Ethics in Statistics

Opportunities and Challenges

Edited by

Hassan Doosti

Ethics in Statistics: Opportunities and Challenges

Edited by Hassan Doosti

This book first published 2024

Ethics International Press Ltd, UK

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

Print Book ISBN: 978-1-871891-65-2

eBook ISBN: 978-1-871891-66-9

Table of Contents

Appendices

Preface

Welcome to "Ethics in Statistics: Opportunities and Challenges." As the editor of this collective endeavor, I am thrilled to introduce a thought-provoking exploration into the ethical dimensions of statistics, featuring a diverse array of perspectives from esteemed experts in the field.

Setting the Tone:

Statistics, often regarded as the silent force shaping our understanding of the world, holds immense power and responsibility. In this collection, we delve into the ethical nuances that underpin statistical methodologies, decision-making processes, and their broader societal impact. From the ethical implications of artificial intelligence in statistics to the challenges of data governance and quality, each chapter contributes to a rich dialogue on how statisticians can navigate challenges while upholding integrity.

Expressing Gratitude:

My sincere gratitude extends to the dedicated contributors who have generously shared their expertise. Each chapter is a testament to their commitment to advancing ethical awareness in statistical practices. Together, we explore the intersections of ethics, social justice, and the evolving landscape where statistics meets technology.

Explaining the Motivation:

The motivation behind this compilation stems from the recognition that as statistics continues to evolve, so too do the ethical considerations surrounding it. From issues of privacy, transparency, and governance to the ethical implications of AI and data involving human variables, this book aims to provoke thoughtful reflection and provide practical guidance for statisticians, educators, and learners navigating complex ethical dilemmas.

Insights into the Editing Process:

The process of curating these chapters involved a captivating journey through a spectrum of ethical perspectives. We engaged in intellectual exploration, fostering collaborative dialogue to weave diverse insights into a cohesive narrative. The result is a book that not only examines ethical challenges but also sparks conversations and prompts critical reflections within the statistical landscape.

This book is more than a collection of chapters; it is a dynamic exploration of the ethical intricacies inherent in the field of statistics. May it inspire conversations, prompt critical reflections, and serve as a guiding compass for those navigating the complex intersection of statistics and ethics.

Hassan Doosti
Senior Lecturer in Statistics
School of Mathematical and Physical Sciences
Macquarie University, NSW 2109, Australia

Chapter 1
Ethics in Statistics: Opportunities and Challenges

Dr. Andrea Vicini*

Introduction

Nowadays, it appears evident even to the more distracted person that our era is characterised by a progressive and continuous digitalisation of many aspects of life, but it is equally evident that this digitalisation corresponds to a vast production of data in every aspect of life. This is collected from the morning when you put on your smartwatch and use your smartcard to access the metro until the evening when you go to the supermarket or restaurant and pay with your credit/debit card. In other words, it is possible to affirm that every aspect of our life is associated with data, information and measurements, consciously or not. Today, data are the basis for every complex decisional process, whether automatic or not, and from the estimation of risk in financial markets to diagnostic hypotheses in medicine and the acceptance or rejection of a pharmaceutical drug or a vaccine, data (big data, machine learning elaborations, and so on) play a central role. It is evident that the handling of these drivers, which could include the summarising of findings and the extending of their results to a population, as highlighted by the editor of this volume, need to be checked via an ethical perspective as well as a statistical one.

From a scientific point of view, the specific weight of the data for statistical analysis and the forecasting and functioning of many devices based on artificial intelligence is not part of the discussion. New scientific acquisitions in public health, medicine and diagnostics processes have frequently been supported and demonstrated by data analysis at a rate that

---

* Senior economist and financial manager with professional experience in reporting, quantitative data analysis and econometric modelling.

was unthinkable until two or three decades ago. However, as a direct consequence, the collection and elaboration of data generate complex ethical questions. This theme involves many aspects, from the security of data to their fair use.

In the last decade, adverse events have produced a big debate on a global level, which flowed into emergency legislation called the EU General Data Protection Regulation (commonly known by the acronym GDPR) (EU Regulation 2016/679).

The importance of the theme is confirmed by the high number of conferences and seminars held and publications realised in the last years about the topic.

In this chapter, we decided to focus our attention on the ethical implications for some macro-aspects of statistics that are apparently disjointed but are in fact intrinsically correlated. In particular, we will attempt to find adequate arguments for the following interrogatives:

- Privacy, profiling and personal data: where is the limit?
- Methods and transparency: why is adequate attention not paid to the methodology adopted in preliminary foresight studies?
- Governance of statistical institutions: what is the role of statistical institutions in society, and how must their relationship with the government be regulated in a democracy?

Privacy, profiling and personal data: where is the limit?

The first key question regarding personal data is synthesisable as: How is the data collected? From an ethical perspective, a person must be aware that a process of collecting personal data is underway during every simple daily operation, as mentioned above. The intention and the perception of the recipient is the key that makes the difference. Very often, digital tools and applications in particular, although they recall the decisive points of the GDPR legislation with every click on the web, do not directly inform the user about the use of their data and what the added value that is produced for them is.

