ISS DEN HAAG
8 3501 00010 3819

of
udies

# Exploratory Data Analysis on Indebtedness in the Third World

Marc Wuyts

Occasional Paper No. 101

1127

# Exploratory Data Analysis on Indebtedness in the Third World

Marc Wuyts

December 1985
Institute of Social Studies

The views expressed in this publication are those of the author and not necessarily those of the Institute of Social Studies.

© Institute of Social Studies
The Hague, 1985

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of the copyright owner.

ISBN 90-6490-015-0

Printed in The Netherlands

# INTRODUCTION

This paper deals with the methods used in *exploring* numerical data as a tool of socio-economic analysis. In applied research, analysis proceeds neither purely deductively nor merely inductively. Rather, it involves a continuous interaction between formulating hypotheses and testing these hypotheses against empirical evidence. It is a process of *searching*, not unlike detective work: it is not just *working out* the answers to already well formulated questions, but also arriving at sharper questions to be investigated subsequently. Data assume a *dual* role within this process of research. On the one hand, a well formulated hypothesis is tested against the data, the selection of which was determined by this hypothesis. On the other hand, however, there is also the fact that the careful exploration of numerical data (the collection of which was determined by some broader hypotheses related to the issue under study) may suggest new avenues of analysis – new, sharper questions which will need further investigation, data gathering and testing. In the process of looking at the data one or another characteristic may strike the researcher: it begs for posing new questions, as yet not fully formulated nor formally tested. The latter role of data within analysis is known to all applied researchers in social analysis, but by its very nature is not formalised. It appears as the product of 'playing around' with data – more an art than a rigorous procedure. But exploring data nevertheless involves systematic, rigorous work which allows one to look at the different aspects manifested by these data.

Most statistics courses taught in social science departments tend to concentrate on the former role of data within research, i.e. the testing and confirmation of hypotheses against the empirical evidence. The focus is, therefore, on such issues as sampling theory, the derivation of different tests and the properties of estimation, *given* the stochastic process which generated the data. It therefore assumes that the researcher is equipped with a well defined hypothesis and knows the central characteristics of the population distribution from which the sample data were generated. One is concerned with specifying the appropriate technique to verify a well defined hypothesis, given clearly specified assumptions about the stochastic characteristics of the data.

In the course of carrying out research, however, the formulation of the basic questions is generally the critical problem. Furthermore, this constitutes the *creative* aspect of the work, which must subsequently be subjected to rigorous testing. The role of data analysis in obtaining a clearer view of the issue – in formulating new and more precise questions – is generally not emphasised in teaching statistics. But in many instances it accounts for a considerable share of the difference between empirical work which leads to a better theoretical grasp of a problem and such work which merely remains descriptive.

Furthermore, more often than not one cannot specify *a priori* the stochastic characteristics of the data generating process. A careful study of the inherent patterns within the data helps in finding out which assumptions are more likely to be satisfied and which are not. This is particularly important, since most economic and social research cannot rely on repeated sampling based on carefully designed experiments, as there is little scope for experimentation. Most of the analysis needs to be based on existing data sources and some additional sampling on a non-experimental basis. Exploratory analysis should encourage one to feel more confident about the assumptions made in setting up formal tests or carrying out estimations.

Exploring data is a more tentative and less formalised process than confirmatory statistics. This does not mean, however, that it does not involve a systematic method or rigour in its approach. It is based on a few common sense principles which underlie the type of approach and the methods it uses.

Exploring data should be based first of all on *quick* statistics, i.e. relatively simple calculations applied to a manageable number of data. Its purpose is to try out various ways of looking at the data, and such a process is only productive if each approach does not require large numbers of data and complex calculations. If this were the case, the researcher would be more likely to end up with a headache than with clearer questions about the phenomena he or she is studying. It could be argued here that the researcher could take recourse to using computer facilities and pre-prepared statistics packages. The aim, however, is to get a *feel* for the data – its patterns and the deviations from those patterns – and this is best done by looking carefully at them. Furthermore, most computer packages are concerned with *confirmatory* statistics rather than with *exploring* data. The summary and synthesis of the data they provide may not be the most appropriate to exploratory analysis for the reasons given below.

Secondly, exploratory data analysis should be based on resistant methods of *summarising* data patterns: i.e. the method used should be sufficiently resistant to incorrect inferences even when the data base is of poor quality or contains some extreme observations (i.e. outliers). Neither should any assumptions be made about the underlying stochastic process which generated the data: i.e. one should not necessarily assume that the data are drawn from a normal distribution.

Finally, the basic method should consist of summarising the basic characteristics of the data – such as level, spread, shape and the presence of clusters – and within this process at each stage studying the 'residuals' left when a basic characteristic is accounted for.

The purpose of this paper is not to survey the techniques used in exploring data, but rather to provide a concrete case of exploratory data analysis by taking a simple example. The example chosen here is the analysis of data on external public debts of developing countries as published by the World Bank. These data are grouped in three categories: low income, lower middle income and upper middle income countries.[1] All data used in this paper are for the year 1981.

It should be clear from the outset that the aim here also is not to provide a deeper analysis of the problem of the debt burden for developing countries in recent years, but rather to serve as an instructive exercise in illustrating exploratory statistics in action.

In the example, therefore, a situation is assumed in which a researcher, interested in the problem of the increasing debt burden of developing countries and of its distribution among them, sets about analysing a set of data to extract as many questions from it as can be obtained. Thus the focus is on the method of investigating numerical data, not on the wider context. Nevertheless, the wider context is needed in order to make sense of the suggestions thrown up by the data themselves.

The methods and techniques to be employed in this analysis are those developed by the statistician John Tukey, a pioneer in exploratory data analysis. References to his work as well as other useful literature for the interested reader are provided in the Bibliographical Appendix (at the end of this paper).

## SUMMARISING THE DATA

Table 1 summarises the World Bank's data on the external public debts of developing countries in 1981. It includes only those countries for which the data are available. One is immediately faced with a question when extracting these data from the World Bank report: should the grouping of countries by income *per capita* categories be preserved, or should they simply be combined into one set of observations? The answer is quite straightforward. Since the main concern is with the importance of the

*Table 1. The External Public Debts of Developing Countries in 1981 (in US$ millions)*

*1. Low Income Economies*

| | | | | | |
|---|---|---|---|---|---|
| Chad | 201 | Bangladesh | 3,850 | Ethiopia | 792 |
| Nepal | 234 | Burma | 1,639 | Mali | 738 |
| Malawi | 685 | Zaire | 3,960 | Uganda | 540 |
| Burundi | 154 | Upper Volta | 296 | Rwanda | 172 |
| India | 17,975 | Somalia | 877 | Tanzania | 1,476 |
| Guinea | 1,255 | Haiti | 360 | Sri Lanka | 1,585 |
| Benin | 549 | Central African Republic | 213 | Sierra Leone | 346 |
| Madagascar | 1,258 | Niger | 605 | Pakistan | 8,814 |
| Sudan | 4,807 | Togo | 860 | Ghana | 979 |

*2. Lower Middle Income Economies*

| | | | | | |
|---|---|---|---|---|---|
| Kenya | 2,228 | Senegal | 944 | Mauritania | 827 |
| Yemen Arab Republic | 1,094 | Yemen PDR | 640 | Liberia | 592 |
| Indonesia | 15,529 | Lesotho | 107 | Bolivia | 2,422 |
| Honduras | 1,223 | Zambia | 2,294 | Egypt | 1,644 |
| El Salvador | 664 | Thailand | 5,169 | Philippines | 7,388 |
| Papua New Guinea | 613 | Morocco | 7,879 | Nicaragua | 1,975 |
| Nigeria | 4,652 | Zimbabwe | 880 | Cameroon | 2,034 |
| Congo (Popular Republic) | 1,105 | Guatemala | 684 | Peru | 5,974 |
| Ecuador | 3,392 | Jamaica | 1,434 | Ivory Coast | 4,497 |
| Dominican Republic | 1,260 | Colombia | 5,123 | Tunisia | 3,171 |
| Costa Rica | 2,246 | Turkey | 13,809 | Syrian Arab Republic | 2,337 |
| Jordan | 1,419 | Paraguay | 707 | | |

*3. Upper Middle Income Countries*

| | | | | | |
|---|---|---|---|---|---|
| Korea | 19,964 | Malaysia | 4,627 | Panama | 2,368 |
| Lebanon | 246 | Algeria | 14,392 | Brazil | 43,821 |
| Mexico | 42,716 | Portugal | 6,313 | Argentina | 10,506 |
| Chile | 2,066 | Yugoslavia | 5,266 | Uruguay | 1,312 |
| Venezuela | 11,352 | Greece | 5,817 | Hong Kong | 309 |
| Israel | 13,868 | Singapore | 1,318 | Trinidad/Tobago | 659 |

*Source:* World Bank, *World Development Report*, Table 16.

external indebtedness of developing countries and of the distribution of debts among different countries, it might be useful to maintain the three groups as ranked by income *per capita* and analyse them separately before comparing them with one another. Indeed, the pattern of indebtedness may differ among countries with respect to the level of income *per capita* (which itself might be taken as a rough proxy for the level of economic development).

How does one go about analysing each group or set of data? To get a better grip on the data it is useful to systematise the information with respect to different characteristics. Just glancing at them and perhaps picking out some exceptional cases (Brazil, Mexico, India etc.) should be avoided, and so should resorting to mean values as the only way to summarise the data. Generally, when looking at a single set of observations of one variable, the researcher is interested in its *level* (or average), its *spread* (or variation), its *shape* (or sample distribution) and the pattern of *clusters* and *outliers*. Each reveals one aspect of the data.

### *Level*

In exploratory data analysis it is preferable to use the median rather than the mean to average the sample. Several reasons may be given for this preference. First, the median is more resistant than the mean with respect to outliers and when the sample is of poor quality. In the latter case, the middle values of the sample may be more reliable than the extreme ones. The mean, which is calculated using all values in the sample, is much more sensitive to the presence of outliers and freak values. In small samples this lack of robustness can be very misleading. Second, when the underlying distribution is unknown, interpreting the mean becomes more difficult. If the unknown distribution is *symmetrical*, then the mean represents the centre of the distribution. If, furthermore, it is normal (or nearly so), then the mean represents a centre around which the observations tend to gravitate: i.e. the further the distance from the mean, the fewer the number of observations. If, however, the underlying distribution is strongly asymmetrical, then interpreting the mean is much less straightforward and becomes much less meaningful. On the other hand, interpreting the median is always straightforward inasmuch as it constitutes the *middle value* of a sample. If the unknown distribution is symmetrical, it will also represent the centre of the distribution. Finally, when comparing different samples with one another, one cannot assume that they all necessarily have underlying distributions of a similar *shape*.Using the means as reference points in the comparison can often lead to nonsensical interpretations, since the location of the means depends on the shape of the sample distribution.

