STATISTICAL ANALYSIS OF ONE-DIMENSIONAL OBSERVATIONS FROM PARTIALLY GROUPED DATA

B. Yu. Lemeshko and S. N. Postovalov

Software to choose the distribution law that best describes sampled data is a further development of the program "System for Statistical Analysis of One-Dimensional Observations of Random Quantities" [1]. The functions of the software make it possible to estimate the parameters and check them for compatibility on the basis of 26 of the distributions most commonly used in practice and mixtures of those distributions.

The nature of the sampled data used to analyze the distributions *may differ*. The most general case is a partially grouped sample [2]. A sample is *not grouped* if the sampled values represent individual values of observations from the domain of definition of a random quantity. The sample is *grouped*, if the domain of definition of the random quantity is divided into k nonintersecting intervals by the boundary points

$$x_0 < x_1 < \ldots < x_{\kappa-1} < x_\kappa,$$

where x_0 is the lower side of the domain of definition of a random quantity ξ, x_k is the higher side of the domain of definition of ξ, and the number n_i of observations in the i-th interval of values. A sample is *partially grouped* if the available information is linked to many nonintersecting intervals, which divide the domain of definition of the random quantity so that each interval is one of two types:

a) the i-th interval is of the first type if n_i is known, but the individual values $x_{ij}, j = \overline{1, n_i}$, are unknown;

b) the i-th interval is of the second type if n_i and all individual values $x_{ij}, j = \overline{1, n_i}$, are known.

The domain of definition of a random quantity in this case can be represented as $X = X_{(1)} \cup X_{(2)}$, where $X_{(1)}$ is the set of intervals of the first type and $X_{(2)}$, of the second type.

When we have to choose the distribution with which our experiments are most compatible, the sequence of our actions is described by the following algorithm.

We limit the class of distributions from which we choose the appropriate law of probability distribution.

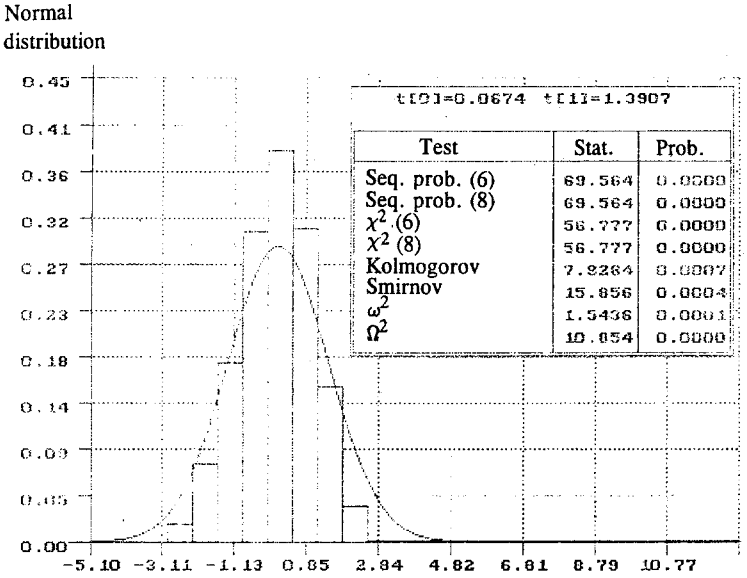
Next, for the chosen distributions we estimate the parameters and verify the compatibility hypothesis.

We choose the distribution that is most compatible with the sample.

The generally accepted procedure for verifying hypotheses by the compatibility test is used. When a compatibility hypothesis with a given distribution is not refuted if the calculated value of the statistic is less than a critical value that corresponds to a prescribed significance level α, it usually turns out that there is no reason to discard a number of distributions. Several possible alternatives remain in that case. We should not dwell on a distribution for which the compatibility is best.

When hypotheses are verified for compatibility with the statistic used S_i, $i = \overline{1, m}$, the software described calculates probabilities of the form $P\{S_i > S_i^*\} = \int_{S_i^*}^{\infty} g_i(s)ds$, where S_i^* is the value found from the sample for the pertinent statistic, $g_i(s)$ is the distribution density function of the statistic S_i, provided that the hypothesis H_0 is true, S_i^* is a functional that depends on the specific sample and the distribution law, i.e., $S_i^* = S_i^*(\bar{X}, f(x, \hat{\theta}))$, where $\bar{X}$ denotes a sample of a random quantity.

