APEC Seminar on Computational Thinking Curriculum for the Digital Economy – 2nd APEC Senior Officials' Meeting (SOM2)
May 2, 2019 – Viña del Mar, Chile

Keynote Lecture 3:
A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy

Dr. Orlando GONZÁLEZ
Graduate School of Human Sciences
Assumption University, Thailand

Dr. Masami ISODA, Dr. Roberto ARAYA and Dr. Maitree INPRASITHA

Outline

This lecture will last 30 minutes, organized as follows:

– Traditional School Statistics vs. Data Science

– Understanding Big Data in the Digital Era

– Statistical Thinking as the Practice of Statistics

– A New Framework for Statistical Thinking for Users of Big Data

– Discussion of an Exemplar Application of the Framework: Aging Population Issues in APEC Countries

– Q & A Session

Traditional School Statistics vs. Data Science

- In today's society (known as "Digital Society", "Information Society" or "Society 4.0"), knowledge, data and technology play fundamental roles in any aspect of life.

- Under this scenario, APEC Inclusive Mathematics for Sustainability in a Digital Economy (InMside) project was proposed by Japan, Chile and Thailand, in order to:

  − develop and share recommendations for high quality curriculum standards on Computational and Statistical Thinking for the Era of Digital Society; and

  − create a communication network among curriculum specialists in APEC economies for further collaboration.

Traditional School Statistics vs. Data Science

- In this lecture, I will share with all of you some recommendations on promoting, at school level, Statistical Thinking for the Era of Digital Society.

- Before that, we must understand:

  − what is being done traditionally in our schools regarding statistics; and

  − what statistical fundamental ideas, thinking processes and competencies Society 4.0 (and the upcoming Society 5.0) demands from people.

Traditional School Statistics vs. Data Science

- The key difference between traditional school statistics and the statistics needed for the Digital Era is the increasing importance of handling unstructured, large-scale data, commonly referred to as Big Data.

- In traditional school statistics, students are either given or asked to collect small data sets to create tables, graphs and calculate statistical measures.

- On the contrary, the statistics needed for the Digital Era demands from people to be "data scientists".

- Data science is defined as the discipline at the intersection of statistics, mathematics and computer science, concerned with extracting meaning from big data.

| APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | 4 |

Traditional School Statistics vs. Data Science

- There are other key differences between traditional school statistics and the statistics needed for the Digital Era:

  - Traditional school statistics is typically concerned with making inferences from datasets that are too small, while data science is concerned with extracting a signal from datasets that are too big.

  - The role of AI and machine learning in traditional school statistics is negligible, while there is no data science without AI and machine learning.

- Are we preparing our students to be "data scientists"?

| APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | 5 |

Traditional School Statistics vs. Data Science

- Of course, literacy in traditional statistics still is, and will be, an important competency for academic, professional and daily-life success.

- However, with Big Data and AI at the core of today's Digital Society, and the uprising of the super-smart Society 5.0, people are expected to be able to identify needs and challenges faced throughout society, and come up with scenarios to solve them, making use of digital technologies and Big Data.

- Then, with Big Data and AI as fundamental ideas for the Digital Era, are we preparing our students to make use of them?

Understanding Big Data in the Digital Era

- In this Digital Era, "Big Data" is understood as data sets usually meeting the following five criteria:

Volume

Variety

Velocity

Veracity

Value

5 V's of Big Data

(Claverie-Berge, 2012; Storey & Song, 2017)

Understanding Big Data in the Digital Era

- **Volume**: This refers to data at rest. Datasets with sizes in the order of terabytes, petabytes and zettabytes.

- **Variety**: This refers to the many forms (e.g., text, images, videos, audio files, emails) and sources (e.g., spreadsheets, databases, social media and monitoring devices) of data.

- **Velocity**: This refers to data in motion: the speed at which data flows in from sources (e.g., real time streaming).

- **Veracity**: This refers to data in doubt. Depending on its origin, processing technologies, and collection methods, data can have biases and inaccuracies attached, which need to be identified and accounted for.

