STATISTICAL THINKING FOR THE ERA OF BIG DATA AND ARTIFICIAL INTELLIGENCE: TOWARD UNDERSTANDING SUSTAINABILITY TRENDS AND ISSUES FOR THE FUTURE SOCIETY

Orlando González, Masami Isoda and Roberto Araya
Assumption University, Thailand
University of Tsukuba, Japan
University of Chile, Chile
ogonzalez@au.edu

This paper aims to characterize the thinking processes needed for the practice of statistics with big data and data analytics platforms driven by artificial intelligence. Such thinking has evolved from a traditional question-then-answer analysis to a more creative approach, which starts with data-first answers from examining opportunistic data, and then works backward to find the questions that should have been asked. Concerned with the rare use of the latter approach at school level, the authors developed a six-phase framework for the statistical thinking needed for the practice of statistics in a technologically- and data-rich world, in the context of the APEC-funded project "Inclusive Mathematics for Sustainability in a Digital Economy" (InMside). An exemplar related to a sustainability issue is provided to briefly illustrate the framework implementation.

INTRODUCTION

The importance and presence of big data in today's society is undeniable (Glaser, 2018; Storey & Song, 2017). Commercial institutions, government agencies and research calls from a diversity of fields, all seem to be interested in capturing and analyzing big data in increasingly powerful ways. Big data—i.e., "datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al., 2011, p.1)—is often associated with five characteristics, known as the 5 V's: volume, velocity, variety, veracity and value (Storey & Song, 2017).

In order to make sense of big data, it is necessary to explore large, complex and seemingly unrelated sets of raw data, looking for significant correlations and new and unanticipated connections among them, using artificial intelligence (AI)-driven data analytics platforms, powerful enough to handle data sets with sizes in the order of terabytes, petabytes and zettabytes (Claverie-Berge, 2012; Glaser, 2018). The way in which big data will be converted into a new type of knowledge is through machine learning, which is an automated process that extracts patterns from data using AI algorithms (Kelleher, Mac Namee & D'Arcy, 2015). This approach to data analysis—the big data analytics—is becoming more and more common among ordinary people, who use AI-driven services (e.g., Google's Siri) to make sense of big data and then make decisions. Big data analytics is defined as the iterative and exploratory process of using advanced technologies and analytical techniques against big data to reveal critical information, such as hidden and/or meaningful data patterns, trends and associations (Claverie-Berge, 2012; Cumming et al., 2017).

STATISTICAL THINKING IN TRADITIONAL AND BIG DATA ANALYTICS

Many researchers and statistics educators consider statistical thinking as the practice of statistics through the enactment of the different thought processes involved in statistical problem solving and statistical investigations. In fact, many frameworks of statistical thinking as the practice of statistics have been integrated in the mathematics curricula of different countries, being the most known ones those developed by Wild and Pfannkuch (1999), Franklin et al. (2007), and Watson (2016) [see Figure 1]. These frameworks are question-then-answer research methods, focused on data gathered systematically for a purpose, using planned processes, and chosen on statistical grounds (e.g., random sampling) to justify certain types of inferences and conclusions. For example, the practice of statistics under question-then-answer frameworks of statistical thinking usually starts with students posing, or being given, a question that digs into data searching for specific metrics, such as "Do brown-eyed Grade 6 students have faster reaction times than other Grade 6 students?" (Watson & English, 2018) or "Are Australian swimming teams improving over time? If so, is Australia likely to win gold in the pool at the 2016 Olympics?" (English & Watson, 2018). This kind of question requires

In: P Arnold (Ed.), *New Skills in the Changing World of Statistics Education*
Proceedings of the Roundtable conference of the International Association for Statistical Education (IASE),
July 2020, Online. Voorborg, The Netherlands. ©2020 ISI/IASE iase-web.org/Conference_Proceedings.php

from students to collect data from/by themselves as the next step of the statistical investigation. This is usually done through the analysis and transformation of a particular quality (e.g., "reaction time" or "improvement over time") into a metric (i.e., taking, inventing and/or revising measures).

