ESTATÍSTICA

Michelle Hanne Soares de Andrade
michellehanne.andrade@gmail.com

Estatística Descritiva

Variáveis

Variável é a característica de interesse que é medida em cada elemento da amostra ou população.

Seus valores variam de elemento para elemento.

Podem ser **características numéricas** sob as quais operações aritméticas podem ser realizadas como salário, altura, ou tempo de vida; ou podem ser **atributos** como sexo, estado civil ou classe social.

Variáveis

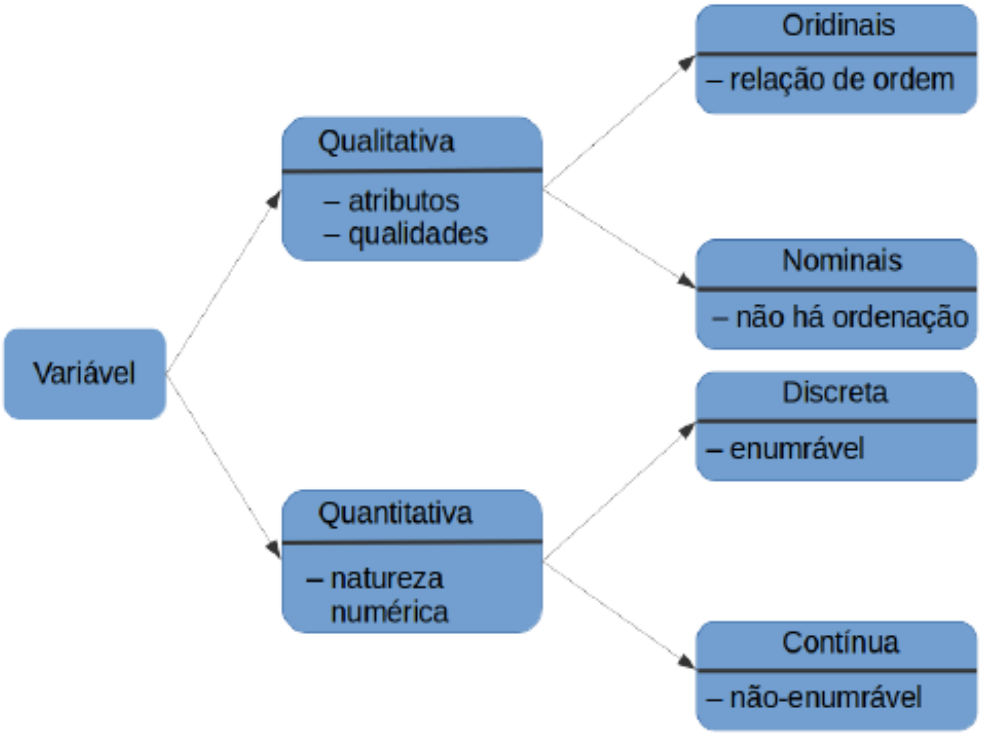
- As variáveis pode ser classificadas em:
  - **Quantitativas:** são características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que nos permite efetuar operações aritméticas diretamente. Podem ser **CONTÍNUAS ou DISCRETAS.**
    - **Variáveis discretas:** características mensuráveis que podem assumir apenas um numero finito ou contável de valores. Geralmente são o resultado de contagens: números de filhos, número de bactérias por litro de agua, número de peças defeituosas por lote, etc.
    - **Variáveis contínuas:** características mensuráveis que assumem valores em uma escala contínua (na reta real). Usualmente devem ser medidas através de algum instrumento: peso, altura, tempo, etc.

Variáveis

- **Qualitativas (ou categoricas):** São características definidas por categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.
  - **Variáveis nominais:** não existe ordenação entre as categorias em estudo: sexo (F ou M), cor dos olhos (castanhos, azuis, verdes), fumante (sim ou n~ao), turma (diurno ou noturno), etc.
  - **Variáveis ordinais:** existe uma ordenação entre as categorias em estudo: escolaridade (ensino medio, tecnico, superior), tamanho (pequeno, medio, grande), classe social (baixa, media, alta).