### *Spread*

As was the case with the mean, the variance – or its square root, the standard deviation – is not very suitable for exploratory data analysis. It is highly sensitive to outliers, since it squares the deviation of each observation from the mean, and hence amplifies the weight of extreme observations. It is, therefore, not resistant. Furthermore, when the shape of a distribution is unknown, its interpretation becomes difficult and comparisons among different samples can lead to very misleading inferences. Preferable is the *midspread* (= interquartile range), which is resistant and whose interpretation is always straightforward. The midspread provides the *range* of the middle 50% of the observations in the sample.

### *Shape*

This aspect is often neglected by many researchers, even though it is of crucial importance for applied work. Shape not only indicates a quantitative dimension of the problem but also incorporates qualitative characteristics which beg for further investigation. For example, pronounced asymmetry in the sample distribution when it is not expected leads one to question the underlying economic or social characteristics of which the numerical data are only a quantitative dimension.

Another important aspect of analysing the shape of sample distributions is of more technical relevance. Many tests or estimation methods used in confirmatory statistics (and especially in classical parametric statistics) are based on clear-cut assumptions about the shape of the population distribution from which the data are drawn. In most cases the underlying distribution is assumed to be normal or at least to approach normality. The test is valid if the assumptions are satisfied, but few researchers take the trouble to verify whether this is indeed the case. In practice, most samples encountered in economic and social research are by no means normal or symmetrical. Nevertheless, often a simple *transformation* of the original data allows the assumptions of the statistical test or estimation technique to be satisfied provided care is taken to analyse the shape revealed by the sample data. Exploratory data analysis helps to improve the standard of confirmatory analysis because one can assess the degree to which its assumptions are satisfied and select the appropriate transformations and/or tests accordingly.

To depict the shape, the tools of exploratory data analysis will be utilised. These include, first, the *stem and leaves diagramme,* which is an improved version of the histogram inasmuch as it keeps the original data intact. Second, use will be made of the *five number summary* of a set of data: the upper and lower extreme values, the upper and lower quartiles and the median. Third, the use of Tukey's *box plot*, a simple graphical representation of the five number summary, will be demonstrated.

### *Clusters and Outliers*

These appear in the patterns and extremes within the overall shape of the distribution of the data. Clusters often indicate sub-groupings within the set of data under consideration. Outliers are not just a statistical nuisance, but may often yield valuable information about the problem under study. To study both characteristics it is necessary to use the stem and leaves diagramme of the data base. To get a rough indication of which observations to consider as outliers, the method suggested by Tukey will be applied. Tukey defines a *step* as one-and-a-half times the midspread, and then considers a value to be an outlier if it is situated more than one step above the upper quartile or below the lower quartile.[2] This, of course, is a rule of thumb which may need to be considered in specific cases. Nevertheless, it provides a good starting-point for defining outliers.

These tools will now be employed to analyse each set of data and subsequently compare the different sets with one another. The reader is reminded that the basic purpose here is *not* to answer questions, but to study the pattern of the data in order to get ideas about the questions to ask about the problem at hand.

## EXTERNAL DEBTS OF LOW INCOME COUNTRIES

The stem and leaves diagramme is an improved version of a histogram in which one can read the original data from the diagramme itself. The stem is equivalent to class widths, and the leaves indicate the frequencies of each class while preserving the individual numbers. To construct such a diagramme for the public debt data on low income countries one must proceed, as in drawing a histogram, by covering the range from the lowest value – US$154 million – to the highest – US$17,975 million. A class width of US$1,000 million will be used, and hence this will constitute the *units* in which the stem is measured. The leaves will then contain 100s, 10s and units. There is a problem here, however, since most observations are less than 2,000 and many fall below 1,000. It seems reasonable, therefore, to halve the class width at the lower end of the scale (below 1,000 and from 1,000 to 2,000). The construction of such a diagramme is shown in Figure 1

| | |
|---|---|
| 17 | 975 |
| 16 | |
| 15 | |
| 14 | |
| 13 | |
| 12 | |
| 11 | |
| 10 | |
| 9 | |
| 8 | 814 |
| 7 | |
| 6 | |
| 5 | |
| 4 | 807 |
| 3 | 850,960 |
| 2 | |
| 1 | 639,585 |
| 1 | 476,255,258 |
| 0 | 792,738,685,540,877,549,605,860,979 |
| 0 | 201,234,154,296,172,360,213,346 |

Stem: 1,000s
Leaves: 100s, 10s, units
Number of observations: 27

*Figure 1. Stem and Leaves Diagramme for Low Income Countries (in US$ millions)*

and should be examined while following the remainder of this explanation. To begin with, a vertical line is drawn to separate the stem (left of the line) from the leaves (right of the line). The *class width* represents the unit in which the stem is measured. In this case the unit is 1,000, but as explained above, this is halved at the lower end of the scale. The stem covers the total range from 17 down to 0, but 1 and 0 are each recorded twice to indicate that the class width is halved at the lower end. It is now possible to fill in the leaves (= the frequencies). The first observation is known from Table 1: Chad, with a US$201 million debt. Where will it be recorded on the stem? Since it is below 1,000 and below 500, the leaf 201 (= 100s, 10s and units) is entered to the right of the vertical line next to the lowest 0. Next comes Bangladesh, with a debt of US$3,850 million: the leaf 850 is entered to the right of the line next to the stem value 3 (1,000s). The figure for Ethiopia is US$792 million (< 1,000 but > 500), so the leaf 792 is entered next to the upper 0 on the stem. Then comes Nepal, with a US$234 million debt. The stem is the lower 0, and its leaf is entered next to the leaf for Chad by separating them with a comma. Continuing in this way, the shape of the histogram appears in the leaves which correspond to the classes enumerated in the stems. It differs from the usual inasmuch as it appears to be turned on its side. A much more important difference, however, is that the stem and leaves diagramme makes it possible to *easily see the original data*, while the histogram loses that information. In the diagrammatical representation the original values can actually be read off from the diagramme by combining stem and leaves.

What does the diagramme reveal? Clearly, the sample distribution is far from being symmetrical: most values are situated at the lower end of the scale, and the distribution then strays upwards with fewer and fewer observations. As a point of interest the countries with extreme values in the diagramme – India, Pakistan, Sudan, Bangladesh and Zaire – may be noted. As has already been mentioned, the lack of symmetry which they create is quite common in social and economic data. One reason for it is that there is often a clear 'floor' in the diagramme of the data (i.e. public debt = 0) but no 'ceiling'. Furthermore, not all countries face the same conditions in their domestic economic situations as they face in terms of access to loans. This point needs further investigation, and often it may be instructive to see why the extreme points differ from the general cluster by looking into specific cases (India, obviously – but also Zaire, for example, and Sudan which are not so large in terms of population as India). Hence, looking at the diagramme and identifying some of its values may help in selecting interesting cases (both those representative of the majority of cases and also the more exceptional ones). Sudan, Zaire and Bangladesh form a *little* cluster: is there any reason for this in terms of similarities in their external payments situation, or is it a coincidence? India stands quite apart from all the other countries in the group, but the size of its population also vastly overshadows theirs. A low income *per capita* in India still implies a large national income relative to the other poor income countries. Questions such as these arise from the diagramme: some may appear to make sense and be worth following up, while others may be rather coincidental. Clearly, one needs to be familiar with the wider context to make sense of these patterns. And this is exactly what the diagramme is used for: to help focus one's attention.

Next, the five number summary will be discussed: this consists of upper and lower extreme values ($X_u$ and $X_l$), upper and lower quartiles ($Q_u$ and $Q_l$) and the median (MD).[3] These data are recorded in Table 2 along with the midspread (MS). Figure 2 shows the corresponding *box plot*: the interquartile range is given by the 'box' in which the location of the median is indicated by a dividing line in this rectangle. The upper and lower extreme values are plotted on the same scale and connected with the box by straight lines. The result is a graphic representation of the major characteristics of shape.

*Table 2. Five Number Summary for Low Income Countries*

| | | |
|---|---|---|
| $X_U$ | 17,975 | (India) |
| $Q_U$ | 1,585 | (Sri Lanka) |
| MD | 792 | (Ethiopia) |
| $Q_L$ | 346 | (Sierra Leone) |
| $X_L$ | 154 | (Burundi) |
| MS | 1,239 | |
| Step | 1,859 | |

Outliers: India, Pakistan, Sudan, Bangladesh, Zaire

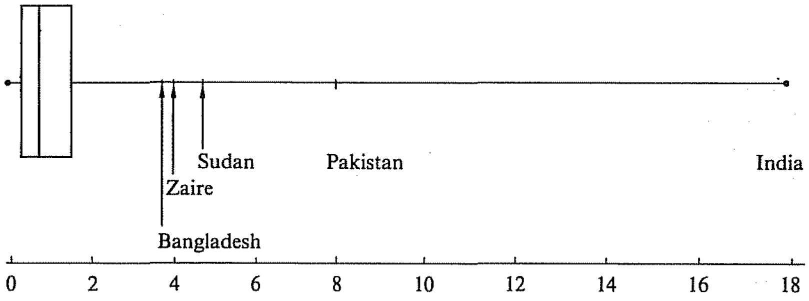

*Figure 2. Box Plot for Low Income Countries*

One can see that the asymmetry in the sample distribution is not merely a result of the outliers at the upper end. Indeed, the median is much closer to the lower quartile than to the upper one, reflecting the fact that the whole distribution straggles upwards. The spread is quite large relative to the level (MS/MD ≅ 1.6). As was already noted, the spread is uneven in both the upward and the downward directions: there is more variation at the upper end, but fewer values. According to the criteria described above for determining outliers, India and Pakistan are *far* outliers (two steps above the upper quartile), and Sudan, Bangladesh and Zaire are normal outliers (one step above $Q_u$).