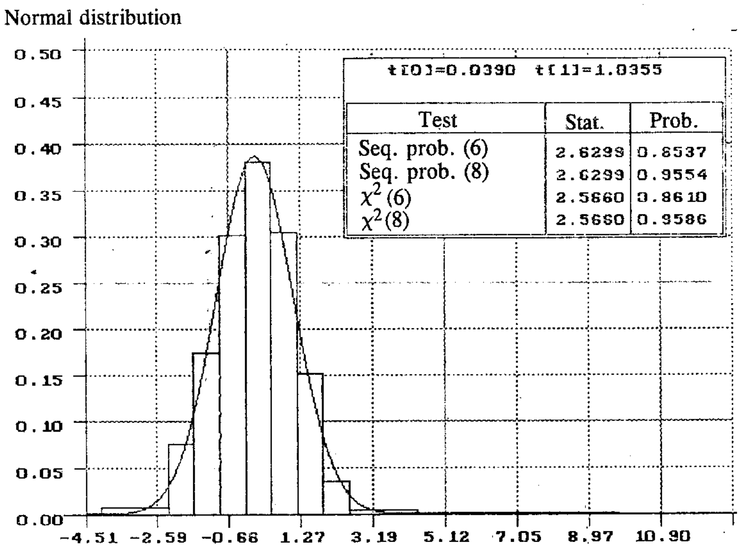
Novosibirsk State Technical University. Translated from Izvestiya Vysshikh Uchebnykh Zavedenii, Fizika, No. 9, pp. 39-45, September, 1995.



t[0]=0.0674 t[1]=1.3907

Test	Stat.	Prob.
Seq. prob. (6)	69.564	0.0000
Seq. prob. (8)	69.564	0.0000
χ^2 (6)	56.777	0.0000
χ^2 (8)	56.777	0.0000
Kolmogorov	7.2264	0.0007
Smirnov	15.856	0.0004
ω^2	1.5436	0.0001
Ω^2	10.054	0.0000

Fig. 1. Results of analysis in the presence of an "anomalous" observation.



t[0]=0.0390 t[1]=1.0355

Test	Stat.	Prob.
Seq. prob. (6)	2.6299	0.8537
Seq. prob. (8)	2.6299	0.9554
χ^2 (6)	2.5660	0.8610
χ^2 (8)	2.5660	0.9586

Fig. 2. Results of working estimation.

Suppose that on the basis of the initial assumptions we distinguished a set of distribution laws, to which the sample under consideration belongs, numbered those laws, using R to denote the set of indices of the density functions $f_j(x, \hat{\theta}), j \in R$, estimated the distribution law parameters from the given sample, and calculated the statistic $S^*_{ij} = S^*_i(\bar{X}, f_j(x, \hat{\theta}))$ and the probability $P\{S_i > S^*_{ij}\} = \alpha_{ij}$. Then during verification of the compatibility hypothesis with the j-th distribution by means of the i-th test, if $\alpha_{ij} > \alpha$, where α is the significance level set by the researcher, there is no reason to refute the compatibility hypothesis with the j-th distribution in accordance with the i-th test. Suppose that according to the tests used there are no grounds for refuting the compatibility hypothesis with the set of laws labeled with indices from $R_1 \subset R$. We should then choose the same distribution law of the random quantity $f_l(x, \hat{\theta})$, for which $\forall i \ \alpha_{il} = \max\limits_{j \in R_1} \alpha_{ij}$.

Usually such a conclusion can be made unambiguously. It is entirely possible, however, (quite often for different but similar distribution laws) that the conclusions from different tests indicate the predominance of one law or other. This means that the solutions of the problem of choosing a distribution by various tests are not compatible. Such "incompatibility" is attributed to the difference in measures used in the tests. Hence we have a natural multiple-test problem for adopting a solution. Since all the tests are measured on the same scale, the problem can be solved by fashioning a simple compromise test of the form $\max\limits_{j \epsilon R_1} \sum\limits_{i=1}^{m} \omega_i \alpha_{ij}$, where ω_i is the weight coefficient of the i-th test, $\sum\limits_{i=1}^{m} \omega_i = 1$.

