SIZE DISTRIBUTION ANALYSIS PACKAGED PROGRAM AND INCOME DISTRIBUTION DATA BASE

By YOSHIRO MATSUDA,* NORIYUKI NOJIMA,**
AYAKO SUGIYAMA** & YASUHIRO TERASAKI***

This paper develops the schematization of the income distribution statistics as a data base and the compilation of the packaged program to analyze the size distribution which are not yet well treated in other statistical packages. This might be one step to compile a more comprehensive micro data sets and to support the empirical researches on the inequality and welfare problems now in fashion.

I. *Introduction*****

Recent revival of the interest in the cross-section data has stimulated the systematization of the micro-data sets fully articulated with the national income statistics like a new SNA.[1] This shift of interest stems from economists' concern over the components of income especially from the welfare aspects of income distribution. This becomes possible through the large scale survey compiled by electronic computer system and the level-up of the accuracy of the income statistics.

The schematization of income statistics requires the clarification of the ambiguity in the definitions of income of various sources and the merging of several surveys like expenditure survey, tax revenue report, etc.[2] This will support the analysis of the growth and

* Assistant Professor (*Jyokyōju*) of the Institute of Economic Research.
** Assistant (*Jyoshu*) of the Computer Laboratory of the Institute of Economic Research.
*** Graduate student, Graduate School of Economics, Hitotsubashi University.

**** This is the abridged and revised version of our research paper entitled *Size Distribution Analysis Packaged Program and Income Distribution Data Base* Tokyo, 1976. (Working Paper. J-2). This paper is mainly based on the results obtained from the project; The Income and Assets Distribution Research Project organized by Professor Toshiyuki Mizoguchi of Hitotsubashi University, supported by the Grant from Toyota Foundation. The opinions here expressed, however, have no relation to Toyota Foundation. After writing the working paper we reported it at the Joint Conference on the Income Distribution in Korea on the 11th of May in 1976 at Seoul National University and later at the Symposium on Computer and Statistical Analysis of the 44th Annual Meeting of the Japan Statistical Society held on the 20th of July in 1976 at Hokkaidō University. Acknowledgements are due to the discusstants of both meetings but any remaining errors are authors' responsibility. Computer facilities used are the Computer Laboratory of the Institute of Economic Research, Hitotsubashi University and the Hokkaido University Computing Centre.

[1] The concepts of micro data sets and their importance in the recent economic analysis are shown in Kurabayashi [18]. The epistemological problems are discussed in Dunn [9].

[2] Jain compiled the income distribution data of eighty-one countries and computed the decile distribution and several inequality measures. Jain [15]. This work whose coverage is most comprehensive will serve to the general international comparison, but, in its nature, lacks detailed cross-tabulations so that it may be insufficient to proceed with the economic analysis. Since it seems to take less consideration for the data availability and the criteria of data selection is not clear, it may not give us adequate information on the data examination.

income redistribution problem and the size distribution. The most serious obstacles to schematize the micro data sets of income statistics are the problem of the privacy protection and the severe conditions to release and disclose of the survey data. Thus it becomes quite difficult to match several surveys directly and in most cases we should search for the compromising solution through the summary tables released by the statistical agencies.[3]

Our main concern is to compile income and assets distribution data base which is suitable for income distribution analysis. Most of the data bases used among us have been compiled and developed for the estimation procedures of large scale macro models of time-series nature and least attention has been paid to the size distribution aspects.[4] The cross-section type data base, especially based on the summary tables, is not the mere piling down of the survey data in the sequence of time but requires the flexible readjustment of the tabulation formats caused by the changes in situations.

It might be more efficient to equip this income distribution data base with the size distribution analysis packaged program.[5] Recent discussions on the measure of size distribution shed the light on the characteristics of the summary measures which require tedious calculation procedures. The fact that the selection of the summary measures highly depends on the orientation of the research program leads us making the package. For example, we should use the lower income sensitive measure for the discussions of the income redistribution policy. Thus it might be convenient to supply unified tabulation formats to each summary measure for size distribution analysis in the package.

Among the packages favoured by the economists there exists few which pay enough attention to the size distribution problems.[6] Our trial is to fill the gap between the computer specialists and economists by compiling the package to analyze the size distribution problem.

---

[3] Budd and Okner attempted to match or merge the micro data sets of some survey data to enrich the information of the household sector. See [5] and [5–a] by Budd and [33] by Okner. Some analyses based on this enriched data are reported in Smith's paper [39]. But there are some unsolved problems associated with micro data sets, particularly confidentiality and the magnetic tape or disk release conditions which prohibit to get direct access to micro data sets. Actually we are not yet in a position to use such kind of data in Japan or in many other countries.

