Análise de dados quantitativos

Análise exploratória, medidas de posição e medidas de dispersão.

Prof. Paulo Henrique Coelho Maranhão

Propósito

Compreender as principais ferramentas de análise exploratória de dados e as medidas de posição e dispersão.

Preparação

Antes de iniciar o conteúdo, tenha em mãos uma calculadora científica ou use a calculadora de seu smartphone/computador.

Objetivos

- Reconhecer as ferramentas de análise exploratória de dados.
- Analisar as medidas de posição ou tendência central.
- Descrever as medidas de dispersão ou variabilidade.

Introdução

Iniciaremos o estudo de uma ferramenta importante da estatística e conheceremos os dados com os quais vamos trabalhar. Veremos desde os conceitos básicos — tais como classificação de variáveis — até as principais ferramentas para apresentar e sintetizar os dados, como distribuição de frequência e representações gráficas.

Além disso, medidas de posição ou tendência central são medidas que visam representar os fenômenos por seus valores centrais, em torno dos quais tendem a concentrar-se os dados. Apresentaremos essas medidas de acordo com os conceitos vistos, ou seja, considerando que os dados podem apresentar-se: agrupados (quando estão dispostos em uma distribuição de frequência) e não agrupados (quando estão dispostos em rol ou dados brutos.

Para concluir, veremos as medidas de dispersão ou variabilidade, as quais possuem grande utilidade, pois avaliam o grau de dispersão dos dados em torno da média. Elas servem para verificar a representatividade da média. Vamos analisar os conceitos das principais medidas de dispersão, tais como variância, desvio-padrão e coeficiente de variação.

De iníco, confira o vídeo, a seguir, que resume os principais assuntos abordados nesse conteúdo, reconhecidos também com *Análise de Métodos Quantitativos*!

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Análise exploratória de dados

Fala, mestre!

De acordo com um levantamento da Associação Brasileira de Empresas de Software (ABES) em parceria com a IDC, até 2023, mais de 30% das empresas brasileiras utilizarão inteligência artificial (IA) em seus negócios. Esta tecnologia já é um recurso importante para muitas empresas, ajudando na automação de processos, contato com consumidores e identificação de experiências do usuário. A IA agrega uma camada adicional de inteligência, possibilitando maior qualidade nas análises e eficiência nos sistemas. Exemplos de uso incluem recomendações de produtos em plataformas como Amazon e Netflix. No Brasil, a pandemia acelerou a adoção digital, com empresas de diversos portes implementando tecnologias para melhorar suas operações e gerar valor. A transformação digital é vista como uma virada econômica essencial para a competitividade e sobrevivência das empresas no futuro.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Definição

A análise exploratória de dados é a parte da estatística responsável pelo primeiro contato com as informações. Essa técnica nos dá um indicativo de como os dados estão distribuídos. Além disso, é útil na detecção de erros, de valores extremos (*outliers*), na verificação de suposições relacionadas à inferência estatística, na seleção preliminar de modelos estatísticos, entre outras utilidades. Aqui veremos os principais conceitos e ferramentas para a exploração correta dos dados.

Métodos quantitativos: análise exploratória de dados

Confira o vídeo, a seguir, sobre as ferramentas de análise exploratória de dados.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Classificação das variáveis

Variáveis são características de interesse em um estudo qualquer. Elas podem ser classificadas em:

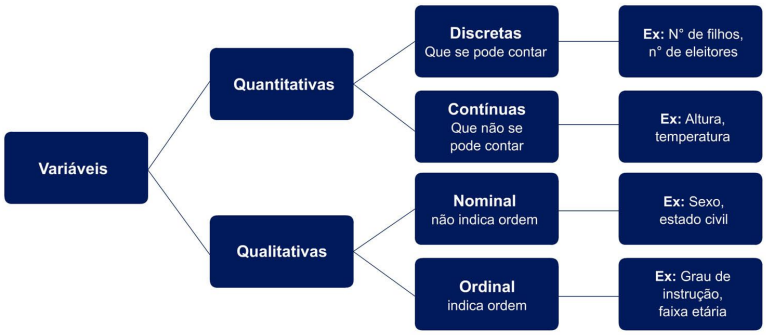
Quantitativas

Quando assumem valores numéricos.

Qualitativas

Quando seus possíveis valores não são numéricos.

A seguir veremos de forma resumida como as variáveis são classificadas:



Fluxograma de classificação das variáveis

Conceitos básicos

Veremos, neste momento, uma série de conceitos que serão importantes, tanto para o melhor entendimento deste conteúdo como também para os seguintes:

População

Conjunto de indivíduos ou objetos com pelo menos uma característica em comum.

Amostra

Uma parte da população.

Dados brutos

Conjunto de dados que não tem uma ordem aparente.

Rol

Conjunto de dados que tem um ordenamento, seja crescente ou decrescente.

Amplitude total

Diferença entre o maior e o menor valor observado no conjunto de dados.

Distribuições de frequência

A distribuição de frequência é uma das formas mais simples e úteis de resumir um conjunto de dados.

Resumindo

Nada mais é do que a apresentação dos dados em classes às suas respectivas frequências absolutas.

As classes são divisões dos valores da variável em estudo.

Exemplo 1

A distribuição de frequência a seguir representa as notas na disciplina de estatística em uma turma de 40 alunos:

Classe (Notas)	F_i	f_i%
$0 \vdash 2$	1	2,5
$2 \vdash 4$	5	12,5
$4 \vdash 6$	12	30,0
$6 \vdash 8$	15	37,5
$8 \vdash 10$	7	17,5
Soma	40	100

Exemplo 2

A próxima distribuição de frequência refere-se à quantidade de famílias que receberam auxílio escolar por número de filhos.

Classe (N° de filhos)	F_i
1	52
2	38
3	18
4	12
Soma	120

Elementos da distribuição de frequência

Veja o conceito e características de cada elemento a seguir:

Limites de classe

São as várias formas de expressar os limites de classe em uma distribuição de frequência. O limite à esquerda é chamado de limite inferior (Li) e o limite à direita é chamado de limite superior (Ls) da classe. Vejamos alguns exemplos:

- Li |---| Ls: indica uma classe que é fechada à esquerda e à direita, em que os limites inferior e superior estão incluídos na classe.
- Li |--- Ls: indica uma classe que é fechada à esquerda e aberta à direita, ou seja, o limite inferior está incluído na classe, mas o limite superior não.
- Li ---| Ls: indica uma classe que é fechada à direita e aberta à esquerda, ou seja, o limite superior está incluído na classe, mas o limite inferior não.

Dentre os limites de classes apresentados, o mais utilizado é o da letra B, isto é, fechada à esquerda e aberta à direita.

Amplitude de classe (AC)

É a diferença entre o limite superior e o limite inferior da classe.

Ponto médio da classe (X_i)

É a média aritmética entre o limite inferior e o limite superior da classe: 2 |---4. Logo o ponto médio dessa classe será:

$$X_i = \frac{2+4}{2} = 3$$

No cálculo do ponto médio da classe (X_i), os limites superior e inferior são considerados, independentemente da classe ser fechada ou aberta nos limites L_i ou L_s.

Frequência relativa (F_i%)

É a frequência relativa dada por: $f_i\% = \frac{F_i}{n}$, em que $n = \sum F_i$. Você pode ver um exemplo clicando aqui.

Observe que para obter a frequência relativa de cada classe, basta dividir a frequência absoluta de cada uma por n, que é o tamanho do conjunto de dados ou tamanho da amostra. Então,

e assim por diante. Veja que o processo é análogo para todas as classes.

Frequência Acumulada (F_{AC})

É o acúmulo das frequências absolutas. A partir da primeira frequência, somam-se as respectivas frequências absolutas. Você pode ver um exemplo clicando aqui.

Veja que, para a primeira classe, a frequência acumulada é igual à frequência absoluta. A partir daí, começamos a somar as frequências absolutas, de forma que a acumulada da segunda classe é a frequência absoluta da primeira classe mais a frequência absoluta da segunda classe, ou seja, $F_{ac}\left(2^{a} \text{ classe }\right)=F_{1}+F_{2}=1+5=6$.

Para determinar a frequência acumulada da terceira classe, somamos a frequência acumulada da segunda com a frequência absoluta da terceira, ou seja, $F_{ac}\left(3^{a} \text{ classe }\right)=F_{ac}\left(2^{a} \text{ classe }\right)+F_{3}=6+12=18$, e assim por diante.

Como construir uma distribuição de frequência

Muitas vezes, a distribuição dos dados é obtida de forma que estão simplesmente dispostos em rol ou mesmo como dados brutos, ou seja, sem ordem aparente.

Dica

Quando a quantidade de dados é muito grande, a melhor forma de apresentá-los é por meio de uma tabela.

No intuito de melhorar a apresentação, é comum dispor os dados em uma distribuição de frequência. Desse modo, veremos a seguir alguns passos práticos para construir essa distribuição de frequência:

1

Passo 1: Determinar o número de classes (k)

Para se calcular o número de classes de uma distribuição de frequência, utilizaremos a seguinte fórmula:

$$
k=\sqrt{n}
$$

Na qual n é o tamanho do conjunto de dados ou da amostra.

2

Passo 2: Determinar a Amplitude Total (AT)

A amplitude de classe é a diferença entre o maior e o menor valor no conjunto de dados.

