Arbeidsnotater

STATISTISK SENTRALBYRÅ

WORKING PAPERS FROM THE CENTRAL BUREAU OF STATISTICS OF NORWAY

IO 70/7 Oslo, 28 April 1970

TWO METHODS FOR SPLITTING DATA INTO HOMOGENEOUS GROUPS

By E. Gilje and I. Thomsen

C O N T E N T S

		Page
0.	Abstract	2
1.	Introduction	2
2.	The homogenizing criterion	3
3.	Method I	3
4.	Method II	5
5.	Comparing method II with a method developed by Dalenius	7
6.	An application	9
7.	References	11

0. Abstract

The optimum classification of a data material is defined variance-analytically; i.e. the variance of a variable X shall be maximized between the groups and minimized within them.

Two methods for obtaining this "best" splitting are given. The latter of these has been programmed in Basic FORTRAN IV, and is used in the examplifications.

In the first of the examples, we have simulated a normal frequency curve. Thus, by splitting the data, we can compare the outcome obtained using our method with theoretical results given by Dalenius (1950). In the next example we have classified all Norwegian municipalities with respect to fertility. The classification variable here is the gross reproduction rate in each municipality.

1. Introduction

It is often desirable to gather observations into groups or classes in such a way that all the elements in each group are as homogeneous as possible with regard to the chosen classification variable(s). At the same time one does not want to split the population into too many groups. (The trivial case is of course one group for each observation.)

In the case of qualitative classification criteria (e.g. geographical or educational criteria), the partitioning into homogenous groups need not raise any problems. Most often in such cases the groups are specified directly by the classification variable. The problem is usually greater when the classification variable is quantitative.

If it is possible to characterize the variable one wishes to use as a classifier by a known theoretical distribution, say a normal or log-normal distribution, the partitioning can be done using methods developed by Dalenius (1950). In many cases, however, this is not possible.

In this article we are therefore going to give two alternative methods for classifying a material. The former is only outlined, while the latter is used in the examples and results compared with those obtained using the more analytical method given by Dalenius (1950).

2. The homogenizing criterion

As a measure for homogeneity we are going to use the sum of the empirical variances of the classification variable within each group.

Let us start with the empirical variance of the observations $X_1, X_2, \ldots, X_n$; $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. As known from the analysis of variance this quantity may, for a given partition, be written as the sum of the variance between the groups and the variance within the groups. As we want the elements within one group to be alike, it seems reasonable to minimize the variance within the groups. Thus, after choosing how many groups L one wants the task is to minimize the expression.

$$(2.\ 1)\quad Q = \frac{1}{n}\sum_{i=1}^{L}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2 \quad \text{where } \bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$$

with regard to $n_1, n_2, \ldots n_L$ under the restriction $n_1 + n_2 + \ldots + n_L = n$. X_{ij} is the j-th value of X in group number i. Our criterion may also be formulated as follows:

As the total variance $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is independent of the choice of partition points, to minimize (2. 1) is equivalent to maximizing the variance between the groups; $\frac{1}{n}\sum_{i=1}^{L} n_i (\bar{X}_i - \bar{X})^2$. This is the same as finding the partitioning that gives the greatest "difference between the groups."

It often happens that we must choose how many partition points, L-1, we want; i.e. L is not determined by considerations outside this particular problem. Obviously, it is important then to choose L in a sensible way. This may be done by studying the change in the variance within the groups for increasing L (see section 6).

3. Method I

In this section we will outline a simple method for splitting a population into two groups.

