Zeszyty Naukowe Wydziału Elektrotechniki i Automatyki Politechniki Gdańskiej Nr 55

XXI Międzynarodowe Seminarium Metrologów

MSM 2017

Rzeszów - Czerniowce, 12-15 września 2017

THE RANGE OF THE FOURTH ORDER MOMENT WHEN THE VALUES OF THE FIRST TWO MOMENTS ARE KNOWN

Grzegorz SMOŁALSKI

Department of Biomedical Engineering, Faculty of Fundamental Problems of Technology, Wrocław University of Science and Technology, phone: 48 71 320 28 61, e-mail: Grzegorz.Smolalski@pwr.edu.pl

Summary: The easy and possibly effortless detection of departures of probability distributions from the normality would be very useful in measurements practice. For this, except for lower order moments, the fourth order moment may also be fruitful. Here, the limits of this moment were found under the assumption that the distribution has finite support and the values of the first two moments are already known. On this basis, the range of the fourth moment was evaluated for all possible values of the mean and mean-square values. Then, an uncertainty reduction coefficient was defined and calculated, that describes how many times the range of the fourth moment is reduced when the values of the first two moments are provided. It was found that for symmetric distributions this uncertainty is reduced at least fourfold for bipolar variables, and 64-fold for the unipolar variables.

Keywords: knowledge-based measurement, kurtosis, finite support distribution.

1. INTRODUCTION

Affording possibilities for a uniform treatment of all the information available about a measurand is one of purposes of the present revision of the Guide to the Expression of Uncertainty in Measurement (GUM), [1]. In this revised approach, a probability distribution describing the measurand becomes a key mathematical tool for depicting both the state of our knowledge and its evolution when new facts are collected.

A random variable is a typical model of the measurand both in its classical meaning as a value of an (almost) stable quantity and as a parameter *E* of the signal [2]. In the case of a signal, the available knowledge may concern not only the estimated parameter *E* directly, but also some other parameter(-s) *M* of the signal.

If the distribution assigned to a measurand appropriately takes into account all physical, technological, and physiological constraints, then the support of distribution usually has a limited support, i.e. the set of values for which the probability is non-zero, is limited: $-\infty < x_l \le X \le x_u < \infty$. This fact may be ignored in many practically important cases when the standard deviation σ is sufficiently small: $\sigma << x_u - x_l$, and additionally the mean value of the distribution is located sufficiently far from both the limits, i.e.: $m_1 - x_l >> \sigma$, and $x_u - m_1 >> \sigma$. In all other cases, the existence of the distribution support limits is important, and it becomes interesting, how to transpose all the known constraints into properties of the estimated parameter *E* distribution. Answering this question is an element of the general Bayesian evaluation of a posterior probability density function (pdf). However, both analytical and Monte Carlo approaches to Bayesian analysis require additional efforts. On the other hand, the possibility of a preliminary determination of some pdf's properties on the basis of the first two moments, would be attractive. It would be especially desirable to detect such cases when a probability differs considerably from a Gaussian, since then a full pdf investigation is necessary.

Distribution asymmetry is often used as a primary property of being non-Gaussian. In the cases, however, when a symmetrical pdf is expected[1], a secondary distinguishing property becomes a kurtosis: $kurt = (\mu_4/\sigma^4) - 3$, defined as a central fourth order moment μ_4 normalised by a standard deviation σ, and shifted by 3, that is the value of μ_4/σ^4 for a normal distribution. Thus, a positive value of kurtosis means that μ_4/σ^4 is greater than that of a normal distribution, which signifies a pdf's fat tails [5, 6]. Even for symmetrical pdfs, for which $\mu_3 = 0$, we have $\mu_4 = m_4 - 6m_1^2 m_2 + 5m_1^4$, and kurtosis usually cannot be calculated accurately solely on the basis of the ordinary moments: m_1 and m_2. Nevertheless, the value of m_4, being a main component of the kurtosis, can be assessed approximately. In this paper, the possible limits of m_4 are investigated for finite supports distributions.

