Some Thoughts on Official Statistics and its Future (discussion paper)

Yves Tillé[*]

September 1, 2020

Abstract

In this article, we share some reflections on the state of statistical science and its evolution in the production systems of official statistics. Data sources and methods are evolving, raising questions about the future of official statistics. The history of the methods used deserves a closer look at the changes that are taking place in the world of official statistics.

Keywords: deduction, induction, Lasso, p-value, registers, sampling, statistical learning

1 Introduction

Official statistics is a somewhat special field of statistics. The methods used there have been developed to deal with original questions that revolve essentially around quality. In this article, we try to glimpse the future of statistical methodology. Currently, statistics is confronted with many epistemological questions, the most emblematic of which is the crisis in the p-value. We describe the specificities of official statistics by tracing the history of sources and the history of controversies around methodology. Then, we analyze the impact of new data sources and new statistical methods. An evolution of the methodology will be necessary and must be supported by quality fundamental research.

2 The Statistical Science

Statistics is a science that aims to study a reality through the processing, analysis, modeling and interpretation of these data. Traditionally, there are several statistical approaches.

1. *Descriptive or exploratory statistics*, which consists of presenting data in a more condensed way by means of tables and graphs. The aim consists of reducing the complexity of the data by means of dimensionality reduction or cluster analysis. Descriptive statistics is above all linked to an interpretive and exploratory process. John Wilder Tukey was a strong supporter of descriptive statistics. His book

[*]University of Neuchatel, Bellevaux 51, 2000 Neuchatel, Switzerland, yves.tille@unine.ch

Exploratory Data Analysis (Tukey, 1977) remains a reference of the approach. Similarly, in France, the school of data analysis initiated by Jean-Paul Benzécri promoted descriptive and exploratory statistics (Benzécri, 1973a,b; Bastin et al., 1980).

2. *Analytical or inferential* statistic which aims to deduce properties on a population from data. This population can be notional and be, for example, a probability distribution or a model. Inferential statistics rely entirely on probability calculations. It includes the theory of statistical hypothesis testing and decision theory. The main founders of this theory are Karl Pearson, William Sealy Gosset, Ronald Fisher and Jerzy Neyman.

3. *Modeling* consists in describing reality by a general model described by one or more equations. A model is necessarily a simpler approximation than reality, as Box & Draper (2007) wrote: "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful". Models can be used either to describe relationships between variables or to make predictions.

3 Interpretation of the Statistical Approaches

One usually distinguishes two scientific approaches: The *inductive approach* is based on observations and data and aims to build a general theory. The *deductive approach* (or hypothetico-deductive) is based on a theory. It aims to deduce particular results from a general theory in order to check whether these particular results can be confirmed by observations.

Exploratory statistics can be associated with an inductive approach. In this approach, we start from the data to try to empirically find a global explanation. Inferential statistics and more particularly the theory of statistical hypothesis testing can be seen as a hypothetico-deductive approach. One formulates a hypothesis and then we decide on that hypothesis by confronting it with data. The two approaches are qualified as complementary. The exploratory analysis allows hypotheses to be formulated. Then the inferential statistic can possibly confirm them. A return to exploratory analysis finally makes it possible to formulate new hypotheses and so on. This is referred to as the induction-deduction cycle which would be the engine of the production of knowledge. Often this cycle has been completely perverted, for example, by using the same data to formulate and to test hypotheses which makes the application of test theory completely lapsed. This approach is ironically called HARKing by Norbert L. Kerr (1998) whose acronym comes from *Hypothesizing After the Results are Known.*

However, this vision of the induction/deduction opposition is far from being unanimously shared. The role of induction has been the subject of controversy between Jerzy Neyman and Ronald Fisher. Fisher (1935) promoted the inductive reasoning. Jerzy Neyman (1957) had a more decision-making approach of the hypothesis testing theory and then he was reluctant to the inductive reasoning. Inferential statistics can be considered as an inductive reasoning because it aims to extrapolate the results of a sample to a population or to a model. It is therefore a generalization process which can be seen as fundamentally inductive (see among others Lehmann, 1993; Capel et al., 1996).

