ISSN 0863–0755

WISSENSCHAFTLICHE SCHRIFTENREIHE
der Technischen Universität
Karl-Marx-Stadt 10/1989

STATISTICS FOR GROUPED OBSERVATIONS

WISSENSCHAFTLICHE SCHRIFTENREIHE

TECHNISCHE UNIVERSITÄT KARL-MARX-STADT

STATISTICS

FOR

GROUPED OBSERVATIONS

Karl-Marx-Stadt

1989

Redaktionsschluß: 27. März 1989

Herausgeber: Der Rektor der Technischen Universität Karl-Marx-Stadt
Redaktion: Wissenschaftliche Zeitschrift der Technischen Universität Karl-Marx-Stadt, Postfach 964, Karl-Marx-Stadt, 9010 DDR
Druckgenehmigungsnummer: K 105/89
Gesamtherstellung: VEB Kongreß- und Werbedruck Oberlungwitz

Summary

Statistical procedures based on continuous random variables are inappropriate if the considered random variables can be observed in a restricted manner only. Then we are forced to use so-called grouped or classified sampling schemes. On the one hand the consideration of grouped sampling schemes requires certain modifications of the corresponding traditional statistical procedures. On the other side we have to take into account that each kind of grouping effects an information loss and, as a rule, the statistical properties of grouped observation procedures change to the worse in comparison with the corresponding non-grouped methods. Then, for instance, answers are needed to the following questions. Which number of groups is necessary and how we have to allocate these groups such that we obtain still enough distribution relevant information solving our statistical problem ? The present paper provides some answers to the questions above and contains corresponding contributions to parameter estimation, parameter testing, goodness of fit tests and regression analysis.

Zusammenfassung

Statistische Verfahren für stetige verteilte Grundgesamtheiten sind ungeeignet wenn die Beobachtungen nur in eingeschränkter Form vorliegen. In solchen Situationen werden sogenannte gruppierte oder klassifizierte Beobachtungsschemata benutzt. Hierfür müssen einerseits Modifikationen der jeweiligen traditionellen statistischen Verfahren gefunden werden. Andererseits ist zu berücksichtigen, daß jede Gruppierung i.a. einen Informationsverlust verursacht, d.h. die statistischen Eigenschaften der gruppierten Verfahren verschlechtern sich gegenüber den nichtgruppierten Verfahren. Von besonderen Interesse sind dabei die folgenden Fragestellungen. Welche Anzahl von Gruppen ist notwendig und wie ist die Gruppeneinteilung vorzunehmen, um noch genügend verteilungsrelevante Information zur Lösung des vorliegenden statistischen Problems zu erhalten ?
Die vorliegende Arbeit vermittelt Antworten zu diesen Fragen und enthält entsprechende Beiträge zur Schätzung und zum Prüfen von Parametern, zu Anpassungstest und zur Regressionsanalyse.

Preface

Statistical procedures based on continuous random variables are inappropriate if the considered random variables can be observed in a restricted manner only. Such restrictions may be given by inaccurate measurements, by digitalization or quantization of a continuous observation variable and of course by certain requirements for a simplification of the observation scheme e.g. for economical or technological reasons.

In such situations we are forced to use so-called grouped or classified sampling schemes. On the one hand the consideration of grouped sampling schemes requires certain modifications of the corresponding traditional statistical procedures. On the other side we have to take into account that each kind of grouping effects an information loss and, as a rule, the statistical properties of grouped observation procedures change to the worse in comparison with the corresponding non-grouped methods. Then, for instance, answers are needed to the following questions. Which number of groups is necessary and how we have to allocate these groups such that we obtain still enough distribution relevant information solving our statistical problem ?

The present volume is intended as an introduction to statistics based on grouped observations. It provides some answers to the questions above and contains corresponding contributions to parameter estimation, parameter testing, goodness of fit tests and regression analysis.

Karl-Marx-Stadt, Novosibirsk
February 1989

K.-H. Eger , E.B. Tsoi

Contents

Estimation of unknown parameters of one-dimensional distributions with partially grouped data

by

V.I.Denisov, B.Yu.Lemeshko, E.B.Tsoi *)

S u m m a r y. The concepts of grouped, ungrouped and censored data are considered. Estimation methods of unknown parameters of one-dimensional distributions under the conditions of grouped data: method of moments, method of chi-square minimum and its modifications, maximum likelihood method are discussed. For a number of distributions the likelihood equations for determining maximum likelihood estimate (MLE) with partially grouped data, as well as expressions of approximate MLE with grouped data for the case of equidistant grouping are presented.

