OMITTING MEANINGLESS DIGITS:
ANALYZING LDR(1), THE STANDARD LEADING-DIGIT RULE

Wheyming T. Song

Department of Industrial Engineering
National Tsing Hua University
Hsinchu, Taiwan, R.O.C.

Bruce W. Schmeiser

School of Industrial Engineering
Purdue University
West Lafayette, IN 47907, U.S.A.

ABSTRACT

The standard leading-digit rule, LDR(1) is to omit point-estimator digits to the right of the leading digit of the point-estimator's standard error. Assuming that the original point estimator is normally distributed, the authors previously showed that LDR(1) guarantees—for all means and for all standard errors—that the truncated estimator's first omitted digit is correct with probability no greater than 0.117, not much greater than the 1-in-10 chance for a random digit. We consider two variations of the previously studied LDR(1) truncated point estimator, which in the worst case has non-negligible bias. The first is the truncated estimator with an implied appended digit "5". The second is the rounded estimator, which truncates after appending the "5". Both point estimators have nearly identical statistical properties, including negligible bias. Because of the omitted digits, however, the statistical quality of the two LDR(1) point estimators cannot be better than that of the original point estimator. In terms of root mean squared error and in terms of correlation with the original estimator, we establish here that the worst-case LDR(1) degradation is about four percent.

1 INTRODUCTION

The classic issue in statistical inference is how to estimate and report the precision or confidence of a conclusion based on a sample from a population. In its simplest form, a point estimator $\hat{\theta}$ computed from the sample is used to estimate the value of an unknown performance measure θ of the population.

We say that the jth digit of $\hat{\theta}$ is correct if it matches the jth digit of θ, where the jth digit is in the 10^j position. Digits to the left of $\hat{\theta}$ have a high probability of being correct; digits to the right of $\hat{\theta}$ are randomly distributed, with each value zero through nine having probability 0.1 of being correct.

Our research issue is how to report the value of the point estimator $\hat{\theta}$. How many of the left-most non-zero digits of $\hat{\theta}$ should be reported? And given the number of reported non-zero digits, what point-estimator value should be reported? Song and Schmeiser (2009) discuss the former question. In this paper we discuss the latter question.

2 PROBLEM DEFINITION

Based on simple classical assumptions, we discuss the two questions. The discussion of Question 1 is a summary of the results in Song and Schmeiser (2009). The discussion of Question 2 defines notation for the point estimators that are studied in Sections 3 and 4.

2.1 Assumptions

Throughout, we assume that the point estimator $\hat{\theta}$ is normally distributed, is unbiased, and has a standard error that decreases with the sample size, say n, which can be either discrete or continuous. These assumptions are true in much of statistical inference, and especially in our motivating application of Monte Carlo simulation, where typically the sample size n is large. The results are general in that the performance measure θ does not need to be a mean; the assumptions are asymptotically true for essentially every reasonable point estimator of a finite θ. Included are sample variances, sample standard deviations,

sample coefficients of variation, sample covariances, sample correlations, and sample quantiles. Also included are both independent and dependent sampling. (See Lewis and Orav 1989.)

We do not consider how the point estimator $\hat{\theta}$ or its standard error σ are obtained. We assume that the value of σ is known. (The effect of using an estimated standard error is a future topic.)

2.2 Question 1: The Number of Reported Digits

Let $\hat{\theta}_l$ denote the truncated point estimator, where l is the position of the left-most digit of the standard error σ. The digits to the right of position l are omitted; the reported digits are those of $\hat{\theta}$. Choosing l too small results in reporting many digits that clutter the report and are misleading when the reader thinks that all reported digits are meaningful. Choosing l too large reports too few digits, degrading the value of the point estimator.

