Journal of Engineering Physics and Thermophysics, Vol. 83, No. 2, 2010

ESTIMATION OF ARBITRARY-DISTRIBUTION PARAMETERS FROM THE DATA OF A REPETITIVE EXPERIMENT

E. V. Chernukho

UDC 519.23:53.089.6

Regular statistics making it possible to efficiently estimate the parameters of arbitrary distribution for the case of a repetitive experiment is described and substantiated. The methodology of engineering mathematics is used for derivation and substantiation of the statistics.

Keywords: *regular statistics, estimation of parameters, repetitive experiment, calibration function.*

Introduction. A major portion of works of mathematical and applied statistics is devoted to estimation of distribution parameters and to the consideration of related problems. To date, the only theoretical distribution for which the problem of estimation of the shift and scale parameters is solved in an optimum manner, has been the distribution by the normal law. We propose regular statistics making it possible to efficiently estimate any parameters of the model of distribution of an arbitrary form. The distribution can be of any kind: theoretical or empirical, with a finite or infinite definition domain, with no variance, continuous or discrete. It is necessary that the distribution be described by a single-valued nonnegative function and be normalized on the probability principle.

Formulation of the Problem and Basic Result. Let point estimation of the parameter b of distribution prescribed by the probability density $p(x, b)$ and determined on the set or space X: $x, b \in X$, be made.

The data $\{d\}_n$ refer to the same experiment: $\forall d \in \{d\}_n \,|\, b = \text{const}$. The data are independent in the sense that the result of estimation should not change because of their interchange in the data set. This condition is necessary for the arithmetic operations used for data processing to correspond to the properties of data, namely, to ensure their commutative and associative properties. The methods of ensuring this condition require a separate discussion.

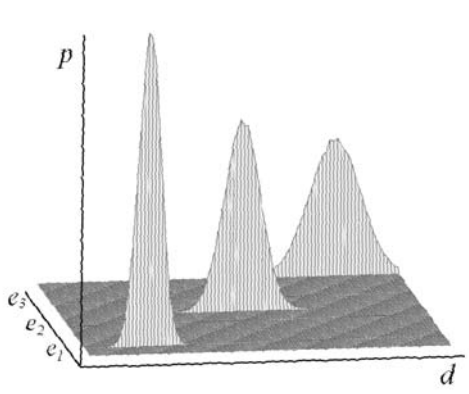
The statistics for point estimation of the parameter in the form of a compact formula is as follows:

$$\dot{b} = \frac{\int_{-\infty}^{\infty} bu(b)\, db}{\int_{-\infty}^{\infty} u(b)\, db}, \quad u(b) = \prod_{i=1}^{n} \frac{n!}{(i-1)!(n-i)!} (P(d_i, b))^{i-1} (1 - P(d_i, b))^{n-i} p(d_i, b). \tag{1}$$

Despite a certain cumbersomeness, the statistics can be represented by a simple computational algorithms each operation of which is easily substantiated. In fact, the proposed statistics has been obtained by successive application of engineering-mathematics methods: on the basis of the comparison principle, we have obtained a measure for estimating the distribution and the data set, have formulated data structures describing the uncertainty of the estimate, and have determined the analytical correction for natural bias. However, as is often the case, we are able to find the method of constructing isomorphisms between the statistics obtained on the traditional principle of substitution and the measures obtained on the comparison principle; this enables us to set forth the solution of the problem in question in a narrow formulation in the ordinary manner without presenting the entire procedure of derivation in detail. We resort to the proof constructed on the structural principle. The proof turns out to be so simple and evident that unnecessary mathematical formalization will be excessive.

A methodological problem exists. In mathematical statistics, there is no tool making it possible to substantiate the proposed statistics but it is available in metrology. This tool is a calibration characteristic (function). On the other

A. V. Luikov Heat and Mass Transfer Institute, National Academy of Sciences of Belarus, 15 P. Brovka Str., Minsk, 220072, Belarus. Translated from Inzhenerno-Fizicheskii Zhurnal, Vol. 83, No. 2, pp. 403–409, March–April, 2010. Original article submitted March 3, 2009.



