The Wide-Sense Parametric Coverage Estimator against the Distribution Mismatch Problem for Sparse Data

Wen-Hui Lo[*], *Sin-Horng Chen*[**]

***Abstract*— Robust parameter estimation of sparse data is generally applied to the tasks when data collection is time-consuming or of high cost. We point out a new problem caused by sparse data. We find that there may exists coverage mismatch between data samples and the population when the sample size is less than 20. We call it the distribution mismatch (DM) problem. In this study, we derive a wide-sense joint *pdf* for coverage, range, the sample of minimum order, and data samples themselves to analyze the DM problem. Based on the formulation, a new algorithm is proposed to compensate the DM problem. Experimental results show that the mean estimate of the algorithm will converge to the population mean if the standard deviation of population is known.**

***Index Terms*: distribution mismatch, sparse data, coverage, time-consuming data collection**

I. Introduction

Parameter estimation for sparse data is sometimes used in electronic device testing for lifetime predication over a small sample size of observations. It needs to face the problem of the sparse input data resulting from time-consuming or high-cost data collection. It is also an important issue in the field of data mining in the computer society. In this paper we address the data sparseness issue in parameter estimation. The study is focused on analyzing the mean and variance estimations of normally distributed random variables under the sparse data constraint. We attack the problem from a new viewpoint via introducing a new variable known as coverage. Coverage is a macro view of the sample data and has been exposited in the past, e.g. for the outlier examination. Hadeed (1990)[1] firstly used it in a classification application for the quality control of integrated circuit products. Real (2000)[2] used it from the viewpoint of "tolerance interval" to do the open set classification with integrated circuit, where tolerance interval was the realistic mapping interval from data samples to its population.

We want to point out that coverage is an important factor when we perform parameter estimation from data with small sample size. This opinion was also examined in [3]. But it adopted a nonparametric estimation approach owing to the fact of distribution free for the probability density function (*pdf*) of coverage. Recently, Chen (2006, 2007)[4][5] suggested a new parametric coverage interval to help to realize a parametric form of coverage *pdf*. In this study, we derive a wide-sense joint *pdf* about the coverage with the same meanings as [5]

[*] Wen-Hui Lo is with the Communication Engineering Department, National Chiao Tung University, Hsinchu, Taiwan. (phone: 886-3-571-2121 ext 54555; fax: 886-3-590-7897; e-mail: hs3341.cm90g@nctu.edu.tw);

[**] Prof. Sin-Horng Chen is with the Communication Engineering Department, National Chiao Tung University, Hsinchu, Taiwan (phone: 886-3-572-1691; fax: 886-3-571-0116; e-mail: schen@cc.nctu.edu.tw; address: 1001 University Road, Hsinchu, Taiwan 300, ROC)

II. Paper Review

Balarkrishnan and Clifford Cohen (1991)[6], Lloyd (1952)[7], Teichroew (1956)[8] have suggested a method, referred to as best linear unbiased estimation (BLUE), for parameter estimation of normal random variables using order statistics. It is a weighted least-square algorithm which is based on the Gauss-Markov least-square theorem. It is known that BLUE was popularly used for sparse data analysis. It is unbiased and more efficient if it takes the censoring sampling scheme.

Let x be a normal random variable with *pdf* $f_x(x) = N(u,\sigma^2)$. Assume that there are n independent observed samples $x_1,\cdots,x_n$ of x. Let $x_{1:n},\cdots,x_{n:n}$ are the ranked samples of $x_1,\cdots,x_n$ in increasing order. The BLUE estimator is calculated as the sum of products of the observations and properly-chosen coefficients. We define the standard normal transformation of the data by $\xi_i=(x_i-u)/\sigma$. Then, we have

$$
\begin{aligned}
&X_n = [x_1,\cdots,x_n]^T \\
&\xi = [\xi_1,\cdots,\xi_n]^T \\
&E\{\xi_{i:n}\} = \rho_{i:n} \\
&Cov(\xi_{i:n},\xi_{j:n}) = \beta_{i,j:n} \quad \text{for } 1 \le i,j \le n \text{ and } i < j \\
&E\{x_{i:n}\} = u + \sigma\xi_{i:n} \\
&E\{x\} = u\mathrm{I} + \sigma\xi \\
&\mathrm{I}_n = [1,\cdots,1]^T_{n\times 1}
\end{aligned} \tag{1}
$$

$$B = \sigma^2 I \tag{2}$$

where I_n is a n-dimensional all-1 vector and B is the

ISBN:978-988-17012-3-7

covariance matrix of *x*. Consider the generalized variance:

$$\left(X_n - u\mathrm{I}_n - \sigma\xi\right)^T B^{-1}\left(X_n - u\mathrm{I}_n - \sigma\xi\right) \tag{3}$$

