# Interpreting the Cumulative Frequency Distribution of Socio-Economic Data

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

## Abstract

The accumulated frequencies of the quantitative variables of socio-economic statistical data, the ogive, are treated , if at all, as a curiosity and only as an instruction of how to rearranging the frequencies of the distribution. The meaning of the result, an ogive, is not explained nor is its interpretation attempted. It is the purpose of this paper to make sense of cumulative frequency distributions, revealing features of the concrete local-historical situations of society, described by the data, that otherwise would remain unnoticed.

Key Words: Cumulative Frequency Distribution, Ogive

## 1. Introducing the Ogive

"Too often scholars teach theory and methods that are relevant to other academics but not to the majority of the students sitting in the classroom before them – The gap between the policy world and the ivory tower is growing wider" †

The ogive, like the original frequency data, are a cross section of simultaneously existing 'statistical counting units' or elements that yield a stationary picture, as if frozen in time, providing a still-picture or snapshot of a dynamic, evolving local-historical situation. This fact must be kept in mind when assessing an ogive.[1] The cumulative distribution of the frequencies of a quantitative characteristic presents that same situation in a new light, revealing features about the underlying socio-economic situation that are not noticed in the corresponding frequency distribution.

## 2. Ogives of grouped Data

Statistical data of a quantitative characteristic are published in frequency classes. A cumulative frequency distribution, an ogive – pronounced variously as 'ojaiw' or 'ojiiiw' -- can be formed in two ways:

1) by adding successively the grouped frequencies of all classes at or below the given class, usually as "the frequencies of all class intervals below the upper limit of the given class", beginning with the first class interval, continuing up to the last, usually open-ended, class interval.

2) by subtracting from the total of the frequency distribution the frequency data of all classes above the given class. This rarely presented alternative form of the ogive starts with the total of all frequencies, reducing that amount by the frequencies of each successive class interval. In the following I will discuss only the first type of ogives.

Cumulative frequency distributions of socio-economic data, to repeat the initially made statement, are shaped by the magnitude and strength of the quantitative

† To this professor Joseph S. Nye, Jr of Harvard University, adds a: quotation of former undersecretary of state David Newsom: " the growing withdrawal of university scholars behind curtains of theory and modeling would not have wider significance if this trend did not raise questions regarding the preparation of new generations of the future influence of the academic community on public and official perceptions of international issues and events Teachers plant seeds that shape the thinking of each new generation, this is probably the academic world's most lasting contribution." The Washington Post, April 13 2009, p. A15

characteristic of each individual case but also by the less obvious but pervasive forces of the institutional, political and economic environment in which the data originated. These historically and geographically unique circumstances determine the individual values and how they are distributed along the values on the horizontal scale. Each one of the n individual 'statistical counting units' (not to be mistaken for 'unit of measurement' such as Kg, or meter in which these values are expressed) represents the strength of each individual case as expressed in its numeric score. Together these individual scores represent, in a more general, less concrete but none-the-less quite descriptive manner the socio-economic setting in which these individual cases have obtained their numeric scores, at that specific time, region and socio-economic setting. These data are to be understood as describing quantitatively socio-economic history, not as abstract algebraic numbers, to be manipulated mathematically. These accumulated data reflect the legal and customary support and constraints of that society, that created a specific environment, favorable or hostile, in which the individual socio-economic actors, each with its individual 'strengths' achieved the recorded values, the statistical data. This awareness enables us to interpret an ogive as describing socio-economic reality.

## 3. An intuitive Approach to Interpretation

To better understand the meaning of the particular shape of a cumulative frequency distribution one can imagine the ogive as the contour of a sea shore. The initial, flat portion of an ogive is like a gently up-sloping sandy beach. The steeper portion of an ogive, corresponding to the higher values of the horizontal scale, is like that of a shore that rises more steeply, perhaps like a rocky cliff. Each individual value in that data set -- value of the cases in that frequency distribution, but also by extension, cases that have not yet occurred, but may happen later -- can be imagined as an individual ocean wave arriving at that shore. Each arriving wave first passes over that flat portion of the shore, and if strong enough, will reach higher up toward the steeper part of the shore. It will run its course, ending there its advance at some point of that shore. Depending on the beach profile and the magnitude of the wave, many waves will not reach very far, ending their run somewhere against the resistance of the increasingly steeper portion of the shore profile. Only few occasional powerful waves will advance over most of the lower portion of that shore profile, reaching high up before having spent their momentum, ending their advancement. Once they have reached the flattening portion on the upper end of that coast, however, the resistance to their further advancement decreases, making it easier for them to advance even further. In the following graphs of ogives I will apply this intuitive approach at interpretation to economic and social data.

The ogive of the income data of persons 15 years or older during 2006, in Figure 1 allows a sensible interpretation of that situation. For the sake of easier comparison, the cumulative frequencies of all ogives are expressed as percentages of the total of all frequencies. The dotted line representing incomes by persons born abroad - mostly immigrants - is higher than the solid line for the income of the entire population. It shows that a higher percentage of all foreign born income recipients (equal 100%) are in the lower income ranges. Their curve also rises steeper than that of the general population, suggesting that overall, immigrants in 2006 found it more difficult to advance his/her income in the range of $10,000 to $25,000 than US-born persons. To make such an upward move on that 'shoreline' faced greater difficulties and required a bigger effort than a similar advancement for the mostly US-born income recipients. Both curves become flatter from incomes of $55,000 on. In the highest income ranges (not shown on that graph) both curve become nearly horizontal, meaning that at that socio-economic level the differences that exist in the lowest income ranges, disappear.. Also note that the horizontal scale would have to be extended to over $1,000,000 to reach the 100% level on top of the graph.

