LA--12114-MS

DE91 018513

Mean Estimation in Highly Skewed Samples

Shane P. Pederson

MASTER

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

MEAN ESTIMATION IN HIGHLY SKEWED SAMPLES

By
Shane P. Pederson

ABSTRACT

The problem of inference for the mean of a highly asymmetric distribution is considered. Even with large sample sizes, usual asymptotics based on normal theory give poor answers, as the right-hand tail of the distribution is often undersampled. This paper attempts to improve performance in two ways. First, modifications of the standard confidence interval procedure are examined. Second, diagnostics are proposed to indicate whether or not inferential procedures are likely to be valid. The problems are illustrated with data simulated from an absolute value Cauchy distribution.

1. Introduction

When data arise from an asymmetric population, and the center of such a population is of interest, many techniques are available for estimation. Most involve either developing an estimator robust to the asymmetry, or transforming the data to create a more symmetric distribution. In these cases, the true mean of the underlying population is not preserved; rather, quantities such as the median are estimated. There are situations, however, in which the mean really is the quantity of interest. For example, in neutron transport the energies of individual particles are measured. Primary interest is often in total energy of the collection of particles: this is merely a rescaling of mean energy per particle. This report details preliminary efforts to develop mean estimation techniques in large samples of highly skewed data.

2. Background

Standard theory states that with a sample of size n from a normal distribution with unknown mean and variance μ and σ^2, the statistic t, where

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{s}, \tag{1}$$

and $\bar{x}$ and s^2 are the sample mean and sample variance, will have a Student's t-distribution with n-1 degrees of freedom. With n large, this distribution is very close to that of a standard normal (Gaussian) random variable. In either case, these distributions can be used to obtain confidence intervals for μ, via appeal to the Central Limit Theorem. It states that the sample mean from any distribution (with at least the first two moments existing and sufficiently regular), will converge to a standard normal random variable.

This result are frequently misapplied to cases in which data are not normal, σ is not known, or both. If moments exist the distribution of t eventually becomes normal, but this may require an enormously large sample. In neutron transport problems, the data are commonly very skewed to the right. This has two effects on mean estimation. First, $\bar{x}$ and s^2 have a high positive correlation; both tend to be large or small together. Second, these estimates are biased in finite samples and highly variable, introducing additional noise in the estimation process. Both adversely affect the standard normal approximation of t.

We use the absolute value of a Cauchy random variable to mimic actual distributions that arise in neutron transport problems. The associated probability density function is

$$f(x) = \frac{2}{\pi}\frac{1}{1+x^2}, \quad 0 \le x < \infty. \tag{2}$$

No moments exists for this distribution. However, if we truncate it at some point $\beta > 0$, all moments exist but can be made arbitrarily large. The mean of a truncated absolute value Cauchy random variable is

$$\mu = \frac{1}{\pi}\left[\log(1+\beta^2) + \beta(\pi - 2\tan^{-1}(\beta))\right]. \tag{3}$$

$(1-\alpha)\times 100\%$ confidence intervals for μ are of the form $\bar{x} \pm z_{(1-\alpha/2)} s/\sqrt{n}$, where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)\times 100$th percentile of the standard normal.

At this point, a pair of illustrations may be helpful. Figure 1 indicates the joint distribution of $\bar{x}$ and s for 1000 Gaussian samples of size 1000. The diagonal lines are the boundaries of the 95% confidence interval for μ. Note that the principal axis of the ellipse formed by the data is horizontal, indicating zero correlation between $\bar{x}$ and s. Confidence intervals that cover μ are represented by points above the "V".

