Unidade IV

7 INTRODUÇÃO À ESTATÍSTICA DESCRITIVA

No mundo ao nosso redor, podemos fazer observações a respeito de elementos que possuem, pelo menos, uma característica em comum, por exemplo, as jogadoras de um mesmo time de vôlei. Quando registramos essas observações, obtemos o que chamamos de dados.

De acordo com Crespo (2009), a estatística é uma parte da matemática aplicada que fornece métodos para a coleta, organização, descrição, análise e interpretação de dados, assim como para a utilização desses recursos na tomada de decisões.

No time de vôlei que mencionamos, a estatística pode ser usada para analisar o desempenho individual e coletivo das jogadoras durante os jogos. O treinador pode, por exemplo, usar a estatística para monitorar o número de bloqueios, de ataques e de passes bem executados por cada jogadora durante a partida. Essa abordagem estatística pode ajudar a definir qual jogadora deve ser mais acionada nas situações mais críticas, ou como ajustar a distribuição das jogadas para maximizar a chance de vitória.

Você, como aluno, provavelmente já utiliza estatística no seu cotidiano, mesmo que não tenha se dado conta disso. Para calcular a média final de uma disciplina, você pode usar o formato de média aritmética utilizado pela sua instituição, que seguirá o formato simples ou o formato ponderado. Nesse caso, os dados com os quais você trabalha são o conjunto de notas que obteve ao longo daquele período.

Segundo Carvalho, Menezes e Bonidia (2024), dados são produzidos a todo momento, por quase todos os eventos e nos mais diversos formatos. Ao serem produzidos, eles carregam informações que podem explicar como e por que foram produzidos. Ao serem analisados, permitem conhecer o que há por trás de alguns fenômenos, possibilitando descrever uma situação, prever uma ocorrência ou apontar o que é necessário para que algo aconteça. Embora a análise de dados seja realizada há muitos séculos, avanços em diferentes áreas de conhecimento – especialmente na computação, na estatística e na matemática – não apenas melhoraram as análises já existentes, mas também permitiram tipos de análises nunca antes imaginadas, o que levou ao surgimento de uma nova área de conhecimento: a ciência de dados.

A ciência de dados é uma área interdisciplinar que combina estatística, programação, aprendizado de máquina e outros conhecimentos específicos para extrair informações úteis e tomar decisões com base em grandes volumes de dados. A estatística é fundamental nesse contexto, pois fornece as ferramentas teóricas e práticas para coletar, organizar, analisar e interpretar os dados. Sem a estatística, o processo de transformar dados em *insights* seria impreciso e desestruturado.

Observação

Insights são percepções ou entendimentos profundos obtidos a partir da análise de dados ou de experiências. No contexto de negócios e da ciência de dados, *insights* representam descobertas significativas que ajudam a identificar padrões, tendências, problemas ou oportunidades.

Vamos começar a nossa introdução ao mundo da estatística definindo alguns conceitos básicos.

7.1 População e amostra

A estatística é frequentemente dividida em três áreas principais: estatística descritiva, probabilidade e estatística indutiva. Para que entendamos melhor o significado delas, vamos antes definir dois importantes conceitos: população e amostra.

Pesquisas, em geral, costumam coletar dados de uma pequena parte de um grupo maior, de forma a estimarmos determinadas características desse grupo a partir dos dados coletados. No contexto da estatística, uma população é a coleção completa de todos os elementos (ou indivíduos) que interessam ao estudo de determinado fenômeno. Na figura a seguir, vemos uma população constituída por 8 elementos.

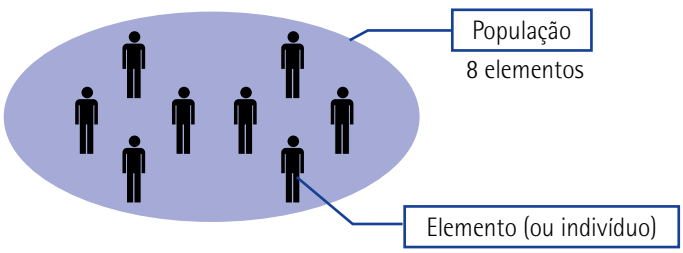

Figura 49 – População estatística

Uma amostra é um subconjunto não vazio da população. Na figura a seguir, vemos uma amostra de 4 elementos, extraída da população composta por 8 elementos.

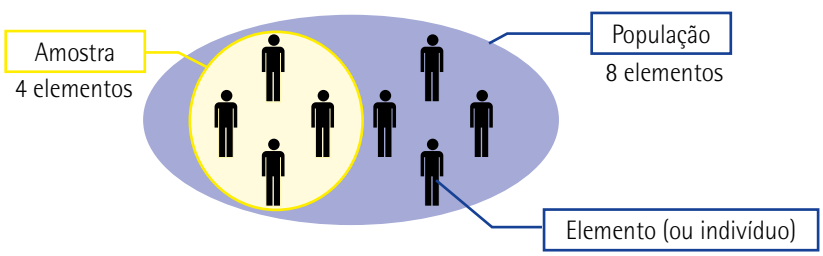

Figura 50 – Amostra

Portanto, a intenção de trabalharmos com dados de uma amostra é estimar características da população que ela representa. De forma geral, ao selecionarmos uma amostra, trabalhamos de uma forma mais rápida, barata e eficiente do que faríamos caso escolhêssemos fazê-lo com todo o conjunto de dados da população.

Vale destacar que nem sempre podemos utilizar a relação "amostra boa é amostra grande". Uma boa amostra é aquela que traz consigo todas as características presentes na população e na proporção em que ocorrem na população.

7.2 Áreas da estatística

Como já comentamos, a estatística é frequentemente dividida em três áreas principais: estatística descritiva, probabilidade e estatística indutiva.

A estatística descritiva, que é o alvo do nosso estudo, se destina a organizar, descrever, explorar, expressar e sintetizar as informações brutas vindas da aplicação de um questionário, da observação de algum evento ou da contagem de ocorrências, por exemplo. Em suma, na estatística descritiva, trabalhamos com conjuntos de dados oriundos de algo certo, que já aconteceu, que pertence ao passado. Por exemplo, se soubermos as idades de 3 pessoas e quisermos calcular a média de idade delas, não há incerteza associada a tal cálculo.

Na parte das probabilidades, lidamos com a chance de algo acontecer, ou seja, estamos no campo da incerteza, dos eventos aleatórios, dos acontecimentos que não podem ser previstos com 100% de exatidão. Por exemplo, sabemos que temos 50% de probabilidade de obtermos cara quando lançamos uma moeda. Logo, não há como dizermos, com certeza, se obteremos cara ou coroa nesse lançamento. Segundo Crespo (2009), o conhecimento dos aspectos fundamentais do cálculo de probabilidades é uma necessidade para o estudo da estatística indutiva.

Na estatística indutiva, trabalhamos com uma amostra, a fim de que, com o uso de técnicas e métodos adequados, possamos obter informações a respeito da população que tal amostra representa. Nesse caso, para dado intervalo de confiança, temos um erro associado. Por exemplo, se uma pesquisa eleitoral diz que certo candidato tem 60% dos votos, com margem de erro de 3% e confiança de 95%, significa que esse candidato tem 95% de chance de ter entre 57% e 63% dos votos na data da pesquisa. A estatística indutiva também estabelece critérios para a seleção de amostras, de modo que consigamos um grupo que represente a população que se quer estudar. Esses critérios levam em consideração o tamanho e a qualidade da amostra, assim como a natureza da pesquisa.

Observação

Muitos autores incluem, na parte de estatística descritiva, descrições, organizações e reduções de dados amostrais. Neste livro-texto, optamos por trabalhar com conjuntos de dados completos, ou seja, que não têm a intenção de estimar características de um grupo maior. Você pode considerar, portanto, que sempre trabalharemos em um contexto populacional.

7.3 Dados brutos e rol

A esse ponto, você já percebeu que a estatística é a ciência que busca extrair informações dos dados. Desse modo, vamos finalmente começar a trabalhar com conjuntos de dados.

Vamos aprender mais um conceito introdutório, mas importante: o significado de dados brutos. Basicamente, eles são uma sequência de valores não organizados, geralmente obtidos da observação direta de um fenômeno. Dados brutos, portanto, são os dados da forma que são obtidos, sem nenhum tratamento ou organização.

Por exemplo, em uma pesquisa de cargos e salários, o conjunto de dados com cargo do funcionário e o salário são dados brutos, se não passaram por nenhuma forma de organização.

Se partirmos de dados brutos e aplicarmos alguma forma de organização, temos o que chamamos de rol. No rol, os dados podem ser organizados de forma crescente, decrescente ou ainda em ordem alfabética.

Exemplo de aplicação

Uma turma de alunos da disciplina Introdução à Programação Estruturada obteve as seguintes notas em uma prova:

Tabela 19 – Desempenho dos alunos na prova

Aluno	Nota
Maria	9
Pedro	4
Otávio	6
Mariana	7
Sheila	8,5
Oswaldo	3
Matheus	8
Guilherme	10
Leonardo	10

Com base nos dados da tabela, determine um rol das notas destes alunos.

Resolução

Se tomarmos as notas dos alunos, sem qualquer organização, temos os dados brutos:

9 4 6 7 8,5 3 8 10 10

Se aplicarmos qualquer processo de organização a esses dados, passamos a ter um rol. Vamos, por exemplo, organizar as notas de forma crescente, conforme exposto a seguir.

3 4 6 7 8 8,5 9 10 10

Dessa forma, organizada, podemos analisar melhor a distribuição de notas dos alunos.

Note que nem todos os dados são de natureza numérica. Na tabela de notas dos alunos, também temos informação sob a forma de nomes, que são classificados como dados alfanuméricos. Poderíamos, portanto, obter um rol organizando o nome dos alunos em ordem alfabética, por exemplo.

Perceba, também, que o rol não omitiu valores repetidos. Partimos de um conjunto de dados brutos composto por 9 numerais, e chegamos a um rol que tem, também, 9 numerais.

O rol, portanto, corresponde à etapa de organização, e não de redução de dados.

Observação

Os indivíduos de um estudo estatístico não precisam ser seres humanos. Podemos, por exemplo, pensar em uma população de canetas produzidas em uma fábrica ou de bovinos em uma região. Os indivíduos são os objetos, as entidades ou os sujeitos sobre os quais os dados são coletados, processados e analisados.

7.4 Tipos de variáveis

No contexto estatístico, uma variável é qualquer característica dos elementos em estudo. Uma variável pode assumir valores distintos para elementos distintos (daí sua semelhança com as variáveis que usamos nas funções, nos capítulos anteriores). Vamos, como exemplo, acompanhar uma tabela de dados.

Tabela 20 – Dados de 8 elementos

	Idade (anos)	Peso (kg)	Altura (m)	Renda mensal (R$)	Sexo (M/F)
1	20	75,2	1,69	1000,00	M
2	25	67,8	1,52	2500,00	F
3	54	92,5	1,86	950,50	M
4	27	54,5	1,62	4000,00	F
5	32	63,0	1,74	10680,00	F
6	43	72,5	1,82	870,00	F
7	17	89,4	1,90	3000,00	M
8	33	52,1	1,55	2000,40	F

Pelos dados da tabela, sabemos que temos 8 indivíduos em estudo. A identificação dada a cada um deles é a própria numeração de 1 a 8, presente na primeira coluna.

A partir da segunda coluna, temos títulos posicionados na linha inicial. Cada título representa uma variável. Cada coluna, portanto, traz o nome da variável e os valores assumidos pela variável para cada um dos 8 elementos. As variáveis que podemos identificar na tabela são: idade, peso, altura, renda mensal e sexo.

Cada linha dessa tabela, por sua vez, traz todos os dados para determinado indivíduo em estudo. Por exemplo, o indivíduo 4 tem 27 anos de idade, 54,5 kg de peso, 1,62 m de altura, R$ 4.000,00 de renda mensal e é do sexo feminino.

Os dados da tabela já nos permitem inferir que existem tipos distintos de variáveis. Segundo Triola (2023), a estatística costuma dividir suas variáveis em dois grandes grupos: variáveis quantitativas e variáveis qualitativas (ou categóricas). Estudaremos esses tipos, a seguir.

Lembrete

No contexto estatístico, os termos elementos e indivíduos representam a mesma coisa: unidades básicas de análise que compõem uma população ou amostra.

7.4.1 Variáveis quantitativas

As variáveis quantitativas são aquelas de característica mensurável. Desse modo, elas apresentam valores numéricos. As variáveis quantitativas se subdividem em dois subgrupos: variáveis quantitativas discretas e variáveis quantitativas contínuas.

Uma variável quantitativa discreta é uma característica mensurável que assume um número contável de valores e, assim, geralmente assume valores inteiros. É muito comum que esse tipo de variável apresente valores oriundos de contagens. Alguns de seus exemplos são: número de filhos, número de batimentos cardíacos em um período, número de cigarros fumados por dia, número de alunos matriculados em uma faculdade, número de peças defeituosas em um lote etc.

Por sua vez, uma variável quantitativa contínua é uma característica mensurável para as quais valores fracionários são possíveis e fazem sentido. Esse tipo de variável pode assumir, teoricamente, qualquer valor dentro de um intervalo contínuo de números reais. Na maioria dos casos, os valores dessas variáveis são associados a uma unidade de medida. Alguns de seus exemplos são: altura, peso, renda mensal, tempo, idade, pressão arterial, nível de colesterol no sangue etc.

Observação

Algumas variáveis quantitativas são intrinsecamente contínuas, mas podem ser tratadas de forma discreta, dependendo do contexto. Tome, como exemplo, a idade de pessoas. Uma idade pode ser medida com qualquer grau de precisão, o que faz com que ela seja considerada uma variável contínua. É muito comum dizermos que uma criança, por exemplo, tem dois anos e meio.

No entanto, no geral, medimos nossas idades em anos completos, o que faz com que a variável assuma apenas valores inteiros. Nesse caso, por conveniência, estamos tratando uma variável contínua de maneira discreta, limitando os valores possíveis ao conjunto dos números inteiros. O tratamento discreto é uma simplificação prática, não uma mudança na natureza da variável.

Exemplo de aplicação

Classifique as variáveis quantitativas a seguir em contínuas ou discretas.

A) Idade

B) Número de reclamações por departamento

C) Número de alunos por classe

D) Volume de líquido em recipiente

Resolução

A) Contínua. Por mais que seja comumente expressa com a precisão de anos completos (resultando em um número inteiro), a idade de um indivíduo pode assumir qualquer valor em um intervalo.

B) Discreta. Os valores da variável são expressos por números inteiros, já que não há como um departamento receber um número fracionário de reclamações.

C) Discreta. Os valores da variável são expressos por números inteiros, já que não há como uma classe ter um número fracionário de alunos.

D) Contínua. O volume de líquido em um recipiente pode assumir qualquer valor em um intervalo e ser medido por diversas precisões, além de estar associado a uma unidade de medida (como mL, por exemplo).

7.4.2 Variáveis qualitativas

As variáveis qualitativas, também chamadas de variáveis categóricas, são aquelas de característica não mensurável. Os valores assumidos por esse tipo de variável são alfanuméricos. Desse modo, eles são definidos por categorias e, assim, representam uma classificação dos indivíduos. As variáveis qualitativas se subdividem em dois subgrupos: variáveis qualitativas nominais e variáveis qualitativas ordinais.

Uma variável qualitativa nominal é uma característica não mensurável para a qual não existe ordenação natural para os valores que podem ser assumidos. Desse modo, não existe uma hierarquia entre as categorias da variável. Alguns de seus exemplos são: sexo, cor dos olhos, fumante/não fumante, doente/sadio, placa de veículos, estado civil, marca de carro, meio de transporte utilizado, nacionalidade, gênero musical etc.

Uma variável qualitativa ordinal é uma característica não mensurável para a qual existe ordenação natural para os valores que podem ser assumidos. Desse modo, existe uma hierarquia entre as categorias da variável. Alguns de seus exemplos são: estágio de determinada doença (inicial, intermediário, terminal), nível de escolaridade (fundamental, médio, superior etc.), mês de observação (janeiro, fevereiro,..., dezembro), risco de crédito (baixo, médio, alto) etc.

Exemplo de aplicação

Classifique as variáveis qualitativas a seguir em nominais ou ordinais.

A) Faixa etária

B) Raça de cachorro

C) Classe social

D) Nacionalidade

Resolução

A) Ordinal. Os valores da faixa etária organizam-se naturalmente em sequência crescente. Nesse caso, houve uma categorização da idade, que é uma variável quantitativa. Os valores categóricos, aqui, podem ser expressos como: infantil (de 0 a 11 anos), adolescente (de 12 a 17 anos), adulto (de 18 a 59 anos) e pessoa idosa (a partir de 60 anos).

B) Nominal. Os valores da variável não demonstram qualquer hierarquia. Nesses casos, é muito comum (mas não necessário) organizarmos os possíveis valores por ordem alfabética em formulários, por exemplo.

