Statistical Thinking & Methodology: Pillars of Data Availability & Quality in the Big Data Era

Pedro Luis do Nascimento Silva

Principal Researcher, ENCE

Contents

- Context
- Data quality
- Quality frameworks
- Total survey error
- Quality in big data context
  - An example regarding coverage bias
- Statistical Science
- Summarizing

Sustainable Development Goals

“On September 27th 2015, 193 world leaders committed to 17 Global Goals to achieve 3 extraordinary things in the next 15 years.

End extreme poverty.

Fight inequality & injustice.

Fix climate change.”

https://www.un.org/sustainabledevelopment/sustainable-development-goals/

Data quality issues

“To reach these Sustainable Development Goals (SDGs), we will need to confront **a crisis** at the heart of solving many of the world’s most pressing issues—a **crisis of poor use, accessibility, and production of high quality data** that is stunting the fight to overcome global challenges in every area—from health to gender equality, human rights to economics, and education to agriculture**.**

The availability and access to **high quality data** is essential to measuring and achieving the SDGs.”

http://www.data4sdgs.org/#intro

Data quality in the Big Data era

More data does not necessarily mean good or better data!

Many of the data available **lack the quality** required for its safe use in many applications.

Challenges cab be even bigger with Big Data!

Data quality

Quality is desirable attribute of all data.

Data quality derives from **quality of the source**(s), **measurement instruments** & **methods**.

Vague concept: **what is data quality**?

Must be defined, so that it can be planned, measured and evaluated.

Frameworks for data quality

Several important organizations have invested in defining frameworks for data quality.

Quality frameworks:

- US Office of Management and Budget (2006);
- Statistics Canada (2009);
- International Monetary Fund (2012);
- OECD (2012);
- UN (2012);
- IBGE (2013).

http://www.ibge.gov.br/home/disseminacao/eventos/missao/codigo_boas_praticas.shtm

OECD Quality Framework

Quality Dimension	Description
Relevance	Statistics and data are relevant if they satisfy user's needs.
Accuracy	Refers to the closeness between the values (estimates) provided and the (unknown) true values.
Credibility	Credibility of data products refers to the confidence that users place in those products.
Timeliness	Timeliness of data products reflects the length of time between their availability and the event or phenomenon they describe.
Acessibility	Accessibility of data products reflects how readily the data can be located and accessed.
Interpretability	Interpretability of data products reflects the ease with which the users may understand and properly use and analyse the data.
Coherence	Coherence of data products reflects the degree to which they are logically connected and mutually consistent.
Cost-efficiency	Cost-efficiency with which a product is produced is a measure of the costs and provider burden relative to the outputs.

OECD Statistics Directorate (2012).

Data quality

Two complementary approaches / trajectories (Lyberg, 2012):

- Models for the **Total Survey Error**;
- **Survey Process & Quality Management** ➔ continuous quality improvement.

Total Survey Error

Four principles guiding design, implementation, evaluation and analysis of surveys:

- **Consider** all known error sources;
- **Monitor** main error sources during implementation;
- **Evaluate** key error sources after completing survey; and
- **Study the effects** of errors on key outputs and analysis.

Total Survey Error

Strength:

Survey is planned to control main error sources.