Finally, to give the reader a feeling for the sensitivity of the arithmetic mean and of the standard deviation to outliers and asymmetry in shape, tabulated below are the means and standard deviations of this sample and two of its sub-samples (tabulated by leaving out far outliers), and they are compared with the behaviour of the median and of the midspread.

| *Sample* | *Mean* | *Median* | *St. Dev.* | *Midspread* |
|---|---|---|---|---|
| n = 27 (all values) | 2,045 | 792 | 3,709 | 1,239 |
| n = 26 (excl. India) | 1,433 | 765 | 1,940 | 1,130 |
| n = 25 (excl. India and Pakistan) | 1,137 | 738 | 1,249 | 909 |

Where level is concerned, leaving out the value for India implies that the mean drops from 2,045 down to 1,433, while the median is hardly affected. Excluding both India and Pakistan from the sample yields a mean of 1,137, or 56% of the mean of the total sample, while the median remains rather close to the overall median. The effect of leaving out India is even more dramatic when spread is examined. Without India the standard deviation drops from 3,709 to 1,940 (i.e. a decrease of nearly 50%), and when Pakistan is excluded as well it drops to 1,249 (or one-third of the original standard deviation of whole sample. The movement of the midspread is much less pronounced. These results confirm the danger of relying only on the mean and the standard deviation to summarise a sample. When outliers are present and the shape of the distribution is unknown, the results can become highly unreliable. In this sample, the mean is located above the upper quartile and provides little information about the centre of the distribution.[4]

## THE EXTERNAL DEBT OF LOWER AND UPPER MIDDLE INCOME COUNTRIES

Having illustrated the basic approach for analysing a set of data, it is now possible to move on more rapidly to the cases of lower and upper middle income countries. It will be left to the reader to pose questions about the patterns: only their characteristics will be noted in the text.

| | |
|---|---|
| 15 | 529 |
| 14 | |
| 13 | 887,809 |
| 12 | |
| 11 | |
| 10 | |
| 9 | |
| 8 | |
| 7 | 388,879 |
| 6 | |
| 5 | 169,974,123 |
| 4 | 652,497 |
| 3 | 392,171 |
| 2 | 228,422,294,034,246,337 |
| 1 | 094,223,975,105,434,260,419 |
| 0 | 944,827,640,592,107,664,613,880,684,707 |

Stem: 1,000s
Leaves: 100s, 10s, units
Number of observations: 35

*Figure 3. Stem and Leaves Diagramme for Lower Middle Income Countries (in US$ millions)*

Figure 3 shows the stem and leaves diagramme for the lower middle income countries. In the case of these countries it was not necessary to halve the stem at the lower end of the scale, since this distribution is not located as close to the 'floor' of the diagramme as it was for the lower income countries. The distribution definitely manifests asymmetry, as is shown especially by the 'tail' at the upper end. It also appears more spread out in the middle, although its upper extreme value does not trail off as much as was noted for low income countries. Moreover, its level or centre is definitely higher compared with that of the low income countries.

Table 3 and Figure 4 show the five number summary and its corresponding box plot. The box plot quite clearly reveals the asymmetry, both within the box and the extremes. Indonesia, Egypt and Turkey constitute

*Table 3. Five Number Summary for Lower Middle Income Countries*

| | | |
|---|---|---|
| $X_U$ | 15,529 | (Indonesia) |
| $Q_U$ | 4,652 | (Nigeria) |
| MD | 2,034 | (Cameroon) |
| $Q_L$ | 880 | (Zimbabwe) |
| $X_L$ | 107 | (Lesotho) |
| MS | 4,545 | |
| Step | 6,818 | |

Outliers: Indonesia, Egypt, Turkey

outliers relative to the main clustering in the sample. The median and midspread are both higher compared with those of the low income countries. This aspect will be considered further in the next section, when the three groups are compared.

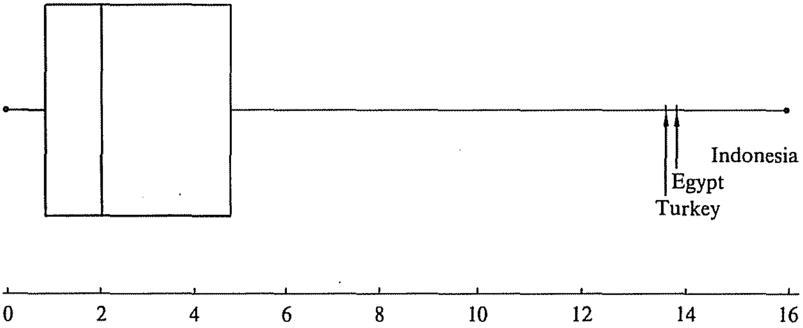

*Figure 4. Box Plot for Lower Middle Income Countries*

If the mean and standard deviation are computed, the values of 3,434 and 3,947, respectively, are obtained. The mean is once again pulled upwards relative to the median, reflecting the asymmetrical shape. The standard deviation appears much less inflated, and in this case it is actually lower than the midspread. In fact, looking back at Figures 1 and 3 quickly reveals why. For the low income countries the asymmetry in the distribution was much more pronounced (it was located on the floor of the dia-gramme), and the outliers were much more distant; as a result, the standard deviation became heavily inflated. Hence, while the midspread in the second case is much greater than that in the low income countries, the standard deviations of both cases are approximately *equal*! Nevertheless, the nature of the variation is by no means similar.

Here again is an example of why blind faith can be very misleading in using mean and standard deviation to summarise data. If one did so, one would conclude that both samples have approximately equal variations but different levels. Using median and midspread, the conclusion is that both the level and the spread of the second sample are greater than those of the first one.

| | |
|---:|:---|
| 40;41;42;43;44 | -;-;716;821 |
| 35;36;37;38;39 | |
| 30;31;32;33;34 | |
| 25;26;27;28;29 | |
| 20;21;22;23;24 | |
| 15;16;17;18;19 | -;-;-;-;964 |
| 14 | 392 |
| 13 | 868 |
| 12 | |
| 11 | 352 |
| 10 | 506 |
| 9 | |
| 8 | |
| 7 | |
| 6 | 313 |
| 5 | 266,817 |
| 4 | 627,423 |
| 3 | |
| 2 | 368 |
| 1 | 312,318 |
| 0 | 246,309,659 |

Stem: 1,000s
Leaves: 100s, 10s, units
Number of observations: 18

*Figure 5. Stem and Leaves Diagramme for Upper Middle Income Countries (in US$ millions)*

The stem and leaves diagramme for upper middle income countries is depicted in Figure 5. In the case of these countries, since the range is larger the stem at the upper end of the scale is grouped by separating the corresponding stem and leaves values by semi-colons. Hence, in Figure 5 the observations at the top read 43,716 and 44,821, respectively, and the next value below them is 19,964. This is equivalent to increasing the class width in a histogram, but by using semi-colons the original data is preserved within the diagramme. The sample is smaller, but a quick glance shows that the level appears to have moved upwards (relative to the other two cases) and that the data are also more spread out. Again, this distribution is by no means symmetrical, but straggles upwards towards some quite extreme values (Brazil and Mexico). The data are more strung out and form some clusters within a wide range of variation. Such a distribution projects a more heterogeneous pattern, certainly in comparison with the low income countries.

*Table 4. Five Number Summary for Upper Middle Income Countries*

| | | |
|---|---|---|
| $X_U$ | 43,821 | (Brazil) |
| $Q_U$ | 13,868 | (Israel) |
| MD | 5,542 | (Yugoslavia/Greece) |
| $Q_L$ | 1,318 | (Singapore) |
| $X_L$ | 246 | (Lebanon) |
| MS | 12,550 | |
| Step | 18,825 | |

Outliers: Brazil, Mexico

Table 4 and Figure 6 present the five number summary and its corresponding box plot. Level and spread are indeed greater than in the previous cases. The asymmetry is clearly shown by the box plot, both within the box and in its extremes. Brazil and Mexico constitute the outliers (but are not far outliers). In this case as well, the arithmetic mean (= 10,515) is much larger than the median, indicating strong asymmetry in the data. The standard deviation is not as inflated as in the first case, which is due to the fact that

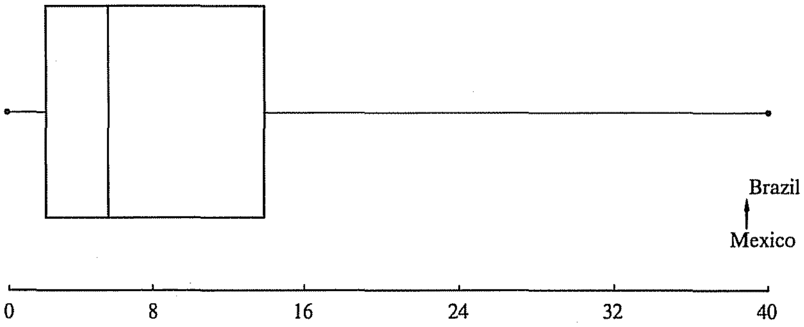

*Figure 6. Box Plot for Upper Middle Income Countries*

there are no far outliers here. The value of the standard deviation is 13,159, as compared with a midspread value of 12,550.

## COMPARING SAMPLES

Having studied each sample on its own, let us now turn to comparing them with each other. An interesting feature of the data is that level and spread tend to move together. To demonstrate this tendency median and midspread are compared below for the three samples in the tables:

| | *LI* | *LMI* | *UMI* |
|---|---|---|---|
| Median | 792 | 2,034 | 5,542 |
| Midspread | 1,239 | 4,545 | 12,550 |

The *covariance* of level and spread becomes quite noticable in these data. An even more vivid impression can be obtained by comparing the box plots for three samples by relating them to the same scale. This is done in Figure 7. The boxes inflate quite considerably as one moves from the sample for low income (LI) countries to that for the upper middle income (UMI) ones. It should also be noted that in this comparison, India's position appears to be much more extreme, relative to its group, than any of the other extreme values. In this respect, it is important to look not only at the absolute distance but also at the distance relative to the midspread.

The fact that level and spread covary may appear strange at first, but it is very common in economic and social data. In economic analysis, this feature is most often found in cross-section analysis (as opposed to time series analysis). For example, a researcher familiar with consumer budget studies is well aware that there is more variation among high income earners than among poorer families. The reason for this is straightforward: poor families generally spend their entire incomes on basic necessities, and hence there is relatively little variation in their consumption behaviour. Richer families have much more scope in their consumption pattern, which implies not only that their consumption levels are higher but also that there is more variation among them. Similarly, if an industrial survey is carried out it is likely that more variation will be found among large firms than among smaller enterprises. Larger firms have more resources, their options are more varied, and therefore level and variance are higher.