C) Ordinal. Os valores da classe social organizam-se naturalmente em uma sequência. Uma forma comum de categorizar essa variável é utilizar os termos A, B, C, D e E, sendo A a classificação para indivíduos de maior renda familiar e E para indivíduos de menor renda familiar. Note que, nesse caso, a variável classe social também é uma categorização de uma variável quantitativa, que é a renda familiar.

D) Nominal. Os valores da variável não demonstram ordem intrínseca e podem ser organizados em ordem alfabética, se necessário.

Observação

É importante conhecermos os tipos de variáveis estatísticas, porque isso influencia diretamente a forma como os dados são coletados, analisados e interpretados. A compreensão dos tipos de variáveis nos permite escolher ferramentas estatísticas mais apropriadas para cada tipo de dado. Por exemplo, ao trabalharmos com variáveis contínuas, podemos usar gráficos de dispersão ou histogramas, enquanto variáveis nominais se enquadram melhor em gráficos de colunas ou de setores.

7.5 Distribuição de frequências

Agora que aprendemos a organizar dados em rol e conhecemos os tipos de variáveis estatísticas, vamos apresentar uma importante etapa de redução de dados.

Uma distribuição de frequências é uma representação de um conjunto de dados de uma variável na qual os valores se apresentam em correspondência com suas repetições, evitando, assim, que apareçam mais de uma vez, como acontece no rol.

Desse modo, a distribuição de frequências mostra como o conjunto de dados é dividido entre todas as classes (que atuam como categorias). Geralmente, as variáveis da qual partimos são quantitativas (mas não necessariamente).

A distribuição de frequências pode ser apresentada de duas maneiras principais:

- **Tabela de frequências**: exibe a contagem de ocorrências (frequências) para cada classe. Trata-se, portanto, de uma forma tabular de representar a distribuição.

- **Gráfico de colunas**: um gráfico que mostra a distribuição de frequências de forma visual. Quando são construídos com colunas justapostas representando as classes, são chamados de histogramas. Cada coluna no histograma indica a frequência de um intervalo específico, o que nos ajuda a visualizar a distribuição de forma contínua e intuitiva. Alguns tipos de distribuição podem ser representados por gráficos de colunas convencionais, que não constituem histogramas, conforme estudaremos.

Vamos começar o nosso estudo montando as tabelas de frequência para, em seguida, aprendermos a construir gráficos.

7.5.1 Classe e frequência simples absoluta

Em uma tabela de frequências, uma classe é uma espécie de categoria formada a partir dos valores da variável em estudo. Há dois tipos principais de classe: classe discreta e classe intervalar.

Uma classe discreta (ou classe unitária), como o próprio nome indica, é geralmente aplicada a variáveis quantitativas discretas (ou que estejam sendo tratadas como uma, como comumente acontece com idades).

Imagine que uma fábrica precisa encomendar calçados de segurança para seus operários. Para isso, a gestora responsável pela encomenda montou a tabela de frequências a seguir. Nesse contexto, os indivíduos em estudo são os operários da fábrica e a característica de interesse é, justamente, o tamanho do calçado de cada um. A tabela, portanto, foi montada a partir do conjunto de dados da variável quantitativa discreta tamanho do calçado.

Tabela 21 – Tabela de frequências com classes discretas

i (índice)	X_i (ponto médio de classe)	F_i (frequência)
1	36	5
2	37	10
3	38	13
4	39	12
5	40	11
6	41	9
7	42	2

Vamos entender a simbologia e a interpretação do conteúdo da tabela. Na primeira coluna, vemos um índice i, que apenas numera as classes. Se chamarmos de k o número total de classes, temos $k=7$. Portanto, o índice i deve contar de 1 até k, o que, nesse contexto, corresponde de 1 a 7.

Esse mesmo índice é utilizado no símbolo à direita, x_i, que representa o valor de cada classe. Cada um desses valores corresponde a um número de calçado, o que nos indica que os operários dessa fábrica calçam de 36 a 42. Se quisermos apontar o valor de uma classe específica, podemos usar o índice para nos auxiliar. Por exemplo, para indicar que a classe 3 corresponde ao calçado de número 38, podemos usar a notação $x_3 = 38$.

Talvez você tenha notado que, na tabela, logo abaixo do símbolo x_i, escrevemos "ponto médio de classe", ao invés de valor de classe, como nos referimos a ele. Entenderemos essa nomenclatura quando virmos classes intervalares. Por enquanto, vamos estudar o próximo símbolo.

O símbolo f_i nos indica a frequência simples absoluta correspondente a cada classe. Uma frequência simples absoluta é o resultado da contagem das ocorrências do valor da classe no conjunto de dados brutos. Não se assuste, é fácil.

Olhando para a classe 1, vemos que 5 operários calçam 36. Isso significa que o valor 36 apareceu 5 vezes no conjunto de dados brutos relativo à variável "tamanho do calçado". Olhando para a classe 4, vemos que 12 operários calçam 39. Isso significa que o valor 39 apareceu 12 vezes.

Com o objetivo de indicar que a frequência da classe 3 corresponde a 13 ocorrências, podemos usar a notação $f_3 = 13$.

Mesmo não tendo acesso ao conjunto de dados brutos, é possível sabermos quantos são os operários dessa fábrica. Basta que façamos o somatório das frequências simples absolutas. Isso nos trará o valor de elementos da nossa população, que chamaremos de N. Esse resultado é demonstrado na tabela a seguir, que replica os mesmos dados da tabela anterior, com a adição do somatório das frequências.

Tabela 22 – Tabela de frequências com classes discretas e somatório das frequências

i (índice)	X_i (ponto médio de classe)	F_i (frequência)
1	36	5
2	37	10
3	38	13
4	39	12
5	40	11
6	41	9
7	42	2
		$\sum_{i=1}^{k} f_i = N = 62$

Perceba a aparição do símbolo de somatório, Σ, acompanhado de uma notação inferior, superior e lateral. Vamos traduzir o significado dessa simbologia. Observe o símbolo completo, reproduzido a seguir.

$$\sum_{i=1}^{k} f_i$$

Podemos fazer essa leitura como: "somatório dos valores de f_i, com i variando de 1 até k". Isso significa que, simplesmente, realizaremos operações de adição entre todos os valores de f_i da tabela. No nosso contexto, no qual $k = 7$, temos a situação a seguir.

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + f_3 + f_4 + f_5 + f_6 + f_7$$

$$\sum_{i=1}^{k} f_i = 5 + 10 + 13 + 12 + 11 + 9 + 2$$

$$\sum_{i=1}^{k} f_i = 62$$

Como o somatório das frequências simples absolutas coincide com o número de elementos da população, N, resumimos essas informações com a notação a seguir.

$$\sum_{i=1}^{k} f_i = N = 62$$

Podemos, também, simplificar a notação de somatório, sem identificar os valores inicial e final de i. Nesse caso, subentende-se que i percorrerá do primeiro ao último elemento.

$$\sum f_i = N$$

Observação

Quando, em uma distribuição de frequências, aparece apenas o termo frequência, assumimos que estamos falando da frequência simples absoluta. Há outros tipos de frequência, que estudaremos em breve.

Na frequência simples absoluta, o termo simples indica que a contagem corresponde exclusivamente aos valores da classe atual, e o termo absoluta indica que o valor será expresso como o próprio resultado da contagem, e não como uma taxa unitária ou percentual.

Vamos entender, agora, como podemos distribuir dados de variáveis quantitativas contínuas em uma tabela de frequências.

Uma classe intervalar, geralmente aplicada a variáveis quantitativas contínuas, apresenta um intervalo de classe e não apenas um valor unitário, como vimos na classe discreta. Classes intervalares também podem ser aplicadas a variáveis quantitativas discretas para as quais o volume de dados brutos é muito extenso ou diverso.

Considere que uma empresa de tecnologia fabrica sensores de temperatura para sistemas de monitoramento climático em diferentes regiões. A companhia tem dados de temperatura registrados por

todos os sensores, com medições coletadas em um mesmo instante específico, para o qual as condições climáticas devem ser estudadas.

Nesse contexto, a característica de interesse é a temperatura registrada pelos sensores, em graus Celsius (°C), que representa uma variável quantitativa contínua. Os elementos da população são os próprios sensores de temperatura da empresa.

O conjunto de dados brutos dessa coleta é apresentado a seguir:

20,1; 22,5; 19,8; 24,3; 21,0; 23,4; 22,1; 25,6; 19,5; 21,8;

22,9; 20,0; 23,3; 19,9; 24,0; 23,1; 21,7; 22,3; 19,7; 21,5;

23,0; 20,4; 22,6; 21,2; 24,7; 25,0; 21,4; 22,8; 23,2; 25,3.

O técnico responsável pelo estudo pretende montar uma tabela de frequências, mas optou por organizar os dados em rol primeiro, para facilitar a visualização dos valores. O rol do conjunto de dados, em ordem crescente, é mostrado a seguir:

19,5; 19,7; 19,8; 19,9; 20,0; 20,1; 20,4; 21,0; 21,2; 21,4;

21,5; 21,7; 21,8; 22,1; 22,3; 22,5; 22,6; 22,8; 22,9; 23,0;

23,1; 23,2; 23,3; 23,4; 24,0; 24,3; 24,7; 25,0; 25,3; 25,6.

Como os dados variam de 19,5 a 25,6 °C, o técnico optou por dividir a tabela de frequências em 7 classes distintas, com amplitude de classe de 1 °C. Isso significa que a diferença entre o limite inferior e o limite superior do intervalo de cada classe será de 1 °C. Nesse cenário, a tabela de frequências montada é descrita a seguir.

Tabela 23 – Tabela de frequências com classes intervalares e somatório das frequências

i (índice)	Intervalo de classe	X_i (ponto médio de classe)	F_i (frequência)
1	19⊢20	19,5	4
2	20⊢21	20,5	3
3	21⊢22	21,5	6
4	22⊢23	22,5	6
5	23⊢24	23,5	5
6	24⊢25	24,5	3
7	25⊢26	25,5	3
			$\sum_{i=1}^{k} f_i = N = 30$

Vamos, com calma, interpretar o conteúdo da tabela. Como temos 7 classes, temos $k = 7$ e índice i variando de 1 a 7.

Cada classe, agora, é representada por um intervalo, e não mais por um único valor. Vamos considerar a classe 1, cujo intervalo é dado por 19⊢20. O que isso significa? Primeiro, identificamos que o limite inferior (o valor mínimo) da classe é 19 e que o limite superior (o valor máximo) da classe é 20. Esses são os limites do intervalo no qual o dado deve se encaixar para que seja considerado pertencente a ela.

O símbolo ⊢, que se encontra entre esses limites, indica o modo de abertura do intervalo. O traço vertical apenas à esquerda indica que o intervalo é fechado à esquerda e aberto à direita. O intervalo ser fechado à esquerda significa que, caso o valor exato do limite inferior ocorra no conjunto de dados brutos, ele fará parte do intervalo. O intervalo ser aberto à direita significa que, caso o valor exato do limite superior ocorra no conjunto de dados brutos, ele não fará parte deste intervalo e sim do próximo.

Parece complicado, mas não é. Para o intervalo 19⊢20, um possível valor 19,0 °C de temperatura integraria essa classe, já que o intervalo é fechado em 19. O valor 19,9 °C também integraria essa classe, já que ele se encontra entre o limite inferior e superior. O valor 19,999 °C (caso a precisão da medição permitisse esse tipo de valor) também integraria a classe, já que ele ainda se encontra entre o limite inferior e superior. Já o valor 20,0 °C não integra essa classe, pois o intervalo é aberto à direita, o que significa que o valor exato do limite superior não é mais contado para a classe. Desse modo, o valor 20,0 °C integra a classe cujo intervalo é descrito como 20⊢21.

A distribuição dos dados do rol nas classes da tabela, portanto, fica conforme descrito a seguir:

19⊢20: 19,5; 19,7; 19,8; 19,9 (4 ocorrências)

20⊢21: 20,0; 20,1; 20,4 (3 ocorrências)

21⊢22: 21,0; 21,2; 21,4; 21,5; 21,7; 21,8 (6 ocorrências)

22⊢23: 22,1; 22,3; 22,5; 22,6; 22,8; 22,9 (6 ocorrências)

23⊢24: 23,0; 23,1; 23,2; 23,3; 23,4 (5 ocorrências)

24⊢25: 24,0; 24,3; 24,7 (3 ocorrências)

25⊢26: 25,0; 25,3; 25,6 (3 ocorrências)

Note que a quantidade de ocorrências de dados em cada uma das classes já representa a sua frequência simples absoluta. O somatório dos valores das frequências resulta em 30, indicando que há 30 sensores participando do estudo. Isso está de acordo com o conjunto de dados brutos, que nos trazia 30 medições de temperatura.

E o tal do x_i, o que faz nessa tabela? Agora sim, explicaremos o conceito de ponto médio de classe. Temos, nesse caso, o valor médio entre o limite inferior e o limite superior do intervalo da classe. É como se quiséssemos representar cada intervalo por um único valor, que aponta para o meio do intervalo da classe. Para calcularmos o ponto médio de classe, usamos a função a seguir.

$$x_i = \frac{Li_i + Ls_i}{2}$$

Na função, x_i é o ponto médio de classe, Li_i é o limite inferior da classe e Ls_i é o limite superior da classe.

Vamos, como exemplo, calcular o ponto médio da 5ª classe da tabela, cujo intervalo é identificado como 23⊢24. Nesse caso, $Li_5 = 23$ e $Ls_5 = 24$. O cálculo é demonstrado a seguir.

$$x_5 = \frac{Li_5 + Ls_5}{2} = \frac{23 + 24}{2} = \frac{47}{2} = 23,5$$

Observação

O intervalo do tipo $Li_i \vdash Ls_i$, que vimos no exemplo, é um dos tipos mais comuns para a distribuição de dados de variáveis quantitativas contínuas. Ele permite que não exista dúvida sobre a qual classe cada dado pertence, mesmo que haja dados com diferentes números de casas decimais no conjunto de dados brutos. No entanto, é possível adotar outras aberturas para os intervalos de frequências simples, havendo quatro possibilidades:

Li_i— Ls_i: intervalo aberto à esquerda e aberto à direita.

$Li_i \vdash Ls_i$: intervalo fechado à esquerda e aberto à direita.

$Li_i \dashv Ls_i$: intervalo aberto à esquerda e fechado à direita.

$Li_i \mapsto Ls_i$: intervalo fechado à esquerda e fechado à direita.

Perceba que, ao montarmos a tabela de frequências, nós categorizamos os 30 dados originais em apenas 7 classes. Com isso, nós classificamos os valores de uma variável quantitativa.

Imagine que um estudo tenha um conjunto de 1.500 dados brutos de uma variável. Ao categorizá-los em uma tabela de frequências, fazemos a redução das informações e conseguimos enxergar com mais clareza quais intervalos apresentam maior frequência. Desse modo, o comportamento dos dados se torna mais visível. Lembre-se: a estatística é a ciência que busca extrair informações dos dados.

Saiba mais

O Google Analytics é uma ferramenta de análise que permite monitorar e analisar o tráfego de um site ou de um aplicativo. Ele oferece informações detalhadas sobre o comportamento dos visitantes, ajudando proprietários de sites e empresas a entenderem como os usuários interagem com o conteúdo e a otimizar suas estratégias de marketing digital.

O Google Analytics faz uso de conceitos de estatística para analisar e interpretar os dados coletados sobre o comportamento dos usuários. A ferramenta coleta grandes volumes de dados e, por meio de técnicas estatísticas, transforma esses dados em informações úteis para ajudar os proprietários a tomar decisões.

GOOGLE. *Google Analytics*. [s.d.]. Disponível em: https://tinyurl.com/ruppkmhh. Acesso em: 5 dez. 2024.

7.5.2 Frequência simples relativa

Agora que conhecemos a frequência simples absoluta, que é o tipo básico de uma tabela de frequências, vamos aprender a calcular um novo tipo. A frequência simples relativa expressa cada uma das frequências como uma taxa, geralmente em formato percentual, em relação ao todo. Para calculá-la, partimos do tipo básico, e o transformamos em uma taxa.

Considere as medidas de tempo de resolução de uma atividade curricular, em minutos, obtidas de um grupo de 40 candidatos em um concurso público.

Tabela 24 – Tabela de frequências do tempo de resolução da atividade curricular

Tempo (min)	F_i (f. s. absoluta)
60⊢70	12
70⊢80	14
80⊢90	11
90⊢100	1
100⊢110	1
110⊢120	0
120⊢130	1
	$\sum f_i = N = 40$

Perceba que, dessa vez, na coluna que indica os intervalos de classe, simplesmente colocamos o nome da variável em questão e sua unidade de medida. Lembre-se sempre: são os dados da variável que serão subdivididos em classes. Por isso, é muito comum que o nome dessa coluna traga o próprio nome da variável.

