Estatítica Descritiva e Exploratória

Gledson Luiz Picharski e
Wanderson Rodrigo Rocha

Universidade Federal do Paraná

3 de Abril de 2008

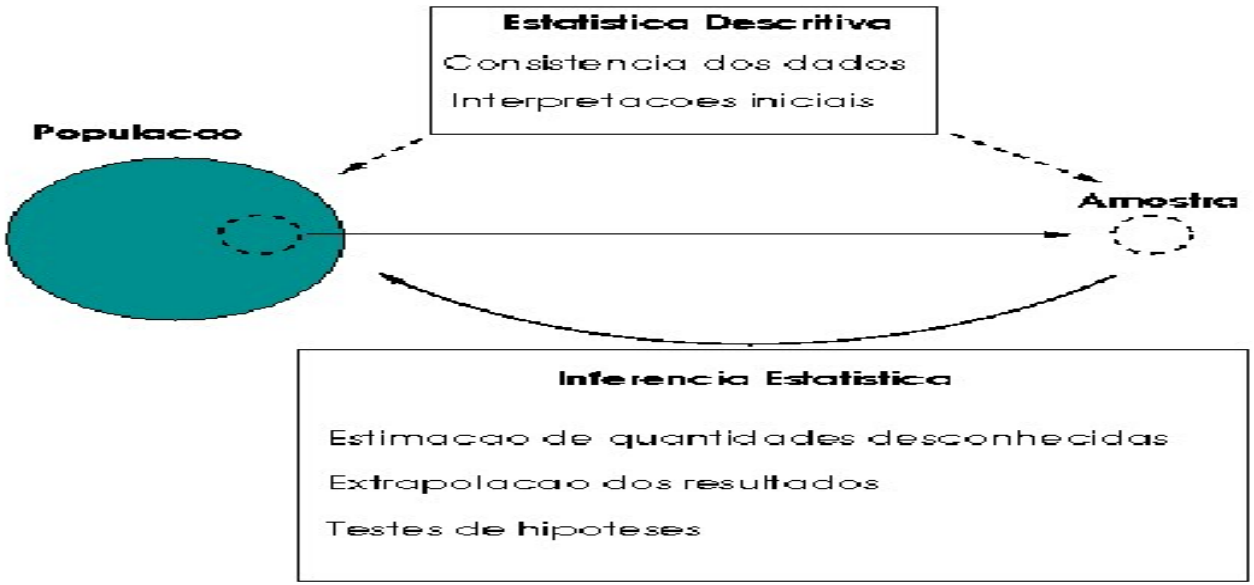
Estatística Descritiva e exploratória

O que é Estatística?

A Estatística é um conjunto de métodos de coleta e descrição de dados, e então a verificação da força da evidência nos dados pró ou contra certas idéias científicas.

- **Estatística descritiva:** conjunto de técnicas destinadas a descrever e resumir dados.
- **Probabilidade:** teoria matemática utilizada para se estudar a incerteza oriunda de fenômenos de caráter aleatório.
- **Inferência estatística:** técnicas que possibilitam a extrapolação, a um grande conjunto de dados (população), dos resultados obtidos a partir de um subconjunto de valores (amostra). Note que se tivermos acesso a todos os elementos que desejamos estudar, não é necessário o uso de técnicas de inferência estatística.

População x amostra



Tipos de amostragem

- **casual simples**: Selecionamos ao acaso, com ou sem reposição, os itens da população que farão parte da amostra.

Se houver informações adicionais a respeito da população de interesse, podemos utilizar outros esquemas de amostragem mais sofisticados.