When solving problems of statistical analysis, in particular when calculating estimates of the distribution parameters, the presence of *anomalous measurements* in the sample becomes extremely important. In the practice of solving such problems it is widely known that even a singly anomalous observation leads to estimates that are not at all compatible with the sample data. Generally speaking, the existence of spikes affects the quality of all the conclusions.

Understandably, every researcher hopes that the estimates have the least possible sensitivity to anomalous observations. Otherwise, before proceeding to make the estimate it is necessary to use a procedure to eliminate gross errors of measurement, which spills over into a rather complicated problem. In the given case we must emphasize the merit of the estimates that use *grouping* of the initial sample data, since obviously they are less sensitive to random spikes. When sample grouping is employed the influence of anomalous observations can be decreased sharply and the influence of random spikes can sometimes be eliminated altogether.

Let us demonstrate this with the following example. We model a sample by means of a normal law with zero mathematical expectation and a single variance, consisting of 500 observations, was modeled. From it we found estimates of the maximum likelihood: $\mu = 1.027$ and $\sigma = 1.017$. Then in this sample we increased the first observation by 20 and carried out the appropriate analysis again. The results are shown in Fig. 1. As was to be expected, the estimate of the mean-square deviation changed most. Compatibility with all tests is rejected.

Next, we grouped a sample with an "anomalous" observation, made an estimate from the grouped sample, and checked for compatibility. The results are shown in Fig. 2. As we see, the "random" error in the data had almost no effect on the estimate of the parameters.

The number of models used to describe real random quantities can be increased substantially by using mixtures of distributions. The distribution function of a mixture of s distributions has the form

$$F(x) = \sum_{i=1}^{s} w_i F_i(x, \theta_i), \quad \sum_{i=1}^{s} w_i = 1,$$

where s is the number of distributions in the mixture, w_i are the mixture parameters, F_i is the i-th distribution function, and θ_i is the vector of its parameters. When the mixture parameters lie in the interval [0, 1], we have a classical mixtures (Fig. 3), which is obtained, e.g., by combining samples. If a parameter $\omega_i \notin [0, 1]$, then one of the distributions enters the mixture with a minus sign and, hence, is subtracted from the other distributions.

When a real sample is indeed a mixture of observations, good results are obtained when the sought law of mixture of distributions is used.

As already stated, the compatibility hypothesis is verified by the Pearson chi-square (χ^2) test, the sequential probability ratio test, the Kolmogorov test, the Smirnov test, and the Mises ω^2 and Ω^2 tests.

The parametric Pearson χ^2 test and the sequential probability ratio test provide for grouping of data. The fact that *asymptotically optimal* grouping, in addition to uniform and equiprobable grouping, is used in the software developed ensures *maximum power* of the sequential probability ratio and Pearson χ^2 tests when the competing hypotheses are competitive [1] and *minimum asymptotic variance* when the parameters are estimated from grouped data.

Construction of asymptotically optimal boundary points of intervals involves solution of the problem

$$\max_{x < x_1 < \ldots x_{K-1} < x_K} \det M_\Gamma(\theta),$$

where $M_\Gamma(\hat{\theta}) = \sum\limits_{i=1}^{K} (\nabla P_i(\hat{\theta}) \nabla^T P_i(\hat{\theta}) / (P_i(\hat{\theta})$ is the Fisher information matrix, and $P_i(\theta) = \int\limits_{x_{i-1}}^{x_i} f(x, \theta) dx$ is the probability of falling within a particular interval.

