DAMNED LIES, HALF TRUTH AND SOCIAL STATISTICS

UZOBO, Endurance*
Department of Sociology and Anthropology,
Niger Delta University, Wilberforce Island,
Bayelsa State, Nigeria
enduzobo@gmail.com, +2348065052298

And

MICHAEL-OLOMU, Oyintonyo
Department of Sociology and Anthropology,
Federal University Otuoke,
Bayelsa State, Nigeria
tonyo_mic@yahoo.com
+2348060912432

*** Correspondence Author**

ABSTRACT
In recent time, the misinterpretation of statistics represents one of the most prevalent tools in misleading the general public. Most politicians and business organisations have used statistics to attract public sympathy, create an outrage or facilitate their personal interest. This is done by either lying or manipulating data to reveal just part of the truth beneficial to them. Therefore, the aim of this paper is to review how statistics could be used to mislead the public. The scope of the paper is limited to some selected descriptive statistics namely: The mean, percentages, and graphs. The study concluded that although social statistics, could be used to deceive the public, it is still a veritable tool for understanding and clarifying the way the social world works. However, it is recommended that before any social statistics is accepted, the source of the statistics must be verified to determine if it is for public or vested interests, the whole data for the statistics must be reviewed to ensure no missing data which might reveal additional findings.

Keywords: *Social Statistics, Descriptive statistics, Mean, Percentages, Graphs*

INTRODUCTION
Recently in Nigeria, the vice-presidential debate during the electoral process raised some questions about the use of social statistics. A particular candidate reeled out important and impressive statistical figures during the debate. However, many Nigerians became angry when it was reported at some quarters that most of these figures are bogus and dubious (News Agency of Nigeria, 2018). For instance, the candidate estimated a trillion naira a year was wasted on fuel subsidy for only two million vehicles in Nigeria. However, the National Bureau of Statistics (NBS), reported that Nigeria has over 11.5million vehicles on the road, as at 2017 (News Agency of Nigeria, 2018). Obi also claimed that intra-African trade is just a mere 9 per cent. However, he was proven wrong, Intra-African trade, according to Afreximbank Africa Trade Report (2018), estimated it to be at 15 per cent. In 2016, this was estimated at 18 per cent of its total exports and imports. IMF even gave an estimate of 20 per cent of the total trade volume of about one trillion dollars (News Agency of Nigeria, 2018). Nevertheless, he went ahead in another post to give the source of his statistical data. These kinds of debate are what led Best (2001:10) to state "…… big numbers (could) arouse public outrage. Big numbers meant there was a big problem." On the other hand, small numbers could also be used by public officials to indicate that they are doing a good job. Hence, statistics can become weapons in political struggles over social problems and social policies (Best, 2001).

Taking all these into consideration, questions such as what does social statistics really tell us? And does social statistics represent the true state of affairs? are at the forefront of current debate about the nature of social statistics. Therefore, the thrust of this paper is to present ways in which descriptive statistics could be distorted not to reflect the trues nature of a phenomenon or event, and this eventually lead the public to making bad decisions. Hence, lots of social scientist are now debunking the results of social statistics (Cardinali,

2016; Best, 2001). In fact, the phrase *"**Lies, damned lies, and statistics**" popularised* by Mark Twain is now been used by critics of statistics to describe the persuasive power of numbers, particularly the use of statistics to reinforce weak propositions. The phrase is also sometimes adapted informally to doubt statistics used to prove the point of a particular debate. Mark Twain himself who claimed to have borrowed the phrase from former British Prime Minister; Benjamin Disraeli captured the whole expression as: "There are three kinds of lies: lies, damned lies, and statistics." To buttress this phrase, he also stated that: *"facts are stubborn things, but statistics are more pliable."* (Twain, 1924).

Ever since Twain made this revelation, others have joined the train in criticising statistical results. Let us look at some of the quotes from scholars similarly highlighting the fallibility of statistics as cited by the University of New York (no date);

> "There are three degrees of falsehood: the first is a fib, the second is a lie, and then come statistics"?.....Now, however, there is a lull in this process of military conversion, the latest statistics would seem to show that the Indians are no longer decreasing in numbers. But this is in reality only another instance of Bagehot's "lies, d—d lies, and statistics...After all, facts are facts, and although we may quote one to another with a chuckle the words of the Wise Statesman, 'Lies—damned lies—and statistics,' still there are some easy figures the simplest must understand, and the astutest cannot wriggle out of."

