Sumário

Capítulo 1

Fundamentos

1.1 Introdução à estatística

a. Introdução: População e Amostra; Atributos e Variáveis; Séries Estatísticas. Distribuições a uma variável;

b. Tabulação de dados; Histogramas, polígonos de frequências e Ogivas;

c. Principais tipos de medidas;

d. Moda e Mediana;

e. Índices de dispersão, assimetria e curtose;

f. Momentos. Números Índices relativos;

g. Índices Simples e Ponderados;

h. Deflacionamento de Séries;

i. Índice Ideal de Fisher;

j. Índice de Bowloy

A origem da estatística remonta às civilizações antigas, em que vários povos coletavam e registravam dados populacionais e econômicos de interesse do Estado.

Nessa época, também eram realizadas estimativas das riquezas individuais e familiares, as quais eram utilizadas para determinar o montante de impostos a serem pagos pela população.

O termo estatística se originou da palavra *status*, que significa Estado em latim. O termo era utilizado para designar um conjunto de dados, relativos aos Estados, que os governantes utilizavam para controle fiscal e segurança nacional. O primeiro a utilizar a palavra foi Schneider, ainda no século XVII, em latim. Depois, foi adotada pelo acadêmico alemão Godofredo Achenwall.

A Estatística pode ser definida como a ciência que estuda os processos de coleta, organização, análise e interpretação de dados numéricos variáveis referentes a qualquer fenômeno. Ou ainda, podemos conceituá-la como um conjunto de técnicas de coleta, organização, análise e interpretação de dados, aplicáveis a várias áreas do conhecimento, que auxiliam no processo de tomada de decisão.

Os avanços computacionais tornaram a Estatística mais acessível e permitiram aplicações mais sofisticadas em diferentes áreas do conhecimento. Nesse cenário, os softwares estatísticos passaram a disponibilizar ferramentas antes inimagináveis, voltadas para planejamento de experimentos, teste de hipóteses, cálculos de confiabilidade, criação de gráficos complexos e elaboração de modelos

preditivos. A Estatística pode ser dividida em três grandes ramos: Estatística Descritiva (ou dedutiva), Estatística Probabilística e Estatística Inferencial (ou indutiva). Alguns autores, porém, consideram a Estatística Probabilística como parte da Estatística Inferencial.

A Estatística Descritiva (ou Dedutiva) é responsável pela coleta, organização, descrição e resumo dos dados observados. A partir de um determinado conjunto de dados, a Estatística Descritiva busca organizá-los em tabelas (ou gráficos) e estabelecer um sumário por meio de medidas descritivas como a média, os valores mínimo e máximo, o desvio padrão, entre outras. A Estatística Probabilística é responsável por estabelecer o modelo matemático probabilístico adotado para explicar os fenômenos aleatórios investigados pela Estatística. Os resultados desses fenômenos aleatórios podem variar de uma observação para outra, dificultando muito a previsão de um resultado futuro. Por isso, a Teoria da Probabilidade é usada para medir a chance de ocorrência de determinados eventos.

A Estatística Inferencial (ou Indutiva) é responsável pela análise e interpretação dos dados. A partir da análise de dados de uma amostra, a Estatística Indutiva estabelece inferências e previsões sobre a população, auxiliando na tomada decisões. Além disso, busca generalizar conclusões a respeito da população a partir de uma amostra, analisando a representatividade, a significância e a confiabilidade dos resultados obtidos.

1.1.1 CONCEITOS INICIAIS

Neste tópico, apresentaremos alguns conceitos iniciais da estatística que costumam ser abordados em provas de concursos públicos, dentre os quais podemos citar: população, amostra, censo, amostragem, parâmetros e estatísticas.

1.1.2 POPULAÇÃO

Uma POPULAÇÃO é um conjunto que contém TODOS OS INDIVÍDUOS, OBJETOS OU ELEMENTOS a serem estudados, que apresentam uma ou mais características em comum. A população pode ser finita, quando apresenta um número pequeno ou limitado de observações; ou infinita, quando apresenta um número muito grande ou ilimitado de observações.

1.1.3 AMOSTRA

Uma AMOSTRA é um SUBCONJUNTO EXTRAÍDO DA POPULAÇÃO para análise, devendo ser representativo daquele grupo. A partir das informações colhidas da amostra, os resultados obtidos podem ser utilizados para generalizar, inferir ou tirar conclusões acerca da população. Como exemplo, podemos citar as pesquisas eleitorais, em que uma amostra de eleitores deve ser extraída conforme a proporcionalidade de gênero, idade, grau de instrução e classe social.

1.1.4 CENSO

O CENSO, ou recenseamento, é um estudo dos dados relativos a TODOS os elementos de uma população. O censo pode custar muito caro e demandar um tempo considerável, de forma que um estudo considerando somente uma parcela da população pode ser uma alternativa mais simples, rápida e menos onerosa. Como exemplos, podemos citar a pesquisa sobre o grau de escolaridade dos habitantes brasileiros, o estudo sobre a renda dos brasileiros e a pesquisa de emprego.

