Metrol. Meas. Syst., Vol. XVII (2010), No. 1, pp. 3–16

PAN POLSKA AKADEMIA NAUK

METROLOGY AND MEASUREMENT SYSTEMS

Index 330930, ISSN 0860-8229

www.metrology.pg.gda.pl

M&MS

EFFECTIVE NUMBER OF OBSERVATIONS AND UNBIASED ESTIMATORS OF VARIANCE FOR AUTOCORRELATED DATA – AN OVERVIEW

Andrzej Zięba

AGH University of Science and Technology, Department of Physics and Applied Computer Science, A. Mickiewicza 30, 30-059 Cracow, Poland (✉ zieba@novell.ftj.agh.edu.pl, +48 12 617 3551)

Abstract

When observations are autocorrelated, standard formulae for the estimators of variance, s^2, and variance of the mean, $s^2(\overline{x})$, are no longer adequate. They should be replaced by suitably defined estimators, s_a^2 and $s_a^2(\overline{x})$, which are unbiased given that the autocorrelation function is known. The formula for s_a^2 was given by Bayley and Hammersley in 1946, this work provides its simple derivation. The quantity named effective number of observations n_{eff} is thoroughly discussed. It replaces the real number of observations n when describing the relationship between the variance and variance of the mean, and can be used to express s_a^2 and $s_a^2(\overline{x})$ in a simple manner. The dispersion of both estimators depends on another effective number called the effective degrees of freedom ν_{eff}. Most of the formulae discussed in this paper are scattered throughout the literature and not very well known, this work aims to promote their more widespread use. The presented algorithms represent a natural extension of the GUM formulation of type-A uncertainty for the case of autocorrelated observations.

Keywords: autocorrelated, time series, estimator, unbiased, variance, effective number of observations.

1. Introduction

The standard statistical analysis of a set of n observations $\{x_i\}$ consists of calculating the mean $\overline{x}$:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1}$$

as well as estimators of the variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} \tag{2}$$

and the variance of the mean:

$$s^2(\overline{x}) = \frac{s^2}{n}. \tag{3}$$

Guide to the Expression of Uncertainty in Measurement (GUM) [1] defines type-A uncertainty as the square root of an unbiased estimator of the variance, hence $u(x) \equiv s(\overline{x}) = \sqrt{s^2(\overline{x})}$. The relative dispersion (defined by Eq. (21b)) of estimators s and $s(\overline{x})$ is given by:

$$s_r(s) = s_r(s(\overline{x})) \cong (2\nu)^{-1/2}. \tag{4}$$

It depends on a parameter $\nu = n - 1$ named the degrees of freedom.

Article history: received on Oct. 15, 2009; received in revised form on Feb. 5, 2010; accepted on Feb. 7, 2010, available online on Feb. 26, 2010; DOI: 10.2478/v10178-010-0001-0.

The estimators (1–3) are unbiased and have the smallest variance assuming that the observations x_i are equivalent, mutually uncorrelated and normally distributed. The adjective "equivalent" means that successive observations have the same statistical properties. In particular they are characterized by the same expected value μ and variance σ^2.

Let us suppose that the observations x_i become autocorrelated, whereas other assumptions mentioned above remain the same. The arithmetic mean $\bar{x}$ remains an unbiased estimator because the expected value of a sum of random variables does not depend on their mutual correlations. This statement is not true for variances. As a result, estimator of the variance defined by Eq. (2) becomes only asymptotically unbiased (namely, understated for small n), whereas the formulae for estimators of the variance of the mean (Eq. (3)) and the dispersion of standard deviation (Eq. (4)) are no longer valid.

This work was prompted by the recent papers of Zhang [2], Dorozhovets and Warsza [3], and Witt [4] aiming to introduce an algorithm representing an extension of the GUM formulation for the case of autocorrelated observations. A subsequent search has shown that similar issues were independently worked out since at least 1935 in various papers usually concerned with areas of science where autocorrelated observations are common (geophysics, meteorology, acoustics, electric signals, *etc*.). This knowledge was, however, only scarcely covered in monographs and is absent in older handbooks on general data analysis. The problem of autocorrelated observations was mentioned by GUM ([1], Section 4.2.7) and certain recent handbooks on data analysis ([5], p. 111 and [6], p. 161), but without providing general methods for the processing of such data.

The objective of this work is to discuss the statistical procedures which represent equivalents of Eqs (2–4) for the case of n autocorrelated observations. It provides a critical synthesis of existing solutions, augmented with some new results. The nomenclature and notation used follows rather closely that of GUM.

The paper is organized as follows. The mathematical description of autocorrelated observations is discussed briefly in Section 2. A presentation of the formalism starts in Section 3 by introducing the relationship between the variance σ^2 and variance of the mean $\sigma^2(\bar{x})$ considered as statistical parameters. This relation is expressed through the use of a quantity named effective number of observations n_{eff}, depending on the real number of observations n and elements ρ_k of the autocorrelation function. Specific estimators of the variance s_a^2 and variance of the mean $s_a^2(\bar{x})$ are introduced in Section 4. It is proved that they are unbiased assuming that the autocorrelation function $\{\rho_k\}$ is known. Statistical properties of both estimators of variance are described in Section 5 using a parameter called effective degree of freedom ν_{eff}. Finally, an application of the discussed formalism to real data and possibilities of its extension are briefly discussed in Sections 6 and 7.

