A non-monotonic convergence analysis of population clusters of random numbers

B. E Obabueki, S. A. Reju.
Department of Mathematics and Statistics. Polytechnic of Namibia.

Abstract

The standard deviation of a population (of size N) is a measure of the spread of the population observations about the mean. A population may be clustered and the standard deviation of each cluster calculated. This paper looked at how the mean of the standard deviations of the clusters of a population of random numbers relate to the standard deviation of the population as the size of the clusters increased. We assumed that all clusters have the same size. As the size n of each cluster increased, the number of clusters $\frac{N}{n}$ decreased, making the population a single cluster when $n = N$. The sequence $\{S_n\}$ of the means of the standard deviations of the clusters converged to the standard deviation S_N of the population. However, this convergence was not monotonic.

Keywords: *standard deviation of a population, population clusters, sequence of means of standard deviations, short-cut estimates, proximity, cluster size, estimation of standard deviation, randomly generated numbers, non-monotonic convergence, convergence simulation.*

Introduction and literature review

Discussions have been on-going about the determination and usage of the standard deviation of a population. Many authors have expressed themselves on the relationship between the standard deviation of a single sample of a population and the standard deviation of the population as the sample size increased. This paper aimed at determining how the size of each cluster affected the proximity between the mean of the standard deviations of the clusters of a population of 5040 randomly generated numbers and the standard deviation of the population.

Short-cut estimates of the standard deviation of a population have their advantages and shortcomings. Sabers and Klausmeier [1] investigated the

accuracy of some short-cut estimates of standard deviation. They found that the loss in accuracy due to short-cut methods versus the conventional method ranged from 0% to 7.8%.

On the other hand, Hargreaves and Samani [2] had the following to say:
A weather simulation procedure utilizing a monthly climatic data base can be substituted for the daily climatic data to produce very comparable results. The weather simulation procedure requires the standard deviation of potential evapotranspiration (ETP). A series of monthly mean values of maximum and minimum temperatures provides the required data for estimating mean ETP and the standard deviation. If only long term mean maximum and minimum temperatures and the mean temperature of a series of years are available, the standard deviation of the mean temperature provides a means for making an estimate of the standard deviation in ETP.

The size of the sample plays a part in the proximity of the standard deviation of a sample to the standard deviation of the population.

Altman and Bland (date unknown), in their response to Nagele [3], wrote that the standard error (SE) of the sample mean depends on both the standard deviation (SD) and the sample size, by the simple relation

$$\mathrm{SE} = \frac{\mathrm{SD}}{\sqrt{sample\ size}} \qquad (1)$$

They further stated that the standard error fell as the sample size increased, as the extent of chance variation was reduced. This idea underlined the sample size calculation for a controlled trial, for example. By contrast the standard deviation would not tend to change as they increased the size of their sample.

Also on the question of sample size, Ziliaka and McCloskeyb [4] wrote the following: We find here that in the next decade, the 1990s, of the 137 papers using a test of statistical significance in the AER fully 82% mistook a merely statistically significant finding for an economically significant finding. A super majority (81%) believed that looking at the sign of a coefficient sufficed for science, ignoring size.

In a response to a question on the relationship between standard deviation and sample size, Professor Mean had this to say: The estimate of the

standard deviation becomes more stable as the sample size increases. But after about 30 – 50 observations, the instability of the standard deviation becomes negligible.[5]

According to Cochran [6], there are four ways of estimating variances for sample size determinations:

(1) by taking the sample in two steps
(2) by the results of a pilot survey
(3) by previous sampling of the same or a similar population, and
(4) by guesswork about the structure of the population, assisted by some mathematical results.

Here again, there is that link between standard deviation and the sample size. However, (4) indicates that the structure of the population plays a role if guesswork is applied.

Methodology and analysis

We generated 5040 random numbers using Excel and calculated the standard deviations s_n, where n is the cluster size and n_i denotes the cluster number for $i = 1,2,\ldots,\frac{5040}{n}$ $i = 1,2,\ldots,\frac{5040}{n}$. The mean of the standard deviations of the clusters was calculated using

$$\overline{s_n} = \frac{n}{5040}\sum_{i=1}^{\frac{5040}{n}} s_{n_i} \qquad (2)$$

For instance, when the cluster size is 504, then

$$\overline{s}_{504} = \frac{1}{10}(s_{n_1} + s_{n_2} + \ldots + s_{10}^{n}) \qquad (3)$$

This procedure was repeated for different sets of 5040 randomly generated numbers. Table 1 shows the non-monotonic convergence of the means of standard deviations of the clusters to the standard deviation of the population for one of the sets. For this set, the standard deviation was 28.6164:

Table 1: Non-Monotonic Convergence Simulation

Serial number	Cluster size n	Mean SD of Clusters $\overline{s}_n$	Difference $\left\lvert s_N - \overline{s}_n \right\rvert$
1	10	26.73438	1.88202
2	15	27.29214	1.32426
3	20	27.64601	0.97039
4	30	27.95824	0.65816
5	35	28.05827	0.55813
6	40	28.11891	0.48749
7	45	28.15516	0.46124
8	60	28.28728	0.32912
9	70	28.31353	0.30287
10	80	28.3693	0.2471
11	90	28.3641	0.2523
12	105	28.45474	0.16166
13	120	28.45288	0.16352
14	140	28.44798	0.16842
15	180	28.51759	0.09881
16	210	28.51448	0.10192
17	240	28.54756	0.06884

18	280	28.54062	0.07578
19	315	28.54933	0.06707
20	360	28.55474	0.06166
21	420	28.56815	0.04825
22	560	28.58982	0.02658
23	630	28.57869	0.03771
24	720	28.59851	0.01789

For each set of 5040 randomly generated numbers, the sequence $\{\overline{s}_n\}$ converged to s_N. However, the convergence was non-monotonic for each set.