Minimizing it with respect to u and σ, we obtain.

$$\begin{aligned} u\mathrm{I}_n^T B^{-1}\mathrm{I}_n + \sigma\mathrm{I}_n^T B^{-1}\xi &= \mathrm{I}_n^T B^{-1} X_n \\ u\xi^T B^{-1}\mathrm{I}_n + \sigma\xi^T B^{-1}\xi &= \xi^T B^{-1} X_n \end{aligned} \tag{4}$$

The solution of Eq.(4) is

$$\begin{aligned} u^* &= \left\{\frac{\xi^T B^{-1}\xi\mathrm{I}_n^T B^{-1} - \xi^T B^{-1}\mathrm{I}_n\xi^T B^{-1}}{(\xi^T B^{-1}\xi)(\mathrm{I}_n^T B^{-1}\mathrm{I}_n) - (\xi^T B^{-1}\mathrm{I}_n)^2}\right\} X_n \\ &= -\xi^T \Delta X_n = \sum_{i=1}^{n} \alpha_{1:i} x_{i:n} \end{aligned} \tag{5}$$

$$\begin{aligned} \sigma^* &= \frac{\mathrm{I}_n^T B^{-1}\mathrm{I}_n\xi^T B^{-1} - \mathrm{I}_n^T B^{-1}\xi\mathrm{I}_n^T B^{-1}}{(\xi^T B^{-1}\xi)(\mathrm{I}_n^T B^{-1}\mathrm{I}_n) - (\xi^T B^{-1}\mathrm{I}_n)^2} X_n \\ &= \mathrm{I}_n^T \Delta X_n = \sum_{i=1}^{n} \alpha_{2:i} x_{i:n} \end{aligned} \tag{6}$$

where u^* and σ^* are the estimated parameters, and $\alpha_{1:i}$ and $\alpha_{2:i}$ are weighting coefficients. These coefficients have been tabulated by Sarhan and Greenberg (1956,1962)[9,10], entries in the 1956 tables are given for sample size up to 10 and in 1962 up to 20.

Generally speaking, BLUE performs well in small sample size. But it needs a table to look up, and this is a shortcoming. The other technique used is the maximum likelihood estimation (MLE) which is often applied to truncated normal distribution in spare data condition. Clifford Cohen (1991)[11] derived the singly truncated and doubly truncated maximum likelihood estimator and found that they outperformed BLUE when the sample size was grater than 20. Cohen recognized the sparse data problem as a truncated normal *pdf* and defined its likelihood by

$$L = \left(\frac{U(x - x_{1:n}) - U(x - x_{1:n} - r)}{\sqrt{2\pi}\sigma(F_x(x_{1:n} + r) - F_x(x_{1:n}))}\right)^n \exp\left(-\sum_{i=1}^{n}\frac{(x_i - u)^2}{2\sigma^2}\right) \tag{7}$$

where $U(\cdot)$ is the unit step function, $x_{1:n}$ is the sample of minimum order, r is the range of the whole sample set, and $F_x(\cdot)$ is the cumulative distribution function (*cdf*) of $\boldsymbol{x}$.

III. DEFINE THE DISTRIBUTION MISMATCH PROBLEM

In this work, we present a new idea different from BLUE and MLE. First, we want to point out that there exists a coverage mismatch between the sample *pdf* and its population *pdf*. Fig. 1 displays the basic relation of normal random variable x, the minimum order $x_{1:n}$, the range of the whole sample r, and the sample's coverage c. All of them are random variables and affected to each other. We then write the coverage *pdf* according to the result of order statistics inference. It is easy to derive the *pdf* of range from order statistics:

$$\begin{aligned} f_{r|n}(r) &= \int_{-\infty}^{\infty} f_{x_{1:n},x_{n:n}}(x_{1:n}, x_{1:n} + r)\,dx_{1:n} \\ &= \int_{-\infty}^{\infty} n(n-1) f_x(x_{1:n}) f_x(r + x_{1:n})(F_x(r + x_{1:n}) - F_x(x_{1:n}))^{n-2}\,dx_{1:n} \end{aligned} \tag{8}$$

where $x_{n:n}$ is the maximum order of random variable. The coverage *pdf* can be found from Eq.(8) and expressed by

$$p_{c|n}(c) = n(n-1)c^{n-2}(1-c) \quad \text{for } c > 0 \tag{9}$$

Fig. 2 displays the coverage *pdf* for some *n*.