Big data is opportunistic, or happenstance, data, which means data already collected by others, not through personally-planned processes conducted by the user, by using AI-driven machine learning platforms. Thus, due to its nature, big data cannot be the object of traditional data handling approaches. This is one of the most relevant criticisms to current statistical thinking frameworks: exploiting opportunistic data requires statistical thinking processes that do not necessarily follow the line of thought sketched by models such as the PPDAC cycle (Wild et al., 2018). In order to start the practice of statistics with big data—i.e., engage in big data analytics— an AI-driven IT platform with a certain degree of access to big data is needed, to enable creative discovery by the users (Claverie-Berge, 2012; Ferguson, 2012). These IT platforms are big data environments in which raw big data sets can be transnumerated under user request (i.e., converted into a new type of representation, within the platform restrictions) by AI. For example, search engines like Google and Google Trends use AI algorithms to analyze vast amounts of text online and determine, as quickly as possible, the most appropriate result for a particular search. Through these AI-driven IT platforms, users must look for connections and significant relationships within the available data, which will result in data-first answers, and then users will work backward to find the questions that should have been asked (Claverie-Berge, 2012; Glaser, 2018).

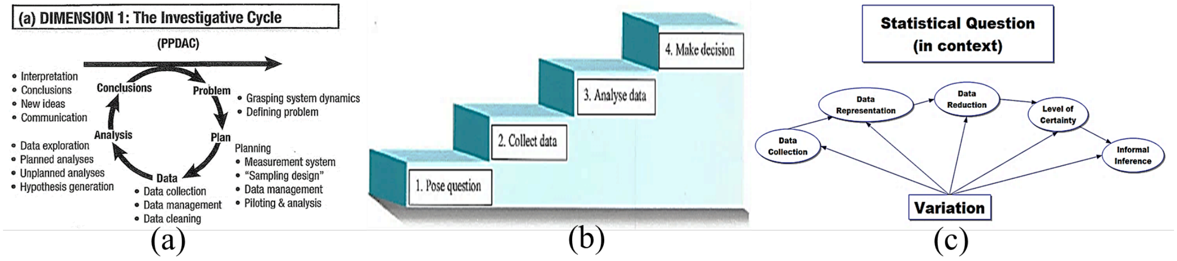


Figure 1. Frameworks of statistical thinking as the practice of statistics: (a) Wild and Pfannkuch (1999), (b) Franklin et al. (2007), and (c) Watson (2016)

A NEW FRAMEWORK OF STATISTICAL THINKING FOR USERS OF BIG DATA

As big data become more readily available, the fact that opportunistic data is rarely, if ever, dealt with at school level (Wild et al., 2018), has implications in the training of future data users. Addressing these implications is one of the aims of the "Inclusive Mathematics for Sustainability in a Digital Economy" (InMside) project, supported by the APEC Secretariat [for a description of the project, refer to https://aimp2.apec.org/sites/PDB/Lists/Proposals/DispForm.aspx?ID=2247].

Through this project, the authors have tried to answer the following questions: What thinking processes are necessary to work with big data at school level? What competencies are required to manage such thinking processes? Then, drawing on the literature, as well as on interviews and project activities with practitioners and curriculum specialists working together for over a year in the InMside project, the following six-phase framework was developed, to describe how a big data user engages in statistical thinking with the support of an AI-driven platform:

- Assessing the quality of big data: In order to start the practice of statistics with big data, an IT platform is needed in order to handle big data databases and then enable creative discovery by the users. However, before engaging in creative discovery, users must assess the quality of the big data at hand, in order to establish, among other things, from where the data came, how it was collected and what survey or questions were used to collect it. In other words, this is a step in which the users assess big data's veracity (i.e., one of the big data's 5 V's, which refers to the assurance of quality or credibility of the collected data for the intended use). In that way, big data users will somehow determine the big data's fitness for use.
- Patterns and relationships: Iteratively look for trends within big data sets, such as patterns and linear or nonlinear relationships between variables, with the support of an AI-driven data analytics platform, based on a particular interest.

