Statistica Sinica **6**(1996), 547-559

CONSERVATIVE, NONPARAMETRIC ESTIMATION OF MEAN CONCENTRATION OF CONTAMINANTS

Ling Chen and Robert W. Jernigan

Florida International University and The American University

Abstract: Statistical estimates and procedures that err on the side of caution are often desirable when dealing with public health issues. The U.S. Environmental Protection Agency, under the Superfund program, takes this approach in its determination of human exposure to soil contaminants at toxic waste sites. Due to uncertainties in estimating the true average contaminant concentration based on site sampling and the need to be conservative in assessing health risk, the EPA advocates the use of the 95% upper confidence limit (UCL) on the mean to provide "reasonable confidence that the true site average will not be underestimated". Skewness of the underlying distribution of the contaminant contributes to the persistent underestimation of the population mean when sample size is small and the need for a conservative procedure as attempted by the EPA. In this article, we propose an estimator for the mean of positively skewed distributions based on a penalized least squares criterion. When sample size is small or moderate, the new estimator has smaller mean square error and greater probability of falling within two standard deviations of the sample mean above the true mean than the UCL estimator currently being used by EPA.

Key words and phrases: Coverage, penalized least squares, penalty constant, superfund program.

1. Introduction

In an effort to protect human health and the environment from threats posed by hazardous substances the U.S. Environmental Protection Agency's Office of Emergency and Remedial Response has developed a process of human health risk assessment. This process is described in EPA (1989) *Risk Assessment Guidance for Superfund: Volume I - Human Health Evaluation Manual* (RAGS/HHEM). It details, in part, procedures for data collection and evaluation, exposure assessment, toxicity assessment, and risk characterization due to human exposure to hazardous substance releases at Superfund sites. A determination of contaminant concentration based on site sampling is required for the estimation of human exposure. Due to uncertainties in estimating the true average concentration and the need to be conservative in assessing health risk, EPA (1989) RAGS/HHEM and EPA (1992) *Supplemental Guidance to RAGS: Calculating the Concentration Term* both advocate the use of the upper 95 percent confidence limit (UCL) of the arithmetic mean to provide "reasonable confidence that the true site average will not be underestimated."

If the extent and level of this average contamination is underestimated then any resulting estimates of the risk of human health will also be underestimated. Such underestimation has been evident in random samples of concentrations at many sites listed in the U.S. Superfund program. The problem lies in the skewed distribution of the contaminant concentrations. Although the sample mean from a skewed distribution is unbiased for the population mean, the population skewness causes substantially more than half of the sample means to fall below the population mean for small to moderate sample sizes. This results in a sample mean with a large probability of falling below the population mean. Such a sample mean we say has low coverage, the probability of falling above the population mean.

Several estimators of the mean have been developed for asymmetric distributions, such as the smearing estimate (Duan (1983)), the retransformed mean obtained through Box-Cox transformations (Taylor (1985, 1986); Shumway, et al. (1989)) and the Once-Winsorized mean (Fuller (1970, 1991)). These procedures do not address the underestimation and coverage problems that concern us here.

The EPA (1992), through various pilot studies and workshops, considered many approaches to correct this underestimation problem and decided that to estimate the mean in a conservative manner, the upper point of a 95 percent normal theory confidence interval (UCL $= \bar{X} + 1.96\sigma_{\bar{X}}$) would be used, irrespective of the underlying distribution. Obviously the essence of the problem here is different from conventional methods that develop estimators through a trade-off between bias and variance. We will formulate an estimator through a trade-off of bias and coverage of the parameter. Thus, we should choose a biased estimator not only with a small mean square error but also with a large probability falling within an allowance range of the estimation. In this particular problem it will be reasonable to choose the allowance range as $[\theta, \theta + 2\sigma_{\bar{X}}]$, where θ is the true mean of the underlying distribution.

In this article, we propose a realistic estimate for the mean of such skewed distributions when high coverage is desired. The new estimate is based on a penalized sum of squares consisting of a squared error loss plus a penalty for each observation that falls above the estimate. The resulting penalized least squares estimator, called the penalized mean, is biased but it has a smaller variance than both $\bar{X}$ and UCL and a smaller mean square error than the UCL estimator while it has greater probability of falling within the allowance range of the estimation.

The new penalized least squares criterion, is proposed in Sections 2. The penalized mean is derived in Section 3 and its large sample properties are discussed.

A choice of the penalty constant is presented in Section 4. The simulation results in Section 5 are based on lognormal, exponential, chi-squared and Weibull distributions. Section 6 is conclusions.