2. Mathematical models for autocorrelated observations

A set of n observations $\{x_i\}$ which are equivalent and autocorrelated represents an experimental reality which can be described by various mathematical models.

2.1. Multidimensional random variable

An elementary approach considers the set $\{x_i\}$ as a particular realization of the multi-dimensional random variable $(X_1, X_2, \ldots, X_n)$. Assuming that its probability distribution function represents an n-dimensional normal distribution, its stochastic properties are fully specified by expected values, variances and correlation coefficients.

Equivalence of the components X_i in the presence of correlations implies that their statistical properties do not depend on the shift of the index i. In the result, the expected

values μ and variances σ^2 are the same for all components, whereas the correlation coefficients relating X_i and X_j depend only on the difference $|i-j|$. Hence, a correlation matrix takes the following form:

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-2} \\ \rho_2 & \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-3} \\ \rho_3 & \rho_2 & \rho_1 & 1 & \cdots & \rho_{n-4} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \rho_{n-4} & \cdots & 1 \end{bmatrix} \tag{5}$$

with only n different coefficients. The information contained in matrix (5) can be conveniently presented as a one-dimensional discrete autocorrelation function $\{\rho_k\}$ of an integer argument $k = 0, 1, ..., n-1$ called lag.

2.2. Stationary time series

The approach outlined above represents a minimal extension of the standard formalism for uncorrelated variables and is sufficient to derive the general results discussed in this work. Nevertheless, the majority of handbooks and papers consider a set $\{x_i\}$ of autocorrelated observations as an n-element sample taken from a stationary time series.

The adjective "stationary" means that the probabilistic structure of a time series is invariant under a shift of an index i (representing a discrete time). It leads to the same consequences as a previous assumption of equivalence of components of multidimensional random variable, *i.e.*, it ensures that both expected value and variance exist and the autocorrelation matrix has the structure given by Eq. (5) ([7], p. 106, [8], pp. 23–29). Both approaches lead to the same results for estimators of variance because they are derived from the same autocorrelation matrix.

Time series theory is really useful when one can assume that the data are described by a specified time series model. Moving average (MA) and autoregressive (AR) processes are the two most simple categories of models of stationary time series ([7–10]). They are exemplified by two specific models defined below.

The simple moving average (SMA) is defined as the arithmetic mean:

$$x_i = \frac{u_i + u_{i-1} + \ldots + u_{i-m}}{m} \tag{6}$$

of m successive numbers from a set of uncorrelated random numbers $\{u_i\}$. The values of x_i are autocorrelated because each u_i is used to calculate m successive elements x_i. Fig. 1 shows a 60-element sample from the series $\{u_i\}$ representing a standard normal distribution ($\mu = 0$, $\sigma = 1$) and the autocorrelated series $\{x_i\}$ calculated using Eq. (6) with $m = 5$. Note that changes of sign of $\{x_i\}$ are less frequent than those for uncorrelated numbers $\{u_i\}$.

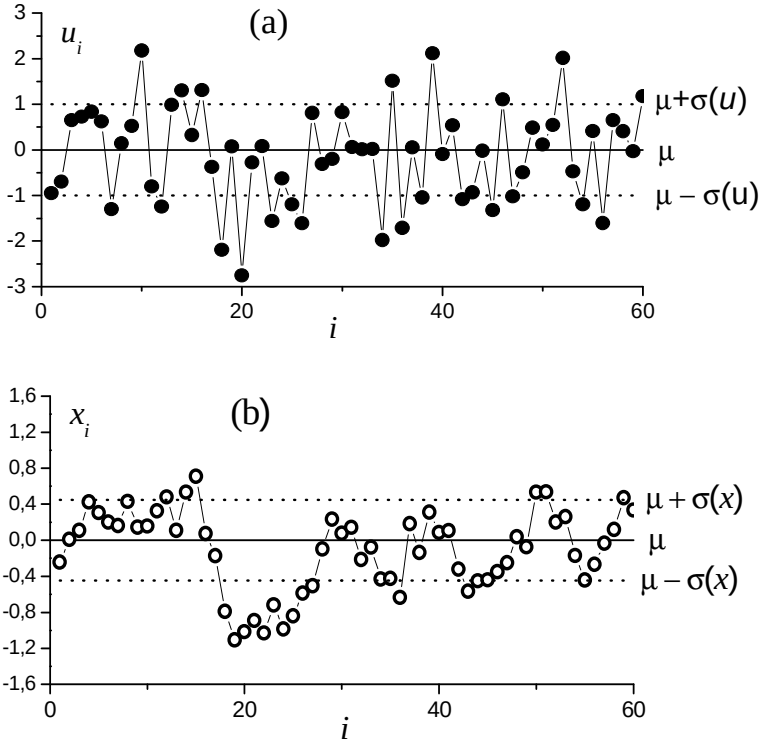

Fig. 1. A comparison of: a) uncorrelated random numbers $\{u_i\}$ with Gaussian distribution, $\mu = 0$, $\sigma = 1$; and b) simple moving average from five elements of the series $\{u_i\}$.