- **Value**: This refers to turning data into profit. Just having big data is of no use unless we can turn it into value.

| APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | 8 |

Statistical Thinking as the Practice of Statistics

BASIC LITERACY	REASONING	THINKING
IDENTIFY DESCRIBE REPHRASE TRANSLATE INTERPRET READ	WHY? HOW? EXPLAIN (THE PROCESS)	APPLY CRITIQUE EVALUATE GENERALIZE (delMas, 2002)

- Many researchers (e.g., del Mas, 2002; Watson et al., 2018) 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 such a way that it leads to the most reliable, logical, and trustworthy conclusions.

| APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | 9 |

Statistical Thinking as the Practice of Statistics

- The school curriculum should provide both
    - experience in the practice of statistics, and
    - opportunities to critically evaluate data from varied sources (e.g., other students, science reports, social media, AI-driven data analytics platforms).
- But, how do we give students the opportunity to experience the "Practice of Statistics" with Big Data in the classroom?
- Before discussing this, let us revise some of the current curriculum frameworks of statistical thinking.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018)	Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy	May 2, 2019	10

Statistical Thinking as the Practice of Statistics

The Practice of Statistics: The PPDAC Model

(a) DIMENSION 1: The Investigative Cycle

(PPDAC)

- Interpretation
- Conclusions
- New ideas
- Communication

- Data exploration
- Planned analyses
- Unplanned analyses
- Hypothesis generation
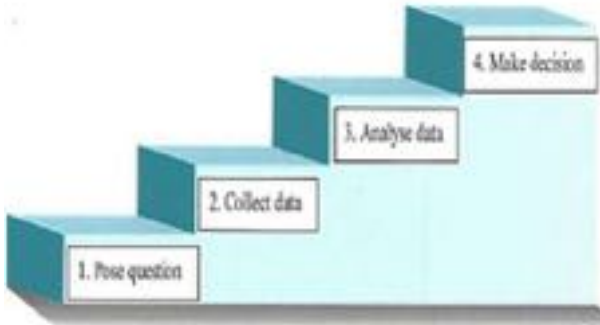
Conclusions — Problem
- Grasping system dynamics
- Defining problem

Analysis — Plan

Data

Planning
- Measurement system
- "Sampling design"
- Data management
- Piloting & analysis

- Data collection
- Data management
- Data cleaning

Wild and Pfannkuch (1999) suggested the Investigative Cycle (better known as PPDAC Model) from their analysis of the work of students of statistics and practicing professional statisticians.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018)	Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy	May 2, 2019	11

Statistical Thinking as the Practice of Statistics

The Practice of Statistics: Franklin et al. (2007) Model

A typical example of the traditional question-then-answer method. From the start, we ask questions like "How many tourists visited Tokyo in the last decade?"
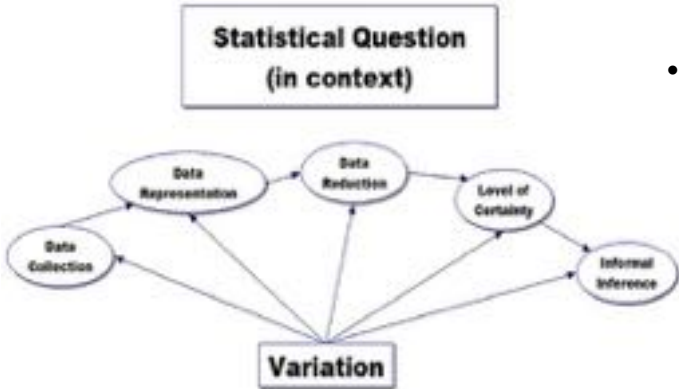
- In the GAISE report of the American Statistical Association, Franklin et al. (2007) suggested the practice of statistics through engaging in four-stage statistical investigations.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | **12**

Statistical Thinking as the Practice of Statistics

The Practice of Statistics: Watson (2009) Model

Another typical example of the traditional question-then-answer method.

- Watson (2009) suggested a way to engage students in the practice of statistics through statistical investigations.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | **13**

A New Framework of Statistical Thinking for Users of Big Data

The Practice of Statistics: Big Data Approach (IBM, 2012)

Big Data Approach
Iterative & Exploratory Analysis

IT
Delivers a platform to enable creative discovery

Users
Explores what questions could be asked

- In order to start the practice of statistics with Big Data, an IT platform that enables creative discovery is needed.
- Then, users look for connections and significant relationships within the available data.
- This will result in data-first answers, and then users will work backward to find the questions that should have been asked.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | **Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy** | May 2, 2019 | **14**

A New Framework of Statistical Thinking for Users of Big Data

- We considered the criticisms given to previous models of statistical thinking, keeping in mind the way in which big data analytics should be carried out.