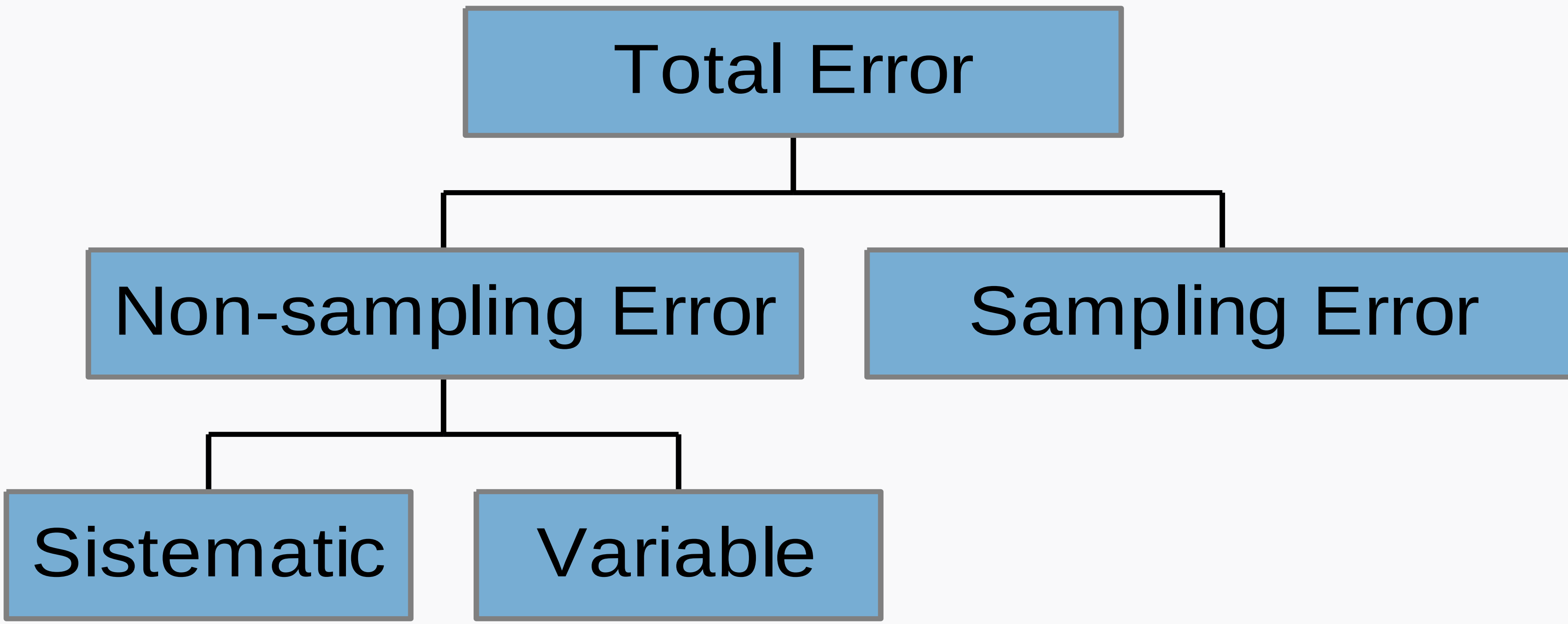
Weakness:

Proper assessment of total survey error is hard and costly to do in practice.

Errors in surveys

“Error” in Estimates

Error = Estimate – True Value



Source: United Nations (2005).

Sampling Error

Easier to control.

Bias (systematic error) may be avoided ➔ use **probability sampling.**

Sample design, sample size and **estimator** defined to make **variable sampling error** as small as required.

Sampling Error

Easier to control.

Bias (systematic error) may be avoided ➔ use **probability sampling.**

Sample design, sample size and **estimator** defined to make **variable sampling error** as small as required.

With ‘Big Data’, there may no longer be sampling error in some applications!

Non-sampling Error

Two broad classes of **non-sampling errors**.

Errors due to ‘**non-observation**’:

- Coverage (frames, populations);
- Non-response (collection).

Errors in **observations**:

- Specification;
- Measurement;
- Processing & estimation.

Non-sampling Error

Two broad classes of **non-sampling errors**.

Errors due to ‘**non-observation**’:

- Coverage (frames, populations);
- Non-response (collection).

Errors in **observations**:

- Specification;
- Measurement;
- Processing & estimation.

With ‘Big Data’, non-sampling errors dominate!

