Issue Brief

No. 13

Statistical Literacy – A Vital Ingredient
for an Informed Indian Citizenry

P. C. Mohanan

Former Acting Chairman
National Statistical Commission (NSC)

THE HINDU CENTRE
— for —
Politics and Public Policy

The Hindu Centre for Politics and Public Policy, Chennai, is an independent platform for exploration of ideas and public policies. As a public policy resource, our aim is to help the public increase its awareness of its political, social, and moral choices. The Hindu Centre believes that informed citizens can exercise their democratic rights better.

In accordance with this mission, The Hindu Centre's publications are intended to explain and highlight issues and themes that are the subject of public debate and aid the public in making informed judgments on issues of public importance.

Photo Caption: Data is being increasingly used in public messaging. However, statistical literacy in India remains far from desirable. File photo: The Hindu.

Statistical Literacy – A Vital Ingredient for an Informed Indian Citizenry

P. C. Mohanan
Former Acting Chairman
National Statistical Commission (NSC)

THE HINDU CENTRE
— *for* —
Politics and Public Policy

TABLE OF CONTENTS

ABSTRACT

Numbers are tricky. They can illuminate as much as they can mislead; inform as much as they can disinform; and reveal hidden problems but also mask real issues. However, one common attribute is that they emit an aura of credibility – even when they are deployed to mislead audiences, distort findings, or conceal facts. Statistics, a field earlier confined to either the ivory towers of research institutions or the obfuscating corridors of governance, is now a common form of public communication. However, a key question is lost in this avalanche of data that is fed to the people: What is the extent of statistical literacy in India?

In this Issue Brief, **P.C. Mohanan, former Acting Chairman, National Statistical Commission (NSC)**, writes on the need to evaluate and enhance statistical literacy – a competence that is required for a knowledge society but has not attracted the attention of policy makers, enumerators, the academia, and pedagogues. Although this form of literacy is elusive to define and measure, the manner in which citizens emerge as active constituents of an informed society depends in good measure on their ability to grapple with the numerals that they encounter on a daily basis. The rising relevance of data journalism, and the widespread use of numbers as a tool to enhance public messaging, should be met by increasing the popular awareness of statistics and its nuances.

This Issue Brief highlights the increased use of data in India's public communications and emphasises the need to ensure that data-based statements are presented in a clear, correct, and unambiguous manner. Mohanan shines the spotlight on some common errors that distort results and emphasises the importance of accuracy of language, logic, and context in conveying statistical results. He highlights the role of data journalists in identifying lapses and correctly conveying the messages revealed by the numbers, as wider dissemination of statistical literacy will result in a better understanding of key issues and facilitate the emergence of a discerning citizenry.

Keywords: Statistical Literacy, Data Journalism, National Statistical Commission

I.   INTRODUCTION

L iteracy is generally defined as the ability to read and write a simple sentence in any language with understanding. This can be tested by asking the respondent to write a dictated sentence to test writing ability, and to answer questions after reading a given paragraph to test reading and understanding. Although literacy skills can be graded on a scale, a simple binary of literate and not literate is what is used to measure the literacy levels of populations. Similarly, numeracy can be defined as the skill necessary to perform simple arithmetic computations using the basic operations of addition, subtraction, multiplication, and division. Simple arithmetic exercises can be devised to measure this ability which is sometimes also termed as arithmetic ability.

These two skills are considered basic abilities not only for carrying out meaningful economic transactions required for modern living but are also considered important for economic and social development. Literacy rate is most often used to understand the stage of development of a country. However, numeracy is not generally assessed or measured on a regular basis essentially due to difficulties in measurement. Literacy can be probed by directly asking the concerned persons or proxy respondents or guessed from the person's educational attainments. Numeracy, however, requires some extent of direct assessment of the respondent's ability.

There are also other kinds of literacy now evolving. Financial literacy is now considered an ability essential for economic decision making by individuals and households. This includes quantitative understanding of one's income, expenditure, saving and similar concepts that would help individuals set financial goals and plan their savings and investments. It is easy to see that financial literacy goes beyond simple literacy and numeracy.