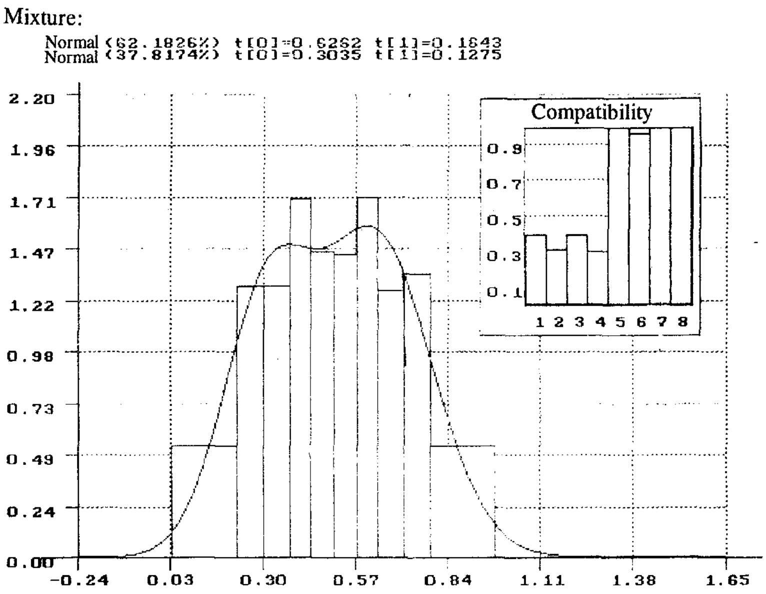


Fig. 3. Estimation of the parameters and verification of the compatibility hypothesis for a mixture of two normal distributions by means of a grouped sample.

The nonparametric Kolmogorov, Smirnov, and Mises ω^2 and Ω^2 tests are difficult to apply when the initial data constitute a grouped or partially group sample [3], since the values of the respective statistics are unknown. The following approach is proposed for such cases. The upper and lower estimates are found for each statistic and statistical conclusions are made on the basis of the upper and lower limits of the probability of compatibility.

The statistic corresponding to the Kolmogorov test has the form

$$D_n = \sup_x |F_n(x) - F(x)|,$$

where $F_n(x)$ is an empirical distribution function, $F(x)$ is the theoretical distribution function, compatibility with which is being checked, and n is the sample size.

Given a partially grouped sample, we introduce the notation

$$N_i = \sum_{j=1}^{i} n_j, \quad N_{-1} = 0, \quad N_0 = n_0, \quad N_{\kappa-1} = n, \quad N_{ij} = N_i + j.$$

The empirical distribution function $F_n(x)$ is fully defined for intervals of the second type,

$$F_n(x) = N_{i-1,j}/n, \quad \forall x \in [x_{ij}, x_{i,j+1}) \subseteq [x_i, x_{i+1}) \subseteq X_{(2)}, \quad j = 1, \ldots, n_i,$$
$$(x_{i,n_i+1} \equiv x_{i+1}),$$

and also at all boundary points x_i, $i = 0, \ldots, \kappa$:

$$F_n(x_i) = N_{i-1}/n.$$

In intervals of the first type we know only that

$$G_\kappa^-(x) = N_{i-1}/n \leq F_n(x) \leq N_i/n = G_\kappa^+(x), \quad x \in [x_i, x_{i+1}) \subseteq X_{(1)}.$$

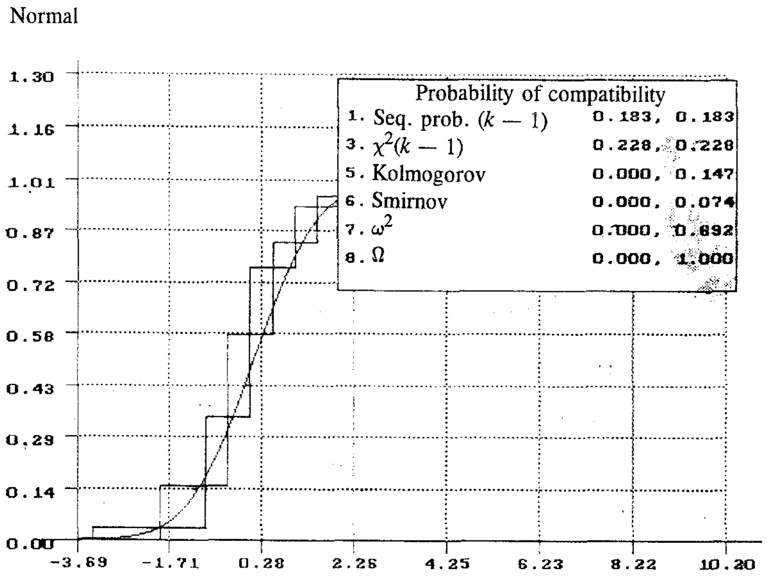


Fig. 4. Verification of the compatibility of the normal distribution with parameters $\mu = 0.1$ and $\sigma = 1.1$ by means of grouped data.