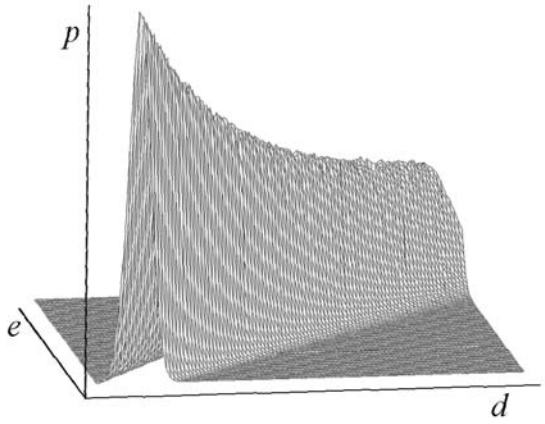


Fig. 1. Illustration of the calibration function for three standards and a large amount of data of the calibration experiment.

Fig. 2. Representation of the correction function in the form of a surface obtained from the calibration of function Fig. 1.

hand, the normal distribution law is of primary importance in metrology. This proposition can be explained. This results in a paradox. In metrology, the problem posed by us appears to be not so topical as in statistics, but there is a tool for its solution. Statistics, conversely, needs solution but has no tool.

A calibration characteristic (function) is obtained in the calibration experiment as the system of equations $\{\{d\} \leftrightarrow e\}$ but is used in the working experiment as the correction function $x = f(d)$. The method of realization of the correction function is dependent on many properties of the data and experiment. We use the classification of types of correction functions which consists of six levels. Substantiation of the obtained result requires the third level of this classification.

At the first level, the correction function is just an ordinary scalar data function. It is useful if the systematic error of an instrument is much larger than the random error; if the spread in experimental data is comparable to the instrumental resolution, it is the only possible one. It is precisely this type of correction function that has received the widest acceptance and is associated with the calibration function.

At the second level, the correction function is constructed as the interval function of the scalar argument $[x] = f(d)$. Legal metrology uses precisely this type. For example, the binomial formula of expression of an error or an uncertainty is nothing but the trapezoidal approximation of a real interval correction function.

At the third level, allowance is made for the real empirical data distribution. This is the maximum level used by theoretical metrology. The set of standards $\{e\}$ is used for calibration as a rule. For each of them, the distribution of experimental data $p(d \mid e)$ is measured. The resulting set of empirical distributions is interpolated to the surface $\{p(d \mid e)\} \rightarrow p(d, e)$, which represents the basis of a correction function. The correction function itself is computed as the section of this surface by the working experimental value $p(x) \mid d = f(p(d, e) \mid (d))$. As a result, each datum is compared to its distribution. The aforesaid is illustrated by Figs. 1–3.

In a repetitive experiment, we obtain the set of distributions $\{p(x) \mid d\}$ by the number of replications. The distributions may differ not only in shift and dispersion but also in form. We seek to evaluate the measured quantity by this data structure.

Theoretical statistics can offer the maximum likehood method (MLM): $\dot{x} = \max \left(\prod_{i=1}^{n} p(x - d_i) \right)$. On the one hand, the correction function may be considered as MLM substantiation. Indeed, sometimes the MLM leads to optimum estimates, as far as the efficiency is concerned, e.g., for the normal distribution. In other cases the efficiency is relatively high if not optimum, e.g., for the Cauchy distribution. Sometimes, the same MLM may lead to inefficient estimates; in this case the method of moments is used. Therefore, common sense enables us to state that the MLM uses important information contained in the form of distribution but ignores some other, no less important, information.

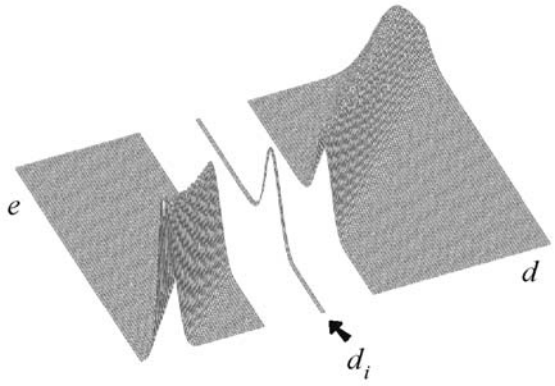


Fig. 3. Realization of the third-level correction function for an arbitrary datum (the surface of the correction function is cut and moved apart).

When the essence is clear, everything becomes simple. The difference in results is caused by the data sequence order. The commutativity of the MLM formula is consistent with the independence of data in the data set. The requirement of data independence is equivalent to the assumption that the use of the distribution is sufficient for estimation.