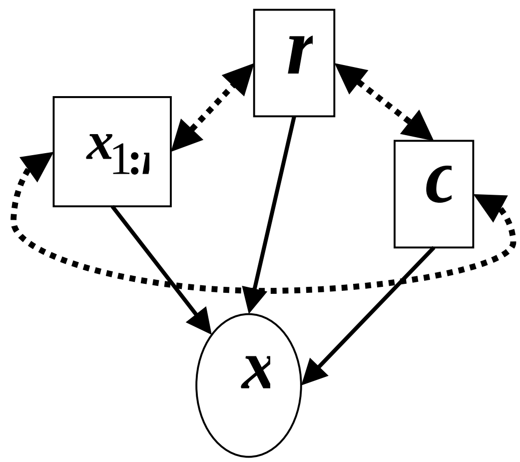

Fig. 1: Relation of variables' interference model

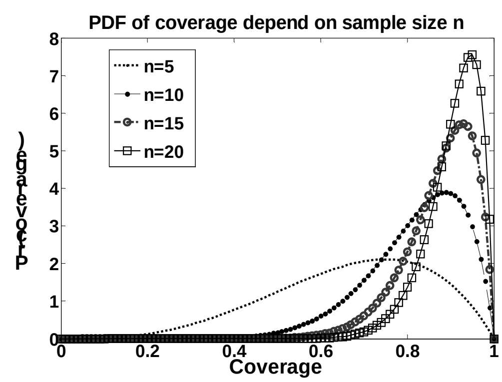

Fig. 2: The coverage *pdf* for some sample size

As shown in Fig.2, the *pdf* of coverage spreads away from 1 as *n* is less than 20, and the situation becomes more serious as *n* decreases. We call this phenomenon as the distribution mismatch (DM) problem.

We then derive a wide-sense normal joint *pdf* for coverage, range, the minimum order, and samples themselves based on Fig. 1 to solve the DM problem. We treat the wide-sense normal joint *pdf* as a variably truncated normal joint (VTNJ) *pdf*. It describes the randomness of the truncated points in the truncated normal distribution. Referring to Fig. 1, we express the joint *pdf* $p_{x,x_{1:n},r,c|n}(x, x_{1:n}, r, c)$ by

$$\begin{aligned} & p_{x,x_{1:n},r,c;u,\sigma|n}(x, x_{1:n}, r, c) \\ &= p_{x|x_{1:n},r,c,n}(x) \cdot p_{x_{1:n}|r,n}(x_{1:n}) \cdot p_{r|c,n}(r) \cdot p_{c|n}(c) \end{aligned} \tag{10}$$

where

ISBN:978-988-17012-3-7

$$p_{x;u,\sigma|x_{1:n},r,c,n}(x) = f_x(x)\frac{U(x-x_{1:n})-U(x-x_{1:n}-r)}{Q(x_{1:n},r)}$$

is the truncated normal *pdf* given with sample size, the truncated points and the sample's coverage, and

$$Q(x_{1:n},r) = F_x(x_{1:n}+r) - F_x(x_{1:n})$$

The other terms in Eq. (10), ($p_{x_{1:n}|r,n}(x_{1:n})$, $p_{r|c,n}(r)$, $p_{c|n}(c)$), must be derived from the original nonparametric order statistics *pdf* of range showing below:

$$p_{r|n}(r) = \int_{dx_{1:n}} n(n-1) f_x(x_{1:n}) f_x(x_{1:n}+r)(F_x(x_{1:n}+r) - F_x(x_{1:n})) \quad (11)$$

We discuss them as follows. First, the minimum order *pdf* can be derived from the Bayes' theorem and expressed by

$$p_{x_{1:n}|r,n}(x_{1:n}) = \frac{f_x(x_{1:n}) f_x(x_{1:n}+r)\left\{Q(x_{1:n},r)\right\}^{n-2}}{\sum_{i=1}^{m}\left\{\left[e^{\gamma_i^2} f_x(\gamma_i) f_x(\gamma_i+r) Q(x_{1:n},r)^{n-2}\right] w_m(\gamma_i)\right\}}$$

Second, $p_{c|n}(c)$ has been given in Eq. (9). Third, $p_{r|c,n}(r)$ can be derived with the Jacobian transform and expressed by

$$p_{r|c,n}(r) = \sum_{j=1}^{k}\left(\frac{f_x(\eta_j) f_x(\eta_j+r)\left\{F_x(\eta_j+r)-F_x(\eta_j)\right\}^{n-2}}{\left|f_x(\eta_j+r)-f_x(\eta_j)\right|}\right) \cdot \frac{1}{Z(r,c,n)}$$

where

$$Z(r,Cc_t,n) = \int_{dr}\left\{\sum_{j=1}^{k}\left[\frac{n(n-1) f_x(\eta_j) f_x(\eta_j+r)\left\{F_x(\eta_j+r)-F_x(\eta_j)\right\}^{n-2}}{\left|f_x(\eta_j+r)-f_x(\eta_j)\right|}\right]\right\}$$