3 Passo 3: Determinar a Amplitude de Classe (Ac)

A amplitude de classe é obtida por:

$$Ac = \frac{AT}{k}$$

4

Passo 4: Construção da distribuição de frequência

A partir dos três elementos vistos nos passos 1, 2 e 3 iniciaremos a construção da distribuição de frequência pelo menor valor do conjunto de dados, que será o limite inferior da primeira classe.

A partir desse valor, acrescentamos a amplitude de classe para obter o limite superior da primeira classe. Esse limite superior da primeira classe será o limite inferior da segunda, independentemente do tipo de classe escolhida. O limite superior da segunda classe será o limite inferior da segunda somada à amplitude de classe.

O limite superior da segunda classe será o limite inferior da terceira e assim por diante. É comum na última classe usarmos a classe fechada tanto no limite inferior como no limite superior.

Exemplo

O rol a seguir representa a altura (em centímetros) de 26 jogadores de uma equipe de futebol.

160 165 166 168 170 170 172 174 175 175 175 178 180 180 182 183 185 185 187 188 188 190 191 195 198 200

Vamos construir uma distribuição de frequência das alturas dos jogadores dessa equipe de futebol. É só seguir os seguintes passos:

Passo 1

Nesse caso, usaremos a regra de arredondamento e consideraremos k igual a 5, mas em alguns casos é interessante arredondar para cima, sempre verificando se a distribuição de frequência contempla todo o conjunto de dados.

A equação fica assim:

$$k = \sqrt{n} = \sqrt{26} = 5,09 \approx 5$$

Passo 2

Em seguida, temos o seguinte: $AT = 200 - 160 = 40$

Passo 3

Depois, a equação a seguir: $Ac = \frac{AT}{k} = \frac{40}{5,09} = 7,86 \approx 8$

Passo 4

Por fim, construindo a distribuição de frequência, temos o seguinte resultado:

Classe (Alturas)	F_i
160 ⊢ 168	3

Classe (Alturas)	F_i
$168 \vdash 176$	8
$176 \vdash 184$	5
$184 \vdash 192$	7
$192 \vdash 200$	3
Soma	26

Representações gráficas

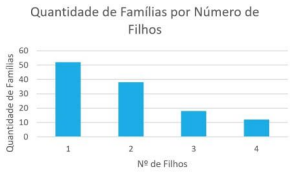
Embora atualmente não seja mais necessário saber as técnicas de construção dos gráficos como se fazia há alguns anos, é importante conhecer os principais tipos de gráficos e quando devem ser empregados, pois ainda hoje são ferramentas indispensáveis para visualização e interpretação de dados.

Por isso, a seguir, veremos alguns meios de sintetizar os dados de uma forma gráfica:

Gráfico de barras ou colunas

É o tipo de gráfico mais utilizado em geral, pois serve para representar quaisquer dados quantitativos.

Exemplo: Considere a distribuição de frequência referente à quantidade de famílias que receberam auxílio escolar por número de filhos.



Histograma

É o gráfico típico da distribuição de frequência. A diferença desse gráfico para o gráfico de barras ou colunas se dá pelo fato de as colunas apresentarem-se justapostas, ou seja, sem espaçamento entre elas. Em geral, a abcissa desse gráfico é representada pelas classes e a ordem é representada pela frequência absoluta ou relativa.

Exemplo: Considere a distribuição de frequência a seguir, que representa as notas na disciplina de estatística em uma turma de 40 alunos.

Gráfico de linhas

É o gráfico mais apropriado quando trabalhamos com uma série de tempo.

Exemplo: Número de acidentes por mês ao longo de um ano.

Setor

É o gráfico mais apropriado quando trabalhamos com porcentagens.

Exemplo: Em uma pesquisa de satisfação sobre determinado produto, 55% dos entrevistados disseram que estavam satisfeitos, 35% disseram que estavam insatisfeitos e 10% disseram que eram indiferentes.

Caixa (boxplot)

É um dos gráficos mais utilizados atualmente, visto que traz várias informações sobre o conjunto de dados. Com esse gráfico é possível verificar a tendência central, a variabilidade e a simetria da distribuição dos dados, conceitos esses que serão vistos de forma mais detalhada posteriormente.

Outra vantagem desse gráfico é que podemos observar a presença de valores atípicos (*outliers*). Para isso é necessário determinar o intervalo interquartil (IQR), que é a diferença entre o 3º e o 1º quartil. Multiplicando esse IQR por 1,5, obtemos a faixa interquartil. Quando subtraímos o 1º quartil dessa faixa e somamos o 3º quartil a esta, encontramos o intervalo no qual seria comum a variação dos dados. Valores acima desse intervalo são considerados *outliers*.

Exemplo: Considere os dados referentes aos preços de aluguéis de imóveis (em reais) em certo bairro do Rio de Janeiro.

1000 1500 1800 2000 2200 2500 2600 3000 3500 7000

Observe que no gráfico a seguir, temos informações como: o primeiro e o terceiro quartis, a mediana e a média. Os traços abaixo do primeiro quartil e acima do terceiro quartil representam o menor e o maior valor dentro do intervalo normal de variação dos dados.

Como vimos, o IQR é calculado subtraindo o terceiro quartil do primeiro quartil, que, nesse caso, é igual a 1400. Note que a faixa interquartil (1,5IQR) é igual a 2100; logo, se somarmos o 3º quartil a essa faixa interquartil, temos o valor limite (5225), que seria considerado normal para variação dos aluguéis. No entanto, como o aluguel de R$7.000,00 reais está acima de R$5.225,00 podemos dizer que se trata de um valor atípico.

Mão na massa

Questão 1

Considere dados sobre o peso de navios. Essa variável é classificada em:

A

Quantitativa discreta

B

Quantitativa contínua

C

Qualitativa nominal

D

Qualitativa ordinal

E

Qualitativa contínua

A alternativa B está correta.

Observe que o peso é uma variável não contável, assumindo valores que pertencem ao conjunto dos números reais. Portanto, é uma variável quantitativa contínua.

Questão 2

Analisando as alternativas a seguir, qual das alternativas é falsa?

A

População é um conjunto de indivíduos com pelo menos uma característica em comum.

B

Amostra é uma porção da população.

C

Rol é um conjunto de dados brutos ordenados.

D

Dados brutos é um conjunto de dados dispostos sem ordem aparente.

E

Distribuição de frequência é o arranjo dos dados em ordem decrescente.

A alternativa E está correta.

A distribuição de frequência é o arranjo dos dados em classes com suas respectivas frequências absolutas.

Questão 3

Dados sobre atendimentos médicos por faixa etária foram coletados e organizados na seguinte distribuição de frequência:

Classes	$0 \vdash 10$	$10 \vdash 20$	$20 \vdash 40$	$40 \vdash 80$	Soma
F_i	280	320	180	220	1000

Determine o ponto médio da 3ª classe.

A

20

B

25

C

30

D

35

E

40

A alternativa C está correta.

Veja que a terceira classe tem limite inferior e limite superior de classe, respectivamente igual a 20 e 30. Portanto, como o ponto médio da classe é a média aritmética entre os limites inferior e superior, temos que Xi é igual a 30.

Questão 4

Considerando a questão anterior, qual a frequência acumulada da terceira classe?

A

280

B

600

C

680

D

780

E

1000

A alternativa D está correta.

Classes (Notas)	0 ⊢ 10	10 ⊢ 20	20 ⊢ 40	40 ⊢ 80	Soma
F_i	280	320	**180**	220	1000
F_{ac}	280	600	**780**	1000	-

Para determinar a frequência acumulada da terceira classe, lembre-se de que, para a primeira classe, a frequência acumulada é igual à frequência absoluta. A partir daí, começamos a somar as frequências absolutas, de forma que a frequência acumulada da segunda classe é a frequência absoluta da primeira classe mais a frequência absoluta da segunda classe, ou seja, $F_{ac}\left(2^a \text{ classe }\right)=F_1+F_2=280+320=600$.

Para determinar a frequência acumulada da terceira classe somamos a frequência acumulada da segunda classe com a frequência absoluta da terceira classe, ou seja,

$$F_{ac}\left(3^a \text{ classe }\right)=F_{ac}\left(2^a \text{ classe }\right)+F_3=600+180=780$$

Questão 5

Julgue as alternativas a seguir e assinale a alternativa verdadeira:

A

O histograma é o gráfico típico das distribuições de frequências.

B

O gráfico ideal para porcentagens é o de barras.

C

Não há diferença entre histograma e gráfico de barras.

D

O gráfico de barras é o mais apropriado para séries de tempo.

E

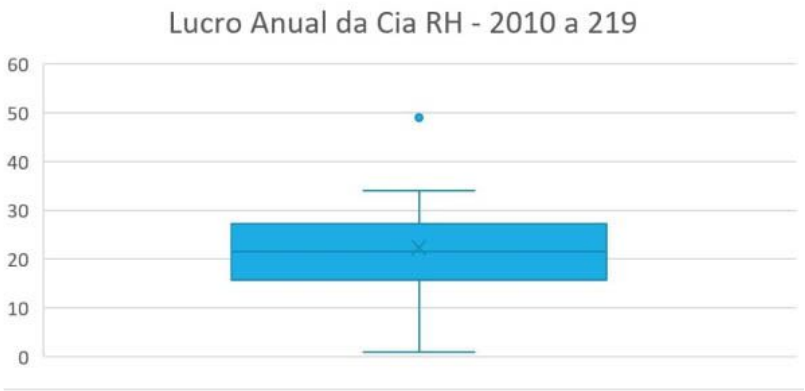
O gráfico de setor não se aplica a porcentagens.