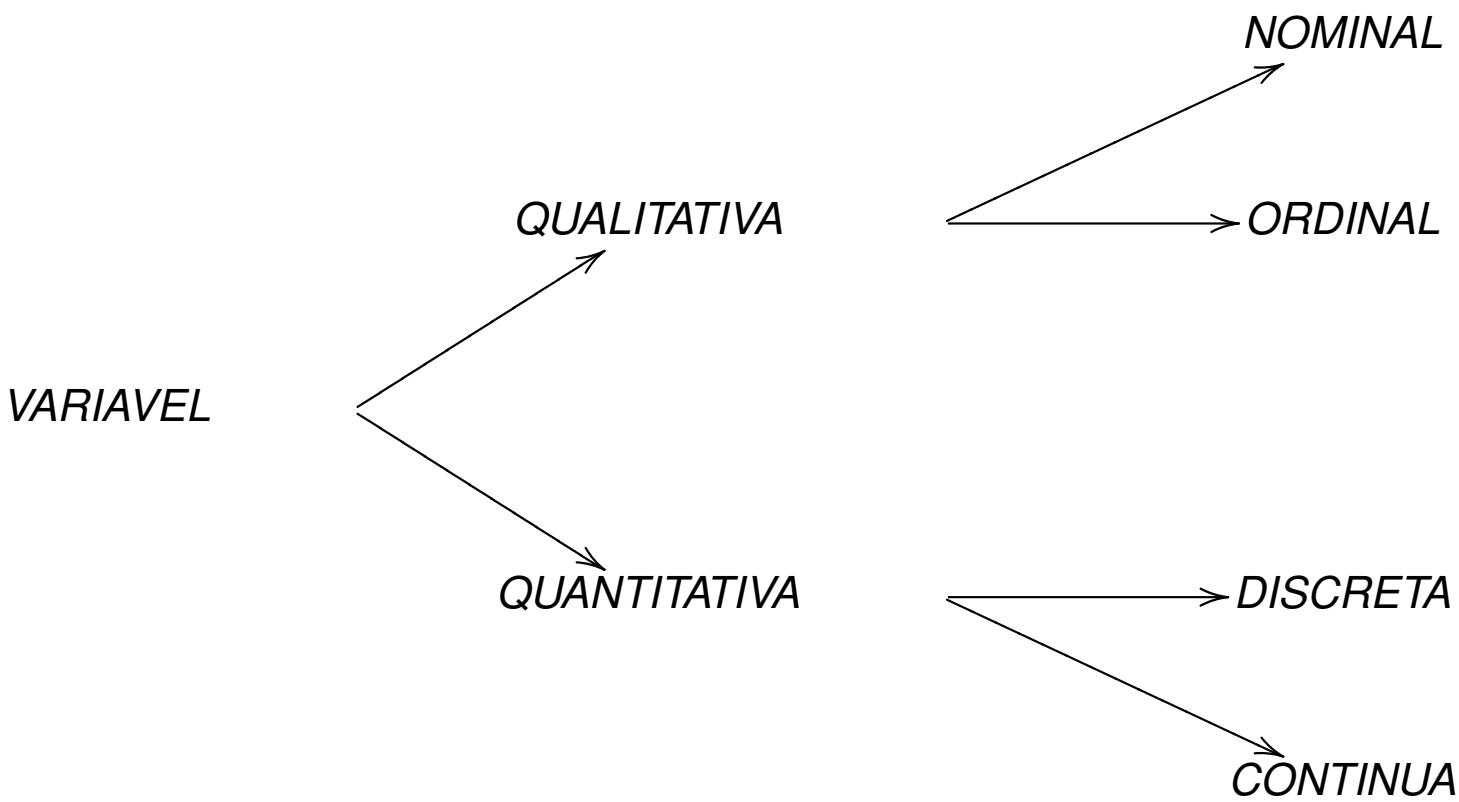
- **Amostragem estratificada:** Por exemplo, se numa cidade, tivermos mais mulheres do que homens, podemos selecionar um certo número de indivíduos entre as mulheres e outro número entre homens.
- **Amostragem sistemática:** Pode existir uma relação numerada dos itens da população (uma lista de referência) que nos permite selecionar os indivíduos de forma pré-determinada, por ex de 8 em 8 ou de 10 em 10.

Organização dos dados

Suponhamos que um questionário seja aplicado a alunos da Universidade fornecendo as seguintes informações:

Id	Turma	Sexo	Idade	Alt	Peso	Filhos	Fuma
1	A	F	17	1.60	60.50	2	NAO
2	A	F	18	1.69	55.00	1	NAO
3	A	M	18	1.85	72.80	2	NAO
4	A	M	25	1.85	80.90	2	NAO
5	A	F	19	1.58	55.00	1	NAO
6	A	M	19	1.76	60.00	3	NAO