Song and Schmeiser (1994) and Song and Schmeiser (2009) discuss the problem of deciding which of the right-most digits of a point estimate $\hat{\theta}$ to omit. LDR, the family of leading-digit rules, omits all digits of $\hat{\theta}$ to the right of the leading digit of σ, the standard error of the point estimator. In particular, LDR(a) omits all digits to the right of the position of the leading digit of $a\sigma$. Because of its reasonable statistical properties and ease of implementation, we focus on LDR(1), the standard LDR.

Song and Schmeiser (2009) show that LDR(1) guarantees that the left-most omitted digit has correctness probability of less than 0.117, conditional that the reported digits are correct. This worst-case bound occurs when the sample size n yields a standard error σ that is a power of ten. At that point, a bit more sampling would increase the value of n, reduce the value of σ, and the current left-most omitted digit of $\hat{\theta}$ would be reported. At this instant, the probability of the left-most omitted digit being correct is maximized. At other times, the probability of the left-most omitted digit being correct is closer to one in ten, the probability that a random digit is correct.

2.3 Question 2: The Reported Point Estimator

LDR determines only which right-most digits to omit. That leaves the question of what to do with the left-most digits that are reported.

Song and Schmeiser's (2009) correct-probability bound of 0.117 is for the truncated point estimator $\hat{\theta}_l$, where l is the position of the left-most non-zero digit of σ. The digits to the right of position l are omitted; the reported digits are those of $\hat{\theta}$.

We consider here two variations of the truncated point estimator $\hat{\theta}_l$. Both point estimators require the same space in an output report as $\hat{\theta}_l$. Both eliminate the first-order bias caused by the truncation in $\hat{\theta}_l$. The first variation, $\hat{\theta}_l^{(5)}$, is to append a "5" in position $l-1$ of $\hat{\theta}_l$; this appended digit is not reported, but is part of the interpretation of the reported $\hat{\theta}$. The second variation, $\hat{\theta}_l^{(r)}$, is to round $\hat{\theta}$ to position l; that is, add $5 \times 10^{l-1}$ to $\hat{\theta}$ and truncate to the lth position.

We compare all four point estimators: the Original point estimator $\hat{\theta}$, the Truncated point estimator $\hat{\theta}_l$, the Append-5 point estimator $\hat{\theta}_l^{(5)}$, and the Rounded point estimator $\hat{\theta}_l^{(r)}$.

The statistical properties of the three LDR(1) point estimators cannot be as good as those of the original point estimator $\hat{\theta}$, because of the digits that are omitted to the right of the LDR position l. The results in Sections 3 and 4 show the amount degradation in terms of mean squared error (mse) and in terms of correlation with $\hat{\theta}$.

3   MEAN SQUARED ERROR

We consider here the mse of three LDR(1) point estimators—the Truncated point estimator $\hat{\Theta}_l$, the Append-5 point estimator $\hat{\Theta}_l^{(5)}$ and the Rounded point estimator $\hat{\Theta}_l^{(R)}$. For most values of θ and σ, the mse of all three LDR(1) point estimators are essentially the same as that of the Original point estimator. Therefore, the mse analyses of this section focus on the worst-case for the three LDR(1) point estimators, when the sample size n yields a value of σ that is a power of ten; that is, just before the left-most omitted digit would be reported.

LDR truncation never increases the absolute value; that is, $|\hat{\Theta}_l| \leq |\hat{\Theta}|$. Therefore, two cases need to be considered. The two cases represent the extremes of the effects of the truncation. In the first case, represented here by $\Theta = 10$ and $\sigma = 1$, $\hat{\Theta}$ always has the same sign, so truncation always is in the same direction, which can cause non-negligible bias. In the second case, represented here by $\Theta = 0$ and $\sigma = 1$, the Truncated point estimator $\hat{\Theta}_l$ is always closer to Θ than the Original point estimator $\hat{\Theta}$. In this second case, the mse of the LDR point estimators can be better than that of the Original point estimator.

So unlike most comparisons of point-estimator behavior, the smallest mse is not necessarily the best. The ideal mse is that of the Original point estimator, with no bias for any value of Θ and standard error σ.

