IBGE
Instituto Brasileiro de Geografia e Estatística

Escola Nacional de Ciências Estatísticas

Metodologia Estatística: Um Pilar da Disponibilidade e Qualidade de Dados

Pedro Luis do Nascimento Silva

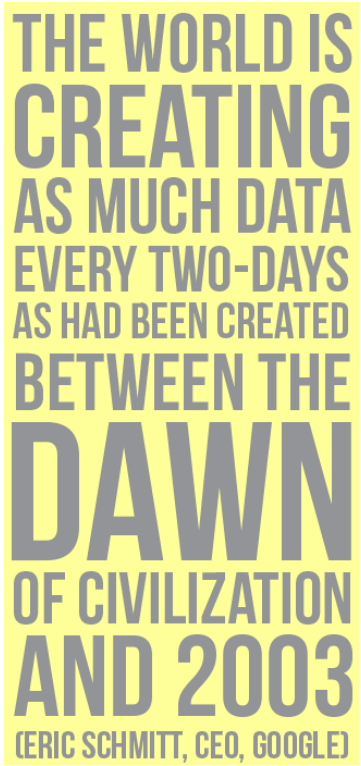
Pesquisador da ENCE/IBGE

A Era dos Dados (‘Big’ ou Não)

Vivemos numa era sem precedentes com respeito à quantidade, disponibilidade e acesso a **dados**.

A Era dos Dados (Big ou Não)

Vivemos numa era sem precedentes com respeito à quantidade, disponibilidade e acesso a dados.



Global Partnership for Sustainable Development Data (GPSDD)

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

A Era dos Dados (Big ou Não)

Vivemos numa era sem precedentes com respeito à quantidade, disponibilidade e acesso a dados.

> "Data in the world is doubling every 18 months."

IBM

http://www-01.ibm.com/software/data/demystifying-big-data/

A Era dos Dados (Big ou Não)

Vivemos numa era sem precedentes com respeito à quantidade, disponibilidade e acesso a dados.

> "O mundo hoje gera mais dados que no passado – de acordo com algumas estimativas, 90% dos dados do mundo foram gerados apenas nos últimos 2 anos."

Paris21: http://datarevolution.paris21.org/the-project

Lacunas nos Dados

"Apesar do 'dilúvio de dados', há claras lacunas. Por exemplo, em países de baixa renda mais de 70% dos nascimentos – cerca de 20 milhões de crianças por ano – não são registrados."

Paris21: http://datarevolution.paris21.org/the-project

Lacunas nos Dados

Uso de internet nos domicílios	%
Brasil	69,3
Urbano	75,0
Rural	33,6

Fonte: IBGE, PNAD Contínua 4o. Trimestre de 2016

Dados e Desenvolvimento

"Em 27 de Setembro de 2015, líderes de 193 países se comprometeram com 17 Objetivos Globais do Desenvolvimento Sustentável (ODS), visando alcançar 3 coisas extraordinárias nos próximos 15 anos:

- Acabar com a pobreza extrema;
- Lutar contra a desigualdade e a injustiça; e
- Resolver a mudança climática."

https://unstats.un.org/sdgs/

Dados e Desenvolvimento

Muitos dos dados disponíveis não têm a qualidade requerida para seu uso seguro em muitas aplicações.

> "Há uma crise no coração dos esforços para resolver os problemas mais críticos do mundo – uma crise de dados ruins. Esta crise está freiando a luta para superar os desafios globais em todas as áreas – da erradicação da pobreza, a acabar com a fome, a empoderar as mulheres, a assegurar saúde, a combater a mudança climática.

Global Partnership for Sustainable Development Data (GPSDD)

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

Dados e Desenvolvimento

No setor privado, a busca por **vantagens competitivas** também não cessa de demandar **mais dados** e '**inteligência**' ou '**conhecimento**' extraídos dos dados.

Ciência Estatística

Pelas razões acima, a **Ciência Estatística** nunca esteve em tanta **evidência** e com tamanha **demanda**.

Metodologia estatística fornece a orientação essencial para obter **dados atuais**, **relevantes**, **precisos** e **custo-efetivos**.

Também guia o saber de **como extrair conhecimento útil dos dados**, para apoiar a tomada de decisões.

