A NEW FRAMEWORK FOR STATISTICAL THINKING IN TIMES OF BIG DATA AND DIGITAL ECONOMY

Orlando González
Assumption University, Thailand

Masami Isoda
University of Tsukuba, Japan

Roberto Araya
University of Chile, Chile

Maitree Inprasitha
Khon Kaen University, Thailand

INTRODUCTION

The interest in big data is growing exponentially in today's society. Commercial insights, government initiatives and even research calls, all seem to be focused on exploiting the potential of technology to capture and analyze massive amounts of data in increasingly powerful ways. Big data, that is, data that are too big for standard database software to process, is everywhere. For some, big data represents a paradigm shift in the ways that we understand and study our world, and at the very least it is seen as a way to better utilize and creatively analyze information for public and private benefit.

The concept of big data "refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al., 2011, p.1). Additionally, big data is often associated with key characteristics that go beyond the question of size, namely the 5 Vs: volume, velocity, variety, veracity and value (Storey & Song, 2017). Big data is dispersed among various platforms that operate with different standards, providers and degrees of access (Ferguson, 2012). For example, a lot of work in big data focuses on Twitter, the blogosphere, and search engine queries. All of these activities are not undertaken equally by the whole population, which raises concerning issues around the question of whose data traces will be analyzed using big data.

There are also a number of practical issues related to working with big data. These include, among others, issues we cannot afford to ignore, such as implications for the training of future teachers regarding handling and analysis of big data.

Due to the fact that big data has recently become mainstream in many research fields, including education, it is important to discuss and answer the following questions:

1. In order to function effectively in a society driven by big data and digital economy, what are the necessary processes of statistical thinking required to handle big data?
2. How can we revise current curriculum frameworks of statistical thinking to incorporate big data for the digital economy?
3. How can we incorporate core ideas of big data for the digital economy into the high school curriculum?
4. What are plausible instructional activities (exemplar applications) for teaching the fundamental ideas of statistics while fostering statistical thinking for big data and the digital economy?

A NEW FRAMEWORK FOR STATISTICAL THINKING

Many researchers (e.g., Wild & Pfannkuch, 1999; delMas, 2002; Watson, 2017) consider statistical thinking as the practice of statistics through the enactment of the different thought processes involved in statistical problem solving and statistical investigations. For us, in this digital era, statistical thinking processes do not follow the Problem-Plan-Data-Analysis-Conclusion (PPDAC) cycle (Wild & Pfannkuch, 1999) anymore, due to the shift in the way we work with data set by the arrival of big data analytics. In fact, the PPDAC cycle is a question-then-answer research method, focused on data gathered for a purpose using planned processes, chosen on statistical grounds to justify certain types of inferences and conclusions. However, in times of big data, this is actually a weakness of the PPDAC cycle, because most of the data available is opportunistic (happenstance or "found") data (including "big data"): huge amounts of data already collected by others and hosted somewhere. Nowadays, many companies have data teams exploring large sets of raw opportunistic data, looking for new connections and identifying significant correlations, while refining their analysis until they arrive at valuable understandings. This approach reverses the question-then-answer process of the PPDAC cycle. It starts with strong, data-first answers, and then works backward to find the questions that should have been asked.

We must acknowledge that any up-to-date framework for statistical thinking must be designed giving consideration to these criticisms to the PPDAC cycle. By doing so, we came up with the following framework to describe how a person engages in statistical thinking while handling big data. The proposed framework understands statistical thinking as a cognitive process comprised of the following five phases:

Patterns and relationships from data: Look for patterns and relationships within the data, based on a particular interest.

Questions: Pose critical and worry questions, in order to find plausible explanations to the patterns and relationships found.

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 the old and new data sources, or introduce new variables for consideration. Data mining can be data-oriented, explanation-oriented, or future-oriented.

Understanding and/or designing: Provide ideas for new activities, based on the understanding of the past, and design plans and strategies for the future, based on the results from the data mining.

A detailed explanation of this framework for statistical thinking (González, Isoda & Araya, 2019) was submitted for publication in the *Statistical Education Research Journal* (SERJ).