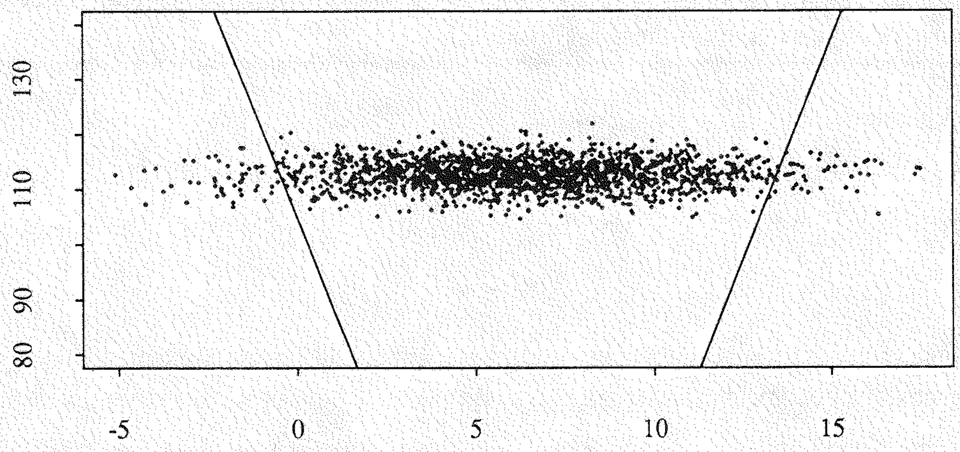

Figure 1. Sample mean vs. sample standard deviation, Gaussian.

Meanwhile, Figure 2 indicates the same plot for data arising from the censored absolute value Cauchy distribution. The differences are obvious. Because of extreme values, both axes are in logarithmic scale, resulting in curved envelope lines. The joint distribution is no longer elliptical (indicating non-normality), and the estimators are biased, highly correlated, and highly variable. Hence, the coverage percentage is no longer the nominal 95% but a value far less. In particular, most confidence intervals that miss μ are too low. This indicates that not enough rare events are being sampled to provide consistent estimates of μ and σ, and sample sizes must be increased. Monitoring the stability of $\bar{x}$ alone is not enough to indicate when inferences are valid, because it is likely that $\bar{x}$ will be fairly stable long before s^2 is, or at least before s^2 is close to σ^2. Both estimates must be stable *and* close to the true population values for nominal coverage rates to be attained.

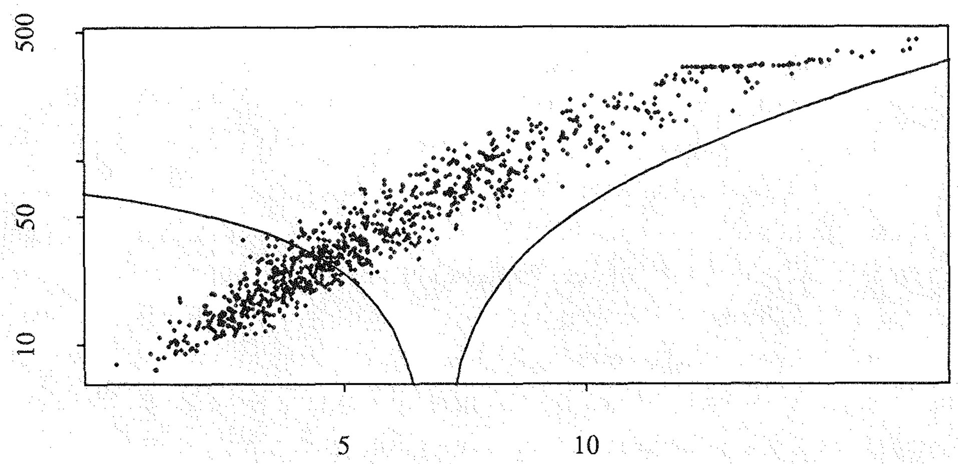

Figure 2. Sample mean vs. sample std. dev., absolute value Cauchy, truncated at 10000.

To improve the coverage rates of mean estimators in these cases, two remedies were considered. The first consists of modifications of the estimation or inference procedure to correct for invalid assumptions. The second is the development of diagnostic procedures to indicate when inferences are or are not likely to be valid. Several methods were tried, with varying degrees of success, and they are recounted here. One important thing to remember about corrections like these is that they are by necessity data dependent, and hence often exhibit the same behavior as the original data.