Estatísticas Oficiais e Públicas

Fontes de dados típicas (estudos observacionais)

- **Censos**
  - Dados obtidos de **todas as unidades da população** de interesse.
- **Pesquisas amostrais**
  - Dados obtidos de **amostras de unidades** da população de interesse.
- **Registros administrativos**
  - Dados obtidos para fins administrativos, e depois usados para fins estatísticos.

Processo de Investigação (Geração de Conhecimento)

Metodologia Estatística

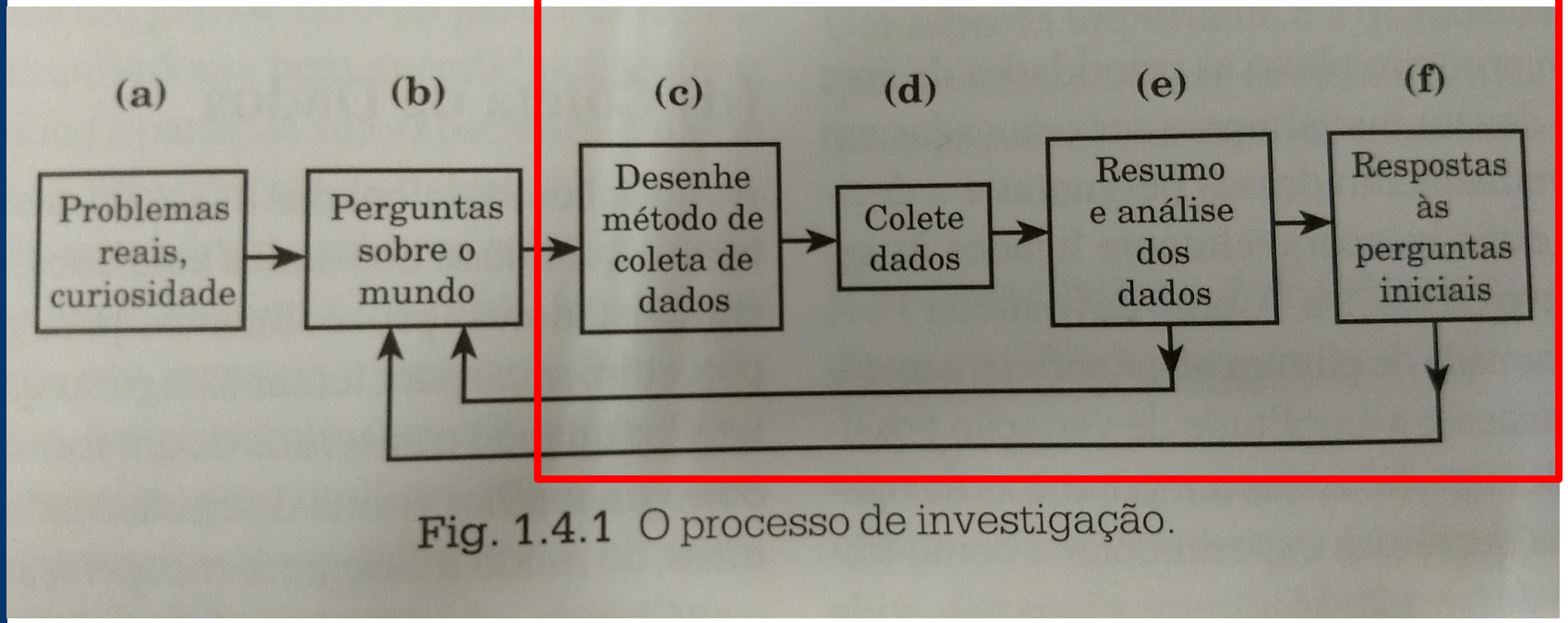


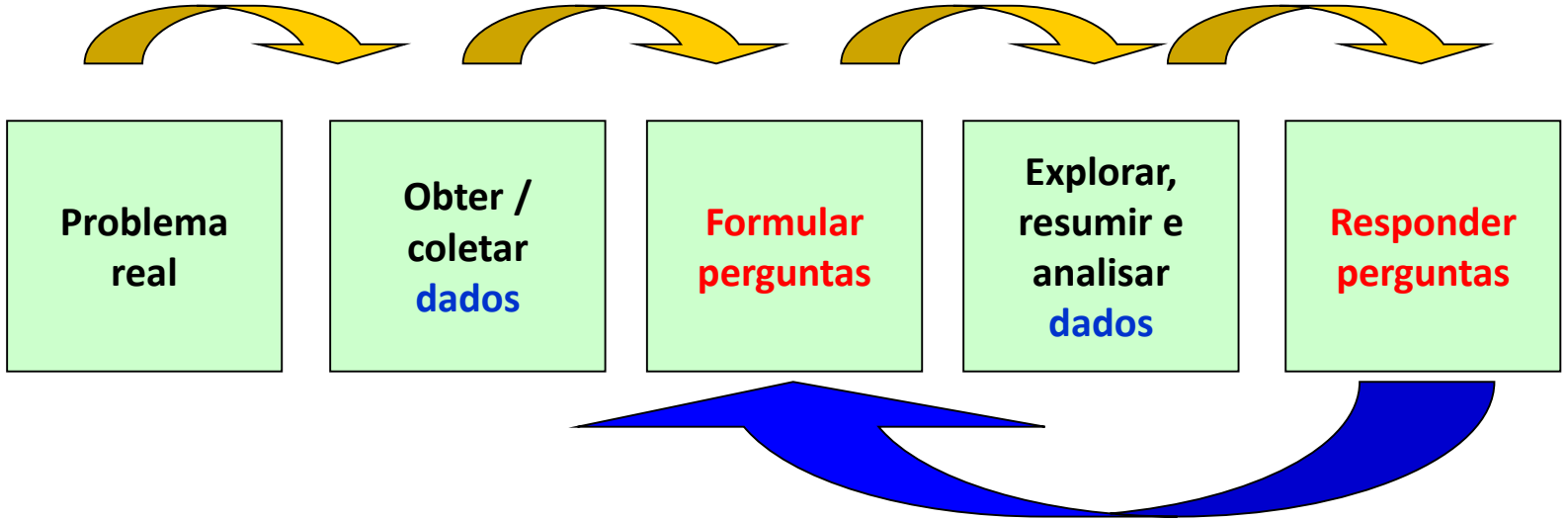
Fig. 1.4.1 O processo de investigação.

Fonte: (WILD; SEBER, 2004, p. 17).

Big Data - Novas Fontes de Dados

- **Tipos de fontes**
  - Transações – p. ex. Nota Fiscal Eletrônica;
  - Comunicações e mensagens;
  - Imagens;
  - Buscas; etc. etc.
- **Características**
  - Grande **Volume**;
  - Grande **Variedade**;
  - Grande **Velocidade**;
  - Dados '**não planejados**' / 'não estruturados' – vem recebendo o nome de '**dados orgânicos**'.

Geração de Conhecimento na Era do '*Big Data*'



Oportunidades para formular **novas perguntas**, mas perguntas formuladas **após** coletar / obter / acessar dados.

Dados podem não ajudar a responder bem **perguntas existentes**.

Big Data nas Estatísticas Oficiais

Muitas ideias de tentar aproveitar novas fontes para **substituir** ou **complementar** fontes tradicionais.

Vários projetos de **experimentação** de uso de novas fontes em andamento no mundo.

Agentes (privados e públicos) pressionam pela **adoção rápida** das novas fontes, sem talvez apreciar com cuidado os riscos envolvidos.

Protocolos e modelos existentes e bem desenvolvidos para **avaliar qualidade** no contexto das fontes 'tradicionais'.

Mesmo não é verdade para novas fontes do '*big data*'.

Big Data – Problemas com Qualidade

- **Variabilidade ou Volatilidade**
  - Inconsistências e/ou instabilidade dos dados ao longo do tempo ou do espaço.
- **Veracidade**
  - Capacidade de confiar que os dados são **acurados** e **completos**.
- **Complexidade**
  - Necessidade de **combinar dados** de múltiplas fontes.
- **Acessibilidade**
  - Necessidade de assegurar que os dados estão e continuarão disponíveis.