Statistical literacy is a term of recent origin but is rapidly gaining currency in the context of data emerging as a key factor in policy making and public debates. In December 2014, then then United Nations Secretary-General, Ban Ki-moon, declared that "the world must acquire a new 'data literacy'". This new skill set was needed, he said, to meet the requirements of the UN's Sustainable Development Goals, covering everything from poverty reduction to gender equality and economic growth. Becoming data literate would help equip the international community with "the tools,

methodologies, capacities, and information necessary to shine a light on the challenges of responding to the new agenda".[1]

In common parlance the words information, data, and statistics are used interchangeably. The more specific contextual meanings are the nuanced understanding that while data consist of numbers collected for a purpose and having specific contextual meaning, information is processed data. The word 'statistics' means "the science of collecting, analysing and interpreting"[2] data following statistical principles. Therefore, one can talk about data literacy, statistical literacy, and information literacy as separate skills. The starting point is data, which are processed to arrive at some summarised figure(s) using statistical concepts. The data and the derived figures are not stand-alone values but have several underlying dimensions attached to it which is now called meta-data. The meta-data add value and meaning to the numbers.

This Issue Brief attempts to understand statistical literacy and discuss its different dimensions. Broadly the term statistical literacy is used to mean a correct understanding of statistical information emanating from data. The data could be on population, economy or societal issues or any other topics of *current* concern. Statistical information could be simple statements that have data content or references to data. This simplification is necessary as any efforts to further segregate the skills will lead to more complex measures impacting the utility of the measure. The level of complexity would also cloud any understanding of the basic skills required to understand statistical measures used in public debates. Unlike literacy or numeracy, the use of statistical literacy occupies a more restricted space. This space often revolves around issues of public interest. Such issues are most often found in the news media, debates or discussions, and other printed works such as propaganda or advertising material or datasheets. For obvious reasons, research based uses of statistics are outside the purview of this Issue Brief as statistical literacy is an inherent requirement. Finally, the discussion on statistical literacy relates to a domain that includes only those with verbal and numerical skills.

---

[1] **United Nations. 2014**. *The Road to Dignity by 2030: Ending Poverty, Transforming All Lives and Protecting the Planet - Synthesis Report of the Secretary-General On the Post-2015 Agenda*, New York, December, p. 38.
[https://www.un.org/disabilities/documents/reports/SG_Synthesis_Report_Road_to_Dignity_by_2030.pdf].
[2] **OECD Statistics Portal. [Online]**. Glossary of Statistical Terms.
[https://stats.oecd.org/glossary/detail.asp?ID=3847].

II. STATISTICAL LITERACY: THE EMERGING CONTEXT

P opular use of the internet as it is known today is over three decades old, although it was available much earlier to restricted groups. According to the World Bank[3], by 2020, an estimated 60 per cent of individuals in the world use the internet. For this purpose internet users are individuals who have used the internet (from any location) in the last three months. The internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV, and similar devices. This was almost nil in 1990. It is now reported by Global Digital Review (2022) that a total of 5.07 billion people around the world use the internet — equivalent to 63.5 per cent of the world's total population.[4] The latest report of the Department of Telecommunications (DoT) notes that in India the number of internet subscribers (both broadband and narrowband put together), which was 776.45 million at the end of September 2020 has increased to 834.29 million by the end of September, 2021.[5]

The emergence of social media and the widespread use of the internet has led to the creation of content beyond what could be conceived in the past. Clearly, a large part of such content is not under any governmental or editorial control. While in the past, carefully edited newspapers and journals provided information to the public, this now flows from a variety of sources with a substantial part of such content being unvetted.

With a drastic lowering of entry costs, access to sources of information has multiplied and the speed of its transmission has increased tremendously. This flood of uncensored information available to individuals round the clock has the potential to impact independent rational opinion making and can become a threat to social cohesion. These could take the shape of false claims, wrong interpretation of data in order to advance specific viewpoints, or other similar misuses, which have now become common recurrences.

---

[3] **The World Bank. [Online]** Individuals using the Internet (% of population).
[https://data.worldbank.org/indicator/IT.NET.USER.ZS?end=2020&start=2020&view=chart].
[4] **Datareportal. 2022**. Digital 2022: October Global Statshot Report, October 20. [https://datareportal.com/global-digital-overview#:~:text=A%20total%20of%205.07%20billion,12%20months%20to%20October%202022].
[5] **Department of Telecommunications. 2022.** Annual Report. 2021-22, Ministry of Communications, Government of India, New Delhi.
[https://dot.gov.in/sites/default/files/Final%20Eng%20AR%20Min%20of%20Tele%20for%20Net%2009-02-22.pdf].

