THE SPIRIT LEVEL REVISITED

REGRESSION LINES, CORRELATION, OUTLIERS and MULTIVARIATE ANALYSIS

A plain man's guide to statistical inference in
THE SPIRIT LEVEL
and in the critique offered by Peter Saunders

Hugh Noble

1. The Spirit Level controversy

1.1 Introduction

Since it was published in 2009, **The Spirit Level** by Wilkinson and Pickett has attracted a deal of praise and some vigorous criticism. Many on the political left have hailed it as a vindication of their long-felt views about the desirability of equality and redistribution of wealth. Many on the political right have gathered their forces for a sustained attack to discredit it.

One such critique, a paper written by Peter Saunders (PS), has been published by the think-tank **Policy Exchange**. It is entitled **Beware False Prophets** and is currently readily available on the Internet. Wilkinson and Pickett (W&P) have responded robustly to their critics and their comments are also available on the Internet [1]. In defence of their thesis - that income inequality correlates (in some cases strongly) with various social problems - they point out that their selection of the dataset from which these results were obtained, was based on sound principles and decided before the data relating to each was analysed. This is in sharp contrast to the datasets preferred by Saunders who has added and subtracted countries to and from the dataset in an effort to get the relationships he prefers.

W&P have also pointed out the wealth of research, most of it reported in peer review journals, which supports their contentions and which also discounts most of the alternative arguments offered by Saunders. I refer the reader to both the Saunders paper and to that response by W&P.

My response to Saunders' arguments is somewhat different from that of W&P. Saunders is an Emeritus Professor of Sociology and so might be expected to have a sound command of statistics. From that position of authority, he accuses W&P of misusing or misunderstanding statistical techniques. When I examined his own methods, however, I found to my surprise, that his own use of statistics is woeful.

So I have written this short paper for the benefit of those who may not be familiar with the theory of statistics. I have tried to avoid the use of mathematics and to explain the essentials using only examples where the underlying rationale is obvious. For example, Saunders relies heavily on a technique called multivariate analysis but he does not ensure that the so-called "independent variables" which he uses, are really independent of each other. Using examples where the variables are obviously not independent, I shall show that the results obtained in these circumstances, can be quite inappropriate. By using more examples which yield obviously invalid results – I shall invalidate his reliance on boxplots to identify "outliers" and his use of time-trends. I shall flatly contradict his view on the linearity of regression lines and I will justify that contradiction by referring to the established authorities in the field who invented many of the techniques Saunders relies upon.

1.2 The Basic Thesis of The Spirit Level

The basic proposition advanced by Wilkinson and Pickett comes in two parts.

(1) **The Diminishing returns of GDP** The first part is their observation that when a poor country becomes more wealthy (measured by GDP/head), the social problems associated with its poverty (identified most clearly by a low lifespan expectancy) will be steadily eliminated - but only up to a point. The improvement does not continue indefinitely. Beyond a certain point, an increase in GDP per head does not result in a significant increase in life expectancy.

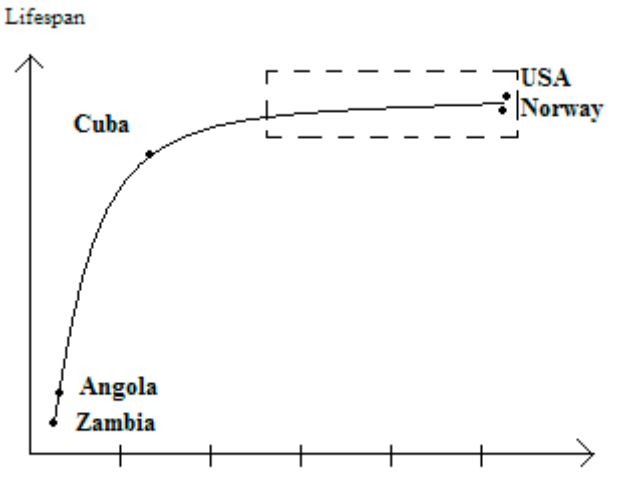

FIG 1

This is clearly shown in FIG 1 which plots life expectancy against GDP per head, for a large number of countries. I have shown only a few of the countries.

The rest are clustered around the line. The point to notice is the knee bend in the curve where Cuba (and several other countries which I have not shown) are located. As GDP/head increases beyond this point the curve flattens.

As time goes on, advances in medical knowledge increase the life expectancy of all countries, but the point made by W&P is that those rich countries do not get significant improvements in life expectancy by increasing GDP. This observation applies to the countries, which are located within that dotted line box in the graph.

Furthermore, when we look at the social problems which are present in the rich countries which lie above that knee-bend in FIG 1, we find that there are considerable differences which are not related to differences in GDP. Indeed, the most striking difference, is between Norway and the USA which have very a similar high level of GDP/head. The USA is afflicted by severe social problems while Norway is blessed by having fewer social problems.

(2) The Influence of Inequality. The factor that is responsible for this disparity (according to W&P), is inequality and the measure of that, which they use, is inequality of income. In that respect, the USA is one of the most unequal countries in the world and Norway is one of the most equal. The USA and Norway are also at opposite ends of the trend line showing the relationship between inequality and the "index" of social problems.

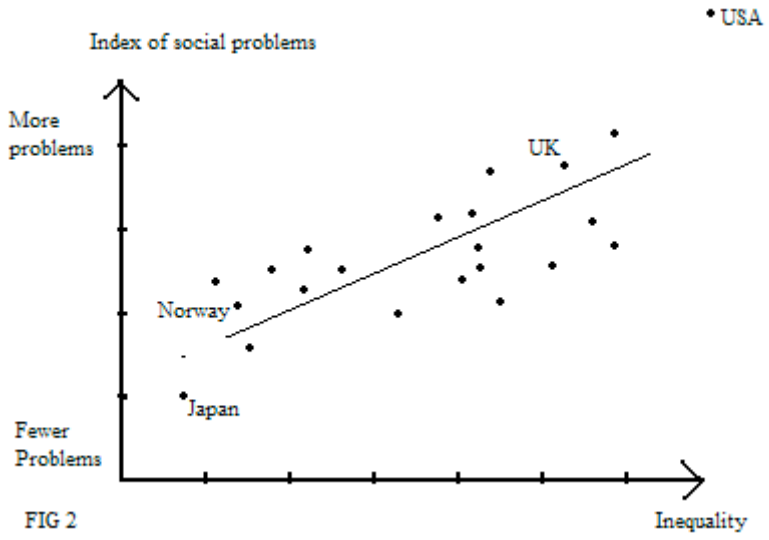

FIG 2

NOTE: Singapore does not appear in this diagram. The apparently anomalous position which Singapore occupies on most of the graphs shown by W&P, is the subject of discussion later.

In their book, W&P examined the statistics for about 20 different types of social problem. These included homicide rates, infant mortality, teenage pregnancies, high rates of imprisonment and also some health problems like adult obesity and mental health. They have demonstrated that the prevalence of those social problems in a group of 23 modern industrial nations (in FIG 1 all of these 23 countries are located within the dotted box above the knee-bend) correlate strongly with income inequality. W&P also examined the same or similar statistics for the 50 individual states of the USA and the same trend line emerged.

The authors also put several of these "problems" together to create what they call "an index of health and social problems and they plotted the relationship between income inequality and that composite index. There were 9 "problems included in the index:

Level of trust,
Mental illness (including drug and alcohol addiction*),
Life expectancy and infant mortality,
Adult obesity, (not child obesity*)
Children's educational performance,
Teenage births,
Homicides,
Imprisonment rates,
Social mobility. (not available for the US states)

* Note: according to the data presented by W&P, alcohol ***abuse*** correlates with income inequality, but alcohol ***use*** does not. Also ***adult*** obesity correlates, but ***childhood*** obesity does not.

W&P based their analysis of the index on data from 20 countries drawn from their dataset (and for which the relevant data were available). They also did it for the 50 individual US states. The graph which emerged for the country data, is shown in FIG 2. The US states gave a similar result. The relationship between the Index and income inequality, was more striking than with any of the individual "problems". It gave a clear correlation line, with all the countries (and also those individual US states) bunched around the line. I copied W&P's diagram by hand. I apologise if there are any discrepancies but I assure the reader that they are negligible. My graph does give the same general impression as W&P's graph particularly with the relative positions of the USA, the UK, Japan and Norway.

Of all the graphs and correlation lines presented in The Spirit Level, this one is the most sharply defined and the most significant. W&P place considerable emphasis on it. Peter Saunders however has questioned its validity.

1.3 Cause and Effect?

After a correlation has been established as statistically significant, the next step is to explain the relationship and hopefully explain it in terms of some kind of mechanism involving cause and effect.

The most obvious assumption is that inequality is a direct and immediate cause of these social problems. Every student of statistics is taught to avoid the temptation to make that simplistic assumption. As we shall see in Section 2, there are other possible explanations for a statistical correlation of that kind. I will discuss all of these other possibilities shortly and I will come to the conclusion that while the central thrust of the W&P thesis is correct (and valuable), the full story may be much more complicated.

1.4 Criticism of The Spirit Level

Much of the argument which has been launched against Wilkinson and Pickett has been focused on the issue of straight regression lines and on the question about whether the USA in particular is a special case which should not be allowed to exert an undue influence on what purports to be a general law.

No one, however - certainly not the detractors - has been able to show that the statistical data about the USA is actually wrong. The USA is a very unequal society and it is afflicted by a plethora of social problems. The argument is only that it is so different from the data relating to other countries, that it gives a false indication of the regression slope.

Even if that is so, however, the USA represents a massively important datum point in the global economy. The evidence available, which is clearly documented in The Spirit Level, and which, ironically, is confirmed by its critics, demonstrates conclusively that the USA we see today does not provide us with an example which any country would be wise to try to emulate. I refer of course to the USA which we see today - what we may call the "Tea-Party" USA, as distinct from the admirable USA of the Marshall Plan, of Martin Luther King, of the Moon landings and of Tom Lehrer).

2. A Plain Man's Guide to Statistical Inference.

2.1 Statistical Correlation

The conventional way to demonstrate that two variables are related in some way, is to draw a graph. Each dot on this graph represents one measured point. Each represents a point where "this much of X" corresponded with "that much of Y". The result is called "a scatter diagram". We can then draw a straight line through those scatted dots so that the line passes as close as possible to as many of the dots as possible. We can do this "by eye" or we can use a computer software package to do it for us. The computer package will be more accurate but in many cases the general lie of the best-fit line is so obvious that precise accuracy scarcely matters.