For these two extreme cases, Table 1 shows the means and standard errors for the three LDR(1) point estimators. The results depend upon our standing assumptions that the Original point estimator is normally distributed with mean Θa and standard deviation σ. All table entries have simple explanations or approximations, as discussed in the Appendix. The table entries were also obtained by Monte Carlo experimentation; the standard errors of the table entries are so small that the digits shown are correct.

As an aside, notice that this table is an example of our topic, the need to decide the number of digits to report. Here the truncation is chosen for practical significance; additional digits are not useful to support the message. In fact, the standard errors are so small that $\hat{\Theta}_l$, $\hat{\Theta}_l^{(5)}$, and $\hat{\Theta}_l^{(r)}$ are equal.

In the worst case, Truncated has non-negligible bias, whereas both Append-5 and Rounded have negligible bias always.

Among the three estimators shown in Table 1, we recommend to use the Rounded estimator $\hat{\Theta}_l^{(r)}$ because it posses the best statistical property and is more natural than the Append-5.

Table 1: Worst-Case Moments of the three LDR(1) variations, where $\hat{\Theta}$ has mean Θ and standard error $\sigma = 10^l$.

cases	$\hat{\Theta}_l$ (Truncated)	$\hat{\Theta}_l^{(5)}$ (Append-5)	$\hat{\Theta}_l^{(r)}$ (Rounded)
$\Theta = 10$	9.50	10.00	10.00
$\sigma = 1$	1.04	1.04	1.04
$\sqrt{\text{mse}} = 1$	1.15	1.04	1.04
$\Theta = 0$	0.00	0.00	0.00
$\sigma = 1$	0.68	1.04	1.04
$\sqrt{\text{mse}} = 1$	0.68	1.04	1.04

4 CORRELATION

Another way to measure the performance of LDR(1) point estimators is the extent to which they predict that value of the Original point estimator $\hat{\Theta}$. Although the LDR(1) point estimators are not normally distributed, assume for the moment that the pairs $(\hat{\Theta}, \hat{\Theta}_l)$, $(\hat{\Theta}, \hat{\Theta}_l^{(5)})$, and $(\hat{\Theta}, \hat{\Theta}_l^{(r)})$ are bivariate normal. Consider the Truncated point estimator. If $\hat{\Theta}_l = c$, the conditional distribution of the Original point estimator is normal with conditional expectation

For most values of Θ and σ, the correlations are close to one. Therefore, we again focus on the two extreme cases of Section 3. Table 2 shows the extreme-case performance of the three LDR(1) point estimators.

As usual, the Truncated point estimator degrades more than the Append-5 and Rounded point estimators. The correlation of 0.88 is substantially worse than the 0.96 correlation obtained by Append-5 and Rounded. Since correlation squared is the fraction of variance explained, both Append-5 and Rounded explain about 92% of the variance of the Original point estimator $\hat{\Theta}$.

5 THE ROUNDED LDR(1) ESTIMATORS $\hat{\Theta}_l^{(r)}$

Among the three estimators shown in Table 1, we prefer the Rounded estimator $\hat{\Theta}_l^{(r)}$ because it possesses good statistical properties in terms of $\sqrt{\text{mse}}/\sigma$ and $\text{corr}(\hat{\Theta}, \hat{\Theta}^{(r)})$, and it is more natural than the Append-5, which requires users sophistication (i.e., knowing to append the digit "5.") In this section, we further investigate the statistical properties of the Rounded LDR(1) estimator $\hat{\Theta}_l^{(r)}$ as functions of σ.

Table 2: Worst-Case correlations of the three LDR(1) variations with the Original point estimator, where $\hat{\Theta}$ has mean Θ and standard error $\sigma = 10^l$.

