DOI: 10.26565/2304-6201-2020-45-08

UDC 519.252+519.226

Empirical probability distribution validity based on accumulating statistics of observations by controlling the moving average and root-mean-square deviation

V.V. Romanuke
O.S. Popov Odessa National Academy of Telecommunications, Ukraine
e-mail: romanukevadimv@gmail.com

Knowing probability distributions for calculating expected values is always required in the engineering practice and other fields. Commonly, probability distributions are not always available. Moreover, the distribution type may not be reliably determined. In this case, an empirical distribution should be directly built based just on observations. So, the goal is to develop a methodology of accumulating and processing observation data so that the respective empirical distribution would be close enough to the unknown real distribution. For this, criteria regarding sufficiency of observations and the distribution validity are to be substantiated. As a result, a methodology is presented that considers the empirical probability distribution validity with respect to the parameter's expected value. Values of the parameter are registered during a period of observations or measurements of the parameter. On this basis, empirical probabilities are calculated, where every next period the previous registration data are used as well. Every period gives an approximation to the parameter's expected value using those empirical probabilities. The methodology using the moving averages and root-mean-square deviations asserts that the respective empirical distribution is valid (i. e., it is sufficiently close to the unknown real distribution) if the parameter's expected value approximations become scattered very little for at least the three window multiple-of-2 widths by three successive windows. This criterion also implies the sufficiency of observation periods, although the sufficiency of observations per period is not claimed. The validity strongly depends on the volume of observations per period.

Key words: *empirical probability distribution; accumulation of statistics; a moving average; root-mean-square deviation; observations; measurements; an expected value approximation.*

Знання розподілів імовірностей для обчислення очікуваних значень завжди потрібно в інженерній практиці та інших сферах. Зазвичай імовірнісні розподіли не завжди доступні. Більше того, тип розподілу може бути невірно визначений. У цьому випадку емпіричний розподіл має бути побудований безпосередньо на основі спостережень. Тому мета полягає у розробці такої методології накопичення та обробки даних спостережень, щоб відповідний емпіричний розподіл був досить близьким до невідомого реального розподілу. Для цього слід обґрунтувати критерії щодо достатності спостережень та обґрунтованості розподілу. В результаті представляється методологія, яка розглядає обґрунтованість емпіричного розподілу ймовірностей відносно очікуваного значення параметра. Значення параметра реєструються протягом періоду спостережень або вимірювань цього параметра. На цій основі обчислюються емпіричні ймовірності, де кожного наступного періоду попередні дані реєстрації також використовуються. Використовуючи ці емпіричні ймовірності, кожен період дає апроксимацію очікуваного значення параметра. Використовуючи ковзні середні та середньоквадратичні відхилення, методологія стверджує, що відповідний емпіричний розподіл є дійсним (тобто він достатньо близький до невідомого реального розподілу), якщо апроксимації очікуваного значення параметра стають дуже мало розсіяними принаймні для трьох вікон, довжини яких кратні 2, упродовж трьох послідовних вікон. Цей критерій також передбачає достатність періодів спостережень, хоча про достатність спостережень за період не стверджується. Обґрунтованість сильно залежить від обсягу спостережень за період.

Ключові слова: *емпіричний розподіл імовірностей; накопичення статистики; ковзне середнє; середньоквадратичне відхилення; спостереження; вимірювання; апроксимація очікуваного значення.*

Знание распределений вероятностей для вычисления ожидаемых значений всегда нужно в инженерной практике и других сферах. Обычно вероятностные распределения не всегда доступны. Более того, тип распределения может быть неверно определён. В этом случае эмпирическое распределение должно быть построено непосредственно на основе наблюдений. Поэтому цель заключается в разработке такой методологии накопления и обработки данных наблюдений, чтобы соответствующее эмпирическое распределение было достаточно близким к неизвестному реальному распределению. Для этого следует обосновать критерии достаточности наблюдений и обоснованности распределения. В результате представляется методология, которая рассматривает обоснованность эмпирического распределения вероятностей относительно ожидаемого значения параметра. Значения параметра регистрируются в течение периода наблюдений или измерений этого параметра. На этой основе вычисляются эмпирические вероятности, где в каждом следующем периоде предварительные данные регистрации также используются. Используя эти эмпирические вероятности, каждый период дает аппроксимацию ожидаемого значения параметра. Используя скользящие средние и среднеквадратические отклонения, методология утверждает, что соответствующее эмпирическое распределение является действительным (то есть оно достаточно близко к неизвестному реальному распределению), если аппроксимации ожидаемого значения параметра становятся очень мало рассеянными по крайней мере для трёх окон, длины которых кратны 2, в течение трёх последовательных окон. Этот критерий также предусматривает достаточность периодов наблюдений, хотя о достаточности наблюдений за период не утверждается. Обоснованность сильно зависит от объёма наблюдений за период.