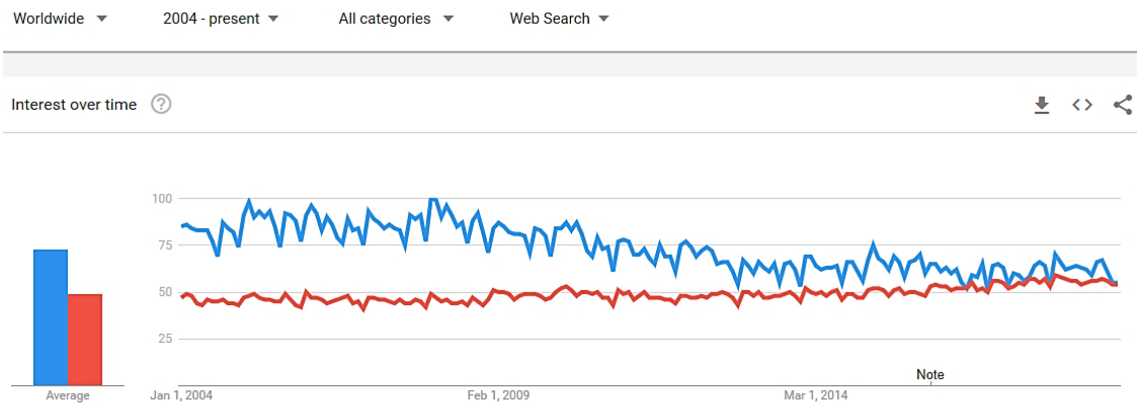
EXEMPLAR APPLICATION OF THE FRAMEWORK: AGING POPULATION ISSUES IN APEC COUNTRIES

For the purpose of exemplifying this framework, let us suppose that we are interested in exploring issues related to population ageing, which is a concerning issue in many societies, such as the Japanese. So, we may check the worldwide trend of web searches for terms such as "social security" and "nursing home", focusing on APEC countries.

For this example, we will use the website "Google Trends" (https://trends.google.com), which is an open-access online platform for big data that enables creative discovery from information on how frequently a given search term was entered into Google's search engine in real-time or within given time and date constraints.

Patterns and relationships from data

Using "Google Trends", we looked for patterns and relationships within the data hosted in the platform, based on our interest in aging population issues on APEC countries. Then, by checking the worldwide trend of web searches for terms related to aging population, such as "social security" and "nursing home", we will be able to identify possible patterns and relationships in the data regarding these terms. Figure 1 shows the evolution of the worldwide search trends for "social security" and "nursing home" in the past 15 years.

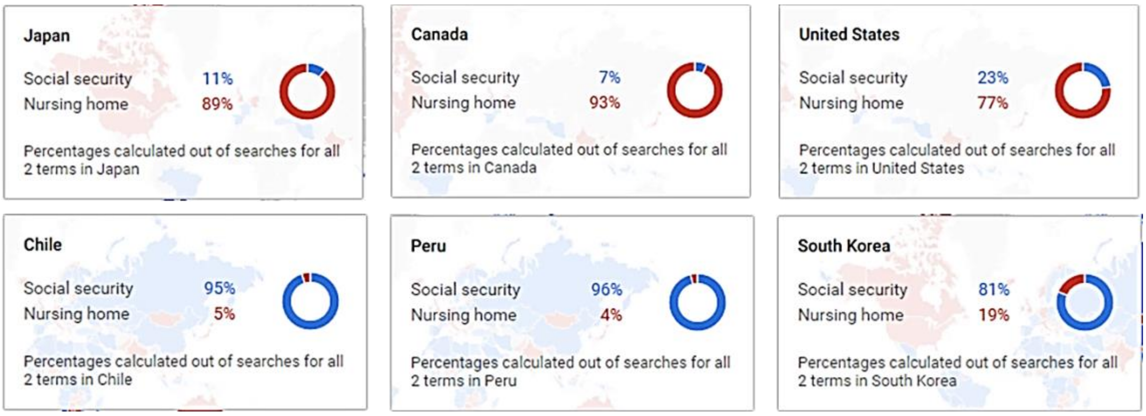


Worldwide ▾ 2004 - present ▾ All categories ▾ Web Search ▾

Interest over time

Figure 1: Evolution of the worldwide Google search trends for "social security" (in blue) and "nursing home" (in red) in the past 15 years.