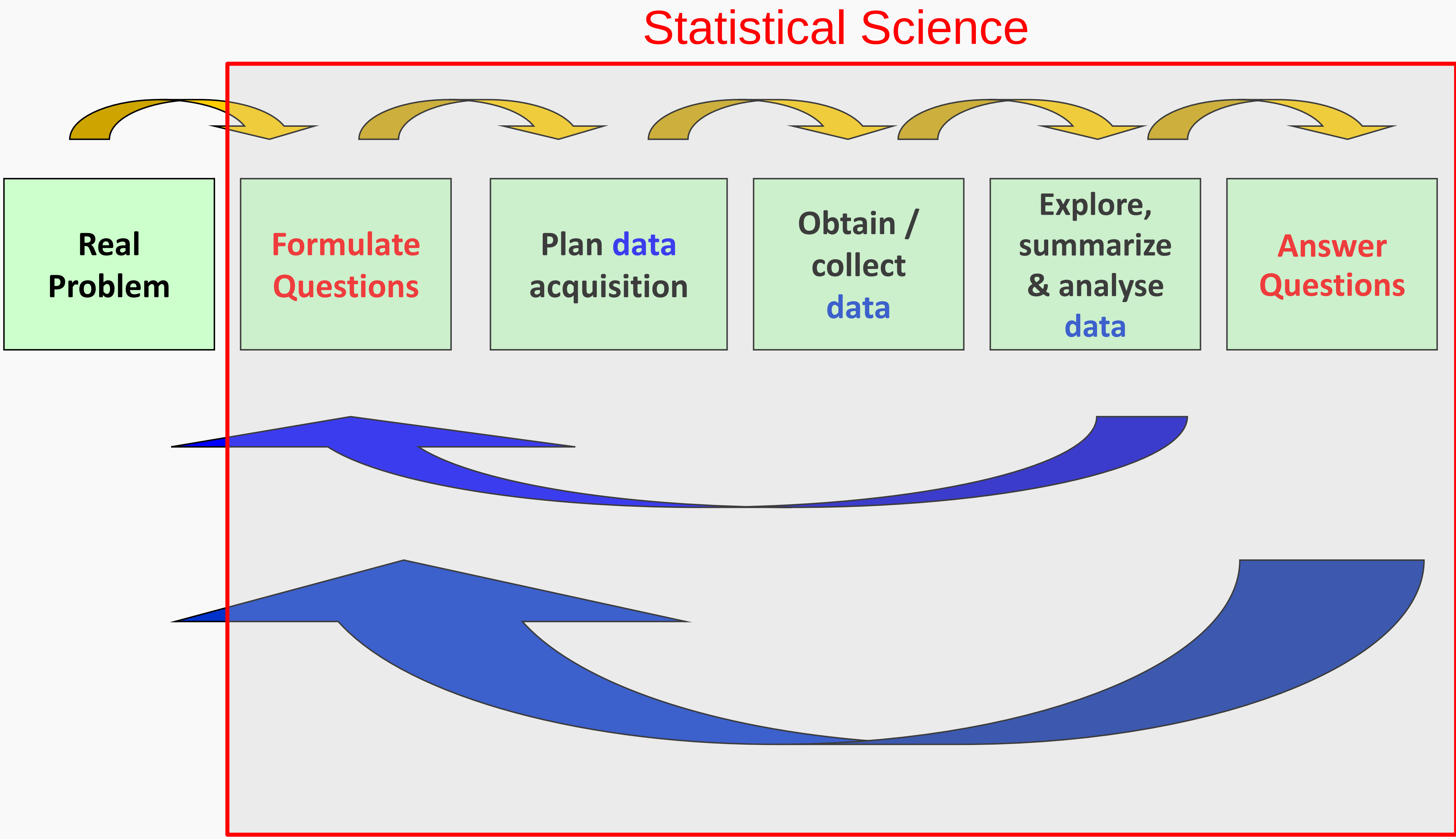
Statistical Science

For all the above reasons, **Statistical Science** has never been in such **evidence** and in such **high demand**.

Statistical thinking & methodology offers the essential guidance to obtaining current, relevant, accurate and cost-effective data.

It also guides the **extraction of useful knowledge** from data, to support decision making.