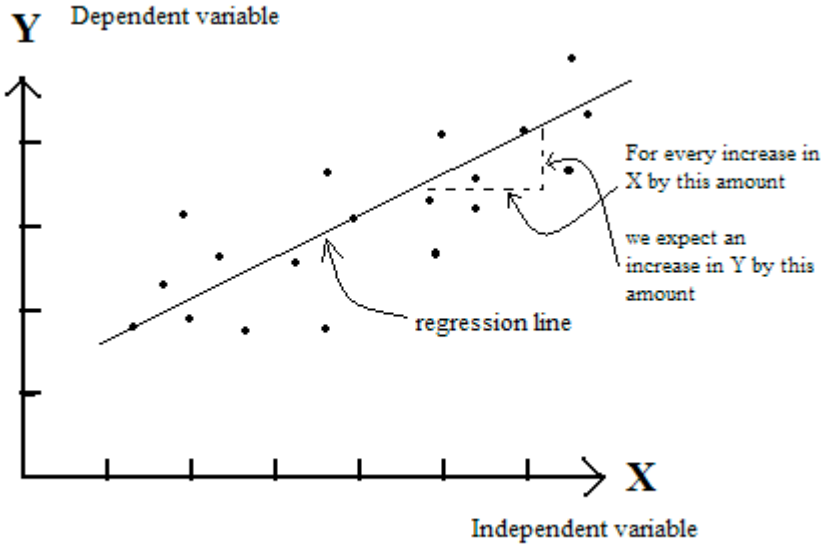

FIG 3

When the slope of the line is upwards to the right that indicates that as X increases, Y also tends to increase. If it slopes downwards to the right, that indicates that as X increases Y tends to decrease. We could call that a reverse or negative correlation. If the line is horizontal (or nearly so) that indicates that the value of X has no influence on the value of Y.

The degree of scatter (away from the regression line) is due either to errors in measurement or to the presence other factors, which influence the value of the Y-variable independently of X. If all the points in the plot were positioned exactly on the regression line, that would indicate a perfect correlation. We could then draw the conclusion that factors (other than X and Y the two factors being plotted against one another) do not exert any influence. There is a special problem which arises when it is possible to draw a line through the datum

points in almost any direction. We shall see later how we can detect that condition.

2.2 Dependent and Independent variables

In the example shown in FIG 3, X is called "the independent variable" and Y is "the dependent variable". That means that we are using the values of X, which are presumably easily measured or fixed, to give us a clue (or to predict) the value of Y that is likely to be associated with that value of X. When we plot the regression line we call that the plot of "Y on X". The dependent/independent designations of X and Y can be switched, which would give us a plot of "X on Y".

2.3 Correlation and Causation

When someone demonstrates a statistical correlation between two variables (Y on X) it is tempting to think that an increase in the value of X must be the direct cause of a corresponding increase in the value of Y. We can write that this way,

"X => Y"

where the symbol "=>" means "causes".

But that is only one possibility. It is also possible that the causal connection goes in the other direction, "Y => X".

2.4 Spurious Correlations

There is a third possible explanation. It could be that there is another unidentified variable (Z perhaps) which is causing both X and Y.

Z => (X & Y)

That third kind of explanation can give rise to all manner of spurious correlations. For example, it is a fact that the quality of a school child's handwriting and the size of his or her big toe, are strongly correlated. The better is the handwriting, the larger is the big toe. That is not because some lobe of the brain, which influences the ability to write is located in the big toe, or because writing properly causes the big toe to grow in size. It is because there is a third variable "age" which influences both. As a child gets older handwriting improves and the big toe grows.

Textbooks, and popular accounts of statistical analysis often simplify this story (about spurious correlations) by saying that a correlation does not necessarily imply a "causal" linkage. That gives us a valid warning, but it is not strictly true. If a correlation between X and Y is genuine and statistically significant (i.e. it is not due to a chance coincidence), then, even when it is "spurious", there will always be ***some kind*** of underlying ***causal*** linkage. But that linkage (as I have shown with the example about writing and toes) is not necessarily direct. It may be indirect. And being indirect, it may also not be of interest to us. But the relationship Z => (X & Y) is not necessarily of no interest. If, for example it is also the case that Y => Z, then we have a much more complex relationship that can be considerable interest.

That point may seem to be pedantic at present, but as we will see later, it does have an impact on the interpretation of the statistical analyses offered us in both The Spirit Level and the Saunders paper.

W&P have shown clearly that there is a statistical correlation between various social ills and income inequality. The task that then confronts us is to find a plausible cause-and-effect explanation for that correlation.

2.5 Other Complications

That range of possibilities - i.e. (X causes Y), (Y causes X) and (Z causes both X and Y) does not exhaust the possibilities. Even if X really is a cause of Y it may not be the only thing which causes Y, so, in addition to

X => Y

we might also have (at the same time) -

A => Y
B => Y
C => Y
etc

It is these additional causal factors, being greater and smaller for different datum points, which often create the scatter effect of the points on a graph plot.

2.6 Causal Chains, feedback and time delays

It is also possible that a causal link involves several intermediate variables so that -

X => => => => => ..... Y

where the long chain of causal connections involves a long sequence of other variables.

Consider this hypothetical scenario - A decrease in direct taxation at the upper end of the income range may cause a decrease in the amount of money available to recruit police. This could cause a decrease in police numbers, which could cause an increase in crime, which could cause public outrage, which could cause an increase in prison sentences for criminal behaviour, which could cause an increase in the amount of money required to build prisons, which might cause an increase in taxation.

I emphasise the word "*could*". I am not suggesting that this chain of events is what inevitably happens. I claim only that it *could* happen that way. However, if you follow that scenario through, you will see that (a) it is a closed loop. A decrease in the first variable eventually causes an increase in itself, and (b) you will recognise that every stage in this complicated causal linkage would almost certainly involve a time delay. That cumulative delay, together with the negative feedback stage, could eventually lead to slow oscillations.

The longer is the causal chain, the more likely it is that time delays and oscillations occur, and that time delay will be longer. The social order in a country (or an individual US state) is not static. So the effect that income inequality may have on any particular social problem may be related, not only to the degree of inequality we observe in that country, at present, but to the length of time it has been in that condition.

It is known that the disparity in incomes within the USA diminished over the period from the Second World War, to the 1970s. This has been called "The Great Convergence". From the 1970s until the present day, the gap became greater ("The Great Divergence") [2]. Until the 1990s, the East European countries (which Saunders wants to be included in the dataset when that produces the result he wants), were part of the Soviet Bloc and subject to a form of enforced income equality. Since the fall of the Berlin Wall, those countries have been subjected to a very rapid switch to capitalism, and we have seen the rise, in these countries, of billionaire oligarchs. Economic and political circumstances may change quite rapidly, but social attitudes will probably change more slowly and that will have an influence on the degree to which inequality of income may be related to social problems.

If we have a circular chain like this,

X => => => => Y => => => => => => =>
<= <= <= <= <= <= <= Z <= <= <= <=

then, by choosing to start counting the circuit from different points we can make a case for all three of the alternatives. "X => Y", "Y => X" and "Z => (X & Y)".

2.7 Straight and curved regression lines

It should not be assumed that because we can draw a "best fit" straight line (or straight regression line) through a scatter of points in a diagram, that the causal relationship between the variables which that line demonstrates, is necessarily straight. That is, it may not be the case that a small increase in X results in exactly the same increase in Y at all points in the graph, irrespective of the value of X to which the increase in X is an increment. In general the relationship between any two variables in the real world, which are causally linked, is seldom straight.

But it is, nevertheless, always possible to draw a "best fit" straight line through any set of points even when they have no causal relationship at all (although, if there is no causal connection, the straight line will have a tendency to be horizontal). Therefore, a significant straight line regression curve between two variables (which is not horizontal) establishes only that there is *some kind of* causal connection between them - even if it is a spurious connection as in the example about handwriting and big toes.

2.8 Heat and Death (an example)

To illustrate that point about the relationship between two variables often being best represented by a curved and not a straight line, let me draw attention to a known causal relationship between death and temperature. The human body is able to cope with changes in temperature. We shiver when we are too cold and we sweat when we are too hot. These and other physiological mechanisms enable us to deal with fluctuations in temperature. But if the temperature falls below or rises above certain limits, we suffer, and a proportion of the population will die. It is widely reported that during a prolonged heat wave, the number of deaths per thousand will rise. Here then is a fictitious (but plausible) set of data -

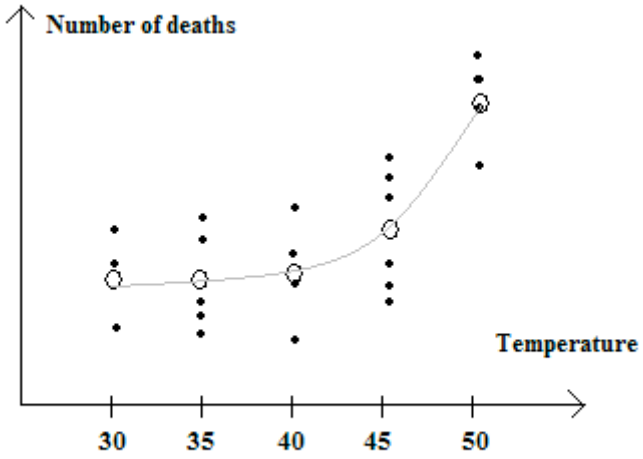

FIG 4

According to this graph (FIG 4) there have been three days which had a noonday temperature of 30 degrees Celsius, six at 35, and so on. When we average the recorded deaths at each of these temperatures, we get an average value (shown as a circle). When we then draw a relationship line by eye through those average values, we get a line which curves upwards to the right. This shows that at normal temperatures the number of deaths remains more or less constant, but begins to curve upwards as the temperature rises above normal body temperature (39.6). Presumably, at some higher temperature (60?) the curve will become vertical. Beyond that point everyone dies. If atmospheric temperature goes beyond some limit, those homeostatic mechanisms are progressively overwhelmed.

The same may well be true of society's ability to limit the damaging effects of inequality. So it would not be surprising if the true shape of the regression line showing the relationship between income inequality and any particular social ill was an upward curve.

In the diagram below (FIG 5) I have drawn a speculative straight regression line through the set of average points and tried to make it pass as close to each of them as possible. I repeat - regardless of the shape of the actual underlying relationship it is always possible to draw such a "best-fit" straight regression line. A straight regression line can, in these circumstances, be regarded as an "average" relationship. It is quite clear however, that a set of figures can have an average value without any individual datum (or any pair of datum points), within the dataset, necessarily having that value.

I drew that straight line in the FIG 5, by eye, but using sophisticated computer software packages makes little difference. One way or another, a best-fit straight line can always be drawn.

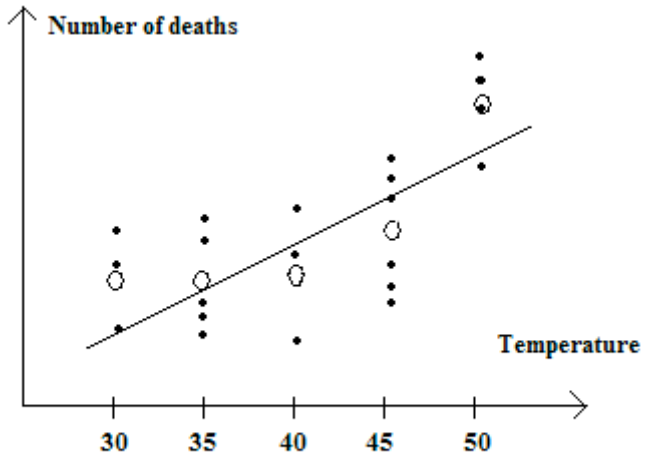

FIG 5

If we do the usual statistical calculations we can show that the slope of this line is statistically significant. Having then established a straight regression line, having shown that it is not horizontal and that its slope is statistically significant, it is incumbent on us that we should sit down and try to work out what that relationship really is.