Costantini & Galavotti (1986) assert that estimation methods, such as the maximum likelihood method, correspond to an inductive approach and theory of hypothesis testing typically corresponds to a deductive approach. Other interpretations exist. For example, Adrew Gelman (2011) discusses a philosophy of statistics that sees frequentist statistics as deductive and Bayesian statistics as inductive.

Faced with these different interpretations, we can only subscribe to the assertion of Capel et al. (1996) who wrote: "In any event, a real understanding of the role of inductive reasoning in the humanities that is common to the statistician, the researcher and the practitioner clearly does not exist, and in particular, as we have seen, with regard to the use of hypothesis tests"[1].

Moreover, the distinction between induction and deduction is sometimes considered outdated. Karl Popper (2005) rejects the description of the scientific process as an induction/deduction cycle. He outright rejects the interest of the inductive approach in science: "Yet even supposing this were the case – for after all, 'the whole of science' might err – I should still contend that a principle of induction is superfluous, and that it must lead to logical inconsistencies." Karl Popper maintains that the scientific method consists in formulating scientific propositions which must be able to be falsified by an experiment. Indeed, no theory can be proven by an experiment. The fact that a model is compatible with data never proves that the model is true. Indeed, another model could also be compatible with the same dataset.

The induction/deduction cycle also completely loses its meaning in certain statistical applications. In official statistics, this distinction can hardly be applied. Indeed, the objective often consists in estimating certain characteristics of a population using scattered sources (surveys, administrative files, censuses). The objective is therefore not to establish a scientific theory but simply to overcome the impossibility of obtaining a complete and correct measurement on all the units of the population of interest.

This induction/deduction cycle also loses its meaning with the advent of new methods known as *statistical learning* (support vector machine, neural network, nearest neighbor, sparse methods, random forests) which make it possible to predict without actually modeling. One could see these methods as typically inductive because they do not imply any a priori formalization of reality (Harman & Kulkarni, 2012). The word *statistical learning* is however a bit misleading because these methods do not really lead to a theorization which would be an automatic modeling of data. The so-called induction/deduction cycle is somehow bypassed. For example, if we do not know the income of an individual in a database, we can assign the income of the statistical unit that most closely resembles her/him. This is a forecast by the nearest neighbor method. Can we however say that this approach consists in modeling the income? At most, we can discuss what it means to "resemble", but we are often forced to use only the variables available in the files to define a distance between the units. However, this approach can be very effective. It is not clear how one could say that this approach would be inductive or deductive. For the statistician, the main question concerning this type of method is to evaluate and estimate its precision.

The new methods of *statistical learning* therefore shake up a basic principle of the scientific process which would like all knowledge to be able to be transmitted through

[1] Translated from French: "En tout état de cause, une compréhension réelle du rôle du raisonnement inductif dans les sciences humaines qui soit commune au statisticien, au chercheur et au praticien n'existe manifestement pas, et en particulier, comme nous l'avons vu, en ce qui concerne l'usage des tests d'hypothèse."

discourse. While forecasts by *neural network, nearest neighbors* or *random forest* can work very well to make a prediction, they do not allow us to establish a general theory or a principle that would allow us to understand the relationships between variables. To make a new forecast, we can at best restart the algorithm which, for some methods with randomization, can provide slightly different results at each run.

4 Crisis of the p-value

In statistical hypothesis testing, the p-value is the probability that under a hypothesis (conventionally called the null hypothesis), we obtain the same value or an even more extreme value than that obtained with observed data. If the p-value is low, then we reject this hypothesis. The error of type 1 is defined as the probability of rejecting the null hypothesis given that it is true. If we perform the test with an error of type 1 of 5% for example, we reject the hypothesis if the p-value is less than 5%.

The p-value has become a decisive argument in many sciences: humanities, economics, finance or biology. However, in scientific journals of statistics, one may be surprised at the very small number of p-values used in published articles. The p-value is often misinterpreted as the probability that the null hypothesis is true, which is obviously not the case.

In many publications several p-values (sometimes tens) appear without any reflection on the probability of having at least one p-value less than 5% if all the null hypotheses were true. Very often, researchers perform a large number of hypothesis tests and only publish results for p-values below 5%. This approach is similar to HARKing. Another methodological error consists in identifying a model by choosing the variables having p-values less than 5% among a very large number of variables. This procedure inevitably leads to over-specification of the model. However, new solutions have been developed to select variables using for example the Lasso method (Tibshirani, 1996, 2011).

