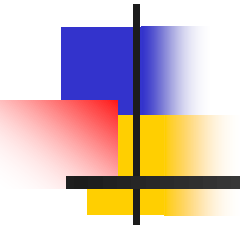
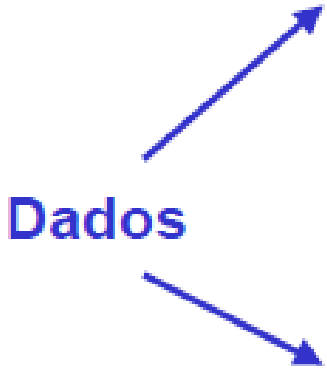


PROBABILIDADE

MEDIDAS ESTATÍSTICAS – Tipos de Variáveis

Tipos de Dados



Cross-sectional (Transversal)

Todas as informações referem-se a um único período de tempo.

Séries Temporais (Longitudinais)

Os dados referem-se ao comportamento de uma variável ao longo do tempo.

COLETA DE DADOS – Tipos de Variáveis

Tipos de Variáveis

Qualitativas

Têm como possíveis realizações características de **natureza** não numérica

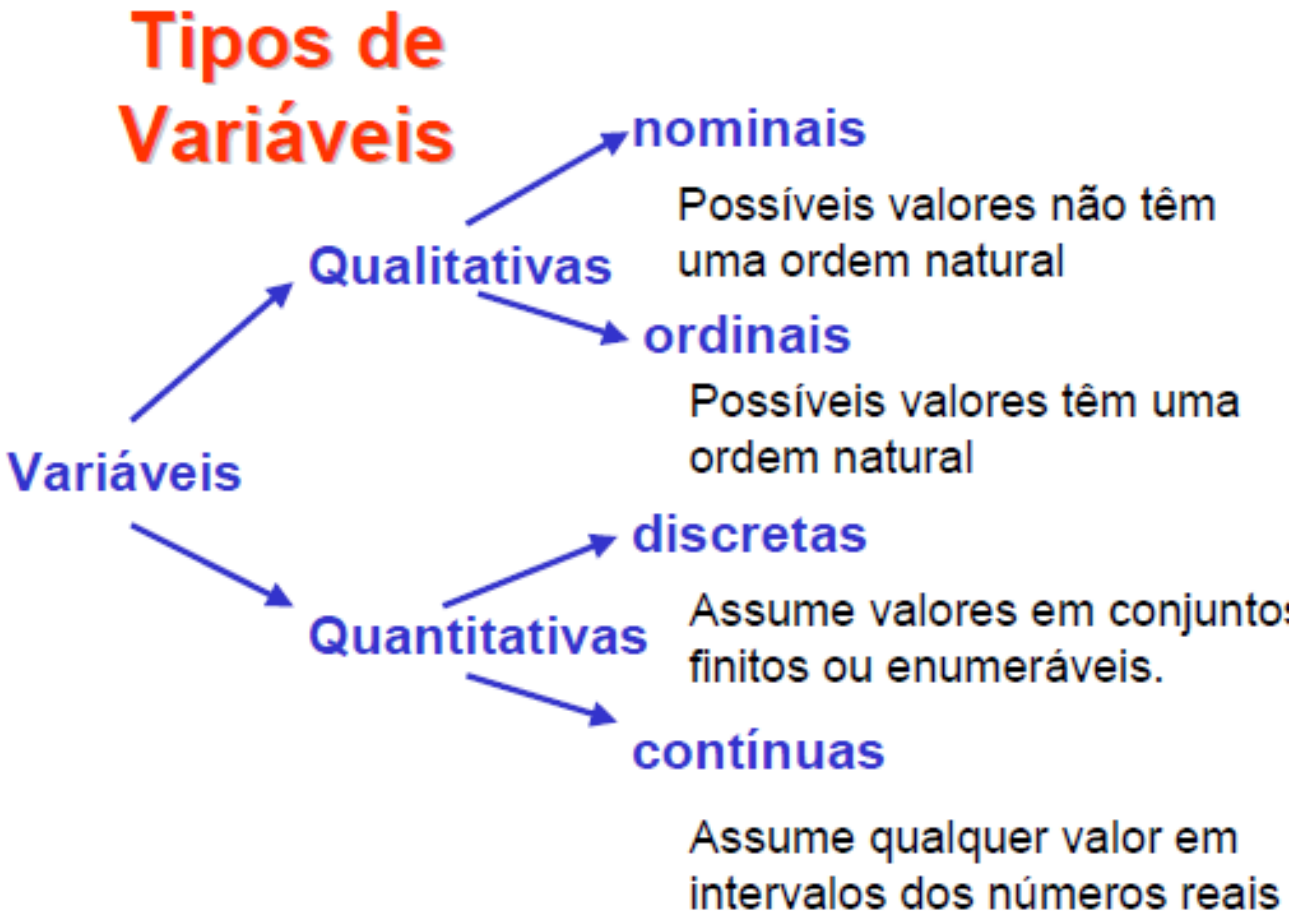
Qualidade ou Atributo

Quantitativas

Têm como possíveis valores números obtidos a partir de contagens ou mensurações.

Contagem ou Numeração

COLETA DE DADOS – Tipos de Variáveis



Categórica ou Nominal: Cor dos olhos (Pretos, Castanhos, Azuis & Verdes)

Ordinal: Grau de Instrução (1º. Grau, 2º. Grau, Superior)

Discreta: Número de Hospitais (ou leitos) de uma cidade

Contínua: Estatura dos Indivíduos

COLETA DE DADOS

EXERCÍCIO: Identifique o tipo de escala (N=Nominal, O=Ordinal, I=Intervalar)

( ) Produtos bancários adquiridos (conta-corrente, poupança, renda fixa, etc.)
( ) Forma de pagamento (à vista, 30 dd, 45 dd, ..)
( ) Data de pagamento
( ) Juros aplicados
( ) Escolaridade
( ) Tipo de canal de vendas utilizado (agência/loja, internet, cx. Eletrônico, ...)
( ) Cargo na empresa em que trabalho
( ) Tipo de residência (apto, casa, ...)
( ) Valor de um imóvel dado em garantia
( ) N° de televisores na residência

MEDIDAS ESTATÍSTICAS