1.1.5 AMOSTRAGEM

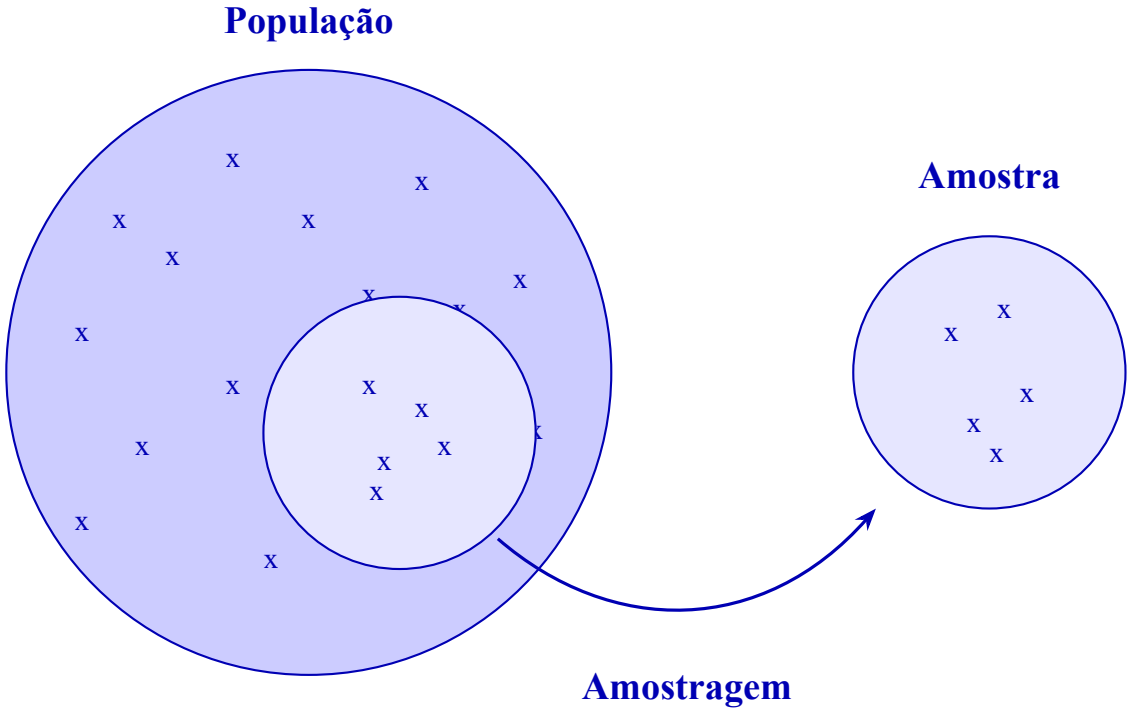
A AMOSTRAGEM é um processo que consiste na SELEÇÃO CRITERIOSA dos elementos a serem submetidos à investigação. Se forem cometidos erros no processo de seleção da amostra, muito provavelmente, o estudo ficará comprometido e os resultados serão tendenciosos. Portanto, devemos garantir que a amostra seja representativa da população. Isso significa que, com exceção de pequenas discrepâncias inerentes à aleatoriedade existente no processo de amostragem, uma amostra deve possuir as mesmas características básicas da população, no que diz respeito às variáveis que desejamos pesquisar.

1.1.6 PARÂMETROS

Os **PARÂMETROS** são **DESCRIÇÕES NUMÉRICAS** de **CARACTERÍSTICAS POPULACIONAIS** que raramente são conhecidas. Em geral, é muito caro ou demorado obter os dados da população inteira. Assim, algumas medidas precisam ser estimadas a partir de critérios ou métodos definidos pelo pesquisador, para representar características desconhecidas de uma população (por exemplo, a proporção de homens e mulheres na população brasileira). Normalmente, os parâmetros populacionais são constantes para uma população.

1.1.7 ESTATÍSTICA (OU ESTIMADOR)

As ESTATÍSTICAS são MEDIDAS NUMÉRICAS OBTIDAS DE AMOSTRAS representativas extraídas da população. A partir das informações colhidas da amostra, as estatísticas amostrais obtidas podem ser utilizadas para inferir ou tirar conclusões acerca dos parâmetros populacionais, como a proporção de homens e mulheres na população brasileira. Resumidamente, as estatísticas (ou estimadores) são descrições numéricas de características amostrais. Normalmente, as estatísticas amostrais diferem de uma amostra para outra.



