ECONOMIC GROWTH CENTER

YALE UNIVERSITY

Box 1987, Yale Station
New Haven, Connecticut

CENTER DISCUSSION PAPER NO. 366

INCOME DISTRIBUTION IN BUENOS AIRES

Hector L. Dieguez and Alberto Petrecolla*

January 1981

Notes: * Visiting Fellow, Economic Growth Center

Center Discussion Papers are preliminary materials circulated to stimulate discussion and critical comment. References in publications to Discussion Papers should be cleared with the author to protect the tentative character of these papers.

1. Introduction

The analysis of the importance of socioeconomic characteristics of households in the determination of income distribution has already won a good reputation in the literature. This paper attempts a contribution on this line by applying a technique, not yet explored[1] in order to evaluate the independent and joint effects of economic and social attributes of households and their contributions to total inequality. The methodology first suggested by Bhattachanya and Malanobis /1/ and developed by Pyatt /11/ and Fei, Ranis and Kuo, /4/ and /5/, is extended to allow for multivariate analysis. The decomposition of the Gini coefficient is carried out in a way that discriminates differences in income distribution supporting and contradicting a set of hypothesis.

The basic information was taken from an as yet unpublished survey of family incomes and expenditures in the Greater Buenos Aires /8/ designed and processed by using a methodology developed in the ECIEL Program[2] for a number of urban centers in Latin America. The survey collected information on incomes, expenditures, and attributes of households and individuals for the period going from July 1969 to June 1970.

Section 2 sketches the decomposition of the Gini coefficients. Section 3 summarizes the main features of income distribution in the Greater Buenos Aires and presents a decomposition of the Gini coefficient in order to link the size and the functional distributions. Section 4 examines the role played by socioeconomic variables in the determination of inequality by means of a multivariate analysis. Section 5 studies the association among the variables. Section 6, finally, states briefly the main features and results of the research.

2. Methodology of Gini Decomposition

The Gini coefficient for any population of size n with m income levels can be expressed as the mean of all possible income differences between units, measured in terms of the population average income, that is,

$$G = \frac{1}{M} \sum_{i=1}^{m} \sum_{j=1}^{m} p_i \cdot p_j \cdot \text{Max}\ (0,\ y_i - y_j) \qquad \text{(E.1)}$$

where C is the Gini coefficient, y_i and y_j are income levels, p_i and p_j are population shares, and M average income.

Since G is a sum of income differences weighted by population shares[3] it can be dissaggregated in many ways. Particularly, when the population is classified into a number of mutually exclusive classes, the coefficient can be decomposed into the sum of weighted income differences between units belonging to the same classes and the sum of income differences between units of different groups. The first set of terms expresses inequality within classes, and it can be written as the sum of the Gini coefficients of the classes weighted by the product of the corresponding population and income shares.

On the other hand, those components obtained by comparing incomes of units belonging tc different classes can be divided into those that can be summarized by differences in the average incomes of classes and those that appear when distributions of classes overlap.

In fact, for any two classes, h and k, it can be shown that D_{kh}, the weighted sum of differences in incomes, as defined above, between units belonging to k with regard to those belonging to h, can be written as

$$D_{kh} = p_k \cdot p_h \cdot (M_k - M_h) + D_{hk} \qquad \text{(E.2a)}$$

where M_k and M_h are the average incomes of classes k and h, relative to the total average income. Expression (E.2a) indicates how inequality originated in income differences between units belonging to different classes can be decomposed. Assuming that $M_k > M_h$, $D_{hk} > 0$ means that there are households in h (the class with lower average income) that have incomes larger than those of some units in k, that is to say, that distributions overlap. At the same time, it shows that income differences between households in k and those in h can be expressed by the difference in average incomes weighted by population shares, plus a term equivalent to D_{hk}. On the other hand, (E.2a) can be rewritten as

$$D_{kh} + D_{hk} = p_k \cdot p_h \cdot (M_k - M_h) + 2D_{hk} \qquad \text{(E.2b)}$$

Hence, inequality accounted for by income differences of units belonging to different classes include an effect of differences in average incomes and an effect of overlapping distributions. It is also clear that half of this last effect is due to income differences emerging because some incomes of h are higher than some of k, and the other half to the opposite situation.

Summing up, the Gini Coefficient can be disaggregated into three effects: the effect of inequality within classes; the effect of average income differences between classes; and the effect of overlapping distributions. We will refer to them simply as inequality effect, differences effect, and overlapping effect, respectively.

This decomposition has several interesting properties. To begin with, any of the three effects can in turn be disaggregated to allow for more detailed analysis. Moreover, this kind of disaggregation makes it possible to test hypothesis.[4] For example, the assertion that individuals belonging to the class k have higher incomes than those belonging to h can be confronted with the results of the disaggregation. The part of the Gini coefficient accounted by the inequality effect neither supports nor contradicts the hypothesis. In turn, the differences effect would support the hypothesis if $M_k > M_h$ and would contradict it if $M_h > M_k$. Finally, half of the overlapping effect would contradict and half would support the hypothesis.