The information deluge and its universal accessibility multiply the possibility of widespread misinterpretation of data. This is something all too familiar now. In sharp contrast to the past when the speed at which information spread was slow and restricted, it is now instant and uncontrolled.

Statisticians are trained to present data in the form of tables, graphs and other forms of visualisation, after ensuring that all relevant information are correctly depicted. However, the public use of such information is to see the story behind the numbers or graphs. This is what appears as headlines in the media. In this, apart from the numbers, the use of language plays a major part. A large part of statistical literacy will have something to do with the use of correct vocabulary while explaining or presenting statistical information.

Then there are clever ways of presenting charts and graphs so as to plant a particular picture in the minds of the viewer. This is also a common abuse of statistics. There is also the intent behind presentation of facts. It can be to present or explain a situation better, to defend or oppose an argument, or, in the worst-case scenario, to deliberately mislead.

The flood of data-based information through different forms of media clearly points to the need to promote statistical literacy among the public. A sustained attempt to spread the concept of statistical literacy would also bring in better analysis and presentation of data by the media and factual appreciation of the issues.

One of the functions assigned to the National Statistical Commission, when it was set up in 2005, was the promotion of public trust in official statistics. This indirectly involves enabling statistical agencies to draw correct conclusions from official data, without which public trust cannot be gained and restored. While the official data themselves are presented in a neutral manner, most often these are used in a way that pays scant attention to data limitations. For example, poverty estimates in the country are usually estimated from the household consumption expenditure surveys, which have a very detailed item list for collecting expenditure data. The last such survey was conducted by the National Statistical Service Organisation (NSSO) in 2011-12.

However, the NSSO also records household consumption expenditure in a highly aggregated form from households in other surveys based on four or five questions. The basic purpose of such data is to rank households according to the level of expenditure. However, researchers have also used

such data to derive poverty estimates disregarding the limitations of collecting expenditure data with a few questions versus the detailed questionnaire in the usual consumer expenditure surveys. Advanced analytical capabilities, now easily accessible and widely taught, are sometimes applied to datasets that are not originally intended for such analysis reminding one of the dictum: 'if you torture data enough it will say anything'.

III. DEFINING AND MEASURING STATISTICAL LITERACY

A direct measure of statistical literacy would require an unambiguous idea of what constitutes 'statistical literacy in terms of some ability' as a starting point. The simple ability to read and write or perform arithmetic operations can be tested or assumed on the basis of educational achievements or qualifications as there is a clear idea of what it means to possess this ability. Statistical literacy combines ordinary literacy, numeracy, and reasoning in some form or the other.

One can, therefore, consider statistical literacy as the ability to comprehend and communicate data-based information. This pre-supposes a somewhat more advanced level of literacy and numeracy. A data-literate person would be expected to understand and apply basic statistical principles, grasp the limitations in data-based statements, and reach conclusions. The W. M. Keck Statistical Literacy Project defines statistical literacy as critical thinking about numbers, about statistics used in arguments, including the ability to read and interpret numbers in statements, surveys, tables, and graphs; and study how statistical associations are used as evidence for causal connections[6]. Statistical literacy as explained in the works of Prof. Milo Schield qualifies knowledge of statistics as different from the approach to statistical literacy by using the conventional cautionary phrase: 'Take CARE'. Each of the four letters in CARE stand for a kind of influence on the size of a statistic:

- **Context** (comparisons, ratios, study design and confounding),
- **Assembly** (how statistics are defined and presented),
- **Randomness** (chance, margin of error and statistical significance) and
- **Error** or bias.

The development of statistical literacy is built on the first of the two influences. His book has substantial sections on explaining commonly used statistical statements and interpretations that can take forward the meaningful and correct understanding of the underlying statistical facts.

---

[6] **Schield, M. 2009**. Statistical Literacy Text book, W. M. Keck Statistical Literacy Project, Augsburg College. [http://www.statlit.org/Schield.htm].

A better understanding of statistical reasoning can contribute to statistical literacy and efforts to do that would advance statistical literacy. This would require some tweaking of the ways in which statistical education is imparted. This is one aspect of the issue. One way to assess the level of statistical literacy would then be developing some tests and administering them on the target population. This can at best be done in a classroom set up or in small groups.