The above represents just a small fraction on the numerous criticisms against statistics. Furthermore, an important question we need to ask ourselves at this juncture is; what is it about statistics that makes it to be termed a "damn lie" This is the whole thrust of this paper. Hence, this study is devoted to demystifying some arguments that shows, that statistical results should not be accepted without reservations. Nevertheless, it should be noted that data used in this paper are fictitious, this is done so as not to discredit statistical bulleting of organisations. Subsequently, the meaning of statistics and what it is utilised for will be briefly discussed in the next paragraphs.

WHAT IS STATISTICS, AND HOW IS IT UTILISED?

Unlike other concepts, statistics is not a hard concept to define. In early societies, statistics was confined to only state affairs. Hence it was used to get the number of viable men for war, and total numbers of people for tax purposes. However, recently, it has been used to embrace almost every sphere of human activity (Uzobo, Akhuetie & Otu 2016). Therefore, a number of old definitions, which were confined to narrow field of enquiry, were replaced by more modern definitions, which are much more comprehensive and exhaustive.

According to A.L. Bowley (cited in Varalakshmi et.al, 2004), Statistics is a numerical statement of facts in any department of enquiry placed in relation to each other. Obviously, this is an incomplete definition as it takes into account only the aspect of collection and ignores other aspects such as analysis, presentation and interpretation. Bowley gives another definition of statistics, and asserts that 'statistics may be rightly called the scheme of averages. This definition is also incomplete, as averages play an important role in understanding and comparing data while statistics provide more measures. Bluman (2009), defined it as: *the science of conducting studies to collect, organize, summarize, analyse, and draw conclusions from data.* Similarly, Triola (2006) defined statistics as *the collection of methods for planning studies and experiments, obtaining data, and then organizing, summarizing, presenting, analyzing, interpreting, and drawing conclusions based on the data.* Secrist, H. (cited in Varalakshmi et.al, 2004) offered a more comprehensive and exhaustive meaning of statistics by defining it as *the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according*

to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other.

In all these definitions, four concepts stand out in the definition of statistics: firstly, statistics has to do with the collection of data, secondly data collected are presented, thirdly, presented data are analysed and finally, the results of analysed data are interpreted to make gleaned meanings from the data.

Historically, statistical thinking revolved around the needs of states to base policy on demographic and economic data. In other words, it designed for development and planning purposes. Overtime, statistics was now meant to influence public debates. People were now meant to base their argument for a particular developmental policy on statistics.

Statistics is not a mere device for collecting numerical data, but a means of developing sound techniques for handling, analysing and drawing valid inferences from the data. Statistics is also applied in every discipline such as Sociology, Biology, Commerce. Education, Planning, Business Management, Information Technology, amongst others. It is almost impossible to find a single department of human activity where statistics cannot be applied.

Statistics are the lifeblood of successful commerce as market projections are often made using statistics. Statistical methods are useful in measuring numerical changes in complex groups and interpreting collective phenomenon. Nowadays the uses of statistics are abundantly made in any economic study. It is based on this note that Alfred Marshall said, “Statistics are the straw only which I like every other economist has to make the bricks”. It may also be rioted that statistical data and techniques of statistical tools are immensely useful in solving many developmental problems such as wages, prices, production, distribution of income and wealth and so on. Statistics is widely used in education (Uzobo, Akhuetie & Otu 2016).

Statistics is necessary for the formulation of policies to start new course, consideration of facilities available for new courses etc. There are many people engaged in research work to test the past knowledge and evolve new knowledge. These are possible only through statistics. Statistics is indispensable in planning. In the modern world, which can be termed as the “world of planning”, almost all the organisations in the government are seeking the help of statisticians to plan for efficient working, formulation of policy decisions and execution of the same. In order to achieve the above goals, the statistical data relating to production, consumption, demand, supply, prices, investments, income expenditure etc. and various advanced statistical techniques for processing, analysing and interpreting such complex data are of importance. Statistics play an important role in planning, commissioning both at the central and state government levels. In Medical sciences, statistical tools are widely used. In order to test the efficiency of a new drug or medicine, the t-test can be used to compare the efficiency of two drugs or two medicines. More and more applications of statistics are at present used in clinical investigation. Recent developments in the fields of computer technology and information technology have enabled statistics to integrate their models and thus make statistics a part of decision-making procedures of many organisations (Uzobo, Akhuetie & Otu 2016).