In addition, the disaggregation just presented can readily be transformed in another that links the functional and the size distribution of income.[5] Since the Gini coefficient is defined as the sum of weighted differences that contradict and of those that support a given hypothesis $(d^+ + d^-)$, we can also compute the net gap of differences $(d^+ - d^-)$, and then define

$$R = \frac{d^+ - d^-}{d^+ + d^-} \qquad \text{(E.3)}$$

R would be equal to 1 if all income differences supported the hypothesis; it would be - 1 if all of them contradicted it. Positive values of R indicate that the income differences supporting the hypothesis outweigh those that contradict it. The opposite is true when R is negative.

Consider now the distribution of income of a given source (k) among all the individuals in a given population and the relation between this particular distribution and that of total income among the same population. The hypothesis that income from this source increases with total income can be tested as previously indicated. If R is positive, it means that inequality in the income distribution from such a source adds to total inequality. On the contrary, a negative value of R would indicate that inequality in the distribution of income from source k diminishes total inequality.

It can be shown that if there are s sources of income, the Gini coefficient can be written as

$$G = \sum_{k=1}^{s} \phi k \cdot R_k \cdot G_k \qquad (E.4)$$

where ϕ_k is the share in income of source k. This decomposition links size distribution and sources of income. We turn now to the consideration of the corresponding results for the Greater Buenos Aires.

3. Size Distribution and Sources of Income

Let us start looking at the size distribution of total family income[6] in the Greater Buenos Aires, shown in Table 1. The Gini coefficient (0.3826) is smaller than any other obtained for the various Latin American cities included in the ECIEL project, as can be seen comparing with results presented in /6/. It also reveals a greater inequality than the one founded in /10/ for Australian urban centers.

The overall inequality includes relatively large differences in both extremes of the distribution and rather small ones in the intermediate intervals, as Table 1 makes clear. The shares of income derived from the several sources considered varies in each bracket.[7] Wages and salaries have a relatively large and decreasing share from the second to the eighth bracket, and a lower participation in the first and especially in the highest income interval. Income from self-employment shows the opposite pattern, with a share that decreases in the first three brackets and then increases, reaching its highest value for upper income families. Transfers are important only in the first three brackets, while imputed rents[8] increase steadily with income. Income from ownership of capital is important only in the highest income group.

The distribution of the different income sources contributes to total inequality as shown in Table 2. Wages and salaries are more evenly distributed than any other kind of income, while incomes from capital and transfers have the largest inequalities. The Gini coefficient for self-employment is also high, essentially because this kind of income is earned both by low income groups and by professionals and others on the top of the distribution.[9]

As it was shown in E.4, these inequalities in the sources of income play

TABLE 1

DISTRIBUTION OF TOTAL HOUSEHOLD INCOMES IN THE GREATER BUENOS AIRES

(July 1969 - June 1970)

Income Intervals (current pesos)	Number of Households	Incomes	Average Income of Each Interval (current pesos)	Percentage Differences of Average Incomes Between Intervals
	(as % of total)			
1 - 4200	10,2	2,4	2928	-
4201 - 5800	10,0	3,9	5004	70,6
5801 - 7000	9,2	4,6	6369	27,4
7001 - 8400	11,8	7,2	7745	21,5
8401 - 9800	9,4	6,8	9176	18,5
9801 - 11800	9,6	8,2	10823	17,9
11801 - 14000	10,4	10,5	12834	18,5
14001 - 16800	9,4	11,3	15354	19,6
16801 - 24000	10,7	17,0	20230	31,8
24001 y más	9,4	28,1	38068	88,1
TOTAL	100,0	100,0	12695	-

Gini coefficient: 0.3826

TABLE 2
DECOMPOSITION OF INEQUALITY BY INCOME SOURCES

Source of Income	Share in Total Income (%)	Correlation with Total Income (R)	Gini Coefficient*	Percentage of House-holds with-out income of the Source	Contribution to Total	
					Value	% of Gini
Wages and Salaries	38.6	.5003	.3181 (.5506)	34.1	.1063	27.8
Self-employment	25.3	.6441	.4822 (.8015)	61.7	.1307	34.1
Capital	4.0	.7615	.6787 (.9747)	92.1	.0305	8.0
Imputed Rent	17.5	.7023	.4074 (.6585)	42.4	.0807	21.1
Transfers	9.4	.2001	.4030 (.7666)	60.9	.0143	3.7
Others	5.2	.4628	.5631 (.8543)	66.7	.0206	5.4
Gini coefficient for total income					.3826	100.-

*The first value refers to households that receive income from the source. The second one (between brackets) to all the households, i.e., includes families not having income from the source. They are related by the expression $G_s = G^*_s \cdot p_s + (1 - p_s)$, where G^*_s is the Gini computed by including only households receiving income from s.

different roles in the determination of total inequality, according to the share of each source in total income and to the magnitude (in this case the sign is always positive) taken by R. In our case, this coefficient is high for incomes derived from capital and self-employment, moderate for wages and salaries, and low for transfers. As a result, the contribution of the distribution of wages and salaries to total inequality is lower than the share of labor. The same happens with transfers even though the Gini for this kind of income is high.[10] On the contrary, the contribution to inequality of self-employment and income from capital results much larger than their income shares.[11]

Another fact deserves consideration. In the Greater Buenos Aires almost 60% of the families live in their own houses and imputed rent accounts for more than 20% of the total inequality. This proportion would increase if the use of other durable goods were included in order to impute rents.