Now, let us focus on the online search trends for the terms "social security" and "nursing home", focusing on APEC countries. In the top row of Figure 2, we can see that Japan, Canada and the US were APEC countries in which, in the last year, online searches for the term "nursing home" were higher in comparison to the term "social security". On the other hand, in the bottom row of Figure 2, we can see that Chile, Peru and the South Korea were APEC countries in which, in the last year, online searches for the term "social security" were higher in comparison to the term "nursing home".



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 2: 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 the visual representations obtained by using Google Trends, it is possible to pose a series of questions regarding the statistical information being displayed, in order to find explanations to the patterns and relationships found. These questions can support the process of critical evaluation of statistical messages and lead to the creation of more data representations, informed interpretations and judgments.

Some questions that might be posed for this example are the following:

- Why do some countries (such as Japan, Canada and USA) seem to show considerably more interest on "nursing home" than on "social security"?
- Why do some countries (such as Chile, Peru and South Korea) seem to show considerably more interest on "social security" than on "nursing home"?
- What are the reasons for the decreasing and increasing trends of the graphs?
- What could be the behavior of these trends for individual APEC countries (such as the ones mentioned in Figure 2) in the next decade?
- In countries such as Japan, Canada and USA, where it seems to be more interest on "nursing home" than on "social security", what is the behavior of related queries, such as "nurse"?
- In countries such as Chile, Peru and South Korea, where it seems to be more interest on "social security" than on "nursing home", what is the behavior of related queries, such as "tax" or "pension"?
- How is the social security policy in countries in showing considerably more interest on "nursing home" than on "social security" (such as Japan, Canada and USA)?
- How is the current state of nursing home services provided to the elderly in countries showing considerably more interest on "social security" than on "nursing home" (such as Chile, Peru and South Korea)?

Objectives

Now, from the posed questions, we are able to set clear objectives to address. In fact, each objective should be associated to at least one question. In this example, some objectives stemming from the questions above are the following:

1. To look for and identify the reasons why some APEC countries seem to show considerably more or less interest on "nursing home" than on "social security".

2. To determine the behavior for individual APEC countries regarding individual queries, such as the mentioned above.
3. To predict the trends of web searches for the terms “nursing home” and “social security” in APEC countries in the next decade.
4. To determine the behavior of related queries (e.g., “nurse” for countries showing more interest on “nursing home”, and “tax” or “pension” in countries showing more interest on “social security”) in APEC countries with a particular search trend.

Data mining

During this phase of the statistical thinking process, addressing the objectives set in the previous phase will lead to a re-examination of the data, from which new insights and knowledge discovery will emerge from three types of data mining: big data-oriented, explanation-oriented, and future-oriented data mining.

For the purpose of exemplifying this phase, let us address the Objective 1 (i.e., to look for and identify the reasons why some APEC countries seem to show considerably more or less interest on “nursing home” than on “social security”). In the case of Japan and other APEC countries, the main reason can be the current structure of the population pyramid (explanation-oriented data mining). In order to construct plausible explanations from the population pyramids, we need to select and transform the necessary data into the required form (big data-oriented data mining), as presented in Figure 3.