The most common model of an autocorrelated time series is, however, the first-order autoregressive model labeled AR(1) ([7–10]) or, sometimes, the Markov chain. Its successive elements are defined as a weighted mean of the previous element x_{i-1} and of the random number u_i:

$$x_i = a\, x_{i-1} + u_i \tag{7}$$

with the parameter $|a| < 1$. (Note that Eqs (6) and (7) define the time series with zero expected value assuming that $\mu = 0$ for $\{u_i\}$. This convention is common in literature. To obtain $\{x_i\}$ with a nonzero expected value one should add a nonzero constant to Eqs (6) and (7).)

For a given time series model one can calculate the autocorrelation function. For a simple moving average (Eq. (6)) one obtains:

$$\begin{cases} \rho_k = 1 - k/m & \text{dla} \quad k < m \\ \rho_k = 0 & \text{dla} \quad k \geq m. \end{cases} \tag{8}$$

For the first-order autoregressive model:

$$\rho_k = a^k. \tag{9a}$$

Example graphs of both $\{\rho_k\}$ functions are shown in Fig. 2. They exemplify two properties occurring for the vast majority of stationary autocorrelated processes occurring in nature. First, correlations are positive, *i.e.*, all coefficients $\rho_k \geq 0$. Second, correlations are characterized by a finite range: for a sufficiently large k all $\rho_k \approx 0$.

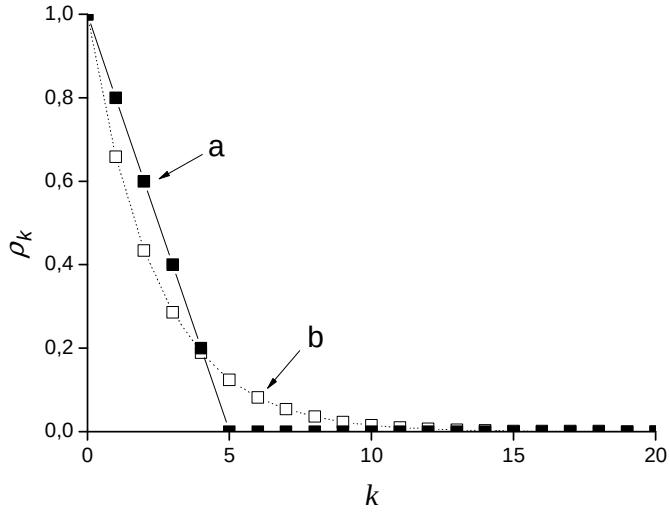

Fig. 2. Autocorrelation functions for two time series models: a) SMA, m = 5 and b) AR(1) with a = 0.659.

2.3. Relation to the stationary stochastic process

A discrete time series $\{x_i\}$ can be considered as the result of sampling a continuous stochastic process $x(t)$ [11, 12] at equal intervals of time Δt. Such a source of autocorrelated data is very often in real word. The sampling is usually realized by using a computerized data acquisition system.

The most common is an AR(1) stochastic process characterized by a continuous autocorrelation function:

$$\rho(\tau) = \exp(-\tau/T_0) \tag{9b}$$

with variable τ representing a continuous counterpart of the discrete lag k. The time constant T_0 is the measure of the correlation range. It is related to the parameter a of the corresponding AR(1) time series (Eq. (9)) by the equation $a = \exp((-\Delta t/T_0)$.

Filtering the thermal noise generated by a resistor through a low-pass RC filter represents an almost ideal physical realisation of an AR(1) stochastic process [4].

3. Variance of the mean. Effective number of observations

3.1. Variance of the mean of autocorrelated observations

For an autocorrelated data set $\{x_i\}$ the relation between variance σ^2 and variance of the mean $\sigma^2(\bar{x})$ is no longer given by Eq. (3). The correct formula reads:

$$\sigma^2(\bar{x}) = \left[n + 2\sum_{k=1}^{n-1} (n-k)\,\rho_k \right] \frac{\sigma^2}{n^2}. \tag{10}$$

The expression inside the square brackets of Eq. (10) represents the sum of all elements of the autocorrelation matrix (5).

This result can be derived using the theorem of variance of linear combination of correlated variables because the arithmetic mean (Eq. (1)) can be considered as a linear combination of x_i with proportionality coefficients $1/n$. (This theorem forms the basis of the law of propagation of uncertainty for correlated input variables ([1], Eq. (16)). Eq. (10) is

present in numerous papers from as early as 1935 [13], as well as in monographs on time series ([7], Eq. (5.3.5), [8], p. 30).

