ \Leftrightarrow ng(\overline{U}_n+n^{-1}) &= S_n+\xi_n, \text{ with } \xi_n = \tfrac{n(\overline{U}_n+n^{-1}-\theta)^2}{(1-\nu_n)^3} + \left\{1+\tfrac{1}{(1-\theta)^2}\right\}, \end{aligned} \quad (6.6)$$

where the suitable random variable ν_n is such that it lies between $\min\{\overline{U}_n, \theta\}$ and $\max\{\overline{U}_n, \theta\}$ for large enough n.

Surely, ξ_n is *slowly changing* (Ghosh et al. (1997, p. 59) or Woodroofe (1982, p.41)). Since $n^{-1}\xi_n$ converges to zero with probability one, we can claim (Lemma 2.9.1 from Ghosh et al. (1997, p. 59) or (4.1) from Woodroofe (1982, p.41)) that $\max_{1\leq k\leq n} n^{-1}|\xi_k|$ also converges to zero with probability one as $n \to \infty$. Thus,

the two sufficient conditions in the Theorem 2.9.3 of Ghosh et al. (1997, p. 62) or Theorem 4.4 in Woodroofe (1982, p.46) hold and hence part (ii) immediately follows.

Now, we are in a position to verify part (iii). Let us denote $Q_N = \Sigma_{i=1}^{N} U_i, W_N = \frac{(Q_N - N\theta)}{\sqrt{\theta(1-\theta)n^*(\theta)}}$, and observe that the sequential risk function $E_\theta[L_N]$ can be rewritten as

$$E_\theta\left[\tfrac{(Q_N-N\theta)^2}{\theta^2 N^2}\right] = E_\theta\left[\tfrac{(Q_N-N\theta)^2}{\theta^2 (n^*(\theta))^2}\right] + E_\theta\left[\tfrac{(Q_N-N\theta)^2}{\theta^2 (n^*(\theta))^2}\left\{\tfrac{(n^*(\theta))^2}{N^2} - 1\right\}\right]. \tag{6.7}$$

Then, using Wald's second lemma (Theorem 2.4.5, Ghosh et al. (1997, p. 27)) and part (ii), we have

$$E_\theta\left[W_N^2\right] = \tfrac{E_\theta(N)\theta(1-\theta)}{\theta(1-\theta)n^*(\theta)} = \tfrac{E_\theta(N)}{n^*(\theta)} \to 1 \text{ as } n^*(\theta) \to \infty. \tag{6.8}$$

Next, using Anscombe's (1952) central limit theorem we can claim that W_N converges in distribution to a standard normal variable as $n^*(\theta) \to \infty$. Now, we can combine this result with (6.8) to conclude that W_N^2 is then uniformly integrable.

Then, from (6.7) we have

$$E_\theta[c^{-1}L_N] = \tfrac{\theta n^*(\theta)}{(1-\theta)} E_\theta[L_N] = E_\theta\left[W_N^2\right] + E_\theta\left[W_N^2\left\{\tfrac{(n^*(\theta))^2}{N^2} - 1\right\}\right]. \tag{6.9}$$

But, since $N \geq k(c)$, observe that for sufficiently large $n^*(\theta)$, we can write

$$W_N^2\left\{\tfrac{(n^*(\theta))^2}{N^2} - 1\right\} \leq W_N^2\left\{\tfrac{(n^*(\theta))^2}{k(c)^2} - 1\right\} = a_{\theta,\theta_0} W_N^2. \tag{6.10}$$

Here, a_{θ,θ_0} is a positive number that does not involve c. Now since W_N^2 is uniformly integrable, it follows from (6.10) that $W_N^2\left\{\tfrac{(n^*(\theta))^2}{N^2} - 1\right\}$ is also uniformly integrable. Now, we know that W_N^2 converges in distribution to a χ_1^2 random variable as $n^*(\theta) \to \infty$ and thus combining with part (i), we can claim that $E_\theta\left[W_N^2\left\{\tfrac{(n^*(\theta))^2}{N^2} - 1\right\}\right] \to 0$ as $n^*(\theta) \to \infty$. At this point, one simply combines (6.8)-(6.9) to complete the proof of part (iii). ■

6.2. The Two-Stage Methodology (4.1)-(4.2): Proof of Theorem 6.1

From (4.2), observe the basic inequality

$$\begin{aligned} & c^{-1}\tfrac{(1-\overline{U}_k - k^{-1})}{(\overline{U}_k + k^{-1})} \leq M < c^{-1}\tfrac{(1-\overline{U}_k - k^{-1})}{(\overline{U}_k + k^{-1})} + kI(M=k) \\ \Rightarrow & \tfrac{(1-\overline{U}_k - k^{-1})}{(\overline{U}_k + k^{-1})}\tfrac{\theta}{1-\theta} \leq \tfrac{M}{n^*(\theta)} < \tfrac{(1-\overline{U}_k - k^{-1})}{(\overline{U}_k + k^{-1})}\tfrac{\theta}{1-\theta} + \tfrac{k}{n^*(\theta)} I(M=k). \end{aligned} \tag{6.11}$$

