Motivação e Importância da Estatística Descritiva

Prof. Walmes M. Zeviani

Departamento de Estatística
Universidade Federal do Paraná

Conceitos fundamentais

Estatística Descritiva vs Inferencial

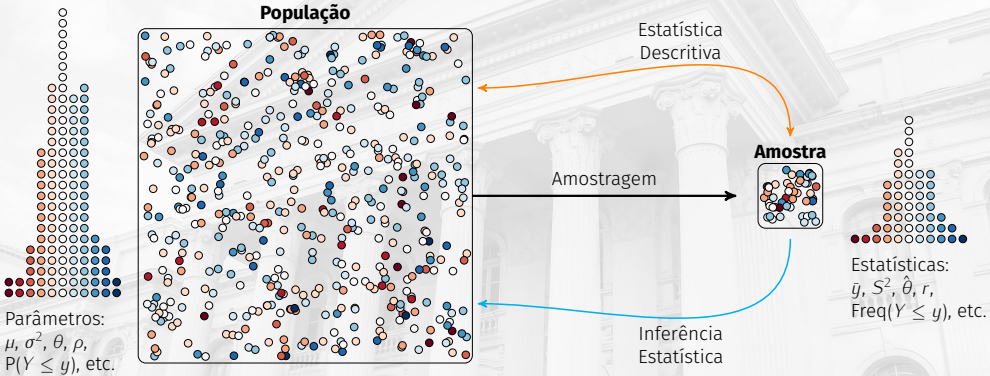

Figura 1. Representação esquemática do funcionamento da estatística descritiva e estatística inferencial.

Estatística Descritiva

A estatística descritiva emprega métodos numéricos e gráficos para investigar padrões em um conjunto de dados, resumir as informações reveladas e apresentá-las de uma forma apropriada.

Estatística Inferencial

A Estatística Inferencial utiliza dados de uma amostra para fazer estimativas, previsões, decisões ou outras generalizações sobre um grande conjunto de dados (a população).

- Diagnóstica ou confirmatória.
- Preditiva.
- Prescritiva.

População

Uma população é um conjunto de unidades amostrais (e.g. pessoas, objetos, transações ou eventos) que estamos interessados em estudar.

Amostra

Uma amostra é um subconjunto das unidades amostrais de uma população.

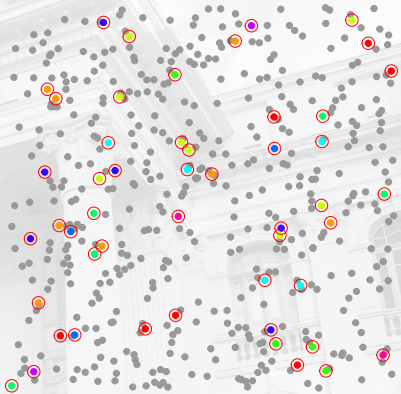

Figura 2. População e amostra.

Unidade amostral (ou experimental)

Uma unidade amostral é um objeto (e.g. pessoa, coisa, transação ou evento) sobre o qual coletamos dados.

Variável ou característica

Uma variável é uma característica ou propriedade de uma unidade amostral individual.

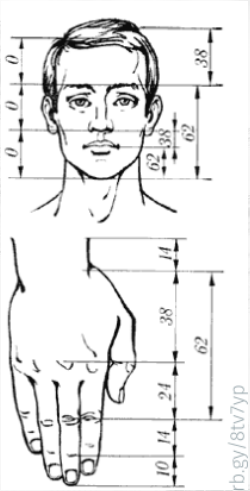

Figura 3. Medidas no corpo humano que apresentam a razão áurea. Fonte: https://rb.gy/8tv7yp.

Motivação e importância da Estatística Descritiva

Qual o comportamento destes dados?

Tabela 1. Os 4 pares de variáveis do quarteto de Ancombe.

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

A visualização de dados

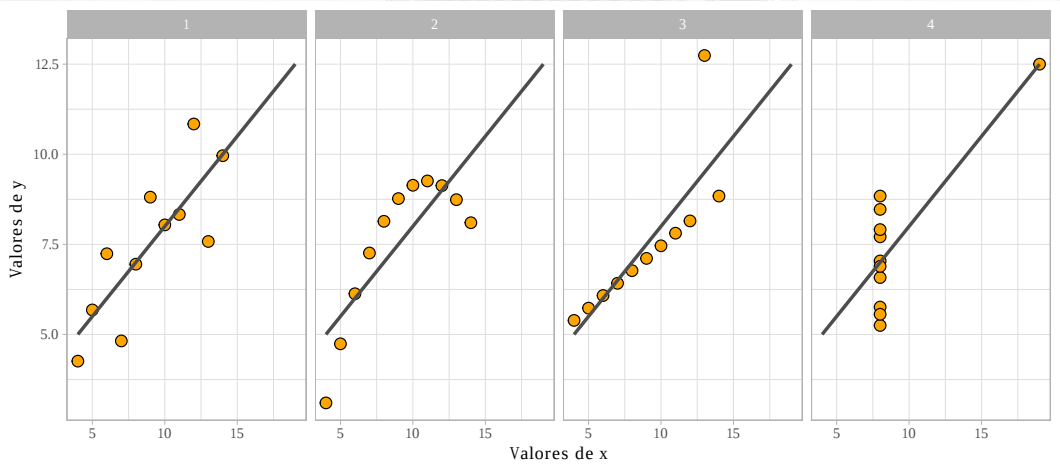

Figura 4. O quarteto de Anscombe em um diagrama de dispersão.