1. Introduction

Let continuous randon variable X be distributed on set $\mathcal{X}$ with distribution funetion $F_\theta(x)$, where θ is an unknown parameter, in general case-vector determined on the open range Θ. The set $\mathcal{X}$ is partitioned in m intervals $\mathcal{X}_1, \mathcal{X}_2, \ldots \mathcal{X}_m$ so that $\mathcal{X}_i \cap \mathcal{X}_j \neq \emptyset$ $i \neq j$, $\bigcup_{i=1}^{m} \mathcal{X}_i = \mathcal{X}$ by the boundary points $X_0^G <$ $< X_1^G < \ldots < X_{k-1}^G < X_k^G < \ldots < X_m^G$, where $X_0^G = \inf \mathcal{X}$, $X_m^G = \sup \mathcal{X}$. Let n be the sample size, n_k - the number of observations occured in the k -th interval, $\sum_{k=1}^{m} n_k = n$. Denote individual values occured in the k -th interval by $X_{k1}, X_{k2}, \ldots, X_{kn_k}$.

D e f i n i t i o n. A sample is called partially grouped, KULLDORF [1], BODIN [2], if the available information is connected with a set of non-intersecting intervals which divide the random variable range so that each interval belongs to one of the two types:

*) Novosibirsky Elektrotechnichesky Institut, Novosibirsk, 92, prospekt Karla Marksa, 20.

a) the k -th interval belongs to the first type, if the number n_k is known but the individual values X_{kj}, $(j = 1, 2, \ldots, n_k)$ are unknown;

b) the k -th interval belongs to the second type, if not only the number n_k, but also all individual values X_{kj}, $(j = 1, 2, \ldots, n_k)$ are known.

Such partially grouped sample is the initial sample for determining the estimate of the unknown distribution parameter. The concept of partially grouped sample combines grouped, ungrouped and censored samples. Thus, we have a grouped sample, if all the intervals belong to the second type, left-censored (or right-censored), if the first (m -th) interval belongs to the first type and all the rest to the second type, and we have a two-sided censored sample, if only the first and the last interval belong to the first type. Further summation and multiplication over all intervals, belonging to the first (or to the second) type, will be denoted accordingly by $\sum_{(1)}$ (or $\sum_{(2)}$) and $\prod_{(1)}$ (or $\prod_{(2)}$). Let us denote integration over all the intervals of the second type by $\int_{(2)}$. In the estimation of parameters with grouped and partially grouped samples we have to consider asymptotic properties of estimates such as consistency and asymptotic efficiency, RAO[2]. In the estimation of distribution parameters with grouped and partially grouped samples different methods may be used. Let us consider some of them.

2. Methods of estimation

Method of moments. In practice grouped samples are often considered as ungrouped ones. All the observations occured in an interval are assigned values, e.g., equal to the midpoint of the interval, and then sampling values of distributions moments are computed. Let the first S distributions moments exist and be explicitly expressed by functions $\alpha_r(\theta_1, \theta_2, \ldots, \theta_s)$, $r = 1, \ldots, S$ of unknown parameters. The sampling values of moments are computed by formulas

$$a_r = \frac{1}{n} \sum_{j=1}^{n} X_j^r = \frac{1}{n} \sum_{k=1}^{m} n_k \left(X_k^c \right)^r. \qquad (1)$$

Setting equal $\alpha_r(\theta_1, \theta_2, \ldots, \theta_s)$ by suitable sampling values derive a system of S equations

$$\alpha_r(\theta_1, \theta_2, \ldots, \theta_s) = a_r, \quad r = 1, 2, \ldots, s. \qquad (2)$$

Solving the system for $\theta_1, \ldots, \theta_s$, derive the estimates of parameters by the method of moments. Under definite conditions the method of moments reduces to consistent estimates, the equations (2) being simple in most cases. The method is unsuitable when theoretic moments of necessary order do not exist. But generally speaking, the estimates by the method of moments are inefficient. Besides, the procedure of assigning for all the observations similar values represents approximation which causes systematic errors and needs corrections. Thus, if the intervals are equal in length, the Sheppard's corrections determined by the relations

$$\begin{aligned}
a_1 &= \bar{a}_1, \\
a_2 &= \bar{a}_2 - \frac{1}{12} h^2, \\
a_3 &= \bar{a}_3 - \frac{1}{4} \bar{a}_1 h^2, \\
a_4 &= \bar{a}_4 - \frac{1}{2} \bar{a}_2 h^2 + \frac{7}{240} h^4, \\
a_5 &= \bar{a}_5 - \frac{5}{6} \bar{a}_3 h^2 + \frac{7}{48} \bar{a}_1 h^4, \\
&\ldots
\end{aligned}$$

where $\bar{a}_i$ are moments determined with grouping by formula (1), h is the length of the interval, are often used for the moments. It should be noted that introducing corrections nmt always brings about satisfactory results. On many occasions the estimate derived by means of moments appears to be further removed from the true value than the estimate without correction. The resuts are especially unsatisfactory, if one of the following situations takes place: the number of groups is few, i.e., rough grouping occurs; the range of determining the random variable is partitioned in intervals of unequal length. In spite of its disadvantages the method of moments as applied to grouped data is under way in practice. First of all it is due to the simplicity and small size of computations. It is rational to use the parameter estimates derived by the method of moments

as initial approximation when searching an estimate by more efficient methods.