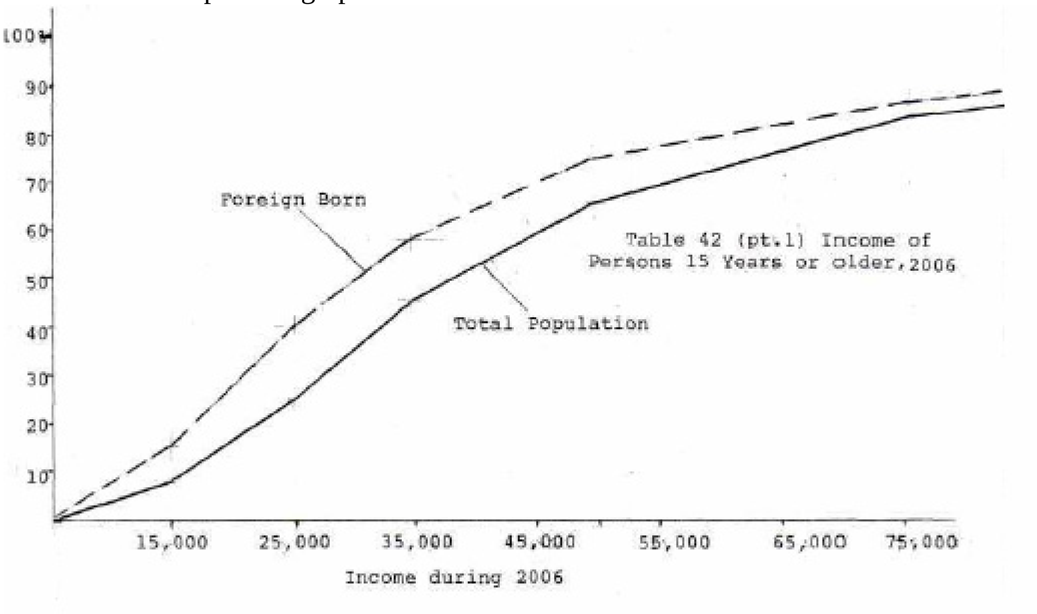

Figure 1 Income of Foreign-born and Total US Population, 2006

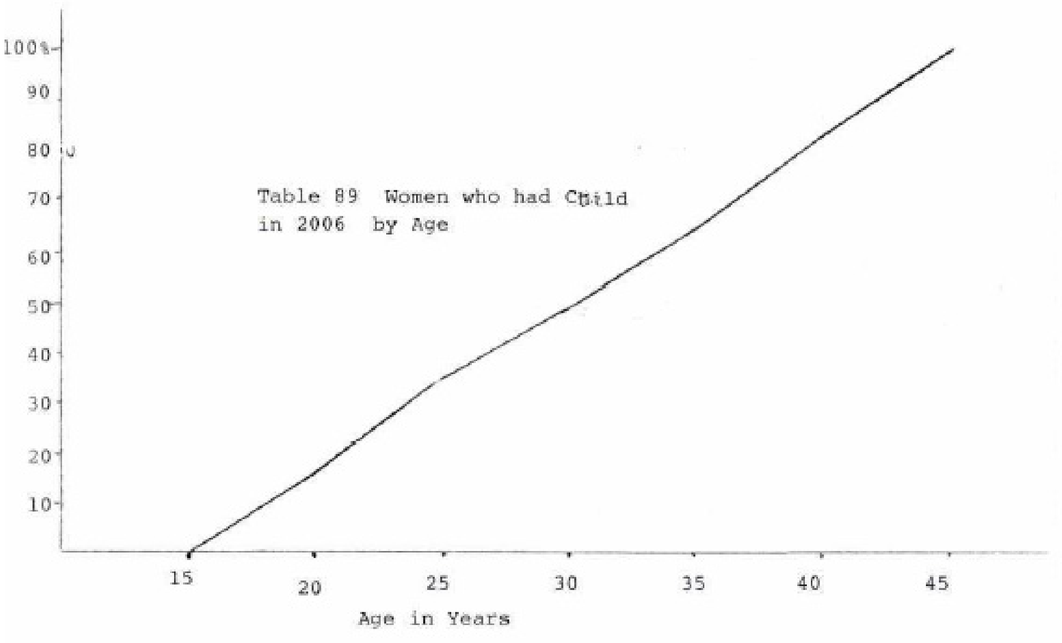

Figure 2 Women by Age who had a child in the US, 2006

The ogive of Figure 2 can approximately be considered to be a straight line except for a very slight increase of young women becoming mothers between the ages 20 and 25 years[2]. Otherwise these women who had a baby in 2006 were surprisingly evenly distributed between the ages 15 to 20 and 25 to 45 years.

If one were to consider this ogive to be a sea shore, and each of these women an ocean wave, these waves would end their run anywhere between the beginning of the beach and its upper end with a slight increase of  waves ending on the stretch of beach between the values 20 and 25 (on the horizontal scale). Otherwise every wave encountered about the same resistance to its advance along that entire up-sloping shore profile.. This surprising description is limited to that particular group of women. It says nothing about other characteristics of these women or their children, the part of the country and the timing of the birth event in that year. Nor can it be extended to all women in that year, or to women in the same age group in  other years.

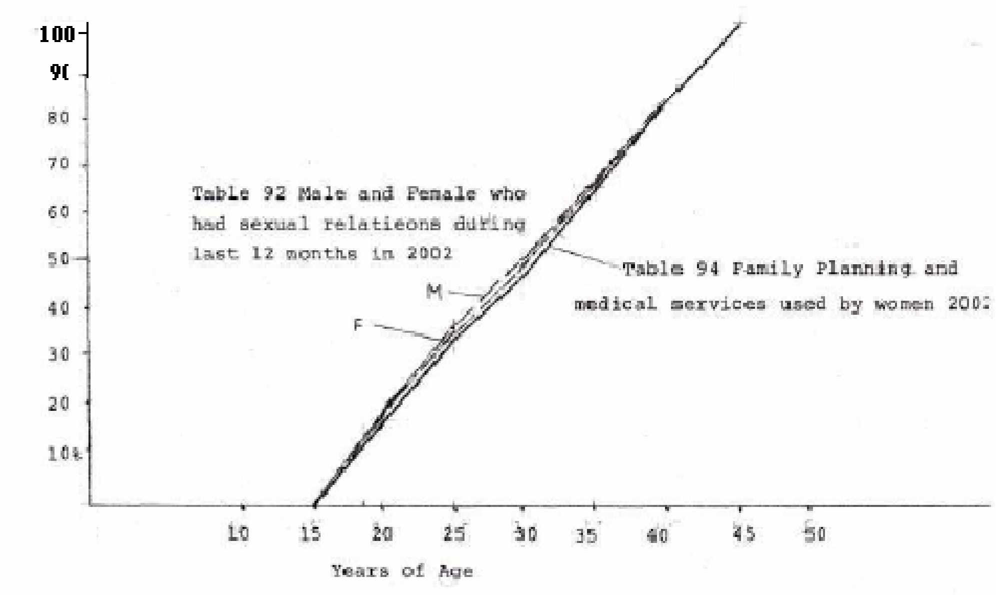

Figure 3 Self-reported Sexual Habits of US Population, 2001

An analogous interpretation applies to the ogive in Figure 3 reporting on related matters. [3]

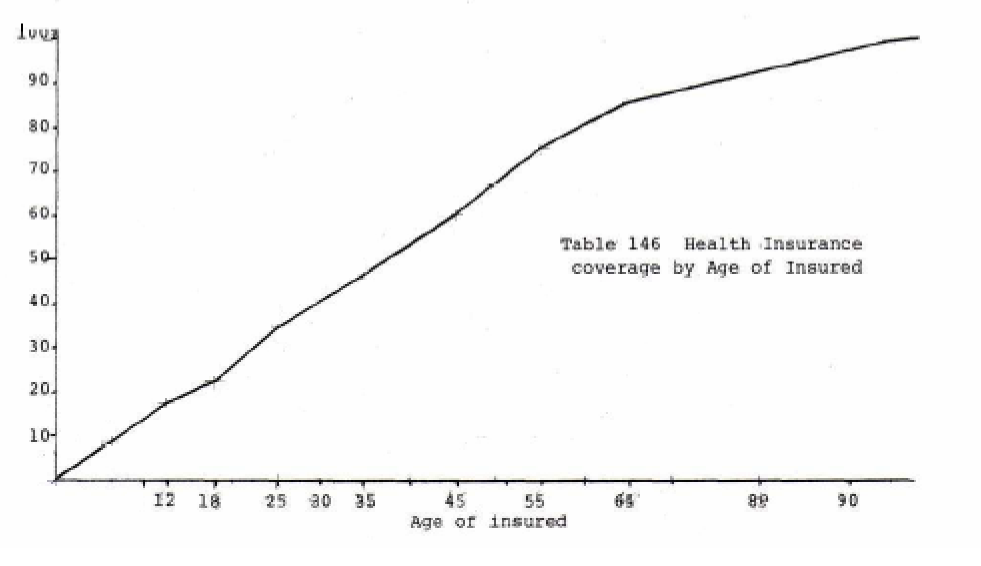

Figure 4 Health Insurance Coverage of US Population, by Age of Insured

This ogive of Figure 4 surprises in the seemingly equal facility of each age group to obtain health insurance (any kind). The value of the ogive e.g. at 20% on the vertical scale, for those 12 year old (on the horizontal scale) indicates that of the entire insured population 20% were that age or younger[4]. If that ogive were the profile of a shore, incoming waves would find the same kind of resistance to their further advance at the low end as well as on the high end of that evenly rising shore profile. In terms of health insurance, there did not seem to exist an institutional or cultural environment that would favor or hinder obtaining health insurance at any age. To further assess this situation, though, this ogive would have to be compared to the ogive of the age distribution of that entire population, like the dotted curve in Figure 6.

