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

Contents

Optimal Grouping in Estimation and Tests of Goodness-of-fit Hypotheses

by

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

S u m m a r y. A problem of asymptotically optimal grouping in estimation of distribution parameters by maximum-likelihood method with grouped data is considered. Involvement of goodness-of-fit tests with Fisher's information matrix from grouped data is discussed. It is shown that solution of the problem of asymptotically optimal grouping makes available maximum efficiency of tests with close competing hypotheses. Examples of the use of tables of optimal grouping in estimation and tests of hypotheses are considered.

1. Introduction

Asymptotic variance matrix of maximum likelihood estimate (MLE) with grouped data is defined by the relation, KULLDORF [1], BODIN [2],

$$D(\theta) = \left[n\, I_F^G(\theta) \right]^{-1},$$

where

$$I_F^G(\theta) = \sum_{k=1}^{m} \frac{\nabla P_\theta^G(k) \nabla^T P_\theta^G(k)}{P_\theta^G(k)} = \left[\sum_{k=1}^{m} \frac{\partial \ln P_\theta^G(k)}{\partial \theta_i} \frac{\partial \ln P_\theta^G(k)}{\partial \theta_j} P_\theta^G(k) \right]$$

is Fisher's information matrix from grouped data,

$$P_\theta^G(k) = \int_{x_{k-1}^G}^{x_k^G} f_\theta(x)\, dx$$

is observation occurence probability in an interval. Elements of this matrix depend on boundary points of intervals. In case, when distribution function is determined by one parameter or estimation of only one distribution parameter of probabilities is carried out, with other parameters known, the aim of the problem of

*) Novosibirsky Elektrotechnichesky Institut,
Novosibirsk, 92, prospekt Karla Marxa, 20

asymptotically optimal grouping is maximization of MLE asymptotic variance from grouped data. And this problem reduces to maximization of Fisher's amount of information about parameter with grouped sample, i.e., to solution of a problem

$$\max_{X^G_{k-1} < X^G_k,\ k=\overline{1,m}} I^G_F(\theta) = \max_{X^G_{k-1} < X^G_k,\ k=\overline{1,m}} \sum_{k=1}^{m} \left(\frac{\partial \ln P^G_\theta(k)}{\partial \theta} \right)^2 P^G_\theta(k). \quad (1)$$

Construction of optimal partition is a problem of design of experiments, and its solution is not a simple one because the amount of lost information for a given partition is a function of unknown parameter θ. Problem (1) for mathematical expectation of normal distribution was solved in the paper of COX [3], for parameters of normal and exponential distribution in the paper of KULLDORF [1]. In estimation of vector of parameters we deal with information matrix. Under such conditions the problem of asymptotically optimal grouping was not considered before. In this case various functionals in asymptotical variance matrix, e.g., the same as applied in theory of optimal design of regression experiments, DENISOV [4], may be chosen as tests of optimality. In this paper optimization is carried out resulting from minimum of generalized asymptotic estimate variance, i.e., determinant of asymptotic variance matrix is minimized or, which is the same, determinant of information matrix in terms of boundary points is maximized

$$\max_{X^G_{k-1} < X^G_k,\ k=\overline{1,m}} \det I^G_F(\theta) = \max_{X^G_{k-1} < X^G_k,\ k=\overline{1,m}} \det \left[\sum_{k=1}^{m} \frac{\nabla P^G_\theta(k)\, \nabla^T P^G_\theta(k)}{P^G_\theta(k)} \right]. \quad (2)$$

Other tests can be used as well. In all cases losses of information from grouped data will be minimized in various ways.

2. Connection of tests of goodness-of-fit hypotheses with information matrix

Fisher's information serves as a measure of interior proximity of random variable distributions, and this interior nature is connected with power of differences between close values of parameter. Statistic reduces sampling data, and therefore power of dif-

ferences with the help of statistic is no more than with the help of the whole sample. It means that, if it is necessary to chose among some statistics, it should be preferable to chose that one, for which losses of Fisher's information are minimal. Thus, with growth of information losses because of grouping power of tests connected with grouping of initial data reduces as well.

Let us consider how Fisher's information matrix is connected with Pearson chi-square test. Statistic

$$\chi^2 = n \sum_{k=1}^{m} (n_k/n - P_\theta^G(k))^2 / P_\theta^G(k), \qquad (3)$$

where $P_\theta^G(k)$ is hypothetic probability of observation occurence in the k -th interval, within the limit, is subject to chi-square distribution with $m-1$ degree of freedom, if the null hypothesis is right, and it is subject to noncentral chi-square distribution with the same number of degrees of freedom and parameter of non-centrality

$$\lambda = n \sum_{k=1}^{m} (P_{\theta_1}^G(k) - P_\theta^G(k))^2 / P_\theta^G(k),$$

if the competing hypothesis is right and the sample agrees with distribution of the same type but with parameter θ_1 (in general case vectorial). Let $\theta_1 = \theta + \Delta\theta$. Expanding $P_{\theta_1}^G(k)$ into Taylor's series and ignoring the terms of the higher order we derive

$$\lambda \approx n \sum_{k=1}^{m} \frac{[P_\theta^G(k) - \nabla^T P_\theta^G(k)\Delta\theta - P_\theta^G(k)]^2}{P_\theta^G(k)} = n \sum_{k=1}^{m} \frac{\Delta\theta^T \nabla P_\theta^G(k)\, \nabla^T P_\theta^G(k) \Delta\theta}{P_\theta^G(k)} =$$

$$= n\, \Delta\theta^T \left(\sum_{k=1}^{m} \frac{\nabla P_\theta^G(k)\, \nabla^T P_\theta^G(k)}{P_\theta^G(k)} \right) \Delta\theta = n\, \Delta\theta^T I_F^G(\theta)\, \Delta\theta. \qquad (4)$$