- Obs.: variáveis qualitativas **podem** ser tratadas como variáveis quantitativas discretas.

Variáveis


Variável
Qualitativa
– atributos
– qualidades
Oridinais
– relação de ordem
Nominais
– não há ordenação
Quantitativa
– natureza numérica
Discreta
– enumrável
Continua
– não-enumrável

Organização dos Dados

- Em um primeiro estágio, conjunto de dados são normalmente apresentados em forma **de tabelas brutas ou listas**.
- Pouco pode ser inferido de maneira imediata nestes casos.
- Algumas formas de organização podem ajudar a evidenciar características relevantes.
- Dentre as principais metodologias **destacam-se a construção de diagramas de ramos e folhas e tabelas de frequências.**

Diagramas de Ramos e Folhas

- Busca oferecer uma apresentação simultaneamente visual e informativa dos dados.
- Aplica-se a **variáveis quantitativas cujos valores têm ao menos dois dígitos.**

Diagramas de Ramos e Folhas

- Construção

  1. seja $x_1, x_1, \ldots x_n$ um conjunto qualquer de valores observados;

  2. divida cada numero x_i em duas partes:

     - 2.1 um ramo consistindo de um ou mais dígitos;
     - 2.2 uma folha consistindo dos dígitos restantes;

  3. liste os valores do ramo em uma coluna vertical;

  4. ao lado do ramo, registre a folha de cada observação;

  5. descreva, em uma coluna mais a direita, o número de folhas contidas em cada ramo.

Diagramas de Ramos e Folhas - Exemplo

- Os dados a seguir correspondem a resistência a compressão (medida em libra-força por polegada quadrada) de 80 corpos de prova de uma certa liga de alumínio em estudo.

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

Diagramas de Ramos e Folhas - Exemplo

ramo	folha	frequência
7	6	1
8	7	1
9	7	1
10	5 1	2
11	5 8 0	3
12	1 0 3	3
13	4 1 3 5 3 5	6
14	2 9 5 8 3 1 6 9	8
15	4 7 1 3 4 0 8 8 6 8 0 8	12
16	3 0 7 3 0 5 0 8 7 9	10
17	8 5 4 4 1 6 2 1 0 6	10
18	0 3 6 1 4 1 0	7
19	9 6 0 9 3 4	6
20	7 1 0 8	4
21	8	1
22	1 8 9	3
23	7	1
24	5	1

- A partir do diagrama observa-se de imediato que:
  - a maioria das resistências está entre 120 e 200;
  - o valor central esta entre 150 e 160;
  - os valores são distribuídos de forma aproximadamente simétrica.

Diagramas de Ramos e Folhas - Exemplo

- Em algumas aplicações, pode ser interessante prover mais ramos. Por exemplo, o seguinte diagrama obtido de outro conjunto de dados:

ramo	folha
6	1 3 4 5 5 6
7	0 1 1 3 5 7 8 8 9
8	1 3 4 4 7 8 8
9	2 3 5

Diagramas de Ramos e Folhas - Exemplo

- Cada ramo poderia ser subdividido em dois outros ramos:
  - um contendo apenas folhas inferiores ou iguais a 4 (L);
  - outro contendo apenas folhas superiores ou iguais a 5 (U);

ramo	folha
6L	1 3 4
6U	5 5 6
7L	0 1 1 3
7U	5 7 8 8 9
8L	1 3 4 4
8U	7 8 8
9L	2 3
9U	5

Diagramas de Ramos e Folhas - Exemplo

- Outra alternativa seria subdividir cada ramo em cinco outros ramos:
  - ramo A contendo as folhas 0 e 1;
  - ramo B contendo as folhas 2 e 3;
  - ramo C contendo as folhas 4 e 5;
  - ramo D contendo as folhas 6 e 7;
  - ramo E contendo as folhas 8 e 9;

ramo	folha
6A	1
6B	3
6C	4 5 5
6D	6
6E	
7A	0 1 1
7B	3
7C	5
7D	7
7E	8 8 9
8A	1
8B	3
8C	4 4
8D	7
8E	8 8
9A	
9B	2 3
9C	5
9D	
9E	

Diagramas de Ramos e Folhas - Exemplo

Em geral, devemos **escolher um número de ramos relativamente pequeno** em comparação com o número de observações.