Let $X_1, X_2, \ldots, X_n$ be the population sorted in ascending sequence. Then choose a dividing point forming the two groups $X_1, X_2, \ldots, X_{n_1}$, and

$X_{n_1+1}, \ldots, X_n$. To test the homogenizing effect of this choice, we will compare it with the splitting $X_1, X_2, \ldots, X_{n_1}, X_{n_1+1}$ and $X_{n_1+2}, X_{n_1+3}, \ldots, X_n$ by investigating the change in the sum of variance within the groups (or the variance between the groups). If D is this change we have

$$D = \sum_{j=1}^{n_1+1} (X_j - \bar{X}_1)^2 + \sum_{j=n_1+2}^{n} (X_j - \bar{X}_1^*)^2 - \sum_{j=1}^{n_1} (X_j - \bar{X}_o)^2 - \sum_{j=n_1+1}^{n} (X_j - \bar{X}_o^*)^2$$

where

$$\bar{X}_1 = \frac{1}{n_1+1} \sum_{j=1}^{n_1+1} X_j, \quad \bar{X}_1^* = \frac{1}{n_2-1} \sum_{j=n_1+2}^{n} X_j,$$

$$\bar{X}_o = \frac{1}{n_1} \sum_{j=1}^{n_1} X_j, \quad \bar{X}_o^* = \frac{1}{n_2} \sum_{j=n_1+1}^{n} X_j \text{ and } n_2 = n - n_1.$$

If D = o the following condition must be satisfied:

$$(3.1)\quad (n_1 + 1)\, \bar{X}_1^2 + (n_2 - 1)\, \bar{X}_1^{*2} = n_1\, \bar{X}_o^2 + n_2\, \bar{X}_o^{*2}.$$

The same condition is found if one alternatively investigates the change in the variance between the groups. By approximating $\frac{n_1+1}{n_1}$ and $\frac{n_2-1}{n_2}$ with 1 the condition (3.1) may be formulated

$$(3.2)\quad X_{n_1+1} \frac{(\bar{X}_o^2 - \bar{X}_o^{*2})}{\frac{n\, X_{n_1+1}}{(n_1+1)(n_2-1)} + 2(\bar{X}_o - \bar{X}_o^*)}$$

For large n, n_1 and n_2 we can simplify this further to

$$(3.3)\quad X_{n_1+1} = \frac{1}{2}(\bar{X}_o + \bar{X}_o^*)$$

This last approximation is dependent on the unit of measurement of X, and therefore it may be an advantage to normalize the observations before using this method.

The results obtained in this section can be generalized to cases where one wants more than one dividing point. If we want the population split

into three groups the following method may be used:

Choose an initial first dividing point X_{n_1+1}. Then $\bar{X}_1$, defined as the mean in the first group, can be found. Insert X_{n_1+1} and $\bar{X}_1$ in (3. 2) ($\bar{X}_1$ replaces $\bar{X}_o$) and solve it with regard to $\bar{X}_o^x$ which in correspondence with $\bar{X}_1$ will be termed $\bar{X}_2$ in the following. Knowing the first two group means, the second dividing point and the corresponding $\bar{X}_3$ are easily found. Now both dividing points are tested against the condition (3. 2) or (3. 3). If this is not fulfilled, the procedure is repeated with a new choice for X_{n_1+1}.

This technique can of counse be used also for a splitting into more than three groups. The method is not tried out on a computer, as we already had a program that was suitable for our purpose. We suspect, however, that at least for large values of L, the method may lead to an extraordinary number of calculations even for a computer.[1)]

4. Method II

The method and the computer program we have used to minimize (2. 1) is described in general by Nelder and Mead (1965). Thus we shall not waste space and time here on a detailed description, but give an abstract of the general principles and our spesial applications of these.

A function of n variables is to be minimized with regard to all the variables. As a starting point we guess at a set of values for the variables. This set ought not to be too far away from the set forming the actual minimum point (we assume that such a point does exist.) Then add an optional constant, δ, to the guess at the first variable while the remaining variables are left unchanged. Proceed by adding the same to the start value of the second variable, the first variable is again given the original value, and leave the remaining variables unchanged. This procedure is continued until (n + 1) such sets of values are formed. The first set is the original guesses. The n following sets consist also of the original guesses except for one of the variables where this value has been given an addition of δ. These (n + 1) sets together are called a simplex.

For each set the function value is computed. By certain methods (the details of which can be found in the article mentioned above) the set giving the highest function value is replaced by a set giving a lesser value, and we have thus formed a new simplex.

1) After finishing the manuscript of this article we have discovered that a method similar to the one described in this section is discussed by Görran Nilsson (1967).

