But Are those Numbers Correct? Some Suggestions for Appraising the Accuracy of Statistics

Simon Scott [1] | *Organisation for Economic Cooperation and Development, Paris, France*

Abstract

Knowing whether data are reliable is of fundamental importance to the establishment of knowledge, the formulation of explanatory hypotheses, and the development of effective policy. Yet there appear to be no standard, established tests to enable users to judge whether they should accept a given statistic as a fact. Numerous internationally agreed documents set out principles and practices to promote sound statistics, but they offer no direct guidance on whether to accept data as presented. Other documents discuss statistical quality, but focus largely on utilitarian considerations such as availability and timeliness; when they do discuss accuracy, they again consider processes (lists of good practices) rather than results (are the data correct?). This paper is a plea for, and a first attempt at, identifying some characteristics of data that may be accepted as true.[2]

Keywords	***JEL code***
Statistical quality, accuracy, reliability, knowledge, truth	*C10*

INTRODUCTION – 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, and also established some rights of privacy, e.g. to refuse to disclose religious belief. Census officials were required to maintain the secrecy of personal data and the confidentiality of commercially sensitive information. Such measures helped ensure an accurate count by compelling the respondent to provide correct information and assuring him that he would not suffer as a result.[3]

[1] Counsellor, Statistics and Data Directorate, Organisation for Economic Cooperation and Development (OECD), 2 rue André-Pascal, 75775 Paris Cedex 16, France. E-mail: simon.scott@oecd.org.

[2] This paper, presented in draft at the *2018 European Conference on Quality in Statistics* held at Krakow, Poland, during 27–29 June 2018, should not be reported as representing the official views of the OECD or of its member countries. The opinions expressed and arguments employed are those of the author only.

[3] Thus, for example, Presidential instructions for conducting the U.S. census of 1840 recognised in the following terms that accuracy required completeness, which in turn would be promoted by respect for privacy: "Objections, it has been

In the 19th and 20th centuries, statistical quality was further promoted through detailed instruction manuals regulating specialised statistical collections. Prominent examples include the standard guides on compiling the national accounts and the balance of payments, or government debt and deficit figures. To the extent that one could be sure that these instructions had been followed, and that no "tricks" had been played, one would be inclined to trust the resulting data – yet it would be difficult for any layman to make judgments on the fidelity with which the instructions had been followed, and in practice, separate judgments would be required on each country's implementation of each set of instructions.

It was in recognition of the generalised need for honesty and probity in statistical activities that, beginning in the 1990s, a new species of document emerged that abstracted from these laws and procedures principles thought to be of general application. The first such document was the *Fundamental Principles of Official Statistics in the Region of the Economic Commission for Europe* (ECE), agreed in 1992 (UNECE, 1992). This aimed to guide the newly free States of the former Soviet bloc to develop statistics that would command public trust because they respected "the fundamental values and principles which are the basis for any democratic society." The emphasis on producing information worthy of trust in a democratic society has persisted in this family of documents down to the present day. The UNECE principles were later adopted as a global standard in the *UN Fundamental Principles of Official Statistics* (UN 2014; original 1994) and the UN also developed *Principles governing international statistical activities* (UN, 2006). Key regional documents in the same group include the *European Statistics Code of Practice* (Eurostat, 2017), the *Code of Good Practice in Statistics for Latin America and the Caribbean* (ECLAC, 2011), and the especially comprehensive *Recommendation of the OECD Council on Good Statistical Practice* (OECD, 2015).

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 guidance on whether to accept data coming under our notice, we might turn in hope to another class of document, namely the presentations by various agencies of *dimensions of statistical quality*. Here again, however, we find no direct answer to the basic question of how to assess whether a number is true, or likely to be true. Indeed a UN study of international agencies' approaches to statistical quality (de Vries, 2002) 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.*

As de Vries' analysis shows, 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 by this throng of virtues, accuracy and reliability have retained only a minor place, and even so are defined in ways that somehow evade the question of whether the data are actually true. For example, de Vries reports

suggested, may possibly arise on the part of some persons to give the statistical information required by the act, upon the ground of disinclination to expose their private affairs. Such, however, is not the intent, nor can be the effect, of answering ingenuously the interrogatories. On the statistical tables no name is inserted – the figures stand opposite no man's name; and therefore the objection can not apply. It is, moreover, inculcated upon the assistant that he consider all communications made to him in the performance of this duty, relative to the business of the people, as strictly confidential" (Wright and Hunt, 1900, p. 145). It is especially easy to trace the evolution of census legislation in the United States of America, as the US Census Bureau has published on its website (US Census Bureau n.d.) detailed documentation, starting with the Constitution itself and the first census Act passed in 1790.

