# But are those numbers correct?:

# Some suggestions for appraising the ACCURACY of statistics

Simon Scott, Counsellor, Statistics Directorate, OECD, Paris, June 2018

## 0. Abstract

Numerous internationally agreed documents set out principles and practices to promote sound statistics, but none provide criteria for assessing whether statistics are actually correct. Other documents discuss statistical quality, but focus largely on utilitarian considerations such as availability and timeliness; and when they do discuss accuracy, they focus on processes (lists of good practices) rather than on results (are the data reliable?). Yet knowing whether data are correct is of fundamental importance to the establishment of knowledge, the formulation of explanatory hypotheses, and the development of effective policy. This paper is a first attempt to identify some characteristics of data that may be accepted as true.

## 1. The evolution of thinking on sound statistics

The first recognisably modern steps towards ensuring statistical quality were taken in census legislation of the 18th and 19th centuries. These laws imposed obligations on citizens to answer questions truthfully, but also established some rights of privacy, e.g. to refuse to disclose religious belief. Census officials were required to maintain the secrecy of personal information and the confidentiality of commercially sensitive information. Such measures helped ensure the accuracy of the count by compelling the respondent to provide correct information and assuring him that he would not suffer as a result.

In the 19th and 20th centuries, statistical quality was further promoted through detailed instruction manuals regulating specialised statistical collections. And from the 1990s, new species of documents have emerged that abstract from these laws and procedures principles thought to be of general application. These include the UN [*Fundamental Principles of Official Statistics*](https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf)(1994, 2014) and [*Principles governing international statistical activities*](https://unstats.un.org/unsd/statcom/37th-session/documents/2006-13-principles-E.pdf)(2006), [*European*](http://ec.europa.eu/eurostat/documents/64157/4392716/Revised_CoP_Nov_2017.pdf) and [*Latin American*](https://repositorio.cepal.org/bitstream/handle/11362/16423/1/FILE_148024_en.pdf) *Codes of Good Practice*, and the OECD Council[*Recommendation on Good Statistical Practice*](http://www.oecd.org/statistics/good-practice-toolkit/Brochure-Good-Stat-Practices.pdf).

All these documents are “how to” guides designed to promote statistics suitable for a democratic society. In this respect they have many excellencies, but they do not attempt to establish the inherent characteristics of statistics that should be accepted as true. For this one might hope for guidance from another class of document, namely the presentations by various agencies of *dimensions of statistical quality*. Here, however, one again finds that the basic question of how to assess whether a number is true, or likely to be true, finds no direct answer. Indeed a [UN study](http://www.oecd.org/std/21687665.pdf) even seems to imply that zeroing in on the question of accuracy is *dépassé*:

*In statistics, quality used to be primarily associated with accuracy. It is now recognised that there are other important dimensions. Even if data are accurate, they do not have sufficient quality if they are produced too late to be useful, or cannot be easily accessed, or conflict with other credible data. Therefore, quality is increasingly approached as a multi-dimensional concept.*

Such an insistence on multi-dimensionality has led to definitions of statistical quality that include timeliness, frequency, accessibility, relevance, coherence, interpretability, comprehensiveness, completeness, serviceability, integrity, credibility and clarity. Jostled among this throng of virtues, accuracy and reliability retain only a minor place, and even then are defined in ways that somehow evade the question of whether the data are actually true. For example, the UN study reports that the IMF defines *reliability* as merely “the closeness of the initial estimated value to the subsequent estimated value”. The Fund’s definition of *accuracy* – “the closeness between the estimated value and the true value that the statistics were designed to measure” – at first looks more promising but it is then stated that “there is no single or overall measure of accuracy”, and no attributes to gauge accuracy are offered. Moreover, insisting on the word “estimate” tends to suggest that absolute accuracy is impossible, no matter how simple the count might be.

Overall, thinking about the soundness of statistics has focused on procedures by which official bodies can generate data useful to society. This is, no doubt, a worthy objective, and governments now have a wealth of advice to follow about how to generate statistics that the public will see as possessing procedural integrity. Yet something has also been lost in the way the discussion has evolved. Nowhere do we find a checklist that citizens can use to judge whether any given statistic should be accepted. It is almost as if the need to secure public trust has discouraged the establishment of quality criteria which individual statistical series might fail. Paradoxically, however, the absence of such criteria may now be undermining that very trust, judging by the frequency in the public square of charges of *dodgy data*, *rubbery figures*, *alternative facts*,or[*GIGO*](https://en.wikipedia.org/wiki/Garbage_in,_garbage_out).