This new simplex is treated in the same manner. We replace the set now giving the highest function value by another set.

The procedure is continued until the process stops automatically in one of two ways. If no minimum point is to be found in the area we are investigating, the absolute function values will eventually exceed the maximum allowed in the computer's registers, and we get a program interruption. If a minimum exists, the simplex contracts on to the final minimum, and the process stops when the absolute difference between the minimum values of two successive computations is less than a certain value.

The algorithm may cause some trouble as the minimum found is a local minimum and therefore not necessarily the absolute minimum of the function. In practice, however, one can usually tell if the result is sensible. If not, other choises of start values and/or δ will eventually lead to the sought minimum.

In our special case (2. 1) is to be minimized with regard to n_1, n_2, ..., n_L. As $\sum_{i=1}^{L} n_i = n$, n_L is given as a linear function of n and n_1, n_2, ..., n_{L-1} there are L-1 unknown variables. We have used

$$(4.\ 1) \qquad n_i = \frac{n}{L} - \frac{\delta}{2} \quad \text{for } i = 1, 2, \ldots, L-1$$

as start values for these. (The choice of n_i for i = 1, 2, ..., L-1 is of course optional, but this procedure seems to be effective.) δ ought to be positive to avoid the sum of the (L-1) first n_i in a row of the simplex exeeding n. It is difficult to give a general rule governing choice of the size of δ as both this and the orientation of the initial simplex have an effect on the speed of convergence. If δ is too small we often get no convergence at all or absurd results (the problem with local minimums), if it is too large, the convergence will be slow. In the examples in the following sections we have used $\delta = 10$ which together with (4. 1) have given sensible results each time.

The algorethm is programmed in Basic Fortran IV and we have used an IBM 360/40 for the computations in the examples. The machine uses slightly more than 1 minute to split a population of 500 into 20 groups. If L is small, the machine will use less time.

5. Comparing method II with a method developed by Dalenius

Assume that the classification variable can be characterized by a theoretical distribution f (x). We want to split the distribution by minimizing

$$(5.\ 1) \qquad V(\bar{x}) = \frac{1}{n} \sum_{h=1}^{L} W_h \sigma_h^2 \quad \text{where}$$

$$W_h = \int_{d_{h-1}}^{d_h} f(x)\, dx, \quad W_h \xi_h = \int_{d_{h-1}}^{d_h} xf(x)\, dx,$$

$$W_h \sigma_h^2 = \int_{d_{h-1}}^{d_h} (x-\xi_h)^2 f(x)\, dx,$$

and d_h and d_{h-1} are partition points number h and h-1 respectively. n is the total number of observations.

Dalenius then proves that in the minimumpoint

$$(5.\ 2) \qquad d_h = \frac{1}{2} (\xi_h + \xi_{h-1}).$$

This result corresponds with the result expressed by (3. 3). As in section 3 and 4 we are compelled to determine the partition points by successive approximations. The following results are taken from Lykke Jensen (1960) p. 386-387.

Table 1: Optimal splitting of the normal distribution with expectation 0 and variance 1

Number of groups	d_1	d_2	d_3	d_4	d_5	$\Sigma W_h \sigma n^2$
2	$u_{0.500}$					0.363
3	$u_{0,270}$	$u_{0,730}$				0.190
4	$u_{0,163}$	$u_{0.500}$	$u_{0.837}$			0.117
5	$u_{0.107}$	$u_{0.352}$	$u_{0.648}$	$u_{0.893}$		0.079
6	$u_{0.074}$	$u_{0.256}$	$u_{0.500}$	$u_{0.744}$	$u_{0.926}$	0.057

u_t is the t-fractile in this distribution.

Table 2: Group - expectations in the normal distribution with expectation o and variance 1 after optimal splitting

Number of groups	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_6
2	-0.798	0.798				
3	-1.225	0	1.225			
4	-1.511	-0.453	0.453	1.511		
5	-1.724	-0.762	0	0.762	1.724	
6	-1.895	-0.999	-0.315	0.315	0.999	1.895

After simulating 500 observations from a normal distribution, we have used method II for splitting these in groups. The outcome is given in table 3 and 4 which corresponds with table 1 and 2 respectively.