In his critique of the Spirit Level, Peter Saunders makes this comment -

... regression techniques are quite demanding. They not only require that the slope of the trend line should not be distorted by a few extreme cases, but also that the association between variables be linear. (i.e. as the value of X increases, so the value of Y should increase or decrease at a fairly steady rate across the whole distribution) ... [PS: 55]

Saunders then goes on to show that the constant gradient condition does not apply to some of the graphs offered in The Spirit Level. He goes on to claim that -

"*a key requirement of regression analysis has been violated*" [PS 57]

Both those statements are technically incorrect. As I have shown with the death and temperature example above, a straight regression line, (which can be shown to be significant) indicates only that a causal relationship of some kind exists. It does NOT imply that that relationship is *necessarily* best represented by a line, which is straight.

Much of Sanders' criticism of The Spirit Level depends upon his understanding (or, as I claim, misunderstanding) of the concept of regression line. In making that claim, I turn for support to the "Dictionary of Statistical Terms" by Kendal and Buckland. [3]. M.G. Kendall was a towering figure in the statistics field. He was also the originator of many of the standard techniques we use today. Here are two entries in that dictionary.

Regression Curve: A diagrammatic exposition of a regression equation. .... The term is sometimes interpreted to mean a regression equation of a higher degree than first, [i.e. not a straight line] *the emphasis then lying on the word "curve" as opposed to a straight line.*

Regression Line: In general this is synonymous with regression curve, but is sometimes (and rather ambiguously) used to denote a linear regression.
[Kendal and Buckland 1957]

Saunders, it appears, has interpreted the expression "*regression line*" in the rather restricted (and ambiguous) way mentioned in the second quotation. He is incorrect, therefore, when he insists that the interpretation, which he prefers, is a "*key requirement of regression analysis*". Furthermore, if his interpretation is abandoned, much of his criticism collapses.

A further point raised by Saunders is that the distribution of "*residuals*" around the regression line (in The Spirit Level diagrams) is not "normal". To examine the validity of this comment we had better think about residuals and normal distributions.

2.9 Residuals

A residual is the (vertical) distance between a datum point and the regression line. A residual can be positive or negative. Residuals are shown on this diagram by dotted lines (FIG 6).

When we look at the diagram above (FIG 5) we can see a clue that the underlying relationship is really not a straight line. Look at the "residuals".

In the (death x temperature) diagram (FIG 5), the residuals are not scattered around the straight regression line in a random kind of way. They are negative in the centre and positive at either end. That gives us a clue about how the "goodness of fit" of a regression line can be tested.

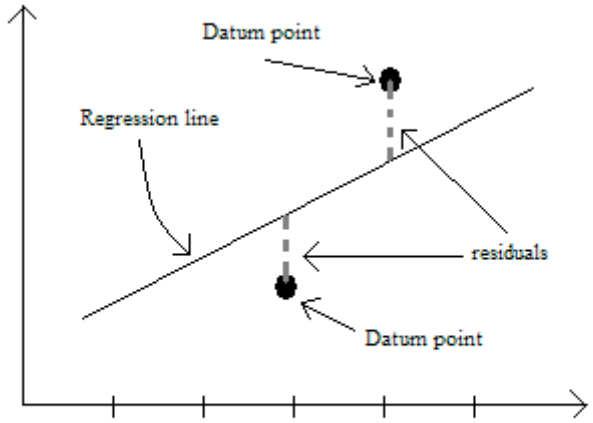

FIG 6

We can do some statistical tests on the residuals and we would like the sum total of all the residuals to be as small as possible. If it is possible to draw another regression line which has a smaller sum total of all its residuals then that alternative line should be preferred.

The calculation of the sum of all the residuals is not simple however. If you add positive and negative residuals it is likely that they will cancel one another out and make it seem as if the regression line is perfect (when it clearly is not). One way to avoid that is to multiply each residual by itself. The result of squaring a number like that, is always a positive number - so the cancelling out effect is avoided.

There is a further advantage gained by squaring the residuals. It weights the result against having large residuals. The square of 1 is 1. The square of 2 is 4. The square of 3 is 9. When we add all those squared residuals together, large residuals count for a lot more than the small ones. That means that to get a minimum total value of all (squared) residuals we should avoid a large residual value like 3 even if that means generating a lot of small ones sized 1. In fact one residual of size 3 is 9 times worse than a residual of size 1.

When a straight regression line is drawn through a set of points, the technique for finding the best-fit line is called the "method of least squares". Computer packages which calculate the best line, use that technique to find it. And the same idea can be applied *to any shape of curve* provided we know a mathematical equation for the curve.

If you throw a collection of different equations at a computer package it can tell you which curve is the "best" one. Unfortunately, however, computer packages cannot tell us what is the underlying mechanism of the relationship between variables. For that we need to understand what the data signify and we

need to come up with a plausible scenario which explains why the curve of our choice is a reasonable one to expect.

There is a snag however. In FIG 7 below, we see a scatter of points which appears to have no preferred direction of regression. If we present this kind of data to a computer package it will draw us a best-fit straight line (the heavy black line), but that line will have little significance, because we could have drawn other lines (dashed lines) in almost any direction through the centroid of the scatter diagram and the sum of the squared residuals associated with each, would be almost the same in every case.

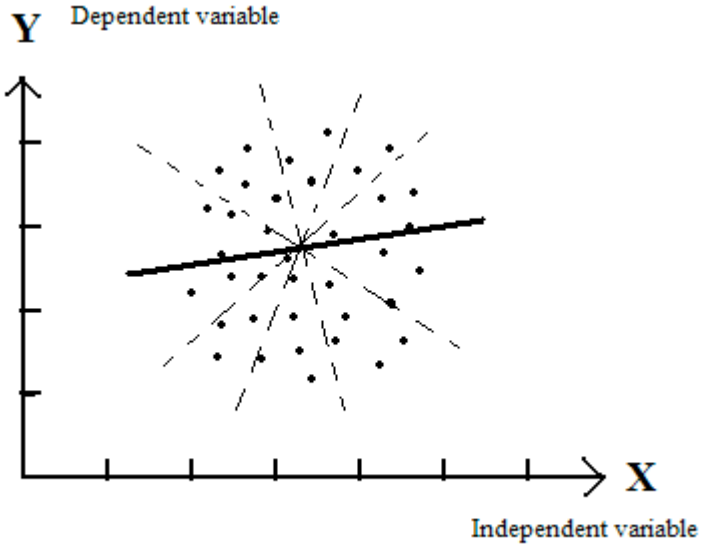

FIG 7

Sophisticated computer software can show us that that is the case. It can show us how sensitive the curve is to rotations of that kind. If any other line orientation results in a very large increase in the sum of the squared residuals, then we can be sure that the best-fit regression line is really the best by a long way. If not, then we can dismiss the regression line as insignificant.

2.10 Normal Distribution When we calculate the sum of the squared residuals and use it as a measure of "goodness of fit" for a regression line, we are making an assumption that the residuals are distributed about the line in a pattern that is called "normal distribution". This means the residuals are mostly bunched close to the regression line and that, as residuals become greater and greater, the frequency of occurrence becomes smaller and smaller. When we draw a diagram of a normal distribution, we get a "bell-curve" like the one illustrated in FIG 8 below.

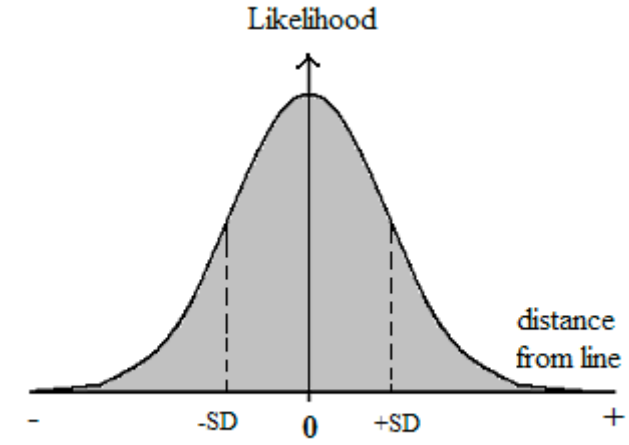

FIG 8

As the values get further and further away from that central zero the frequency of occurrence drops off to zero. Theoretically a normal distribution tails off to zero at infinity (and negative infinity) but we can usually ignore that because it gets close enough to zero within a measurable distance. If we square the value of each deviation (distance from the central value), take the average of all those squared values and then find the square root of that average value, we get what is called "the standard deviation" (or SD) of the distribution. You can find an SD for any kind of distribution but for a genuinely normal distribution you will find that some 68% of deviations will be less than one SD away from that central zero. Less than 5% of the deviations will be more than 2 SDs away from zero and less than 1% will be more than 3 SDs away from zero. These values are true no matter what the curve represents, how flattened it is, or how peaked in the centre, so long as it is a normal distribution. Most of the tests which we can do on regression lines assume that residuals are distributed around the regression line in that way.

2.11 P-Values and the Null Hypotheis.

Let's say that we have two sets of results - showing, for example, the growth of two groups of plants. One group has been treated with a special kind of fertilizer and the other group has not been treated. We have the average growth rate for each group and we have the standard deviations of both. The question we now ask is - does that amount of difference in growth indicate that the fertilizer has really worked? Is the difference between them significant?

When we are confident that the data we are looking at corresponds to a normal distribution (or is sufficiently close to it), we can use the regularity of the distribution of SDs to tell us when there is something significant about our observations.

The idea is based on what is called "The Null Hypothesis" and it goes like this. We say to ourselves - "How likely is it that this data could just be a coincidence?" If we had used the roll of dice, the drawing of cards, or the spin of a roulette wheel (or any other "random" method of generating data) to produce a set of results like that, and if we did that millions of times, in what proportion of cases would we find that the data was similar to our actual observations (or are even more extreme)?

If the distribution of the data is "normal" then we can look at a published table of the normal distribution and read there the value of "p" (p = "probability") of getting such a result "by accident". The conventional decision is that if fewer than 5% of outcomes would show a similar discrepancy, the real observation is said to be "significant" (i.e. it is not likely to be an accident). If fewer than 1% would show such a difference, the real observation is said to be "highly significant".

What goes for the growth of plants, also goes for other data such as the slope of our regression line and the scatter of points around it. What we are saying is, how likely is it that a similar scatter of points (but one in which the scatter is produced by some entirely random method) would exhibit a similar regression line and a similar scatter of points away from the line?

Nothing in this world is absolutely certain, but by following that convention we can be sure that we are not allowing personal prejudice to influence our decisions.