- Questions: Pose critical and worry questions, in order to find plausible explanations to the patterns and relationships found. These questions are not digging into data searching for specific metrics, as in the traditional data handling approach.
- Objectives: Set objectives related to the posed questions, in order to analyze the data.
- Data mining: Re-examine the data in the light of the objectives, explore old and new data sources, or introduce new variables for consideration, using AI, machine learning and statistics. Data mining can be data-oriented, explanation-oriented, or future-oriented.
- Designing: Provide ideas for new activities, based on the understanding of the past and present, and design plans and strategies for the future, based on the results from the data mining.

EXEMPLAR APPLICATION: POPULATION AGEING IN APEC COUNTRIES

In order to exemplify this framework, let us suppose that we are interested in exploring issues related to population ageing, a concern for many APEC societies.

Assessing the Quality of Big Data

For this exemplar application, we chose to use "Google Trends" (https://trends.google.com), an AI-driven data analytics platform, which reflects online searches people make on Google every day. According to Google, Google Trends data is comprised of an unbiased and random sample of Google search data by people from all over the world with access to Google.

Patterns and Relationships

Based on our interest in population ageing issues on APEC countries, we might check the worldwide trend of web searches for related terms, such as "social security" and "nursing home", to identify possible patterns and relationships in the data regarding these terms [see Figure 2]. Although for this example we chose to identify the correlations and connections between two related terms, we could also have tried to identify the correlations and connections between "social security" and other terms that seem to be unrelated to this concept.

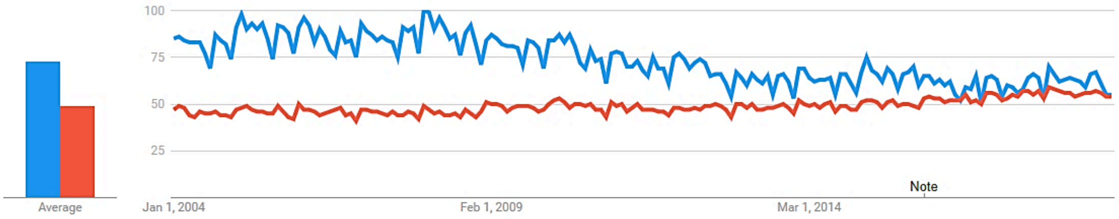
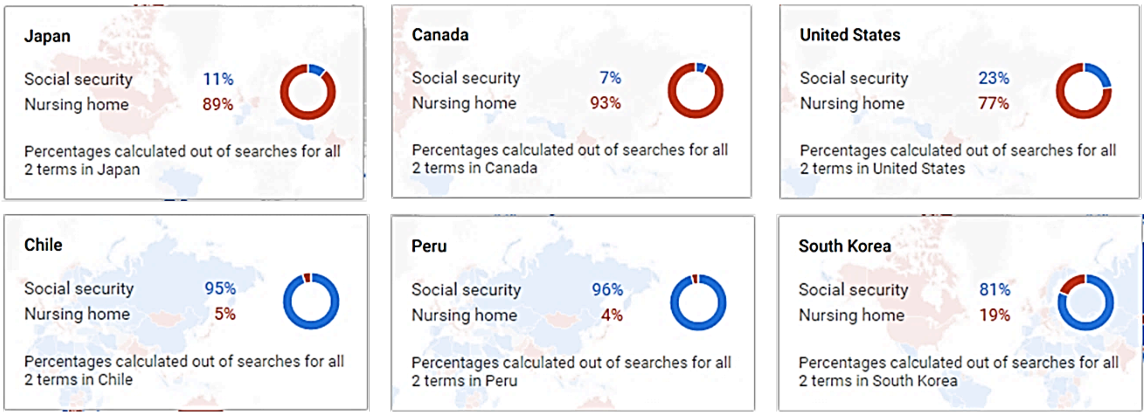


Figure 2. Evolution of the worldwide Google search trends for "social security" (in blue) and "nursing home" (in red) from 2004 to 2019

Now, let us focus on the online search trends for the terms "social security" and "nursing home" in APEC countries. In 2018, in the countries shown in the top row of Figure 3, online searches for the term "nursing home" were higher in comparison to the term "social security". On the other hand, in the countries appearing in the bottom row of Figure 3, the opposite happened.