We can limit D_n from below as follows:

$$D_n = \sup_x |F_n(x) - F(x)| = \max\{\sup_{X_{(1)}} |F_n(x) - F(x)|,\ \sup_{X_{(2)}} |F_n(x) - F(x)|\},$$

$$D_n \geq \max\{\max_{i=0,\ldots,\kappa} |N_{i-1}/n - F(x_i)|,\ \sup_{X_{(2)}} |F_n(x) - F(x)|\} = \underline{D_{n\kappa}}.$$

Next we find the upper estimate. The functions $G_\kappa^+(x)$ and $G_\kappa^-(x)$ are constructed so that $\forall x \in X_{(1)}$ $G_\kappa^-(x) \leq F_n(x) \leq G_\kappa^+(x)$. Then

$$G_\kappa^-(x) - F(x) \leq F_n(x) - F(x) \leq G_\kappa^+(x) - F(x),$$

$$F(x) - G_\kappa^+(x) \leq F(x) - F_n(x) \leq F(x) - G_\kappa^-(x).$$

Next, upon denoting $A = \{x \in X_{(1)}\colon F_n(x) \geq F(x)\}$ and $B = \{x \in X_{(1)}\colon F_n(x) < F(x)\}$, we find that

$$D_n = \max\{\sup_{A \subseteq X_{(1)}} (F_n(x) - F(x)),\ \sup_{B \subseteq X_{(1)}} (F(x) - F_n(x)),\ \sup_{X_{(2)}} |F_n(x) - F(x)|\},$$

$$D_n \leq \max\{\sup_{A \subseteq X_{(1)}} (G_\kappa^+(x) - F(x)),\ \sup_{B \subseteq X_{(1)}} (F(x) - G_\kappa^-(x)),\ \sup_{X_{(2)}} |F_n(x) - F(x)|\},$$

$$D_n \leq \max\{\sup_{X_{(1)}} |G_\kappa^+(x) - F(x)|,\ \sup_{X_{(1)}} |F(x) - G_\kappa^-(x)|,$$
$$\sup_{X_{(2)}} |F_n(x) - F(x)|\} = \overline{D_{n\kappa}}.$$

Hence, $D_{n\kappa} \leq D_n \leq \overline{D_{n\kappa}}$ and, since the distribution function increases monotonically we have $p_{\min} \leq p \leq p_{\max}$, where $p = 1 - K(g(D_n))$, $p_{\min} = 1 - K(g(\overline{D_{n\kappa}}))$, $p_{\max} = 1 - K(g(\underline{D_{n\kappa}}))$, $K(\lambda)$ is the Kolmogorov distribution function, and $g(y) = \sqrt{(6ny + 1)^2/36n}$ [3].

Similar estimates were obtained for the other statistics.

The following conclusions, therefore, can be made for the tests under consideration, for the given significance level α: the compatibility hypothesis must be put aside if $p_{\max} \leq \alpha$: the compatibility hypothesis must not be rejected if $p_{\min} > \alpha$.

Figure 4 shows an example with a normal distribution. The step functions denote the upper and lower limits for the unknown empirical distribution function. For a given significance level $\alpha = 0.15$ the hypothesis of compatibility with a normal distribution having the parameters $\mu = 0.1$ and $\sigma = 1.1$ passes the Pearson χ^2, sequential probability ratio, and Mises ω^2 and Ω^2 tests and is rejected by the Kolmogorov and Smirnov tests.

REFERENCES

1. V. I. Denisov, B. Yu. Lemeshko, and E. B. Tsoi, Optimal Grouping, Parameter Estimation, and Planning of Regression Experiments [in Russian], Novosbirsk (1993).
2. G. Kulldorf, Introduction to the Theory of Estimation from Grouped and Partially Grouped Samples [Russian translation], Nauka, Moscow (1966).
3. N. L. Bol'shev and N. V. Smirnov, Mathematical Statistics Tables [in Russian], Nauka, Moscow (1983).