[4] The philosophy to share common data file cannot be traced back so old as the developments of macro-econometric model building. We do not have adequate literature on this topics. Mori [31] analyzed the impacts of computer on macro model building only through estimation procedures and Matsuda [25] traced the European impacts on Japanese economists. Overall review of the situations of data base in Japan is under preparation. Apart from the problem of compilation of data base, cross-section type analysis itself has a long history, especially the analysis of Paretian distribution of income. It is surveyed in Mizoguchi's survey article [30]. We here only refer to Hayakawa's pioneering paper [13] and his final summing up [14] as the latter is not listed in Mizoguchi [30].

[5] With the propagation of the so to speak third generation electronic computer system the development of statistical packages has become fashion among computer specialists. There are many literatures discussing the merits and demerits of the existing packages but here we restrain ourselves from the overall survey of the present situations. Scucany & Minton [38] surveyed about 55 packages. Mori [31] surveyed the situations in Japan paying much attention on the econometric analysis. More extensive survey will be published by Professor C. Asano in near future.

[6] For example, BMD established in 1963, although designed not for economists but favoured by economists, SPSS established in 1970 and its Japanese extension are not appropriate for this purpose. See Miyake [29]. As to the Japanese packages which are not the exception, we refer only to KEMP and ASTRO FOILE. See Sugiura [40].

II. *Income and Assets Distribution Data Base*

i) Scope and Aim of our Data Base

Our purpose to compile an income and assets distribution data base is closely related to the analysis of the impacts of economic growth on income distribution. Its coverage is limited to some Asian countries. Because it is fairly easy to get access to various data sources containing cross tabulated data which are necessary for examination of the accuracy of the data and further detailed analysis. Thus we will call our data base IADDEC (abbreviation for Income and Assets Distribution of Developing Countries) where we have already stored postwar Japanese data and Korean data and are now accumulationg data of prewar Japan, the Philippines, Taiwan and Hong Kong. (See Appendix.) With this IADDEC data base we will secure full posession of micro and macro income data base for developing countries since we already have the data base of national income statistics of developing countries compiled by Professor Kurabayashi and Ms. Ayako Sugiyama.[7]

ii) Definition and Qualification of Data Base in General

Before explaining our IADDEC data base, it might be useful to sketch the general characteristics of data base, and the kinds of data base.

The term "data base" is used in various contexts with different and obscure connotations. Here we define it in the following way. First, the data base is a kind of data file which is processed by computers of different kinds of hardware and software. Second, this data file can be utilized by different users for their own uses. Third, the contents of the file are easily retrieved by its specified processing program. The first and second qualifications are realized through the dispute of building the data bank which has not presupposed such qualifications. Pioneering runners of the data bank or data archives thought that the data bank could be organized by the similar methods with libraries of documents or monographs even for the data file of each statistical reports. With the rapid propagation of the statistical reports by magnetic tape or similar devices they realized the difficulties of conversions of various data formats into a single unified format which can be used by various users. This situation is accelerated with the propagation of the packaged programs, which now lessen the burden of the researchers. Thus the third qualification is required.[8]

The main components of data base are 1) data file, either numeric or characters, 2) program for the IR of the data file and 3) the maintenance system of the data file, for example, updating or error correcting.

As for the contents of the data file, there are three kinds of numeric data records, such as:

1) individual records of the survey or data by questionnaire base,
2) summary tables of the survey and
3) aggregated or recompiled data.

The packaged programs like SPSS are designed mainly to treat the first category because

[7] See Kurabayashi [18].

[8] There are many serious technical problems to construct the data bank. See, [4] ed. by Bisco. As to the American experience in constructing the statistical data bank, Dunn's survey may be most comprehensive. See Dunn [9].

sociologists or political scientists often carry out sampling survey by themselves. But econometricians mostly utilize the data of the second or third categories, for statistical agencies release the results of their survey in cross-tabulation format to protect the privacy of the reporters. This situation is the same for the income distribution data mentioned in the previous section. Thus our IADDEC is designed to manage this kind of data, i.e. summary tables.

Time-series data base consists of the third category data but the fully articulated micro data sets require the first category data, which will make it possible to adopt more efficient data structure. However, IADDEC data base depends on simple tree logic data structure because of the lack of the first category data. IADDEC data base might be regarded as a compromise of the actual situations.

III. *Description of our Data Base and its File Structure*

The IADDEC data base is based on the recompilation of the summary tables into an unified data base suitable for our packaged program for size distribution analysis. As we have not yet completed the accumulation of the data we want to store, the data are stored only by survey base of each country. Apart from the data storage structure, the data structure of the survey stored is as follows:

Data base consists of three different files, 1) contents of the stored data, 2) definitions of codes and labels and 3) the statistical data with codes. File 1) will act as a linkage of the file 2) and 3). The record of the File 1) consists only of codes and that of the File 2) of codes of one alphabet letter and three numerics and definitions in 150 alphabet letters.[9] Thus the master file compiled by them will give us all the information of the survey in natural language. The record of the File 3) consists of codes of one alphabet letter and three numerics followed by real number of F13. 2 type as an element of data.

To illustrate the data structure of each survey we use the table for National Survey of Family Income and Expenditure of Japan (Zenkoku Shohi Jittai Chosa).