- Especialmente para Análise Exploratória de dados.
- São valores calculados para um conjunto de dados, e usados de alguma forma, para descrever e resumir estes dados. As medidas estatísticas são divididas em medidas de posição e medidas de dispersão.

Observação: quando as medidas de tendência central e as de dispersão são calculadas sobre a população, elas são chamadas de *parâmetros*. Por outro lado, quando essas medidas são obtidas considerando-se uma amostra retirada de uma população, elas são chamadas de *estatísticas*.

MEDIDAS DE POSIÇÃO

São as medidas estatísticas, obtidas para um conjunto de dados, cujos valores, geralmente, estão localizados em torno do centro deste conjunto de dados, estando eles ordenados.

São também chamados de medidas tendência central. As mais importantes medidas de posição são: Mínimo, Máximo, Média Aritmética, Mediana, Moda e os Quartis.

DADOS NÃO AGRUPADOS

- Quando os dados NÃO estão agrupados em uma distribuição de frequências, tem-se o valor individual da variável.

Dados Não Agrupados

2	4	2	1	2
3	1	0	5	1
0	1	1	2	0
1	3	0	1	2

Dados Agrupados

Distribuição de frequências		
x_i	f_i	fr_i
0	4	0,20
1	7	0,35
2	5	0,25
3	2	0,10
4	1	0,05
5	1	0,05
	20	1

MÉDIA AMOSTRAL

$$\overline{x} = \frac{\sum X}{n}$$

- n é o número total de observações

Exercício 1:

- Um processo de geração de faturas está sendo mensurado com relação ao número de erros por semana, nos últimos dois meses, o número de erros semanais foi: 8,11,5,14,8,11,16,11.