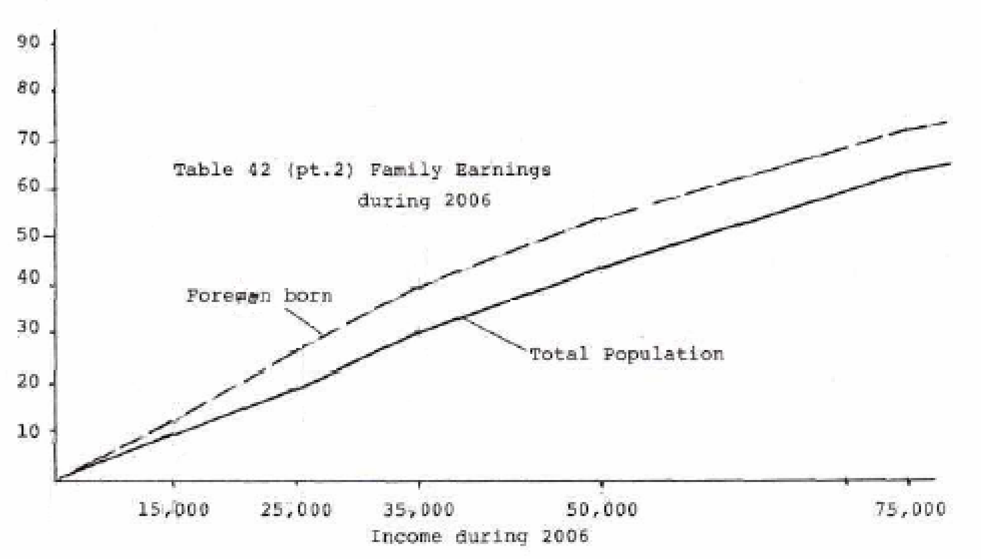

Figure 5 Family Earnings of Foreign born and US born Population, 2006

The curves in Figure 5 represent the ogives of family earnings. Notice that the horizontal scale is cut off at incomes of $75,000. These ogives of the income situation reach the top of the graph, the height of 100%, beyond the horizontal scale value of "$1,000,000 or larger" (imagine how far to the right you would have to extend this graph so as to be able to plot incomes of that magnitude!). These apparently flat curves indicate a relatively even environment for the income of families to advance from one income level to the next higher – or decline to a lower -- level of the incomes in this limited range. The dotted line of the ogive of the incomes of foreign born immigrant families indicates that a higher percentage of those families earned in the lower income ranges, and encountered more resistance to advancement.

In Figure 6 the lines are reversed: the dotted line represents the ogive of the percentage of persons of certain ages of the entire US population. The solid line represents the ogive of the foreign born population. The explanation, particularly of the initial portion of that curve, should be left to demographers and sociologists. It appears, though, that a smaller percentage of children and young persons are foreign born than the comparable age groups in the general population. At the riper ages, from about 40 years of age on, both ogives seem identical, the ogive of the foreign born only a trifle higher. In general it seems to indicate that a higher percentage of the foreign born immigrant population entered the USA as adults, while a smaller percentage entered as children. The contrast becomes clear when comparing this ogive with that of the dotted-line ogive of the population born in the US. Then also, before interpreting these ogives as the dynamic image of "waves moving up against the beach profile", it should be recalled, that this is a static still picture, a snapshot of the situation in 2006. The actual dynamics of that situation may have developed differently. On the face of it one might be inclined to conclude, not necessarily in line with the actual dynamics of the situation, that a person in the immigrant population faced fewer difficulties to grow up than the comparable age groups in the general population. But it appears that from age 45 on the immigrant population had the same opportunity to grow older as the general population. It also seems to be somewhat counter-intuitive, to further interpret the flatter slopes of these ogives, that it is easier for an older person to grow older than for a younger person to get older. This seems to confirm the increasing life expectancies of older people according to tables of vital statistics.

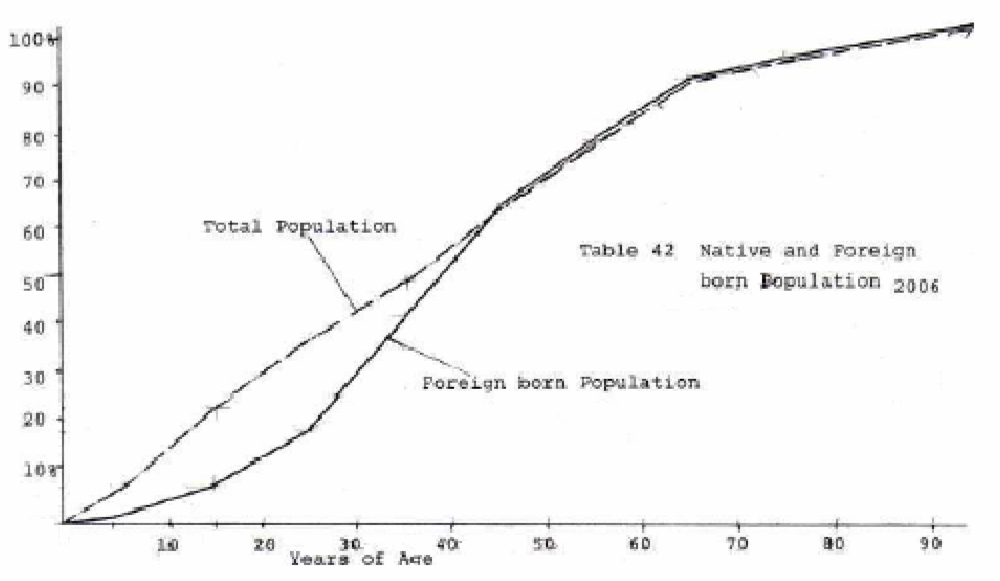

Figure 6 Total Population and Foreign born Population by Age, US 2006

Most socio-economic frequency distributions are right-skewed: the bulk of the data occurring at the beginning, the lower end of the horizontal scale, with the typical long, thin tail-end to the right. The image of a coastline, that would match the ogive of many socio-economic frequency distributions, would more resemble a cliff, rising steeply, then leveling off toward the higher values on the horizontal scale.