An alternative approach to study statistical literacy or general interest in statistics-based information would be through direct surveys of individuals asking them if they understand statistical measures like average and know about some ordinary and publicly available statistical indicators such as inflation, GDP, population, and other important macro indicators. There are currently no indications of such surveys having been conducted anywhere.

The question of defining a standard measure of statistical literacy, therefore, hinges on a functional definition of the term without which comparable measures cannot be developed. This, as seen from the above discussions, is not readily possible. Statistics is often used to support a viewpoint and the level of statistics used would depend on the context and the settings.

As an indirect understanding of the level of statistical literacy, one can think of public engagements with data-related issues in the media. An approximate way of understanding this could be through the news media by looking at the reporting of data-based news.

One effort to define and measure statistical literacy is the project undertaken by PARIS 21 as a follow up of the Busan Action Plan for Statistics[7]. The statistical literacy indicator attempted here measures the use of and critical engagement with statistics in national newspapers. The target population are journalists and newspaper readers. The source materials used are the RSS [Really Simple Syndication] feeds of national newspapers, primarily based on the global news aggregator Google News. The indicator used by them is a three-dimensional composite indicator of the equally weighted percentages of national newspaper articles that contain references to statistics at three different levels. The first level is the consistent, non-critical use of statistics based on keywords in the articles that refer to statistical data sources such as census, surveys or statistical

---

[7] **Klein, T, Galdin, A, and Mohamedou, E.I. 2016**. *An Indicator for Statistical Literacy Based on National Newspaper Archives*, Partnership in Statistics for Development in the 21st Century (PARIS21), France. [https://iase-web.org/documents/papers/rt2016/Klein.pdf].

indicators like CPI, GDP or any reference to statistical projects or institutions. Levels 2 and 3 are critical engagement with statistics, and critical mathematical engagement with statistics.

To derive the indicator, the project analysed a total of 8,880 articles during a three-month period in 2016 for the use of statistics in general news (Level 1). This corresponds to an average of 261 articles per country. For Levels 2 and 3, a total of 3,067 articles with explicit references to specific words such as 'statistics', 'data', 'study', 'research' or 'report' were analysed. For each of the three levels of statistical literacy, the resulting score gives the percentage of articles that contain at least one search term from the keyword lists. The score for each level thus ranges between 0 and 100 and the maximum total score over all three levels is 300. The results place Mexico and the UK jointly at the top slot, while the Philippines is ranked third for the Anglophone developing countries. The explanations for these somewhat unexpected results is possibly due to three reasons.

Firstly, two statistical institutes — el Instituto Nacional de Estadística, Geografía e Informática (INEGI) [the National Institute of Statistics, Geography and Informatics] in Mexico and the Philippine Statistics Authority (PSA) in the Philippines — are very engaged in monitoring the use of statistics by journalists. The INEGI reports the impact and value of statistics based on daily monitoring of newspapers and media resources, and the PSA tracks references to their institutions via Google news subscriptions, and engages with the media.

A second explanation is the differences in the nature of the audience of the main newspapers by country. In such instances, a good degree of 'scurrilous coverage' may explain a lower score. Finally, but importantly from the perspective of this Issue Brief, in many of the developing countries newspapers use press releases from statistical/governmental agencies verbatim, without making them digestible for a general audience through simple and meaningful explanations. This points to a weakness in this indicator, in that it rewards top level keywords related to the critical mathematical category but good journalism should actually avoid verbatim reproduction. The detailed findings can be seen in the source quoted above.[8]

Newspapers are a very good source to understand statistical awareness or interest in society. Despite variations in reach and readership of newspapers in societies, readership is a function of the levels of the overall literacy in the society. For example, as per data published by the Registrar

---

[8] **Ibid**.

of Newspapers in India (RNI)[9], the number of Dailies published in India as on March 31, 2021 was 9,750. The claimed circulation of Dailies was 22.6 crores copies per publishing day for a population of around 135 crores. Hindi had 4,349 Dailies, claiming a circulation of 10.4 crore copies, while 1,107 Urdu dailies, 1,083 Telugu dailies, and 820 English dailies claimed circulation figures of 2.2, 1.5 and 2.1 crore copies per publishing day respectively. Out of the total 22,930 periodicals, 19,608 mainly covered News and Current Affairs.