- Calcule o número médio de erros por semana.

Mediana

Calcula média usando um critério diferente.
Medida Central dos Dados ordenados.

- A Mediana divide um grupo ordenado de valores em 2 partes iguais (50% acima e 50% abaixo da Mediana).
- Se o número de itens for **ímpar**, a Mediana será o valor do meio.
- Se o número de itens é **par**, a Mediana será a média dos 2 valores do meio.

Interpretação: Número de Observações maior que a mediana é igual ao número de observações menor que a mediana.

Exemplo: Determine a Mediana e a Média

- 1
- 5
- 8
- 9
- 10

- Posição da Mediana:(n+1)/2
- (5+1)/2= 3 lugar
- Mediana= 8

- Média?

Determine a Mediana e a Média

- 8
- 11
- 5
- 14
- 8
- 11
- 16
- 11

Determine a Mediana e a Média

- Ordenar
- 5 11
- 8 14
- 8 16
- 11
- 11

- Posição: (n+1)/2
- (8+1)/2
- 4,5
- Med=11

- Média ?

E se ao invés do 16 houvesse o Número 99?

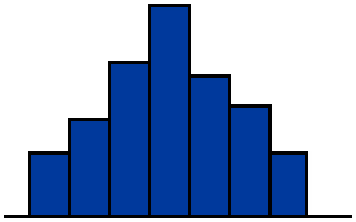
- 8
- 11
- 5
- 14
- 8
- 11
- ~~16~~ → 99
- 11

Mediana = ?

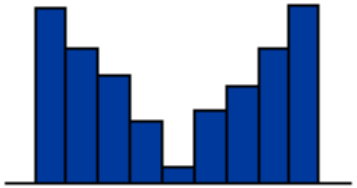
Média = ?

Moda

- A Moda é o valor que mais se repete em um conjunto de dados. Frequência Máxima, ou Valor mais frequente.
  - Pode-se ter:



uma moda:unimodal



duas modas: bimodal

Moda

- Para o nosso Exemplo (2 slides atrás), determine a moda para o número de faturas.

DADOS AGRUPADOS

- Quando os dados estão agrupados em uma distribuição de frequências, tem-se o valor individual da variável.

Esperança Estatística

Considere a seguinte variável discreta e sua respectiva função de probabilidade.

x	0	1	2
p(x)	½	¼	¼

Assim, teremos a esperança $E(X) = (0.1/2) + (1.1/4) + (2.1/4) = 3/4$

Exemplo 1 – por Idade

Tabela 1 – Pacientes com hipertensão, segundo a idade em anos completos

Idade em anos completos	Número de indivíduos (frequência - f_i)	$x_i . f_i$	Idade em anos completos	Número de indivíduos (frequência - f_i)	$x_i . f_i$
22	1	22	47	1	47
27	1	27	48	1	48
30	1	30	50	2	100
31	1	31	53	3	159
34	1	34	56	1	56
35	3	105	58	1	58
36	5	180	59	2	118
40	1	40	60	1	60
42	1	42	61	1	61
43	1	43	63	1	63
44	2	88	65	3	195
45	1	45	67	2	134
46	2	92			
			Total	40	1 878

$$\bar{X} = \frac{\sum_{i=1}^{k} x_i . f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} x_i . f_i}{n}$$

$$\bar{X} = \frac{22+27+30+31+\cdots+65+65+65+67+67}{40}$$

$$\bar{X} = \frac{22.1+27.1+30.1+31.1+\cdots+65.3+67.2}{40} = \frac{1878}{40} = 46,95 \text{ anos} = 46 \text{ anos}$$

e 11 meses, ou seja, a idade média dos hipertensos é igual a 46 anos e 11 meses.