2. A Penalized Least Squares Criterion

Our goal is to find a nonparametric estimator for the mean of a positively skewed distribution by tolerating some bias in the estimator in order to improve its coverage. To accomplish this goal we extend the usual least squares approach since a traditional decision theoretic approach would be extremely difficult in this nonparametric setting. We define a penalized squared error in order to balance bias, variance, and coverage as follows:

$$L^*(x_i, t) = (x_i - t)^2 + 2\lambda_n \mathrm{pr}(t < X < x_i), \tag{2.1}$$

where $\lambda_n > 0$, $\lambda_n = o(1)$. Here, the penalty we impose is proportional to the probability that X falls between the observation and the estimate t. Based on the form of the penalty term in (2.1), the more extreme an observation, the greater the penalty it imposes on our estimator. To minimize the penalty, the estimator will tend to be larger, and thus have a small probability of falling below the population mean, θ.

The penalized sum of squares is

$$R(t) = \sum L^*(x_i, t) = \sum (x_i - t)^2 + 2\lambda_n \sum \mathrm{pr}(t < X < x_i). \tag{2.2}$$

We shall find an estimator of θ by minimizing this penalized sum of squares. The estimator is called penalized least squares estimator, or simply called penalized mean.

3. One-step Iterate of the Penalized Mean

To find the penalized least squares estimator of the mean, let us examine the properties of the penalized sum of squares function $R(t)$.

Let X be a continuous random variable with positive skewness, $f(x)$ and $F_n(x)$ be the probability density function and the empirical distribution of X, respectively. For ease of notation, we will denote the order statistics by a notation that is more commonly used for the sample. Let $X_1 \leq \cdots \leq X_n$ denote the order statistics of the sample. Then we have the following results.

(1) *$R(t)$ is a continuous function of t.*

(2) *$R(t)$ is piecewise differentiable. The derivative does not exist at $t = x_i$, and $R'(x_i^-) < R'(x_i^+)$, $i = 1, \ldots, n$.*

(3) *If $|f'(t)| \leq M$ for some M and $\lambda_n = o(1)$, then the minimum value of $R(t)$ exists and is unique.*

All the proofs in this paper are given in Appendix I.

Based on these results we have the following theorem.

Theorem 3.1. *Suppose $|f'(t)| \leq M$, for some $M < \infty$ and $\lambda_n = o(1)$. Then the penalized sum of squares function $R(t)$ is minimized by the solution to the equation*

$$t = \bar{x} + \lambda_n f(t)\{1 - F_n(t)\}, \tag{3.1}$$

if it exists, where $F_n(t)$ is the empirical distribution function of X, and $f(\cdot)$ is the probability density function of X.

The solution of (3.1) is not a viable penalized least squares estimator of θ, since (3.1) includes the unknown probability density function of the population. To find a suitable estimator, a density estimator $\hat{f}(t)$ is substituted into Equation (3.1), and a new estimator of the mean is defined as follows.

Definition 3.1. If $(X_1, \ldots, X_n)$ denotes a random sample from a positively skewed distribution, then the solution $\hat{\theta}$ to the equation

$$\hat{\theta} = \bar{X} + \lambda_n \hat{f}(\hat{\theta})\{1 - F_n(\hat{\theta})\} \tag{3.2}$$

is called a penalized mean, where $\hat{f}(\cdot)$ is a density estimator of X, $F_n(\cdot)$ is the empirical distribution function of X and λ_n is a penalty constant.

Equation (3.2) defining $\hat{\theta}$ includes an estimate of the underlying density function, whose complete knowledge would make estimation of θ unnecessary. But complete knowledge of the density is not needed. We need an estimate of the density function only at one point which we approximate by $\hat{f}(\hat{\theta})$. Since $\hat{\theta}$ is an estimate of the mean, it will not be in the tail of the distribution, and therefore estimating $f(\hat{\theta})$ is not difficult.

The penalized mean is defined implicitly. Ideally, it can be obtained iteratively. However, due to the substitution of $\hat{f}(\theta)$ in (3.1), the recursive algorithm utilizing an initial estimate $\hat{\theta}_0$,

$$\hat{\theta}_{k+1} = \bar{X} + \lambda_n \hat{f}(\hat{\theta}_k)\{1 - F_n(\hat{\theta}_k)\}, \quad k \geq 0,$$

may not be convergent. In simulations, we have found that if we choose $\bar{x}$ as a starting point, most of the improvement of the estimate occurs after one step. Thus we define the one-step iterate of the penalized mean as

$$\hat{\theta} = \bar{X} + \lambda_n \hat{f}(\bar{X})\{1 - F_n(\bar{X})\}, \tag{3.3}$$

where λ_n is a constant and $F_n(\cdot)$ is the empirical distribution function of X.