CODE: W001 is nation code of the world registered in IADDEC for Japan. (Numeric order is according to the order of the data input process and has no specific meaning.)

CODE: S002 is survey code for the National Survey of Family Income and Expenditure of Japan.

Variables used in IADDEC are cross-tabulated by categories and attributes. Attributes are both numeric (shown by scale measure) and characteristics (shown by attribute codes). Variables are defined as follows: CODE: V001 to V003 are figures on the object or characteristics of the samples surveyed, such as:

CODE: V001 is for the estimated number of households,
V002 is for the number of households surveyed, and
V003 is for the number of persons per household.

Variables V004 to V010 are for income and assets concepts.

CODE: V004 is for earning income,
V005 is for primary income defined as the sum of earning income and property income,

[9] Documentation to explain the nature of the survey affords much longer expressions.

V006 is for pre-tax income defined as the sum of primary income and transfer,
V007 is for disposable income defined as the rest of pre-tax income after tax is reduced from it,
V008 is for consumption expenditures,
V009 is for assets in money value, and
V010 is for net assets in money value.

If the survey does not supply exactly the same concepts as defined above, we make necessary adjustments combining various concepts obtained in the reports.

CODES: V011 to V020 are left free for the definitions by each survey.

Now each variable is shown in summary table classified by Class Category to calculate the size distribution like CODES: C001 to C008. Originally they are regarded as multiway classifications but here we condensed the other classification scheme as mere attributes. (CODES: X001 to X0. . , CODES: Y001 to Y0. .) The most crucial point is that these classification schemes vary by the date of survey because in the case of quantitative classification scheme scale varies due to growth or decrease of the economy (for example, money income grows with the inflationary process) and in the case of qualitative classification scheme scale varies due to the changes of the economy (for example, some kinds of occupation disappear due to the changes of the society).

Thus the class measure (CODE: M001 etc.) has various scale measure (CODE: N001 to etc.) illustrated as follows:

CODE: M001: Measure by money income class
N001: Numeric scale by Yen in 16 classes such as:
0-, 5000-, 10000-, 15000-, ... , 90000-, 100000-
N002: Numeric scale by Yen in 19 classes such as;
0-, 5000-, 10000-, ... 100000-, 120000-, 140000-, 160000-.

Other attributes like X001, etc., have subclasses like Y001, etc., which may vary by the date of the survey (shown by the CODE; D590, D640, D690. .) Thus X, M are fixed codes and Y, N are variable codes. They are expressed with fixed code D. The IR program is written with this code system and data base will be called into work area with this code system.

IV. *Size Distribution Analysis Packaged Program*

i) Characteristics of Size Distribution and its Analytical Methods.

A lot of empirical economic researches tell us that most of the distribution by size show the common characteristics, i.e., positive skewness, whatever variable is taken. The typical examples are the income distribution among persons and the distribution of private assets. Also the distribution of products sales among firms or market shares of a brand commodity are well-known.

Two approaches to the analysis of size distribution are often used by the researchers to shed the light on the different aspects of the distribution, though closely related to each other.

The first approach is the one to fit the positively skewed distribution of the observations

to the well-known parametric distribution function. The principal aims of this approach are firstly to represent the degree of inequality (concentration) by the parameter of the fitted distribution and secondly to examine whether the observed distribution could be explained by some of the distribution family for a long period, and if so, to investigate the generation mechanisms of the distribution. The Paretian distribution and the lognormal distribution are two major contenders which represent the skewed family of distribution and Pareto coefficient and Gibratian coefficient are taken as the measure of inequality (concentration) respectively. As to the generation mechanism of the distribution, many models proposed so far are based on the theory of stochastic process such as "the law of proportionate effect" or "Markoff chain model" but lack economic content behind them except a few ones. Thus, for the time, the degree of fitting to the distribution and the estimation of the parameters are main analytical tools of this approach which we can choose.[10]

The second approach is to measure the degree of inequality (concentration) of the skewed distribution by certain criteria and, further, to decompose this degree of inequality into factors to find the causes of inequality. The inequality (concentration) measure is important in evaluating the economic or industrial policies and many measures are proposed. Most of them are closely related to the statistics about the degree of dispersion but some are derived in the economic context. They are generally distribution free measures. Those frequently used are Lorenz curve, Gini concentration ratio (or simply Gini coefficient), coefficient of variation, variance of a logarithmically transformed variable, Oshima index, Kuznets index, Theil's entropy measure, Atkinson coefficient, Herfindahl index and so on.