IV. THE PROPOSED METHOD

The proposed method adopts an approach different from BLUE to use VTNJ marginal likelihood to formulate the closed-form equations for estimating mean and standard deviation. Because coverage is also a random variable in our formulation architecture, it is important to take into account the coverage of samples as interval estimation. The coverage interval is set in the form of the percentile of the relative order statistics and 0.15 percent fluctuation is added when we applied interval estimation for coverage in this study.

The whole estimation interval is represented in the following form:

$$[a,b] = [F_x(x_{n:n}) - F_x(x_{1:n}) - 0.15, F_x(x_{n:n}) - F_x(x_{1:n}) + 0.15] \quad (12)$$

The VTNJ *pdf* considers implemented by chain rules but such a decision make it must fact a difficult problem which there is no explicit transformation from coverage to range with the term $Q(x_{1:n},r) = F_x(x_{1:n}+r) - F_x(x_{1:n})$ in $p_{r|c,n}(r)$. To overcome this problem, the sampling concept is a natural solution. We first apply the Gauss Legendre Integration (GLI) to $p_{c|n}(c)$ for interval estimation where the sampling points may be decided by Legendre polynomials computing. If we apply GLI to $p_{c|n}(c)$, $p_{r|c,n}(r)$ is naturally a profile-conditional *pdf* as shown in Fig.3. Here k represents a constant for coverage.

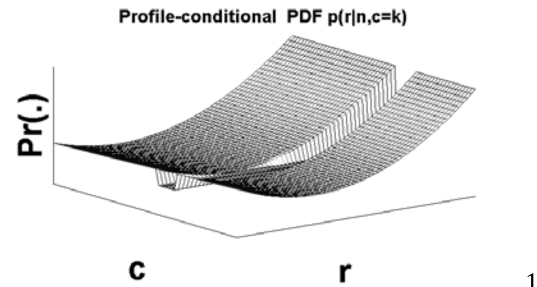

Fig.3: Profile-conditional *pdf* for $p_{r|c,n}(r)$ resulting from applying GLI to $p_{c|n}(c)$.

By using the GLI to implement the $p_{c|n}(c)$ of VTNJ *pdf*, we can approximate the joint distribution of x, $x_{1:n}$, r, and c by

$$\begin{aligned} p_{x,x_{1:n},r,c|n}&(x,x_{1:n},r,c=Cc_t) \\ &= f_x(x)\frac{U(x-x_{1:n})-U(x-x_{1:n}-r)}{Q(x_{1:n},r)} \\ &\cdot \frac{f_x(x_{1:n}) f_x(x_{1:n}+r)\left\{Q(x_{1:n},r)\right\}^{n-2}}{\sum_{i=1}^{m}\left\{\left[e^{\gamma_i^2} f_x(\gamma_i) f_x(\gamma_i+r) Q(x_{1:n},r)^{n-2}\right] w_m(\gamma_i)\right\}} \\ &\cdot \sum_{j=1}^{k}\left(\frac{f_x(\eta_j) f_x(\eta_j+r)\left\{F_x(\eta_j+r)-F_x(\eta_j)\right\}^{n-2}}{\left|f_x(\eta_j+r)-f_x(\eta_j)\right|}\right) \\ &\cdot \frac{1}{Z(r,c,n)} n(n-1)(c^{n-2}-c^{n-1}) \end{aligned} \quad (13)$$

where a and b are the endpoints of coverage interval shown in Eq.(14); $Cc_t = \frac{b-a}{2}\xi_t + \frac{b+a}{2}$, for $-1 < \xi_t < 1$, are the sampling points of coverage; t is the sampling index; γ_i is the *i*-th root of the *m*-th order Hermite polynomial; $w_{Hm}(\gamma_i) = \frac{2^{m-1} m! \sqrt{\pi}}{m^2\left[\text{Hermite}_{m-1}(\gamma_i)\right]^2}$ is the weighting coefficient for the *i*-th root of the *m*-th order Hermite polynomial; and η_j is the *j*-th root of $F_x(\eta_j+r) - F_x(\eta_j) - Cc_t = 0$; η_j must satisfy the following constrains:

$\eta_j \in$ □ ,
$f_x(\eta_j+r) - f_x(\eta_j) \neq 0$,
$lb \le \eta_j \le ub$, and
$Q(x_{1:n},r) = F_x(x_{1:n}+r) - F_x(x_{1:n})$;

ISBN:978-988-17012-3-7

$Cc_t = \frac{b-a}{2}\xi_t + \frac{b+a}{2}, -1 < \xi_t < 1$; $lb \le \eta_j + r \le ub$, lb and ub are empirically set to -4σ and 4σ , respectively;

$$Z(r, Cc_t, n) = \int_{dr} \left\{ \sum_{j=1}^{k} \left[\frac{n(n-1) f_x(\eta_j) f_x(\eta_j + r)\left\{F_x(\eta_j + r) - F_x(\eta_j)\right\}^{n-2}}{\left| f_x(\eta_j + r) - f_x(\eta_j) \right|} \right] \right\};$$

ξ_t is the t-th root of Legendre polynomial; and

$$P_v(x) = \frac{1}{2^v v!} \frac{\partial^v}{\partial x^v}(x^2 - 1)^v, \text{ for } v = 0, 1, 2, \cdots.$$

If we want to directly calculate the VTNJ *pdf* in $p_{r|c,n}(r)$, we will face the problem that the mean and standard deviation of the population must be known in advance. But this is unrealistic in our mission. We therefore adopt an alternative approach to construct a new bridge to conjoint with these variables. The idea is to transform the observed data into the standard normal domain. The suggestion is shown in Fig. 4. As shown in the figure, we transform the observed ranked samples into the domain of standard normal by $\xi_{i:n} = (x_{i:n} - u)/\sigma$. Each transform pair is marked with the same digit number. The coverage is also transformed by $r_s = \xi_{n:n} - \xi_{1:n}$.

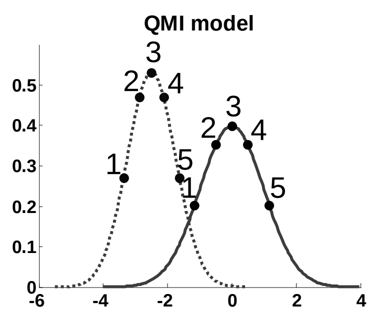

Fig. 4: Relative quantile mapping invariance based on their percentiles. Dash-line represents the original normal *pdf* and solid-line represents the standard normal *pdf*.

We then apply GLI to the VTNJ *pdf* to obtain the marginal log likelihood by

$$MLL(\cdot) \propto \sum_{t=1}^{v} \left\{ \frac{b-a}{2} n(n-1)(Cc_t^{\,n-2} - Cc_t^{\,n-1}) w_{P_v}(\kappa_t) \int_{dr_s} \int_{d\xi_{1:n}} G \right\} \quad (15)$$

where

$$G = \log\left\{ \left(\frac{1}{\sqrt{2\pi}\sigma\, \left(\Phi_\xi(\xi_{1:n} + r_s) - \Phi_\xi(\xi_{1:n})\right)} \right)^n \exp\left\{ -\sum_{i=1}^{n} \frac{(x_i - u)^2}{2\sigma^2} \right\} \right\} \cdot p_{\xi_{1:n}|r_s,n}(\xi_{1:n}) \cdot p_{r_s|c=Cc_t,n}(r_s)$$

An example of profile-conditional *pdf*, $p_{r|c,n}(r)$, is plotted in Fig. 5. It is to demonstrate the fact that if we would like to guarantee the coverage of the estimation large enough to above a lower bound, then there will be much more tolerance intervals qualified as the solution. Let we return to Eq.(9) to inspect the *pdf* of coverage which is distribution free. We find that its form is inconvenient for parameter estimation due to the no use of derivative operator. Fortunately, Chen [5] gave a good suggestion to the computation of coverage. In accordance with the conclusion of Chen [5], the *pdf* of coverage can be parametric if we constrain the tolerance interval (range) in its minimum case of all possible values. The plot shown in Fig. 5 demonstrates that $p_{r|c,n}(r)$ looks like an impulse with its distribution concentrating near the minimum-case. It is hence reasonable to take $Min\{r_s\}$ to substitute all other possible values of r_s . Fig. 5 also proves that our result is consistent to that of Chen [5].