For simplicity, the one-step iterate of the penalized mean is called the pmean.

Although we are interested in the small sample properties of the pmean, the following theorem shows that even though the pmean is biased, it has other large sample properties similar to the sample mean.

Theorem 3.2. *Let $\hat{f}_n(x)$ be a density estimator of $f(x)$ and $\hat{\theta}_n$ be the one-step iterate of the penalized mean. Assume that*
(1) $\|\hat{f}_n - f\| \to_{a.s.} 0$, *as* $n \to \infty$; ($\|f\| = \sup_x |f(x)|$)
(2) $\lambda_n \to 0$, *as* $n \to \infty$.
Then
(i) $\hat{\theta}_n \to_{a.s.} \theta$, *as* $n \to \infty$;
(ii) *If* $\lambda_n = cn^{-1/2}$, *then,* $n^{1/2}(\hat{\theta}_n - \theta) \to_L N(cf(\theta)\{1 - F(\theta)\},\ \sigma^2)$, *as* $n \to \infty$, *where c is a constant and $F(\cdot)$ is the distribution function of X.*

Theorem 3.2 shows that the pmean is strongly consistent for estimating the true mean. Its asymptotic bias is related to $f(\theta)\{1 - F(\theta)\}$. The values of $f(\theta)\{1 - F(\theta)\}$ for the distributions lognormal(0,1), lognormal(0,1.25), lognormal(0,1.5), $\chi^2(1)$, exponential(1) and Weibull(1,0.5) are listed in Table 1. In Section 4, we shall propose a choice of the penalty constant $\lambda_n = 4.5\sigma^2 n^{-1/2}$. The asymptotic bias, $4.5\sigma^2 g(\theta)$, and the probability for $\hat{\theta}$ to fall in the allowance range $[\theta, \theta + 2\sigma n^{-1/2}]$ of the pmean using such a penalty constant are also shown in the table.

Table 1. Asymptotic results of the PMEAN

	logN(0,1)	logN(0,1.25)	logN(0,1.5)	$\chi^2(1)$	E(1)	Weib(1,0.5)
θ	1.6487	1.8682	2.1170	1.0000	1.0000	2.0000
σ^2	4.6708	8.6922	15.604	2.0000	1.0000	20.000
$g(\theta)$	0.0659	0.0435	0.0289	0.0768	0.1353	0.0209
$4.5\sigma^2 g(\theta)$	1.3849	1.6998	2.0299	0.6910	0.6091	1.8807
p_2	0.6795	0.6412	0.6269	0.6224	0.6468	0.6057

Note: θ and σ^2 denote mean and variance of the distribution, respectively.
$g(\theta) = f(\theta)\{1 - F(\theta)\}$ and $p_2 = \text{pr}(\theta \le \hat{\theta} \le \theta + 2\sigma n^{-1/2})$

Note that the asymptotic bias of the UCL estimator is 1.96σ. Thus, the probability of the UCL falling within the allowance range is only 0.491 which is more than 10% lower than that of the pmean. The asymptotic bias of the UCL is much larger than that of the pmean. Therefore, when the sample size is large, the pmean is preferable.

4. A Choice of the Penalized Constant

We have found a point estimator of the population mean for a positively skewed distribution, the one-step iterate of the penalized mean – pmean, under penalized least squares. Although the pmean has large sample properties similar to those of the sample mean, in our particular application the small sample behavior of the pmean is of greater concern. For a study of the small sample behavior of the pmean we consider the lognormal distribution, which has received special attention in modeling pollutant concentrations.

Figure 1 shows the scatter plot of the penalty factor $\hat{f}(\bar{x})\{1-F_n(\bar{x})\}$ versus $\bar{x}$, in 1,000 repetitions with sample size 25 from the lognormal(0,1) distribution, where $\hat{f}(\cdot)$ is the k-NN (Nearest Neighbor) density estimator with $k=n^{1/2}$ (See Appendix II). From the figure, we can see the negative correlation between the sample mean and the penalty term for a fixed penalty constant λ_n. Note that

$$\begin{aligned}\mathrm{Var}\,(\hat{\theta}) \;&=\; \mathrm{Var}\,(\bar{X}) + 2\lambda_n \,\mathrm{Cov}\,(\bar{X}, \hat{f}(\bar{X})\{1-F_n(\bar{X})\}) \\ &\quad + \lambda_n^2 \mathrm{Var}\,(\hat{f}(\bar{X})\{1-F_n(\bar{X})\}).\end{aligned}$$

For certain λ_n, it is possible for the variance of the pmean to be smaller than the variance of $\bar{X}$.