Exemplo 2 – por faixas

Tabela 2 – Pacientes com hipertensão, segundo a idade em anos completos

Classes	Ponto Médio (Pm_i)	Número de pacientes (f_i)	Produto $Pm_i . f_i$
20 |— 30	25	2	50
30 |— 40	35	11	385
40 |— 50	45	10	450
50 |— 60	55	9	495
60 |— 70	65	8	520
Total		40	1 900

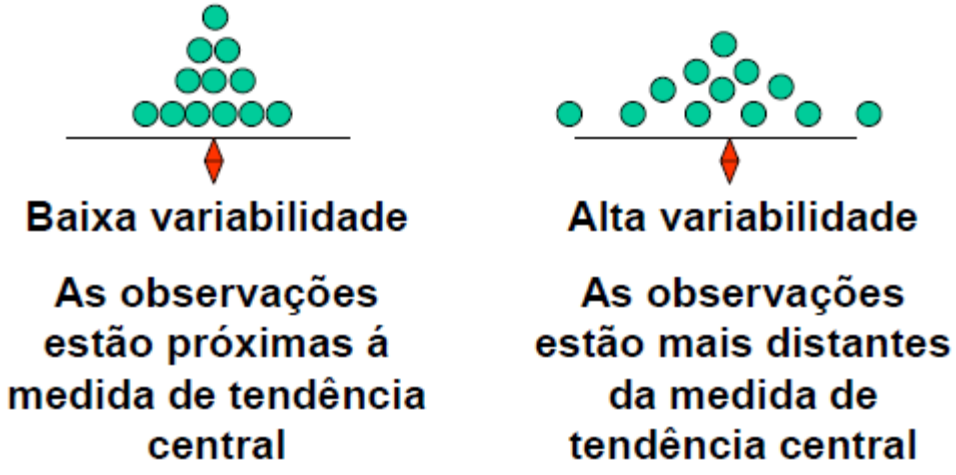
$$\overline{X} = \frac{\sum_{i=1}^{k} PM_i . f_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} PM_i . f_i}{n}$$

$$\overline{X} = \frac{1\,900}{40} = 47{,}5 \text{ anos} = 47 \text{ anos e } 6 \text{ meses ou } 47 \text{ anos (completos).}$$

Observação: a idade média calculada a partir dos dados da tabela 2 não coincide com a idade média verdadeira dos 40 hipertensos, calculada a partir dos dados da Tabela 1. Isso se deve ao fato de ter sido suposto, para o cálculo da média aritmética com os dados da Tabela 2, que todos os indivíduos de uma determinada classe tinham a idade dada pelo ponto médio da classe, o que, em geral, não corresponde à realidade.

MEDIDAS DE DISPERSÃO

O grau no qual os dados numéricos tendem a dispersar-se em torno de um valor médio chama-se variação ou dispersão desses dados. Essa variação pode ser calculada através das chamadas medidas de dispersão ou de variabilidade. As mais importantes medidas de dispersão são: Variância absoluta, Desvio padrão e Coeficiente de Variação.



Amplitude de Variação (R)

Uma das medidas mais elementares é a *amplitude*, a qual é definida como sendo a diferença entre o maior e o menor valor do conjunto de dados:

$$R = x_{max} - x_{min}$$

Evidentemente que essa medida é muito precária, pois a amplitude não dá informe algum a respeito da maneira pela qual os valores se distribuem entre os valores extremos.

Por exemplo, nos dois conjuntos de valores:

4, 6, 6, 6, 8

4, 5, 6, 7, 8

Amplitude Semi-Quartil ou Desvio Quartil

Esta medida, que se baseia na posição ocupada pelos 50% centrais da distribuição, é definida por:

$$Q = \frac{Q_3 - Q_1}{2},$$

onde Q_1 e Q_3 são o primeiro e o terceiro quartis.

Essa medida, conquanto se baseia também em apenas dois valores, apresenta sobre a anterior a vantagem de não estar tão sujeita às flutuações amostrais quanto os valores extremos.