3.2. Effective number of observations

It follows from the structure of Eq. (10) that the effect of many coefficients ρ_k may be accounted for by introducing a single parameter. It can be rewritten as:

$$\sigma^2(\overline{x}) = \frac{\sigma^2}{n_{eff}}, \tag{11}$$

where the effective number of observations n_{eff} is given by:

$$n_{eff} = \frac{n}{1 + 2\sum_{k=1}^{n-1} \frac{n-k}{n} \rho_k}. \tag{12}$$

The concept of the effective number of observations was introduced by different authors, largely independently and in various ways. Many of them were concerned with meteorology because in this field the time-dependent observations are notorious for being autocorrelated. Not surprisingly, n_{eff} appears in the literature under similar but not identical names: equivalent number of subsequent coordinates, reduced number of coordinates [13], effective number of independent observations ([14, 12] p. 222), equivalent number of independent data [15], equivalent number of uncorrelated samples [16], effective independent sample size [17], effective sample size [18], equivalent independent process effective number [19], equivalent number of independent observations ([7] p. 320, [20]), effective number of uncorrelated observations [3]. The term adopted in this work follows the nomenclature of GUM, in which a (repeated) measurement is composed of observations, and the term effective is preferred over "equivalent".

A closely related parameter:

$$r = \frac{n}{n_{eff}} = 1 + 2\sum_{k=1}^{n-1} \frac{n-k}{n} \rho_k \tag{13}$$

has been used by a few authors [13, 2, 4]. The advantage of the term "effective numbers of observations" is that its name provides a heuristic understanding of the fact that the statistical properties of autocorrelated data are similar to a suitably defined number of independent observations. As will be shown in Section 3.3, n_{eff} also remains a well defined quantity for a continuous stochastic process, for which the ratio $r = n/n_{eff}$ becomes meaningless.

The properties of n_{eff}, considered as a fixed parameter defined using an *a priori* known autocorrelation function $\{\rho_k\}$, are as follows:

- effective number of observations is a real number within the interval [1, ∞). Its lower limit $n_{eff} \geq 1$ results from the fact that the values of ρ_k cannot be larger than unity. The upper limit $n_{eff} < \infty$ arises because the correlation matrix is positive-definite;
- for uncorrelated variables $n_{eff} = n$;
- for positive correlation coefficients ($\rho_k > 0$) $n_{eff} < n$;
- in the limit of strong positive correlations (all $\rho_k \to 1$) n_{eff} approaches unity. This means that such an n-element sample represents effectively a single observation;
- the case $n_{eff} > n$ may occur when some autocorrelation coefficients are negative.

3.3. Effective number of observations for specific models of time series and for continuous stochastic process

Assuming that the range of correlation is much smaller that the length of the sample n the factor $(n - k)/n$ in Eq. (12) may be approximated as unity. This leads to the approximate formula [21]:

$$n_{eff} \cong \frac{n}{1 + 2\sum_{k=1}^{n-1} \rho_k}. \tag{14}$$

Eq. (14) can be used as a starting point to derive simple formulae for n_{eff} for specific time series models. For the AR(1) model the sum in Eq. (14) can be approximated as the sum of an infinite geometrical series, $\Sigma a^k = a/(1 - a)$. The resulting expression for AR(1):

$$n_{eff} \cong n\,\frac{1-a}{1+a} \tag{15}$$

was derived in a different way by Priestley ([7], p. 320). For the SMA model one obtains:

$$n_{eff} \cong \frac{n}{m}. \tag{16}$$

An important property of n_{eff} concerns autocorrelated data obtained as a result of sampling a continuous stochastic process at equal intervals of time Δt. Let us assume that the total measurement time T is fixed and Δt tends to zero, thus the number of observations $n = T/\Delta t$ increases without limit. However, because ever denser sampling cannot provide new information able to reduce the variance of the mean, the effective number of observations n_{eff} remains a finite number. This general property of n_{eff} is illustrated for the AR(1) model by the function:

$$n_{eff} \cong n \tanh\frac{T}{2nT_0} \tag{17}$$

obtained by inserting $a = \exp(-\Delta t/T_0)$ with $\Delta t = T/n$ into Eq. (15)). The evolution of n_{eff} as a function of n is shown in Fig. 3.

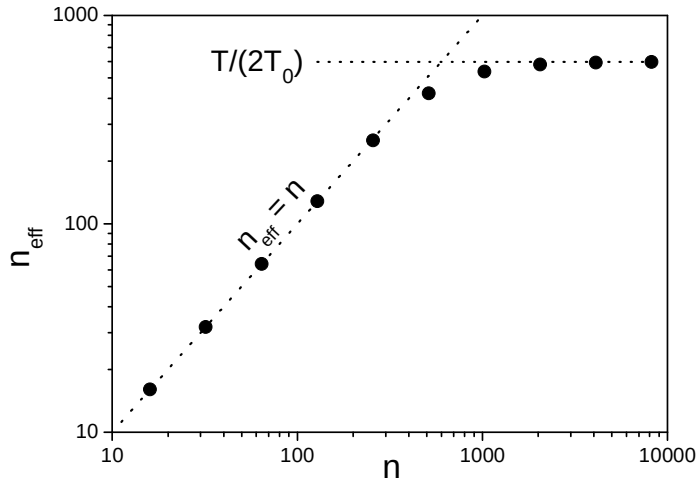

Fig. 3. The function n_{eff} vs. n for the AR(1) model. The parameters of the plot, T_0 = 239 ms and T = 286 s, are adopted from Ref. [4].

