Corrado Gini

Sulla misura della concentrazione e della variabilit`a dei caratteri 1914

On the measurement of concentration and variability of characters[1]

METRON - *International Journal of Statistics,* vol. LXIII, n. 1, 2005

Summary - **1**. The concept of concentration of a character and various indices proposed for its measurement. Purposes of the present note. **2**. On a measure of the concentration that does not depend on the distribution of the character. The *concentration ratio*. **3**. Various arithmetic procedures for the determination of the concentration ratio according to the available statistical information. **4**. Concentration for some human physiological characters and for some economic characters. **5**. Concentration ratio for general cases and for positive cases. Transition formula. **6**. Relationship between the concentration ratio and some graphical representation of the distribution of wealth. **7**. Graphical procedures for determining the concentration ratio. **8**. Further considerations on the graphical representation of the concentration ratio. **9**. Relationship between the concentration ratio and the *mean difference*. **10-11**. Related considerations. **12**. Concentration ratio for truncated seriations and concentration ratio for the whole seriation. Their relationships. **13**. Conclusions.

1. Let us consider n quantities that measure the intensity of a certain character in n different cases. Let us rank such quantities so that each of these is less than or equal than the following one, and let a_k $(k = 1, 2, \dots , n)$ denote the k-th element of the sequence obtained.

For two arbitrary values (i, l) of k such that $i < l$, we have

$$a_i \leq a_l \tag{1}$$

$$\frac{1}{i}\sum_{k=1}^{i} a_k \leq \frac{1}{l}\sum_{k=1}^{l} a_k. \tag{2}$$

From this relationship we have

$$\frac{\sum_{k=1}^{i} a_k}{\sum_{k=1}^{l} a_k} \leq \frac{i}{l} \tag{3}$$

[1] Originally published in Italian in *Atti del Reale Istituto Veneto di Scienze, Lettere ed Arti*. 1914, Vol. LXXIII - Part II, pages 1203-1248.
Translation: Fulvio De Santis.

and, as a specific case, for $l = n$,

$$\frac{\sum_{k=1}^{i} a_k}{\sum_{k=1}^{n} a_k} \leq \frac{i}{n}.$$

For simplicity, assume that $\sum_{k=1}^{i} a_k = A_i$; $\sum_{k=1}^{n} a_k = A_n$; $\frac{A_i}{A_n} = q_i$; $\frac{i}{n} = p_i$; $\frac{A_i}{i} = M_i$; $\frac{A_n}{n} = M_n$; $\frac{A_n - A_i}{n-i} = M_{n-i}$, where we always have that $i < n$.

We say that *the stricter the inequality*

$$p_i > q_i \tag{4}$$

for the $n - 1$ *values of* i*, the stronger the concentration of the character.*

In other words, we say that *the smaller the part of the total amount of the character owned by those cases whose intensity of the character itself is below a certain level, the stronger the concentration of the character.*

The *concentration of the character* is said to be *perfect* when the intensity of the character is $= 0$ in $n - 1$ cases and $= A_n$ in just one case. For the remaining $n - 1$ values of i we then have $q_i = 0$ and, hence, $p_i - q_i = p_i$. If the intensity of the character is the same for all the cases, so that $p_i = q_i$ for all the $n - 1$ values of i, we say that the concentration of the character is null or, in other words, that the character is *equally distributed.*

If, using one or several constants, we were able to express, for all the values of i and with a certain degree of approximation, a valid relationship between the two terms of inequality (4) (or between the terms of an inequality implied by (4)), then we could assume this constant (or these constants) as a *concentration index* for the character.

To this end, in the place of inequality (4), it is sometimes convenient to consider the inequalities

$$1 - p_i < 1 - q_i \tag{5}$$

$$M_{n-i} > M_n \tag{6}$$

that follow directly from (4). In a previous paper of mine, I made a distinction between *simple* and *complex* concentration indices, depending on the number of constants involved; for both, some examples have been given.

The concentration indices for individual and family incomes, and that of family dwelling rent values, can almost always be derived with sufficient accuracy by the following equation

$$1 - p_i = (1 - q_i)^{\delta}; \tag{7}$$

the concentration index of prolificacy of married couples can be obtained from equation

$$q_i = p_i^{\delta} \tag{8}$$

or, with more accurate approximations, either from

$$q_i = p_i^{\delta+\varepsilon(a_i-1)} \tag{9}$$

or from

$$M_{n-i} = M_n + \delta a_i + \frac{a_i^2}{100}. \tag{10}$$

The concentration index of the value of successions in Italian provinces and French departments can be derived as well from equation (8).([2])

All these indices have an undeniable usefulness, since they allow one to compare the concentration of a given character in different populations and in different times (for instance, the concentration of individual incomes in different countries and in different years); sometimes they also allow to compare the concentration of different characters, that nevertheless present similar distributions whose index can then be derived from the same formula (for instance, the concentration of individual income and that of dwelling rent values). However, on the one hand the number of characters for which a formula that allows us to obtain a concentration index is quite limited; on the other hand, even if we could find formulas for other characters, nevertheless, comparisons between characters whose indices are obtained using different formulas would be impossible.

([2]) For all this, see *Indici di concentrazione e di dipendenza*, in *Biblioteca dell'Economista*, Series V, vol. XX, Torino, Unione Tipografica Editrice, 1910. A summary of this memory, presented at the *Terza riunione della Società italiana per il progresso delle scienze* (*Third meeting of the Italian Society for the progress of sciences*, October 1909), was published in the *Atti* of the same Society (Rome, Bertero, 1910). The concept of concentration and the concentration index for income and dwelling rent values had already been presented at the *Seconda riunione della Società italiana per il progresso delle scienze* (*Second meeting of the Italian Society for the progress of sciences* (October 1908) one year before, in a communication whose title was *Il diverso accrescimento delle classi sociali e la concentrazione della ricchezza*, that appeared on the *Giornale degli Economisti* (January 1909).

On the use of the concentration indices, see G. Del Vecchio, *Ricerche statistiche sui depositi a risparmio.* Udine. Tosolini, 1910; L.V. Furlan, *Neue Literatur zur Einkommensvereilung in Italien*, in *Jahrbücher für Nationalökonomie und Statistik* (August 1911); F. Corridore, *Relazioni tra affitto reale e valore locativo fiscale nel Belgio.* Roma, Loescher 1911; E. Porru, *La concentrazione della ricchezza nelle diverse regioni d'Italia*, in *Studi economico-giuridici pubblicati per cura della Facoltà di giurisprudenza della R. Università di Cagliari.* Vol. IV, Part I, 1911-1912; F. Savorgnan, *La distribuzione dei redditi nelle provincie e nelle grandi città dell'Austria,* in *Pubblicazioni del Museo commerciale di Trieste.* 1912; C. Gini, *Variabilità e Mutabilità, contributo allo studio delle distribuzioni e delle relazioni statistiche.* Fasc. I. *Introduzione - Indici di Variabilità - Indici di Mutabilità,* in *Studi economico-giuridici pubblicati per cura della Facoltà di giurisprudenza della R. Università di Cagliari.* Vol. III, Part II. Bologna, Cuppini. 1912.

The goal of the present note is to propose a measure of concentration that is independent of the distribution curve of the character and that allows one to make comparisons between concentration of the most different characters.

2. The larger the difference $p_i - q_i$, the stronger inequality (4) in absolute value; the higher the ratio

$$R_i = \frac{p_i - q_i}{p_i},$$

the stronger in relative value. R_i represents the coefficient by which the fraction of the cases p_i must be multiplied, in which the intensity of the character does not exceed a given limit, in order to obtain the difference between this given fraction and the total sum of the character corresponding to said cases. The value of R_i ranges from 1, in the case of perfect concentration, to 0, in the case of equidistribution of the character.

As a measure of concentration of the character, we can assume the weighted average of the $n - 1$ values of R_i, each term R_i with a weight proportional to p_i.

This average is

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}. \tag{11}$$

We call R *concentration ratio*.

Multiplying R by a value p_i randomly chosen, we obtain the probable value of the corresponding difference $p_i - q_i$. In the case of equidistribution of the character, we have $p_i = q_i$ for all the values of i and hence that $R = 0$; in the case of maximum concentration, in which the total amount of the character is owned by a single unit, we have $q_i = 0$ for all the $n - 1$ values of i and hence that $R = 1$.

Therefore, the concentration ratio is an index of concentration that satisfies the following conditions:

a) it increases as the concentration increases,
b) it is equal to 1 in the case of maximum concentration,
c) it is equal to 0 in the case of minimum concentration,
d) it is the coefficient that, multiplied by the fraction (p_i) of the cases, in which the character presents an intensity below a given threshold, gives the probable value of the difference ($p_i - q_i$) between the fraction itself and the fraction (q_i) of the total amount of the character owned by these cases.