We noted earlier that $\lim_{n^*(\theta)\to\infty} k = \infty$, but $k = O(n^*(\theta))$. Using the weak law of large numbers it follows that the estimator $\overline{U}_k$ of θ converges to θ in probability as $n^*(\theta) \to \infty$. Also, essentially the earlier proof of (6.2) can be cited here to claim that $I(M=k) \xrightarrow{P} 0$ as $n^*(\theta) \to \infty$. Thus, part (i) will follow from (6.11).

Next, we turn to part (ii). Along the line of (6.3), one can easily verify that $P_\theta\{M=k\} = O([n^*(\theta)]^{-p})$ for any fixed $p(>0)$. Also, Fatou's lemma and part

(i) together will let us conclude that $\liminf_{n^*(\theta)\to\infty} E_\theta[\frac{M}{n^*(\theta)}] \geq E_\theta[\liminf_{n^*(\theta)\to\infty} \frac{M}{n^*(\theta)}] = 1$. Thus, in view of the upper bound given by (6.11), part (ii) will immediately follow if we can verify the following result:

$$\lim_{n^*(\theta)\to\infty} E_\theta\left[\frac{1}{\overline{U}_k+k^{-1}}\right] = \frac{1}{\theta} \text{ for every fixed } 0<\theta<1. \tag{6.12}$$

Now, with some appropriate random variable ψ_k between θ and $\overline{U}_k + k^{-1}$, we can express

$$\begin{aligned} & E_\theta\left[\frac{1}{\overline{U}_k+k^{-1}}\right] \\ &= \frac{1}{\theta+k^{-1}} - \frac{1}{\theta^2}E_\theta\left[(\overline{U}_k-\theta+k^{-1})\right] + E_\theta\left[(\overline{U}_k-\theta+k^{-1})^2\frac{1}{\psi_k^3}\right] \\ &= \frac{1}{\theta+k^{-1}} - \frac{1}{k\theta^2} + E_\theta\left[(\overline{U}_k-\theta+k^{-1})^2\frac{1}{\psi_k^3}\right]. \end{aligned} \tag{6.13}$$

But, on the set where $\overline{U}_k + k^{-1} > \frac{1}{2}\theta$ is observed, we must have $\psi_k > \frac{1}{2}\theta$ so that with the help of the Cauchy-Schwartz inequality we can write

$$\begin{aligned} & E_\theta\left[(\overline{U}_k-\theta+k^{-1})^2\frac{1}{\psi_k^3}I(\overline{U}_k+k^{-1}>\tfrac{1}{2}\theta)\right] \\ & \qquad \leq \frac{8}{\theta^3}E_\theta[(\overline{U}_k-\theta+k^{-1})^4] = o(1). \end{aligned} \tag{6.14}$$

Also, along the line of (6.3), one can easily verify that $P_\theta\{\overline{U}_k \leq \frac{1}{2}\theta\} = O(k^{-p})$ for any fixed $p(>0)$. Thus, for sufficiently large k, on the set where $\overline{U}_k + k^{-1} \leq \frac{1}{2}\theta$ is observed, we must have $\psi_k > \overline{U}_k + k^{-1} \geq k^{-1}$ so that

$$\begin{aligned} & E_\theta\left[(\overline{U}_k-\theta+k^{-1})^2\frac{1}{\psi_k^3}I(\overline{U}_k+k^{-1}\leq\tfrac{1}{2}\theta)\right] \\ & \quad \leq k^3 E_\theta[(\overline{U}_k-\theta+k^{-1})^2 I(\overline{U}_k+k^{-1}\leq\tfrac{1}{2}\theta)] \\ & \quad = k^3 E_\theta^{1/2}[(\overline{U}_k-\theta+k^{-1})^4]P_\theta^{1/2}[\overline{U}_k\leq\tfrac{1}{2}\theta] \\ & \quad = k^3 O(k^{-1})O(k^{-p}) \\ & \quad = o(1), \text{ if } p \text{ is chosen to exceed 2.} \end{aligned} \tag{6.15}$$

At this point, combining (6.13)-(6.15), we immediately conclude the validity of (6.12), which in turn completes the proof of part (ii).

The part (iii) can be verified in exactly the same way it was handled in the sequential case. ■

ACKNOWLEDGMENTS

The authors wish to thank Timothy Mills of Rutgers University's Computer Services Department for conversations on comparing computers running at different speeds. Two referees independently pointed out some confusing statements and caught a number of typographical errors in a previous version that needed substantial tightening. Their critical assessments have led to a considerably clearer presentation for which we remain grateful and we take this opportunity to thank them.