Para calcular a frequência simples relativa a partir desses dados, precisamos, primeiramente, saber que há dois formatos possíveis: formato unitário e formato percentual. No formato unitário, f_{ri}, calculamos a razão entre a frequência simples absoluta de cada classe pelo número total de elementos do conjunto. No formato algébrico, temos o que segue:

$$f_{ri} = \frac{f_i}{N}$$

No formato percentual, $f_{ri\%}$, transformamos a frequência simples relativa unitária em uma taxa percentual, multiplicando esse valor por 100. Temos, portanto, o formato a seguir:

$$f_{ri\%} = \frac{f_i}{N} \cdot 100 = f_{ri} \cdot 100$$

Na tabela seguinte, calculamos a frequência simples relativa unitária e a frequência simples relativa percentual dos tempos de resolução da atividade.

Tabela 25 – Tabela de frequências do tempo de resolução da atividade curricular, com os cálculos das frequências relativas

Tempo (min)	F_i (f. s. absoluta)	F_{ri} (f. s. relativa unitária)	$F_{ri\%}$ (f. s. relativa percentual)
60⊢70	12	0,300	30,0%
70⊢80	14	0,350	35,0%
80⊢90	11	0,275	27,5%
90⊢100	1	0,025	2,5%
100⊢110	1	0,025	2,5%
110⊢120	0	0	0
120⊢130	1	0,025	2,5%
–	$\sum f_i = N = 40$	$\sum f_{ri} = 1$	$\sum f_{ri\%} = 100\%$

Como exemplo, faremos dois cálculos cujos resultados já estão presentes na tabela: f_{r2} e $f_{r3\%}$.

$$f_{r2} = \frac{f_2}{N} = \frac{14}{40} = 0,35$$

$$f_{r3\%} = \frac{f_3}{N} \cdot 100 = \frac{11}{40} \cdot 100 = 0,275 \cdot 100 = 27,5\%$$

Vamos, agora, interpretar algumas informações presentes na tabela. 12 candidatos resolveram a atividade curricular entre 60 e 70 min, o que corresponde a 30% dos candidatos. 35% dos candidatos foram capazes de terminar a atividade entre 70 e 80 min, isso corresponde a 14 candidatos. A partir de 90 min, a contagem de frequência cai bastante, o que significa que, a esse ponto, a maior parte dos candidatos já havia resolvido a atividade. Além disso, nenhum candidato resolveu a atividade entre 110 e 120 min, o que trouxe uma frequência nula para a tabela.

Observe que, muitas vezes, é mais confortável interpretarmos informações relativas que foram expressas como uma taxa percentual, já que esse formato é muito presente no nosso cotidiano. No entanto, tanto a frequência simples relativa unitária quanto a frequência simples relativa percentual nos trazem a mesma informação. Apenas o formato de entrega dessa informação é diferente.

Perceba, também, que o somatório de frequências unitárias deve resultar em 1 (daí usarmos o termo unitária). Já o somatório de frequências percentuais deve resultar em 100, o que representa o nosso todo.

Observação

Algumas vezes, o cálculo de frequências relativas resultará em números irracionais, ou em números com várias casas decimais. Nesses casos, adote pelo menos 4 casas para frequências unitárias e 2 casas para frequências percentuais.

7.5.3 Frequência acumulada absoluta

Agora que aprendemos os dois tipos de frequências simples (absoluta e relativa), vamos passar para as frequências acumuladas. O termo acumulada significa que vamos considerar as contagens não apenas da classe atual, mas também de todas as classes que vieram anteriormente. Desse modo, acumulamos frequências, conforme avançamos na tabela.

Mas por que trabalhamos com frequências acumuladas? De forma geral, as frequências acumuladas complementam a tabela de frequências simples ao fornecer uma visão progressiva da distribuição dos dados. Elas permitem observar rapidamente o número de ocorrências abaixo de um determinado valor ou dentro de uma faixa específica. Além disso, são úteis para determinar algumas medidas, como a mediana, os quartis, os percentis e outras métricas relacionadas à posição dos dados.

Uma frequência acumulada absoluta acumula resultados de contagens, partindo da tabela de frequência simples absoluta. Se partirmos dos mesmos dados da tabela 24, podemos calcular a frequência acumulada absoluta, F_i, conforme mostrado a seguir.

Tabela 26 – Tabela de frequências do tempo de resolução da atividade curricular, com cálculo da frequência acumulada absoluta

Tempo (min)	F_i (f. s. absoluta)	F_i (f. ac. absoluta)
60⊢70	12	12
70⊢80	14	26
80⊢90	11	37
90⊢100	1	38
100⊢110	1	39
110⊢120	0	39
120⊢130	1	40
–	$\sum f_i = N = 40$	–

Como fizemos esses cálculos? Simplesmente, fomos acumulando resultados da frequência simples, linha a linha. Para a 1ª classe, cujo intervalo é dado como 60⊢70, houve apenas a repetição do mesmo valor, 12, já que não há classes anteriores com as quais podemos acumular contagens. A partir da 2ª classe, começamos a fazer o somatório de todos os valores de frequência simples absoluta anteriores. Como 12 + 14 = 26, esse é o resultado da frequência acumulada da 2ª classe. Para chegarmos ao valor 37, da 3ª classe, basta fazermos 12 + 14 + 11 = 37. Alternativamente, podemos somar o resultado acumulado anterior (26) à contagem absoluta atual (11) e, com isso, temos que 26 + 11 = 37. Se generalizarmos esse último modo de cálculo, temos, no formato algébrico, o que segue:

$$F_i = F_{i-1} + f_i, \text{ com } F_1 = f_1$$

Os cálculos, cujos resultados estão expostos na tabela, são realizados a seguir:

$$F_1 = f_1 = 12$$
$$F_2 = F_1 + f_2 = 12 + 14 = 26$$
$$F_3 = F_2 + f_3 = 26 + 11 = 37$$
$$F_4 = F_3 + f_4 = 37 + 1 = 38$$
$$F_5 = F_4 + f_5 = 38 + 1 = 39$$
$$F_6 = F_5 + f_6 = 39 + 0 = 39$$
$$F_7 = F_6 + f_7 = 39 + 1 = 40$$

Repare que, pelo fato de acumularmos os resultados anteriores de contagem, a última frequência acumulada absoluta (F_7) apresenta o mesmo valor do número de elementos do conjunto de dados (N) que, nesse caso, vale 40. Desse modo, para uma tabela de k classes, temos a propriedade a seguir:

$$F_k = \sum f_i = N$$

Finalmente, faremos algumas interpretações dos valores da tabela. Os resultados da frequência acumulada absoluta nos indicam que 12 candidatos resolveram a atividade curricular em menos de 70 min. 26 candidatos resolveram em menos de 80 min. 39 candidatos resolveram em menos de 120 min e assim por diante.

É possível, inclusive, mudarmos a notação de intervalo de classe, a fim de deixar mais claro esse significado. Para frequências acumuladas, os limites de classe podem ser substituídos por "menor do que" ou pelo símbolo <, que descrevem o novo intervalo.

Tabela 27 – Tabela de frequências do tempo de resolução da atividade curricular, com frequência acumulada absoluta e nova notação de intervalo

Tempo (min)	F_i (f. ac. absoluta)
< 70	12
< 80	26
< 90	37
< 100	38
< 110	39
< 120	39
< 130	40

7.5.4 Frequência acumulada relativa

Do mesmo modo que podemos calcular frequências simples relativas, conseguimos, também, calcular frequências acumuladas relativas. Seguindo o mesmo conceito que já aprendemos, vamos transformar a frequência acumulada absoluta em uma taxa, seja unitária, seja percentual. Para isso, partiremos dos dados da tabela 26. Vamos continuar os cálculos, dessa vez, calculando a frequência acumulada relativa unitária, F_{ri}, e a frequência acumulada relativa percentual, $F_{ri\%}$.

Tabela 28 – Tabela de frequências do tempo de resolução da atividade curricular, com cálculo de frequências acumuladas relativas a partir da frequência acumulada absoluta

Tempo (min)	F_i (f. ac. absoluta)	F_{ri} (f. ac. relativa unitária)	$F_{ri\%}$ (f. ac. relativa percentual)
< 70	12	0,300	30,0%
< 80	26	0,650	65,0%
< 90	37	0,925	92,5%
< 100	38	0,950	95,0%
< 110	39	0,975	97,5%
< 120	39	0,975	97,5%
< 130	40	1	100%

Os cálculos foram feitos de forma análoga aos que fizemos para calcular as frequências simples relativas, mas, dessa vez, partimos dos dados absolutos da frequência acumulada e não da frequência simples.

No formato unitário, F_{ri}, calculamos a razão entre a frequência acumulada absoluta de cada classe pelo número total de elementos do conjunto. Algebricamente, temos o que segue.

$$F_{ri} = \frac{F_i}{N}$$

No formato percentual, $F_{ri\%}$, transformamos a frequência acumulada relativa unitária em uma taxa percentual, multiplicando esse valor por 100. Temos, portanto, o formato algébrico a seguir.

$$F_{ri\%} = \frac{F_i}{N} \cdot 100 = F_{ri} \cdot 100$$

No caso da frequência unitária, sempre devemos chegar ao valor 1 na última classe. Isso acontece porque a última classe abrange todos os elementos do conjunto de dados com o qual trabalhamos. Algo semelhante acontece com o formato percentual, no qual devemos atingir 100% na última classe, afinal, nesse contexto, todos os candidatos concluíram a atividade em menos de 130 min. Se olharmos para outros dados da tabela, podemos inferir que 95% dos candidatos concluíram a atividade em menos de 100 min, sendo que apenas 30% foram capazes de terminá-la em menos de 70 min.

7.5.5 Histograma e polígono de frequências

É muito comum distribuições de frequências com classes intervalares serem apresentadas na forma de gráficos, como o histograma, que nos trazem uma ideia visual bastante clara sobre a distribuição dos dados.

Um histograma é um gráfico de colunas de mesma largura, geralmente aplicado a distribuições com classes intervalares. As colunas são desenhadas de forma adjacente umas às outras. Para classes discretas, é mais comum representarmos os dados por um gráfico de colunas convencional (que veremos ainda neste capítulo).

A escala horizontal representa os valores nas quais as classes estão subdivididas (geralmente são os limites de classe, mas é possível usar os pontos médios de classe, alternativamente) e a escala vertical representa um tipo de frequência simples (seja absoluta, seja relativa).

Na figura a seguir, vemos um histograma que representa os dados da tabela 24, que utiliza frequências simples absolutas. Ele é a representação visual dos dados do tempo de resolução da atividade curricular por parte dos candidatos. Os dados da tabela foram posicionados novamente à esquerda, apenas para comparação.

Tempo (min)	F_i
60⊢70	12
70⊢80	14
80⊢90	11
90⊢100	1
100⊢110	1
110⊢120	0
120⊢130	1

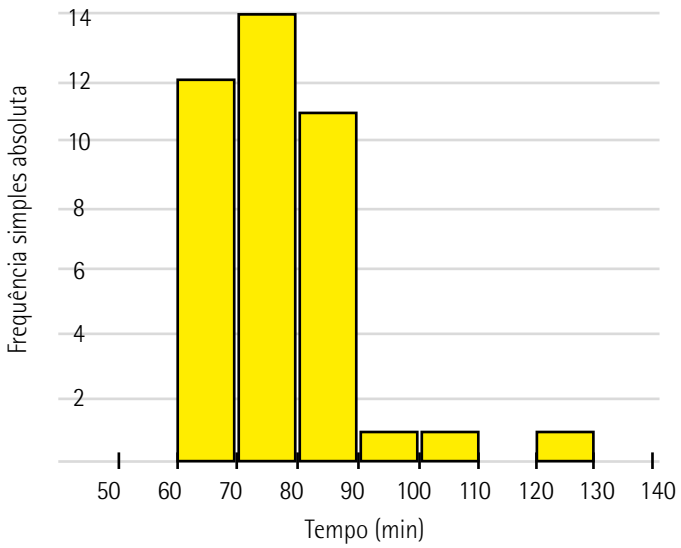

Figura 51 – Histograma criado a partir da coluna de frequência simples absoluta, encontrada na tabela 24

Perceba que o eixo horizontal representa os valores da variável, com a exibição dos valores usados como limites de classe. Cada coluna, portanto, representa uma classe da distribuição de frequências. Sua base se estende ao longo dos limites de classe e a sua altura representa a frequência. A coluna mais alta do gráfico é justamente a que representa a classe de intervalo 70⊢80, cuja frequência é 14 (o valor mais alto da tabela). A ausência de uma coluna entre os valores 110 e 120 do eixo horizontal indica a frequência 0, observada para a classe de intervalo 110⊢120.

Alternativamente, é possível construirmos histogramas utilizando um dos tipos de frequência simples relativa. Nesse caso, o formato do gráfico de frequências absolutas se mantém, apenas os valores da escala vertical indicarão a taxa unitária ou percentual associada a cada classe. Para montar esse gráfico, disponível na figura a seguir, utilizamos os dados de frequência simples relativa percentual da tabela 25, reproduzidos à esquerda.

Tempo (min)	$F_{ri\%}$
60⊢70	30,0%
70⊢80	35,0%
80⊢90	27,5%
90⊢100	2,5%
100⊢110	2,5%
110⊢120	0
120⊢130	2,5%

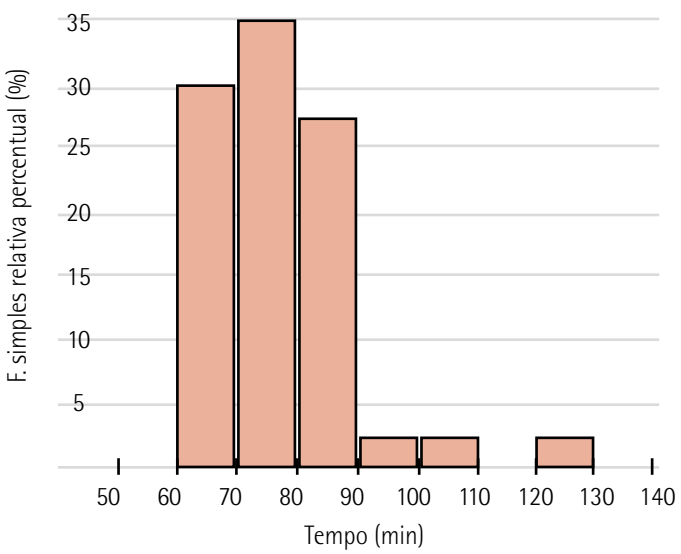

Figura 52 – Histograma criado a partir da coluna de frequência simples relativa percentual, encontrada na tabela 25

Representações gráficas das distribuições de frequência não se resumem a histogramas. A partir de frequências simples, também podemos construir um polígono de frequências, que é um gráfico de linhas que une os pontos médios (x_i) de classes adjacentes por segmentos de reta.

Lembrete

O ponto médio de classe corresponde ao valor médio entre o limite inferior e o limite superior do intervalo da classe.

Na figura a seguir, um polígono de frequências foi construído acima do histograma apresentado na figura 51. Os segmentos do polígono, em vermelho, constituem esse tipo de gráfico. Os valores posicionados em roxo mostram os pontos médios de classe.

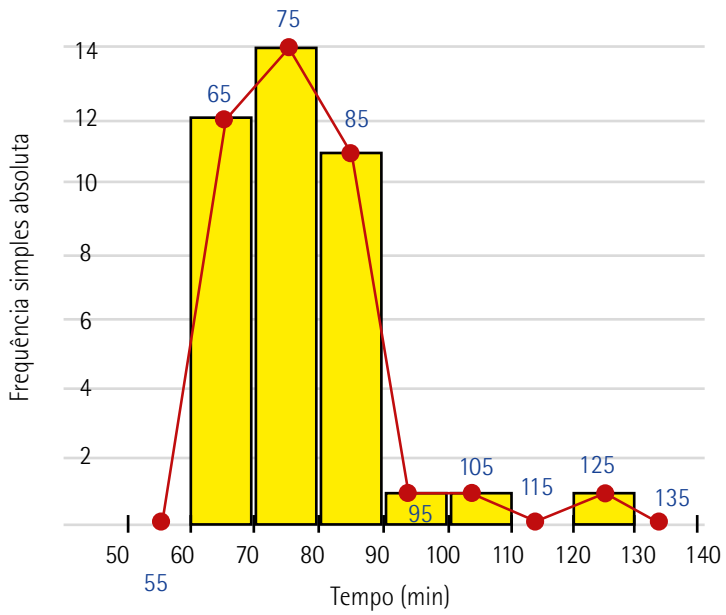

Figura 53 – Polígono de frequências (em vermelho) construído acima do histograma da figura 51

Devemos iniciar a construção do polígono uma classe antes da primeira classe da tabela, e terminarmos uma classe depois da disponível na tabela. Note que aparecem dois pontos médios de classe adicionais: 55 e 135. É como se criássemos uma classe anterior, de intervalo 50⊢60, e uma classe posterior, de intervalo 130⊢140. Isso é feito para que a imagem do polígono fique fechada no eixo horizontal, ou seja, ele é iniciado e encerrado na altura da frequência 0.

