INTERPRETING SOCIO-ECONOMIC DATA[1]

Othmar W. Winkler
Professor emeritus
Georgetown University - Washington, DC, 20057 – USA
winklero@georgetown.edu

1. Introduction. At the beginning of any statistical investigation is the need to know about a social or economic phenomenon. It needs to be defined, what it is, where and when it can be found, and how it should be captured statistically. Obviously the phenomena in society are quite different from phenomena in the natural sciences. They also differ in the manner in which „real-life-objects" project socio-economic phenomena; they accomplish this like iron filings projecting the shape of magnetic forces of a magnet. Socio-economic phenomena also differ from phenomena in the sciences by their great variability, their rapid change, their unpredictable evolution. Nothing can be expected to remain the same from one social stratum to another, from one country or culture to another, from one month to the next. Statistical data[2] must keep up with this dynamism, but statistical theorist have not been comfortable approaching this topic, as if discussing the phenomena in society did not belong to statistics. But it is important to realize that socio-economic phenomena are complex, more elusive than is commonly realized and the manner of capturing them is quite different from phenomena in the sciences.

The term "measurement" implies scientific, precise and objective determination of the value of some measurable characteristic. In socio-economic statistical data "measurement" evokes images of professionals in white lab coats taking precise readings of social and economic facts with electronic measuring devices. Users of published data, even statisticians holding this view, seem unaware of how unrealistic this image is. The data about the economy and society are not determined by objective trained outside observers but are reported to a survey taker by each informant according to his/her subjective judgment, quality of memory, intelligence and readiness to cooperate. Few quantities in socio-economic statistics are objectively determined -- measured -- by a qualified external observer, such as the assessment of the condition of a building by a professional appraiser. In most instances, however, statistics must rely on what the respondents report in questionnaires, face-to-face or in telephone interviews or on the internet. Responses are checked only by small samples for the reliability of the reported information. The veracity and accuracy of the responses, even the willingness to cooperate, varies considerably for different topics and from one informant to the next. Information about a business firm is rendered by an insider, not by an objective outside observer.

Most data are obtained by approaching the "real-life- objects" that portray those social and economic phenomena in society while others are byproducts of administrative activities. Statistics registers these facts „out there in society" documenting them in paper or electronic form. It is these records that become the „statistical counting units" (scu"s) in the published data, not the „real-life-objects" themselves. The scu"s play the role in statistics comparable to that which atoms play in physics.These scu"s are not yet the data themselves but are gathered into aggregates. These then are the data in published tabulations.

These aggregates, the data, describe what happened, when and where, and can be thought of as tri-dimensional structures, like cardboard boxes. Most of the published statistical figures can be imagined as such boxes, or ratios between them.

---

1 This presentation is based on the ideas expressed in my book "Interpreting Economic and Social Data – A Foundation of Descriptive Statistics", Springer, Heidelberg 2009

2 Here, and in the referred book "Data" is treated as a plural word when referring to the numbers of statistical information- ("data are") . The term "data" in the singular (data is) refers to general, non-numeric information.

2. The Frame of an Aggregate. Data are tabulated by only two features: their subject matter and time, or their subject matter and geographic distribution or the geographic distribution over time. Every statistical figure has these three components that determine the „dimensions" of an aggregate because socio-economic phenomena deal with a subject matter that happen in time and in a geographic region. These three dimensions are present in all statistical data, a fact easily overlooked, because data published in statistical tabulations usually present only two of these three dimensions, either the subject-matter and time, or the subject matter and geographic extension, and occasionally, time and geographic distribution for a single subject matter category.

As the development over time is of greatest interest, "time"- semesters, quarters, months, or weeks.- should be marked on the horizontal dimension facing the observer. On the vertical dimension is the subject matter as a one-dimensional listing, and on the other, horizontal, geographic dimension, the administrative regions listed as a one-dimensional sequence, Figure 1.