Chi-square method of minimum and its modifications

These methods assume that the sample by which parameters are grouped is adequately grouped. Under the method of minimum an estimate is determined as a value of parameter, minimizing the statistic

$$\chi^2 = n \sum_{k=1}^{m} \frac{(n_k/n - p_\theta^G(k))^2}{p_\theta^G(k)},$$

where $\sum_{k=1}^{m} n_k = n$, $p_\theta^G(k)$ is the probability of observation occurence in the k-th interval. Upon using modified chi-square statistic the value is minimized

$$mod\ \chi^2 = \sum_{k=1}^{m} \frac{(n_k - n\, p_\theta^G(k))^2}{n_k},$$

where n_k is replaced with 1, if $n_k = 0$. Related statistics are Hellinger's distance

$$H.D. = \arccos \sum_{k=1}^{m} \sqrt{(n_k/n)\, p_\theta^G(k)},$$

Kullback-Leibler divergence

$$K.L.S. = \sum_{k=1}^{m} p_\theta^G(k)\, \ln [p_\theta^G(k)/(n_k/n)]$$

and Haldane's measure of divergence

$$D_j = \frac{(n+j)!}{n!} \sum_{k=1}^{m} \frac{n_k! (p_\theta^G(k))^{j+1}}{(n_k+j)!}, \quad j \neq -1,$$

$$D_{-1} = -\frac{1}{n} \sum_{k=1}^{m} n_k \ln p_\theta^G(k).$$

Under the regularity conditions all these methods give consistent and asymptotically efficient estimates [3]. But between these methods there are differences as well, appearing in view of efficiency of the second order introduced in [4]. It was shown that asymptotic estimate variance is determined by the relation

$$D_\theta^2 = \frac{1}{n\, I_F(\theta)} + \frac{\psi(\theta)}{n^2} + O\left(\frac{1}{n^2}\right),$$

where $I_F(\theta)$ is Fisher's parameter information amount, and value $\psi(\theta)$ is determined by method of estimation and non-negative. If we denote $\psi(\theta)$ for maximum likelihood method by ψ_M, then for the method of minimum χ^2 $\psi(\theta) = \psi_M + \delta$, where δ is a non-negative value, being equal to zero only in special cases for modified χ^2 $\psi(\theta) = \psi_M + 4\delta$, for Hellinger's distance $\psi(\theta) = \psi_M + 8/4$, for Kullback-Leibler divergence $\psi(\theta) = \psi_M + \delta$, for Haldane's measure of divergence $\psi(\theta) = (j+1)^2\delta + \psi_M$. Hence, in view of efficiency of the second order maximum likelihood method is the best one.

Maximum likelihood method. Likelihood function for practically grouped sample takes a form of

$$L(n,\theta) = \prod_{(1)} \left(P_\theta^G(k)\right)^{n_k} \prod_{(2)} \prod_{j=1}^{n_k} f_\theta(x_{kj}),$$

where $P_\theta^G(k)$ is the probability of observation occurence in the k-th interval, $f_\theta(x)$ is probability density function. Maximum likelihood estimate (MLE) of parameter θ is determined as such value of θ, which transforms the function $L(n,\theta)$ into absolute maximum. This MLE is usually found as a solution of likelihood equation derived by differentiation of likelihood function logarithm by θ and setting the derivative equal to zero

$$\sum_{(1)} n_k \frac{\partial \ln P_\theta^G(k)}{\partial \theta} + \sum_{(2)} \sum_{j=1}^{n_k} \frac{\partial \ln f_\theta(x_{kj})}{\partial \theta} = 0. \qquad (3)$$

In case of vectorial parameter we derive a system of likelihood equations. When solving likelihood equations, in particular with grouped data, one has to dwell upon such questions as existence of solution, its uniqueness as well as upon the fact whether this solution transforms the likelihood function into maximum. Testing the conditions of existence and uniqueness allows to give up a useless process of estimate computation, if it does not exist. During experimental research, e.g., reliability, knowledge of conditions of existence and uniqueness allows to make a decision whether to continue or to stop the experiment, if according to the data received it is impossible to find an estimate of parameter distribution. It's curious to make a note of statistic data structure the methods considered above deal with. Method of moments calls for transformation of grouped data into ungrouped ones, and only then parameters are estimated using if nece-

ssary (or possible) corrections for grouped data. On the contrary, method of chi-square minimum and related ones use only grouped data: if only individual observations are available for the researcher, they are subject to grouping. Unlike other methods, maximum likelihood method allows to determine MLE of parameters with ungrouped, partially grouped and grouped samples. Generally speaking, maximum likelihood method calls for large size of computations. In connection with this a large number of papers are devoted to deriving various approximations of MLE of distribution parameters with grouped data at limited computational expenses.