Table 3: The number of elements in each group after splitting a normal distribution optimally

Number of groups	Group number 1	2	3	4	5	6	7	8	$\frac{1}{500}\sum_{i,j}(x_{ij}-\bar{x}_j)^2$
2	211	229							0.371
3	156	232	112						0.194
4	86	185	161	68					0.116
5	70	125	126	132	47				0,082
6	33	97	123	123	86	38			0.056
7	29	65	93	103	100	76	34		0.042
8	29	68	73	86	57	82	71	34	0.036

Table 4: Group-averages in the groups from table 3

Number of groups	Group number 1	2	3	4	5	6	7	8
2	-0.79	0.81						
3	-1.18	0.03	1.31					
4	-1.52	-0.45	0.47	1.61				
5	-1.63	-0.70	-0.05	0.70	1.81			
6	-1.96	-1.05	-0.38	0.25	0.94	1.92		
7	-2.02	-1.23	-0.64	-0.15	0.39	1,03	1,98	
8	-2.02	-1.21	-0.68	-0.26	0.09	0.47	1,06	1.99

6. An application

While calculating regional fertility-tables we have a dual problem. We want regions as small as possible to make sure that regions with significantly differant fertility get differnet tables. This is particularly important when these tables are to be used for the calculations of regional

population projections. On the other hand the regions can be too small. Then the diagram of a set of agespecific fertility rates tends to show a rather rugged curve, and these "irregularities" will often dominate over the "true" underlying fertility. (Gilje, 1969).

A solution of this problem is to merge regions with "similar" fertility to a larger area. Such an area needs not to be geographically coherent. The homogenizing technique described in this article has been used for this merging.

For each of Norway's 451 municipalities, age-specific fertility rates for women between 15 and 44 years of age, have been computed. The sum of these age-apecific rates is the gross joint reproduction rate, GRR (i.e. the average number of live children that would be born to a hypothetical female birth cohort which would be subjected to current age-specific fertility on the assumption that mortality before the end of the reproductive age is zero.) We will use GRR as a measure of the total fertility in a municipality. The data used gives us average GRRs for the years 1966 to 1968.

After sorting the municipalities according to GRR, this variable will correspond with X in (2. 1). The results of the homogenizing process are shown in table 5.

The histogram in figure 1 shows the distribution of municipalities by GRR.

In figure 2 $\sum_{i=1}^{L} \sum_{j=1}^{n_i} (GRR_{ij} - \overline{GRR}_i)^2$, i.e. the botton row in table 5, is plotted as a function of L. From this diagram we conclude that a splitting into more than 7 or 8 groups, gives a relatively small reduction in the sum of variances within the groups.

Table 5: The number of municipalities in each group after homogenizing by the gross joint reproduction rates

n_i	Number of groups (L)									
i	2	3	4	5	6	7	8	10	15	20
1	253	183	96	53	48	42	7	36	6	4
2	198	178	158	150	101	96	39	61	24	11
3		95	145	139	102	52	57	64	27	19
4			52	84	107	81	94	37	39	24
5				25	74	85	75	47	37	24
6					19	76	86	55	38	36
7						19	74	45	29	28
8							19	53	25	22
9								42	33	29
10								11	38	25
11									36	16
12									44	22
13									39	14
14									27	23
15									9	56
16										24
17										28
18										25
19										19
20										2
$\sum_{i=1}^{L} \sum_{j=1}^{n_i} (GRR_{ij} - \overline{GRR}_i)^2$:	45.57	25.05	16.98	10.75	8.03	6.91	5.35	4.31	2.16	1.52

7. References

(1) Dalenius, T. (1950) The problem of optimum stratification. Skandinavisk Aktuarietidsskrift. 203-13.

(2) Gilje, E. (1969) Fitting curves to age-specific fertility rates: Some examples. Statistisk Tidsskrift (Statistical Review) 2 (2): 118-134.

(3) Lykke Jensen, E. (1960) Representative undersøkelsers teori og metode. København.