It is useful to remember that on 13 June 2019, on the occasion of a stock-taking event to mark the first year of the application of GDPR, the European Commission published the results of a special Eurobarometer survey on data protection. On that occasion, Věra Jourová, Commissioner for Justice, Consumers and Gender Equality, declared:

> Helping Europeans regain control over their personal data is one of our biggest priorities. But, of the 60% Europeans who read their privacy statements, **only 13% read them fully**. This is because the statements are too long or too difficult to understand. I once again urge all online companies to provide privacy statements that are concise, transparent and easily understandable by all users. I also encourage all Europeans to use their data protection rights and to optimise their privacy settings. (European Commission, 2019)

Figure 1 below reports the results of the Eurobarometer survey carried out in 2019.

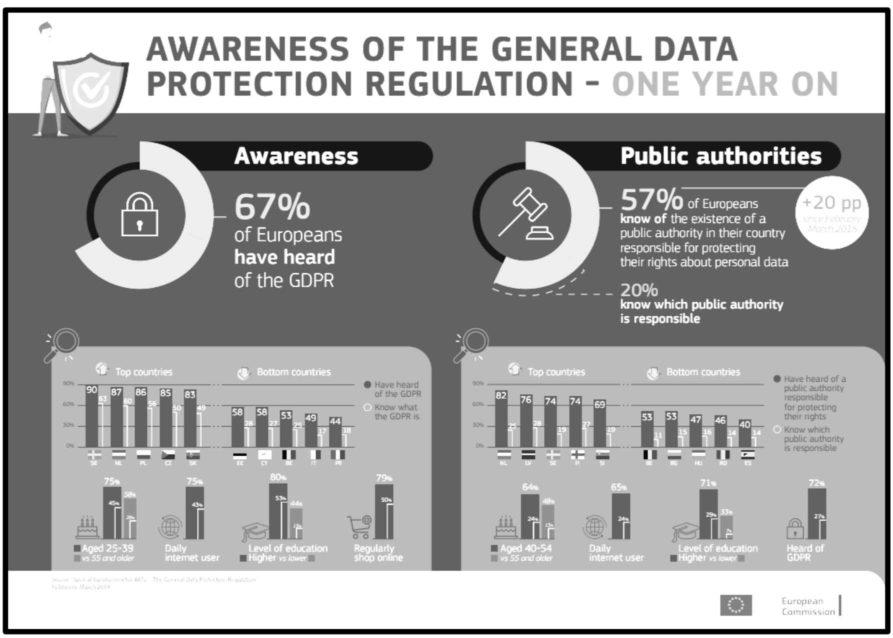


Figure 1: *Awareness of the General Data Protection and Regulation. Source: Special Eurobarometer 487a - "The General Data Protection Regulation" Fieldwork (March 2019).*

The infographic derived from the Eurobarometer survey shows the level of awareness of GDPR. This survey confirms that it is necessary to evolve the current legislation towards a pragmatic approach that is more oriented towards the ethical use of data. As in every context, transparency improves confidence. An organisation that collects personal data about an individual's lifestyle, consumption, habits, health and social conditions for profiling must provide them with rapid feedback about the benefits that are derived from this operation. This is an ethical theme, a sort of recognition.

Data are assets for a company, which can build profiles and statistical reports and patterns, whereas for a person, they are perceived as elements without a specific value. In this asymmetry, there is a vast space for businesses related to data within which they can work. To this end, it is a useful preface to state that the value of data is not intrinsic in itself, as with a common good, but rather depends on two factors.

The first is the user; for example, for an insurance agent, it is relevant to have information about a specific driver and his risk assessment, for a financial planner, the gold mine is to know about one's available income and the propensity to save, while for an insurance company dedicated to health service, the relevant information is the user's personal health condition.

The second factor is the structure of data. An ordered, structured and viable database with common software has more added value with respect to a set of unstructured data or which has been built without common criteria; this affects its comparability and the possibility to elaborate on it such as by building indexes and compiling statistical indicators, which is the real aim.

Thus, to reduce this gap, which is exacerbated by the asymmetry in the perception of value by the two sides, in accordance with an ethical vision, it should be possible to ask the organisation that collects the data to provide feedback/a certificate to the recipient that shows, in short, the added value produced by the data collected. In my view, this proposal would solve and reconcile the relationship between the individual and the organisation (firms, companies, public administration, ministerial departments). Another key question is whether the sale of personal data from one organisation to another can be permitted. Very often, organisations that collect personal data sell them to another organisation and so on, following the principle that the value of data as specified above depends on the user, so it becomes a saleable asset for a specific market. This aspect, although permitted by the legislation of many countries, is particularly insidious and does not respect ethical principles because, with every step, responsibility diminishes; moreover, the individual does not know the real aim of the secondary and subsequent collector(s) (generally an agency with unclear ends), and the data disappears into a black hole. To this end, it is necessary to maintain this practice in very restricted and limited cases. The proposal mentioned above (providing feedback/a certificate) could be a useful tool to mitigate this effect and help to build an ethical relationship in the use of data.

Naturally, the information contained in the certificate must be checked by a specific authority, with the goal of providing certification for the ethical use of data to the organisation that issues the feedback/certificate.

In conclusion, it is possible to propose a first set of principles and behaviours that are associable with an ethical use of data so that this can be adopted to formulate guidelines associable with potential certification.

First of all, an organisation must clearly inform the user about the collection of their data. It is necessary to stress that this aspect is for the most part provided for by the current legislation, which is similar in many countries because it covers a global space that has no borders, specifically the digital world. The organisation must provide feedback to the user about the use of their data and the added value produced for itself. The organisation must not forward the data to another organisation.