These uses of p-values were denounced in a provocative article: *Why Most Published Research Findings Are False* (Ioannidis, 2005). The subject has become controversial to such an extent that one can say that there is a crisis in statistics (Gelman & Loken, 2014; Fraser & Reid, 2016) and that the *American Statistical Association* felt compelled to issue a statement on the p-value (Wasserstein & Lazar, 2016).

5 What about Official Statistics?

The Swiss Official Statistics Charter is available on the website of the Ethics Council for Official Statistics (Conseil d'éthique de la statistique, 2012). In this charter, the mission of official statistics is defined: "The mission of official statistics is to meet the needs for statistical information of general interest of society as well as those relating to the conduct of public policies"[2].

The mission therefore does not consist of interpreting, modeling, establishing knowledge, or deciding. The induction-deduction cycle for the production of knowledge is

[2] Translated from French: "La statistique publique a pour mission de répondre aux besoins d'informations statistiques d'intérêt général de la société ainsi qu'à ceux relatifs à la conduite des politiques publiques."

thus not an appropriate framework in official statistics. In fact, in official statistics, the approach is neither exploratory nor decisional. The statistical methods developed are specific, because it is not a question of carrying out scientific research to establish new knowledge, but of meeting information needs by providing reliable, lasting and high-quality statistics.

There exists however a cycle between civil and political society and official statistics in order to determine the needs. Official statistics must ensure continuity but also renew their statistical production by opening up to new themes such as gender inequalities or environmental concerns. The statistical methods used in official statistics are therefore neither exploratory nor decision-making. Official statisticians have focused on quality. An exemplary document is *The Statistics Canada's Quality Assurance Framework* whose basic principles are relevance, accuracy, timeliness, accessibility, coherence, interpretability (Statistics Canada, 2017).

6 Data Sources in Official Statistics and their Integration

Official statistics have therefore developed somewhat specific methods to meet their objectives. The advantage is that official statistics are immune to the p-value crisis because its use is relatively limited. The history of official statistics methods can be summed up in a few eras (see among others Hansen & Madow, 1974; Kruskal & Mosteller, 1980; Hansen, 1987; Bellhouse, 1988; Bethlehem et al., 2009; Tillé, 2020). First, there is the era of censuses which covers the entire 19th century. During this period, only comprehensive data compilation was considered scientific. This doctrine is clearly stated by the statistician Adolphe Quételet (1846). The rupture is initiated by the director of the Norwegian Institute of Statistics: Anders Nicolai Kiær (1896, 1899, 1903, 1905) who proposed to use partial data and therefore a sample. After a long controversy, the idea of using samples was finally accepted by the International Statistical Institute (Jensen, 1926). Ken Brewer (2013) interprets this debate as the first controversy in survey sampling.

This opens the era of sampling, which from the start has been the subject of assiduous and fruitful research, (notably on the part of Jerzy Neyman, 1934, 1938, 1952). He shows in particular that it is necessary to over-represent in the sample the categories where the dispersion is larger. This result renders the still too often used concept of representativeness obsolete. With the development of information technologies, then came the era of registers and administrative files in the 1970s. Administrative data are often presented as the new source of data. However, it should not be forgotten that some countries like Finland have had a population register for more than 50 years. The technical means for creating a register have been around for a very long time. The obstacles to the use of registers are above all organizational, political and legal. For a long time, they have obviously not been technical.

The development of new data sources has not put an end to previous practices. Many National Institutes of Statistics still carry out censuses and almost all of them carry out sample surveys. The integration of administrative files into official statistics is far from obvious. The registers depend on the administrative specificities of the countries, which does not facilitate statistical harmonization. Administrative files often contain many errors, as evidenced by the Serafe crisis in Switzerland.

Serafe has been a private company responsible for collecting audiovisual license fees in Switzerland since 2019. During the first invoices sent, a large number of errors were noted. Serafe had sent the invoices on the basis of the "control of inhabitants" registers updated by the municipalities. Many errors appeared to be related to household composition in multi-unit buildings or to obsolescence of data. These population control registers are also used for administrative and statistical purposes, but the impact of these errors is relatively small for these applications. In administrative files, it is often observed that only the variables which are necessary for immediate administrative operation are correctly updated.