3. Let us consider how to determine the value of R in practice. Replacing p_i, q_i with their respective values $\frac{i}{n}$ and $\frac{A_i}{A_n}$ and taking into account that $\sum_{i=1}^{n-1} i = n(n-1)/2$, (11) can be written in the following, computationally convenient way

$$R = 1 - \frac{2}{(n-1)A_n} \sum_{i=1}^{n-1} A_i, \tag{12}$$

or, taking into account that $\sum_{i=1}^{n-1} A_i = \sum_{i=1}^{n-1}(n-i)a_i$, it can be written as follows

$$R = \frac{2\sum_{i=1}^{n}(i-1)a_i}{(n-1)A_n} - 1. \tag{13}$$

This expression can be simplified further, when several values of a_i are equal. Denoting by x_l $(l = 1, 2, \dots, s)$ one of the s values that the n quantities a_i take on, and by f_i the corresponding frequency, by i_l the number of quantities less than or equal to x_l and by $i_{l-1} = i_l - f_l$ the number of quantities less than x_l, it is straightforward to show that (13) becomes

$$R = \frac{2\sum_{l=1}^{s} x_l \sum_{i=i_{l-1}+1}^{i_l} (i-1)}{(n-1)A_n} - 1$$

or, taking into account that

$$\frac{1}{f_l} \sum_{i=i_{l-1}+1}^{i_l} (i-1) = \frac{i_{l-1} + i_l - 1}{2},$$

(13) becomes[3]

$$R = \frac{\sum_{l=1}^{s}(i_{l-1} + i_l - 1) f_l x_l}{(n-1)A_n} - 1. \tag{15}$$

This formula allows us to determine the value of R quickly, when the values of the character are given for each case. The following example (Tab. 1), that reports the number of heartbeats of 263 Native Americans (males, adults) shows how computations should be organized and performed.

[3] Expression (12) can be simplified similarly. Assuming that $A_l = \sum_{i=1}^{i_l} A_i$, $A_{l-1} = \sum_{i=1}^{i_{l-1}} A_i$ and taking into account that $\sum_{i=1}^{n-1} A_i = \sum_{i=1}^{n} A_i - A_n$,

$$\frac{1}{f_l} \sum_{i=i_{l-1}+1}^{i_l} A_i = \frac{A_l + A_{l-1} + x_l}{2},$$

(12) can be written in the following

$$R = 1 - \frac{\sum_{l=1}^{s} f_l (A_l + A_{l-1}) - A_n}{(n-1)A_n}. \tag{14}$$

Tab. 1: Determination of the concentration ratio using formula (15). *Heartbeats of 263 Native Americans (*Hrdlicka) ([4]).

x_l	f_l	$f_l x_l$	i_l	$i_{l-1} - 1$	$i_l + i_{l-1} - 1$	$(i_l + i_{l-1} - 1) f_l x_l$
44	1	44	1	−1	0	0
45	1	45	2	0	2	90
48	3	144	5	1	6	864
49	1	49	6	4	10	490
50	4	200	10	5	15	3000
51	3	153	13	9	22	1166
52	5	260	18	12	30	7800
53	2	106	20	17	37	3922
54	12	648	32	19	51	33048
55	4	220	36	31	67	14740
56	19	1064	55	35	90	95760
57	7	399	62	54	116	46284
58	24	1392	86	61	147	204624
59	7	413	93	85	178	73514
60	23	1380	116	92	208	287040
61	2	122	118	115	233	28426
62	19	1178	137	117	254	299212
63	11	693	148	136	284	196812
64	19	1216	167	147	314	381824
65	3	195	170	166	336	65520
66	32	2112	202	169	371	783552
67	5	335	207	201	408	136680
68	18	1224	225	206	431	527544
69	1	69	226	224	450	31050
70	12	840	238	225	463	388920
71	2	142	240	237	477	67734
72	12	864	252	239	491	424224
73	1	73	253	251	504	36792
74	3	222	256	252	508	112776
75	1	75	257	255	512	38400
76	2	152	259	256	515	78280
78	3	234	262	258	520	121680
80	1	80	263	261	524	41920
Totals	263	16343	—	—	—	4533688

[4] A. Hrdlicka. *Physiological and medical observations among the Indians of Southwestern United States and Northern Mexico*. Smithsonian Institution Bureau of American Ethnology. Washington. Government printing office. 1908, pages 348-371.

In this case

$$\begin{aligned} n-1 &= 262\\ A_n &= 16343\\ R &= \frac{4533688}{262 \times 16343} - 1 = 5.88\%. \end{aligned}$$

In a similar way I have also computed the value of R for the heartbeats of 94 Egyptians (males, adults) of Kharga Oasis([5]). It turned out that $R = 6.73\%$.

Often the intensities of the character are grouped in more or less wide classes and sometimes, but not always, the sum of the intensities of the cases is also provided for each given class.

Let us first consider the situation in which, for each given class, the sum of intensities is known. Assume that the n quantities are split into r classes and let k denote the k-th ordered class, $k = 1, 2, \dots, r$. Let f_k be the number of quantities in the k-th class, S_k the sum of such quantities, l_k and l_{k-1} the upper and lower bounds of class k, respectively, i_k the number of quantities less than l_k and i_{k-1} the number of quantities less than l_{k-1}.

Let $\delta_{kl} = a_i - \frac{S_k}{f_k}$ $(l = 1, 2, \dots, f_k)$ be the deviation of a_i (of class k) from the arithmetic mean of the analogous quantities f_k in the same class and let

$$\varepsilon_{kl} = (i-1) - \frac{i_k + i_{k-1} - 1}{2}$$

be the deviation of $(i-1)$ from the arithmetic mean of the analogous values f_k in the same class. We have that

$$\sum_{l=1}^{f_k} \delta_{kl} = 0, \qquad \sum_{l=1}^{f_k} \varepsilon_{kl} = 0,$$

and hence that

$$2 \sum_{i=i_{k-1}+1}^{i_k} (i-1)a_i = (i_k + i_{k-1} - 1)S_k + 2\sum_{l=1}^{f_k} \varepsilon_{kl}\delta_{kl}.$$

Using the above equality in (15) we obtain

$$R = \frac{\sum_{k=1}^{r}(i_k + i_{k-1} - 1)S_k + 2\sum_{k=1}^{r}\sum_{l=1}^{f_k} \varepsilon_{kl}\delta_{kl}}{(n-1)A_n} - 1. \tag{16}$$

([5]) A. Hrdlicka, *The natives of Kharga Oasis Egypt.* Smithsonian Miscellaneous Collections. Vol. 59. Number 1, City of Washington, 1912. Pages 112-115.

The quantity $\sum_{k=1}^{r} \sum_{l=1}^{f_k} \varepsilon_{kl}\delta_{kl}$ is always positive([6]). Let

$$R' = \frac{\sum_{k=1}^{r}(i_k + i_{k-1} - 1)S_k}{(n-1)A_n} - 1, \tag{17}$$

we always have that

$$R' < R.$$

The difference $R - R'$ increases with the values $\varepsilon_{kl}/(n-1)$ and δ_{kl}/A_n. Now, the values $\varepsilon_{kl}/(n-1)$ and, *coeteris paribus*, the values δ_{kl}/A_n, increase with f_k/n; furthermore for the same values f_k/n, the values δ_{kl}/A_n increase as the differences between the values a_i become larger. Hence, we can say that the difference $R' - R$ increases with the concentration of the character and with the wideness of the classes the intensities of the character are grouped in.

However, in practice, when one has ten classes, whose wideness are not too much different one from the other, R' can be considered a sufficiently accurate approximation of R. The following table (Tab. 2) reports the steps for computing R' using formula (17) for the heartbeat data of the 263 Native Americans, grouped in 7 classes([7]).

([6]) In fact, the values a_i increase with the values $(i-1)$, that are an arithmetic progression of common difference equal to 1. Hence, in each class the median of the values $(i-1)$ coincides with the mean of the same values, that is equal to $(i_{k-1} + i_k - 2)/2$.

a) If also the median of the values a_i of the k-th class coincides with the arithmetic mean S_k/f_k, then all the f_k product terms $\varepsilon_{kl}\delta_{kl}$ are positive, since the two terms ε_{kl} and δ_{kl} have the same sign.

b) If the median of the values a_i of the k-th class does not coincide with the arithmetic mean S_k/f_k, let v be the number of terms between the two mean values. Of the f_k product terms $\varepsilon_{kl}\delta_{kl}$, v are negative, since v is the number of times the terms ε_{kl} and δ_{kl} are of different sign; while $f_k - v$ are positive.

α) Suppose that the arithmetic mean is greater than the median of the terms a_i. Then, there are v negative products in which ε_{kl} is positive and δ_{kl} is negative, $(f_k/2) - v$ positive products in which ε_{kl} and δ_{kl} are both positive, $f_k/2$ positive products in which both ε_{kl} and δ_{kl} are negative. Note that all the v values δ_{kl} in the negative products are, in absolute values, less than any value δ_{kl} in the positive products in which both ε_{kl} and δ_{kl} are negative and note also that the v values of ε_{kl} in the negative products are equal, term by term, to the v smallest values of ε_{kl} in the positive products in which both ε_{kl} and δ_{kl} are negative. Therefore, the absolute value of the sum of the v negative product terms is less than the sum of the v smallest product terms in which both the terms are negative and, *a fortiori*, smaller than the sum of all the positive terms.

β) The proof is analogous if we assume the median of the terms a_i to be greater than the arithmetic mean.

([7]) For sufficiently large values of n, instead of using (17), R' can be obtained by using the following formula

$$R'_2 = \frac{\sum_{k=1}^{n}(i_k + i_{k-1})S_k}{nA_n} - 1. \tag{19}$$

It is straightforward to check that $R'_2 = \frac{n-1}{n}R'$.