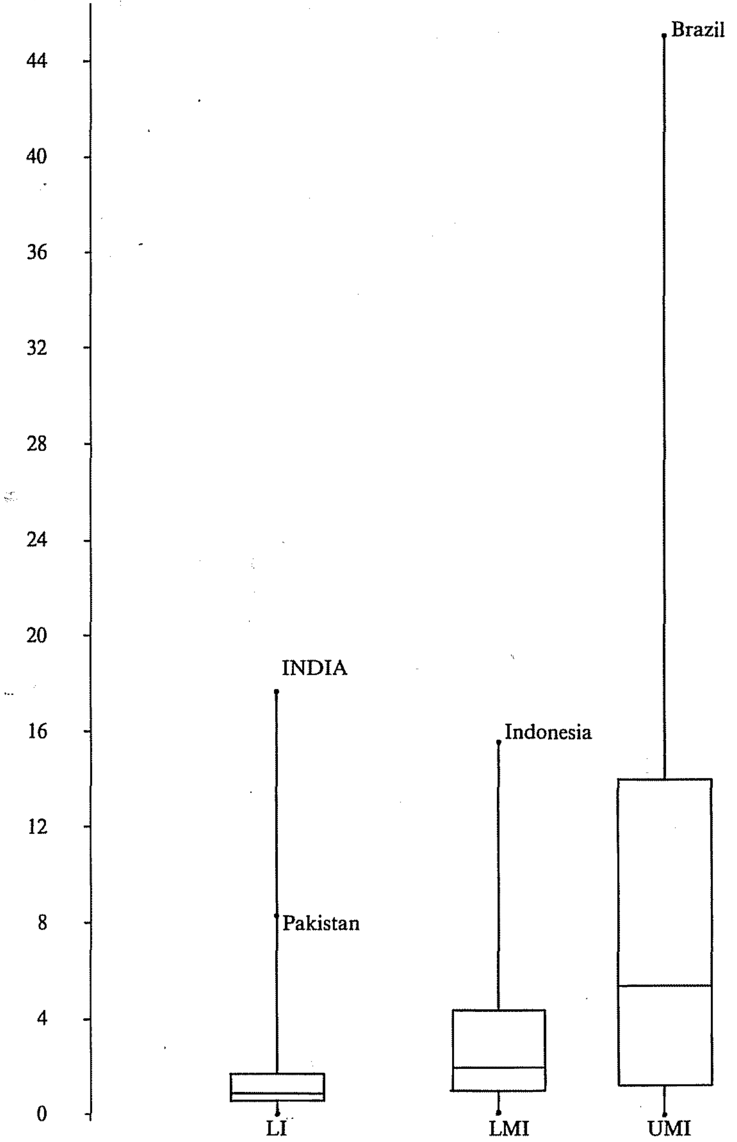

*Figure 7. Comparison of the Box Plots of the Three Groups of Countries (in US$ millions)*

In the World Bank data chosen as an example here, the covariance of level and spread is similarly not so surprising. Countries with low incomes *per capita* generally have a weaker economic base: their position in the international division of labour often depends on a few export crops; their industrial development is rather limited; and it is constrained by foreign exchange shortages. As such, their capacity to borrow is much lower and less heterogeneous. In this respect, the far outlier presents an interesting exception. India's income *per capita* is low, but the sheer size of its economy places it in a different category altogether. In terms of GDP and the size of its population it vastly overshadows all the other low income countries with the exception of China (which is not included in the sample for lack of data). Its options and capacities when external loans are needed are, therefore, out of reach for other low income countries. On the other hand, at the upper end of the scale, the capacity to borrow is likely to be much higher since the countries appearing there have more developed economic bases. It is also likely that there will be more variation depending on the economic situation of the respective countries.

In applied econometrics, the covariance of level and spread constitutes a specific case of *heteroscedasticity* which is common in cross-section data. An econometrician who comes across this often transforms his or her original data in order to eliminate this 'scaling up' effect which level exerts on variance. In exploratory data analysis, one often also uses transformations of the data to effect a change in the shape of the distributions being compared. The reasons for such transformations may be varied and will not be discussed here in detail. In the example taken here, however, the usefulness of this type of transformation will be explored in connection with further analysis of the patterns inherent in the data.

To do this, let us start with a simple question: it is known that spread and level covary, but by how much? If level goes up, does spread increase *proportionally* with level or not? There are various ways to check this relationship; here a method suggested by Tukey will be used. First, the logarithms of the medians and the midspreads of the samples are computed:

| | *LI* | *LMI* | *UMI* |
|---|---|---|---|
| Log median | 2.8987 | 3.3084 | 3.7437 |
| Log midspread | 3.0931 | 3.6575 | 4.0986 |

Next, the log of the midspread is plotted against the log of the median for each sample. This is shown in Figure 8. The three points clearly line up quite well. This indicates that the relationship between level and spread is fairly *stable*. That is, spread increases with level from the LI sample to the LMI sample in much the same way as it does from the LMI sample to the UMI sample. Hence, the relationship between spread and level appears rather systematic and stable.

Then the two extreme points are connected by a straight line, and the slope of that line is computed. Note that the slope = 1.19. Now it can be shown that if the slope equals one, spread is *proportional* to level. If the slope is less than one, spread increases at a slower rate than level as the level rises, and *vice versa* when the slope is greater than one. In the present case, therefore, spread tends to change at a rate that is slightly more than proportional with level. For practical purposes one could say that the slope is approximately one.

It is now clear that one can ascertain whether the covariance between spread and level is stable from sample to sample, and how *strong* the covariance is (proportional, less than proportional or more so). What does this have to do with transformation of the data? In fact, once the degree of covariance is known as measured by the slope, one can assess how the data could be 'scaled down' so as to eliminate the covariance of spread and level.

Why, however, would one want to eliminate this covariance? It would certainly not be because one wishes to get rid of it so that it could be forgotten after having transformed the data! As has been observed, the covariance of spread and level is an important feature of the data and raises many theoretical questions about the problem being studied. While neglecting this would weaken the analysis, what is needed here, however, is something different. Having noted the presence of covariance of spread with level and having analysed its stability and degree, let us take it out of the data by transforming them so as to see whether new elements may be discovered in the patterns in them. To do this it is necessary to remove the scaling up effect, which is so striking that it continuously catches one's eye. In this way some other features may be seen more clearly as well.

How does one scale down the data in order to eliminate the covariance of spread and level? Data can be scaled down in many ways: the most common are the square root operator, logarithms, negative reciprocals $-1/x$ (with a minus sign to preserve the order) etc. Square roots scale down less than logarithms, and logarithms less than negative reciprocals, etc. In deciding which one to use, the value of the slope of the line drawn in Figure

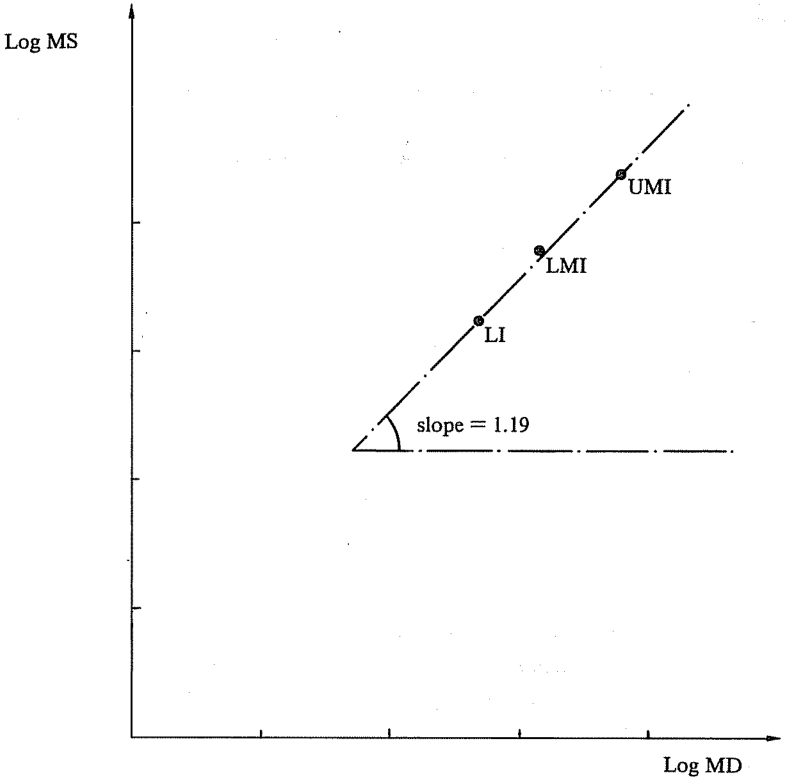

*Figure 8. Plotting Logs of Midspreads against Logs of Medians for the Three Groups of Countries*

8 is quite useful. Depending on the value of this slope, Tukey suggests the following transformations:

| *Slope* | *Transform with* |
|---|---|
| ±0.5 | $\sqrt{}$ |
| ±1 | log |
| ±1.5 | $-\frac{1}{x}$ |

Hence, in this case it was decided to try with logarithms.[5]

## TRANSFORMING THE DATA

To get an idea of the impact of transforming the data by using logarithms one need not start by looking at the logarithms of all the data. It is quicker to transform the five number summary for each sample and then check the effect of the transformation on shape and on spread. This effect is presented in Table 5, while Figure 9 shows the box plots of the three samples. As is seen in the table, the midspread of the new data diverges much less between samples, although it increases from sample to sample. The reason is that the slope of the line in Figure 8 equals 1.19. which is greater than one. Hence, the log transformation will not wholly eliminate the covariance of spread with level. Nevertheless, it has been scaled down considerably.