As classes adicionais devem ter a mesma amplitude daquelas existentes na tabela. A amplitude de classe corresponde à diferença entre o limite superior e inferior de cada classe. No caso, temos que a amplitude de cada classe vale 10.

A altura de cada ponto do polígono de frequências, por sua vez, corresponde à frequência da classe representada.

Se isolarmos o polígono de frequências do histograma, temos um gráfico composto por linhas conectadas, que pode ser observado na figura a seguir.

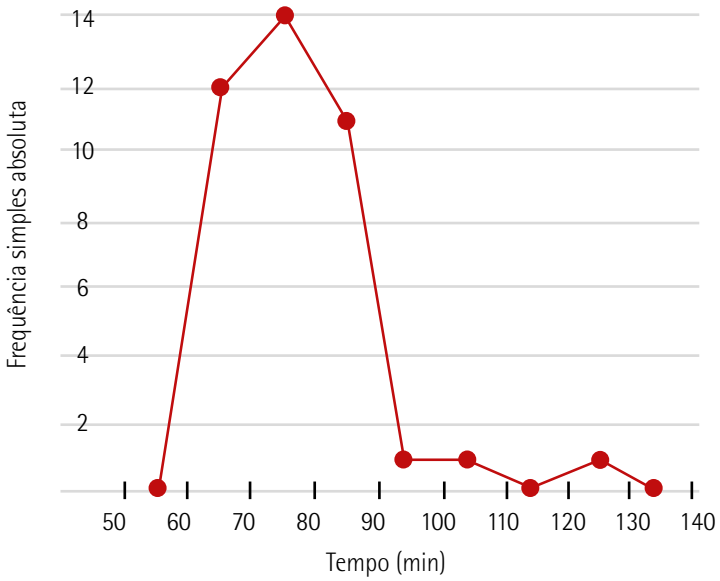

Figura 54 – Polígono de frequências da figura 53, isolado do histograma

Saiba mais

As frequências acumuladas podem ser representadas graficamente por uma ogiva, um gráfico de linhas que conecta os pontos correspondentes às frequências acumuladas. A ogiva é particularmente útil para análises rápidas e interpretativas, sendo uma ferramenta comum em estudos descritivos e na apresentação de dados.

Com a finalidade de saber mais sobre essa ferramenta gráfica, consulte o capítulo 2 do livro:

TRIOLA, M. F. *Introdução à estatística*. Rio de Janeiro: LTC, 2023.

Ainda, existe uma ferramenta da Social Science Statistics que nos auxilia a construir histogramas.

Disponível em: https://www.socscistatistics.com/. Acesso em: 7 dez. 2024.

Outra possibilidade é utilizar o Statdisk Online.

Disponível em: https://www.statdisk.com/. Acesso em: 7 dez. 2024.

7.6 Outros tipos de gráficos

Na seção anterior deste livro-texto, vimos que histogramas e polígonos de frequência são gráficos que representam visualmente a distribuição de frequências de determinado conjunto de dados. Nesta seção, veremos outros tipos de gráficos que nos ajudam a visualizar informações oriundas de conjuntos de dados, que não necessariamente estão divididos em uma tabela de frequências. Vários desses gráficos você, provavelmente, já viu em jornais, revistas, noticiários televisivos, portais de notícias, apresentações corporativas, livros didáticos ou boletins informativos. Além disso, gráficos são uma ferramenta estatística muito utilizada em artigos científicos, e contribuem com a disseminação do conhecimento científico.

É muito importante que você, profissional de tecnologia, conheça os princípios construtivos de gráficos e saiba interpretar os dados que eles nos trazem.

7.6.1 Gráfico de setores

Segundo Triola (2023), um gráfico de setores, popularmente conhecido como gráfico de pizza, é uma representação visual de dados geralmente oriundos de variáveis qualitativas. Esses dados são retratados como setores (ou "fatias") de um círculo, sendo que cada setor tem área proporcional à contagem de frequência da categoria. É muito comum que essa frequência seja expressa de maneira relativa percentual.

Observe a figura a seguir, nela vemos um gráfico de setores que traz as disciplinas preferidas dos alunos do 3º ano do Ensino Médio de uma escola fictícia, chamada Stephen Hawking. Nesse contexto, cada aluno respondeu à pergunta: "Qual é a sua disciplina preferida?". Temos, desse modo, a variável qualitativa nominal "disciplina preferida". Ao organizarmos as respostas por frequência, chegamos ao gráfico a seguir:

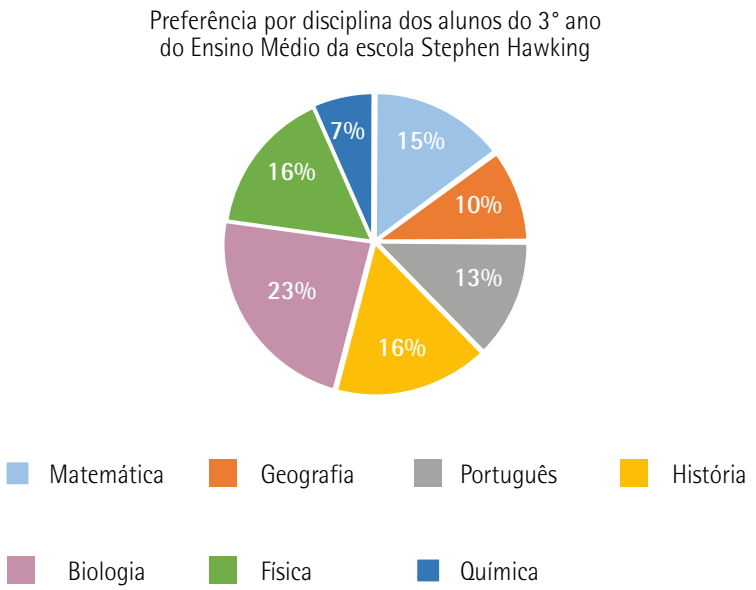

Figura 55 – Gráfico de setores no formato percentual

Observe que o gráfico de setores traz uma legenda que identifica cada um dos setores. No caso, cada setor representa uma disciplina, ou seja, uma categoria da variável. Lendo os dados do gráfico, podemos concluir que 23% dos alunos participantes da pesquisa responderam que consideram Biologia como sua disciplina favorita. Apenas 7% dos alunos consideram Química. Já Matemática, por exemplo, é a disciplina escolhida por 15% dos alunos respondentes.

Note que não conseguimos inferir a quantidade de alunos respondentes apenas lendo os dados do gráfico, já que as informações foram trazidas exclusivamente no formato relativo. No caso, esperamos que o somatório das taxas resulte em 100%, assim como acontece com a frequência simples relativa percentual das tabelas de frequência.

7.6.2 Gráfico de colunas

Já vimos que um histograma é um tipo de gráfico de colunas, no qual colunas representam os dados. Agora, seremos apresentados à representação mais usual desse tipo de gráfico, que se adequa muito bem a dados oriundos de variáveis qualitativas. Nesse formato, também são expressas distribuições de frequências com classes discretas, geralmente oriundas de variáveis quantitativas discretas.

Observe a figura a seguir, na qual vemos um gráfico de colunas que traz, novamente, a disciplina preferida dos alunos do 3º ano do Ensino Médio da escola fictícia Stephen Hawking. No entanto, aqui, esses dados são apresentados em outro formato, nos indicando o resultado da contagem de alunos que preferem cada uma das disciplinas. Desse modo, optamos por apresentar os dados no formato absoluto, não no formato relativo.

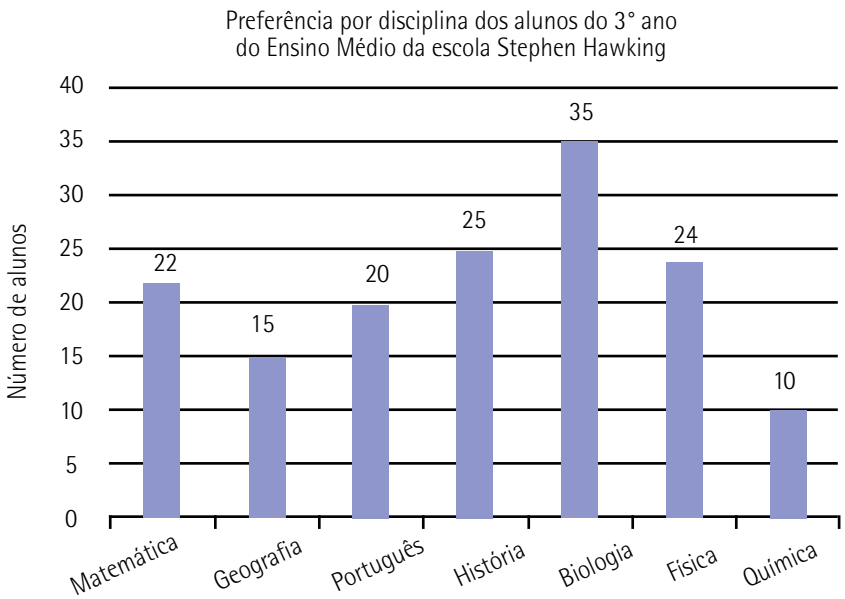

Figura 56 – Gráfico de colunas no formato absoluto

Perceba que, quando não correspondem a histogramas, os gráficos de colunas geralmente vêm com espaços de separação entre as colunas, já que, dessa vez, cada coluna apresenta uma categoria nominal ou discreta, e não um intervalo.

Assim como todos os outros gráficos vistos, um gráfico de colunas pode trazer dados no formato absoluto ou no formato relativo. O eixo horizontal traz as categorias ou valores discretos, enquanto o eixo vertical traz as frequências.

Vamos, agora, fazer algumas interpretações dos dados presentes no gráfico. Perceba que a coluna da disciplina Biologia é a mais alta de todas, indicando que essa disciplina foi a escolhida com mais frequência como resposta. Podemos inferir que 35 alunos consideram Biologia como disciplina preferida. 25 alunos escolheram História, 24 alunos escolheram Física, até chegarmos a Química, que foi escolhida por apenas 10 alunos. Naturalmente, a coluna referente à disciplina Química é a mais baixa do gráfico.

Por termos dados absolutos, podemos inferir quantos são os alunos respondentes:

22 + 15 + 20 + 25 + 35 + 24 + 10 = 151.

7.6.3 Gráfico de barras

Um gráfico de barras é semelhante a um gráfico de colunas, com a diferença na orientação das barras, que, nesse caso, se estendem na horizontal. Esse tipo de gráfico também é adequado para variáveis qualitativas, mas costuma ser mais adequado para dispor categorias com nomes longos, já que permite uma visualização mais clara para rótulos textuais grandes.

Considere que, devido a uma mudança na gestão, o conselho escolar de uma instituição privada de ensino assumiu a responsabilidade de escolher um novo nome para a escola. Para isso, foi realizada uma pesquisa entre os membros do conselho, na qual cada um indicou o nome de um entre cinco cientistas sugeridos, considerando critérios como relevância acadêmica e legado científico. O cientista mais votado será o escolhido para dar nome à instituição. O resultado da pesquisa foi divulgado ao conselho, por meio do gráfico de barras mostrado na figura a seguir.

Figura 57 – Gráfico de barras no formato absoluto

Perceba que, nesse caso, temos a variável qualitativa nominal "cientista favorito". Os cinco possíveis valores assumidos pela variável são: Alan Turing, Marie Curie, Isaac Newton, Charles Darwin e Stephen Hawking. Como cada uma dessas categorias traz um nome relativamente longo, dispor esses dados em um gráfico de colunas seria possível, mas desconfortável. Desse modo, as categorias foram dispostas no eixo vertical e a contagem no eixo horizontal, de forma a compor um gráfico de barras.

Pelos dados do gráfico, podemos inferir que o vencedor foi Stephen Hawking que, após a pesquisa, passou a nomear a instituição. 12 membros do conselho consideraram que ele é o melhor nome, entre os cinco, em termos de relevância acadêmica e legado científico. Desse modo, a barra que acompanha o rótulo "Stephen Hawking" é a que mais se estende no gráfico, atingindo o valor 12.

7.6.4 Diagrama de dispersão

De acordo com Triola (2023), um diagrama de dispersão é um gráfico de pares de dados quantitativos no formato (x, y), com um eixo x horizontal e um eixo y vertical. O eixo horizontal é usado para a primeira variável (x) e o eixo vertical para a segunda variável (y). Assim, temos a disposição de um plano cartesiano, nos quais pontos relativos ao estudo serão marcados.

O padrão dos pontos marcados no gráfico é, em geral, útil para determinar a existência, ou não, de uma correlação entre as variáveis. Esse tipo de gráfico é amplamente utilizado em artigos científicos, especialmente os que envolvem pesquisa experimental.

Considere que uma engenheira de materiais faz, em laboratório, deposições de um material semicondutor chamado óxido de índio dopado com estanho (ITO) em lâminas de vidro. Essa camada de ITO tem a intenção de atuar como eletrodo transparente no vidro. Para caracterizar o equipamento de deposição, a engenheira mediu a taxa de deposição do ITO (em nanômetros por minuto – nm/min) em função da potência de radiofrequência (RF) ajustada no equipamento. Os dados desse estudo foram plotados no gráfico de dispersão constante na figura a seguir.

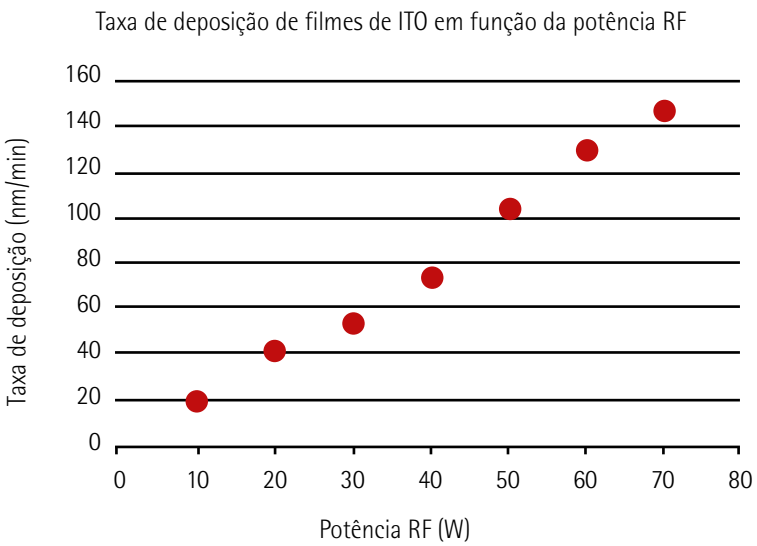

Figura 58 – Diagrama de dispersão

Nesse caso, temos duas variáveis quantitativas contínuas se relacionando no gráfico: a potência RF (no eixo x) e a taxa de deposição (no eixo y). Note que, ao aumentar a potência, a taxa de deposição tende a aumentar, formando um padrão de pontos aproximadamente linear. Isso indica uma correlação positiva entre as variáveis potência RF e taxa de deposição. Como o gráfico mostra um padrão crescente, aproximadamente linear, é possível observar uma tendência que poderia ser modelada por uma linha reta. Isso significa que uma função de 1º grau pode descrever, aproximadamente, a correlação observada experimentalmente.

Saiba mais

O ajuste de curvas é uma técnica estatística utilizada para encontrar uma função matemática que melhor descreve o relacionamento entre um conjunto de dados experimentais ou observados.

Quando coletamos dados e os plotamos em um gráfico, os pontos podem não formar uma reta ou curva exata, mas, muitas vezes, seguem um padrão. O ajuste de curvas cria uma função que representa esse padrão de maneira simplificada e previsível, por meio de uma lei de formação.

Para saber mais sobre esse assunto, consulte o capítulo 10 do livro:

TRIOLA, M. F. *Introdução à estatística*. Rio de Janeiro: LTC, 2023.

Diagramas de dispersão podem, muitas vezes, trazer dados um pouco mais confusos do que os vistos na figura anterior. Observe o diagrama de dispersão a seguir, cujos dados trouxeram as circunferências da cintura (em cm) e as circunferências dos braços (em cm) de um grupo de 40 homens. O diagrama nos mostra os pontos das medidas emparelhadas cintura/braço.