2. POSSIBLE VALUES OF m_1 AND m_2

Pieces of information assumed here to be known, i.e.: the lower x_l and upper x_u limit of the distribution support, the first two moments m_1, m_2, and the symmetry of the pdf, are not independent. The mean value satisfies $m_1 \in [x_l; x_u]$. For some given value of m_1 and the assumption of a pdf

[1] The assumption about pdf's symmetry may be a consequence of prior knowledge about symmetry of all factors causing variability of the results and, additionally, about sufficient symmetry or linearity of transfer functions of transducing paths. When such assumptions are not plausible, pdf's symmetry may also be verified statistically on the basis of a sample.

symmetry, the value of the second moment is constrained from the bottom by a parabola $m_{2\min} = m_1^2$ that corresponds to a random variable taking values other than m_1 with a zero probability, i.e. very rarely. From the top m_2 is constrained by one of the following parabolas:

$$m_{2\max} = 2m_1^2 - 2cm_1 + c^2 , \qquad (1)$$

where $c = x_l$ for $m_1 \le (x_l + x_u)/2$, or $c = x_u$ for $m_1 \ge (x_l + x_u)/2$. These parabolas correspond to random variables assuming two values with equal probabilities: x_l and $2m_1 - x_l$ (for $m_1 \le (x_l + x_u)/2$) or $2m_1 - x_u$ and x_u (for $m_1 \ge (x_l + x_u)/2$). A resulting set of possible pairs of (m_1, m_2) values is depicted in figure 1 as a shaded area. In this graph, also a straight line $(x_l + x_u)\, m_1 - x_l x_u$ is placed constituting an upper limit of m_2 in the case when a pdf symmetry is not assumed. As may be observed, a symmetry assumption markedly constrains a set of possible pairs (m_1, m_2).

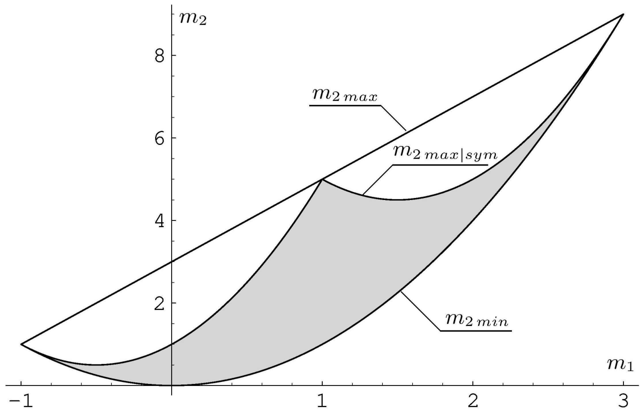

Fig. 1. Bounding parabolas of the mean-square value m_2 when the mean value m_1 is known. The straight line above depicts $m_{2\max}$ when the assumption about pdf's symmetry is omitted. Here, exemplary values $x_l = -1$ and $x_u = 3$ are taken.

3. THE LIMITS OF THE FOURTH MOMENT

Next, distributions were found that, having a support limited to $[x_l; x_u]$, reveal a maximum or minimum value of m_4. Successively, the following values were assumed to be known: the mean value m_1, the mean-square value m_2, both m_1 and m_2 at the same time, the values m_1 and m_2 and the fact that the pdf is symmetrical. Because the value of m_4 is used for a pdf shape evaluation mainly when the pdf is symmetric, therefore this case is presented here in detail.

It was found that a maximum value of m_4 reveals a discrete distribution taking three values: $P(X = m_1) = p$, $P(X = c) = (1 - p)/2$, and $P(X = 2m_1 - c) = (1 - p)/2$, where c assumes values defined with equation (1), while the value of probability p must be selected so as to ensure the observed value of m_2 (for a similar approach cf. [7]). On the other hand, a minimum value of m_4 is revealed by a random variable having a two-point distribution. Both values $m_1 - \delta$ and $m_1 + \delta$ should be taken with equal probabilities, and the value of δ should assure the known value of m_2 and, of course, should not exceed the range: $0 \le \delta \le \min\{x_u - m_1, m_1 - x_l\}$.