LYING WITH DESCRIPTIVE STATISTICS

Descriptive statistics are the most commonly used social statistics. This form of social statistics describes the main features of a collection of information, or the quantitative description itself. Descriptive statistics are distinguished from inferential statistics (or inductive statistics) in that descriptive statistics aim to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent. This generally means that descriptive statistics, unlike inferential statistics, are not developed on

the basis of probability theory. Even when a data analysis draws its main conclusions using inferential statistics, descriptive statistics are generally also presented (Uzobo, Akhuetie & Otu 2016).

Descriptive statistics provides simple summaries about the sample and about the observations that have been made. Such summaries may be either quantitative, i.e. summary statistics, or visual, i.e. simple-to-understand graphs. These summaries may either form the basis of the initial description of the data as part of a more extensive statistical analysis, or they may be sufficient in and of themselves for a particular investigation. Some common measures used to describe a data set are measures of central tendency and measures of variability or dispersions. Measures of central tendency include the mean, median and mode, while measures of variability include the standard deviation or variance, kurtosis and skewness. Nevertheless, in this paper only a few descriptive statistics will be explored.

The Mean (Averages)

The mean is one of the most persuasive descriptive statistics that can be used. It is one of the many ways that *central tendency* (a single number that describes an entire set of data by a single central or typical value) can be measured. In summarising statistical data, the mean is extremely popular and is often the first choice for data analysis. Its advantages lie in the fact that the mean can give a rationally clear and easily discernible picture of what is happening to a dataset, and also, most people have some familiarity with its meaning. Nevertheless, if care is not taken, sometimes averages could be very misleading and lead to poor decision or policy formation. Let us take the example below using fictitious data. The table below represents the average contraceptive use between the Southern and Northern region of a particular country in 2018:

Region	Average Contraceptive use
Southern	50,000
Northern	50,000

If we take this data on the face value, it can be interpreted that regions have the same averages with regards to contraceptive use. Assuming the target for contraceptive use for that year is 15,000,000 for that year, it will be assumed that both regions are performing well. Nonetheless, a critical look at the averages below the surfaces of this data and an addition in individual region performance, the tables below revealed something different.

Southern Region	**Individual Contraceptive use**
South-East	51,000
South-South	54,000
South-West	45,000
Average Contraceptive use	**50,000**

Northern Region	**Individual Contraceptive use**
North-Central	76,000
North-East	25,000
North-West	49,000
Average Contraceptive use	**50,000**

Looking at this new tables, we can now see that everyone in the southern region have a high prevalence of contraceptive use at relatively similar levels, but there's considerable variation in the northern region. Hence, while there is the need for little action to be taken in the southern region, policy makers and development planners might want to find out why the north-east have an under-utilisation of contraception compared to other zones within the region, while sustaining its efforts in the north-central zone. Thus, by looking only at the averages, we might have lost meaningful information about contraceptive use in that country.

Nevertheless, it should be noted that this does not amount to lying. This example simply shows instances when policy makers use wrong statistical tools. Hence, in a situation like this, it is wrong to use the mean as average because the data is skewed. Whenever a normal distribution cannot be guaranteed as is the case with the northern region which has extreme value, the median becomes the most accurate measure of central tendency. So, in this case if the right tool is used it will not mislead.

Therefore, measures of central tendency can over-simplify and under-inform us, leading to poor decisions making. This is why when trying to make making decisions, it is expedient that we understand details such as the *distribution* (or shape) of the data, and know if there are any *outliers* (extreme scores) distorting our picture of the data. Consequently, while averages can be a useful high-level indicator of a dataset, it is imperative to make sure you have got all the critical information before taking action.

Consequently, the mean is best used with both discrete and continuous data, although its use is most often with continuous data, and when the distribution is symmetrical. However, the mean is susceptible to the influence of outliers (values that are unusual compared to the rest of the data set by being especially small or large in numerical value). In a situation such as this, the median would be a better measure of central tendency. Again, when the data is skewed (i.e., the frequency distribution of data is tilted), the use of the mean could be misleading, as the mean loses its ability to provide the best central location for the data because the skewed data is dragging it away from the typical value.