4. Multivariate Analysis

4.1 First Decomposition of the Gini Coefficient

In a previous paper[12] an univariate analysis was presented. Variables such as education, occupation, family size, age, ownership of capital, sex, and others were considered one at the time. Here we propose a way to extend the method to multivariate analysis, aiming at a better understanding of the independent and joint effects of the variables. For this purpose we selected the four variables that in the previous study were found to be the most improtant, that is to say, that showed the largest effect of differences among average incomes of the groups. Three of them refer to attributes of the family head (education, occupation, and age) and the fourth to the household (family size). For each variable, classes were given values that correspond to the ranking as regards avefage income in the univariate analysis. For size, education and occupation the ranking coincides exactly with a priori judgement. Such a kind of judgement is instead less clear for the age of the family head.[13]

For the multivariate analysis the population was divided into 300 classes by combining all the classes of the four variables. Average income, population and income shares, and the Gini coefficient of every group was computed. An additional class was also defined in order to include families on which no valied information could be obtained for any of the variables.[14] The results of the decomposition of the Gini using this multivariate classification are presented in Table 3. The discriminatory power of the chosen classification and the large number of classes taken into consideration explain the high relative importance of the differences effect and the very small (practically negligible) of the inequalities effect.

Table 3

Multivariate decomposition of the Gini Coefficient into three effects

	Contributions	
	Values	% of Gini
Inequalities effect	0.0033	0.86
Differences effect	0.2720	71.10
Overlapping effect	0.1073	28.04
TOTAL	0.3826	100.00

4.2 The Hypothesis

It is often assumed that differences between groups show the amount of inequality "explained" by the classification adopted. In general, this is not correct since the direction of the differences must be taken into account. Moreover, even if income differences run in some expected direction on the average, there could be some households not following that pattern. As we have already pointed out half of the income differences composing the overlap effect run in one direction and the other half in the opposite.[15]

For these reasons we believe it is necessary to build first a set of hypothesis and only thereafter to decompose inequality distinguishing differences of incomes that support it from those that contradict it.

Let us start by a simplified set of hypothesis. When two classes of households (or two individual households) differ in the values of the four variables taken into account and all these differences run in the same direction then the class (the household) showing higher values is expected to have higher income.[16] Similar hypothesis is assumed for the cases in which there are one, two, or three control variables[17] and the remaining ones have higher values in one of the classes(cases 1.1 to 1.4 in Table 4). In these four cases the effect of any variable reinforces the effect of the others. We call them "cases without opposite variables". The highest proportion of differences supporting the hypothesis is expected to be found among these groups, decreasingly as the number of control variables increases. Results are expected to be less conclusive when differences in the values of the non-control variables run in different directions ("cases with opposite variables").

Four additional cases have to be distinguished. When some variables have higher values in one class and some in the other, then the class having more variables with higher values is expected to have higher average income (cases 2.1 and 2.2 in Table 4). If the classes are opposed two to two (that is to say, two variables have higher values in one class and the other two in the other class, case 2.3), the class with a higher value in education is expected to have higher income. Finally, if there are two control variables and the other two oppose one to one (that is to say, one variable has a higher value in each of the two classes under comparison, case 2.4 in Table 4), the class whose head has higher education is assumed to have higher income; if education is one of the control variables, the higher income will correspond to the class with a higher value in occupation; finally, if both education and occupation are control variables, the higher income will be expected in the class with higher value in size.[18]

Since we have postulated the hypothesis in terms of classes of households, in what follows we limit our attention to the differences effect. An operational difficulty in the Gini decomposition applied to a multivariate classification is the large number of terms in this effect. In our present case there are 45.150 terms, so that it is crucial to find a suitable way to group them. As a first step we have divided them into 8 sets of terms, corresponding precisely to the cases distinguished in the hypothesis, as detailed in Table 4.[19]

On the whole, for the eight cases considered together, 87.6% of the contributions to the inequality support the hypotheses and 12.4% contradict it.[20] However, the pattern is quite different as we move along the lines of the table, fully in agreement with the qualifications formulated to the hypothesis. On one extreme (cases 1.1 to 1.3, corresponding to minimum opposition) we find the higher proportion of values supporting the hypothesis; on the other extreme (cases 2.2 to 2.4, with maximum opposition) we find the lower proportions ;

and there is an intermediate zone (cases 1.4 and 2.1) where the proportion of contributions supporting the hypothesis takes values between those of the extremes. Roughly speaking we may say that the first five cases (1.1 to 1.4 and 2.1) support rather satisfactorily the hypothesis, representing more than 70% of the differences effect. In the other three cases the results are less neat. They will be reexamined below (in 4.3).