Ключевые слова: *эмпирическое распределение вероятностей; накопление статистики; скользящее среднее; среднеквадратическое отклонение; наблюдения; измерения; аппроксимация ожидаемого значения.*

The problem of probability distribution estimations

The engineering practice and many other fields dealing with uncertainties always require knowing probability distributions for calculating expected values. Commonly, probability distributions are not always available, unless a presumption about a distribution is given a priori. Even the type of the probability distribution may be unknown [1]. In particular, however, if the probability distribution type is presumed to be given, the distribution parameters still need to be estimated [2]. This leads to a series of additional tasks like substantiation of the estimation procedure for each parameter, suggestion of a criterion of reliability or validity, control of stability, preservation of unbiasedness, etc. On the other hand, when the distribution type is unknown or its determination requires resources which exceed available or reasonable expenses, the problem of probability distribution estimation might be solved without learning the distribution type [3]. In this case, an empirical distribution is built directly from the observations [1, 2, 4]. Nevertheless, this approach can lead to multiple probability distribution estimations inasmuch as criteria of validity and sufficiency of observations have not been suggested and studied yet [5].

Goal

Due to the lack of a theoretical approach to building valid and stable empirical distributions directly, the goal is to develop a methodology of accumulating and processing observation data so that the respective empirical distribution would be close enough to the unknown real distribution. Therefore, a criterion of sufficiency of observations should be formulated. Besides, a supporting criterion of the distribution validity is to be substantiated. The methodology will be thoroughly discussed and practical aspects of its implementation will be underlined.

Empirical probabilities obtained from observations

Let x be a value of a parameter whose probability distribution is to be estimated. The step along the abscissa axis of the distribution is defined by the accuracy of measuring or observing this parameter. Indeed, if $x_{\min}$ is the minimal value of the parameter, whose accuracy is α, then only probabilities of values

$$x_{\min},\ x_{\min}+\alpha,\ x_{\min}+2\alpha,\ ...,\ x_{\min}+h_{\max}\alpha \tag{1}$$

are of interest, where $h_{\max}$ is an integer and $x_{\max}=x_{\min}+h_{\max}\alpha$ is the maximal value of the parameter.

Suppose that value $x=x_{\min}+h\alpha$ is registered $u_h^{(1)}$ times during a period of observations or measurements of the parameter. Obviously, $u_h^{(1)}\in\mathbb{N}\bigcup\{0\}$ by $h=\overline{0,h_{\max}}$. Then the very first rough distribution estimation is a set of relative frequencies

$$P_h^{(1)}=u_h^{(1)}\Big/\sum_{i=0}^{h_{\max}}u_i^{(1)}\quad\text{by}\quad h=\overline{0,h_{\max}}. \tag{2}$$

Relative frequencies (2) are empirical probabilities after the first (initial) observation period.

Accumulation of statistics

The next period the parameter is continued to be observed (measured), and value $x=x_{\min}+h\alpha$ is registered $u_h^{(2)}$ times, $u_h^{(2)}\in\mathbb{N}\bigcup\{0\}$. The second estimation of the distribution can use now both counts $\left\{u_h^{(2)}\right\}_{h=0}^{h_{\max}}$ and $\left\{u_h^{(1)}\right\}_{h=0}^{h_{\max}}$. Therefore, a set of empirical probabilities

$$P_h^{(2)}=\left(u_h^{(1)}+u_h^{(2)}\right)\Big/\sum_{i=0}^{h_{\max}}\left(u_i^{(1)}+u_i^{(2)}\right)\quad\text{by}\quad h=\overline{0,h_{\max}} \tag{3}$$

becomes the second distribution estimation.