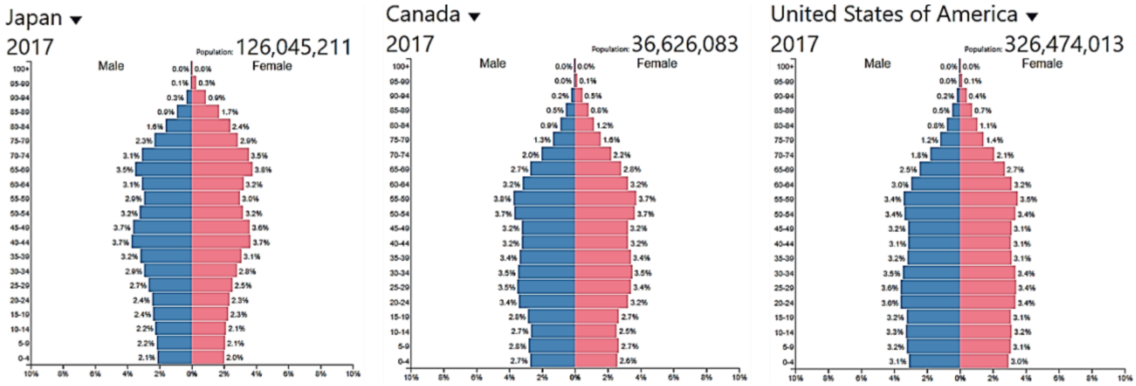


Figure 3: Population pyramids in 2017 for Japan, Canada and the USA.

So, if we look at the population pyramids of the APEC countries in the top row of Figure 2, we can see different structures among the countries, regardless the fact that, in 2018, online searches in 2018 for the term “nursing home” were higher in Japan, Canada and the US, in comparison to the term “social security” (see Figure 3).

In Japan, we can observe a large proportion of the population close to or over retirement age. In Canada, we can observe a large proportion of the population close to retirement age. In the US, this phenomenon is not an issue, because a large proportion of the population is below 50 years-old, far from retirement age. Thus, plausible explanations of why some APEC countries seem to show considerably more or less interest on “nursing home” than on “social security” may vary from country to country. In the cases of Japan and Canada, a large group of elderly people close to retirement, or already retired, might be planning to live in a nursing home. In the case of the US, young people, starting to make a live by themselves, might be looking information on nursing homes for their elderly parents.

From these hypotheses, we might make inferences, imagining the future of “nursing homes” in Japan, Canada and the US (future-oriented data mining), using big data to support our plausible explanations for the future to come (see Figure 4).

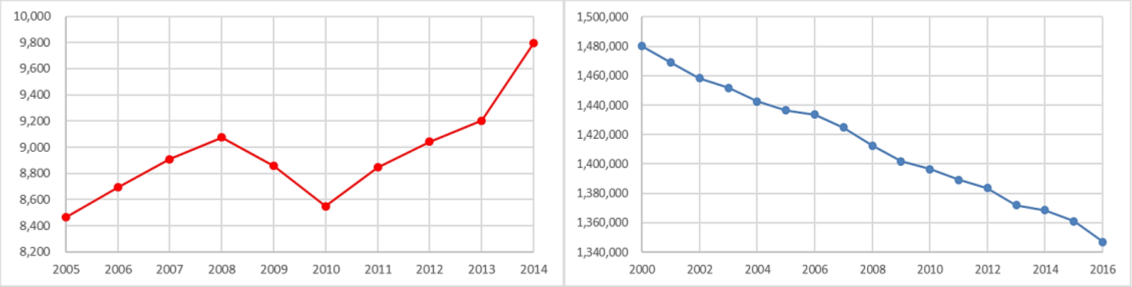
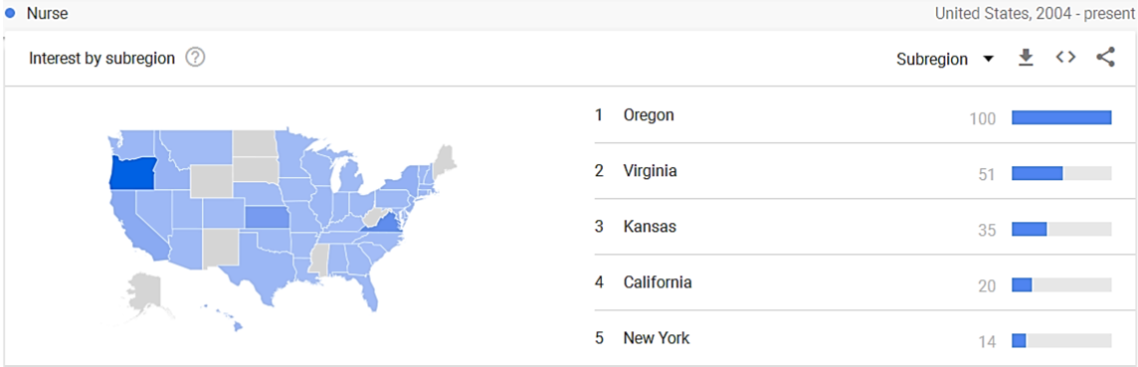