The power of Pearson chi-square test is the non-decreasing function in λ. Matrix of information losses caused by grouping $\Delta I = I_F(\theta) - I_F^G(\theta)$, where $I_F(\theta)$ is Fisher's information matrix for ungrouped observations, is a non-negatively definite one and hence $\Delta\theta^T \Delta I \Delta\theta \geq 0$. And as $\Delta\theta^T I_F^G(\theta) \Delta\theta = \Delta\theta^T I_F(\theta) \Delta\theta - \Delta\theta^T \Delta I \Delta\theta$, it is clear that with the growth of information losses the power

of test under close alternative hypotheses decreases as well. These losses may be reduced by fitting the boundary points so that $I_F^G(\theta)$ would tend to $I_F(\theta)$. Thus, in the given case we came to the same problem of asymptotically optimal grouping as in the estimation of parameter. Partition of the range, in which sampling values of random variables occured in intervals of equal space with the following combination of adjacent intervals, if a small number of observations occured in them, or partition of the range of definition of random variable in intervals of equal probability, and these procedures are basically applied in practice, in general case are far from optimal ones.

Not only for Pearson chi-square test, but also for a number of other statistics used in tests of hypotheses, suitable measures of proximity of distributions are directly defined by Fisher's information matrix with grouped data and grow with decreasing of information losses from grouping, and hence, powers of suitable tests increase. Statistic of likelihood relation for goodness-of-fit tests with certain distribution takes a form of

$$\ell = n^n \prod_{k=1}^{m} \left(P_\theta^G(k)/n_k \right)^{n_k} = \prod_{k=1}^{m} \left(\frac{P_\theta^G(k)}{n_k/n} \right)^{n_k},$$

where $P_\theta^G(k)$ are hypothetic probabilites of observation occurence in the k -th interval. Goodness-of-fit hypothesis is rejected, if ℓ is sufficiently small. Exact distribution of this statistic is unknown. However, if the null hypothesis is right, then with $n \to \infty$ the variable $-2\ln\ell$ is distributed asymptotically as chi-square with $m-1$ degrees of freedom. Moreover, in this case statistic $-2\ln\ell$ is asymptotically equivalent to chi-square statistic If the competing hypothesis is right and the sample belongs to the analogous distribution but with parameter θ_1, then statistic

$$-2\ln\ell = 2n \sum_{k=1}^{m} P_{\theta_1}^G(k) \ln \left(\frac{P_{\theta_1}^G(k)}{P_\theta^G(k)} \right)$$

is the measure of proximity of distributions considered. With its growth the power of the test increases. Having denoted $\theta_1 = \theta + \Delta\theta$, expand $P_{\theta_1}^G(k)$ into Taylor's series, ignoring the terms of the higher order; we have

$$-2\ln l \approx 2n\sum_{k=1}^{m}\left(P_\theta^G(k)+\nabla^T P_\theta^G(k)\Delta\theta\right)\ln\left(1+\frac{\nabla^T P_\theta^G(k)\Delta\theta}{P_\theta^G(k)}\right).$$

Further, expanding $\ln(1+x)$ by Taylor's formula and ignoring the terms which are higher than the second order, we obtain

$$-2\ln l \approx 2n\sum_{k=1}^{m}\left(P_\theta^G(k)+\nabla^T P_\theta^G(k)\Delta\theta\right)\left[\frac{\nabla P_\theta^G(k)\Delta\theta}{P_\theta^G(k)}-\frac{\Delta\theta^T\nabla P_\theta^G(k)\nabla^T P_\theta^G(k)\Delta\theta}{2\left(P_\theta^G(k)\right)^2}\right]\approx$$

$$\approx 2n\sum_{k=1}^{m}\left[\Delta\theta^T\nabla P_\theta^G(k)+\frac{\Delta\theta^T\nabla P_\theta^G(k)\nabla^T P_\theta^G(k)\Delta\theta}{2P_\theta^G(k)}\right]=$$

$$=n\Delta\theta^T\left[\sum_{k=1}^{m}\frac{\nabla P_\theta^G(k)\nabla^T P_\theta^G(k)}{P_\theta^G(k)}\right]\Delta\theta=n\Delta\theta^T I_F^G(\theta)\Delta\theta. \quad (5)$$