3. Exact maximum likelihood estimates

When computing maximum likelihood estimates with grouped and partially grouped samples, it's possible to face a situation, when MLE simply does not exist, and, if existing, it's then not unique. In real problems it takes place though not often. That's why before the change-over to parameter estimation with grouped data it is useful to test whether MLE exists and will be unique. In most cases the conditions of existence and uniqueness of MLE take a simple form and are easily tested. Thus, for distribution parameters of exponential, Raley's, Maxwell's, modulus of multidimensional normal distribution, scale parameter of Weibull's distribution, mathimatical expectation of normal distribution and a number of others these conditions are defined by the following Theorem.

Theorem 1. MLE of distribution parameter with partially grouped samples exists if and only if, when $\sum_{(2)} n_k > 0$ or for intervals of the first type $n_1 < n$ and $n_m < n$. Moreover, MLE is derived as one and only one solution of likelihood equation

$$\sum_{(1)} n_k \frac{\partial \ln p_\theta^G(k)}{\partial \theta} + \sum_{(2)} \sum_{j=1}^{n_k} \frac{\partial \ln f_\theta(x_{kj})}{\partial \theta} = 0, \qquad (4)$$

where $p_\theta^G(k)$ is the probability occurence in the k-th interval, $f_\theta(x)$ is function density.

Likelihood equations for some distribution are presented in table 1. Conditions for the key parameter of Weibull's distribution, scale parameters of logistic distribution, Cauchy's, parameter of Γ normal distribution and a number of others are more complex. For instance, for the key parameter of Weibull's distributi-

on with probability density function $f_\theta(x)=\frac{\theta}{\theta_1}\left(\frac{x}{\theta_1}\right)^{\theta-1}exp\left\{-\left(\frac{x}{\theta_1}\right)^{\theta}\right\}$ we have the following theorem.

Theorem 2. MLE of the key parameter of Weibull's distribution with partially grouped sample exists if and only if, when $\sum_{(2)} n_k > 0$ or for intervals of the first type one of the following conditions is satisfied:

a) at $m=2$ $n_1>n(1-e^{-1})$ for $t_1>1$ or
$n_1<n(1-e^{-1})$ for $t_1<1$;

b) at $m>2$, $n_1<n$, $n_m<n$, $n_1+n_m=n$ and
$n_1>n_m(e-1)\ln t_{m-1}/\ln t_1$ for $t_1>1$ or
$n_1<n_m(e-1)\ln t_{m-1}/\ln t_1$ for $t_1\le 1$;

c) at $m>2$, $n_1+n_m<n$ and for some k such that $t_k<1$ or $t_{k-1}>1$, $n_k>0$.

Moreover, MLE is determined as solution of likelihood equation (4). Here $t_k=(x_k^G/\theta_1)^{\theta}$.

There exists a positive probability that MLE of the key parameter is not unique, if the k-th interval, such that $t_{k-1}\le 1\le t_k$, belongs to the first type and a considerable number of observations falls within it. Conditions of existence and uniqueness of MLE for some other distributions are presented in the papers of DENISOV [5], LEMESHKO [6]. In most cases likelihood equations (3) appear too complex to rely on their explicit solution, though in particular cases of grouped sample at the number of intervals $m=2$ for certain parameters MLE with grouped data are derived in explicit form. For instance, parameter estimate of Raley's distribution with probability density function

$$f_\theta(x)=\frac{x}{\theta^2}exp\left\{-\frac{x^2}{2\theta^2}\right\},$$

where $x>0$, $\theta>0$, MLE with grouped data is determined by an expression

$$\hat{\theta}=x_1^G/\sqrt{2\ln(n_1<n_2+1)},$$

and for Weibull's distribution with probability density function

$$f_\theta(x)=\frac{\theta}{\theta_1}\left(\frac{x}{\theta_1}\right)^{\theta-1}exp\left\{-\left(\frac{x}{\theta_1}\right)^{\theta}\right\},$$

where $x>0$, $\theta,\theta_1>0$, MLE of one parameter, where the other one is known, are derived accordingly from expression

Table 1.