Table 4

Test of hypothesis: The effect of average incomes differences disaggregated in eight sets of terms.

	Contributions to the effect		Supporting the hypothesis		Contradicting the hypothesis	
1. Cases with no opposite variables[1/]	Value	%	Value	%	Value	%
1.1 No control variables	.0486	18.6	.0485	99.8	.0001	0.2
1.2 One control variable	.0566	21.7	.0556	98.9	.0011	1.1
1.3 Two control variables	.0372	14.2	.0348	93.6	.0024	6.4
1.4 Three control variables	.0114	4.3	.0095	84.0	.0018	16.0
2. Cases with some opposite variables						
2.1 Three variables vs. one[2/]	.0308	11.8	.0267	86.6	.0041	13.4
2.2 Two variables vs. one[3/]	.0491	18.8	.0356	72.5	.0135	27.5
2.3 Two variables vs. two[4/]	.0123	4.7	.0091	74.0	.0032	26.0
2.4 One variable vs. one[5/]	.0157	5.9	.0095	60.8	.0061	39.2
TOTAL	.2617		.2293		.0323	

1/- All the variables taking different values in the two classes under comparison have higher values in the same class.

2/- Three variables have higher values and the fourth a smaller value in one class.

3/- Two variables have higher values and another a smaller value in one class (the fourth is a control variable).

4/- Two variables have higher values and the other two smaller in one class.

5/- One variable has a higher value and another a smaller value in one class (the other two are control variables).

A second natural step in the disaggregation process[21] consists in the consideration of 40 cases by distinguishing the variables. For instance, case 1.1 in table 4 (one control variable, the other three taking higher values in one class) is disaggregated in four, according to which is the control variable. This further disaggregation of the figures suggests the strength of education and the weakness of age as explanatory variables. The joint effect of age and the other variables appears feeble and in the other extreme it is easily appreciated the power of education and occupation running together in the same direction.[22]

As regards the cases with opposite variables we already noticed that they are characterized by the higher proportion of contributions contradicting the hypothesis. Now, at the new level of disaggregation (40 cases) it can be seen that in five cases the contributions contradicting the hypothesis overpowered the contributions supporting it. This finding reinforces the need to improve the set of hypothesis. We explore this line in 4.3, limiting our effort to the consideration of cases 2.2 to 2.4, where the results are less satisfactory.

In order to complete the consideration of the multivariate analysis based upon the hypothesis formulated in their simple form, we try to assess the relative importance of the variables. Given the strength of the joint effects, any way of imputing values is somehow arbitratary, so that we need to make clear the criteria to be followed.

To impute values to individual variables in the cases without opposite variables (1.1 to 1.4) we proceed to divide equally the contributions supporting the set of hypothesis among the non-control variables while contributions rejecting it are considered non-imputable. When there are opposite variables (2.1 to 2.4) contributions supporting the set of hypothesis are divided equally among the vatiables whose effect as assumed to prevail, while those contributions contradicting it are attributed to the variables assumed weaker. The results obtained are presented in panel A of Table 5. The proportion of non-imputable differences is quite small (only 2% of the differences effect). The variables rank as assumed in the hypothesis: education, occupation, size and age.

It seems to be also relevant to impute the effect of differences in average incomes to groups of variables. In panel B of the same table the results for couples of variables are presented, using imputing criteria similar to those used for individual variables. The non-imputable contributions are in this case larger since there are cases in which it is not at all possible to make imputations to pairs of variables (as in cases 1.4 and 2.4). The joint effect of education and occupation is considerably higher than any other, while the lower values correspond to age combined with any of the other three variables. It does not seem necessary to show results for combinations of three variables. It is enough to point out that the most important combination is education-occupation-size.

Table 5

The differences effect and the relative importance of the socioeconomic variables

A. Individual variables

Cases	Size	Age	Education	Occupation	Non-imputable
1.1	.0121	.0121	..0121	.0121	.0001
1.2	.0120	.0126	.0147	.0164	.0011
1.3	.0057	.0052	.0118	.0121	.0024
1.4	.0021	.0014	.0031	.0028	.0018
2.1	.0080	.0059	.0085	.0084	-
2.2	.0104	.0063	.0182	.0143	-
2.3	.0026	.0016	.0046	.0036	-
2.4	.0040	.0022	.0061	.0033	-
TOTAL	.0569	.0473	.0791	.0730	.0054

B. Pairs of variables

Cases	Size Age	Size Educ.	Size Occup.	Age Educ.	Age Occup.	Educ. Occup.	Non- imputable
1.1	.0081	.0081	.0081	.0081	.0081	.0081	.0001
1.2	.0060	.0082	.0099	.0087	.0104	.0125	.0011
1.3	.0033	.0042	.0038	.0030	.0041	.0163	.0024
1.4	-	-	-	-	-	-	.0114
2.1	.0033	.0043	.0054	.0034	.0045	.0056	.0041
2.2	.0040	.0045	.0035	.0029	.0026	.0181	.0135
2.3	.0017	.0025	.0010	.0009	.0005	.0057	-
2.4	-	-	-	-	-	-	.0157
TOTAL	.0264	.0318	.0317	.0270	.0302	.0663	.0483

4.3 Three alternative ways for further consideration of the hypothesis.

In order to examine in greater detail the cases in which the evidence supporting the hypothesis is weaker, three alternative roads are explored, mainly in order to indicate possible extensions of the research.