Furthermore, the variation of the share that the upper or lower percentile has is used as an important measure. Take the income distribution for example, Lorenz curve, coefficient of variation and Gini index are frequently used among others. But there seem to be no consensus on the choice of measures. In fact, recent work by Atkinson tells us that most of the measures give the same ranking if the Lorenz curves do not intersect, but that, if they do intersect, the ranking depends on the weight structure which each measure has. Thus, for example, the coefficient of variation gives the same weight to every class, Gini coefficient heavier weight to the modal class and Theil's measure heavier weight to the lower class. Atkinson's coefficient can even vary weight by the parameter. Atkinson's work implies that in actual situations 1) it is necessary to choose the measures appropriate for the analytical purpose and 2) interpretation of the results requires researchers' careful attention.[11]

We can also extend the simple analysis of comparing the inequality (concentration) measures above-mentioned to the more elaborate one. That is, the decomposition of the measures into factors enables us to seek for the causes of inequality (concentration). The adequate measures for this purpose are Theil's measure and the variance of a logarithmically transformed variable. They are decomposed into the proportion due to the factor difference and the proportion due to the differences within each factor.[12]

[10] The law of proportionate effect referred to in the text is discussed in Kalecki [17] and the Markoff chain model in Champernowne [6].

[11] See Atkinson [3].

[12] Theil's measure and "log variance" are two convenient measures for the decomposition analysis of inequality. Recently several researchers have tried the decomposition of other measures but they still cannot obtain full approval. As an example of the controversy, see Mangahas [24].

Our basic philosophy of making the size distribution packaged program (hereafter abbreviated as SDAPP) is to provide these techniques in a simplified format at the same time.

ii) System of the SDAPP

A system of the packaged program consists of the following four parts or subprograms:

1) control of the total system,
2) computational procedures for the statistical analysis,
3) tabulation of the results and the translation of codes into natural language and
4) error messages informing the users of their error found in the control specification.

Needless to say, the core of the statistical packaged program lies in the computation for the statistical analysis but the merits as a package largely depend on the operational easiness. Thus the design of reading the data and writing out the computational results plays essential role in the package. We aim to adopt free field format for instruction to the control program to imitate the natural language like SPSS unlike BMD which uses numeric parameters. But at this stage we do not fully realize this philosophy.

The second important aspect of the package is the system support for the varying hardware and also level-up of the computer system. It might be desirable to write the package using PL / 1 language because this package needs to handle natural language for the code translation, etc., effectively. But due to the restriction of our hardware we write the program by FORTRAN of 7000 level or FORTRAN IV.[13] Using assembler or machine language may save our computation time but after the recent fashion to make system support easily we dare to adopt the above-mentioned process.[14]

iii) Packaged Techniques

Not only computing summary measures or parameters of some distribution, but also the preliminary treatments such as the treatment of missing values, aggregation, the arrangement of classes and graphic plotting of the Lorenz curve are important tools in analyzing the size distribution. We have packaged the techniques which are theoretically adequate and frequently used. Actually, we are not fully satisfied with the present packaged techniques which are rather heuristic and not yet well constructed on the probability theory. But since there are only a few papers which treat the estimation problems of the size distribution, we had to follow the traditional procedures used at present.[15] We are planning to carry out more comprehensive investigation of this problem in the near future.

Now, we will show the packaged techniques below. Some preliminary treatments for the data transformation are shown in section A and analytical methods are shown in section B.

[13] Our host computer is first NEAC 3100 and then FACOM 230–25 and further development will be carried on by higher level computer system.

[14] It might be well acknowledged that the cost of rewriting the package to adapt the change of computer system sometimes exceeds the increase of computation time due to the adoption of FORTRAN or PL / 1.

[15] Some works on this problem are found in [1], [2], [10], [11], [14], [16] and [36]. The literatures on income distribution in general are listed in the comprehensive bibliography by Windmuller & Mehran [44].

First, we refer to the type of available data. Generally, we are supposed to have the histogram type of data of T classes and know the class limits $x_1<x_2<\ldots\ldots<x_{T-1}$ and the relative frequencies $f_1, f_2, \ldots, f_T$ and, if possible, we also know the class means $m_1, m_2 \ldots m_T$. If the class means are not known, they are estimated in the way mentioned below.

A. Preliminary Data Transformations and Recompilations

A.1. Class mean

If we do not know the class means as in most of the tax data, we can estimate them in the following way. For the bottom class, $m_1=x_1/\sqrt{2}$, from the 2nd to T-1th class the geometric mean of each class limit are used and for the top class, by fitting Pareto law, we have $m_T=\alpha \cdot x_{T-1} / (\alpha-1)$, where α is estimated from the data of T-1th and T-2th class.[16]

A.2. *Missing values*

We classifies the observed frequencies in four types, non-zero frequency, zero frequency, missing and concealed. These are indicated in the output list. In the actual calculations, only non-zero frequency is used and the frequency in the latter three cases is put equal to 0.

A.3. *Arrangement of the class*

It is sometimes desirable that all factors have the same class limits and no zero frequency. We can arrange the class to have the same class limits and no zero frequency according to the indication.

A.4. *Aggregation*

If the class limits are the same, we can aggregate several distributions into one.

A.5. *Information on the shape of the distribution*

Since the shape of the distribution and the Lorenz curve are elementary tools, we can get the information on the distribution such as relative frequency, cumulative frequency, income share and cumulative income share and if necessary, the graphic plottings of the distribution and the Lorenz curve by X-Y Plotter are also available.