(4) Nelder, J.A. and Mead, R. (1965). A Simplex Method for Function Minimization. The Computer Journal. 2: 308-313.

(5) Nilsson, G. (1967) Optimal Stratification according to the Method of least Squares. Skandinavisk Aktuarietidsskrift: 128-36.

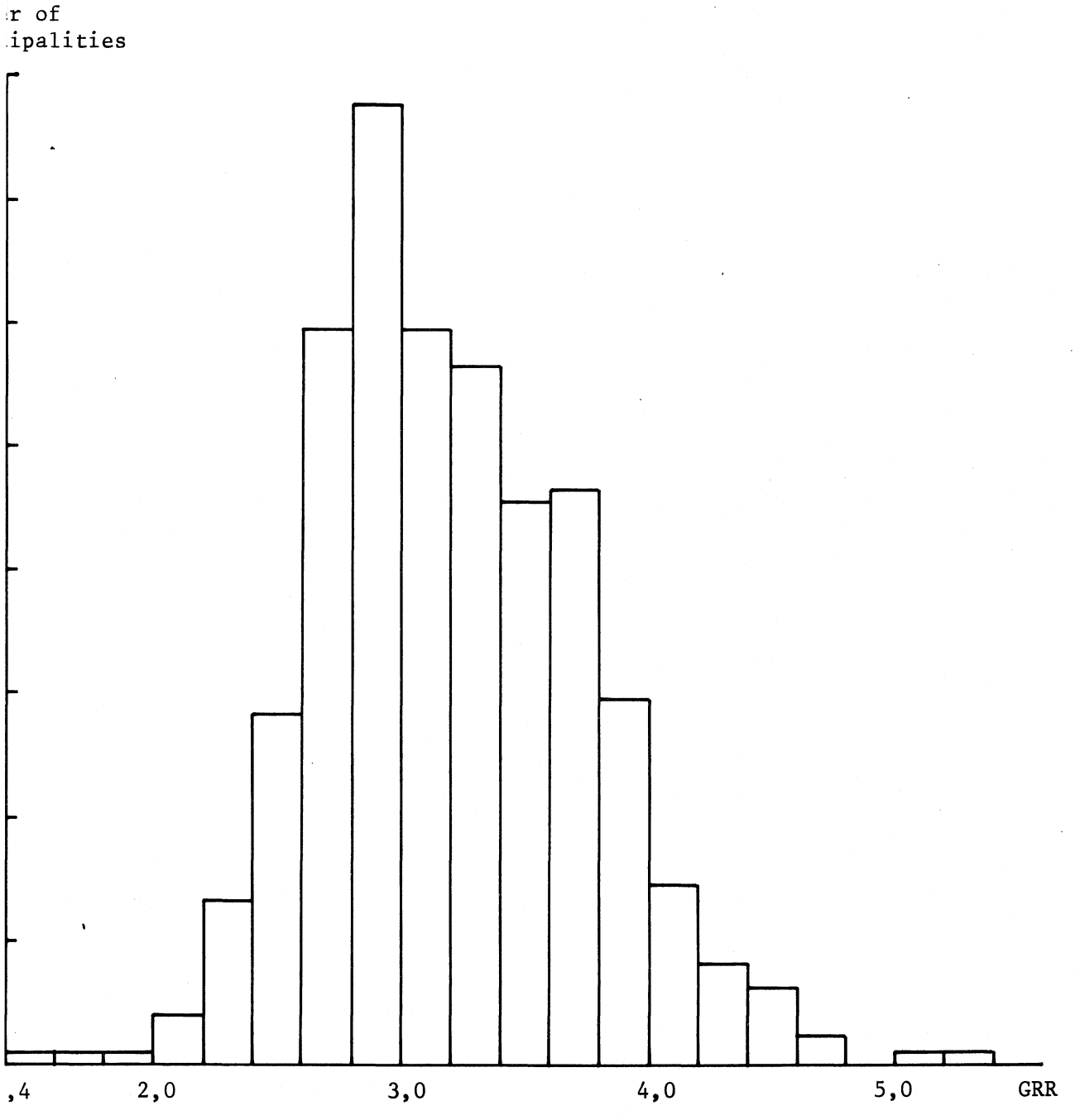

Fig. 1: The municipalities distributed by GRR.