In the first place, cases 2.2 to 2.4 of table 4 were reconsidered by giving only two values to every variable.[23] The rationale behind this procedure is quite simple. The hypothesis stated above took into account only the fact that the value of a given variable was higher or lower in one class, but no consideration was given to the magnitude of the difference. However this could be done in different ways. We have followed this line postulating a very simple weighting pattern: differences in the values of a variable were given a zero weight if both units belonged to the same consolidated class, while the weight was one for differences in attributes of units corresponding to different new classes.

A certain improvement results from this neoclassification: the sum of differences supporting the hypothesis increased from 0.2293 to 0.2335, and that contradicting it diminished from 0.0323 to 0.0218. A small proportion (0.0063) neither supports nor contracts it because it corresponds to previous differences in attributes that were consolidated.

The transformation to dichotomous variables reversed the five cases that previously contradicted the hypothesis. All of the 40 cases examined register higher contributions supporting the hypothesis than the ones contradicting it.

A second possible way to refine the hypothesis consists in taking into account the association that exists among the variables. As an example we have considered different patterns of incomes along the life-cycle for different occupational groups. The cycle for the whole population is also observed in the three occupational classes of lower incomes, while for executives, entrepreneurs, professionals and technicians, on the one hand, and merchants, on the other, incomes tend to increase with age. Taking this into account the age intervals were assigned different values according to the occupational group. Combining this approach with the first one we get a further - even if small-improvement of the results. The sum of contributions to the differences effect supporting the hypothesis increases to 0.2345, the sum of those contradicting it decreases to 0.0210 and the non imputable add up .0061.

A third possibility consists in desaggregating further some of the 40 groups.[24] For instance, for every control variable the corresponding group can be subdivided into as many subgroups as there are possible control levels for that variable. Let us consider an example. In nine out of the forty cases age is a control variable. However in six cases the analysis is not necessary.[25] In the other three cases the consideration of the levels at which the variable is controlled suggests that the importance of education declines relatively to occupation and size as age increases.[26] The results bring out the possibility of introducing qualifications to the hypothesis. For instance, in one of the subcases in 2.2, age as a control variable and one class has higher values in occupation and size, and the other in education. The hypothesis indicates that the class with higher values in two variables will be expected to have higher average income. The qualification would be "except if the family heads, have less than 35 years; in such a case, the class with a higher value in education will have higher income", because of the importance of education for the youngers. In the other extreme, consider the case in which age and size are control variables and education is opposed to occupation.

The hypothesis says the class with a higher value in education will have higher cincome. The qualification here could be "except if family heads are old people, having more than 65 years, in whose case occupation will predominate over education."

5. Association among Variables

The results presented in the preceding section suggest that the joint effect of the variables is quite important. A large proportion of contributions to the differences effect derives from cases without opposite variables and with only one or none control variables. So, it seems to be quite necessary to investigate further such association.

Let us begin by using standard statistical techniques. The values of C (Cramer) and T_c (Kendal)[27/] for pairs of variables show that occupation-age, education-occupation, and size-age have the highest values. On the other hand, education appears to be associated rather weekly with both size and age. The association between size and occupation takes an intermediate place.

Global indexes of association could be misleading when applied to contingency tables larger that two by two, because the association may be positive in some part of the table and negative in another. For this reason we applied the analysis of residuals developed by Haberman. It has the advantage of allowing at the same time local analysis and significance tests[28/].

Table 6 presents the results. Positive adjusted residuals correspond to cases in which there are more households than the number that there would be in case of no association between the variables. Symmetrically, negative residuals indicate that there are less families than in the case of no association. For reasons of space we prefer to omit [29/] a detailed analysis of the table: only as an example, let us take a look at panel F. Being either a blue collar worker or out of the labor force is negatively associated with high levels of education.

In the other extreme, the occupational class with higher incomes has positive association with high levels of education. As expected, white collars are in an intermediate position.[30/]

Table 6

Population contingency tables: adjusted residuals

A: Age-size

	65 and more	12-34 yrs old	50-64 yrs old	35-49 yrs old
1-2 members	16.33	- 0.92	1.89	- 12.52
3-4 members	- 9.41	- 0.28	1.92	5.07
5 and more	- 5.67	1.27	- 4.18	6.87

B Education-size

	None	Some of primary	Primary complete	Some of secondary	Some of University
1-2 members	4.23	1.57	- 2.84	- 1.18	0.31
3-4 members	- 3.42	- 2.65	2.58	2.98	- 1.85
5 and more	- 0.33	1.50	- 0.11	- 2.28	1.84