This paper is a first attempt to identify some potential tests by which data users might judge whether to accept the numbers under their notice as knowledge. The discussion is divided into sections on *measurability* (how susceptible the target variable is to exact measurement), *measure* (the role of concepts and definitions in arriving at a correct representation of the target), and *measurement* (how the process of gathering and processing data may affect the accuracy of the resulting numbers).

## 2. Measurability

The simplest form of statistic is an *enumeration*. If I count the toes on my feet, or the apples in a barrel, then I shall arrive at an exact number, and if I do the job diligently, I may expect this number to be correct.

At the other end of the scale, some things cannot be enumerated, although attempts may be made to give them numerical expression. This especially applies to qualities rather than quantities. Business “confidence”, employers’ “willingness” to hire staff, the “liveability” of cities, as well as optimism, happiness, well-being, generosity, and other moods, intentions, or moral or ethical states, are not countable. However, they are of interest, and must be expressed in numbers if they are to be compared over time and between parties. Hence they may be worked into figures by one technique or another, though the results must remain largely arbitrary.

The general rule is that *simplicity* and *tangibility* of objects improve their measurability. The most accurate statistics relate to countable objects. Objects in this sense may include animate objects, as long as their living nature does not impede their identification as objects of measurement. If one of my toes, or an apple in the barrel, is split or deformed, the question may arise whether it should be counted, or perhaps counted twice.

It is also important to appreciate the *temporality* of measurability. Measuring is an instantaneous act bringing together the measurer, the measure and the object of measurement. Only objects present at the moment of measurement may be apprehended. This means, first, that the past cannot be directly measured. Only the present evidence of that past is available to be measured. Moreover, the quality of this evidence generally decreases with time, so that a count made from the present evidence of a past state becomes less reliable as that state recedes further into history. Nevertheless, the same target may eventually be estimated more accurately if new techniques improve the quality of evidence available about the past.

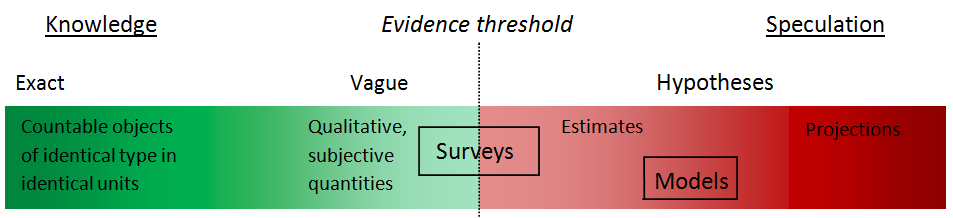
A further implication of the fact that measurability exists only in the present is that, once they have been performed, measurements already relate to the past. Measurability does not exist for future objects, since those objects do not yet exist. It follows that all *projections* should be treated from a scientific point of view as hypotheses rather than findings. Since direct measurement of the future is not possible, projections are often derived from models, which are themselves often composed largely of hypotheses about how variables relate to one another. Model outputs should be viewed as speculations, to be confirmed by actual measurements in future. In essence, they are not statistics at all.

In sum, from the point of view of the measurability of their targets, we may regard published figures as falling into one of two broad categories: knowledge, and hypotheses or speculation. Within these categories are certain gradations. Knowledge may be exact or vague, and hypotheses may be more or less grounded in existing observations.

The two categories are not quite watertight, and some forms of statistics straddle the divide. This particularly applies when the dimension of accessibility of information is taken into account. For many variables, the true figure is not known, and resort is made to *surveys*. Surveys are here taken to mean the collection of actual data, but which do not cover the whole population concerned. The raw results of such surveys may be regarded as knowledge, but knowledge of a limited value since it does not relate to the totality of the category under study. Survey data are often presented as percentages, with the suggestion that the percentages can be taken to apply to the whole population, perhaps with an error margin based on the size of the sample and its share of the estimated total population. In fact, applying survey percentages to the whole population produces estimates, or speculations, the reliability of which depends on the size of the sample, the extent to which it appears representative of the whole, the clarity of the measure and the diligence of the measurement.