A careful analysis of the newspaper feeds can be used to tabulate data related news reports and these newspaper feeds can give an idea of the general interest in data-based news/topics. These would depict one aspect of the statistical literacy. Although it can be indicative of the public interest in data related news, it may not be very helpful in arriving at the actual readership and reader interest in data related issues. Often, newspapers merely republish the government releases, most of which contain statistical information or conclusions. This is noted by the authors of the PARIS 21 study as a weakness of using newspaper feeds to study statistical literacy.

Broadly speaking, defining and then measuring statistical literacy is multi-layered. While one can roughly gauge public engagement with data through their presence in the media, the depth of this engagement cannot be easily quantified. The efforts should, therefore, be focussed on ensuring that such engagement provides unbiased conclusions from the underlying data.

The next section discusses some examples on how statistical information could present a biased picture either by economising or twisting facts to support a specific argument or sometimes just to make the resulting story interesting.

---

[9] **Registrar of Newspaper for India. [n.d]** Press in India – 2020-21 65th Annual Report – Volume-I, Ministry of Information and Broadcasting, Government of India. (Vide: Chapter 5) [http://rni.nic.in/all_page/pin202021.html].

IV. THE IMPORTANCE OF STATISTICAL LITERACY AND SOME ASPECTS OF STATISTICAL CONTENT

News media are indeed the main source for public engagement with statistical information. Usually, there is little external check on what information is presented by newspapers or how it is portrayed, and correctly so, although it is expected that they do conform to ethical principles while publishing, keeping in mind the legal consequences of spreading inaccuracies or falsehoods. That said, nothing stops them from presenting publicly available data and information in a way that projects a particular view or leads the reader to a particular conclusion, sometimes statistically untenable.

The growing public interest in data-based issues is evident from major newspapers now publishing specialised sections providing data insights by qualified data journalists. These are usually meant for experts or those who have special interest in the relevant fields and not generally for ordinary readers. In recent years, there has been an increasing interest in data-based reports. During the recent COVID-19 crisis, large sections of the public keenly followed the infection related data as it impacted their livelihood directly. On the flip side, in the process of creating exciting headlines, the media quite often give misleading interpretations to data. Usually, they are not expected to critically examine the data collection and aggregation methodology or the manner of administering the relevant questions in the case of survey data.

This heightened public interest in data is persistent and needs to be studied to understand the depth of public involvement with statistical information and possible ways of improper use. This will lead to developing appropriate dissemination/communication strategies at different levels. Correct appreciation of statistical information would require the introduction of statistical literacy beyond the usual statistics theory taught at the school level. One might say that such modules should become part of journalism courses.

In a paper presented in the World Statistical Congress[10] titled *Statistical Literacy for Policy Makers*, Milo Schield proposed seven simple questions while presenting and understanding statistical data: How big? Compared to what? Why not rates? Per what? Defined, counted or measured how? What

---

[10] **Schield, M. 2021**. *Statistical Literacy for Policy Makers*, Proceedings - 63rd ISI World Statistics Congress, 11 - 16 July 2021[Virtual]. [https://www.isi-web.org/files/docs/papers-and-abstracts/225-day5-ips087-statistical-literacy-for-polic.pdf].

was controlled for? What should have been controlled for? Although these questions are fundamental in understanding and explaining data, statistical theory classes do not engage their students on such issues.

Reporting findings of surveys

Survey results are fertile grounds for making headlines. Understanding any statistics requires some knowledge of the process of generating these numbers. This is more so in the case of sample surveys where reliability of the results is highly dependent on the correct application of statistical theory. In particular, survey estimates have sampling and non-sampling errors. Most often, while reporting the findings, people tend to forget this and quote the numbers with an air of unquestionable authority. For example, we see how public debate hinges on small changes in the unemployment rates derived from labour force surveys, conveniently forgetting the error margins of the estimates. Changes can even follow from simple rounding off while authoring the report.

There are other questions about survey estimates that are important. Some examples include possible exclusion of a section of the population in the survey coverage, definitional, and conceptual changes. For the sake of brevity, this Issue Brief does not go into the details.

Understanding the context of questions

Users of survey data are generally advised to carefully study the survey questionnaire and the instructions for the field data collectors or other material used during the survey operations — what is now called the meta-data of the survey. All such information are expected to be part of the survey catalogue or archive and disseminated along with the survey data. Understanding this is often a pre-requisite for researchers before analysing the survey data. Quite often media and the public use the information from the published survey reports where all these meta-data are not usually provided. There are many examples of survey responses getting influenced by the way the questions are framed or the response options managed by the survey designers. A little more on this.