However, the distribution itself is not commutative, as far as the measured parameter is concerned: we cannot interchange buckets. Consequently, we cannot directly compare the distributions and data sets obtained immediately in the experiment. It is necessary to preorder the data set by using order statistics. Now it turns out to be unallowable to compare the order datum (numbered datum of the ordered data set) to the entire distribution as a unified structure. We have to compare each order datum only to its order distribution $d_{i/n} \leftrightarrow p_{i/n}(x)$, whence the estimate's kernel which can be interpreted as the uncertainty figure of the estimate of the shift parameter has the form

$$u(x) = \prod_{i=1}^{n} p_{i/n}(x - d_{i/n}), \tag{2}$$

where $p_{i/n}(x)$ is the density distribution of the ith datum from an ordered data set of length n on the density distribution $p(x)$.

Indeed, if we record the multiplicity of experiment, we will have to construct its own calibration function for each number of the order datum. The resulting distribution will be the empirical distribution of the order statistics. Figures 1–3 remain valid but only for each order datum individually.

A distinctive feature of order distributions is that their dispersion is much lower than the dispersion generating the distribution. Now each order datum is related to its (order) correction function. The parameter should be estimated by the set of order distributions obtained from these individual correction functions. The problem is substantially simplified by the fact that the formula for computing the distributions of the order statistics from the distribution of the generating process is known:

$$p_{i/n}(x) = \frac{n!}{(i-1)!(n-i)!} (P(x))^{i-1} (1 - P(x))^{n-i} p(x).$$

We illustrate the aforesaid by Fig. 4 which shows the initial distribution $p(x)$, three order statistics $p_{1/3}(x)$, $p_{2/3}(x)$, and $p_{3/3}(x)$, and their product $p_{\Pi}(x) = \dfrac{p_{1/3}(x), p_{2/3}(x), p_{3/3}(x)}{\int_{\Omega} p_{1/3}(x), p_{2/3}(x), p_{3/3}(x) dx}$ interpreted as the maximum dense uncertainty (ambiguity) function which can be obtained in the context in question. Experimental uncertainty functions will be more dispersed. We mention in passing that the average of the centers of gravity of the order densities coincides with the center of gravity of the initial distribution, i.e., the average of the gravity of distributions is an invariant in a sense.

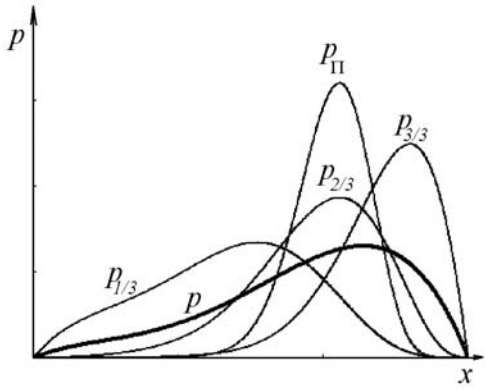


Fig. 4. Order statistics for an arbitrary distribution.

The problem will be simplified still further if the data set has a small dispersion, as far as a variation in the correction function is concerned, with the dispersion being so small that both the form and value of the distribution dispersion cannot substantially change. We will be able to assume that the distribution of the generating process is independent of the datum value. In this case computation of the uncertainty figure is reduced to the sequence of simple operations $\{d\} \to n \to \{p_{i/n}(x)\} \to u(x)$: we determine the value of the obtained data set, construct, from it, a set of order-statistics distributions, and compute the uncertainty function. For convenience of interpretation, the uncertainty function is normalized to the probability norm. Now the uncertainty figure could be interpreted as evidence for the probability that the sought parameter has the value of the argument.

The uncertainty function is, in essence, the complete estimate of the parameter measured in a repetitive working experiment. The completeness is understood as the use and conversion of all available a priori and experimental information on the value of the measured parameter. The experimental information is, naturally, the data set. The priori information is primarily contained in the correction function; also, observance of all the assumptions under which this correction function has been constructed is of importance.

The results of estimation is necessary for making a decision; in actual practice, the complete estimate is often excessive, and we used only a certain simplified estimate as a rule. Point and interval estimates have gained acceptance. We seek to obtain a point estimate $\dot{x} = s(u(x))$ from the uncertainty figure. In the context in question and in statistical terms, we are dealing with estimation of the shift parameter of distribution of the generating process.