Thus, we obtained an expression analogous with (4). Statistic of modified chi-square test in defined by the expression

$$mod\,\chi^2=\sum_{k=1}^{m}\left(n_k-nP_\theta^G(k)\right)^2/n_k,$$

where n_k is replaced with 1 if $n_k=0$. Let us consider how the given measure depends on Fisher's information matrix. Applying, as before, expansion into Taylor's series and ignoring the terms with $\Delta\theta$ higher than the second order, we obtain

$$mod\,\chi^2\approx n\sum_{k=1}^{m}\frac{\left(P_\theta^G(k)+\Delta\theta^T\nabla P_\theta^G(k)-P_\theta^G(k)\right)^2}{P_\theta^G(k)+\Delta\theta^T\nabla P_\theta^G(k)}=$$

$$=n\sum_{k=1}^{m}\frac{\Delta\theta^T\nabla P_\theta^G(k)\nabla^T P_\theta^G(k)\Delta\theta}{P_\theta^G(k)\left[1+\Delta\theta^T\nabla P_\theta^G(k)/P_\theta^G(k)\right]}.$$

Futher, using the expansion into $(1+x)^{-1}$ series we have

$$mod\,\chi^2\approx n\Delta\theta^T\left(\sum_{k=1}^{m}\frac{\nabla P_\theta^G(k)\nabla^T P_\theta^G(k)}{P_\theta^G(k)}\right)\Delta\theta=n\Delta\theta^T I_F^G(\theta)\Delta\theta, \quad (6)$$

expression analogous with(4) and (5), i.e., this measure also increases with decreasing of information losses from grouping. Statistic of Hellinger's distance takes a from of

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

Operating analogous with the above and expanding $P^G_{\theta_1}(k)$ into Taylor's series by ignoring the terms of the higher, order we have

$$H.D. \approx \arccos \sum_{k=1}^{m} \sqrt{(P^G_\theta(k) + \Delta\theta^T \nabla P^G_\theta(k))\, P^G_\theta(k)} =$$

$$= \arccos \sum_{k=1}^{m} P^G_\theta(k) \sqrt{1 + \Delta\theta^T \nabla P^G_\theta(k) / P^G_\theta(k)}.$$

Futher, using expansion in series for $\sqrt{1+X}$, we obtain

$$H.D. \approx \arccos \sum_{k=1}^{m} P^G_\theta(k) \Big[1 + \frac{1}{2} \Delta\theta^T \nabla P^G_\theta(k) / P^G_\theta(k) -$$

$$- \frac{1}{8} \frac{\Delta\theta^T \nabla P^G_\theta(k)\, \nabla^T P^G_\theta(k) \Delta\theta}{(P^G_\theta(k))^2} \Big] = \arccos \Big[1 - \frac{1}{8} \Delta\theta^T I^G_F(\theta) \Delta\theta \Big]. \quad (7)$$

From (7) it is obvious that with the decrease of information losses in grouping Hellinger's distance increases.

Kulback-Leibler divergence statistic is defined by the expression

$$K.L.S. = \sum_{k=1}^{m} P^G_\theta(k)\, \ln \left(P^G_\theta(k) / (n_k/n) \right).$$

In the similar way we derive

$$K.L.S. \approx \sum_{k=1}^{m} P^G_\theta(k)\, \ln \left[\frac{P^G_\theta(k)}{P^G_\theta(k) + \nabla^T P^G_\theta(k) \Delta\theta} \right] = - \sum_{k=1}^{m} P^G_\theta(k)\, \ln \left[1 - \frac{\nabla^T P^G_\theta(k) \Delta\theta}{P^G_\theta(k)} \right].$$

Futher, using expansion for $\ln(1+x)$ we obtain

$$K.L.S. \approx - \sum_{k=1}^{m} P^G_\theta(k) \left[\frac{\nabla^T P^G_\theta(k) \Delta\theta}{P^G_\theta(k)} - \frac{1}{2} \frac{\Delta\theta^T \nabla P^G_\theta(k)\, \nabla^T P^G_\theta(k) \Delta\theta}{(P^G_\theta(k))^2} \right] =$$

$$= \frac{1}{2} \Delta\theta^T I^G_F(\theta) \Delta\theta. \quad (8)$$

Hence, it obvious that by minimizing the information losses in grouping we increase this measure as well.

Relations (4-8) show that the problems of asymptotically optimal grouping (1-2) increase the quality of statistic inferences in all cases considered. Solution of the problem of asymptotically optimal grouping is obtained for a series of continuous distributions most often used in practice: exponential, normal, Weibull's, Raley's, Maxwell's, Gauchy's, logistic, extreme values, double exponental, Laplace's distributions, modulus of multidimentional normal distribution, gamma-distribution.

The practical value of the tables obtained consists in the fact that in most cases it has become possible to derive the solution in an invariant form about parameters of distributions.