The vertical line in the graph indicates the position of the population mean. It is easy to see that the factor $\hat{f}(\bar{x})\{1-F_n(\bar{x})\}$ gives small weight to λ_n when $\bar{x}$ is larger than the population mean. Thus a large value of the sample mean would get a smaller penalty. This would result in more shrinkage of the distribution of pmean towards the true mean.

Figure 2 shows the sampling distribution of $\bar{X}$ and the pmean in 1,000 repetitions with $\lambda_n = 1, \ldots, 6$. (Note: when $\lambda_n = 0$, pmean $= \bar{X}$.) It is interesting that when the λ_n increase from 1.0 to 4.0, the maximum value of the distribution does not increase much but the distribution of the pmean shrinks from the left towards to the true mean. The minimum value, MIN, maximum value, MAX, variance, VAR, bias, BIAS, the ratio of the mean square error of the pmean to that of the $\bar{X}$, RATIO, and the coverage, P_c of the pmeans are calculated and listed in Table 2. From Figure 2 and Table 2, we can see the changes of the sampling distribution along with the penalty constant. The variance of the pmean is less than that of the $\bar{X}$ even when the coverage of pmean reaches 0.983 in the case of $\lambda_n = 6$. When $\lambda_n = 4$, the coverage of the pmean reaches 90.7% which is twice as large as that of the $\bar{X}$. The ratio of the mean square errors is only 1.7162. Compared with $\bar{X}$, the maximum value of pmean is only 0.0232 units greater than that of $\bar{X}$, meanwhile the minimum value of the pmean increases to 1.3449 from 0.7651. Furthermore, when $\lambda_n = 1$, or 2, the ratio of the mean square error is less than one.

The penalty constant λ_n should be specified before applying the penalized mean. In sampling from strongly positively skewed distributions, the distribution of the sample mean remains quite positively skewed for small to moderate sample sizes. The pmean is a modification of the sample mean imposing a non-constant penalty factor $\hat{f}(\bar{x})\{1 - F_n(\bar{x})\}$ to improve coverage of the population mean. This penalty factor allows us to minimize the mean square error while retaining a large probability of falling into the allowance range $[\theta, \theta+2\sigma_{\bar{X}}]$ of the estimation. According to Tchebysheff's theorem, at most 5% of the observations are below 4.5 standard deviations of the mean. Since the density $f(x)$ is generally inversely proportional to the scale parameter σ, we define λ_n to be $4.5\sigma^2 n^{-1/2}$ so that the additive penalty term is proportional to σ. Because the penalty factor $\hat{f}(\bar{x})\{1 - F_n(\bar{x})\}$ is negatively correlated with $\bar{x}$, we expect that the pmean, with this penalty constant, will have a small variance with the desirable coverage. Here we assume the population variance is known.

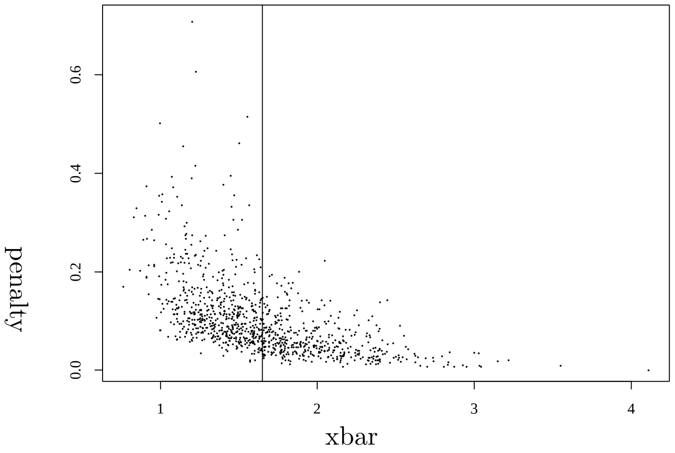

Figure 1. Penalty plot for Lognormal(0,1) with $n = 25$.

