Increase of mean and variance estimates reliability for limited data size

V. Volkovas, J. Dulevičius, M. Eidukevičiūtė

Kaunas University of Technology

Kęstučio 27, Kaunas

Introduction

When identifying unknown parameters of mathematical models, experimental measurements should be made, which are often technologically sophisticated. Then the price of diagnostic technology depends on the number of measurements. Due to inadequate number of measurements and statistically unjustified data, wrong results of the simulation may be obtained. In order to diagnose the state of the object reliably, the result of the identification or measurement should be reliable enough, and this increase the number of the experiments, time necessary for measurements and so on. Because of the small number of experiments wrong result of the diagnostics procedure may be obtained also.

Measurement data or identification results may be treated as limited size sample of random quantities from which statistically reasoned value should be picked, e.g., mean. Randomness may be related to the distribution of the measured/identified parameter in the substance volume, e.g., to the density of the heterogeneous substance, experimental errors and other. Then both in physical and mathematical identification cases it is essential correctly assess limited size of the data. For that point theoretical technique [1] were suggested, practical use of which requires additional investigation.

Theoretical base of the method

Let us examine two random quantities X_t and Y_t, with the quantity X_t being normally distributed. Its distribution function $F(x) = P(x_t < x)$ has the form

$$F(x) \dot{\sim} N(m_x, \sigma_x^2), \tag{1}$$

where m_x is the mean value; σ_x^2 is variance; , $\dot{\sim}$ is symbol of a distribution, whose mean value Y_t is distributed according to the law *F(y)* other than normal.

It is known [2] that quantities of characteristics m_y and σ_y are tabulated only for some types of the distribution *F(y)*. In general, for a random value Y_t a non-normal distribution, for a given confidence level α, confidence intervals for quantities m_y and σ_y cannot be determined by analytic methods. In order to determine quantities m_y and σ_y using the measured data $y^T = (y_1, y_2, \ldots, y_s)$ and empirical distribution function $\hat{F}_s(y) \dot{\sim} N(\hat{m}_y, \hat{\sigma}_y^2)$ let us use normalizing transformations of the quantities $y^T = (y_1, y_2, \ldots, y_s)$ with subsequent computer processing of the results.

The essence of normalizing transformations is as follows [3]: $\hat{F}_s(y) \dot{\sim} N(\hat{m}_y, \hat{\sigma}_y^2)$, then the data $y^T = (y_1, y_2, ..., y_s)$ are normalized using the transformation

$$\varphi(x) = f(y)\left|\frac{dy}{dx}\right|, \frac{dx}{dy} \neq 0, \tag{2}$$

where $\varphi(x), f(y)$ are distribution density functions of the random quantities X_t and Y_t, respectively.

Expression (1) establishes the rule for transforming random quantities from one distribution pattern into another; namely if $F^{-1}(y)$ is a function reciprocal to function $F(y)$, then $z = F(y)$ is uniformly distributed, while the random quantity $x = \Phi^{-1}(z)$ is distributed in normal pattern, i.e.

$$z = F(y) \dot{\sim} R[0,1], \tag{3}$$

$$x = \Phi^{-1}(z) \dot{\sim} N[0,1]. \tag{4}$$

In addition, the following correlations are valid:

$$x = \Phi^{-1} \otimes F(y) = H^{-1}(y), \tag{5}$$

$$y = F^{-1} \otimes \Phi(x) = H(x), \tag{6}$$

where symbol $\otimes$ denotes superposition of functions.

The transformation procedure defined by the last two expressions is called the normalization process and serves as the basis for obtaining estimates $\hat{H}(\cdot)$ and $\hat{H}^{-1}(\cdot)$, i.e.

$$x = H^{-1}(y), \tag{7}$$

where by $\hat{H}^{-1}(\cdot)$ transforms a random quantity with the empirical distribution function $\hat{F}_s(y)$ into a random quantity with the distribution function $\Phi(y)$, i.e. into normal distribution pattern.