Example 1. Density function of exponential distribution is described by an expression $f_\theta(x) = \theta \exp(-\theta x)$, where $x>0$, $\theta > 0$. Fisher's information amount about parameter θ with grouped data is

$$I_F^G(\theta) = \frac{1}{\theta^2} \sum_{k=1}^{m} \frac{(t_k e^{-t_k} - t_{k-1} e^{-t_{k-1}})^2}{e^{-t_{k-1}} - e^{-t_k}},$$

where $t_k = \theta X_k^G$, and with ungrouped ones $I_F(\theta) = E_\theta(\partial \ln f_\theta(x)/\partial\theta) = 1/\theta^2$. In table 1 optimal boundary points, maximizing optimal asymptotic information, being equal to $A = I_F^G(\theta)/I_F(\theta)$, are presented. Table 2 presents suitable values of optimal probabilities.

Let us note that value A allows to make inferences on the quality of the grouping carried out. Computations of values of relative asymptotic information A with partition in intervals of equal probability have been carried out for comparison. It turned out that with m =10 A=0,8928, and with m =20 A=0,9462, as value A=0,9798 agrees with optimal grouping with m =10.

Example 2. Density function of normal distribution is described by an expression

$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

where $x \in (-\infty,\infty)$, $\sigma>0$. Asymptotically optimal boundary points of grouping intervals, maximizing the determinant of Fisher's matrix with grouped data

$$I_F^G = \begin{bmatrix} I_F^G(\mu) & I_F^G(\mu,\sigma) \\ I_F^G(\mu,\sigma) & I_F^G(\sigma) \end{bmatrix},$$

where

$$I_F^G(\mu) = \frac{1}{\sigma^2} \sum_{k=1}^{m} \frac{(\varphi(t_k) - \varphi(t_{k-1}))^2}{\Phi(t_k) - \Phi(t_{k-1})},$$

where

$$t_k = (X_k^G - \mu)/\sigma, \quad \varphi(t) = e^{-t^2/2}/\sqrt{2\pi}, \quad \Phi(t) = \int_{-\infty}^{t} \varphi(u)\,du,$$

Table 1

Optimal boundary points of intervals of grouping for estimation of parameters of exponential distribution in the form of $t_k = \theta x_k^{\epsilon}$ and tests of hypotheses with Pearson chi square tests and suitable values of relative asymptotic information A.

k	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}	A
2	I.5936										0.6476
3	I.0I76	2.6II2									0.8203
4	0.754I	I.77I6	3.3652								0.89I0
5	0.6004	I.3545	2.3720	3.9657							0.9269
6	0.4993	I.0997	I.8538	2.87I4	4.4650						0.9476
7	0.4276	0.9269	I.5273	2.28I3	3.2989	4.8925					0.9606
8	0.3739	0.80I5	I.3008	I.90I2	2.6553	3.5729	5.2665				0.9693
9	0.3323	0.7063	I.I338	I.633I	2.2336	2.9876	4.0052	5.5988			0.9754
I0	0.2990	0.63I4	I.0053	I.4329	I.9322	2.5326	3.2866	4.3042	5.8979		0.9798
II	0.27I6	0.5695	0.90I4	I.2746	I.70I5	2.I989	2.7955	3.5429	4.5480	6.II76	0.9832

Table 2

Optimal frequences with estimation of parameters of exponential distribution, tests of hypotheses with Pearson chi-square test and suitable values of relative asymptotic information A.

k	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	A
2	0.7968	0.2032										0.6476
3	0.6385	0.2880	0.0735									0.8203
4	0.5296	0.3004	0.1355	0.0345								0.8911
5	0.4514	0.2905	0.1648	0.0744	0.0189							0.9269
6	0.3930	0.2740	0.1763	0.1000	0.0451	0.0116						0.9476
7	0.3479	0.2563	0.1787	0.1150	0.0652	0.0294	0.0075					0.9606
8	0.3120	0.2394	0.1763	0.1229	0.0791	0.0449	0.0202	0.0052				0.9693
9	0.2827	0.2238	0.1717	0.1265	0.0882	0.0567	0.0322	0.0145	0.0037			0.9154
10	0.2584	0.2097	0.1659	0.1273	0.0938	0.0654	0.0421	0.0239	0.0107	0.0028		0.9798
11	0.2378	0.1964	0.1598	0.1264	0.0971	0.0715	0.0498	0.0322	0.0183	0.0083	0.0024	0.9832

$$I_F^G(\sigma) = \frac{1}{\sigma^2} \sum_{k=1}^{m} \frac{(t_k \varphi(t_k) - t_{k-1} \varphi(t_{k-1}))^2}{\Phi(t_k) - \Phi(t_{k-1})},$$

$$I_F^G(\mu,\sigma) = \frac{1}{\sigma^2} \sum_{k=1}^{m} \frac{(\varphi(t_k) - \varphi(t_{k-1}))(t_{k-1} \varphi(t_{k-1}) - t_k \varphi(t_k))}{\Phi(t_k) - \Phi(t_{k-1})}$$

in the form of $t_k = (X_k^G - \mu)/\sigma$ are presented in table 3. In the same table suitable values of relative asymptotic information A equal to $A = |I_F^G| / |I_F|$, where I_F is Fisher's information matrix of parameters of normal distribution with ungrouped sample and $|I_F| = 2/\sigma^4$, are presented.