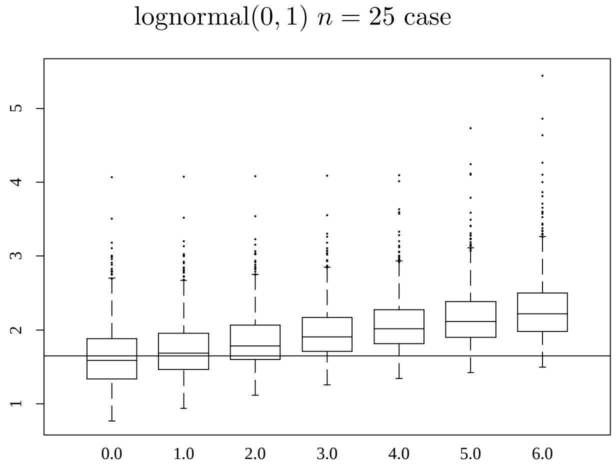

Figure 2. Boxplot comparison by varying penalty constant

Table 2. Statistical summary of the PMEAN for the Lognormal(0,1) distribution with $n = 25$

λ_n	MIN	MAX	VAR	BIAS	RATIO	P_c
0	0.7651	4.1095	0.1817	$-.0073$	1.0000	0.4250
1	0.9398	4.1153	0.1521	0.1000	0.8916	0.5400
2	1.1145	4.1211	0.1343	0.2073	0.9752	0.6960
3	1.2578	4.1269	0.1283	0.3146	1.2497	0.8150
4	1.3449	4.1327	0.1340	0.4219	1.7162	0.9070
5	1.4188	4.7692	0.1515	0.5291	2.3735	0.9650
6	1.4924	5.4827	0.1808	0.6364	3.2228	0.9830

5. Monte Carlo Study

5.1. Description of the Monte Carlo experiment

The three procedures studied are $\bar{X}$, UCL and pmean. Random variables were generated from positively skewed distributions which included lognormal distributions with parameters $\mu = 0$ and $\sigma^2 = 1.00, 1.25$, and 1.50, χ_1^2, exponential($\lambda = 1$) and Weibull($a = 1, b = 0.5$). The notation of Mood, Graybill, and Boes (1974), appendix B has been used to denote the distribution parameters. Because we are more concerned with the small or moderate large sample behavior of the estimators, only sample sizes of 15, 25, 35 and 50 were used in the simulation. The sampling distributions of the estimators were simulated using a Monte Carlo sample of size 5,000.

5.2. Simulation results

The empirical results presented in Tables 3 and 4 are variance, bias, mean square error and the proportion of the estimates falling in the allowance range $[\theta, \theta + 2\sigma n^{-1/2}]$ for each estimator in our study. It can been seen from the tables that except for $n = 15$, the variance of the pmean is smaller than that of $\bar{X}$. The bias of the pmean decreases when the sample size increases and it is smaller than that of the UCL. Because we have assumed the variance of the population is known, the variance of the UCL is the same as that of $\bar{X}$. Hence, the UCL has larger mean square error than the pmean. For an estimator which has a normal limiting distribution when the sample size is large, a proper location shift will result in a proportion of 68.26% of the estimates falling in the allowance range. However, the proportion of the values of pmean falling in the allowance range can reach 85% (see lognormal(0,1.50), $n = 35$). It should be noted that when the skewness of a population is large, such as for lognormal(0,1.25), lognormal(0,1.50) and Weibull(1,0.5), the results for the pmean are surprisingly good — compared to UCL, the pmean has lower mean square error and higher coverage within the allowance range. Even for distributions with small skewness, the results for the pmean are still much better than that of the UCL.

Figure 3 shows a comparison among three estimators of their variance, bias, mean square error and percentage of estimates falling within the allowance range (denoted by p_2) in lognormal(0,1) case.

5.3. Notes on computational techniques

Fortran 77 was used in the Monte Carlo studies reported in this article. CMLIB random-generating functions UNI and RNOR were used to generate pseudo uniform and normal random numbers. The other random variates were created from uniform random numbers or normal random numbers using standard techniques. All programs were run on a VMS Dec-Alpha-7620.