A alternativa A está correta.

O histograma é um gráfico de colunas ou barras, que tem por objetivo ilustrar como determinada amostra ou população de dados está distribuída.

Questão 6

De acordo com o diagrama de caixa a seguir, julgue a alternativa verdadeira:



A

O primeiro quartil é aproximadamente 16.

B

A mediana é a aproximadamente 25.

C

Os dados estão simétricos.

D

O maior valor do conjunto de dados é 34.

E

O ponto em azul é considerado um *outlier* simplesmente porque está fora da caixa.

A alternativa A está correta.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Teoria na prática

O conjunto de dados, a seguir, representa o número de horas extras mensais trabalhadas por 17 funcionários de um banco de investimentos:

10 12 12 14 15 16 16 18 19 20 20 21 24 24 25 28 30

Organize os dados em uma distribuição de frequência.

Chave de resposta

Assista ao vídeo a seguir e entenda como, a partir de um conjunto de dados, construímos a distribuição de frequência.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Dados sobre evasão escolar em determinado município estão exibidos na distribuição de frequência a seguir:

Classes	$0 \vdash 6$	$6 \vdash 10$	$10 \vdash 14$	$14 \vdash 17$	Soma
F_i	20	44	64	72	200

Determine a amplitude entre as frequências relativas.

A

4

B

10

C

16

D

26

E

36

A alternativa D está correta.

Observe que, inicialmente, devemos determinar as frequências relativas com base nos dados da distribuição de frequência. Lembre-se de que a frequência relativa para cada classe é dada por: $f_i\% = \frac{F_i}{n} \times 100$.

Classes	0 ⊢ 6	6 ⊢ 10	10 ⊢ 14	14 ⊢ 17	Soma
F_i	20	44	64	72	200
$f_{1\%}$	10	22	32	36	100

Como desejamos a amplitude entre as frequências relativas, basta calcular a diferença entre a maior e a menor frequência relativa. Assim, a amplitude entre as frequências relativas é igual a 36-10=26.

Questão 2

Considerando novamente a distribuição de frequência da questão anterior:

Classes	0 ⊢ 6	6 ⊢ 10	10 ⊢ 14	14 ⊢ 17	Soma
F_i	20	44	64	72	200

Qual o gráfico mais apropriado para representar esse conjunto de dados?

A

Barra

B

Histograma

C

Linha

D

Setor

E

Caixa

A alternativa B está correta.

Vimos que o gráfico que representa os dados em distribuição de frequência é o histograma, que nada mais é do que um gráfico de barras justapostas, cujas classes se encontram ao longo do seu eixo horizontal (abcissa) e as frequências absolutas ou relativas são apresentadas no eixo vertical (ordenada).

Medidas de posição ou tendência central

São medidas que visam representar os fenômenos por seus valores centrais, em torno dos quais tendem a concentrar-se os dados. Apresentaremos essas medidas de acordo com os conceitos vistos no módulo anterior, ou seja, considerando que os dados podem apresentar-se:

Agrupados

Quando estão dispostos em uma distribuição de frequência.

Não agrupados

Quando estão dispostos em rol ou dados brutos.

Medidas de Posição: Média, Mediana e Moda

Confira o vídeo, a seguir, sobre as medidas de posição ou tendência central.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

A seguir, você verá as principais medidas de posição.

Média \((X)\)

Para dados não agrupados

Média é a medida de posição mais conhecida e mais usada na prática para verificar o comportamento central dos dados. Vejamos a definição da média considerando a forma como os dados são apresentados.

Quando os dados estão dispostos como dados brutos ou rol, a média é definida por:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Exemplo

Vamos determinar a média para o seguinte Rol de dados: 1, 2, 3, 4, 5. Veja sua solução a seguir:

Resolução

$\bar{X} = \frac{1+2+3+4+5}{5} = 3$

Logo, a média é igual a 3.

Para dados agrupados

Quando os dados se apresentam em distribuição de frequência, a média é definida por:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i F_i}{n}$$

Em que:

- X_i é ponto médio da classe i;
- F_i é a frequência absoluta da classe i;
- n é tamanho do conjunto de dados ou da amostra.

Exemplo

Considere os dados sobre o peso (em Kg) de recém-nascidos de certa maternidade, dispostos na distribuição de frequência abaixo:

Classe	F_i
2,0 ⊢ 2,5	2
2,5 ⊢ 3,0	4
3,0 ⊢ 3,5	7
3,5 ⊢ 4,0	5
4,0 ⊢ 4,5	5
4,5 ⊢ 5,0	7
Soma	30

Veja sua solução a seguir:

Resolução:

Vimos que, para calcular a média para dados agrupados, usamos a seguinte expressão:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i F_i}{n}$$

Dessa forma, precisamos determinar o produto de $\frac{21}{4}\frac{5}{8}$ $\frac{21}{4}\frac{5}{8}$ $\frac{21}{4}\frac{5}{8}$. Assim, podemos utilizar a própria distribuição de frequência acima para obter esse produto. Daí,

Classe	F_i	X_i	X_i, F_i
2,0 ⊢ 2,5	2	2,25	4,5
2,5 ⊢ 3,0	4	2,75	11
3,0 ⊢ 3,5	7	3,25	22,75

Classe	F_i	X_i	X_i, F_i
3,5 ⊢ 4,0	5	3,75	18,75
4,0 ⊢ 4,5	5	4,25	21,25
4,5 ⊢ 5,0	7	4,75	33,25
Soma	30	-	111,5

Portanto, a média é calculada da seguinte forma:

$$< br > \bar{X} = \frac{\sum_{i=1}^{n} X_i F_i}{n} = \frac{111,5}{30} = 3,72 < br >$$

Logo, a média do peso dos recém-nascidos dessa maternidade é de aproximadamente 3,72 kg.

Mediana \(\left(M_D\right)\)

Para dados não agrupados

Disposto o conjunto de dados em ordem crescente ou decrescente, a mediana é o elemento que ocupa a posição central, isto é, divide o conjunto de dados em duas partes iguais, de forma que metade dos dados está acima e a outra metade está abaixo da mediana.

Para o cálculo da mediana para dados não agrupados, serão levados em consideração os dois seguintes fatores:

1. Se o tamanho da amostra é ímpar ou par;
2. Se refere ao elemento mediano, que é o elemento que nos dá a posição ocupada pela mediana.

Para os dois casos, a seguir, temos o seguinte:

Se n é ímpar

A mediana será o valor dado pela posição determinada pelo elemento mediano.

$$E_{M_d} = \frac{n+1}{2}$$

Exemplo: Considere o seguinte conjunto de dados: 2, 5, 7, 9, 10. Assim, como n = 5, temos:

$$n = 5 \rightarrow E_{M_d} = \frac{5+1}{2} = 3 \rightarrow M_d = 7$$

Veja que o elemento que ocupa a terceira posição no conjunto de dados é justamente o número 7. Portanto, a mediana é igual a 7.

Se n é par

A mediana será a média aritmética entre as medianas obtidas pela posição dos elementos medianos $E_{M_{d_1}}$ e $E_{M_{d_2}}$ que são determinados da seguinte forma:

$$< br > \left.\begin{array}{c} < br > E_{M_{d_1}} = \frac{n}{2} \rightarrow M_{d_1} \\ < br > E_{M_{d_2}} = \frac{n}{2} + 1 \rightarrow M_{d_2} < br > \end{array}\right\} M_d = \frac{M_{d_1} + M_{d_2}}{2} < br >$$

Exemplo: 3, 4, 6, 8, 10, 11

$$< br > n = 6 \rightarrow \left\{\begin{array}{c} < br > E_{M_{d_1}} = \frac{n}{2} = 3 \rightarrow M_{d_1} = 6 \\ < br > E_{M_{d_2}} = \frac{n}{2} + 1 = 4 \rightarrow M_{d_2} = 8 < br > \end{array}\right. \rightarrow M_d = \frac{6+8}{2} = 7 < br >$$

Observe que, nesse caso, o primeiro elemento mediano é o elemento que ocupa a terceira posição e o segundo elemento mediano é o elemento que ocupa a quarta posição, o que equivale aos valores 6 e 8, respectivamente. Então, a mediana é a média aritmética entre esses valores (6 e 8), que é igual a 7.

Para dados agrupados

Para o cálculo da mediana para dados agrupados, vamos seguir os seguintes passos:

1

Passo 1

Determinar o elemento mediano.

$$E_{M_d} = \frac{n}{2}$$

2

Passo 2

Determinar a classe mediana ($\square_{\square\square}$), que é a classe que contém o elemento mediano.

3

Passo 3

Aplicar a fórmula:

$$M_d = L_{M_d} + \left(\frac{E_{M_d} - F_{aac}}{F_{M_d}}\right) \times h$$

Em que:

$\square_{\square\square}$= Limite inferior da classe mediana.
$\square_{\square\square}$= Frequência acumulada anterior à classe mediana.
$\square_{\square\square}$= Frequência absoluta da classe mediana.
h= Amplitude da classe mediana (diferença entre as amplitudes de classe superior e inferior).

