GEOPHYSICAL TRANSACTIONS 1993
Vol. 38. No. 4. pp. 231-238

COMMENT ON AN OLD DOGMA: 'THE DATA ARE NORMALLY DISTRIBUTED'

Péter SZŰCS*

Attention is called to the dangers applying the χ^2-test in normality investigations. As is well known, the χ^2-test is one of the most frequently used methods for normality investigations when the hypothetical distribution is Gaussian. The Monte-Carlo simulations carried out show that the χ^2-test at the usual significance levels find different distributions (significantly differing from the Gaussian one) from the Gaussian distribution. This situation is termed the 'trap of the χ^2-test' and it may further strengthen the lack of credibility of the predominant presence of Gaussian mother distributions.

Keywords: χ^2-test, normality investigation, significance level, probability

1. Introduction

Depending on the type of probability distribution some authors directly reject the appearance of Gaussian distributions as being mother ones [MOSTELLER, TUKEY 1977, TUKEY 1977]. For example we can read on p. 661 of TUKEY [1977]: 'When the underlying distribution, as always, is nongaussian...'.

We can use several so called normality tests to check whether a sample originates from Gaussian distribution or not. One of the most frequently used methods for normality investigations is the χ^2-test. In this we almost always utilize the sample mean and the standard deviation as parameters, i.e. we carry out the test of goodness of fit [VINCZE 1968]. The question arises whether the level of probability of the χ^2-test finds some distributions different from the normal one — as is Gaussian distribution. We performed Monte-Carlo investigations to answer the question. Taking our results into consideration we

* University of Miskolc, Department of Geophysics, H-3515 Miskolc-Egyetemváros
Manuscript received: 22 November, 1993

suggest, as a first step, another test [CSERNYÁK 1989] instead of the χ^2-test for a given distribution family.

2. Dangers of the χ^2-test

HAJAGOS [1988] carried out Monte-Carlo investigations that indicated the dangers of the χ^2-test. At that time however the investigations could not have been expanded to sufficiently great sample and repetition numbers because of the limitations of the domestic computer field. We therefore felt justified in carrying out similar investigations as the present level of computer sciences can now offer us far more scope.

What type of distributions do we submit to the χ^2-test? We investigated three different representatives of the $f_a(x)$ supermodel. We can define the supermodel in the following manner [STEINER 1990]:

$$f_a(x) = n(a) \cdot \frac{1}{(\sqrt{x^2+1})^a} \quad (a>1)\,. \tag{1}$$

where a is the type parameter, since the tails of the distribution functions are wider when the values of a are small. When the values of a are great, the tails will be much shorter and the maximum will be flatter. It can be proved that for $a \to \infty$ the standard form approaches the Gaussian distribution function. The $n(a)$ figuring in (1) is a normalization factor and can be calculated as follows:

$$n(a) = \frac{\Gamma\left(\frac{a}{2}\right)}{\sqrt{\pi} \cdot \Gamma\left(\frac{a-1}{2}\right)}\,. \tag{2}$$

The $f_a(x)$ model-family is able to model the cases that may occur in practice. If we have no preliminary information about the type of data distribution, the application of a=5 can be offered for geostatistical tasks [STEINER 1991, page 298, fig.1]. Let us take this a=5 type as one of our investigated distributions. The a=9 distribution was named after JEFFREYS [1961]. This is a representative of the distributions with the shortest tails, which are likely to occur in the geosciences. Thus, our second investigated distribution will be the Jeffreys one. Our third distribution will be the $f_3(x)$. This represents a distribution with wide tails, but it is still not Cauchy type.

During the Monte-Carlo investigations we created samples with 100 and 400 elements from the above mentioned distributions with the aid of a random generator. We repeated the sampling a thousand times. After finishing the χ^2-tests we were able to calculate probability values to an accuracy of two decimal places for different significance levels. These values show with how much probability the χ^2-test would accept the given type and size samples as normally distributed ones at the given significance level. The thousandfold sampling proved to be reliable. When we repeated the investigations, there was only a negligible fluctuation in the third decimal figure of the probability values. We can see the detailed results of the investigation in Figs. 1 and 2. The curves have great probability values. For the samples with 100 elements (see *Fig. 1*) we accept our data originating from geostatistical (a=5) distribution as normally distributed ones in half of the instances at the 90 percentile significance level. In the case of a=9 the situation is even worse: the concrete probability value is 0.842 at the 90 percentile significance level. For the samples with 400 elements the situation is slightly better although the probabilities remain high enough henceforward (*Fig.* 2). In the case of a=3 there was no 'acceptance'. Based on the χ^2-test we would even say, with high probability, that our samples with 400 elements originated from the Jeffreys distribution as normally distributed ones.