Beyond data generated by surveys conceived for statistical purposes, much use is now also made of data already available in existing records or through automatic logging of actions or transactions. Nevertheless, similar considerations of measurability apply to these sources, whether they be “administrative data”, “big data”, or data from scanners, webclicks, webscraping, or other sources. In all cases, the part of the variable accessible to measurement, and its relation to the whole, must be carefully assessed.

The essential features of the measurability scale implied by the above discussion are depicted below, with the “evidence threshold” marking the boundary between knowledge and speculation.



## 3. Measure

This section deals with how statistical concepts and definitions can promote or impede accurate measurement.

Statistical measures are instruments which translate phenomena into numbers. So the first step in ensuring the accuracy of a measure is to make its relation to the target phenomenon clear. The measure must define its object of measurement in a way that leaves no doubt what will be counted and what will not.

Sometimes a mere term will be sufficient. “Persons”, “tonnes” or “dollars” are readily identifiable by all sane observers. Until recently the same might have applied to “men” and “women”, despite some admitted marginal cases, but political discussions now cloud these categories. Wherever vagueness or ambiguity is present, mere terms will have to be supplemented by definitions that impose objective tests to consistently identify the objects of measurement. Good and effective definitions possess both *exhaustiveness* and *exclusiveness*: they identify all and only those objects that are to be measured. It helps if the objects of measurement themselves form a logical and homogenous whole.

Measures will typically also require specifications of time. *Stock* measures relate to a moment in time; *flow* measures to a period of time. Locations or *points of measurement* must also be defined so as to avoid multiple counting of the same item. For “stock” objects such as persons or commodities, this requires their unique localisation at the instant of measurement. For “flow” objects – and especially for money, which can pass through many hands before being exchanged for goods or services – careful thinking may be required to fix the point of measurement in a way that avoids unwarranted multiple recording.

Many different problems may arise in relation to *units of measurement*. A common error with money measures is to express them in “real” terms without specifying the base year. And data on technological subjects may be clouded by a misplaced urge to simplify units of measurement. Thus one sees the output of power plants expressed in terms of the number of “homes” that it could serve, ignoring the fact that households’ use of electricity varies by season and time of day and in any case accounts for only part of total demand. Press articles also often confuse megawatts, which measure instantaneous *power*, with megawatt-hours of *energy*, or temperature (the intensity of heat at a point) with enthalpy (the heat content of a system). The use of inappropriate measures vitiates measurement.

To sum up, the following may be considered as potential tests of the soundness of a statistical measure and the hence of the reliability of associated data:

1. A good statistical measure starts with a sound and well-understood concept expressed in a definition which precisely identifies the target of measurement.
2. If a definition requires multiple dimensions, then it must deal with all possible combinations of these dimensions in a way that clearly includes or excludes all potentially concerned phenomena.
3. In general, the definition needs to be clear, unambiguous, exclusive and exhaustive. This may require sub-definitions of terms used, and explicit instructions about special cases.
4. Units of measurement, points of measurement, the moment of a stock measurement, and the period of a flow measurement must all be specified.

## 4. Measurement

*Certainty of identification* remains an issue at the measurement stage. If identification is by the enumerator, then some level of consistency may be expected, though the number and competence of the enumerators will also play a role. But if the targets identify themselves, the prospects of a strictly accurate count are compromised. The degree of inaccuracy introduced may vary with the parameter involved. Statistics by age or sex may only be affected to the extent that respondents lie, are incapable of correctly identifying themselves, refuse to answer, or are of ambiguous sex. Statistics on religious faith or on other beliefs will generally be more inaccurate, as the categories are more open to interpretation, and the self-image of respondents may diverge from the assessment of an objective enumerator. Even more inaccuracy is to be expected in responses on matters which may be the subject of pride, shame, reward or penalty.[[1]](#footnote-1)

The method of measurement also has important implications for accuracy. As already mentioned, statistical measurements have traditionally been of two essential types: *censuses*, where the whole population is recorded, and *sample surveys*. In principle, censuses produce more reliable data, since estimation is limited to filling gaps created by non-responses. However, census data could only be perfectly accurate if all the target population were reached.