This process can be continued until a stop criterion fires or by other critical circumstances (events). In general, when value $x=x_{\min}+h\alpha$ is registered $u_h^{(m)}$ times during the m-th period of observations, a set of empirical probabilities

$$P_h^{(m)}=\sum_{j=1}^{m}u_h^{(j)}\Big/\sum_{i=0}^{h_{\max}}\sum_{j=1}^{m}u_i^{(j)}\quad\text{by}\quad h=\overline{0,h_{\max}} \tag{4}$$

becomes the m-th distribution estimation, $m=1,2,3,...$ (there is no constraint to the number of observation periods).

A sequence of the parameter's expected value approximations

It is uncertain whether empirical probabilities (4) obtained from a series of periods of observations are close enough to the unknown real distribution. This is so because they cannot be compared, even roughly, to an available pattern. Nevertheless, such a pattern may exist for the parameter. Therefore, it is reasonable to consider an approximation to the parameter's expected value using empirical probabilities (4):

$$\tilde{x}_m = \sum_{h=0}^{h_{\max}} \left(x_{\min} + h\alpha\right) P_h^{(m)} = \sum_{h=0}^{h_{\max}} \left(\left(x_{\min} + h\alpha\right) \sum_{j=1}^{m} u_h^{(j)} \right) \Bigg/ \sum_{i=0}^{h_{\max}} \sum_{j=1}^{m} u_i^{(j)} \quad \text{by} \quad h = \overline{0, h_{\max}} \quad \text{and} \quad m = 1, 2, 3, \ldots \quad (5)$$

Eventually, the expected value approximations $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, ... can be used to develop and substantiate the criteria of sufficiency of observations and the distribution validity.

The moving average and root-mean-square deviation

Obviously, using the law of large numbers, the sequence of the parameter's expected value approximations $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, ... is expected to converge to the unknown expected value of this parameter (if the observations or measurements are performed methodologically and instrumentally unbiased). Although some information about the unknown expected value may be available, it is still impossible to confidentially claim how close the approximations are to the unknown value. However, it is possible to study how badly approximations $\tilde{x}_1$, $\tilde{x}_2$, $\tilde{x}_3$, ... are scattered, and whether the scattering decreases as m increases.

Let Δt be a window between measurement periods $(l-1)\Delta t + 1$ and $l\Delta t$, where $\Delta t \in \mathbb{N} \setminus \{1\}$ (the case with $\Delta t = 1$ does not make sense as then the window is "singular"), $l = 1, 2, 3, \ldots$ (there is no constraint to the number of such windows). Then the average of the parameter's expected value approximations across this window is

$$\tilde{x}\left[(l-1)\Delta t + 1, l\Delta t\right] = \frac{1}{\Delta t} \sum_{m=(l-1)\Delta t+1}^{l\Delta t} \tilde{x}_m \quad (l = 1, 2, 3, \ldots). \quad (6)$$

The root-mean-square deviation of this average (across window Δt) is the square root of its variance:

$$\sigma_{\tilde{x}}\left[(l-1)\Delta t + 1, l\Delta t\right] = \sqrt{\frac{1}{\Delta t} \sum_{m=(l-1)\Delta t+1}^{l\Delta t} \left(\tilde{x}_m - \tilde{x}\left[(l-1)\Delta t + 1, l\Delta t\right]\right)^2} \quad (l = 1, 2, 3, \ldots). \quad (7)$$

The moving average (6) and the respective root-mean-square deviation (7) depend on the window. Fig. 1 shows an example of calculating values (6) and (7) by $\Delta t = 400$ along 2000 observation periods. The averages in windows 2 — 5 are rather close. Fig. 2 shows how the moving averages and deviations change when the window is twice as narrow as the previous one.