E x a m p l e 3. Weibull's distribution density function takes the form of

$$f_\theta(x) = \frac{\theta}{\theta_1}\left(\frac{x}{\theta_1}\right)^{\theta-1} exp\left\{-\left(\frac{x}{\theta_1}\right)^{\theta}\right\}, \quad x > 0, \ \theta, \theta_1 > 0.$$

By maximizing the determinant of Ficher's information matrix with grouped data

$$I_F^G = \begin{bmatrix} I_F^G(\theta) & I_F^G(\theta,\theta_1) \\ I_F^G(\theta,\theta_1) & I_F^G(\theta_1) \end{bmatrix},$$

where

$$I_F^G(\theta) = \frac{1}{\theta^2} \sum_{k=1}^{m} \frac{(t_k e^{-t_k} \ln t_k - t_{k-1} e^{-t_{k-1}} \ln t_{k-1})^2}{e^{-t_{k-1}} - e^{-t_k}},$$

$$I_F^G(\theta_1) = \frac{\theta^2}{\theta_1^2} \sum_{k=1}^{m} \frac{(t_{k-1} e^{-t_{k-1}} - t_k e^{-t_k})^2}{e^{-t_{k-1}} - e^{-t_k}},$$

$$I_F^G(\theta,\theta_1) = \frac{1}{\theta_1^2} \sum_{k=1}^{m} \frac{(t_{k-1} e^{-t_{k-1}} - t_k e^{-t_k})(t_k e^{-t_k} \ln t_k - t_{k-1} e^{-t_{k-1}} \ln t_{k-1})}{e^{-t_{k-1}} - e^{-t_k}}$$

we derive optimal boundary points of grouping intervals in the form of $t_k = (X_k^G/\theta_1)^\theta$, which are presented in table 4. Table 5 presents suitable optimal probabilities of observation occurence in an interval. The tables present values of optimal asymptotic information $A = |I_F^G| / |I_F|$, where $|I_F| = 1.6449340\theta_1^2$. Figure 1 illustrates the gain of the chi-square test power with optimal grouping in comparison with partition in intervals of equal pro-

Table 3

Optimal boundary points of intervals of grouping in the form of $t_k = (x_k^G - \mu)/\sigma$ for simultaneous estimation of two parameters of normal distribution and tests of goodness of fit with Pearson chi-square test, and suitable value of relative asymptotic information A.

k	t_1	t_2	t_3	t_4	t_5	t_6	t_7
3	-1.1106	1.1106					
4	-1.3834	0.0	1.3834				
5	-1.6961	-0.6894	0.6894	1.6961			
6	-1.8817	-0.9970	0.0	0.9970	1.8817		
7	-2.0600	-1.2647	-0.4918	0.4918	1.2647	2.0600	
8	-2.1954	-2.4552	-0.7863	0.0	0.7863	1.4552	2.1954
9	-2.3188	-1.6218	-1.0223	-0.3828	0.3828	1.0223	1.6218
10	-2.4225	-1.7578	-1.2046	-0.6497	0.0	0.6497	1.2046
11	-2.5167	-1.8784	-1.3602	-0.8621	-0.3143	0.3143	0.8621
12	-2.5993	-1.9828	-1.4914	-1.0331	-0.5334	0.0	0.5534
13	-2.6746	-2.0762	-1.6068	-1.1784	-0.7465	-0.2669	0.2669
14	-2.7436	-2.1609	-1.7092	-1.3042	-0.9065	-0.4818	0.0
15	-2.8069	-2.2378	-1.8011	-1.4150	-1.0435	-0.6590	-0.2325

k	t_8	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	A
3								0.4065
4								0.5527
5								0.6826
6								0.7557
7								0.8103
8								0.8474
9	2.3188							0.8753
10	1.7578	2.4225						0.8960
11	1.3602	1.8784	2.5167					0.9121
12	1.0331	1.4914	1.9828	2.5993				0.9247
13	0.7465	1.1784	1.6068	2.0762	2.6746			0.9348
14	0.4818	0.9065	1.3042	1.7092	2.1609	2.7436		0.9430
15	0.2325	0.6590	1.0435	1.4150	1.8011	2.2378	2.8069	0.9498

Table 4

Optimal boundary points of intervals of grouping in the form of $t_k = (x_k^{\theta}/\theta_1)^{\theta}$ for simultaneous estimation of two parameters of Wribull's distribution and suitable values of relative asymptotic information A.