*Table 5. Five Number Summaries of Transformed Data for the Three Groups of Countries*

| | *LI* | *LMI* | *UMI* |
|---|---|---|---|
| $X_U$ | 4.25467 | 4.19114 | 4.6417 |
| $Q_U$ | 3.20003 | 3.66764 | 4.1420 |
| MD | 2.89873 | 3.30835 | 3.7437 |
| $Q_L$ | 2.53908 | 2.94448 | 3.1199 |
| $X_L$ | 2.18752 | 2.02938 | 2.3909 |
| MS | 0.66095 | 0.72316 | 1.0221 |
| Outliers: | India | none | none |

The box plots reveal much more symmetry in the transformed distributions. There is even a slight tendency in those of the lower and upper middle income countries to straggle downwards. With the exception of India,

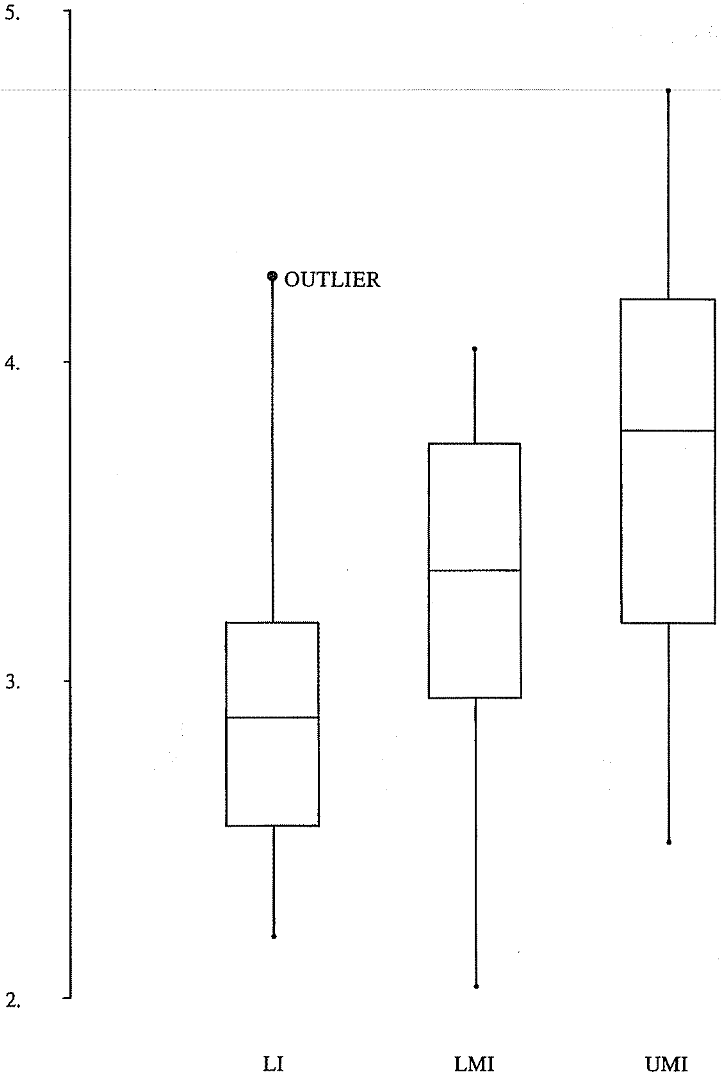

*Figure 9. Comparison of Box Plots with Log Transformed Data*

there are no longer any outliers: i.e. after taking account of the fact that variation increases with level, many former outliers no longer appear as unusual as they seemed before, though India remains the exception. Even if one bears in mind that variation increases at the upper end within each distribution with different levels, India is still very clearly an outlier, while Brazil and Mexico do not appear to be. The importance of the size of India's population is confirmed here. In fact, in the World Bank's publications of data on low income countries, India and China are normally separated from the other countries when computing totals or averages, which clearly shows recognition of their outlier status.

The respective sample distributions of the transformed data have become much more symmetrical and nearly free of outliers. What happens if the arithmetic mean and the standard deviation for the transformed data are computed? In Table 6 the values obtained for the means and the standard deviations of the transformed data are tabulated, and the anti-logs of the means of the transformed data are compared with the medians and the means of the original data.[6] This table provides some interesting insights. First, the log transformation could be deduced from transforming the five number summary only. The standard deviations of the transformed data are rather close to each other. With symmetrical distributions and with only one outlier (which is no longer a far outlier), the information in these standard deviations is now reliable. In the absence of the outlier, the mean of the first sample would be 2.8964 and its standard deviation 0.466. Comparing the latter with the standard deviations of the other samples also shows a gentle rise in the standard deviations, but they are of quite the same order of magnitude.

*Table 6. Comparative Analysis of Original and Log-Transformed Data*

| | *LI* | *LMI* | *UMI* |
|---|---|---|---|
| 1. Means of logs | 2,9467 | 3,3007 | 3,6588 |
| 2. Standard deviations of logs | 0,5265 | 0,4701 | 0,6630 |
| 3. Anti-log means of transforming data | 885 | 1,999 | 4,558 |
| 4. Medians of original data | 792 | 2,034 | 5,542 |
| 5. Means of original data | 2,045 | 3,434 | 10,515 |

Moving now to the bottom half of Table 6, one sees that the anti-log of the means of the transformed data corresponds quite closely to the medians of the original data. In fact, if India is removed from the first sample the anti-log of the mean of the transformed data becomes 788. Note, however, that neither of these measures of level correspond to the means of the original data, both in their order of magnitude and in their general movement from sample to sample.

At this juncture it is useful to remind the reader that none of the results are obtained mechanically. Transforming samples by using logarithms is not a safeguard against wrongly interpreting the patterns in the data. Whether a mean is meaningful depends on the context provided by the data, i.e. the shape of its distribution and the presence or absence of outliers. Whether a transformation is meaningful depends on careful examination of the patterns revealed by the data. One is not dealing here with a box of magic tricks, but with methods of looking at data in order to discern patterns within them which may give rise to questions about the problem being analysed. All too often one reads articles on empirical analysis in which the author transforms the data without showing much understanding of them. Frequently, the conclusions drawn are products of the misspecified transformation rather than insights into the problem at hand. At the other extreme, many social scientists do not analyse their data: they often suffer from an extreme degree of 'dataphobia' and merely use data to illustrate the text – as paintings on the wall.

## MOVING ON TO HYPOTHESIS TESTING

The preceding sections tried to show how exploratory data analysis can help in gaining a better understanding of the World Bank data that was taken as an example here. By itself, it does not produce theoretical insights. It assists in raising questions and putting these in the wider context of the isssue being studied. Some of the patterns that were noted may confirm some prior hunches and others may be puzzling, since they contradict what was expected. As such the empirical analysis comes to take an active part in furthering the analysis of the problem.

Yet, exploratory analysis is also important for testing hypotheses and for estimation. Statistical tests can be powerful instruments in analysis, since they make it possible to reject or maintain hypotheses and estimate numerical parameters which characterise the distributions or relationships hypothesised. The power of a test, however, depends on whether its assumptions are satisfied *in practice*. Some tests or estimation techniques can be fairly robust – i.e. even if some of their underlying assumptions are not strictly satisfied, they continue to yield reliable results. That does not mean, however, that they can be applied under any circumstances. It is still necessary to find out whether the assumptions of the test and/or estimation technique are reasonably satisfied, and since this is often not known *a priori*, one has to check the samples to see whether the assumptions are likely to be satisfied. For example, if a test requires that the underlying population be normal, one cannot apply it to samples which clearly show pronounced asymmetry.

Let us return to the World Bank case study to illustrate this point. Suppose a researcher wants to test the hypothesis that 'the capacity to borrow externally is much less for countries with lower *per capita* incomes'. This hypothesis is quite plausible since one often reads – in both scientific journals and the newspapers – that not all the developing countries were equally strongly affected by the world economic crises of the 1970s and 1980s. The worsening terms of trade and the growing foreign exchange crisis slowed down growth in the poorer countries, which have limited capacities to borrow their way out of the problem, but the developing countries with higher incomes have had more scope for maintaining their growth rates in that they have had more opportunities to borrow. Hence, a researcher may wish to test whether the levels of external public debts are higher for countries with a higher income *per capita*. The appropriate technique for testing such a hypothesis would be to apply the analysis of variance with respect to the three categories of developing countries: i.e. one would test whether the differences between the means of the three groups are significant.

Analysis of variance is, however, based on the assumption that the population distributions in the various groups of countries are normal and, furthermore, that they have equal variances. It can be shown that the analysis of variance is a robust technique, which means that its conclusions remain valid even if not all of its hypotheses are fully satisfied. Hence, if the underlying population distributions are reasonably symmetrical (and somewhat approaching a bell shape), and if the variances do not diverge widely, applying this technique will yield reliable results. In fact, the technique has a good reputation for robustness when faced with deviations from normality in the underlying distributions. This does not mean, though, that it can be applied to any form of distribution, however asymmetrical it may be.

In this excursion into the exploratory data analysis of the data on external debts, it was suggested that the underlying pattern in the distributions seems to be far from symmetrical. For example, the sample of the low income countries clearly revealed a 'floor' but also had some far outliers. Furthermore, although it seemed that the standard deviations were equally large in the first two samples (but not in the third), the analysis suggested that the real patterns of variation in the three samples are actually different. The conclusion must therefore be that the assumption underlying the analysis of variance method are not likely to be satisfied.

This does not necessarily imply, however, that we cannot rely on this technique at all. The analysis also showed that transforming the original data by using logarithms resulted in much more symmetrical distributions that had sample variances which were more alike in terms of the order of their magnitude. Hence, applying the analysis of variance method to the transformed data would be a valid procedure.

Should one therefore proceed by applying variance analysis to the three samples in order to check the significance of the differences between the means in the transformed data? Although this paper is not concerned with confirmatory statistics as such, this question raises an important issue about the relationship between exploratory data analysis and formal hypothesis testing.[7] Can one use the set of data which served as a basis for exploration when one tests hypotheses?

Clearly, doing so would present the danger of pure 'data mining'. First, the patterns in a set of data would have to be carefully diagnosed, and this information would then be used to test hypotheses against that same set of data! *Generalising* conclusions derived from such results would definitely not be a valid procedure.

Hence, the correct way to proceed would be to select a new set of data (carefully sampled to assure randomness) and formally test the hypothesis with respect to this new set of data. In socio-economic research, however, the issue is not always so easy to resolve. More often than not one cannot sample repeatedly, and therefore the available data basis may be rather restricted. Economic research offers little scope for data analysis based on designed experiments which can be repeated whenever necessary.

Clearly, whenever data can be sampled in order to test a hypothesis – which itself is derived in part from careful exploration of existing data sources – this should be done. In the specific case at hand, one could solve the problem by checking the pattern in external debts of developing countries over a series of years and by testing the hypothesis of different means on the transformed data for each year.