Japan
Social security 11%
Nursing home 89%
Percentages calculated out of searches for all 2 terms in Japan

Canada
Social security 7%
Nursing home 93%
Percentages calculated out of searches for all 2 terms in Canada

United States
Social security 23%
Nursing home 77%
Percentages calculated out of searches for all 2 terms in United States

Chile
Social security 95%
Nursing home 5%
Percentages calculated out of searches for all 2 terms in Chile

Peru
Social security 96%
Nursing home 4%
Percentages calculated out of searches for all 2 terms in Peru

South Korea
Social security 81%
Nursing home 19%
Percentages calculated out of searches for all 2 terms in South Korea

Figure 3. Percentage comparison of Google searches for the terms "social security" (in blue) and "nursing home" (in red) in six APEC countries in 2018

Questions

From Figures 2 and 3, it is possible to pose questions intended to find explanations to the patterns and relationships found. These questions can lead to the creation of more data representations. Some questions that might be posed for this example are the following:

- Why do some countries seem to show considerably more interest on "nursing home" than on "social security", or vice versa?
- In countries where it seems there is more online search interest on "nursing home" than on "social security", what is the behavior of related queries, such as "nurse"?
- In countries where it seems there is more online search interest on "social security" than on "nursing home", what is the behavior of related queries, such as "tax" or "pension"?

Objectives

From the posed questions, we are now able to set clear objectives, such as the following:

- To look for and identify the reasons why some APEC countries seem to show considerably more or less online search interest on "nursing home" than on "social security".
- To determine the behavior of related queries (e.g., "nurse" for countries showing more online search interest on "nursing home", and "tax" or "pension" in countries showing more online search interest on "social security") in APEC countries with a particular search trend.

Data Mining

Data mining is defined as the process of using AI, machine learning and statistics to extract information, such as patterns and knowledge, from data sets (Cumming et al., 2017). To exemplify this phase, let us address the first objective previously stated. In the case of Japan and other APEC countries, the main reason could be the current structure of the population pyramid. Thus, explanation-oriented data mining is needed to provide informed and reasonable explanations of the current situation of the phenomena at hand [see Figure 4].

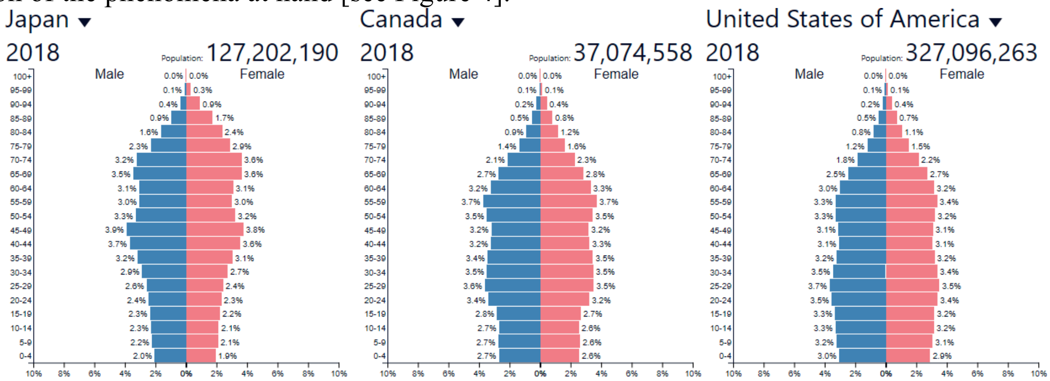


Figure 4. Population pyramids in 2018 for Japan, Canada and the USA

In Japan and Canada, a large group of people close to, or already in, retirement age, might be planning to live in a nursing home. In the US, young people might be looking for information on nursing homes for their elderly parents. These two hypotheses should be supported by data, so we need to engage in data-oriented data mining in order to accept or reject them. After addressing the validation of these hypotheses, we can engage in further data mining, aimed at making informed inferences, imagining the future of "nursing homes" in these countries (i.e., future-oriented data mining, which is the information extraction process aimed at clarifying where a particular phenomenon might be headed to).