Tab. 2: Determination of the concentration ratio using formula (17). *Heartbeats of 263 Native Americans (*HRDLICKA)

l_{k-1}	l_k	f_k	S_k	i_k	$i_{k-1}-1$	$i_k+i_{k-1}-1$	$(i_k+i_{k-1}-1)S_k$
44	49	6	282	6	−1	5	1410
50	54	26	1367	32	5	37	50579
55	59	61	3488	93	31	124	432512
60	64	74	4589	167	92	259	1188551
65	69	59	3935	226	166	392	1542520
70	74	30	2141	256	225	481	1029821
75	80	7	541	263	255	518	280238
Totals		263	16343	–	–	–	4525631

$$R' = \frac{4525631}{262 \times 16343} - 1 = 5.69\%.$$

The value obtained for the concentration ratio is then 3% less than the exact value (5.88%), determined using (15). For the Egyptians of Kharga, grouping the data in 9 classes, we obtain $R' = 6.59\%$, 2% less the exact value, 6.73%.

In order to point out the influence of the number of classes on R', dr. G. Pietra (Direzione generale della Statistica, Rome, Italy), computed the value of R' for private land property in the State of Victoria in 1910([8]), first considering the more refined classification into 30 categories provided by official statistics, and then progressively restricting the number of categories to 25, 20, 15, 10, 8, 6, 5, 4.

Tab. 3: Victoria, 1910. Private land property.

Number of classes	Value of R'
30	69.0%
25	68.9%
20	68.8%
15	68.4%
10	67.8%
8	66.5%
6	65.8%
5	64.6%
4	58.1%

([8]) The data are from the *Statistical Register of the State of Victoria for the year 1911.* Melbourne, Mullet. Part VIII. Production, page 107.

In Table 3, the value of R' decreases as the number of classes decreases; with 10 classes, the reduction of R' (2% compared to the value of R' for 30 classes) can still be deemed negligible.

In other cases, of course, the reduction of R' might be quite different. In the following example (Tab. 4), the reduction is much more relevant.

Tab. 4: Belgium, 1911. Stipends of public employees[9].

Number of classes	Value of R'
32	33.9%
12	31.5%

The value of R' has a 4% reduction as the number of classes is restricted from 32 to 12[10].

Note that it is not the number of classes itself that matters but rather the fact that, as this number decreases, these classes become wider and wider.
Of course, the values of R' might result different if the quantities are grouped in the same number of classes but in different ways. It could be the case that with different groupings of the initial quantities, one can obtain smaller values of R' with a larger number of classes, if the classes become more unequal one from the other. For the Victoria example, dr. Pietra computed the values of R', grouping the 30 classes in 22 classes in three different ways (Tab. 5). The resulting values of R' are different from each other and from the value found using 20 classes of more homogeneous wideness.

Tab. 5: Values of R'.

on 22 classes obtained grouping the classes			on 20 classes obtained grouping the classes
from 1st to 5th and from 26th to 30th	from 6th to 10th and from 21th to 25th	from 11th to 15th and from 16th to 20th	8th–9th; 11th–13th; 14th–16th; 17th–18th; 19th–20th; 23th–24th; 25th–26th; 29th–30th
0.6844	0.6698	0.6875	0.6879

[9] *Ministère des Finances - Secrétariat général. - Tableau statistique du nombre et des traitments des magistrats, fonctionnaires et employés civils de l'Etat e des Ministres des cultes retribués par l'Etat. 1-1-1911.* Bruxelles, G. Piquart. 1913. The data used to compute R' do not include the clergy's stipends.

[10] Computation of R' for the Belgium data is due to prof. de' Stefanis, University of Padova, Italy.

A second approximation of R can be obtained by computing the value of

$$\frac{2}{(n-1)A_n}\sum_{k=1}^{r}\sum_{l=1}^{f_k}\varepsilon_{kl}\delta_{kl},$$

under specific hypotheses.

For instance, let us suppose that the f_k values of a_i in the same class form an arithmetic progression. Then, for each l and for each k, $\delta_{kl} = \varepsilon_{kl}\frac{c_k}{f_k}$, where $c_k = l_k - l_{k-1}$; hence

$$\sum_{k=1}^{r}\sum_{l=1}^{f_k}\varepsilon_{kl}\delta_{kl} = \sum_{k=1}^{r}\sum_{l=1}^{f_k}\varepsilon_{kl}^2\frac{c_k}{f_k}.$$

It can be shown that([11])

$$\sum_{l=1}^{f_k}\varepsilon_{kl}^2 = \frac{f_k(f_k^2-1)}{12}$$

and that

$$\frac{2}{(n-1)A_n}\sum_{k=1}^{r}\sum_{l=1}^{f_k}\varepsilon_{kl}\delta_{kl} = \frac{1}{6(n-1)A_n}\sum_{k=1}^{r}(f_k^2-1)c_k.$$

Under the above hypothesis, we obtain the following approximation for R([12])

$$R'' = \frac{1}{(n-1)A_n}\left[\sum_{k=1}^{r}(i_k+i_{k-1}-1)S_k + \frac{1}{6}(f_k^2-1)c_k\right] - 1. \qquad (21)$$

Note that, in approximating $\sum_{k=1}^{r}\sum_{l=1}^{f_k}\varepsilon_{kl}\delta_{kl}$, we assume that all the product terms are positive, which is true only if S_k/f_k coincides with the median of the class. This makes us understand why R'' is often larger than R.

[11] see *Variabilità e Mutabilità*, pages 51-52.

[12] For a sufficiently large value of n, instead of using (21), R'' can be determined using the following expression

$$R_2'' = \frac{1}{nA_n}\left[\sum_{k=1}^{r}(i_k - i_{k-1})S_k + \frac{1}{6}f_k^2 c_k\right] - 1. \qquad (20)$$

It is easy to check that

$$R_2'' = \frac{n-1}{n}R'' + \frac{D}{6(n-1)A_n},$$

where $D = \sum_{k=1}^{r} c_k$ is the difference between the largest and smallest a_i.

In the Native Americans and in the Egyptians examples, we can assume that, for any k, $c_k = 5$. Then, for the Native Americans seriation,

$$\sum_{k=1}^{r} \left(f_k^2 - 1 \right) = 14332$$

$$R'' = 5.97\%;$$

for the Egyptians seriation

$$\sum_{k=1}^{r} \left(f_k^2 - 1 \right) = 1499$$

$$R'' = 6.78\%.$$

The exact values are 5.88% and 6.73%, just negligibly smaller than the approximations.

Sometimes statistics provide only the maximum and the minimum values of a seriation: this is often the case with biological data; in other circumstances, as for economic data relative to incomes, rents, patrimonies, this information is not available. In these cases we do not know neither the upper limit of the last class nor the lower limit of the first class and the values c_1 and c_r can only be determined approximatively, using some possible artifices.

In order to determine c_1, sometimes one can assume that the lower limit of the first class is equal to zero. For instance, this assumption is realistic for the minimum extension of a land property.

An artifice that can often be used for determining c_1 or c_r is assuming that the average intensity of the class coincides with the class midpoint, the mean value of its extreme values. Solving with respect to l_k and l_{k-1} the equality

$$\frac{S_k}{f_k} = \frac{1}{2}(l_k + l_{k-1})$$

then, given the other terms in the equality, one can easily determine the value of l_k or l_{k-1}. For example, in the classification of the land property in Victoria, year 1910, the last category includes two properties of 116486 acres.

From equation

$$\frac{116486}{2} = \frac{1}{2}(50000 + x)$$

we obtain the upper limit

$$x = 66486.$$

In the corresponding classification for the year 1906, the last class includes 6 properties with 366766 acres. From equation

$$\frac{366766}{6} = \frac{1}{2}(50000 + x)$$

the upper limit is

$$x = 72255.$$

The approximation obtained using this approach depends on the distribution of the quantities within the first and the last class.

The approximations obtained for heartbeats of Native Americans and of Kharga Egyptians are good. Following this procedure for the first and the last classes of Table 2, which is relative to Native Americans, we find 79,6 and 45 as extreme values; while the true values are 80 and 44. For Egyptians of Kharga, the estimated values are 104, 3 and 57 and the true ones are 105 and 54. Even relevant errors in the determination of c_1 and c_r have, in general, a limited effect on the resulting value of R''.

For the land property in the State of Victoria, 1910, dr. Pietra has also computed the values of R'' using formula (20) for several number of classes (Tab. 6). Note that, as the number of classes is larger than 4, the values of R'' are stable, with differences only at the third digit.

Tab. 6: Victoria, 1910. Private land property.

Number of classes	Value of R''
30	69.1%
25	69.2%
22	$69.5 - 69,3 - 69,2\%$[13]
20	69.2%
15	69.1%
10	69.2%
8	69.1%
6	69.6%
5	69.5%
4	71.1%

Similarly, for the Belgian public employees example, the reduction of the number of classes from 32 to 12 affects the value of R'' only at the third digit. Using formula (20), Prof. de' Stefanis obtained the following results (Tab. 7).

[13] These values have been obtained using different groupings of the classes. The corresponding values of R' are 67.0%, 68.8%, 68.4%. See Tab. 5.

Tab. 7: Belgium, 1911. Stipends of public employees.

Number of classes	Value of R'
32	34.1%
12	34.3%

Statistics often provide, for each class, just the value of f_k but not the value of S_k. In these cases one can determine an approximate value of R assuming that, for each class, the average intensity corresponds to the mean of the extreme values of the class([14]). Assuming that

$$S_k = \frac{1}{2} f_k (l_k + l_{k-1})$$

and

$$A_n = \sum_{k=1}^{r} S_k = \frac{1}{2} \sum_{k=1}^{r} f_k (l_k + l_{k-1}),$$

(21) becomes([15])

$$R''' = \frac{\sum_{k=1}^{r} (i_k + i_{k-1} - 1) f_k (l_k + l_{k-1}) + \frac{1}{3} (f_k^2 - 1) c_k}{(n-1) \sum_{k=1}^{r} f_k (l_k + l_{k-1})} - 1 \tag{23}$$

In this case, the knowledge of lower and upper limits of the classes is necessary not only to compute the correction term $\frac{1}{6} \sum_{k=1}^{r} (f_k^2 - 1) c_k$, but also to determine approximatively the value of S_k. Of course, ignoring the minimum and maximum values of a seriation might represent in this case a much more serious problem than for determining R''; the inconvenient is much more serious since one cannot resort to the method described above for determining the smallest and the largest terms when the values S_k are not given. When the minimum and maximum values of a seriation, as well as the values of S_k, are not given, one cannot determine the lower bound of the first class and the upper bound of the last class without introducing some arbitrariness; but the error in determining such values has a relevant influence on the value of R''' only if the frequencies of the first and the last classes are large.