The "normalization" procedure for the empirical distribution function $\hat{F}_s(y)$ can be performed in various ways: by tabulation, by graphic constructions or by analytic means. For the purpose of computer processing, the graphic method is the most convenient one. In that case, a discrete normalizing function $\hat{H}^{-1}(y_i)(i = \overline{1,s})$ is constructed in accordance with Eq.4 for data $y^T = (y_1, y_2, ..., y_s)$ using an empirical distribution function $\hat{F}_s(y_i)\left(i = \overline{1,s}\right)$. Thus, for each point $y_i\left(i = \overline{1,s}\right)$,

corresponding quantities $x_i\left(i=\overline{1,s}\right)$ are normally distributed, i.e., $N(0,\hat{\sigma}_y^2)$. According to Eq.4, the transformation $\Phi^{-1}(\cdot)$ is performed with parameters $(0,\sigma_y^2)$, i.e. with mean equal to zero and dispersion

$$\hat{\sigma}_y^2=\frac{1}{s}\sum_{i=1}^{s}(y_i-\hat{m}_y)^2;\ \hat{m}_y=\frac{1}{s}\sum_{i=1}^{s}y_i\,. \qquad (8)$$

In order to evaluate the mean and variance more reliably for limited size s experimental data, having sequence $\{x_i\}$ $(i=\overline{1,s*})$, $s*>>s$ distributed normally, which is generated by $\hat{H}^{-1}(y)$. The corresponding sequence $\left\{y_i^*\right\}$ $i=\overline{1,s*}$ is computed according to Eq.7; in such a way we get statistical continuation for y^T [1], which can be used for more reliable evaluation of experimental data mean value $\hat{m}_y$ and variance $\hat{\sigma}_y^2$. For this in Eq.8 instead of s size sequence generated statistically based new sequence is used where $s*>>s$. The success of this methodology depends on function the $H^{-1}(y_i)$, $i=\overline{1,s}$, approximated with the determined function $H(y)$, precision and method effectiveness depends on minimal data size s_0, sufficient for statistical method validity.

The task will be examined on the base of statistical simulation by generating random non-normally distributed quantities y_i.

The results of statistical simulation

Statistical data simulation to test effectiveness of methodology was carried out. The sample of 1000 random quantities of Gamma distribution were generated: $y^T=(y_1, y_2, ..., y_{1000})$, where

$$F(y)\sim G(\lambda,\eta). \qquad (9)$$

Here: λ - distribution shape parameter, η - distribution scale parameter.

For this simulation $\lambda=1$, $\eta=1$ were chosen.

The samples of $Y^{(1)}$, with size $s_1=100$; $Y^{(2)}$, with size $s_2=50$; $Y^{(3)}$, with size $s_3=10$; $Y^{(4)}$, with size $s_4=5$ were randomly picked from the original sample Y. Each sample was analyzed separately to examine effectiveness of methodology while the sample size is decreasing. Empirical estimation of the parameters of the samples Y, $Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$, $Y^{(4)}$ are presented in the Table 1.

Table 1. **Empirical means and variations of the samples Y, $Y^{(1)}$, $Y^{(2)}$, $Y^{(3)}$, $Y^{(4)}$**

s_j	$\hat{m}_y$	$\hat{\sigma}_y^2$
$s=1000$	1,017419	1.047865
$s_1=100$	1,023478	1.130951
$s_2=50$	0,884406	0.649623
$s_3=10$	0,4963623	0.046860
$s_4=5$	4,653040	0.511671

Samples of the normal random quantities $X^{(j)}$ with empirical mean $\hat{m}_y^{(j)}$ (the mean of the corresponding $Y^{(j)}$ sample) and standard deviation $\hat{\sigma}_y^{(j)}$ (the standard deviation of the corresponding $Y^{(j)}$ sample) are generated $\left(j=\overline{1,4}\right)$. The size of the $X^{(j)}$ size corresponds to the size of the parallel sample $Y^{(j)}$

$$X^{(j)}\sim N(\hat{m}_y^{(j)},\hat{\sigma}_y^{(j)})\,. \qquad (10)$$

Normalizing function $H(x)$ is expressed by n order polynomial:

$$H(x)=a_0+a_1x+a_2x^2+\ldots+a_nx^n\,, \qquad (11)$$

where a_0, a_1, .. a_n – polynomial coefficients, x – random quantity, distributed normally.

