

ON THE THEORY OF GROUPED SAMPLES

N. A. Bodin

1968

SovietRxiv

View the original and related papers at <https://soviextrxiv.org/items/ru-196801.43920>

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.

Abstract

Full Text

UDC 519.281

MATHEMATICS

N. A. BODIN

ON THE THEORY OF GROUPED SAMPLES

(Presented by Academician Yu. V. Linnik on 4 III 1967)

In the present note a refinement and generalization to the multidimensional case is given of some results of Kulldorff ⁽¹⁾ on the theory of grouped samples. For the history of the question see ⁽¹⁾.

Let there be given a random variable with a continuous n -dimensional distribution function

$$F(\mathbf{x}) = F(\mathbf{x}, \vec{\theta}) = F(x_1, \dots, x_n; \theta_1, \dots, \theta_s), \quad (1)$$

depending on an unknown vector parameter $\vec{\theta} = (\theta_1, \dots, \theta_s)$ with s -dimensional range of variation Ω . The problem is to estimate the parameter $\vec{\theta}$ on the basis of N observations of this random variable. For this purpose the following procedure is proposed (a method of grouping observations). Let $\mathfrak{X}$ be the range of variation of the variable $\mathbf{x} = (x_1, \dots, x_n)$. This region may be infinite and, in particular, may coincide with the whole n -dimensional space. Partition the region $\mathfrak{X}$ into pairwise nonintersecting subregions $\mathfrak{X}_1, \dots, \mathfrak{X}_k$. Usually $\mathfrak{X}$ is a finite or infinite rectangular parallelepiped; the regions $\mathfrak{X}_i$ are also parallelepipeds formed by planes parallel to the faces of the parallelepiped $\mathfrak{X}$.

The function $F(\mathbf{x}, \vec{\theta})$ and this partition define a finite probability scheme with k outcomes having probabilities

$$p_i = p_i(\vec{\theta}) = p_i(\mathfrak{X}_i, \vec{\theta}) \quad (i = 1, \dots, k), \quad (2)$$

where $p_i(\mathfrak{X}_i, \vec{\theta})$ is the probability of falling in the region $\mathfrak{X}_i$. The method of grouping a sample for estimating the parameter $\vec{\theta}$ consists in reducing this problem to estimating $\vec{\theta}$ from observations in a finite probability scheme. In what follows we shall estimate the parameter $\vec{\theta}$ from observations in the finite probability scheme (2).

Consider a sample $\mathfrak{M}_N$ consisting of N observations in our finite scheme. Let N_i be the number of observations of the i -th outcome of the scheme. In the case where the p_i are determined by formulas (2), N_i is the number of observations

falling in the region $\mathfrak{X}_i$ ($i = 1, \dots, k$). To estimate the parameter $\vec{\theta}$ we use the maximum likelihood method. To the sample $\mathfrak{M}_N$ there corresponds the likelihood function

$$L = L_{\mathfrak{M}_N}(\vec{\theta}) = \prod_{i=1}^k p_i^{N_i}. \quad (3)$$

By a **maximum likelihood estimate (m.l.e.) in the strict sense** we mean such a value $\vec{\theta} \in \Omega$ at which the likelihood function $L = L(\vec{\theta})$ assumes its absolutely largest value.

Let the functions $p_i = p_i(\vec{\theta})$ have partial derivatives with respect to all arguments. Any root of the system of likelihood equations

$$\frac{\partial}{\partial \theta_t} \log L = 0 \quad (t = 1, \dots, s), \quad (4)$$

lying in Ω , will be called an m.l.e. in the **broad sense**. An m.l.e. may fail to exist, but then it can be replaced by a certain random variable

(see (1)), for which the results given below will also be valid.

Theorem 1. Let a sample $\mathfrak{M}_N$ be fixed, and suppose that, for every i ($i = 1, \dots, k$) for which $N_i > 0$, the functions $p_i = p_i(\vec{\theta}) = p_i(\theta_1, \dots, \theta_s)$ have continuous first partial derivatives in the open parallelepiped $\vec{\Omega}$ defined by the inequalities

$$\omega'_i < \theta_i < \omega''_i \quad (i = 1, \dots, s) \quad (5)$$

(possibly $\omega'_i = -\infty$, $\omega''_i = +\infty$). Suppose that there exist points $\vec{\bar{\theta}} = (\bar{\theta}_1, \dots, \bar{\theta}_s)$ and $\vec{\underline{\theta}} = (\underline{\theta}_1, \dots, \underline{\theta}_s)$ of the parallelepiped $\vec{\Omega}$ satisfying the following conditions: for each index $t = 1, \dots, s$,

$$\sum_{i=1}^k N_i \frac{\partial \log p_i}{\partial \theta_t} \bigg|_{\theta_t = \bar{\theta}_t} > 0, \quad (6a)$$