1.2 NOÇÕES INICIAIS SOBRE ESTATÍSTICA

1.3 MÉTODO EXPERIMENTAL VS. MÉTODO ESTATÍSTICO

Para a investigação de um fenômeno, temos a nossa disposição dois métodos: experimental e estatístico. Resumidamente, o **MÉTODO EXPERIMENTAL** consiste em manter constantes as causas (fatores), com exceção de uma, variada para que seus efeitos sejam descobertos. Contudo, nem sempre poderemos aplicar o método experimental, pois os fatores que afetam um fenômeno podem não permanecer constantes enquanto variamos a causa que nos interessa. Por exemplo, para analisarmos uma queda nas vendas de uma empresa nacional que produz chocolates finos, teríamos que considerar vários fatores que não necessariamente permanecerão constantes durante toda a investigação do fenômeno, tais como a região, o fluxo de turistas na localidade; a temperatura média; o preço do concorrente; o mês de férias, etc.

Assim, diante da impossibilidade de manter as causas ou fatores constantes, o **MÉTODO ESTATÍSTICO** admite e registra todas as possíveis variações das causas presentes, procurando determinar a influência de cada fator no resultado. Dessa forma, o método estatístico descobrirá relações entre os fatores, como, por exemplo, a influência da temperatura média e do fluxo de turistas na venda de chocolates finos.

1.4 DADOS ESTATÍSTICOS

Os dados estatísticos constituem os valores resultantes da coleta de dados. Os dados referem-se a um conjunto de valores, os quais são organizados por meio de variáveis (a característica está sendo medida) e observações (elementos da amostra/população). É o caso, por exemplo, dos valores obtidos na pesquisa de peso, altura, idade e sexo de uma determinada amostra de indivíduos/população.

Com relação ao número de observações coletadas, os dados são classificados em univariados, bivariados ou multivariados:

1. dados univariados: quando uma única observação de cada indivíduo é registrada.
2. Por exemplo: peso;
3. dados bivariados: quando duas observações de cada indivíduo são registradas. Por exemplo: peso e altura: A;
4. dados multivariados: quando mais duas observações acerca de cada indivíduo são registradas. Por exemplo: peso, altura, sexo e idade.
5. Quanto à forma de apresentação, os dados podem ser classificados em dados brutos ou rol.

1.4.1 DADOS BRUTOS

Os dados brutos são aqueles que não foram numericamente organizados em ordem crescente ou decrescente, ou seja, estão na forma como foram coletados. Como exemplo de dados brutos, podemos citar uma relação dos tempos médios de estudo diário, em minutos, de 50 alunos do Estratégia, na qual a seleção dos alunos ocorreu de forma aleatória, não havendo nenhuma ordenação de valores.

Esse tipo de tabela, onde os elementos não aparecem numericamente ordenados, é denominada de tabela primitiva. A tabela primitiva, em geral, oferece pouca ou nenhuma informação ao leitor, sendo necessário haver uma organização dos dados, a fim de torná-los mais expressivos.

Aluno	**Tempo** min.	**Aluno**	**Tempo** min.	**Aluno**	**Tempo** min.	**Aluno**	**Tempo** min.	**Aluno**	**Tempo** min.
1	143	11	113	21	170	31	124	41	105
2	142	12	143	22	158	32	137	42	154
3	161	13	159	23	123	33	153	43	99
4	126	14	168	24	96	34	129	44	114
5	134	15	123	25	98	35	148	45	161
6	137	16	135	26	135	36	173	46	128
7	171	17	135	27	129	37	126	47	175
8	85	18	175	28	126	38	104	48	137
9	155	19	115	29	103	39	157	49	165
10	171	20	89	171	40	50	115	150	170

Tabela 1.1: Dados brutos

1.4.2 ROL

O rol é a organização dos dados brutos em ordem de grandeza crescente ou decrescente. Com os dados organizados em rol, podemos saber, com facilidade, qual o menor e o maior elemento de um conjunto de dados. Os dados do nosso exemplo, isto é, os tempos médios de estudo diário, podem ser organizados em ordem crescente ou decrescente:

85	115	129	143	161
89	115	129	143	165
96	123	134	148	168
98	123	135	153	170
99	124	135	154	171
104	126	137	157	171
105	126	137	158	173
113	127	137	159	175
114	128	142	161	175

Tabela 1.2: Rol em ordem crescente

175	161	142	128	114
175	159	137	127	113
173	158	137	126	105
171	157	137	126	104
171	155	135	126	103
171	154	135	124	99
170	153	135	123	98
168	148	134	123	96
165	143	129	115	89
161	143	129	115	85

Tabela 1.3: Rol em ordem decrescente

1.5 VARIÁVEIS ESTATÍSTICAS

1.5.1 VARIÁVEIS ESTATÍSTICA.

A variável estatística consiste no conjunto de características que desejamos averiguar estatisticamente. Ela também pode ser definida como o objeto da pesquisa estatística. Por exemplo, se nosso interesse é conhecer quantas horas os alunos do Estratégia estudam diariamente, então nossa variável é o número de horas estudadas por dia. As variáveis estatísticas podem ser classificadas, inicialmente, em duas categorias: qualitativas e quantitativas.