Results and discussion

We found that as the sample size of each cluster increased, the mean of the standard deviations of the clusters tended to the standard deviation of the population. However, for each n, the difference between the means of the standard deviations and the standard deviation of the population does not necessarily decrease as n increased.

The following charts illustrate the relationship between the means of the standard deviations of the clusters and the standard deviation of the population:

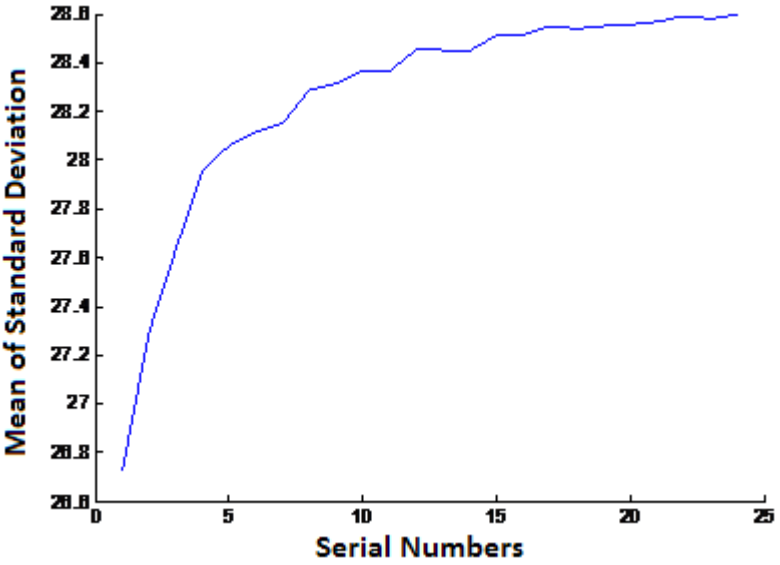

Figure 1: Means of standard deviations of clusters

In Figure 1, as the serial numbers (and consequently the cluster sizes) increased, the mean of the standard deviations of the clusters also increased generally. However, when serial number is 14, $\overline{S}_{140}$ is 28.44798, When serial number increased to 15, $\overline{S}_{180}$ increased to 28.51759 whereas when serial number further increased to 16, $\overline{S}_{210}$ decreased to 28.51448. The same trend could be noticed for the serial numbers 21, 22 and 23. These indicated that the convergence was not monotonic.

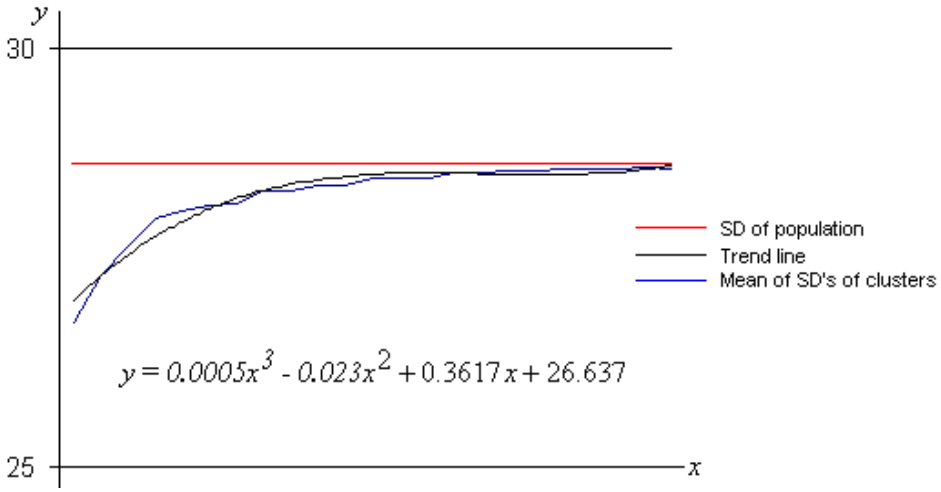

Figure 2: Means of standard deviations of clusters and the trend line

Figure 2 shows the relationship between the mean of standard deviations of clusters and a trend line of degree 3. It also shows how these two approached the standard deviation of the population as the cluster sizes increased.

Conclusion

The results indicate that the mean of the standard deviations of clusters of a population may be used to estimate the standard deviation of the population by making the size of the clusters large enough.

References

[1] Sabels, D.L and Klausmeier, R.D: Accuracy of short-cut estimates for standard deviation. Journal of Educational Measurement, Vol. 8 No. 4, 1971

[2] Hargreaves, G.H and Samani, Z.A: Estimation of Standard Deviation of Potential Evapotranspiration. Journal of Irrigation and Drainage Engineering, Vol. 144, Issue 1

[3] Nagele, P: http://www.bmj.com/content/331/7521/903

[4] Ziliaka, S.T and McCloskeyb, D.N: Size matters: the standard error of regression in the American Economic Review. The journal of Social-Economics. Vol. 33 pages 527 - 546

[5] http://www.children-mercy.org/stats/weblog2006/StandardDeviation.asp

[6] Cochran, W.G: Sampling Techniques (3rd edition). John Wiley & Sons, 1999. Page 78