2.12 The Theorem of Central Limits

There is also a mathematical theorem which proves that measurements which we may make in many walks of life - using a tape measure for example, or a theodolite, or a protractor to measure an angle, or a barometer or a thermometer, or any of a wide variety of measuring instruments - the measurements we get are subject to small random errors, and those errors will be distributed in a way that approximates very closely to the normal distribution, especially if those measurements are actually the average values obtained from lots of individual measurements. This is because it is assumed that the error obtained from each of those individual measurements is the sum total of lots and lots of even smaller errors. Perhaps the temperature caused our measuring tape to expand very slightly. Perhaps our theodolite was very slightly off true horizontal. Perhaps the gradations on our thermometer were not perfectly marked on its surface. These and a myriad other circumstances, added together, contributed to our total error in the measurement. The **Theorem of Central Limits** proves mathematically that such an accumulation of small errors will yield a distribution of errors which is so close to normal that the difference is negligible.

However, and this is very important to our consideration of the arguments offered us by both W&P and by PS, when we are dealing with measurements that arise from the answers which people give to questionnaires and from other common ways of gathering data on social conditions, it is not at all clear that that assumption about the distribution of errors being normal, is valid. But a great many of the standard statistical tests we use to compare samples to see if the difference between them is significant, assume that normal distribution of errors pertains. So the results that those tests yield, may not be valid. This should not lead us to discount the statistical analysis of either W&P or PS, but it should make us cautious. More about this later.

2.13 Outliers

The issue of outliers features strongly in the critique offered by Saunders. He draws our attention to the fact that the USA, on most of the graphs used by W&P, is stuck out on its own. It is obviously the most unequal society (with the exception of Singapore) and the most heavily affected by the social problems identified by W&P. By being in that position, with no other datum points anywhere near it, it exerts a considerable influence on the best-fit regression line. This, PS claims, is sufficient justification to remove it from the graphs where it has this effect. When this is done (along with some other judicious deletions of datum points), the apparent regression between inequality and the social problem being addressed, disappears or is greatly diminished. W&P have responded to this by producing still more statistical support for their thesis and by challenging the legitimacy of this kind of selective deletion (and addition) of data.

W&P are right to be sceptical about the claims made by Saunders on this "outlier" argument. I want to add weight to their argument and I want to do that by coming at the problem from a somewhat different direction. I want to use an easily understood example to show when it is legitimate to delete an apparent outlier and when it is not.

2.14 A case of mistaken identity (The Nevis-Everest Mistake)

Example-1 Imagine that we are in the Scottish Highlands and that we are trying to make an accurate measure of the height of Ben Nevis (Britain's highest peak). We lay out a baseline, we measure angles with a theodolite and we measure the inclination angle. To ensure accuracy, we do this 6 times and we do it from different locations and using different baselines. We note all of these measurements in a notebook and take it back home. We then get out a calculator and start to do the necessary calculations. Five of these calculations

yield results which are all close to 4406 feet. The sixth calculation gives us a height of 6044 feet.

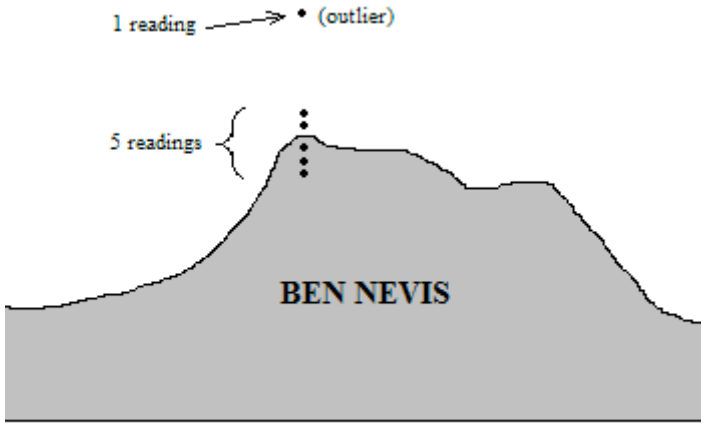

FIG 9

What should we do about that? If we include this anomalous value we will get an average value which is very different from the average of all the other results. Since all of these results are estimates of a ***single*** mountain, we can reasonably expect them all to yield results, which are very similar. That is true for 5 of the results. But it is not true for the sixth observation. We can argue therefore (and plausibility) that that 6th reading was a fluke which was the result of some kind of silly mistake. Perhaps when we wrote down the numbers in the notebook we reversed the positions of two digits. (It happens!). Perhaps we mistook the top of a white cloud for the snowbound peak of Ben Nevis. Whatever the reason may be, we are justified in declaring that result to be an "outlier" and ignoring it completely.

Example-2 The Himalayan foothills. For this example we change location. We are again measuring the heights of mountains but now we are in the northern plains of India with the foothills of the Himalayas lying a few miles to the north of us. We are in a hurry, so this time we make only one measurement of each mountain. We do 20 measurements in all. Again we go home to do the required calculations.

This time we find that we have 19 measurements which yield similar but by no means identical results. That is because each measurement is of a single (different) mountain. They are similar only because all these mountains are part of the same mountain range. So those 19 measurements give us mountain heights ranging from 6000 feet to 7000 feet.

But the 20th reading tells us that the mountain concerned is 29,141 feet high. That is more than 4 times higher than any of the other peaks. So is it an outlier? Would we be justified in eliminating that set of readings from our set of results? It could be an error due to some silly mistake. But then again it could

just be that through a valley in the foothills and through a gap in the clouds, we have actually taken measurements of Mount Everest some 150 miles further north. What is certainly the case is that we would NOT be justified in eliminating it from our readings without further checks.

What is the difference? When we were dealing with Ben Nevis all our readings were of a single mountain and we therefore had a reasonable expectation that they would all be similar. The group of 5 measurements gave us an estimate of what that reading should be. The sixth reading could therefore be identified as an outlier.

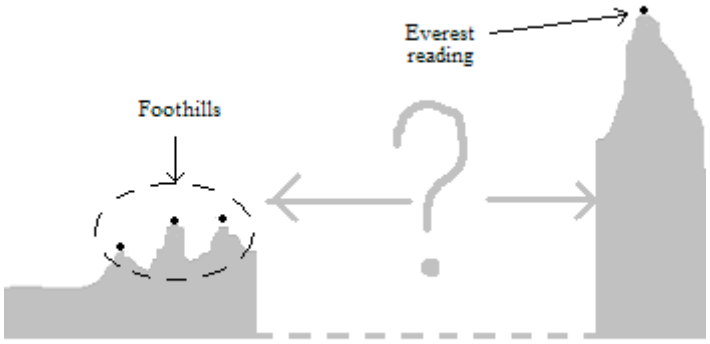

FIG 10

For the Himalayan data, however, all the readings refer to ***different*** mountains and therefore we have no reason to expect them all to be clustered around a common average value. In these circumstances the Everest reading might be a little startling but we have no justifiable reason to reject it as an outlier. The least we could do would be to go back to the location where we made the readings and check again.

In the Spirit Level all the datum points shown in its graphs, refer to different countries (or different states in the USA). So we have no prior reason to expect them all to have similar values (of whatever "social ill" is being examined) unless they have similar levels of inequality.

But we are not finished. The Spirit Level data corresponds to a third example.

Example-3 The Himalayan Slope. We are now back in the plains of India and we are trying to check the validity of a geological theory about how mountain ranges are formed. This theory suggests that the successive ranges of the Himalayas rise in a series, from the lower foothills to the highest peaks at the most northerly edge of the range. The diagram below (FIG 11) illustrates.

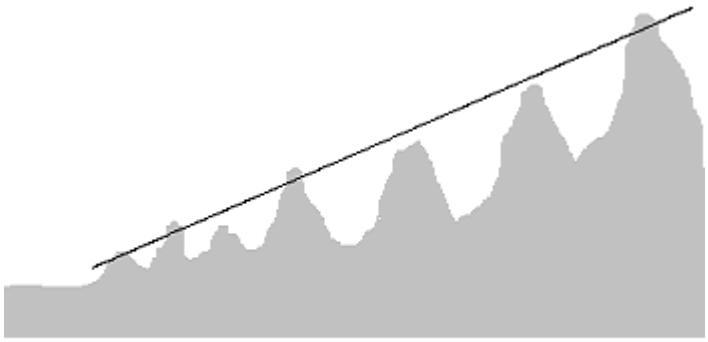

FIG 11

Now suppose that because of cloud we are unable to see the intermediate ranges so that the picture we get looks like this.

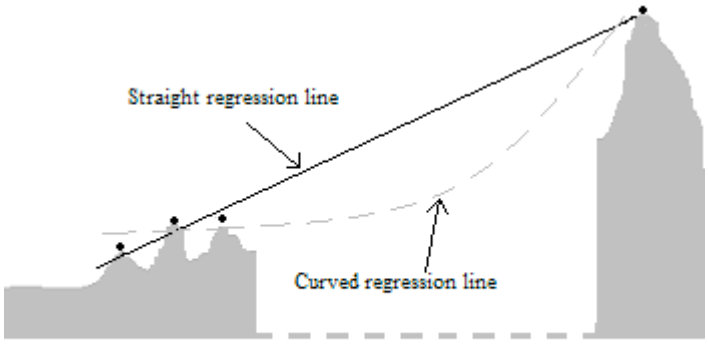

FIG 12

Mount Everest is now isolated from the other readings. In that position, and because there are no other readings from that part of the range, it will dictate the slope of line from foothills to the highest peaks (the solid black line). But a curved regression line is also plausible. There is simply not enough information in the data to reach a reliable decision on the issue. If we eliminate Mt Everest we may get a very different slope of line which will be determined by the foothills alone and one which will probably be flat. This is much closer to the situation we have in the Spirit Level, with the USA at the extreme right hand side of the graph and dominating the slope of the line.

Since publication of The Spirit Level, W&P have introduced new data which lend strength to their contention. They have also drawn attention to the correlation they have found between inequality and social ills among the various states of the USA. They have, therefore, theoretical reasons for expecting a correlation of the kind they claim. In the absence of intermediate data, however, it is also possible that the true regression line might be better shown by the curved dotted line on the diagram.

NOTE: There may be legitimate reasons to exclude Mt Everest from the data set. We could, for example, be interested only in the foothills. In that case,

however, exclusion would be justified by its geographical location, not because its height was different from the others.

2.15 Residuals and the Identification of Outliers.

When we are dealing with data of this kind we cannot assign significance to the disparity between the norm of the majority of readings and one or two isolated readings like the USA (or Mt Everest) if they are not estimates of the same datum value as the other readings. However, if all these points are expected to lie on a single regression line, then we can use the regression line itself (rather than the average of all the other points) as the common value to which we can expect them all to conform. That is, we can use residuals (deviations from the regression line) and not deviations from the average value of all points, as a criterion for identifying (possible) outliers. So while we cannot compare the height of Mount Everest with the heights of the foothills, we can look at the way it differs in height from the value we would expect if it did lie exactly on the regression line. In other words we can look at the residuals.