Likelihood equations with partially grouped samples

No	Distributions	Likelihood equations
1	2	3
1	**Exponential** $f_\theta(x) = \theta e^{-\theta x}$, $x>0,\ \theta>0$	$\frac{1}{\theta}\left\{\sum_{(1)} n_k \frac{t_k e^{-t_k} - t_{k-1} e^{-t_{k-1}}}{e^{-t_{k-1}} - e^{-t_k}} + \sum_{(2)}\sum_{j=1}^{n_k} (1 - t_{kj})\right\} = 0,$ $t_k = \theta X_k^G,\ t_{kj} = \theta X_{kj}$
2	**Raley's** $f_\theta(x) = \frac{x}{\theta^2} e^{-x^2/2\theta^2}$, $x>0,\ \theta>0$	$\frac{1}{\theta}\left\{\sum_{(1)} n_k \frac{t_{k-1}^2 e^{-t_{k-1}^2/2} - t_k^2 e^{-t_k^2/2}}{e^{-t_{k-1}^2/2} - e^{-t_k^2/2}} + \sum_{(2)}\sum_{j=1}^{n_k} (t_{kj}^2 - 2)\right\} = 0,$ $t_k = X_k^G/\theta,\ t_{kj} = X_{kj}/\theta$
3	**Maxwell's** $f_\theta(x) = \frac{2x^2}{\theta^3\sqrt{2\pi}} e^{-x^2/2\theta^2}$	$\frac{1}{\theta}\left\{\sum_{(1)} n_k \frac{t_{k-1}^3 \varphi(t_{k-1}) - t_k^3 \varphi(t_k)}{\Phi_0(t_k) - t_k \varphi(t_k) - \Phi_0(t_{k-1}) + t_{k-1}\varphi(t_{k-1})} - \sum_{(2)}\sum_{j=1}^{n_k} (t_{kj} - 3)\right\} = 0,$ $t_k = X_k^G/\theta,\ t_{kj} = X_{kj}/\theta,\ \varphi(t) = e^{-t^2/2}/\sqrt{2\pi},\ \Phi_0(t) = \int_0^t \varphi(u)\,du$

1	2	3
4	**Gamma distribution** $f_\theta(x) = \theta_1^{\theta} x^{\theta-1} e^{-\theta_1 x}/\Gamma(\theta)$, $x>0$, $\theta,\theta_1>0$	for θ $\sum\limits_{(1)} n_k \dfrac{\int_{t_{k-1}}^{t_k} t^{\theta-1}e^{-t}\ln t\,dt}{\int_{t_{k-1}}^{t_k} t^{\theta-1}e^{-t}dt} + \sum\limits_{(2)}\sum\limits_{j=1}^{n_k} \ln t_{kj} - n\psi(\theta) = 0$, $t_k = \theta_1 x_k^G$, $t_{kj} = \theta_1 / x_{kj}$, $\psi(\theta) = \Gamma'(\theta)/\Gamma(\theta)$, for θ_1 $\frac{1}{\theta}\left\{\sum\limits_{(1)} n_k \dfrac{t_k^{\theta} e^{-t_k} - t_{k-1}^{\theta} e^{-t_{k-1}}}{\int_{t_{k-1}}^{t_k} t^{\theta-1}e^{-t}dt} + \sum\limits_{(2)}\sum\limits_{j=1}^{n_k}(\theta - t_{kj})\right\} = 0$
5	**Normal** $f_\theta(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\dfrac{(x-\mu)^2}{2\sigma^2}\right\}$, $x\in(-\infty,\infty)$, $\sigma>0$, $\mu\in(-\infty,\infty)$	for μ $\frac{1}{\sigma}\left\{\sum\limits_{(1)} n_k \dfrac{\varphi(t_{k-1}) - \varphi(t_k)}{\Phi(t_k) - \Phi(t_{k-1})} + \sum\limits_{(2)}\sum\limits_{j=1}^{n_k} t_{kj}\right\} = 0$, $t_k = (x_k^G - \mu)/\sigma$, $t_{kj} = (x_{kj} - \mu)/\sigma$, $\varphi(t) = \dfrac{1}{\sqrt{2\pi}} e^{-t^2/2}$, $\Phi(t) = \int_{-\infty}^{t} \varphi(u)\,du$, for σ $\frac{1}{\sigma}\left\{\sum\limits_{(1)} n_k \dfrac{t_{k-1}\varphi(t_{k-1}) - t_k\varphi(t_k)}{\Phi(t_k) - \Phi(t_{k-1})} + \sum\limits_{(2)}\sum\limits_{j=1}^{n_k}(t_{kj} - 1)\right\} = 0$.

$$\hat{\theta} = \ln(\ln(n_1/n_2 + 1)) / \ln(X_1^G/\theta_1),$$
$$\hat{\theta}_1 = X_1^G / \{\ln(n_1/n_2+1)\}^{1/\theta}.$$

In general cases MLE are derived as a result of solving likelihood equations or a system of likelihood equations by numerical methods.

To solve likelihood equations, methods of searching a solution on an interval by the type of dichotomy, "golden section", Fibonacci's numbers, etc., may be applied. But Newton's method or its modifications are the most preferable ones, HIMMELBLAU [7]. In cases when density function is defined by more than one parameter there arises a necessity to solve a system of nonlinear equations. Though MLE of vector of distribution parameters may be found by direct search of likelihood function maximum or its logarithm as well. Both in the first case and in the second one it is more preferable to use the searching methods of minimization of functions of many variables.