We cannot therefore expect that all official statistics will be produced from the same type of source. Each type of source contains specific errors, whether they are data produced by statistical offices or not. Problems to be addressed can be sampling errors, errors due to non-response, under-coverage, over-coverage, duplicates, measurement errors, errors due to fraud. Official statistics are a perpetual struggle against these errors.

In each source, one can find a particular reliability. One type of reliability is the correct identification of statistical units, which can often be obtained through an administrative file or register. However, this register may contain poor quality or obsolete variables. A sample survey, even one with non-response, may contain more reliable measures for the variables of interest. The problem then comes down to combining what is most reliable in the different sources. The crucial issue in official statistics is thus the integration of data from different sources in order to best enhance the reliability of each source.

From the beginning, research in official statistics has been concerned with the problem of data integration. Almost a century ago, the controversy between, on the one hand, Corrado Gini and Luigi Galvani and, on the other hand, Jerzy Neyman (Gini & Galvani, 1929) revolves around balanced sampling and random sampling. Gini and Galvani had selected a sample of 29 districts (*circondari*) out of 214 so as to return the same means as those of the census for several known variables. It is therefore a question of using a source (the census) to improve the collection of a sample. Neyman criticizes this way of doing things, because the selection of the sample is not random and therefore it is not possible to make an inference. We obviously now know that it is possible to select a sample that is both balanced and random (Deville & Tillé, 2004). Ken Brewer (2013) interprets this discussion as the second controversy in survey sampling.

Another issue is the adjustment of survey data to census data. On this subject, the article by W. Edwards Deming and Frederick F. Stephan of 1940 is considered as a founding text in official statistics (Deming & Stephan, 1940). The problem dealt with is the adjustment of a table obtained by sampling on marginal totals known by a census. The integration of different sources is therefore present from the start of the use of sampling methods. For the record, the article is however mathematically false since it maintains that the *raking ratio* method is obtained by minimizing the chi-square distance under the constraints given by known marginal totals. This is obviously not true since the Kullback-Leibler divergence must be minimized. The fact remains that this article and all those that follow seek to optimize the integration of survey and census data.

The article of Deville & Särndal (1992) which defines the general method of calibration is the outcome of this research. These authors provide a methodology for calibrating survey data on census data while freeing itself from modeling. The originality is that the processing is carried out in order to obtain a weighting system applicable to any variable, which makes its application extremely practical. Deville and Särndal also provide a method for measuring the accuracy of the estimates obtained. The method has become essential and typical of official statistics. It is now applied by all survey statisticians (see also Särndal, 2007; Devaud & Tillé, 2019a,b).

A developing problem is the integration not only of two sources but of a multitude of distinct sources. The calibration methods can be generalized to harmonize several sources. These sources can be two samples or several samples and a census (see among others Guandalini & Tillé, 2017). Yang & Kim (2020) provide an overview of modern methods for integrating data from different sources.

The use of sources may change quite rapidly. Many European countries have abandoned large population censuses. Statistics based on these censuses are now produced by using registers and new surveys complete these registers. National Institutes of Statistics still carry out a lot of sample surveys. Sampling methods are increasingly used to check or improve the quality of censuses or registers. Thus, sampling may become more of a quality control tool than a direct method of producing data.

The evolution of practices does not go without raising new questions. Most emblematic has been the resignation of Martha Farnsworth Riche from the head of the United States Census Bureau. According to the Wikipedia contributors (2020) site: "Although she cited only personal reasons in her resignation, it was seen as a sign that Congressional Republicans were winning in their fight to prevent the Census Bureau from using sampling techniques to correct for persistent undercounting of minorities and other underrepresented groups." This political intervention in the methodology is obviously extremely worrying and goes against a principle of methodological independence of the National Institutes of Statistics.

