GEOPHYSICAL TRANSACTIONS 1997
Vol. 40. No. 3–4. pp. 191–216

MFV-corrected variances

Ferenc STEINER*, Béla HAJAGOS*, Gábor HURSÁN*

The first part of the paper shows and quantitatively characterizes the probability distortions of the 'corrected empirical variance' (denoted by $\sigma^2_{emp,corr}$) which can be considerable for some, in practice often occurring, parent distribution types and sample sizes, respectively. Even for the pure Gaussian distribution (as the parent one) such probability distortions of the conventionally corrected empirical variance are present.

The second part of the paper deals with the proper correction method called 'MFV-correction' or 'M_k-correction' (M_k means the most frequent value, the actual values of k being 1/2; 1; 2 and 3). The correction constants $C(F,k;n)$ are tabulated for three types of parent distributions (see Table IIIa). Table IIIb gives the coefficient-triplets $A_{F,k}$, $B_{F,k}$ and $C_{F,k}$ to calculate with very good approximation $C(F,k;n)$ for an arbitrary value of $n < 1000$.

Eight tables (Table IVa to Table IVh) give detailed information about the most important probability characteristics of the 'MFV-corrected empirical variances' $\sigma^2_{M_k corr}$ as random variables. The values are convincing in demonstrating that if one single $\sigma^2_{M_k corr}$ is calculated (this is the ordinary case in practice), then with great probability this estimate is near to the true value of the variance.

In the overwhelming majority of the discussed cases $c(F,k;n) = C(F,k;n)(n-1)/n$ is significantly greater than unity, showing that though using the simple correction factor $n/(n-1)$, $E(\sigma^2_{emp,corr}) = \sigma^2$ really holds, this requirement does not, however, result in such a distribution of the estimates which is acceptable in practice; in asymmetrical cases the fulfilment of the equation '*E(estimate) = true value*' is rather misleading in respect of the occurrence probabilities of the bias. It seems better to speak about 'unbiased estimate' if the equation *MFV(estimate) = true value* holds.

Keywords: variance, estimates of the variance, corrected empirical variances, unbiased estimate, most frequent value

* University of Miskolc, H-3515 Miskolc-Egyetemváros

Manuscript received: 20 November, 1995

1. Introduction

In practice, based on the sample x_1, x_2, …, x_i, … ,x_n, generally the expression

$$\sigma^2_{emp} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \tag{1}$$

is used as an estimate of the true value of the variance (i.e., of the square of the scatter) defined by

$$VAR(\xi) \equiv \sigma^2 = \int_{-\infty}^{\infty}(x - E)^2 \cdot f(x); \tag{2}$$

where $\bar{x}$ is the arithmetic mean of the data x_i, ξ is the random variable from where the sample was taken, $f(x)$ is the probability density function of ξ, and E is its 'expected value':

$$E(\xi) = \int_{-\infty}^{\infty} x \cdot f(x)\mathrm{d}x\,. \tag{3}$$

$VAR(\xi)$ plays a fundamental role in classical geostatistics for calculating the variograms [see e.g. MATHERON 1965], and their square root, i.e., σ itself is widely used as an error-characteristic. The question arises as to whether or not Eq. (1) satisfactorily estimates the true value of variance σ^2 (and its square root the scatter σ)?

In every standard book of probability theory and mathematical statistics [see e.g. CRAMÉR 1945] it is proven that

$$E\left(\sigma^2_{emp}\right) \neq \sigma^2 \tag{4}$$

and therefore we speak about distortion. The correction is, however, very simple: in the same books it is proven that the 'corrected empirical variance' $\sigma^2_{emp,corr}$ defined by

$$\sigma^2_{emp,corr} = \frac{n}{n-1}\sigma^2_{emp} \tag{5}$$

fulfils the equation

$$E\left(\sigma^2_{emp,corr}\right) = \sigma^2 \tag{6}$$

and therefore $\sigma^2_{emp,corr}$ is usually known as an 'unbiased' estimate of σ^2, i.e., of the true value of the variance.

HAJAGOS and STEINER [1994], however, pointed out that even $\sigma^2_{emp,corr}$ shows distortions but of other type: these estimates are less (or even significantly less) with greater (eventually significantly greater) probability than 0.5 than the true value of σ^2. Such distortions may be called 'probability distortions' and are extremely important in the practice. A proposal for their quantitative characterization is given in STEINER [1991] (see page 247). More intensive investigations of these probability distortions are reported in KIS and HURSÁN [1995]. In the next point a fairly complete characterization of this type of distortion of the $\sigma^2_{emp,corr}$ estimates will be given for three parent distribution types: geostatistical, Jeffreys, and Gaussian (for the formulae of their probability density functions see in Appendix C/ of the present paper, where the distribution functions are also given). Comments are given on the choice of two parent distribution types: a/ DUTTER [1986/87] states that in geostatistic practice, geostatistical distribution generally serves as an appropriate model (this type is not only characteristic of geostatistics, see e.g. the practical: on one hand geodetical on the other hand astronomical examples given in HAJAGOS and STEINER [1994]); b/ after Jeffreys [cited by KERÉKFY 1978] shorter flanks than those of the Jeffreys-distribution never or very rarely occur in practice.

All investigations of the present paper were made with the Monte Carlo method characterized by the repetition number $N=10000$ for both of the above- mentioned parent distributions. Although the Gaussian case as a parent distribution is far from the realities of the geosciences, for comparison purposes this classical case is also treated in this paper. The Monte Carlo method is not needed for investigating the Gaussian case (only integrating or differentiating is neccessary) as the probability density function of $\sigma^2_{emp,corr}$ can be given in this case also analytically:

$$f\left(\sigma^2_{emp,corr}\right)=\frac{n-1}{\sigma^2}\frac{y^{\frac{n-1}{2}-1}\cdot e^{-\frac{y}{2}}}{2^{\frac{n-1}{2}}\cdot\Gamma\left(\frac{n-1}{2}\right)}, \tag{7}$$

where $y=\sigma^2_{emp,corr}$ [see e.g. CRAMÉR 1945].

2. Characterization of the probability distortions

Figure 1 shows the frequency function of $\sigma^2_{emp,corr}/\sigma^2$ for the sample size $n=9$ and for the geostatistical parent distribution. The already cited proposal of STEINER [1991] for measuring the probability distortion will be denoted by *pd*; the definition is very simple:

$$pd = \frac{P\left(\sigma^2_{emp,corr} < \sigma^2\right)}{P\left(\sigma^2_{emp,corr} > \sigma^2\right)} - 1 \,. \tag{8}$$

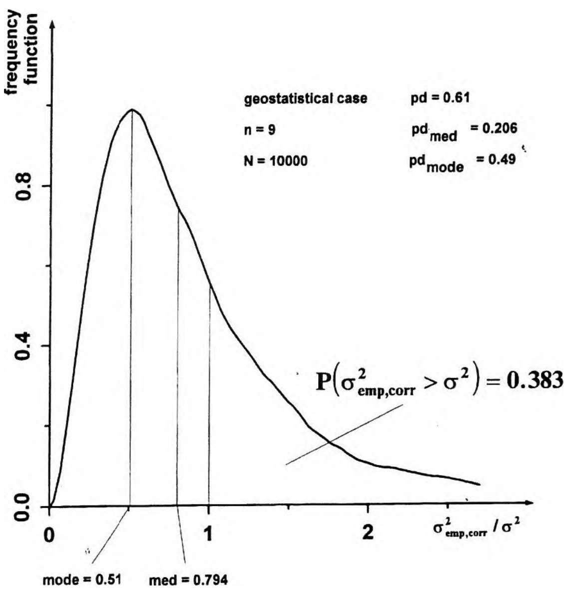


Fig. 1. Three alternatives for characterizing 'probability distortions' of the so-called 'corrected variance' ($\sigma^2_{emp,corr}$): the value of the mode and that of the median of the estimations are significantly less than the true value of the variance (denoted by σ^2), and the probability of the event $\sigma^2_{emp,corr} > \sigma^2$ is significantly less than 0.5. The parent probability distribution calculating this curve was of geostatistical type but these probability distortions occur also if the parent distribution is pure Gaussian, nevertheless in the latter case they are not so considerably (even shockingly) large as those presented here

1. ábra. A „korrigált empirikus variancia" ($\sigma^2_{emp,corr}$) háromféleképpen lehet 'valószínűségi értelemben torzított': a módusz és a medián értékei szignifikánsan kisebbek lehetnek a variancia elméletileg helyes (σ^2-tel jelölt) értékénél, valamint $\sigma^2_{emp,corr} > \sigma^2$ a reláció előfordulásának valószínűsége szignifikánsan kisebb lehet 1/2-nél. Az ábra esetében geostatisztikus típusú volt az anyaeloszlás, de ezek a „valószínűségi torzulások" fellépnek akkor is, ha steril Gauss típusú az anyaeloszlás, csak utóbbi esetben kisebb mértékűek ezek a torzulások az ábrán bemutatottaknál

The value of the denominator is given in Fig. 1. (=0.383) and the other probability value figuring in Eq. (8) is clearly 1-0.383=0.617; therefore *pd* equals 0.61.