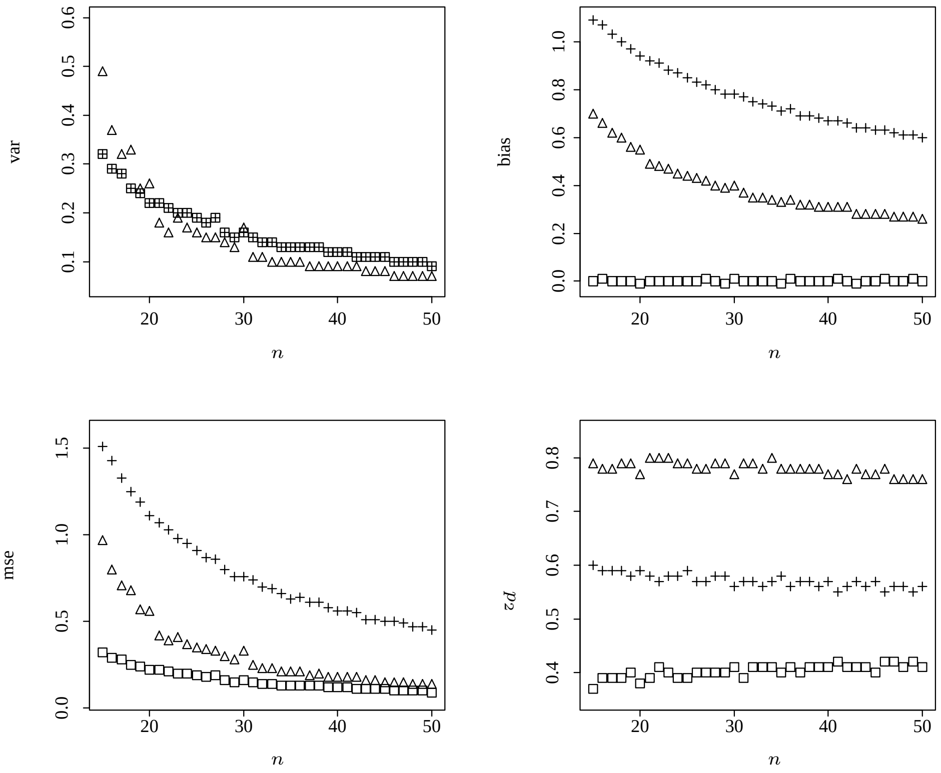

Figure 3. Comparison of $\bar{X}$, UCL and pmean in lognormal(0,1) case

Note: "□" denotes $\bar{X}$, "+" denotes UCL and "△" denotes pmean. The upper left, lower left, upper right and lower right figures are scatter plots of variance (var) vs sample size, mean square error (mse) vs sample size, bias vs sample size and the proportion of the estimates falling within allowance range (p_2) vs sample size respectively.

Table 3. Summary of behaviors of $\bar{X}$, UCL and pmean in lognormal cases

		logN(0,1.00)				logN(0,1.25)				logN(0,1.50)			
n	est.	var	bias	mse	p_2	var	bias	mse	p_2	var	bias	mse	p_2
15	$\bar{X}$	0.32	0.00	0.32	0.37	0.58	-0.01	0.58	0.36	1.16	0.02	1.16	0.37
	UCL	0.32	1.09	1.51	0.60	0.58	1.48	2.78	0.62	1.16	2.02	5.25	0.61
	pmean	0.49	0.70	0.97	0.79	0.99	1.01	2.01	0.80	2.11	1.50	4.35	0.79
25	$\bar{X}$	0.19	0.00	0.19	0.39	0.34	-0.01	0.34	0.39	0.58	-0.02	0.58	0.37
	UCL	0.19	0.85	0.91	0.59	0.34	1.15	1.66	0.59	0.58	1.53	2.92	0.61
	pmean	0.16	0.44	0.35	0.79	0.29	0.62	0.68	0.83	0.55	0.90	1.35	0.84
35	$\bar{X}$	0.13	-0.01	0.13	0.40	0.27	0.00	0.27	0.39	0.46	0.01	0.46	0.38
	UCL	0.13	0.71	0.63	0.58	0.27	0.98	1.22	0.59	0.46	1.32	2.20	0.59
	pmean	0.10	0.33	0.21	0.78	0.19	0.48	0.42	0.82	0.32	0.66	0.76	0.85
50	$\bar{X}$	0.09	0.00	0.09	0.41	0.18	-0.01	0.18	0.39	0.33	0.00	0.33	0.39
	UCL	0.09	0.60	0.45	0.56	0.18	0.81	0.83	0.58	0.33	1.09	1.52	0.59
	pmean	0.07	0.26	0.14	0.76	0.12	0.35	0.24	0.79	0.22	0.49	0.46	0.83

Note: For each distribution, the last column p_2 indicates the proportion of estimates that fell into $[\theta, \theta + 2\sigma n^{-1/2}]$.

