XXI IMEKO World Congress "Measurement in Research and Industry"
August 30 – September 4, 2015, Prague, Czech Republic

SPECIFICATION OF THE ASSUMPTIONS FOR THE TYPE-A EVALUATION OF STANDARD UNCERTAINTY

Andrzej Zięba [1]

[1] Faculty of Physics and Applied Computer Science, AGH University of Science and Technology, Kraków, Poland, *andrzej.zieba@fis.agh.edu.pl*

Abstract – The formulation of Type A uncertainty in the GUM does not include the necessary assumptions. The paper reviews the common set of four assumptions and, in particular, the replacing the identity of distributions (i. d.) by a more general assumption of equivalence of observations. This assumption shows several advantages including its more natural relation to the measurement conditions. Possibilities of validating the assumptions, by an analysis of experiment and using the statistical tests, are briefly discussed.

Keywords: Type A uncertainty, assumptions, equivalence, autocorrelated

1. INTRODUCTION

The GUM formulae for evaluating the Type A uncertainty [1] of repeated measurement,

$$x \equiv \bar{x} = \sum_{i=1}^{n} x_i \tag{1}$$

and

$$u(x) \equiv s(\bar{x}) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n\,(n-1)}}\,, \tag{2}$$

are well-known expressions for estimators of the expected value and the standard deviation of the mean for an n-element sample.

The formulae (1) and (2) are nothing else but the theses of some mathematical theorems, hence they are valid under certain assumptions. The fact that formulation of these assumptions is not trivial is confirmed by the fact that they are not specified by *Guide* [1]. This document says barely that Eqs. (1) and (2) are valid "in most cases" (see 4.2.1 in [1]). This statement implies that there exist situations when another mathematical formalism should be used to calculate both estimate of the measurement value and its uncertainty.

The arithmetic mean (1) and sample standard deviation of the mean (2) are, from the point of view of probability theory, the particular estimators. Contrary to the majority of mathematical formulae, they cannot be labelled as "true" or "false". Validity of estimators is a more subtle issue. There exist several estimators of, say, location and scale parameters, and one should make a choice of optimal estimator in the given circumstances. The properties of consistency, lack of bias, and high effectiveness are considered the most important ones in classical mathematical statistics. The robust statistics adds the requirement of robustness. One should bear in mind that estimators (1) and (2) are not robust.

The standard set of assumptions will be discussed in Section 2. It ensures that the mean $\bar{x}$ as an estimator of expectation and the variance of both single measurement s^2 and the variance of the mean $s^2(\bar{x})$ are unbiased and most effective. The sample standard deviation s and the standard deviation of the mean $s(\bar{x})$ are biased-due to a nonlinearity introduced by the operation of calculating the square root. We stand for that (small) bias in order to preserve the simplicity of the formula (2).

The same mathematical formula or theorem can often be valid under various assumptions. The well-known example is the central limit theorem. Its thesis of normality of an infinite sum of random variables was proved, from de Moivre to Lyapunov [2], using more and more general assumptions. In Section 3 there is discussed the assumption of equivalence of elements of the sample as an alternative to the more common assumption of identity of probability density function of elements x_i of the sample. The equivalence assumption is more general and, in spite of generality, enables a better connection with the conditions of measurement.

One can ask a question: what is a reason to discuss these mathematical issues in metrology? The choice of assumption in mathematics is usually important to provide a proof of a given theorem. The choice of assumptions in applied sciences, including metrology, has a wider significance. The given set of assumptions can be more or less suitable to check whether they are fulfilled for a given measurement. This check can be made in two ways. The first method consists in identifying the relevant conditions by a competent investigator. The second one is the use of statistical tests aimed at checking the properties of a set of observations. They can be labelled, respectively, as the Type B and the Type A methods of validating the assumptions. Tests of the assumptions using both approaches are discussed in Sections 4 and 5, respectively.