Big Data – Problemas com Qualidade

- Ter **mais dados não** necessariamente **implica** ter **dados** bons ou **melhores**!
- Muitas das fontes de dados disponíveis **carecem da qualidade** requerida para seu **uso seguro**.
- Os **desafios** da qualidade **são** ainda **maiores** com '*Big Data*'.

Qualidade de Dados e de Pesquisas

- Objetivo a ser perseguido / alcançado.
- Qualidade do dado deriva da **qualidade da fonte** / **instrumento de medida** / **pesquisa**.
- Qualidade é **atributo desejável** de dados e de pesquisas.
- Conceito vago: **o que é qualidade de um dado**?
- Precisa ser definida, de modo que possa ser 'planejada', 'medida' e 'avaliada'.

Sistemas de Referência para Qualidade

- Organizações importantes têm investido em definir **sistemas de referência** para qualidade de suas pesquisas.
- *'Quality frameworks'*:
  - *US Office of Management and Budget* (2006);
  - *Statistics Canada* (2009);
  - *International Monetary Fund* (2012);
  - OECD (2012);
  - UN (2012);
  - IBGE (2013).

Sistema de Referência para Qualidade da OECD

Dimensão de qualidade	Descrição
Relevância	Estatísticas são relevantes se satisfazem necessidades dos usuários.
Acurácia	Proximidade entre o valor final da estatística e o verdadeiro, mas desconhecido, valor populacional.
Credibilidade	Grau de confiança que os usuários têm nas Estatísticas com base na imagem do produtor.
Atualidade	Intervalo de tempo entre a disponibilização do dado e o evento ou fenômeno que o dado descreve.
Acessibilidade	Quão facilmente os dados podem ser localizados e acessados pelos usuários.
Interpretabilidade	Facilidade com que os usuários dos dados podem entender, usar e analisar apropriadamente os dados.
Coerência	Reflete o grau com que diferentes dados são logicamente conectados e mutuamente consistentes.
Custo-benefício	Uma medida dos custos e carga dos respondentes relativamente ao valor dos resultados.

OECD Statistics Directorate (2012).

Qualidade nos INEs

- Discussão internacional e muitos avanços sobre o gerenciamento e a avaliação da qualidade nos INEs.
- Eventos internacionais sobre o tema:
  - 2000 – *Statistical Quality Seminar* – Coréia;
  - *European Conference on Quality in Official Statistics*:
    - Q2001 – Stockholm / Suécia;
    - Q2004 – Mainz /Alemanha;
    - Q2006 – Cardiff / Reino Unido;
    - Q2008 – Roma / Itália.
    - ...
    - Q2016 – Madrid / Espanha (http://www.q2016.es/)
    - Q2018 – Kraków / Polônia (https://www.q2018.pl/)
- Rica bibliografia sobre *"boas práticas"* .

Qualidade de Dados e de Pesquisas

Duas abordagens / trajetórias complementares (Lyberg, 2012):

- Modelos para **Erro Total da Pesquisa;**
- **Gerenciamento de processos** e da qualidade em pesquisas ➔ melhoramento contínuo da qualidade.