Table 4. Summary of behaviors of $\bar{X}$, UCL and pmean in other cases

		$\chi^2(1)$				$E(1)$				Weibull(1,0.5)			
n	est.	var	bias	mse	p_2	var	bias	mse	p_2	var	bias	mse	p_2
15	$\bar{X}$	0.14	0.00	0.14	0.41	0.06	0.00	0.06	0.44	1.30	-0.01	1.30	0.36
	UCL	0.14	0.72	0.66	0.56	0.06	0.51	0.32	0.53	1.30	2.26	6.38	0.62
	pmean	0.13	0.32	0.23	0.71	0.07	0.25	0.13	0.71	2.25	1.33	4.01	0.81
25	$\bar{X}$	0.08	-0.01	0.08	0.42	0.04	0.00	0.04	0.45	0.77	-0.01	0.77	0.38
	UCL	0.08	0.55	0.38	0.55	0.04	0.39	0.19	0.53	0.77	1.75	3.82	0.60
	pmean	0.07	0.20	0.11	0.69	0.04	0.17	0.06	0.69	0.57	0.72	1.09	0.79
35	$\bar{X}$	0.06	0.00	0.06	0.43	0.03	0.00	0.03	0.44	0.57	0.00	0.57	0.39
	UCL	0.06	0.47	0.28	0.54	0.03	0.33	0.13	0.54	0.57	1.48	2.76	0.58
	pmean	0.05	0.16	0.07	0.69	0.03	0.13	0.04	0.69	0.38	0.54	0.67	0.76
50	$\bar{X}$	0.04	0.00	0.04	0.44	0.02	0.00	0.02	0.44	0.39	0.00	0.39	0.40
	UCL	0.04	0.39	0.19	0.53	0.02	0.28	0.10	0.53	0.39	1.24	1.92	0.57
	pmean	0.03	0.12	0.05	0.66	0.02	0.11	0.03	0.67	0.27	0.39	0.43	0.70

Note: For each distribution, the last column p_2 indicates the proportion of estimates that fell in to $[\theta, \theta + 2\sigma n^{-1/2}]$.

6. Conclusions

In this article, we have developed a new estimator, the pmean, for the mean of skewed distributions when underestimation of the mean is undesirable. The pmean involves a penalty constant λ_n. For a given allowance range $[\theta, \theta+2\sigma n^{-1/2}]$, we suggest choosing $\lambda_n = 4.5\sigma^2 n^{-1/2}$. The simulation results show that when the sample size is small, the pmean with this choice of λ_n is better than the UCL estimator currently being used by EPA in the sense of having smaller mean square error and larger probability of falling in the allowance range.

Under some mild assumptions, we have derived the large sample properties of the pmean. The pmean is asymptotically normally distributed. If we choose $\lambda_n = 4.5\sigma^2 n^{-1/2}$, the asymptotic bias of the pmean is less than σ in all the cases we have studied. This implies that the probability of the pmean falling within the allowance range decreases when n increases. However, from Table 1, one can see that the asymptotic probability of the pmean falling within the allowance range can still reach more than 60%.

The choice of the penalty constant is not unique. It can vary based on different applications. However, we point out that it must be related to the variance of the underlying distribution; and to keep its normal limiting distribution, its order must be $O(n^{-1/2})$.

The pmean is a very attractive biased estimator. The choices of the penalty constant for the pmean are under further investigation.

Acknowledgement

The first author's research was supported by U.S. EPA grant No. R821153-01-0. The authors wish to thank an anonymous referee for his/her helpful comments and suggestions.

Appendix I. Proofs

1. Proof of the properties of the sum of squares function $R(t)$ defined in (2.2):

(2) Let $Q(t) = \sum_{i=1}^{n} \mathrm{pr}(t < X < x_i)$. The result follows from the fact that

$$Q'(t) = \sum_{n=k+1}^{n} \{-f(t)\} = -f(t)(n-k), \quad x_k < t < x_{k+1},$$

and hence $Q'(x_k^-) < Q'(x_k^+)$.