Alternative statistical methods are to be used when the assumptions related to standard formalism are not fulfilled. A short overview of alternative methods is given in Section 6. They can be used when one of four assumptions discussed in this work is violated, whereas the remaining ones are valid.

2. STANDARD SET OF ASSUMPTIONS

The mathematical model of measurement represents a basis for the data processing. The subsequent result x_i of a repeated measurement is considered as the sum of the true value x_0 and a random variable e_i representing the measurement error,

$$x_i = x_0 + e_i \, . \tag{3}$$

One assumes that

a) expected value of e_i is zero, *i. e.*, $\mu_e = 0$,

b) elements e_i are statistically independent,

c) elements e_i have an identical probability distribution, and

d) distribution is normal (Gaussian).

The number of four assumptions can be considered excessive. For the choice of the mean and sample variance s^2 one can use a single assumption that the measurement error is modelled by a normal distribution with an unknown standard deviation and zero expectation. Specification of four assumptions is, however, useful because it turns attention to investigate various features of the measurement. It is particularly useful to classify the alternative methods (Sec. 6).

Assumptions b) and c) are known in statistical literature under the acronym i. i. d., after *independent, identically distributed*. Estimators $\bar{x}$, s^2 and $s^2(\bar{x})$ are consistent, unbiased and most effective in the class of linear estimators when the i. i. d. assumption is fulfilled (Gauss-Markoff theorem). The additional assumption of normality is necessary to ensure that the mean is most effective in the class of all estimators of location (Cramer-Rao lower bound). The most effective estimators of location for a given distribution can be derived from the maximum likehood principle. That way one can obtain, *e. g*., that the median is the most effective estimator of location for the double exponential distribution.

The assumption "independent" means that the probability of obtaining the value x_i (in an arbitrary interval of x variable) does not depend on the previous values. Alternatively, one can state that the elements of the sample are *uncorrelated*, *i. e*., all correlation coefficients $\rho(x_i, x_j)$ are equal zero. These two conditions are related but not identical. Nonzero value of $\rho(x_i, x_j)$ implies an existence of statistical dependence but there are examples of statistically dependent variables with $\rho(x_i, x_j) = 0$. This remains rather a mathematical curiosity because statistically dependent observations are also correlated (with positive values of correlation coefficients) in a vast majority of practical circumstances.

3. EQUIVALENCE OF OBSERVATIONS AS A GENERALISATION OF THE ASSUMPTION OF IDENTITY OF DISTRIBUTIONS

The mathematical condition of the identity of probability distribution (i. d.) cannot be verified in most measurements. It has been proposed, in conjunction with an analysis of autocorrelated observations, to replace it by a postulate of *equivalence* of observations [3]. The term "equivalence" is commonly understood as a fact that different objects have the same properties (*e. g*., equivalence of drugs with various brand names). *The condition of equivalence can be defined as a requirement that all stochastic properties of a given sample element x_i, including the relation to neighbouring elements, do not depend on the shift of the index i*. This condition is used, *e. g*., in crystallography. The translation by a discrete lattice vector, starting from a given atom, leads to an equivalent atom with the same surrounding.

The principle of equivalence can be applied to a series of observations. One can list several advantages when compared to the common assumption of identity of distributions:

a) The equivalence condition is more general. For independent observations it implies identity of probability distributions $g_i(x) = g_j(x)$, the corresponding parameters like expectation μ and variance σ (Fig. 1). Thus the notion of identical distribution (i. d.) is a consequence of the equivalence of a series of observations.