3. Modifications

The first modification attempt deals with inflation of the variance estimate. (Recall s^2 is more variable than $\bar{x}$.) By adding an arbitrary constant to all nonzero data, the resulting augmented distribution (and hence the augmented distribution of $\bar{x}$) becomes more symmetric, while still enabling recovery of the original mean. Unfortunately, the amount of shift needed to reduce the skewness to approximately zero depends on the higher, unknown moments of the distribution.

Too much shift results in overly conservative error bounds. Because the higher moments are poorly estimated, no reliable estimator of the amount of shift necessary to achieve nominal coverage was found.

A second method considered batching means and developing variance estimators from these quantities. In highly skewed samples, the variance estimator from a collection of batched means may be less correlated with the sample mean than is the original sample variance; the tradeoff is that this variance estimate is more variable. Simulations have indicated that for sufficiently large batch sizes, coverage rates are improved. Batch sizes of at least m = 20,000-50,000 appear to work, but the improvement is most marked in cases where the initial coverage is very low (e.g., 50% in a nominal 95% situation). A heuristic rule may be to use the maximum of the two variance estimators, if the validity of the procedure is in doubt, although asymptotically this will bias the estimate.

The third modification examined approximations to the joint distribution of $\bar{x}$ and s^2, or equivalently to the distribution of t. Direct density estimation does not appear to be feasible, again because of the poor representation of the right-hand tail of the distribution by a finite sample. Alternatively, the distribution of t can be represented by the inversion of an Edgeworth expansion (Hall [3]). In this method, cumulative probabilities (of the t-statistic, for example) can be represented as a weighted sum of standard normal distributions and density functions; the weights depend on moments of the true underlying distribution. Problems arise in practice, because the moments again must be estimated from the available data. Also, it is not clear how many terms of expansion need to be utilized for the representation to be valid. Simulations indicated that adding the third moment correction term improved coverage rates up to 20%; again, the biggest improvement occurred in cases of poor initial coverage. Higher order correction factors did not appreciably improve this performance. These factors will not improve 50% coverage to a nominal 95%, and their biggest value may be in fine-tuning an estimator with nearly nominal coverage rates. The refined confidence interval, using the first correction factor, is

$$\bar{x} \pm \frac{s}{\sqrt{n}} z_{(1-\alpha/2)} + \frac{1}{6n}\frac{\hat{\mu}_3}{s^2}\left(1 + 2z^2_{(1-\alpha/2)}\right), \tag{4}$$

where μ_3 is the third moment of the parent distribution and is estimated by the third sample moment $\hat{\mu}_3$.

In summary, modification procedures appear to be of limited usefulness, for the same reason the usual inference procedures are invalid; they use estimators based on assumptions that do not hold for the sample sizes considered. In the next section, diagnostics for determining whether these assumptions are violated will be examined.

4. Diagnostics

We now turn to diagnostics for detecting when inferences are valid. Initially, tests of symmetry or normality were considered. Some of these are computationally intensive, and none were found to have strong power in predicting when inferences are valid. Cases in which normality tests (such as the Wilk-Shapiro test) indicate normality do correspond to approximately valid inferences; however, in simulation, tests such as these rarely indicated normality (Beckman [1]). Correspondingly, visual examination of histograms of batched means may be a useful graphical technique but is difficult to quantify.

A more fruitful result was found by considering an expansion of the distribution of t. Rather than an Edgeworth expansion, as in Section 3, a Taylor series expansion was computed to determine approximations to the first few moments of t. From Geary's [2] expressions for the semi-invariants of t, the first four (central after the first) moments of t are, to $o(n^{-1})$,

$$E(t) = \frac{-\mu_3}{2\sqrt{n}\sigma^3} \tag{5}$$

$$\mathrm{var}(t) = 1 + \frac{2}{n} + \frac{7}{4n}\frac{\mu_3^2}{\sigma^6} \tag{6}$$

$$E(t-E(t))^3 = \frac{-2\mu_3}{\sqrt{n}\sigma^3} \tag{7}$$

$$E(t-E(t))^4 = 3 + \frac{18}{n} - \frac{2(\mu_4-3\mu_3^2)}{n\sigma^4} + \frac{45}{2n}\frac{\mu_3^2}{\sigma^6}. \tag{8}$$

Note that in the case of standard normal data, the first four moments of t become, to $o(n^{-1})$, 0, $1 + 2/n$, 0, and $3 + 18/n$, respectively.