With increasing n, the effective number of observations approaches a value defined by:

$$n_{eff}(n \to \infty) = T/(2T_0). \tag{18}$$

By taking the limit $\Delta t \to 0$ in Eq. (12) one obtains an integral expression:

$$n_{eff} = \frac{T}{2\int_0^T [1-\tau/T]\rho(\tau)d\tau}, \tag{19}$$

which defines n_{eff} for the stationary stochastic processes characterized by a continuous autocorrelation function $\rho(\tau)$. Eq. (19) was used to derive the formula for n_{eff} for a continuous AR(1) process [17, 18].

4. Unbiased estimators of variance for autocorrelated observations

The approximate value of a given statistical parameter can be calculated from a finite sample $\{x_i\}$ using various estimators. We would like, however, to use estimators, which are unbiased, efficient, and expressed by formulae without numerical coefficients. The last property is rarely explicitly formulated but often applied. Estimator of variance s^2 defined by (Eq. (2) (alternatively named the sample variance) is widely accepted because it is better to use an unbiased estimator rather than the biased one:

$$s_b^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} \tag{20}$$

(known as the population variance), in a situation when each of them is expressed by a simple formula without numerical coefficients. The bias and dispersion of estimators considered in this work can be characterized by dimensionless quantities:

– relative bias:

$$bias_r(e) = \frac{\mathrm{E}(e) - \varepsilon}{\varepsilon}, \tag{21a}$$

– relative dispersion:

$$s_r(e) = \frac{\sqrt{\mathrm{Var(e)}}}{\varepsilon}. \tag{21b}$$

Symbols ε and e denote statistical parameter and corresponding estimator, respectively, and E and Var are standard statistical symbols for expected value and variance.

The well-known derivation of the unbiased estimator of variance for an uncorrelated variable will be adopted here for the case of autocorrelated observations. It starts by defining an estimator s_b^2 (Eq. (20)). To check whether it is unbiased or biased one should calculate its expected value $\mathrm{E}[s_b^2]$ and compare it to σ^2.

Firstly, the use of algebraic manipulations (see, *e.g.*, [11]) allows Eq. (20) to be transformed as follows:

$$s_b^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 - (\bar{x} - \mu)^2.$$

One may calculate now the expected value:

$$\mathrm{E}\left[s_b^2\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2 - (\bar{x}-\mu)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[(x_i-\bar{x})^2\right] - \mathrm{E}\left[(\bar{x}-\mu)^2\right].$$

From the very definition of variance the first term of the obtained expression represents the variance of the variable x_i, and the second the variance of the mean:

$$\mathrm{E}[s_b^2] = \sigma^2 - \sigma^2(\bar{x}). \tag{22}$$

Eq. (22) shows that the estimator s_b^2 is biased because its expected value is smaller than σ^2. For uncorrelated observations $\sigma^2(\bar{x}) = \sigma^2/n$, hence $\mathrm{E}[s_b^2] = (1-1/n)\sigma^2$. To compensate for this bias, Eq. (20) should be multiplied by a factor $n/(n-1)$. In this way one obtains the well-known formula (2).

What is to be changed in this derivation for the case of autocorrelated observations? Eq. (22) also holds for such a case because (i) calculation of the expected value of the sum of random variables does not depend on the presence of correlations, and (ii) in the derivation leading to Eq. (22) the relation $\sigma^2(\bar{x}) = \sigma^2/n$, specific to uncorrelated observations, was not used.

The remaining part of derivation should, however, be modified because the variance of the mean for an autocorrelated variable is now given by Eq. (11). Hence, Eq. (22) reads:

$$\mathrm{E}[s_b^2] = \sigma^2 - \sigma^2/n_{eff}. \tag{23}$$

The correction factor, which is to be applied to transform the biased estimator (Eq. (20)) into an unbiased one equals $n_{eff}/(n_{eff}-1)$. The resulting formula for the estimator of variance for autocorrelated data reads:

$$s_a^2 = \frac{n_{eff}}{n(n_{eff}-1)}\sum_{i=1}^{n}(x_i-\bar{x})^2. \tag{24a}$$

(The index *a* is used to distinguish it from other estimators of variance). It can alternatively be written as:

$$s_a^2 = C\,s^2 \ \text{with}\, C = \frac{n_{eff}(n-1)}{n(n_{eff}-1)}, \tag{24b}$$

i.e., as a product of the sample variance for uncorrelated data s^2 (Eq. (2)) and the correction factor $C = n_{eff}(n-1)/[n(n_{eff}-1)]$. An example graph of the dependence C *vs.* n is shown in Fig. 4.

The variance of the mean is n_{eff} times smaller (Eq. (10)). Hence, the estimator of variance of the mean is given by the formula:

$$s_a^2(\bar{x}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n(n_{eff}-1)}. \tag{25}$$

Both s_a^2 and $s_a^2(\bar{x})$ can alternatively be expressed by the parameter r (Eq. (13)) or by using elements ρ_k of the autocorrelation function.