A dispersão poderia ser medida pela *amplitude quartil*, ou seja, $Q_3 - Q_1$; todavia, a divisão por 2 dá a distância média pela qual os quartis se desviam da mediana.

Soma dos Desvios em Relação à Média

Tabela 3 – Diferenças entre as observações e a respectiva média

x_i	$(x_i - \overline{X})$
1	$1 - 3 = -2$
2	$2 - 3 = -1$
3	$3 - 3 = 0$
4	$4 - 3 = 1$
5	$5 - 3 = 2$
Total	$\Sigma (x_i - \overline{X}) = 0$

VARIÂNCIA

- A variância baseia-se nos desvios em torno da média aritmética, porém determinando a média aritmética dos quadrados dos desvios.

x_i	$(x_i - \bar{X})$	$(x_i - \bar{X})^2$
1	$1 - 3 = -2$	4
2	$2 - 3 = -1$	1
3	$3 - 3 = 0$	0
4	$4 - 3 = 1$	1
5	$5 - 3 = 2$	4
Total	$\Sigma (x_i - \bar{X}) = 0$	10

e a medida de variabilidade seria

$$\frac{\Sigma(x_i - \bar{X})^2}{n} = \frac{10}{5} = 2$$

a qual recebe o nome de *variância* (*Var* ou σ^2).

- Assim, representando a variância por s2, temos:

$$s^2 = \frac{\sum (x_i - x)^2}{n - 1}$$

VARIÂNCIA

Para as amostras 3, 4, 5, 6, 7 e

1, 3, 5, 7, 9

As variâncias seriam:

$S_1^2 = (3-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2/4$ **$S_1^2 = 2,5$**

$S_2^2 = (1-5)^2 + (3-5)^2 + (5-5)^2 + (7-5)^2 + (9-5)^2/4$ **$S_2^2 = 10$**

A amostra 3, 4, 5, 6, 7 é mais homogênea.

DESVIO PADRÃO

Desvio Padrão é a Raiz quadrada da Variância, ou seja,

$$s = \sqrt{\frac{\sum (x_i - x)^2}{n-1}}$$

- O desvio-padrão é uma quantidade essencialmente positiva.
- O desvio-padrão só é *nulo* se todos os valores da distribuição forem iguais entre si, isto é, se não houver variabilidade.
- O desvio-padrão é da mesma natureza da variável X e depende também de sua magnitude.

O DESVIO PADRÃO DAS AMOSTRAS 3, 4, 5, 6, 7 e 1, 3, 5, 7, 9 seria:

S1= $\sqrt{2,5}$ **=1,58**

S2= $\sqrt{10}$ **=3,16**

COEFICIENTE DE VARIAÇÃO

Para comparar duas distribuições quanto à variabilidade, deve-se usar *medidas de variabilidade relativa*, tais como o *coeficiente de variação de Pearson* (*CV*), o qual é dado por:

- O desvio padrão por si só tem suas limitações. Assim, um desvio padrão de duas unidades pode ser considerado pequeno para uma série de valores cujo valor médio é 200; no entanto, se a média for igual a 20, o mesmo não pode ser dito. Para contornar essa dificuldade, podemos caracterizar a dispersão ou variabilidade dos dados em termos relativos a seu valor médio, medida essa denominada coeficiente de variação (CV):

$$CV = \frac{s}{\bar{x}} \times 100$$

Exemplo de Coeficiente de Variação

Exemplo: para duas emissões de ações ordinárias da indústria eletrônica, o preço médio diário, no fechamento dos negócios, durante um período de um mês, para as ações A, foi de R$150,00 com um desvio-padrão de R$5,00. Para as ações B, o preço médio foi de R$50,00 com um desvio-padrão de R$3,00. Em termos de comparação absoluta, a variabilidade do preço das ações A foi maior, devido ao desvio-padrão maior. Mas em relação ao nível de preço, devem ser comparados os respectivos coeficientes de variação:

$$CV(A) = \frac{S_A}{\bar{X}_A} = \frac{5}{150} = 0,033 \text{ ou } 3,3\%$$

$$CV(B) = \frac{S_B}{\bar{X}_B} = \frac{3}{50} = 0,060 \text{ ou } 6\%$$

Portanto, relativamente ao nível médio de preços das ações, podemos concluir que o preço da ação B é quase duas vezes mais variável que o preço da ação A.