A.6. *Decile data*

It is impossible to get the decile points only with the data $f_1, f_2, \ldots, f_T$; $x_1, x_2, \ldots, x_{T-1}$, or $m_1, m_2, \ldots, m_T$. Besides these data, we need to specify the shape of the distribution by some type of distribution family or interpolate between the observed cumulative points of the distribution by some method. Our package takes the latter way and makes the interpolation by Gini's law. The decile points thus obtained is used to make the data of the ten classes, i.e. decile data.

A.7. *Quintile data*

Quintile data is obtained simply by averaging the decile data.

B. Analytical Methods

Packaged analytical methods are shown below with the computational formula and the kinds of output list from the computer.

B.1. *Pareto distribution*

[16] We have no theoretical reason for the computation formula of the class mean. Geometric mean was adopted simply because it might be better approximation than the mid-point. Obviously mid-point would be better in the lower class, though. We followed the widely used method for the estimation of the top class mean. But in case the estimated α is less than 2, we tentatively put α equal to 2. (Note that when α is equal to or less than 2, the distribution has no longer the variance.) When the population share of top class is large, possibly because of the coarse classification, the fitting of Pareto law seems to cause a large bias. The estimation of the bottom class was purely based on our intuition. Since we have no alternatives so far, these rather crude method were employed. We are now investigating this problem. Mehran [27] may be helpful.

Denote income by y and its density (or frequency) by f, Pareto's law is

$$f(y)=\alpha m^{\alpha}/y^{\alpha+1} \qquad (y \geq m)$$

where α, m are constants. α is called Pareto coefficient.

Computational formula

From the relation $w_\tau=\int_{x_\tau}^{\infty} f(y)dy=m^{\alpha}\cdot x_\tau^{1-\alpha}$, $\log(w_\tau)=\beta-\log x_\tau\alpha\cdot\log(x_\tau)$ $(\beta=\alpha\cdot\log(m))$. Thus define $\overline{w}_\tau=\sum_{i=\tau}^{T} f_i$ and regress low w_τ linearly on $\log x_\tau$ $(\tau=1, 2, \ldots, T\text{-}1)$, we obtain the estimates of α. As for the constant m, we try several x values. Specifically, for each $x_1, x_2, \ldots, x_{T-3}$, we calculate the Paretian coefficient.

Output list

1) Pareto coefficient and the multiple correlation coefficeint.
2) Gini coefficeint (GI) induced by the relation $GI=1/(2\cdot\alpha\text{-}1)$.
3) Graphical plotting of the Paretian distribution.

B.2. Lognormal distribution

Put $z=\alpha+\beta\log(y)$, where $z\sim N(0, 1)$ and α, β are constants, then y is said to follow the lognormal distribution.

Computational formula

Define $w_\tau=\sum_{i=1}^{\tau} f_i$ $(\tau=1,2, \ldots, T)$. Plot the points $(w_\tau, \log(x_\tau))$, $(\tau=1, 2, \ldots, T)$ on the normal probability paper and do the linear regression analysis on it. The regression coefficient of log x is taken as the estimate of the Gibratian coefficient.[17]

Output list

1) Gibratian coefficient, Gibratian mean and the multiple correlation coefficient.
2) The graphical plottings on normal probability paper.

B.3. Gini Coefficient (GI)

Computation formula

Define $w_\tau=\sum_{i=1}^{\tau} f_i\cdot m_i$ and $\overline{w}_\tau=w_\tau/w_T$, then we get

$$GI=1-\sum_{i=1}^{T} f_i\cdot(\overline{w}_i+\overline{w}_{i-1}),$$

where $\overline{w}_0=0$.

Output list

1) Gini coefficient.
2) Paretian coefficient induced by Gini coefficient.

B.4. Coefficient of variation (CV)

Computation formula

Calculate mean and variance by $Y=\sum_{i=1}^{T} f_i\cdot m_i$, $V=\sum_{i=1}^{T}(m_i-Y)^2\cdot f_i$, then we get

$$CV=\sqrt{V}/Y.$$

Output list

1) Coefficient of variation.
2) Mean and variance.

B.5. Variance of logarithmically transformed variable (LOGV)

Computation formumla

[17] More sophisticated estimation on lognormal distribution would be a quantile method. For the grouped data, however, the most efficient quantile points would not be available and have to be got by interpolation. Thus the graphical method may be rather reliable.

Calculate $LOGV$ by the formula,

$$LOGV=\sum_{i=1}^{T}(\log m_i-w)^2\cdot f_i$$

where $w=\sum_{i=1}^{T} f_i\cdot\log m_i$.

Output list

1) Mean and variance of a logarithmically transformed variable.

B.6. Kuznets index (KI)

Computation formula

Calculate by the formula,

$$KI=\sum_{i=1}^{T}|f_i/w-m_i/w|,$$

where $w=\sum_{i=1}^{T} f_i$ and $w=\sum_{i=1}^{T} f_i\cdot m_i$.