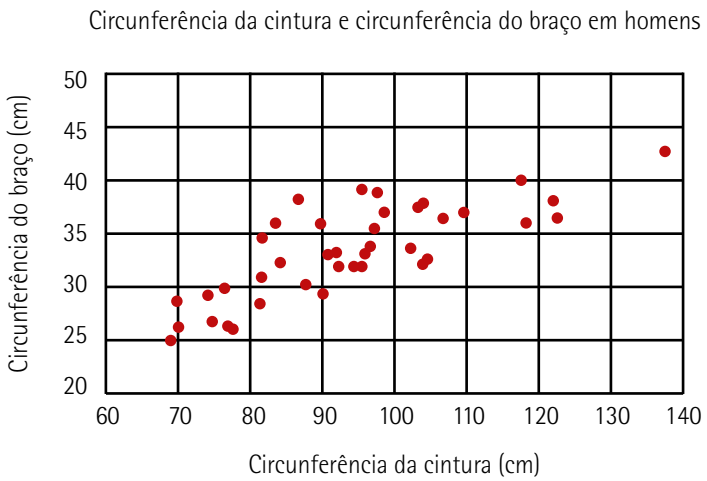

Figura 59 – Diagrama de dispersão

Adaptada de: Triola (2023, p. 59).

Mesmo sendo mais numerosos e com um padrão menos evidente de forma visual, os pontos mostram um padrão de valores crescentes da esquerda para a direita, assim como observamos no gráfico da figura anterior. Esse padrão sugere que existe uma correlação direta entre a circunferência da cintura e a circunferência do braço nos homens: quando uma aumenta, a outra tende a aumentar também. Do mesmo modo, esses dados podem ser ajustados por uma curva, que trará uma lei de formação de função que descreve essa correlação.

7.6.5 Gráfico de série temporal

Um gráfico de série temporal utiliza dados quantitativos que foram coletados ao longo de um período em determinada frequência, como diária, semanal, mensal, trimestral ou anualmente. Nesse caso, é muito comum que o gráfico seja disposto com pontos unidos entre si por segmentos de reta, tal qual fizemos com o polígono de frequências.

De acordo com Sharpe, Veaux e Velleman (2011), os diagramas de séries temporais, comumente, mostram uma grande variação de ponto a ponto e, muitas vezes, é difícil identificar uma tendência apenas de forma visual.

O gráfico de série temporal da figura a seguir mostra o faturamento mensal da empresa fictícia ConnectAI Software, ao longo do ano de 2024.

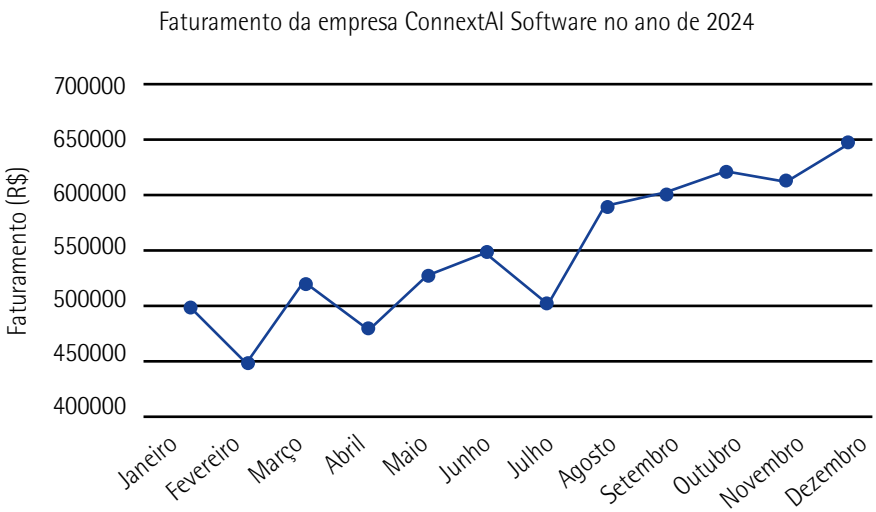

Figura 60 – Gráfico de série temporal

Nesse caso, o período de coleta de dados foi de um ano, pois ela ocorreu ao longo de todo o ano de 2024. Já a frequência de coleta foi mensal.

Pelas informações do gráfico, podemos inferir que o menor faturamento do ano de 2024 ocorreu no mês de fevereiro, cujo valor foi de aproximadamente R$ 450.000. Já o maior faturamento foi observado no mês de dezembro, com valor de, aproximadamente, R$ 650.000.

Gráficos de séries temporais são amplamente utilizados para descrever dados que variam ao longo do tempo, como preços de ações, taxas de câmbio, índices de inflação, índices meteorológicos, evoluções de taxas de mortalidade, pesquisas de intenção de voto, crescimento populacional, entre outros.

8 MEDIDAS DE TENDÊNCIA CENTRAL E DE DISPERSÃO

O capítulo 8 do nosso livro-texto será dedicado ao estudo das medidas de tendência central e das medidas de dispersão. Vamos começar nossos estudos calculando alguns tipos de tendência central, como média ou mediana, e, posteriormente, passaremos aos cálculos de dispersão, como desvio-padrão.

8.1 Medidas de tendência central

Uma medida de tendência central é uma medida que tenta representar o centro de um conjunto de dados quantitativos com apenas um valor. Existem diversos tipos de medidas de tendência central, sendo que a média aritmética é a principal delas. Veremos, a seguir, o cálculo da média aritmética em dois formatos: simples e ponderado.

8.1.1 Média aritmética simples

A média aritmética simples é uma medida de tendência central na qual realizamos o somatório dos valores dos elementos do conjunto de dados e, em seguida, dividimos o resultado pelo número de elementos. Há uma brincadeira que ilustra bem o sentido da média: se você come uma barra de chocolate e eu não como nenhuma, na média, cada um de nós comeu meia barra de chocolate.

A média aritmética é geralmente representada pela letra grega μ. Sendo:

- N o número de elementos do conjunto de dados.
- X uma variável que representa os elementos da variável quantitativa estudada.
- i um índice de contagem de elementos (que vai de 1 até N).

Temos, algebricamente, o que segue:

$$\mu = \frac{X_1 + X_2 + X_3 + ... + X_N}{N}$$

Essa mesma fórmula pode ser representada de uma forma mais enxuta e genérica, utilizando o símbolo de somatório no numerador:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

O que a fórmula diz é o seguinte: faça o somatório dos valores X, cujo índice i varia de 1 até N, e divida o resultado desse somatório por N. Se você não entendeu muito bem, acompanhe o exemplo na sequência.

Exemplo de aplicação

Uma turma do Ensino Fundamental da escola Stephen Hawking é composta por 10 alunos. Os alunos fizeram uma prova de conhecimentos gerais, cujas notas são apresentadas na tabela a seguir.

Tabela 29

Notas dos alunos									
9,0	8,5	7,0	9,5	6,5	5,0	6,5	7,5	8,0	8,5

Nesse contexto, qual é a média que a turma obteve na prova?

Resolução

O enunciado pede para calcularmos a média aritmética simples que a turma obteve na prova de conhecimentos gerais. Temos as notas de todos os indivíduos da turma e, portanto, conhecemos toda a população do contexto. Com isso, conseguimos calcular a média populacional μ.

Como temos 10 elementos na população, temos N=10 nesse contexto, X é uma variável que representa as notas disponíveis. Lembrando que i é um índice de contagem dos elementos que vai de 1 até N, nossos dados X_1 variam de X_1 até X_{10}. Observe a tabela das notas expandida a seguir.

Tabela 30 – Notas dos alunos associadas a valores da variável X

9,0	8,5	7,0	9,5	6,5	5,0	6,5	7,5	8,0	8,5
X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}

Logo, no contento em estudo, temos $X_1 = 9{,}0$, $X_2 = 8{,}5$, $X_3 = 7{,}0$ e assim por diante, até chegarmos a $X_{10} = 8{,}5$. Vamos, portanto, fazer o somatório desses valores e dividir o resultado por 10, conforme mostrado a seguir:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7 + X_8 + X_9 + X_{10}}{10}$$

$$\mu = \frac{9{,}0 + 8{,}5 + 7{,}0 + 9{,}5 + 6{,}5 + 5{,}0 + 6{,}5 + 7{,}5 + 8{,}0 + 8{,}5}{10} = \frac{76}{10} = 7{,}6$$

Logo, a nota média da turma como um todo foi de 7,6 pontos. Note que, em vez de pensarmos em fórmulas, simplesmente podemos pensar em somar todos os valores das observações e dividir o resultado obtido pelo número de elementos somados. É exatamente isso o que a fórmula nos pede para fazer.

Observação

A unidade de medida da média aritmética acompanha a unidade de medida dos dados da amostra ou da população. Isso valerá para todas as medidas de tendência central que veremos. Por exemplo, se temos dados da altura de uma população de indivíduos, medida em metros, podemos calcular a média, cujo resultado será dado também em metros.

8.1.2 Média aritmética ponderada e de distribuições de frequência

A média que acabamos de estudar é chamada de média aritmética simples porque atribui a mesma importância a cada um dos dados do conjunto. Existe outro tipo de média que estudaremos agora, que é a média aritmética ponderada, na qual esse cenário será modificado. Nesse tipo de medida de tendência central, damos pesos diferentes para determinadas medidas. Isso denota que damos importâncias distintas para diferentes dados da nossa variável.

Temos, portanto, dois conjuntos de dados com o qual lidar: o conjunto "Valores", que nos trazem os elementos X_i da variável quantitativa em estudo, e o conjunto "Pesos", que indicam a importância dada a cada valor X_i. Esses conjuntos são representados a seguir.

Valores = $\{X_1, X_2, ..., X_N\}$ (conjunto de valores da variável)

Pesos = $\{P_1, P_2, ..., P_N\}$ (conjunto de pesos de cada X_i)

Para calcularmos uma média ponderada, multiplicamos cada valor X_i pelo seu respectivo peso P_i. Depois, fazemos o somatório desses produtos. Por fim, o resultado é dividido pelo somatório de todos os pesos. Algebricamente, a média aritmética ponderada μ_P pode ser descrita conforme exposto a seguir.

$$\mu_P = \frac{(X_1 \cdot P_1) + (X_1 \cdot P_1) + ... + (X_N \cdot P_N)}{P_1 + P_2 + ... + P_N}$$

Essa mesma fórmula pode ser representada de uma forma mais enxuta e genérica, utilizando o símbolo de somatório, da seguinte maneira:

$$\mu_P = \frac{\sum_{i=1}^{N} X_i \cdot P_i}{\sum_{i=1}^{N} P_i}$$

Vamos acompanhar, a seguir, um exemplo de aplicação que nos ajudará a entender melhor esse conceito.

Exemplo de aplicação

Durante um curso, um aluno obteve as notas 7, 6 e 5 nas três provas que constituem a avaliação geral do seu aprendizado. Como as provas tinham conteúdos cumulativos, foram atribuídos pesos 2, 3 e 5, respectivamente, a cada uma delas. Qual foi a média final deste aluno?

Resolução

Nesse contexto, temos os dados descritos a seguir:

Valores = $\{X_1, X_2, X_3\} = \{7, 6, 5\}$

Pesos = $\{P_1, P_2, P_3\} = \{2, 3, 5\}$

$$\mu_P = \frac{\sum_{i=1}^{N} X_i \cdot P_i}{\sum_{i=1}^{N} P_i} = \frac{7 \cdot 2 + 6 \cdot 3 + 5 \cdot 5}{2 + 3 + 5} = \frac{14 + 18 + 25}{10} = \frac{57}{10} = 5{,}7$$

Portanto, a média final do aluno no curso foi 5,7.

Repare que, quanto maior é o peso, maior é a importância dada ao valor X_i. Nesse caso, a 3ª prova é a mais importante de todas, já que seu peso é 5 (o maior valor entre os três pesos). Isso significa que ela contribuirá mais com o valor da média do que as outras provas. A prova menos importante é a 1ª, já que seu peso é 2 (o menor valor entre os três pesos). Isso significa que ela contribuirá menos com o valor da média do que as outras provas.

E se tivermos acesso apenas a uma distribuição de frequências e não conhecermos os dados brutos, conseguimos calcular a média aritmética? Sim, e o formato de cálculo é análogo ao formato de média ponderada que acabamos de aprender. Para isso, precisamos dos dados da frequência simples absoluta, f_i.

Utilizaremos, nesse caso, os pontos médios de classe, x_i (atuando como os valores X_i da média ponderada) e as frequências simples absolutas, f_i (atuando como pesos P_i).

Algebricamente, a média aritmética de uma distribuição de frequências μ_f de k classes pode ser descrita conforme exposto a seguir:

$$\mu_f = \frac{(x_1 \cdot f_1) + (x_1 \cdot f_1) + \ldots + (x_k \cdot f_k)}{f_1 + f_2 + \ldots + f_k}$$

Essa mesma fórmula pode ser representada de uma forma mais enxuta e genérica, utilizando o símbolo de somatório, da seguinte maneira:

$$\mu_f = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{\sum_{i=1}^{k} f_i}$$

Como o somatório das frequências corresponde ao número de elementos do conjunto de dados, N, também podemos expressar o formato algébrico da média aritmética da seguinte forma:

$$\mu_f = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N}$$

Vamos acompanhar, a seguir, um exemplo de aplicação.

Exemplo de aplicação

1) Calcule a média aritmética dos dados da tabela de frequências a seguir.

Tabela 31 – Tabela de frequências

X_i	F_i
30	2
31	5
32	1
33	6
34	5
35	1

Resolução

Temos, nesse caso, uma tabela de frequências com classes discretas. Nesse caso, o próprio valor associado a cada classe é o ponto médio de classe. Devemos, então, multiplicar cada valor x_i por sua frequência correspondente, f_i. Faremos isso na própria tabela, criando uma coluna adicional.

Tabela 32 – Tabela de frequências com cálculo de $X_i \cdot F_i$

X_i	F_i	$X_i \cdot F_i$
30	2	60
31	5	155
32	1	32
33	6	198
34	5	170
35	1	35

Agora que já temos cada valor individual de $x_i \cdot f_i$, faremos o somatório desses valores, cujo resultado deverá compor o numerador de μ_f. Vamos aproveitar, também, para fazer o somatório dos valores de f_i, cujo resultado deverá compor o denominador de μ_f.

Tabela 33 – Tabela de frequências com cálculo do somatório de $X_i \cdot F_i$ e com o cálculo do somatório de F_i

X_i	F_i	$X_i \cdot F_i$
30	2	60
31	5	155
32	1	32
33	6	198
34	5	170
35	1	35
	$\sum f_i = N = 20$	$\sum x_i \cdot f_i = 650$

Com isso, já sabemos que o valor do numerador de μ_f é 650, bem como que o valor do denominador de μ_f é 20. Calcular a média aritmética, agora, ficou fácil. Acompanhe a seguir.

$$\mu_f = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N} = \frac{650}{20} = 32,5$$

Portanto, a média aritmética do conjunto de dados distribuídos na tabela de frequências é 32,5.

2) Calcule a média aritmética dos dados da tabela de frequências a seguir.

Tabela 34 – Tabela de frequências

Classes	F_i
10⊢20	8
20⊢30	6
30⊢40	10
40⊢50	6
50⊢60	10

Resolução

Temos, nesse caso, uma tabela de frequências com classes intervalares. A coluna chamada de "Classes" nos traz o intervalo de cada classe da tabela. Vamos, em um primeiro momento, calcular o ponto médio (x_i) de cada classe, lembrando que o ponto médio se trata da média aritmética simples entre o limite inferior e superior de cada classe.

Tabela 35 – Tabela de frequências com cálculo do ponto médio de classe, X_i

Classes	X_i	F_i
10⊢20	15	8
20⊢30	25	6
30⊢40	35	10
40⊢50	45	6
50⊢60	55	10

Agora, faremos a multiplicação $x_i \cdot f_i$ e o somatório da coluna $x_i \cdot f_i$ com a coluna f_i.

Tabela 36 – Tabela de frequências com cálculo de $X_i \cdot F_i$ e somatórios

Classes	X_i	F_i	$X_i \cdot F_i$
10⊢20	15	8	120
20⊢30	25	6	150
30⊢40	35	10	350
40⊢50	45	6	270
50⊢60	55	10	550
		$\sum f_i = N = 40$	$\sum x_i \cdot f_i = 1440$

Já sabemos que o valor do numerador de μ_f é 1440, bem como que o valor do denominador de μ_f é 40. Calcular a média aritmética, agora, ficou fácil.

$$\mu_f = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N} = \frac{1440}{40} = 36$$

Portanto, a média aritmética do conjunto de dados distribuídos na tabela de frequências é 36.

Lembrete

A unidade de medida da média aritmética acompanha a unidade de medida dos dados da amostra ou da população. Nos exemplos anteriores, não colocamos unidades de medida nos valores de média porque não conhecemos a unidade dos dados, já que se tratam de exemplos sem contextualização.

8.1.3 Mediana

A média aritmética é, sem dúvidas, a medida de tendência central mais utilizada na análise de dados. No entanto, em alguns casos, a mediana é calculada por trazer valores mais condizentes com o valor central do conjunto de dados, principalmente em conjuntos nos quais existem valores atípicos, também conhecidos como *outliers*.

Observação

Valores atípicos são observações ou dados que se desviam significativamente dos demais do conjunto, podendo indicar algo anômalo ou excepcional.