‘Conventional’ Knowledge Generation Process



Official and Public Statistics

Typical data sources (observational studies)

Censuses

Data obtained from **every unit** in the target population.

Sample surveys

Data obtained from **samples of units** in the target population.

Administrative records

Data obtained for admin purposes, but later used for statistical purposes.

Big Data

New and emerging **data sources**:

“Big Data are data sources that can be – generally – described as: high **volume**, **velocity** and **variety** of data that demand cost-effective, innovative forms of processing for enhanced insight and decision making.”

UNECE Definition 2013

Types of sources:

Social networks (communications; images; searches);

Traditional business data (transactions; records);

‘Internet of things’ (sensor data).

UNECE Classification:
http://www1.unece.org/stat/platform/display/bigdata/Classification+of+Types+of+Big+Data

Big Data Quality Issues for Official Statistics

Variability or Volatility

Inconsistence and/or instability of data across time.

Veracity

Ability to trust that data is accurate and/or complete.

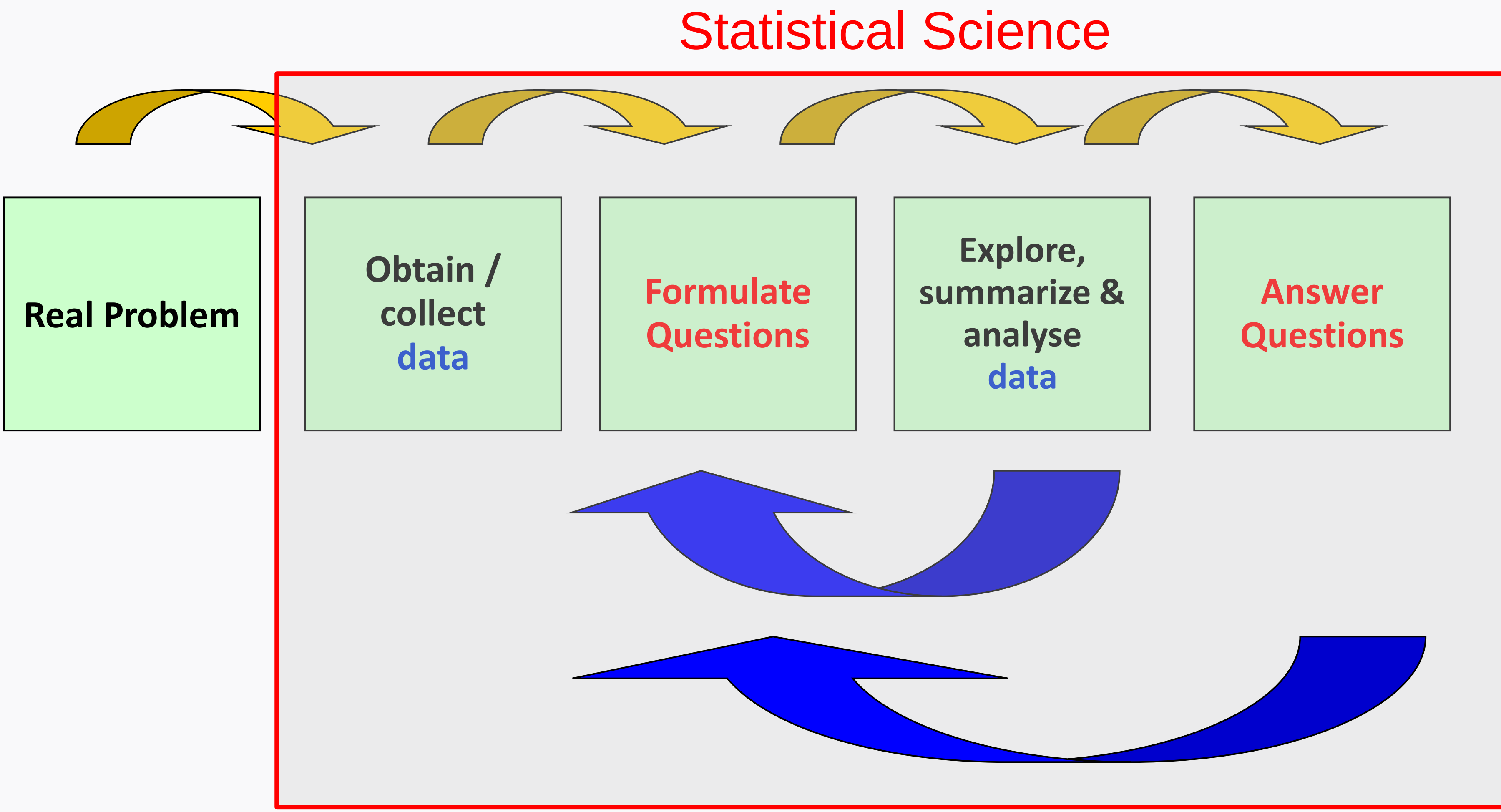
Complexity

Need to link multiple data sources.

Accessibility

Need to ensure that data is and will be available.

Knowledge Generation Process in the Big Data Era



Core quality issues

Coverage bias

Available data does not fully cover target population

Missing data

Data of interest not available for all records on file

Measurement error

Data available may be poorly recorded or measured

Specification error

Data available may be different from concept of interest

A self-monitoring social and economic eco-system is emerging

- Designed (or traditional survey) data
  - Data produced to discover the unmeasured
- Organic (or big) data
  - Data produced auxiliary to processes, to record the process

Blending these two types of data is the future.

6

GEORGETOWN UNIVERSITY

Robert Groves
http://directorsblog.blogs.census.gov/2011/05/31/designed-data-and-organic-data/

Coverage bias - Meng (2018)

Mean square error for estimating the population mean from a Register:

$$EQM(\bar{y}_{\mathcal{R}}) = E_R(\bar{y}_{\mathcal{R}}-\bar{Y})^2 = E_R\left(\rho^2_{R,y}\right) \times \left(\frac{1-c}{c}\right) \times \sigma^2_y$$