Exemplo

Considerando os dados dos pesos dos recém-nascidos na tabela adiante, vamos determinar o valor da mediana.

Classe	F_i
2,0 ⊢ 2,5	2
2,5 ⊢ 3,0	4
3,0 ⊢ 3,5	7
3,5 ⊢ 4,0	5
4,0 ⊢ 4,5	5
4,5 ⊢ 5,0	7
Soma	30

Vimos que para calcular a mediana é necessário obter a classe mediana. Dessa forma, para determinar quem é a classe mediana é necessário obter a frequência acumulada (Fac) para cada classe.

Relembrando

Para determinar as frequências acumuladas de cada classe, basta ir acumulando (somando) as frequências absolutas.

Basta vermos a tabela formada a seguir:

Classe	F_i	F_{ac}
2,0 ⊢ 2,5	2	2
2,5 ⊢ 3,0	4	6
3,0 ⊢ 3,5	7	13
3,5 ⊢ 4,0	5	18
4,0 ⊢ 4,5	5	23
4,5 ⊢ 5,0	7	30
Soma	30	-

Seguindo os passos para o cálculo da mediana para dados agrupados, temos:

Determinar o elemento mediano

Note que esse elemento mediano ocupa a décima quinta posição no conjunto de dados.

$$E_{M_d} = \frac{30}{2} = 15$$

Determinar a classe mediana

É a classe que contém $\circ E_{M_d}$.

A classe que contém o elemento mediano é a quarta classe, visto que ela contém o décimo quinto elemento. Note ainda que essa classe contém do décimo quarto ao décimo oitavo elemento, conforme visto a seguir:

Classe	F_i	F_{ac}
2,0 ⊢ 2,5	2	2
2,5 ⊢ 3,0	4	6
3,0 ⊢ 3,5	7	13
3,5 ⊢ 4,0	**5**	**18**
4,0 ⊢ 4,5	5	23
4,5 ⊢ 5,0	7	30
Soma	30	-

A classe mediana é a quarta classe:

3,5 ⊢ 4,0	**5**	**18**

Aplicar a fórmula

Temos a seguinte fórmula:

$$M_d = L_{M_d} + \left(\frac{E_{M_d} - F_{\text{aac}}}{F_{M_d}} \right) \times h = 3,5 + \left(\frac{15 - 13}{5} \right) \times 0,5 = 3,7$$

Por isso, a mediana do peso dos recém-nascidos é igual a 3,7 Kg.

A interpretação da frequência acumulada é feita da seguinte forma: note que na primeira classe, temos o primeiro e o segundo elementos do conjunto de dados, pois os dados estão em ordem crescente, conforme definição de mediana.

⚠

Atenção

Verifique que o valor da frequência acumulada da última classe deve ser igual à soma da frequência absoluta.

Na segunda classe, temos do terceiro ao sexto elemento. Na terceira classe, temos do sétimo ao décimo terceiro elemento, e assim por diante, de forma que, a última classe contém do vigésimo quarto ao trigésimo elemento.

Moda \(\left(M_O\right)\)

Para dados não agrupados

Moda é o valor mais mais frequente no conjunto de dados. Para determinar a moda nesse caso, basta ver o valor que mais se repete no conjunto de dados.

Dica

Caso nenhum valor se repita, dizemos que o conjunto de dados é amodal.

Exemplo 1

$1, 2, 2, 3, 5 \rightarrow M_o = 2$, veja que o número 2 é o valor que mais se repete.

Exemplo 2

$1, 3, 6, 7 \rightarrow$ Amodal, pois não temos nenhuma repetição dos valores.

Exemplo 3

$1, 2, 3, 3, 4, 5, 5 \rightarrow M_o = 3$ e $5($ bimodal $)$.

Para dados agrupados

Nesse caso, para obter a moda, seguiremos os seguintes passos:

Passo 1

Determinar a classe modal C_{M_o} que é a classe com maior frequência, seja ela absoluta ou relativa.

Passo 2

Calcular a moda a partir da fórmula de Czuber:

$$M_o = L_{I_{M_0}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times h$$

Em que:

$L_{I_{M_o}} =$ Limite inferior da classe modal.
$\Delta_1 =$ Frequência absoluta da classe modal - Frequência absoluta da classe imediatamente anterior.
$\Delta_2 =$ Frequência absoluta da classe modal - Frequência absoluta da classe imediatamente posterior.

Observe que os valores obtidos para as modas são apenas aproximações. Eles devem ser obtidos desse modo somente se não for possível dispor dos dados originais.

Exemplo

Considerando novamente os dados dos pesos dos recém-nascidos, obtenha o valor da moda.

Classe	F_i
2,0 ⊢ 2,5	2
2,5 ⊢ 3,0	4
3,0 ⊢ 3,5	7
3,5 ⊢ 4,0	5
4,0 ⊢ 4,5	5
4,5 ⊢ 5,0	7
Soma	30

Seguindo os passos para determinação da moda, temos que localizar primeiro, onde há a maior frequência. Observando a tabela a seguir, veremos que se encontra entre os pontos 3,0 ⊢ 3,5, e 4,5 ⊢ 5,0, o que nos da uma distribuição bimodal.

Classe	F_i
2,0 ⊢ 2,5	2
2,5 ⊢ 3,0	4
3,0 ⊢ 3,5	**7**
3,5 ⊢ 4,0	5
4,0 ⊢ 4,5	5
4,5 ⊢ 5,0	**7**
Soma	30

Veja que, nesse caso, temos duas classes com maiores frequências absolutas. Por isso, teremos duas classes modais e, assim, duas modas associadas a esse conjunto de dados.

Calcular as modas

Temos a seguinte fórmula:

$$M_{o_1} = L_{I_{M_0}} + \left(\frac{\Delta_1}{\Delta_1+\Delta_2}\right) \times h = 3 + \left(\frac{3}{3+2}\right) \times 0,5 = 3,3$$

Na qual:

$$L_{I_{M_0}} = 3, \Delta_1 = 7 - 4 = 3, \quad \Delta_2 = 7 - 5 = 2 e h = 0,5$$

$$M_{o2} = L_{I_{M_0}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times h = 4,5 + \left(\frac{2}{2+7}\right) \times 0,5 = 4,61$$

Na qual:

$$L_{I_{M_0}} = 4,5, \quad \Delta_1 = 7 - 5 = 2, \quad \Delta_2 = 7 - 0 = 7 \text{ e } h = 0,5$$

Note que há um caso especial, pois a segunda classe modal está na última classe e, para o cálculo de Δ_2, considera-se que a classe imediatamente posterior é igual a zero. O mesmo procedimento deve ser adotado quando a classe modal está na primeira classe, com a diferença que, nesse caso, para o cálculo de Δ_1 considera-se que a classe imediatamente anterior é igual a zero.

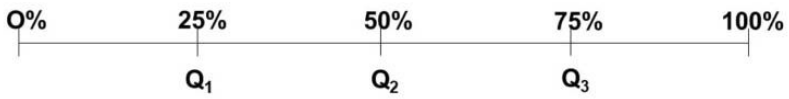
Separatrizes

Definição

As separatrizes têm a função de dividir o conjunto de dados em certo número de partes iguais. A mediana que divide os dados em duas partes iguais é um caso particular de separatriz. No entanto, outras separatrizes têm papel de destaque na estatística, como é o caso dos quartis, decis e percentis, cujos conceitos serão vistos a seguir.

Quartil

O quartil divide o conjunto de dados em quatro partes iguais, conforme visto a seguir:



Divisão em quatro partes iguais

Cada parte representa o seguinte:

- Q_1 = 1º quartil, representa 25% dos elementos;
- Q_2 = 2º quartil, coincide com a mediana, representa 50% dos elementos;
- Q_3 = 3º quartil, representa 75% dos elementos;
- Q_4 = 4º quartil, representa 100% dos elementos.

Para o cálculo dessas medidas, serão adotados os mesmos procedimentos realizados para o cálculo da mediana para dados agrupados. Assim, seguiremos os passos a seguir.

Passo 1

Determinar o elemento quartil.

$$E_{Q_i} = i \cdot \frac{n}{4}$$

Lembre-se de que, neste caso, há quatro quartis, ou seja, i = 1, 2,3,4. Observe que se i = 2, equivale à mediana. Além disso, não importa se n é ímpar ou par, pois estamos trabalhando com dados agrupados.

Passo 2

Determinar a classe quartil i (Q_{i}), que é a classe que contém o elemento quartil i.

Passo 3

Aplicar a fórmula:

$$Q_i = L_{Q_i} + \left(\frac{E_{Q_i} - F_{aac}}{F_{Q_i}}\right) \times h$$

Em que:

$L_{Q_i} =$ Limite inferior da classe quartil i.
$F_{aac} =$ Frequência acumulada anterior à classe quartil i.
$F_{Q_i} =$ Frequência absoluta da classe quartil i.
$h =$ Amplitude da classe quartil i. (diferença entre as amplitudes de classe superior e inferior).