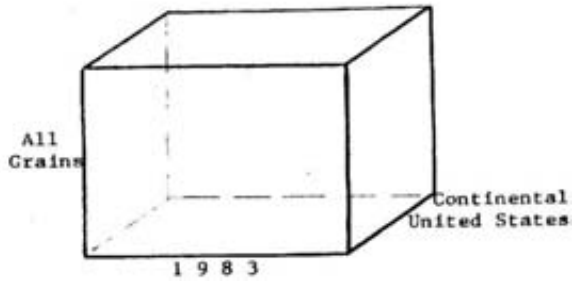

Figure 1 The Conceptual Frame

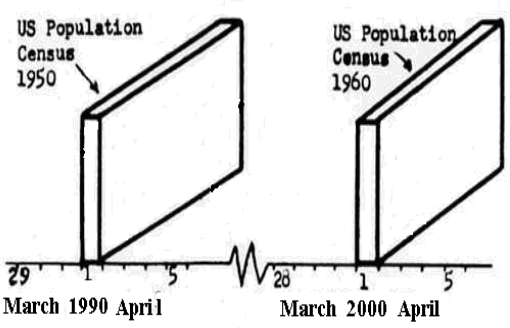

Figure 2 Placement of Aggregates along Time

When the time dimension is small, like in a census or inventory, the tri-dimensional character of a statistical aggregate shrinks to a seemingly two-dimensional sheet, capturing a specific socio-economic phenomenon in that particular point in time. The placement of the resulting aggregates on the time vector, Figure 2, is important, because it connects them with other statistical and non statistical materials.

The aggregate space can be subdivided in a variety of different ways. Sub-aggregates can be formed