$\bar{y}_{\mathcal{R}} = \frac{1}{m}\sum_{k\in\mathcal{R}} y_k$ estimates population mean $\bar{Y}=\frac{1}{N}\sum_{k\in U} y_k$;

$\rho_{R,y}=\frac{Cov_1(R_k;y_k)}{\sqrt{V_1(R_k)V_1(y_k)}}$ is the correlation between Register inclusion indicator and the variable of interest y;

$\sigma^2_y=\frac{1}{N}\sum_{k\in U}(y_k-\bar{Y})^2$ and $c = m / N$ is the Register's coverage rate.

Coverage bias - Meng (2018)

Size of SRS needed for smaller MSE than that of register based estimate for population mean $\bar{Y}$

N	c	m	rho_R,y		
			0,01	0,05	0,1
200.000.000	50%	100.000.000	10.000	400	100
	80%	160.000.000	40.000	1.600	400
	95%	190.000.000	190.000	7.600	1.900

Potential remedies (Kim & Wang, 2018)

Three alternative ideas to tackle **coverage bias**:

- Make sure that $\rho_{R,y}$ - not realistic in general;
- Apply **inverse sampling** to Register so that ‘SRS-like’ inference from selected sample is unbiased;
- Estimate Register **selection propensity scores** (PS), and use these to perform PS-inverse weighted estimation – should be approximately unbiased under mild assumptions.

Statistical Science

Offers solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;

Statistical Science

Offers solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;
- Exploratory analysis and data cleaning and preparation;

Statistical Science

Offers solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;
- Exploratory analysis and data cleaning and preparation;
- Formulation and fitting of statistical models to describe data in synthetic form;

Statistical Science

Offers solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;
- Exploratory analysis and data cleaning and preparation;
- Formulation and fitting of statistical models to describe data in synthetic form;
- Using fitted models to answer formulated questions (inference);

Statistical Science

Offers solutions for research and knowledge discovery via:

- Careful planning and realization of data & measurement acquisition operations regarding phenomena of interest;
- Exploratory analysis and data cleaning and preparation;
- Formulation and fitting of statistical models to describe data in synthetic form;
- Using fitted models to answer formulated questions (inference); and
- Creating visual displays of data, summaries and key findings revealed from the data.