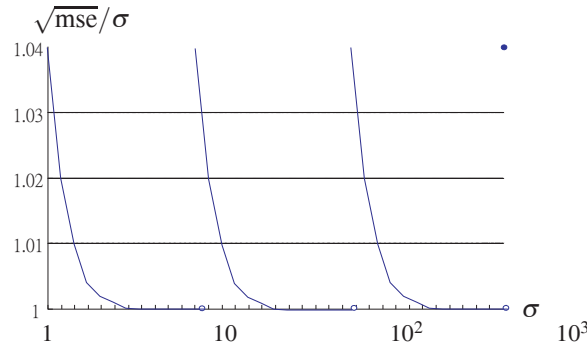| cases	$\hat{\Theta}_l$ (Truncated)	$\hat{\Theta}_l^{(5)}$ (Append-5)	$\hat{\Theta}_l^{(r)}$ (Rounded)
$\theta = 10$	9.50	10.00	10.00
$\sigma = 1$	1.04	1.04	1.04
corr	0.96	0.96	0.96
$\theta = 0$	0.00	0.00	0.00
$\sigma = 1$	0.68	1.04	1.04
corr	0.88	0.96	0.96

Figure 1: Relative root mse as a function of σ for the rounded estimator $\hat{\Theta}^{(r)}$

Figure 1 shows that the root relative mse ($\sqrt{\text{mse}}/\sigma$) of $\hat{\Theta}^{(r)}$ is maximized with a value of 1.04 when the standard error σ is a power of 10. Moreover, we observe that the $\sqrt{\text{mse}}/\sigma$ dramatically decreases when $\log\sigma$ is in the range $(i, i+0.5]$, and decreases slowly when $\log\sigma$ is in the range $[i+0.5, i+1)$, where $i = 0, 1, 2, \ldots$.

Figure 2 shows that the correlation between $\hat{\Theta}_l$ and $\hat{\Theta}^{(r)}$ is minimized with a value of 0.96 when the standard error σ is a power of 10. Moreover, we observe that the correlation dramatically increases when $\log\sigma$ is in the range $(i, i+0.5]$, and increases slowly when $\log\sigma$ is in the range $[i+0.5, i+1)$, where $i = 0, 1, 2, \ldots$.

According to Figures 1 and 2, we conclude that Rounded LDR(1) is a good rule for omitting meaningless digits for statistical estimates because it is easy to implement and has a good statistical property in terms of $\sqrt{\text{mse}}/\sigma$ and $\text{corr}(\hat{\Theta}, \hat{\Theta}^{(r)})$.

6   SUMMARY AND FUTURE RESEARCH

The Truncated, Append-5, and Rounded LDR(1) estimators require the same reporting space, but the Append-5 and Rounded estimators have better statistical properties. Always, the worst-case comparison to the original point estimator $\hat{\Theta}$ is when the standard error σ is a power of ten, just before the next digit is to be reported.

Because both the Append-5 and Rounded estimators have negligible bias their worst-case root mse performance is substantially better than that of the Truncated estimator. They both have about a four-percent worst-case degradation compared to the original estimator$\hat{\Theta}$. They both have worst-case correlation of about 0.96 with the original point estimator. Other than when σ is only a bit larger than a power of ten, the root mse is essentially the same as that of the original point estimator and the correlation with the original point estimator is essentially one.
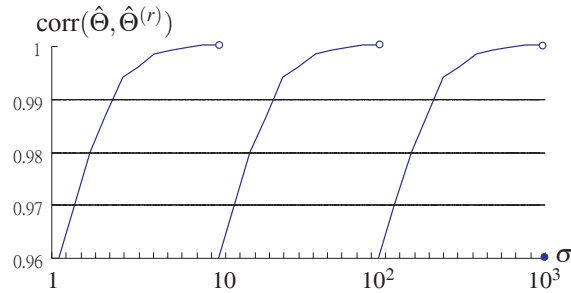
Figure 2: Correlation of the estimators $\hat{\Theta}$ and $\hat{\Theta}^{(r)}$ as a function of σ

Because the Append-5 point estimator requires reader sophistication—knowing to append the digit "5"—the Rounded point estimator is preferable. We advocate the use of LDR(1) with the point estimator $\hat{\Theta}_l^{(r)}$.