4. Approximate maximum likelihood estimates

There are many papers, e.g., HARTLEY [8], LINDLEY [9], TALLIS [10], in which computation of approximate maximum likelihood estimates are considered. Moreover, the technique first described in the parer of LINDLEY [9] as applied to normal and gamma distributions, then generalized for multidimensional distributions in the paper of TALLIS [10] mostly gained ground. It consists in replacing the initial grouped sample by an ungrouped one, in which individual values are assigned values of interval centre of grouping at their equal length, and then expressions for corrections derived in research of such procedure of estimates are deduced.

Here we confine the discussion to the case of one-dimentional probability density $f_\theta(x)$, having continuous derivatives up to the third order included, scalar parameter θ and grouping of data, in which all intervals have equal length ΔX, such that

$$\bigcup_{k=1}^{m} \mathcal{X}_k = \bigcup_{k=1}^{m} (\bar{X}_k^G - \Delta X/2,\ \bar{X}_k^G + \Delta X/2) = \mathcal{X},$$

where $\bar{X}_k^G = (X_{k-1}^G + X_k^G)/2$ is the value of the midpoint of the k-th interval $\mathcal{X}_k$, $k=\overline{1,m}$.

Then the probability of observation occurence in the k -th interval will be

$$p_\theta^G(k) = \int_{x_{k-1}^G}^{x_k^G} f_\theta(x)dx = \int_{\bar{x}_k^G - \Delta x/2}^{\bar{x}_k^G + \Delta x/2} f_\theta(x)dx \qquad (5)$$

Let us expand (5) into Taylor's series in the neighbourhood of $\bar{x}_k^G$, confining the discussion to the terms of $O((\Delta x)^3)$- order infinitesimal. We derive

$$p_\theta^G(k) = \Delta x \cdot f_\theta(\bar{x}_k^G) + (\Delta x)^3 \cdot f_\theta''(\bar{x}_k^G)/24 + (\Delta x)^5 R,$$

where $f_\theta''(x) = \partial^2 f_\theta(x)/\partial x^2$, R is the remaining term. It is obvious that

$$\ln p_\theta^G(k) = \ln[\Delta x \cdot f_\theta(\bar{x}_k^G)] + \ln\left[1 + \frac{(\Delta x)^2}{24}\frac{f_\theta''(\bar{x}_k^G)}{f_\theta(\bar{x}_k^G)} + \frac{(\Delta x)^4 R}{f_\theta(\bar{x}_k^G)}\right].$$

Hence,

$$\frac{\partial \ln p_\theta^G(k)}{\partial\theta} = \frac{\partial}{\partial\theta}\ln f_\theta(\bar{x}_k^G) + \frac{(\Delta x)^2}{24}\frac{\partial}{\partial\theta}\left(\frac{f_\theta''(\bar{x}_k^G)}{f_\theta(\bar{x}_k^G)}\right) + O((\Delta x)^3).$$

And the likelihood equation takes the form of

$$\sum_{k=1}^m n_k \frac{\partial \ln f_\theta(\bar{x}_k^G)}{\partial\theta} + \frac{(\Delta x)^2}{24}\sum_{k=1}^m n_k \frac{\partial}{\partial\theta}\left(\frac{f_\theta''(\bar{x}_k^G)}{f_\theta(\bar{x}_k^G)}\right) + O((\Delta x)^3) = 0.$$

Let θ^0 be the maximum likelihood estimate computed on the basis of ungrouped data, which are centres of intervals $\bar{x}_k^G$ as observed values . Then

$$\sum_{k=1}^m n_k \frac{\partial \ln f_{\theta^0}(\bar{x}_k^G)}{\partial\theta} \equiv 0$$

and the unknown approximate maximum likelihood estimate $\hat{\theta}_n$ is derived as $\hat{\theta}_n = \theta^0 + \Delta\theta$, where $\Delta\theta$ is a correction for grouping computed as a result of one iteration of Newton's method, i.e.,

$$\Delta\theta = -\frac{(\Delta x)^2}{24}\left[\sum_{k=1}^m n_k \frac{\partial}{\partial\theta}\left(\frac{f_\theta''(\bar{x}_k^G)}{f_\theta(\bar{x}_k^G)}\right)\right] \Big/ \sum_{k=1}^m n_k \frac{\partial \ln f_{\theta^0}(\bar{x}_k^G)}{\partial\theta}$$

or

$$\Delta\theta = -\frac{(\Delta x)^2}{24}\frac{E_\theta\, \partial[f_{\theta^0}''(x)/f_{\theta^0}(x)]/\partial\theta}{E_\theta\, \partial^2 \ln f_{\theta^0}(x)/\partial\theta^2},$$

where "E" is as formely a symbol of mathematical expectation. Asymptotic variance of estimate $\hat{\theta}_n$ will be found as

$$D(\theta) = n^{-1}\left[I - (\Delta x)^2/24 \cdot E_\theta \left\{ \partial^2 A / \partial \theta^2 + A\, \partial^2 \ln f_\theta(x)/\partial\theta^2 - \right.\right.$$