([14]) One can resort to this hypothesis for the computation of R when statistics provide, for each class, only the value of S_k and not the value of f_k. This is a quite rare circumstance; this is the case, for instance, of statistics of Swiss farms size in 1905 *(Betriebe)*.

([15]) For values of n sufficiently large, one can use the following expression

$$R_2''' = \frac{\sum_{k=1}^{r} (i_k + i_{k-1}) f_k (l_k + l_{k-1}) + \frac{1}{3} f_k^2 c_k}{n \sum_{k=1}^{r} f_k (l_k + l_{k-1})} - 1 \tag{22}$$

which is obtained from (20) as expression (23) is obtained from (21).

When the wideness of intermediate classes is the same, a plausible hypothesis is that extreme classes have the same wideness as well. This is the case of the classification we have adopted for Native Americans and Egyptians heartbeats. The following table (Tab. 8) contains the elements needed for determining R''' for the Native Americans example; the maximum (79) and the minimum (45) of the seriation have been determined according to the assumption described above.

Tab. 8: Determination of the concentration ratio using formula (23). *Heartbeats of 263 Native Americans (*Hrdlicka).

l_{k-1}	l_k	$l_{k-1}+l_k$	f_k	$f_k(l_{k-1}+l_k)$	$i_k+i_{k-1}+1$	$f_k(l_{k-1}+l_k)$ $(i_k+i_{k-1}+1)$
45	49	94	6	564	5	2820
50	54	104	26	2704	37	100048
55	59	114	61	6954	124	862296
60	64	124	74	9176	259	2376384
65	69	134	59	7906	392	3099152
70	74	144	30	4320	481	2077920
75	79	154	7	1078	518	558404
Totals		–	263	32702	–	9077024

We have

$$\sum_{k=1}^{r} f_k(l_{k-1}+l_k) = 32702$$

$$\sum_{k=1}^{r} f_k(l_{k-1}+l_k)(i_k+i_{k-1}-1) = 9077024$$

and, as found above,

$$\sum_{k=1}^{r}(f_k^2-1) = 14332 \qquad c_k = 5.$$

It follows that

$$R''' = \frac{9077024+\frac{5}{3}14332}{262\times 32702} - 1 = 6.22\%.$$

For the Egyptians example, we have that $R''' = 6.90\%$.

In both the examples above, the value of R''' turns out to be larger than the exact value of R determined using (15) (= 5.88% for the Native Americans and 6.73% for the Egyptians).

In other cases it might be convenient to resort to alternative hypotheses; and these should naturally change, depending on the specific situations, according to the distribution of the character investigated.

For land property in Victoria, it seemed sensible to set equal to zero the lower limit of the first class and to compute approximately the upper limit of the last class, assuming that in the last three classes the character's frequency curve is a portion of an equilateral hyperbola. By reducing progressively the number of classes, and by computing the upper limit of the last class each time, dr. Pietra obtained, under these hypotheses, the following values for R_2''' (Tab. 9).

Tab. 9: Victoria, 1910. Private land property.

Number of classes	Value of R_2'''
30	69.5%
25	69.7%
22	70.4 – 71.2 – 67.6%
20	69.7%
15	69.5%
10	70.1%
8	71.6%
6	70.2%
5	68.1%
4	52.5%

Except for the case in which 4 classes are considered, the values of R_2''' do not depart significantly from those of R'' given in Table 6. Values of R_2''' had been also determined by computing the upper limit of the last class under hypotheses that differ from the previous one, reaching similar results.

These uses of formulas (23) and (22) are quite important as they show that one can typically determine accurately enough the concentration ratio for a character even in those situations in which statistics provide just a classification of the character intensity and, for each class, give only the number of cases.

4. We report below some values of the concentration ratio, ranked increasingly, that have been computed by dr. Pietra, prof. Savorgnan and myself, for some human physiological characters (Tab. 10) and for some economic characters (Tab. 11).

For each character, only individuals with such a character are reported (for instance, for the number of children, only families with children; for patrimonies and successions, only those included in the census or dead people who have been assessed a patrimony). In the following we will consider what results are obtained by also including those individuals with intensity of the character equal to zero.

Comparison of the several concentration ratios considered above may suggest interesting considerations, both from the biological and the economical point of view. Notice how much different the concentration is for economic and physiological characters: for most of the physiological characters, the value of R is smaller than 10%; for most of the economic characters R is greater than 50%. It is of particular interest the comparison between the concentration of the number of children left at marriage dissolution (34%) and the concentration of patrimony amongst the wealthies (71-93 %), because the concentration of the surviving children at the marriage dissolution represents a minimum limit of the patrimony concentration, which would take place in a generation, for the sole fact that families are differently prolific, in case each family of the preceding generation, at its dissolution, would leave its children the same patrimony.[16]

Tab. 10: Concentration of some human physiological characters [17]

Character		Value of R
Sublingual temperature	Kharga Egyptians[18]	0.4%
Cephalic module	” ”	1.2%
Head antero-posterior diameter	” ”	1.5%
Auditory meatus to bregma distance	” ”	1.6%
Height	” ”	1.7%
Maximum head cross-sectional diameter	” ”	1.8%
Cephalic index	” ”	2.0%
Left foot length	” ”	2.3%
Height	Italian conscripts born in 1890[19]	2.5%*
Left foot second finger	Kharga Egyptians[18]	2.9%
Left ear length	” ”	3.0%
Left foot width	” ”	3.2%
Left ear height	” ”	3.5%
Heartbeats	Native Americans[20]	5.9%*
Heartbeats	Kharga Egyptians[18]	6.7%*
Respiration	” ”	7.7%
Right hand pressure strength	” ”	9.7%
Traction strength	” ”	14.8%

(continue)

[16] See *Indici di concentrazione*, etc. in *Biblioteca dell'Economista*, V Serie. Vol. XX, pp. 78-79. Considerations therein about concentration index δ can be repeated for the concentration ratio. The concentration ratio of the number of surviving children at the marriage dissolution can in fact also be retained equivalent, in a sufficiently large number of observations, to the patrimony concentration ratio that would result amongst the children in the case all the couples at the moment of marriage dissolution would leave the children the same patrimony. That can easily be seen graphically by depicting concentrations of the two characters using the method indicated in Section 6.

[17] The values of R with an asterisk had been computed by the author; the remaining ones by prof. de' Stefanis and by his fellows at the Technical Institute in Vicenza.

[18] Hrdlicka. *The natives*, etc..

[19] Ministero della Guerra. *Della leva di terra sui giovani nati nell'anno 1890.* Roma, Voghera, 1913.

[20] Hrdlicka. *Physiological and medical observations*, etc.

Character		Value of *R*
N. children remaining at marriage dissolution	Budapest 1903-1908[21]	33.5%*
N. children living at the moment of the census	France 1901[22]	34.0%*
N. children born up to the moment of the census	France 1906[22]	37.4%*
N. children born up to marriage dissolution	Budapest 1903-1908[21]	37.5%*

Tab. 11: Concentration of some economic characters [23].

Character			Value of *R*
Farms size	Germany	1907[24]	23%**
”	Belgium	1895[25]	25%**
”	France	1892[26]	28%**
Public employees stipends	Belgium	1911[27]	34%**
Owner run farms size	Switzerland	1905[28]	43%**
”	Denmark	1905[29]	43%**
Labor incomes	Argovia	1892[30]	44%**
Incomes	Denmark	1909[31]	46%
Farms size	Serbia	1905[32]	50%
Incomes	Sassony	1910[33]	50%
”	Norway	1906[34]	52%
Dwelling rent values	Paris	1911[35]	63%*
Farms size	Holland	1887[36]	63%**
Land property value	Victoria	1911[37]	64%

(continue)

[21] *Statistisches Jahrbuch der Haupt-und Residenzstadt Budapest.*1903-1908.

[22] Statistique Général de la France. *Statistique des familles en 1906.* Paris. Imprimerie nationale, 1906.

[23] The values of R with an asterisk have been computed by the Author; those with two asterisks by dr. Pietra; the remaining ones by prof. de' Stefanis and his fellows at the Techical Institute of Vicenza. For the concentration ratio of postal saving in Austria, see note 42.

[24] *Statistisches Jahrbuch für das Deutsche Reich.* 1913.

[25] *Annuaire statistique de la Belgique.* 1912.

[26] *Statistique agricole de la France. Résultats généraux del l'enquéte décennale de 1892.*

[27] *Tableau statistique*, etc..

[28] *Résultats du recensement fédéral des enterprises agricoles, industr. et comm. du 9 Aoùt 1905.*

[29] *Danmarks jordbrug, 1850-1905, udgivet af statens statistikt bureau.* Copenhagen, 1907.

[30] J. Kristler. *Erhebungen über Vermögen, Shulden und Erwerb im Kanton Aargau in den Jahren in 1892, 1886 und 1872.* Bern. Stämpfli u. Cie. 1895.