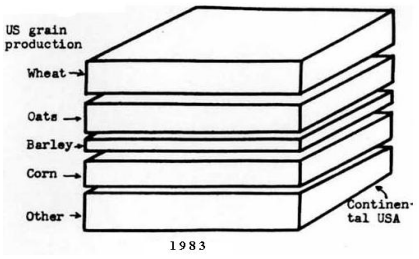

Figure 3 Forming Sub-Aggregates by Subject Matter

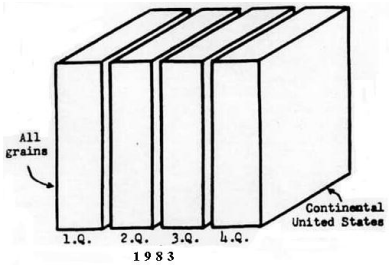

Figure 4 Forming Sub-aggregates by Quarterly Subdivision

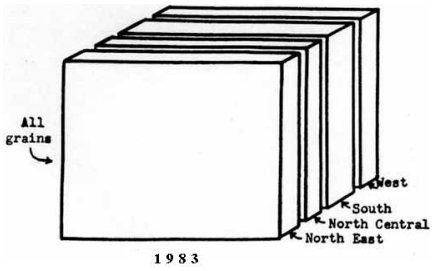

Figure 5 Forming Sub-aggregates by Geographic Region

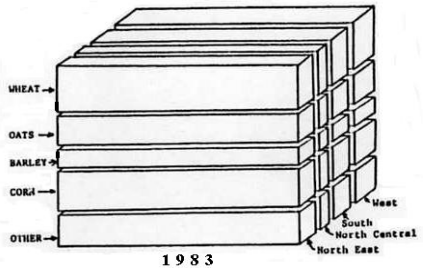

Figure 6 Disaggregation by Region and Subject-matter

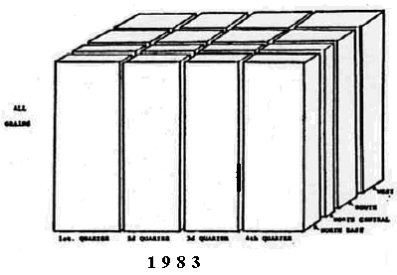

Figure 7 Disaggregation by Region and Time

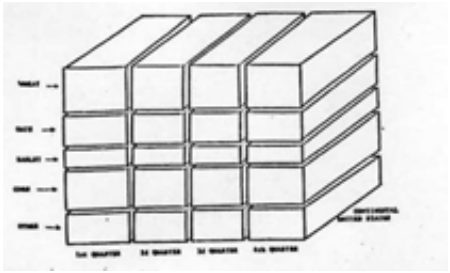

Figure 8 Disaggregation by Time and Subject-matter

simultaneously in any **two** of the vectors, Figures 6,7,8 or simultaneously in all **three** of these vectors, Figure 9. Consider, for example, the conceptual box for the statistical aggregate "Total U.S. Grain Production, 2005." A breakdown along the vertical subject-matter vector, e.g. by economic categories, creates thinner horizontal spaces as separate -- still empty -- sub-aggregates. Each extends over the entire USA, for the entire year. Each of these separate horizontal boxes indicates the different species of grain that are harvested, whose separate knowledge is of interest and of which sufficient quantities can be expected to be produced during the entire year in the entire territory of the United States, such as wheat, barley, oats, corn, sorghum, etc., Figure 3. The original space of the aggregate “all grains” is taken apart into sub-aggregates, each of which becomes meaningful, independent statistical information. If one were to study the timing when grain was produced, one would form sub-aggregates along the time vector according to quarters or months, breaking up he entire space of the original aggregate “Total Grain Production 2005,” into vertical slices, Figure 4. The geographic vector can also be subdivided along that dimension, Figure 5, to study that grain production for the regions of the country. Sub-aggregate spaces can also be formed along any two dimensions. Figure 6 shows a breakdown by subject matter and geographic subdivisions for the entire year. Figure 7 shows the subdivision according to geographic regions and shorter time intervals without detail of the grain harvest. Figure 8 deals with the „what; and „when“ but not „where“ grain was produced. Sub-aggregates can also be formed **along all three dimensions** at the same time, Figure 9, like when presenting the production of the different grain species by region and by shorter time intervals. Note that qualitative, non- measurable characteristics or attributes of the scu“s are the important subset-forming criteria. The quantitative or measurable characteristics – often misleadingly referred to as random variables -- play a similar, but minor role in forming sub-sets but in contrast to statistics in the sciences where they play a major role**.**

3. The Nature of Socio-Economic Statistical Aggregates.

The empty conceptual boxes represent the defined frame of a statistical aggregate. When scu“s -- the simplified records of people, things or events as the elements of statistical data – are placed into this three-dimensional conceptual space of an aggregate it becomes a statistical figure ready to be tabulated. Through comparison with other, similarly defined aggregates the features of a socio-economic phenomenon are revealed and can be interpreted. That overview, however, is achieved by sacrificing detail of information, a fact about which statistical theory and textbooks have nothing to say. The internal structure, the distribution of the scu“s inside of an aggregate is of no interest and is treated as if it were unknown. It is often assumed to be equally spread out inside the aggregate. That internal structure can only be revealed by decomposing the aggregate. This is the key to understanding and properly interpreting data in statistical tabulations. One can also form a larger statistical aggregate of which the given aggregate becomes a sub-aggregate. Such manipulation is limited by their usefulness and practicality.

Depending on the "size" of an aggregate one focuses on different aspects of a given socio-economic phenomenon. This statistical perception differs by its fuzziness from the manner in which phenomena in the sciences are perceived. Features which are not explicitly stated in the definition of the aggregate are treated as not existing. The larger a statistical aggregate, the broader and more inclusive the definition of subject matter, region and time interval, the fewer of the features of the socio-economic phenomenon remain

recognizable. That allows a better overview of a socio-economic phenomenon but leaves fewer details for the interpretation.

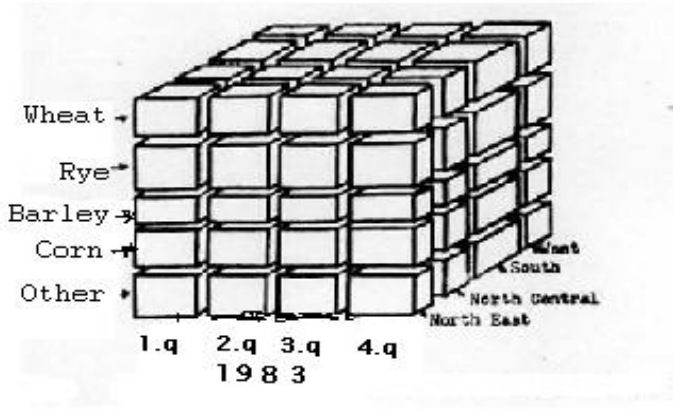

Figure 9 Disaggregation by Subject-Matter, Time and Region

4. The Interpretation of Aggregates. The larger a statistical aggregate, viz the broader and more inclusive its definition is with regard to any or all of its dimensions, the subject matter, time interval, or region, the fewer of the features of the individual scu‟s included in that aggregate, are available to reveal features of the economic or social phenomenon. An aggregate, a statistical figure alone cannot be interpreted. It must be linked to the time periods or geographic regions or subject matter categories of other, related aggregates, to reveal the big picture of a socio-economic phenomenon, providing information that is not available otherwise. The scu‟s, are the building elements of the data in our field, like the atoms in the physical sciences. But a large number of scu‟s in an aggregate neither implies a greater validity of a statement, nor establishes the socio-economic phenomenon with greater certainty. The Law of Large Numbers -- or Central Limit Theorem – does not apply to aggregates and the number of scu‟s in them.