Designing

This stage is strongly related to the fifth V in big data: value. After engaging in the previous four phases of the practice of statistics with big data, data users should be able to understand how they and others could benefit, in relation to particular aspects of the phenomenon at hand. Thus, from understanding the rising need for nursing homes in countries like Japan, someone could design business plans to service senior citizens, such as in-home care services, e-commerce stores for the elderly, transportation services, and foreign nurse recruitment agencies.

CONCLUSIONS

It is important to develop students' statistical thinking skills for engaging them in the practice of statistics through question-then-answer research methods, as "data detectives": collecting and producing evidence using planned processes, while addressing a statistical question. However, with the undeniable importance and presence of big data in today's society, it is also necessary to develop students' big data analytics skills, which require statistical thinking processes that do not necessarily follow the line of thought sketched by traditional statistical inquiry frameworks: students must be "data forensic scientists" when they test and analyze evidence collected by others, in an attempt to generate and discuss findings. Considering the implications in the training of future data users on handling and analyzing big data, we propose a six-phase framework for statistical thinking to support big data analytics using AI-driven platforms at school level. Some of the key features of these phases are illustrated with an exemplar application, addressing a sustainability issue relevant to many countries. Since the big data approach is an iterative process, phases and within-phase steps may have to be repeated, depending on the knowledge, connections and behaviors revealed. One possibility is that, after the "Question" phase is completed, the traditional data analytics approach could be performed.

We hope that this framework—and similar research efforts, such as the "International Data Science in School" project, http://www.idssp.org—will provide future users of statistics with the skills in big data handling and use of AI-driven platforms that are required to actively participate in today's digital society as capable statistical thinkers. By doing so, this article might help address the educational challenges related to the effective instructional use of big data at school level.

REFERENCES

Claverie-Berge, I. (2012, March). *Solutions Big Data IBM*. Lecture Presented at the IBM Petit Déjeuner Marketing Digital Meeting, Paris, France.

Cumming, G. P., French, T., Hogg, J., McKendrick, D., Gilstad, H., Molik, D., & Luciano, J. S. (2017). Trust and Provenance in Communication to eHealth Consumers. In L. Menvielle, A. F. Audrain-Pontevia, & W. Menvielle (Eds.), *The Digitization of Healthcare* (pp. 189-203). London: Palgrave Macmillan.

English, L. D., & Watson, J. M. (2018). Modelling with Authentic Data in Sixth Grade. *ZDM Mathematics Education*, *50*(1-2), 103-115.

Ferguson, R. (2012). Learning Analytics: Drivers, Developments and Challenges. *International Journal of Technology Enhanced Learning*, *4*(5/6), 304-317.

Franklin, C., Kader, G., Mewborn, D., Moreno, J., Peck, R., Perry, M., & Scheaffer, R., (2007). *Guidelines and Assessment for Instruction in Statistics Education (GAISE) Report: A Pre-K-12 Curriculum Framework*. Alexandria, VA: American Statistical Association.

Glaser, H. (2018, August 8). Are You Asking Your Data the Wrong Questions? [Online post]. *https://www.forbes.com/sites/forbestechcouncil/2018/08/08/are-you-asking-your-data-the-wrong-questions/*

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics*. Cambridge, Massachusetts: The MIT Press.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. Seoul: McKinsey.

Storey, V. C., & Song, I. Y. (2017). Big Data Technologies and Management: What Conceptual Modeling Can Do? *Data and Knowledge Engineering*, *108*, 50-67.

Watson, J. M. (2016). Linking Science and Statistics: Curriculum Expectations in Three Countries. *International Journal of Science and Mathematics Education*, *15*, 1057-1073.

Watson, J. M., & English, L. D. (2018). Eye Color and the Practice of Statistics in Grade 6: Comparing Two Groups. *Journal of Mathematical Behavior*, *49*, 35-60.

Wild, C. J., & Pfannkuch, M. (1999). Statistical Thinking in Empirical Enquiry. *International Statistical Review*, *67*(3), 223-265.

Wild, C. J., Utts, J. M., & Horton, N. J. (2018). What is Statistics? In D. Ben-Zvi, K. Makar & J. Garfield (Eds.), *International Handbook of Research in Statistics Education* (pp. 5-36). Cham, Switzerland: Springer.