1.5.2 VARIÁVEIS QUALITATIVAS

As variáveis qualitativas são as características que não podem ser descritas de forma numérica, mas que podem ser definidas por meio de qualidades (atributos ou categorias) do indivíduo pesquisado. Elas podem ser classificadas em nominais ou ordinais:

a. variável qualitativa nominal (ou categórica), as possíveis categorias não podem ser ordenadas. Por exemplo, a cor dos olhos dos moradores de uma determinada cidade (pretos, castanhos, azuis e verdes).

b. Variável qualitativa ordinal, as possíveis categorias podem ser ordenadas de alguma forma. Por exemplo, o grau de instrução dos funcionários de um determinado órgão (fundamental, médio, superior).

1.5.3 VARIÁVEIS QUALITATIVAS

As variáveis quantitativas são características que podem ser descritas em termos de quantidades (valores numéricos), obtidas por meio de contagem ou mensuração. Elas podem ser classificadas em discretas e contínuas:

a. variáveis quantitativas discretas, os possíveis valores formam um conjunto finito ou enumerável de números e, geralmente, resultam de um processo de contagem. O número de ocorrências da característica em análise pode ser contado. Por exemplo, o número de leitos abertos em hospitais de uma determinada cidade;

b. variáveis quantitativas contínuas, os possíveis valores formam um intervalo de números reais e, normalmente, resultam de um processo de mensuração. A característica pode ser medida em uma escala contínua, a qual podem ser associados um número infinito de possíveis valores, de modo a não haver lacunas ou interrupções. Por exemplo, a altura dos moradores de uma determinada cidade.

1.6 SÉRIES ESTATÍSTICAS

Uma série estatística consiste em um conjunto de dados organizado com base em uma característica comum, ou seja, uma mesma variável. Normalmente, uma série estatística é representada por meio de uma tabela ou de um gráfico, conforme ficar melhor representado, a fim de sintetizar os dados estatísticos observados e torná-los mais compreensivos.

a. Corpo — conjunto de linhas e colunas com as informações sobre a variável em estudo;

b. cabeçalho — parte superior que especifica o conteúdo das colunas;

c. coluna indicadora — parte que indica o conteúdo das linhas;

d. linhas — traços que facilitam a leitura dos dados;

e. célula — espaço onde os dados são armazenados;

f. título — identificação da tabela, contendo as informações sobre seu conteúdo;

g. fonte — referência de onde os dados foram obtidos, localizada no rodapé.

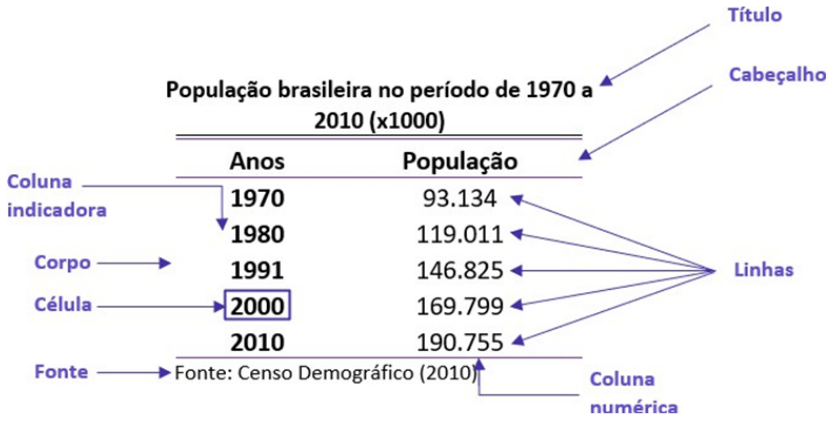


Figura 1.1: Série estatística

Um gráfico é uma forma clara e objetiva de apresentar uma série estatística. O objetivo é proporcionar uma compreensão mais rápida do fenômeno em estudo. Para isso, o gráfico deve ser destituído de detalhes sem importância (ser simples); permitir a correta intepretação dos valores representativos do fenômeno (ser claro); e transmitir a verdade sobre o fenômeno (ser verossímil). A série estatística apresentada na tabela anterior pode ser representada graficamente da seguinte forma:

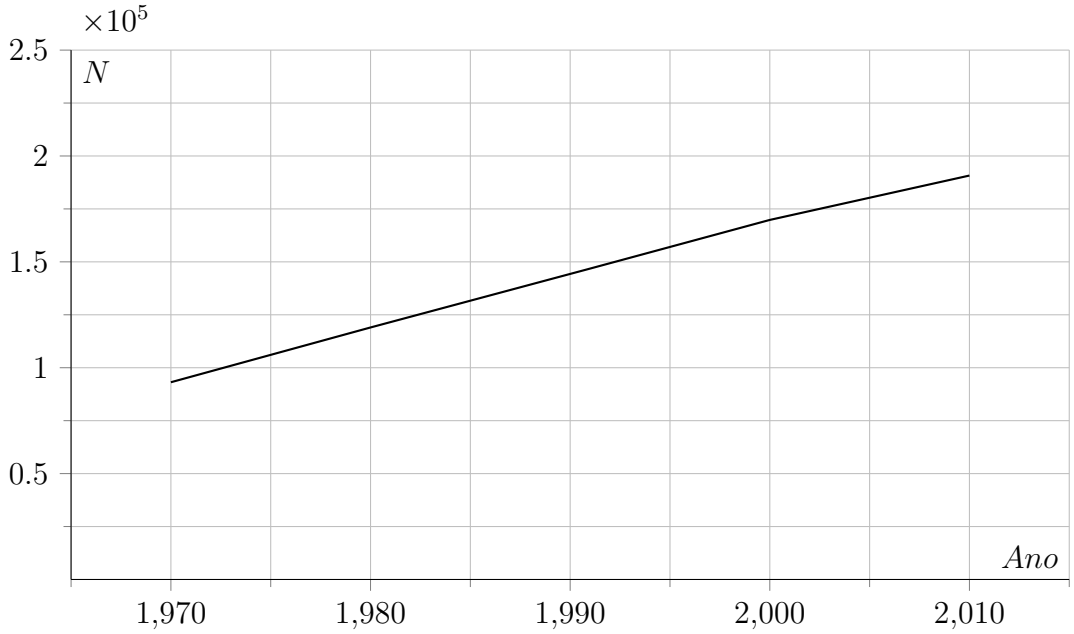


Figura 1.2: Variação da População.

Tabela.

- Quadro que resume um conjunto de observações.
- Composta de cabeçalho, corpo, coluna indicadora, linhas, células, título e fonte.

Gráfico.

- Forma simples e clara de apresentar uma série estatística.
- Proporcionar uma compreensão mais rápida do fenômeno em estudo.
- Deve ser simples, claro e verossímil.

Finalmente, podemos verificar a presença de três elementos nas séries estatísticas: o tempo, o espaço e a espécie. Conforme os elementos variem, a série pode ser classificada em temporal (ou cronológica), geográfica (ou territorial) e específica.

1.6.1 SÉRIES TEMPORAIS (CRONOLÓGICAS)

É a série cujos dados são dispostos segundo a época de ocorrência. Isto é, enquanto o tempo varia, o fato e o local permanecem constantes. Também são chamadas de séries históricas ou evolutivas. A principal característica é o fator cronológico variável.

A seguir temos a série histórica da população residente no Brasil no período de 1970 a 2010, com frequência decenal. Percebam que o tempo varia; contudo, o fato que está sendo analisado (quantidade populacional) e o local alvo da pesquisa (Brasil), continuam iguais.

Anos	População
1970	93143
1980	119011
1991	146825
2000	169799
2010	190755

Tabela 1.4: Censo 2010

1.6.2 SÉRIES GEOGRÁFICAS (OU TERRITORIAIS)

É a série cujos dados são dispostos segundo a localidade de ocorrência. Isto é, enquanto o local varia, o fato e o tempo permanecem constantes. Também são chamadas de séries espaciais ou de localização. A principal característica é o fator geográfico variável.

A seguir temos a série geográfica da população urbana residente em cada uma das regiões brasileiras no ano de 2010. Percebam que o local (região) varia; contudo, o fato que está sendo analisado (quantidade populacional) e o tempo (ano de 2010) permanecem constantes.

Região	População
norte	11664
nordeste	32821
sudeste	74696
sul	23260
centro-oeste	12482

Tabela 1.5: Censo 2010

1.6.3 SÉRIES ESPECÍFICAS

É a série cujos dados são dispostos segundo a modalidade de ocorrência. Isto é, enquanto o fato varia, a época e o local permanecem constantes. Também são chamadas de séries categóricas. A principal característica é o fator especificativo variável.

A seguir temos uma série específica das populações urbana e rural residentes no Brasil no ano de 2010. Percebam que os fatos analisados variam (população urbana x população rural); contudo, o tempo (2010) e o local de análise (Brasil) são onstantes.

Zona	População
Urbano	93134
Rural	119011

Tabela 1.6: Censo 2010

1.6.4 SÉRIES MISTAS (OU COMPOSTAS)

Muitas vezes, podemos ter a necessidade de apresentar, em uma única tabela, a variação de valores de mais de uma variável, isto é, combinar duas ou mais séries. As séries resultantes desse processo de combinação são chamadas de séries mistas (ou compostas) e apresentadas por meio de tabelas de dupla entrada.