The cumulative frequencies of murder victims in Figure 7 gives an idea of such a shore profile. There were few young murder victims, while it became abruptly more likely for a person in the entire USA in 2005 to be murdered between the ages 18 to 35. If such a murder victim in the US were considered to be an ocean wave moving up against the seashore, that wave quite likely would end its advancement in that steep portion of that shore. In other words, a person in the US could easily reach age 18 without being murdered. That changed rather abruptly at the ages 18 to 35 where it became more likely to be murdered. After that it was less likely to be murdered beyond the age of about 55. This graph would illustrate a seashore that begins rather flat, like a low, near sea-level, narrow sandy beach that is clearly separated from the back country by a high and steep shore profile that would protect residents living on the high side of that shore even from very strong waves e.g. in a big storm.

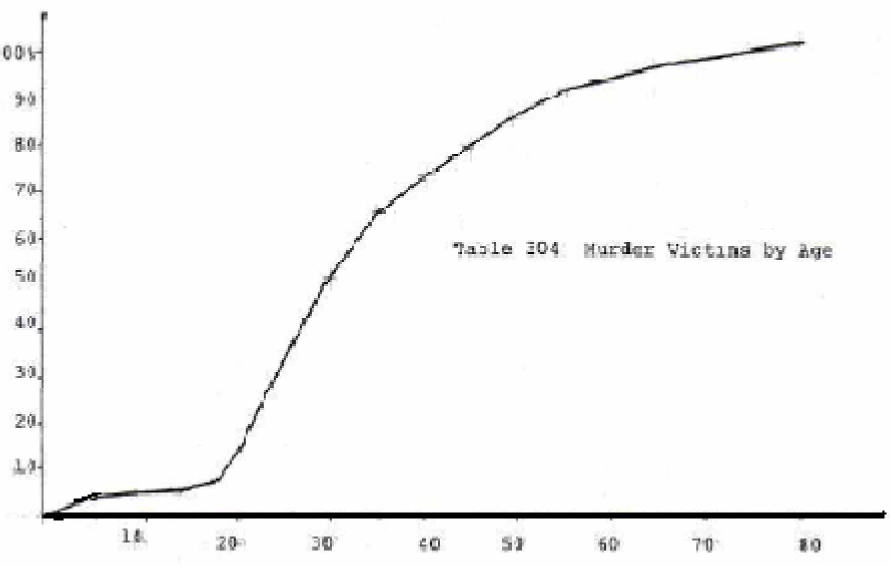

Figure 7 Murder Victims in the US during 2005, byAge

More extreme is the ogive of the US Federal Income Tax Returns arranged according to the declared adjusted income in 2004, in Figure 8. Income distributions in general are strongly skewed to the right, with long tail ends of their frequency distribution. An example of a left-skewed Frequency Distribution (not shown) could be an ogive of the income distribution in a society with an income tax structure, that favors small income recipients but makes it increasingly more difficult for persons and corporations who would have earned large after-tax incomes in the environment of a less punitive tax structure that favors the wealthy.

The steep rise of this ogive in the low income levels indicates the difficulty of those income receivers to advance to a higher level of a larger adjusted taxable income. If that were a seashore most waves would break in the lower, steepest portion of that shore, up to $20,000. Between this point and the next change in the shore profile, at $50,000, there appears to be a reduced resistance to the advancement of stronger waves that made it up to that level. Once a wave made it up to the higher level of the shore, at $100,000, the resistance of the contour of the shore gradually decreases further, becoming less and less of a hindrance to further advancement, allowing such a wave to advance with ever less resistance. The parallel to 'adjusted incomes' should be obvious. Although this is a picture 'frozen in time', it does give a vivid illustration of the difficulties as well as the advantages the recipients of such incomes encountered in this environment of a capitalist economy. Those making incomes between 25,000 and 60,000 appear to have somewhat less difficulty to advance to a somewhat higher income. This gets easier from about $80,000 on, and seemed to have become increasingly easier to advance income-wise in the higher income ranges. To stay with the image of a wave rolling against that seashore profile, if it were strong enough to reach high up to the flattening portion of that shore, there is nearly nothing that would stop such a wave to progress even much further. Our capitalistic society seems to place fewer and fewer obstacles to income recipients to the further advancement of their large incomes into the even larger million dollar ranges.

The income distribution in a truly socialist society would look like the Figures 2, 3 or 4 that deal with other social issues, not incomes[5].They indicate a smooth progression from one value on the horizontal scale to the next. In other words, it would be equally easy, or difficult, to advance from any given income level, on the horizontal scale, to the next, regardless of the level of that income.

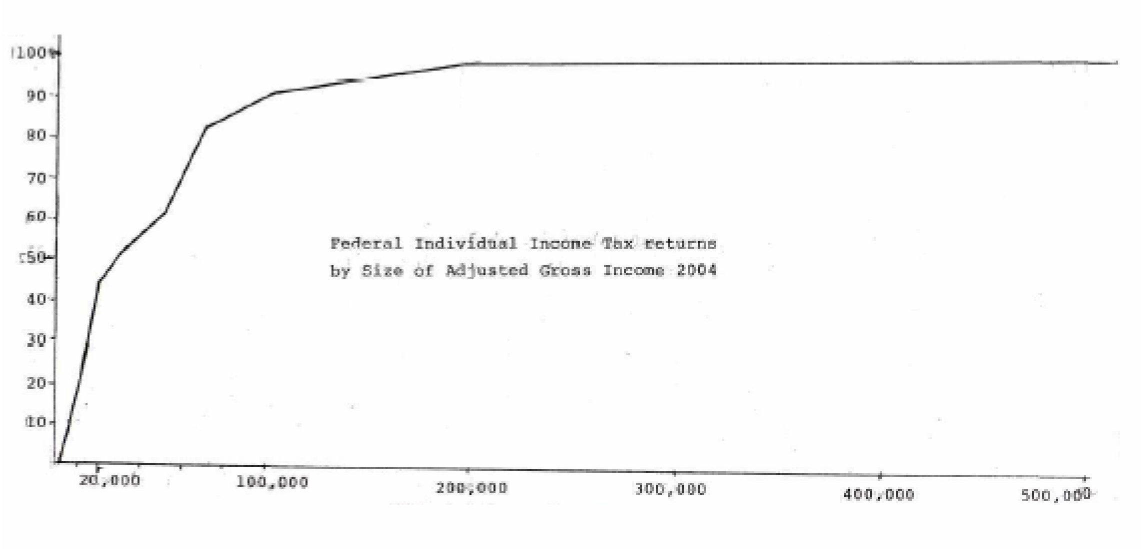

Figure 8 Adjusted Gross Incomes reported to the US Income Tax Authority in 2004

The data in many frequency distributions seem to be rather symmetrically distributed. That is deceptive because the frequencies in the ever-widening class-intervals need to be adjusted because in a frequency histogram the frequencies indicate their value by their area, not just by the height of the histogram bar. To make the bars in such a frequency graph really comparable they must be adjusted .e.g. as ‘frequency per one scale unit in x’ to properly reveal the true asymmetric nature of such a frequency distribution.. The ogive, on the other hand, automatically takes care of such class unevenness. No adjustment of the frequencies of their classes is needed..

In Figure 9 the uneventful, smooth line between ages 5 and 18 indicates the routine advancement of children and young people through the school system from Kindergarten to Senior level in high School. The dynamics after age 18 is changing given that this data-set refers to all kind of schools, including advanced technical and trade schools, Colleges and Universities, with undergraduate and graduate students.