The social or economic situation is revealed through the frequencies with which scu‟s with certain characteristics occur in specific locations within this three-dimensional hierarchical structure of a statistical aggregate and its sub-aggregates. In the sciences, in contrast, each individual measurement represents an observation that gives account of the physical or chemical phenomenon. Measurement errors tend to cancel each other out, the more of such individual measurements are obtained. The scientific phenomenon then emerges with less distortion, hence, the importance of the Gaussian distribution, also tellingly referred to as the error distribution or normal curve.

Keep in mind that in the socio-economic field neither the scu‟s nor the aggregates are the counterparts of measurements in the sciences. Rather, the scu‟s portray socio-economic phenomena such as unemployment, production, foreign trade, inflation, or sex discrimination, like the small mosaic stones that compose a mural. Each individual stone represents something quite different than the picture which these stones compose collectively. Not one of these mosaic stones, individually, gives the slightest idea of that picture that they crater together. Socio-economic phenomena are like such murals: the scu‟s function like the small mosaic stones, and their aggregates represent clusters of similar mosaic stones in that mural. Consider the socio-economic phenomenon "unemployment," that can be perceived only when all the persons in search of a job but lacking one at the time of a survey, are viewed together. The intensity and structure of that socio-economic phenomenon is given by the clustering of the unemployed in certain geographic locations, industries, occupations, gender, racial and age groups. The economically and politically important socio-economic phenomenon, "unemployment" represents something that is different from each one of those persons without a job. Knowledge of the subject area is more valuable to interpret statistical aggregates than

proficiency in probability calculus.

When dealing with the analytical value of a large aggregate, consider the movements over time of the value of a market basket of goods. The recorded shelf price/unit in a sample of stores, multiplied with the quantities of each item in a base period, is currently used for the measurement of price movements. Implied is the erroneous belief that the products in that basket also remain separate entities in a statistical aggregate, such as a quart of milk, packaged carrots, butter, eggs, bread, etc. But statistical aggregation homogenizes these products like a food blender, converting them into a uniform pulp in which these items have disappeared. It follows that the characteristics of the resulting aggregate are different from those of the original ingredients. The more varied the characteristics of the items included in that aggregate, the less clearly defined will be the resulting pulp -- the statistical aggregate. The view of an economy e.g. through the highly aggregated data of the GDP deals with different aspects of the economic situation, than the view conveyed by less aggregated statistical data, an important issue that has to be kept in mind when interpreting aggregate data in statistical tabulations. The customary presentation, such as in Figure 10, stresses the totals as if they were closer to reality. But this is misleading because these are aggregates of different magnitude, that ought to be interpreted correspondingly as indicated in the more appropriate presentation in the tabulations of Figure 11A or 11B.

Total Expenditures…....719,031,000,000

Education….....170,685,000,000
Public Welfare……..... ..265,105,000,000
Health/Hospitals……...68,374,000,000
Highways…….72,455,000,000
Police Protection……... 9,860,000,000
Correctional Facilities……. …36,938,000,000
Natural Resources…….. ….17,110,000,000
Parks/Recreation……4,636,000,000
Governmental Administration… .42,846,000,000
Interest on Government Debt…...31,295,000,000

Figure 10 Customary Way of Presenting Data

Total Expenditures...719,031,000,000	
Education	**170,685,000,000**
Public Welfare	**265,105,000,000**
Health/Hospitals	**68,374,000,000**
Highways	**72,455,000,000**
Police Protection	**9,860,000,000**
Correctional Facilities	**36,938,000,000**
Natural Resources	**17,110,000,000**
Parks/Recreation	**4,636,000,000**
Governmental Administration	**42,846,000,000**
Interest on Government Debt	**31,295,000,000**