Obtaining Data

Methods for careful planning and conducting of cost-effective data gathering studies

- Sampling;
- Design of experiments;
- Design for observational studies;
- Measurement protocols (questionnaires, instruments, etc.)
- Data checking, cleaning, storage and sharing protocols.

Analysis / discovery

Methods for exploratory and confirmatory data analysis:

- Exploratory data analysis;
- Data mining;
- Hypothesis formulation and testing;
- Model formulation, fitting, selection, diagnostics and interpretation;
- Data summarization, presentation & visualization.

Seven Pillars Core Ideas

#1 Targeted reduction/compression of data

#2 Diminishing value of more data

#3 Putting a probability measure to inferences

#4 Doing this based upon internal data variation

#5 Different perspectives give different answers

#6 The essential role of planning / designing studies

#7 How to explore in nested families of models

Stigler (2015)

Statistical Science

“These Seven Pillars **are not** *Mathematics* and are not *Computer Science*.

They do centrally constitute the important **core ideas** underlying the **Science of Statistics**.”

Stigler (2015)

Summarizing

Data quality remains fundamental concern.

Statistical thinking & methodology is essential pillar for promoting data quality.

Big data era will require **more statistical development**, not less:

- In the past, small n & small p;
- With Big Data, large n or large p or both!

Thanks for your attention.

References

1. European Foundation for Quality Management (1999). *The EFQM Excellence Model*. Van Haren.
2. IBGE (2013). Código de Boas Práticas das Estatísticas do IBGE. Rio de Janeiro: IBGE.
3. International Monetary Fund. 2012. *Data Quality Assessment Framework - Generic Framework*.
4. KIM, J.K.; WANG, Z. (2018). Sampling techniques for big data analysis in finite population inference. Int. Stat. Rev. https://doi.org/10.1111/insr.12290 .
5. Lyberg, Lars. 2012. “Survey Quality.” *Survey Methodology* 38 (2): 107–130.
6. Meng, S. L. (2018) Statistical paradises and paradoxes in Big Data (I): law of large populations, big data paradox, and 2016 US Presidential Election. Ann. Appl. Stat., v. 12, n. 2, 685-726.

Instituto Brasileiro de Geografia e Estatística IBGE

References

7. Office of Management and Budget. 2006. Standards and Guidelines for Statistical Surveys. Federal Register. Washington, DC.
8. Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth edition. Ottawa, Canada: Statistics Canada.
9. Statistics Directorate, OECD. 2012. *Quality Framework and Guidelines for OECD Statistical Activities*.
10. Stigler, Stephen M. (2015). The seven pillars of statistical wisdom. Talk at LSHTM on 29 January 2015.
11. Stigler, Stephen M. (2016). The seven pillars of statistical wisdom. Harvard University Press.

Instituto Brasileiro de Geografia e Estatística IBGE

References

12. United Nations. 2005. *Household Sample Surveys in Developing and Transition Countries*. Ed. Department of Economic and Social Affairs. *Studies in Methods*. Vol. F No. 96. New York: United Nations.

13. United Nations. 2005. *Designing Household Survey Samples: Practical Guidelines*. Ed. Statistics Division Department of Economic and Social Affairs. *Studies in Methods*. Vol. F No. 98. New York: United Nations Statistics Division.

14. United Nations. 2012. Guidelines For The Template For A Generic National Quality Assurance Framework (NQAF). http://unstats.un.org/unsd/dnss/qualityNQAF/nqaf.aspx

15. Vale, Steven. 2009. *Generic Statistical Business Process Model*.

16. Weisman, Ethan, Zdravko Balyozov, and Louis Venter. 2010. *IMF ' s Data Quality Assessment Framework 1*.