O nome da nova série deve considerar pelo menos dois elementos. Assim, se for uma série mista de fato e tempo, denominaremos de série específico-temporal. A seguir temos uma série específicotemporal representando as populações de homens e mulheres residentes no Brasil, no período de 1970 a 2010, com variação decenal.

Tabela 1.7: População residente no Brasil por sexo (1970–2010).

	Sexo	
Ano	Homens	Mulheres
1970	46327	46807
1980	59142	59868
1991	72485	74340
2000	83602	86270
2010	93406	97348

Por sua vez, se tivermos uma série mista de local e tempo, denominaremos de série geográfica-temporal. A seguir temos uma série geográfico-temporal representando as populações residentes em cada região brasileira, no período de 1970 a 2010, com variação decenal.

	Regiões				
Anos	N	NE	SE	S	CO
1970	3105	28111	3980	16496	5072
1980	5880	34815	51737	19031	7545
1991	10030	42497	62470	22129	9427
2000	12900	47741	72412	25107	11636
2010	15864	53081	80364	27386	14058

Tabela 1.8: Censo 2010

Por fim, devemos notar que podem existir séries compostas de três ou mais entradas, embora isso raramente aconteça, por conta da dificuldade de representação.

1.7 DISTRIBUIÇÃO DE FREQUÊNCIAS

Vimos anteriormente que, logo após a coleta de dados, temos o que denominamos de dados brutos. Como exemplo de dados brutos, citamos uma pesquisa de tempo médio de estudo diário, em minutos, envolvendo 50 alunos do Estratégia, onde os alunos foram escolhidos de maneira aleatória, não havendo nenhuma organização dos valores observados. Por serem apresentados na forma em que foram coletados, são denominados de dados brutos.

Aluno	Tempo	Aluno	Tempo	Aluno	Tempo	Aluno	Tempo	Aluno	Tempo
1	143	11	113	21	170	31	124	41	105
2	142	12	143	22	158	32	137	42	154
3	161	13	159	23	123	33	153	43	99
4	126	14	168	24	96	34	129	44	114
5	134	15	123	25	98	35	148	45	161
6	137	16	135	26	135	36	173	46	128
7	171	17	135	27	129	37	126	47	175
8	85	18	175	28	126	38	104	48	137
9	155	19	115	29	103	39	157	49	165
10	171	20	89	30	171	40	127	50	115

Tabela 1.9: Dados brutos.

Normalmente, esses dados fornecem pouca informação ao leitor, sendo necessário organizá-los, com o propósito de aumentar sua capacidade informativa. A simples organização dos dados em um rol crescente já ajuda bastante nesse sentido. Com os dados organizados em rol, conseguimos verificar que o menor tempo observado foi de 85 minutos, e o maior, de 175 minutos, o que nos fornece uma amplitude total ($AT = 175 - 85 = 90$) de variação da ordem de 90 minutos.

Rol em ordem crescente

85	115	129	143	161
89	115	129	143	165
96	123	134	148	168
98	123	135	153	170
99	124	135	154	171
104	126	137	157	171
105	126	137	158	173
113	127	137	159	175
114	128	142	161	175

Outra informação que conseguimos extrair dos dados organizados em rol crescente é que alguns tempos, como 126 min, 135 min, 137 min e 171 min, foram mais frequentes, ou seja, apareceram mais vezes durante a pesquisa. Uma maneira mais concisa de mostrar os dados do rol é apresentar cada valor juntamente com o número de vezes em que ocorre, em vez de repeti-los. O número de ocorrências de um determinado valor recebe o nome de frequência. A tabela que contém todos os valores com suas respectivas frequências é denominada de distribuição de frequências.

Uma distribuição de frequências também pode ser definida como uma série estatística na qual permanecem constantes o fato, o local e a época. Ela pode ser classificada em dois tipos: distribuição

de frequências pontual (ou discreta) e distribuição de frequências intervalar (ou contínua). Na distribuição de frequências pontual, são apresentados todos os dados coletados juntamente com suas respectivas frequências, não havendo perda de valores. Contudo, esse processo pode exigir muito espaço, especialmente quando o número de valores da variável tende a aumentar.

Tempo (min)	**Freq.**	**Tempo (min)**	**Freq.**	**Tempo (min)**	**Freq.**	**Tempo (min)**	**Freq.**
85	1	114	1	135	3	158	1
89	1	115	2	137	3	159	1
96	1	123	2	142	1	161	2
98	1	124	1	143	2	165	1
99	1	126	3	148	1	168	1
103	1	127	1	153	1	170	1
104	1	128	1	154	1	171	3
105	1	129	2	155	1	173	1
113	1	134	1	157	1	175	2

Tabela 1.10: Distribuição de frequências do tempo (min).