Exercícios 1

1. Em uma determinada empresa X, a média dos salários é 10 000 unidades monetárias e o 3º quartil é 5 000. Pergunta-se:

   a) Se você se apresentasse como candidato a esta empresa e se o seu salário fosse escolhido ao acaso entre todos os possíveis salários, o que seria mais provável: ganhar mais ou menos que 5 000 unidades monetárias? Justifique.

   b) Suponha que na empresa Y a média dos salários é 7 000 unidades monetárias e a variância é praticamente zero, e lá o seu salário também seria escolhido ao acaso. Em qual empresa você se apresentaria para procurar emprego X ou Y? Justifique.

2. A média aritmética é a razão entre:

   a) o número de valores e o somatório deles.

   b) o somatório dos valores e o número deles.

   c) os valores extremos.

   d) os dois valores centrais.

   e) nenhuma das alternativas anteriores.

3. Na série 60, 90, 80, 60, 50 a moda é:

   a) 50

   b) 60

   c) 66

   d) 90

   e) nenhuma das anteriores.

4. A estatística que possui o mesmo número de valores abaixo e acima dela é:

   a) a moda.

   b) a média.

   c) a mediana.

   d) o elemento mediano.

   e) nenhuma das anteriores.

Exercícios 2

5. A soma dos desvios entre cada valor e a média sempre será:

   a) positiva.

   b) negativa.

   c) zero.

   d) diferente de zero.

   e) nenhuma das alternativas anteriores.

6. Considere a série 6, 5, 7, 8, 9 o valor 7 será:

   a) a média e a moda.

   b) a média e a mediana.

   c) a mediana e a moda.

   d) a média, a mediana e a moda.

   e) nenhuma das alternativas anteriores.

7. Quando desejamos verificar a questão de uma prova que apresentou maior número de erros, utilizamos:

   a) moda.

   b) média.

   c) mediana.

   d) qualquer das anteriores.

   e) nenhuma das anteriores.

8. O coeficiente de variação é uma estatística denotada pela razão entre:

   a) desvio-padrão e média.

   b) média e desvio-padrão.

   c) mediana e amplitude interquartílica.

   d) desvio-padrão e moda.

   e) nenhuma das alternativas anteriores.

Exercícios 3

9. Uma prova de estatística foi aplicada para duas turmas. Os resultados seguem abaixo

Turma 1: média = 5 e desvio-padrão = 2,5

Turma 2: média = 4 e desvio-padrão = 2,0

Com esses resultados podemos afirmar:

a) a turma 2 apresentou maior dispersão absoluta.

b) a dispersão relativa é igual à dispersão absoluta.

c) tanto a dispersão absoluta quanto a relativa são maiores para a turma 2.

d) a dispersão absoluta da turma 1 é maior que a turma 2, mas em termos relativos as duas turmas não diferem quanto ao grau de dispersão das notas.

e) nenhuma das alternativas anteriores.

10. Uma empresa possui dois serventes recebendo salários de R$250,00 cada um, quatro auxiliares recebendo R$600,00 cada um, um chefe com salário de R$1.000,00 e três técnicos recebendo R$2.200,00 cada um. O salário médio será:

a) R$1.050,00.

b) R$1.012,50.

c) R$405,00.

d) R$245,00.

e) nenhuma das alternativas anteriores.

11. O cálculo da variância supõe o conhecimento da:

a) média.

b) mediana.

c) moda.

d) ponto médio.

e) desvio-padrão.