UENF
Universidade Estadual do Norte Fluminense Darcy Ribeiro

CCT
LCMAT

Probabilidade e Estatística

Estatística Descritiva

Curso: Ciência da Computação

Prof. Fermín A. Tang Montané

Estatística

- É uma parte da Matemática Aplicada que fornece métodos para a coleta, organização, descrição, análise e interpretação de dados quantitativos e para a utilização dos mesmos na tomada de decisões acertadas.

Estatística Descritiva

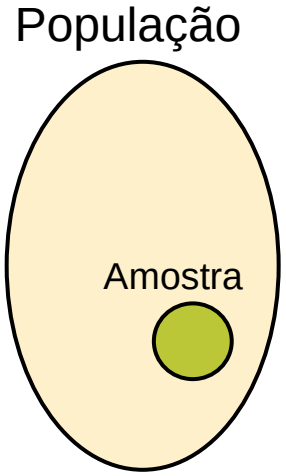
- É a parte da estatística que se preocupa com a coleta, organização e descrição dos dados observados, porém sem tirar conclusões mais genéricas.

Estatística Indutiva ou inferencial

- É a parte da estatística que trabalha com a análise e interpretação de dados.
- Ela tem como base os resultados obtidos de uma amostra, procurando inferir ou tirar conclusões para o comportamento da população, dando a precisão dos resultados e com que probabilidade se pode confiar neles.

População e Amostra

- **População** é o conjunto de todos os elementos (indivíduos ou objetos) que têm pelo menos uma característica em comum e que está sob investigação ou estudo.
- **Amostra** é qualquer subconjunto de uma população. O processo de obter amostras recebe o nome de amostragem.

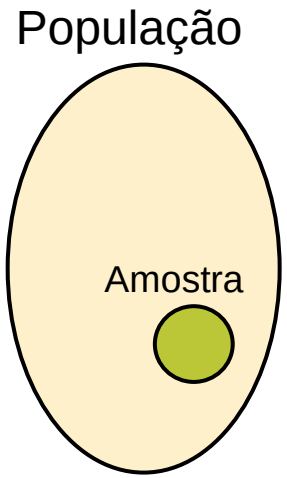


População e Amostra

- **Exemplo:**
- Suponha que você está interessado em avaliar a altura média dos alunos de uma escola. Para conhecer esta característica, deverá medir a altura dos alunos (dados).
- Como o seu interesse atinge somente uma determinada escola, então todos os alunos desta escola formam a população da pesquisa e o conjunto dos alunos de uma determinada sala representa uma amostra.

Parâmetros e Estatísticas

- As medidas estatísticas obtidas com base em uma população são chamadas de **parâmetros** (representadas por letras gregas)
- Já, as medidas baseadas em amostras são chamadas **estatísticas** (representadas por letras do alfabeto latino).



Parâmetros populacionais:
μ: Média
σ: Desvio padrão

Estatísticas amostrais:
$\bar{x}$: Média
s: Desvio padrão

Amostragem

- As análises estatísticas geralmente são realizadas através de amostras, já que em geral a maior parte das populações são representadas por um número muito grande de indivíduos ou objetos, o que ocasiona um número muito grande de dados.
- Para que se possa fazer inferências estatísticas válidas sobre uma população apartir de uma amostra, deve-se definir cuidadosamente a população de interesse, selecionar a característica que irá pesquisar e cuidar para que a amostra seja representativa.

Variáveis

- Variável é a característica que é objeto de estudo numa pesquisa.
- **Variável qualitativa:** é aquela que não pode ser medida numericamente.
- Exemplos: cor dos olhos, marca de refrigerante, etc.
- **Variável quantitativa:** é aquela que pode ser medida numericamente.
- Exemplos: peso, altura, número de defeitos,etc.

Variáveis

- **Variável discreta:** é a variável quantitativa que pode assumir um número infinito enumerável de valores.
- Exemplo: número de filhos: 0,1,2,3,...
- **Variável contínua:** é a variável quantitativa que pode assumir um número infinito não enumerável de valores.
- Exemplo: altura dos alunos: 1,73m, 1,84m,...