A mediana (Me) de um conjunto de dados é definida como o valor numérico que separa as metades superior e inferior de um conjunto de dados ordenados em rol.

Considere o conjunto de valores a seguir:

Valores = {9, 3, 11, 3, 6}

Primeiramente, vamos ordenar os cinco valores em rol:

Rol = {3, 3, 6, 9, 11}

A mediana, nesse caso, coincide com o valor central, que separa a metade inferior da metade superior do conjunto:

$$Rol = \{3, 3, 6, 9, 11\} \rightarrow Me = 6$$

Temos, portanto, que a mediana desse conjunto de dados vale 6, já que esse valor se encontra na posição central do rol. Esse tipo de raciocínio funciona para conjuntos de dados que apresentam N **ímpar** (nesse caso, N=5). Em conjuntos cujo número de elementos é ímpar, há apenas um valor central, que coincide com o valor da mediana.

Algebricamente, para N ímpar, a mediana é o valor X_i que ocupa a posição i do rol dada por:

$$i' = \frac{N+1}{2}$$

Logo, temos que:

$$Me_{ímpar} = X_{i'}$$

Vamos acompanhar os exemplos, a seguir.

Exemplo de aplicação

1) A equipe de testes da empresa iSoftGames, produtora do jogo Crazy Pigeons, coletou a pontuação, que poderia variar de 0 a 100, de cada um dos 7 desenvolvedores que participaram do teste alfa do jogo. As pontuações conquistadas foram as seguintes: 52, 41, 37, 82, 24, 63 e 68. Encontre a média aritmética e a mediana desta série de dados.

Resolução

Como temos 7 pontuações, temos $N = 7$, que representa um número ímpar. Vamos, a princípio, montar o rol dos valores da série.

$$Rol = \{24, 37, 41, 52, 63, 68, 82\}$$

Vamos, a seguir, calcular a média aritmética.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{24+37+41+52+63+68+82}{7} = 52,8$$

Agora, vamos encontrar a posição da mediana.

$$i' = \frac{N+1}{2} = \frac{7+1}{2} = \frac{8}{2} = 4$$

Isso significa que o elemento X_4=52 do rol representa a mediana. Logo, $Me = 52$.

Portanto, a média de pontuação encontrada no teste alfa do jogo foi de 52,8 pontos, e a mediana foi de 52 pontos.

Note que os valores são relativamente parecidos, já que se tratam de um conjunto de dados com valores relativamente bem distribuídos, sem a existência de valores atípicos.

2) Sabe-se que o número de casos de certa doença nos meses de janeiro dos últimos 7 anos foi: 52, 41, 37, 1.000, 24, 63 e 68. Encontre a média aritmética e a mediana deste conjunto de dados.

Resolução

$N = 7$ (ímpar)

Rol = {24, 37, 41, 52, 63, 68, 1000}

Média aritmética:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{24+37+41+52+63+68+1000}{7} = \frac{1285}{7} = 183{,}6$$

Mediana:

$$i = \frac{N+1}{2} = \frac{7+1}{2} = \frac{8}{2} = 4$$

$$Me = X_4 = 52$$

Portanto, em janeiro dos últimos 7 anos, a média de casos da doença foi de 183,6 casos, e a mediana de 52 casos.

Note que os valores de tendência central, dessa vez, foram muito diferentes. Isso aconteceu por conta da existência do valor atípico 1.000, que difere muito dos outros valores encontrados no conjunto de dados. Esse valor afeta muito a média, que foi puxada para cima. No entanto, a mediana permaneceu inalterada em relação ao conjunto de dados do exemplo anterior.

A mediana de um conjunto de dados cujo N é par é dado pela média aritmética dos dois valores centrais do rol. Considere o conjunto de valores a seguir:

Valores = {9, 12, 3, 11, 3, 5}

Primeiramente, vamos ordenar os seis valores em rol.

Rol = {3, 3, 5, 9, 11, 12}

A mediana, nesse caso, coincide com a média dos dois valores centrais já que, nesse caso, não há apenas um único valor que separa a metade inferior da metade superior do conjunto.

$$\text{Rol} = \{3, 3, 5, 9, 11, 12\} \rightarrow \text{Me} = \frac{5+9}{2} = 7$$

Temos, portanto, que a mediana desse conjunto de dados vale 7, já que esse é o valor da média entre os dois valores centrais do rol.

Algebricamente, para N par, a mediana é a média aritmética dos valores X_i que ocupam as posições i do rol dadas por:

$$i' = \frac{N}{2} \text{ e } i'' = \frac{N+2}{2}$$

Logo, temos que:

$$\text{Me}_{par} = \frac{X_{i'} + X_{i''}}{2}$$

Exemplo de aplicação

Um médico endocrinologista está avaliando os efeitos de uma dieta alimentar, associada a exercícios físicos, em 10 pacientes que se submeterão ao programa oferecido pela clínica. Um dos resultados esperados é a diminuição do nível de colesterol total (em mg/dL) nesses pacientes. Para analisar esse resultado, o médico considera os exames de colesterol total de todos eles, utilizando esses dados para identificar padrões de resposta ao tratamento. Os dados coletados são mostrados a seguir:

200 180 150 200 190 141 160 170 174 165

Nesse cenário, qual é o valor da mediana do nível de colesterol dos pacientes submetidos ao programa?

Resolução

Nesse contexto, temos $N = 10$ (par).

O rol dos valores de colesterol é mostrado a seguir:

Rol = {141, 150, 160, 165, 170, 174, 180, 190, 200, 200}

Os posicionamentos dos valores centrais são calculados a seguir:

$$i' = \frac{N}{2} = \frac{10}{2} = 5$$

$$i'' = \frac{N+2}{2} = \frac{10+2}{2} = 6$$

Logo, os elementos $X_5 = 170$ e $X_6 = 174$ são os valores centrais. Com isso, calculamos a média aritmética entre eles para encontrar a mediana.

$$Me_{par} = \frac{X_{i'} + X_{i''}}{2} = \frac{170 + 174}{2} = 172$$

Portanto, a mediana do nível de colesterol para esse conjunto de pacientes é de 172mg/dL.

Observação

Não precisamos, necessariamente, calcular a posição dos elementos centrais. Se conseguirmos identificá-los visualmente, uma tarefa fácil em conjuntos pequenos de dados, podemos encontrar o valor de mediana diretamente.

Saiba mais

É possível calcular a mediana a partir de uma distribuição de frequências, assim como fizemos com a média. Nesse caso, utilizamos dados da frequência acumulada absoluta (F_i). Para saber sobre esse procedimento, consulte o capítulo 6 do livro:

CRESPO, A. A. *Estatística fácil*. Rio de Janeiro: Saraiva Uni, 2009.

8.1.4 Moda

Vamos, agora, conhecer mais uma das medidas básicas de tendência central. Por sua definição tradicional, a moda (Mo) de uma série de dados coincide com o valor que se repete com maior frequência nesse conjunto. Dessa maneira, a moda busca o valor mais comum da série. Considere o conjunto de valores a seguir:

Valores = {3, 4, 10, 8, 9, 8, 10, 8, 14, 13}

Apenas para facilitar a leitura, vamos organizar esses valores em rol:

Rol = {3, 4, 8, 8, 8, 9, 10, 10, 13, 14}

Podemos notar que o valor que mais se repetiu foi o 8, pois ele apareceu 3 vezes no conjunto de dados. Nesse caso, a moda do conjunto será expressa como Mo=8.

A moda nem sempre existe e nem sempre é única. Devido à sua instabilidade, dentre as medidas de tendência central que vimos, costuma ser a menos utilizada.

No entanto, a moda pode funcionar como uma medida imediata de tendência central, já que, muitas vezes, é identificada apenas em uma breve observação dos valores do conjunto, principalmente em conjuntos pequenos.

Além disso, ela pode ser útil na identificação de estratos (subgrupos específicos) dentro de um conjunto de dados. Por exemplo, se em uma análise de alturas de estudantes de uma escola, encontramos dois picos na distribuição – um em torno de 1,50 m e outro em torno de 1,80 m – isso pode indicar que há duas modas ou estratos no conjunto de dados. Nesse caso, talvez haja um grupo de crianças mais novas e outro de adolescentes.

De acordo com Sharpe, Veaux e Velleman (2011), para dados quantitativos, especialmente os contínuos, faz mais sentido usar a palavra moda no sentido de pico do histograma do que como um único valor que mais se repete. Desse modo, se um histograma ou um gráfico de colunas apresenta dois picos aparentes, a distribuição é considerada bimodal, ou seja, apresenta duas modas.

Exemplo de aplicação

1) Considere cada uma das séries de dados a seguir, já organizadas em rol. Para cada uma delas, identifique os valores da moda, de acordo com a definição tradicional do termo.

A) Rol 1 = {1, 1, 3, 5, 7, 7, 7, 11, 13}

B) Rol 2 = {3, 5, 8, 11, 13, 18}

C) Rol 3 = {3, 5, 5, 5, 6, 6, 7, 7, 7, 11, 12}

Resolução

A) Rol 1 = {1, 1, 3, 5, 7, 7, 7, 11, 13} → Mo = 7

O valor 7 se repetiu três vezes, sendo essa a maior frequência da série. Portanto, esse é o valor da moda.

B) Rol 2 = {3, 5, 8, 11, 13, 18} → O conjunto não tem moda.

Nessa série de valores, cada numeral aparece somente uma vez, o que indica ausência de moda no conjunto.

C) Rol 3 = {3, 4, 5, 5, 5, 6, 6, 7, 7, 7, 9, 11, 12} → Mo' = 5 e Mo'' = 7

Nesse caso, temos um conjunto de dados bimodal, ou seja, que apresenta duas modas distintas. Isso indica que existem dois valores que ocorrem com maior frequência no conjunto de dados, o que pode sugerir a presença de duas tendências ou grupos distintos dentro do conjunto.

2) Uma loja de eletrônicos deseja analisar a quantidade de smartphones vendidos durante o último mês para otimizar o estoque. Para isso, o gerente coletou os dados sobre o número de dispositivos vendidos por dia. Os resultados são mostrados na tabela a seguir.

Tabela 37 – Tabela de frequências

Número de smartphones vendidos por dia	F_i
8	1
9	2
10	4
11	5
12	7
13	5
14	4
15	2

Qual é a moda de vendas de smartphone dessa loja?

Resolução

Vamos, primeiro, entender os dados dispostos. Somando as frequências, concluímos que N=30, ou seja, a coleta de dados foi feita por 30 dias. Se olharmos para a 1ª classe, podemos inferir que, durante a coleta, em apenas 1 dia (dentre os 30) foram vendidos 8 smartphones. Em 3 dias, foram observados a venda de 9 smartphones, e assim por diante.

Para achar a moda, devemos buscar o valor que se repete com mais frequência. Como, nesse caso, já temos o valor de frequências disposto na coluna de f_i, essa tarefa é simples. Note que a maior frequência da tabela corresponde ao valor 7, que ocorreu na 5ª classe da tabela, correspondente à classe de 12 smartphones vendidos por dia. Logo, em 7 dos 30 dias o número de vendas diárias foi de 12 smartphones. Logo, a moda de vendas dessa loja é de 12 smartphones por dia, ou seja, Mo=12.

É notável, também, que essa distribuição não pode ser considerada bimodal, pois, se montarmos o gráfico de colunas referente aos dados, não há dois picos aparentes, como podemos ver na figura a seguir. Observe que o pico ocorre justamente na classe cujo ponto médio é X_i=12, conforme esperado.

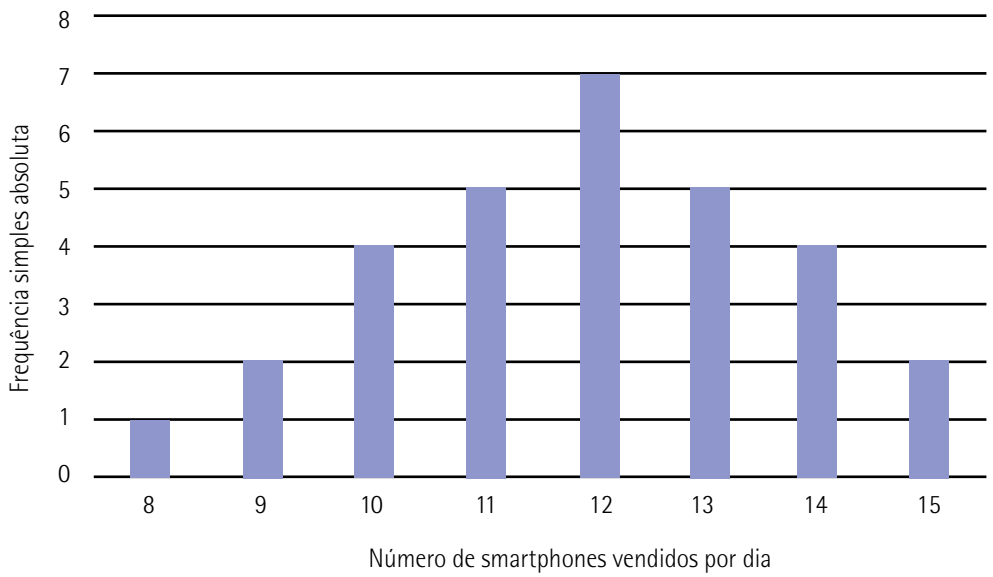

Figura 61 – Gráfico de colunas representando a distribuição dos dados do enunciado

8.2 Medidas de dispersão

Depois de termos discutido as medidas de tendência central, daremos sequência ao nosso estudo conhecendo as medidas de dispersão. Uma medida de dispersão é um valor que descreve a variação ou o espalhamento dos dados em torno de uma medida de tendência central, como a média aritmética.

Considere os três conjuntos de dados a seguir:

Conjunto 1 = {8, 8, 8}

Conjunto 2 = {7, 8, 9}

Conjunto 3 = {6, 8, 10}

Perceba que o conjunto 1 é formado por três elementos iguais, que o conjunto 2 é formado por três elementos distintos, mas próximos, e que o conjunto 3 é formado por três elementos distintos e um pouco mais distantes entre si do que os do conjunto 2. Se calcularmos a média aritmética (μ) de cada uma dessas séries de dados, chegaremos ao valor 8 para cada uma delas, conforme mostrado a seguir:

Conjunto 1 = {8, 8, 8} $\rightarrow \mu = 8$

Conjunto 2 = {7, 8, 9} $\rightarrow \mu = 8$

Conjunto 3 = {6, 8, 10} $\rightarrow \mu = 8$

A média, como medida de centro, nos traz um único valor que pretende descrever todo um conjunto de dados numéricos. No entanto, ela não traz qualquer informação a respeito do quão diversos eram os dados dos quais partimos. Para termos uma ideia do quão dispersos são os valores que compõem nosso conjunto de dados, podemos trabalhar com medidas de dispersão.

8.2.1 Desvio-padrão

A principal medida de dispersão é chamada de desvio-padrão. O desvio-padrão de um conjunto de dados é um valor que indica se a variabilidade de um conjunto de valores é grande ou pequena em relação à média aritmética dos valores. Assim, vejamos as situações a seguir:

- Se o desvio-padrão de um conjunto de dados é zero, significa que esse conjunto é formado por valores idênticos.

- Se o desvio-padrão de um conjunto de dados é pequeno quando comparado à média, significa que esse conjunto é formado por valores pouco dispersos, que variam pouco.

- Se o desvio-padrão de um conjunto de dados é grande quando comparado à média, significa que esse conjunto é formado por valores muito dispersos, que variam muito.

Se estamos tratando de dados de toda uma população, o desvio-padrão é geralmente expresso pela letra grega σ. Seu formato de cálculo é dado conforme exposto a seguir, em que N é o número de elementos da população, X é a variável que representa os elementos da característica quantitativa estudada, i é o índice de contagem dos elementos (que vai de 1 até N) e μ é a média aritmética populacional.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}}$$

As etapas de cálculo, indicadas na própria fórmula, são as seguintes:

- Calcule a média aritmética populacional.
- Calcule a diferença entre cada elemento X_i e a média μ.
- Eleve cada uma dessas diferenças ao quadrado. Isso fará com que a distância entre X_i e μ se torne um valor sempre positivo, independentemente de X_i ser maior ou menor do que μ.
- Faça o somatório dos valores elevados ao quadrado.
- Divida o resultado por N. Até essa etapa, temos uma medida de dispersão chamada de variância.
- Extraia a raiz quadrada. O desvio-padrão, portanto, é a raiz quadrada da variância. Como elevamos, anteriormente, as diferenças ao quadrado, extrair a raiz nos permitirá voltar à mesma ordem de grandeza das diferenças, e também fará com que a unidade de medida do desvio-padrão seja a mesma dos dados dos quais partimos.