Why pay so much attention to the data? Because, to use a metaphor, data are the bricks in the building, which can be represented by a statistical report in a general sense. Moreover, it is opportune to highlight that the latest frontier of innovation, namely artificial intelligence, which takes up more and more space in our lives, is continuously fuelled by the data produced with every click. Thus, it is necessary and urgent to make new students on academic courses aware of the correct and ethical use of data because, often, a data set represents a person and an individual.

Methods and transparency: why is adequate attention not paid to the methodology adopted in preliminary foresight studies?

In a global world in which tastes, desires and lifestyles seem to be converging, such as with technological devices, the statistical methods that are part of the mathematics applied to real events seem destined to converge processes. With respect to the collection of data and statistical methodologies, there are no exceptions; a standardisation process is underway.

First of all, harmonisation in data collection is fundamental considering that there is always a potential error in the data collection stage, as mentioned below in Table 1. It is evident that absolute certainty with respect to the quality of data collected is not possible and the degree of certainty depends on the data's scope, but it is evident that stressing the importance of accuracy to reduce dirty data is a key question that must be emphasised and learned during academic statistics courses.

In an interesting article, a group of Korean researchers belonging to different academic institutions contributed to classifying so-called dirty data, generating a detailed taxonomy that considers dirty data's impact on the overall results (Kim et al., 2003). A broad analysis of dirty data is beyond the scope of this paper, but a short summary, starting from the most common causes of the data errors, is in order. For the interested reader, we suggest consulting the articles mentioned in the references in the paper by Kim et al. (2003).

In short, dirty data is simply data that contains errors. The following table (Table 1) classifies the most common typologies and situations in which the generation of dirty data occurs. For every situation, the remediations/ suggestions that can be adopted to reduce the generation of dirty data are correspondingly presented.

Situation of dirty data generation	Remediations/suggestions
Human error: This is one of the major problem areas for data care. Many of the processes require the manual input of data, and this can cause errors. The errors get worse if those inputting data have no technical knowledge. A lack of knowledge and the absence of ranges or characters' acceptance could be a problem because this information is passed on to the rest of the organisation for successive elaboration.	This can be eliminated with the automatisation of the process, which is already underway in many sectors; however, it is opportune to recall the fact that, although it should be reduced, humans' interface with machines (PCs, sensors, or other devices) remains crucial in the analysis stage, and an error in a similar step could be amplified during the successive steps.
Incomplete data: This is data that has been left blank. Without it, some functions could be compromised.	This can be eliminated with an automated process that issues an alert with respect to the blank.
Duplicate data: Duplicate records can produce different results in analytical terms.	This situation can be eliminated with the automatisation of the process ex ante.
Incorrect data: This kind of error occurs when, during input, the creation of values outside of a valid range happens. It could also be due to spelling variations, typos, formatting problems, and transpositions. Incorrect data result in the wrong interpretation of data.	This can be eliminated with the automatisation of the process and the inclusion of an accepted range for the data input manually.
Inaccurate data: Data can be correct but inaccurate. This can, in turn, cause unwanted interruptions in activity. For example, this can happen when customers/users fail to input some data when filling in online forms.	This can be reduced with the presence of automatic checks.
Inconsistent data: This happens when the same data values are stored in different locations, causing	This is reducible if adequate attention is paid to the compatibility between IT software, hardware, and the

Situation of dirty data generation	Remediations/suggestions
redundancies and inconsistencies. For example, an organisation may have the same company information stored in different, non-synced systems or apps.	synchronisation phase.
Storage failure: Hardware failure can also cause data corruption or loss. Such loss and corruption can lead to disruption and, ultimately, the loss of part of the database, and the loss of data leads to failure in the successive steps of the data's elaboration.	This needs continuous maintenance, considering that the obsolescence period for hardware and software solutions is decreasing, and the vulnerability of complex IT systems is increasing.

Table 1: *Dirty data typologies and remediation suggestions. Source: Author's elaboration.*

Paradoxically, dirty data also occurs during the elaboration or the "cleaning step" and commonly when organisations are linking data across sets. If there is no unique identifier for the data, linking them creates problems. Sometimes data are combined incorrectly, whereby data belonging to two different records with the same name are mixed. This kind of data problem can occur from data migration if the company is using old technology or when multiple databases (referred to different times) are merged with the intention to combine data. Furthermore, the same problems occur during the condensing of data into manageable forms. Other situations can include minor problems such as duplicate data, missing fields, invalid figures or ranges, or failed logical code verification (Haughton et al., 2003).

Why is so much importance attributed to these aspects? Because it is logical that if you have dirty data in your input, you obtain dirty results in your output, and it is not possible to operate an adjustment on the results ex-post (Haughton et al., 2003). It is appropriate in statistics courses to pay attention to the importance to have an accurate database for subsequent elaborations. Adequate attention in the data collection phase reduces this risk, although estimating the dimension of the potential error remains a fundamental driver for attributing the correct dimension of the risk (Kim et al., 2003). One hundred percent data accuracy is not required for all projects; in fact, it might not even be possible.

Therefore, it is advisable that the statistician/researcher be able to reply to the following question: What is the effect of the error rate on the results produced? Simulations can increase the knowledge of the risk impact and reduce uncertainty (Haughton et al., 2003).