It can be more informative in respect of the probability distortion to determine the difference between the median of the estimates (i.e., of the value $med(\sigma^2_{emp,corr})$) and the true σ^2 value. Denoting this difference by pd_{med}, this other characteristic of the probability distortion can be calculated simply as

$$pd_{med} = 1 - med\left(\sigma^2_{emp,corr}\right)/\sigma^2 . \tag{9}$$

The third characteristic of the probability distribution expresses our spontaneous wish: it would be desirable that the mode (maximum) of the $\sigma^2_{emp,corr}$-curve should coincide with the true value σ^2. In this respect it is understandable to define this third characteristic of the probability distortion as

$$pd_{mode} = 1 - mode\left(\sigma^2_{emp,corr}\right)/\sigma^2 . \tag{10}$$

Table I presents the values of all three probability distortions for the geostatistical, Jeffreys and Gaussian parent types — and for the sample sizes n=4; 9; 16; 25; 49; 100; 400; 900. *Figure 2* shows the results as curves; naturally all probability distortions increase with decreasing sample size. The distortion values themselves are considerable even if the sample sizes are not extremely small. Neglecting this distortion is eventually justified at the Gaussian but only at very great sample sizes.

3. Most frequent values (M_k-s) of the estimates $\sigma^2_{emp,corr}$

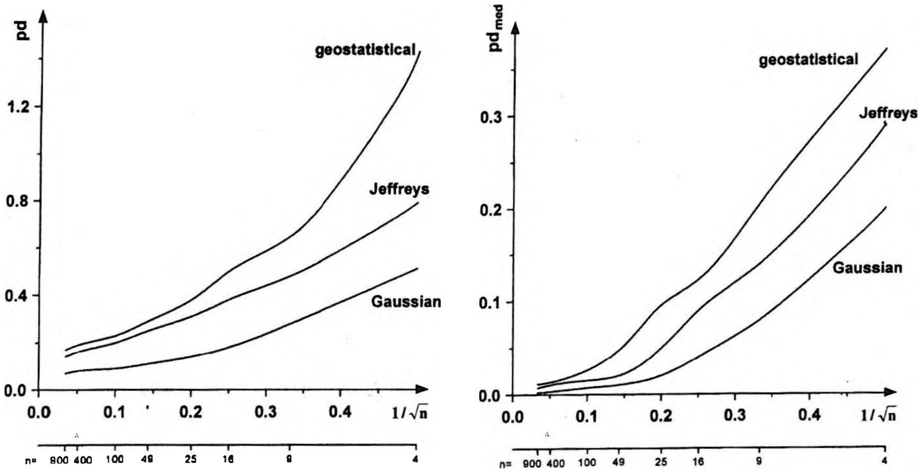
The formulae for determining the so-called 'most frequent values' (MFVs) denoted by M_k simultaneously with the corresponding dispersion-characteristics ε_k called dihesions (k=1/2; 1; 2 and 3) are given, for example, in STEINER [1991], and in STEINER [1989]. These formulae are reproduced in Appendix A/ both for samples and for given density functions.

Figure 3 shows the probability density curve for the same case as was dealt with in Fig. 1. In Fig. 3, however, besides the frequency function the M_k-value of the random variable for k=3 is also given as well as the interval ($M_k-\varepsilon_k$, $M_k+\varepsilon_k$) with the corresponding value of the probability of the event

parent distribution type	Values of the probability distortions for different parent distribution types and various sample sizes								
	characteristics of the probability distortions (see Eqs.8., 9. and 10)	size of the sample (n)							
		$n=4$	$n=9$	$n=16$	$n=25$	$n=49$	$n=100$	$n=400$	$n=900$
geostatistical	$pd=$	1.43	0.65	0.50	0.38	0.29	0.23	0.19	0.17
	$pd_{med}=$	0.371	0.206	0.123	0.096	0.048	0.026	0.013	0.011
	$pd_{mode}=$	0.80	0.49	0.24	0.16	0.12	0.07	0.03	0.015
Jeffreys	$pd=$	0.79	0.48	0.38	0.31	0.25	0.20	0.16	0.14
	$pd_{med}=$	0.29	0.14	0.09	0.05	0.02	0.015	0.01	0.007
	$pd_{mode}=$	0.78	0.34	0.21	0.14	0.09	0.04	0.02	0.01
Gaussian	$pd=$	0.51	0.28	0.18	0.14	0.11	0.09	0.08	0.07
	$pd_{med}=$	0.2	0.08	0.04	0.02	0.01	0.007	0.003	0.002
	$pd_{mode}=$	0.61	0.32	0.12	0.07	0.03	0.01	0.005	0

Table I. Probability distortions of the variances $\sigma^2_{emp,corr}$ (see Eq. 5)

I. táblázat. A $\sigma^2_{emp,corr}$ varianciák valószínűségi torzítottságai (ld. (5) egyenlet)



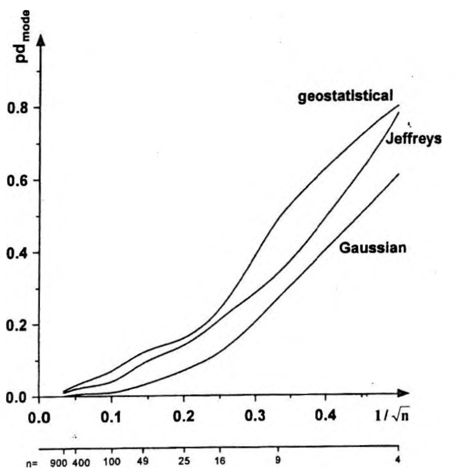


Fig. 2. Curves of the probability distortions pd, pd_{med} and pd_{mode} (see Eqs. 8, 9 and 10) versus $1/\sqrt{n}$ where n is the size of the sample. Note that at some frequently occurring parent distributions these distortions cannot be neglected even if the sample size is 400 or 900 (see also Table I)

2. ábra. A (8), (9) és (10) egyenletekkel definiált és pd-vel, pd_{med}-del valamint pd_{mode}-dal jelölt valószínűségi torzulások görbéi az $1/\sqrt{n}$ függvényében, ahol n a mintaelemszámot jelenti. Látható, hogy e torzulások viszonylag gyakran előforduló eloszlástípusokra még elég nagy n-eknél sem hanyagolhatók el: a pd esetében például még $n=400$, sőt $n=900$ esetén sem

$M_k-\varepsilon_k<\sigma^2_{emp,corr}/\sigma^2<M_k+\varepsilon_k$. It is by no means a surprise that the values of a random variable concentrare around a most frequent value (MFV), i.e. around $M_{1/2}$, M_1, M_2 or M_3 with different but considerable probability. In respect of the appropriate estimate of the true value of the variance (σ^2) naturally MFV = 1 would be desirable — but the estimate $\sigma^2_{emp,corr}/M_k$ trivially fulfils this demand (*Table II* gives the corresponding M_k-values for a great variety of cases). In *Fig. 4*, M_k means the most frequent value of this newly corrected estimate if the *E(estimate)*$\neq\sigma^2$ holds: according to the traditional definition of unbiasedness this new estimate should be considered as a biased one. However, Fig. 4 clearly shows that this once more corrected estimate

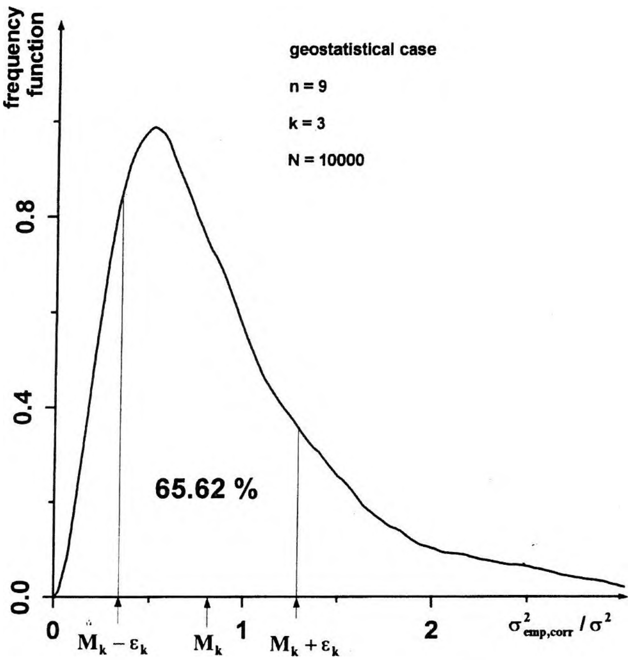


Fig. 3. For just the same case as Fig. 1, symmetrical interval of $2\varepsilon_3$ length (i.e., twice that of the dihesion) marked around the most frequent value M_3. The probability that $\sigma^2_{emp,corr}/\sigma^2$ occurs in this interval is also given in per cent

3. ábra. Az M_3 leggyakoribb érték körüli, $2\varepsilon_3$ (azaz kétszeres dihéziónyi) hosszúságú intervallum az 1. ábrán már megismert esetben; százalékosan adott annak a valószínűsége, hogy a $\sigma^2_{emp,corr}/\sigma^2$ arány ebbe az intervallumba esik

fully meets our practical demands: a considerable percentage of the estimates are concentrated around the true value. Consequently, it is indifferent in practice if our estimates are 'unbiased' in the conventional use of Eq. (8), or generally written: in all asymmetrical cases it is completely indifferent from the point of view of practice if

$$E(estimate) = true\ value \tag{6a}$$

holds or not. The classical notion of 'unbiased estimate' is of no use: it has no practical meaning according to our present investigations. It is logical to demand

$$MFV(estimate) = true\ value \tag{11}$$

and if for the estimate Eq. (11) is fulfilled, the estimate can be called 'unbiased' in the new sense and this adequately corresponds to the demands of the practice.