Resumos numéricos

Expressão da equação da reta ajustada

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

A notação será trabalhada nos próximos vídeos.

Tabela 2. Resumo do ajuste da regressão linear simples com cada par de variáveis do quarteto de Anscombe.

	$\overline{x}$	$\overline{y}$	correl.	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2	Valor p
1	9.0	7.5	0.82	3.00	0.50	0.67	0.0022
2	9.0	7.5	0.82	3.00	0.50	0.67	0.0022
3	9.0	7.5	0.82	3.00	0.50	0.67	0.0022
4	9.0	7.5	0.82	3.00	0.50	0.67	0.0022

Por que saber métodos de análise descritiva?

- Porque “correr” o olho sobre a planilha de dados é **impraticável**.
- Porque abordagens subjetivas não são **escaláveis**.
- Para dispor de formas apropriadas de **síntese**.
- Para não fazer julgamentos ocasionados por **problemas** nos dados.
- Para **interpretar** corretamente a informação contida nos dados.
- Para tomar decisões **seguras**.
- Para fazer inferência estatística de forma **qualificada**.

O que vamos aprender?

- Com que **tipo de informação** estamos lidando?
- Como **sintetizar** números que representem as **tendências**, **variabilidade**, a **relação** entre variáveis e demais propriedades?
- Como verificar existência de **anomalias** ou inconsistências?
- Como **apresentar** adequadamente a informação contida nos dados?

Procedimento para a análise de dados

O que é feito e pra que serve?

O que é feito?

- Análise de dados históricos.
- O que aconteceu/está acontecendo?
- Explorar e descrever os dados brutos.
- Ter impressões preliminares.

Qual a utilidade?

- Recursos para comunicação: *data storytelling*.
- Auxilia no pré-processamento e curadoria de dados.
- Determinante para o processo de inferência estatística.

Aspectos da qualidade dos dados

- **Validade**: grau de **conformidade** com o mundo real.
  - Fora de escala: pessoa com 180 m de altura.
  - Fora do conjunto: tipo sanguíneo = vermelho.
  - Fora de lógica: data de alta médica antes da internação.
- **Acurácia**: valores próximos dos valores **verdadeiros**.
  - Dados de sensores, avaliações sensoriais.

Fonte: https://rb.gy/7caksz.

- **Completude**: a quantidade de valores **preenchidos** frente ao esperado.
  - Valores ausentes, suas razões e implicações.
- **Uniformidade**: dados expressos com os mesmos **padrões**.
  - Pressão em *psi*, *bar* ou *atm*?
  - Data no formato *dd/mm/yyyy* ou *yyyy-mm-dd*?
- **Unicidade**: se não existem registros **duplicados**.

Processos de análise de dados

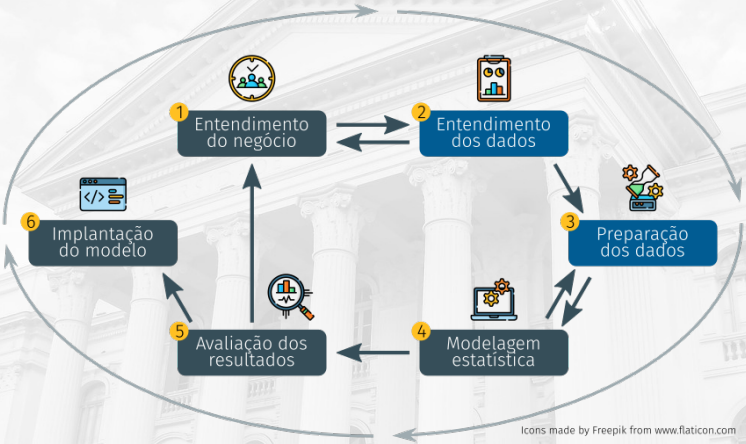

Figura 5. O CRISP-DM (Cross Industry Standard Process for Data Mining) estabelece etapas para a análise de dados. As caixas em azul são as etapas que envolvem análise exploratória de dados.

Considerações finais

Considerações finais

Revisão

- Importância da Estatística Descritiva.
- Para que serve.
- Aspectos da qualidade dos dados.

Os números têm uma importante história para contar. Eles dependem de você dar-lhes uma clara e convincente voz.

– StephenFew