Exemplo

A corretora XYZ tem em seu portifólio 60 imóveis que foram distribuídos de acordo com seu valor de venda (em milhares de reais). Os dados estão representados na distribuição de frequência a seguir:

Classe	F_i	F_{ac}
0 ⊢ 100	5	5
100 ⊢ 200	17	22
200 ⊢ 300	20	42
300 ⊢ 500	13	55
500 ⊢ 800	5	60
Soma	60	-

A partir de que valor estão os 25% dos imóveis mais caros dessa corretora?

Solução: Note que o problema pede para determinar o valor que divide os 25% mais caros dos 75% mais baratos, o que é equivalente a determinar o 3º quartil (Q3). Então:

Determinar o elemento quartil 3

Temos a seguinte fórmula:

$$E_{Q_3} = 3.\frac{60}{4} = 45$$

Determinar a classe quartil

Observe pela frequência acumulada que a classe que contém o elemento quartil 3 é a quarta, visto que essa classe contém do $43°$ elemento ao $55°$, de acordo com a tabela a seguir:

Classe	F_i	F_{ac}
0 ⊢ 100	5	5
100 ⊢ 200	17	22
200 ⊢ 300	20	42
300 ⊢ 500	**13**	**55**
500 ⊢ 800	5	60
Soma	60	-

Aplicar a fórmula

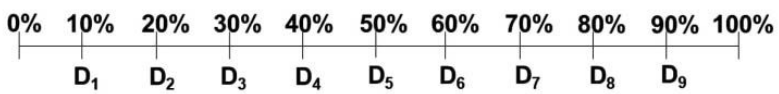
Temos a seguinte fórmula:

$$Q_3 = L_{Q_3} + \left(\frac{E_{Q_3 - F_{\text{ant}}}}{F_{Q_3}}\right) \times h = 300 + \left(\frac{45 - 42}{13}\right) \times 200 = 346,15$$

Portanto, o valor pedido é de R$346.150,00 reais.

Decil

Divide o conjunto de dados em 10 partes iguais.



Divisão em 10 partes iguais

O cálculo dos decis é análogo ao dos quartis. Dessa forma, os seguintes passos serão realizados:

Passo 1

Determinar o elemento decil i.

$$E_{D_i} = i \cdot \frac{n}{10}$$

Aqui temos 10 decis, ou seja, $i = 1, 2, 3, \ldots, 10$. Observe que se $i = 5$, equivale à mediana e ao segundo quartil.

Passo 2

Determinar a classe decil i (C_{D_i}) , que é a classe que contém o elemento decil i.

Passo 3

Aplicar a fórmula:

$$D_i = L_{D_i} + \left(\frac{E_{D_i} - f_{aac}}{F_{D_i}}\right) \times h$$

Na qual:

$L_{D_i} =$ Limite inferior da classe decil i.
$F_{aac} =$ Frequência acumulada anterior à classe decil i.
$F_{D_i} =$ Frequência absoluta da classe decil i.
$h =$ Amplitude da classe decil i (diferença entre as amplitudes de classe superior e inferior).

Exemplo

Considerando o exemplo da corretora, a partir de que valor estão os 10% dos imóveis mais baratos?

Solução: Veja que o problema pede para determinar o valor que representa justamente o 1º decil.

Determinar o elemento decil 1

Temos a seguinte fórmula:

$$E_{D_1} = 1.\frac{60}{10} = 6$$

Determinar a classe decil

Veja que a classe que contém o elemento decil 1 é a segunda classe, conforme mostra na seguinte tabela:

Classe	F_i	F_{ac}
0 ⊢ 100	5	5
100 ⊢ 200	**17**	**22**
200 ⊢ 300	20	42

Classe	F_i	F_{ac}
$300 \vdash 500$	13	55
$500 \vdash 800$	5	60
Soma	60	-

Aplicar a fórmula

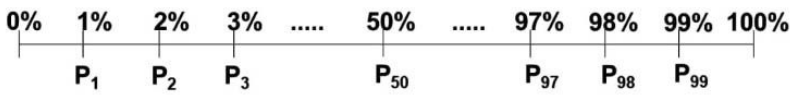
Temos a seguinte fórmula:

$$D_i = L_{D_i} + \left(\frac{E_{D_i} - f_{\text{aac}}}{F_{D_i}} \right) \times h = 100 + \left(\frac{6-5}{17} \right) \times 100 = 105,88$$

Logo, o valor pedido é R$105.880,00 reais.

Percentil

O percentil divide o conjunto de dados em 100 partes iguais.



Divisão em 100 partes iguais

Portanto, devem ser feitos os seguintes passos:

Passo 1

Determinar o elemento decil i.

$$E_{P_i} = i \cdot \frac{n}{100}$$

Neste caso, temos 100 percentis, ou seja, i = 1, 2,3, ... ,100. Observe que se i = 50, equivale à mediana, ao segundo quartil e ao quinto decil.

Passo 2

Determinar a classe decil i (C_{P_i}) , que é a classe que contém o elemento decil i.

Passo 3

Aplicar a fórmula:

$$P_i = L_{P_i} + \left(\frac{E_{P_i} - f_{aac}}{F_{P_i}}\right) \times h$$

Na qual:

$L_{P_i} =$ Limite inferior da classe percentil i.
$F_{aac} =$ Frequência acumulada anterior à classe percentil i.
$F_{P_i} =$ Frequência absoluta da classe percentil i.
$h =$ Amplitude da classe percentil i (diferença entre as amplitudes de classe superior e inferior).

Exemplo

Considerando o exemplo da corretora, a partir de que valor estão 1% dos imóveis mais caros?

Solução: Veja que o problema pede para determinar o valor que representa justamente o 99º percentil. Então:

Determinar o elemento percentil 99

Temos a seguinte fórmula:

$$E_{P_{99}} = 99.\frac{60}{100} = 59,4$$

Determinar a classe percentil

Veja que classe que contém o elemento percentil 99 é a quinta classe, segundo consta na tabela a seguir:

Classe	F_i	F_{ac}
0 ⊢ 100	5	5
100 ⊢ 200	17	22
200 ⊢ 300	20	42
300 ⊢ 500	13	55
500 ⊢ 800	**5**	**60**
Soma	60	-

Aplicar a fórmula

Temos a seguinte fórmula:

$$P_i = L_{P_i} + \left(\frac{E_{P_i} - f_{aac}}{F_{P_i}}\right) \times h = 500 + \left(\frac{59,4-55}{5}\right) \times 300 = 764$$

Logo, o valor pedido é R$764.000,00 reais.

Mão na massa

Questão 1

O rol a seguir representa os valores de itens vendidos (em reais) em uma loja de produtos alimentícios durante um dia de trabalho.

5, 8, 10, 10, 12, 15, 18, 20, 20, 24, 25, 25, 25, 30, 38, 45, 52, 52, 60, 65, 70, 70, 79, 84, 90

Determine a média, a mediana e a moda de vendas nesse dia de trabalho.

A

25, 25 e 25

B

38, 30 e 25

C

38, 25 e 25

D

30, 25, 20

E

38, 34 e 25

A alternativa C está correta.

Veja que, como os dados estão em rol, a média é dada por:

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{952}{25} = 38,08$

Portanto, a média de vendas nesse dia de trabalho é de aproximadamente R$38,00 reais.

Para o cálculo da mediana levamos em consideração o tamanho da amostra (n). Dessa forma, como o tamanho da amostra é igual a 25, n é ímpar. Portanto:

$E_{M_d} = \frac{n+1}{2} = \frac{26}{2} = 13^0$

Logo, a mediana é o elemento que ocupa a décima terceira posição:

$M_d = 25$

Para determinarmos a moda, basta verificarmos no conjunto de dados qual o valor que mais se repete. Assim, verificamos que o valor que representa a moda é 25.

Questão 2

A Secretaria de educação de certo município coletou dados sobre o número de evasão escolar no ensino fundamental durante os últimos 5 anos. Os dados estavam salvos em uma planilha eletrônica, mas por um descuido do digitador, os dados foram multiplicados por 2. Sobre esse deslize do digitador é correto afirmar:

A

A média e a mediana ficam multiplicadas por 2, mas a moda não.

B

A mediana e a moda ficam multiplicadas por 2, mas a média não.

C

A média e a moda ficam multiplicadas por 2, mas a mediana não.

D

A média, mediana e moda ficam multiplicadas por 2.

E

Nenhuma medida de posição é afetada por essa multiplicação equivocada.

A alternativa D está correta.

Observe que se multiplicarmos os dados por 2, todos os valores serão multiplicados por 2. Por exemplo, se considerarmos os valores 1, 2, 2, 4 e 6, é fácil ver que a média é igual a 3 e a mediana e moda são iguais a 2. Se multiplicarmos os dados por 2, os seus novos valores serão 2, 4, 4, 8 e 12 e agora a sua média é 6 e a mediana e moda são iguais a 4. Portanto, as novas medidas de posição (média, mediana e moda) ficam multiplicadas por 2. Logo, a opção correta é a D).

Questão 3

Um plano de saúde fez um levantamento da quantidade de famílias associadas levando em conta o número de dependentes. Os dados foram resumidos na distribuição de frequência a seguir:

N° de dependentes	Quantidade de Famílias
0	800
1	1200
2	350
3	150
Soma	2500

A média e a mediana do número aproximado de dependentes dessas famílias são:

A

0 e 1

B

1 e 1

C

1 e 0

D

1 e 2

E

2 e 2

A alternativa B está correta.