parent distribution type	M_k and ε_k value-pairs of $\sigma^2_{emp,corr}/\sigma^2$ for different parent distribution types and various sample sizes (see Eqs. A1,…,A4 in the Appendix)								
		size of the sample (n)							
		$n=4$	$n=9$	$n=16$	$n=25$	$n=49$	$n=100$	$n=400$	$n=900$
geostatistical	$M_{1/2}=$	0.433	0.633	0.852	0.882	0.950	0.965	0.986	0.988
	$\varepsilon_{1/2}=$	0.372	0.371	0.346	0.37	0.27	0.187	0.085	0.06
	$M_1=$	0.513	0.684	0.874	0.903	0.958	0.970	0.987	0.988
	$\varepsilon_1=$	0.428	0.398	0.353	0.304	0.236	0.170	0.085	0.06
	$M_2=$	0.631	0.758	0.889	0.918	0.964	0.974	0.987	0.988
	$\varepsilon_2=$	0.519	0.445	0.366	0.320	0.240	0.171	0.085	0.06
	$M_3=$	0.704	0.821	0.906	0.931	0.973	0.978	0.988	0.989
	$\varepsilon_3=$	0.638	0.476	0.375	0.329	0.243	0.172	0.085	0.06
Jeffreys	$M_{1/2}=$	0.568	0.798	0.874	0.909	0.965	0.973	0.988	0.992
	$\varepsilon_{1/2}=$	0.4787	0.424	0.343	0.289	0.222	0.158	0.082	0.054
	$M_1=$	0.663	0.842	0.898	0.928	0.977	0.978	0.989	0.993
	$\varepsilon_1=$	0.5404	0.443	0.350	0.294	0.225	0.159	0.082	0.054
	$M_2=$	0.781	0.900	0.933	0.952	0.990	0.983	0.991	0.993
	$\varepsilon_2=$	0.623	0.471	0.363	0.303	0.229	0.159	0.083	0.054
	$M_3=$	0.847	0.932	0.953	0.965	0.996	0.986	0.991	0.994
	$\varepsilon_3=$	0.673	0.489	0.372	0.308	0.231	0.160	0.083	0.054
Gaussian	$M_{1/2}=$	0.691	0.876	0.940	0.959	0.983	0.986	0.996	0.997
	$\varepsilon_{1/2}=$	0.550	0.428	0.323	0.258	0.184	0.132	0.064	0.043
	$M_1=$	0.769	0.908	0.954	0.968	0.986	0.988	0.997	0.998
	$\varepsilon_1=$	0.590	0.438	0.325	0.260	0.184	0.132	0.064	0.043
	$M_2=$	0.866	0.946	0.973	0.980	0.991	0.992	0.998	0.998
	$\varepsilon_2=$	0.646	0.451	0.329	0.261	0.184	0.132	0.064	0.043
	$M_3=$	0.866	0.965	0.983	0.987	0.994	0.994	0.998	0.998
	$\varepsilon_3=$	0.679	0.458	0.332	0.263	0.185	0.133	0.064	0.043

Table II. MFV (most frequent value) of the variances $\sigma^2_{emp,corr}$ (see Eq. 5)
II. táblázat. A $\sigma^2_{emp,corr}$ varianciák legyakoribb értékei (lásd (5) egyenlet)

4. MFV-corrected estimates of the variance ($\sigma^2_{M_k corr}$)

Empirical variances are defined both in Eq. (1) and Eq. (5); in the first case the estimate is denoted by σ^2_{emp}, in the second one by $\sigma^2_{emp,corr}$. To have M_k-corrected estimates, the new correction factors in both cases depend equally upon F (the parent distribution, characterized here by its distribution function), upon k (which was chosen from the values 1/2; 1; 2 and 3) and naturally from the sample size n. For the M_k-values that are known (see

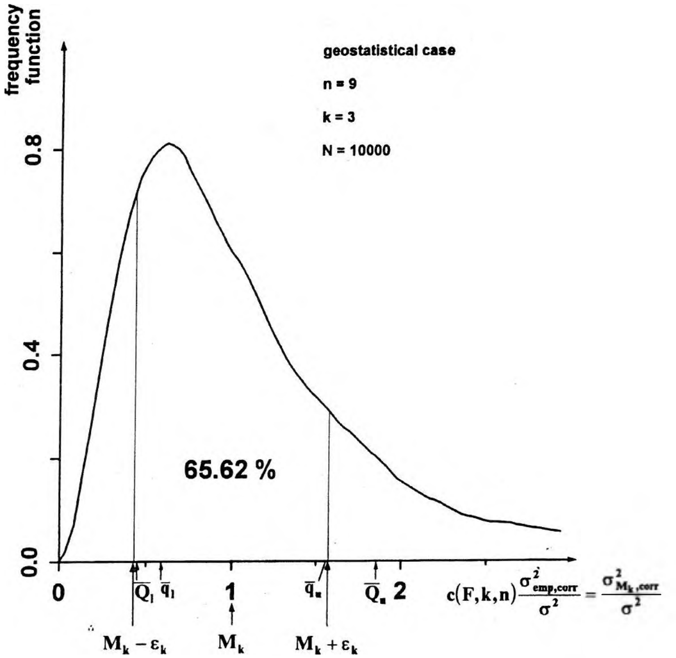


Fig. 4. MFV-corrected variances ($\sigma^2_{emp,cor}$) estimate the true value of the variance (σ^2) in the interval (M_k−ε_k, M_k+ε_k) with considerable probability

4. ábra. Az MFV-korrigált varianciák ($\sigma^2_{emp,cor}$) az (M_k−ε_k, M_k+ε_k) intervallumban jelentős valószínűséggel adnak becslést a variancia σ^2-tel jelölt, elméletileg helyes értékére

Table II), *Table IIIa* presents the coefficients of this new correction in the sense that *c* and *C* values are given for calculating in two ways:

$$\sigma^2_{M_k corr} = c(F,k;n) \cdot \sigma^2_{emp,corr} \tag{12}$$

or

$$\sigma^2_{M_k corr} = C(F,k;n) \cdot \sigma^2_{emp} . \tag{13}$$

To be able to execute this correction not only for the sample sizes n=4, 9, 16, 25, 49, 100, 400 or 900 given in Table IIIa, *Table IIIb* gives the coefficients of simple n-dependent polynomials both for *c(F,k;n)* and *C(F,k;n)*. Calculations for the coefficients of a simple approximation were made on the basis of the supposition that a polynomial of the form

$$c(F,k;n) = 1 + A_{F,k}/n^{1/2} + B_{F,k}/n + C_{F,k}/n^{3/2} \tag{14}$$

parent distribution type	k	sample sizes (n)							
		$n=4$	$n=9$	$n=16$	$n=25$	$n=49$	$n=100$	$n=400$	$n=900$
geostatistical	½	3.077	1.776	1.320	1.175	1.066	1.046	1.016	1.012
	1	2.596	1.643	1.218	1.153	1.062	1.041	1.015	1.012
	2	2.114	1.484	1.181	1.133	1.058	1.036	1.014	1.012
	3	1.893	1.370	1.177	1.117	1.049	1.032	1.013	1.011
Jeffreys	½	2.346	1.432	1.220	1.145	1.058	1.037	1.014	1.008
	1	2.010	1.356	1.188	1.122	1.045	1.033	1.013	1.008
	2	1.706	1.270	1.143	1.093	1.031	1.027	1.012	1.008
	3	1.573	1.226	1.119	1.078	1.023	1.020	1.011	1.008
Gaussian	½	1.928	1.303	1.134	1.086	1.037	1.024	1.006	1.003
	1	1.733	1.258	1.117	1.076	1.035	1.021	1.005	1.003
	2	1.539	1.208	1.096	1.062	1.021	1.018	1.005	1.003
	3	1.453	1.184	1.085	1.055	1.027	1.016	1.004	1.002

Table IIIa. Correction factors $C(F,k;n)$ to calculate MFV-corrected variances denoted by $\sigma^2_{M_k corr}$ as $C(F,k;n)\cdot\sigma^2_{emp}$

IIIa. táblázat. A $\sigma^2_{M_k corr}$-al jelölt MFV korrigált varianciák σ^2_{emp} alapján történő kiszámításához szükséges $C(F,k;n)$ korrekciós faktorok

is able to give satisfactory approximations of the n-dependency of the MFV-correction factor if $\sigma^2_{emp,corr}$ is to be corrected: that $A_{F,k}$, $B_{F,k}$, $C_{F,k}$ coefficient-triplets which are the best for the actual k-value and for the F parent distribution in question (in the sense of the L_2-norm,) can be easily calculated using the simple least squares techniques, based on the c-values given in Table IIIa.

Obviously the C correction constant, also defines the same value-triplet, as

$$C(F,k;n) = \frac{n}{n-1} c(F,k;n) \tag{15}$$

can easily be proven on the basis of Eqs. (1), (5) and (14). The $A_{F,k}$, $B_{F,k}$, $C_{F,k}$ value-triplets obtained are presented in Table IIIb, and *Fig. 5* shows the curves of the C-approximations (some numerical differences from the true C-values are given — in per cent — in the last column of Table IIIb).