Usually in surveys, questions are either open-ended, where respondents provide a response in their own words, or are closed-ended, where they choose from a list of choices provided by the individual/organisation conducting the survey. The Pew Research Center, the internationally

known survey agency, in their methodological note on survey questionnaire design, gives some examples. In a poll[11] conducted after the U.S. 2008 Presidential election, people responded very differently to two versions of the same question: "What one issue mattered most to you in deciding how you voted for President?" One was closed-ended and the other open-ended. In the closed-ended version, respondents were provided five options and could volunteer an option not on the list.

When explicitly offered 'the economy' as a response, more than half of respondents (58 per cent) chose this answer; only 35 per cent of those who responded to the open-ended version mentioned 'the economy'. Moreover, among those asked the closed-ended version, fewer than one-in-ten (8 per cent) provided a response other than the five they were read. In sharp contrast, as high as 43 per cent of those asked the open-ended version provided a response that was not listed in the closed-ended version of the question. All of the other issues were chosen at least slightly more often when explicitly offered in the closed-ended version than in the open-ended version. There are other similar examples in the referred link.[12]

The National Family Health Survey (NFHS) is one of the most relied upon official sources for data on a wide range of health and behavioural information of the population in India. Domestic violence, for instance, is a topic covered by the NFHS and usually not covered in other national surveys. Based on one of the questions in the latest NFHS survey, newspaper headlines gave the startling finding that wife-beating was justified by a significant number of wives/husbands. The actual question was: "In your opinion, is a husband justified in hitting or beating his wife in the following situations:". Seven circumstances were then listed. Each of the seven situations in the closed-ended question had three possible responses: "Yes", "No", "Do not know". The findings were headlines in the media, with most news reports reproducing the findings more or less verbatim as given in the survey report.

The news reports suggested that both men and women had almost identical opinions on the issue. The survey did not explicitly ask the husband/wife if they justified husbands beating wives nor whether they ever beat or received beatings. This would have provided an unconditional opinion on husbands beating wives or on the actual incidence of wife-beating. However, in this context,

---

[11] **Pew Research Center [Online]**. *Writing Survey Questions,* Washington. [https://www.pewresearch.org/our-methods/u-s-surveys/writing-survey-questions/].

[12] **Ibid.**

the responses are only for the given situations, which could have resulted in the almost identical reporting by men and women. Although these responses pertain to the 'justification' in the opinion of the respondents; the results presented could confuse the readers giving them the wrong impression that the numbers reflect the actual state of violence against wives. The other side of the argument could be that these are sensitive questions unlikely to elicit truthful responses. The following is a news report based on the survey:

Forty-five per cent of women and 44 per cent of men believe that a husband is justified in beating his wife in at least one of seven specified circumstances. Women and men are both most likely to agree that a husband is justified in hitting or beating his wife if she shows disrespect for her in-laws (32% and 31%, respectively), and are both least likely to agree that a husband is justified in hitting or beating his wife if she refuses to have sex with him (11% and 10%, respectively). For both women and men, agreement with wife-beating is lower in urban than rural areas and it tends to decrease with schooling and wealth.

Sources:

Sengar, S. 2021. [An Alarming Number Of Women Justify Getting Beaten By Husbands In The Country](https://www.indiatimes.com/news/india/indian-women-justify-getting-beaten-by-husbands-555421.html), *Indiatimes.com*, November 29.
[https://www.indiatimes.com/news/india/indian-women-justify-getting-beaten-by-husbands-555421.html].

Sengar, S. 2022. [45% Of Women Justify Husband Beating Wife If She Argues, Doesn't Cook Properly Or Refuses Sex](https://www.indiatimes.com/news/india/45-of-women-justify-husband-beating-wife-if-she-argues-national-family-health-survey-569175.html), *Indiatimes.com*, May 10.
[https://www.indiatimes.com/news/india/45-of-women-justify-husband-beating-wife-if-she-argues-national-family-health-survey-569175.html].

Opinion polls, therefore, require carefully constructed questions and are poor substitutes for providing statistics depicting actual reality.