Figure 4: Number of nursing homes in Japan (red line, 2005–2014) and United States (blue line, 2000–2016).

Understanding and/or designing

From the discovered knowledge, inferences and plausible explanations generated in the previous phase, we have gained valuable understanding about the topics of interest ("nursing home" and "social security"). This understanding provides us with ideas to develop new activities related to the topics or variables of interest. Understanding the past allow us to design for the future, all supported on the data mining results.

In our exemplar application, from understanding the rising need for nursing homes and nurses in Japan, someone could design business plans targeting senior citizens: in-home care services, senior citizen transportation services, e-commerce store for the elderly, wheelchair manufacturing, foreign nurse recruitment agency, and so on. In the USA, most of these ideas (e.g., a foreign nurse recruitment agency to provide care service for the elderly) will not work everywhere, but could be successful in states such as Oregon and Virginia, as shown in Figure 5.



1	Oregon	100
2	Virginia	51
3	Kansas	35
4	California	20
5	New York	14

Figure 5: Interest in Google searches for the term "nurse" in United States in the last 15 years.

CONCLUSIONS

In times of big data, artificial intelligence and digital economy, statistical thinking has evolved from a traditional question-then-answer analysis, through which we ask questions and then collect and analyze the data to arrive at a conclusion that can be used to make decisions. Now, we should use a more disruptive and creative approach, which starts with data-first answers from the examination of opportunistic data (which is data that just happen to be available in electronic form because they have accumulated for other reasons by other people), and then works backward to find the questions that should have been asked.

Under this scenario, previous frameworks of statistical thinking, such as the PPDAC cycle, is not appropriate to explain the richness and complexity of thinking involved in real-world statistical investigations dealing with big data. On that regard, we developed a new framework for statistical thinking, comprised of five phases or cognitive processes (i.e., patterns and relationships from data, questions, objectives, data mining, and understanding and/or designing), in order to appropriately describe how a person engages in statistical thinking while handling big data. An exemplar application of the framework illustrated the richness and complexity of thinking involved in handling big data with Google Trends, exploring issues related to population ageing, by checking the worldwide trend of web searches for terms such as "social security" and "nursing home", focusing on APEC countries. From this application, it was possible to identify countries with similar and contrasting characteristics, and classify APEC countries based on such characteristics.

Although we have illustrated our proposed new framework for statistical thinking with an exemplar application, issues regarding how we can incorporate core ideas of big data for the digital economy into the high school curriculum of each APEC economy, as well as what plausible instructional activities (exemplar applications) can be designed for teaching the fundamental ideas of statistics while fostering statistical thinking for big data and the digital economy, are expected to be fundamental outcomes from the discussion of the current document among all the participants to this APEC-Tsukuba International Conference 2019.

REFERENCES

delMas, R. C. (2002). Statistical literacy, reasoning and learning: A commentary. *Journal of Statistics Education*, *10*(3). Retrieved from http://jse.amstat.org/v10n3/delmas_discussion.html

Ferguson, R. (2012). Learning analytics: Drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, *4*(5/6), 304–317.

González, O., Isoda, M., & Araya, R. (2019). A new framework for statistical thinking in times of big data and digital economy. Manuscript submitted for publication to *Statistical Education Research Journal*.

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 Global Institute.

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. (2017). Linking science and statistics: Curriculum expectations in three countries. *International Journal of Science and Mathematics Education*, *15*, 1057–1073.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, *67*(3), 223–265.