Distribuição de Frequência

- Feita a coleta, os dados originais ainda não se encontram prontos para análise, por não estarem numericamente organizados. Por essa razão, são chamados de dados brutos.
- Para obter informações de interesse sobre a característica em estudo, deve-se agrupar os dados obtidos em uma **distribuição de freqüência**, onde os valores observados não mais aparecerão individualmente.

Distribuição de Frequência

- Os dados abaixo representam as idades(em anos) dos alunos de Estatística de um determinado curso da UENF do ano de 2024.

20	21	21	21	22	22	22
22	23	23	23	23	23	23
23	24	24	24	24	24	24
24	24	24	25	25	25	25
25	25	26	26	26	26	28

Distribuição de Frequência

Idade (x_i)	Número de Alunos (f_i)	f_{ac}	f_r
20	1	1	0,0286
21	3	4	0,0857
22	4	8	0,1143
23	7	15	0,2000
24	9	24	0,2571
25	6	30	0,1714
26	4	34	0,1143
27	0	34	0,0000
28	1	35	0,0286
Total	**35**		**1**

Onde:

x_i: valor observado;

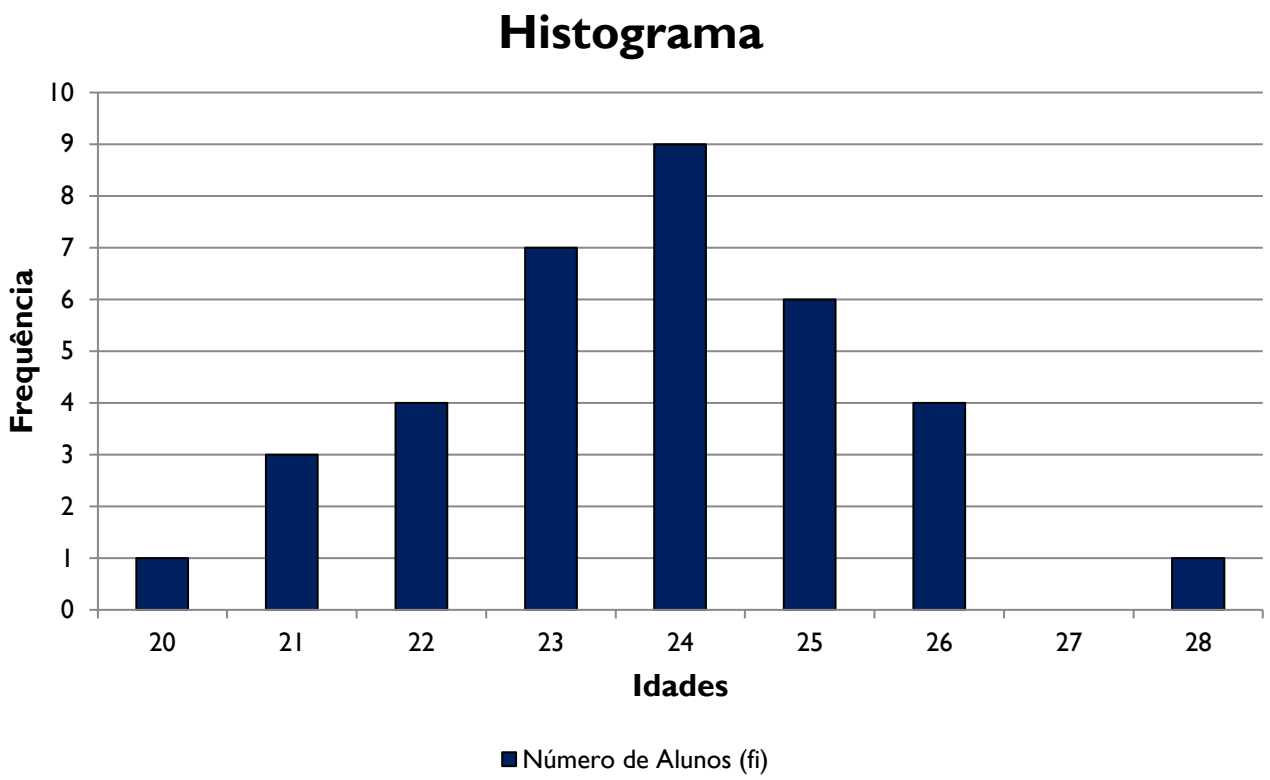
f_i: frequência observada ou absoluta;

f_{ac}: frequência acumulada (F_i);

f_r: frequência relativa.