C: Occupation-size

	Not in the labor force	Blue collar	White collar	Merchents	Executive, entrepreneu etcetera.
1-2 members	13.34	- 5.15	-3.85	- 0.12	- 3.97
3-4 members	- 6.67	1.62	2.38	1.43	1.41
5 and more	- 5.83	3.36	1.15	- 1.54	2.40

D: Education-age

	None	Some of primary	Primary complete	Some of secondary	Some of University
65 and more	6.26	3.77	- 1.37	- 4.82	- 2.43
12-34 years old	- 3.25	- 2.58	- 3.10	6.18	3.69
50-64 years old	0.85	0.07	1.99	- 2.63	- 0.64
35-49 years old	- 2.66	- 0.71	1.54	0.99	0.58

E: Occupation-age

	Not in the labor force	Blue collar	White collar	Merchants	Executives, entrepreneu etcétera.
65 and more	24.54	- 8.36	- 7.12	- 3.07	- 6.16
12-34 years old	- 9.23	6.15	2.36	2.23	1.87
50-64 years old	4.97	- 4.87	- 0.02	2.35	- 1.66
35-49 years old	-14.66	5.57	3.19	1.73	4.41

F: Ocupation-education

	Not in the labor force	Blue collar	White collar	Merchants	Executives, entrepreneu etcétera.
None	3.84	2.48	- 3.01	- 0.87	- 2.99
Some of primary	3.25	6.41	- 2.30	- 1.57	- 7.61
Primary complete	1.04	0.44	2.40	2.67	- 7.22
Some secondary	- 4.98	- 4.54	3.17	0.29	7.25
Some university	- 3.35	- 6.56	- 2.82	- 2.18	17.84

We do not pursue further the standard statistical consideration of association among the variables. Instead, we prefer to explore it in the context of the Gini decomposition. The idea is to compare first the theoretical population values that would have resulted in the case of no association with those observed in the sample, and then these observed values with the results of the Gini decomposition. Table 7 shows the results.

The first column details the relative values that the population weights should show if there was no association among the variables. The weights correspond to the differences effect, that is $P_i.P_j$ for all i and j. The values of this first column were computed as the product of the marginal values of rows and columns divided by the sum of weights. The second column gives the relative values of the weights actually observed in the sample. Finally, the third column shows the relative values of the contributions to the differences effect.

The comparison of the two first columns indicates the association among the variables. In the cases with no opposite variables, the values of the observed relative weights are higher than those expected in the case of no association. It means that when two classes are compared and one of them has a higher value in one variable the probability of finding for the same class larger values in the other variables is higher than that of finding lower ones. The highest discrepancy between expected and observed weights corresponds to case 1.1, where there are not control variables and all the variables have higher values in one class. On the other hand, in the cases where there are opposite variables, the higher values correspond to the expected weights had there been no association.

The distance is shorter in the cases with less opposition and longer in the cases with more opposition.

In order to understand better the meaning of comparing columns 2 and 3, it is convenient to think the values in column 3 as the sum of population shares weighted by income differences. It is then clear that differences in the values of columns 3 and 2 are determined by differences in average incomes: when these are high, column 3 has larger values. Looking at the table we can see than column 3 registers higher values (relative to column 2) in case 1.1; at the other extreme the lower value of column 3 (relative to column 2) corresponds to case 2.4. That is to say, the larger income differences correspond to comparisons in which one the classes has higher values in the four variables. The smaller differences to one of the cases with most opposition(one against one and two control variables 2.4).

Table 7

Association among the variables and the Gini deccmposition

	Population weights Theoretical relative values (%)	Actual relative values (%)	Shares in the differences effect (%)
1. Cases with no opposite variables			
1.1 No control variables	3.12	10.18	18.58
1.2 One control variable	10.52	16.36	21.65
1.3 Two control variables	12.99	14.89	14.19
1.4 Three control variables	6.96	8.24	4.34
	33.59	49.67	58.76
2. Cases with opposite variables			
2.1 Three variables vs. one	12.48	11.87	11.77
2.2 Two variables vs. one	31.56	22.66	18.78
2.3 Two variables vs. two	9.36	5.94	4.70
2.4 One variable vs. one	12.99	9.85	5.99
	66.39	50.32	41.24

The first five cases taken together have expected relative weights adding up to 46% of the total, observed weights of 61.5%, and contributions to the differences effect adding up to 70.5%. If only the first two cases are considered, we find expected weights adding up to 13.6%; observed values, 26.6%; and contributions to the differences effect, 40.2%. These findings strongly support the conclusions that there are positive association among the variables and that the differences in the average income of two classes is greater the less is the opposition among variables.

6. Summary

Inequality in income distribution is measured with the Gini coefficient. The analysis of the inequality is carried out through a decomposition of the coefficient that discriminates an effect of inequality within classes, an effect of differences in average income among classes, and an effect of overlapping among classes. The method allows to distinguish contributions to the inequality that support and contributions that contradict a hypothesis, as well as to link sources of income and size distribution. The association among the variables is examined using both standard statistical techniques and the Gini decomposition as framework of reference.