Nesse caso, quando a variável é contínua, o mais recomendável é agrupar os valores por intervalos de classe. Desse modo, em vez de listar cada um dos valores que ocorrem, utilizamos uma distribuição de frequências intervalar, listando os intervalos de classe e as frequências correspondentes.

Tabela 1.11: Distribuição de frequências do tempo médio.

Tempo médio (X_i)	**Frequência (f_i)**
$85 \leq x < 100$	5
$100 \leq x < 115$	5
$115 \leq x < 130$	12
$130 \leq x < 145$	10
$145 \leq x < 160$	7
$160 \leq x < 175$	9
$175 \leq x < 190$	2

Procedendo dessa forma, perdemos a informação detalhada dos tempos médios, mas ganhamos em termos de praticidade, o que simplifica o processo de análise de dados. Examinando a Tabela 1.13, percebemos facilmente que a maioria dos alunos estuda diariamente entre 115 e 130 minutos, enquanto uma minoria alcança entre 175 e 190 minutos.

Para identificar uma classe, temos que conhecer os valores dos limites inferior e superior da classe, que delimitam um intervalo de classe. Desse modo, precisamos definir a natureza do intervalo de classe, se aberto ou fechado. Portanto, temos as seguintes notações para os diferentes tipos de intervalos:

Tabela 1.12: Tipos de intervalos e suas notações.

Tipo de intervalo	**Notação matemática**	**Notação estatística**	**Significado**
Intervalo aberto	$a < x < b$	(a, b)	Engloba todos os elementos entre a e b, mas não engloba a nem b.
Fechado à esquerda e aberto à direita	$a \leq x < b$	$[a, b)$	Engloba todos os elementos entre a e b, inclusive a mas não b.
Aberto à esquerda e fechado à direita	$a < x \leq b$	$(a, b]$	Engloba todos os elementos entre a e b, inclusive b mas não a.
Intervalo fechado	$a \leq x \leq b$	$[a, b]$	Engloba todos os elementos entre a e b, inclusive a e b.

Por fim, é importante salientarmos que, em análises estatísticas, constantemente encontramos distribuições de frequências intervalares, pois o objetivo da estatística é justamente fazer um apanhado geral das características de um conjunto de dados, sem adentrar em detalhes de casos particulares.

1.7.1 ELEMENTOS DE UMA DISTRIBUIÇÃO DE FREQUÊNCIAS

Agora, analisaremos detalhadamente cada elemento de uma distribuição de frequências.

Tabela 1.13: Distribuição de frequências do tempo médio.

Tempo médio (X_i)	**Frequência (f_i)**
$85 \leq x < 100$	5
$100 \leq x < 115$	5
$115 \leq x < 130$	12
$130 \leq x < 145$	10
$145 \leq x < 160$	7
$160 \leq x < 175$	9
$175 \leq x < 190$	2

1.7.2 CLASSE

As classes são os intervalos nos quais o fenômeno é subdividido. Podemos dizer que as classes são os intervalos ou subdivisões dos elementos que compõem um conjunto de dados. Na Tabela 1.13, as classes são: $85 \leq x < 100$, $100 \leq x < 115$, $115 \leq x < 130$, $130 \leq x < 145$, $145 \leq x < 160$, $160 \leq x < 175$, $175 \leq x < 190$.

Existem duas maneiras de determinar o número ”ideal”de classes, k, em função do número de

dados da tabela, n. A primeira consiste em utilizar a fórmula de Sturges:

$$k = 1 + 3{,}3 \log n.$$

Outra abordagem, utilizada quando o número de dados é menor ou igual a 50, é:

$$k = \sqrt{n}.$$

1.7.3 LIMITE DE CLASSE

Cada classe tem um limite inferior de classe l_{inf}, sendo o menor número que pode pertencer à classe, e um limite superior de classe l_{sup}, sendo o maior número que pode pertencer à classe. Os limites de uma classe são seus valores extremos.

Tabela 1.14: Classes e respectivos limites inferior e superior.

Classes		**Limite Inferior (l_{inf})**	**Limite Superior (l_{sup})**
Pimeira classe	$85 \leq x < 100$	85	100
Segunda Classe	$100 \leq x < 115$	100	115
Terceira Classe	$115 \leq x < 130$	115	130
Quarta Classe	$130 \leq x < 145$	130	145
Quinta Classe	$145 \leq x < 160$	145	160
Sexta Classe	$160 \leq x < 175$	160	175
Sétima Classe	$175 \leq x < 190$	175	190

1.7.4 AMPLITUDE DE UM INTERVALO DE CLASSE

A amplitude de um intervalo de classe, ou simplesmente intervalo de classe, é a distância entre os limites inferiores (ou superiores) de classes consecutivas. Ela é obtida pela diferença entre dois limites inferiores (ou superiores) consecutivos:

$$h = l_{sup} - l_{inf},$$

em que l_{inf} é o limite inferior do intervalo de classe e l_{sup} é o limite superior do intervalo de classe.