As noted before, sample moment estimators of these quantities are not reliable - they tend to underestimate the true parameter values and are highly variable. In simulations, however, a strong correlation was observed between the coefficient of variation of s^2 (i.e., $\mathrm{cv}(s^2) = \sqrt{\mathrm{var}(s^2)}/E(s^2)$ and the observed coverage rate. To utilize this, it will be useful to rewrite the expressions for the moments of t.

The first two moments of the joint distribution of $\bar{x}$ and s^2 are $E(\bar{x}) = \mu$, $E(s^2) = \sigma^2$, $\mathrm{var}(\bar{x}) = \sigma^2/n$, $\mathrm{var}(s^2) = (\mu_4-\sigma^4)/n$, and $\mathrm{cov}(\bar{x},s^2) = \mu_3/n$. From these we can obtain the the squared correlation between $\bar{x}$ and s^2 as

$$\rho^2 = \mathrm{corr}^2(\bar{x},s^2) = \frac{\mu_3^2}{\sigma^2(\mu_4-\sigma^4)}. \tag{9}$$

The squared coefficient of variation of s^2 is

$$\gamma = \mathrm{cv}^2(s^2) = \frac{(\mu_4-\sigma^4)}{n\sigma^4} \tag{10}$$

The correlation can hence be written as

$$\rho = \frac{\mu_3}{\sigma^3} \frac{1}{\sqrt{n}\ \mathrm{cv}(\bar{x}, s^2)}. \tag{11}$$

This allows us to write the moments of t, to $o(n^{-1})$, as

$$E(t) = \frac{-\rho\gamma^{1/2}}{2} \tag{12}$$

$$\mathrm{var}(t) = 1 + \frac{2}{n} + \frac{7}{4}\rho^2\gamma \tag{13}$$

$$E(t-E(t))^3 = -2\rho\gamma^{1/2} \tag{14}$$

$$E(t-E(t))^4 = 3 + \frac{22}{n} + \frac{\gamma}{2}(45\rho^2-4). \tag{15}$$

The dependence of these terms on the coefficient of variation of s^2 is now obvious. In highly skewed data, ρ is near 1. When γ is sufficiently small, the variance of t will be near 1, the skewness near zero, the kurtosis near three, and inferences can proceed with the standard normal as a reference distribution. In practice, it was found that $\hat{\mu}_4/s^4$ is a poor estimator of μ_4/σ^4; as in other moment estimators, it is biased low in finite samples. Hence, these moment expansions are not useful, at least in the censored absolute value Cauchy problem, in developing correction factors to be used in confidence intervals; however, they may be of use as diagnostics.

From given values of ρ and γ, approximate moments of t can be computed. These can be compared to moments of a standard normal to determine the proximity of the two distributions. Calculations based on Pearson curves (Johnson *et al.* [4]) indicate that γ values of less than about 0.4 correspond to reasonable matching of t percentage points with standard normal ones up to the 1% point. Simulations conducted on the censored absolute value Cauchy distribution indicate that for γ values of less than 0.2, matching of the percentage is good up to the 5% point. The discrepancy in the two results is due to the approximate nature of the Pearson curve calculation. A conservative recommendation is that once γ is in the 0.1 - 0.2 range, inferences on the mean based on t-statistic confidence intervals are approximately valid, if the third-moment correction term from Section 3 is used.