It was found, in the first place, that there is a significant positive association among the variables considered. This means that the probability of finding a class with a higher value for a variable is greater if the class already have other variables with higher values. Moreover, income differences between two classes are greater as more variables take higher values for the same class. The combination of this two facts explain the relatively large contribution to inequality emerging from income differences between classes with none or one control variable while all the others take higher values in the same class.

The relative importance of the variables in their independent contribution shows education and occupation - in that order - as the most significant. The size of the households is in an intermediate position, and age of the family head is the weaker explanatory variable. The consideration of the joint effect of the variables taken by pairs concludes that the combination education-occupation is by far the most powerful.

Three ways of refining the hypothesis are explored. As a simple way of weighting differences in values of the variables, a transformation to dichotomous variables is presented. The association among variables gives place to a reformulation of the pattern of incomes along the life-cycle, so that instead of a single pattern, two different ones are assumed, depending on the occupational class. Finally, the possibility of further disaggregation is considered: when there is a control variable it may be important to distinguish at which level it is controlled. It is shown that when age is a control variable then the relative importance of education decreases along the life-cycle.

A final word of caution. As any research using a new methodology for a particular case, it is not at all-easy to evaluate the results and findings, because of the lack of a comparative framework of reference. For instance we have emphasized the association among the variables, but if a similar methodology was applied to other Latinoamerican urban centres if would not be impossible that the results showed still larger association. This is what we found in our univariate analysis. Looking only at Beunos Aires, we stressed the importance of education and occupation and the weaker explanatory power of age. But when compared with other Latinoamerican cities, we found exactly the same pattern, still magnified. So that the important conclusion for Buenos Aires is that education and occupation have less importance than in the other cities and age more. To improve the understanding of the interrelationships between economic development and income distribution the results of the present research as least a comparative reference.

References

/1/ N. Bhattachrya and B. Mahalanobis, Regional disparities in household consumption in India, Journal of the American Statistical Association, Vol. 62, 1967.

/2/ H. L. Dieguez and A. Petrecolla, Distribución de ingresos en el Gran Buenos Aires, Instituto Torcuato Di Tella, Buenos Aires, 1979

/3/ H. L. Dieguez and A. Petrecolla, Desigualdad y concentración de depósitos bancarios en la Argentina, Ensayos Económicos, No. 9 (la. parte), 1979

/4/ J. C. H. Fei, G. Ranis, S. W. Y. Kuo, Growth and family distribution of income by factor components, Quarterly Journal of Economics, Vol. 92, 1978

/5/ J.C. H. Fei, G. Ranis, W. W. Y. Kuo, Growth with equity (The Taiwan case), Oxford University Press, 1980.

/6/ R. Ferber, Distribución de ingresos y desigualdad de ingresos en algunas areas urbanas, Ensayos ECIEL, No. 3, Washongton D.C., 1976

/7/ S.J. Haberman, The analysis of residuals in cross-classified tables, Biometrics, 1973

/8/ Instituto Nacional de Estadistica y Consos, Encuesta de presupuestos familiares (unpublished)

/9/ M.G. Kendal and A. Stuart, The advanced theory of statistics, Vol. II, Hafner Publishing Co., New York, 1963

/10/ N. Podder, Distribution of household income in Australia, The Economic Record, 1972

/11/ G. Pyatt, Disaggregation of Gini coefficients, The Economic Journal, 1976

FOOTNOTES

1- Except in a very simple case. See /3/.

2- Programa de Estudio Conjuntos Sobre Integración Económica Latinoamericana.

3- For simplicity of exposition it is assumed that income differences are expressed in units of average income.

4- This idea has been introduced by Fei, Ranis and Kuo, /4/ and /5/

5- For a full development of this decomposition and its relationship with growth theory, see /4/ and /5/

6- This paper limits its attention to total incomes but the survey provides information by five sources of income, as can be seen in Table 2. In Ch. III of /2/ some univariate analysis are carried out, focussing attention on comparisons among inequalities in total income, and incomes from wages and self-employment.

7- In this connection, figures are no reported here. See /2/, Table 3, p. 45.

8. They are assigned to families that own the houses where they live.

9. Underreporting of income is always supposed to be present in household surveys. In our case there are reasons to believe that underreporting was relatively larger in the higher brackets, especially as regards the incomes from capital and self-employment. This of course suggests than inequality income from these two sources, as well as total inequality, are larger than the Gini coefficients indicate.

10- Transfers are mainly payments by the social security system (old age benefits). Even though their contribution to total ineouality is small, it is positive, that is to say, inequality of transfers increase the inequality of total income.

11- Of course the contribution to inequality of income from capital and self-employment would be still higher if the presumption explained in 9 was true.

12- See /2/, Chapter II.