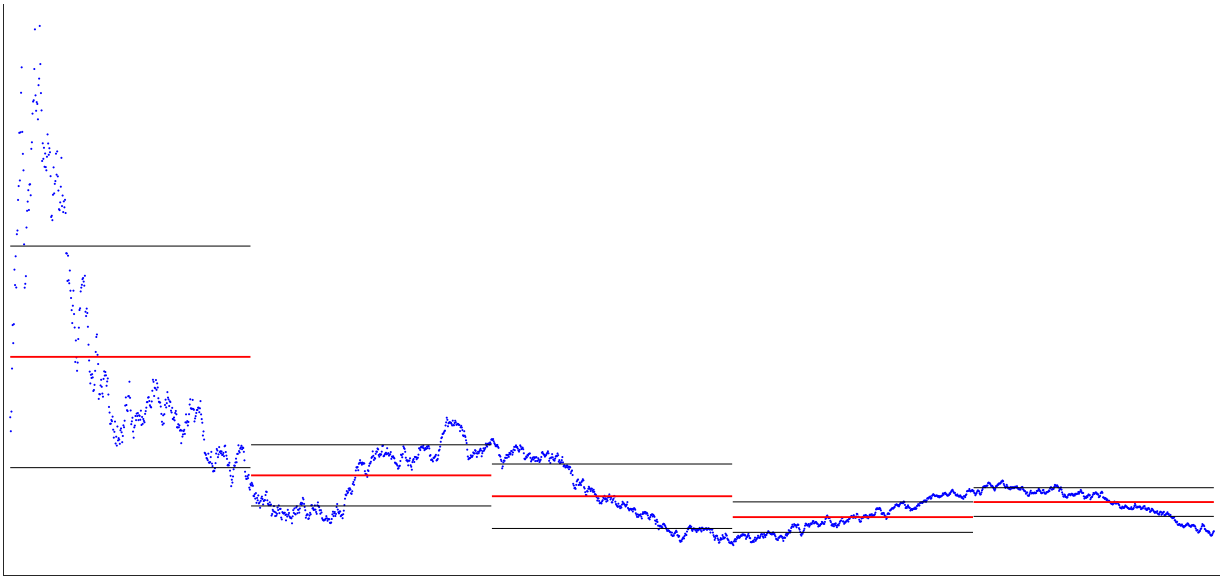

Fig. 1. An example of the moving averages and the respective deviations bounding the averages by $\Delta t = 400$

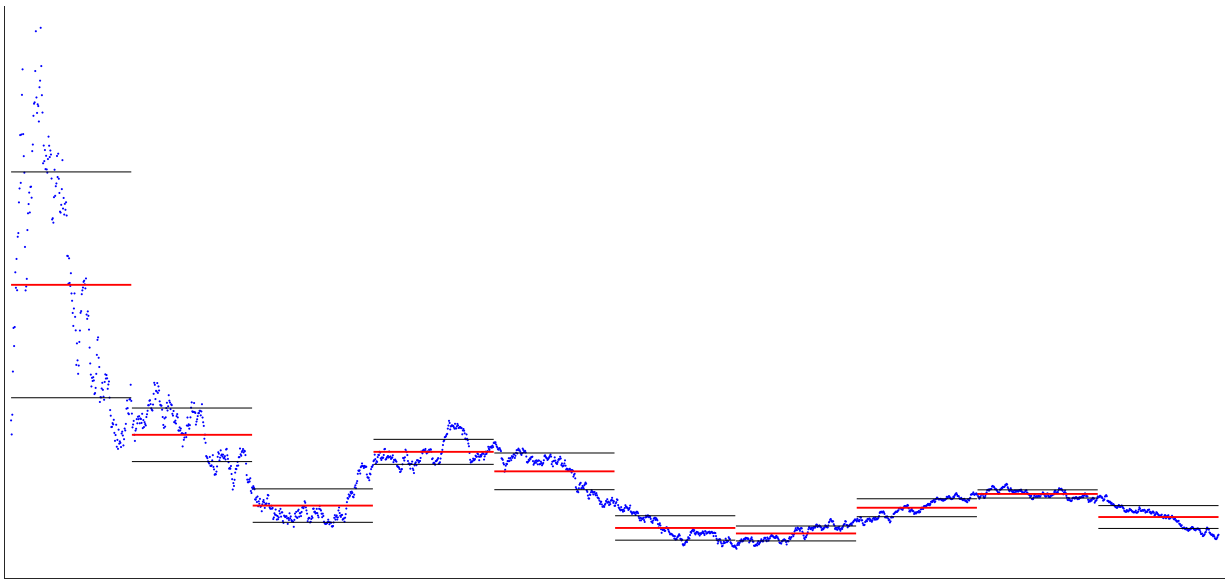

Fig. 2. The example (see Fig. 1) of the moving averages and the bounding deviations by $\Delta t = 200$

It is well seen from the shown example how the moving averages may disperse as the window is made narrower. Indeed, the example supplemented with Fig. 3 by $\Delta t = 50$ (an extremely short length for this example) shows that the narrow moving averages roughly reproduce the wave with two maxima and the minimum between them.