In the present case the data base could be enlarged by looking at different years, but in other cases there would be no scope for enlargement. This would happen particularly when time series data were analysed, since exploratory analysis and confirmatory statistics would converge on the same data base and the danger of incorrect inferences would thus be much the greater.

It is not the purpose of this paper to address this specific topic in any more detail. Suffice it to say that within the econometrics of time series data, the need for diagnostic testing of the stability of the estimated model and of its prediction performance are important tools in avoiding incorrect inferences, such as those described above.

## HETEROSCEDASTICITY AND REGRESSION ANALYSIS

In the foregoing analysis the pattern of the external public debts of developing countries has been investigated by grouping them into three broad categories based on their incomes *per capita*. It was found that level and spread tend to increase as one moves from the low income group to the higher ones. It was also observed that the size of the population is significant: i.e. it is not just income *per capita* but also the absolute level of income.

To account for size one could now proceed by investigating the distributions of the external debts *per capita* for the three groups of countries. However, it will be left to the reader to explore the data on debts *per capita* for each of the three groups. Instead of taking that course, let us proceed in a somewhat different manner which leads to similar results.

In the discussion above, countries were grouped into three categories, each of which was defined by a given range in which the income *per capita* is situated: e.g. low income countries had incomes *per capita* of US$400 or less. As such, the analysis consisted of exploring the distribution of a numerical variable (= external debt) for different categories (= ordinal variable). Income *per capita*, however, is a numerical variable as well, and hence it is possible to explore the relationship between two numerical variables: debt *per capita* and income *per capita*. Regression analysis is the appropriate technique to deal with this type of relationship.

It is proposed, therefore, to estimate the following relationship:[8]

$$D_i = a + b \cdot I_i + e_i$$

where:

$D_i$ = external public debt *per capita* of country i
$I_i$ = income *per capita* of country i
$e_i$ = unexplained residual for country i

Estimating the regression line yields the following results:

$$D_i = \underset{(55)}{103} + \underset{(0.031)}{0.183}\, I_i$$

|  |  |  |
|---|---|---|
| Standard error of residuals |  | = 361.5 |
| t-statistic of | a | = 1.85 |
|  | b | = 5.84 |
| F-statistic (1,77) | F | = 34.11 |

$R^2 = 0.31$
DW= 2.48

Clearly, the coefficient of $I_i$ is highly significant, and $I_i$ accounts for about 30% of the total variation in $D_i$. In this case one would not expect the coefficient of determination to be much higher, because there are considerable differences among countries with similar income levels and population sizes with respect to their external indebtedness. Saying that income level is significant is not the same as saying that it explains all there is to explain.

Too many researchers, however, tend to interpret regression results only on the basis of the values of the $R^2$. In fact, when *exploring* a problem it is much more interesting to look at the pattern of residuals. This pattern can provide hints about the direction in which to look for further explanations.

What would be a useful way to look at residuals? A good way to start is to make a stem and leaf diagramme of them so that outliers, asymmetry etc., can be spotted, as shown in Figure 10.[9]

As can be seen, the residuals are scattered asymmetrically, and this is true of both the middle values and the outliers. The fact that the mean (=0) is higher than the median confirms this asymmetry. An interesting feature that is highlighted by this scatter is that all the outliers (far and normal)

*Table 7. Data Base for the Regression Line $D_i = a + b\, I_i + e_i$ (in US$)*

| *Country* | $D_i$ | $I_i$ |
|---|---|---|
| Chad | 44.7 | 110 |
| Bangladesh | 42.5 | 140 |
| Ethiopia | 24.8 | 140 |
| Nepal | 15.6 | 150 |
| Burma | 48.1 | 190 |
| Mali | 107.0 | 190 |
| Malawi | 110.5 | 200 |
| Zaire | 132.9 | 210 |
| Uganda | 41.5 | 220 |
| Burundi | 36.7 | 230 |
| Upper Volta | 47.0 | 240 |
| Rwanda | 32.5 | 250 |
| India | 26.0 | 260 |
| Somalia | 199.3 | 280 |
| Tanzania | 77.3 | 280 |
| Guinea | 224.1 | 300 |
| Haiti | 70.6 | 300 |
| Sri Lanka | 105.7 | 300 |
| Benin | 152.5 | 320 |
| Central African Republic | 88.8 | 320 |
| Sierra Leone | 96.1 | 320 |
| Madagascar | 139.8 | 330 |
| Niger | 106.1 | 330 |
| Pakistan | 104.3 | 350 |
| Sudan | 250.4 | 380 |
| Togo | 318.5 | 380 |
| Ghana | 83.0 | 400 |
| Kenya | 128.0 | 420 |
| Senegal | 160.0 | 430 |
| Mauritania | 516.9 | 460 |
| Yemen Arab Republic | 149.9 | 460 |
| Yemen Popular Democratic Pepublic | 320.0 | 460 |
| Liberia | 311.6 | 520 |
| Indonesia | 103.9 | 530 |
| Lesotho | 76.4 | 540 |
| Bolivia | 249.5 | 600 |
| Honduras | 321.8 | 600 |
| Zambia | 395.5 | 600 |
| Egypt | 320.7 | 650 |
| El Salvador | 141.3 | 650 |
| Thailand | 107.7 | 770 |

| | | |
|---|---|---|
| Phillipines | 149.0 | 790 |
| Papua New Guinea | 197.7 | 840 |
| Morocco | 377.0 | 860 |
| Nicaragua | 705.4 | 860 |
| Nigeria | 53.1 | 870 |
| Zimbabwe | 122.2 | 870 |
| Cameroon | 233.8 | 880 |
| Congo, People's Republic | 650.0 | 1110 |
| Guatamala | 91.2 | 1140 |
| Peru | 351.4 | 1170 |
| Ecuador | 394.4 | 1180 |
| Jamaica | 651.8 | 1180 |
| Ivory Coast | 529.1 | 1200 |
| Dominican Republic | 225.0 | 1260 |
| Colombia | 194.1 | 1380 |
| Tunisia | 487.8 | 1420 |
| Costa Rica | 976.5 | 1430 |
| Turkey | 303.5 | 1540 |
| Syrian Arab Republic | 251.3 | 1570 |
| Jordan | 417.4 | 1620 |
| Paraguay | 228.1 | 1630 |
| Korea (Republic of) | 513.2 | 1700 |
| Malaysia | 325.8 | 1840 |
| Panama | 1246.0 | 1910 |
| Algeria | 734.3 | 2140 |
| Brazil | 363.7 | 2220 |
| Mexico | 599.9 | 2250 |
| Portugal | 644.2 | 2520 |
| Argentina | 372.6 | 2560 |
| Chile | 391.4 | 2560 |
| Yugoslavia | 234.0 | 2790 |
| Uruguay | 452.4 | 2820 |
| Venezuela | 737.1 | 4220 |
| Greece | 599.7 | 4420 |
| Hong Kong | 59.4 | 5100 |
| Israel | 3467.0 | 5160 |
| Singapore | 549.2 | 5240 |
| Trinidad/Tobago | 549.2 | 5670 |

*Source*: World Bank, *World Development Report*, Tables 1 and 16. All data are for the year 1981.

| | | |
|---|---|---|
| 24 | 20 ISRAEL | $X_U = 2420$ |
| 9 | | |
| 8 | | |
| 7 | 94 PANAMA | |
| 6 | 12 Costa Rica | |
| 5 | | |
| 4 | 45 Nicaragua | |
| 3 | 30,44,33 | |
| 2 | 07,40 | |
| 1 | 83 | |
| 1 | 46,33,14,09,17,25 | |
| 0 | 67,78,99,76,99,85,80 | $Q_U = 80$ |
| 0 | 45,37,37,35,18 | |
| –0 | 30,29,08,09,23,21,30 | |
| –0 | 79,85,89,77,87,52,73,65,57,63,93,52,96,80,98,59,81 | MD = –65 |
| –1 | 04,15,01,08,00,16,24,25,36,40,08,39,14,46,38 | $Q_L = -115$ |
| –1 | 61,73,99,80,67 | |
| –2 | 09,20 | |
| –3 | 79,12 | |
| –4 | | |
| –5 | 13,92 Singapore, Trinidad & Tobago | |
| –6 | | |
| –7 | | |
| –8 | | |
| –9 | 77 HONG KONG | $X_L = -977$ |

| | |
|---|---|
| Stem: | 100s |
| Leaves: | 10s, units |
| MS | 195 |
| Step | 293 |

*Figure 10. Stem and Leaf Diagramme of Residuals of the Regression Line $D_i = 103 + 0.183\ I_i$*

are countries which have not yet shown themselves to be exceptional! Of course one must ask why this is the case.

Some reflection on the meaning of these residuals shows that the answer is simple. Here they represent the residual variation in foreign public debts *after* taking account of differences in income *per capita* and in population size. This difference was taken into account by relating external debt *per capita* to income *per capita* and not to the absolute value of the debt. As a result, many of the former outliers no longer appear as exceptional values.

The new outliers which now emerge, however, are most interesting! This is because these countries have exceptional levels of debt after *discounting* the differences in income *per capita* and in size. 'Positive outliers' are those countries which have high levels of debt relative to their income level and population size. Conversely, 'negative outliers' are those with low levels of debt after removing the influence of income level and population size.

Israel sticks out 'like a sore thumb', but this result should not surprise anyone familiar with the present-day political situation. At the opposite end of the scale is Hong Kong: it will have to be left to the reader, however, to figure out the possible reasons for this and other outliers, and also to question whether certain patterns can be inferred from their presence. One problem that remains in interpreting these residuals is that a relatively low level of debt can mean two things: either the country in question has little borrowing capacity (it needs financing but cannot get it), or its economic situation is such that it does not need as much finance (and possibly could get it if it wanted it).

There is another striking feature about these outliers: the far outliers (on both sides) are all upper middle income countries, and the other ones are either from that group or among the lower middle income countries. None of the outliers is a low income country! This provides a definite hint as to what patterns might be inferred.

In the course of this analysis of the three samples, the covariance of level with spread was noted. The regression analysis confirmed that income *per capita* is significant in explaining differences in the level of external debt *per capita*. It is, however, quite likely that level also moves together with spread in this case, and hence the residuals would be more spread out at the upper end (i.e. the higher levels of income *per capita*) than at the lower one. This would explain why the outliers at the top and the bottom mainly consist of upper middle income countries.