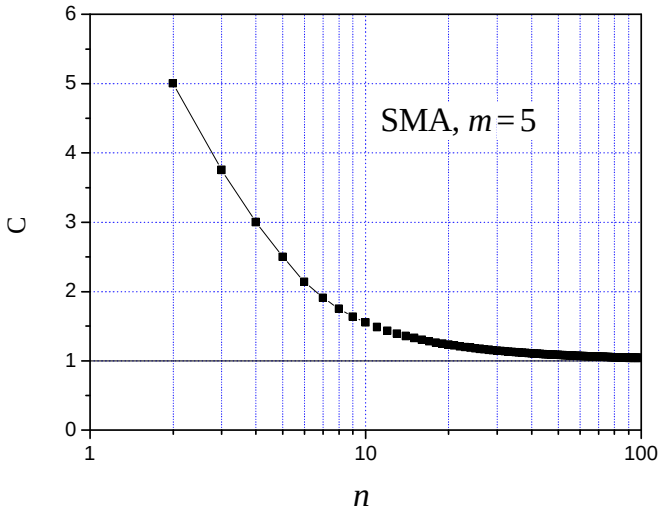

Fig. 4. The factor $C = n_{eff}(n-1)/[n(n_{eff}-1]$ as a function of the sample size n for the SMA model, $m = 5$.

Estimators s_a^2 and $s_a^2(\bar{x})$ are unbiased assuming that the effective number of observations is calculated (via Eq. (12)) from the autocorrelation function $\{\rho_k\}$. This property does not depend on the probability distribution $g(x)$ because this function is not involved in the presented derivation. The only assumption which is made is that both μ and σ^2 *exist* for the given $g(x)$.

The above derivation was given by the author [22]. A subsequent search has shown that Eq. (24b) was published by Bayley and Hammersley in 1946 ([14], Eq. (10)). Later on Anderson calculated the bias $\mathrm{E}[s_b^2]$ for an autocorrelated sample ([10], Eq. (52) on p. 448). Using this result, the formulae equivalent to Eqs. (24) and (25) were given by Law and Kelton [23] and in Wikipedia [24]. Independently, Şen [19] calculated s_a^2 for a special case of the ARIMA(1,0,1) time series model. In these works the derivation was not provided and it seems that the existence of unbiased estimators of variance for autocorrelated observations is rather not noticed in the literature. Instead, it is common to combine Eqs (2) and (11) which leads to the estimator:

$$s^2(\bar{x}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n_{eff}\,(n-1)} \tag{26}$$

which is biased (note the difference with respect to Eq. (25)). Relative bias (defined by Eq. (21a)), which is removed by the introduction of unbiased estimators s_a^2 and $s_a^2(\bar{x})$ (when compared to corresponding estimators given by Eq. (2) and Eq. (26)), equals approximately $-1/n_{eff}$. This is similar to the relation between estimators s_b^2 (Eq. (21)) and s^2 (Eq. (2)) for uncorrelated data: the use of the latter removes the relative bias $-1/n$.

From Eq. (23) it follows that the estimator s_a^2 is consistent, *i.e.*, it converges to the value of the parameter σ^2 for $n \to \infty$, but only when the limit $n_{eff} \to \infty$ is simultaneously fulfilled. For autocorrelated data resulting from the sampling of a continuous stochastic process this limit can be accomplished by increasing the total time of measurement. A mere increase of the number of samples with the measurement time kept constant is not sufficient.

5. Dispersion of the estimators of variance. Effective degrees of freedom

Relative dispersion of the estimator of variance for uncorrelated data (Eq. (4)) is a consequence of the formula:

$$\mathrm{Var}(s^2/\sigma^2) = 2/\nu. \tag{27}$$

It is independent of the probability distribution of the variable considered.

Eq. (27) is not valid for autocorrelated data. However, an equivalent formula to Eq. (27) can be formally written as:

$$\mathrm{Var}(s_a^2/\sigma^2) = 2/\nu_{eff}, \tag{28}$$

i.e., with the parameter ν replaced by the effective degrees of freedom ν_{eff}. Alternatively, Eq. (28) can be expressed as $\mathrm{Var}(s_a^2/\sigma^2) = 2/(n_{eff}^* - 1)$, *i.e.*, by using another "effective number" n_{eff}^* (Refs [7, 14], and [20]). The term "effective degrees of freedom" (introduced by Taubenheim [18]) seems to be better because it reflects the meaning of this quantity and follows the nomenclature of GUM. (Another type of effective degrees of freedom is used in the Welch-Satterthwaite formula, see [1], pt. G.4.1.) Any confusion with n_{eff} is now avoided. Note also that $\nu_{eff} \neq n_{eff} - 1$!

An estimator of the variance of the mean $s_a^2(\bar{x})$ is obtained from s_a^2 by dividing it by n_{eff} (representing a fixed number). Hence, the statistical properties of s_a^2 and $s_a^2(\bar{x})$ are the same. In particular, Eq. (28) remains valid also for $s_a^2(\bar{x})/\sigma^2(\bar{x})$.