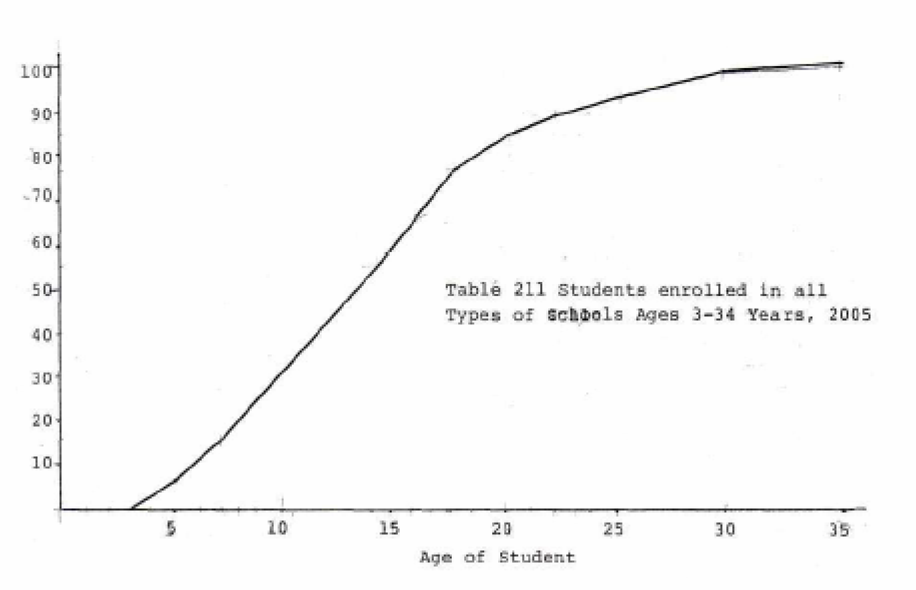

Figure 9 Age of Students Enrolled in all Types of Schools, US, 2005

Then consider the Enrollment of students in colleges and graduate schools in Figure 10. The ogive of those at and above the age of 18, looks more demanding. Entering higher education, at a University or a College, obviously undergraduate studies, is concentrated in the short time between the ages of 18 and 22. as indicated by the steep part of the ogive. Most of the older persons obviously are graduate students, who have reached that stretch on that curve. The moderate incline of that portion suggests that to study seems to be less concentrated on certain ages but was spread out evenly to the older, more mature population in 2005.

A somewhat different ogive arises from the discrete quantitative variable "Number of Persons in a household", Figure 11. This step-wise ogive has an analog interpretation. From a one-person household to become a 2-person household is a major step. The hesitation shown in the step to 3-persons in a household is nearly the same as that from a 3 to a 4-person household. As one can see, the steps become smaller and smaller, implying that it becomes easier to move form a 4-person household to one with 5 persons and even easier to move on to one with 6 persons. There seems to be nearly no difference between a household with 6 and 7 persons. It is tempting to extend this to conclude that from then on it is nearly indifferent how many people beyond 7 are in a household. To add a sense of real-life dynamism to this static picture of discrete steps, the dotted line indicates a kind of gradual transition from one family size to the next. In most instances the addition of one more family member, usually a child, does not happen abruptly as the addition of a full-grown adult, but as a baby, that only gradually alters the family dynamics, as suggested by the dotted line, converting this discontinuous graph into one of smoother continuity. It also makes it easier to imagine this as a shore profile in which each household proceeds like an ocean wave approaching that shore. The ogive in Figure 12 about the population size of county governments -- this statistic of county and city governments as "statistical counting units or elements" is unique to the USA -- presents a similar shape stemming form the fact that there are many more counties with smaller populations than there are those with very large populations. It seems sensible that a small sized county can change only to one that is slightly larger or smaller, while a county with a very large population can easily become one with an even larger

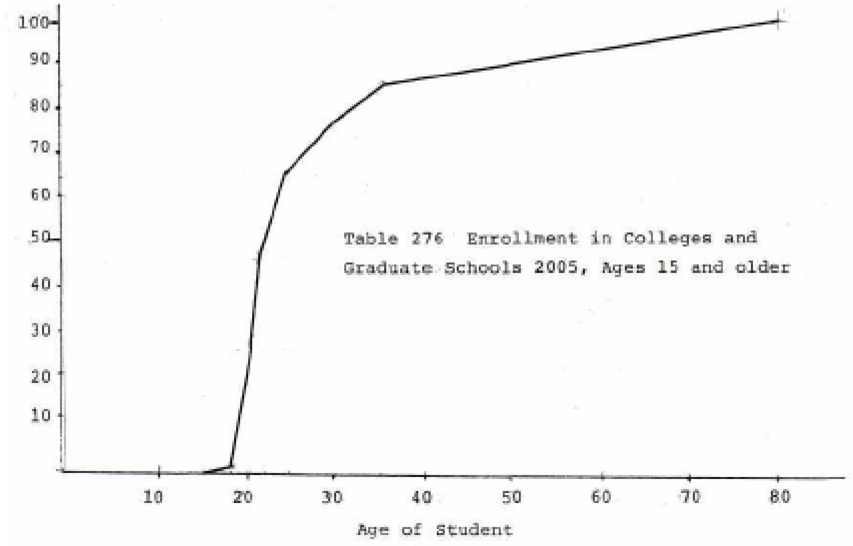

Figure 10 Enrollment in Colleges and Universities US 2005 by Age of Student

population. The interpretation of this ogive in terms of a sea shore and ocean waves reaching it follows the already established pattern.

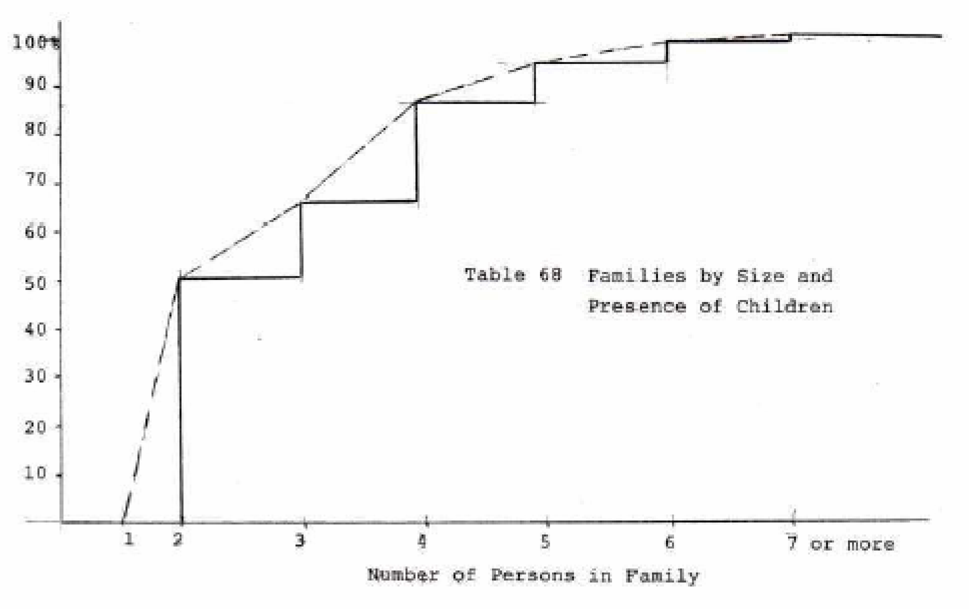

Figure 11 US Families by Number of Person in Family, 2006

Figure 13 deals with townships that are also regional administrative units, but with much smaller populations. The steepness of that ogive indicates the high density of accumulation many townships with very small populations. For most of them only small changes in their population are at issue, as expressed in the steepness at the beginning of their ogive. If a township were an ocean wave, it would most likely end somewhere on that near-vertical part of that shore, while a strong wave, representing a very populous township, can expand with little resistance to become an even larger township. Here we dealt only with ogives of frequency distributions that are skewed to the right. When I prepared this presentation I could not find a single socio-economic frequency distributions that was skewed to the left, with an elongated tail to the left of the highest frequencies on the upper end of its horizontal scale.