Accordingly, the initiatives proposed by international institutions (OSCE, Eurostat, World Bank, IMF) are useful for arriving at a harmonisation of the data collected to reduce the risk of dirty data, improve accuracy, and ease comparisons among countries. This aspect must be adequately evidenced and learned in statistics courses because the entire team, from the simple human detector to the researcher, that works on a statistics or quantitative project must be oriented towards reaching the same objective, which in general terms must be synthesised as an improvement to data accuracy, which, ultimately, will increase the quality of every decision, and thus allow humans to progress.

Secondly, another big issue that can be the cause of biased results in statistics works or an inferior manipulation of the results obtained is the methodology adopted. Here the issue assumes a smoother profile.

The recent revolution in statistics and data science includes the now well-known phenomenon of "big data analysis" (high speed, high volatility, high uncertainty). This phenomenon has raised many questions, one of which is founded on the inferential methods that, in turn, are based on different and alternative probability definitions: objectivist vs. subjectivist.

This difference involves the quintessence of the methods in which statistical inference is conducted; for this reason, the debate was enlarged to experts from many areas, not least from philosophy. The statistical literature on this debate is endless (James et al., 2021). It is interesting to note that the discussion and the concerns of statisticians are about the methodology being crucial and investing the interest of "professional users" of statistical tools, who are continually increasing in number. A general idea of the vivacity of this debate is deducible from the discussions held in 2005 at the conference "Statistical Problems in Particle Physics, Astrophysics and

Cosmology", organised by Prof. Lousie Lyons, a famous physician from Oxford University, who also co-edited the proceedings (Lyons & Ünel, 2006).

After this introduction to the debate, it is useful to outline the main differences between the objectivist and subjectivist approaches and their influence on statistical inference, without losing sight of the main question that concerns us, that is, the ethical aspect.

These approaches are the two macro-categories of interpretations of probability, whose adherents hold different visions about the fundamental nature of probability:

- Objectivists assign numbers to describe some objective or physical state of affairs. The most noteworthy concept of objective probability is frequentist probability, which asserts that the probability of a random event denotes the relative frequency of occurrence of an experiment's outcome when the experiment is repeated indefinitely.
- Subjectivists assign numbers per subjective probability, that is, as a degree of belief. The degree of belief has been interpreted as "the price at which you would buy or sell a bet that pays 1 unit of utility if E, 0 if not E", although this interpretation is not universally agreed upon (de Finetti, 2006). The most popular version of subjective probability is Bayesian probability, which includes expert knowledge as well as experimental data to produce probabilities. Expert knowledge is represented by some subjective prior probability distribution. This information is incorporated in a likelihood function. The product of the prior and the likelihood results in a posterior probability distribution that incorporates all the information known to date.

The subjectivist and, in particular, the Bayesian approach, which was conceived in the XVIII century by Thomas Bayes, are currently universally applied within big data analysis and software applications (Agresti & Franklin, 2007).

For this reason, the statistical investigation approach has changed over the last two decades. This is because, in particular in economic and financial investigations, we have passed from an approach based on a theoretical hypothesis at the first stage (which was the golden rule in time series and econometrics studies) to an approach based on data mining (atheoretical), in which the modelisation is deducted from the behaviour of data filtered by machine learning algorithms, which is only apparently objective.

The availability of "big data" pushed the application of statistics tools towards the direction indicated above, but with a high number of shortcomings. Although it is widely used every day to formulate forecasts in financial markets, betting odds, or infographics applied to a great number of fields, this approach, which seems neutral at first sight, does not consider the role played by the theoretical knowledge of the topics or object(s) of statistical analysis with the right attention. In other words, there are multiple risks of bias, including the inappropriate use of the variables available that require a specific focus.

The vast diffusion of this approach has reduced the importance of the correct application of inference theory, which seems to be reduced to a sub-discipline of software development and computational analysis. In a similar vein, common sense suggests the quality of the statistical results (forecasts) is dependent on the statistical tool adopted. Although the computational part is fundamental in statistics and enormous benefits have been produced thanks to the exponential development of the software industry, we are convinced that knowledge of the sector and topic analysed remain an irreplaceable pillar for the ethical use of statistics. This aspect must be adequately stressed to offer students a realistic vision of the potential and limits of statistics, rather than suggesting it as a magical method.

Following our previous discussion, a provocation: Would you trust yourself to a medical diagnosis carried out by a physician whose background was focused on the use of excellent diagnostics software?

Even if data mining as an approach is powerful, easy in terms of time and attractive, deep knowledge of the topic examined remains an irreplaceable part of ethical statistical analysis. In my view, presentations of this kind of approach (data mining on big data) must focus their attention within the right limits and not be compared with serious statistical analysis because, in big data works (which are generally destined for a wide public audience), the results obtained could be oriented in many directions and used to influence opinions, without ethical respect for the truth and knowledge. In the long run, a similar use of statistics would impact the reputation of statistical analysis itself and open the door to further scientific relativism, which is already a vast plague today.

In this regard, the origin of the famous expression "[There are] lies, damned lies, and statistics", although vaguely attributed to the former British prime minister Benjamin Disraeli, remains unclear. However, what is clearer is the meaning that describes the persuasive power of numbers, and particularly the use of statistics to reinforce weak argumentations (Velleman, 2008).