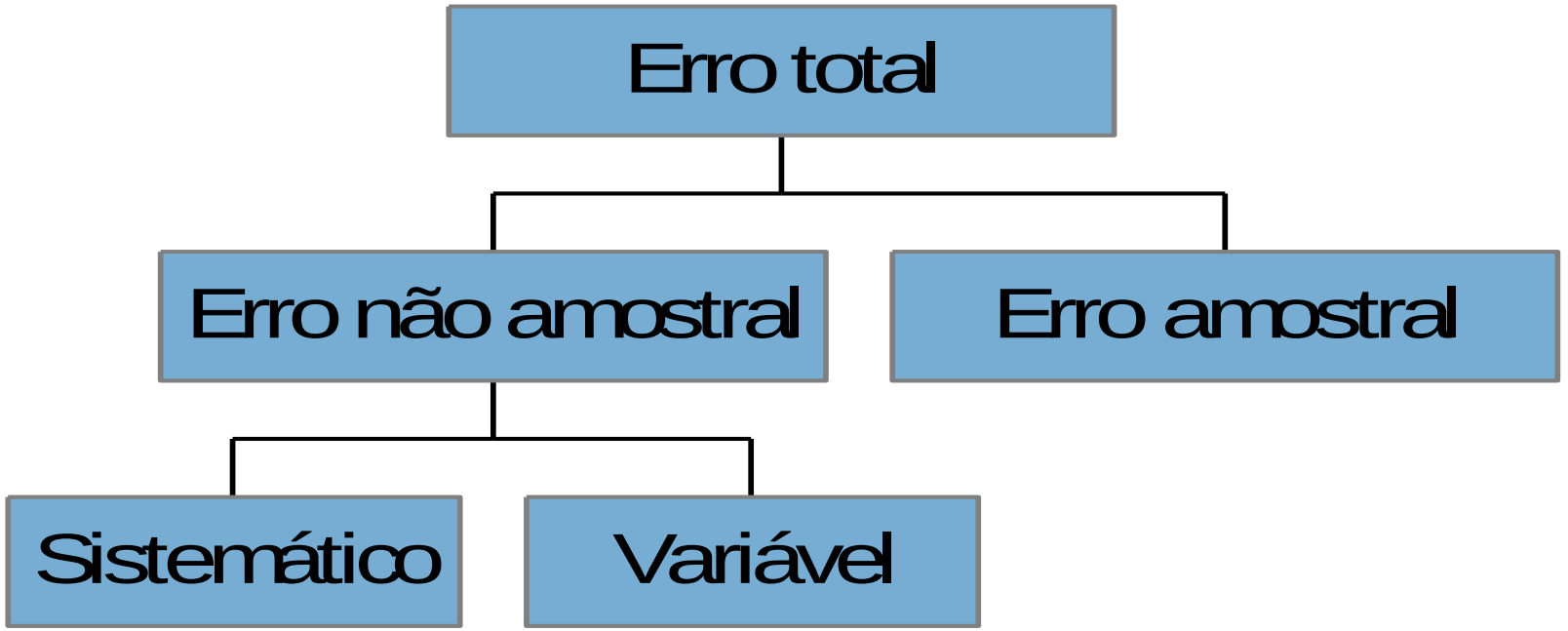
Modelos para Erro Total da Pesquisa

- Quatro princípios guiam o planejamento, implementação, avaliação e análise de pesquisas.
- É importante:
  - **Considerar** todas as fontes de erro conhecidas;
  - **Monitorar** as principais fontes de erro durante a implementação da pesquisa;
  - **Avaliar** periodicamente as principais fontes de erro e combinações destas após concluir a pesquisa; e
  - **Estudar** os **efeitos dos erros** nas análises da pesquisa.

Erros em Pesquisas

“Erro” de Estimativas

Erro = Estimativa - Valor Verdadeiro



Fonte: *United Nations* (2005).

Erro amostral

- Mais fácil de controlar.
- **Vício** (erro sistemático) pode ser evitado → usar métodos de **amostragem probabilística**.
- **Plano amostral**, **tamanho da amostra** e **estimador** definidos para tornar **erro amostral variável** tão pequeno quanto seja necessário.
- Algumas vezes, foco exclusivo no controle do erro amostral em vez do erro total pode ser problema.
  - Ex. Amostras 'grandes demais'.

Erro amostral

- Mais fácil de controlar.
- **Vício** (erro sistemático) pode ser evitado → usar métodos de **amostragem probabilística**.
- **Plano amostral**, **tamanho da amostra** e **estimador** definidos para tornar **erro amostral variável** tão pequeno quanto seja necessário.
- Algumas vezes, foco exclusivo no controle do erro amostral em vez do erro total pode ser problema.
  - Ex. Amostras 'grandes demais'.
- Com '*Big Data*', pode não mais haver erro amostral em muitas aplicações!

Erros não amostrais

- Duas classes amplas de **erros não amostrais**.
- Erros devidos à '**não observação**':
  - Cobertura (cadastro, população);
  - Não resposta (coleta).
- Erros nas **observações**:
  - Especificação;
  - Medida;
  - Processamento e estimação.

Erros não amostrais

- Duas classes amplas de **erros não amostrais**.
- Erros devidos à '**não observação**':
  - Cobertura (cadastro, população);
  - Não resposta (coleta).
- Erros nas **observações**:
  - Especificação;
  - Medida;
  - Processamento e estimação.
- Com '*Big Data*', erros não amostrais dominam! Pior: podem não desaparecer com n grande!