Observe que os dados são apresentados em uma distribuição de frequência. Portanto, utilizaremos as fórmulas para dados agrupados para o cálculo das medidas pedidas. Tendo a média:

N° de dependentes	F_i	X_i , F_i
0	800	0
1	1200	1200
2	350	700
3	150	450
Soma	2500	2350

Daí,

$$X = \frac{\sum_{i=1}^{n} X_i F_i}{n} = \frac{2350}{2500} = 0,94 \approx 1$$

Para o cálculo da mediana, considere os seguintes passos:

1) Determinar o elemento mediano.

$$E_{M_d} = \frac{n}{2} = \frac{2500}{2} = 1250$$

2) Determinar a classe mediana (C_{M_d}) , que é a classe que contém o E_{M_d} .

N° de dependentes	F_i	$X_i . F_i$	F_{ac}	
0	800	0	800	
1	1200	1200	2000	→ Classe Mediana
2	350	700	2350	
3	150	450	2500	
Soma	2500	2350	-	

A classe que contém o elemento mediano é a segunda, visto que essa classe contém o elemento de ordem 1250, que é o elemento mediano.

3) Aplicar a fórmula:

$$M_d = L_{M_d} + \left(\frac{E_{M_d} - F_{aac}}{F_{M_d}}\right) \times h$$

Veja que a amplitude das classes é zero. Então, a fórmula para o cálculo da mediana se reduz a:

$$M_d = L_{M_d} = 1$$

Na qual L_{M_d} é o limite inferior da classe mediana.

Questão 4

A distribuição de frequência a seguir representa a faixa etária dos funcionários de certa empresa.

Classe	X_i	F_i	X_i , F_i
20 ⊢ 30	25	30	750
30 ⊢ 40	35	40	1400
40 ⊢ 50	45	25	1125
50 ⊢ 60	55	17	935
60 ⊢ 70	65	13	845
Soma	-	125	5055

Sobre a média, mediana e moda, podemos afirmar que:

A

Média < Mediana < Moda

B

Moda < Média < Mediana

C

Mediana < Moda < Média

D

Moda < Mediana < Média

E

Mediana < Média < Moda

A alternativa D está correta.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Questão 5

Considere agora que a distribuição de frequência a seguir representa a faixa etária dos funcionários da empresa XYZ.

Classe	X_i	F_i	F_{ac}
20 ⊢ 30	**25**	**25**	**25**
30 ⊢ 40	35	35	60
40 ⊢ 50	45	20	80
50 ⊢ 60	55	12	92
60 ⊢ 70	65	8	100
Soma	-	100	-

Qual o valor do primeiro quartil?

A

30

B

35

C

45

D

50

E

55

A alternativa A está correta.

1) Determinar o primeiro elemento quartil.

$E_{Q_1} = 1.\frac{100}{4} = 25$

2) Determinar a classe quartil $1\,(C_{Q_1})$.

Classe	X_i	F_i	F_{ac}
20 ⊢ 30	**25**	**25**	**25**
30 ⊢ 40	35	35	60
40 ⊢ 50	45	20	80
50 ⊢ 60	55	12	92
60 ⊢ 70	65	8	100
Soma	-	100	-

Observe, pela frequência acumulada, que a classe que contém o elemento quartil 1 é a primeira classe, visto que essa classe contém do primeiro ao 25º elemento.

3) Aplicando a fórmula, temos:

$Q_1 = L_{Q_1} + \left(\frac{E_{Q_1} - F_{ant}}{F_{Q_1}}\right) \times h = 20 + \left(\frac{25-0}{25}\right) \times 10 = 30$

Logo, o primeiro quartil é igual a 30, o que significa que 25% dos funcionários têm menos de 30 anos e 75% deles têm mais de 30 anos.

Questão 6

O histograma a seguir representa o número de funcionários de uma consultoria jurídica por tempo de serviço em anos.

Com base no histograma, julgue as alternativas e marque a incorreta:

A

A média é maior do que a mediana.

B

A mediana é maior do que a moda.

C

25% dos funcionários tem menos de 4,16 anos na empresa.

D

75% dos funcionários tem menos de 11,25 anos na empresa.

E

25% dos funcionários têm mais de 7,5 anos na empresa.

A alternativa E está correta.

Observe que, para responder a essa questão, é necessário determinarmos a média, mediana e a moda para dados agrupados. Assim, para facilitar o cálculo vamos transportar os dados do histograma para uma distribuição de frequência.

Tempo de Serviço (Anos)	X_i	F_i	X_i , F_i	F_{ac}
$0 \vdash 5$	2,5	15	37,5	15
$5 \vdash 10$	7,5	20	150	35
$10 \vdash 15$	12,5	10	125	45

Tempo de Serviço (Anos)	X_i	F_i	X_i , F_i	F_{ac}
15 ⊢ 20	17,5	5	87,5	50
Soma	-	50	400	-

Calculando a média:

$$X = \frac{\sum_{i=1}^{n} X_i F_i}{n} = \frac{400}{50} = 8$$

Calculando a mediana:

$$E_{M_d} = \frac{n}{2} = 25$$

$$M_d = L_{M_d} + \left(\frac{E_{M_d} - F_{anc}}{F_{M_d}}\right) \times h = 5 + \left(\frac{25-15}{20}\right) \times 5 = 7,5$$

Calculando a moda:

Observe que a classe modal é a segunda classe, uma vez que possui a maior frequência absoluta. Daí, aplicando a fórmula de Czuber, temos:

$$M_0 = L_{I_{M_0}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times h = 5 + \left(\frac{5}{5+10}\right) \times 5 = 6,67$$

Calculando o primeiro quartil:

$$E_{Q_1} = i.\frac{n}{4} = 1.\frac{50}{4} = 12,5$$

$$Q_1 = L_{Q_1} + \left(\frac{E_{Q_1} - F_{anc}}{F_{Q_1}}\right) \times h = 0 + \left(\frac{12,5-0}{15}\right) \times 5 = 4,17$$

Calculando o terceiro quartil:

$$E_{Q_3} = 3 \cdot \frac{n}{4} = 3.\frac{50}{4} = 37,5$$

$$Q_3 = L_{Q_3} + \left(\frac{E_{Q_3} - F_{anc}}{F_{Q_3}}\right) \times h = 10 + \left(\frac{37,5-35}{10}\right) \times 5 = 11,25$$

Portanto, a opção correta é a letra E, pois a mediana que é 7,5 divide o conjunto de dados em duas partes iguais, ou seja, 50% estão abaixo e 50% estão acima de 7,5.

Teoria na prática

Uma loja de produtos naturais tem suas vendas (em reais) do mês de janeiro apresentadas na distribuição de frequência a seguir:

Venda (R$)	F_i
0 ⊢ 20	24
20 ⊢ 60	52
60 ⊢ 100	80
100 ⊢ 200	38
200 ⊢ 400	6
Soma	200

Qual a média de vendas da loja no mês de janeiro?

Chave de resposta

Assista ao vídeo e veja a resolução do que se pede na atividade.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Foram coletados dados sobre o número de passageiros, em certa companhia aérea nos aeroportos do brasil, de janeiro a outubro conforme a seguir:

Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out
8522	12630	7453	6005	5874	6612	8439	7531	6430	4986

O número médio de passageiros dessa empresa é de aproximadamente:

A

5814

B

6438

C

6840

D

7034

E

7448

A alternativa E está correta.

Note que as informações estão em forma de dados brutos, visto que não seguem uma ordem aparente. Assim, o cálculo da média é dado por:

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{74482}{10} = 7448,2$

Questão 2

A distribuição de frequência a seguir representa o lucro líquido de 50 empresas do setor petroquímico (em milhares de reais).

Lecro Líquido (R$)	F_i
100 ⊢ 300	8
300 ⊢ 500	10
500 ⊢ 1000	12
1000 ⊢ 2000	15
2000 ⊢ 5000	5
Soma	50

Com base nesses dados podemos afirmar que:

A

O valor do primeiro quartil se encontra na primeira classe.

B

A média é menor do que a mediana.

C

A mediana é maior do que a moda.

D

A média é menor do que a moda.

E

25% das empresas com maior lucro tem lucro aproximado de R$2.000.000,00.

A alternativa D está correta.

Veja que, para responder a essa questão, precisamos determinar a média, a mediana, a moda, o primeiro quartil e o terceiro quartil.