As mentioned previously, knowing the value of γ for which valid inferences occur may not be useful in practice, unless γ is well estimated. In simulations on the Cauchy problem, these estimates were often biased for γ until roughly the point at which $\hat{\gamma}$ is less than 0.25; γ is underestimated, sometimes severely, until then. Table 1 contains results of a simulation for a specific censored Cauchy model (censoring point equal to 10,000). 800 replications at each sample size were simulated and tabled values are averages of those replications. Coverage rates for upper confidence intervals using the standard and 3rd-moment corrected intervals are given, as well as corresponding average values for γ and $\hat{\gamma}$. Similar results were observed for other values of the censoring point β.

Table 1

Observed coverage rates and average γ values for 800 replications
(upper confidence interval, nominal level = 0.95)

n	Standard	Corrected	γ	$\hat{\gamma}$
1000	0.59	0.66	5.23	0.49
5000	0.74	0.79	2.63	0.44
10000	0.82	0.84	0.52	0.36
20000	0.87	0.92	0.26	0.21
35000	0.90	0.94	0.15	0.15
50000	0.93	0.94	0.11	0.10
100000	0.93	0.94	0.05	0.05

For values of γ less than 0.2, confidence intervals based on t do an adequate job of covering the mean the proper fraction of the time; the 3rd-moment correction factor does improve performance. However, $\hat{\gamma}$ is not as valid an indicator as γ of performance. To use as a diagnostic, the following procedure is suggested. For $\hat{\gamma}$ less than 0.1, we will assume that inferences based on the standard normal distribution are valid (using the correction factor). If sample statistics are sequentially computed as sample size increases, then the rate of decrease of $\hat{\gamma}$ can be monitored. If the decrease in $\hat{\gamma}$ is primarily due to increase in sample size, this indicates that estimation of γ has stabilized and $\hat{\gamma}$ is approximately unbiased for γ. Thus, if $\hat{\gamma}$ is less than 0.2, and it appears $\hat{\gamma}$ is a good estimator of γ, inference for μ should be valid. As an example (using the parameters for Table 1), increasing n from 25000 to 50000 decreases γ by one-half. If the ratio $\hat{\gamma}_{25000}/\hat{\gamma}_{50000}$ is near 2, this indicates a decrease in $\hat{\gamma}$ mainly due to sample size. If this ratio is considerably less than 2, however, this is an indication that estimation of γ has not yet stabilized. In this instance, the validity of inferences on γ may be in doubt. This is an *ad hoc* procedure and more examination is needed to determine its efficiency, but is offered as a possible suggestion when $\hat{\gamma}$ lies between 0.1 and 0.2. Values greater than 0.2 may correspond to acceptable cases, but there is not currently a method to verify this.

5. Conclusions and Recommendations

It is clear that drawing inferences about the mean is difficult when using standard methods in situations in which usual assumptions do not hold, e.g., highly skewed data. Most modifications of these procedures are themselves of limited utility, as they are based on estimates which share the same properties as those being modified. A diagnostic was found that indicated when inferences could proceed by using normal probability points, but it is less useful in indicating when inferences are of questionable validity.

Recommendations to users faced with this sort of data are the following: If it is possible to focus on quantities other than the mean, for which more robust techniques are available, do so. If not, several things may be done. First, the third-moment correction factor should be utilized. Second, estimates of $\gamma = \mathrm{var}(s^2)/\mathrm{E}^2(s^2)$ should be monitored, with values of $\hat{\gamma}$ less than 0.1 indicating probable correctness of nominal confidence intervals. Values in the range (0.1-0.2) may be tracked to indicate the appropriateness of normal theory confidence statements. It should be noted, however, that these diagnostics are not foolproof, and are only a representation based on the data at hand. If the tail region has not been sampled sufficiently, inference procedures may not perform at nominal levels.

References

1. Beckman, R. J. (1991). Group A-1, Los Alamos National Laboratory. Personal communication.
2. Geary, R. C. (1947). Testing for Normality. *Biometrika* 34, 209-242.
3. Hall, P. (1985). Inverting an Edgeworth Expansion. *Ann. Statist.* 11, 569-576.
4. Johnson, N. L., Nixon, E., Amos, D. E., and Pearson, E. S. (1963). Table of Percentage Points of Pearson Curves. *Biometrika* 50, 459-498.