Muitos ramos podem gerar uma **perda significativa** de informação em uma primeira analise visual.

E **comum** considerarmos algum valor entre **5 e 20 ramos**.

Tabelas de Frequência

- Oferecem um resumo mais compactado dos dados na comparação com os diagramas de ramos e folhas.
- Ajudam a explicar a distribuição dos dados empiricamente.
- A ideia é **agrupar** os valores observados em **classes** e em seguida indicar suas **frequências (absolutas ou relativas).**

Tabelas de Frequência

- Dados Discretos:
  - Cada observação constitui naturalmente uma classe **(uma linha da tabela).**
  - Basta então contar o número de repetições de cada um dos diferentes valores observados e agrupá-los em uma tabela.

Tabelas de Frequência - Exemplo

- Em um questionário estudantil, avaliou-se sexo, idade, altura, peso, número de filhos, habito de fumar (sim ou não) e tolerância ao cigarro (indiferente, incomoda pouco e incomoda muito) dos 25 indivíduos de uma turma.

Identificação	Sexo	Idade	Altura	Peso	Filhos	Fuma	Toler
1	F	17	1,60	60,5	2	NAO	P
2	F	18	1,69	55,0	1	NAO	M
3	M	18	1,85	72,8	2	NAO	P
4	M	25	1,85	80,9	2	NAO	P
5	F	19	1,58	55,0	1	NAO	M
6	M	19	1,76	60,0	3	NAO	M
7	F	20	1,60	58,0	1	NAO	P
8	F	18	1,64	47,0	1	SIM	I
9	F	18	1,62	57,8	3	NAO	M
10	F	17	1,64	58,0	2	NAO	M
11	F	18	1,72	70,0	1	SIM	I
12	F	18	1,66	54,0	3	NAO	M
13	F	21	1,70	58,0	2	NAO	M
14	M	19	1,78	68,5	1	SIM	I
15	F	18	1,65	63,5	1	NAO	I
16	F	19	1,63	47,4	3	NAO	P
17	F	17	1,82	66,0	1	NAO	P
18	M	18	1,80	85,2	2	NAO	P
19	F	20	1,60	54,5	1	NAO	P
20	F	18	1,68	52,5	3	NAO	M
21	F	21	1,70	60,0	2	NAO	P
22	F	18	1,65	58,5	1	NAO	M
23	F	18	1,57	49,2	1	SIM	I
24	F	20	1,55	48,0	2	SIM	I
25	F	20	1,69	51,6	2	NAO	P

Tabelas de Frequência - Exemplo

- Tabelas de frequência absoluta para as variáveis sexo, idade, número de filhos, habito de fumar e tolerância ao cigarro:

Sexo	freq.
F	20
M	5
Total	25

Idade	freq.
17	3
18	11
19	4
20	4
21	2
25	1
Total	25

Filhos	freq.
1	12
2	8
3	5
Total	25

Fuma	freq.
S	5
N	20
Total	25

Toler	freq.
I	6
P	10
M	9
Total	25

Tabelas de Frequência – Caso Contínuo

Variáveis Continuas assumem valores em conjuntos não-enumeráveis, logo, é pouco provável, mas não impossível, observarmos valores repetidos.

- Neste caso, uma tabela de frequências pode ser obtida através de um agrupamento em classes (intervalos).
- O número de classes é escolhido de acordo com a quantidade de observações.
- Uma pratica usual consiste de escolher o número de classes como aproximadamente $\sqrt{\boldsymbol{n}}$, onde $\boldsymbol{n}$ é o **número total de observações**.
- Uma vez especificado o número de classes, o limite de cada classe é escolhido de modo a garantir homogeneidade nas amplitudes.