[31] Statistique du Danemark. *Revenus et fortunes d'après la taxation pour 1909-10.* Kiobenhavn. Bianco Lunos. 1912.

[32] On the basis of hand written communications to the “Istituto Int. di Agricoltura” in Rome.

[33] L. Dugé de Bernonville. *Distribution de salaires et de revenus en divers pays. Bull. de la Stat. générale de la France.* Juillet 1913.

[34] A. N. Kiaer *Répartition des revenus en Norvège.* Kristiania Bjoernstads. 1907-1910.

[35] *Annuaire statistique de la Ville de Paris.* 1911.

[36] *Handwörterbuch der Staatswissenschaften.* Zweite Auflage, Zweiter Band. pages 437 and following.

[37] *Statistical Register of the State of Victoria for the year 1912. Part II. Finance.*

Character			Value of R
Farms size	Ireland	1896([36])	66%**
”	Great Britain	1895([36])	68%**
Personal work income	Victoria	1911([37])	69%
Property incomes	Victoria	1911([37])	69%
Land property size	Victoria	1910([38])	69%**
”	West Australia	1911([39])	70%**
Patrimonies	Argovia	1892([40])	71%
Land property size	South Australia	1911([39])	72%**
”	Denmark	1901([41])	76%**
Postal savings	Austria	1900([42])	78%
Successions	Victoria	1908-1910([43])	81%
Patrimonies	Zurich Canton	1909([44])	82%
Land property size	Tasmania	1911([45])	83%**
”	N. South Galles	1911([45])	85%**
Successions	France	1904([46])	88%
Patrimonies	Denmark	1909([47])	93%

5. There are some characters that, in a large or small number of cases, present a positive intensity and, in the remaining cases, a null intensity; as an example, consider the individual income. We will call *positive cases* those of the first group, *null cases* those of the second group and *total cases* the totality of the cases.

We might consider the measurement of the concentration of a character among the total cases or just among the positive cases. Let n be the positive cases and v the null cases.

Once we have determined the concentration ratio of the character for the positive cases - that will be denoted as R_p - it is straightforward to obtain the concentration ratio for the total cases - denoted as R_t - with the following expression

$$R_t = R_p \ m + (1 - m) \tag{24}$$

([38])*Statistical Register of the State of Victoria for the year 1911. Part VIII, Production.*

([39])*Official Year Book of the Commowealth of Australia.*

([40])J.Kistler. *Erhebungen über Vermögen, Schulden und Erwerb im Kanton Aargau in den Jahren 1892, 1886 und 1872.* Bern. Stämpfli u. Cie. 1895.

([41])*Annuaire statistique du Danemark.*

([42])Concentration ratio taken from F. Savorgnan. *Il risparmio postale in Austria dal 1883 al 1912. Ricerche statistiche.* Trieste, 1914. Balestra.

([43])A. M. Laughton. *Victorian Year Book 1910-11.* Part IV. *Accumulation.*

([44])*Uebersicht der Vermögens - und Einkommenssteuerpflichtigen des Kantons Zürich,* etc. Zürich. G. Meyer. 1910.

([45])*Official Year Book of the Commonwealth of Australia.*

([46])*Bullettin de Statistique et de Législation comparée. 1905.*

([47])Statistique du Danemark. *Revenus et fortunes d'après la taxation pour 1909-10.* Kiobenhavn. Bianco Lunos. 1912.

where m is the ratio $(n-1)/(n+v-1)$. We always have that $R_t > R_p$ and the difference increases as m and R_p decrease[48].

The values of R_p and R_t are compared in Table 12 for several characters.

Tab. 12

	m	R_p	R_t
N. children at census. France:			
Surviving children (1901)	0.841	34.0%	44.5%
Children born (1906)	0.885	37.4%	44.6%
N. children at marriage dissolution. Budapest (1903-1908):			
Surviving children	0.686	33.5%	54.4%
Children born (dead or living)	0.737	37.5%	53.9%
Successions. Victoria (1908-1910)	0.371[49]	80.5%	92.8%
Successions. France (1903-1904)	0.687[49]	87.8%	91.6%
Patrimonies. Argovia (1892)	0.860[50]	71.2%	75.2%
Patrimonies. Denmark (1909)	0.836[50]	92.7%	93.9%

These results make us realize that it can be highly misleading to draw conclusions on the concentration of wealth in different countries on the basis of just the wealth concentration for wealthy people. The concentration of the successions amongst those whose inheritance is taxed, turns out to be ($R = 80.5\%$) in Victoria, much lower than the patrimony concentration amongst those who are assessed a patrimony in Denmark ($R = 92.7\%$); but if one takes into account the persons that statistically result as propertyless, the concentration ratio turns out to be only slightly different in the two countries (Victoria=92.8 %; Denmark=93.9 %).

[48] The value of R_p is obtained from (15), that can be written as

$$R_p = \frac{\sum_{l=1}^{s}(i_{l-1} + i_l - 1) f_l x_l - (n-1)A_n}{(n-1)A_n},$$

the value of R_t is

$$R_t = \frac{\sum_{l=1}^{s}(i_{l-1} + v + i_l + v - 1) f_l x_l - (n+v-1)A_n}{(n+v-1)A_n},$$

that can also be written as

$$R_t = R_p \frac{n-1}{n+v-1} + \frac{2v\sum_{l=1}^{s} f_l x_l - vA_n}{(n+v-1)A_n}.$$

This expression is easily shown to be equal to (24), noting that

$$A_n = \sum_{l=1}^{s} f_l x_l, \qquad m = \frac{n-1}{n+v-1}.$$

[49]The ratio of successions to adult dead.

[50]The ratio of the persons that result patrimony owners to the persons that would result in the case of maximum wealth diffusion.

6. The ratio that we are proposing in this note as the appropriate measure of concentration, can also be obtained by improving a graphical method already introduced by some authors, as Lorenz, Chatelain, Séailles([51]), in order to evaluate inequality in the distribution of wealth. If in a Cartesian diagram, we report the values p_i on the abscissa and the values q_i on the ordinate and we connect the points (p_i, q_i), the resulting curve is the *concentration curve* (Fig. 1), which is increasing and convex([52]).

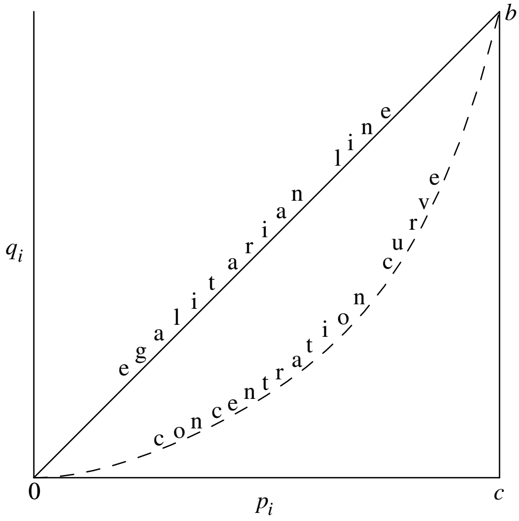

Figure 1.

The less unequal is the wealth distribution, the less accentuated is the concentration curve, that tends to a straight line (*egalitarian line*) in the case of equidistribution.

([51]) See M.O. Lorenz, *Method of measuring the concentration of wealth,* in Publications of the American Statistical Association. N. 70, June 1905; see also G.P Watkins, *Comment on the method of measuring concentration of wealth.* Ibidem. N. 72, December 1905; G.P Watkins, *An interpretation of of certain statistical evidence of concentration of wealth.* Ibidem. N. 81, March 1908; W.M. Person. *The variability in the distribution of wealth and income,* in *The Quarterly Journal of Economics*. Vol. XXIII, N.3. May 1909; G.P Watkins, *The measurement of concentration of wealth.* Ibidem. Vol. XXIV, N.1. November 1909; W.M. Person. Ibidem; W.J. King. *The elements of Statistical Methods.* New York. The Macmillan Company, 1912.
See also E. Chatelain, *Les succession déclarées en 1905,* in *Revue politique et parlamentaire*, 1907; *Le tracé de la courbe des successions en France,* in *Journal de la Soc. de Stat. de Paris,* 1910. Pages 352 and following; *La fortune française d'après les successions en 1909,* in *La Démocraties*. 20 Janvier 1911; J. Séailles, *La réparition des Fortunes en France,* Paris. Alcan, 1910.

([52]) The fact that the concentration curve is an increasing and convex function follows respectively from (1) and (2).

The above mentioned authors relied on this property of the concentration curve for comparing the distribution of wealth.

Drawing on the same plot several curves, relative to different times or places, they were able to assess in what time or place the wealth was more concentrated.

This graphical approach presented two drawbacks, promptly acknowledged by Lorenz and by King:

a) it does not provide a precise measurement of concentration;
b) it does not allow to assess, not even in some circumstances, when or where concentration is stronger. In fact, if two curves cross each other (Fig. 2), it is not always possible to say if one denotes a stronger concentration than the other. This drawback, that can be deemed not relevant for the comparison of phenomena of the same nature (e.g. the concentration of incomes for two different years or countries), is particularly serious for comparing phenomena of different nature, whose distribution's shape differs.