How does one check this? One method is to plot the residuals in the regression against income *per capita* and then study the pattern obtained. Since there are 79 observations, however, the resulting graph becomes rather cumbersome, and it will not be reproduced here. Instead the residuals will be grouped into three groups corresponding to the low income, lower middle and upper middle income countries. This makes it possible to refer back to the analysis above. In Figure 11 the overall stem and leaves diagramme of the residuals is broken down into these groups.

The resulting pattern is quite interesting but not surprising. Spread

| LI | | LMI | | UMI | |
|---|---|---|---|---|---|
| | | | | 24 | 20 |
| 9 | | 9 | | 9 | |
| 8 | | 8 | | 8 | |
| 7 | | 7 | | 7 | 94 |
| 6 | | 6 | 12 | 6 | |
| 5 | | 5 | | 5 | |
| 4 | | 4 | 45 | 4 | |
| 3 | | 3 | 30,44,33 | 3 | |
| 2 | | 2 | 07 | 2 | 40 |
| 1 | | 1 | 83 | 1 | |
| 1 | 46 | 1 | 33,14,09,17,25 | 1 | |
| 0 | 67,78 | 0 | 99,76 | 0 | 99,85,80 |
| 0 | 45 | 0 | 37,37,35,18 | 0 | |
| 0 | 30,29,08,09,23 | 0 | 21,30 | 0 | |
| 0 | 79,85,89,77,87,52,73,65,57,63,93 | 0 | 52,96,80,98,59,81 | 0 | |
| 1 | 04,15,01,08,00,16,24 | 1 | 25,36,40,08,39 | 1 | 14,46,38 |
| 1 | | 1 | 61,73 | 1 | 99,80,67 |
| 2 | | 2 | 09,20 | 2 | |
| 3 | | 3 | | 3 | 79,12 |
| 4 | | 4 | | 4 | |
| 5 | | 5 | | 5 | 13,92 |
| 6 | | 6 | | 6 | |
| 7 | | 7 | | 7 | |
| 8 | | 8 | | 8 | |
| 9 | | 9 | | 9 | 77 |

Stem: 100s
Leaves: 10s, units

*Figure 11. Stem and Leaves Diagramme of Residuals, Disaggregated by the Samples from the Three Groups of Countries*

increases as level rises. Note that what is meant here by 'level' is *not* the level of the residuals but the level of the dependent variable $D_i$ (which varies with the independent variable $I_i$). This is a case of heteroscedasticity in regression analysis, which implies that one of the basic assumptions of the linear regression model is not satisfied, namely that the residual term has a constant variance for all i.

Before turning to this problem, however, it is worth noting some other features of the disaggregated stem and leaves diagramme. In the sample of low income countries, only four out 27 residuals are positive. This means that in general, the level of public debt is still *lower* than what is explained by the lower income *per capita* (even after taking account of population size). Here the limits on the capacity to borrow are more significant, and the pattern of residuals indicates that these countries are indeed at a disadvantage in terms of their ability to obtain finance, even if their lower income *per capita* is accounted for. The middle income countries are more evenly distributed on either side of the regression line, and there appears to be a bit of a bimodal shape which might be interesting to pursue. The upper middle income countries are so heterogeneously spread out that it may be more useful to analyse them by examining specific cases on the spectrum rather than 'averaging' them. In other words, the nature of a scatter often helps in deciding whether groups make sense, or whether it would be better to look at certain cases within a pattern.

The preceding analysis has shown that the variation in the residuals tends to be greater as the level of the independent variable increases. In this sense the assumptions of the linear regression model are not satisfied. The estimation of the regression line, however, has as yet been only part of *exploring* the data and their interrelationships. By doing this, it was possible to ascertain that the assumptions were not likely to be satisfied, but this required a study of the residuals – and not a mere glance at the coefficient of determination – to see whether the fit was good. All too often researchers make regression analyses with a view to getting a high $R^2$, but they take little notice of the really interesting features revealed by the data. Regression can be used to explore data if the researcher concentrates on both the variations which are explained *and* those which are not.

In fact, the method of least squares regression is not always the most appropriate one for exploring data. As was the case with both the mean and the standard deviation, it uses all the data to derive the regression coefficients, and thus it is sensitive to outliers. More robust regression techniques do exist, however, and if the data base is small and outliers are present, it is better to use these methods for exploratory analysis. In the analysis above this was not done because it would have involved elaborating on the technique of estimating the coefficients using robust estimation. Also, as mentioned above, the data base was rather large (79 observations) and hence the distorting impact of a few outliers is nowhere near as strong.

Having explored the data by using regression techniques it is also possible to get a better idea of the specific model which will subsequently be used for estimation. In this example, it is clear that it will be necessary to correct for heteroscedasticity in the residuals. In some cases, transforming the data may help. In others, it may be preferable to try estimating the variances at different intervals along the scale (of the independent variable) and subsequently to use a more appropriate technique such as generalised least squares. Often, exploratory analysis can help in more clearly specifying the model and the method used to estimate it. In this case, for example, since many countries are grouped together around narrow intervals of income *per capita*, one could try to estimate the variance in the residuals from such groups at different points along the scale, and thus to get a better idea of the pattern of heteroscedasticity.

## CONCLUSIONS

As was made clear from the start, the purpose of this paper is not to put forward a finished theory on the indebtedness of developing countries. In fact, far too few factors were integrated into the analysis for tackling that problem. Only the cross-section pattern of the external indebtedness of developing countries in a given year was examined, and that discussion referred only to their incomes *per capita* category and, later on, their population sizes as well. No analysis was made of the pattern of the debts as they evolved over time, nor of such factors as the evolution of the terms of trade, differences in the pattern of export dependency and in the degree of industrialisation etc.

The aim of this paper is different and much more modest. It tries to show that careful analysis of relatively few data can provide many insights and reveal questions to be pursued further. Too often, researchers do not analyse their data or think that they can only be analysed by starting with interrelationships that are already very complex. Hence, it is not uncommon that an elaborate and sophisticated model is tested against the data while little or no prior analysis of the data base has been done. Modifications to the model are then made only to improve the fit, but not necessarily to arrive at a better understanding of the problem. This type of procedure implies that one's knowledge is not really enriched by the process of empirical analysis. A hypothesis is tested for its fit, but the process of progressing towards clearer hyphotheses about the problem being studied does not take place.

This analysis was developed from a simple set of data: the pattern of the developing countries' public external debts was grouped into income *per capita* categories. It was obviously assumed from the outset that size of the population of a country matters just as much as the level of income *per capita.* Nevertheless, since each income category includes large and small countries, it was considered preferable to start with the raw data. Careful analysis revealed the asymmetry in the pattern within each group and the phenomena of the covariance of spread and of level among groups. Outliers were noted in passing for further reference later. The analysis was then extended by relating external debts (scaled down by size measured by population) to income *per capita*, and this confirmed that the latter factor was significant in explaining the pattern of the distribution of external debt. It was shown that analysis of the pattern of its residuals can give further hints as to which elements could be important for further investigation.

Finally, it was shown that a careful exploration of the patterns which the data reveal can help to ascertain whether it is possible to satisfy the assumptions of the method of hypothesis testing, or of the estimation techniques which might be applied subsequently. In a sense, this method of analysis provides some diagnostic tests before the actual tests are carried out.

## NOTES

1. See World Bank, *World Development Report* (Oxford University Press, 1983), pp. 187-179, Table 16.

2. A 'far outlier' is defined as having a value that is two steps or more away from its nearest quartile.

3. Tukey has suggested an easy way to determine the location of the median and upper and lower quartiles. If n = sample size, n/2 and n/4 may be computed. If the resulting value is *not* an integer, the position of the median (or quartile) is obtained by rounding up the result to the highest unit. For example, if n/4 = 2.25 or 2.5, the rank is 3 in both cases. If n/2 or n/4 yields an integer, the median (or quartile) is calculated as the arithmetic mean of the $\frac{n}{2}$-th and $(\frac{n}{2} + 1)$ th values.

4. Note that the standard deviation and the midspread cannot be compared in the same manner as the mean and median. A normal symmetrical distribution yields an equal mean and median, but its standard deviation and its midspread are not equal in size, since they do not measure the same thing.

5. The interested reader will find a more formal explanation of the rationale behind using logarithms to transform the data in Appendix I.

6. The anti-log of the mean of the transformed data is the *geometric* mean of the original data. See Appendix I.

7. I am indebted to Rudolf Teekens of the Institute of Social Studies for drawing my attention to this fundamental point about the process of data analysis.

8. For the data base see Table 7.

9. In Figure 10 the stem contains positive and negative values. This is laid out by recording positive values above the vertical dividing line and negative values below it. The stem has been halved for lower residuals (with respect to their absolute value).

## APPENDIX: TRANSFORMING BY USING LOGARITHMS

In this Appendix some of the propositions made on the use of logarithms in transforming data will be proved. This will be done by starting with an analysis of the effect of stretching (or shrinking) a set of data by multiplying them with a constant scaling factor.

Let $X_i$, $i = 1$ to n be a sample of data. It is assumed that the data have already been ranked in order of magnitude from small to large. As an exercise, the data is stretched by multiplying each observation with a constant scaling factor. Hence a new set of data

$$Y_i = a \cdot X_i; \; i = 1 \text{ to } n$$

is obtained where:

$a$ = the constant scaling factor.

The question may now be posed of how different measures of level and spread for the set of $Y_i$s relate to those for the $X_i$s.

*Level*

*1. The median*: to obtain the median of the $X_i$, n/2 is computed. If the result is an integer, the median is computed by taking the arithmetic mean of the (n/2)th and the (n/2) + 1th observations. If not, the median is the value of the observation whose rank corresponds to the *upward* rounding off of n/2. For example, if n = 25, then n/2 = 12.5, and hence the median is the 13th observation. The reader can easily see that in both cases, the median of the $Y_i$s selects the corresponding values, and hence:

$$MD_y = a \cdot MD_x$$

*2. The mean*

$$\bar{x} = \frac{1}{n} \sum_{i=n}^{n} X_i$$

and

$$\begin{aligned}\bar{y} &= \frac{1}{n} \sum_{i=1}^{n} (a \cdot X_i) \\ &= \frac{1}{n} (a \cdot X_1 + a \cdot X_2 + \ldots + a \cdot X_n) \\ &= a\bar{x}\end{aligned}$$

Hence, if a set of data is multiplied with a constant scaling factor, the level of the newly obtained set of data will be equal to the prior level multiplied by the scaling factor.