Percentages

Percentages are one of the most simple but misleading statistics. We must have probably heard the reports like: "The 86.9 million Nigerians now living in extreme poverty represents nearly 50% of its estimated 180 million population" and "Nigeria's unemployment rate rose to 20.9% of the workforce by the end of September 2018, up from 17.6% in 2017". Though the sound of these statistics feels great in principle, however, statements containing impressive percentages are most times flawed.

Firstly, there are conflicting measures of poverty. While some place emphasis on average income per day, others have used access to health, nutrition, drinkable water etc as its measures. Cardinali (2016) succinctly captured the problem with poverty statistics thus;

> "First, we're not even sure how to count poverty. The government's official measure is increasingly suspect (sic) because it doesn't adjust for safety-net programs, federal tax policies and regional variations in cost of living. Those items can make a big difference".

Christopher Jencks, a Harvard sociologist, in contrast to American official statistics of 45.3 million persons living in poverty found out that only 15 million Americans were living in poverty when all the proper adjustments have been made (Cardinali, 2016).

Again, the percentages of poverty and employments might be flawed on the ground that the experience of poverty and employment is relative, rather than absolute. For this reason, the Organisation for Economic Co-operation and Development (OECD) counts as "poor" anyone with an income that's less than half the national median level. Thus, using the 50 percent measure, the poverty line for a family of four in the United States was put at $25,974 rather than $22,811, according to the Center for Economic and Policy Research (2013), meaning that many more families would be counted as officially poor

Furthermore, no matter how you count poverty and employment, any positive signs usually do not appear. This was the situation between the National Bureau of Statistics and the Federal government of Nigeria. While according to the NBS the unemployment rate was 20.9%, to the federal government the figure was incorrect because over 12 million jobs have been created for rice farmers alone as declared by the Rice Producers Association of Nigeria which was not counted in the employment statistics of the NBS. This is another example of wrong use of statistics. It is a situation of garbage in and garbage out. Indeed, percentages are most accurate for accounting for the sample and not the population.

In a the same vein, a July study from the Pew Research Center (2015) reported that, about 1.6 million U.S. children were lifted out of poverty between 2010 and 2013. As excited as the news was, unemployment rates are still rising according to statistics even in the United States. Hence, percentages like the above discussed could be very misleading for a number of reasons:

- They look like absolute figures, so it's difficult to question them
- They can sound like they're based on much more data than they often are
- They rarely have enough data to decide if something statistically significant
- They can lead us to compare two things unfairly

Thus, statistics such as above, can skew our understanding of the information we are being presented with, and ultimately stop us from making good decisions.

In essence, percentages are good in the generalization of results, and to apply the result in a population. In contrast, data expressed in percentages without the actual total mentioned may be misleading. For example; "During the Second World War, there were 100 soldiers and 2 nurses in a camp. The nurses were girls anyway. One day, one of the nurses got married to one of the soldiers. Incidentally, on the Notice Board there appeared a notice: 1% of the soldiers got married to 50% of the nurses." Hence, despite the fact that percentages are very convenient for comparisons, a table with only percentages in it does not give you as much information as a table with the actual counts, which might not really the actual information sorted for.

The Simpson's Paradox

The Simpson's paradox describes a phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined. This situation is often most prevalent in social and medical sciences statistics and is particularly problematic when frequency data is unduly given causal interpretations. The paradoxical elements disappear when causal relations are brought into consideration. It has been used to try to inform the non-specialist or public audience about the kind of misleading results mis-applied statistics can generate. Let us look at the example below using the Cambridge admissions statistics for 1996:

	Women	Men

	Applied	Accepted	%	Applied	Accepted	%
Computer Science	26	7	**27**	228	58	**25**
Economics	240	63	**26**	512	112	**22**
Engineering	164	52	**32**	972	252	**26**
Medicine	416	99	**24**	578	140	**24**
Veterinary Medicine	338	53	**16**	180	22	**12**
Total	**1184**	**274**	**23**	**2470**	**584**	**24**

From the table above, in all five subjects, women have an equal or better success rate in applications than do men. However, taken overall, 24% of men are successful but only 23% of women are successful.