$\mu, \sigma, g(x)$

$\downarrow$

$x_1, \quad x_2, \quad x_3, \quad x_4, \quad x_5, \quad x_6, \quad x_7, \quad x_8, \quad x_9$

$\uparrow \cdots\cdots \rho_{i,\, i+3} \cdots\cdots \uparrow$

$\mu, \sigma, g(x)$

$\downarrow$

$x_1, \quad x_2, \quad x_3, \quad x_4, \quad x_5, \quad x_6, \quad x_7, \quad x_8, \quad x_9$

$\uparrow \cdots\cdots \rho_{i,\, i+3} \cdots\cdots \uparrow$

Fig. 1. Illustration of the equivalence of observations

b) When observations are autocorrelated, the equivalence ensures that correlation coefficients linking elements x_i and x_j can only depend on the difference $i - j$. As a result, it is the equivalence condition which determines the specific structure of a correlation matrix for serially correlated data

$$\rho_{ij} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{bmatrix} . \tag{4}$$

The information contained in this 2-d matrix is uniquely specified by a one-dimensional autocorrelation function $\{\rho_k\}$, $k = 0, 1, \ldots, n-1$. This structure of autocorrelation matrix cannot be deduced from the i. d. assumption alone.

c) The equivalence assumption, due to its generality, can also be used in the interval theory of uncertainty, where the very notion of probability is not used.

d) In spite of its wider generality, it is the equivalence assumption which can be linked more easily to the characteristics of measurement.

e) Replacement of the i. d. assumption by the equivalence assumption leads to a more coherent set of assumptions, in which two of them, namely independence and equivalence, are expressed in a similar manner.

4. MATHEMATICAL ASSUMPTIONS vs. CHARACTERISTICS OF MEASUREMENT

Metrology is an applied science. Application of any mathematical formalism, including its underlying assumptions, should be related to the more-or-less known features of measurement.

The condition that the expected value of the error is zero corresponds to the case that the systematic error is small enough to be neglected. Our knowledge concerning that issue stems either from our knowledge of the measurement or from the dedicated control measurement with the use of a well-characterised standard.

Independence or statistical dependence of observations can be deduced, in most cases, from an understanding of both the measured object and the measuring device. This condition is fulfilled in most situations but its violation is not uncommon. Almost all meteorological and economic data (say, a daily temperature and a value of stock index) are autocorrelated because their actual value is related to yesterday's or earlier recordings.

One can expect that *equivalence* of observations is fulfilled when both the object and the measuring apparatus do not change during the measurement. This condition can be linked to a metrological term of *repeatability* (of results of measurement) ([1], Annex B).

Normal distribution of error is known a priori only in some types of measurement. One of them is the number of counts in a detector of particles of quanta or radiation. In other cases the normality is commonly assumed as long as the evidence of non-normality can be proved by an experiment.

5. STATISTICAL TESTS FOR ASSUMPTIONS

The evaluation of four assumptions discussed in Section 4 can be named as „Type B" since it is based on the analysis of the measurement condition made by an experienced experimenter. One should also discuss both possibilities and limits of validation of the assumptions by application of statistical tests to investigate the given sample $x_1, x_2, \ldots, x_n$. This option can be labelled as „Type A" analysis of assumptions.

No statistical procedure allows detecting the systematic error by the analysis of a set of observations alone. One should know a reference standard x_0 with a known uncertainty (e. g., voltage standard when checking a voltmeter) clearly smaller than the determined $u(x)$. The systematic error is detected when

$$|\bar{x} - x_0| > U(x) \, . \tag{4}$$

where $U(x) = k\,u(x)$ is the expanded uncertainty.

The condition of identity of distribution (i. d.) cannot also be deduced from the data. It seems that only two remaining assumptions can really be tested for the given sample, namely, the presence of autocorrelation and the normal distribution of errors.

Figure 1 shows an example of real autocorrelated data [4]. This graph differs qualitatively when compared to the "white noise": it is less rugged and the change of the sign in $x_i - \bar{x}$ occurs less frequently.

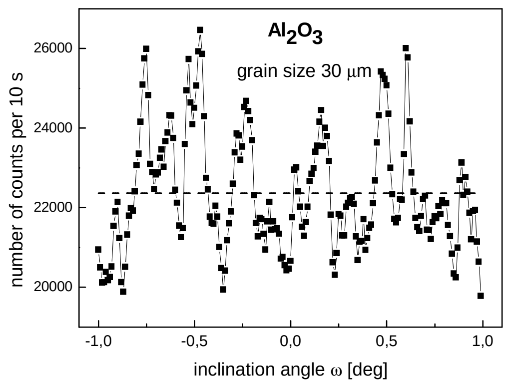

Fig. 2. Intensity of selected Bragg reflection of corundum (Al_2O_3) as a function of inclination angle ω. Dashed line denotes the mean.