$$\sum_{i=1}^k N_i \frac{\partial \log p_i}{\partial \theta_t} \bigg|_{\theta_t = \underline{\theta}_t} < 0 \quad (6b)$$

for all $\theta_1, \dots, \theta_{t-1}, \theta_{t+1}, \dots, \theta_s$ satisfying inequalities (5). Then an m.l.e. exists in the broad sense. If, moreover, for any vector $\vec{\theta} = \vec{\theta}'$ satisfying the system of likelihood equations (4), the matrix

$$\left\| \sum_{i=1}^k N_i \frac{\partial^2 \log p_i}{\partial \theta_t \partial \theta_u} \right\|_{\vec{\theta} = \vec{\theta}'} \quad (t, u = 1, \dots, s) \quad (7)$$

is the matrix of a negative definite quadratic form, then there exists a unique m.l.e. in the strict sense.

Let $\vec{\theta}^{(0)}$ be the true value of the parameter $\vec{\theta}$. An m.l.e. (in the strict sense or, respectively, in the broad sense) $\vec{\theta}^{(\mathfrak{M}_N)}$ is called **consistent** if, as $N \rightarrow \infty$, the vector $\vec{\theta}^{(\mathfrak{M}_N)}$ converges in measure to $\vec{\theta}^{(0)}$. The m.l.e. $\vec{\theta}^{(\mathfrak{M}_N)}$ is called **asymptotically efficient** (in Wald's sense) if

$$(a) \quad \lim_{N \rightarrow \infty} E \left[\sqrt{N} (\vec{\theta}^{(\mathfrak{M}_N)} - \vec{\theta}^{(0)}) \right] = 0; \quad (8)$$

$$(b) \quad \lim_{N \rightarrow \infty} \| E \{ [\sqrt{N} (\theta_t^{(\mathfrak{M}_N)} - \theta_t^{(0)})] [\sqrt{N} (\theta_u^{(\mathfrak{M}_N)} - \theta_u^{(0)})] \} \| =$$

$$= \left\| \sum_{i=1}^k \left(p_i \frac{\partial \log p_i}{\partial \theta_t} \cdot \frac{\partial \log p_i}{\partial \theta_u} \right) \right\|_{\vec{\theta} = \vec{\theta}^{(0)}}^{-1}; \quad (9)$$

(c)

the vector random variable

$$\sqrt{N} (\vec{\theta}^{(\mathfrak{M}_N)} - \vec{\theta}^{(0)})$$

is asymptotically normal as $N \rightarrow \infty$ (the parameters of the limiting distribution are determined by the right-hand sides of (8) and (9)).

Theorem 2. Suppose that, for sufficiently large N , the system of maximum-likelihood equations (4) has a unique* solution $\vec{\theta}^{(\mathfrak{M}_N)}$, and suppose that, in some neighborhood $\vec{\Omega}_0 \subset \vec{\Omega}$ of the point $\vec{\theta}^{(0)}$, there exist second partial derivatives of the functions $p_i = p_i(\vec{\theta}) = p_i(\theta_1, \dots, \theta_s)$ ($i = 1, \dots, k$), and, moreover,

$$\sum_{i=1}^k \left(\frac{\partial p_i}{\partial \theta_t} \right)_{\theta_t = \theta_t^{(0)}}^2 \neq 0 \quad (t = 1, \dots, s); \quad (10)$$

* If, as mentioned above, we generalize the concept of an m.l.e., then it is sufficient to require that the system of equations (4) have no more than one root.

suppose, moreover, that there exists a positive differentiable function $g(\vec{\theta}) = g(\theta_1, \dots, \theta_s)$ such that the functions

$$\frac{\partial}{\partial \theta_t} \left(g(\vec{\theta}) \frac{\partial \log p_i}{\partial \theta_t} \right) \quad (i = 1, \dots, k; t = 1, \dots, s) \quad (11)$$

are continuous. Then, as $N \rightarrow \infty$, the estimate $\vec{\theta}^{\mathfrak{M}_N}$ is consistent and asymptotically efficient.

Let p_N be the probability that, for a given N , the m.l.e. exists in the strict sense. Then $p_N \rightarrow 1$ as $N \rightarrow \infty$.