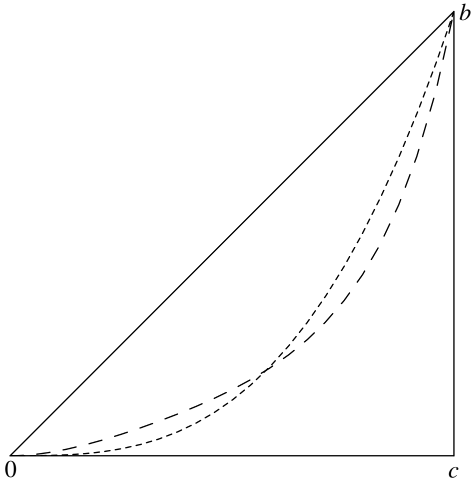

Figure 2.

Both the above drawbacks disappear if, as measurement of the concentration, we consider the ratio between the area limited by the concentration curve and the egalitarian line (*concentration area*) and the area of the triangle *obc*, that represents the concentration area in the case of maximum concentration. It is now straightforward to show that this ratio is the limit the concentration ratio R tends to, when the number n of cases increases and the distribution of the character is unchanged. Consider the following plot (Fig. 3), where $n = 14$.

The $n-1$ small rectangles limited by the p_i-axis and by the concentration curve have height equal to q_i and basis equal to $1/n$; the sum of their areas is then equal to $\frac{1}{n}\sum_{i=1}^{n-1} q_i$. The $n-1$ small rectangles limited by the p_i-axis

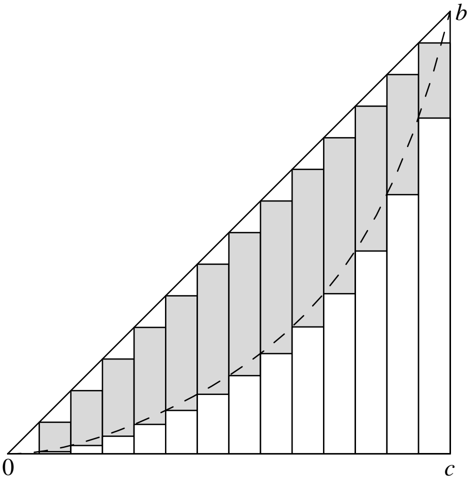

Figure 3.

and by the egalitarian line have basis equal to $1/n$ and height equal to p_i, since each ordinate forms with the egalitarian line and the p_i-axis an isosceles triangle: the sum of their areas is $\frac{1}{n}\sum_{i=1}^{n-1} p_i$. The difference between the two sums is $\frac{1}{n}\sum_{i=1}^{n-1}(p_i - q_i)$, that is equal to the sum of the areas of the $n-1$ triangles limited by the egalitarian line and the concentration curve. Now, as n increases, the areas of the small surfaces limited by the egalitarian line and the upper side of the rectangles of height p_i decreases and, analogously, the area of the small surfaces limited by the concentration curve and the upper sides of the rectangles of height q_i decreases as well. Hence, as n increases, the area of the *obc* triangle tends to the value $\frac{1}{n}\sum_{i=1}^{n-1} p_i$, the concentration area tends to $\frac{1}{n}\sum_{i=1}^{n-1}(p_i - q_i)$ and their ratio tends to the concentration ratio $R = \sum_{i=1}^{n-1}(p_i - q_i)/\sum_{i=1}^{n} p_i$.

Note that, in order to describe the concentration curve, it is not necessary to know all the values of p_i and q_i. The knowledge of 4 or 5 of these values is enough to approximate the curve with sufficient accuracy.

7. The above considerations suggest a further procedure for determining the value of R in practice.

Let us plot the concentration curve and let us consider the segment *oc* as unit of measurement. The area of the triangle *obc* is equal to $1/2$; the concentration area can be easily computed using an integraph or a planimeter or, more coarsely, by reporting the diagram on millimeter paper employing a very large scale and by counting the number of squared millimeters and of their fractions within the concentration area.

This method, that is quite simple and fast, can be in principle followed not only when we know the intensity of the character for each single case, but also when we only know the number of cases and their total amount for category, even if these are not numerous. In practice, however, this method is not safe when the concentration curve is not accentuated and the concentration area is small, since, in this case, small mistakes in constructing the diagram might have a not negligible influence of the results. And, even when this drawback does not show up, we are still left with the unavoidable problem that the measurement of the area obtained using the integraph or the planimeter or the millimeter paper is affected by computational or by accidental causes with the final effect that two people, or the same person in two different occasions, will always obtain more or less different results. When the intensity of the character for all the cases is given or when the classes in which the cases are grouped are sufficiently numerous, so that we can assume that (21) provides a good approximation of R, in my opinion one should not avoid the arithmetic computation of the concentration ratio, and the graphical computation should be used for an initial and rapid inspection. Of course, the smaller the number of classes, the more important the graphical computation with respect to arithmetic computation will be; the relevance of the former approach will be overwhelming when the number of classes is so small that (21) cannot be trusted to provide a good approximation of R.

8. The graphical representation of the concentration curve is useful for visualizing the relationship between the exact value of the concentration ratio, R, computed using formula (15), and its approximation from the bottom, R', computed using (17) (Fig. 4).

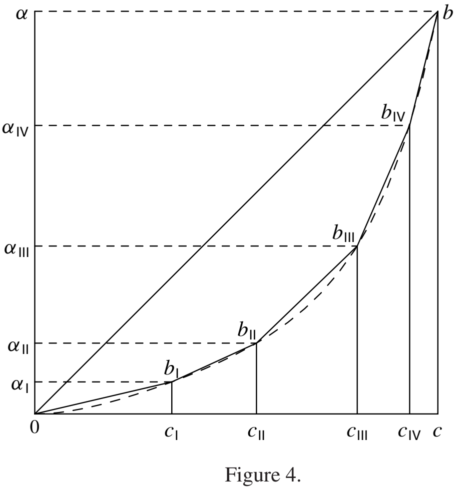

Figure 4.

Suppose that the n intensities of the character are grouped in 5 classes and let $\overline{a_I b_I}$, $\overline{a_{II} b_{II}}$, $\overline{a_{III} b_{III}}$, $\overline{a_{IV} b_{IV}}$, $\overline{ab}$ be abscissas proportional to the corresponding 5 values of i_k and let $\overline{b_I c_I}$, $\overline{b_{II} c_{II}}$, $\overline{b_{III} c_{III}}$, $\overline{b_{IV} c_{IV}}$, $\overline{bc}$ the ordinates proportional to the corresponding values or A_k. It is straightforward to show that the ratio between the area delimited by the line ob and the piecewise line $ob_I b_{II} b_{III} b_{IV} b$ and the area of the triangle obc is the limiting value of R' as n increases([53]); while, as shown in section 6, the ratio between the area delimited by the line ob and the curve and the area of triangle obc is the limiting value of R as n increases. Since the concentration curve is convex, the piecewise line $ob_I b_{II} b_{III} b_{IV} b$ is inscribed inside the concentration curve and the area delimited by the piecewise linear function and the line ob is less than the area delimited by the concentration curve and the line ob. Examining the plot, it can be argued that the smaller is the number of the classes, the larger is such difference and that, typically, for a fixed number of classes, the more unequal the number of cases of each class, the more accentuated this difference.

9. Let us now examine the relationship between the concentration ratio and the indices of variability, used to characterize the distribution of the variables investigated.

Variability indices can be classified in two categories:

a) indices which measure variability of the character by means of an average

([53]) Let a_k denote one of the r values of a and b_k the corresponding value of b. The area of the trapezium $a_k a_{k-1} b_{k-1} b_k$ is equal to

$$\overline{a_k a_{k-1}} \times \frac{1}{2}(\overline{a_k b_k} + \overline{a_{k-1} b_{k-1}}).$$

Note that

$$\overline{a_k a_{k-1}} = \frac{S_k}{A_n}; \qquad \overline{a_k b_k} = \frac{i_k}{n}; \qquad \overline{a_{k-1} b_{k-1}} = \frac{i_{k-1}}{n}.$$

Hence, the area of the trapezium is equal to $[(i_k + i_{k-1})S_k]/(2nA_n)$; the area of the surface delimited by segments $\overline{oa}$ and $\overline{ab}$ and the piecewise line is equal to $[\sum_{k=1}^{r}(i_k + i_{k-1})S_k]/(2nA_n)$; the area of the surface delimited by the egalitarian line and the piecewise line is equal to $[\sum_{k=1}^{r}(i_k + i_{k-1})S_k]/(2nA_n) - 1/2$ and the ratio between this last value and the area of of triangle obc is equal to

$$\frac{\sum_{k=1}^{r}(i_k + i_{k-1})S_k - nA_n}{nA_n}. \tag{25}$$

Recalling that $\sum_{k=1}^{r} S_k = A_n$, from (17) it follows that

$$R' = \frac{\sum_{k=1}^{r}(i_k + i_{k-1})S_k - nA_n}{(n-1)A_n},$$

that, up to the value $1/n$, coincides with the value given by (25), its limiting value as n increases.

of (absolute values of) deviations of each intensity from an average intensity of the character. Depending on the average used, different variability indices are obtained. Typically, one chooses either the arithmetic mean, obtaining the *simple mean deviation*; or the quadratic mean([54]), obtaining the *quadratic mean deviation*; or the median, obtaining the *probable deviation*. Note that one can use different average intensity for defining each of the above variability indices: most typically, deviations of intensities are considered from the arithmetic mean and, sometimes, from the median of the character.

b) indices based on averages of (absolute value of) differences between the intensities of the character. If the arithmetic mean is used as average, the resulting variability index is called *mean difference*.