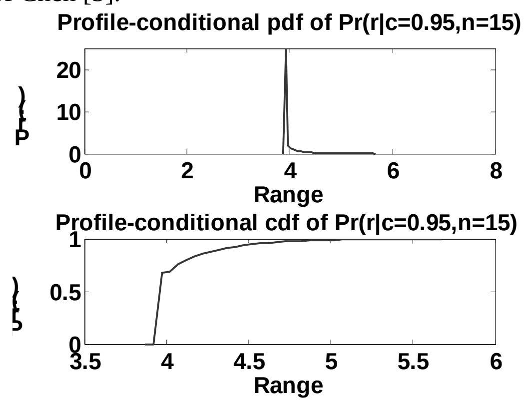

Fig. 5: Exemplified profile-conditional *pdf* to show the impulse properties for the sample size $n = 15$ and coverage=0.95 of standard normal *pdf*.

Eq.(16) can be optimized and simplified as a quadratic equation of variables σ and u . Take the roots of the quadratic equation will result in the following solutions:

$$\sigma^* = \frac{B_\sigma \pm \sqrt{(B_\sigma)^2 + 4\left(\sum_{t=1}^{v} nD_t\right) C_\sigma}}{2\left(\sum_{t=1}^{v} nD_t\right)} \quad (17)$$

where

$$B_\sigma = \left(\sum_{t=1}^{v} D_t \left(E_{\xi_{1:n}|c=Cc_t, Min\{r_s\}, n}\left\{\xi_{1:n}\right\} \right) \left(\sum_{i=1}^{n} (x_i - x_{1:n}) \right) \right),$$

$$C_\sigma = \left(\sum_{t=1}^{v} \left(D_t \sum_{i=1}^{n} (x_i - x_{1:n})^2 \right) \right),$$

$$D_t = \left(\frac{b-a}{2} n(n-1)(Cc_t^{\,n-2} - Cc_t^{\,n-1})\left(w_{P_v}(\kappa_t)\right)\right);$$

and

$$u^* = \frac{-B_u \pm \sqrt{B_u^2 - 4(\sum_{t=1}^{v} D_t) C_u}}{2(\sum_{t=1}^{v} D_t)} \quad (18)$$

Where

ISBN:978-988-17012-3-7

$$B_{u}=\sum_{t=1}^{v}\left\{D_t\left[\left(\overline{x}-x_{1:n}\right)\left(E_{\xi_{1:n}|c=Cc_t,Min\{r_s\},n}\left\{\xi_{1:n}^2\right\}\right)-2x_{1:n}\right]\right\},$$

$$C_u=\sum_{t=1}^{v}\left\{D_t\left[x_{1:n}^2+\left(\overline{x}x_{1:n}-\overline{x^2}\right)\left(E_{\xi_{1:n}|c=Cc_t,Min\{r_s\},n}\left\{\xi_{1:n}^2\right\}\right)\right]\right\},$$

$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ is the sample mean, $\overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2$ is the mean of sample square, $w_{P_v}(\kappa_t)$ is the weighting coefficient of the t-th root of the v-th order Legendre polynomial $\frac{(b-a)}{(1-\kappa_t^2)(P_v'(\kappa_t))^2}$, $[a,b]$ is the coverage estimation interval, and $\Phi_\xi(\xi)$ is the *cdf* of the standard normal distribution. The same strategy can be applied to the order statistics random variable via replacing $\xi_{n:n}$ by $\xi_{1:n}$.

V. Simulation Results

By checking Eqs.(17) and (19), we find that they are mainly affected by the sample mean, $\overline{x}$, and the individual order statistics random variables, $x_{i:n}, 1\le i\le n$. Our strategy is to adjust the coverage to make it approach to the real coverage, generated from $\overline{x}$ and $x_{i:n}, 1\le i\le n$, in order to compensate the DM effects. We examine two methods. One is to view the joint effect of $\overline{x}$ and $x_{i:n}$ under our suggestion, QMI. The other is to realize the QMI based only on the real coverage. Its purpose is to see only the effect of sample mean.

1) *Test the results with consistency to sample mean under the QMI principle—case of the default percentile*

We first formed an interval estimation for coverage by performing a coverage estimation from the expectation of order statistics by $F_x(E\{\xi_{n:n}\})-F_x(E\{\xi_{1:n}\})$ and adding fluctuation of ± 0.015. We then examined the accuracy of the conventional sample mean estimator. Two different conditions for sample mean were considered. One was to constrain the sample means in the interval of $-0.3\sigma+u\le\overline{x}\le 0.3\sigma+u$. It was referred to as the good sample mean case. The other was to constrain the sample means in the interval of $-2.3\sigma+u\le\overline{x}\le -1.3\sigma+u$ or $1.3\sigma+u\le\overline{x}\le 2.3\sigma+u$, and was referred to as the bad sample mean case. Three estimators were compared: A represented the conventional sample mean estimator; B was the coverage-based estimator defined below