As an application of the results obtained, let us consider the m.l.e. for the mathematical expectation $\vec{\mu}$ of an n -dimensional normal distribution with known covariance matrix $\vec{\Sigma}$, on the basis of a grouping of the observation results. Let $\mathfrak{X}$ coincide with n -dimensional space. Partition it into $k = k_1 \cdots k_n$ pairwise disjoint parallelepipeds ($k_i \geq 2$ ($i = 1, \dots, n$))

$$\mathfrak{X}_{\nu_1, \dots, \nu_n} \quad (\nu_1 = 1, \dots, k_1; \dots; \nu_n = 1, \dots, k_n), \quad (12)$$

defined by the inequalities

$$\begin{aligned} x_1^{(\nu_1-1)} &< x_1 \leq x_1^{(\nu_1)}, \\ &\dots\dots\dots \\ x_n^{(\nu_n-1)} &< x_n \leq x_n^{(\nu_n)}, \end{aligned} \quad (13)$$

where

$$-\infty = x_i^{(0)} < x_i^{(1)} < \dots < x_i^{(k_i)} = +\infty \quad (i = 1, \dots, n) \quad (14)$$

are certain fixed numbers. Then system (4) takes the form

$$\frac{\partial}{\partial \mu_t} \log L = 0 \quad (t = 1, \dots, n), \quad (15)$$

where

$$L = L_{\mathfrak{M}_N}(\vec{\mu}) = \prod_{\nu_1=1}^{k_1} \cdots \prod_{\nu_n=1}^{k_n} p_{\nu_1, \dots, \nu_n}^{N_{\nu_1, \dots, \nu_n}} = \prod_{\nu_1=1}^{k_1} \cdots \prod_{\nu_n=1}^{k_n} \left(\int_{\mathfrak{X}_{\nu_1, \dots, \nu_n}} f(\mathbf{x}, \vec{\mu}) d\mathbf{x} \right)^{N_{\nu_1, \dots, \nu_n}}. \quad (16)$$

where

$$f(\mathbf{x}, \vec{\mu}) = (2\pi)^{-1/2n} |\vec{\Sigma}|^{-1/2} \exp \left[-\frac{1}{2} (\mathbf{x} - \vec{\mu})^T \vec{\Sigma}^{-1} (\mathbf{x} - \vec{\mu}) \right]; \quad (17)$$

here $N_{\nu_1, \dots, \nu_n}$ is the number of observations in the sample $\mathfrak{M}_N$ falling into the region $\mathfrak{X}_{\nu_1, \dots, \nu_n}$.

Theorem 3. Suppose that for the sample $\mathfrak{M}_N$ the conditions

$$\sum_{\nu_1=1}^{k_1} \dots \sum_{\nu_{i-1}=1}^{k_{i-1}} \sum_{\nu_{i+1}=1}^{k_{i+1}} \dots \sum_{\nu_n=1}^{k_n} N_{\nu_1, \dots, \nu_{i-1}, 1, \nu_{i+1}, \dots, \nu_n} \neq N, \quad (i = 1, \dots, n) \quad (18)$$

$$\sum_{\nu_1=1}^{k_1} \dots \sum_{\nu_{i-1}=1}^{k_{i-1}} \sum_{\nu_{i+1}=1}^{k_{i+1}} \dots \sum_{\nu_n=1}^{k_n} N_{\nu_1, \dots, \nu_{i-1}, k_i, \nu_{i+1}, \dots, \nu_n} \neq N.$$

Then there exists a unique m.l.e. in the strict sense. It is computed as the unique root of system (15). This estimate, as $N \rightarrow \infty$, is consistent and asymptotically efficient.

Theorem 3 follows directly from Theorems 1 and 2 if we use the following integral inequality.

Lemma. Let $A(\mathbf{x}) = A(x_1, \dots, x_n)$ be a negative definite quadratic form, and let D be an n -dimensional parallelepiped in the space $(x_1, \dots, x_n)$, defined by the inequalities (13); let $D(\vec{\mu})$ be the result of a parallel translation of the parallelepiped D by the vector $\vec{\mu} = (\mu_1, \dots, \mu_n)$. Denote

$$p(\vec{\mu}) = \int_{D(\vec{\mu})} e^{A(\mathbf{x})} d\mathbf{x}.$$

Then the matrix

$$\left\| \frac{\partial}{\partial \mu_i \partial \mu_j} \log p(\vec{\mu}) \right\| \quad (i, j = 1, \dots, n)$$

is the matrix of a negative definite quadratic form.

In the case $n = 2$, this proposition was proved by the author and V. A. Zalgaller⁽²⁾. In the general case it was proved geometrically by V. A. Zalgaller.

The results of the study can be generalized to the case of partial grouping of samples (see⁽¹⁾). In addition to the example of the maximum-likelihood estimate $\vec{\mu}$ under the normal distribution, one may consider problems of estimating the parameters of other specific multivariate distributions.

I express my gratitude to Yu. V. Linnik for his attention to this work.

Leningrad Branch
of the V. A. Steklov Mathematical Institute
Academy of Sciences of the USSR

Received
28 II 1967

REFERENCES

1. T. Kulldorff, *Introduction to the theory of estimation from grouped and partially grouped samples*, Moscow, 1966.
2. N. A. Yudin, V. A. Zalgaller, *Lithuanian Mathematical Collection*, 7, No. 3 (1967).

Note: Figure translations are in progress. See original paper for figures.

Source: Math-Net.Ru and CyberLeninka. Machine translation. Verify with the original.