Sometimes, we can use the mode of the uncertainty function to obtain a point estimate. Indeed, the uncertainty function is conveniently normalized to the probability measure; this can be reasonably explained. The value of the measured parameter is necessarily existent and must be unique (single). There appears a wish to interpret it as the probability density of the estimated parameter having the corresponding probability. It is natural that the best estimate is at the maximum of this probability. In particular, this is the recommendation of the MLM, which is justified for it.

However, usually the maximum statistics leads to a bias of the estimate or, in metrological terms, to a systematic error. Sometimes, the bias can be computed analytically or the systematic error can largely be compensated for with the correction function. The situation is complicated in the presence of several maxima.

The proposed statistics of estimation of the parameter of a repetitive experiment (1) gives no bias for the point estimate since the center of gravity of the uncertainty $\dot{x} = \int xu(x)dx$ is computed. This also eliminates the problem of ambiguity of the estimate but increases the volume of computations, since the tails of the figure are of importance. For symmetric figures, the center of gravity coincides with the maximum value.

We cannot give a theoretical substantiation of selecting the statistics of the center of gravity for the point estimation by the uncertainty figure on the basis of the axioms of probability theory, but it is not difficult to show the efficiency of the proposed statistics by a numerical experiment modeling the calibration experiment on obtaining a point estimate.

Now the structure of the statistics (1) is clear. The distribution densities of the order statistics are under the product sign. The integral of the numerator computes the center of gravity of the uncertainty figure. The denominator

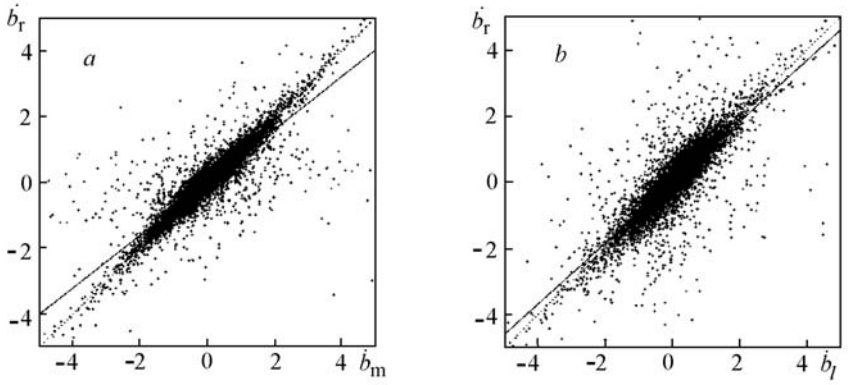


Fig. 5. Example of estimation of the relative efficiency (Cauchy distribution C(x, 0, 1), 10,000 experiments, size of the data set 5): a) the efficiency of the median with respect to the rank measure is ef = 0.768, the bias is sh = –0.0006, and the correlation coefficient is 0.782; b) the efficiency of the MLM with respect to the rank measure is 0.92 and the correlation coefficient is 0.82.

normalizes the uncertainty figure. In our opinion, the term "rank measure" will be appropriate for this statistics and for the method in general.

Formally, constructive derivation requires axioms, rules of derivation, and the chain of arguments itself. The method is based on the axiom of independence of data in a data set. It can be formulated as the requirement $\forall s$ [$\{b = s(\{d\})\}$:$s(\{d\}) \equiv s(\{d\})$.sort) that all statistics of the class of parametric estimation be independent of the interchange of data in a data set, in particular, be efficient for an ordered data set. As a consequence, the diagram $\{\{d\}\} \to p(x)$ is commutative: the empirical order distributions obtained from the data directly and computed from the

$\downarrow \qquad \downarrow$

$\{p_{i/n}(x)\}$

resulting empirical distribution are equivalent in a statistical sense.

We have used two rules of derivation: the rule of construction of a correction function and the rule of multiplication of probabilities of independent events. The derivation chain is elementary $\{d\} \xrightarrow{f} \{p_{i/n}(x - d_i)\} \xrightarrow{\Pi} u(x)$: from the experimental data, we compute the sections of order calibration functions and multiply them together to obtain the uncertainty function of the calibrated parameter. Besides, this is the only possible derivation chain and no alternative can be offered.

