A SEQUENTIAL STOPPING RULE FOR DETERMINING THE NUMBER OF REPLICATIONS NECESSARY WHEN SEVERAL MEASURES OF EFFECTIVENESS ARE OF INTEREST

October 2004

Anthony J. Quinzi
TRADOC Analysis Center
White Sands Missile Range, NM

Part I. Stopping Rule

1. Introduction

Historically, TRAC analysts have relied on a fixed-sample-size[1] procedure (the "n = 21 rule-of-thumb") to estimate the mean value μ of an output measure of battle effectiveness. For example, μ may represent the mean number of friendly losses.[2] The "n = 21 rule-of-thumb" is based on the assumptions that the replications are independent and produce a sequence of independent, identically distributed random variables $X_1, X_2, X_3, ..., X_n$**.** Confidence intervals and tests of hypothesis can then be obtained based on an application of the Central Limit Theorem, namely that for n sufficiently large, the distribution of the random variable

$$\frac{\overline{X}}{s_n / \sqrt{n}} \qquad \textbf{(1.1)}$$

is approximately normally distributed, where $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $s_n = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}}$. It follows that for sufficiently large n, an *approximate* $100 \times (1 - \alpha)\%$ confidence interval for μ is given by

$$\overline{X} \pm t_{n-1,1-\alpha/2} \sqrt{\frac{s_n^{\ 2}}{n}}, \qquad \textbf{(1.2)}$$

where $0 < \alpha < 1$ and $t_{n-1,1-\alpha/2}$ is the upper $1 - \alpha/2$ critical point for the t distribution on $n - 1$ degrees of freedom. A value typically chosen for α is .05 yielding a confidence level of $(1 - \alpha)$ or 95%. If it is further assumed that $X_1, X_2, X_3, ..., X_n$ are *normally* distributed, it follows that the confidence interval (1.2) is *exact for any sample size n* > 1. One drawback of the fixed-sample-size approach is that the analyst has no control on the precision of the estimate $\overline{X}$.

2. Notions of precision

Law and Kelton [2000], hereafter referred to as LK [2000], define a number of ways of measuring the error in $\overline{X}$. Suppose that n replications resulted in a mean $\overline{X}$ = 99.7 when the (unknown) true value of $\mu = 100$. The *absolute error* of estimation β would be

$$\beta = \left|\overline{X} - \mu\right|$$

or 0.3. The *relative* error of estimation γ would be

$$\gamma = \frac{\left|\overline{X} - \mu\right|}{\mu}$$

or 0.003 which can be thought of as a *percentage error* of 0.3% in $\overline{X}$.

The sample mean $\overline{X}$ of a random sample of size n from a population with mean μ and standard deviation σ has a standard deviation which can be estimated by $s_n / \sqrt{n}$, where s_n is defined as above. Because s_n is a consistent estimator of the population variance, the sample mean becomes "stable" for large n. By specifying a degree of accuracy, say relative error, the

[1] That is, a fixed number n of replications

[2] Generally, analysts are not in the business of obtaining precise estimates of battle parameters for the sake of estimation alone, but rather to be able to compare these estimates across study alternatives.

analyst is able to formulate a stopping rule for a sequence of replications and be assured that with high probability (specifically, 1 – α), the sample mean has been estimated within the specified degree of accuracy.

3. The work of Cherolis

Cherolis [1992] suggested a sequential procedure based on (1.2) to determine the number of replications necessary to estimate μ with a specified degree of accuracy. The procedure is in the form of a stopping rule which can determine, after a small number of replications have been performed, how many *subsequent* replications are necessary to be able to estimate μ with a specified accuracy. One drawback of Cherolis' result is that it applies to a single measure of effectiveness. Because of recent simulation work involving the Army's Future Combat System (FCS), it is of interest to determine a stopping rule that can determine how many replications are necessary to be able to estimate a number of output parameters simultaneously.