Identifying Outliers (Saunders style) In his critique of W&P, Saunders looked at the graph showing the plot of homicides per 100,000 population against inequality. It looks like this

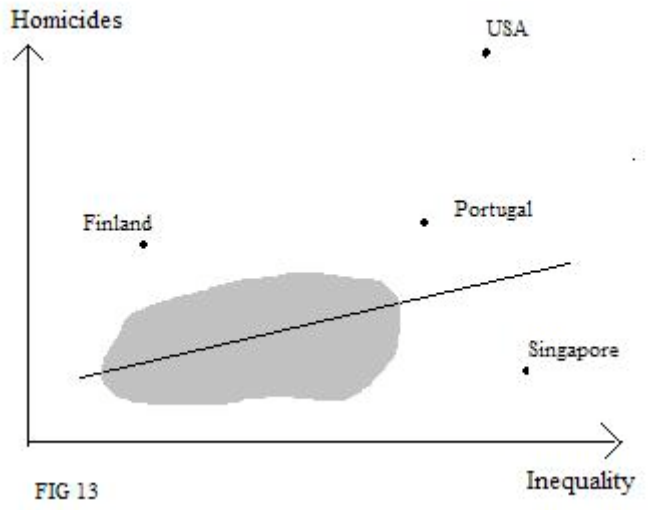

FIG 13

The grey blob represents a cluster of datum points which correspond to Sweden, France, Canada, Australia, UK, Norway, Germany, Greece, New Zealand, Italy, Switzerland, Belgium, Japan, Austria, Denmark, Israel, Spain and Ireland. I have not shown these individually because the exact position they occupy is not important to the point I am making here. The important

point is that they are clustered in that way and collectively (like the Himalayan foothills) and they do not show a noticeable regression line sloping upwards to the right. I have shown the approximate positions of four countries Finland, Portugal, Singapore and the USA. The effect of two of these (Finland and Singapore) is to make the regression line more horizontal and the effect of the other two (USA and Portugal) is to give the regression line a more steeply incline slope upwards to the right.

Saunders is concerned about the undue influence which the position of the USA and Portugal exert on the slope of the regression line but does not mention Singapore or Finland. Here are Saunders' exact words -

But look at the scatter of countries on the vertical (y) axis in figure 5a. Most of them seem to have homicide rates which are compressed in a range between about 10 to 20 murders per 100,000. The glaring exception is the USA ... with its homicide rate of over 60 per 100,000. Judging by this graph we might expect that the USA is a unique case, and that its exceptionally high homicide rate is being caused by factors which are specific to that one country alone (the laxity of gun control laws is an obvious explanation). [PS p29]

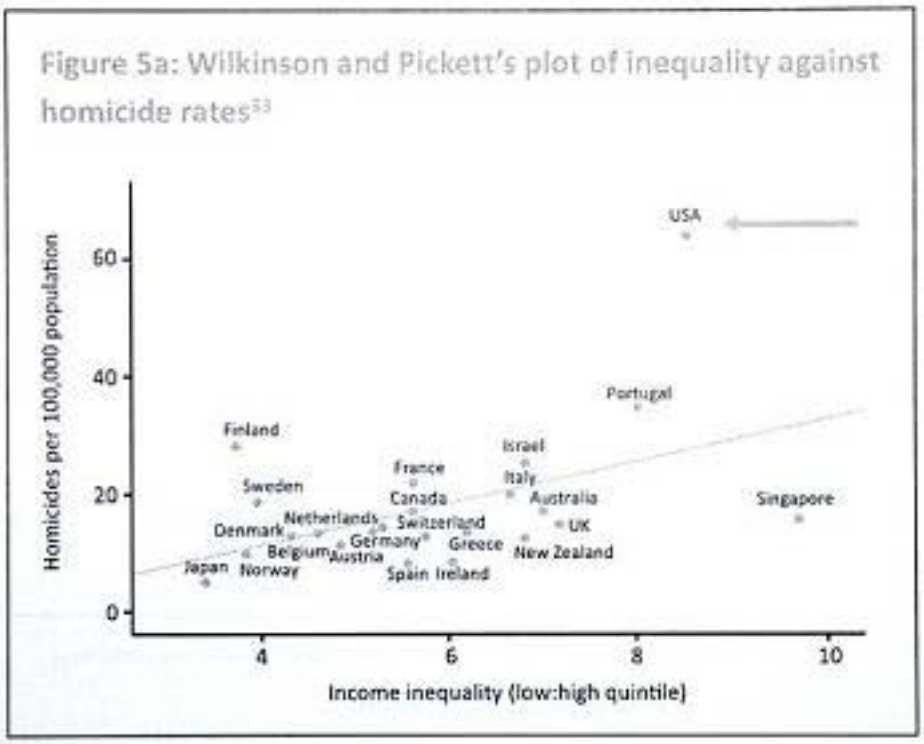

FIG 14 (my Fig 14)

To ensure that the reader does not think I am misinterpreting Saunders' argument I show here a photograph of his illustration (my FIG 14).

What he appears to be doing here is comparing the US homicide rate with the average homicide rates of all the other countries. That's like comparing the height of Mount Everest with the average height of the Himalayan foothills. Let's call that "The Nevis-Everest Mistake.

What he should be doing, is comparing the discrepancy between the US homicide rate and the homicide rate predicted for it *by the regression line*, with the average *discrepancy* of the other points. That is, he should be comparing the residual of the USA datum point with the average *residual* of the other data.

"Saunders repeats that mistake several times. See for example, page 66 where he discusses the elimination of Singapore (which he describes as an outlier), even although it sits squarely on the regression line.

2.16 Boxplots

Saunders again -

There is a simple test we can run to detect what statistician [sic] *call 'outliers' in any distribution of data. It is called a 'boxplot', and it provides a visual representation of how cases are distributed on any given variable.*

[PS p29]

He continues with these words -

There is no need to go into details of how to interpret a boxplot, other than to note that 'outliers' are identified by a circle and 'extreme outliers' by an asterisk. We can see from this example that Portugal is an 'outlier' and the USA is an 'extreme outlier' when it comes to murder rates. [PS p30]

Here is a re-drawing of his boxplot -

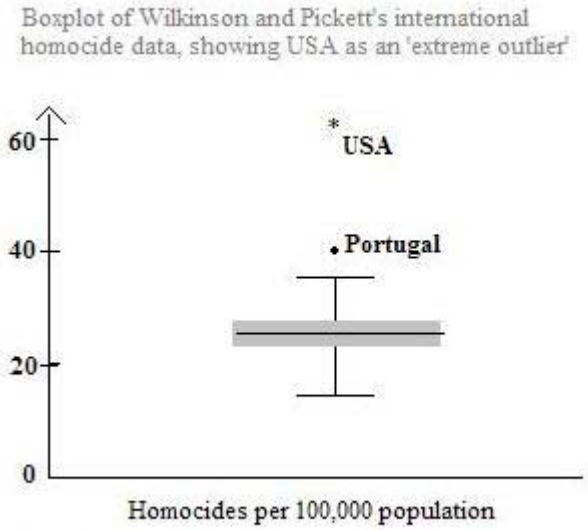

FIG 15

2.17 Boxplots and the SatNav Mistake

According to Saunders the boxplot is "*a test*" which we can use to identify outliers. That is not true. A boxplot is NOT a *test* of anything. It is a way of presenting data for ease of visual inspection - like a piechart or a histogram. A boxplot shows the datum points which lie beyond certain limits (relative to the standard deviation of the distribution). But since the boxplot knows nothing at all about what the data signify, it cannot decide for us which points are outliers and which are not. That decision remains our own responsibility. All a boxplot can do is to identify the points which are *candidates* for detailed consideration on the criteria *which have been chosen by ourselves*. Saunders however regards it as a test which identifies outliers without the need for our own contribution to the decision. I quote -

a boxplot identifies the USA as an outlier. [PS p49] and [PS p52]

Sure enough, a bloxplot confirms that these two [USA and Singapore] *are indeed outliers.* [PS p66]

This is equivalent to thinking that a SatNav device can not only help us to reach our destination, but is also able, in some mysterious way, to choose that destination for us. Nasal electronic voice: "*You have input the postcode for London. I have changed your destination to Glencoe in the Scottish Highlands. The scenery is better.*"

SatNavs are helpful, but they are not *that* helpful. Then again, perhaps Microsoft would approve of a device (like a bug-eyed paperclip) which kept changing automatically the postcode of you destination. Let's call that "The SatNav Mistake".

As we have seen from the paragraphs above, USA and Portugal have been identified by Saunders as outliers (but not Singapore or Finland).

Finland and Singapore. It is quite clearly seen in the diagram (FIG 14) that Finland and Singapore are both further from the regression line than Portugal.

2.18 The USA data are not wrong.

Note that there is no suggestion that the datum point relating to the USA is wrong. Its unusual location on the graph is not caused (as was that rogue 6th reading of Ben Nevis) by some trivial error in measurement. The USA really does have that degree of inequality and it really does have that number of homicides. There is also no reason to suspect that homicides data should have a

normal distribution. So the use of a boxplot which assumes normal distribution, is quite inappropriate.

The USA is a valid datum point which merits its position in the graph plot. Like Mount Everest, it is simply different from the rest. Saunders' protest is that it occupies that location not because of its inequality but because of "other factors". We should note however that "other factors" are also present in every other datum point in the graph plot. Different countries have different types of gun laws. Think for a moment about Singapore.

If we ignored Singapore we would be able to draw a very simple curved line through the scatter diagram. It would pass very close to all of the other points on the graph and it would actually pass right through both Portugal and the USA. Singapore is spoiling that relationship. So why might Singapore be different from the rest?

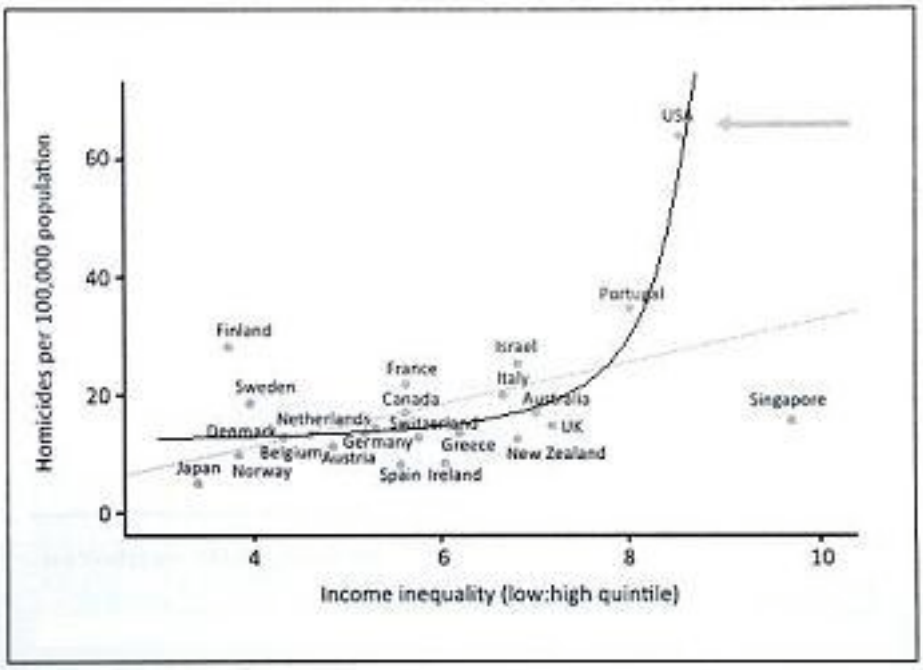

FIG 16

Singapore. Unlike all the other countries in the sample of 23 used by W&P, Singapore consists almost entirely of a single large city. Its population is just over 3 million (the cut-off point used by W&P to eliminate tax havens) so it was very nearly excluded. It has a very strict regime of law enforcement (including a death penalty for drug trafficking). Note that New York is also an anomalous datum point on the graphs relating to the various US states with a relatively low rate of crime (for the USA) despite its high level of inequality. Until recently New York had a very high rate of urban crime. But a recent "clamp down" has reversed that position. Clearly strict enforcement of law can have an effect (at a cost).