Comparing absolute numbers

The RNI report quoted earlier states that,

"Even among dailies also, Uttar Pradesh (U.P.), with a total circulation of 3,83,98,144 copies per publishing day retained **its top position** and was followed by Maharashtra with 3,10,70,720 copies per publishing day ...".

As U.P. and Maharashtra are two of the top States in terms of population, this statement is not anything unexpected. This is a case of comparing absolute numbers that are not comparable without some normalisation.

Index-based ranking and rating

Nowadays a lot is heard about different kinds of indices to summarise multi-dimensional issues. These catch the public attention very quickly and are used in debates as they show advancement or regression of outcomes. One familiar index is the Human Development Index (HDI), which was constructed after very careful practical and technical considerations as

> "a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and have a decent standard of living. The HDI is the geometric mean of normalised indices for each of the three dimensions."[13]

There are many other indices released nowadays of which the World Hunger Index was recently in the news essentially because, among other things, it showed India in a very bad light. The fact that the index has severe technical limitations to depict hunger was then brought to the fore. It is anyone's guess whether the index would have been acceptable by its critics, as it is, if it had shown India in a very bright position. Another case in point is the Ease of Doing Business Index that the World Bank used to compile. This was one highly complimented and oft-quoted index as the country's position showed significant improvements over the years. However, the World Bank[14] subsequently discontinued the index after internal investigations reported data irregularities, although these 'irregularities' did not pertain to India.

Preparing a multi criteria based index has advantages in comparing rankings of States or countries or time periods, but the basis for selection of variables and the weights assigned for combining the variables are not usually highlighted.

---

[13] **United Nations Development Programme [UNDP]. [Online]**. *Human Development Reports: Human Development Index* [https://hdr.undp.org/data-center/human-development-index].

[14] **The World Bank. 2021**. World Bank Group to discontinue Doing Business Report, September 16. [https://www.worldbank.org/en/news/statement/2021/09/16/world-bank-group-to-discontinue-doing-business-report].

What is the denominator or base?

Shoppers are all too familiar with advertisement like 'SALE: 50% off'. Nothing would be said further like 50 per cent of what price or for what items, if it is a flat 50 per cent or "up to 50 per cent", and so on. Clearly these are attention grabbers and not meant to be taken seriously. At times there are claims that a party's vote share has increased by 50 per cent. This could be an increase from 10 per cent to 15 per cent or 2 per cent to 3 per cent.

A leading claim in the newspapers from the CEO of the eyewear company, Lenskart[15] was, "In the next five years, we aspire to have 50 per cent of India wearing our specs." He further elaborated,

> "At Lenskart, we are obsessed with our customers, technology, and making the world a better place through easily accessible, best-quality eyewear. More than 600 million people in India and 4.5 billion people globally need vision correction, but only a fraction of them use it due to a lack of access, awareness, and high-quality, affordable solutions."

This, of course, sounds to be a very laudable objective; but from where do these numbers come or is it just business optimism?

A Government of India press release on December 15, 2022,[16] quoting a reply given to a question in Parliament on the Jal Jeevan Mission programme for providing tap water to every rural household, states:

> "At the time of announcement of Jal Jeevan Mission, 3.23 Crore (17%) households were reported to have tap water connections. So far, around 7.48 Crore (38%) rural households have been provided with tap water connections in last 3 years. Thus, as on 12.12.2022, out of 19.36 Crore rural households in the country, around 10.71 Crore (55%) households are reported to have tap water supply in their homes".

It then quoted the findings of an assessment survey,

---

[15] **Gupta, V. 2021**. "In the next five years, we aspire to have 50 pc of India wearing our specs," says Lenskart, IndianRetailer.com, May 17. [https://www.indianretailer.com/news/lenskart-to-enhance-digital-offerings-for-customers-omnichannel-experience.n10820].

[16] **Ministry of Jal Sakthi. 2022**. *Impact of Jal Jeevan Mission,* Press Information Bureau, Government of India. December 15. [https://pib.gov.in/PressReleasePage.aspx?PRID=1883851].

"Department of Drinking Water & Sanitation undertakes annual assessment of the functionality of household tap water connections provided under the Mission, through an independent third-party agency, based on standard statistical sampling. During the functionality assessment 2021-22, it was found that 86% of households had working tap connections. Out of these, 85% were getting water in adequate quantity, 80% were getting water regularly as per the schedule of water supply for their piped water supply scheme, and 87% of households were receiving water as per the prescribed water quality standards."