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 “estimated” tends to suggest that absolute accuracy is impossible, no matter how simple the count might be.

Overall, it is clear that thinking about the soundness of statistics has focused on procedures by which official bodies can generate data useful to society.[4] 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 it is difficult to avoid the impression that something has 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 – at least if we are to judge by how frequently we hear charges of *fake news, dodgy data*, *rubbery figures*, *alternative facts*, or *GIGO*.

This paper represents an initial 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). These are loose categories, and they overlap; they should not be seen as an attempt at typology but only as a means of giving structure to the argument.

1 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. Nevertheless, 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.

[4] For further references to relevant literature on statistical quality, focusing on international macroeconomic data, see the IMF’s useful Data Quality Reference Site (IMF, n.d.). An anonymous reviewer has also rightly pointed to the existence of extensive broader literatures on data quality assessment (for a review, see Batini et al., 2009), as well as on metrology, on mathematical measurement theory, and on measure theory as a branch of mathematics. Each of these is a specialised and sophisticated discipline, geared towards the improvement of quantification through the application of professional skills and knowledge. While their discoveries and methods are, as a rule, beyond the grasp of the average citizen attempting to judge the reliability of a given statistic, they do repose on logical and empirical principles the study of which may suggest further practical tests to those sketched in this paper.

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 a measurement may be apprehended in that measurement. This means, first, that past states cannot be directly measured. Only the surviving evidence of that past state 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 or measurability of evidence available about the past.

A further implication of the fact that measurability is a finite act occurring at a specific time is that, once they have been performed, measurements already relate to the past. By the same token, measurability does not exist now 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 often include 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 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 from only a sample of 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 involved. 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 represents the whole, the clarity of the measure and the diligence of the measurement.[5]

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. Yet 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.

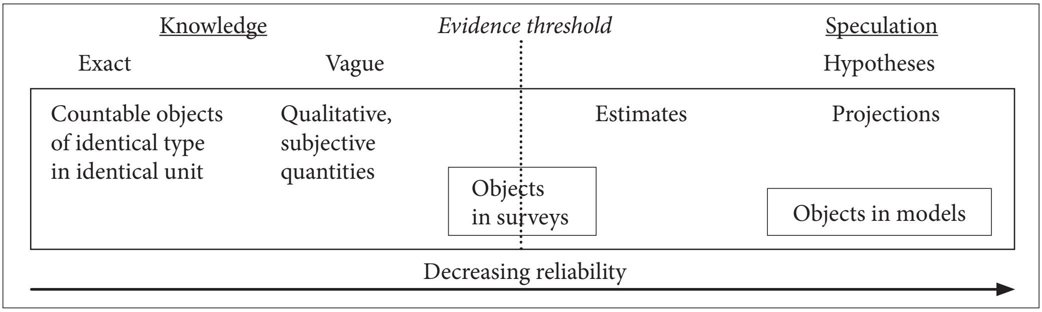
The essential features of the measurability scale implied by the above discussion are depicted in Figure 1. Items shown in the box are statistical quantities of various types. The line marked "evidence threshold" roughly marks the boundary between knowledge and speculation.

2 MEASURE

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

[5] US President Donald Trump evinced a realistic attitude to survey percentages in a speech at Wheeling, West Virginia on 29th September 2018. After citing numbers from an opinion poll showing strong support for a course of action he favoured, he observed "Hey, it's a poll. But we love those polls, don't we? I love polls. Only when they're good; when they're not good, I don't talk about them" (Trump, 2018).