Indices in the first category answer the question: what is the difference between the observed intensities of the character and a certain average value of the same intensities? These indices play a special role in those contexts in which it is sensible to assume that such an average corresponds to a quantity that actually exists and that the observed intensities deviate from this average due to errors of measurement (as those occurring in Physics, Astronomy or in Geodesy), or it corresponds to a quantity that can be regarded as typical and such that the observed intensities can be assumed to depart from it due to accidental oscillations (as it is often the case in surveys relative to characters of a species).

The second category indices give an answer to the problem: what is the difference between the various intensities of the character? These indices have a particular research value when one of the two above mentioned circumstances is not satisfied, for example in surveys related to income, property, renting, newborns, marriages, deaths; in general, in economic statistics and in demography. In *Variabilità e Mutabiltà* we have pointed out the relationship between indices of categories *a*) and *b*), and we have analyzed in detail the research contexts in which either the one or the other are appropriate.

We now show that the concentration ratio coincides with the ratio between the mean difference and its maximum value, or, in other words, with the ratio of the mean difference with twice the arithmetic mean of the character.

We know that the mean difference is given by the following formula

$$\Delta = \frac{2}{n(n-1)} \sum_{i=1}^{\frac{n+1}{2}} (n+1-2i)(a_{n-i+1} - a_i)$$

and that its maximum is equal to $2M_n = 2A_n/n$([55]). Hence, it is sufficient to

([54]) With the expression quadratic mean of different quantities I mean the square root of the arithmetic mean of the squared quantities.

([55]) See *Variabilità e Mutabiltà*, page 22 and page 80.

show that the following equality holds

$$R = \frac{\sum_{i=1}^{\frac{n-1}{2}} (n+1-2i)(a_{n+1-i} - a_i)}{(n-1)A_n} \tag{26}$$

or, recalling that from (12) we have

$$R = \frac{(n-1)A_n - 2\sum_{i=1}^{n-1} A_i}{(n-1)A_n},$$

to show that

$$\sum_{i=1}^{\frac{n+1}{2}} (n+1-2i)(a_{n+1-i} - a_i) = (n-1)A_n - 2\sum_{i=1}^{n-1} A_i. \tag{27}$$

Let us start by noting that

$$\begin{gathered}
(n-1)(a_1 + a_2 + \ldots + a_n) = \\
(n-1)a_1 + (n-2)a_2 + (n-3)a_3 + \ldots + 2a_{n-2} + a_{n-1} + \\
+ a_2 + 2a_3 + \ldots + (n-3)a_{n-2} + (n-2)a_{n-1} + (n-1)a_n.
\end{gathered}$$

By subtracting from both sides the sum

$$2[(n-1)a_1 + (n-2)a_2 + (n-3)a_3 + \ldots + 2a_{n-2} + a_{n-1}],$$

the following equality is obtained

$$\begin{gathered}
(n-1)(a_1 + a_2 + \ldots + a_n) - 2\{(n-1)a_1 + (n-2)a_2 + \ldots + a_{n-1}\} = \\
= a_2 + 2a_3 + \ldots + (n-2)a_{n-1} + (n-1)a_n + \\
-(n-1)a_1 - (n-2)a_2 - (n-3)a_3 - \ldots - a_{n-1}
\end{gathered}$$

and from this, taking into account the previous equalities, we obtain

$$(n-1)A_n = (n-1)(a_1 + a_2 + \ldots + a_n)$$

$$\sum_{i=1}^{n-1} A_i = (n-1)a_1 + (n-2)a_2 + \ldots + 2a_{n-2} + a_{n-1}$$

$$\begin{aligned}
\sum_{i=1}^{\frac{n+1}{2}} (n+1-2i)(a_{n+1-i} - a_i) = (n-1)a_n + (n-2)a_{n-1} + \ldots + 2a_3 + a_2 - \\
- (n-1)a_1 + (n-2)a_2 - \ldots - 2a_{n-2} - a_{n-1}
\end{aligned}$$

that leads to (27).

10. The relationship between R and Δ allows us to provide a further proof of the relationships between R and its approximations, R' and R''.

Grouping in r classes the n quantities whose mean difference is Δ, it is straightforward to show that the mean difference of the n quantities obtained by replacing each initial quantity with the mean of the class it belongs to is

$$\Delta' = \Delta - \frac{\sum_{k=1}^{r} f_k(f_k - 1)\Delta_k}{n(n-1)}, \tag{28}$$

where Δ_k $(k = 1, 2, \dots, r)$ denotes the mean difference of the f_k quantities in the k-th class([56]).

Furthermore in the previous section we have shown that $R = \Delta/(2M_n)$ and similarly that $R' = \Delta'/(2M_n)$.

It follows that

$$R' = R - \frac{\sum_{k=1}^{r} f_k(f_k - 1)\Delta_k}{2n(n-1)M_n}. \tag{29}$$

This equality shows that the smaller is Δ_k (variability in each class) with respect to M_n (general mean) and the smaller is the sum $\sum_{k=1}^{r} f_k(f_k - 1)$ with respect the total number of differences, $n(n-1)$, the closer is the approximate value R' to the true value R. It can also be argued that the smaller the number r of classes in which the n observations are grouped in and the more unequal the numbers f_k of cases of each class, the larger the sum $\sum_{k=1}^{r} f_k(f_k - 1)$. The second term in the right-hand-side of (29) can be approximated assuming that the values f_k of the k-th class form an arithmetic progression of common difference equal to c_k/f_k, where c_k is the wideness of the class. Under this hypothesis([57])

$$\Delta_k = \frac{(f_k + 1)c_k}{3f_k}.$$

Replacing the values of Δ_k given by this equality in (29) we obtain the approximation R'' for R that is exactly that given by formula (21).

[56] Note that the $n(n-1)$ differences, obtained by the n quantities and having mean difference equal to Δ, can be split into two categories: *a*) differences between quantities in the same class; the number of these differences is $\sum_{k=1}^{r} f_k(f_k - 1)$; *b*) differences between quantities that belong to different classes.
If each quantity is replaced by the class mean, the sum of the differences of category *b*) is unchanged, while the differences of category *a*) are all equal to zero. In order to obtain the value $n(n-1)\Delta'$ we have to subtract from the value $n(n-1)\Delta$ the sum of the differences between quantities in the same class, that is $\sum_{k=1}^{r} f_k(f_k - 1)\Delta_k$.

[57] See *Variabilità e Mutabiltà*, page 52.

11. The relationship established between R and Δ is important since it allows to exploits the conclusions made on Δ in *Variabilità e Mutabiltà*, where:

a) we examined the influence that the random choices with and without replacement have on the probable value of Δ([58]);
b) we examined the relationship between values of Δ obtained for parallel seriations([59]);
c) we examined the relationships that, for a given distribution, exist between the value of Δ and the simple or quadratic mean deviations from the arithmetic mean or the median. These relationships vary according to the distribution of the character. We have considered the relationships of the above mentioned variability indices when the intensities of the character are distributed as an arithmetic progression, as a geometric progression, as an accidental errors curve, as an hyperbolic curve in the case of maximum inequality in distribution([60]);
d) we checked the discrepancies that might turn out in the measurement of the variability of a character by using either the mean difference or one of the other variability indices([61]). Further considerations in this regard have been made by dr. Giovanni Dettori([62]) ;
e) we have pointed out the relationships between the ratio of the mean difference over the arithmetic mean, the index α of income's distribution, introduced by Pareto, and the index δ of concentration of incomes according to formula (7)([63]);
f) we have examined when, in measuring the variability of characters, it is more convenient the use of the mean difference or of another variability index([64]);
g) we have examined when and to what extent in the comparison of the variability of characters their mean values should be taken into account.

All these conclusions can be extended to the theory of concentration ratio as well, taking into account that the concentration ratio is equal to the ratio between the mean difference and twice the arithmetic mean of the character.

12. Sometimes we do not know the entire seriation of the character but just a truncated seriation, that contains all the cases in which the character shows

([58]) Pages 37-46.
([59]) Pages 46-49.
([60]) Pages 49-83.
([61]) Pages 83-86.
([62]) *Contributo allo studio della variabilità dei prezzi* in *Studi economico-giuridici della R. Università di Cagliari.* Vol. IV. Part I. Cagliari, Dessì, 1912.
([63]) Pages 70-80.
([64]) Pages 86-99

up with an intensity larger than a given threshold. This happens for instance in economic statistics, considered for fiscal purposes, and relative to incomes, patrimonies, salaries, rents, for which an exemption limit is admitted. Let us examine the impact on the concentration ratio of considering the truncated seriation rather than the complete seriation of the character.

Let k be the number of cases ignored out of the n available, p_k the fraction of the cases ignored in the truncated seriation and q_k the fraction of the total amount of the character owned by these cases.

In the following plot (Fig. 5), where, as usual, $\overline{oc} = p_n = 1$; $\overline{cb} = q_n = 1$, we have that $\overline{og} = p_k$; $\overline{dg} = q_k = \overline{ec}$; $\overline{gc} = 1 - p_k = \overline{de}$; $\overline{be} = 1 - q_k = \overline{ef}$; $\overline{df} = (1-q_k)-(1-p_k) = p_k - q_k$; area $obc = \frac{1}{2}$; area $bde = \frac{1}{2}(1-p_k)(1-q_k)$; area $fbd = \frac{1}{2}(p_k - q_k)(1 - q_k)$; area $dfo = \frac{1}{2}(p_k - q_k)q_k$.