Parliamentary replies are carefully vetted at many levels of the bureaucracy. These two paragraphs look somewhat contradictory in terms of the numbers given. Possibly the survey findings refer only to the connections provided under the scheme and not to the whole of rural households. Can someone using only the second part of the reply be faulted?

In a report on Breast Cancer related issues, The Hindu[17] quoted an ICMR report which said,

"100.5 out of 1,00,000 women were being diagnosed with breast cancer. From the approximately 1,82,000 cases of breast cancer at present, the report has projected cases to rise to 2,50,000 by 2030."

Most probably the number 100.5 refers to those getting screened and not to the whole population, in which case the expected number should be around 6.5 lakhs for a female population of roughly 65 crores in the country now. The use of correct syntax while using per cent and percentages generally receive less attention than it deserves.

Question of causality

In statistics classes, students are always reminded that correlation does not imply causality. In many cases there are claims of achievement directly linked to some new government initiatives. These statements, however, do not provide any information about what would have been the

---

[17] **Koshy, J. 2022**. *Government says breast cancer not a matter of 'national' or 'extreme' urgency*, The Hindu, December 18. [https://www.thehindu.com/news/national/breast-cancer-not-a-health-emergency-govt/article66272052.ece].

achievement without these initiatives in place; possibly the new initiative has brought only a marginal improvement.

Use of language is an important part of statistical explanation either in statistical tables or statements. Statistics education revolves around textbook descriptions using statistical terminology. This does not provide scope for specialising in proper communications of statistical results. At the same time, the use of statistics in the media has to be in a way that captures the reader's attention. The media, therefore, cannot be faulted if they move away from drab presentation of facts that will not engage the audience. Most Government statistical reports come with paragraphs that are just table reading and escape being caught with incorrect conclusions from wrong syntax. On the other extreme are eye-catching headlines from journalists that are far removed from what the data actually says.

V. CONCLUSION

T he idea of statistical literacy has acquired some importance in recent times. The widespread use of official statistics to show how a government is doing better than the previous one is now an all-too-familiar tool of propaganda.

However, a clear cut definition of statistical literacy is still elusive, and measuring this important metric of a knowledge society at the individual level remains problematic. That said, the existing literature on the topic provides the following conclusion: the extent of citizens' engagement with data needs to be understood to improve the usage of statistics, particularly official statistics, to convey information. Needless to emphasise, developing statistical literacy is crucial for healthy public debates in a knowledge driven society. It is, therefore, an urgent requirement to improve public trust in data, especially official data.

Currently, the starting point for building statistical literacy rests with data users, in particular with those who present data-based arguments in media, in public forums, and in official statements. The need to start with data-users is because the subject of statistical literacy is not adequately addressed in basic statistics courses taught in institutions. This would involve understanding the proper grammar of the language while making data-based statements. This Issue Brief highlights the increased use of data in India's public communications and emphasises the need to ensure that data-based statements are presented in a clear, correct, and unambiguous manner. A wider discussion on the correct use of data in statements is beyond its scope, except to the extent of citing a few examples.

Although the case of creating awareness about statistical literacy is very strong, measuring it can only be attempted indirectly at present. Currently, data journalists have a major role to play in the correct use of data and to point out statistically invalid claims in public debates and media. The role of the data journalist gains centre-stage as official statisticians will naturally be constrained in going public on any possible improper usage and interpretation of government data by data users.

Finally, this Issue Brief is exploratory in its scope, and is meant to encourage wider discourse on the topic. The examples given are only indicative and more serious examples of incorrect usages of data or misleading statements based on data can be cited by the readers.

About the Author

P.C. Mohanan is former Acting Chairman, National Statistical Commission (NSC). He was earlier a Member of the NSC from June 2017, and the Acting Chairman of the Commission from October 2018 until his resignation from the position in January 2019. He entered the Indian Statistical Service ranked Second in the 1979 batch and worked in both the NSSO and the CSO until his retirement in 2015. He has been a member of important technical committees that have addressed issues in India's socio-economic sectors and has held international consultancy assignments in the Asian Development Bank, UNDP, Food and Agricultural Organisation, and International Labour Organisation. He is currently Chairman of the Kerala State Statistical Commission. He can be contacted at **parachottilmohanan@gmail.com**

THE HINDU CENTRE
for
Politics and Public Policy