Figure 1 Gradiations of the measurability of objects



Source: Own construction

Statistical measures are instruments which translate phenomena into numbers. So the first step in ensuring the reliability 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.[6] Good and effective definitions possess both *exhaustiveness* and *exclusiveness*: they identify all and only those objects that are to be measured.

Tight definition is easier to achieve if the objects of measurement themselves form a logical and homogenous whole. Absolute homogeneity, or identity, is not required, but the objects counted under a measure must all possess some identifiable property which distinguishes them as a group and which separates them from other things which will not be counted. This identifiable quality must also be expressed in a single unit of measurement, such as tonnes, dollars, numbers of persons, etc. A single measure should never include within it different quantities, such as currency units of different countries, or real and nominal monetary units. A measure must always have one and only one *unit of measurement*; otherwise, the resulting number is meaningless, as it relates to no identifiable quantity.

Many different problems may arise in relation to units of measurement. A common error with money measures is to express them in "real" terms – i.e. at constant prices – without specifying the base year.[7] 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" they could serve, ignoring the fact that households' use of electricity varies by season and time of day and in any case accounts

[6] An example well-known to the author is that of "official development assistance" (ODA) – which has become the standard measure of government foreign aid. At first this was merely a descriptive term, qualified only by the observation that ODA was "intended to be concessional in character". Initial attempts to sharpen the definition focused on the source of the funds, but attention then shifted to the need to define "concessional". At one point, a qualitative definition requiring that the terms of ODA transactions be "significantly softer than the terms normally available for commercial transactions" was almost agreed. But in the end it was found necessary to introduce a strict mathematical test – loans would be reportable as ODA only if they embodied a grant element of at least 25 per cent, using a 10 per cent discount rate. The whole process took nearly four years, from early 1969 to late 1972 (see Scott, 2015).

[7] A Google search for "images" containing "constant prices" will disclose dozens of examples. See, for example, the Trading Economics page on Czech GDP (Trading Economics n.d.) which fails to supply the base year (2010) in the introductory statement, the chart, or the table – though it is finally mentioned in a box.

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 units vitiates measurement.

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 and after 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.

The following may be considered as potential tests of the soundness of a statistical measure and hence of the reliability of associated data:

a. A good statistical measure starts with a sound and well-understood concept expressed in a definition which precisely identifies the target of measurement.
b. 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.
c. 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.
d. A measure must never mix quantities: it should always possess a single, clear unit of measurement.
e. Units of measurement, points of measurement, the moment of a stock measurement, and the period of a flow measurement must all be specified.

3 MEASUREMENT

Certainty of identification remains an issue at the measurement stage. If identification is done 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 of the enumeration identify themselves, then 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 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 enumerator. Even more inaccuracy is to be expected in responses on matters which may be the subject of pride, shame, reward or penalty.[8]

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

[8] Huff (1954, pp. 132–133) gives the example of a survey of "8 000 representative British homes" which asked British men and women to say how often they took a bath. He rightly points out that "saying and doing may not be the same thing at all".

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 9th 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"[9] (ABC News, 2016).

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. Attaining representativity of a sample in all relevant dimensions is thus vital to ensure 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* (OECD, 2015) 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* (UN, 2014), 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:

a. The comprehensiveness of the count.
b. The number and competence of the enumerators.
c. The ease or difficulty in practice of making an unmistakeable identification.
d. Whether the identification is performed by the enumerator or the enumerated.
e. The presence or absence of institutional incentives or biases.
f. Personal biases towards obtaining one result or another.
g. The extent to which results are corroborated by other reliable measurements.

CONCLUSION

Current lists of statistical principles and good practices, instructions on how to collect specific statistics, and statements of the dimensions of statistical quality, do not provide – and do not attempt to provide – comprehensive guidance as to which statistics should be accepted as knowledge.

Yet in an era of "fake news", the public has never been in greater need of a set of objective criteria by which to judge the reliability of data as presented. This paper has therefore made a first attempt at suggesting potential aids to judgment. It has been organised according to three broad elements or

[9] The Prime Minister was speaking metaphorically as Australia had abolished the death penalty for federal offences in 1973.

stages of the statistical process, so as to offer guidance relating to the inherent measurability of the objects being quantified, the soundness of the statistical measure being applied, and the diligence and faithfulness of the act of measurement.