Figure 11A Data Appropriately Presented

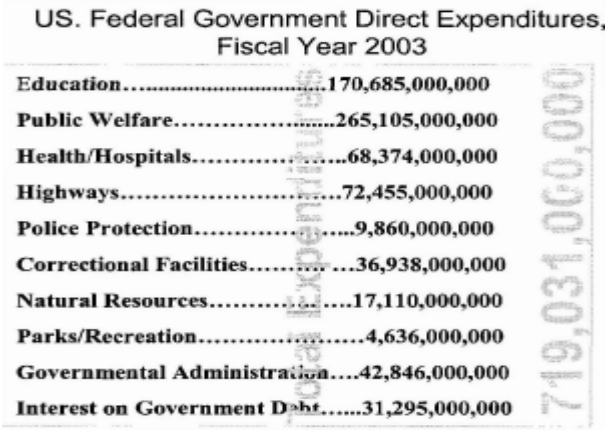

Figure 11B The Same Data Even More Appropriately Presented

There are other kinds of statistical materials in business, economics and the social sciences that are neither aggregates nor ratios, but are based on aggregates such as the difference between an aggregate of beginnings, and an aggregate of endings, representing a class of flow data such as the amount of currency in circulation. Their interpretation poses no new problems

5. Interpreting the Ogive. Figure 12 shows the frequency distribution of the family income of white and black Americans in 2007. The ogives of these data, Figure 13, had to be converted to % for the sake of comparability. The numbers of Asian families was too small to be included in Figure 12. The ogive of white families is presented by the solid line, the ogive of the black families by the dotted line on top and that of the Asian families by the lowest, dotted line. These curves reveal the institutional and cultural shape of the economic and social environment which the three groups encounter, while the frequency distributions do not show it. The slopes of these ogives show the different degrees of resistance against the advancement of families of different income levels and racial groups. An ogive can be imagined as a shore profile against which individual ocean waves break – in this example, the family incomes - of varying economic strength. Most will not reach very high on that shore profile arriving with average strength. Very few family incomes of economically powerful families, like arriving ocean waves that have the strength to push high up on that shore. Once reaching the high end of that shore they encounter less resistance to their further advancement. Such a „wave" will find it easy to roll far into the hinterland. A family with such a big income can further increase it with far less difficulties than a family struggling with a much smaller income, facing far greater difficulties to

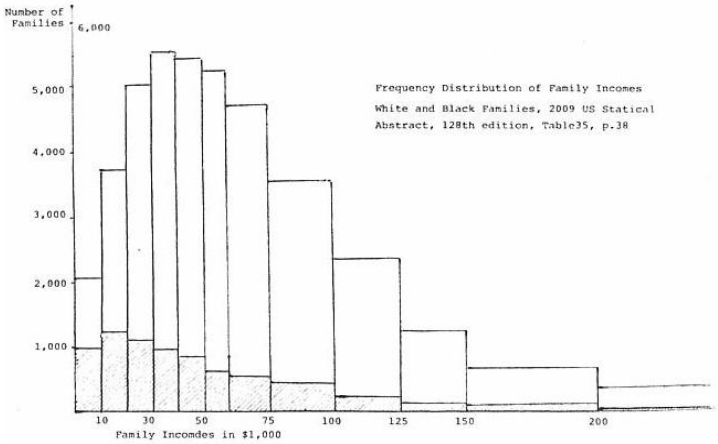

Figure 12 Frequency Distribution of Family Incomes –USA 2007

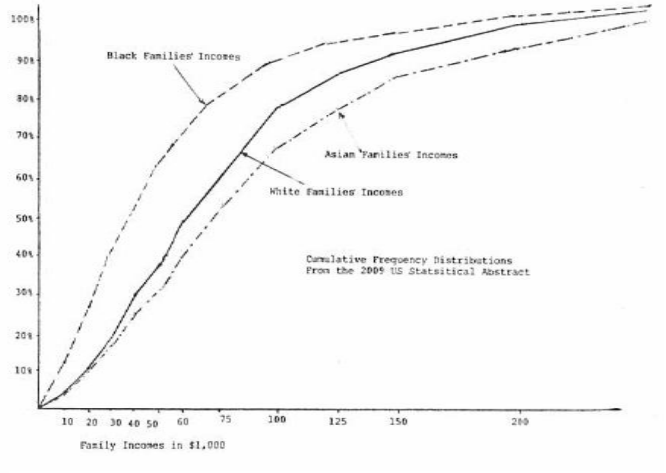

Figure 13 Cumulative Frequency Distribution of Family Incomes –USA

increase its family income. Most frequency distributions of socio-economic data are skewed to the right. Hardly any socio-economic frequency distributions are left-skewed, with their highest frequencies on the upper end of its horizontal scale. Their ogive would resemble a flat beach that rises steeply toward the end, making it easy for most oncoming waves to roll over that flat portion of the beach, further along the horizontal scale. In such a situation most of the imaginary ocean waves would have no difficulty rolling over that initial, flat part of the shore profile, breaking further back, where that shore is becoming steep.