Alternatively, the presence of autocorrelation can be detected using statistical tests. The popular Durbin-Watson test is based on an investigation of the test variable

$$d = \frac{\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \, . \tag{5}$$

The comparison of (5) to a standard expression for the elements of the sample autocorrelation function,

$$r_k = \frac{\sum_{i=1}^{n-k} (x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \, . \tag{6}$$

indicates that Durbin-Watson test is equivalent to testing the significance of the first element of $\{r_k\}$, since the correlations between neighbouring observations are usually the largest. Nowadays one should calculate all elements of the full estimate $\{r_k\}$ of the ACF to rule out the scarce possibility that the r_1 term does not differ significantly from zero but the presence of autocorrelation can be detected in next terms, using the results of the time series theory.

The tests discussed above depend on the assumption that the error distribution is Gaussian. The Wald-Wolfowitz run test represents an example of a robust test of autocorrelation.

The deviation from Gaussian distribution of errors can nowadays be tested using normality tests implemented in statistical packages. Figure 2 shows the comparison of the power of normality tests available in ORIGIN 7. Samples $\{x_i\}$ of *rectangular distribution* with various sizes (between n = 20 and n = 500) were investigated with the use of three tests.

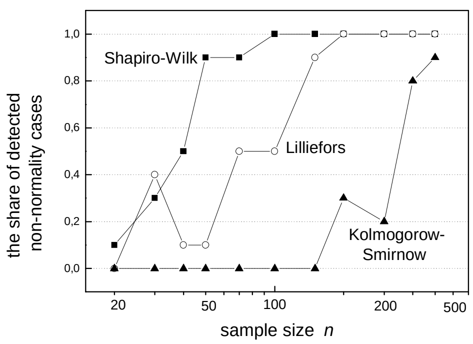

Fig. 3. Comparison of the power of three tests of normality with the use of the Monte Carlo method.

This experiment shows substantial differences in the power of various tests. The Shapiro-Wilk test is the best normality test (at least for this particular case), nevertheless one should use the sample size of about n = 50 to detect that a rectangular distribution is not normal. A dedicated experiment is often necessary to obtain a sufficiently large sample to test a normality of the random error.

The presence of outliers can also be classified as a departure from normality. As a result, they are also detected by a normality test (like the Shapiro-Wilk test), also when the number of observation is relatively small. Nevertheless it is better to use statistical tests dedicated to detecting outliers (like the Grubbs' test).

6. CLASSIFICATION OF ALTERNATIVE TYPE A METHODS

Type A evaluation of uncertainty remains possible when one or more assumptions discussed previously is violated. One should just use other statistical methods. The set of four assumptions seems convenient for classification of alternative methods. We should discuss shortly which branch of statistical theory can be used for processing the result of replicate measurements when *one of the discussed assumptions is violated and three remaining are valid*.

The case of simultaneous occurrence of both systematic and random error is the only one which is discussed in the GUM. The uncertainty related to a systematic error can be evaluated using the Type B method and should be summed geometrically with the Type A uncertainty [5].

Violation of the equivalence assumption alone means that independent observations have Gaussian distribution of error with the same expected value but with various standard deviations. This case occurs in inter-laboratory comparisons when the same sample is measured by various laboratories with an uneven accuracy. An application of the maximum likelihood method leads to the conclusion that, under the assumption of normality, an optimal estimate of measurand is the *weighted mean*, with the weights proportional to the inverse square of uncertainty.

The statistical dependence of observation can be described, for a Gaussian error, with the use of one additional quantity, namely the autocorrelation function. An optimal estimator of location can be derived using generalised least squares [6]. The minimum extension of standard formalism consists in retaining the arithmetic mean as an estimator of location but the ways of estimation of standard and expanded uncertainties are to be modified [3], [4].