The initial steep shore profile of the ogive in Figure15 -- of which Figure 14 is an amplification of the asset values at the beginning of that scale.-- indicates how difficult it must be in the US economy in order to establish a small business enterprise to acquire the necessary assets, or for an existing business corporations with a small asset size to increase its assets. Although this is a static, cross section of the firms in that industry of the US economy at that point in time, one can see how difficult it appeared to be for any of the existing small firms to advance from one asset level to the next higher level. In these lower ranges - on the horizontal scale - even a big effort would not allow them to advance by much in that steep part of the cumulative distribution.

Relatively few individual 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 that 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 the flat portion of that ogive.

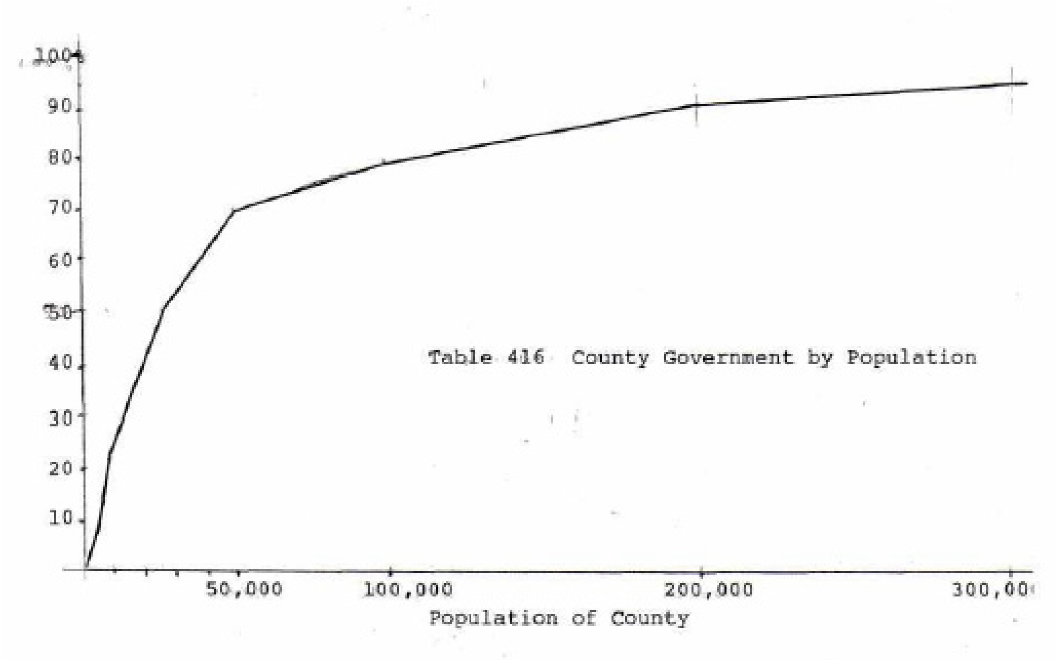

Figure 12 US County Governments by Size of its Population, 2006

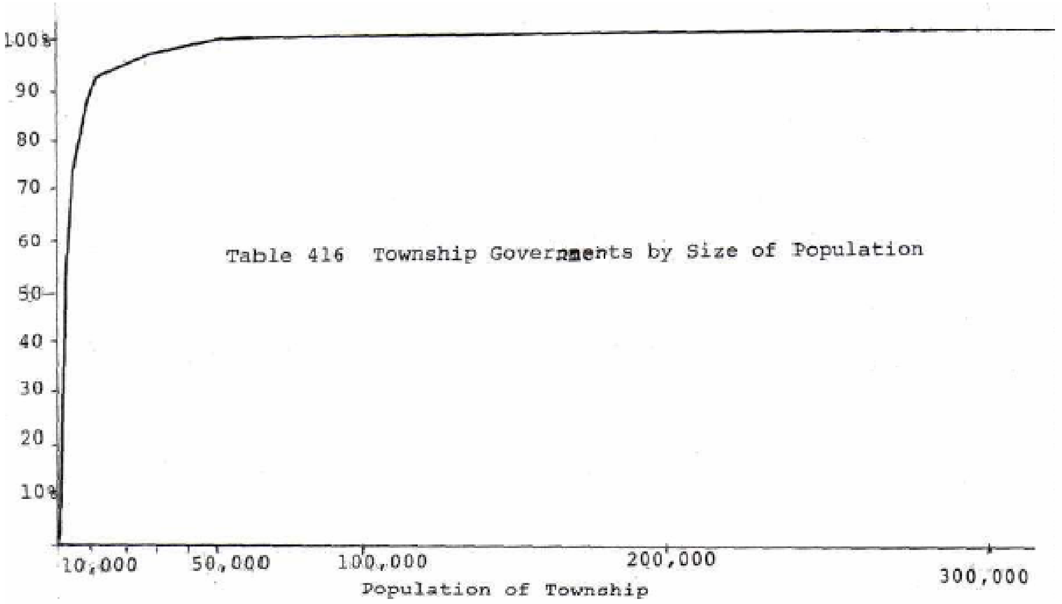

Figure 13 US Township Governments by size of their Population

In terms of the distribution of firms by size of their assets, and quite likely also the size of their revenue, to stay with that analogy of the profile of a sea coast and ocean waves approaching it, once reaching the upper, flattened portion of that coast, even a minor, additional effort on the part of such a 'wave' can produce an effect equivalent to an added income of a few more million dollars. Every frequency distributions of socio-economic data is the result of such a selection process that is shaped by the institutional and economic constraints of that society that can encourage but also resist the advancement of the individual units in a society.