Graphs/Charts

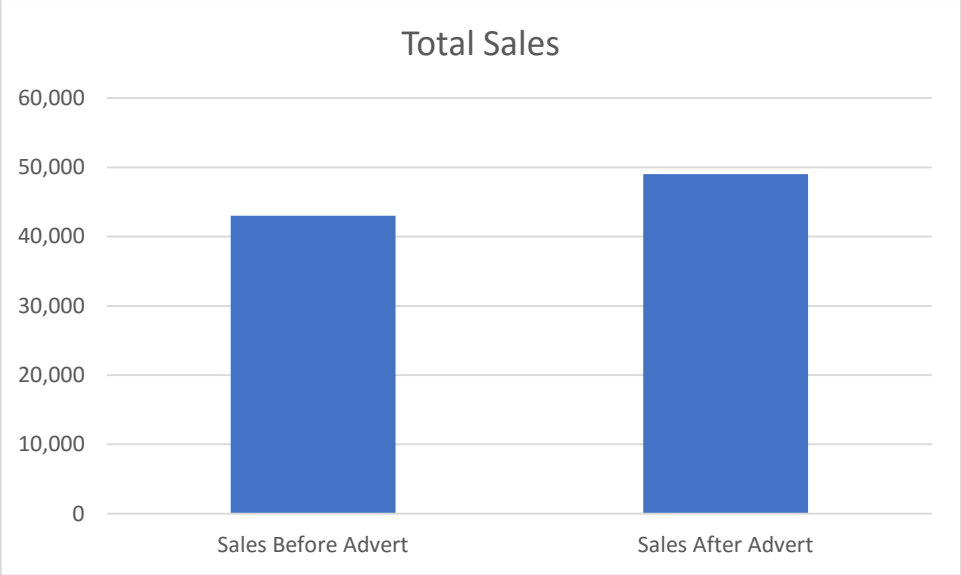
When used properly, graphs are clearer and more concise in presenting data. The purpose of a graph is to present data that are too numerous or complicated to be described adequately in the text and in less space. Graphs are also adequate if the data shows pronounced trends or reveals relations between variables. Also, choosing which type of graph to use is determine by the level of measurement. In statistics, the basic rules are as follows:

- For nominal/ordinal variables, use pie charts and bar charts
- For interval/ratio variables, use histograms (bar charts of equal interval)

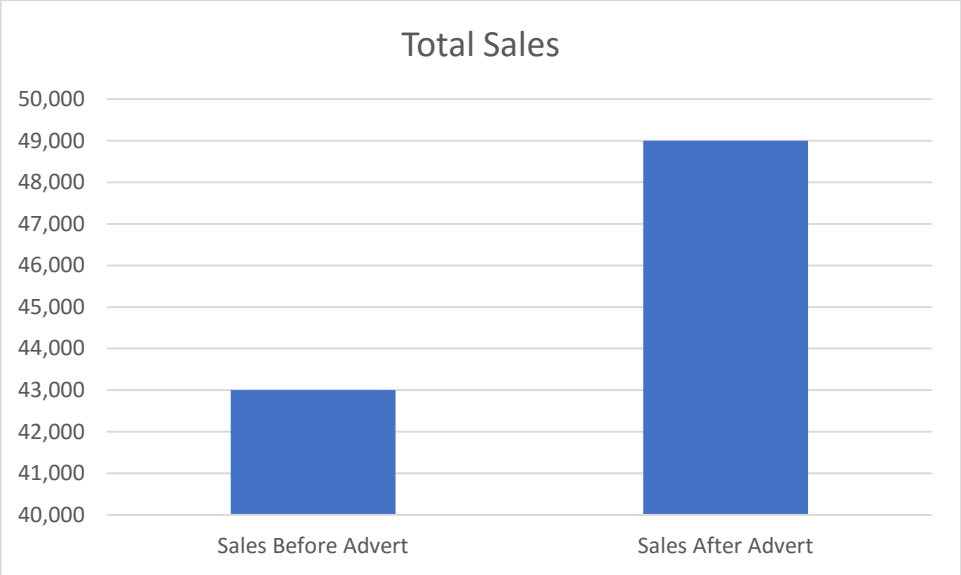
Some ways writers use graph in misleading the public include;