Big Data – Uma Rápida Análise (Meng, 2018)

- **Registro Administrativo** cobre uma fração c=(m/N) da população.
- **Amostra Aleatória Simples** de fração f=n/N da população.
- f << c (Amostra é muito menor que o Registro).
- Quão grande deve ser n (ou f) antes que um estimador baseado na AAS domine o baseado somente no Registro em termos de EQM?

Qualidade dos Dados de Fontes Orgânicas e Registros

Tabela 1 - Tamanhos de amostra necessários para que uma AAS tenha EQM menor que um Registro Administrativo ao estimar a média populacional

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

Resultado **surpreendente**, não é??

Big Data – Uma Rápida Análise (Meng, 2018)

- U = {1, 2, ..., N} (População).
- $\mathcal{A} \subset U$ é subconjunto de n unidades selecionadas de U por AAS.
- $\mathcal{R} \subset U$ é subconjunto de m unidades de U cobertas pelo Registro Administrativo.
- **Alvo:** estimação da **média populacional**

$$\bar{Y} = \frac{1}{N} \sum_{k \in U} y_k$$

Big Data – Uma Rápida Análise (Meng, 2018)

- Estimadores para média populacional
- Usando **Registro Administrativo**

$$\bar{y}_{\mathcal{R}} = \frac{1}{m}\sum_{k\in\mathcal{R}} y_k = \frac{1}{m}\sum_{k\in U} R_k y_k$$

- Usando **AAS**

$$\bar{y} = \frac{1}{n}\sum_{k\in\mathcal{A}} y_k = \frac{1}{n}\sum_{k\in U} A_k y_k$$

Big Data – Uma Rápida Análise (Meng, 2018)

Sob amostragem de populações finitas, únicas quantidades aleatórias nos estimadores são indicadores de pertinência (à amostra ou ao registro).

No caso da **amostra**, o mecanismo de inclusão é aleatório e induzido pelo pesquisador.

No caso do **registro**, o mecanismo está fora do controle do pesquisador.

Big Data – Uma Rápida Análise (Meng, 2018)

Erro total sob AAS:

$$EQM_n(\bar{y}) = V_n(\bar{y}) = \frac{1-f}{n}\frac{N}{N-1}\sigma_y^2$$

Erro total sob Registro:

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

Big Data – Uma Rápida Análise (Meng, 2018)

Para garantir que o erro é menor usando AAS:

$$n \geq \left(\frac{m}{N-m}\right)\frac{1}{E_R\left(\rho_{R,y}^2\right)}$$

Considerando N = 200 milhões, $\rho_{R,y}$ = 0,05, e c = 50%, o EQM de uma AAS com n = 400 seria **menor ou igual** ao obtido usando o Registro Administrativo com m = 100 milhões!

Big Data e Qualidade

Resultado é claro: '**fé cega**' no Registro pode ser perigosa e levar a resultados de qualidade pior que a viabilizada por pequenas amostras.

Mensagem: novas fontes precisam ser avaliadas com mesmo rigor que amostras e censos...

Big Data - Novas Fontes de Dados

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

Ciência Estatística

Oferece soluções para os problemas de investigação e geração de dados e conhecimento mediante:

- Cuidadoso **planejamento** e **realização** de operações de **obtenção de dados e medidas** sobre os fenômenos de interesse (**levantamentos**);

Ciência Estatística

Oferece soluções para os problemas de investigação e geração de conhecimento mediante:

- Cuidadoso **planejamento** e **realização** de operações de **obtenção de dados e medidas** sobre os fenômenos de interesse (**levantamentos**);
- **Análise exploratória** e **tratamento preparatório** dos dados coletados (**observações**);

Ciência Estatística

Oferece soluções para os problemas de investigação e geração de conhecimento mediante:

- Cuidadoso **planejamento** e **realização** de operações de **obtenção de dados e medidas** sobre os fenômenos de interesse (**levantamentos**);
- **Análise exploratória** e **tratamento preparatório** dos dados coletados (**observações**);
- **Formulação e ajuste** de **modelos estatísticos** para **descrever os dados de forma sintética**, para obtenção de **respostas** às perguntas de interesse (**inferência**);

Ciência Estatística

Oferece soluções para os problemas de investigação e geração de conhecimento mediante:

- Cuidadoso **planejamento** e **realização** de operações de **obtenção de dados e medidas** sobre os fenômenos de interesse (**levantamentos**);
- **Análise exploratória** e **tratamento preparatório** dos dados coletados (**observações**);
- **Formulação e ajuste** de **modelos estatísticos** para **descrever os dados de forma sintética**, para obtenção de **respostas** às perguntas de interesse (**inferência**);
- **Apresentação (visualização)** das respostas e resumos de padrões revelados pelos dados.