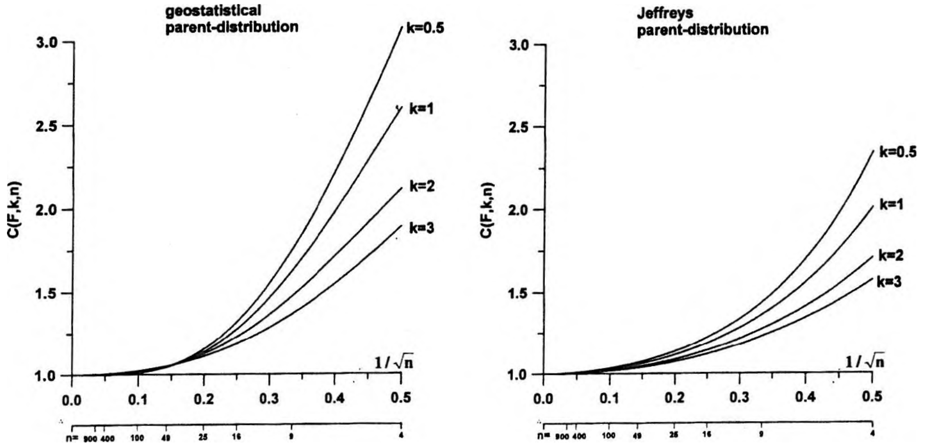
parent distribution type	k	$A_{F,k}$	$B_{F,k}$	$C_{F,k}$	maximum of the difference between the true value of C and its approximation in the sample size interval $4 \leq n \leq 900$
geostatistical	½	-0.75	40.86	-37.41	2.3
	1	-0.3	34.79	-35.89	1.9
	2	0.253	18.97	-21.12	1.7
	3	1.74	6.49	-5.68	1.1
Jeffreys	½	3.944	-6.157	18.07	1.2
	1	2.98	-2.08	8.41	1.3
	2	2.03	0.490	2.198	1.2
	3	1.586	1.247	0.334	1.2
Gaussian	½	1.491	1.966	4.958	0.8
	1	1.368	2.02	2.22	0.8
	2	1.03	2.79	-1.09	0.7
	3	0.83	3.17	-2.43	0.7

Table IIIb. $A_{F,k}$, $B_{F,k}$ and $C_{F,k}$ coefficients to have $C(F,k;n)$ for arbitrary n (between $4 \leq n \leq 900$) using Eqs. (14) and (15)

IIIb. táblázat. A $C(F,k;n)$ meghatározásához szükséges $A_{F,k}$, $B_{F,k}$, $C_{F,k}$ koefficiensek tetszőleges ($4 \leq n \leq 900$) n-ekre a (14) és (15) egyenletekben

5. Detailed numerical characterization of the random variable $\sigma^2_{M_k corr}$

The purpose of this point is to characterize in a sufficiently detailed way the measure of concentration of the estimates $\sigma^2_{M_k corr}$ around the true value of the variance (σ^2). For each studied sample size n (see *Table IVa* to *Table IVh*) the corresponding occurrence probabilities are given for 5 lengths of the interval around the correct σ^2-value . Also given are the $\overline{q}_l$ and $\overline{q}_u$ values defined by the following probabilities (denoted for short $\sigma^2_{M_k corr}/\sigma^2$ by y): $P(\overline{q}_l \leq y \leq 1) = 1/4$ and $P(1 \leq y \leq \overline{q}_u) = 1/4$. The differences $\overline{q}_u - \overline{q}_l$ are also given (being more or less independent of k). These latter intervals belonging to the probability of 50 % are enhanced with each density curve of *Figs. 6a, 6b.* and *6c.* (In Figs. 6a, 6b, and 6c, for the discussed three parent distributions and for all eight studied sample sizes the density curve for that k is shown where the difference $\overline{q}_u - \overline{q}_l$ is minimum, see the last column in each part of Table IV.)



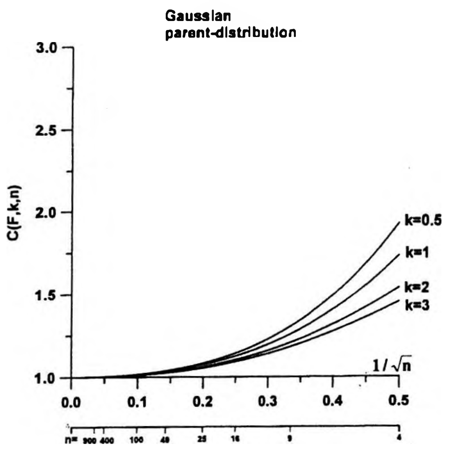


Fig. 5. Curves of coefficients $C(F,k;n)$ to calculate MFV-corrected variances in the following way: $\sigma^2_{M_k corr} = C(F,k;n) \cdot \sigma^2_{emp}$ (Table IIIa gives the correct values

5. ábra. A $\sigma^2_{M_k corr}$ értékek, azaz az MFV-korrigált varianciák számításához szükséges, $C(F,k;n)$-nel jelölt korrekciós tényezők görbéi; a korrekció $\sigma^2_{M_k corr} = C(F,k;n) \cdot \sigma^2_{emp}$ szerint történik. (A C pontos értékeire vonatkozóan ld. a IIIa. táblázatot.)

* * *

It is hoped that our results are found to be convincing: working on the basis of the MFV-concept, i.e., using the P_k-norms instead of the L_2 one, all conventional definitions and statements must be checked. This is very laborious work but unavoidable. It is also hoped, that by introducing the MFV-corrected estimate of the variance, the variograms in the classical geosciences and the scatters (as error-characteristics) in many fields of science and its applications can be determined much more reliably than with the conventional formulae.

Concentration of the MFV corrected variances around the true value of the variance for the sample size n=4							
parent distribution type	k	probability of the event $1-\Delta<\sigma^2_{M_k corr}<1+\Delta$					$\bar{q}_u-\bar{q}_l$
		Δ=0.2	Δ=0.4	Δ=0.6	Δ=0.8	Δ=1.0	
geostatis-tical	0.5	0.1385	0.2836	0.4247	0.5632	0.6584	1.5447
	1	0.1525	0.295	0.4547	0.6108	0.7218	1.4974
	2	0.1562	0.3101	0.4828	0.6571	0.7877	1.5411
	3	0.152	0.3181	0.4885	0.6737	0.8192	1.5931
Jeffreys	0.5	0.1453	0.3016	0.451	0.5973	0.6934	1.4432
	1	0.1562	0.3185	0.4805	0.6463	0.7545	1.3723
	2	0.1669	0.3295	0.5101	0.692	0.8119	1.3632
	3	0.1669	0.3361	0.5194	0.7046	0.8378	1.3969
Gaussian	0.5	0.1695	0.3319	0.4945	0.6424	0.7515	1.2704
	1	0.1768	0.3506	0.521	0.6805	0.7946	1.2191
	2	0.182	0.3684	0.5463	0.7159	0.8407	1.2096
	3	0.1853	0.3734	0.5525	0.7316	0.8579	1.2026

Table IVa. Concentration of the MFV-corrected empirical variances around the true value σ^2 for the sample size $n=4$

IVa. táblázat. Az MFV korrigált varianciáknak a σ^2 körüli koncentrációja $n=4$ esetén

Concentration of the MFV corrected variances around the true value of the variance for the sample size n=9							
parent distribution type	k	probability of the event $1-\Delta<\sigma^2_{M_k corr}<1+\Delta$					$\bar{q}_u-\bar{q}_l$
		Δ=0.2	Δ=0.4	Δ=0.6	Δ=0.8	Δ=1.0	
geostatis-tical	0.5	0.2266	0.4393	0.6214	0.7372	0.7993	0.9664
	1	0.2326	0.4492	0.6432	0.7726	0.8312	0.9679
	2	0.234	0.4545	0.6652	0.8125	0.8643	0.9811
	3	0.2304	0.4584	0.6758	0.8257	0.8807	0.9967
Jeffreys	0.5	0.2592	0.4966	0.687	0.8067	0.8627	0.8321
	1	0.2631	0.5066	0.7086	0.8316	0.8839	0.8165
	2	0.2652	0.5181	0.7292	0.8564	0.9056	0.8159
	3	0.265	0.522	0.7364	0.8698	0.9158	0.8324
Gaussian	0.5	0.2838	0.5629	0.7722	0.8847	0.9305	0.7263
	1	0.2895	0.5663	0.7724	0.885	0.9328	0.7118
	2	0.2952	0.5745	0.7871	0.9026	0.9445	0.7061
	3	0.3001	0.5761	0.7922	0.9104	0.9499	0.7080

Table IVb. Concentration of the MFV-corrected empirical variances around the true value σ^2 for the sample size $n=9$

IVb. táblázat. Az MFV korrigált varianciáknak a σ^2 körüli koncentrációja $n=9$ esetén