Sample $Z^{(j)}$ is generated then. It's size is $k_j=s_j$:

$$Z=H(X). \qquad (12)$$

It was noticed that when approximating the inverse normalizing function $H^{(j)}(x)$ $\left(j=\overline{1,4}\right)$ with higher order of polynomial empirical variance $\hat{\sigma}_z^2$ and mean $\hat{m}_z$ of the sample $Z^{(j)}$ enlarges. So the order of the polynomial is determined in iterative way by comparing the mean $\hat{m}_y$ and variance $\hat{\sigma}_z^2$ of the sample Z with the mean $\hat{m}_z$ and variance $\hat{\sigma}_y^2$ of the sample Y and the order in which minimum of the parameters difference is reached is considered to be optimal. i.e.

$$n:\ \min_{n=s_j-K_j,1}\left|(\hat{m}_x^{(j)}-\hat{m}_z^{(j)})+(\hat{\sigma}_x^{(j)}-\hat{\sigma}_z^{(j)})\right|. \qquad (13)$$

Here n – polynomial order, s_j – size of samples $X^{(j)}$ and $Z^{(i)}$, K_j – constant, which depends on the s_j. If $s_j<7$ then $K_j=1$.

Sample X, consisting of $s=1000$ random normal quantities with empirical mean $\hat{m}_y$ and variance $\hat{\sigma}_y^2$ is generated:

$$X\sim N(\hat{m}_y,\hat{\sigma}_y^2). \qquad (14)$$

This sample X is transformed into samples $Y^{*(1)}$, $Y^{*(2)}$, $Y^{*(3)}$, $Y^{*(4)}$ by using four normalizing functions $H^{(j)}(x)$, $j=1,2,3,4$ i. e.:

$$Y^{*(1)}=H^{(1)}(X), \qquad (15)$$

$$Y^{*(2)}=H^{(2)}(X), \qquad (16)$$

$$Y^{*(3)}=H^{(3)}(X), \qquad (17)$$

$$Y^{*(4)}=H^{(4)}(X). \qquad (18)$$

Negative values, obtained during transformation are omitted as they cannot have any influence to parameters.

Relative errors of empirical means $\delta_m^{(j)}$ and standard deviations $\delta_\sigma^{(j)}$ of the samples $Y^{*(1)}$, $Y^{*(2)}$, $Y^{*(3)}$, $Y^{*(4)}$ are presented in the Table 2 calculated according to formula:

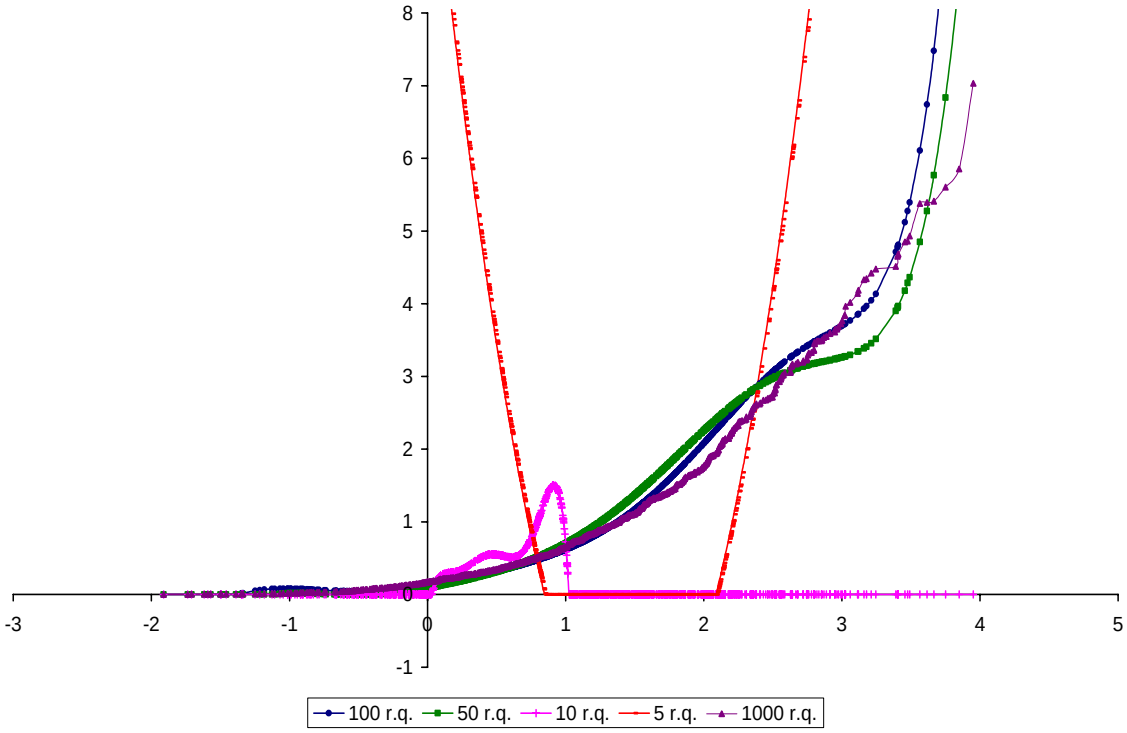

Fig. 1. **Graphs of random quantity sequence Y and normalized sequences Y^{*}, Y^{**}, Y^{***}, Y^{****}**

$$\delta_m^{(j)} = \frac{\hat{m}_y - \hat{m}_y^{*(j)}}{\hat{m}_y}, \qquad (19)$$

$$\delta_\sigma^{(j)} = \frac{\hat{\sigma}_y - \hat{\sigma}_y^{*(j)}}{\hat{\sigma}_y}, \qquad (20)$$

where $\hat{m}_y$ is the empirical mean of the original sample Y; $\hat{m}_y^{*(j)}$ is the empirical mean of the extended sample $Y^{*(j)}$; $\hat{\sigma}_y$ is the standard deviation of the original sample Y; $\hat{\sigma}_y^{*(j)}$ is the empirical standard deviation of the extended sample $Y^{*(j)}$.

Table 2. **Relative errors of the means and standard deviations**

$Y^{(j)}$	$\delta_m^{(j)}$	$\delta_\sigma^{(j)}$
$Y^{(1)}$	-0.06542	-0.21389
$Y^{(2)}$	-0.14203	-0.08255
$Y^{(3)}$	-4.89901	-7.44377
$Y^{(4)}$	-8.86441	-9.9541

The smallest relative error of the mean is obtained with $H^{(1)}$, i.e., when the sample of 100 random quantities were approximated. The best estimation of variance was obtained with $H^{(2)}$, i.e. when the sample of 50 random quantities was approximated.

The experiment was repeated 12 times enlarging Gamma distribution shape parameter λ with step 0,25 (from 1 to 4), then the variance of the distribution is increasing. In such a way dependence of effectiveness of methodology on the sample variance can be examined.

When sample size is decreasing, the sensitiveness of normalizing function polynomial order increases and when wrong order is chosen, the relative error enlarges in tens of times. In Fig. 2 relative errors of means are presented separating different normalizing function $H^{(j)}$.

As we see in Fig. 2 and 3 relative errors $\delta_m^{(j)}$ and $\delta_\sigma^{(j)}$ do not depend on the growth of distribution variance, but with a very small sample size the optimal polynomial order may not be obtained. The estimations $\hat{m}_y^{*(j)}$ of the mean $\hat{m}_y^{(j)}$ are more precise than the estimations $\hat{\sigma}_y^{*(j)}$ of standard deviation $\hat{\sigma}_y^{(j)}$. While comparing Figures 2 and 3, one can see, that the relative errors $\delta_m^{(j)}$ and $\delta_\sigma^{(j)}$ both get high values when j = 10, λ = 2 and j =5, λ = 1, 2. Then we can do an assumption that when such a conjunction occurs, new sample of random variables could be generated to eliminate such deviation.

As we see from the Tables 3 and 4 the possibility of large deviance of mention quantities when optimizing polynomial order decreases, though when approximating the sample of 5 random quantities order of optimization has a lesser impact.