For the found distributions the values of $m_{4\max}$ and $m_{4\min}$ were calculated. Their difference $\Delta m_4|\{m_1, m_2, sym\} = m_{4\max} - m_{4\min}$ constitutes in fact a doubled expanded uncertainty of the fourth moment when for a symmetrical pdf the values of m_1 and m_2 are both known. A coverage probability corresponding to the interval $[m_{4\min}; m_{4\max}]$ equals unity, because one can be sure that the value of m_4 lies within this range. The dependence of the width $\Delta m_4|\{m_1, m_2, sym\}$ of the coverage interval on all possible values of m_1 and m_2 is presented in figure 2.

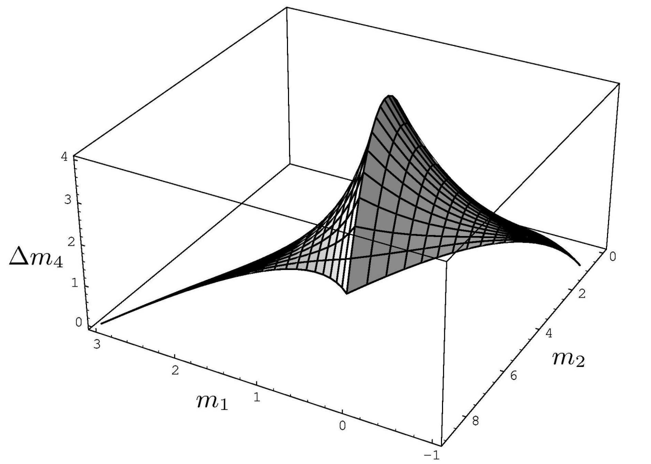

Fig. 2. The width Δm_4 of the coverage interval of m_4 when both the mean m_1 and the mean-square m_2 values are known, and additionally the pdf's symmetry is assumed. Here $x_l = -1$ and $x_u = 3$.

As can be seen, the uncertainty vanishes at the borders of the possible values of m_2 since there is then a single possible distribution fulfilling the requirements. Although the uncertainty is relatively small in a substantial part of the area of possible values (m_1, m_2), nevertheless for assessing a general suitability of the first two moments for the fourth moment evaluation the greatest, i.e. worst case, values of $\Delta m_4|\{m_1, m_2, sym\}$ must be taken into account.

As an example of the application of the above analysis for the detection of the marked departure of the pdfs from the normality, a two-parameter beta distribution was considered. This distribution has a finite support in the form of an interval $[0;1]$. When both the shape parameters α and β of the beta distribution are equal and acquire a value from the interval $(0;1)$, its pdf is symmetrical and U-shaped. When $\alpha = \beta \to 0_+$, the beta distribution tends to a two-point distribution, while for $\alpha = \beta \to 1_-$ it approaches a uniform distribution. For all symmetrical beta distributions $m_1 = 1/2$ and $m_2 = (1+\alpha)/(2+4\alpha)$. In this special case, also the value of m_4 might be calculated directly on the basis of α. For the present purpose, however, the limits $m_{4\min}$ and $m_{4\max}$ of the fourth moment were calculated on the basis of the first two moments and on the support limits x_l, x_u. Then, the limits of μ_4 and of the kurtosis were obtained in turn. It was expected that for a significant departure of the pdf's shape from the normality all the possible values of the kurtosis will be of the same sign. Indeed, it was found that for $\alpha = 0.1$ (that corresponds to $m_2 \cong 0.458$) and on the assumption that $x_l > -0.291$ and $x_u = 2$, the upper limit of the kurtosis is $kurt_{\max} \le 0$. This means that for this distribution a value of the kurtosis must be negative. It is worth noting that this fact was detected here for a quite rough outer bounding of the support: $(0;1) \subset (-0.291;2)$. For $\alpha = 0.9$, however, when $m_2 \cong 0.339$ and the distribution is almost uniform, such a detection, while still possible, requires already a much tighter evaluation of the lower limit of the support: it must be $x_l > -0.00175$ then.