Tipos de Variáveis



Exemplos de tabelas

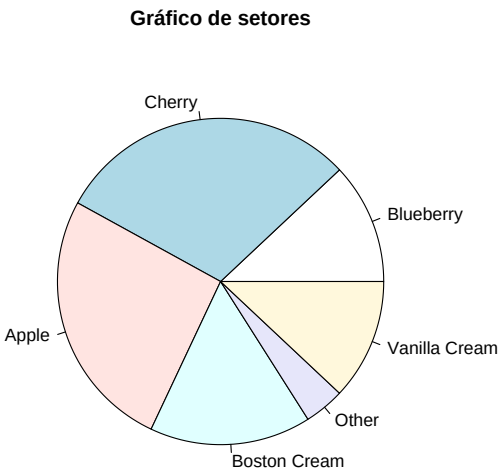
Idade	ni	fi	fac
17	9	0,18	0,18
18	22	0,44	0,62
19	7	0,14	0,76
20	4	0,08	0,84
21	3	0,06	0,90
22	0	0	0,90
23	2	0,04	0,94
24	1	0,02	0,96
25	2	0,04	1,00
total	n=50	1	

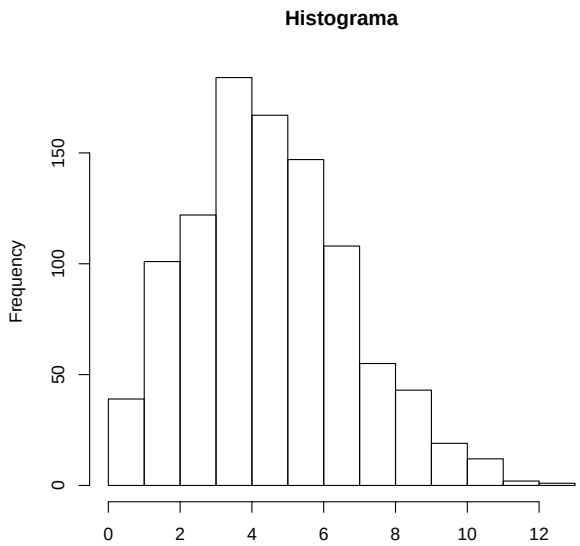
Tabela: Frequências para Idade

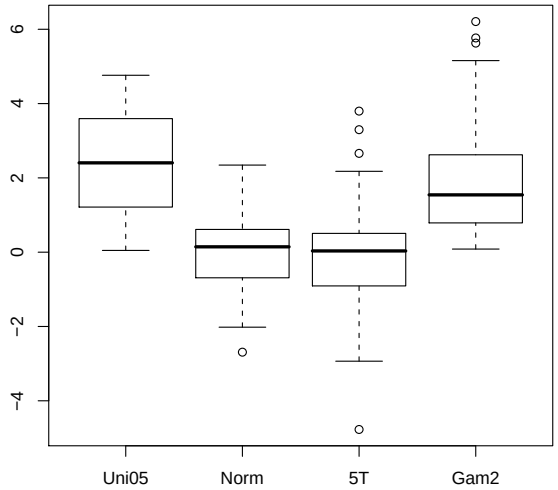
Peso	ni	fi	fac
40,0 —- 50,0	8	0,16	0,16
50,0 —- 60,0	22	0,44	0,60
60,0 —- 70,0	8	0,16	0,76
70,0 —- 80,0	6	0,12	0,88
80,0 —- 90,0	5	0,10	0,98
90,0 —- 100,0	1	0,02	1,00
total	50	1	

Tabela: Freqüências para Peso

Exemplos de Gráficos







Estatística Descritiva e exploratória

Introdução

Medidas resumo são técnicas que nos auxiliam a sumarizar informações disponíveis sobre o comportamento de uma variável.

Medidas de Posição para um conjunto de dados

- **Média**

$$\bar{x}_{obs} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Soma dos valores que a variável assume dividida pelo número de observações.

- **Média para conjunto de dados organizados em tabela de Frequência**

$$\bar{x}_{obs} = \frac{n_1 x_1 + n_2 x_2 + \ldots + n_k x_k}{n_1 + \ldots + n_k} = \frac{\sum_{i=1}^{k} n_i x_i}{n}$$

Média dos k diferentes valores ponderada pelas respectivas frequências de ocorrências.

- **Mediana**
  Representada por:

  $$md_{obs}$$

  É o valor que ocupa a posição central dos dados ordenados. Coloca-se os dados em ordem crescente, se o número de elementos no conjunto de dados é ímpar selecionamos o dado central. No caso em que o número de elementos for par a mediana será a média dos dois valores que ocupam a posição central.
- **Moda**
  Representada por:

  $$mo_{obs}$$

  É dada pelo valor mais frequente do conjunto de dados.