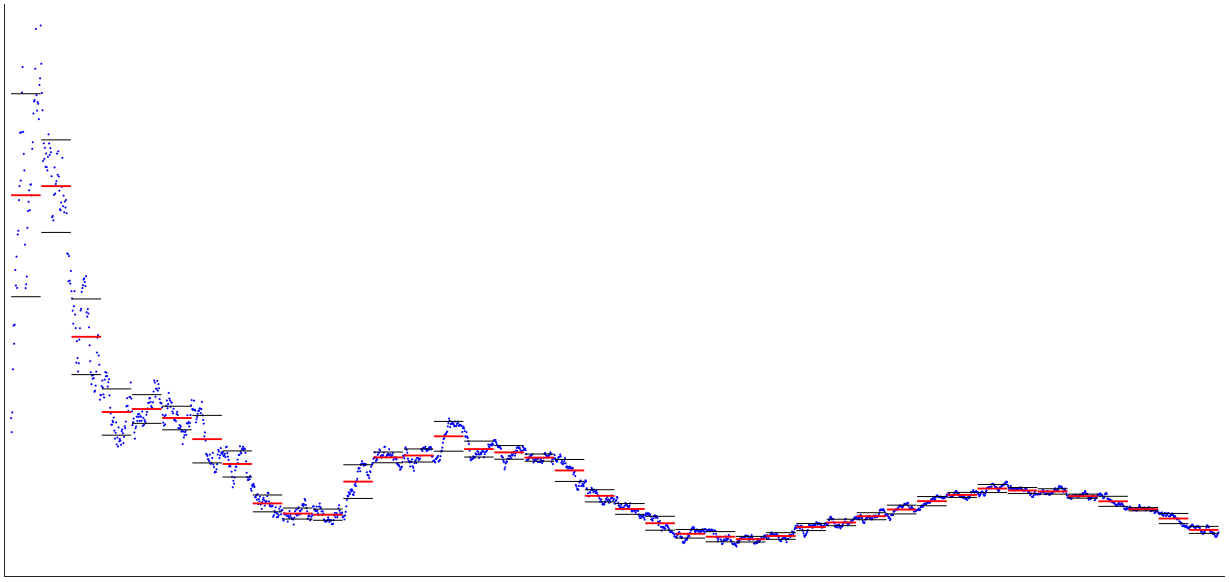

Fig. 3. The example (see Fig. 1) of the moving averages and the bounding deviations by $\Delta t = 50$

So, how to select an appropriate window width in order to conclude on the sufficiency of observations? Obviously, a single window width can hardly be selected and, therefore, a few window widths are to be studied until the most appropriate window width is empirically found.

Validity of the empirical probability distribution

At first glance, the most preferable relationship here is to have a descending sequence of deviations:

$$\sigma_{\tilde{x}}\left[(l-1)\Delta t+1, l\Delta t\right] > \sigma_{\tilde{x}}\left[l\Delta t+1, (l+1)\Delta t\right] \qquad (8)$$

(not for every l but starting with some l). Another potential condition is to require that the succeeding neighboring averages be not farther from each other than the preceding neighboring averages, i.e.

$$\left|\tilde{x}\left[l\Delta t+1, (l+1)\Delta t\right] - \tilde{x}\left[(l-1)\Delta t+1, l\Delta t\right]\right| \geqslant \left|\tilde{x}\left[(l+1)\Delta t+1, (l+2)\Delta t\right] - \tilde{x}\left[l\Delta t+1, (l+1)\Delta t\right]\right|. \qquad (9)$$

Nevertheless, requirements (8) and (9) are too primitive and can be satisfied only in special cases with wide windows. Even the example in Fig. 1, which seems to fit for (8) and (9), satisfies neither (8) nor (9). So, requirements (8) and (9) must be converted into more flexible conditions. Thus, the following

two inequalities should additionally hold starting at some l_* for the narrowest window τ_*:

$$\eta(l_*)=\frac{\sigma_{\tilde{x}}\left[(l_*-1)\tau_*+1,\, l_*\tau_*\right]}{\tilde{x}\left[(l_*-1)\tau_*+1,\, l_*\tau_*\right]}<\varepsilon \quad \text{and} \quad \mu(l_*)=\frac{\max\limits_{m=(l_*-1)\tau_*+1,\, l_*\tau_*}\left|\tilde{x}_m-\tilde{x}\left[(l_*-1)\tau_*+1,\, l_*\tau_*\right]\right|}{\tilde{x}\left[(l_*-1)\tau_*+1,\, l_*\tau_*\right]}<\lambda\varepsilon \qquad (10)$$