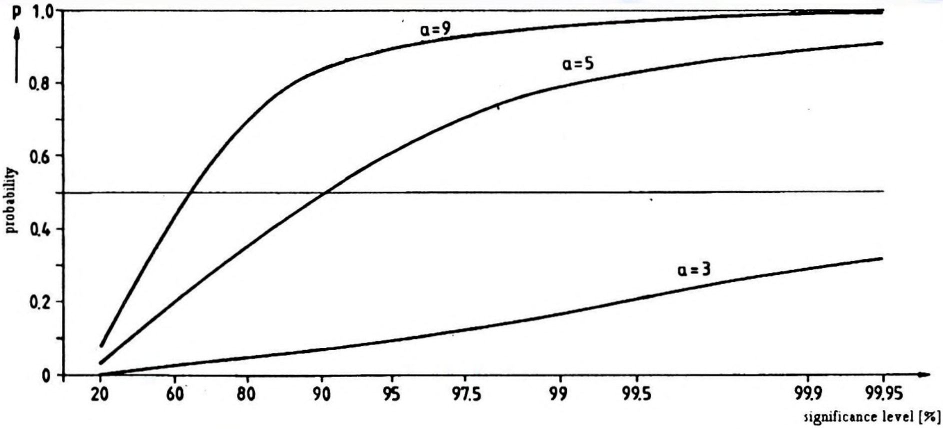

Fig. 1. Probabilities of acceptance of the Gaussian hypothesis at the given significance levels (χ^2-test, n=100)

1. ábra. A Gauss-hipotézis elfogadásának valószínűségei az adott szignifikancia szinteken (χ^2-próba, n=100)

These findings can be termed the 'trap of the χ^2-test' that may further strengthen the lack of credibility of the predominance of Gaussian mother distributions. From the practical aspect this situation has a harmful effect on those users who apply the least squares method without deeper consideration and investigation. From the theoretical aspect this can lead to the general acceptance of the standard deviation as a universal uncertainty property, and we may wrongly take into account the message and the validity domain of the Heisenberg relation [CSERNYÁK, STEINER 1991].

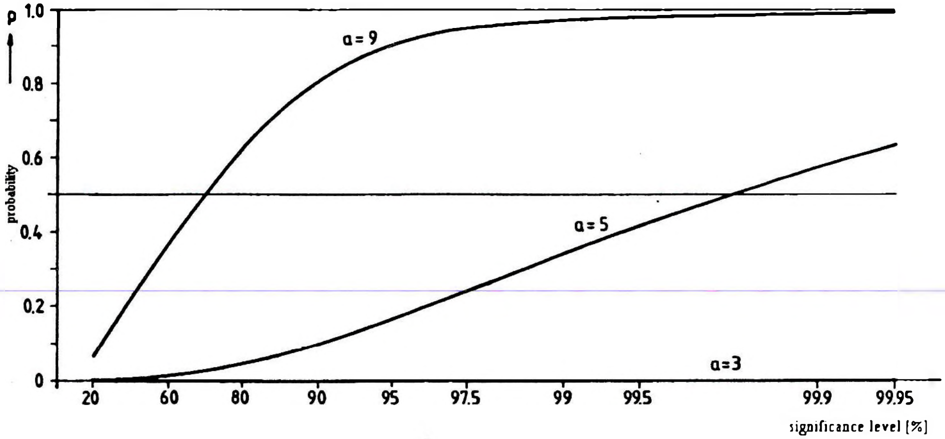

Fig. 2. Probabilities of acceptance of the Gaussian hypothesis at the given significance levels (χ^2-test, n=400)

2. ábra. A Gauss-hipotézis elfogadásának valószínűségei az adott szignifikancia szinteken (χ^2-próba, n=400)

The question may arise, with how much probability we would accept the Gaussian hypothesis for the χ^2-test if our samples originated from any member of the $f_a(x)$ supermodel. To answer the question we should know with what degree of probability the different a values in the $f_a(x)$ supermodel would occur. During our investigation we applied two different distribution functions that are able to model the occurrence probabilities [see Fig. 4, and Eqs. 10 and 11 of STEINER, HAJAGOS 1993]. These are as follows:

$$\text{I.} \qquad f_D\left(\frac{1}{a-1}\right)=\frac{16}{a-1}\cdot e^{-\frac{4}{a-1}}, \tag{3}$$

$$\text{II.} \qquad f_J\left(\frac{1}{a-1}\right)=\frac{64}{a-1}\cdot e^{-\frac{8}{a-1}}. \tag{4}$$

We summarize our results in *Table I.* Naturally the results of the table were not calculated from infinite different distributions. We obtained the numerical values in a similar way to the way in which we completed the χ^2-tests for eleven different distributions of the $f_a(x)$ supermodel, and we integrated numerically the results weighted with (3) and (4) probability distributions.