The initial, very steep shore profile of the ogive of the asset size of business firms, Figure 14, -- the amplification of the beginning of that scale of the asset values.-- indicates how difficult it is in the US economy for a small business enterprise to acquire the necessary assets to increase its business. Although this is a static cross section of the situation in the US economy at that point in time, one can

see how difficult it appeared to have been for any of the existing small firms to advance from one asset level to higher level. In these lower ranges even a big effort would allow them to advance only very little - on the horizontal scale - in that steep part of the cumulative distribution.

Only few business firms, perhaps under a particularly gifted and driven leader, can move a great deal up against this "asset profile", overcoming the obstacles posed by the institutional and cultural environment of the US capitalistic society. It would be like that occasional powerful wave that seems to flood with apparent ease over the initially steep part of that coast. Once over that steep portion, the flattening-out shore profile places a decreasing resistance to the oncoming wave that has arrived upon that portion of that ogive. Few business firms have the strength to reach the higher values on the horizontal scale representing asset size. Every frequency distribution of socio-economic data is the result of such a process of selection that is shaped by the institutional and economic constraints of that society that can resist but also encourage the advancement of individual persons and business firms in a society. Such highly asymmetric distributions do not only occur in the social sciences, but also in the physical sciences. The ogive of oil-spill, Figure 15, shows that most reported oil spills are of minor size.

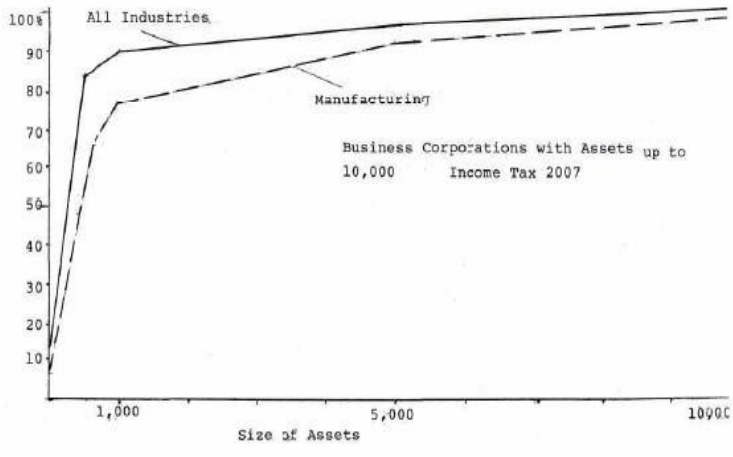

Figure 14. Manufacturing and all Corporations with Assets up to $10,000 - Income Tax 2007