$$u^*=u_p=x_{p:n}-\frac{\sum_{t=1}^{v}\left\{D_t\left[E_{\xi_{p:n}|Cc_t,Min\{w_p\},n}\left\{\xi_{p:n}\right\}\right]\right\}}{\sum_{t=1}^{v}D_t}\sigma_p \qquad (20)$$

where p was constrained to be either 1 or n which corresponded to the endpoints of the range; and C was the estimator defined in Eq.(21). If $p=n$, then the term $E_{\xi_{p:n}|Cc_t,Min\{w_p\},n}\left\{\xi_{p:n}\right\}$ can be computed by $(-1)E_{\xi_{1:n}|Cc_t,Min\{w_1\},n}\left\{\xi_{1:n}\right\}$. The results are displayed in Fig. 6. It can be found that the MSE were very small if the sample mean is near the population mean. This implies that if we want to obtain a guaranteed coverage, then the difference between the estimated mean and the sample mean should be small.

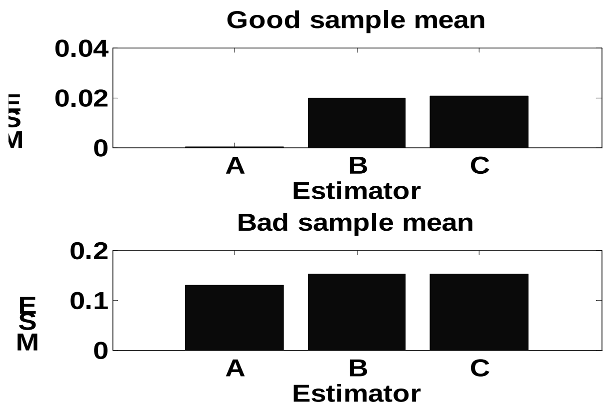

Fig. 6: Comparison of the conventional sample mean estimator and two coverage-based mean estimators.

2) *Test the results with consistency to sample mean under the QMI principle—case of realistic percentile*

In the test phase, we eliminated the effects caused by the QMI mapping mismatch for $\xi_{1:n}$ to $x_{1:n}$ or $\xi_{n:n}$ to $x_{n:n}$. In such a case, $\xi_{1:n}=(x_{1:n}-u)/\sigma$ and $\xi_{n:n}=(x_{n:n}-u)/\sigma$ were known. But, we pretended that we did not know u and σ. The fluctuation assumption for coverage was therefore not needed. So, the previous formulation could be simplified and expressed by

$$\sigma^*=\frac{\xi_{p:n}\left(\sum_{i=1}^{n}(x_i-x_{p:n})\right)}{2n}\pm\frac{\sqrt{\left(\xi_{p:n}\left(\sum_{i=1}^{n}(x_i-x_{p:n})\right)\right)^2+4n^2\left(\sum_{i=1}^{n}(x_i-x_{p:n})^2\right)}}{2n},\quad \text{for } \sigma^*>0 \qquad (22)$$

and

$$u^*=\frac{-\left(\left(\overline{x}-x_{p:n}\right)\left(\xi_{p:n}^2\right)-2x_{p:n}\right)}{2}\pm\frac{\sqrt{\left[\left(\overline{x}-x_{p:n}\right)\left(\xi_{p:n}^2\right)-2x_{p:n}\right]^2-4\left[x_{p:n}^2+\left(\overline{x}x_{p:n}-\left(\overline{x^2}\right)\right)\left(\xi_{p:n}^2\right)\right]}}{2} \qquad (23)$$

where p was constrained to be either 1 or n. Actually, Eq.(22) is equivalent to Eq.(23) because $u^*=x_{p:n}-\xi_{p:n}\sigma^*$. We generated 1,000 trials to examine the new estimator and used MSE as the score of comparison. The results are listed in Table 1.

Table 1: Performance of realistic QMI analysis

Item	Sample mean	Realistic QMI
MSE	0.0765	0.0252

ISBN:978-988-17012-3-7

Notice that the MSE of realistic QMI was defined by $\frac{1}{1000}\sum\left(\frac{(u_1+u_n)}{2}-u\right)^2$, where u_1 and u_n were the estimated results for $x_{1:n}$ and $x_{n:n}$, respectively. It can be found from Table 2 that the realistic QMI mean estimator performed better than the sample mean estimator.