Obtenção de Dados

Métodos para cuidadoso planejamento e realização de levantamentos custo-efetivos para obtenção de dados:

- Amostragem;
- Planejamento e condução de experimentos;
- Planejamento e condução de estudos observacionais;
- Protocolos de mensuração (questionários, instrumentos, coleta, etc.);
- Protocolos de verificação, limpeza, armazenamento e compartilhamento de dados.

Análise / Descoberta

Métodos para **análises exploratória** e **confirmatória** de dados:

- Análise exploratória de dados;
- Formulação e teste de hipóteses;
- Formulação, ajuste, seleção, diagnóstico e interpretação de modelos;
- Sumarização, visualização, e apresentação de dados; e mais recentemente
- Mineração de dados, aprendizado por máquinas, etc.

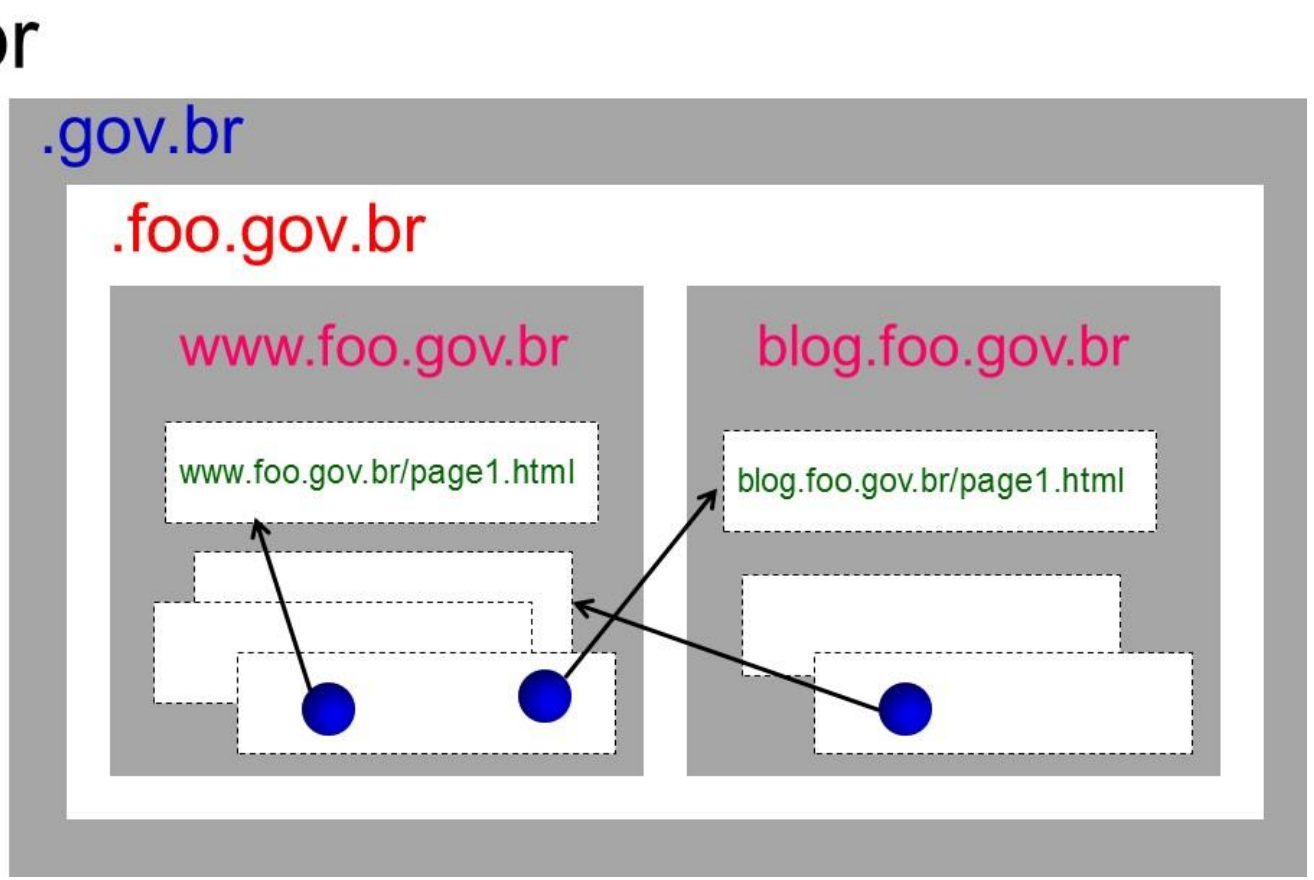
Ajudando a Preencher as Lacunas

Métodos estatísticos têm papel predominante na busca pelo preenchimento das lacunas de dados.

- **Amostragem** é largamente usada para obter dados de forma rápida e custo-efetiva.
- Ideias básicas da **amostragem probabilística** desenvolvidas e aperfeiçoadas ao longo do Século XX.
- **Métodos para estimação em pequenos domínios** agora mais e mais usados para obter estimativas para domínios com pequenas amostras.
- **Ligação de registros**, **combinação de dados** e **meta-análise** são outras abordagens disponíveis para **combinar dados e resultados** de diferentes estudos ou fontes.

Amostragem – Um Exemplo

Interesse: pesquisa sobre domínios da internet no Brasil.



Amostragem – Um Exemplo

- Projeto piloto: ***".gov.br"***
- **Censo** dos sítios e páginas com domínios registrados a pedido do setor público brasileiro.
- 'Robô' visitou **todas** os sítios e páginas encontrados ou conectados.
- Conjunto inicial continha cerca de **12 mil domínios** registrados.
- Coleta dos dados durou cerca de **3 semanas**.

Amostragem – Um Exemplo

- Desafio: ***".com.br"***