In conclusion, our best hope for maintaining an adequate scientific level of statistics is education. If we teach statistics as a mechanistic disorder of magical methods, our students and the "users" will conclude for themselves that statistics are a pack of damn lies, useful only to influence public opinions. However, it is possible to do better:

- We must tell our students that they must make judgments to use statistics and that there may be no guaranteed method to arrive at the truth. This means promoting an open discussion with them and rejecting approaches based on mechanistic procedures.
- We should advise students to know the motivating reason for the analysis because this will guide them in making these judgments. They should know who (or what) the cases are in the data, what has been measured or recorded about them, and when that was done. Even definitions that seems obvious should be questioned.
- We should teach that the driver in statistics judgments is a search for truth about the world. Thus, the choices adopted in terms of patterns must be inspired to support our efforts to model or understand the world as it is. Where that choice is not clear, we must make an honest attempt (approximation) to make the best

choice. It is fine to maintain alternative or contradictory models until there are no data that allow us to choose them.

- We should teach students and analysts to resist jumping to rapid conclusions, based on arbitrary extrapolations, that assume causation without the right assumptions. We should teach them to be sceptical of statistical reports that do not meet high standards in inference and extrapolating analysis (Velleman, 2008).
- Last but not least, we must present statistics as a search for understanding about the world and the decisional processes under conditions of uncertainty, thereby coming back to the original line of our argument (Velleman, 2008).

The diffusion of forecasts based on inference without a rigorous explanation or details about the hypotheses and methods adopted is a form of biased and not ethically correct communication. These aspects are important, although often not adequately emphasised. Confirmation bias and falsification bias are common when there is a deficiency in the topics or methodology adopted to explore a phenomenon.

Today, data are the basis for every complex decisional process, and from the estimation of risk in financial markets to diagnostic hypotheses in medicine and the acceptance or rejection of a pharmaceutical drug or a vaccine, data (big data, machine learning elaborations, and so on) play a central role. It is evident that the handling of these drivers, which could include the summarising of findings and the extending of their results to a population, as highlighted by the editor of this volume, need to be checked via an ethical perspective as well as a statistical one.

Governance of statistical institutions: what is the role of statistical institutions in society, and how must their relationship with the government be regulated in a democracy?

The birth of statistics as an autonomous discipline was in the XVII century with the Enlightenment (Harari, 2014). In its origin, the scope of this new discipline was to rationally address the public decisions of the government in uncertain conditions and so, in short, became a source of significant support for every decision-maker (Porter, 2020). During the XX century, the majority of countries established an official statistical institution dedicated specifically to collecting data and elaborating complex reports about socio-economic and demographic facts and other features of relevance to the nation (Porter, 1993).

In this regard, it is important to remember that before the declaration of the Greek default in 2009, the country was not equipped with an independent statistical authority dedicated to monitoring the dynamic of the macro variables, in particular the public debt. The consequences of that historical event were known to all: there was the real risk of an implosion in the Eurozone, the tensions inside the Eurozone were discharged into the financial markets, and this situation sparked a big macroeconomic debate which is still ongoing. However, it is useful to remember that this aspect contributed to an unprecedented global crisis of confidence. In fact, it is useful to recall that one of the conditions that the European Union (EU) imposed upon Greece at the beginning of 2010 during the early days of the Greek debt crisis was a significant reform of its statistical office, which belonged to the Ministry of National Economy before its reform.

Statistical institutions play a significant role in citizens' perception of economic variables such as inflation and unemployment, just to cite the most commonly diffused data, as well as demographic data. The role of a statistical institution for a country is similar to a dashboard of a vehicle: to provide state-of-the-art information and, in many circumstances, to enable not only policymakers but also the establishment and civil society to track a path for future scenarios. In a democratic state, statistical reporting activity must be inspired by the picture of the reality; consequently, the picture could contain lights and shadows. However, it turns out that the crucial question is this: who is the individual or institution that can nominate the president of the statistical institution and, consequently, what comprises adequate governance for maintaining its necessary ethics and independence?

This is a problem of corporate governance but which also influences the quintessence of these institutions, which must have the possibility to express uncomfortable judgments to the government. As aptly expressed in a short sentence by Corrado Gini, an influential Italian statistician of the XX century,

> The scope of science is not to reach appreciated conclusions but founded conclusions. (Gini, 1939)

From another point of view, we can say that a government that wants to preserve the prestige and credibility of the statistical institution (and ultimately itself) must be conscious that the institution and its board of directors must be free to express reports with pictures and predictions that could generate antipathy for the government's actions and the policy it has adopted. This is the essence of the democratic system.

To this end, it is evident that the institution have independence with respect to the political power and, in particular, its government must be clearly written in its constitution or in basic legislation. Moreover, the prerequisite of independence can be reached only with a president and a board of directors recruited for their high technical capability and a personality that expresses the right distance from the political powers. This is to say that the statistical institution's board must be able to make reports and express judgments without suffering pressure from the external or internal environment, and here the role played by the rules is crucial.