Singapore became an independent country as recently as 1965 when it parted from the Malaysian Nation. Its economy is very heavily dependent upon international trade. Unlike the USA Singapore has a healthcare system, which is accessible to all of its citizens. Generally speaking, the style of government is paternalistic and its efforts are aided by the relatively homogeneous environmental conditions. Policy does not have to include provision for a very large agrarian hinterland. Nearly everyone is employed in commercial enterprises related to international trade. Singapore may be unequal in terms of income, but it has a remarkable degree of equality in some other respects. These factors make it a very "different" social community from the others in the sample. If there is a reason therefore for excluding any country from the sample set based on "other factors", Singapore is the obvious candidate.

Note this - I am not suggesting that we have sufficient evidence to say that the true regression line is curved like the one shown in FIG 16. What I am saying is that the case for removing Singapore from the dataset is every bit as valid as that for removing the USA from the scatter diagram. I am also saying and that a curved regression line, like the one shown, is quite plausible.

But perhaps the safest policy, is the one adopted by W&P. Having chosen a sample set on fixed criteria (without regard to any theorizing about inequality) W&P stuck to that set and accepted the results which emerged. It is really not legitimate to remove points from a graph when there is no reason to think that the associated data are somehow in error. It is doubly unacceptable to remove points after it is found that those points somehow spoil a preferred interpretation of the data.

A boxplot *is* able to identify *potential* outliers only by comparing their residuals with the standard deviation of the distribution of *residuals* and assuming the they are normally distributed about the regression line. Since Saunders, at a later point in his document, casts doubt of the assumption of normal distribution, he appears to be using the assumption when it suits him and abandoning it when it does not.

Normal distribution for the residuals is only one possibility. In some circumstances it is found that the standard deviation (of the dependent variable - Y) increases with the value of Y or with the value of X. If that was the case then we would expect the standard deviation to be much greater when we are dealing with countries which are more unequal.

2.19 Multivariate Regression.

A claim made repeatedly by Saunders is that the various social ills which W&P have associated with inequality, are actually caused by "other factors". He also

claims that he can show this to be the case by using a form of analysis called a multivariate regression. Just as it is possible to plot the regression line of one variable against another, it is possible to extend that approach to three or even more variables.

Consider, for example, the correlation which undoubtedly exists between the age of a school pupil and the size of that pupil's vocabulary. We might also argue that the size of a pupil's vocabulary is also influenced by the number of books that pupil has read. So we could draw two scatter diagrams -

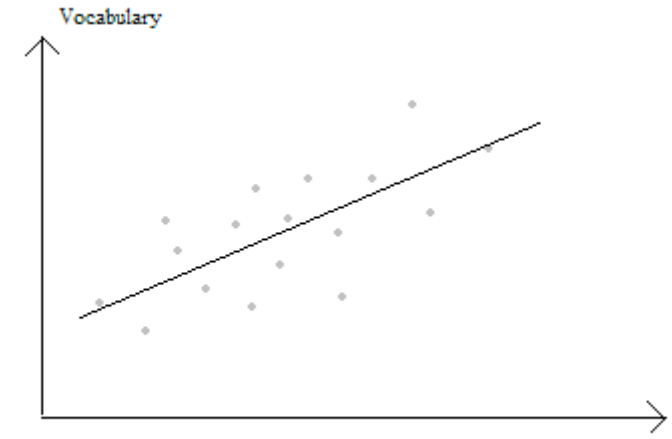

FIG 17

(All pupils have read the same number of books)

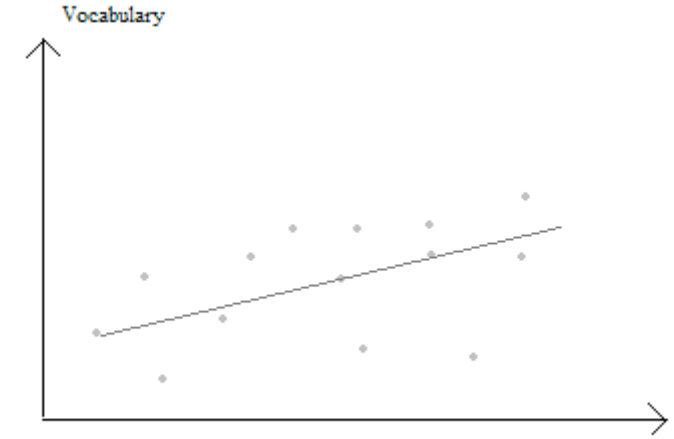

FIG 18

(All pupils are the same age).

Now let's see what happens when these two graph plots are put together.

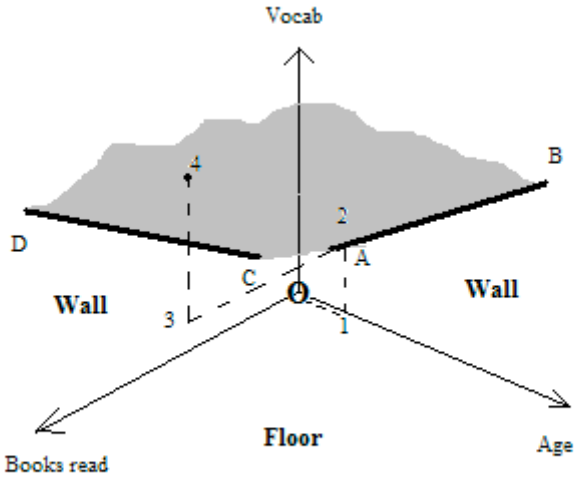

FIG 19

To "read" this graph plot you must imagine that the central point marked "O" is the far corner of a room and that the "Vocab" axis is the vertical corner rising from that point. The surface bounded by the two axes "Age" and "books read", is the floor of the room, and the two other surfaces (age, vocab) and (books-read, vocab) are the walls which meet at that far corner. The two lines A-B and C-D are the two individual graph-plots shown on the two diagrams above. These have been drawn on the two walls which are at right angles to each other. What you must then imagine is that there is a rubber sheet (shaded grey) stretched between the two lines A-B and C-D. This sheet is the plot of the values of a child's vocabulary plotted against the possible values of age and books-read. To get the position of a single point (4) on the rubber sheet, for a given child, follow the journey along the dashed lines 0->1->2->3->4. The first stage (0->1) is the distance along the age axis (corresponding to his age). 1->2 represents the size of his vocabulary (for a child of that age) as indicated by the regression line A-B. 2->3 is a line drawn parallel to the books-read axis. The length of that line represents the number of books the child has read (on the book-reading axis). We can then see from the C-D regression line how far we must travel upwards 3->4 for that number of books read.

Point 4 represents a prediction of the size of that child's vocabulary, given his age and number of books read. The contribution of each factor to that total, can be read directly from the position of the point.

The variables represented by the two axis at floor level (age and books-read) are called the "independent" variables. The vertical axis represents the value of the "dependent variable" (vocabulary). To use multivariate analysis you go in the reverse direction. You start with a scatter of points in three dimensional space, you find the "best-fit" flat-sheet (or a curved sheet) through those points and then you see where that sheet cuts the two walls. The two lines of

intersection with the two walls are the regression lines of each of the independent variables (while the other is held fixed).

In practice a graphical method like this is not used. The complexity of trying to represent a three-dimensional relationship on two-dimensional paper is too great. Even worse would be an attempt to draw a graph in four dimensions if there were three independent variables. The data are usually presented in the form of tables and with that format we can extend the method to more than 3 variables.

2.20 The Independence of "Independent" Variables

The diagram above, however, illustrates a principle that underlies multivariate analysis. It shows why it is important that the two "walls" of the three dimensional plot should be at right angles to one another. If they are not then that implies that the two so-called "independent variables" are not really independent at all. When that is the case, movement along one of these "independent" axes, automatically causes movement along another. The contributions made by each then become entangled and are hard to separate.

In this example a child's age and the number of books read are not truly independent. You would expect a child to read more books as age increases. As an illustration of good practice, therefore, my example is not a good one but it does illustrate the problem quite clearly caused by "independent" variables which are not really independent.

Consider this more extreme example - A statistician has counted the number of people who have died each year in a certain coastal holiday resort. He has also counted the number of people who have fallen over the edge of a high cliff in that location, during the same periods. He has plotted the figures for each year and declared that there is a correlation between the number of "fallings over a cliff" and the number of deaths, over several years. (Not an unreasonable proposition you may think). The correlation, of course, is not perfect because quite a lot of people will have died from other causes - like being frozen to death while sitting in deck chairs, being poisoned by boarding-house cuisine and by having consumed an excessive amount of ice cream. However, another statistician claims that even that degree of correlation is spurious. "There is no causal relationship between falling over a cliff and death," he claims. "Those deaths were caused by another factor. What really kills some people is not falling over a cliff but landing at the bottom of a cliff."

The statement is obviously true but it does not amount to a refutation of the first claim because it ignores the close causal relationship between falling and landing. These are not independent variables. A count of the number of fallings

is a pretty good measure of the number of landings. It may ignore the small number of people who saved themselves by catching on to a bush halfway down, but the two counts will be very close.

Although the claim made by Saunders about "other factors" causing the extreme position of the USA in the graphs shown by W&P, may not be as absurd as my example of falling over cliffs, it does have an element of the same faulty logic. The fact that the USA has lax gun laws is not unrelated to the fact that the USA is a very unequal society. There is a cause-and-effect linkage between them. Given the horrifying rate of gun-related homicides in the USA it might have been expected that the population as a whole would rise up and demand strict gun control by law. That is what happened in the UK after the Dunblane school massacre. The fact that there is no effective popular clamour for gun control in the USA, despite the Columbine massacre and a number of copycat killings, speaks volumes about the prevalent attitude about guns in the USA - an attitude which places reliance on individual self-help rather than collective action. It is characterized by a general disdain for those who, for one reason or another, are unable or unwilling to achieve what is seen as an adequate level of self-help. This, of course, is not true of the whole population in the USA, but it does seem to be the attitude of a section of the population which has great political influence. An abhorrence of gun control seems to be a common feature among those who are intolerant of any collective action to promote social welfare (and equality).