Output list

1) Kuznets index.

B.7. Oshima index (OI)

Computation formula

Using the decile data, denote the ith decile by w_i and $w=\sum_{i=1}^{10} w_i$, we get

$$OI=\sum_{i=1}^{10}|w_i/w-0.1|/1.80.$$

Output list

1) Oshima index.

B.8. Theil's measure (TI)[18]

Computation formula

Calculate by the formula,

$$TI=\sum_{i=1}^{T}\overline{w}_i\cdot\log(\overline{w}_i/f_i),$$

where $w=\sum_{i=1}^{T} f_i\cdot m_i$ and $\overline{w}_i=f_i\cdot m_i/w$.

Output list

1) Theil's measure.

B.9. Atkinson coefficient (AI)

Computational formula

Calculate by the formula,

[18] Theil's measure is to measure the gap between the population share and the income share of each income class by employing the concept of information content. We can compute two measures; one (TI) from the population share to income share and the other (TI^*) from the income to the population share. Our computation formula is the former one. The difference between them lies in their weight structure. TI is averaged by income share and TI^* by population share. In fact, TI essentially corresponds to the Atkinson coefficient with $\epsilon=0$ and TI^* can be transformed to the Atkinson coefficient with $\epsilon=1$ in such way that $AI=1\text{-}\exp(-TI^*)$. In other words, TI^* puts relatively more weight to lower income class. TI is derived straight from the entropy concept, while the interpretation of TI^* from the entropy concept is indirect. The decomposition formula for TI^*, however, corresponds to that of log-variance. For example,

	TI	TI^*
1962	0.26390	0.44990
1963	0.23888	0.53358
1964	0.24333	0.52878
1965	0.21874	0.35726

Source data: Kokumin seikatsu jittai chosa. [Zensetai bunpu]. (Survey of People's Living Conditions. [whole households])

$$AI=1-\{\sum_{i=1}^{T}(m_i/\ Y)^{1-\varepsilon}f_i\}^{1/1-\varepsilon} \qquad \varepsilon\neq 1$$
$$=1-\exp[\sum_{i=1}^{T}f_i\cdot\log(m_i/\ Y)] \qquad \varepsilon=1$$

where $Y=\sum_{i=1}^{T}m_i\cdot f_i$ and is an arbitrary parameter ($\varepsilon>0$).

Output list

1) The value of the parameter ε and the associated Atkinson coefficient.

B. **10.** *Decomposition of the variance of logarithm*

Computation formula

Denote the relative frequency of the jth group and the ith class by f_{ij} and its class mean by m_{ij}, then $\sum_{j=1}^{n}\sum_{i=1}^{T_j}f_{ij}=1$, where $j=1, 2, \ldots, n$, $i=1, 2, 3, \ldots, T_j$. Put $f_{.j}=\sum_{i=1}^{T_j}f_{ij}$, $Y=\sum_{i=1}^{T}f_{ij}\cdot\log m_{ij}$, $Y_{.j}=\sum_{i=1}^{T_j}(f_{ij}/f_{.j})\log m_{ij}$. Then,

$$LOGV=\sum_{j=1}^{n}\sum_{i=1}^{T_j}f_{ij}(\log m_{ij}-Y)^2$$
$$=\sum_{j=1}^{n}f_{.j}(Y_{.j}-Y)^2+\sum_{j=1}^{n}f_{.j}\sum_{i=1}^{T_j}(f_{ij}/f_{.j})(\log m_{ij}-Y_{.j})^2$$
$$=LOGV_b+LOGV_w,$$

where $LOGV_b$ is the between variance and $LOGV_w$ is the weighted average of the within variance, $LOGV_j=\sum_{j=1}^{T_j}(f_{ij}/f_{.j})(\log m_{ij}-Y_{.j})^2$

Output list

1) Total variance, between variance, its percentage, and within variance.
2) Variance of each group and its percentage to the $LOGV_w$.

B. **11.** *Decomposition of Theil's measure*

Computation formula

Define $SS=\sum_{j=1}^{n}\sum_{i=1}^{T_j}f_{ij}\cdot m_{ij}$, $S_j=\sum_{i=1}^{T_j}f_{ij}\cdot m_{ij}$, $f_{.j}=\sum_{i=1}^{T_j}f_{ij}$, then we get the decomposition formula,

$$TI=\sum_{j=1}^{n}\sum_{i=1}^{T_j}(f_{ij}\cdot m_{ij}/\ SS)\log(m_{ij}/\ SS)$$
$$=\sum_{j=1}^{n}(S_j/\ SS)\log(SS/(S_j\cdot f_{.j}))$$
$$+\sum_{j=1}^{n}(S_j/\ SS)\sum_{i=1}^{T_j}(f_{ij}\cdot m_{ij}/\ S_j)\log(f_{.j}m_{ij}/\ S_j)$$
$$=TI_b+TI_w,$$

where $TI_w=\sum_{j=1}^{n}(S_j/\ SS)\,TI_j$, TI_j is the Theil's measure of the jth group. TI_b implies the between inequality (concentration) and TI_w, the weighted average of TI_j, is the within inequality (concentration).