Estimation of the Relative Efficiency of the Rank Measure. Since we are dealing with estimation of the efficiency of determining the parameter for an arbitrary distribution, numerical experiment will be the only appropriate method to compare the efficiency of statistics. For this purpose, we generate a collection of test data sets and estimate the parameter by different methods. We compare the collection of estimates using the appropriate algorithm, computing efficiency, and bias. Although many algorithms for estimating the efficiency can be proposed, the algorithm of linear regression is the most obvious. The regression coefficient is interpreted as the efficiency, whereas the constant term is interpreted as the bias. From our observations, the regression criterion is of little importance, since the method is differential. The least-squares method is also suitable, since the estimate distribution is normal-like, even if with a relatively large correlation between the distributions. Finally, we follow the scheme $\{p(x, b) \to \{d\} \to (\dot{b}_r, \dot{b}_m)\} \xrightarrow{\text{LR}}$ (ef, sh). The aforesaid is illustrated by Fig. 5.

The dispersion figure appears typical. The main contribution to the increase in the efficiency is made by the points poorly identified by the media and satisfactorily by the rank statistics. They are to the right and left of the center in Fig. 5. This occurs if two ejections created by heavy tails are found to be on one side of the true value of the parameter.

A comparison of the arithmetical-mean statistics and the rank measure for normal distribution shows a near-unity efficiency and an almost total correlation. For a uniform distribution, the rank statistics is only fractions of a per-

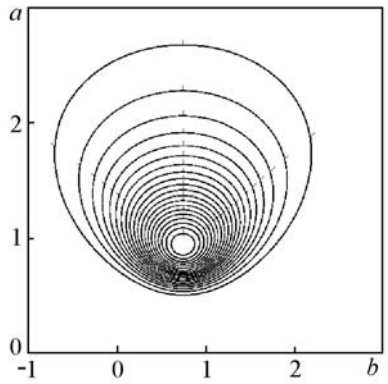


Fig. 6. Contour diagram of the uncertainty function from the rank measure for the data set {2.097, –0.023, –0.252, 1.625, 0.256} with respect to the normal distribution $N(x, b, a)$. The point estimate of the parameters is $(\dot{b}, \dot{a}) = (0.757, 1.412)$.

cent more efficient than the sampling-center statistics and 5% better than the estimate by the mean. It turns out that the proposed rank statistics has an efficiency no lower than the well-known optimum statistics and even somewhat higher than the existing efficient empirical statistics.

The confidence interval can be computed and interpreted in different manners according to the observer's needs. In the context of validity of the calibration function, the confidence interval is simply a quantum of distribution which is a section of the correction function when the value of the data's argument is equal to the value of the point estimate if, naturally, the corresponding calibration function has been constructed.

We can estimate the confidence interval by the uncertainty figure, but since it is not a correct probability density, we should carefully select the value of the confidence probability; usually the required value turns out to be very close to unity, much closer that for the distribution. Here there is a small terminological problem. We have to speak of the confidence probability as applied to an object which is not the probability density in the strict sense. The problem can be solved by using the expression "quantum parameter" as the synonym of the term "confidence probability" for the true distribution.

Legal metrology seeks to estimate the confidence interval directly from experimental data, motivated it by the fact that, in the course of the working experiment, errors can be both larger and smaller than those observed in the calibration experiment. In fact, this means the assumption that it is only the form of the error distribution that is reproduced, whereas the parameters of shift and dispersion (scale) must be estimated. The problem of estimation becomes bifactorial. This complication is often excessive, since, having a limited number of data, we have to use estimation on the worst-case principle or by the highest probability, which is due to the difficulty in differentiating the action of the case with a known dispersion and the action of change in the noise level. Here the graduation experiment on estimating the spread in data for the group of working measurements can be of substantial help.

The developed method is applicable in the case of multifactorial estimation but now as a measure estimating the similarity of the data set and the a priori form of distribution whose parameters are estimated, rather than in the form of a statistics. This subject requires separate consideration.

Generalization to Multifactorial Estimation. For multifactorial estimation of the distribution parameters, the initial a priori information is a function $\rho(x)$ interpreted as the reproducible form of an identified distribution (canonical distribution). The identified distribution is obtained by certain transformation from this function with the estimated parameters. For example, for the most widespread case of estimation of the parameters of shift b and dispersion a, we use the linear transformation $p(x, b, a) = \rho\left(\frac{x}{ka} - b\right)$, where k is the coefficient determined by the distribution type and by the method of standardization of description of the canonical distribution; for the normal distribution and reduction of the canon to the interval $[0 \pm 1]$ for a quantum parameter (confidence probability) of 0.95, we have $k \approx 0.33$. The number of the parameters can be increased by a more complex transformation.