13- This is the detail of variables, classes, and values:

VARIABLE	CLASS	VALUE
Size of household	One-two members	1
	Three-four members	2
	Five or more members	3
Age of family head	12-34 years old	2
	35-49 years old	4
	50-64 years old	3
	65 or more years old	1
Education of family head	None	1
	Some of primary	2
	Primary complete	3
	Some of secondary	4
	Some of university	5
Occupation of family head	Not in the labor force	1
	Blue collar workers	2
	White collar workers	3
	Merchants	4
	Executives, entrepreneurs, professionals and technicians	5

14- This additional class ("invalid answers") represents less than 3% of the families, and it was not taken into account in most of of the analysis.

15- Observe that this fact is hidden when the decomposition is carried on with indexes that compare classes only by considering their average incomes.

16- In what follows, only for simplicity of exposition, we are going to study classes of households, so that the hypothesis are referred to the average income of a class. As it was explained above it is easy to extend the analysis to individual households, because we need only to split the overlapping effect in two halves, one supporting and the other contradicting the hypothesis.

17- As usual, we consider a control variable the one having the same values in the two classes considered.

18- The hypothesis implies the assumption that the rank of the variables is education, occupation, size, and age, such as it was found in the univariate analysis. See /2/, Ch. III.

19- We are going to consider first the disaggregation into 8 groups of terms and then to comment some results - without fully reporting the figures - of a disaggregation into 40 groups. Looking the figures from another perspective we could, for instance, ask which particular classes out of the 300 taken into consideration make the main contributions to the overall inequality. Ordering the classes by the importance of their contributions we found that - to mention only the first four - the first two have incomes well over the average of the population. In the two cases the heads are entrepreneurs, excutives, professionals, and technicians, with ages in the second bracket (between 35 and 49 years). One class is composed by families of large size (5 or more members) whose heads attained higher education. The other has medium size (3 to 5 members) with the head having secondary education. The third and fourth classes are in the other extreme of the distribution, with incomes well below the average. They are small size families (1-2 members) and the head is old (65 years and more) and out of the labor force (passive). In one of the classes the head has some primary education and in the other primary complete.

20- I we had to choose priorities for future research in the line explored by this paper, we would select an statistical research on confidence intervals for the Gini coefficient and the component we have called "differences effect" in order to test hypothesis with previously determined rejection intervals. In what follows we carry out the analysis in a loose way, exploiting the descriptive posibilities of the Gini decomposition both without reaching a level of statistical inference. The help of statisticians on this respect would be warmly welcome.

21- Figures are commented but not reported here. See /2/, and Table 19, pp. 108-110.

22- We will return to this comment in a more precise way.

23- Intermediate and large family size were grouped into one class, leaving small families in the other. The two extreme age groups - the younger and the older - were consolidated in one class and the two intermediate groups in another. Executives, professionals, etcétera, on the one hand, and merchants, on the other, formed one of the consolidated occupational class while the other one was blue and white collars and those not in the labor force. Finally, the variable education took a value for households whose heads received up to complete primary education and another for those having received secondary or higher education.

24- A disaggregation across the board for the 40 groups is not advisable in our case, because the size of the sample does not allow for such fragmentation.

25- For instance one of the cases in 1.2, age is a control variable, and occupation, size, and education take higher values in one class. This a clear case that does not require further elaboration.

26- The following table outlines the results for these three cases:

Age levels	Control: Age Opposite: Education to Occupation and size	Control: Age and size Opposite: Education to Occupation	Control: Age and Occupation Opposite: Education and size
	(% of contributions contradicting the hypothesis)		
12-34 years old	61.6	20.9	6.6
35-49 years old	28.9	28.9	49.6
50-64 years old	18.7	29.1	77.5
65 and more	9.2	61.9	79.4
TOTAL	26.3	30.3	53.7

27- C is based on the chi-square distribution and T_c on rank correlation concepts. Values are not given here. See /2/, Table 15, p. 97.

28- If n_{ij} is the value of a cell in a contingency table, the expected value in the case of no-association is $E_{ij}=\frac{n_{i\cdot}n_{\cdot j}}{n}$ that is, the product of marginal values of row and collumn divided by the sample population. The standarized residuals are then $e_{ij}=\frac{n_{ij}-E_{ij}}{\sqrt{E_{ij}}}$ and the variance can be estimated by $v_{ij}=(1-\frac{n_i}{n})(1-\frac{n_j}{n})$ so that the adjusted residuals can be cumputed by $d_{ij}=\frac{e_{ij}}{\sqrt{v_{ij}}}$.

For a detailed reference see [7].

29- When commenting above the global indexes of association, we said that three pairs of variables had the most significant values: education-occupation, occupation-age, and size-age. The first case has an obvious interpretation but not the other two. Table 6 allows a better understanding of these cases. Panel E shows that the association between occupation and age is mainly due to the classification in the variable occupation, since people not in the labor force constitute a class there. As they are chiefly retired old people there is an strong association with the class "65 years and more". Panel A, on the other hand, shows the fact that old people (65 and more) are predominantely heads of small families and that heads between 35 and 49 belong to medium and large families.

30- The analysis of adjusted residuals for the multivariate classification did not add any substantial insight to the bivariate case here considered. The largest positive residuals appear in the groups of small family size, whose heads were old, not in the labor force and with a level of education not exceeding primary school.