Lucro Líquido (R$)	X_i	F_i	X_i, F_i	F_{ac}
100 ⊢ 300	200	8	1600	8
300 ⊢ 500	400	10	4000	18
500 ⊢ 1000	750	12	9000	30
1000 ⊢ 2000	1500	15	22500	45
2000 ⊢ 5000	3500	5	17500	50
Soma	-	50	54600	-

Calculando a média:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i F_i}{n} = \frac{54600}{50} = 1092$$

Calculando a mediana:

$$E_{Md} = \frac{n}{2} = 25$$

$$M_d = L_{M_d} + \left(\frac{E_{M_d} - F_{aac}}{F_{M_d}}\right) \times h = 500 + \left(\frac{25-18}{12}\right) \times 500 = 791,67$$

Calculando a moda:

Veja que a quarta classe é a classe modal, visto que tem a maior frequência absoluta.

$$M_0 = L_{I_{M_0}} + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) \times h = 1000 + \left(\frac{3}{3+10}\right) \times 1000 = 1230,77$$

Calculando o primeiro quartil:

$$E_{Q_1} = i \cdot \frac{n}{4} = 1.\frac{50}{4} = 12,5$$

$$Q_1 = L_{Q_1} + \left(\frac{E_{Q_1} - F_{aac}}{F_{Q_1}}\right) \times h = 300 + \left(\frac{12,5-10}{8}\right) \times 200 = 362,5$$

Calculando o terceiro quartil:

$$E_{Q_3} = 3.\frac{n}{4} = 3.\frac{50}{4} = 37,5$$

$$Q_3 = L_{Q_3} + \left(\frac{E_{Q_3} - F_{aac}}{F_{Q_3}}\right) \times h = 1000 + \left(\frac{37,5-30}{15}\right) \times 1000 = 1500$$

Variância (S2)

Desvio quadrático total e médio

A variância e o chamado desvio-padrão são as medidas mais conhecidas, servindo de base para medir o quanto os dados estão dispersos com relação à média, ou seja, o quanto os dados estão afastados da média.

Medidas de dispersão: variância e desvio-padrão

Confira o vídeo, a seguir, sobre as medidas de dispersão: variância e desvio-padrão.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Uma forma de calcular o afastamento de cada dado da média é calcular a diferença entre o dado e a média, mas, se calcularmos todas estas diferenças e as somarmos, obteremos ZERO. Ou seja, tal soma não avalia o afastamento total...

Veja:

$$\sum_1^n (x_i - \bar{x}) = \sum_1^n x_i - \sum_1^n \bar{x} = \text{soma} - n.\bar{x} = \text{soma} - n.\frac{soma}{n} = 0$$

Então, uma forma de evitar que o cálculo do desvio **total** simplesmente nos conduza ao valor zero é, por exemplo, elevar ao quadrado os desvios anteriores, **antes** de somá-los!

De fato, chamamos de **desvio quadrático total (DQ)**, ao somatório dos desvios subtrativos simples, previamente elevados ao quadrado, ou seja:

$$\boldsymbol{DQ} = \sum_1^n (x_i - \bar{x})^2$$

Para calcularmos um desvio quadrático médio que é a medida de dispersão de nosso interesse, chamada de **variância**, devemos analisar duas situações:

1. Dispomos de todos os dados que temos interesse em analisar, ou seja, toda a **população**;
2. Os dados fornecidos correspondem apenas a uma **amostra** dos dados totais de interesse.

Além disso, é útil distinguir as situações em que os dados estão agrupados ou não.

O estudo mais avançado da estatística nos indica que o **desvio quadrático médio** da população e de uma amostra devem ser calculados como se segue:

População

Variância:

$S^2 = \frac{1}{n} \times \sum_1^n (x_i - \overline{x})^2 = \frac{\sum_1^n (x_i)^2 - \frac{(\sum_1^n x_i)^2}{n}}{n}$

Amostra

Variância:

$S^2 = \frac{1}{n-1} \times \sum_1^n (x_i - \bar{x})^2 = \frac{\sum_1^n (x_i)^2 - \frac{(\sum_1^n x_i)^2}{n}}{n-1}$

Observe que no caso da população, dividimos o **desvio quadrático total** pela própria quantidade n de objetos da **população** em análise.

Entretanto, no caso de analisarmos uma amostra, devemos realizar um ajuste, justificado pela estatística, e que envolve o conceito de graus de liberdade, que não é abordado neste nível de curso.

Devemos, então, neste caso, dividir o **desvio quadrático total** por **n-1**, e não por **n**, a quantidade de elementos da nossa amostra.

Se tivermos os dados agrupados em uma tabela de frequência, podemos escrever:

População

Variância:

$S^2 = \sum \frac{(X_i - \mu)^2 \times F_i}{n}$

Amostra

Variância:

$S^2 = \sum \frac{(X_i - \mu)^2 \times F_i}{n-1}$

Onde μ é a média do conjunto de dados e F_i a frequência das amostras.

Exemplo

Considere a seguinte amostra de dados: 1, 3, 5, 7 e 9. Determine o valor da variância. Solução: Veja que, para determinar a variância, é necessário inicialmente calcular a média dos dados.

Dados agrupados

O cálculo da variância para dados agrupados, ou seja, quando os dados estão dispostos em distribuição de frequência, levam em consideração o ponto médio da classe e a frequência absoluta. Tecnicamente, supõe-se que o ponto médio de cada classe é um bom representante dos dados de cada classe.

Desse modo, temos:

População

Variância:

$$S^2 = \frac{1}{n} \times \sum_1^n (x_i - \bar{x})^2 f_i = \frac{1}{n} \times \left[\sum_1^n x_i^2 f_i - \frac{(\sum_1^n x_i f_i)^2}{n} \right]$$

Amostra

Variância:

$$S^2 = \frac{1}{n-1} \times \sum_1^n (x_i - \bar{x})^2 f_i = \frac{1}{n-1} \times \left[\sum_1^n x_i^2 f_i - \frac{(\sum_1^n x_i f_i)^2}{n} \right]$$

Onde xi é o ponto central da i=ésima classe e f¬i a frequência absoluta da i-ésima.

Exemplo

Considere os dados sobre o peso (em Kg) de uma amostra de recém-nascidos de certa maternidade dispostos na distribuição de frequência a seguir:

Classe	F_i
2,0 ⊢ 2,5	2
2,5 ⊢ 3,0	4
3,0 ⊢ 4,0	5
4,0 ⊢ 4,5	5
4,5 ⊢ 5,0	7
Soma	30

Solução: Veja que, para calcularmos a variância, vamos precisar dos produtos $X_i F_i$ e $X_i^2 F_i$. Assim, podemos utilizar a própria distribuição de frequência anterior para obter esses produtos. Daí:

Classe	F_i	X_i	X_i, F_i	X^2_i, F_i
2,0 ⊢ 2,5	2	2,25	4,5	10,125
2,5 ⊢ 3,0	4	2,75	11	30,25
3,0 ⊢ 3,5	7	3,25	22,75	73,9375
3,5 ⊢ 4,0	5	3,75	18,75	70,3125
4,0 ⊢ 4,5	5	4,25	21,25	90,3125
4,5 ⊢ 5,0	7	4,75	33,25	157,9375
Soma	30	-	111,5	432,875

Logo,

$$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 F_i - \frac{\left(\sum_{i=1}^{n} X_i F_i\right)^2}{n}\right] = \frac{1}{29}\left[432,875 - \frac{(111,5)^2}{30}\right] = \frac{1}{29}[432,875 - 414,408] = \frac{18,47}{29} = 0,64$$

Note que essa medida tem a interpretação prejudicada, visto que eleva ao quadrado os desvios $(X_i - \bar{X})$. Esse fato faz com que a unidade da variável com que estamos trabalhando fique ao quadrado.

Se a variável de interesse é medida em quilogramas (Kg), como no exemplo acima, a variância nos dará a resposta em quilogramas ao quadrado (Kg^2) e isso compromete a interpretação da referida medida. Para solucionar esse problema, foi criado o desvio-padrão, que será a próxima medida que veremos.

Desvio-padrão (S)

Definição

O desvio-padrão é simplesmente definido como a raiz quadrada da variância, pois sua unidade, então, é exatamente a unidade dos dados! Assim, fica claro porque indicamos a variância por S^2. Ou seja, S é o desvio-padrão da população ou da amostra.

Considerando o exemplo da variância para dados não agrupados, vimos que o resultado da variância foi igual a 10. Logo,

$$S = \sqrt{10} \cong 3,2$$

Lembre-se de que o desvio-padrão é uma medida de variabilidade dos dados em torno da média.

Então nesse caso, podemos dizer que a dispersão em torno da média é de 3,2. Portanto, com essa medida, é possível verificar como se concentram os dados em torno da média.

Dados agrupados

Temos a seguinte fórmula:

$$S = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 F_i - \frac{\left(\sum_{i=1}^{n} X_i F_i\right)^2}{n}\right]} = \sqrt{S^2}$$

Considerando os dados sobre o peso (em Kg) de recém-nascidos, temos:

$$S = \sqrt{S^2} = \sqrt{0,64} = 0,8 \approx 1,0$$

Interpretação: A variabilidade em torno da média de 1 kg.

Note que, apesar de o desvio-padrão ser uma boa medida da variabilidade, é uma medida absoluta, e nem sempre conseguimos ver com clareza se o seu valor reflete maior ou menor variabilidade dos dados.

Agora imagine um analista que trabalha com uma quantidade enorme de dados, por exemplo, na casa dos milhares ou mesmo milhões de dados: seria quase impossível dizer se a dispersão desses dados é alta ou

baixa, simplesmente observando o valor do desvio-padrão. Para resolver tal problema, foi criado o coeficiente de variação, que veremos a seguir.

Coeficiente de variação (CV%)

Definição

É uma medida de dispersão relativa, sendo muito útil quando temos uma quantidade expressiva de dados ou quando queremos fazer comparações entre variáveis que são medidas em diferentes amostras ou populações. Essa medida é definida por:

$$CV\% = \frac{S}{\bar{X}} \times 100$$

Em que:

- S é o desvio-padrão;
- $\bar{X}$ a média da amostra.