Inequality of income and a lack of gun control laws are in effect, proxy measures (or proxy indications) of the same thing - an attitude of tolerance towards social inequality and a willingness to rely on self-help rather than collective social responsibility.

Multivariate analysis is effective only when the so-called "independent" variables are genuinely independent of one another, or nearly so. This is a point which Saunders appears to have ignored. Let's call this the "Falling-Landing Mistake".

2.21 Charitable Donations

In the Spirit Level, W&P compared the donations which each of the 23 rich countries in their sample, made to third world underdeveloped countries. The data (in terms of donation/head of population) showed quite clearly that the more unequal the country, the smaller was the charitable foreign aid donation. Saunders has challenged this conclusion. He attributes the significance of the regression line to the Scandinavian countries. When these are removed from consideration, he claims, the significance disappears.

Here again we see Saunders' approach to these data. When the USA produces the result he does not want it is removed. When the Scandinavian countries produce results he does not want they are removed. The diagram in question is FIG 20 below (re-drawn from The Spirit Level).

Saunders wants the Scandinavian countries removed and claims that that would eliminate the significance of the regression line drawn by W&P. I point out that removal of the UK, Japan and Finland would restore the significance of the reverse correlation between inequality and generosity of donations. I am not suggesting that that is a reasonable thing to do. I point out only that it is no less reasonable than removing the Scandinavian countries.

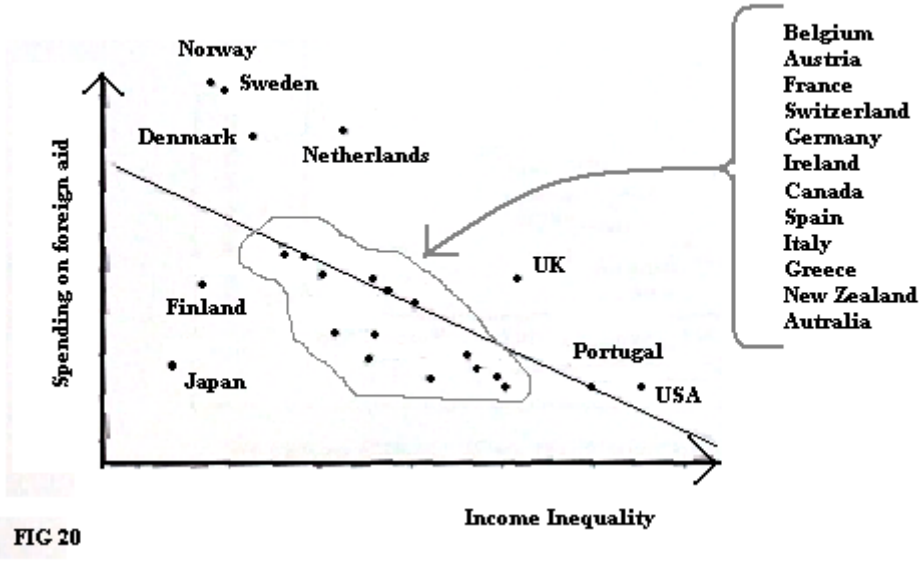

FIG 20

To support his argument Saunders has presented data about the charitable donations made by the citizens of each country, as *individuals*, to charitable causes. These data are presented to us as a histogram, which indicates the complete reverse of the graph shown above. Based on these data, the USA would appear to be very much more generous than other countries. Saunders obtained the information on which his analysis and conclusions are based from a report published by CAF (Charities Aid Foundation) in 2006. The reference to this was given in Saunders' in a side-note (in very small print). I used a magnifying glass and took the trouble to look up that report and found a number of caveats which warn us not to interpret the data in the way that Saunders has chosen to do. Here are Saunders' words -

The generosity of the people has nothing to do with how much their politicians spend and when it comes to voluntary activity the Anglophone cultures appear to be the most generous in the world. [PS p39]

And here are two quotations from the CAF report -

The evidence ... suggests that personal tax might well be an important factor in giving levels; however, it is the level of social security contribution and not personal taxation which seemed most significant. Amongst EU members in the survey an inverse relationship between average social security contribution as a proportion (%) of GDP was noted. [CAF p8]

and

Differences are also to be explained by the importance attached to charitable giving in different cultures. In countries such as the Netherlands, France, Sweden (not included in this survey), there is a strong belief that governments rather than charities should provide for social needs, whereas in the US, and increasingly in the UK, charities assume an important role in meeting the needs of socially excluded groups. [CAF p12]

It is probably the case, as Saunders says, that the *generosity* of people has nothing to do with how much their politicians spend, but according to the report on which he has based his argument, the amount individuals actually *donate* and the *causes to which they choose to donate*, appears to take cognizance of the way national governments (through taxation) give support to various causes. People give where they see a need to give. In the USA the need is considerable and close to home. So individuals who are generous, try to make good the lack of generosity in their government's provision for disadvantaged groups within their own population - and perhaps they also try to compensate for a lack of generosity in the rest of the population who voted for such restricted generosity in terms of tax and benefits for the underprivileged.

The CAF report notes that a large proportion of individuals giving within the USA consists of donations to local religious organizations. Much of the giving is selective. Among the rich nations, however, the USA (as individuals) gives the lowest proportion of its income per head, in the form of foreign aid.

I feel we should also note that in the USA there is a culture of making ostentatious personal gifts to local charitable causes at public functions. The narrow and non-universal nature of these gifts makes me suspect that less attention is being paid to the needs of the underprivileged (in general), than to the social kudos being gained by the donor.

In an interesting aside the CAF reports -

... there is evidence in the UK that poorer people give away higher proportions of their income than the rich. [CAF p12]

2.22 Time Trend Analysis

Towards the end of his critique, Peter Saunders presents a number of graphs, most of which relate to various types of recorded crime (homicide, violence robberies, burglaries and total crime) in the UK. These are plotted against time. He compares these trend-lines with the trend lines over the same time period, for life expectancy, and income inequality. Over 4 decades, all of these graphs show an upward trend. But the upward slope is punctuated here and there by dips. Saunders draws our attention to the fact that these dips do not correspond closely with changes in income inequality.

This observation, however, does not contradict the basic W&P thesis. W&P do not claim that inequality is the only factor which influences well being. Health, in particular (and therefore life expectancy) is strongly influenced by medical advances and improvements to public sanitation. Better vaccination regimes and new cancer screening programmes will also improve things, just as new viral epidemics, or super-bugs in hospitals, will make things worse. This is to be expected. Temporary fluctuations in a general upward trend are of little significance to any examination of the long-term relationship such as that between income inequality and social welfare. Short-term dips and wiggles are no more relevant to that kind of analysis than short-term cold spells are of relevance to the general upward trend of global warming. A sudden dip in car theft statistics, for example, could be caused by the introduction of better anti-theft devices into the standard inventory of mass produced cars.

So time-trends are not the best way to examine these relationships. No one disputes that "other factors" are involved in the occurrence of social problems. It is, moreover, impossible to control or eliminate those other factors. The best we can do is to take a sample which is large enough and sufficiently representative to make it likely that those other factors will average themselves out in the manner of swings and roundabouts. That canceling-out effect, however, does not apply to a time trend analysis.

For a time trend graph, time itself varies along the line and with it, any other factor like the discovery of a new drug or the introduction of a new car theft device. The datum points which occur before and after such an event are not directly comparable. Unlike a sample snapshot, a time trend graph is strongly influenced by these "other factors" in a way that totally obscures the factors we are trying to analyse. A time-trend graph has its uses, but, I repeat, it is not an appropriate way to analyse these issues.

3. Parametric vs Non-Parametric Statistics

3.1 Parametric Statistics

A parameter is a constant variable (or a variable constant). Like a variable its value can change, but unlike the quantities for which we usually reserve the term "variable", it does not change during a particular investigation or analysis. A parameter is a value (usually numeric) which defines the context of some set of circumstances. In conventional statistical analysis, the values of the parameters define the type of populations and distributions with which we are dealing. Those parameters therefore define certain assumptions which we are obliged to make if we want to use the standard techniques of parametric statistics. Sometimes, we make these assumptions unwittingly (like the assumption that residuals have a normal distribution), when this is quite inappropriate. If, for example, we invert data (as we do when we use miles-per-gallon rather than gallons-per-mile) it is the case that if one of those variables is distributed normally the other cannot be.

Note: If you find that hard to believe, consider this simple arithmetic example. Take the values (1,2,3,4,5) as the datum point values. Average =3 The datum points are distributed symmetrically about the average value.

1……2……(3)…….4…….5

Now invert those data to get the data values (1, 0.5, 0.333, 0.25, 0.2). Average = 0.4566. The datum points are not distributed symmetrically.

…….a…b…..c…()……d……e

where a=0.2, b=0.25, c=0.33, d=0.5, e=1 and () = 0.4566 (average value)

So - should we be talking in terms of (homicides per 100,000 people) or (persons per homicide)?

3.2 Non-Parametric Statistics

Non-parametric statistics have been developed to provide tests of significance which avoid most of these basic assumptions. Non-parametric statistics are particularly useful in the social sciences where the data collected do not consist of numeric measurements found by using measuring tapes or any other kind of instrument. In the social sciences it is often the case that the data represent human judgements on the relative values of various quantities. Every time a panel of judges tries to assess the relative merits of say paintings, or piano recitals, or the acceptability of political policies, they are making judgements

about the relative value of these things. They cannot assign numeric values to the merit of a pianist's playing. All they can do is to say that pianist A was better than pianist B, and so on.

Sometimes we can make these relative value judgements look as if they were numeric measurements by asking the judges to place each performance on (say) a scale of 1 to 10.. But these are really not absolute numeric measurements since each judge is using his or her own (non-standardized) measuring scale.

3.3 Rank Correlation Tests

Suppose we have four pianists (called A,B,C and D) and one judge. Each pianist plays two pieces of music called M1 and M2 and for each piece of music our solitary judge tries to place each pianist in order of merit. For the first piece of music (M1) the order is (A,B,C,D) and for the second piece of music (M2) the order is (B,C,A,D). What we need is a way to decide whether or not there is some consistency about these two orders of merit. That is, we want to know if that judge is really judging the merit of the pianists in a way that is independent of the music they are playing, or whether the apparent merit is entirely dependent on the piece involved or even a completely random selection.

The data we have at our disposal provides us with a number of partial orderings. That is, we can say (for M1) that -

M1 = (A,B,C,D)

A is better than B
A is better than C
A is better than D
B is better than C
B is better than D
C is better than D

That is six partial orders.

We have exactly the same number of partial orders for M2. These are

M2 = (B,C,A,D)

B is better than C
B is better than A
B is better than D
C is better than A

C is better than D
A is better than D

Now, if we assume that such agreement that there is, could easily have arisen by chance, that would be the same thing as assuming that each of these partial orders has been decided by the toss of a coin. Each partial order in list M2 is then either the same or the opposite of one of the partial orderings in list M1. So we could say it is either a "*head*" or a "*tail*" - a *head* if it is the same as in M1 and a *tail* if it is reversed.