- By doing so, we came up with a framework that understands Statistical Thinking in the Era of Big Data for the Digital Economy as a cognitive process comprised of the following five phases:
  − *Patterns and relationships from data*.
  − *Questions*.
  − *Objectives*.
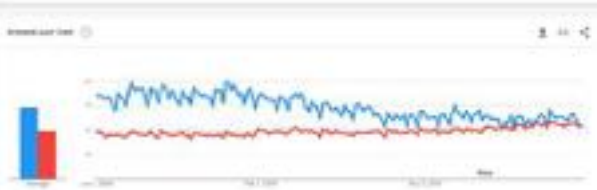 − *Data mining*.
  − *Understanding and/or designing*.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | **Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy** | May 2, 2019 | **15**

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

- Let us suppose that we are interested in exploring issues related to population ageing, focusing on APEC countries.
- To do so, we checked the worldwide trend of web searches for the terms "social security" and "nursing home", using the online AI-driven platform "Google Trends" (https://trends.google.com).

Patterns and Relationships

Using Google Trends platform, we look for patterns (trends) and relationships within the data, based on our particular interest: online search terms for "social security" and "nursing home" in APEC countries.

Worldwide search trends in the last 15 years (2004-2019)

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | 16

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

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.

What questions would you pose? Let's hear them!

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | 17

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Questions

From the obtained visual representations, we can pose questions about to find explanations to the patterns and relationships found.

Some questions that might be posed for this example are:

– 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 could be the behavior of these trends for individual APEC countries in the next decade?

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | **Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy** | May 2, 2019 | **18**

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Objectives

From the posed questions, we can set clear objectives to address.

Some objectives stemming from the questions we posed earlier 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.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | **Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy** | May 2, 2019 | **19**

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Data mining

Addressing the previous objectives will lead to a re-examination of the data, from which new insights and knowledge discovery will emerge from three types of data mining:

– *Explanation-oriented data mining*.

– *Confirmatory data mining*.

– *Future-oriented data mining*.

Data mining is the process of using artificial intelligence, machine learning and statistics to extract information (patterns and knowledge) from data sets. (Cumming, French, Hogg, McKendrick, Gilstad, Molik, & Luciano, 2017)

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | **Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy** | May 2, 2019 | **20**

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Data mining

– For APEC countries showing considerably more interest on searching for the term "nursing home", the main reason could be the current structure of their population pyramids.
– This is *explanation-oriented data mining*: providing reasonable explanations for the new knowledge (patterns and relationships) extracted from big data.
– To do this, old and new sets of raw big data will be explored, looking for identifying significant correlations and new connections, in order to arrive at insights and extract new knowledge (patterns and relationships).
– This is big data-oriented *confirmatory data mining*.
– This is not the traditional data exploration (i.e., digging into data searching for specific metrics). Here, identifying new patterns and relationships is the goal.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | **Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy** | May 2, 2019 | **21**

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Data mining

Figure 3: Population pyramids in 2017 for Japan, Canada and the USA, APEC countries with more online searches for "nursing home" than for "social security".

From the big data-oriented confirmatory data-mining:

– In Japan, a large proportion of the population close to or over retirement age.

– In Canada, a large proportion of the population close to retirement age.

– In the US, a large proportion of the population is below 50 years-old.

Remember: Explanations might not be found in the data initially examined!

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018)	Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy	May 2, 2019	22

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Data mining

Figure 3: Population pyramids in 2017 for Japan, Canada and the USA, APEC countries with more online searches for "nursing home" than for "social security".

Supporting your explanations will require more *confirmatory data mining*, possibly using new big data sources.

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018)	Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy	May 2, 2019	23

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Data mining

From our hypothetical explanations, we might make inferences, imagining the future of "nursing homes" in Japan, Canada and the US (*future-oriented data mining*), using big data.

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

- In which country nurses for elderly care will be a good career option in the near future?
- In which country will be the establishment of nursing homes a more profitable business?

| APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | 24 |

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Understanding and/or designing

From the new knowledge, inferences and plausible explanations generated in the previous phase, we have gained valuable understanding about the topics of interest.

This understanding provides us with ideas to develop new activities related to the topics or variables of interest, aiming to turning them into value of some sort.

For example, 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, etc.

Any more ideas about how to turn this knowledge into value?

| APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018) | Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy | May 2, 2019 | 25 |

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

Understanding and/or designing

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 like Oregon and Virginia.

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

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018)	Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy	May 2, 2019	26

Exemplar Application of the Framework: Aging Population Issues in APEC Countries

After this presentation, what have we learned about 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?

APEC Seminar at APEC-SOM2 (InMside Project - APEC HRDWG 01-2018)	Keynote Lecture 3: A Framework of Statistical Thinking in the Era of Big Data for the Digital Economy	May 2, 2019	27