Tabelas de Frequência – Caso Contínuo

- Exemplo:
  - Consideremos o questionário estudantil do exemplo anterior.
  - Afim de construir uma tabela de frequências podemos considerar k = 5 classes, já que p $\sqrt{25} = 5$.
  - A amplitude das classes, com respeito as variáveis altura e peso, serão dadas respectivamente por:

$$\text{amplitude}_{alt} = \frac{\text{amplitude total}}{k} = \frac{1,85 - 1,55}{5} = 0,06;$$

$$\text{amplitude}_{peso} = \frac{\text{amplitude total}}{k} = \frac{85,2 - 47,0}{5} = 7,64.$$

Tabelas de Frequência – Caso Contínuo - Exemplo

- Intervalos resultantes:
  - altura { [1,55; 1,61); [1,61; 1,67); [1,67; 1,73); [1,73; 1,79); [1,79; 1,85];
  - peso { [47,0; 54,64); [54,64; 62,28); [62,28; 69,92); [69,92; 77,56); [77,56; 85,2]:
- **Agora basta contar o número de ocorrências em cada intervalo.**

Tabelas de Frequência – Caso Contínuo - Exemplo

- **Tabelas de frequência relativa** diferem dos Ramos e Folhas pelo fato de dividirem as frequências absolutas pelo número de observações, oferecendo assim uma noção de proporção. Tabelas para as variáveis altura e peso:

Altura	freq.
$1,55 \vdash 1,61$	0,2
$1,61 \vdash 1,67$	0,28
$1,67 \vdash 1,73$	0,28
$1,73 \vdash 1,79$	0,08
$1,79 \vdash\dashv 1,85$	0,16
Total	1,00

Peso	freq.
$47,00 \vdash 54,64$	0,32
$54,64 \vdash 62,28$	0,40
$62,28 \vdash 69,92$	0,12
$69,92 \vdash 77,56$	0,08
$77,56 \vdash\dashv 85,2$	0,08
Total	1,00

Resumo dos Dados

- O objetivo é descrever numericamente, e de forma sucinta, características importantes dos dados. Em algumas situações, o simples armazenamento de Medidas Resumo é suficiente para o desenvolvimento da análise estatística de um problema.
- **Resumos incluem basicamente medidas de posição, separação e dispersão.**

Medidas de Posição

- Indicam alguma tendência central ou comportamento esperado com respeito extraídos de uma amostra qualquer.
- Dentre as principais medidas, destacam-se: **media, mediana e moda.**

Medidas de Posição

- Sejam $x_1, x_2, \ldots x_n$ n observações de um fenômeno aleatório qualquer. Denominamos media aritmética da amostra.

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

Média Aritmética

- Pode ser interpretada como o centro de massa (baricentro das observações).

- Caso os dados estejam organizados em uma tabela de frequências, isto é, se para cada x_i foi verificado um número f_i relativo a sua frequência, então

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i}.$$

Exemplo

- Em um estudo informal, desejamos consultar a opinião de 30 colegas a respeito do perfil social da população de uma dada cidade. Cada colega deve responder se acredita que a maioria da população é de classe baixa (B), media (M) ou alta (A). As respostas foram sumarizadas na seguinte tabela de frequências:

classe	freq
B	5
M	10
A	15
Total	30

Exemplo

Uma pesquisa avaliou a idade dos 25 alunos de uma turma de engenharia. Os dados foram organizados na seguinte tabela de frequências:

Idade	freq.
19	3
20	5
21	1
22	8
23	4
24	1
25	0
26	3
Total	25

A média aritmética pode ser obtida diretamente do dispositivo (1):

$$\overline{x} = \frac{\sum_{i=1}^{8} x_i f_i}{\sum_{i=1}^{8} f_i} = \frac{548}{25} = 21,92.$$

Exemplo

Os dados a seguir correspondem ao tempo de vida, em anos, de cada elemento de uma amostra de lâmpadas produzidas por uma fábrica qualquer.