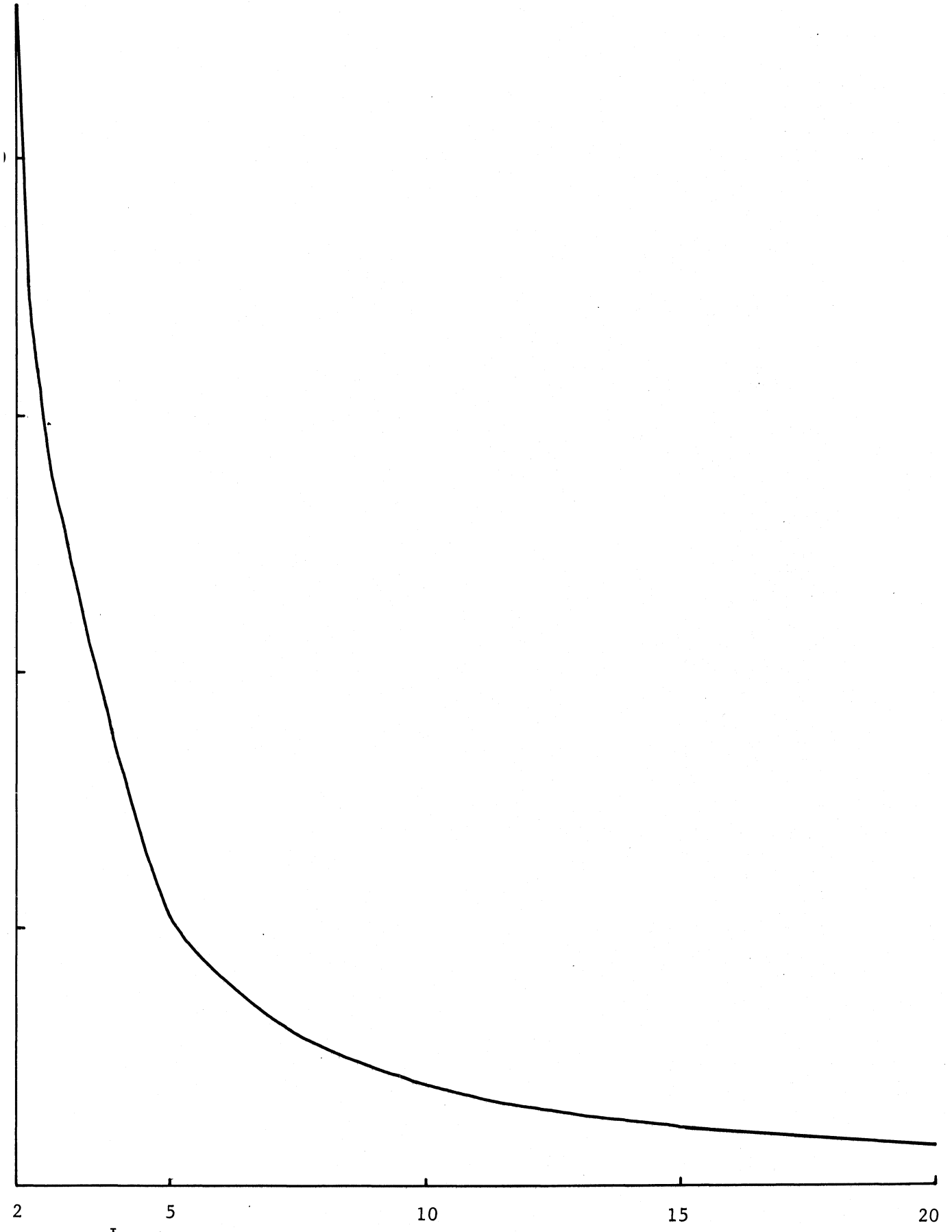

Fig. 2: $\sum_{i=1}^{L} \sum_{j=1}^{n_i} (GRR_{ij} - \overline{GRR}_i)^2$ as a function of L.