4. The case for a single measure of effectiveness

LK [2000] suggest the following *sequential* procedure for obtaining an estimate of μ with a specified relative error of γ, $0 < \gamma < 1$, that takes only as many replications as are actually needed:

Suppose X_1, X_2, X_3, ... is a sequence of independent, identically distributed random variables. It is important to note that the random variables need not be normally distributed. These may represent, for example, the numbers of friendly losses in replications 1, 2, 3, etc.. Choose an initial number of replications $n_0 \geq 2$. The actual number chosen should depend on the amount of replication-to-replication variability. If the variability is not large, then $n_0 = 5$ replications may be sufficient. If the variability is large, then at least $n_0 = 10$ replications should be made. The choice of relative precision γ may have to be adjusted when there are not sufficient resources to perform the required number of replications.[3]

Step 1. Make n_0 replications of the simulation and set $n = n_0$.

Step 2. Compute $\overline{X}$ and the quantity $\delta(n,\alpha) = t_{n-1,1-\alpha/2}\sqrt{\frac{s^2}{n}}$, where s is defined in (1.1) and the level of confidence is $100\times(1-\alpha)$ %, $0 < \alpha < 1$.

Step 3. If $\delta(n,\alpha)/\left|\overline{X}\right| \leq \gamma'$, where $\gamma' = \gamma/(1+\gamma)$, use $\overline{X}$ as the point estimate for μ and stop. If $\alpha = .05$, for example, then the interval

$$I(.05,\gamma) = [\overline{X} - \delta(n,.05), \overline{X} + \delta(n,.05)]$$

is an approximate 95% confidence interval for μ with the desired precision. If the inequality fails, replace n by $n + 1$, make one additional replication of the simulation, go to step 2 and repeat the process.

[3] LK[2000] show that it is possible to obtain *rough estimates* (a table of values) of the number of replications required to estimate μ with desired levels γ of relative precision.

5. The case of multiple measures of effectiveness

Let μ_1, ..., μ_k represent the means of k measures of effectiveness. For each mean μ_s, a $(1 - \alpha_s) \times 100\%$ confidence interval is determined, $s = 1, ..., k$. Suppose that $\sum_{s=1}^{k} \alpha_s = \alpha$. Then the joint probability that *all k* confidence intervals *simultaneously* contain their respective true means is at least

$$\mathbf{1 - \sum_{s=1}^{k} \alpha_s} \ . \qquad \mathbf{(5.1)}$$

This result is known as the Bonferroni inequality and (5.1) is called the Bonferroni bound. It should be noted that the α_s need not be equal. For example, given four measures of effectiveness, suppose that a 99% confidence interval were computed for μ_1, a 98% confidence interval were computed for μ_2, a 97% confidence interval were computed for μ_3 and a 96% confidence interval were computed for μ_4. In this case, it may be that the first measure ($s = 1$) is most important and so the highest level of confidence (99%) is chosen for that measure. If the confidence level is 99%, then $\alpha_1 = 1 - .99 = .01$. For confidence level 98%, $\alpha_2 = 1 - .98 = .02$, and so on. Using the Bonferroni bound (5.1), the joint probability that all 4 confidence intervals simultaneously contain their respective means would be at least $[1 - (.01 + .02 + .03 + .04)]$ or 0.90. In order to extend the above stopping rule to include multiple measures of effectiveness, it is necessary to specify a relative precision for each measure. The 3-step procedure outlined above would have to be performed for *each* measure. At any stage of the process, it may occur that some measures require an additional replication and some not. The procedure will stop when *every* inequality in Step 3 of the above procedure holds. Because the procedure requires more data than in the case of a single measure of effectiveness, LK[2000] recommend that the number k of measures be no greater than 10.

Part II. Measures of Effectiveness

The following four measures of performance were of interest: friendly (BLUE) system (vehicle) losses, friendly individual soldier (dismounted) losses, threat (RED) system (vehicle) losses, and threat individual soldier (dismounted) losses. It was desired to apply the stopping rule to all four measures simultaneously.

Part III. Application

Because one replication of the full Caspian scenario takes approximately 60 hours of computer run time, it was recommended that the sequential procedure suggested in Part I be tested in scaled down version of the same scenario whose run time is considerably less, about 6 hours. The sequential procedure was tabulated in an Excel spreadsheet. A portion of the spreadsheet is reproduced here.