Traditional censuses have been the mainstay of official statistics throughout modern times but may now be dying out, as technology provides governments with all they need to know. Denmark has been a pioneer in this regard. Every individual, business and dwelling in the country is numbered, and data can be matched or extrapolated across the governmental system to produce information on population, employment, use of transport and government services etc. Other countries are heading in the same direction, but the digital transition is proving problematic. Australia still ran a census in 2016, but encouraged respondents to complete the forms online on the evening of 9 August. However, the system crashed at the vital time, leaving millions unable to file their returns. Access was not restored until nearly two days later, and the Prime Minister ordered an enquiry to determine “[which heads roll, where and when](http://www.abc.net.au/news/2016-08-11/census-malcolm-turnbull-slams-abs-over-failure/7718584).”[[2]](#footnote-2)

The contrasting experiences of these two countries show the advantages for statistical reliability of adopting a *single consistent approach to data collection*. This also applies in censuses of businesses, industries, or agricultural activities.

Sample surveys introduce issues of *representativity*: as already mentioned, any figures presented for the whole population from which the sample is drawn are merely estimates that depend for their accuracy not just on the extensiveness of the survey, but on the degree of conformity of the sample to the whole. Assuring representativity of a sample in all relevant dimensions is thus key to the reliability of a survey-based estimate.

Especially in surveys, it is important for accuracy that those collecting data do not have *personal or institutional incentives* to either exaggerate or minimise the phenomenon they are counting. In particular, data which violate the provision of the [OECD Recommendation on Good Statistical Practice](http://www.oecd.org/statistics/good-practice-toolkit/Brochure-Good-Stat-Practices.pdf) that statisticians need to be “professionally independent from other policy, regulatory or administrative departments and bodies” should be treated with caution, especially if the measure in question has been made the subject of a *target*. Raising or spending predicted volumes of money, reducing waiting times for government services, improving clean-up rates for reported crime, or making the trains run on time may all become matters of announced targets, and figures showing whether the targets have been achieved will be more reliable to the extent that they are collected by officials with no incentive to “cook the books”.

Sometimes no incentives are required for bias to be present. It is sufficient for enumerators to have a *firm opinion about the subject* of their count. If this is the case, one will almost always find that the figures published support the enumerators’ prior opinion. This is the opposite of the “scientific principles and professional ethics” mentioned in the [UN Fundamental Principles](https://unstats.un.org/unsd/dnss/gp/FP-New-E.pdf), but it is common in academic debates and the bespoke data collections of think-tanks and lobby groups.

To sum up, accuracy of measurement can be assessed by examining:

1. The comprehensiveness of the count
2. The number and competence of the enumerators
3. The ease or difficulty in practice of making an unmistakeable identification
4. Whether the identification is performed by the enumerator or the enumerated
5. The presence or absence of institutional incentives or biases
6. Personal biases towards obtaining one result or another
7. Whether results agree with other reliable measurements.

## 5. Conclusion

This paper is a first attempt to identify some characteristics of statistics that may be accepted as true. It has followed the stages of statistical production, first examining features affecting the measurability of objects, then considering what makes a good statistical measure, and finally focusing on the act of measurement itself.

It concludes that current lists of statistical principles and good practices, instructions on how to collect specific statistics, and statements of the dimensions of statistical quality, provide only limited guidance on the question which statistics should be accepted as knowledge.

The paper has suggested a number of possible criteria for application at each of the three stages of statistical production to help judge whether a given statistic is true. If further work is done in this area, it may be possible to arrive at agreed checklists of statistical reliability, perhaps differentiated by the type of data or field of enquiry.

One might hope for several benefits from such checklists. First, they could contribute to improving knowledge, especially by removing from consideration statistics that failed the criteria. Second, they could contribute to the elaboration of new and better data, by incorporating the desiderata on the checklists into the design of statistical collections. Both of these benefits could help improve the basis on which new hypotheses, research strategies, and policies are constructed.

Beyond these simple benefits, any patterns that emerge from determining which statistics pass reliability tests may also contribute to eventual revisions of the existing general codes of principles and good practices.

1. Huff (op. cit., pp. 132-3) gives the example of a survey of “8 000 representative British homes” which asked British men and women how often they took a bath. He rightly points out that “saying and doing may not be the same thing at all.” [↑](#footnote-ref-1)
2. The Prime Minister was speaking metaphorically as Australia had abolished the death penalty for federal offences in 1973. [↑](#footnote-ref-2)