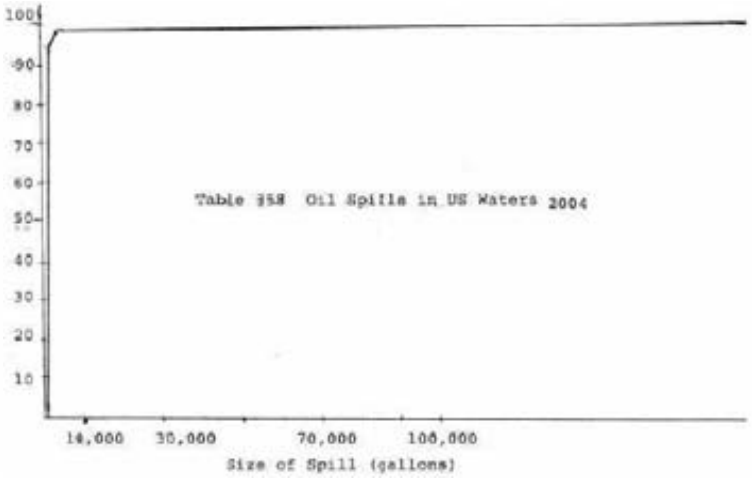

Figure 15. Recorded Oil Spills In US Waters - 2004

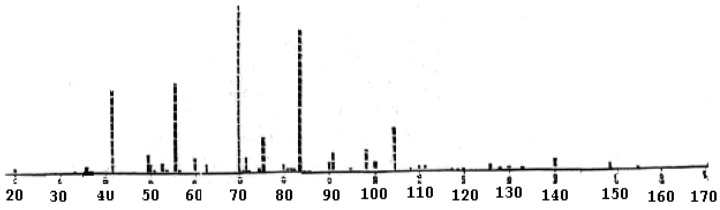

Figure 16. Wages of 462 Textile workers Caracas, Venezuela 1954

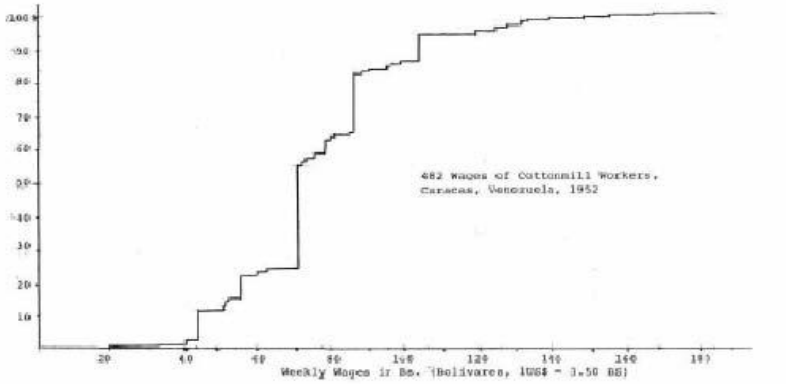

Figure 17. Cumulative Frequency Textile Worker- Caracas, Venezuela 1954

The rare large oil spill that causes the occasional huge environmental disaster is like a powerful ocean wave that can wash easily over the steep first part of that shore profile where they encounter little resistance.. Ogives can also be constructed from ungrouped data, Figure 16. Their ogive shows a serrated profile, Figure 17 in which each case forms a small „riser". The interpretation is analogous to that of grouped frequency distributions.

6. The Case that Provoked the Re-Interpretation of Regression.

The following personal experience represented a milestone in my thinking about regression and about socio-economic data in general. The data were from the sex-discrimination lawsuit in which I was a statistical expert-witness for the plaintiff. The group of female employees claimed to be paid less than male employees at entry level, to have received smaller pay raises and fewer promotions, yet to have the same qualifications and identical work requirements as their male counterparts. The plaintiff's lawyer approached me to statistically establish the facts of their claim. The Court ordered the unrestricted access to the employment records of the professional men and women in ERDA (Energy Resources Development Agency) a large agency of the Federal Government. For the discovery of such alleged differential treatment of equally qualified female employees, the group of 32 librarians, 18 male and 18 female professionals, seemed ideally suited to explore the alleged discriminatory treatment of female employees. All of these professional men and women held comparable academic degrees and were charged with the same duties. I started with a simple linear regression between salary and length of service for men and an analogous, separate study for women. This first approach provided unexpected, puzzling information. Given the salary structure of the federal government, I expected a positive association between length of service and salary, regardless of gender. The regression equation, computed separately for the male librarians[3] was SAL(M) = $16,900 + $1,380*ERDAEMPL Obviously it meant: „The average beginning (at entering) salary for men was $16,900 with a yearly raise of approximately $1,380 for each additional year employed as a librarian at ERDA" which seemed to be a sensible interpretation, Figure 18. I expected to discover a similar situation among the equally qualified female librarians, but probably on a lower salary level. I was unprepared, however, for what I

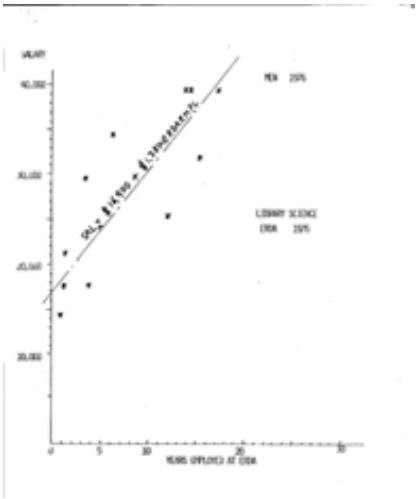

Figure 18

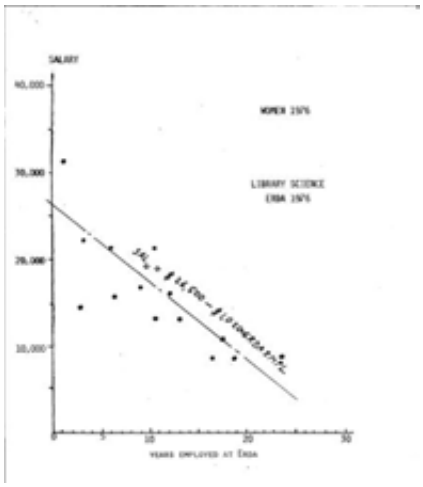

Figure 19

3. Of the18 records of the male librarians only 12 could be used in this regression analysis. In six records one or more of its variables were missing. Of the 18 women librarians only 14 records were used omitting the four records with variables missing