		Reference: Law & Kelton Ed. 3, pp. 513-514									
REP					Blue Losses Vehicles		Blue Losses Dismounts		Red Losses Vehicles		Red Losses Dismounts
1					35		32		75		110
2					43		29		83		133
3					41		32		81		122
4					32		43		82		141
5					47		24		75		115
6					38		35		81		118
7					34		36		83		122
				delta(n, a)		delta(n, a)		delta(n, a)		delta(n, a)	
	d1	2.969	80% (Each .05)	6.04	CONTINUE	6.67	CONTINUE	3.94	STOP	11.98	CONTINUE
	d2	3.707	92% (Each .02)	7.54	CONTINUE	8.33	CONTINUE	4.92	STOP	14.96	CONTINUE
	d3	4.317	96% (Each .01)	8.78	CONTINUE	9.70	CONTINUE	5.73	STOP	17.42	CONTINUE
	d4	5.959	99.2% (Each .002)	12.12	CONTINUE	13.39	CONTINUE	7.91	CONTINUE	24.05	CONTINUE
8					39		60		78		112
				delta(n, a)		delta(n, a)		delta(n, a)		delta(n, a)	
	d1	2.841	80% (Each .05)	5.01	CONTINUE	11.07	CONTINUE	3.34	STOP	10.67	STOP
	d2	3.499	92% (Each .02)	6.17	CONTINUE	13.63	CONTINUE	4.12	STOP	13.14	CONTINUE
	d3	4.029	96% (Each .01)	7.10	CONTINUE	15.70	CONTINUE	4.74	STOP	15.13	CONTINUE
	d4	5.408	99.2% (Each .002)	9.53	CONTINUE	21.07	CONTINUE	6.36	STOP	20.31	CONTINUE
9					52		31		87		118
				delta(n, a)		delta(n, a)		delta(n, a)		delta(n, a)	
	d1	2.752	80% (Each .05)	5.92	CONTINUE	9.60	CONTINUE	3.61	STOP	9.18	STOP
	d2	3.355	92% (Each .02)	7.21	CONTINUE	11.70	CONTINUE	4.41	STOP	11.20	CONTINUE
	d3	3.833	96% (Each .01)	8.24	CONTINUE	13.36	CONTINUE	5.03	STOP	12.79	CONTINUE
	d4	5.041	99.2% (Each .002)	10.84	CONTINUE	17.58	CONTINUE	6.62	STOP	16.82	CONTINUE

Explanation of Spreadsheet Calculations

1. We begin with an initial n_0 = 7 replications and test the inequality in Step 3.
2. The inequality is tested for each parameter. I tried four different experimentwise values for α: .2, ,08, .04 and 0.008 for use in the Bonferroni bound. The corresponding critical values for the *t*–distribution are listed in the next column.
3. The word "CONTINUE" appears in green if the respective inequality fails, and the word "STOP" appears in red, otherwise.
4. The quantities delta(n, a) refer to the quantities $\delta(n,\alpha)$ in the sequential procedure.

Replications were continued until *n* = 30, and the inequality for measure 2, friendly individual soldier losses, was never realized. It is clear that the most variable of a collection of parameters will always be the one which determines the necessary sample size. This was not a satisfactory result for the scaled down scenario, and would have been impossible to determine for the full blown scenario because of time limitations. There has to be a better way.

Questions for the Panel Members

1. **Given that we are dealing with purely discrete distributions (numbers of losses) each of whose underlying distributions results from a large number of random draws in the model, should we really be considering a procedure based on the *t*-distribution which assumes population is Gaussian, especially in view of small sample sizes?**

2. Should we instead be trying to estimate the underlying discrete probability mass functions [cf. Chiu, S.T. (1991) "Bandwidth selection for kernel density estimation", *Ann. Stat. Vol. 19*, pp. 1883-1905] or use some other methodology so that we might be able to improve on the Bonferroni bounds?

3. Are bootstrap methods really appropriate in a PURELY discrete context such as this?

4. Is this question crying for some sort of Bayesian approach?

REFERENCES

Chelis, George T. (1992). A Sequential Stopping Rule for Reducing Production Times during CASTFOREM Studies, Master's Thesis, New Mexico State University, Las Cruces, NM.

Chiu, S. T. (1991). Bandwidth selection for kernel density estimation, *Ann. Stat. Vol. 19*, pp. 1883-1905.

Law, Averill M. and Kelton, W. David (2000). *Simulation Modeling and Analysis, Third Edition*, McGraw-Hill, Boston, Massachusetts.