If further work is done in this area – whether by officials, academics, or civil society groups – then it may be possible over time to arrive at widely accepted checklists of statistical reliability, perhaps differentiated according to broad types of data or fields of enquiry.

One might hope for at least two benefits from such checklists. First, they could contribute to improving knowledge, especially by removing from consideration statistics that failed the criteria. Second, they could foster 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 the work of determining which statistics pass reliability tests may also contribute to eventual revisions of the existing general principles and codes of good practice.

References

ABC NEWS. *Census: Survey back online as angry Malcolm Turnbull warns "heads will roll"* [online]. Australian Broadcasting Corporation, 11th August 2016. <http://www.abc.net.au/news/2016-08-11/census-malcolm-turnbull-slams-abs-over-failure/7718584>.

BATINI, C., FRANCALANCI, C., CAPPIELLO, C., MAURINO, A. Methodologies for Data Quality Assessment and Improvement [online]. *ACM Computing Surveys*, 41(3), July 2009. <https://www.researchgate.net/publication/220565749_Methodologies_for_Data_Quality_Assessment_and_Improvement/download>.

DE VRIES, W. *Dimensions of Statistical Quality: A discussion note about the quality initiatives of some international organisations* [online]. New York: United Nations, 22nd August 2002. <https://unstats.un.org/unsd/accsub/2002docs/sa-02-6add1.pdf>.

ECLAC. *Code of Good Practice in Statistics for Latin America and the Caribbean adopted by the Statistical Conference of the Americas of the ECLAC* [online]. 6th meeting, Bávaro, Dominican Republic, November 2011. <https://repositorio.cepal.org/bitstream/handle/11362/16423/1/FILE_148024_en.pdf>.

EUROSTAT. *European Statistics Code of Practice* [online]. Luxembourg: Eurostat, 2017 [orig. 2005]. <https://ec.europa.eu/eurostat/documents/64157/4392716/Revised_CoP_Nov_2017.pdf>.

HUFF, D. *How to Lie with Statistics*. New York and London: W. W. Norton, 1954.

IMF. *Data Quality Reference Site* [online]. Washington: International Monetary Fund. <https://dsbb.imf.org/dqrs/approaches-to-data-quality>.

OECD. *Recommendation of the OECD Council on Good Statistical Practice* [online]. Paris: OECD, 2015. <http://www.oecd.org/statistics/good-practice-toolkit/Brochure-Good-Stat-Practices.pdf>.

SCOTT, S. *The Accidental Birth of Official Development Assistance* [online]. Paris: OECD, 2015. <https://www.oecd-ilibrary.org/development/the-accidental-birth-of-official-development-assistance_5jrs552w8736-en>.

TRADING ECONOMICS. *Trading Economics Website* [online]. <https://tradingeconomics.com/czech-republic/gdp-constant-prices>.

TRUMP, D. J. *President Trump EXPLOSIVE Speech at MASSIVE Rally* [online]. Wheeling, West Virginia, 29th September 2019. <https://youtu.be/x5sLPW0yYkI?t=553>.

UN. *Principles governing international statistical activities* [online]. New York: United Nations, 2006. <https://unstats.un.org/unsd/statcom/37th-session/documents/2006-13-principles-E.pdf>.

UN. *Fundamental Principles of Official Statistics* [online]. New York: United Nations, 2014. [cit. 29.1.2014]. <https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx>.

UNECE. *The Fundamental Principles of Official Statistics in the Region of the Economic Commission for Europe* [online]. Geneva: United Nations, 1992. <https://www.unece.org/fileadmin/DAM/stats/documents/e/1992/32.e.pdf>.

US CENSUS BUREAU. *Website of US Census Bureau* [online]. Washington. <https://www.census.gov/history/www/through_the_decades/overview/>.

WRIGHT, C. D. AND HUNT, W. C. *The History and Growth of the United States Census* [online]. Washington, Government Printing Office, 1900. <https://www.census.gov/history/pdf/wright-hunt.pdf>.