$$\left.\left. - f_\theta^{-1}(x) \cdot \Lambda_x^2 \left[f_\theta(x) \cdot \partial^2 \ln f_\theta(x)/\partial\theta^2 \right] \right\} \right]^{-1}, \qquad (6)$$

where

$$I = - E_\theta \left\{ \partial^2 \ln f_\theta(x)/\partial\theta^2 \right\},$$

$$A = f''_\theta(x)/f_\theta(x), \quad \Lambda_x = \partial/\partial x .$$

E x a m p l e. For exponential distribution with the density function

$$f_\theta(x) = \theta e^{-\theta x}, \quad x>0, \ \theta>0$$

maximum likelihood estimate θ for ungrouped data will take the form of

$$\hat{\theta} = \left[\frac{1}{n} \sum_{j=1}^{n} x_j \right]^{-1}.$$

Then as θ° let us take

$$\theta^\circ = \left[\frac{1}{n} \sum_{k=1}^{m} n_k \bar{x}_k^G \right]^{-1}. \qquad (7)$$

Now let us find the value of correction $\Delta\theta$. We have

$$\ln f_\theta(x) = \ln\theta - \theta x,$$

$$\partial \ln f_\theta(x)/\partial\theta = 1/\theta - x,$$

$$\partial^2 \ln f_\theta(x)/\partial\theta^2 = -1/\theta^2,$$

$$\partial f_\theta(x)/\partial x = -\theta^2 \exp(-\theta x),$$

$$\partial^2 f_\theta(x)/\partial x^2 = \theta^3 \exp(-\theta x).$$

Hence, $E_\theta \left\{ \partial^2 \ln f_\theta(x)/\partial\theta^2 \right\} = -1/\theta^2$,

$$\frac{(\Delta x)^2}{24} E_\theta \left\{ \frac{\partial}{\partial\theta} \left(\frac{f''_\theta(x)}{f_\theta(x)} \right) \right\} = \frac{\theta(\Delta x)^2}{12}.$$

Table 2

Approximate maximum likelihood estimates with grouped data

No.	Distribution	Estimate	Estimate variance
1	2	3	4
1	**Weibull's** $f_\theta(x)=\frac{\eta}{\theta}\left(\frac{x}{\theta}\right)^{\eta-1}e^{-\left(\frac{x}{\theta}\right)^{\eta}}$, $x>0,\ \theta>0,\ \eta>0$	$\left(\frac{1}{n}\sum_{k=1}^{m}n_k\left(\bar{X}_k^G\right)^{\eta}\right)^{1/\eta}-\frac{(\Delta x)^2(\eta-1)\Gamma\left(2-\frac{2}{\eta}\right)}{24\left(\frac{1}{n}\sum_{k=1}^{m}n_k\left(\bar{X}_k^G\right)^{\eta}\right)^{1/\eta}}$ with $\eta>1$	$\frac{1}{n}\left(\frac{\theta}{\eta}\right)^2\left[1-\frac{(\Delta x)^2\eta(\eta-1)\Gamma\left(2-\frac{2}{\eta}\right)}{6\theta^2}\right]^{-1}$
2	**Gamma** $f_\theta(x)=\frac{x^{\alpha-1}}{\theta^{\alpha}\Gamma(\alpha)}e^{-x/\theta}$, $x>0,\ \alpha>0,\ \theta>0$	$\frac{1}{n}\sum_{k=1}^{m}n_k\bar{X}_k^G$	$\frac{\theta^2}{n\alpha}\left[1-\frac{(\Delta x)^2}{12\alpha\theta^2}\right]^{-1}$
3	**Laplace's (double exponential)** $f_\theta(x)=\frac{\theta}{2}e^{-\theta\lvert x-\mu\rvert}$, $x\in(-\infty,\infty),\ \theta>0,$ $\mu\in(-\infty,\infty)$	$\frac{n}{\sum_{k=1}^{m}n_k\lvert\bar{X}_k^G-\mu\rvert}+\frac{(\Delta x)^2}{12}\left(\frac{n}{\sum_{k=1}^{m}\lvert\bar{X}_k^G-\mu\rvert}\right)^3$	$\frac{\theta^2}{n}\left[1-\frac{(\Delta x)^2\theta^2}{12}\right]^{-1}$
4	**Maxwell's** $f_\theta(x)=\frac{2x^2}{\sqrt{2\pi}\theta^3}e^{-x^2/2\theta^2}$, $x>0,\ \theta>0$	$\sqrt{\frac{1}{3n}\sum_{k=1}^{m}n_k\left(\bar{X}_k^G\right)^2}-\frac{(\Delta x)^2}{72\sqrt{\frac{1}{3n}\sum_{k=1}^{m}n_k\left(\bar{X}_k^G\right)^2}}$	$\frac{\theta^2}{6n}\left[1-\frac{(\Delta x)^2}{18\theta^2}\right]$