Output list

1) Total inequality, between inequality, its percentage to the total inequality and within inequality.
2) Inequality of each group and its percentage to the within inequality.

V. *Concluding Remarks*

Now we have sketched briefly the essential parts of our system. The total system will be operated with the combination of the IR through our IADDEC data base and the calling the necessary parts of the data base into work area with appropriate instructions to the SDAPP. The IR processes are based on the specified questionnaire system for the sake of speed up of searching time.

Needless to say our system is still on the way to a well refined packaged program. We will point out several points which seem important. 1) Further elaboration is necessary especially in the form of combining the well supported packaged program which has much wider scope than ours. 2) The IR process still requires so much processing time and so this system needs further refinement. As to the IADDEC data base 3) the transformation system should be developed when we have a chance to get survey data by MT base from the statistical agencies. And 4) as to the practical spread of our data base in the future it is necessary to have an agreement between supplier and the user on the following points, such as: i) to prohibit the further reproduction by the user or to declare no responsibility of the supplier as to the second transfer to the data base, and ii) updating system of the data base which requires much input cost.

Bibliography

[1] Aiger, D.J. & A.S. Goldberger, "Estimation of Pareto's Law from Grouped Observations," *Journal of the American Statistical Association*, 65, 1970, pp. 712-723.

[2] Aitchison, J. & J.A.C. Brown, *The Lognormal Distribution.* Cambridge, 1957.

[3] Atkinson, A.B., "On the Measurement of Income Inequality," *Journal of Economic Theory*, 2, 1970, pp. 244-263.

[4] Bisco, R.L. (ed.), *Data Base, Computers, and Social Sciences.* New York, 1970. (Information Science Series)

[5] Budd, E.C. and D.B. Radder, "The Bureau of Economic Analysis and Current Population Survey Size Distributions; Some Comparisons for 1964" in ed. J.D. Smith [38].

[5-a] Budd, E.C., "The Creation of a Microdata File for Estimating the Size Distribution of Income," *Review of Income and Wealth*, Series 17 No. 4, 1971, pp. 317-334.

[6] Champernowne, D.G., *The Distribution of Income Between Persons.* Cambridge, 1973.

[7] Chenery, H. et al. (eds.), *Redistribution with Growth.* Oxford, 1974.

[8] Cramer, J.S., *Empirical Econometrics.* Amsterdam, 1969.

[9] Dunn, E.S. Jr., *Social Information Processing and Statistical Systems; change and reform.* New York, 1974. (Wiley Interscience)

[10] Gastwirth, J.L., "The Estimation of the Lorenz Curve and Gini Index" *Review of Economics and Statistics*, 63, 1972, pp. 306-316.

[11] _____, "The Estimation of a Family of Measures of Economic Inequality," *Journal of Econometrics*, 3, 1975, pp. 61-70.

[12] Hayakawa, M., "The Application of Pareto's Law of Income to Japanese Data," *Econometrica*, 19, 1951.

[13] _____, *Pareto Hosoku ni yoru Shotoku to Zaisan no Bunpu ni Kansuru Kenkyu.* (On the Distibution of Income and Assets by Pareto's Law) (Ph.D. Dissertation), 1960.

[14] Hussain, A., "A Note on the estimation of the quantile of Pareto's Distribution," *International Economic Review*, 12, 1971, pp. 153-156.

[15] Jain, Shail, *Size Distribution of Income: a compilation of data.* Washington D. C., 1975.

[16] Kakwani, N.C. & N. Podder, "On the Estimation of Lorenz Curve from Grouped Observations," *International Economic Review*, 14, 1973, pp. 278-293.

[17] Kalecki, M., "On the Gibrat Distribution," *Econometrica*, 13, 1945, pp. 161-170.

[18] Kurabayashi, Y., "Use of National Accounts as a Basis of Economic Data System," *Hitotsubashi Journal of Economics*, 1973.

[19] _____ & Y. Matsuda, "A System Approach to the Statistics of Retail and Wholesale Trades as Micro-Data Sets." (Mimeo.) 1975.

[20] _____ & _____. "The Data Base for the Analysis of Non-Profit Private Institutions." (Mimeo.) 1976.

[21] _____ in collaboration with A. Sugiyama, *Series of GDP by Expenditure for Developing Countries, 1958-1967.* Tokyo, 1974. (Data List Series, no. 2)

[22] Kuznets, S., "Distribution of Income by Size," *Economic Development and Cultural Change*, 11, Part II, 1963, pp. 1-80.

[23] _____, "Demographic Components in Size Distributions of Income," in *Income Distribution Employment & Economic Development in South East Asia: papers & proceedings of the Seminar by JERC & CAMS*, 1975. (hereafter abbreviate as (JERC-CAMS)

[24] Mangahas, M., "Income Inequality in the Philippines; a decomposition analysis," in JERC-CAMS.