The case of Andreas Georgiou, President of the Greek Statistical Authority after its reform in 2010, is emblematic in this respect. His defence in terms of his ethical behaviour is that he arrived from the International Statistical Institute, a non-governmental organisation that promotes the understanding, development and good practice of statistics worldwide. The declaration of his commendation follows:

> On 18 September 2018, during a Special Meeting on *National Statistical Offices' Professional Independence: Threats and Responses* immediately prior to the XVI IAOS Conference in Paris, a special Commendation was awarded to Andreas Georgiou, former and inaugural President of ELSTAT (the Greek Statistical Authority).

> This Commendation was given to acknowledge Andreas Georgiou's upholding of the highest professional standards in his public service in the pursuit of integrity of statistical systems.
>
> During his Presidency at ELSTAT, Andreas Georgiou committed himself to ensuring that the production of all official statistics in Greece should be undertaken in strict conformity with international and European statistical principles and standards. In particular, he insisted that all statistics, and the processes that underpinned their production and dissemination, should conform to the UN Fundamental Principles of Official Statistics and the European Statistics Code of Practice. (ISI, 2018)

Andreas Georgiou was recruited to head the reform of the national statistical office, the Greek Statistical Authority (ELSTAT), with the mission to modernise the production of official statistics, fully apply the EU rules for the production of these statistics, and ensure ELSTAT's independence and overall implementation of statistical ethics.

In conclusion, this shows that the independence, and thus the prestige, of the institution must be founded on a few irrevocable pillars. The first is the independence of the statistical institute that is realised with its segregation from governmental institutions, established with constitutional or similar legislation. The second is that it has a board of directors that is highly professional and with the right distance from political pressures. The third is linked to the adoption of standards and principles conforming to the UN Fundamental Principles of Official Statistics, which are drivers for appreciable results.

References

Agresti, A., & Franklin, C. (2007). *Statistics: The Art and Science of Learning from Data.* Upper Saddle River: Pearson Prentice Hall.

De Finetti, B. (2006). *L'invenzione della verità*. Milan: Raffaello Cortina Editore.

EUR-Lex. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).* Available at: http://data.europa.eu/eli/reg/2016/679/oj.

European Commission. (2019). *Data Protection Regulation one year on: 73% of Europeans have heard of at least one of their rights.* Available at: https://ec.europa.eu/commission/presscorner/detail/en/IP_19_2956.

Gini, C. (1939). "I pericoli della Statistica", in *Relazione inaugurale della prima riunione scientifica della Società Italiana di Statistica*, Pisa, 9 October 1939.

Harari, Y. N. (2014). *Sapiens: A Brief History of Humankind*. London: Harvill Secker.

Haughton, D. Robbert, M. A., Senne, L. P., & Vismay, G. (2003). "Effect of Dirty Data on Analysis Results." Paper given at Eighth International Conference on Information Quality (IQ 2003), November 7-9. 2003.

ISI. (2018). *Special Meeting on National Statistical Offices' Professional Independence: Threats and Responses.* Available at: https://www.isi-web.org/files/docs/statements-and-letters/press-release-commendation-for-andreas-georgiou-aug-2018.pdf. Accessed 27 December 2018.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). "Introduction", in James, G., Witten, D., Hastie, T., & Tibshirani, R. (eds.) *An Introduction to Statistical Learning.* Springer Texts in Statistics. New York: Springer, pp. 1-14.

Kim, W., Choi, B.-J., Hong, E.-K., Kim, S.-K., & Lee, D. (2003). "A Taxonomy of Dirty Data", *Data Mining Knowledge Discovery*, 7, pp. 81-99.

Lyons, L. & Ünel M. K. (2006). *Statistical Problems in Particle Physics, Astrophysics and Cosmology: Proceedings of PHYSTAT05, Oxford, UK, 12 – 15 September 2005.* London: Imperial College Press.

Porter, T. M. (1993). *Le origini del moderno pensiero statistico (1820-1900).* Florence: La Nuova Italia.

Porter, T. M. (2020). *Probability and statistics* Available at: https://www.britannica.com/science/probability.

Velleman, P. (2008). "Truth, Damn Truth, and Statistics", *Journal of Statistics Education*, 16(2). doi: 10.1080/10691898.2008.11889565.

Chapter 2
Learning Outcomes Supporting the Integration of Ethical Reasoning into Quantitative Courses: Three Tasks for Use in Three General Contexts

Rochelle E. Tractenberg*

Abstract

This paper gives a brief overview of cognitive and educational sciences' perspectives on learning outcomes (LOs) to facilitate the integration of LOs specific to ethical reasoning into any mathematics or quantitative course. The target is undergraduate (adult) learners but these LOs can be adapted for earlier and later stages of learning. Core contents of Ethical Reasoning are: 1. its six constituent knowledge, skills, and abilities; 2. a stakeholder analysis; and 3. ethical practice standards or guidelines. These are briefly summarized. Five LOs are articulated at each of three levels of cognitive complexity (low/med/high), and a set of assignment features that can be adapted repeatedly over a term are given supporting these LOs. These features can support authentic development of the knowledge, skills, and abilities that are the target of ethical reasoning instruction in math and quantitative courses at the tertiary level. Three contexts by which these can be integrated are *Assumption* (what if the assumption fails?), *Approximation* (what if the approximation does not hold?), and *Application* (is the application appropriate? what if it is not?). One or more of the three core contents of Ethical Reasoning can be added to any problem already utilized in a course (or new ones) by asking learners to apply the core to the context. Engagement with ethical reasoning can prepare students to assume their

---

* Georgetown University, Washington, DC, USA
This synthesis was funded by a National Science Foundation Collaborative Grant (ER2 Standard Grant Award # 2220314)

responsibilities to promote and perpetuate the integrity of their profession across their careers using mathematics, statistics, data science, and other quantitative methods and technologies.