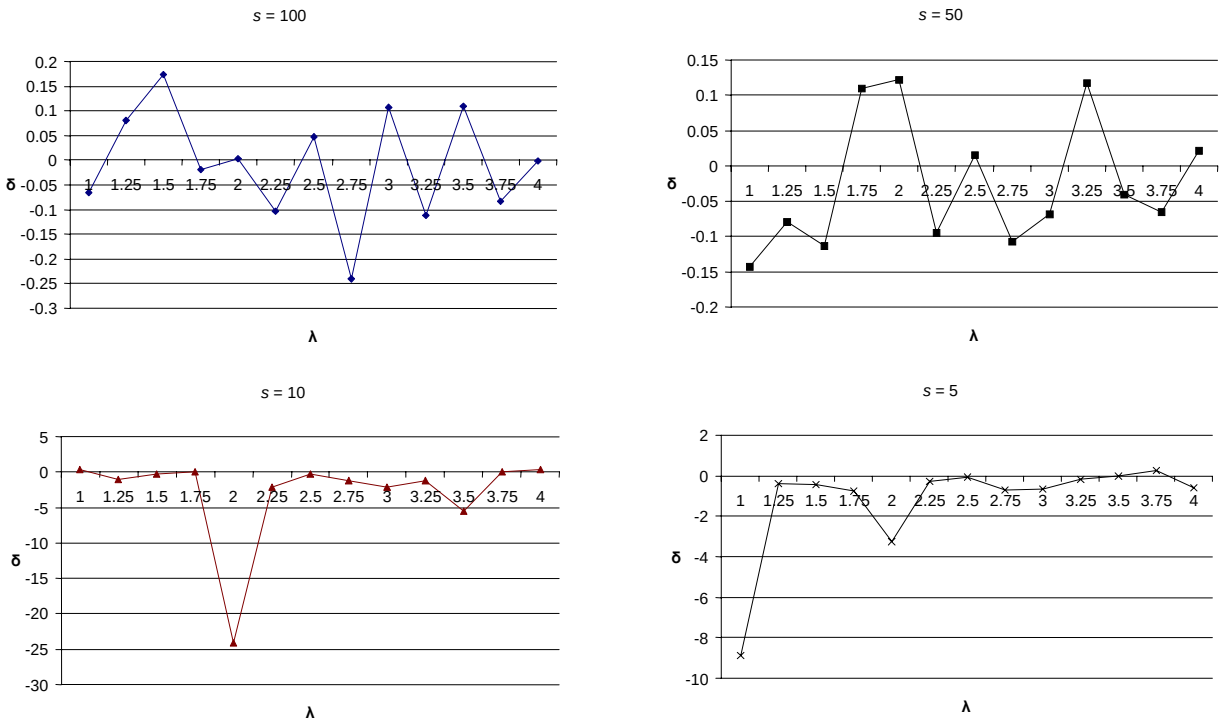

Fig. 2 **Graphs of empirical mean relative errors for each $H^{(j)}$ (here λ – shape parameter which increase enlarges variance σ^2)**

Table 3. **Relative error δ_m of empirical mean for different normalizing functions $H^{(j)}$ (here λ – shape parameter indicating growth of variance σ^2) showing dependence of error on the optimizing polynomial order of the normalizing function**

λ	H1		H2		H3		H4	
	Optimized	Non-Optimized	Optimized	Non-Optimized	Optimized	Non-Optimized	Optimized	Non-Optimized
1	-0.06542	-0.66511	-0.14203	-0.7013	0.320982	0.320982	-8.86441	-8.86441
1.25	0.081355	-0.14621	-0.07856	-0.03548	-0.96506	-2.41839	-0.3802	0.198634
1.5	0.174123	0.344674	-0.11312	0.03057	-0.30985	-0.30985	-0.44165	-0.07408
1.75	-0.01936	-0.03432	0.110459	-0.55837	0.005114	-8.61969	-0.76036	-0.5941
2	0.002534	0.002534	0.122371	0.09957	-24.1761	-24.1761	-3.27175	-3.27175
2.25	-0.10215	0.290004	-0.09484	-0.18509	-2.06032	-12073.6	-0.29675	-0.29675
2.5	0.046583	-0.18637	0.015984	0.015984	-0.23924	-66.4247	-0.05869	-3.21513
2.75	-0.24047	-0.24047	-0.1075	-0.51344	-1.27055	-1.27055	-0.68336	-0.84839
3	0.106482	-13.2287	-0.06906	-0.05251	-2.10603	-7.93042	-0.64599	-1.06406
3.25	-0.11156	-0.11156	0.116959	0.116959	-1.13861	-1.13861	-0.19184	-0.19184
3.5	0.108819	0.642857	-0.04012	-0.04012	-5.47519	-5.47519	-0.01838	-0.01838
3.75	-0.08338	-60.4805	-0.06474	-0.26427	-0.01763	-16677.2	0.240447	0.240447
4	-0.00211	-1.1065	0.021657	0.021657	0.41233	0.41233	-0.60802	-0.60802
$\hat{m}_{\delta m}$	**-0.00804**	**-5.76305**	**-0.02481**	**-0.15891**	**-2.8477**	**-2220.6**	**--1.2293**	**-1.43137**