Distribuição de Frequência

- Representação Gráfica mediante Histograma.



Distribuição de Frequência em Classes

- Exemplo:
- Considere que as alturas(em metros) de 30 alunos de uma sala de aula são os seguintes:

1,50	1,53	1,68	1,51	1,63	1,65
1,54	1,55	1,65	1,56	1,57	1,50
1,60	1,48	1,61	1,52	1,63	1,47
1,52	1,50	1,52	1,46	1,45	1,66
1,65	1,59	1,51	1,58	1,62	1,60

- Chama-se classe ao intervalo considerado para agrupar os dados observados.

Distribuição de Frequência em Classes

- Para se construir uma distribuição de freqüência utilizando classes, deve-se determinar:
  - a) Número de classes (k);
  - b) Amplitude total dos dados (A);
  - c) Amplitude do intervalo de classe (h);
  - d) Limite inferior (LI_i) e Limite superior (LS_i) da classe i.

Distribuição de Frequência em Classes

- a) Número de classes (k):
  - Sugere-se utilizar a fórmula de Sturges: $k = 1 + 3{,}32.\log n$
  - onde:
  - n é o número de dados e
  - k deve ser um número inteiro positivo.

- b) Amplitude total dos dados (A):
  - $A = X_{max} - X_{min}$,
  - onde:
  - X_{max} é o valor máximo da amostra e
  - X_{min} é o valor mínimo da amostra.

Distribuição de Frequência em Classes

- c) Amplitude do intervalo de classe (h):
  - $h = A/k$
  - h deve ser um valor de modo que as classes acomodem todos os dados da amostra.
- d) Limite inferior (LI_i) e Limite superior (LS_i) da classe:
  - LI_i é o menor valor aceito na classe i;
  - $LS_i = LI_i + h$.

Observação:

- Os intervalos de classe são definidos de maneira que possuam igual amplitude. Para isso, sugere-se ajustar o limite inferior da primeira classe e/ou o limite superior da última classe, sempre que conveniente.

Distribuição de Frequência em Classes

- c)

Intervalo de Classe Alturas (m)			Nº Alunos f_i	f_{ac}	x_i
[1,45	1,49	>	4	4	1,47
[1,49	1,53	>	8	12	1,51
[1,53	1,57	>	4	16	1,55
[1,57	1,61	>	5	21	1,59
[1,61	1,65	>	4	25	1,63
[1,65	1,69	>	5	30	1,67
Total			**30**		

Ponto médio da classe

Medidas de Tendencia Central

- Medidas de tendência central são medidas estatísticas, cujos valores estão próximos do centro de um conjunto de dados dispostos ordenadamente em sentido crescente ou decrescente.
- As mais conhecidas são:
  - Média aritmética
  - Média geométrica
  - Mediana
  - Moda

Media Aritmética

a) Dados não agrupados

- A média aritmética de um conjunto de n valores:

$x_1, x_2, x_3, \ldots, x_n$ é definida por:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Exemplo:

- As idades (em anos) de 5 jogadores de futebol são:

18, 16, 15,17, 17. A média aritmética das idades destes jogadores é:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{18+16+15+17+17}{5} = 16,6$$

Media Aritmética

a) Dados agrupados

- No caso em que os dados estão agrupados em classes discretas. Se x_1, x_2, x_3, ... ,x_n ocorrem com as frequências f_1, f_2, f_3, ... ,f_n, a média aritmética é calculada por:

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots + x_n f_n}{n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}$$

- No caso em que os dados estão agrupados em intervalos de classes. Os valores x_1, x_2, x_3, ... ,x_n correspondem aos pontos médios de cada classe, assim:

$$x_i = \frac{LI_i + LS_i}{2}$$

Media Aritmética

a) Dados agrupados em classes discretas

Idade (x_i)	Número de Alunos (f_i)	$x_i f_i$
20	1	20
21	3	63
22	4	88
23	7	161
24	9	216
25	6	150
26	4	104
27	0	0
28	1	28
Total	**35**	**830**