4. THE COEFFICIENT OF UNCERTAINTY REDUCTION

To quantify the suitability of m_1 and m_2 for Δm_4 reduction, a coefficient was defined [3,4]: $r = \Delta m_4 \big| \{x_l, x_u, PK\} / \Delta m_4 \big| \{x_l, x_u\}$. In this definition, the symbols $\Delta m_4 \big| \{\ldots\}$ stand for the uncertainty of the fourth moment assuming that values or facts (constituting prior knowledge *PK*) enumerated in braces are known. Thus, the denominator of this fraction signifies the uncertainty of the fourth moment when the limits of the distribution support are solely known. This uncertainty equals:

$$\Delta m_4 \big| \{x_l, x_u\} = \begin{cases} \max\{x_l^{\,4}; x_u^{\,4}\} - \min\{x_l^{\,4}; x_u^{\,4}\}; & x_l x_u > 0 \\ \max\{x_l^{\,4}; x_u^{\,4}\} & ; x_l x_u < 0 \end{cases} \tag{2}$$

Here, the symbols $\min\{\cdots\}$ and $\max\{\cdots\}$ stand for minimum and maximum value listed in braces, respectively.

In the numerator of the *r* definition it is assumed that additional prior knowledge *PK* is available comprising here the knowledge of m_1, m_2, or both of them, without or with the symmetry of the pdf. Since usually the value of $\Delta m_4 \big| \{x_l, x_u, PK\}$ depends on the value of the known moment(-s) included in the prior knowledge, the coefficient *r* can be considered as a function of the known value(-s) or, as here, the maximum of the numerator can be found, thus providing a worst-case value of *r*:

$$r_{m_4|PK} = \frac{\max \Delta m_4 \big| \{x_l, x_u, PK\}}{\Delta m_4 \big| \{x_l, x_u\}} \ . \tag{3}$$

The value of r always satisfies $0 \le r_{m_4|PK} \le 1$ since additional knowledge never results in the increase of uncertainty. The value of *r* equal to 1 doesn't mean that prior knowledge *PK* is completely useless for the evaluation of m_4 but solely for some case(-s) considered in the numerator maximization, the reduction of uncertainty is not achieved.

The value of $r_{m_4|PK}$ was analysed for all four cases of prior knowledge mentioned above, and the results are depicted in figure 3. It was found that $r_{m_4|PK}$ doesn't depend directly on the values of the support limits, but rather depends on their quotient: $k = -x_u / x_l$. Negative values of *k* correspond to "unipolar" random variables, i.e. taking solely values of one sign. The case of $k = -1$ stands for random variables assuming with a non-zero probability only a single value $x_u = x_l$, while a vicinity of the value $k = -1$ describes unipolar random variables of a relatively small variability range. Positive *k* describes "bipolar" random variables, i.e. taking the values of both signs. A special case of $k = +1$ corresponds to variables the support of which has limits placed symmetrically about zero: $x_u = -x_l$.