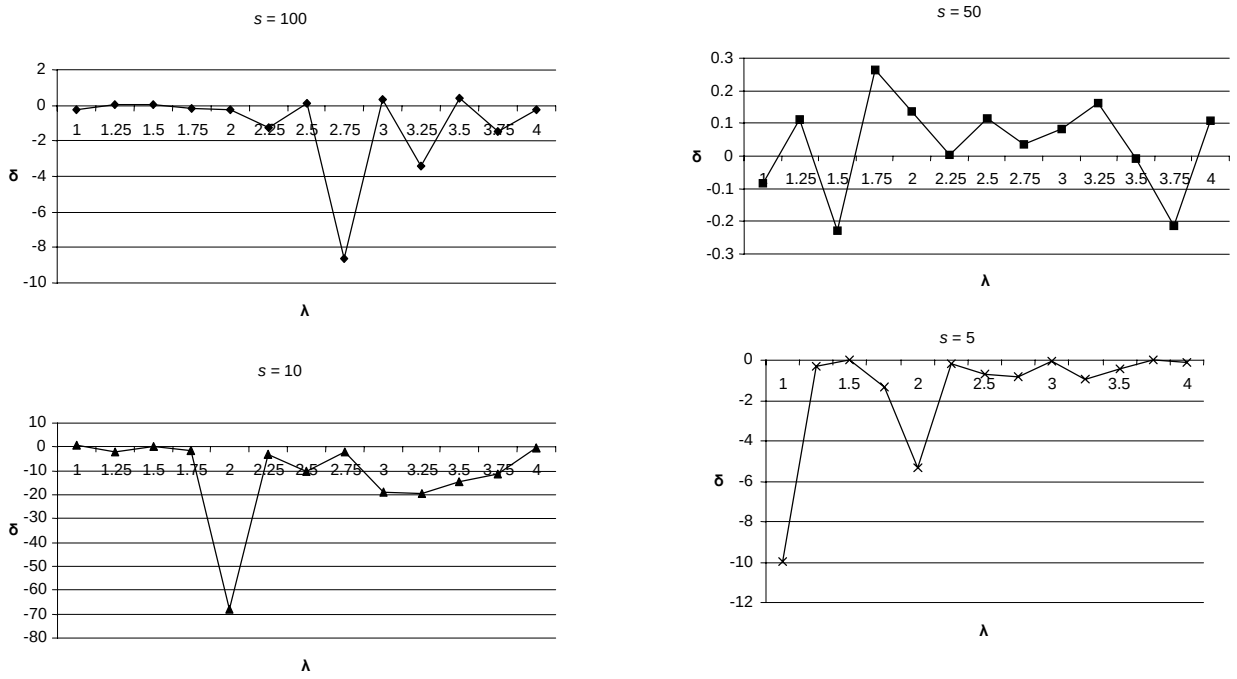

Fig. 3 **Graphs of empirical standard deviation relative errors for each $H^{(j)}$ (here λ – shape parameter which increase enlarges variance σ^2)**

Table 4. **Relative error δ_m of empirical standard deviation for different normalizing functions $H^{(j)}$ (here λ – shape parameter indicating growth of variance σ^2) showing dependence of error on the optimizing polynomial order of the normalizing function**