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{830}{35} = 23,71429$$

$$\overline{x} = 23,7 \text{ anos}$$

Media Aritmética

a) Dados agrupados em intervalos de classes

Intervalo de Classe Alturas (m)			Número de Alunos (f_i)	x_i	$f_i x_i$
[1,45	1,49	>	4	1,47	5,88
[1,49	1,53	>	8	1,51	12,08
[1,53	1,57	>	4	1,55	6,20
[1,57	1,61	>	5	1,59	7,95
[1,61	1,65	>	4	1,63	6,52
[1,65	1,69	>	5	1,67	8,35
Total			**30**		**46,98**

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{46,98}{30} = 1,5666$$

$\bar{x} \sim 1,57$ metros

Media Geométrica

a) Dados não agrupados

- A média geométrica de um conjunto de n valores:

$x_1, x_2, x_3, \ldots, x_n$ é definida por:

$$M_g = \sqrt[n]{x_1.x_2.x_3.\cdots.x_n} = 10^{\frac{\sum_{i=1}^{n} \log x_i}{n}}$$

Lembre que:
$\log b^c = c \log b$

Exemplo:

- A média geométrica das idades dos 5 jogadores de futbol citados anteriormente é:

$$M_g = \sqrt[5]{x_1.x_2.x_3.x_4.x_5} = \sqrt[5]{18.16.15.17.17} = 16{,}56823 \sim 16{,}6 \text{ anos}$$

Media Geométrica

a) Dados agrupados

- No caso em que os dados estão agrupados em classes discretas. Se $x_1, x_2, x_3, \ldots, x_n$ ocorrem com as frequências $f_1, f_2, f_3, \ldots, f_n$, a média geométrica é calculada por:

$$M_g = \sqrt[n]{x_1^{f_1}.x_2^{f_2}.x_3^{f_3}.\cdots.x_n^{f_n}} = 10^{\frac{\sum_{i=1}^{n} f_i \log x_i}{n}}$$

- No caso em que os dados estão agrupados em intervalos de classes. Os valores $x_1, x_2, x_3, \ldots, x_n$ correspondem aos pontos médios de cada classe, assim:

$$x_i = \frac{LI_i + LS_i}{2}$$

Media Geométrica

a) Dados agrupados em classes discretas

Idade (x_i)	Número de Alunos (f_i)	$f_i.\log(x_i)$
20	1	1,30
21	3	3,97
22	4	5,37
23	7	9,53
24	9	12,42
25	6	8,39
26	4	5,66
27	0	0,00
28	1	1,45
Total	**35**	**48,09**

$$M_g = 10^{\frac{\sum_{i=1}^{n} f_i \log x_i}{n}} = 10^{\frac{48,09}{35}} = 23,6592$$

$$M_g = 23,7 \text{ anos}$$

Media Geométrica

a) Dados agrupados em classes discretas

Intervalo de Classe Alturas (m)			Número de Alunos (f_i)	x_i	$f_i.\log(x_i)$
[1,45	1,49	>	4	1,47	0,67
[1,49	1,53	>	8	1,51	1,43
[1,53	1,57	>	4	1,55	0,76
[1,57	1,61	>	5	1,59	1,01
[1,61	1,65	>	4	1,63	0,85
[1,65	1,69	>	5	1,67	1,11
Total			**30**		**5,83**

$$M_g = 10^{\frac{\sum_{i=1}^{n} f_i \log x_i}{n}}$$

$$= 10^{\frac{5,83}{30}} = 1,5643$$

$$M_g = 1,56 \text{ metros}$$

Mediana

a) Dados não agrupados

- A mediana M_e de um conjunto de n valores ordenado $x_1, x_2, x_3,...,x_n$ é representada pelo valor central do conjunto para n ímpar e pela média aritmética dos dois valores centrais para n par.