7 Modeling in Official Statistics

The debate on the place of the model in survey sampling theory is interpreted by Ken Brewer (2013) as the third controversy in survey sampling. The idea of introducing a model to exploit the auxiliary information resulting from a census in a survey was initially proposed by this same Ken Brewer (1963). However, this idea was mainly developed by Royall (1970, 1971, 1976) (see also Valliant et al., 2000; Chambers & Clark, 2012). The model-based approach involves building a model that is estimated using the sample, and then predicting population data that is not in the sample.

The model approach is opposed to the so-called design-based approach which consists in weighting the sample units by the inverse of their probabilities of being selected. The initial survey weights are the inverse of the inclusion probabilities. These weights are then slightly modified by means of the calibration technique of Deville and Särndal in order to return exactly the totals of the variables known by a register or by a census.

Official statisticians are inherently reluctant to model. The reason is this: Modeling is somehow expressing an opinion on observations. Modeling can therefore be interpreted as contrary to the principle of impartiality of official statistics. However, model-based and design-based approaches are not necessarily contradictory. Särndal et al. (1992) advocate a model-assisted and nonetheless valid design-based approach. It is therefore possible to construct the concept of double robustness in the sense of an estimate that would be valid either when the model is correct or when the inclusion and response probabilities are correctly identified.

However, there are issues that cannot be solved without some modeling. Research concerning estimation for small areas has been very active during the last decades. These methods consist in producing estimates at very low levels (districts, municipalities) from a sample survey and a register or census. Once again, the problem is to make the best use of the two sources of information. The sample contains the variable of interest. The register contains the list of all units of the population and auxiliary variables. Composite estimators are often used, which are mixtures of direct estimates computed from the sample and estimates obtained from a model linking the auxiliary variables. These composite estimators can be obtained by means of mixed models for which the domains are random effects (see the very complete book of Rao & Molina, 2015).

The National Institutes of Statistics, however, have been relatively cautious and have rarely published estimates obtained by these methods. One of the reasons is that there may be peculiarities well known at the level of local communities that would not be taken into account by a model. For example, Molina & Strzalkowska-Kominiak (2020) proposed labor force estimates in Swiss districts using the Swiss Structural Survey. A border Swiss district of Lichtenstein was peculiar because a large number of Swiss border workers work in Lichtenstein. This district has therefore undergone special treatment. However, if it is not possible to identify the singularities before modelling, the estimates may turn out to be very far from reality.

Another area where it is difficult to work without modeling is non-response. There are two ways to do this. Either we can predict the missing values (imputation), or we can estimate the probability that these values are missing in order to weight the responding observations. The models can be very simple. One can impute by a closest neighbor. One can also impute by a simple ratio or by a prediction by regression. One can also predict by taking at random an individual belonging to a small homogeneous stratum. In order to avoid a bad specification of the model, one often seeks to treat the non-response in a doubly robust way. Two models are used: the first one allows us to predict the missing value and the second one allows us to predict the probability of being missing. Double robustness means that the estimation is then approximately unbiased if at least one of the two models is well specified (see on this subject Kang et al., 2007; Han & Wang, 2013; Kim & Haziza, 2014; Boistard et al., 2016; Chen & Haziza, 2017).

8 Big Data and Official Statistics

Tim Harford (2014) gives a strong opinion on big data: "As with so many buzzwords, "big data" is a vague term, often thrown around by people with something to sell. Some emphasise the sheer scale of the data sets that now exist – the Large Hadron Collider's computers, for example, store 15 petabytes a year of data, equivalent to about 15,000 years' worth of your favorite music." Indeed, we can often attend presentations of "experts" announcing the era of "big data" whose slides only contain lists of words starting with "V", numbers in yottabytes and potatoes linked by arrows without ever giving a real and concrete application. The slides are embellished with pontificating quotes and cartoons. Making lists of words beginning with a letter is the opposite of a serious scientific approach.

For official statistics, the era of big data is above all a myth. The fascination with big data can be irritating in several ways. Talking about "big data" in official statistics is not appropriate. Large administrative files such as population or business registers cannot be qualified as big data. They can contain millions of records but are processable over any desktop computer. These large files have been around in some countries for over fifty years and these countries have always found the computer means to manage them. They are not big data.