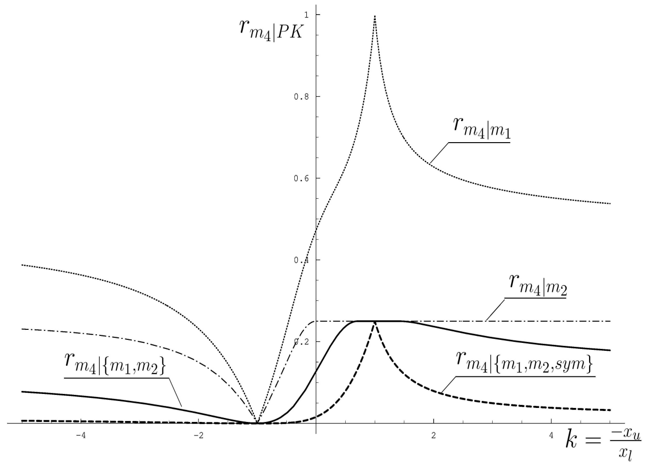

Fig. 3. Values of $r_{m_4|PK}$ coefficients (3) plotted as a function of the support limits' quotient $k = -x_u / x_l$.

The graphs in figure 3 reveal many general and interesting properties discussed already in [4]. The main thing they reveal is that each of the graphs consists of two branches: for $k < 0$ and $k > 0$. On the other hand, each of the branches reveals a special kind of symmetry about the values $k = \mp 1$, namely: $r(k) = r(1/k)$.

As seen from the graphs, the mean value is usually the least appropriate for m_4 estimation. This may be caused both by a large difference of both moments orders and also by the fact that the order of the estimated moment is even while the order of the known - is odd. The graph of $r_{m_4|m_2}$, however, always takes values equal to or below 0.25. It is seen that the second moment is much more suitable for the m_4 estimation than m_1. For bipolar random variables ($k > 0$) appending the knowledge of m_1 when the value of m_2 is already known gives only a weak further reduction of uncertainty. For unipolar variables ($k < 0$), however, an additional knowledge of m_1 is already visibly advantageous because it decreases the coefficient *r* from 0.25 to 0.125, in the worst cases.

Revealing a pdf symmetry together with simultaneous knowledge of both m_1 and m_2, reduces the uncertainty about m_4 at least 4 times for bipolar variables, and at least 64 times for unipolar random variables. It is important that $r_{m_4|\{m_1,m_2,sym\}}$ not only vanishes for $k = -1$ (that is obvious), but is negligible also for a relatively large vicinity of this point, i.e. also for variables of considerable variability.

5. CONCLUSIONS

Kurtosis is a distribution parameter used mainly for verification how fat the tails of the pdf are [5, 6]. In measurements kurtosis may be, e.g., useful for the detection of significant departures of the symmetric pdfs from a normal distribution. An ordinary moment of the fourth order is an important component of the kurtosis. In this work, values of $m_{4\,\min}$ and $m_{4\,\max}$ were found for every possible distribution revealing the finite pdf's supports, assuming that additionally the value(-s) of m_1 or/and m_2 are known. On this basis, the range Δm_4 was determined for the assumed knowledge. It was also studied, how the knowledge about m_1 or/and m_2 reduces the value of Δm_4 in relation to the case when solely the limits of the pdf's support are known. It was found that if the pdf is symmetric, and both values of m_1 and m_2 are known at the same time, then Δm_4 is reduced at least fourfold for bipolar random variables, and at least 64-fold for unipolar ones.

The obtained results will be further applied for finding such pairs of $\{m_1, m_2\}$ for which normality of the distribution is excluded and therefore a detailed investigation into the pdf's shape is required, e.g. using a Monte Carlo method. The values of m_1 and m_2 are never known exactly. Therefore, in practical applications, the extreme values of the fourth moment $m_{4\,\min}$ and $m_{4\,\max}$ should not be calculated for a single pair of values $\{m_1, m_2\}$, but rather they should be found out from the entire rectangle: $m_1 \pm \Delta m_1$ and $m_2 \pm \Delta m_2$.

The presented results were obtained using a method of probability support constraining [3,4]. As it is not Bayessian, this method fails to provide a full pdf description. Instead, this method propagates all available limits of the pdf supports to the limits of the interesting pdf. For some purposes, the limits of such support prove quite sufficient.