λ	H1		H2		H3		H4	
	Optimized	Non-Optimized	Optimized	Non-Optimized	Optimized	Non-Optimized	Optimized	Non-Optimized
1	-0.21389	-5.20939	-0.08255	-6.82872	0.623245	0.623245	-9.9541	-9.9541
1.25	0.062972	-1.11812	0.111475	0.182038	-2.16803	-4.40735	-0.33466	0.381536
1.5	0.062251	0.254907	-0.22939	-0.47482	0.058803	0.058803	-0.02429	-0.12895
1.75	-0.1328	-0.02971	0.26278	-0.84665	-1.60884	-107.096	-1.33862	-0.29481
2	-0.24557	-0.24557	0.138856	-0.04647	-67.7444	-67.7444	-5.31439	-5.31439
2.25	-1.25475	0.52876	0.004473	-0.89824	-3.07416	-99812	-0.18666	-0.18666
2.5	0.127576	-14.1555	0.115959	0.115959	-10.2201	-1009.85	-0.70077	-1.5921
2.75	-8.62536	-8.62536	0.034977	-9.00251	-2.15941	-2.15941	-0.85636	-0.38831
3	0.362547	-109.296	0.084063	0.138201	-18.9681	-82.1352	-0.05598	-3.11678
3.25	-3.41526	-9.0665	0.162755	0.162755	-19.5473	-19.5473	-0.97896	-0.97896
3.5	0.426292	0.821976	-0.00693	-0.00693	-14.4665	-52.2098	-0.4316	-0.4316
3.75	-1.47466	-1083.33	-0.21349	-0.79527	-11.407	-80124.6	-0.01459	-0.01459
4	-0.21512	-21.7746	0.107433	0.107433	-0.37327	-0.37327	-0.13814	-0.13814
$\hat{m}_{\delta\sigma}$	**-1.11814**	**-96.2493**	**0.037724**	**-1.39948**	**-11.6196**	**-13944.7**	**-1.56378**	**-1.70445**

Example of application

The obtained result will be practically used to process experimental measurements of heterogeneous pollution layer thickness in technological pipe [4]. Measurement method based on the Lamb waves interference is the only one at the meantime, which allows to indicate state of the internal surface of the cylindrical systems. Experimental data of heterogeneous layer measurement are provided in the Table 5.

Table 5. **Experimental data of the heterogeneous pollution layer thickness in technological pipe**

$Y^{(1)}$	$Y^{(2)}$	$Y^{(3)}$
1.4	2	1.9
1.9	1.8	2.3
2	1.7	2.1
1.8	1.8	1.9
1.6	2	2.1
2	1.6	2
1.4	1.5	1.8
1.8	1.9	1.8
1.8	1.9	1.7
1.9	1.9	2
1.3	1.8	2
1.8	1.8	1.9

First of all empirical means $\hat{m}_y^{(j)}$ and standard deviations $\hat{\sigma}_y^{(j)}$ for the samples were calculated: $\hat{m}_y^{(1)}=1{,}725$, $\hat{\sigma}_y^{(1)}=0{,}231391$, $\hat{m}_y^{(2)}=1{,}808333$, $\hat{\sigma}_y^{(2)}=0{,}144097$, $\hat{m}_y^{(3)}=1{,}958333$, $\hat{\sigma}_y^{(3)}=0{,}155233$. With these parameters sequences X^i, i=1,2,3 of normal random variables were generated and polynomial orders were optimized. The data are presented in Tables 6 and 7.

Table 6. **Generated normal random quantities with the mean $\hat{m}_y^{(j)}$ and standard variance $\hat{\sigma}_y^{(j)}$**

$X^{(1)}$	$X^{(2)}$	$X^{(2)}$
1.10263	1.520382	1.701714
1.243807	1.532066	1.804623
1.451602	1.584084	1.848453
1.600224	1.685052	1.96436
1.623691	1.717358	1.97961
1.631994	1.736624	1.989646
1.707283	1.755504	2.042627
1.743742	1.762074	2.051461
1.767008	1.806566	2.075077
1.816357	1.864855	2.08291
1.829156	1.908473	2.121007
2.028317	1.931201	2.197536

Table 7. **Estimated coefficients of the optimal order polynomials**

	$H^{(1)}$	$H^{(2)}$	$H^{(3)}$
a_0	-12.0077	-76.3211	-32.7262
a_1	40.3893	132.7531	55.348
a_2	-45.479	-75.4585	-29.9082
a_3	22.345	14.3547	5.4494
a_4	-3.9748		

After generation of normal samples of 1000 random quantities empirical means and $\hat{m}_y^{*(j)}$ standard deviations $\hat{\sigma}_y^{*(j)}$ were revised (Table 8).