(3)
$$R(t) = \sum_{i=1}^{n} (x_i - t)^2 + 2\lambda_n \sum_{i=1}^{n} P(t < X < x_i).$$

$$\frac{\partial R(t)}{\partial t} = -2\sum_{i=1}^{n} (x_i - t) - 2\lambda_n \sum_{i=1}^{n} f(t) I_{\{t<x_i\}}. \tag{A.1}$$

$$\frac{\partial^2 R(t)}{\partial t^2} = 2n - 2\lambda_n f'(t) \sum_{i=1}^{n} I_{\{t<x_i\}}.$$

For $f'(t) < 0$, we have $\partial^2 R(t)/\partial t^2 > 0$. Otherwise,

$$\frac{\partial^2 R(t)}{\partial t^2} \geq 2n - 2\lambda_n n f'(t) \geq 2n(1 - \lambda_n M).$$

Since $\lambda_n \to 0$, *as* $n \to \infty$, by choosing n large so that $\lambda_n < 1/M$, we have $\partial^2 R(t)/\partial t^2 > 0$, where $t \neq x_i,\ i = 1, \ldots, n$. Combining the above and assertion (2), we conclude that $R(t)$ is strictly convex and continuous. Hence, the minimum value of $R(t)$ exists and is unique.

2. Proof of Theorem 3.1.:

Let $\partial R(t)/\partial t$ in (A.1) be equal to zero. Solving for t yields $t = \bar{x} + \lambda_n f(t)\{1 - F_n(t)\}$. By result (3), this result holds.

3. Proof of Theorem 3.2.:

(i) According to Condition (1), with probability one for large n, we have

$$\|\hat{f}(\bar{X})\{1 - F_n(\bar{X})\}\| \leq \|\hat{f}(\bar{X})\| < M < \infty.$$

Thus,

$$\lambda_n \hat{f}(\bar{X})\{1 - F_n(\bar{X})\} \to_{a.s.} 0.$$

Note that when $\bar{X} \to_{a.s.} \theta$, the result holds.

(ii) $n^{1/2}(\hat{\theta}_n - \theta) = n^{1/2}(\bar{X} - \theta) + n^{1/2}\lambda_n \hat{f}(\bar{X})\{1 - F_n(\bar{X})\}$.

By $\lambda_n = cn^{-1/2}$, we have

$$n^{\frac{1}{2}} \lambda_n \hat{f}(\bar{X})\{1 - F_n(\bar{X})\} \to_{a.s.} cf(\theta)\{1 - F(\theta)\}.$$

Note that $n^{1/2}(\bar{X} - \theta) \to_L N(0, \sigma^2)$. Thus the result holds.

Appendix II. Definition of k-NN Density Estimator

Let k_n be an integer between 1 and n. Given any x, let

$$\begin{aligned} a_n(x) &= a_n(x;\ X_1, \ldots, X_n) \\ &= \min\{\ a \mid \#\ k_n \text{ of } X_1, \ldots, X_n \in [x - a,\ x + a]\ \}, \end{aligned}$$

then $\hat{f}_n(x) = k_n/(2na_n(x))$. If k_n satisfies
(1) $k_n/n \to 0$;
(2) $\sum e^{-ck_n} < \infty$, for any $c > 0$, (A.2)
then, $\lim_{n\to\infty} \hat{f}_n(x) = f(x)$, a.s. for all continuous points x of $f(x)$, where $f(x)$ is the probability density function of the population.

Note that an important case for satisfying (A.2) is $\lim_{n\to\infty} \log(n)/k_n = 0$. Therefore, in our simulation, we chose $k_n = n^{1/2}$.

References

Duan, Naihua (1983). Smearing estimate: A nonparametric retransformation method. *JASA* **78**, 605-610.

EPA (1992). *Supplemental Guidance to RAGS: Calculating the Concentration Term*, Intermittent Bulletin, Vol.1, No.1, Publication 92-081.

EPA (1989). *Risk Assessment Guidance for Superfund: Volume I - Human Health Evaluation Manual (Part A)*, EPA/540/1-89/002.

Fuller, W. A. (1970). Simple estimators for the mean of skewed populations. *Report to the U.S. Bureau of the Census*, Iowa State University, Ames, Iowa.

Fuller, W. A. (1991). Simple estimators for the mean of skewed populations. *Statist. Sinica* **1**, 137-158.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the Theory of Statistics.* John Wiley, New York.

Shumway, R. H., Azari, A. S. and Johnson, P. (1989). Estimating mean concentrations under transformation for environmental data with detection limits. *Technometrics* **31**, 347-356.

Taylor, Jeremy M. G. (1985). Measures of location of skew distributions obtained through box-cox transformations. *JASA* **80**, 427-432.

Taylor, Jeremy M. G. (1986). The retransformed mean after a fitted power transformation. *JASA* **81**, 114-118.

Department of Statistics, Florida International University, Miami, FL 33199, U.S.A.

Department of Mathematics and Statistics, The American University, 4400 Massachusetts Ave. N. W., Washington, DC 20016, U.S.A.

(Received May 1994; accepted July 1995)