[25] Matsuda, Y., "Nihon-keizai no keiryo-bunseki (Econometric analysis of Japanese economy)" in I. Yamada (ed.) *Keiryo-keizaigaku kogi* (Lecture notes on econometrics), Tokyo, 1972.

[26] _____, *Bibliography; Works on Econometrics*, Otaru, 1973. (KWIC index series for social sciences, No. 1)

[27] Mehran, F., "Dealing with Grouped Income Distribution Data," (World Employment Programme Research of ILO), 1975. (Working Paper)

[28] Meyer. J. & E. Kuh, *Investment Decisions.* Cambridge (Mass.), 1975.

[29] Miyake, S., *Shakai-kagaku no tame no Tokei Program Package* (SPSS). (in Japanese) Tokyo, 1972.

[30] Mizoguchi, T., *A Review of Income and Assets Distribution in Japan.* (Mimeo.) 1976. (J-1).

[31] Mori, T., "Keiryo-keizaigaku to Computer," (econometrics & computer), (in Japanese) *Jyoho-shori*, 15, 1974.

[32] Okamoto, M., "The Role of Computer for Statistical Analysis," (in Japanese) Résumé for the report presented at the 44th Annual Meeting of the Japan Statistical Society.

[33] Okner, B.A., "Constructing a New Data Base from Existing Microdata Sets: the

1966 Merge File," *Annales of Economic and Social Measurement*, 1-3, 1972, pp. 325-342.
[34] Oshima, H.T., "Income Inequality and Economic Growth; the postwar experience of Asian countries," *Malayan Economic Review*, 15, 1970.
[35] Prais, S.J. & H.S. Houthaker, *The Analysis of Family Budgets.* Cambridge, 1955.
[36] Quandt, R., "Old and New Methods of Estimation and the Pareto Distribution, *Metrika*, 10, 1966, pp. 55-82.
[37] Schultz, T.P., "Secular Trends and Cyclical Behaviour of Income Distribution in the United States: 1944-1965," in *Six Papers on the Size Distribution of Wealth and Income*, ed. by L. Soltow, New York, 1969.
[38] Scucany, W.R., Paul D. Minton and B. Stanley Shannon, Jr., "A Survey of Statistical Packages, "*Computing Surveys*, 4-2, 1972.
[39] Smith, J.D. (ed.), *The Personal Distribution of Income and Wealth.* NBER, New York, 1975.
[40] Sugiura, I., *ASTRO FOIL.* (in Japanese) Tokyo, 1972.
[41] Takahashi, C., *Dynamic Changes of Income and its Distribution of Japan.* Tokyo, 1959.
[42] Theil, H., *Statistical Decomposition Analysis.* Amsterdam, 1972.
[43] Ura, S., *Data kozo* (Data structure). (in Japanese) Tokyo 1974.
[44] Windmuller, T.S. & F. Mehran, *Income Distribution and Employment Programme Bibliography on Income Distribution.* Geneva, 1975.

Appendix: Data Stored in IADDEC

1) Japanese data

Japan, Bureau of Statistics, Office of the Prime Minister

[1] *Zenkoku shohi jittai chosa.* (National Survey of Family Income and Expenditure, 1959, 1964, 1969.)
[2] *Kakei chosa.* (Family Income and Expenditure Survey.) 1953-1973. (yearly)
[3] *Kojin kigyo keizai chosa.* (Unincorporated Enterprise Survey.) 1955-1968, 1970-1971. (yearly).
[4] *Chochiku doko chosa.* (Family Saving Survey.) 1959-67, 1969-1970. (yearly)
[5] *Shugyo kozo kihon chosa.* (Employment Status Survey.) 1956, 1959, 1962, 1965, 1968, 1971.

Economic Planning Agency

[6] *Shohi doko yosoku chosa.* (Consumer Behaviour Survey.) 1960-1964, 1966-1972. (yearly)

Ministry of Agriculture and Forestry

[7] *Noka keizai chosa.* (Farm Household Economy Survey.) 1950, 1952, 1957, 1962, 1967.
[8] *Noka seikeihi chosa / Noka sozei-koka shofutan chosa.* (Cost of Living Survey of Farm Household.) 1950, 1051, 1953-1955, 1957-1968. (yearly)

Ministry of Labour

[9] *Chingin kozo kihon chosa.* (Basic Survey of Wage Structure.) 1955-1971.

(yearly)

2) Korean data

Korea Economic Planning Board

[1] *Toshi kage chosa.* (Annual Report on the Family Income and Expenditure Survey.) 1963-1971. (yearly)

Ministry of Agriculture and forestry

[2] *Nong-ga kyong-ji chosa.* (Farm Household Economic Survey.) 1963-1972. (yearly)

The Institute of Social Science, Chung Ang University

[3] *Income Distribution and Consumption in Korea.* 1966.

Bank of Korea

[4] *Report on Wage Survey.* 1967.

The Institute of Industrial Development in Korea

[5] *Report on Wage Survey.* 1970.