*Spread*

*1. Midspread:* the reasoning here is similar to that used to obtain the median, and the result is the same, i.e.:

$$MS_y = a \cdot MS_x$$

*2. Standard deviation*

$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{x})^2$$

and:

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (a \cdot X_i - a \cdot \bar{x})^2$$
$$= a^2 \cdot s_x^2$$

and therefore:

$$s_y = a \cdot s_x$$

As with the level, the spread of the new data is equal to the prior spread multiplied by the scaling factor. Hence, the result of scaling a set of data is that *spread moves proportionally with the level* as determined by the scaling factor.

The reader will recall that this was exactly the situation in which using logarithm transformations was recommended. Clearly, in analysing the external debts of developing countries we were obviously not concerned with scaling one set of data but with comparing different sets. The theoretical example of the scaling of data does, however, provide an important insight into the interpretation of differences among samples. If it is noted that spread moves proportionally with level in comparing samples (as in the example of external debts), then there is good reason to conclude that the samples are not drawn from identical population distributions. Rather, there is an important element of scale operating here: *scale matters*! Furthermore, this is a significant point in furthering better theoretical understanding. It was noted that it matters, but why and how this can explain other aspects of the problem needs to be investigated.

What, then, is the logic behind transforming by using logarithms? Let us return to the theoretical example and transform both sets of data with logarithms. Let:

$$X_i' = \log X_i$$
$$Y_i' = \log Y_i$$
$$= \log (a \cdot X_i)$$
$$= \log a + \log X_i$$

Hence, it is clear that each $Y_i'$ differs from $X_i'$ with a constant factor (log a). This is indeed normal, since using logarithms implies transforming multiplication into addition.

What about the spread and level of $Y_i'$ relative to $X_i'$? Let us consider these one by one.

*Level*

*Mean*

$$\bar{x}' = \frac{1}{n} \sum_{i=1}^{n} \log X_i$$
$$\bar{x}' = \frac{1}{n} \log \prod_{i=1}^{n} X_i$$

This is because the sum of logarithms of the $X_i$ sample equals the logarithm of their product. And furthermore:

$$\bar{x}' = \log \left(\prod_{i=1}^{n} X_i\right)^{1/n}$$
$$= \log \bar{\bar{x}}$$

where $\overline{\overline{x}}$ equals the *geometric mean* of the original sample of the Xis.[1] Now:

$$\begin{aligned}\bar{y}' &= \frac{1}{n} \sum_{i=1}^{n} \log a \cdot X_i \\ &= \frac{1}{n} \log \prod_{i=1}^{n} a \cdot X_i \\ &= \frac{1}{n} \log a^{n} (X_1 \cdot X_2 \ldots \cdot X_n) \\ &= \log a + \overline{\overline{x}}'\end{aligned}$$

This result can also be expressed as follows:

$$\log \overline{\overline{y}} = \log a + \log \overline{\overline{x}}$$

As can be seen, the means of the transformed data differ by a constant equal to the logarithm of the scaling factor.

*Median*: If n/2 is not an integer, the result is straightforward since the medians of X′ and Y′ data will be the middle values of both samples. It was already shown that the corresponding $X_i'$ and $Y_i'$ differ by a constant equal to the logarithm of the scaling factor.

If n/2 is an integer, one interpolates between the two middle values of each sample. The reader can easily see that the arithmetic mean of $X'_{n/2}$ and of $X'_{(n/2)+1}$ is equal to the logarithm of the geometric mean of the middle values of the original $X_i$ sample. Using this result the reader can easily work out that the medians of the $X_i'$ and the $Y_i'$ samples differ by a constant factor equal to log a. Hence in general:

$$MD_{y'} = MD_{x'} + \log a$$

This result is identical to the one obtained for the mean.

*Spread*

*Midspread*: Using exactly the same logic as was used for the medians, one can deduce that:

$$Q_{U_{y'}} = \log a + Q_{U_{x'}}$$

and similarly for the lower quartiles of both sets. Therefore:

$$MS_{y'} = MS_{x'}$$

i.e. both sample have equal spreads! This result may appear strange at first, but in fact it is quite comprehensible. One only needs to remember that the difference between two logarithms equals the logarithm of the ratio of the two original values. Hence, the midspread of the Y′ data equals the logarithm of the ratio of the upper and lower quartile of the Y data. The latter ratio is the same as the ratio of the upper and lower quartiles of the X data! Hence, when using logarithms on the original data, the spread becomes equalised, since the ratios remain constant with scaling! Indeed, for any i and j:

$$\frac{Y_j}{Y_i} = \frac{a \cdot X_j}{a \cdot X_i} = \frac{X_j}{X_i}$$

*Standard deviation*
It is clear that:

$$s^2_{y'} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i' + \cancel{\log a} - \bar{x}' - \cancel{\log a})^2$$

since:

$$Y_i' = X_i' + \log a$$

and:

$$\bar{y} = \bar{x}_i' + \log a$$

and hence:

$$s^2_{y'} = s^2_{x'}$$

Therefore, both distributions have equal variances.

*Conclusion*
The above demonstrates that spread moves proportionally with level across different samples if they differ with respect to a scaling factor. Using logarithms to transform the data preserves the differences between means (or, in general, between levels), but it eliminates the differences in spread.

The reason for this equalisation of spread is quite easy to understand if one reflects on the properties of logarithms. Spread is always defined in terms of level. It always concerns variation around a basic theme – level. The midspread takes the median as its reference point and is obtained by *subtracting* the lower quartile from the upper one. The variance and its square root – the standard deviation – take the mean as their reference point. In calculating the variance the mean is subtracted from each value in the sample. Spread therefore involves *differencing*. Subtracting logarithms is equivalent to taking the logarithm of the ratios! Scaling down does not affect the *ratios* of the different values relative to each other and to their common mean (or median)!

This point is important in understanding how to correctly interpret the spread of transformed data. Taking another look at Table 5 demonstrates this. This table tabulates the midspreads of the transformed data in the samples on external indebtedness of developing countries. For example, the midspread of the transformed sample for low income countries equals 0.66. What does one get if the anti-log of this number is taken? This anti-log equals 4.57. Certainly, this bears no relationship to the midspread of the original data, which equalled 1,239. Looking at Table 2, however, one sees that $Q_U = 1{,}585$ and $Q_l = 346$ and, furthermore, their ratio equals:

$$Q_U/Q_L = 4.58$$

This is exactly equal to the anti-log of the midspread of the transformed data (the slight difference being due to rounding off errors in calculations). Hence, because logarithms are used the spread refers to *ratios* and no longer to *absolute values.*

A final point to be made here is that this type of transformation is used more frequently than may be apparent from this example. There is an interesting parallel, however, with analysing growth processes at a *constant* rate. If a growth rate is constant, the absolute addition from period to period is proportional to the *level* attained in the previous period. Indeed, if we let:

$$r = \frac{\Delta X_t}{X_t} = \text{cte}$$

where $r = $ growth rate
$X_t = $ a time series variable

then it follows that:

$$\begin{aligned} \Delta X_t &= \text{cte} \cdot X_t \\ &= r \cdot X_t \end{aligned}$$

And, hence, the absolute increment rises proportionally with the level of X attained in the previous period. Here, we have covariance of increment with level, and not surprisingly the logarithm transformation is also very popular in this case.

[1] The reader should note that

$$\overline{\overline{y}} = a \cdot \overline{\overline{x}}$$

i.e. the geometric mean of the $Y_i$ sample equals the geometric mean of the $X_i$ sample multiplied by the scaling factor. This result is in accordance with those obtained for the arithmetic mean and for the median.

## BIBLIOGRAPHICAL APPENDIX

It is only in recent years that exploratory data analysis has developed as a subject in its own right within the wider field of statistical analysis. Previously, descriptive statistics was a particularly undervalued branch of statistics. It was usually relegated to the first chapter in the textbooks and consisted of a quick survey of various ways to summarise data, presented before moving on to what were considered to be the really interesting parts of statistics. As such its content was very much determined by the requirements of confirmatory statistics. It was looked upon as a mere *prelude* to, rather than being a *complementary* branch of, confirmatory statistics.

Mainly due to the methods developed by the statistican John Tukey, exploratory data analysis has now asserted itself in its own right. The standard reference work on these methods is J.W. Tukey, *Exploratory Data Analysis* (Reading, MA, Addison-Wesley, 1979).

For a lucid account of exploratory analysis and its relationship to confirmatory statistics in applications within the social sciences, see B.H. Erichson and T.A. Nosanchuk, *Understanding Data* (Open University Press, 1979). This text is largely inspired by the methods developed by Tukey.

Economists may find it helpful to consult a general textbook on statistical methods which incorporates techniques in exploratory data analysis, in which case a good choice is T.H. and R.J. Wonnacott, *Introductory Statistics for Business and Economics* (Third Edition, Chichester, John Wiley & Sons, 1984).

Finally, for a complementary source which surveys non-parametric methods in statistical analysis and may be very handy for carrying out quick checks and tests in the process of data analysis, see P. Sprent, *Quick Statistics* (Harmondsworth, Penguin, 1981).

**Recently published Occasional Papers**

No. 97 C.A.O. van Nieuwenhuijze, 'Culture and Development: the prospects of an afterthought', 56 pp. (December 1983)

No. 98 Jan J.P. van Heemst, 'National Accounting and Subsistence Production in Developing Countries: some major issues', 27 pp. (March 1984)

No. 99 Sulabha Brahme, 'Producers' Cooperatives: experience and lessons from India', 39 pp. (June 1984)

No. 100 Wim Burger, 'Yugoslavia's Economic Crisis: the Price of Overexpansion', 23 pp. (October 1984)

No. 101 Marc Wuyts, 'Exploratory Data Analysis on Indebtedness in the Third World', 44 pp. (December 1985)

No. 102 Phil Wright, 'The Political Economy of the Yugoslav Revolution', 63 pp. (December 1985)

No. 103 Farooq Haroon, 'The Federalisation of Cooperative Banking in Pakistan and Rural Cooperatives in Punjab Province', 53 pp. (January 1985)

A full list of ISS publications can be obtained from the Publications Office, Institute of Social Studies.