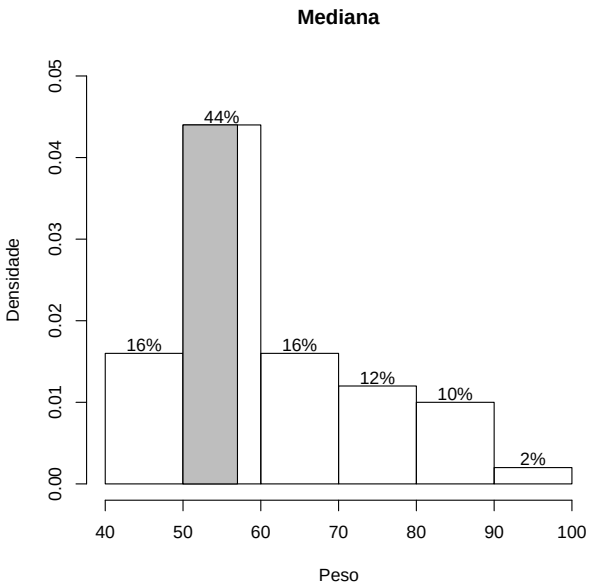
Quartis

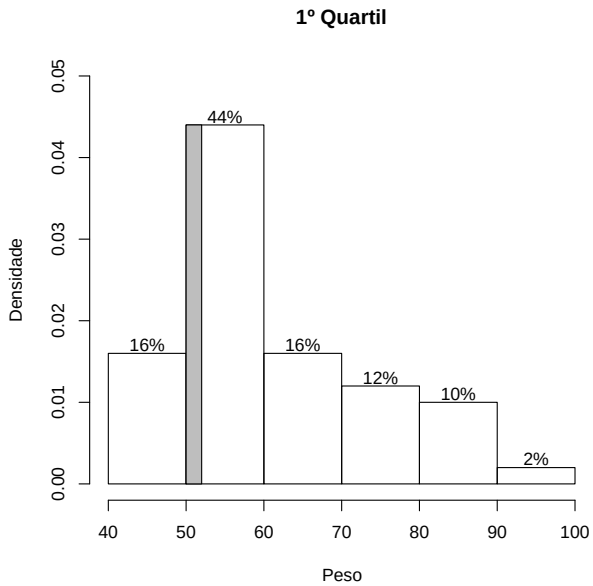
Definições:

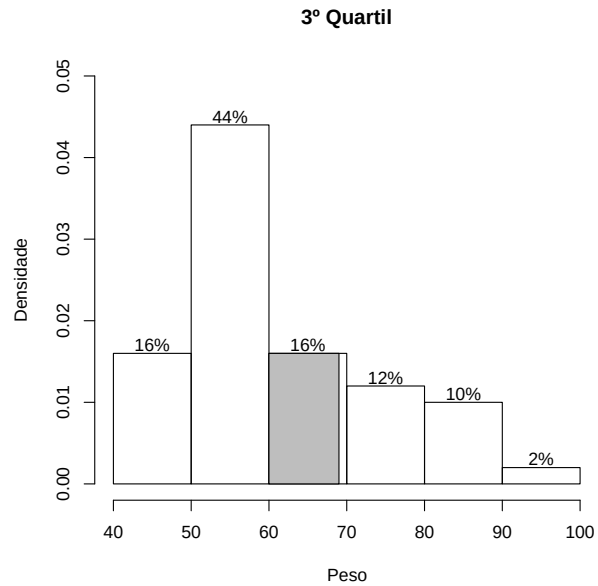
- Q_1: Valor que deixa 25% das observações ordenadas abaixo dele.
- Q_2: Valor coincidente com a mediana.
- Q_3: Valor que deixa 75% das observações ordenadas abaixo dele.
- **Amplitude Interquartílica:** É a diferença entre Q_3 - Q_1.

Exemplo

Calcular Q_1, md_{obs} e Q_3 a seguir:







Assim Q_1 seria dado por: $\frac{Q_1 - 50}{0.09} = \frac{60-50}{0.44} \Rightarrow = Q_1 = 52,05.$

Medidas de posição para variáveia aleatórias Discretas

- **Valor Esperado** Valor Esperado de uma variável aleatória X é dada pela expressão:

$$E(X) = \sum_{i=1}^{k} x_i p_i$$

  Outra notação possível seria μ_X

- **Mediana**

  A mediana é o valor Md que satisfaz ás seguintes condições: $P(X \geq Md) >= \frac{1}{2}$ e $P(X \leq Md) >= \frac{1}{2}$ Em algumas situações, as desigualdades serão satisfeitas por qualquer valor num intervalo.