Other point estimators based on $\hat{\Theta}_l$ are useful. In a future paper we consider appending a random digit or a random fraction to $\hat{\Theta}$ to allow confidence-interval procedures to have reasonable coverage probabilities.

APPENDIX

The results in Table 1 are based on Monte Carlo experimentation. Some of the results also are obtained by approximation, as discussed in the Appendix.

Case 1: $\Theta = 10, \sigma = 1$

- $\mathrm{se}(\hat{\Theta}_l) \approx \sqrt{1 + \frac{1}{12}} \approx 1.04$
- $\mathrm{se}(\hat{\Theta}_l^{(5)}) \approx \sqrt{1 + \frac{1}{12}} \approx 1.04$
- $\mathrm{se}(\hat{\Theta}_l^{(r)}) \approx \sqrt{1 + \frac{1}{12}} \approx 1.04$

Case 2: $\Theta = 0, \sigma = 1$

- $\mathrm{se}(\hat{\Theta}_l) = \sqrt{2 \sum_{i=0}^{\infty} i^2 [\Phi(i+1) - \Phi(i)]} \approx 0.684$
- $\mathrm{se}(\hat{\Theta}_l^{(5)}) = \sqrt{2 \sum_{i=0}^{\infty} (i+0.5)^2 [\Phi(i+1) - \Phi(i)]} \approx 1.04$
- $\mathrm{se}(\hat{\Theta}_l^{(r)}) = \sqrt{2 \sum_{i=0}^{\infty} (i+0.5)^2 [\Phi(i+1) - \Phi(i)]} \approx 1.04$

REFERENCES

Lewis, P.A.W. and E.J. Orav. 1989. *Simulation Methodology for Statisticians, Operations Analysts, and Engineers*. Volume 1. Wadsworth & Brooks/Cole, California: Pacific Grove.

Song, W.T. and B.W. Schmeiser. 1994. Reporting the precision of simulation experiments. In New Directions in Simulation for Manufacturing and Communications,ed. S. Morito, H. Sakasegawa, K. Yoneda, M. Fushimi, and K. Nakano, 402V407. Tokyo: Operations Research Society of Japan.

Song, W.-M. T and Bruce Schmeiser 2009. Omitting Meaningless Digits in Point Estimates: the Probability Guarantee of Leading-Digit Rules. *Operations Research* 57: 109 - 117.

ACKNOWLEDGMENTS

AUTHORS BIOGRAPHIES

WHEYMING TINA SONG is a professor in the Department of Industrial Engineering at the National Tsing Hua University in Taiwan. She received her undergraduate degree in statistics and masters degree in industrial engineering at Cheng-Kung University in Taiwan in 1979. She then received masters degrees in applied mathematics in 1983 and industrial engineering in 1984, both from the University of Pittsburgh. Dr. Song received her Ph.D. from the School of Industrial Engineering at Purdue University in 1989. She joined Tsing Hua in 1990 after spending one year as a visiting assistant professor at Purdue IE. Her email address is <wheyming@ie.nthu.edu.tw>.

BRUCE W. SCHMEISER is a professor at Purdue University. He received his Ph.D. from the School of Industrial and Systems Engineering at Georgia Tech in 1975; his undergraduate degree in the Mathematical Sciences and Masters degree in Industrial Engineering are from The University of Iowa. His research interests include stochastic root finding, simulation output analysis, input modeling, random-variate generation, variance-reduction techniques, Markov chain Monte Carlo methods, and applied operations research. He has served in a variety of roles for the IIE, INFORMS and the Winter Simulation Conference. His email address is <bruce@ecn.purdue.edu>.