k	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
3	0.2731	2.6067						
4	0.2109	1.3979	3.4137					
5	0.1044	0.5123	1.9590	3.8606				
6	0.0772	0.3649	1.2269	2.5726	4.4096			
7	0.0501	0.2318	0.6758	1.7192	2.9922	4.7949		
8	0.0377	0.1740	0.4837	1.1904	2.2041	3.4285	5.2049	
9	0.0275	0.1269	0.3431	0.7829	1.6027	2.5713	3.7667	5.5273
10	0.0213	0.0988	0.2638	0.5770	1.1805	1.9932	2.9269	4.1024
11	0.0165	0.0771	0.2046	0.4359	0.8560	1.5344	2.3192	3.2319
12	0.0123	0.0618	0.0638	0.3434	0.6517	1.1789	1.8570	2.6163
13	0.0106	0.0500	0.1326	0.2754	0.5106	0.9030	1.4807	2.1401
14	0.0087	0.0412	0.1094	0.2261	0.4126	0.7116	1.1798	1.7608
15	0.0072	0.0344	0.0913	0.1881	0.3394	0.5734	0.9387	1.4426

k	t_9	t_{10}	t_{11}	t_{12}	t_{13}	t_{14}	A
3							0.4079
4							0.5572
5							0.6836
6							0.7571
7							0.8109
8							0.8480
9							0.8756
10	5.8478						0.8963
11	4.3930	6.1270					0.9123
12	3.5103	4.6589	6.3853				0.9248
13	2.8810	3.7623	4.9016	6.6208			0.9349
14	2.4019	3.1286	3.9997	5.1314	6.8444		0.9431
15	2.0116	2.6381	3.3538	4.2169	5.3425	7.0506	0.9498

Table 5

Optimal frequencies with simultaneous estimation of two parameters of Weibull's distribution or tests of goodness of fit with Pearson chi-square test and suitable values of relative asymptotic information A.

k	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8
3	0.2390	0.6872	0.0738					
4	0.1901	0.5628	0.2142	0.0329				
5	0.0991	0.3017	0.4581	0.1199	0.0211			
6	0.0743	0.2314	0.4011	0.2169	0.0641	0.0122		
7	0.0489	0.1581	0.2843	0.3295	0.1290	0.0419	0.0083	
8	0.0370	0.1227	0.2238	0.3124	0.1938	0.0779	0.0269	0.0055
9	0.0271	0.0921	0.1712	0.2525	0.2557	0.1250	0.0533	0.0191
10	0.0211	0.0729	0.1379	0.2065	0.2545	0.1708	0.0827	0.0371
11	0.0164	0.0578	0.1108	0.1683	0.2218	0.2101	0.1164	0.0589
12	0.0131	0.0468	0.0912	0.1395	0.1882	0.2136	0.1515	0.0830
13	0.0105	0.0383	0.0754	0.1165	0.1592	0.1947	0.1779	0.1099
14	0.0087	0.0317	0.0632	0.0988	0.1357	0.1710	0.1836	0.1354
15	0.0072	0.0266	0.0635	0.0842	0.1166	0.1486	0.1725	0.1548

k	P_9	P_{10}	P_{11}	P_{12}	P_{13}	P_{14}	P_{15}	A
3								0.4079
4								0.5572
5								0.6836
6								0.7571
7								0.8109
8								0.8480
9	0.0040							0.8756
10	0.0136	0.0029						0.8963
11	0.0271	0.0102	0.0022					0.9123
12	0.0432	0.0204	0.0078	0.0017				0.9248
13	0.0615	0.0329	0.0158	0.0061	0.0013			0.9349
14	0.0814	0.0467	0.0255	0.0124	0.0048	0.0011		0.9431
15	0.1025	0.0623	0.0365	0.0203	0.0099	0.0039	0.0009	0.9498

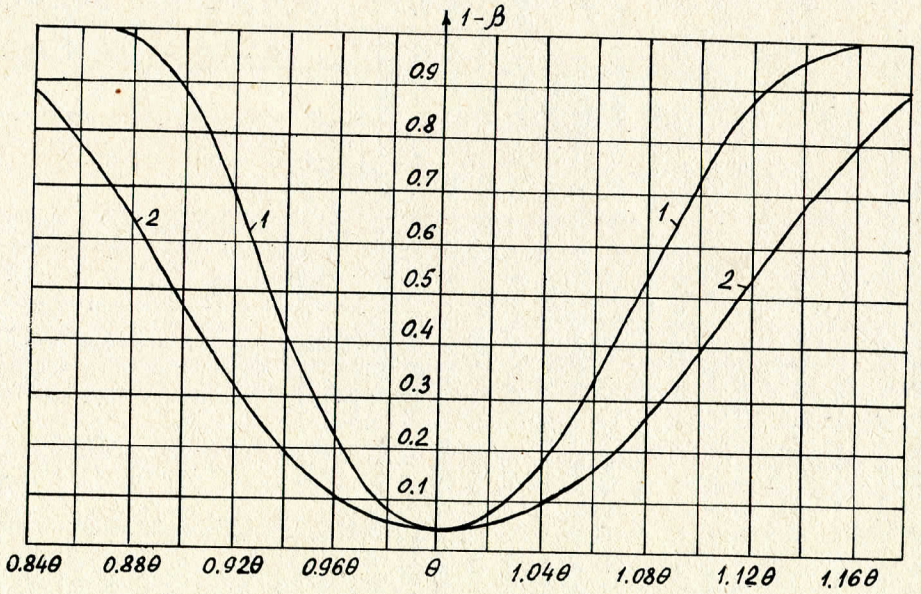

Fig. 1. Power function of chi-square test in the verification of hypotheses about the main parameter of Weibull's distribution: significance level α =0.05 the sample size n =1000, the number of intervals m =3, 1- in the optimal grouping, 2- in the equidistant one.

bability at test of hypotheses about the main parameter θ for $m=3$.