We will not use *web scrapping* to produce official statistics either. For example, it is a fantasy to think that we will calculate inflation better by going automatically to find prices on the Internet. First, because official statistics must be sustainable and the Internet is not stable in times. Secondly, because official statistics must have a methodology and documented sources. Because it is not enough to look at a price on the internet, you must also be able to verify them. And especially because the National Institutes of Statistics did not wait for the big data fashion to directly use the price files of the main distributors.

The huge data flows of social networks are not valuable either. They belong to the giants of the web. This data has immense commercial value for performing advertising profiling. It would also be extremely dangerous to become dependent on these giants. Above all, these data are not reliable in terms of identifying statistical units and potentially usable variables.

However, there are areas where the data is really massive, such as image analysis for territorial statistics. In this case, the data can be produced (aerial or satellite photos, cartography) by the country which uses them. However, image analysis is an area of research that emerged with the early days of computer science. Notable progress has been made in particular by using neural networks.

9 Statistical Learning in Official Statistics

We often talk about new statistical methods. Most of these methods are actually not so new and were all developed in the 20th century. Several methods (random forests, support vector machine, neural network, nearest neighbors) make it possible to make predictions without having to think about the relationships between the dependent variables and the variable of interest. Can we follow the famous sentence of Deng Xiaoping who said "it doesn't matter whether a cat is black or white, if it catches mice it is a good cat"? Can we use in official statistics methods that make it possible to predict without understanding? Indeed, the Statistical Charter of the Swiss Ethics Council for Official Statistics specifies that "Statistical information is documented in order to facilitate its understanding and allow its correct use"[3]. Is it compatible with statistical learning methods? Is it sufficient to specify the method used?

We believe that *statistical learning* methods can be used, but with certain precautions. The generalized regression estimator (see Särndal et al., 1992) allows a prediction to be incorporated into the estimate while remaining approximately unbiased under the sampling design. This estimator thus avoids slippage due to poor specification of the

[3] Translated from French: "Les informations statistiques sont documentées afin d'en faciliter leur compréhension et leur utilisation correcte."

model. So, official statisticians use fairly simple models that can be integrated in a method that remains valid under the design.

Official statisticians have also long been familiar with light non-parametric methods for dealing with non-response, such as imputation by the nearest neighbor or by an individual selected from a homogeneous stratum. Methods like *support vector machine* or *random forests* ultimately consist in splitting the space of explanatory variables to define one or more neighborhoods of the unit in order to make a forecast.

The generalized regression estimator has enabled Breidt & Opsomer (2000), for example, to construct estimates assisted by a model whose predictions are made by the local polynomial method. By following this same approach, all forecasting methods can be used without introducing disproportionate risks into the estimates. Thus, the recent work of Dagdoug et al. (2020a,b) which integrates the methods of *statistical learning* in the treatment of non-response, probably shows the way forward on future research in official statistics.

The so-called *shinkage* methods like the Lasso make it possible to choose variables in a modeling or a calibration. Survey statisticians often tend to over-calibrate surveys because more and more variables are available in the registers. However, calibration is above all an estimation technique which aims to reduce the variances of the estimates. In general, both calibration and modeling must be subject to a principle of parsimony. It consists of finding the most efficient model as possible while being as simple as possible. The Lasso method makes it possible to reduce the number of calibration variables. It has already been applied to survey data by McConville et al. (2017).

10 Conclusions

The multiplication of sources can be misleading because it does not necessarily imply an improvement in quality (see among others Deville, 1997). As Tim Harford (2014) reminds us, abundance of data is not and never will be synonymous with quality. Administrative records often contain a large number of errors because they were not designed to be used for statistical purposes. These new sources must be combined with all other available sources. The integration of sources is an issue that researchers have tackled from the very beginning of official statistics. These methods must be further developed, because more and more sources will be available. It would be useful to have a general theoretical framework for data integration.

The new statistical methods must also be integrated into existing methods or rather combined with existing methods. We think that there will certainly not be a clean slate of the methodology but an evolution towards more varied practices and more targeted to the types of data to be processed. Official statistics must more than ever promote and develop methodological research around its challenges.

References

Bastin, C., Benzécri, J. P., Bougarit, C. & Cazès, P. (1980). *Pratique de l'Analyse des Données*. Paris: Dunod.