6. REFERENCES

1. Lira I.: The GUM revision: the Bayesian view toward the expression of measurement uncertainty, European Journal of Physics, Vol. 37 (2016), p. 025803.
2. Smołalski G.: Parameter of the signal – A useful construct of a measurand, Measurement, Vol. 56 (2014), pp. 163-169.
3. Smołalski G.: Measurability conditions of the signal parameter for a given prior knowledge, Measurement, Vol. 42 (2009), pp. 583-603.
4. Smołalski G.: A method of probability support constraining and its application to the evaluation of the third order moment, Problems and progress in metrology PPM'16, Szczyrk, 5-8 June 2016, ISBN 978-83-939486-4-2, pp. 35-38.
5. Westfall P.H.: Kurtosis as Peakedness, 1905-2014. R.I.P., The American Statistician, Vol. 68 (2014), No. 3, pp. 191-195.
6. Moszczyński L.: The interpretation of the kurtosis statistics (in Polish), Przegląd Elektrotechniczny, Vol. 79 (2003), No. 9(1), pp. 558-560.
7. Wong F.S.: First-Order, Second-Moment Methods, Computers & Structures, Vol. 20 (1985), pp. 779-791.

ZAKRES ZMIENNOŚCI MOMENTU CZWARTEGO RZĘDU GDY ZNANE SĄ WARTOŚCI MOMENTÓW DWÓCH PIERWSZYCH RZĘDÓW

W metrologii często występuje potrzeba szybkiej, i możliwie nie wymagającej dodatkowych analiz, oceny czy rozkład prawdopodobieństwa wyniku pomiaru jest choćby w przybliżeniu podobny do normalnego. W przypadku znacznych odchyleń funkcji gęstości od krzywej normalnej oszacowanie przedziału niepewności wyniku wymaga pełnego zbadania funkcji gęstości, np. metodą Monte Carlo. Najprostszą metodą wykrywania takich odchyleń od normalności jest detekcja asymetrii funkcji gęstości, do czego zwykle wykorzystuje się unormowany trzeci moment centralny. Gdy jednak spodziewany kształt funkcji gęstości jest symetryczny, trzeba wykorzystywać momenty czwartego rzędu. W pracy omówiono możliwość wnioskowania o wartości momentu zwykłego czwartego rzędu m_4 na podstawie wartości średniej m_1 i średniokwadratowej m_2 w przypadku rozkładów prawdopodobieństwa o ograniczonym nośniku. Określono wartości $m_{4\,\min}$ i $m_{4\,\max}$ dla czterech przypadków wiedzy apriorycznej: gdy znana jest wyłącznie wartość jednego z momentów: m_1 albo m_2, gdy znane są obie te wartości jednocześnie, oraz przypadek gdy oprócz tego wiadomo, że rozkład ma symetryczną funkcję gęstości. Dla tego ostatniego przypadku zaprezentowano zależność niepewności Δm_4 od wartości znanych momentów m_1 i m_2. Podano przykład wykorzystania wartości granicznych $m_{4\,\min}$ i $m_{4\,\max}$ momentu czwartego rzędu. Zbadano także zależność stopnia redukcji niepewności Δm_4 uzyskiwanej dzięki znajomości momentów m_1 albo/i m_2 w najmniej korzystnych warunkach. Stwierdzono, że redukcja niepewności zależy wtedy od ilorazu wartości granicznych nośnika funkcji gęstości. W przypadku, gdy funkcja gęstości jest symetryczna i znane są oba pierwsze momenty rozkładu, wówczas Δm_4 jest co najmniej czterokrotnie mniejsze niż wtedy, gdy znamy tylko granice nośnika funkcji gęstości, a dla zmiennych losowych przyjmujących wyłącznie wartości jednego znaku, jest nawet 64-krotnie mniejsze.

Słowa kluczowe: pomiar z wykorzystaniem wiedzy apriorycznej, kurtoza, rozkład o ograniczonym nośniku.