Getting "ethics" into quantitative courses: neither simple nor straightforward

Ethics is defined as "the principles of conduct governing an individual or a group (e.g., professional ethics)"[1]. "Ethics" is recommended content for statistics and data science curricula in higher education (e.g., American Statistical Association Undergraduate Guidelines Workgroup 2014; DeVeaux et al. 2017; National Academies 2018; Association of Computing Machinery Data Science Task Force, 2021), although notably not in undergraduate curricula in "business analytics" (Wilder & Ozgur, 2015) or mathematics (Saxe & Braddy, 2015), and also not in the European Union-focused EDISON Data Science Framework (Demchenko, Belloum & Wiktorski, 2017). One reason why some disciplinary curricular guidelines do not specify that "ethics" should be integrated might be that it is a very vague term. Even defined as professional guidelines, there may be more than one set that is relevant for an individual's practice (e.g., for federal statistics in the United States, a practitioner working with data might be guided by the *Data Ethics Tenets* (Office of Management and Budget, 2020), applicable to every federal employee working with data, or by the *Principles and Practices for Federal Statistical Agencies* (National Academies of Science, 2021) which are specifically utilized by designated agencies[2]. There are no alternative federal guidelines for mathematical practice in the United States.

---

[1] https://www.merriam-webster.com/dictionary/ethics

[2] The Principles and Practices apply specifically to 13 U.S. principal federal statistical agencies: Bureau of Economic Analysis (Department of Commerce); Bureau of Justice Statistics (Department of Justice); Bureau of Labor Statistics (Department of Labor); Bureau of Transportation Statistics (Department of Transportation); Census Bureau (Department of Commerce); Economic Research Service (Department of Agriculture); Energy Information Agency (Department of Energy); National Agricultural Statistics Service (Department of Agriculture); National Center for Education Statistics (Department of Education); National Center for Health Statistics (Department of Health and Human Services); National Center for Science and Engineering Statistics (National Science Foundation); Office of Research, Evaluation, and Statistics (Social Security Administration); and Statistics of Income (Department of Treasury). There are

In the context of data science statistical science, we can consider ethical practice standards that derive from statistics (American Statistical Association (ASA), 2022), computing (Association of Computing Machinery (ACM), 2018), and the area of specialization in which the data/statistical scientist is applying their statistical and computational expertise and methodologies. Critically, both the ASA and ACM assert the applicability of their ethical practice standards for *all* who utilize their domain knowledge, skills, and technologies. Whenever an individual - irrespective of membership in these professional organizations, degree or training, or job title- uses statistical practices or computing, the ASA and ACM ethical practice standards are relevant. “Upon entry into practice, all professionals assume at least a tacit responsibility for the quality and integrity of their own work and that of colleagues. They also take on a responsibility to the larger public for the standards of practice associated with the profession.” (Golde & Walker, 2006: p. 10) This responsibility for the quality and integrity of data and statistical sciences work is a responsibility to follow the professional ethics articulated for that type of work by these professional organizations. There is no single code of ethics for mathematical practices (but see American Mathematical Society, 2019 for a recently-updated version of their Code of Ethics; and Tractenberg et al., 2024 for *Mathematical Ethical Proto-Guidelines*).

It could be argued by instructors in any of the individual courses that make up undergraduate curricula in math, statistics, data science, and computing that theirs is not the most appropriate course into which "ethics" - especially not "ethical practice standards" - should be integrated. The argument might be that, since it is unclear if any of the students will in fact go on to be practitioners in the field, the professional practice standards are not yet relevant. Recognizing the potential for these objections, this paper hopes to empower instructors to overcome them. The solutions recognize that (and how) "ethics" is very difficult to teach -and even harder to assess in the typical undergraduate quantitative course. Quantitative-heavy courses do

also three recognized federal statistical units: Microeconomic Surveys Unit (Federal Reserve Board); Center for Behavioral Health Statistics and Quality (Substance Abuse and Mental Health Services Administration; Department of Health and Human Services); and National Animal Health Monitoring System (Animal and Plant Health Inspection Service, Department of Agriculture).

not facilitate engagement with "ethics" content. Moreover, instructors with quantitatively oriented materials would have to make room in the syllabus and in the course schedule to teach, and assess, engagement with ethics content. Within departments, each instructor might choose different teaching methods (e.g., discussion or case analysis), contents (e.g., case studies or online programs), and grading (e.g., rubrics; quizzes or tests). Thus, there are many problems that any instructor might encounter when considering the integration of ethics into a quantitative course or curriculum. These problems appear in Table 1. Solutions are also given in Table 1, and are elaborated throughout this paper.