for some $\varepsilon>0$ and $\lambda\geqslant 1$. In practice, it is relevant to set $\varepsilon=0.005$, $\varepsilon=0.001$, or even less. Now, the example in Fig. 3 seemingly satisfies requirements (10), but that wave is a sign of an instable empirical distribution. To spot such cases, it is better to widen the window and see whether both inequalities in (10) are still true. Having satisfied requirements (10) by $\tau_*=\Delta t$, the widening can be fulfilled for $\tau_*=2\Delta t$ and $\tau_*=4\Delta t$ (this can be named a rule of three windows). If the rule of three windows is satisfied at windows l_*, l_*+1, l_*+2, the observations are sufficient to estimate the probability distribution. Therefore, validity of the empirical probability distribution is ensured by requirements (10) which should hold for at least the three window widths by three successive windows.

An example based on the example in Fig. 1 — 3 is presented in Fig. 4 showing how the validity is achieved for the requirements by $\varepsilon=\lambda=0.001$ (there are 4400 observation periods altogether, for which window widths of 100, 200, and 400 are used). It should be noted that the observation period in this case comprises from 800 to 1000 measurements for $x_{\min}=1$, $\alpha=1$, $x_{\max}=60$. The expected value of the parameter (unknown to the observer) is 12.6021, which is shown in Fig. 1 as the horizontal line on the plots of the parameter's expected value approximations, the moving averages and deviations.