- **Moda** A moda Mo será dada pelo valor que apresentar a maior probabilidade de ocorrência.

Exemplo aplicado a uma Variável Aleatória X discreta

Função densidade:

X	2	5	8	15	20
p	0.1	0.3	0.2	0.2	0.2

Analizando a variável teremos:

$\mu = 10,3$; Md = 8; Mo = 5.

Medidas de dispersão

Apesar de muitas vezes as medidas de tendência central nos fornecerem uma idéia do comportamento da variável, precisamos de ferramentas que quantifiquem a dispersão dos valores que a variável assume. Esse é o papel das medidas de dispersão.

- **Amplitude dos dados** A amplitude dos dados será dada pela diferença entre o maior e o menor valor do conjunto de dados, representada por Δ.
- **Desvio Médio** Calcula-se o somatório dos desvios dos valores absolutos em relação á média.

$$\frac{1}{n}\Sigma_{i=1}^{n}|x_i - \bar{x}_{obs}|$$

Embora seja muito útil, torna-se complicada pelas propriedades da função módulo.

- **Variância e desvio-padrão em um conjunto de dados**. A variância referente a uma variável aleatória X em um conjunto de dados é dada por:

$$var_{obs} = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x}_{obs})^2$$

ou pela expressão alternativa:

$$var_{obs} = \frac{1}{n}\Sigma_{i=1}^{n}x_i^2 - \bar{x}_{obs}^2$$

- **Desvio-padrão** Usado para representar a variabilidade nas mesmas unidades de medida dos nossos dados. É dado pela raiz quadrada da variância: $\sqrt{var_{obs}}$

Variância de uma va. Discreta

Se X é uma variável aleatória com $P(X_i = x_i)$ $i = 1, 2, \ldots, k$ e média μ. Então sua variância é dada pela expressão:

$$Var(X) = \Sigma_{i=1}^{k}(x_i - \mu)^2 p_i$$

Ou pela expressão:

$$Var(X) = E(X^2) - \mu^2$$

Exemplo de cálculo da Variância de va. Discreta

Admita que X possua a seguinte função de probabilidade:

X	4	5	6
p	0.3	0.4	0.3

Então teremos $\mu = 5$, $E(X^2) = 25.6$ e $\mu^2 = 25$. A partir daí concluiremos que $var_{obs} = 0.6$

Estatística Descritiva e exploratória

- **Fenômeno Aleatório**
  São fenômenos não deterministicos, eventos que não podem ser previstos com absoluta certeza.

- **Fenômeno Aleatório**
  São fenômenos não deterministicos, eventos que não podem ser previstos com absoluta certeza.
- **Espaço amostral -** Ω
  Conjunto de todos os eventos possiveis para um determinado experimento.

Axiomas de Probabilidade

i) $0 \leq P(A) \leq, \ \forall A \subset \Omega;$

Axiomas de Probabilidade

i) $0 \leq P(A) \leq,\ \forall A \subset \Omega;$

ii) $P(\Omega) = 1;$

Axiomas de Probabilidade

i) $0 \leq P(A) \leq, \ \forall A \subset \Omega;$

ii) $P(\Omega) = 1;$

iii) $P(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} P(A_j)$, com os $A_{j's}$ disjuntos.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A \cap B) = P(A).P(B)$ (Para eventos independentes).

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A \cap B) = P(A).P(B)$ (Para eventos independentes).
- $P(A \cup A^c) = P(A) + P(A^c)$.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A \cap B) = P(A).P(B)$ (Para eventos independentes).
- $P(A \cup A^c) = P(A) + P(A^c)$.
- $P(A|B) = \dfrac{P(A \cap B)}{P(B)}, \ P(B) > 0$.

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- $P(A \cap B) = P(A).P(B)$ (Para eventos independentes).
- $P(A \cup A^c) = P(A) + P(A^c)$.
- $P(A|B) = \dfrac{P(A \cap B)}{P(B)}, \ P(B) > 0$.
- $P(A) = P(A \cap B) + P(A \cap B^c)$.

Teorema de Bayes

$$P(C_j|A) = \frac{P(A|C_j)P(C_j)}{\sum_{i=1}^{k} P(A|C_i)P(C_i)},$$

$$j = 1, 2, \ldots, k.$$

B O A

P R O V A