- 1. Omitting baselines, or the axis of a graph, to make one group look better than another.
- 2. Axis changing to blow out the scale of a graph to minimize or maximize a change.
- 3. Cherry picking by including certain parts of the data in the misleading charts or graphs
- 4. Using the wrong graph usually by picking a type of graph or chart that does not fit the data presented

Good graphs can bring a dataset to life and enable us to visualise what is going on. But as we shall see in the illustrations below, graphs could also be used to distort statistical information and mislead us. Imagine an advertisement agency shows you the following graph as evidence to show the success of advert on sales.

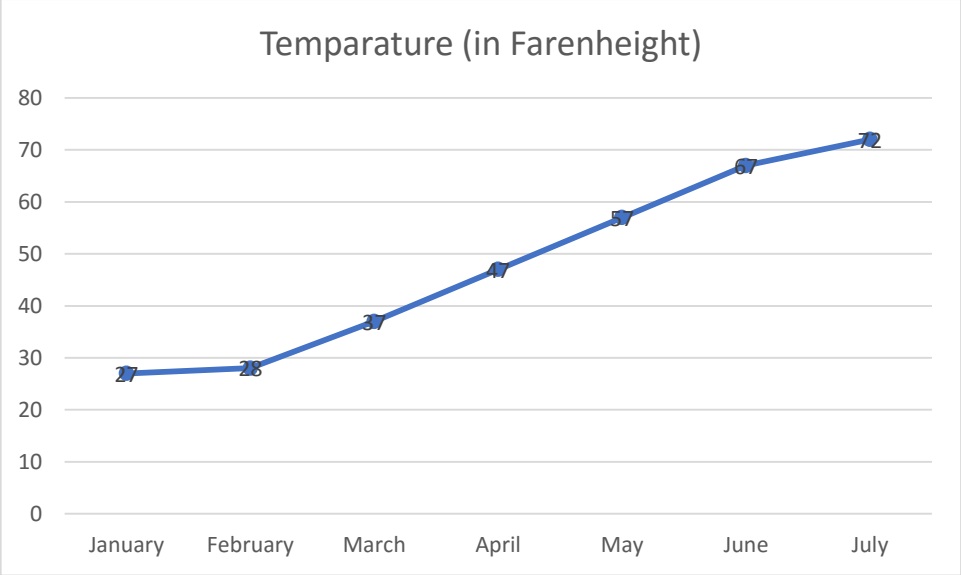


As we can see, the graph shows that the advert seems to have led to a moderate improvement in sales performance, but it is nothing drastic. Therefore, you might be tempted to ask yourself if advert was an effective investment? However, a different graph using the same dataset as can be seen below might convince you otherwise.



The same sets of data which was produced using the first suddenly shows a much more impressive effect of advert on sales. Though it is an exact information displayed by the first and second graph, nonetheless, here have been distortions which makes the two graphs provide different results and interpretation. By not starting the sales numbers from zero and decreasing the increments to 1000 each, the gap between the before and after figures looks bigger, making the training appear better. Without closer scrutiny, it will not be easy to accept the apparently impressive results at face value.

In a rather disturbing manner, a line graph like the below was used as proof of global warming:



It shows a trend of rapidly rising temperature levels that would be alarming, were it not for the fact that only data from January to July is shown. The creator of the graph has removed information that is not in-line with the message they want to deliver, making the graph highly misleading. If it showed the full year, we would have seen a drop in temperature towards the winter months, and there would be very little of interest. Better still, showing a comparison of temperature year on year would have given a much better picture of global warming, without being misleading. Thus, when being presented with graphs, it is important to focus in the detail, as well as the broader picture.

CAN AN ALTERNATIVE TO SOCIAL STATISTICS BE CONSIDERED?

It can be agreed that statistics can be misinterpreted to suit personal interest. Their point of origin is bad because most come from wild guesses or dubious data used to buttress personal interest. Other statistics become *mutant* (they become bad after mangled or reworded). Anyway, bad social statistics are potentially important because it could be used to stir up a public outrage or fear; they can distort our understanding of our world and lead to poor policy choices (Best, 2002). Thus, statistics have a bad reputation. In fact, there is a saying that *"you can prove anything with statistics."* People who use bad statistics maybe trying to manipulate the public by using numbers to somehow distort the truth.