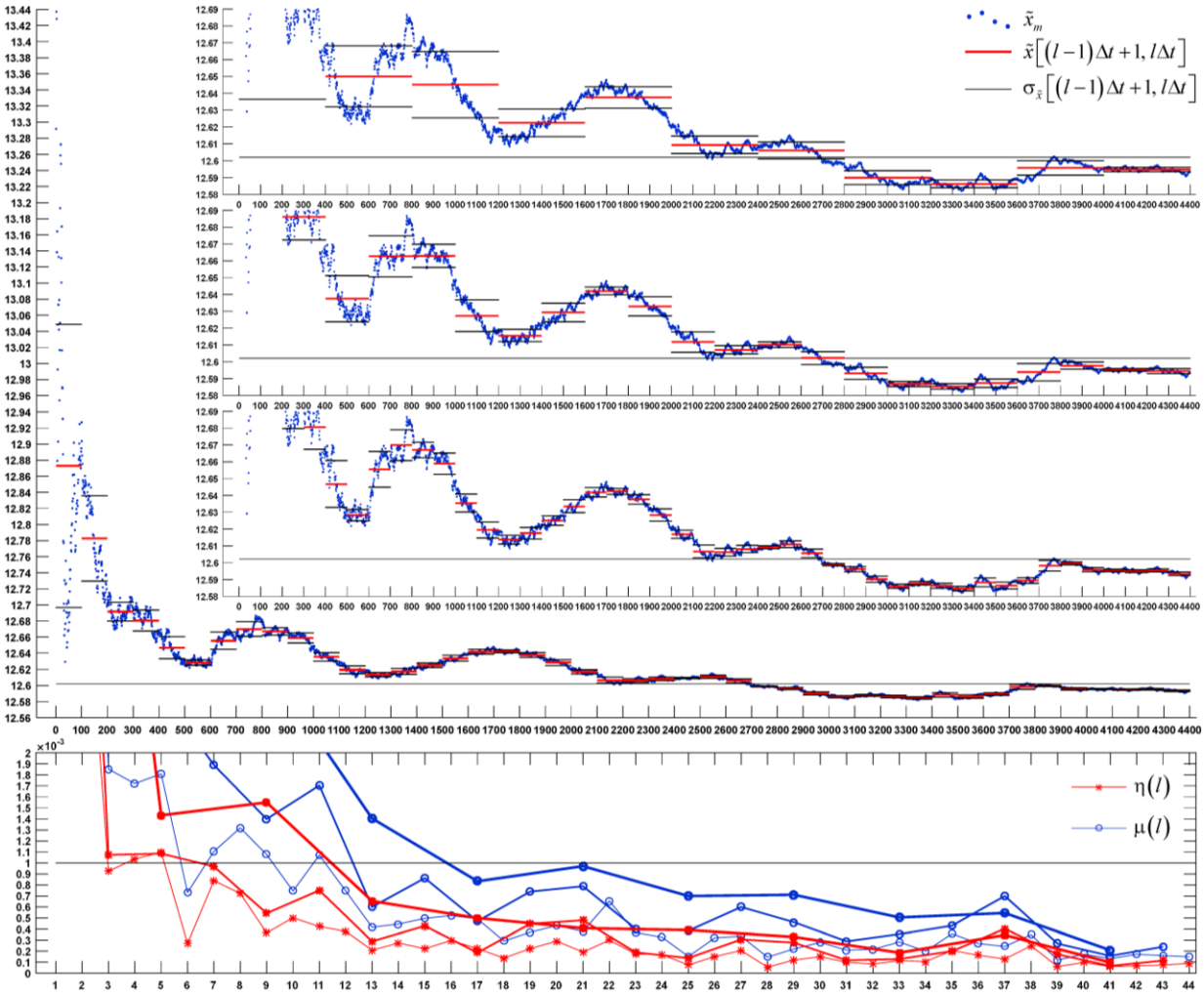

Fig. 4. The averages and the bounding deviations for the three window widths, where the functions in requirements (10) are plotted below (the thicker line corresponds to the wider window) along with the horizontal level of $\varepsilon=\lambda=0.001$ showing that 1600 observation periods (in this case) are sufficient to obtain a valid empirical probability distribution (starting off 1600 observation periods, the difference between the moving average and the unknown expected value of the parameter does not exceed 0.3648 %)

Discussion and conclusion

In practice, it is worth remembering that the validity strongly depends on the volume of observations per period. The bigger is this volume, the faster is the convergence. Otherwise, when the number of observations per period is relatively small, a much greater number of observation periods may be required. Besides, the empirical probability distribution validity herein is substantiated with respect to the parameter's expected value. So, if the parameter is badly influenced by a lot of weakly controllable factors, then either ε and λ should be increased or the duration of observations should be prolonged.

In general, the presented methodology of accumulating and processing observation data is based on the rule of three windows, where the moving averages and root-mean-square deviations are used. It asserts that the respective empirical distribution is valid (i.e., it is sufficiently close to the unknown real distribution) if the parameter's expected value approximations become scattered very little for at least the three window multiple-of-2 widths by three successive windows. This criterion also implies the sufficiency of observation periods, although the sufficiency of observations per period is not claimed.

REFERENCES / ЛІТЕРАТУРА

1. M. Melucci, "A brief survey on probability distribution approximation", *Computer Science Review*, vol. 33, pp. 91 – 97, 2019.
2. P. Samui, D. Tien Bui, S. Chakraborty, R. C. Deo, *Handbook of Probabilistic Models*. Butterworth-Heinemann, 2020, 590 p.
3. V. V. Romanuke, "Evaluating validity of the statistic frequencies distribution of a variate with undefined mathematical expectation and variance", *Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling*, no. 21, pp. 152 – 161, 2010.
4. D. S. Wilks, "Empirical Distributions and Exploratory Data Analysis", in: *Statistical Methods in the Atmospheric Sciences (Fourth Edition)*. D. S. Wilks (Ed.), Elsevier, 2019, pp. 23 – 75.
5. V. V. Romanuke, "Wind farm energy and costs optimization algorithm under uncertain parameters of wind speed distribution", *Studies in Informatics and Control*, vol. 27, iss. 2, pp. 155 – 164, 2018.

Romanuke Vadim Vasylyovych *— doctor of technical sciences, professor; professor of department of information technologies, O. S. Popov Odessa National Academy of Telecommunications, Kuznecznaya str., 1, Odessa, Ukraine, 65029; e-mail: romanukevadimv@gmail.com;*
ORCID: http://orcid.org/0000-0003-3543-3087.

Романюк Вадим Васильович *— доктор технічних наук, професор; професор кафедри інформаційних технологій, Одеська національна академія зв'язку ім. О. С. Попова, вул. Кузнечна, 1, Одеса, Україна, 65029; e-mail: romanukevadimv@gmail.com;*
ORCID: http://orcid.org/0000-0003-3543-3087.

Романюк Вадим Васильевич *— доктор технических наук, профессор; профессор кафедры информационных технологий, Одесская национальная академия связи им. А. С. Попова, ул. Кузнечная, 1, Одесса, Украина, 65029; e-mail: romanukevadimv@gmail.com;*
ORCID: http://orcid.org/0000-0003-3543-3087.

Надійшла – 12.03.2020.