REFERENCES

[1] Anscombe, F. J. (1952). Large sample theory of sequential estimation. *Proc. Camb. Philos. Soc.*, **48**, 600-607.

[2] Aoshima, M. (2000). Second-order properties of improved two-stage procedure for a multivariate normal distribution. *Commun. Statist. Theory & Methods*, **29**, 611-622.

[3] Aoshima, M. and Mukhopadhyay, N. (1999). Second-order properties of a two-stage fixed-size confidence region when the covariance matrix has a structure. *Commun. Statist. Theory & Methods*, **28**, 839-855.

[4] Bratley, P., Fox, B. L. and Schrage, L. E. (1988). *A Guide to Simulation*, 2^{nd} edition. Springer Verlag, New York.

[5] Chow, Y. S. and Robbins, H. (1965). On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *Ann. Math. Statist.*, **36**, 457-462.

[6] Ghosh, M., Mukhopadhyay, N. and Sen, P. K. (1997). *Sequential Estimation.* Wiley and Sons, New York.

[7] Hoeffding, W. (1961). The strong law of large numbers for U-statistics. *Univ. of North Carolina Inst. of Statist. Mimeo Series*, No. 302.

[8] Khan, R. A. (1969). A general method of determining fixed-width confidence intervals. *Ann. Math. Statist.*, **40**, 704-709.

[9] Martinsek, A. T. (1983). Sequential estimation with squared relative error loss. *Bull. Inst. Math. Acad. Sinica*, **11**, 607-623.

[10] Martinsek, A. T. (1995). Estimating a slope parameter in regression with prescribed proportional accuracy. *Statistics & Decisions*, **13**, 363-377.

[11] Mukhopadhyay, N. (1999). Second-order properties of a two-stage fixed-size confidence region for the mean vector of a multivariate normal distribution. *J. Multivar. Anal.*, **68**, 250-263.

[12] Mukhopadhyay, N. and Diaz, J. (1985). Two-stage sampling for estimating the mean of a negative binomial distribution. *Sequential Anal.*, **4**, 1-18.

[13] Mukhopadhyay, N. and Duggan, W. T. (1997). Can a two-stage procedure enjoy second-order properties? *Sankhya, Ser. A*, **59**, 435-448.

[14] Nadas, A. (1969). An extension of a theorem of Chow and Robbins on sequential confidence intervals for the mean. *Ann. Math. Statist.*, **40**, 667-671.

[15] Posten, H. O., Yeh, H. C., and Owen, D. B. (1982). Robustness of the two-sample t-test under violations of the homogeneity of variance assumption. *Commun. Statist. Theory & Methods*, **11**, 109-126.

[16] Robbins, H. and Siegmund, D. (1974). Sequential estimation of p in Bernoulli trials. In *Studies in Probab. Statist., E. J. G. Pitman Volume* (E. J. Williams, ed.), 103-107. Jerusalem Academic Press, Jerusalem.

[17] Ross, S. M. (1999). *Simulation*, 2^{nd} edition. American Press, San Diego.

[18] Sproule, R. N. (1974). Asymptotic properties of U-statistics. *Trans. Amer. Math. Soc.*, **199**, 55-64.

[19] Sriram, T. N. (1991). Second order approximation to the risk of a sequential procedure measured under squared relative error loss. *Statistics & Decisions*, **9**, 375-392.

[20] Stein, C. (1945). A two sample test for a linear hypothesis whose power is independent of the variance. *Ann. Math. Statist.*, **16**, 243-258.

[21] Stein, C. (1949). Some problems in sequential estimation (abstract). *Econometrica*, **17**, 77-78.

[22] Tiku, M. L. and Singh, M. (1981). Robust test for means when population variances are unequal. *Commun. Statist. Theory & Methods*, **10**, 2143-2159.

[24] Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*, (I. Olkin, ed.).Stanford University Press, California.

[25] Tukey, J. W. and McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization. *Sankhya*, **25**, 331-352.

[26] Willson, L. J. and Folks, J. L. (1983). Sequential estimation of the mean of the negative binomial distribution. *Sequential Anal.*, **2**, 55-70.

[27] Woodroofe, M. (1982). *Nonlinear Renewal Theory in Sequential Analysis*. SIAM, Philadelphia.

Addresses for Communications:

NITIS MUKHOPADHYAY, Department of Statistics, UBox 4120, University of Connecticut, Storrs, CT 06269-4120, U.S.A. E-mail: mukhop@uconnvm.uconn.edu; Web site: http://merlot.stat.uconn.edu/~nitis/
GREG CICCONETTI, Department of Mathematical Sciences, Muhlenberg College, Allentown, PA 18104, U.S.A. E-mail: cicconet@muhlenberg.edu; Web site: http://mathcs.muhlenberg.edu/~cicconet/