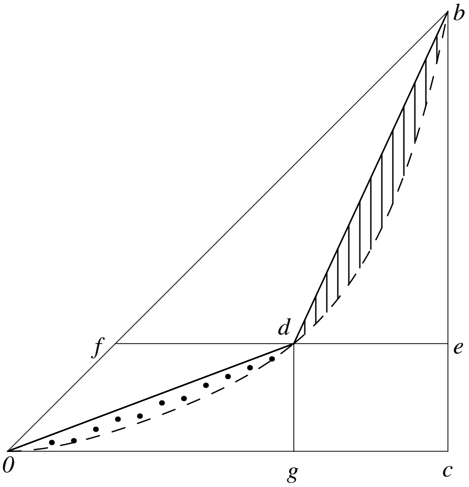

Figure 5.

Let C denote the concentration area for the complete seriation, C' the concentration area for the truncated seriation (dashed area), T the dotted area in the plot, R the concentration ratio for the complete seriation and R_{-p_k} the concentration ratio for the truncated seriation in which the fraction p_k of cases is ignored.

From the equalities

$$R = \frac{C' + \text{area } fdb + \text{area } fdo + T}{\text{area } obc}$$

and

$$R_{-p_k} = \frac{C'}{\text{area } bde}$$

it follows that

$$R = R_{-p_k}(1 - p_k)(1 - q_k) + p_k - q_k + 2T. \tag{30}$$

The value of T can be approximately computed under the hypothesis that character intensities for the ignored cases form an arithmetic progression. This hypothesis might lead to very different approximations, depending on the situation. When it can also be supposed that the first term of the progression is very close to zero, we can assume that[65]

$$\frac{T}{\text{area } odg} = \frac{1}{3}$$

and, as a consequence, (30) becomes

$$R = R_{-p_k}(1 - p_k)(1 - q_k) + p_k - q_k + \frac{1}{3} p_k q_k \tag{31}$$

implying that

$$R \gtrless R_{-p_k} \qquad \text{if} \qquad R_{-p_k} \gtrless 1 - \frac{q_k\left(2 - \frac{4}{3}p_k\right)}{p_k + q_k(1 - p_k)}. \tag{32}$$

[65] In fact, if Δ_k denotes the mean difference of the k ignored cases and M_k their arithmetic mean, then

$$\frac{T}{\text{area } odg} = \frac{\Delta_k}{2M_k}.$$

However, if the intensities of the character in the k ignored cases form an arithmetic progression of common difference equal to H, then (see *Variabilità e Mutabilità*, page 52)

$$\Delta_k = \frac{H(k+1)}{3}$$

and

$$M_k = a_1 + \frac{H(k-1)}{2},$$

where a_1 is the minimum intensity of the character and it follows that

$$\frac{T}{\text{area } odg} = \frac{k+1}{3(k-1)} \frac{M_k - \frac{1}{2}a_1}{M_k}$$

that, for a sufficiently large k and a_1 very close to zero, can be assumed to be equal to $\frac{1}{3}$.

Inequality (32) shows that, in different occasions, it might be either $R > R_{-p_k}$ or $R < R_{-p_k}$; in other words, the concentration ratio computed on the truncated seriation can differ and, depending on the circumstances, it can be either larger or smaller than the concentration index computed on the entire seriation.

In one situation, remarkably relevant in practice, we have approximatively that $R = R_{-p_k}$. If

$$f_l = V x_l^{-h} dx_l \tag{33}$$

denotes the number of times the character assumes a value between x_l and $x_l + dx_l$, and the following conditions hold: *a*) the minimum value of x_l is very small compared to its maximum, *b*) $h - 2$ is positive and not too close to zero, *c*) the number $n - k$ of cases considered is sufficiently large so that the term $(n - k)/(n - k - 1)$ can be neglected, the mean difference of the $n - k$ values considered for the character can be written as[66]

$$\Delta = \frac{2}{2h - 3} M_{n-k}$$

and the concentration ratio is

$$R = \frac{1}{2h - 3}, \tag{34}$$

regardless of the minimum value of x_l and of the number k of ignored cases.

Under the above hypotheses, the concentration ratio for the truncated seriation might be assumed approximately equal to that of the complete seriation, regardless of the fraction of ignored cases. Now, the above mentioned conditions hold almost always for the global incomes seriations. In fact, as pointed out by Pareto, global incomes follow (33) in their distribution, and almost always with a good approximation; the values of h determined so far range from 2.15 to 2.9, the number of income receivers from official statistics is always very large and the minimum taxed income is always very small compared to the maximum income.

Hence, for global incomes we can expect that the concentration ratio for a truncated seriation is approximately representative of the concentration ratio for the complete seriation.

Prof. de' Stefani checked this claim using Danish incomes (1909). He first computed the value of R'' (=45.6 %) and then the value of R''_{-p_k}, starting from ignoring the 1-st class, then the 1-st and the 2-nd classes, then the 1-st, the 2-nd and the 3-rd classes and so on. The results obtained are reported in Table 13.

[66] See *Variabilità e Mutabilità*, page 63.

Tab. 13: Denmark, 1909. Incomes.

Values of p_k	Values of R''_{-p_k}
0.63	41.3%
0.68	40.3%
0.76	40.1%
0.83	40.3%
0.89	40.9%
0.93	41.8%
0.97	41.6%
0.98	41.0%
0.990	41.1%
0.994	40.6%
0.996	40.1%
0.998	38.9%

The concentration ratio decreases remarkably if the first class is neglected, and this fact, as already noticed by Pareto, confirms that the smallest incomes do not follow (33) very closely. As the following classes are ignored one by one, changes in the concentration ratio are not relevant: in order to have a significative reduction in its value we have to consider only the 0.2% of the cases. This example confirms the claim that, for global incomes, the concentration ratio computed for a truncated seriation does not differ significantly from that computed for the entire seriation.

The conclusion is quite different for the seriations of patrimonies and salaries. The values of R''_{-p_k} have been computed by de' Stefanis for Danish patrimonies (1909) (Tab. 14) and for Belgian public employees (1911) (Tab. 15).

Tab. 14: Denmark, 1909. Patrimonies.

$R'' = 92.7\%$

Values of p_k	Values of R''_{-p_k}
0.78	64.8%
0.90	57.4%
0.93	55.9%
0.97	54.7%
0.986	54.1%
0.995	52.0%

Tab. 15: Belgium, 1911. Stipends of public employees.

$R'' = 34.3\%$

Values of p_k	Values of R''_{-p_k}
0.009	33.9%
0.054	37.2%
0.640	26.0%

For Danish patrimonies, as the truncation fraction increases the decrease of the concentration ratio is continuous and quite relevant; the behavior is more irregular for Belgian stipends. However, for these two characters the concentration ratio computed for a truncated seriation does not even provide an approximation of the value for the entire seriation.

Let us now show how well the value of R can be approximated, using (31), by means of the values of p_k, q_k and R_{-p_k} (Tab. 16, 17, 18). We consider here the values of p_k and q_k and the values computed of R for Danish patrimonies and incomes and for Belgian salaries; the corresponding values of R_{-p_k} can be found in the previous tables.

Tab. 16: Denmark, 1909. Incomes.

$R'' = 45.6\%$

Values of p_k	Values of q_k	Values of R computed $using(31)$
0.633	0.302	50.0%
0.683	0.339	50.6%
0.762	0.414	50.8%
0.828	0.491	50.8%
0.885	0.573	5.1%
0.933	0.661	48.7%
0.968	0.749	46.4%

Tab. 17: Denmark, 1909. Patrimonies.

$R'' = 92.7\%$

Values of p_k	Values of q_k	Values of R computed $using(31)$
0.783	0.052	87.8%
0.900	0.181	82.1%
0.972	0.419	69.8%
0.995	0.644	56.6%

Tab. 18: Belgium, 1911. Stipends of public employees.

$R'' = 34.3\%$

Values of p_k	Values of q_k	Values of R computed using (31)
0.009	0.0016	34.3%
0.054	0.019	38.1%
0.640	0.388	39.2%

For the seriations of Danish incomes and Belgian stipends, the values of R computed using (31) do not differ essentially form those of R'', determined for the whole seriation, even though the values of R are remarkably larger; on the contrary, for Danish patrimonies the values of R are always lower than those of R'', and the larger is the fraction of truncation p_k, the larger is the difference.

In my opinion it would be worthwhile extending this analysis to other characters and, for each of them, to many seriations, in order to establish what is the effect of using a truncated seriation on the concentration ratio. It would also be useful to determine the level of accuracy reached in approximating the value of the concentration ratio for the whole seriation by using the value found for the truncated one, on the basis of (31) or of other formulas that can be proposed under suitable hypotheses.

13. Let us now summarize the results presented in this article.

1. Starting from the concept of *concentration index* (that we have developed and used since September 1908 and more widely in September 1909) we have proposed a concentration index (named *concentration ratio*) that is valid for any character, regardless of its distribution.
2. We have shown that the concentration ratio can be obtained by improving a graphical method proposed by Lorenz (1905), Chatelain (1907), Séailles (1910) for representing wealth distribution.
3. We have shown that the relationship between the concentration ratio and the mean difference for the intensities of a character is quite simple: the former is equal to the ratio between the latter and the mean difference computed for the case of maximum inequality or, in other words, it is equal to the mean difference over twice the arithmetic mean of the character.
4. We have indicated arithmetic and graphical procedures suitable for computation of the concentration ratio, both in the cases in which statistics provide complete information on the intensity of the character and also in the cases in which information are not complete.

5. We have considered several examples of use of the concentration ratio, both for human physiological characters and for economic characters.

In this way, the theories of concentration and variability indices are connected, showing the meaning that graphical methods, proposed for representing the distribution of wealth, have with respect to these theories.

(*Corrected proofs sent to print on June* 25, *1914*)