A rather complex exact formula for ν_{eff} for the unbiased estimator s_a^2 was given by Bayley & Hammersley ([14]):

$$\nu_{eff} + 1 = \frac{n^3(n-1) - 4n^2\Sigma_1 + 2\Sigma_2 + 8\Sigma_5 - 4n\Sigma_6 - 8n\Sigma_7}{n^2(n-1) - 4n\Sigma_1 + 2\Sigma_3 - 8\Sigma_4 - 4n\Sigma_6 - 8n\Sigma_7}, \tag{29}$$

where:

- $\Sigma_1 = \sum_{k=1}^{n-1} (n-k)\,\rho_j$;
- $\Sigma_2 = \sum_{k=1}^{n-1} (n-k)\,(n^2 + 2n - 4k)\rho_k^2$;
- $\Sigma_3 = \sum_{k=1}^{n-1} (n-k)\,(n^2 - 2k)\rho_k^2$;
- $\Sigma_4 = \sum_{j=2}^{n-1} \sum_{k=1}^{j-1} (n-k)\,k\,\rho_j\,\rho_k$;
- $\Sigma_5 = \sum_{j=2}^{n-1} \sum_{k=1}^{j-1} (n-j)(n-k)\,\rho_j\,\rho_k$;
- $\Sigma_6 = \sum_{k=1}^{[(n-1)/2]} (n-2k)\,(n^2 - 2k)\rho_k^2$;
- $\Sigma_7 = \sum_{j=2}^{n-1} \sum_{k=1}^{K} (n-j-k)\,\rho_j\,\rho_k$.

The symbol $[(n-1)/2]$ denotes the largest possible integer not greater than $(n-1)/2$, and K is equal to $n-j-1$ or $j-1$ whichever is less [14].

An approximate expression for the effective degrees of freedom can be obtained by retaining the highest order terms with respect to n, *i.e.*, $\propto n^4$ in the numerator, and $\propto n^3$ in the denominator of Eq. (29). In this way one obtains an approximate formula [25]:

$$\nu_{eff} \cong \frac{n}{1+2\sum_{k=1}^{n-1}\rho_k^2} - 1 \tag{30}$$

which is sufficient for practical applications. It follows from Eq. (30) that ν_{eff} is less than $n-1$ for both positive and negative correlations. This fact reflects an intuitive understanding that any additional constraints, both deterministic and statistical (correlations) result in a decrease in the degree of freedom.

The expressions for ν_{eff} can be derived for specific models of time series. For the AR(1) model, inserting the sum $\Sigma(a^k)^2 \cong a^2/(1-a^2)$ into Eq. (30) leads to a formula:

$$\nu_{eff} \cong n\,\frac{1-a^2}{1+a^2} - 1, \tag{31}$$

derived in a different way by Bartlett ([26], Eq. (2)) and Priestley ([7], Eq. (5.3.30)).

6. Application of the formalism to experimental data

It has been assumed so far that the autocorrelation function $\{\rho_k\}$ is known *a priori*. This situation occurs for artificial data generated using a given time-series model. The presented theory should, however, be applicable also to the processing of real autocorrelated data. This requires the knowledge of $\{\rho_k\}$ which can be gained, respectively, from a statistical analysis of the investigated set of data, or by making use of other available information.

The autocorrelation function $\{\rho_k\}$ can be known exactly when primarily uncorrelated data are smoothed using the moving average or other well-defined procedure. The same applies to numerical differentiation of uncorrelated data. (For the two-point numerical derivative $\rho_1 = -1/2$ and the subsequent coefficients equal zero. From Eq. (14) one obtains $n_{eff} \cong n^2$, *i.e.*, it is larger than n.)

The autocorrelation function can sometimes be derived for electronic circuits with known characteristics. It can also be determined to a high degree of accuracy when a large amount of data is available. Examples of autocorrelation functions for autocorrelated noise and for meteorological data are given, respectively, in Refs. [4] and [17].

The possibilities mentioned above can be labeled, using the vocabulary of GUM, as type-B methods. The type-A methods aim to determine $\{\rho_k\}$, n_{eff} and ν_{eff} from the investigated set of observations $\{x_i\}$. This issue will be the subject of a forthcoming paper [27].

7. Conclusions

Guide to the Expression of Uncertainty in Measurement [1] defines type-A standard uncertainty as the square root of the unbiased estimator of variance of the mean. Estimators of variance s_a^2 and variance of the mean $s_a^2(\bar{x})$ for autocorrelated variables presented in this work are unbiased assuming that the autocorrelation function is known. Hence the square root of $s_a^2(\bar{x})$ represents type-A uncertainty for autocorrelated observations: $u(x) \equiv [s_a^2(\bar{x})]^{1/2}$.

In general, the formalism discussed in this work can be used to extend the application of GUM to the case of equivalent autocorrelated observations.

The approach presented in this work is based on an investigation of the properties of estimators. It is equivalent to the formalism of the least-squares method for correlated entry data. An application of this formalism for the case of weighted mean was recently discussed by Cox *et al*. [28].

Several aspects of this work are open to further studies. One is the calculation of expanded uncertainty and the use of statistical tests in the case of autocorrelated observations [18, 29]. The presented ideas can be extended to the problem of fitting straight line and other functions. The concept of n_{eff} can be generalized to the case of an autocorrelated time-varying field [30]. Finally, the applicability of the presented formalism to various experimental situations should be tested.

The processing discussed here depends on the assumption that the statistical parameters μ, σ and $\{\rho_k\}$ exist and do not depend on the sample size n. This assumption is not fulfilled for nonstationary stochastic processes like the random walk (in theory) or the difference of time measured by two atomic clocks (in experiment). The use of other statistical tools, such as the Allan variance, is necessary in such cases [31]. However, when the investigated autocorrelated process has well-defined classical statistical parameters, the presented approach is simpler and more adequate.