3.Recommendations on the use of tables of asymptotically optimal grouping

In the matching of distributions of a class limited by preliminary reasonings it is customary to follow the next sequence: parameters are defined for each distribution with empirical data; goodness of fit of the derived theoretical curve with empirical data with one of the goodness-of-fit test in verified; distribution function yielding the best goodness of fit is chosen.

The most acceptable goodness of fit test with a large number of observations is Pearson chi-square test.

The test of goodness of fit with chi-square test is carried out according to the following scheme. The range of change of random variables which is defined by empirical data is partitioned in m intervals by boundary points. Futher frequencies of n_k observation occurences in each interval are computed. For hypothetic distribution defined according to theoretic considerations, probabilities $P_\theta^G(k)$ of observation occurences in k -th interval are computed, and after that the value of chi-square statistic is computed (3). This statistic has distribution of χ^2 with $S=m-1$ degrees of freedom, if estimation of parameters in terms of the given sample has not been carried out, and it has a distribution of χ^2 with $m-r-1$ degrees of freedom, if from grouped data r parameters of distribution were estimated, and it has a distribution intermediate between chi-square distribution with the numbers of $m-1$ and $m-r-1$ degrees of freedom, if the estimation was carried out with ungrouped data.

Depending on the fact, whether the distribution parameters were estimated in terms of the given sample and according to what type of data, grouped or ungrouped ones, the estimates were defined, by the given significance level α from the tables of chi-square distribution with the suitable number of degrees of freedom the critical value of $\chi^2_{S,\alpha}$ is defined. If the value of chi-square statistic is less than $\chi^2_{S,\alpha}$, then the goodness-of-fit hypothesis is not rejected. In the estimation of distribution parameters with ungrouped data one has to be convinced that the value of χ^2

does not exceed the values of $\chi^2_{m-1,\alpha}$ and $\chi^2_{m-r-1,\alpha}$.
In case of the use of tables of asymptotically optimal grouping in the verification of hypotheses in terms of chi-square test the maximum power of the test with close alternative hypotheses is guaranteed.
In this case it is necessary to follow the sequence of operations: depending on the form of distribution boundary points of grouping intervals (number of intervals is defined in such a way that the product of the sample size by the probability of observation occurence in each interval would be more than one, and if possible $\geqslant$ 10) are chosen; the number of values of random variables n_k occurred in the interval is computed; the value of probabilities corresponding to the theoretical distribution are taken from the tables of optimal frequences; chi-square statistic is computed and compared with critical values of $\chi^2_{s,\alpha}$.

<u>Example 4.</u> It is necessary to test the goodness of fit of empirical data with exponential distribution with parameter θ =1. Ordered by increasing the sample from 50 random numbers is equal to 0.01, 0.01, 0.01, 0.04, 0.17, 0.18, 0.22, 0.22, 0.25, 0.25, 0.29, 0.42, 0.46, 0.47, 0.56, 0.59, 0.67, 0.68, 0.70, 0.72, 0.76, 0.78, 0.83, 0.85, 0.87, 0.93, 1.00, 1.01, 1.01, 1.02, 1.03, 1.05, 1.33, 1.34, 1.37, 1.47, 1.50, 1.52, 1.54, 1.59, 1.71, 1.90, 2.10, 2.35, 2.38, 2.46, 2.50, 3.73, 4.07, 6.03.
From the table 2 we find that we have to choose 4 grouping intervals as $n \cdot P^G_\theta(4) = 50 \cdot 0.0345 > 1$. From table 1 define the boundary points at $m=4$, $t_1=0.7541$, $t_2=1.7716$, $t_3=3.3652$, $X^G_k = t_k/\theta$ and compute the number of observation occurences in each interval: $n_1=20$, $n_2=21$, $n_3=6$, $n_4=3$. Theoretical probabilities from table 2 are $P^G_\theta(1)=0.5296$, $P^G_\theta(2)=0.3004$, $P^G_\theta(3)=0.1355$, $P^G_\theta(4)=0.0345$. Compute statistic χ^2 =4.566. The number of degrees of freedom $S = m-1 = 3$. At the significance level α =0.1 find the critical value of $\chi^2_{3,\alpha}$ =6.251 and, consequently, the sampling data agree with exponential distribution.
In the estimation of parameters of distribution with the application of maximum likelihood method or chi-square minimum the use of optimal grouping allows to reduce asymptotic estimate

variance.
As it is seen from the tables of asymptotically optimal grouping the optimal boundary points depend on the true value of parameter which as a matter of fact is unknown. This difficulty is overcome by carrying out grouping with predicted value of parameter on the basis of a priori information.

We may suggest a second approach which consists in the partition of the initial sample in groups such that the number of realizations of a random variable in each group be proportional to probabilities at optimal grouping. With the large size of the sample this approach is more preferable. It appears to be the only possible one, if a priori information about parameters of distribution is missing, and only one form of distribution is assumed.