Bellhouse, D. R. (1988). A brief history of random sampling methods. In *Handbook of Statistics Volume 6: Sampling*, P. R. Krishnaiah & C. R. Rao, eds. New York, Amsterdam: Elsevier/North-Holland, pp. 1–14.

Benzécri, J.-P. (1973a). *L'analyse des données : tome 1 : La taxinomie*. L'analyse des données. Paris: Bordas.

Benzécri, J.-P. (1973b). *L'analyse des données : tome 2 : L'analyse des correspondances*. L'analyse des données. Paris: Bordas.

Bethlehem, J. et al. (2009). The rise of survey sampling. The Hague, Statistics Netherlands.

Boistard, H., Chauvet, G. & Haziza, D. (2016). Doubly robust inference for the distribution function in the presence of missing survey data. *Scandinavian Journal of Statistics* **43**, 683–699.

Box, G. E. P. & Draper, N. R. (2007). *Response Surfaces, Mixtures, and Ridge Analyses*, vol. 649. Hoboken: John Wiley & Sons.

Breidt, F. J. & Opsomer, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *The Annals of Statistics* **28**, 1026–1053.

Brewer, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics* **5**, 5–13.

Brewer, K. R. W. (2013). Three controversies in the history of survey sampling. *Survey Methodology* **39**, 249–262.

Capel, R., Monod, D. & Müller, J.-P. (1996). Essai sur le rôle des tests d'hypothèse en sciences humaines. *Actualités Pédagogiques* **1**, 1–51.

Chambers, R. L. & Clark, R. G. (2012). *An Introduction to Model-Based Survey Sampling with Applications*. Oxford: Oxford University Press.

Chen, S. & Haziza, D. (2017). Multiply robust imputation procedures for the treatment of item nonresponse in surveys. *Biometrika* **104**, 439–453.

Conseil d'éthique de la statistique (2012). Charte de la statistique publique de la Suisse. Office fédéral de la statistique, Neuchâtel.

Costantini, D. & Galavotti, M. C. (1986). Induction and deduction in statistical analysis. *Erkenntnis* **24**, 73–94.

Dagdoug, M., Goga, C. & Haziza, D. (2020a). Imputation procedures in surveys using nonparametric and machine learning methods: an empirical comparison. ArXiv 2007.06298.

Dagdoug, M., Goga, C. & Haziza, D. (2020b). Model-assisted estimation through random forests in finite population sampling. ArXiv 2002.09736.

Deming, W. E. & Stephan, F. F. (1940). On a least square adjustment of sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* **11**, 427–444.

Devaud, D. & Tillé, Y. (2019a). Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *TEST* **4**, 1033–1065.

Devaud, D. & Tillé, Y. (2019b). Rejoinder on: Deville and Särndal's calibration: revisiting a 25 years old successful optimization problem. *TEST* **28**, 1087–1091.

Deville, J.-C. (1997). *Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?* Document de travail – Institut national de la statistique et des études économiques. Insee.

Deville, J.-C. & Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87**, 376–382.

Deville, J.-C. & Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893–912.

Fisher, R. A. (1935). The logic of inductive inference. *Journal of the royal statistical society* **98**, 39–82.

Fraser, D. A. S. & Reid, N. (2016). Crisis in science? or crisis in statistics! mixed messages in statistics with impact on science. *Journal of Statistical Research* **48**, 1–9.

Gelman, A. (2011). Induction and deduction in Bayesian data analysis. *Rationality, Markets and Morals* **2**, 1999.

Gelman, A. & Loken, E. (2014). The statistical crisis in science: data-dependent analysis–a "garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American scientist* **102**, 460–466.

Gini, C. & Galvani, L. (1929). Di una applicazione del metodo rappresentativo al censimento italiano della popolazione (1. dicembre 1921). *Annali di Statistica* **Series 6, 4**, 1–107.

Guandalini, A. & Tillé, Y. (2017). Design-based estimators calibrated on estimated totals from multiple surveys. *International Statistical Review* **85**, 250–269.

Han, P. & Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika* **100**, 417–430.

Hansen, M. H. (1987). Some history and reminiscences on survey sampling. *Statistical Science* **2**, 180–190.