Despite these challenges, we cannot do without social statistics. We need social statistics to summarise and clarify the nature of our complex world. This is mostly true when we want to talk about social problems which demands that we use statistical analysis to show the endemic nature of the problem.

Conclusively, the problem of 'damned lies' statistics should not be to ignore all statistics or assume numerical data are false. Though some statistics are bad, there exist pretty good statistics that can actually inform and educate the public genuinely. The way forward the, is not to give up on social statistics, but to become better judges of the numbers we encounter.

Thus, when presented with a statistical data, certain questions need to be addressed in our mind before either accepting or rejecting the result of the statistics. First and foremost, we need to ask ourselves if we understand what the data is telling us. If there

are areas that needs clarifications, we must ask for explanations especially if it feels like we are being blinded by statistics.

Secondly, we must take note of the source of the statistics. That is, who is providing the information. Then we ask ourselves what they stand to gain by providing these statistics. Hence, we ask ourselves if the statistics is provided purely as an academic venture for public good or if it has some vested interest in its provision. Thus, you find out if the source of the statistics is from a reliable source or a likely biased source. Furthermore, you need to find out if the evidence provided in the statistics can be applicable to all situations or for just specific purposes.

In addition, we need to ask ourselves what is missing. Therefore, the some of the following questions needs to be addressed before accepting the veracity of the statistics. questions that can likely address the missing link include: Is there a piece of data that is conspicuously absent? Is there something more you need to know before you accept what the statistics is saying? Is one of the pitfalls being used? If you have experience in the area, does the information match with your own knowledge and expertise?

On a final note, though we have already established that it is not all social statistics that has the intention of misleading the public, however, adopting a cautious mindset and an attitude of healthy scepticism may prevent the public from falling prey to pitfalls, and help the public make well-informed decisions. On a general note, statistics can be manipulated on the following grounds; Faulty polling, flawed correlations, data fishing, misleading data visualization, purposeful and selective bias, and using percentage change in combination with a small sample size. Thus, spotting these at the beginning of any statistical presentations helps in identifying which data are manipulated.

REFERENCES

Best, J. (2001). *Damned Lies and Statistics: Untangling Numbers from the Media, Politicians, and Activist.* California: University of California Press.

Bluman, A.G (2009). *Elementary Statistics: A step by step Approach* 7th edition, Boston: McGraw Hill Higher Education.

Triola, M.F (2006). *Elementary Statistics10th Edition*, Boston: Pearson Education Inc.

Uzobo, E., Micheal, O.T., & Akhuetie, E.R (2016). *Basic Descriptive Statistics in the Social Sciences,* Benin: Dimaf Publishers.

Cardinali, D. (2016). *The Problem with Poverty Statistics.* Retrieved from: https://www.huffingtonpost.com/dan-cardinali/the-skirmish-on- poverty_b_7969794.html

Centre for Economic and Policy Research (2013). *Relative Poverty Measures Can Help Paint a More Accurate picture of Poverty in 21st Century America.* Retrieved from: http://cepr.net/blogs/cepr-blog/relative-poverty-measures-can-help-paint-a-more- accurate-picture-of-poverty-in-21st-century-america)

Pew Research Centre (2015). *Black Child Poverty rate holds steady, even as other groups see declines.* Retrieved from: http://www.pewresearch.org/fact-tank/2015/07/14/black-child-poverty-rate-holds-steady-even-as-other-groups-see-declines/

News Agency of Nigeria, (2018). *Knocks for Peter Obi for quoting dubious statistics*. Retrieved from: https://www.nan.ng/news/knocks-for-peter-obi-for-quoting-dubious- statistics/

Varalakshmi, V., Sundaram, G.G., Indrani, B., Suseela, N., & Ezhilarasi, S. (2004). *Statistics: Higher Secondary- First Year.* Chennai: Tamilnadu Test Book Corporation.

Twain, Mark, (1924), *Autobiography*, in; A. B. Paine (ed), New York and London: Harper Brothers Vol. I.
University of New York (no date). *"Lies, Damned Lies and Statistics."* Retrieved from: https://www.york.ac.uk/depts/maths/histstat/lies.htm