discovered. The linear regression equation for women was SAL(W) = $26,500 - 1,020*ERDAEMPL , Figure 19. An analogous interpretation of the regression line of the female librarians in that government agency would state that these female librarians were paid an average entry salary of $26,500 making those with the least time on the job the highest paid employees in that department. The slope of the linear regression line for female librarians, β_1 = -$ 1,020. indicated that the longer these female librarians had served at ERDA, the less they got paid, reducing their beginning salary by $1,020 for each additional year at ERDA.-- the Greek symbol β to indicate that these 14 women librarians were treated as a „population," not as a sample. The salary of these women apparently decreased, on average, by $ 1,020 per year. According to these data, they apparently got paid less the longer they worked. That obviously made neither sense nor appeared to be a credible description of the situation. It defied what is known about the US Federal Government"s pay scale, and failed to conform to the general experience of the business world. What could have caused this unexpected, puzzling statistical result? The plotted data for the male librarians revealed a sensible picture: lower salary for those who had been a shorter time with ERDA, and correspondingly higher salaries for the male librarians who had served longer on the job. The slope of the linear regression line was β_M = +$1,380 per year of service. The data for women librarians, however, did not make sense. Who would continue to work for an employer who seemed to penalize seniority on the job? This obviously had to be some error or data mix-up. I contacted the person in the government department responsible for these figures and got the promise of a careful review made credible by the court-ordered nature of this investigation. The next day I was assured that no errors were involved and that the data represented the situation correctly. This excluded the possible explanation of an error in the data and made the situation even more puzzling**.** Then it dawned on me that the issue was not an error in the data or in the calculations. The error was in the interpretation! A change in the interpretation, then, resolved this conundrum and began to make sense of the situation. It was also an έυρεκα moment of enlightenment, understanding that in all regression analyses, not only this one, the data are treated as cross-sections, as if the scu"s existed simultaneously, side-by-side. β_1 is not a dynamic factor of "change -- growth or decline -- in Y for a one-unit increase in X". Instead, **the slope** is **the difference** between simultaneously existing scu"s, the data points, assumed to be contemporaries at the point in time when these data were recorded. It was only a small further step to realize, that every scu is really situated at the intersection of the present with the past, usable for both, a cross sectional, static view of the situation, as well as a longitudinal, developmental perspective of that same situation..

A close look at individual librarians" employment histories revealed that the female librarians who entered this government agency after the promulgation of the law "Title VII" -- one of the laws against sex-discrimination was promulgated in 1969 and expanded in 1974 -- were hired at starting salaries that were substantially higher than the starting salaries of the newly hired male librarians and of course, also much higher than the salaries at which women librarians had been hired years before the promulgation of these anti-sex-discrimination laws. At the time of that report in 1976 these old-timers were still with that agency. This imbalance was the result of management"s reaction to these anti-sex discrimination laws and the threat of law suits. Such a sex discrimination class-action law suit was actually filed by women from all departments against that government agency, not just the librarians. The situation reflected management"s inadequate and inappropriate reaction to that legislation that was intended to correct discriminatory treatment of women. Apparently management attempted to upgrade only the **average salaries** of these female librarians but made no attempt to adjust the salaries of its older, aggrieved employees. Management obviously assumed that only salary averages for each department would be checked for compliance with that law.

Scatter-diagrams and regression equations always represent a cross-sectional view of a situation, not

its longitudinal aspects, even if they deal with characteristics involving time, such as the „employee length of service". Although the dynamism of the salary situation through this cross-sectional view of the situation is indirectly revealed through regression analysis, the coefficient β_1 **is not to be interpreted as** the amount by which the salary of an employee **increased** for each additional year of service. Instead the slope β_1 is to be interpreted as the amount by which the salaries of any two female librarians employed by that agency, existing side by side in 1976, differed overall by negative $1,020 for a positive difference of one year of affiliation with that agency. The farther back a woman had been hired at ERDA as a librarian, the lower was their salary at the time of this report, on average by $1,020 per year, a static cross section view of the income situation in that agency in 1976.

I hope that these few points of my presentation will open a new understanding when using published data about the economy and society.