t	freq.
$0 \vdash 0,5$	3
$0,5 \vdash 1,0$	12
$1,0 \vdash 1,5$	20
$1,5 \vdash 2,0$	5
$2,0 \vdash$	0
Total	40

Quando buscamos uma tendência central e os dados estão agrupados em intervalos, devemos selecionar um elemento específico como representante de cada classe a fim de efetuarmos os cálculos necessários. Algumas das escolhas comuns são: centro ou extremos do intervalo. Neste caso, dizemos que a média foi **estimada** e não calculada.

Exemplo

Tomando o ponto médio dos quatro primeiros intervalos e o extremo esquerdo do último intervalo, obtemos:

$$\overline{x} = \frac{\sum_{i=1}^{5} x_i f_i}{\sum_{i=1}^{5} f_i} = \frac{43,5}{40} = 1,0875 \text{ anos.}$$

A Moda

- A moda é o valor que mais se repete na amostra. Assim como a mediana, e pouco sensível a valores atípicos. Contudo, pode não ser única. Pode ser aplicada a variáveis nominais.
- Resultado coincide com o obtido em termos da mediana.
- Tanto a mediana quanto a moda apresentam interpretações mais razoáveis, em relação a media, quando desejamos estudar variáveis ordinais.

Sensibilidade dos Valores das Médias

- Médias aritméticas são bastante sensíveis a valores extremos.
- Pode levar a conclusões equivocadas.
  - Suponhamos, por exemplo, que os salários dos funcionários de uma empresa de pequeno porte são: R$2500,00, R$2500,00, R$1000,00, R$1800,00 e R$20000.
  - A media salarial da empresa

$$\overline{x} = \frac{2500 + 2500 + 1000 + 1800 + 20000}{5} = 5560$$

- Conclusão equivocada a respeito da tendência salarial da empresa.

Mediana

A mediana corresponde ao valor que ocupa a posição central dos dados após sua ordenação.

Desta forma, a mediana desconsidera os valores observados em si, já que basta analisar a posição das observações perante sua ordenação.

Informalmente, esta medida trata os valores das observações de forma indireta, já que estes são úteis apenas para o desenvolvimento da ordenação.

Pode apresentar uma medida de tendência mais honesta na presença de valores atípicos.

Mediana - dados discretos ou não agrupados

- Caso os dados de interesse sejam discretos ou, sejam contínuos e ainda não tenham passado por um agrupamento, então o cálculo da mediana é mais simples.
- Basta observar a paridade do número n de observações:
  - **se n e ímpar, então a mediana e exatamente o valor central das observações.**

$$x^{\left(\frac{n+1}{2}\right)};$$

  - caso contrario, a mediana e dada pela media aritmética dos dois valores centrais

$$\frac{x^{\left(\frac{n}{2}\right)} + x^{\left(\frac{n}{2}+1\right)}}{2}.$$

Mediana - dados discretos ou não agrupados

- Exemplos:
  - No exemplo dos salários, temos a seguinte ordenação: R$1000,00, R$1800,00, R$2500,00, R$2500,00, R$20000. Como o número de observações e ímpar, temos que a mediana e dada pelo elemento **central x(3) = 2500.**
  - No exemplo das classes sociais, temos a seguinte ordenação: B, B, B, B, B, M, M, M, M, M, M, M, M, M, M, A, A, A, A, A, A, A, A, A, A,A, A, A, A, A. Como o número de observações é par, segue que a mediana é dada pelo ponto médio de x(15) e x(16). A mediana então ficaria entre os valores M e A.

Exercício

- Um borracheiro anotou a vida útil dos pneus dos carros de seus clientes

2,4	5,2	3,5	4,9	3,5	6
4,5	3,2	3,7	2,9	4,7	5,2
3,8	5,4	6,2	5,5	4,5	6

- Faça um Diagrama de Ramos e Folhas, calcule a Amplitude e faça a Tabela de Frequência.