- Exemplos:

a) 3, 3, 4, 5, 7, 8, 9, 10, 12

Como n = 9, então, $M_e = 7$

b) 3, 3, 4, 5, 7, 7, 9, 10

Como n = 8, então, $M_e = \frac{5+7}{2} = 6$

Mediana

b) Para dados agrupados em classes discretas considera-se:

- Se o número de elementos é impar, calcula-se a posição da mediana e identifica-se a classe que a contêm. O valor da mediana corresponderá ao valor da classe. Calcula-se a posição da mediana como:

$$P = \left\lfloor \frac{n}{2} \right\rfloor + 1$$

- Se o número de elementos é par, a mediana será calculada como a média dos dois valores centrais. Neste caso, a mediana não pertence a uma classe. Calcula-se a posição do primeiro elemento central como:

$$P = \frac{n}{2}$$

Mediana

- Para dados agrupados em classes discretas.

Idade (x_i)	Número de Alunos (f_i)	f_{ac}
20	1	1
21	3	4
22	4	8
23	7	15
24	9	24
25	6	30
26	4	34
27	0	34
28	1	35
Total	**35**	

$$P = \left\lfloor \frac{35}{2} \right\rfloor + 1 = 17 + 1 = 18$$

$$M_e = 24 \text{ anos}$$

Mediana

b) Para dados agrupados em intervalos de classe utiliza-se a expressão:

$$M_e = LI_e + \left(\frac{P - f'_{ac}}{f_{M_e}} \right).h$$

onde:

- LI_e: limite inferior da classe mediana;
- P: posição da mediana;
- f'_{ac}: frequência acumulada da classe anterior a classe mediana;
- f_{Me}: frequência da classe mediana;
- h: amplitude do intervalo de classe.

Mediana

- Para dados agrupados em intervalos de classe.

Intervalo de Classe Alturas (m)			Número de alunos f_i	f_{ac}
[1,45	1,49	>	4	4
[1,49	1,53	>	8	12
[1,53	1,57	>	4	16
[1,57	1,61	>	5	21
[1,61	1,65	>	4	25
[1,65	1,69	>	5	30
Total			**30**	

$$M_e = LI_e + \left(\frac{P - f'_{ac}}{f_{M_e}} \right).h$$

$$P = \frac{30}{2} = 15$$

$$M_e = 1,53 + \left(\frac{15 - 12}{4} \right).0,04$$

$$= 1,56 \text{ metros}$$

Moda

a) Dados não agrupados

- Moda M_o de um conjunto de n valores x_1, x_2, x_3,...,x_n é o número desse conjunto que possuir a maior repetição.
- Se o conjunto não tiver valores repetidos não existirá moda (amodal).
- Se dois valores estiverem igualmente repetidos, tem-se então duas modas e o conjunto será dito bimodal.
- A moda é o valor ao qual está associado a freqüência mais alta.

Moda

- b) Para dados agrupados em classes discretas.

Idade (x_i)	Número de Alunos (f_i)
20	1
21	3
22	4
23	7
24	9
25	6
26	4
27	0
28	1
Total	**35**

Moda é a idade que mais se repete, ou seja, a que têm maior frequência absoluta.

$$M_o = 24 \text{ anos}$$

Moda

b) Para dados agrupados em intervalos de classe utiliza-se a formula de Czuber:

$$M_o = LI_o + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right).h$$

onde:

- LI_o: limite inferior da classe modal; Chama-se classe modal a classe com maior frequência absoluta.
- Δ_1: diferença entre a frequência da classe modal e a classe imediatamente anterior;
- Δ_2: diferença entre a frequência da classe modal e a classe imediatamente posterior;
- h: amplitude do intervalo de classe.

Moda

- Para dados agrupados em intervalos de classe.

Intervalo de Classe Alturas (m)			Número de alunos f_i	f_{ac}
[1,45	1,49	>	4	4
[1,49	1,53	>	8	12
[1,53	1,57	>	4	16
[1,57	1,61	>	5	21
[1,61	1,65	>	4	25
[1,65	1,69	>	5	30
Total			**30**	

A segunda classe é a classe modal.

$$M_o = LI_o + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right).h$$

$$M_o = 1,49 + \left(\frac{4}{4+4} \right).0,04$$

$$= 1,51 \text{ metros}$$