Now each data set can be compared to the uncertainty function $\left.\begin{matrix}\{d\}\\ p(x, b, a)\end{matrix}\right\} \to \Pi\ \{p_{i/n}(d_i, b, a)\} \to u(b,\ a)$, dependent on two parameters. The point and interval estimates of each parameter can be obtained independently by any appropriate method. From our data, the most adequate statistics for the proposed rank uncertainty is the center of gravity of the uncertainty figure. The estimate is unambiguous and relatively simple to compute and fully uses the information contained in the uncertainty function. Figure 6 illustrates the aforesaid.

Finally, we dwell on the separation of the field of application of our rank measure and the MLM. For the above reasons, the MLM is inapplicable to analysis of repetitive-experiment data in a strict sense. On the other hand, when the data set consists of only one datum ($n = 1$), we obtain $p_{1/1}(x) = p(x)$, and the methods become nearly equivalent. The difference is preserved in the method of obtaining a point estimate, more precisely, in interpretation of the uncertainty figure. In our opinion, the main reason why the operator of computation of the center of gravity can be used is the preserved value of the calibration function in obtaining the uncertainty figure. In the cases where interpretation of the data densities is relatively arbitrary, interpretation of the uncertainty figure becomes arbitrary, too. There may occur a situation where estimation by the maximum of the uncertainty density turns out to be preferable, e.g., in terms of computations, especially as the coordinates of the center of gravity for symmetric distributions with one maximum coincide with the coordinates of this maximum. In the general case the choice is the observer's.

Conclusions. The proposed method of estimating the distribution parameters solves the posed problem under the assumption of observance of the condition of independence of the result from the position of the datum in the data set. If the form of distribution of the instrumental error is reproduced, the form of all distribution densities of the order statistics will be reproduced, too. Under these conditions, the individual calibration function for each datum of the ordered data set will correspond to the order distribution density. Any substantial (in a statistical sense) disagreement will indicate the violation of the assumptions made.

On the efficiency of the proposed statistics, we can make the following remarks. The reason why the statistics of a repetitive experiment can be used is the absence of the statistics making it possible to evaluate the parameter by the only one datum more accurately than it is allowed by the knowledge of the distribution of the generating process, namely, $u(x) = p(d - x)$.

To each position of the datum in the data set, there unambiguously corresponds its order probability density. Therefore, the proposed method can be interpreted as the statistics making an estimate by the data upon the transformation $\{u_i(x) = p_{i/n}(d_i - x)\}_n$. The transformation is unambiguous and the only possible under the assumptions made, which follows from the structure of the calibration function for the repetitive experiment. Consequently, the most efficient (optimum from the viewpoint of the balance of available information) statistics for combining the order uncertainties $\{u_i(x)\}_n$ is the operator of their product as the operation over independent statistical objects.

Exhaustive data on the terms used can be found in the papers of "Probability and Mathematical Statistics." Encyclopaedia/Editor-in-Chief Yu. V. Prokhorov, Moscow: Bolshaya Rossiiskaya Éntsiklopediya, 1999.

NOTATION

a, dispersion (scale) parameter; b, distribution-shift parameter; $\dot{b}$, point estimate of the shift parameter; $\dot{b}_l$, point estimate by the MLM; $\dot{b}_\mathrm{m}$, point estimate by the median; $\dot{b}_\mathrm{r}$, point estimate by the rank statistics; d, datum: unit element of experimental information, e.g., record of the measuring instrument's reading; $\{d\}_n$, data set: collection of n data processed as a unified data structure; e, standard value of the measured quantity, used for construction of a calibration function; ef, estimation of the relative efficiency of the statistics; LR, algorithm of linear regression; $P(x)$, cumulative distribution function; $p(x)$, distribution-density function; $p_{i/n}(x)$, probability density of the ith order statistics from the data set of length n; s, statistics used for estimating the parametric value; sh, estimate of the relative bias of the statistics; $\dot{x}$, point estimate of the unknown quantity; $[x]$, interval estimate of the unknown quantity; $u(x)$, uncertainty density; Ω, definition domain of the distribution function. Subscripts: l, estimation by the MLM; m, estimation by the median; r, estimation by the rank measure.