	Significance levels									+
	20%	60%	80%	90%	95%	97.5%	99%	99.5%	99.9%	99.95%
n=100 I.,f_D	0.027	0.168	0.297	0.378	0.456	0.525	0.563	0.594	0.651	0.668
n=400	0.018	0.072	0.130	0.183	0.228	0.267	0.315	0.349	0.419	0.441
n=100 II.,f_J	0.045	0.263	0.499	0.562	0.656	0.718	0.769	0.797	0.844	0.858
n=400	0.042	0.154	0.262	0.346	0.410	0.460	0.516	0.556	0.637	0.662

Table I. Probability values for the acceptance of the Gaussian hypothesis when using the χ^2-test at the given significance levels if our distribution originated from the $f_a(x)$ supermodel with f_D or f_J probability distributions

I. táblázat. Valószínűségek a Gauss-hipotézis elfogadására χ^2-próba alkalmazása esetén az adott szignifikanciaszinteken, ha eloszlásuk az $f_a(x)$ szupermodellből származik f_D vagy f_J valószínűségsűrűséggel

The rows belonging to I were calculated with the help of (3), the values belonging to II were calculated with the aid of (4). For (3) the geostatistic distribution (a=5) occurs with the greatest probability whereas in the case of (4) the most probable distribution is the Jeffreys one (a=9). The large probabilities we find in the table tend to underline the dangers of applying the χ^2-test. For example, even for samples with 400 elements the probabilities of acceptance of the Gaussian hypothesis are 0.315 and 0.516. These are very great probability values, especially if we take it into consideration that in the case of (3) and (4) the occurrence probability of Cauchy distribution is still not negligible.

3. The Csernyák test

It is a well known result of mathematical statistics that the distribution function of the 'extent' of the sample with n elements

$$R = X_{\max} - X_{\min} \tag{5}$$

is associated with the type of mother distribution [CRAMÉR 1946]. The sample size cannot be regarded as statistics that characterize the distribution because R is obviously proportional to the scale parameter (S) as well as to the sample size. We neglect S if we compare R to the empirical interquartile range determined from the same sample in the following manner:

$$C = \frac{R}{2q_{emp}} \tag{6}$$

We can accept this as the statistical function of the test for type determination [CSERNYÁK 1989]. This expression is suitable for normality investigations so we refer to the procedure as the Csernyák test.

On the basis of our calculations it can be stated that the Csernyák test is more reliable in the applied type range. Our results are shown in *Figs. 3* and *4*. If these figures are compared with Figs. 1 and 2 it can be realized that in case of the Csernyák test we accept the samples as Gaussian type with much less probability than in the case of the χ^2-test.

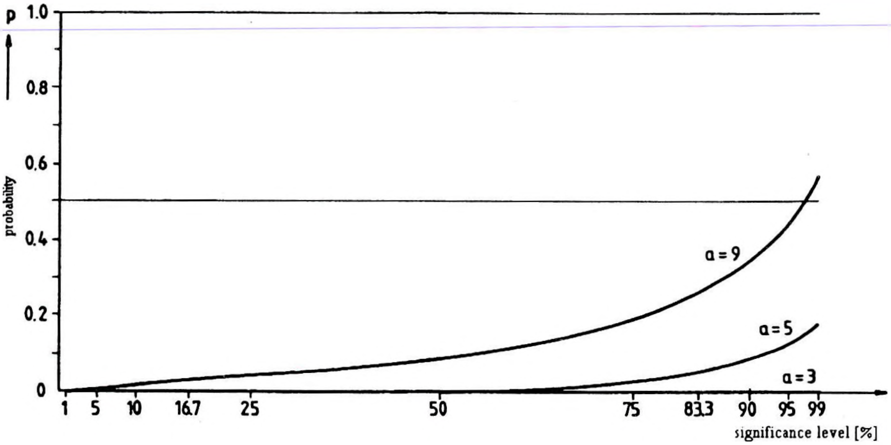

Fig. 3. Probabilities of acceptance of the Gaussian hypothesis at the given significance levels (Csernyák test, n=100)

3. ábra. A Gauss-hipotézis elfogadásának valószínűségei az adott szignifikancia szinteken (Csernyák teszt, n=100)