Table 8. **Revised empirical means $\hat{m}_y^{*(j)}$ and standard deviations $\hat{\sigma}_y^{*(j)}$**

	j = 1	j = 2	j = 3
$\hat{m}^{(j)}$	1.725	1.808333	1.958333
$\hat{m}^{*(j)}$	1.762064	1.92523	1.935729
$\hat{\sigma}^{(j)}$	0.231391	0.144097	0.155233
$\hat{\sigma}^{*(j)}$	0.219549	0.228393	0.205222

Conclusions

While analyzing the technique, a new aspect was examined. It was noticed that effectiveness of the methodology depends on the order of the polynomial, which approximates the normalizing function. However, due to random approximation of each sample, there is possibility that optimal order of the polynomial will not be obtained. While comparing approximation of large and small samples it was noticed that the mentioned possibility for very small samples is larger. But the sample of more than 10 random variables is enough optimal to use the methodology and revised estimate of mean and variance can be obtained. Effectiveness of methodology does not depend on the sample variance growth; the standard deviation is more sensitive for the polynomial order than the sample mean.

References

1. **Baltrūnas I. I., Volkov V. V.** Reliability of Experimental Results For a Limited Number of Measurements. Vibration engeneering. 1990. Vol. 4. P. 215-224.
2. **Gnedenko B. V.** Kurs teorii verojatnostei. (in Russian)- M.:Nauka. 1969. 399 s.
3. **Baltrūnas J. J., Nakutis E. J.** Identifikatsija nelineinich statsionarnich stochasticheskich protsesov v klase modeley L ⊕ H, H ⊕ L (1. metodika identifikatsii). (in Russian). Тр. АН Лит ССР. Сер. Б. 1984. Т. 5(144). S. 110-120
4. **Jonušas R., Jurkauskas A., Volkovas V.** Rotoriniu sistemu dinamika ir diagnostika. Kaunas:Technologija. 2001. P. 213-228.

V. Volkovas, J. Dulevičius, M. Eidukevičiūtė

Ribotos apimties duomenų vidurkio ir dispersijos įverčio patikimumo didinimas

Reziumė

Diagnostikos technologijos kaina priklauso nuo eksperimentinių matavimų skaičiaus. Tačiau dėl nepakankamo matavimų skaičiaus bei statistiškai nepagrįstų duomenų galima gauti klaidingus modeliavimo rezultatus, technologija gali būti neefektyvi. Fizikinio ir matematinio identifikavimo atvejais būtina statistiškai teisingai įvertinti ribotą duomenų apimtį. Tam tikslui buvo pasiūlytas teorinis metodas. Jo praktiniam naudojimui įvertinti atlikti papildomi tyrimai.

Metodas pagrįstas atvirkštinės normalizacijos funkcijos *H(x)* įvertinimu. Jos dėka galima išplėsti pradinę statistinę imtį ir patikslinti nagrinėjamus parametrus: vidurkį ir dispersiją. Šio metodo panaudojimo sėkmė priklauso nuo funkcijos $H^{-1}(y_i)$ aproksimacijos determinuota funkcija *H(y)* tikslumo, o metodo efektyvumas - nuo minimalios duomenų apimties s_0, pakankamos metodui statistiškai pagrįsti.

Užduotis nagrinėjama statistinio modeliavimo pagrindu, generuojant atsitiktinius dydžius y_i, pasiskirsčiusius ne pagal normalinį dėsnį.

Pastebėta, kad metodikos efektyvumas priklauso nuo polinomo, optimizuojančio normalizuojančiąją funkciją, eilės. Tačiau dėl atsitiktinės imčių aproksimacijos yra tikimybė, kad optimalus polinomo laipsnis nebus gautas. Lyginant didelių ir mažų imčių aproksimaciją, nustatyta, kad ši tikimybė mažoms imtims yra didesnė. Bet didesnė nei 10 a.d. imtis yra pakankamai optimali metodui naudoti ir patikslintiems vidurkio ir dispersijos įverčiams gauti. Metodikos efektyvumas nepriklauso nuo imties dispersijos didėjimo, tačiau standartinis nuokrypis daug jautresnis polinomo eilės pokyčiui nei imties vidurkis.

Pateikta spaudai 2002 02 22