Hansen, M. H. & Madow, W. G. (1974). Some important events in the historical development of sample survey. In *On the History of Statistics and Probability*, D. B. Owen, ed. New York: Marcel Dekker, pp. 75–102.

Harford, T. (2014). Big data: A big mistake? *Significance* **11**, 14–19.

Harman, G. & Kulkarni, S. (2012). *Reliable Reasoning: Induction and Statistical Learning Theory.* Cambridge, Massachusetts: MIT Press.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS MED* **2**, e124.

Jensen, A. (1926). Report on the representative method in statistics. *Bulletin of the International Statistical Institute* **22**, 359–380.

Kang, J. D., Schafer, J. L. et al. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science* **22**, 523–539.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review* **2**, 196–217.

Kiær, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique* **9**, 176–183.

Kiær, A. N. (1899). Sur les méthodes représentatives ou typologiques appliquées à la statistique. *Bulletin de l'Institut International de Statistique* **11**, 180–185.

Kiær, A. N. (1903). Sur les méthodes représentatives ou typologiques. *Bulletin de l'Institut International de Statistique* **13**, 66–78.

Kiær, A. N. (1905). Discours sans intitulé sur la méthode représentative. *Bulletin de l'Institut International de Statistique* **14**, 119–134.

Kim, J. K. & Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica* **24**, 375–394.

Kruskal, W. & Mosteller, F. (1980). Representative sampling, IV: The history of the concept in statistics, 1895-1939. *International Statistical Review* **48**, 169–195.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association* **88**, 1242–1249.

McConville, K. S., Breidt, F. J., Lee, T. C. & Moisen, G. G. (2017). Model-assisted survey regression estimation with the lasso. *Journal of Survey Statistics and Methodology* **5**, 131–158.

Molina, I. & Strzalkowska-Kominiak, E. (2020). Estimation of proportions in small areas: application to the labour force using the swiss census structural survey. *Journal of the Royal Statistical Society* **A183**, 281–310.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society* **97**, 558–606.

Neyman, J. (1938). Contribution to the theory of sampling human population. *Journal of the American Statistical Association* **33**, 101–116.

Neyman, J. (1952). *Lectures and Conferences on Mathematical Statistics and Probability.* Washington: Graduate School, U. S. Department of Agriculture.

Neyman, J. (1957). "Inductive Behavior" as a basic concept of philosophy of science. *Revue de l'Institut International de Statistique* , 7–22.

Popper, K. (2005). *The logic of scientific discovery.* London: Routledge.

Quételet, A. (1846). *Lettres à S. A. R. le Duc régnant de Saxe-Cobourg et Gotha sur la théorie des probabilités appliquées aux sciences morales et politiques.* Bruxelles: M. Hayez.

Rao, J. N. K. & Molina, I. (2015). *Small Area Estimation.* New York: Wiley.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika* **57**, 377–387.

Royall, R. M. (1971). Linear regression models in finite population sampling theory. In *Foundations of Statistical Inference*, V. P. Godambe & D. A. Sprott, eds. Toronto, Montréal: Holt, Rinehart et Winston, pp. 259–279.

Royall, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association* **71**, 657–664.

Särndal, C.-E. (2007). The calibration approach un survey theory and practice. *Survey Methodology* **33**, 99–119.

Särndal, C.-E., Swensson, B. & Wretman, J. H. (1992). *Model Assisted Survey Sampling.* New York: Springer.

Statistics Canada (2017). Statistics Canada's Quality Assurance Framework. Documentation of the internet site of Statistics Canada. Statistics Canada, Ottawa.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society* **B58**, 267–288.

Tibshirani, R. J. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society* **B73**, 273–282.

Tillé, Y. (2020). *Sampling and Estimation From Finite Populations.* Hoboken: Wiley.

Tukey, J. W. (1977). *Exploratory Data Analysis*, vol. 2. Reading, MA: Addison-Wesley.

Valliant, R., Dorfman, A. H. & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* New York: Wiley.

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician* **70**, 129–133.

Wikipedia contributors (2020). Martha Farnsworth Riche – Wikipedia, the free encyclopedia. [Online; accessed 24-August-2020].

Yang, S. & Kim, J. K. (2020). Statistical data integration in survey sampling: A review. ArXiv:2001.03259v1.