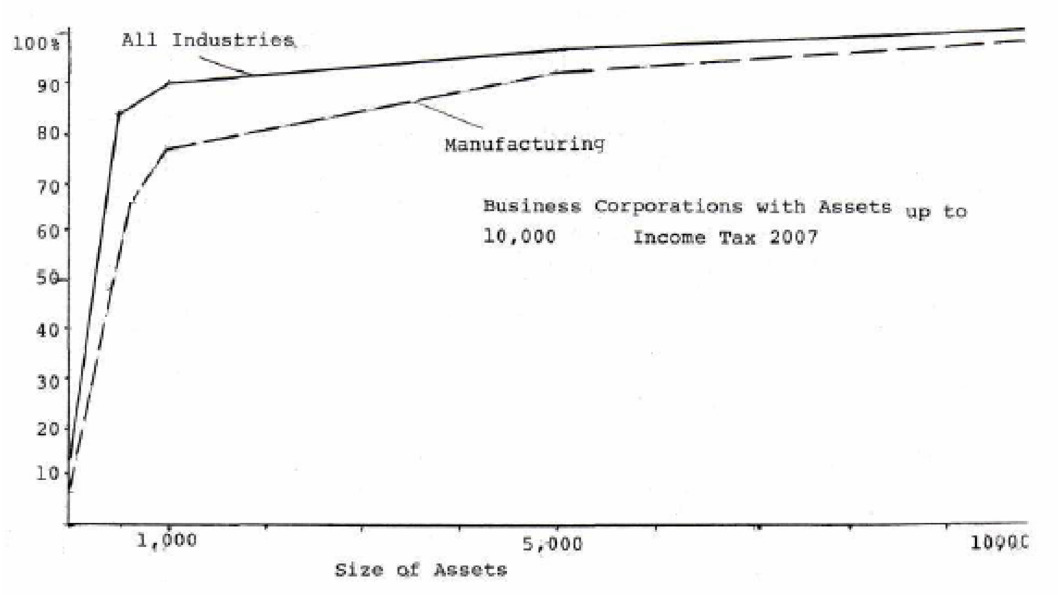

Figure 14 US Small Business Corporations by Size of their reported Assets, 2007

The corresponding ogive would look like a shore profile that initially rises gently, like a gradually up-sloping sandy beach, that becomes steeper only at the end. In such a situation most of the imaginary ocean waves have no difficulty rolling over that initial, flat part of the shore profile, breaking further back, higher up where the steeper shore puts an increasingly greater resistance to its further advancement. To show that such extremely asymmetric distributions do not only occur in the social sciences, Figure 16 is such an example of asymmetric distributions in the sciences and engineering. The oil-spill ogive seems to show that large oil spills are not frequent. Comparing a given oil-spill to an ocean wave approaching this shore profile such waves have a very difficult time to advance beyond the very beginning of that beach. Most of the oncoming waves barely appear on that cliff. That means that most oil spills if they happen at all, have been quite small. But once such an oil spill of some magnitude does occur, it has easily spread into a huge spill indicated by the near horizontal trace of that ogive, posing little resistance to such an event becoming a major disaster. When Interpreting a de-cumulating ogive, which was not discussed in this paper, the general socio-economic environment is the same, but there is a difference: the de-cumulating ogive, F(x ∃X) represents a profile of the strength of the individual ‘statistical counting units’. At the outset all appear equally strong, sufficiently strong to be included in this FD. But fewer have the strength or power to reach the higher values on the x-scale. This form of de-accumulation refers to the same general situation but views it from the individual’s point of view. Returning to the image of an ocean wave, it shows how many ocean waves have the strength to reach the higher values on the horizontal scale.

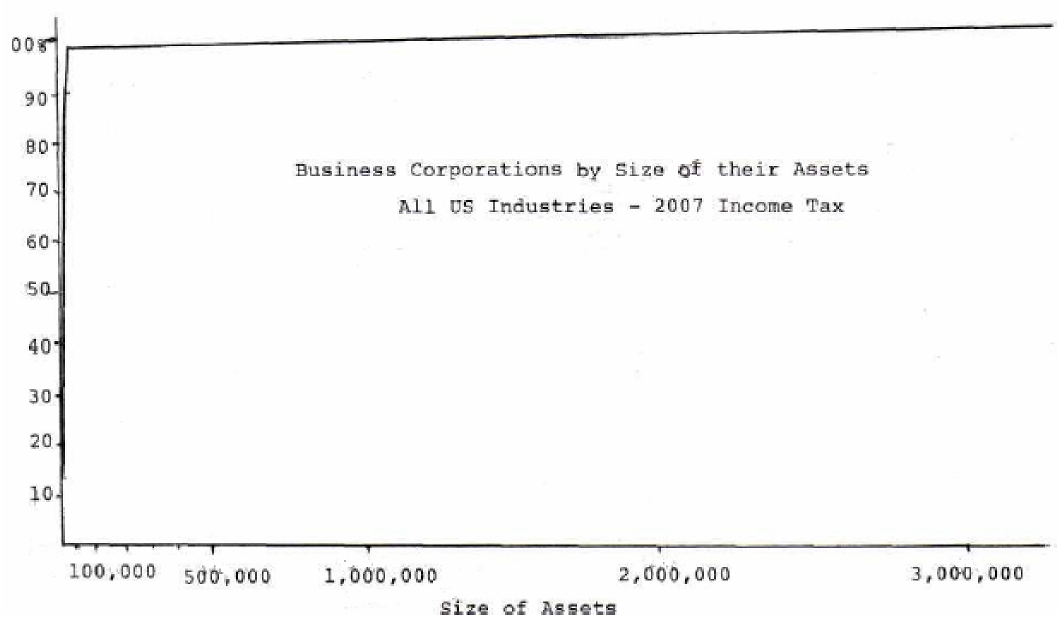

Figure 15  All US Business Corporations by their reported Asset Size, 2007

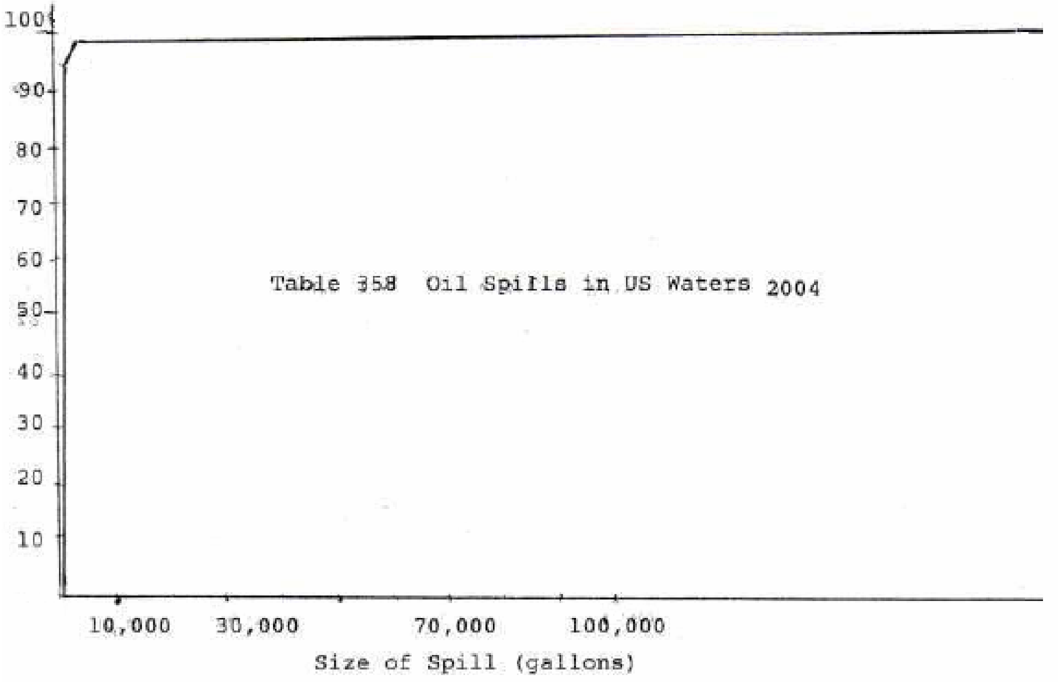

Figure 16  Recorded Oil Spills in US Waters, 2004

## 4. Ogives of ungrouped Data

It is a rare instance when the ungrouped individual values become available. In that case the  Ogive can also be constructed from these ungrouped data, that must be sorted by size from smallest to largest value. Their ogive presents a serrated profile in which each case forms a small 'riser' in the stepwise picture. The following graph shows the frequencies of an original set of 482 wages of factory workers in a cotton mill, Caracas, Venezuela 1952 (see Appendix). Each 'statistical-counting-unit', a workers weekly wage, is added as the 1/N th case, forming a small 'riser' of the many steps that establish that cumulative frequency distribution.