- Na ocasião, **2,5 milhões** de domínios registrados.
- Censo seria inviável com tecnologia da ocasião → coleta dos dados duraria **≈ 11 anos**.
- Abordagem empregada: amostra estratificada e conglomerada de domínios (**n ≈ 4.000**).
- Coleta concluída em **45 dias**.

Resumindo

Metodologia Estatística é **pilar essencial** para promoção da qualidade de dados e pesquisas.

- Abordagem do **Erro Total da Pesquisa** oferece modelo para guiar a mensuração e busca da qualidade.
- Abordagem do **Gerenciamento do Processo de Pesquisa** permite atuar na melhoria 'contínua' da qualidade.

➔ Abordagens são complementares, não excludentes.

Obrigado por sua atenção.

www.ence.ibge.gov.br
pedro-luis.silva@ibge.gov.br

Referências

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. Lyberg, Lars. 2012. “Survey Quality.” *Survey Methodology* 38 (2): 107–130.
5. Meng, X. L. (2018). Statistical paradises and paradoxes in Big Data (I): law of large populations, big data paradox, and 2016 US Presidential Election. Submitted to Annals of Applied Statistics.
6. Office of Management and Budget. 2006. *Standards and Guidelines for Statistical Surveys*. *Federal Register*. Washington, DC.
7. Statistics Canada (2009). Statistics Canada Quality Guidelines, fifth edition. Ottawa, Canada: Statistics Canada.

Referências

8. Statistics Directorate, OECD. 2012. *Quality Framework and Guidelines for OECD Statistical Activities*.
9. 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.
10. 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.
11. United Nations. 2012. Guidelines For The Template For A Generic National Quality Assurance Framework (NQAF). http://unstats.un.org/unsd/dnss/qualityNQAF/nqaf.aspx
12. WILD, Christopher J; SEBER, George A F. Encontros com o acaso: um primeiro curso de análise de dados e inferência. Rio de Janeiro: LTC - Livros Técnicos e Científicos Editora S.A., 2004. 411 p.