So the question boils down to this - If we toss a coin four times, how likely is it that we would get four heads? How likely is it that we would get four tails? Alternatively, how often would we get one head and three tails? Two heads and two tails? Three heads and one tail?

We can work out the probability of each of these outcomes quite easily by counting how many different ways we could get each of these outcomes. There is, for example, only one way that we could obtain four heads. But there are four ways in which we could obtain one head and three tails. So this is a way of deciding how likely it is that we would get any given degree of agreement between list M1 and list M2. And that is exactly how these non-parametric rank order tests work.

3.4 Rank Ordering of the Spirit Level data.

I decided to try a non-parametric test on a sample of the data used by Wilkinson and Pickett. Here again we have two lists. In this case we don't have a judge making decisions, we have the actual data about the income inequality of several countries and we have a list of countries ordered by their homicide rates (for example). So again we have two lists and we can still call them M1 and M2 if we want. For the first of our two lists (on **income inequality)**, we can say that -

USA is more unequal than Portugal
USA is more unequal than UK
USA is more unequal than Australia
USA is more unequal than NZ
USA is more unequal than Israel
USA is more unequal than Italy
and so on, down to
Finland is more unequal than Japan

Our second list refers to a particular **social problem.** It goes like this -

USA is worse than Portugal

USA is worse than UK
USA is worse than Greece
USA is worse than NZ
and so on, down to
Sweden is worse than Japan

So now we can do exactly the same test as we did for the piano players by asking how likely is it that the degree of agreement we see between these two lists is down to chance (or the toss of a coin). Note that this test does not involve the use of any raw numeric values at all. There are no residuals. So we are not obliged to make any assumptions about how the residuals are distributed. There is no assumption about the linearity of the relationship and the concept of an "outlier" is not relevant.

There are two well known forms of rank correlation tests - one attributed to Spearman and the other to Kendall. Both are famous statisticians of the last century. For this exercise I have used the Kendall Rank Correlation Test. The software I have used will be found online at:

www.wessa.net/rwasp_kendall.wasp

or perhaps more easily by supplying Google with the keywords/phrases:

"Kendall rank correlation", "free statistics and forecasting software"
(You should include the quotation marks, as shown.)

Another advantage of these rank correlation tests is that it is relatively easy to extract the required data from the graphs published in The Spirit Level. The rank order of countries with respect to inequality is provided in an appendix. To get a second list representing rank order in terms of (say) prison populations, or homicides per 100,000, or the level of trust in fellow citizens, one need only lay a ruler horizontally on the page at the top, slide it down, and write down the name of each country as it appears from under the ruler's edge. When two datum points appear to be equal, the test can cope with that, but all I did was to toss a coin to decide which came first. Reversal of the relative positions of one or two equal points does not make a significant difference to the result.

3.5 TAU

I did not examine all of the data used by W&P but I did calculate the Kendall Rank Correlation or "Tau" value for

(1) The Index of Social Problems
and
(2) Homicide rates

Saunders questioned the validity of the index of social ills but could not deny the correlation with inequality with respect to the Index. W&P have since demolished his criticisms. Saunders also claimed that the correlation found by W&P with respect to homicide rates was spurious and due to the "outlier" position of the USA. He did this by removing the USA from the dataset (but not Finland or Singapore).

In both these cases, using the Kendall Rank Correlation test I found that the value of Tau was positive and significant. For both, the value of "p" (the probability that the result could have arisen by chance) was significant at the 1% confidence level. For the homicide data, the value of p was 0.00014 while for the Index of Social Problems, the value of p was $3.5 \times e^{-5}$ These values leave no room for doubt. The correlations are statistically highly significant.

I repeat, the Kendall Rank Correlation Test does not make any prior assumptions about linearity. The concept of residuals, of residual distribution and of outliers are not applicable. Furthermore, the calculation of Tau does not tell us what the underlying causal relationship could be. But this is clear - there is some kind of significant and positive relationship between these social ills and inequality of income and between homicide rates and income inequality.

4. Discussion

4.1 Saunders - his critique

The critical analysis of The Spirit Level which Peter Saunders has offered us cannot be taken seriously because it contains so many serious technical flaws. He describes the Spirit Level as "bad sociology". I would describe his account as "very bad statistical analysis". The errors I identified were -

(1) Arbitrary removal of datum points to produce the results he prefers.
(2) Using of multivariate analysis on so-called "independent" variates, which are not in fact independent at all (the falling-landing mistake)
(3) Using boxplots on raw data instead of residuals (Nevis-Everest mistake).
(4) Claiming a boxplot can *identify* an "outlier". (The SatNav mistake)
(5) He uses time-trend analysis in an inappropriate way.

There are other mistakes but Wilkinson and Pickett have identified these and given a robust response which is available on the EQUALITY TRUST web site. These mistakes discredit Saunders' analysis. I conclude that the thesis offered us by Wilkinson and Pickett, in their book The Spirit Level, remains standing in the face of this criticism and stands largely unscathed.

4.2 Reservations

I do have some reservations of my own concerning The Spirit Level.

(1) Linearity. W&P did not claim explicitly that the relationships between income inequality and each of the various social problems are linear. The alternatives, however, were not discussed by them.

(2) Datum Points. I am uncomfortable with regression lines which depend so much on a few datum points which stand out from the rest. I do not suggest, as Saunders does, that these (or the other ones which spoiled his preferred thesis) should be removed from the analysis. But I do think some discussion is required into why there are such marked differences. I am thinking particularly of the difference between the USA and Singapore despite having a similar level of income inequality. I am also thinking of New York, which frequently departs from the trend line in W&P's various analyses of the US states. As a measure of income inequality, W&P used the ratio - income of the top 20 percentile / bottom 20 percentile. This is a figure which is available internationally. I suspect, however, that while (on that measure) the difference between the USA and Singapore, is not great, a much greater difference might be shown if we compared them using a figure given by the top 1 percentile compared with the rest of the population.

(3) Time. I wish W&P had considered the effect of time on the relationship between inequality and the various social ills. There is almost certainly not an immediate causal connection in relation to every factor. It takes nearly two decades for the disadvantage which a child suffers during his or her education, to have an easily observed effect on that person's potential earning power as an adult, whereas the effect on the child's educational performance will be immediate. It is likely that each of the social ills has a different time-course.

(4) Causal Chains. Earlier (in section 2.6) I discussed the possibility of "causal chains" which could connect inequality with one or more of the social ills in question. I also gave a hypothetical example of such a chain linking inequality with taxation, crime, prisons, and so on, back to income inequality. The chain was therefore circular. I suspect, however, that even the concept of a causal-chain is too simplistic. What is more likely, is a causal-web - a complex interconnected network of cause and effect. It is also the case that once a causal chain has been established, it is very difficult to see how circumstances can be changed, if only one factor is modified. It looks to me that all the elements on the chain would need to be changed at the same time.

(5) Better for everyone. The subtitle of The Spirit Level is "Why Equality is Better for Everyone." It was that claim - that equality not only benefits those at the poor end of the income spectrum, but those at the top as well - that lifted the book out of the general run of left-wing political commentaries, and galvanized the left-right debate. However, any judgement about what constitutes a "better" society is highly subjective. Some agree with John Donne that "*Any man's death diminishes me, because I am involved in Mankind.*" (For death read impoverishment, loss of job, mortgage foreclosure etc.) Others take the view that another man's death is that man's problem, not mine. The extent to which one of those attitudes predominates over the other must have a role in determining the extent to which a society tolerates or even promotes inequality.

(6) The psychological effects of inequality on the better off. In The Spirit Level, W&P discussed the psychological effects of inequality on disadvantaged people. In particular they quoted evidence about the dispiriting effect that the perception of a low-caste status has on children's educational performance.

What they did not discuss, however, is the effect that the perception of high-caste status has on those at the top of the income pecking order. One obvious effect is that they tend to attribute their own success to "working hard and taking risks". Another effect - and this one seems to go quite far down the pecking-order - is to develop a disdain for those further down the gradient. The maintenance of an acceptable self-image for some, appears to require the presence of some other group which is worse off. We might call that a Reverse Dependency Culture. That disdain, moreover, can become vindictive.

The Stanford Prison experiment illustrated that effect very clearly [4]. A group of students was divided, arbitrarily, into two sub-groups. One sub-group was placed in detention in a simulated "correctional facility". The second sub-group was given to role of prison warders. The experiment had to be terminated prematurely when some members of the "warder" sub-group became over-zealously authoritative and showed signs of becoming sadistic.

(7) The Ultra-rich. Not everyone who reaches the top of the income pecking-order succumbs to that reverse dependency culture. Some opt for high profile philanthropy. But there is a much larger number whose raison d'etre is the maintenance of their privileged position. The reason that this is important in the context of inequality and social problems, is that the ultra-inequality which they enjoy enables them to -

(i) buy political influence,
(ii) distort (by their purchasing power) the supply side of the free market, and
(iii) take ownership of the mass news media.

Once these distortions have been established, it is difficult to see how any government can break free from the stranglehold the ultra-rich can exert on political decisions. In a recent article published in the New York Times (Secretive Republican Donors are Planning Ahead, 19th Oct 2010) Kate Zernike described an ultra-right-wing group of the ultra-rich which meets secretively to plan campaigns to influence public attitudes - in their words - "to build education channels to establish widespread belief in the benefits of a free and prosperous society" and in order to counteract policies which "threaten to erode our economic freedom and transfer vast sums of money to the state."

(8) Cause and/or Effect? When we consider some of the social problems which W&P identified as being correlated with inequality, it is clear that some of these are themselves interrelated. It has often been claimed, for example, that a very significant proportion of crime is "drug related". But this interrelationship could be broken if drug use was not a criminal offence and if drug dependency was dealt with by the health care system rather than by draconian law enforcement. So it seems to me that it may not be inequality alone that we should see as the basic cause of these problems. I suspect there may be a toxic group of social characteristics (of which inequality is a major component) which are collectively responsible for a range of social problems - which then become self-perpetuating.

In conclusion - nearly a hundred years ago, The Glasgow Herald (as it was then called) published this poem by D.M.MacKenzie (who, I think, had all this suss'd).

THE SOPHIST

"The uses of adversity
Are sweet," when poor they said to me,
But now that I am rich they say
I ought to give my wealth away.
I cannot tell which way to turn,
Though to do good my heart doth yearn.
For if good to the poor I be,
He knows no more adversity,
Whose uses sweet, so oft extolled,
Him in his adversity consoled.
Shall I, to gain a selfish joy,
The sweetness of his lot destroy?
Nay, nay! for now I clearly see
'Tis best for him and best for me
That I the poor man's cry ignore,
Keep all my wealth and gather more.

References

1. The Equality Trust
http://www.equalitytrust.org.uk/
2. SLATE. The United States of Inequality
http://www.slate.com/
3. Dictionary of Statistical terms by Kendal and Buckland, 2nd Ed
Oliver and Boyd Ltd 1960
4. The Stanford Prison Experiment
http://www.prisonexp.org/

Copyright © Hugh Noble, 2010

(Version dated 30th Dec 2010)