Consequences of non-Gaussian distribution of error can be different. The use of the maximum likelihood principle allows deriving the efficient estimators for a given probability density function (M-estimators). The type of distribution of error is not known in most measurements and the departure from Gaussian distribution consists in a presence of outliers. Such data can be handled with the use of various methods of *robust statistics*.

The presented discussion contains only the most common methods. Several other alternative statistical methods can also be used.

One can also consider the effect of violation of two or more assumptions. The case of correlated observations which are non-equivalent was discussed by Cox *et al*. [7]. The most of interlaboratory comparisons concern an issue of non-equivalent observations (with various standard deviation attributed to each of them) and the possible presence of outlying results. The optimal interpretation of such data is a subject of the ongoing debate.

6. OUTLOOK

The standard formalism of Type A methods can be safely used when suitable mathematical assumptions are fulfilled. They san be conveniently specified as follows: the subsequent observations are *equivalent*, mutually *independent*, burdened with a *random error*, and the probability distribution of the error is *normal* (*Gaussian*).

The present paper aims to promote the *equivalency* of observations instead of the assumption of identity of distributions (abbr. i. d.). This assumption shows several advantages including its more natural link to the measurement condition.

The lack of specification of any assumptions in *Guide* is the shortcoming of that document and one should discuss how to improve the present situation. This specification can be included into some new version of the GUM. This opens an issue of discussing alternative methods of data analysis when one or more assumptions are not fulfilled. These methods are so numerous that probably a general statement should be left in official documents like the statements promoted by NIST: "It is understood that any valid statistical method that is technically justified under the existing circumstances may be used to determine the equivalent of u_i, u_c, or U. " [8] and "A Type A evaluation of standard uncertainty may be based on any valid statistical method for treating data" [9].

The arithmetic mean or other estimator of location can be derived by fitting the constant function to the data with the use of the maximum likelihood method. Almost all the discussed problems can be adapted to the case of fitting the other functions to the two-dimensional data $\{x_i, y_i\}$. One can use, in particular, the same set of assumptions, the same methods of their validation in measurement and an analogous classification of alternative methods.

REFERENCES

[1] Guide to the Expression of Uncertainty in Measurement (GUM), revised and corrected version of GUM 1995, BIPM_JCGM 100:2008.

[2] Adams W. J., *The Life and Times of the Central Limit Theorem: Second Edition*. American Math. Society, 2009.

[3] A. Zięba, "Effective number of observations and unbiased estimators of variance for autocorrelated data – an overview". *Metrol. Meas. Syst.* , vol. 17, pp. 3–16, 2010.

[4] A. Zięba and P. Ramza, "Standard deviation of the mean for autocorrelated observations estimated with the use of the autocorrelated function estimated from the data." *Metrol. Meas. Syst.* , vol. 18 , pp. 529–542, 2011.

[5] S. D. Phillips, K. R. Eberhardt, and B. Parry, "Guidilines for Expressing the Uncertainty of Measurement Results Containing Uncorrected Bias", *J. Res. Natl. Inst. Stand. Technol.*, vol. 102, pp. 577–584, 1997.

[6] J. S. Chipman, K. R. Kadiyala, A. Madansky, and J. W. Pratt, „Efficiency of the sample mean when residuals follow a first-order stationary Markoff process". *J. Amer. Statist. Assoc.*, vol. 63, pp. 1237–1246, 1968.

[7] M. G. Cox, C. Eiø, G. Mana, and F. Pennecchi, „The generalized weighted mean of correlated quantities". *Metrologia*, vol. 43, pp. S268–S275, 2006.

[8] *NIST Technical Note 1297* (1994). Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, by B. N. Taylor and C. E. Kuyatt.

[9] *The NIST Reference on Constants, Units, and Uncertainty*, http://physics.nist.gov/cuu/