Lembra dos nossos três conjuntos de dados? Vamos calcular o desvio-padrão para cada um deles a seguir:

Conjunto 1 = {8, 8, 8} $\rightarrow \mu = 8$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}} = \sqrt{\frac{(8-8)^2 + (8-8)^2 + (8-8)^2}{3}} = \sqrt{\frac{0^2 + 0^2 + 0^2}{3}} = \sqrt{\frac{0}{3}} = \sqrt{0} = 0$$

Conjunto 2 = {7, 8, 9} $\rightarrow \mu = 8$

$$\sigma = \sqrt{\frac{(7-8)^2 + (8-8)^2 + (9-8)^2}{3}} = \sqrt{\frac{(-1)^2 + 0^2 + 1^2}{3}} = \sqrt{\frac{1+0+1}{3}} = \sqrt{\frac{2}{3}} = 0{,}82$$

Conjunto 3 = {6, 8, 10} $\rightarrow \mu = 8$

$$\sigma = \sqrt{\frac{(6-8)^2 + (8-8)^2 + (10-8)^2}{3}} = \sqrt{\frac{(-2)^2 + 0^2 + 2^2}{3}} = \sqrt{\frac{4+0+4}{3}} = \sqrt{\frac{8}{3}} = 1{,}63$$

Em resumo, temos os dados expostos a seguir:

Conjunto 1 = {8, 8, 8} $\rightarrow \mu = 8 \rightarrow \sigma = 0$

Conjunto 2 = {7, 8, 9} $\rightarrow \mu = 8 \rightarrow \sigma \approx 0{,}82$

Conjunto 3 = {6, 8, 10} $\rightarrow \mu = 8 \rightarrow \sigma \approx 1{,}63$

Intuitivamente, já sabíamos que o conjunto 1 deveria apresentar variabilidade nula, que o conjunto 2 deveria apresentar variabilidade maior do que zero e que o conjunto 3 deveria apresentar variabilidade maior do que a do conjunto 2. Agora, os valores calculados de desvio-padrão nos mostram esses resultados, de maneira quantitativa.

Lembrete

A unidade de medida da variância é a unidade dos dados dos quais partimos elevada ao quadrado. O desvio-padrão representa a raiz quadrada da variância. Assim, a unidade de medida do desvio-padrão é a mesma unidade dos dados do conjunto.

Observação

Um dos critérios de arredondamento mais simples que existem é o que estamos utilizando neste conteúdo.

A última casa decimal que queremos preservar no resultado tem seu algarismo mantido caso o próximo algarismo seja 0, 1, 2, 3 ou 4. A última casa decimal que queremos preservar no resultado tem seu algarismo aumentado em uma unidade caso o próximo algarismo seja 5, 6, 7, 8 ou 9.

Por exemplo: 62,6492, arredondado para duas casas decimais, fica 62,65, já que o próximo algarismo é o 9. Já o número 3,45378, arredondado para duas casas decimais, fica 3,45, já que o próximo algarismo é o 3.

Talvez você tenha achado os cálculos que fizemos anteriormente um pouco confusos, devido à complexidade da fórmula do desvio-padrão. Vamos acompanhar um exemplo contextualizado, no qual utilizaremos uma tabela para nos auxiliar nas etapas de cálculo, o que tornará o raciocínio mais claro, bem como, detalhar todas as etapas. Em seguida, resolveremos o mesmo exemplo simplesmente aplicando a fórmula do desvio-padrão, como fizemos nos exemplos iniciais.

Exemplo de aplicação

Leila realizou, ao longo do ano, quatro provas de Ciências, cujas notas são apresentadas na tabela a seguir.

Tabela 38 – Notas da Leila

Estudante	Notas em Ciências			
Leila	1,0	3,0	9,0	3,0

Calcule o desvio-padrão das notas da Leila em Ciências.

Resolução 1

Primeiramente, devemos encontrar a média aritmética do conjunto de dados. Sabemos que $N = 4$ e que cada elemento X_i está disposto na tabela.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + X_3 + X_4}{4} = \frac{1,0 + 3,0 + 9,0 + 3,0}{4} = \frac{16}{4} = 4,0$$

Já sabemos que a média aritmética das notas de Leila vale 4,0. Portanto, o próximo passo é acharmos a diferença entre cada uma das notas X_i e o valor médio $\mu = 4,0$. Para isso, vamos utilizar a tabela a seguir, em que a diferença é indicada como $X_i - \mu$.

Tabela 39 – Primeira tabela de apoio

X_i	$X_i - \mu$
$X_1 = 1,0$	$X_1 - \mu = 1 - 4 = -3$
$X_2 = 3,0$	$X_2 - \mu = 3 - 4 = -1$
$X_3 = 9,0$	$X_3 - \mu = 9 - 4 = 5$
$X_4 = 3,0$	$X_4 - \mu = 3 - 4 = -1$

Repare que encontramos resultados negativos quando os valores X_i eram menores do que o valor médio, assim como encontramos resultado positivo para X_3, que era maior do que o valor médio.

A próxima etapa é elevarmos o resultado de cada linha ao quadrado. Para isso, criaremos uma nova coluna, indicada como $(X_i - \mu)^2$. Vamos omitir as etapas de cálculos anteriores, mantendo na tabela apenas os resultados.

Tabela 40 – Segunda tabela de apoio

X_i	$X_i - \mu$	$(X_i - \mu)^2$
$X_1 = 1,0$	-3	$(X_1 - \mu)^2 = (-3)^2 = 9$
$X_2 = 3,0$	-1	$(X_2 - \mu)^2 = (-1)^2 = 1$
$X_3 = 9,0$	5	$(X_3 - \mu)^2 = 5^2 = 25$
$X_4 = 3,0$	-1	$(X_4 - \mu)^2 = (-1)^2 = 1$

Perceba que, independentemente de termos partido de uma diferença positiva ou de uma diferença negativa, os valores que encontramos nessa etapa foram todos positivos. Elevar ao quadrado cada uma das diferenças tem esse efeito. Agora, não importa mais se os valores X_i eram maiores ou menores do que μ. Estamos interessados apenas em quão distantes eles estão do valor médio.

A próxima etapa será realizar o somatório dos valores da 3ª coluna, que acabamos de calcular. Com isso, já conseguimos tudo o que o numerador da fórmula do desvio-padrão populacional nos pede. Vamos adicionar uma linha para representar esse somatório.

Tabela 41 – Terceira tabela de apoio

X_i	$X_i - \mu$	$(X_i - \mu)^2$
$X_1 = 1,0$	-3	9
$X_2 = 3,0$	-1	1
$X_3 = 9,0$	5	25
$X_4 = 3,0$	-1	1
		$\sum$=36

Portanto, sabemos que $\sum_{i=1}^{N}(X_i-\mu)^2=36$. Agora, basta dividirmos esse somatório por $N=4$, pois assim, vamos dividir o somatório dos quadrados das distâncias entre o número de elementos do conjunto de dados. O resultado dessa divisão é a variância dos dados. Em seguida, vamos extrair a raiz quadrada da variância, para obtermos o desvio-padrão. Como elevamos as diferenças ao quadrado para nos livrarmos dos sinais negativos, extrair a raiz nos levará novamente a um valor próximo à média das distâncias. Pense assim: se elevarmos 2 ao quadrado, chegaremos a 4. Se extrairmos a raiz de 4, voltaremos ao 2. É algo parecido ao que estamos fazendo aqui. Vamos ao cálculo:

$$\sigma=\sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}=\sqrt{\frac{36}{4}}=\sqrt{9}=3$$

Finalmente, sabemos que Leila tirou média 4,0 nas provas de Ciências, com desvio-padrão de 3,0 pontos. De maneira simples, esse resultado indica que a maior parte dos valores que compõem nossa população estão a uma distância de até 3,0 pontos do valor da média, seja para mais ou para menos.

Resolução 2

O método da tabela é interessante para entendermos as etapas de cálculo e quando estamos lidando com um conjunto de dados relativamente grande. Como nosso conjunto era composto por apenas quatro elementos, podemos facilmente aplicar os cálculos diretamente na fórmula do desvio-padrão. Desse modo, no numerador de nossa fração, cada parcela corresponderá ao cálculo de um elemento X_i.

Já sabemos que a média $\mu=4,0$ e que $N=4$. Com isso, vamos substituir esses valores diretamente na fórmula, conforme exposto a seguir.

$$\sigma=\sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}=\sqrt{\frac{(X_1-\mu)^2+(X_2-\mu)^2+(X_3-\mu)^2+(X_4-\mu)^2}{N}}$$

$$\sigma=\sqrt{\frac{(1-4)^2+(3-4)^2+(9-4)^2+(3-4)^2}{4}}$$

$$\sigma=\sqrt{\frac{(-3)^2+(-1)^2+5^2+(-1)^2}{4}}$$

$$\sigma = \sqrt{\frac{9+1+25+1}{4}} = \sqrt{\frac{36}{4}} = \sqrt{9} = 3$$

Observação

Você deve imaginar que, na maior parte das vezes, precisamos extrair raízes quadradas de valores não exatos nos cálculos de desvio-padrão. Para isso, pode-se contar com o auxílio de uma calculadora que realize essa operação. Vários modelos de calculadoras simples, como a Casio HL-4A, possuem essa função. Você também encontrará essa operação em qualquer calculadora científica, como a HP 10s+.

Saiba mais

Em um vídeo no YouTube, o professor Rafael Procopio ensina uma dica para encontrarmos valores aproximados de raízes quadradas não exatas, utilizando apenas operações de adição, multiplicação e divisão.

RAIZ quadrada não exata. 2018. 1 vídeo (4 min). Publicado pelo canal Matemática Rio com Prof. Rafael Procopio. Disponível em: https://tinyurl.com/4urv55nn. Acesso em: 10 dez. 2024.

E se tivermos acesso a uma distribuição de frequências, mas não ao conjunto de dados brutos? Nesse caso, podemos calcular o desvio-padrão σ_f utilizando os pontos médios de classe (x_i) e as frequências simples absolutas (f_i). Nessa situação, usamos um formato algébrico parecido com o do desvio-padrão para dados brutos, com uma modificação fundamental: o quadrado da diferença entre x_i e a média μ_f é multiplicado por f_i. Lembre-se que N representa o número de elemento do conjunto e que k representa o número de classes. Acompanhe:

$$\sigma_f = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \mu_f)^2 \cdot f_i}{N}}$$

Utilizaremos essa fórmula na seção seguinte do nosso livro-texto, depois que aprendermos o conceito de coeficiente de variação.

8.2.2 Coeficiente de variação

Agora que já conhecemos o desvio-padrão, podemos estabelecer uma taxa dessa medida em relação à média. Essa taxa constitui uma nova medida de dispersão, que estudaremos a seguir.

Dois conjuntos de dados podem ter médias de magnitudes muito distintas. Nesse caso, o desvio-padrão desses dois conjuntos não é comparável. Observe os conjuntos de dados a seguir. Qual deles você considera mais homogêneo, ou seja, qual tem valores mais parecidos entre si?

Conjunto 1 = {4, 5, 6, 7, 8}

Conjunto 2 = {40, 50, 60, 70, 80}

Vamos analisar a situação. Se calcularmos a média aritmética e o desvio-padrão desses conjuntos, chegaremos aos resultados a seguir:

Conjunto 1 = {4, 5, 6, 7, 8} $\rightarrow \mu = 6 \rightarrow \sigma = \sqrt{2} \approx 1{,}41$

Conjunto 2 = {40, 50, 60, 70, 80} $\rightarrow \mu = 60 \rightarrow \sigma = \sqrt{200} \approx 14{,}14$

O desvio-padrão do conjunto 1 é muito menor do que o desvio-padrão do conjunto 2, o que poderia nos indicar que o conjunto 1 é mais homogêneo. Porém, precisamos notar que a média aritmética do conjunto 1 é 10 vezes menor do que a do conjunto 2. Como as médias têm valores muito distantes, a comparação entre os dois conjuntos utilizando apenas o valor do desvio-padrão não é confiável.

A solução, nesse caso, é usar o coeficiente de variação, que representa uma taxa entre o desvio-padrão e a média. Algebricamente, temos os possíveis formatos a seguir, nos quais CV nos entrega uma taxa unitária e $CV_{\%}$ nos entrega uma taxa percentual.

$$CV = \frac{\sigma}{\mu} \text{ ou } CV_{\%} = \frac{\sigma}{\mu} \cdot 100$$

Se considerarmos o formato percentual, o coeficiente de variação nos indica qual porcentagem o desvio-padrão de um conjunto de dados representa em relação à sua média aritmética. Esse valor, sim, pode ser comparado ao de outro conjunto de dados, independentemente da magnitude de suas médias aritméticas. Quanto menor for o valor do coeficiente de variação, mais homogêneo é o conjunto de dados.

Vamos, agora, resolver alguns exemplos de aplicação do conteúdo estudado.

Exemplo de aplicação

1) Considere dois grupos de alunos que estudam em duas filiais distintas, A e B, da mesma escola. Para cada grupo, foi calculada a média e o desvio-padrão das notas de um exame preparatório para o vestibular, conforme os dados da tabela a seguir.

Tabela 42 – Notas do exame

Filial	Nota média	Desvio-padrão
A	40	4
B	20	4

Calcule o coeficiente de variação para os grupos de alunos de ambas as filiais e indique qual grupo apresenta menor variabilidade relativa, ou seja, qual grupo é mais homogêneo.

Resolução

Vamos calcular os coeficientes de variação em taxas percentuais, já que é mais confortável interpretarmos os resultados nesse formato. Os cálculos são apresentados a seguir:

$$\text{Filial A: } CV_{\%A} = \frac{\sigma_A}{\mu_A} \cdot 100 = \frac{4}{40} \cdot 100 = 10\%$$

$$\text{Filial B: } CV_{\%B} = \frac{\sigma_B}{\mu_B} \cdot 100 = \frac{4}{20} \cdot 100 = 20\%$$

O coeficiente de variação do grupo da filial A apresentou menor valor (10%) do que o da filial B (20%). Isso indica que o grupo da filial A é mais homogêneo, ou seja, os alunos tiveram um desempenho mais parecido entre si.

2) A tabela de frequências a seguir apresenta a distribuição das idades dos 60 jovens participantes de um projeto social.

Tabela 43 – Distribuição das idades dos jovens

Idades (anos)	X_i	F_i
14⊢16	15	10
16⊢18	17	20
18⊢20	19	25
20⊢22	21	5

Qual é o coeficiente de variação desses dados?

Resolução

Sabemos que temos 60 jovens como elementos do estudo, portanto, N=60, valor que coincide com o somatório das frequências. Como a tabela foi dividida em 4 classes, temos k=4. Os pontos médios de classe, x_i, já foram posicionados na própria tabela do enunciado, então, não precisamos calculá-los.

Nosso primeiro cálculo será encontrar a média aritmética da tabela de frequências, μ_f. Sabemos que:

$$\mu_f = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N}$$

Para nos auxiliar, vamos posicionar uma coluna nova na tabela, $x_i \cdot f_i$, e realizar o seu somatório.

Tabela 44 – Primeira tabela auxiliar

Idades (anos)	X_i	F_i	$X_i \cdot F_i$
14⊢16	15	10	150
16⊢18	17	20	340
18⊢20	19	25	475
20⊢22	21	5	105
		$\sum f_i = N = 60$	$\sum x_i \cdot f_i = 1070$

Agora, vamos ao cálculo da média.

$$\mu_f = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{N} = \frac{1070}{60} \approx 17{,}8333$$

Logo, a média de idade dos jovens é de 17,8 anos. Como os cálculos sequentes dependem do valor da média, usaremos 4 casas decimais nos cálculos, para não prejudicar muito a precisão dos próximos resultados.

Nossa próxima etapa é calcular o desvio-padrão σ_f. Para isso, usaremos o formato a seguir:

$$\sigma_f = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \mu_f)^2 \cdot f_i}{N}}$$

Note que, no numerador, teremos um somatório de 4 parcelas das diferenças entre o ponto médio de classe x_i e a média aritmética μ_f elevadas ao quadrado. Em sequência, esses quadrados devem ser multiplicados pelas frequências f_i. Essas etapas são feitas na tabela a seguir. Para reproduzir os resultados, utilize uma calculadora para ajudar.

Tabela 45 – Segunda tabela auxiliar

Idades (anos)	X_i	F_i	$X_i \cdot F_i$	$(X_i - \mu_i)^2 \cdot F_i$
14⊢16	15	10	150	$(15-17{,}8333)^2 \cdot 10 \approx 80{,}2759$
16⊢18	17	20	340	$(17-17{,}8333)^2 \cdot 20 \approx 13{,}8878$
18⊢20	19	25	475	$(19-17{,}8333)^2 \cdot 25 \approx 34{,}0297$
20⊢22	21	5	105	$(21-17{,}8333)^2 \cdot 5 \approx 50{,}1399$
		$\sum f_i = N = 60$	$\sum x_i \cdot f_i = 1070$	$\sum (x_i - \mu_i)^2 \cdot f_i = 178{,}3333$

A partir do resultado do somatório da última coluna da tabela, podemos finalmente chegar ao desvio-padrão dos dados.

$$\sigma_f = \sqrt{\frac{\sum_{i=1}^{k}(x_i - \mu_f)^2 \cdot f_i}{N}} = \sqrt{\frac{178{,}3333}{60}} \approx \sqrt{2{,}9722} \approx 1{,}7240$$

Portanto, a média de idade dos jovens é de 17,8 anos, com desvio esperado de 1,7 ano.