Concentration of the MFV corrected variances around the true value of the variance for the sample size n=16							
parent distribution type	k	probability of the event $1-\Delta<\sigma^2_{M_k\,corr}<1+\Delta$					$\overline{q}_u-\overline{q}_l$
		Δ=0.2	Δ=0.4	Δ=0.6	Δ=0.8	Δ=1.0	
geostatis-tical	0.5	0.3504	0.6379	0.8184	0.8999	0.9346	0.6089
	1	0.3539	0.6436	0.8293	0.9099	0.9413	0.5970
	2	0.3569	0.6533	0.8442	0.9231	0.952	0.6095
	3	0.3572	0.6553	0.8522	0.9292	0.9566	0.6188
Jeffreys	0.5	0.3531	0.6355	0.82	0.8979	0.9353	0.5939
	1	0.357	0.6436	0.8326	0.9088	0.9435	0.6013
	2	0.3578	0.6499	0.8478	0.9224	0.952	0.6059
	3	0.3593	0.6559	0.8551	0.9297	0.9564	0.6118
Gaussian	0.5	0.4067	0.7221	0.8924	0.9559	0.9804	0.5015
	1	0.4101	0.727	0.8984	0.9596	0.9824	0.5035
	2	0.4146	0.7325	0.9062	0.9663	0.9848	0.5040
	3	0.4143	0.7356	0.9106	0.9693	0.9859	0.5041

Table IVc. Concentration of the MFV-corrected empirical variances around the true value σ^2 for the sample size $n=16$

IVc. táblázat. Az MFV korrigált varianciáknak a σ^2 körüli koncentrációja $n=16$ esetén

Concentration of the MFV corrected variances around the true value of the variance for the sample size n=25							
parent distribution type	k	probability of the event $1-\Delta<\sigma^2_{M_k\,corr}<1+\Delta$					$\overline{q}_u-\overline{q}_l$
		Δ=0.2	Δ=0.4	Δ=0.6	Δ=0.8	Δ=1.0	
geostatis-tical	0.5	0.3661	0.6611	0.8258	0.8991	0.9387	0.5777
	1	0.3669	0.6734	0.8468	0.914	0.9496	0.5766
	2	0.3657	0.6827	0.862	0.9286	0.961	0.5848
	3	0.3652	0.6853	0.87	0.9369	0.9681	0.5953
Jeffreys	0.5	0.4329	0.7407	0.8902	0.9492	0.9754	0.4764
	1	0.4354	0.7475	0.9022	0.956	0.9783	0.4763
	2	0.4347	0.7547	0.913	0.9628	0.9819	0.4834
	3	0.4348	0.7594	0.9175	0.9662	0.9841	0.4858
Gaussian	0.5	0.5026	0.8263	0.9507	0.9854	0.996	0.3986
	1	0.5049	0.8315	0.9535	0.9866	0.9963	0.3949
	2	0.5107	0.8362	0.958	0.988	0.9967	0.3929
	3	0.5142	0.8382	0.9597	0.9888	0.997	0.3956

Table IVd. Concentration of the MFV-corrected empirical variances around the true value σ^2 for the sample size $n=25$

IVd. táblázat. Az MFV korrigált varianciáknak a σ^2 körüli koncentrációja $n=25$ esetén

Concentration of the MFV corrected variances around the true value of the variance for the sample size n=49							
parent distribution type		probability of the event $1-\Delta<\sigma^2_{M_k\,corr}<1+\Delta$					$\bar{q}_u-\bar{q}_l$
	k	Δ=0.2	Δ=0.4	Δ=0.6	Δ=0.8	Δ=1.0	
geostatistical	0.5	0.4892	0.8101	0.9335	0.9743	0.9902	0.4134
	1	0.4916	0.8172	0.9398	0.9776	0.9915	0.4096
	2	0.4981	0.8269	0.946	0.9813	0.9935	0.4099
	3	0.4979	0.8297	0.9495	0.982	0.994	0.4132
Jeffreys	0.5	0.5752	0.8831	0.9689	0.9915	0.9992	0.3384
	1	0.5776	0.89	0.9727	0.9929	0.9994	0.3387
	2	0.5797	0.8957	0.9763	0.9941	0.9995	0.3416
	3	0.5801	0.8978	0.9775	0.9951	0.9995	0.3434
Gaussian	0.5	0.6676	0.9424	0.9941	0.9996	1	0.2733
	1	0.6687	0.9439	0.9941	0.9996	1	0.2732
	2	0.6706	0.947	0.9943	0.9996	1	0.2716
	3	0.6712	0.9476	0.9947	0.9997	1	0.2722

Table IVe. Concentration of the MFV-corrected empirical variances around the true value σ^2 for the sample size $n=49$

IVe. táblázat. Az MFV korrigált varianciáknak a σ^2 körüli koncentrációja $n=49$ esetén

Concentration of the MFV corrected variances around the true value of the variance for the sample size n=100							
parent distribution type		probability of the event $1-\Delta<\sigma^2_{M_k\,corr}<1+\Delta$					$\bar{q}_u-\bar{q}_l$
	k	Δ=0.2	Δ=0.4	Δ=0.6	Δ=0.8	Δ=1.0	
geostatistical	0.5	0.6553	0.9378	0.9867	0.998	0.9997	0.2876
	1	0.6587	0.9412	0.988	0.9983	0.9998	0.2872
	2	0.6644	0.9455	0.9904	0.9985	0.9998	0.2877
	3	0.6629	0.9434	0.9893	0.9987	0.9998	0.2901
Jeffreys	0.5	0.7389	0.9705	0.9964	0.9998	1	0.2382
	1	0.7411	0.9719	0.9969	0.9999	1	0.2382
	2	0.7442	0.9735	0.9974	0.9999	1	0.2396
	3	0.745	0.9742	0.9975	0.9999	1	0.2391
Gaussian	0.5	0.8375	0.9926	0.9999	0.9999	1	0.1952
	1	0.8397	0.9931	0.9999	0.9999	1	0.1944
	2	0.8415	0.9932	0.9999	0.9999	1	0.1950
	3	0.8419	0.9934	0.9999	0.9999	1	0.1949

Table IVf. Concentration of the MFV-corrected empirical variances around the true value σ^2 for the sample size $n=100$

IVf. táblázat. Az MFV korrigált varianciáknak a σ^2 körüli koncentrációja $n=100$ esetén

parent distribution type	k	probability of the event $1-\Delta<\sigma^2_{M_k\,corr}<1+\Delta$					
		Δ=0.2	Δ=0.4	Δ=0.6	Δ=0.8	Δ=1.0	$\bar{q}_u-\bar{q}_l$
geostatis-tical	0.5	0.9447	0.9995	1	1	1	0.1410
	1	0.9446	0.9995	1	1	1	0.1409
	2	0.9447	0.9995	1	1	1	0.1401
	3	0.9442	0.9994	1	1	1	0.1391
Jeffreys	0.5	0.9749	1	1	1	1	0.1217
	1	0.9751	1	1	1	1	0.1210
	2	0.9757	1	1	1	1	0.1210
	3	0.9756	1	1	1	1	0.1208
Gaussian	0.5	0.9953	1	1	1	1	0.0937
	1	0.9953	1	1	1	1	0.0936
	2	0.9953	1	1	1	1	0.0936
	3	0.9954	1	1	1	1	0.0937

Table IVg. Concentration of the MFV-corrected empirical variances around the true value σ^2 for the sample size n=400

IVg. táblázat. Az MFV korrigált varianciáknak a σ^2 körüli koncentrációja n=400 esetén

parent distribution type	k	probability of the event $1-\Delta<\sigma^2_{M_k\,corr}<1+\Delta$					
		Δ=0.2	Δ=0.4	Δ=0.6	Δ=0.8	Δ=1.0	$\bar{q}_u-\bar{q}_l$
geostatis-tical	0.5	0.9396	1	1	1	1	0.0953
	1	0.9391	1	1	1	1	0.0952
	2	0.9389	1	1	1	1	0.0950
	3	0.9386	1	1	1	1	0.0949
Jeffreys	0.5	0.9989	1	1	1	1	0.0795
	1	0.9989	1	1	1	1	0.0794
	2	0.999	1	1	1	1	0.0794
	3	0.999	1	1	1	1	0.0793
Gaussian	0.5	1	1	1	1	1	0.0628
	1	1	1	1	1	1	0.0628
	2	1	1	1	1	1	0.0628
	3	1	1	1	1	1	0.0629

Table IVh. Concentration of the MFV-corrected empirical variances around the true value σ^2 for the sample size n=900

IVh. táblázat. Az MFV korrigált varianciáknak a σ^2 körüli koncentrációja n=900 esetén