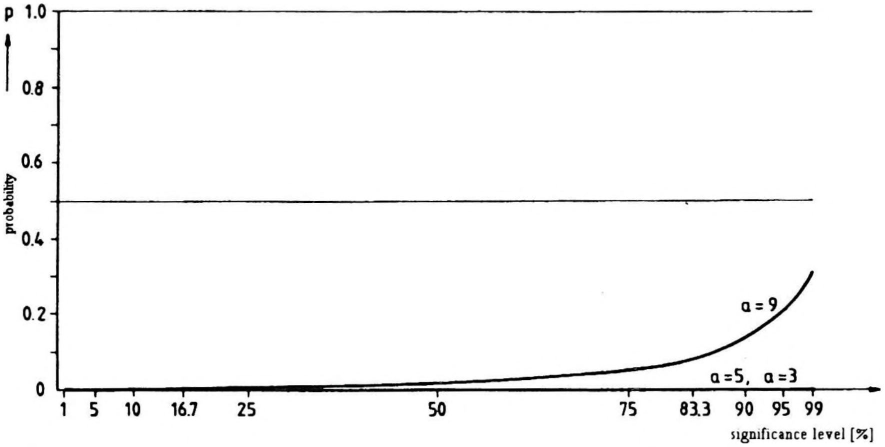

Fig. 4. Probabilities of acceptance of Gaussian hypothesis at the given significance levels (Csernyák test, n=400)

4. ábra. A Gauss-hipotézis elfogadásának valószínűségei az adott szignifikancia szinteken (Csernyák teszt, n=400)

It might well be said that the Csernyák test supposes freedom from outliers. Although this may be true, it does not alter the situation: it makes no difference whether the great value of R is caused by outlier free types with heavier tails than the tails of normal distribution, or by the appearance of outliers. The rejection of the hypothesis calls attention in both cases to the need to handle the methods of traditional statistics cautiously.

4. Conclusions

Based on the results of Monte-Carlo investigations we can establish the following facts:

— the χ^2-test cannot be recommended for the normality tests of different distributions occurring in the practice of geosciences. Even if our samples are quite different from the Gaussian distribution, the χ^2-test accepts them as normally distributed ones with large probabilities at the most frequently used significance levels;

— when applying the χ^2-test the lack of credibility of the predominant presence of Gauss mother distribution may contribute to the survival of the traditional (not robust and not resistant) statistical algorithms;

— for measured data sets we would suggest the use of the Csernyák test as a first step if our distribution originates from the f_a supermodel.

REFERENCES

CRAMER H. 1946: Mathematical methods of statistics. Princeton University Press, Princeton, IV. J.

CSERNYÁK L. 1989: Determination of type using sample range. Acta Geodaetica, Geophysica et Montanistica, Acad. Sci. Hung. **24**, 3-4, pp. 441-447

CSERNYÁK L., STEINER F. 1991: The inadequacy of the Heisenberg relation in generally posed questions of uncertainties. *In*: The Most Frequent Value: Introduction to a modern conception of statistics. (ed. F. STEINER) Appendix IX. pp. 271-295, Akadémiai Kiadó, Budapest

HAJAGOS B. 1988: Normalitätsuntersuchungen mit Hilfe der χ^2-Probe an Stichproben, die aus Student-schen Mutterverteilungen stammen. Publications of the Technical University for Heavy Industry, Miskolc Series A. Mining, Vol. **44**. pp. 217-230

JEFFREYS H. 1961: Theory of Probability. Clarendon Press, Oxford

MOSTELLER F., TUKEY J. W. 1977: Data analysis and regression. Addison — Wesley Reading, Mass

STEINER F. 1990: A geostatisztika alapjai. Tankönyvkiadó, Budapest, 363 p.

STEINER F. (ed.) 1991: The Most Frequent Value. Akadémiai Kiadó, Budapest (Hungary), 315 p.

STEINER F., HAJAGOS B. 1993: Practical definition of robustness. (present issue)

TUKEY J.W. 1977: Exploratory data analysis. Addison- Wesley, Reading, Mass

VINCZE I. 1968: Matematikai statisztika ipari alkalmazásokkal. Műszaki Könyvkiadó, Budapest, 352 p.

MEGJEGYZÉS EGY RÉGI DOGMÁHOZ: „AZ ADATOK GAUSS-ELOSZLÁSÚAK"

SZŰCS Péter

Ez a cikk a χ^2-próba normalitásvizsgálatbeli alkalmazásának a veszélyeire szeretné felhívni a figyelmet. Mint jól ismert, az egyik leggyakrabban alkalmazott módszer a normalitásvizsgálatra a χ^2-próba, amikor a hipotetikus eloszlás a Gauss-féle. Az elvégzett Monte-Carlo vizsgálatok azt mutatják, hogy a χ^2-próba a szokásos szignifikanciaszinteken nagy valószínűséggel Gauss-eloszlásúnak talál attól szignifikánsan különböző eloszlásokat. Ezt akár a „χ^2-próba csapdájának" is nevezhetnénk, ami tovább erősítheti a Gauss-eloszlás anyaeloszlásként való túlnyomó előfordulásának a tévhitét.