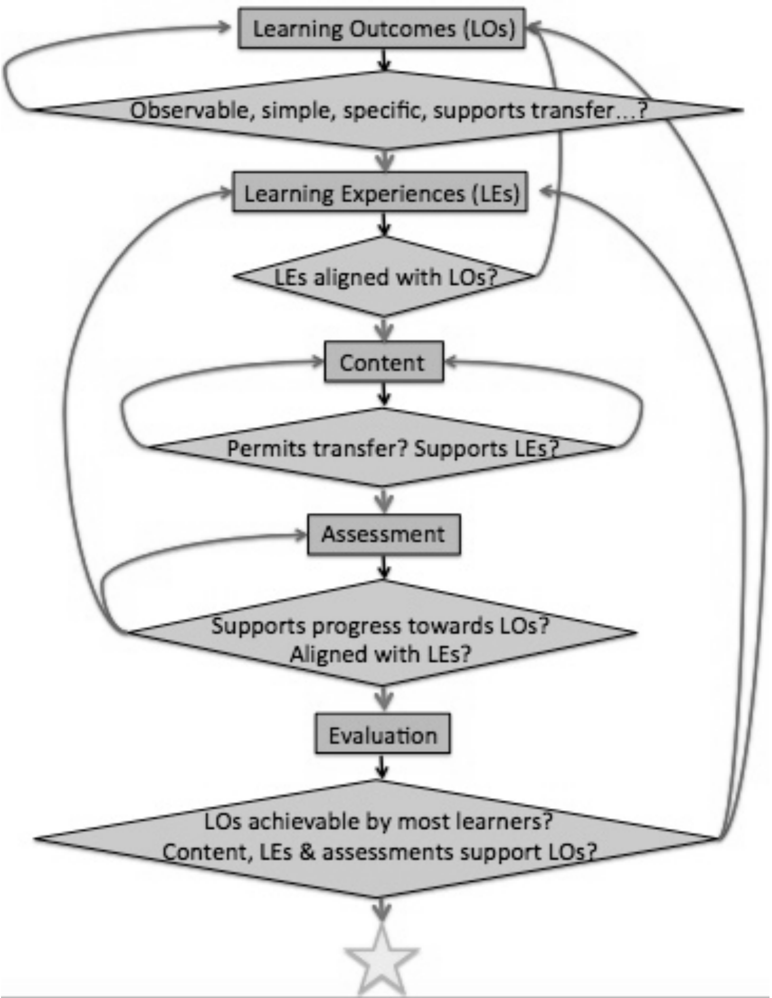
PROBLEM	**SOLUTION(s)**
"Ethics" is recommended content for statistics and data science curricula in higher education (National Academies; ASA; NIH; NSF, etc.)	Teach ethical reasoning, not "ethics". Leverage – and teach – Ethical Practice Standards in codes and guidelines, not a topics list. Formulate & share specific learning outcomes.
"Ethics" is hard to teach! Harder to assess.	Formulate & share specific learning outcomes that promote ethical reasoning, not "ethics". Use stakeholder analysis, not "discussion". Leverage – and teach – Ethical Practice Standards, not a topics list.
Quantitative-heavy courses do not facilitate "ethics" content.	Use stakeholder analysis, not "discussion" based on homework problems and proofs. Leverage – and teach – Ethical Practice Standards, not a topics list.
Quantitative materials, and instructors, have to make room to teach, and assess, engagement with ethics content.	Leverage – and teach – Ethical Practice Standards, not a topics list. Maximize the case analysis/case study teaching method – to facilitate both teaching and assessing.
Each instructor, in different contexts, might choose different methods, contents, student work, grading. Creating consistency within/across programs is difficult.	Leverage – and teach – Ethical Practice Standards, not a topics list. Use stakeholder analysis, not "discussion". Teach ethical reasoning, not "ethics".

Table 1. *Problems with "integrating ethics" into quantitative courses, with solutions.*

In order to support the solutions listed in Table 1, some background is needed to contextualize the suggestions.

Background: Education Sciences

Figure 1. *Five phases of curriculum and instructional design (from Tractenberg et al. 2020, with permission; after Nichols, 2002).*



What is a learning outcome (LO)?

"Learning outcomes are statements of the knowledge, skills and abilities individual students should possess and can demonstrate upon completion of a learning experience or sequence of learning experiences." (Stanford University, n. d.)

https://web.stanford.edu/dept/pres-provost/irds/assessment/downloads/CLO.pdf.

LOs are a starting point for both instructional design (e.g., Nicholls, 2002; Diamond 2008; Nilson 2016; Tractenberg et al. 2020) and considerations of the integration of new content or ideas into existing curricula or courses.

Role of LOs in curriculum and instructional design

Five phases of curriculum and instructional development (Nicholls, 2002) are widely recognized:

1. Identify aims and learning outcomes (LOs);
2. Identify/create learning experiences (LEs, lectures, exercises, readings, etc.) that will help students achieve the aims and outcomes;
3. Select content that is relevant to outcomes;
4. Identify/develop assessments to ensure learner is progressing towards outcomes;
5. Evaluate the effectiveness of the learning experiences for leading/developing learners to the outcomes.

This figure, showing the five phases of curriculum and instructional design (in blue boxes), also shows how these phases are iteratively informed by LOs. In fact, LOs drive all decisions in curriculum and in course development (Tractenberg et al. 2020; Nilson 2016).

Nilson (2016) noted, "your student learning outcomes provide the foundation for every aspect of your course, and you should align all the other components with them." (p. 129). Note that even course content is derived indirectly from LOs as a function of the learning experiences (teaching and classroom activities). Figure 1 shows how, at each phase, support for LOs is key in decision making about all other aspects of instruction (including assessment of learning and evaluation of teaching).