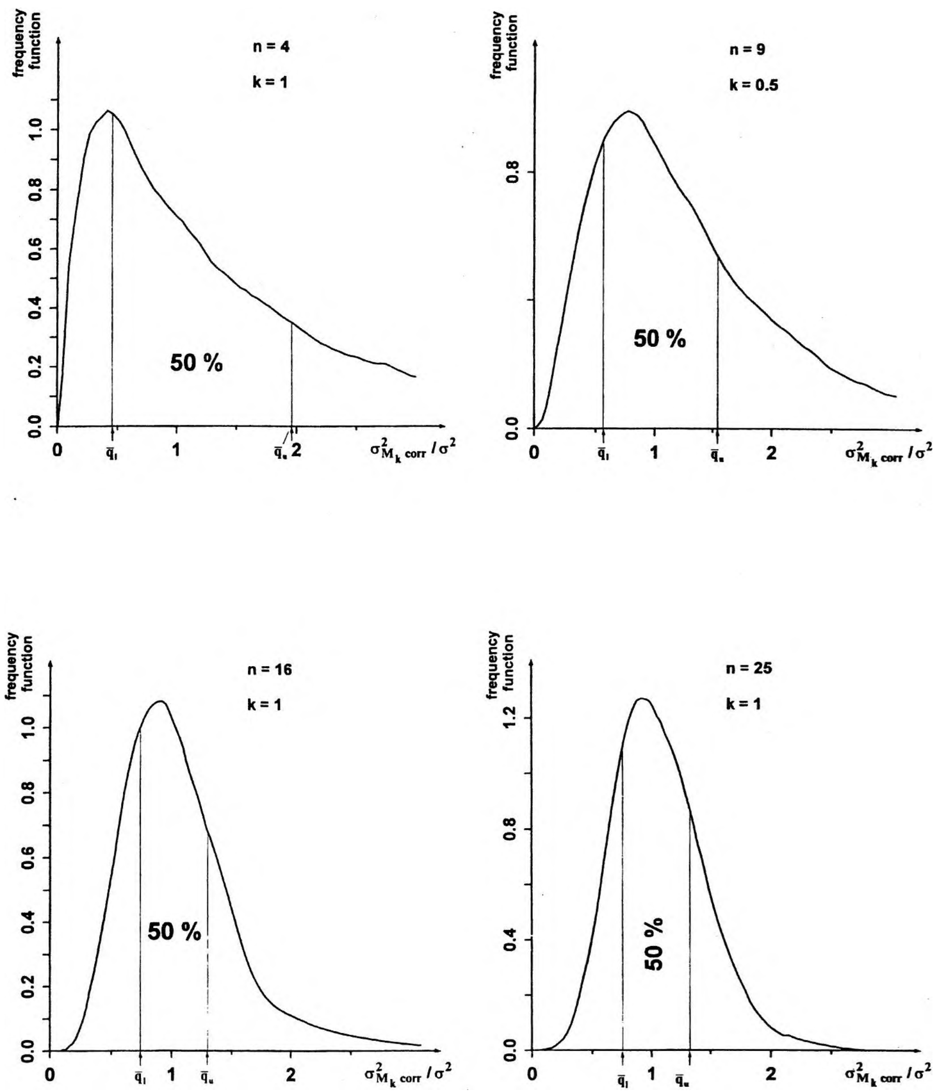


Fig. 6a. (the figure caption is given on the next page)
6a. ábra. (az ábrafelirat a következő oldalon található)

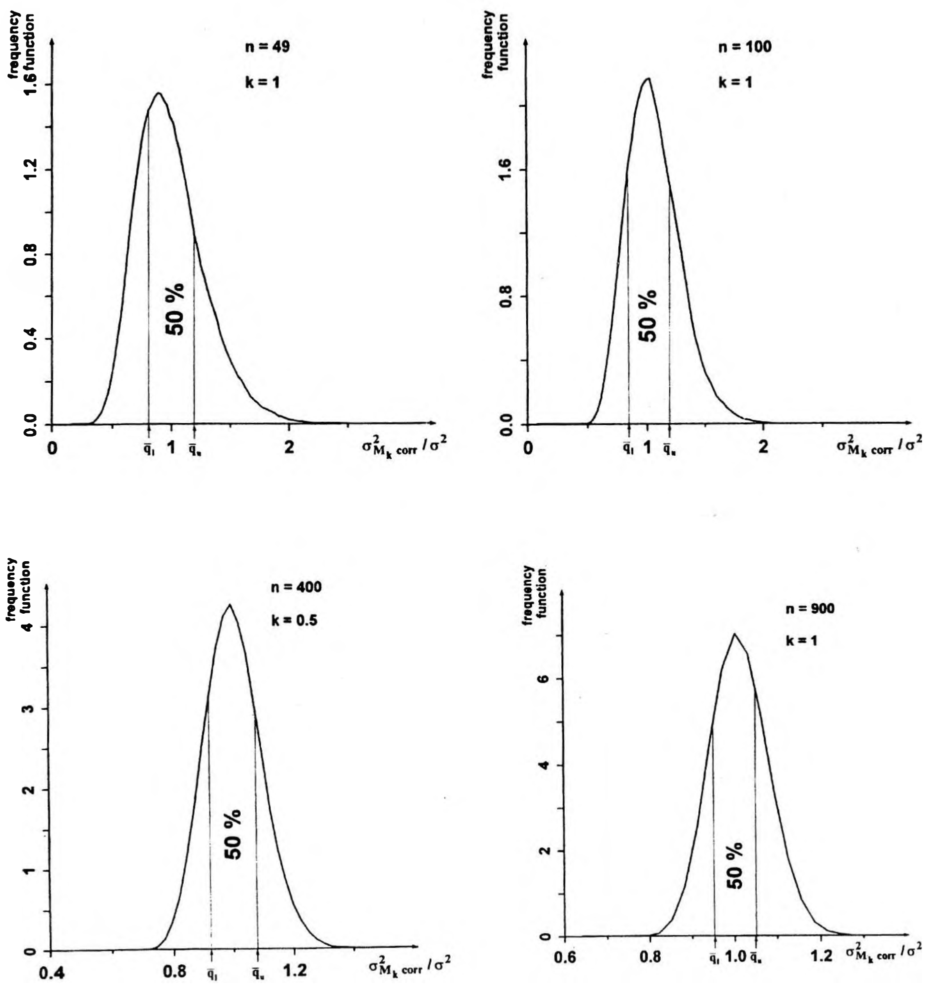


Fig. 6a. Curves of MFV-corrected variances ($\sigma^2_{M_k corr}$) for eight sample sizes (from $n=4$ to $n=900$); the parent distribution is of geostatistical type. The areas between $\bar{q}_l$ and $\bar{q}_u$ are characterized by the probability 0.5; the probability of both events $\bar{q}_l < \sigma^2_{M_k corr}/\sigma^2 < 1$ and $1 < \sigma^2_{M_k corr}/\sigma^2 < \bar{q}_u$ is equally 1/4

6a. ábra. MFV-korrigált varianciák görbéi (azaz $\sigma^2_{M_k corr}$-görbék) nyolcféle mintaméretre ($n=4$-től $n=900$-ig); az anyaeloszlás geostatisztikus típusú. A görbék alatti területeket $\bar{q}_l$ és $\bar{q}_u$ között 1/2 valószínűség jellemzi. Mind a $\bar{q}_l < \sigma^2_{M_k corr}/\sigma^2 < 1$, mind az $1 < \sigma^2_{M_k corr}/\sigma^2 < \bar{q}_u$ valószínűsége 1/4-del egyenlő

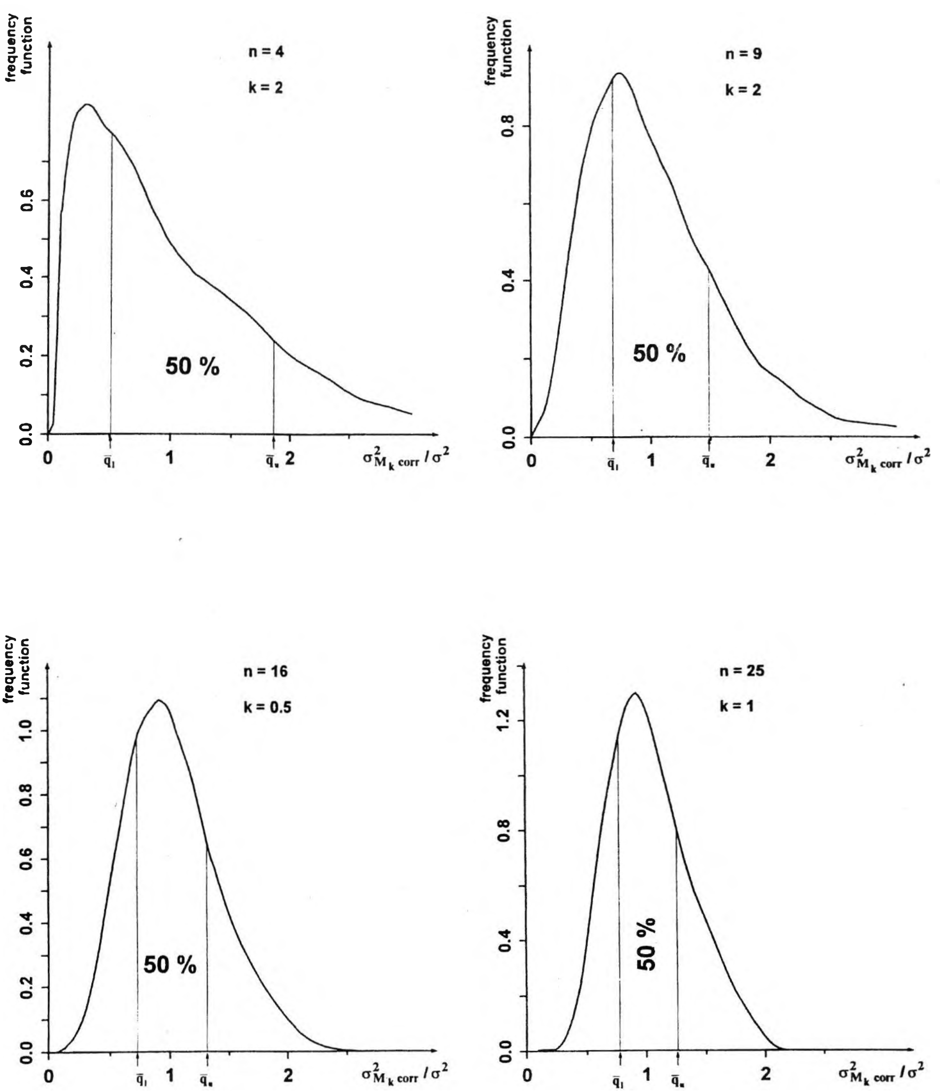


Fig. 6b. (the figure caption is given on the next page)
6b. ábra. (az ábrafelirat a következő oldalon található)