VI. APPLICATION OF USING THE RESULTS OF REALISTIC QMI

The above testing results of realistic QMI show us that if we are able to take the relative coverage for the range, then we can probably reduce the bias of the sample mean. Now we utilize the above result for analysis in depth. The transform $u = x_{p:n} - \xi_{p:n}\sigma$ has only two degree of freedom. So, if $\xi_{p:n}$ is known, the degree of freedom will be reduced to 1. We will have an opportunity to approach the real value by iteration.

We performed 1,000 trials. In each trial, 13 samples of the normal random variable of $N(10,1)$ were generated. We assumed that the standard deviation was known to be 1 and the quantile $\xi_{p:n}$ was unknown.

In the beginning, a pseudo mean was tried. Then, we got a $\xi_{p:n}$. The estimator of u^* was then found. The pseudo MSE could be computed by $\frac{1}{1000}\sum\left(u^*-u_s\right)^2$, where u_s was the pseudo mean. As the pseudo MSE decreased, the pseudo mean would be closer to the true mean. Table 2 listed the experimental results. It can be seen from Table 2 that the pseudo MSE became low when the pseudo mean was close to the true population mean (=10).

Table 2: Results of realistic QMI via pseudo mean

Pseudo mean	Pseudo MSE	Pseudo mean	Pseudo MSE
9.70	0.0164	10.00	0.0059
9.72	0.011	10.02	0.0058
9.74	0.009	10.04	0.008
9.76	0.0117	10.06	0.0056
9.78	0.0114	10.08	0.0072
9.80	0.0107	10.10	0.0079
9.82	0.0063	10.12	0.0067
9.84	0.0113	10.14	0.0066
9.86	0.0096	10.16	0.0092
9.88	0.0097	10.18	0.0086
9.90	0.0056	10.20	0.0146
9.92	0.004	10.22	0.0112
9.94	0.0066	10.24	0.0118
9.96	0.0087	10.26	0.0156
9.98	0.0061	10.28	0.0109

VII. CONCLUSIONS

In this paper, the detail proof of parametric coverage has not been presented before and we show it in Fig. 5. In the other hand we have discussed the DM problem encountered in parameter estimation using data samples with size less than 20. The problem is addressed from the coverage of data samples. We referred the old skills and develop the advance truncated normal distribution, the variably truncated normal joint *pdf*, to formulate the DM problem. In the realization consideration, parametric coverage and QMI are also the core for successfully passing the test. In the default QMI test, we have showed that our coverage-based mean estimator follows the sample mean. It outperforms the sample mean in the realistic QMI case.

This conclusion of the realistic QMI implies that if either σ or u is known in our estimation process, we are able to approach the real mean by iteration. Finally let us mention a thing that the result of realistic QMI is also accepted for large sample size condition. The reason is that it is free to coverage.

VIII. REFERENCE

[1] Y. T. Hadeed, and K. T. Lewis, “The use of tolerance intervals in the characterization of semiconductor devices”, Inter. Test conf. proc., pp.924-928, 1990.

[2] Real, E. C. and Baumann, A. H., “Open set classification using tolerance intervals”, Signals, systems and computers, the 34th Asilomar conf. , vol. 2, pp.1217-1221, 2000.

[3] H. Chen and T.-K. Yang, “Estimation of the sample and coverage for guaranteed-coverage nonormal tolerance intervals”, Winter simulation conf., pp.593-600, 1998

[4] L-A. Chen and H-N. Huag, “Extending the discussion on coverage intervals and statistical coverage intervals”, Metrologia, vol. 43, L43-L44, 2006.

[5] L-A. Chen, J-Y. Huang and H-C Chen, “Parametric coverage interval”, Metrologia, vol. 44, L7-L9, 2007.

[6] N. Balarkrishnan, A. Clifford Cohen , “Order statistics and inference estimation methods”, Academic Press, Inc., 1991.

[7] E. H. Lloyd, “Least-square estimation of location and scale parameters using order statistics”, Biometrika, 39, pp.88-95, 1952.

[8] D. Teichroew, “Tables of expected values of order statistics for samples of size twenty and less from the normal distribution”, The Ann. of Math. Stat., vol. 27, pp.410-426, 1956.

[9]A. E. Sarhan and B. G. Greenberg, “Estimation of location and scale parameters by order statistics from singly and doubly censored samples, part one. The normal distribution up to size 10”, The Ann. of Math. Stat., vol. 27, pp.427-451, 1956, (correction , vol. 40, p.325)

[10] A. E. Sarhan and B. G. Greenberg. eds., “Contributions to order statistics”, Wiley, New York, 1962.

[11] A. Clifford Cohen, “Truncated and censored samples — theory and applications”, New York, Marcl Dekker, 1991.

ISBN:978-988-17012-3-7