End of table 2

1	2	3	4
5	**Multidimentional normal vector modulus** $f_\theta(x)=\frac{2x^{r-1}\exp(-x^2/2\theta^2)}{(2\theta^2)^{r/2}\,\Gamma(r/2)}$, $x>0$, $\theta>0$, $r=1,2,\ldots$	$\sqrt{\frac{1}{rn}\sum_{k=1}^{m} n_k(\bar{X}_k^G)^2}-\frac{(\Delta X)^2}{24r\sqrt{\frac{1}{rn}\sum_{k=1}^{m} n_k(\bar{X}_k^G)^2}}$	$\frac{\theta^2}{2rn}\left[1-\frac{(\Delta X)^2}{6r\theta^2}\right]^{-1}$
6	**Least extreme value** $f_\theta(x)=\frac{1}{\sigma}\exp\left\{\frac{x-\theta}{\sigma}-\exp\left[\frac{x-\theta}{\sigma}\right]\right\}$, $x\in(-\infty,\infty)$, $\sigma>0$, $\theta\in(-\infty,\infty)$	$\sigma\ln\left[\frac{1}{n}\sum_{k=1}^{m} n_k\exp(\bar{X}_k^G/\sigma)\right]-\frac{(\Delta X)^2}{24}$	$\frac{\sigma^2}{n}\left[1-\frac{(\Delta X)^2}{6\sigma^2}\right]^{-1}$
7	**Exponential** $f_\theta(x)=\theta\exp(-\theta x)$, $x>0$, $\theta>0$	$\frac{n}{\sum_{k=1}^{m} n_k\bar{X}_k^G}+\frac{(\Delta X)^2}{12}\left(\frac{n}{\sum_{k=1}^{m} n_k\bar{X}_k^G}\right)^3$	$\frac{\theta^2}{n}\left[1-\frac{(\Delta X)^2\theta^2}{12}\right]^{-1}$

Finally we derive that $\Delta\theta = (\Delta X)^2\theta^3/12$. So $\hat{\theta}_n = \theta^0 + (\Delta x)^2(\theta^0)^3/12$ and expression for approximate maximum likelihood estimate involves Sheppard's correction. Or, remembering relation (7), we derive that

$$\hat{\theta}_n = \left[\frac{1}{n}\sum_{k=1}^{m} n_k \bar{X}_k^G\right]^{-1} \cdot \left\{1 + \left[\frac{1}{n}\sum_{k=1}^{m} n_k \bar{X}_k^G\right]^{-2} \frac{(\Delta x)^2}{12}\right\}.$$

Using (6) we derive $D(\hat{\theta}_n) = n^{-1}\theta^2\left[1 - (\Delta X)^2\theta^2/12\right]^{-1}$.

And if we let the number of intervals tend to infinity, $m \to \infty$, which is equivalent to $\Delta X \to 0$, and change from grouped data to ungrouped ones, then $\hat{\theta}_n \to \hat{\theta}$, and $D(\hat{\theta}_n) \to D(\theta) = n^{-1}\theta^2$, which is in agreement with the estimate and variance of maximum likelihood estimate with ungrouped data. Approximate maximum likelihood estimates computed for a number of distributions are compiled in table 2.

References

[1] KULLDORF,G. (1966) Vvedeniye v teoriyu otsenivaniya po gruppirovannym i chastichno gruppirivannym vyborkam. Nauka, Moskva.-176 s.

[2] BODIN,N.A. (1970) Otsenka parametrov raspredeleniya po gruppirovannym vyborkam. Tr.matem.instituta im.V.A.Steklova AN SSSR- t. 111.-s.110-154

[3] RAO,C.R. (1962) Lineinye statisticheskiye metodi i ikh primeneniye. Nauka, Moskva.-548s.

[4] RAO,C.R. (1962) Criteria of estimation in large samples. Sankhus, vol.25,-p.189-206

[5] DENISOV,V.I. (1977) Matematicheskoye obespecheniye sistem EVM-experimentator. Nauka, Moskva. - 251 s.

[6] LEMESHKO,B.Yu. (1977) Otsenivaniye parametrov po gruppirovannym nablyudeniyam. Voprosi kibernetiki. Moskva,Vip.30 s. 80-90

[7] HIMMELBLAU,D. (1975) Pricladnoye nelineinoye programmirovaniye. Mir, Moskva.-535 s.

[8] HARTLEY,H.O. (1950) A simplified form of Sheppard's correction formulae. Biometrika, vol.37.-p.145-148

[9] LINDLEY,D.V. (1950) Grouping corrections and maximum likelihood equations. Proceedings of the Cambridge Philisophical Society. Vol.46, No.1 - p.106-110

[10] TALLIS,G.M. (1967) Approximate maximum likelihood estimates from grouped data. Technometrics, vol.9,No.3- p.599-606