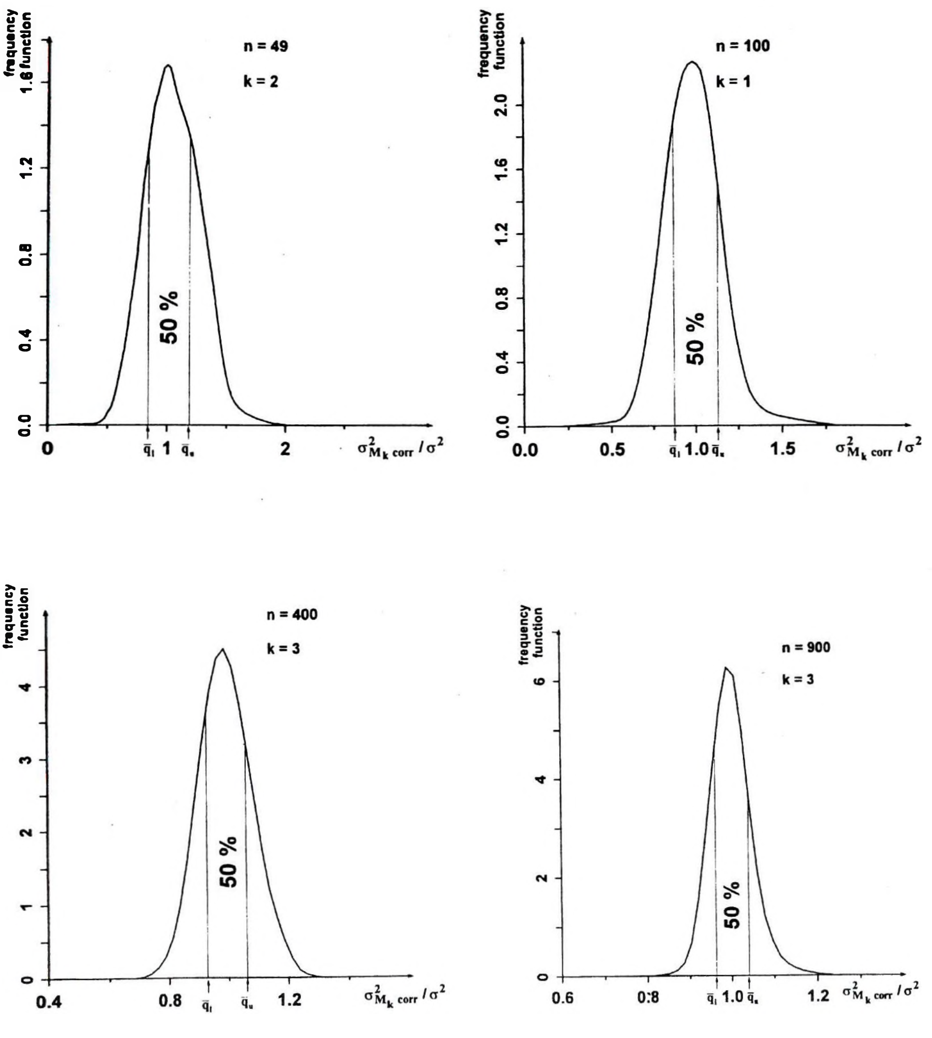


Fig. 6b. Curves of MFV-corrected variances ($\sigma^2_{M_k corr}$) for eight sample sizes (from $n=4$ to $n=900$); the parent distribution is of Jeffreys type. The areas between $\bar{q}_l$ and $\bar{q}_u$ are characterized by the probability 0.5; the probability of both events $\bar{q}_l < \sigma^2_{M_k corr}/\sigma^2 < 1$ and $1 < \sigma^2_{M_k corr}/\sigma^2 < \bar{q}_u$ is equally 1/4

6b. ábra. MFV-korrigált varianciák görbéi (azaz $\sigma^2_{M_k corr}$-görbék) nyolcféle mintaméretre ($n=4$-től $n=900$-ig); az anyaeloszlás Jeffreys típusú. A görbék alatti területeket $\bar{q}_l$ és $\bar{q}_u$ között 1/2 valószínűség jellemzi. Mind a $\bar{q}_l < \sigma^2_{M_l corr}/\sigma^2 < 1$, mind az $1 < \sigma^2_{M_k corr}/\sigma^2 < \bar{q}_u$ valószínűsége 1/4-del egyenlő

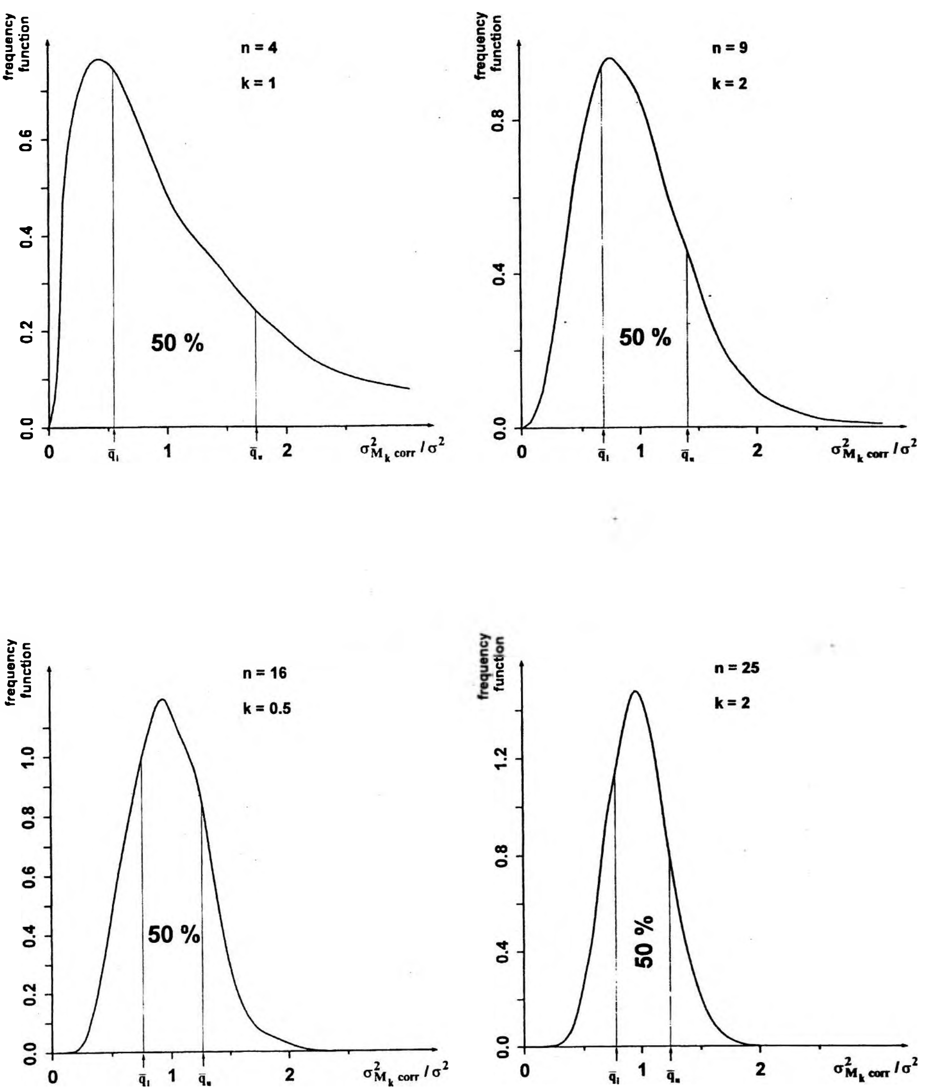


Fig. 6c. (the figure caption is given on the next page)
6c. ábra. (az ábrafelirat a következő oldalon található)