In this case we have a chance by means of the tables of asymptotically optimal grouping to derive approximate maximum likelihood estimates of parameters of distribution.
Using the suitable table of optimal frequencies, partition the initial sample into groups, such that the number of observation n_k occurred in each one be proportional to optimal frequency, such that $n_k = n\, p_\theta^G(k)$. The choice of the number of intervals m is defined by the condition $n \cdot p_\theta^G(k) > 1$, better $\geqslant$ 10. As a result, optimal boundary points $\hat{x}_k^G$, partitioning the groups, will be approximately chosen. For instance, by way of $\hat{x}_k^G$ we may take the average between adjacent sampling values occurred in different groups.
From the tables of optimal boundary points with the given number of intervals the values t_k which are connected with x_k^G with quite definite dependence are taken. Hence, the approximate estimate of the unknown parameter is already easily found from the equations $t_k = \psi(x_k^G, \theta)$ and averaged over all k.
In particular, the estimate of parameter θ of <u>exponential distribution</u> is defined by the formula

$$\hat{\theta} = \frac{1}{m-1} \sum_{k=1}^{m} t_k / \hat{x}_k^G ,$$

where m is the number of intervals and t_k is chosen from table 1. Moreover, the boundary points $\hat{x}_k^G$ are defined in the partition of the sample into groups proportionally to the probabilities from table 2.

In the estimation of two parameters of Weibull's distribution t_k from table 4 are used, and the estimates are defined by the relations

$$\hat{\theta} = \frac{1}{m-2} \sum_{k=2}^{m-1} \frac{\ln t_{k-1} - \ln t_k}{\ln \hat{x}_{k-1}^G - \ln \hat{x}_k^G}, \tag{9}$$

$$\hat{\theta}_1 = \frac{1}{m-2} \sum_{k=2}^{m-1} \exp\left\{ \frac{\ln t_{k-1} \ln \hat{x}_k^G - \ln t_k \ln \hat{x}_k^G}{\ln t_{k-1} - \ln t_k} \right\}. \tag{10}$$

The partition into groups of the sample is carried out proportionally to the probabilities from table 5.

E x a m p l e 5. Assuming, that the sample presented in example 4 corresponds to the Weibull's distribution, let us estimate its parameters.

According to table 6 we find that it is possible to choose 5 grouping intervals as minimal product $n \cdot p_\theta^G(5) = 50 \cdot 0.0211 = 1.055 > 1$. The sample has to be partitioned into groups proportionally to the values $n \cdot p_\theta^G(1) = 50 \cdot 0.0991 = 4.955$, $n\, p_\theta^G(2) = 15.09$, $n\, p_\theta^G(3) = 22.905$, $n\, p_\theta^G(4) = 5.995$, $n\, p_\theta^G(5) = 1.055$. Hence, $n_1 = 5$, $n_2 = 15$, $n_3 = 23$, $n_4 = 6$, $n_5 = 1$. In the capacity of $\hat{x}_k^G$ we'll take the average between the sampling values occurred in the adjacent groups. For instance, $\hat{x}_1^G$ between the fifth and the sixth sampling value: $\hat{x}_1^G = 0.175$. Further, $\hat{x}_2^G = 0.515$, $\hat{x}_3^G = 2$, $\hat{x}_4^G = 5.05$. From table 5 at $m = 5$ choose $t_1 = 0.1044$, $t_2 = 0.5123$, $t_3 = 1.959$, $t_4 = 3.8606$. Now by formula (9) define $\hat{\theta} = 1.065$ and by formula (10) $\hat{\theta}_1 = 0.874$. Verify the goodness of fit by chi-square test. The number of intervals $m = 5$. From table 4 in view of $t_k = (x_k^G/\theta_1)^\theta$ define from relation $x_k^G = \exp\{(\ln t_k)/\theta - \ln \theta_1\}$, $x_1^G = 0.1371$, $x_2^G = 0.6105$, $x_3^G = 2.1512$, $x_4^G = 4.0677$.

By the sample define the number of observations occurred in each interval: $n_1 = 4$, $n_2 = 13$, $n_3 = 27$, $n_4 = 4$, $n_5 = 2$. Take the probabilities from table 5. Compute chi-square statistic equal to 2.716. The number of degrees of freedom $s = m - r - 1 = 5 - 2 - 1 = 2$. With the significance level $\alpha = 0.2$ find the critical value $\chi^2_{2,\alpha} =$

=3.219 and, consequently, the sampling data agrees well with the Weibull's distribution with estimated parameters.
For some other distributions expressions for estimates and tables of asymptotic distribution are presented in the papers of DENISOV [4] , LEMESHKO [5] .

R e f e r e n c e s

[1] KULLDORF,G.(1966) Vvedeniye v teoriyu otsenivaniya po gruppirovannym i chastichno gruppirovannym 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. Vol. 111.-s.110-154

[3] COX,D.R.(1957) Note on grouping. Journal of the American Statistical Association. Vol.52,Nn 280, p.543-547

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

[5] LEMESHKO,B.Yu.(1977) Otsenivaniye parametrov raspredeleniya po gruppirovannym nabludeniyam. Voprosi kibernetiki. Moskva. Vip.30, s. 80-96