Agora, calcular o coeficiente de variação da distribuição, que chamaremos de $CV_{\%f}$, ficou fácil. Basta encontrarmos a razão entre σ_f e μ_f e multiplicar por 100, se desejarmos ter o resultado em taxa percentual.

$$CV_{\%f} = \frac{\sigma_f}{\mu_f} \cdot 100 = \frac{1{,}7240}{17{,}8333} \cdot 100 \approx 9{,}67\%$$

Desse modo, o coeficiente de variação das idades dos jovens que participam do projeto social é de 9,67%.

Resumo

A unidade IV do nosso livro-texto foi dedicada ao estudo introdutório da estatística, que é a ciência que busca extrair informações dos dados. O foco do nosso estudo foi a área descritiva da estatística. Iniciamos o conteúdo aprendendo os conceitos de população e amostra, vimos as áreas nas quais a estatística pode ser subdividida e aprendemos a forma mais simples de organização de dados, o rol.

Depois, continuamos nossos estudos falando sobre as distribuições de frequência, na qual aprendemos a calcular os quatro tipos de frequência: simples absoluta, simples relativa, acumulada absoluta e acumulada relativa. Na sequência, interpretamos histogramas e vimos alguns outros tipos de gráficos utilizados para mostrar visualmente conjuntos de dados.

Em seguida, exibimos os conceitos e os formatos algébricos de algumas das principais medidas de tendência central: média aritmética, mediana e moda. As medidas de tendência central procuram descrever um conjunto de dados com um único valor numérico, que aponta para o centro do conjunto.

Por fim, estudamos as medidas de dispersão, que nos indicam quão diversos são os valores da nossa série de dados. Nessa parte, apresentamos a variância, o desvio-padrão e o coeficiente de variação. Aprendemos, também, a calcular a dispersão de dados distribuídos em frequências.

Exercícios

Questão 1. A empresa Software para Você fornece soluções computacionais para empresas que atuam em várias áreas do comércio. No gráfico a seguir, temos a distribuição do tempo, em horas, que os desenvolvedores dessa empresa levaram para responder às demandas dos 200 clientes atendidos no último mês.

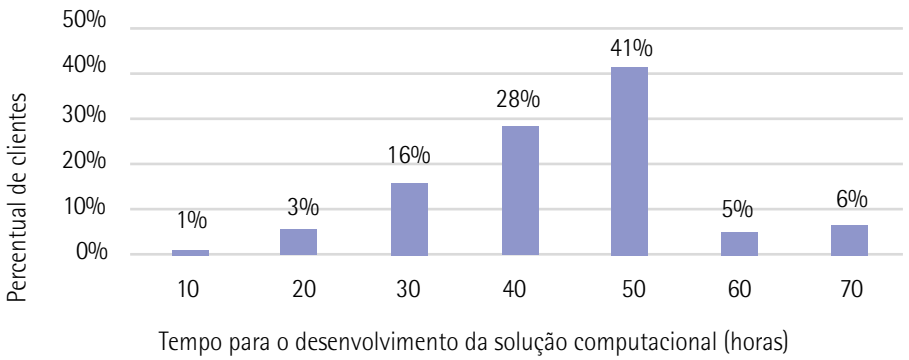

Figura 62 – Gráfico de distribuição de tempo, em horas

O tempo médio que os desenvolvedores da Software para Você levaram para responder às demandas dos 200 clientes atendidos no último mês é igual a:

A) 44,4 horas.

B) 35,0 horas.

C) 50,5 horas.

D) 37,8 horas.

E) 50,5 horas.

Resposta correta: alternativa A.

Resolução

A quantidade total de clientes atendidos no último mês foi igual a 200.

Pela leitura do gráfico do enunciado, podemos concluir que, dos 200 clientes:

- 2 deles (1% de 200) utilizaram 10 horas para a resposta às demandas;
- 6 deles (3% de 200) utilizaram 20 horas para a resposta às demandas;

- 32 deles (16% de 200) utilizaram 30 horas para a resposta às demandas;
- 56 deles (28% de 200) utilizaram 40 horas para a resposta às demandas;
- 82 deles (41% de 200) utilizaram 50 horas para a resposta às demandas;
- 10 deles (5% de 200) utilizaram 60 horas para a resposta às demandas;
- 12 deles (6% de 200) utilizaram 70 horas para a resposta às demandas.

Os cálculos feitos podem ser resumidos na tabela a seguir:

Tabela 46 – Tempo médio de atendimento para cada cliente

Tempo (horas)	Frequência de clientes
10	2
20	6
30	32
40	56
50	82
60	10
70	12
Total	200 clientes

Na tabela, temos o total de 200 clientes. Para calcularmos o tempo médio, precisamos somar todos os 200 valores e dividir essa soma por 100. Observe que, com base na tabela:

- 10 (horas) é um valor que precisa ser somado 2 vezes;
- 20 (horas) é um valor que precisa ser somado 6 vezes;
- 30 (horas) é um valor que precisa ser somado 32 vezes;
- 40 (horas) é um valor que precisa ser somado 56 vezes;
- 50 (horas) é um valor que precisa ser somado 82 vezes;
- 60 (horas) é um valor que precisa ser somado 10 vezes;
- 70 (horas) é um valor que precisa ser somado 12 vezes.

Logo, o tempo médio é igual a 44,4 horas, conforme calculado a seguir:

$$\text{Tempo médio} = \frac{10.2 + 20.6 + 30.32 + 40.56 + 50.82 + 60.10 + 70.12}{200}$$

$$\text{Tempo médio} = \frac{20 + 120 + 960 + 2240 + 4100 + 600 + 840}{200} = \frac{8880}{200} = 44,4$$

$$\text{Tempo médio} = 44,4$$

Questão 2. O responsável pela ouvidoria da empresa ABC fez um levantamento sobre o número de reclamações recebidas pelos funcionários do setor no mês corrente e resumiu as informações obtidas na tabela a seguir.

Tabela 47 – Levantamento feito pelo responsável pela ouvidoria da empresa ABC

Nome do funcionário	Número de reclamações recebidas
Ana	3
Bianca	2
Beatriz	3
Catarina	2
Diego	1
Elsa	5
Fábio	1
Gabriela	2
Júlia	3
Laila	2
Marcelo	0
Mariana	1
Patrícia	2
Paulo	2
Rafael	3
Sofia	2
Tobias	2

Com base na tabela e nos seus conhecimentos, assinale a alternativa que indica correta e respectivamente a moda, a média e a mediana do levantamento apresentado.

A) 2; 2; 2

B) 2; 2,12; 2

C) 5; 2,12; 2,5

D) 3; 2; 5

E) 5; 2,12; 2

Resposta correta: alternativa B.

Análise da questão

Vamos começar nossa análise respondendo às perguntas a seguir.

- Há funcionários que não receberam reclamações no mês corrente? Sim, apenas um funcionário, Marcelo.
- Há funcionários que receberam uma reclamação no mês corrente? Sim, 3 funcionários, Diego, Fábio e Mariana.
- Há funcionários que receberam duas reclamações no mês corrente? Sim, 8 funcionários, Bianca, Catarina, Gabriela, Laila, Patrícia, Paulo, Sofia e Tobias.
- Há funcionários que receberam três reclamações no mês corrente? Sim, 4 funcionários, Ana, Beatriz, Júlia e Rafael.
- Há funcionários que receberam quatro reclamações no mês corrente? Não, nenhum ("0 funcionário").
- Há funcionários que receberam cinco reclamações no mês corrente? Sim, uma funcionária, Elsa.

Com essas respostas, podemos elaborar a tabela a seguir, que mostra as quantidades de funcionários que receberam 0, 1, 2, 3, 4 ou 5 reclamações no mês corrente. Além disso, adicionamos os nomes dos funcionários.

Tabela 48 – Quantidades de reclamações recebidas e quantidades de funcionários

Quantidade de reclamações	Quantidade de funcionários	Nomes dos funcionários
0	1	Marcelo
1	3	Diego, Fábio e Mariana
2	8	Bianca, Catarina, Gabriela, Laila, Patrícia, Paulo, Sofia e Tobias
3	4	Ana, Beatriz, Júlia e Rafael
4	0	-
5	1	Elsa
Total	1+3+8+4+0+1=17	-

Pela tabela anterior, vemos, por exemplo, que, dos 17 funcionários, 3 receberam uma reclamação e nenhum recebeu 4 reclamações.

Vamos chamar de frequência absoluta de cada medida, indicada por FA, a quantidade de funcionários que recebeu dado número de reclamações, indicado por x. Vejamos:

- a FA de 0 reclamações é igual a 1 (se x=0, FA=1);
- a FA de 1 reclamação é igual a 3 (se x=1, FA=3);
- a FA de 2 reclamações é igual a 8 (se x=2, FA=8);
- a FA de 3 reclamações é igual a 4 (se x=3, FA=4);
- a FA de 4 reclamações é igual a 0 (se x=4, FA=0);
- a FA de 5 reclamações é igual a 1 (se x=5, FA=1).

Podemos calcular a frequência relativa, indicada por FR, de cada quantidade de reclamações recebidas pelos funcionários. Para isso, dividimos a frequência absoluta (FA) pelo número total N de funcionários, que é 17. Ou seja:

$$FR = \frac{FA}{N}$$

Na tabela a seguir, temos as frequências absolutas e relativas do caso em estudo.

Tabela 49 – Quantidade de reclamações (x), frequência absoluta (FA) e frequência relativa (FR)

Quantidade de reclamações (x)	Frequência absoluta (FA)	Frequência relativa (FR), sendo FR=FA/N
0	1	1/17 = 0,05882
1	3	3/17 = 0,17647
2	8	8/17 = 0,47059
3	4	4/17 = 0,23529
4	0	0/17 = 0
5	1	1/17 = 0,05882
Total	N=1+3+8+4+0+1=17	$\text{Soma} = \frac{1}{17} + \frac{3}{17} + \frac{8}{17} + \frac{4}{17} + \frac{0}{17} + \frac{1}{17} = 1$

Vale notar que, em qualquer conjunto de dados, a soma de todas as frequências relativas dá 1.

Podemos fazer um cálculo bastante semelhante ao feito para determinarmos a frequência relativa, multiplicando-a por 100%. Desse modo, obtemos os percentuais de cada quantidade de reclamações recebidas, indicada por P%. Ou seja:

$$P\% = FR.100$$

Na tabela a seguir, temos as frequências absolutas, as frequências relativas e os percentuais do caso em estudo.

Tabela 50 – Quantidade de reclamações, frequência absoluta, frequência relativa e percentual

Quantidade de reclamações (x)	Frequência absoluta (FA)	Frequência relativa (FR)	Percentual (P%), sendo P%=FR.100
0	1	0,05882	5,882%
1	3	0,17647	17,647%
2	8	0,47059	47,059%
3	4	0,23529	23,529%
4	0	0	0%
5	1	0,05882	5,882%
Soma	N=17	1	100%

Podemos, de certa forma, resumir o conjunto de dados em valores como moda, média e mediana, conhecidas como medidas de tendência central.

A observação do conjunto de dados que aparece mais vezes, ou seja, a de maior FA, é a moda do conjunto de dados. No caso em estudo, vemos, pela tabela anterior, que o valor que aparece mais vezes é 2 reclamações, com FA=8. Logo, a moda da quantidade de reclamações recebidas no mês corrente pelos funcionários da empresa ABC é 2.

A fim de acharmos a média, fazemos assim: somamos as quantidades multiplicadas pelas respectivas frequências e dividimos essa soma pelo total. Com base na tabela anterior, concluímos que a média do número de reclamações é 2,12, pois:

$$\text{Média} = \frac{0x1 + 1x3 + 2x8 + 3x4 + 4x0 + 5x1}{17} = \frac{36}{17}$$

$$\text{Média} = 2{,}12$$

Essa média de 2,12 é um valor teórico, pois não há número fracionário de reclamações. O valor 2,12 corresponde ao número de reclamações que cada funcionário teria recebido se todos os funcionários tivessem recebido o mesmo número de reclamações.

Para acharmos a mediana, ordenamos todas as observações e indicamos o valor central. Visto que há o total de 17 observações, a mediana é o valor central, que corresponde à nona observação, conforme indicado na tabela a seguir. Ou seja, no caso em estudo, a mediana da quantidade de reclamações recebidas no mês corrente pelos funcionários da empresa ABC é 2.

Tabela 51 – Quantidade (ordenada) de reclamações e quantidade de observações

Quantidade (ordenada) de reclamações	Quantidade de observações
0	8 observações
1	
1	
1	
2	
2	
2	
2	
2	Valor central (9ª observação): 2
2	8 observações
2	
2	
3	
3	
3	
3	
3	

Logo, no caso em estudo, a moda é 2, a média é 2,12 e a mediana é 2.

REFERÊNCIAS

Audiovisuais

RAIZ quadrada não exata. 2018. 1 vídeo (4 min). Publicado pelo canal Matemática Rio com Prof. Rafael Procopio. Disponível em: https://tinyurl.com/4urv55nn. Acesso em: 10 dez. 2024.

TRAJETÓRIA da bola. 2024. 1 vídeo (7 min). Publicado pelo canal Como Pode Cair no Enem. Disponível em: https://tinyurl.com/4s3c8fwm. Acesso em: 18 nov. 2024.

WHY IS 'x' the unknown. 2012. 1 vídeo (3 min). Publicado pelo canal Ted. Disponível em: https://tinyurl.com/mubcfhyc. Acesso em: 4 set. 2024.

Textuais

ALBERTO, J. *Matemática para concursos públicos e vestibulares*. São Paulo: Digerati Books, 2008.

CARVALHO, A. C. P. L. F de; MENEZES, A. G.; BONIDIA, R. P. *Ciência de dados*: fundamentos e aplicações. Rio de Janeiro: LTC, 2024.

CASTANHEIRA, N. P. *Noções básicas de matemática comercial e financeira*. Curitiba: IBPEX, 2008.

CRESPO, A. A. *Estatística fácil*. Rio de Janeiro: Saraiva Uni, 2009.

DANTE, L. R.; VIANA, F. *Matemática*: Contexto & aplicações. São Paulo: Ática, 2019.

DEMANA, F. D. *et al. Pré-cálculo*. São Paulo: Pearson Education do Brasil, 2013.

EMH. *Cofactor matrix calculator*. [s.d.]. Disponível em: https://tinyurl.com/ymuu45a5. Acesso em: 17 dez. 2024.

ESTRUTURAS de dados. *Python*. [s.d.]. Disponível em: https://tinyurl.com/4aydpkn6. Acesso em: 16 out. 2024.

FERNANDES, M. P. M. Matriz inversa: inversão por matriz adjunta. *InfoEscola*. [s.d.]. Disponível em: https://tinyurl.com/3k6rcmvj. Acesso em: 5 jul. 2024.

GOMES, F. M. *Pré-cálculo*: operações, equações, funções e trigonometria. São Paulo: Cengage Learning, 2019.

GOOGLE. *Google Analytics*. [s.d.]. Disponível em: https://tinyurl.com/ruppkmhh. Acesso em: 5 dez. 2024.

GUEDES, S. *Lógica de programação algorítmica*. São Paulo: Pearson Education do Brasil, 2014.

JACQUES, I. *Matemática para economia e administração*. São Paulo: Pearson Prentice Hall, 2010.

MARGUTI, A. L. *Minimanual de matemática*: Enem, vestibulares e concursos. São Paulo: Rideel, 2017.

MINAWI, F. Virtual oscilloscope. *PhysicsZone*, 2020. Disponível em: https://tinyurl.com/ypyrv8mx. Acesso em: 13 nov. 2024.

MINIWEBTOOL. *Calculadora de log*. [s.d.]. Disponível em: https://tinyurl.com/2jp4m9us. Acesso em: 11 nov. 2024.

OLIVEIRA, C. A. M. *Matemática*. Curitiba: InterSaberes, 2016.

RAMAKRISHNAN, R.; GEHRKE, J. *Sistemas de gerenciamento de banco de dados*. Porto Alegre: AMGH, 2011. E-book.

SHARPE, N. R.; VEAUX, R. D. D.; VELLEMAN, P. F. *Estatística aplicada*. Porto Alegre: Bookman, 2011.

TRIOLA, M. F. *Introdução à estatística*. Rio de Janeiro: LTC, 2023.

WAGNER, E. *Matemática 1*. Rio de Janeiro: FGV, 2011.

Informações:
www.sepi.unip.br ou 0800 010 9000