Na prática, considera-se a seguinte regra para dizer se os dados são poucos ou muito dispersos:

Pouca dispersão

$0\% \leq CV\% \leq 10\%$

Dispersão moderada

$10\% < CV\% \leq 30\%$

Alta dispersão

$CV\% \geq 30\%$

Exemplo

Foi aplicada uma prova de conhecimentos gerais em duas turmas, digamos A e B. A turma A obteve média 8 e desvio-padrão 2, a turma B obteve média 6,5 e desvio-padrão 1,8. Qual turma teve maior dispersão em torno da média?

Solução: Aparentemente, a turma que teve maior dispersão foi a A, pois tem desvio-padrão igual a 2, enquanto a turma B teve desvio-padrão igual a 1,8. Porém, para sabermos de fato qual turma teve maior variabilidade, temos que calcular o coeficiente de variação. Assim:

$$CV\%(A) = \frac{S}{\bar{X}} \times 100 = \frac{2}{8} \times 100 = 25\%$$

$$CV\%(B) = \frac{S}{\bar{X}} \times 100 = \frac{1,8}{6,5} \times 100 = 27,69\%$$

Portanto, a turma B teve maior dispersão do que a turma A. Note que ambas as turmas apresentam dispersão moderada.

Mão na massa

Questão 1

O rol a seguir representa os valores dos itens vendidos (em reais) em uma loja de produtos alimentícios durante um dia de trabalho.

5, 8, 10, 10, 12, 15, 18, 20, 20, 24, 25, 25, 25, 30, 38, 45, 52, 52, 60, 65, 70, 70, 79, 84, 90

Qual o desvio-padrão de vendas nesse dia de trabalho?

A

22,24

B

26,44

C

28,16

D

30,14

E

32,66

A alternativa B está correta.

Veja que, para determinar o desvio-padrão para dados não agrupados, temos que inicialmente calcular a média, que nesse caso é dado por:

$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{952}{25} = \mathbf{38,08}$

Assim, $S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}} = \sqrt{S^2} = 26,44$

Questão 2

Considerando o resultado do desvio-padrão obtido na questão anterior, podemos afirmar que:

A

A variabilidade dos dados é em torno de $\sqrt{26,44} = 5,1$.

B

A medida mais adequada de variabilidade para medir a variabilidade dos dados é a variância porque possui a mesma unidade dos dados.

C

A dispersão dos itens vendidos em torno da média é de aproximadamente R$26,00 reais.

D

A concentração dos dados em torno da média é aproximadamente R$676,00 reais.

E

A maior concentração dos dados em torno da média está entre R$12,00 e R$65,00 reais.

A alternativa C está correta.

Como vimos, o desvio-padrão é dado por:

$S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}} = \sqrt{S^2} = 26,44$

Logo, podemos dizer que a dispersão em torno da média para o valor dos itens vendidos é de aproximadamente R$26,00 reais.

Questão 3

Ainda considerando o enunciado da primeira questão. Determine o valor do coeficiente de variação.

A

50,33%

B

58,34%

C

60,26%

D

69,43%

E

75,62%

A alternativa D está correta.

Vimos que, para determinar o Coeficiente de Variação, podemos usar a seguinte expressão:

$CV\% = \frac{S}{\bar{X}} \times 100 = \frac{26,44}{38,08} \times 100 = 69,43\%$

Assim, podemos dizer que esses dados têm uma dispersão relativa em torno da média de 69,43%.

Questão 4

A distribuição de frequência a seguir representa a faixa etária de funcionários de certa empresa. O desvio-padrão da idade desses funcionários é de aproximadamente:

Classe	F_i
$20 \vdash 30$	25
$30 \vdash 40$	35
$40 \vdash 50$	20
$50 \vdash 60$	12
$60 \vdash 70$	8
Soma	100

A

12

B

15

C

18

D

20

E

25

A alternativa A está correta.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Questão 5

Considerando o enunciado da questão anterior, qual seria o valor do coeficiente de variação?

A

21%

B

25%

C

28%

D

31%

E

35%

A alternativa D está correta.

Observe que, para o cálculo do coeficiente de variação, precisamos determinar a média. Logo:

$\bar{X} = \frac{\sum_{i=1}^{n} X_i F_i}{n} = \frac{3930}{100} = 39,3$

Assim,

$CV\% = \frac{S}{\bar{X}} \times 100 = \frac{12,17}{39,3} \times 100 = 30,97\%$

Questão 6

Considerando o resultado obtido do coeficiente de variação da questão anterior, é possível afirmar que:

A

Os dados têm baixa dispersão.

B

Os dados têm dispersão moderada.

C

Os dados têm alta dispersão.

D

Os dados têm dispersão irrelevante.

E

Não é possível medir a dispersão.

A alternativa C está correta.

Vimos que quando $CV\% \geq 30\% \Rightarrow$ Alta dispersão. Como o $\mathrm{CV}\%$ é igual a 30,97\%, podemos dizer que os dados têm alta dispersão.

Teoria na prática

Considere que os dados informados a seguir se referem à idade dos alunos de duas turmas de inglês.

Turma 1: 8, 8, 8, 9, 9, 10, 10, 12
Turma 2: 17, 17, 19, 19, 20, 20, 21, 23

Em qual das duas turmas as idades dos alunos é mais espalhada?

Chave de resposta

Assista ao vídeo e veja a resolução do que se pede na atividade.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Foram coletados dados sobre o número de passageiros, em determinada companhia aérea nos aeroportos do brasil, de janeiro a outubro, conforme o quadro a seguir.

Jan	Fev	Mar	Abr	Mai	Jun	Jul	Ago	Set	Out
8522	12630	7453	6005	5874	6612	8439	7531	6430	4986

O desvio-padrão de passageiros, ao longo desses 10 meses nessa companhia aérea, é de aproximadamente:

A

2144

B

2200

C

2224

D

2340

E

2451

A alternativa A está correta.

Note que as informações estão em forma de dados brutos, visto que não seguem uma ordem aparente. Desse modo, o cálculo do desvio-padrão é dado por:

$$S = \sqrt{\sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}} = \sqrt{S^2} = 2143,52 \approx 2144$$

Questão 2

A distribuição de frequência a seguir representa o lucro líquido de 50 empresas do setor petroquímico (em milhares de reais).

Lucro Líquido (R$)	F_i
100 ⊢ 300	8
300 ⊢ 500	10
500 ⊢ 1000	12
1000 ⊢ 2000	15

Lucro Líquido (R$)	F_i
2000 ⊢ 5000	5
Soma	50

Com base nesses dados podemos afirmar que:

A

A variância é igual a 1.000.000.

B

O desvio-padrão é igual a 2.000.

C

O coeficiente de variação é de aproximadamente 87%.

D

A dispersão absoluta é baixa.

E

A dispersão relativa é moderada.

A alternativa C está correta.

Veja que, para responder a essa questão precisamos determinar a média, o desvio-padrão e o coeficiente de variação:

Lucro Líquido (R$)	X_i	F_i	X_i, F_i	X^2_i, F_i
100 ⊢ 300	200	8	1600	320000
300 ⊢ 500	400	10	4000	1600000
500 ⊢ 1000	750	12	9000	6750000
1000 ⊢ 2000	1500	15	22500	33750000
2000 ⊢ 5000	3500	5	17500	61250000
Soma	-	50	54600	103670000

Calculando a média:

$$X = \frac{\sum_{i=1}^{n} X_i F_i}{n} = \frac{54600}{50} = 1092$$

Calculando a variância:

$S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 F_i - \frac{\left(\sum_{i=1}^{n} X_i F_i\right)^2}{n}\right] = \frac{1}{49}\left[103670000 - \frac{(54600)^2}{50}\right] = 898.914,29$

Calculando o desvio-padrão:

$$S = \sqrt{S^2} = \sqrt{898.914,29} = 948,11$$

Calculando o coeficiente de variação:

$$CV\% = \frac{S}{\overline{X}} \times 100 = \frac{948,11}{1092} \times 100 = 86,82\%$$

Considerações finais

Divididos em três módulos, abordamos conceitos fundamentais de análise de dados quantitativos. No primeiro, vimos conceitos associados a análise exploratória de dados, que tem por objetivo oferecer um panorama inicial sobre o conjunto de dados. Além disso, aprendemos formas de organizá-los e representá-los graficamente. No segundo e no terceiro módulos, respectivamente, trabalhamos as principais medidas de tendência central e variabilidade, medidas essas com vastas aplicações práticas no dia a dia.

Todos os conceitos adquiridos contêm grande aplicabilidade e servem para dar continuidade ao seu aprendizado de estatística.

Explore +

Para saber mais sobre os assuntos tratados neste tema, assista:

Instituto de Matemática Pura e Aplicada – IMPA, Youtube.

Referências

FONSECA, J. S.; MARTINS, G. A. **Curso de Estatística.** 6. ed. São Paulo: Atlas, 1996.

MORETTIN, P. A.; BUSSAB, W. O. **Estatística Básica.** 9. ed. São Paulo: Saraiva, 2017.

OVALLE, I. I.; TOLEDO, G. L. **Estatística Básica.** 2. ed. São Paulo: Atlas, 2010.

SICSU, A. L.; DANA, S. **Estatística Aplicada** – Análise Exploratória de Dados. São Paulo: Saraiva, 2012.