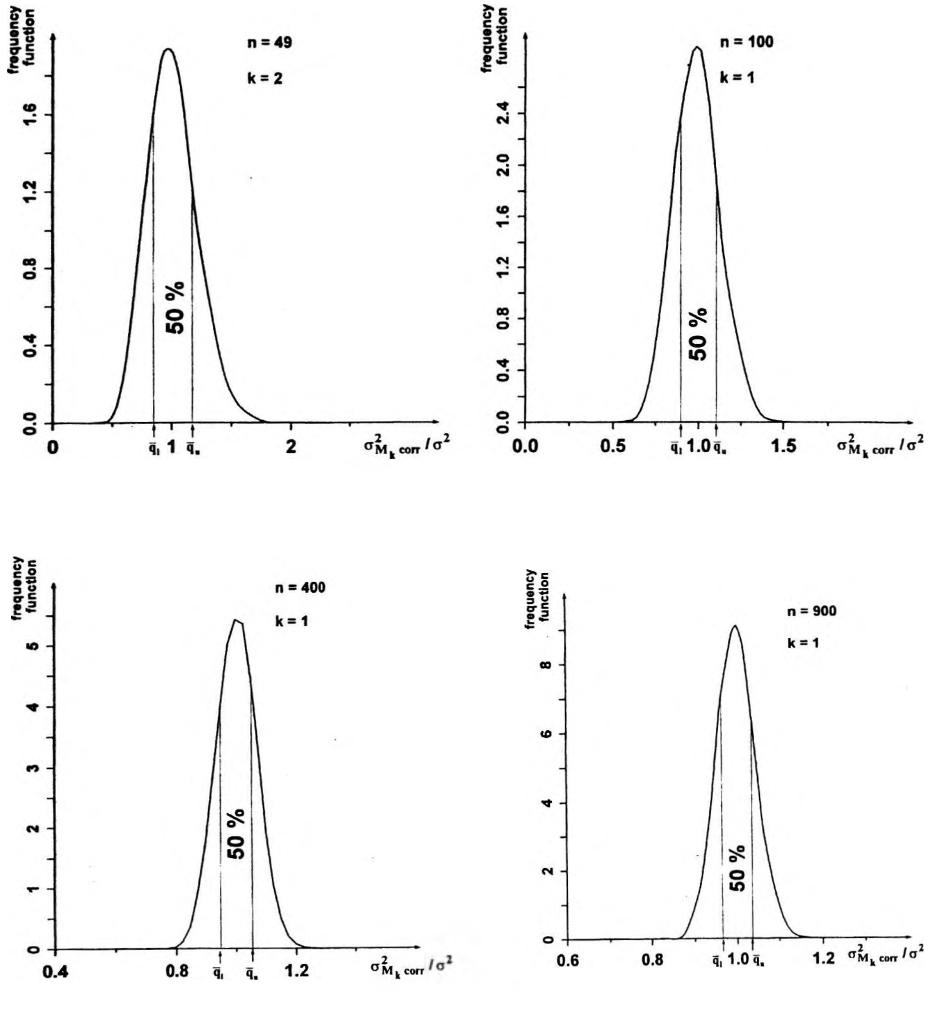


Fig. 6c. Curves of MFV-corrected variances ($\sigma^2_{M_k corr}$) for eight sample sizes (from $n=4$ to $n=900$); the parent distribution is of Gaussian type. The areas between $\bar{q}_l$ and $\bar{q}_u$ are characterized by the probability 0.5; the probability of both events $\bar{q}_l < \sigma^2_{M_k corr} / \sigma^2 < 1$ and $1 < \sigma^2_{M_k corr} / \sigma^2 < \bar{q}_u$ is equally 1/4

6c. ábra. MFV-korrigált varianciák görbéi (azaz $\sigma^2_{M_k corr}$-görbék) nyolcféle mintaméretre ($n=4$-től $n=900$-ig); az anyaeloszlás Gauss típusú. A görbék alatti területeket $\bar{q}_l$ és $\bar{q}_u$ között 1/2 valószínűség jellemzi. Mind a $\bar{q}_l < \sigma^2_{M_k corr} / \sigma^2 < 1$, mind az $1 < \sigma^2_{M_k corr} / \sigma^2 < \bar{q}_u$ valószínűsége 1/4-del egyenlő

Appendix

A/ Formulae for calculating M_k and ε_k based on samples: the equations

$$M_k = \frac{\sum_{i=1}^{n} \frac{x_i}{(k\varepsilon_k)^2 + (x_i - M_k)^2}}{\sum_{i=1}^{n} \frac{1}{(k\varepsilon_k)^2 + (x_i - M_k)^2}} \qquad (A1)$$

and

$$\varepsilon_k^2 = 3\frac{\sum_{i=1}^{n} \frac{(x_i - M_k)^2}{\left[\varepsilon_k^2 + (x_i - M_k)^2\right]^2}}{\sum_{i=1}^{n} \frac{1}{\left[\varepsilon_k^2 + (x_i - M_k)^2\right]^2}} \qquad (A2)$$

must be simultaneously fulfilled. (These formulae were applied in the Monte Carlo investigation of the present paper executed for estimates of variances obtained for samples from geostatistical and Jeffreys parent type distributions.)

B/ Formulae for calculating M_k and ε_k based on given probability density functions $f(x)$: the equations

$$M_k = \frac{\int_{-\infty}^{\infty} \frac{x}{(k\varepsilon_k)^2 + (x - M_k)^2} f(x)\mathrm{d}x}{\int_{-\infty}^{\infty} \frac{1}{(k\varepsilon_k)^2 + (x - M_k)^2} f(x)\mathrm{d}x} \qquad (A3)$$

and

$$\varepsilon_k^2 = 3\frac{\int_{-\infty}^{\infty} \frac{(x - M_k)^2}{\left[\varepsilon_k^2 + (x - M_k)^2\right]^2} f(x)\mathrm{d}x}{\int_{-\infty}^{\infty} \frac{1}{\left[\varepsilon_k^2 + (x - M_k)^2\right]^2} f(x)\mathrm{d}x} \qquad (A4)$$

must be simultaneously fulfilled.

C/ The density- and distribution functions (for the standard case, i.e., if for the parameter of location (*T*) and that of scale (*S*) $T=0$ and $S=1$ hold) for the geostatistical and Jeffreys types of probability distribution and for the Gaussian are the following:

$$f_{st}(x) = \frac{3}{4}\cdot\frac{1}{\left(1+x^2\right)^2\cdot\sqrt{1+x^2}} \tag{A5}$$

$$F_{st}(x) = \frac{1}{2} + \frac{x}{2\sqrt{1+x^2}} + \frac{x}{4\left(1+x^2\right)\sqrt{1+x^2}} \tag{A6}$$

$$f_J(x) = \frac{35}{32}\cdot\frac{1}{\left(1+x^2\right)^4\cdot\sqrt{1+x^2}} \tag{A7}$$

$$F_J(x) = F_{st}(x) + \frac{3x}{16\left(1+x^2\right)^2\sqrt{1+x^2}} + \frac{5x}{32\left(1+x^2\right)^3\sqrt{1+x^2}} \tag{A8}$$

$$f_G(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} \tag{A9}$$

$$F_G(x) = \int_0^x f_G(y)\mathrm{d}y \tag{A10}$$

calculated by computer with the approximation formula given, for example, in STEINER [1990] as Eq.(1–45).

The maximum difference between these approximate values and the correct values of $F_G(x)$ equals 10^{-7}. For generating samples of geostatistical and Jeffreys types, random numbers distributed equally between 0 and 1 are based on the distribution functions $F_{st}(x)$ and $F_J(x)$ — in the sense demonstrated in STEINER [1990] (Fig. 5.2) .

REFERENCES

CRAMÉR H. 1945: Mathematical Methods of Statistics. Almqvist & Wiksells, Uppsala, 575. p.

DUTTER R. 1986/87: Mathematische Methoden in der Montangeologie. Vorlesungsnotizen. Manuscript, Leoben

HAJAGOS B., STEINER F. 1994: Bestimmungsfehler der statistischen Momente. Publ. Univ. Miskolc Series D, **35**, No.4. pp.65–95

KERÉKFY P. 1978: On robust estimates. (in Hungarian) Alkalmazott Matematikai Lapok, **4**, pp. 327-357

KIS M., HURSÁN G. 1995: Current distortions of the corrected empirical variance. (in Hungarian) Poster presentation. V. Conference for Geomathematics 5th Oct. 1995 Szeged

MATHERON G. 1965: Les variables regionalisées et leur estimation. Masson & Cie, Paris, 305. p.

STEINER F. 1990: Principles of geostatistics. (in Hungarian) Tankönyvkiadó, Budapest, 363. p.

STEINER F. ed. 1991: The most frequent value. Akadémiai Kiadó, Budapest, 315. p.

MFV-korrigált varianciák

STEINER Ferenc, HAJAGOS Béla és HURSÁN Gábor

A konvencionálisan korrigált empirikus varianciának ($\sigma^2_{emp,corr}$-nak) még az adatok steril Gauss-típusú eloszlása esetén is vannak ún. „valószínűségi torzulásai". Ezek ellen úgy védekezhetünk, hogy MFV-korrekciót alkalmazunk; az ehhez szükséges korrekciós faktorokat a IIIa. táblázat mutatja be három anyaeloszlás-típusra és különböző n mintaméretekre (görbékkel az 5. ábra szemlélteti ezeket). A IIIb. táblázat együttható-hármasai segítségével e korrekciós faktorok jó közelítéssel számíthatók bármely n mintaméretre (megjegyzendő, hogy ilyen korrekcióra $n<1000$ esetén a tárgyalt anyaeloszlásoknál már nincs is szükség). A $\sigma^2_{M_t,corr}$-eket, mint valószínűségi változókat, a gyakorlat szemszögéből a legfontosabb adataikkal a IVa.–IVh. táblázatok mutatják be.

A gyakorlati célú vizsgálatok felvetették azt a kérdést, hogy a b becslések torzítatlanságának általánosan elfogadott elméleti definíciója (nevezetesen hogy ti. b-t torzítatlan becslésnek nevezzük, ha a becslések számértékeinek várható értéke azonos az elméletileg helyes értékkel) kritika nélkül alkalmazható-e a gyakorlatban. A torzítatlanság kritériumaként a gyakorlat inkább azt igényelné, hogy a dolgozat szerint a becslések számértékeinek leggyakoribb értéke legyen egyenlő a valódi értékkel.