Acknowledgment

I would like to thank Z. Warsza for inspiration and providing a variety of information, to M.G. Cox for insisting on a more comprehensive literature search, and to K. Różański for critical reading of the manuscript.

References

[1] ISO/IEC. *Guide to the Expression of Uncertainty in Measurement* . (1995). Geneva.

[2] Zhang, N.F. (2006). Calculation of the uncertainty of the mean of autocorrelated measurements. *Metrologia*, 43, S276–S281.

[3] Dorozhovets, M., Warsza, Z.L. (2007). Upgrading calculating methods of the uncertainty of measurement results in practice. *Przegląd Elektrotechniczny*, 83, 1–13. (in Polish)

[4] Witt, T.J. (2007). Using the autocorrelation function to characterize time series of voltage measurements. *Metrologia*, 44, 201–209.

[5] Kirkup, L., Frenkel, B. (2006). *An Introduction to the Uncertainty in Measurement*. Cambridge: Cambridge University Press.

[6] Freund, R.J., Wilson, W.J., Sa, P. (2006). *Regression Analysis. Statistical Modeling of a Response Variable.* Amsterdam: Elsevier.

[7] Priestley, M.B. (1981). *Spectral Analysis and Time Series.* Amsterdam: Elsevier.

[8] Box, G.E.P., Jenkins, G.M., Reinsel, G.C. (1944). *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs.

[9] Brockwell, P.J., Davis, R.A. (1991). *Time series: theory and methods*. New York: Springer.

[10] Anderson, T.W.(1971). *The Statistical Analysis of Time Series*. New York: Wiley.

[11] Bendat, J.S., Piersol, A.G. (1971). *Random data: Analysis and measurement procedures.* New York: Wiley.

[12] Yaglom, A.M. (1987). *Correlation theory of stationary and related random processes*. Berlin: Springer.

[13] Bartels, J. (1935). Zur Morphologie geophysikalischer Zeitfunktionen. *Sitz.-Ber. Preuss. Akad. Wiss.*, 30, 504–522. (in German)

[14] Bayley, G.V., Hammersley, G.M. (1946). The Effective Number of Independent Observations in an Autocorrelat-ed Time-Series. *J. Roy. Stat. Soc. Suppl.*, 8, 184–197.

[15] Bagrov, N.A. (1969). On the equivalent number of independent data. *Tr. Gidrometeor. Cent.*, 44, 3–11. (in Russian).

[16] Lubman, D. (1969). Spatial Averaging in a Diffuse Sound Field. *J. Acoust. Soc. Am.*, 46, 532–534.

[17] Leith, C.E. (1973). The standard error of time-averaged estimates of climatic means. *J. Appl. Meteorol.*, 12, 1066–1069.

[18] Taubenheim, J. (1974). On the significance of the autocorrelation of statistical tests for averages, mean-square deviations and superposed epochs [geophysical data]. *Gerlands Beitr. Geophysik*, 83, 121–128. (in German)

[19] Şen, Z. (1998). Small sample estimation of the variance of time-averages in climatic time series. *Int. J. Climatol.*, 18, 1725–1732.

[20] Fortus, M.I., (1999). Equivalent Number of Independent Observations: A Review. *Izvestia AN. Fizika Atmosf. Okeana*, 35, 725–733. (in Russian)

[21] That useful approximate form of Eq. (12) was introduced by Bealey and Hammersley [14]. However, because of an error by a factor of two, approximate formulae for n_{eff} and ν_{eff} given at p. 185 of their paper are incorrect.

[22] Zięba, A. (2008). Uncertainty of the mean of correlated observations". *Podstawowe Problemy Metrologii, Conference materials*, Sucha Beskidzka, Poland, 15–24. (in Polish)

[23] Law, A.M., Kelton, W.D. (2000). *Simulation Modelling and Analysis*. New York: McGraw-Hill, 251–252.

[24] *http://en.wikipedia.org/wiki/Unbiased_estimation_of_standard_deviation.*

[25] Incorrect version of Eq. (31) is given in [14]. See remark [21].

[26] Barlett, M.S. On the theoretical specification and sampling properties of autocorrelated time-series. *J. Roy. Stat. Soc. Suppl.*, 8, 27–41.

[27] Zięba, A., Ramza, P. In preparation.

[28] Cox, M.G., Eiø, C., Mana, G., Pennecchi, F. (2006). The generalized weighted mean of correlated quantities. *Metrologia*, 43, S268–S275.

[29] Cliff, A.D., Ord, J.K. (1975). The comparison of means when samples consist of spatially autocorrelated observations. *Environment and Planning A*, 7, 725–734.

[30] Bretherton, C.S., Widmann, M., Dymnikov, V.P., Wallace, J.M., Bladé, I. (1999). The Effective Number of Spatial Degrees of Freedom of a Time-Varying Field. *J. Climate*, 12, 1990–2009.

[31] Zhang, N.F. (2008). Allan variance of time series models for measurement data. *Metrologia*, 45, 549–561.