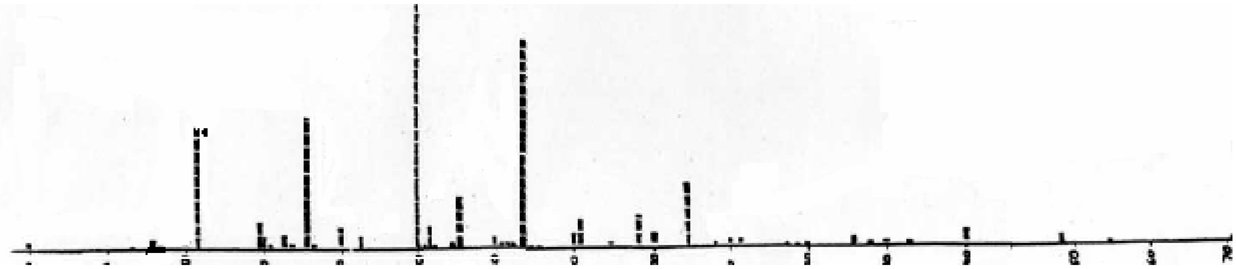

Figure 17 Weekly Wages of a Group of Cotton Mill Workers, Caracas, Venezuela 1954

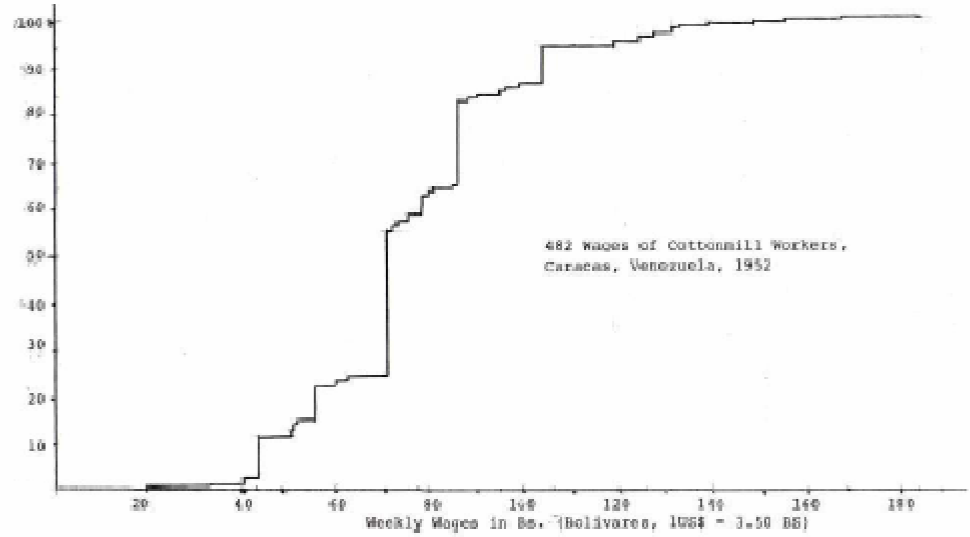

Figure 18 Ogive of the Weekly Wages of Cotton Mill Workers, Caracas, Venezuela 1954

The interpretation of this serrated profile is following the same reasoning as the ogives of grouped data. It gives a more detailed view of the exact values at which difficulties for the further advancement arise of a wage earner in this corporation, at that time and place arose. This picture of the entire workforce does not explain why the wages of e.g. 70Bs.and 84 BS seemed to be such hurdles. In reality these are workers with different skill levels and specialties. This is an important detail that only a finer breakdown of that total can reveal and explain.

## 5. Summary

.Given that the socio-economic situations must be expected to change in every respect, such a profile as provided by the ogive, should not be considered as a reliable predictor of future incidents, as the situation for future cases quite likely will be different. The given ogive should be taken as the description of the situation in that particular population, at the time and in that region.

In social and economic data asymmetries are the rule, symmetric frequency distributions are the rare exception. Gaussian distributions simply do not occur. Throughout this study the respective vertical scale values were converted to percent figures, to allow reasonable comparisons of different ogives. By converting to percentages the values of the cumulative frequencies – the vertical scale - one can make the ogives of different frequency distributions somewhat comparable. This has been done in all ogives in this paper.

Finally, it should be kept in mind that socio-economic phenomena are portrayed mainly through qualitative variables and the geographic location of the 'statistical-counting-units'. Quantitative variables, presented in frequency distributions, though having historically drawn attention early on and are well suited for numeric manipulation, are less important, acquiring meaning only in conjunction with some qualitative characteristics as their complement.

From here it is only a small step to extract additional information from a standardized ogive by using the cumulative total value on the horizontal axis -- e.g. personal income --to construct one of the measure of concentration[6]

## Notes

[1] Mathematical statistics had the correct intuition, giving priority to the cumulative distribution function F(x). The probability density function f(x) is then considered as secondary, as the first derivative of the cumulative density function, usually as a probability density function,. But this was not accompanied by an explanation of the reasons for preferring the cumulative over the simple frequency distribution.. – Even when the data are treated as if they were from a dynamic, evolving situation for the purpose of illustration, it must be kept in mind that the data of the actual dynamics over time of that same situation may show a different picture.

[2] The data of this and of subsequent figures are taken from the Statistical Abstract of the US, 2008.The text inside each graph lists the number of the tabulation from which the data were taken. The Figures 14 and 15 are from the "Source book, Statistics of Income 2004, corporate Income tax returns, IRS, Publication 1053, 2-2007". Figures 17 and 18 were data supplied by the accountant of that corporation,at that time a student in the authr's Economic Statistics I class in 1952, Facultad de Economia, Universidad Central de Venezuela, Caracas, Venezuela

[3] The data of this Figure are taken from tables #92 and 94, surveyed in 2002.

[4] This figure is not to be interpreted as indicating that 80% in that age group were not insured because this tabulation does not deal with the issue of the failure to obtain health insurance.

[5] Note that the ogives in both Figures 1 and 5 also dealing with income aspects, but amplify the initial portion of Figure 8. Neither ogives in Figures 1 or 5 reach the top level of 100%. That happens because the horizontal scale is far too short to cover the extended, flattening portion of those curves to the multi-million incomes that are more fully presented in Figure 6.

[6] See e.g. Bruckmann, Gerhart "Konzentrationsmessung" as chapter 26, The Measurement of Concentration, in: Bleymüller, Josef, Gehlert, Günther and Gülicher, Herbert, *Statistik für Wirtschaftswissenschaftler* München, Verlag Franz Vahlen, 1981, zweite überarbeitete und erweiterte Auflage, pp. 185-190.