Tabela 1.15: Limites e amplitude das classes.

Classes	**Limite Inferior (l_{inf})**	**Limite Superior (l_{sup})**	**Amplitude (h)**
$85 \leq x < 100$	85	100	15
$100 \leq x < 115$	100	115	15
$115 \leq x < 130$	115	130	15
$130 \leq x < 145$	130	145	15
$145 \leq x < 160$	145	160	15
$160 \leq x < 175$	160	175	15
$175 \leq x < 190$	175	190	15

Embora seja desejável, a amplitude do intervalo de classe nem sempre será constante ao longo de toda a distribuição de frequências intervalar.

1.7.5 AMPLITUDE TOTAL

A amplitude total é a diferença entre o limite superior da última classe (limite superior máximo) e o limite inferior da primeira classe (limite inferior mínimo). Portanto, corresponde à diferença entre o último e o primeiro elemento de um conjunto de dados ordenado crescentemente:

$$AT = l_{max} - l_{min}.$$

Note que, quando todas as classes possuem a mesma amplitude, também podemos determinar o valor da amplitude total multiplicando o valor do intervalo de classe h pela quantidade de classes da distribuição k:

$$AT = h \times k.$$

1.7.6 PONTO MÉDIO DA CLASSE

O ponto médio é a média aritmética simples dos valores extremos de uma classe, ou seja, a soma dos limites inferior e superior dividida por dois. Esse ponto divide a classe em duas partes iguais. Também costuma ser chamado de marca ou representante da classe:

$$PM = \frac{l_{sup} + l_{inf}}{2}.$$

Tabela 1.16: Limites e ponto médio das classes.

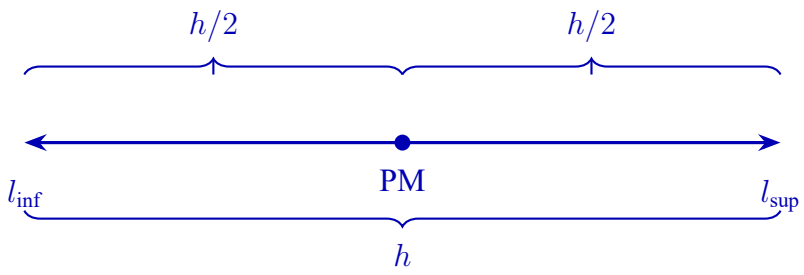
Classes	**Limite Inferior** (l_{inf})	**Limite Superior** (l_{sup})	**Ponto médio** (PM)
$85 \leq x < 100$	85	100	92,5
$100 \leq x < 115$	100	115	107,5
$115 \leq x < 130$	115	130	122,5
$130 \leq x < 145$	130	145	137,5
$145 \leq x < 160$	145	160	152,5
$160 \leq x < 175$	160	175	167,5
$175 \leq x < 190$	175	190	182,5

Veja que os pontos médios formaram uma progressão aritmética, pois a diferença entre dois pontos médios consecutivos foi constante e igual a 15. Isso ocorreu porque o intervalo de classe h também foi constante (e igual a 15) em toda a distribuição. Assim, quando o intervalo de classes é constante, a diferença entre os pontos médios também será constante e igual ao intervalo de classe.

Adicionalmente, sabendo que o ponto médio divide a classe em duas partes iguais, podemos derivar outras relações envolvendo o próprio ponto médio, a amplitude de classe e os limites inferior e superior.

Dada a figura anterior, podemos obter os limites de uma classe por meio das seguintes expressões:

$$l_{inf} = PM - \frac{h}{2} \qquad \text{e} \qquad l_{sup} = PM + \frac{h}{2}.$$



Além disso, podemos encontrar o ponto médio de uma classe a partir das seguintes relações:

$$PM = l_{inf} + \frac{h}{2} \qquad \text{e} \qquad PM = l_{sup} - \frac{h}{2}.$$

1.7.7 FREQUÊNCIA

Ao longo dessa aula, em várias oportunidades abordamos conceitos relacionados à frequência, isto é, ao número de ocorrências de um determinado valor ou de uma certa classe. Esse conceito é de grande relevância para a estatística descritiva e deve ser estudado de forma mais aprofundada. Nesse contexto, é importante sabermos que existem quatro tipos de frequência, os quais serão analisados nas subseções seguintes:

a. Frequência absoluta simples (f_i)

b. Frequência absoluta acumulada (f_{ac} e f_{ad})

c. Frequência relativa simples (F_i)

d. Frequência relativa acumulada (F_{ac} e F_{ad})

1.7.8 FREQUÊNCIA ABSOLUTA SIMPLES

A frequência absoluta simples corresponde ao número de observações correspondentes a uma determinada classe ou a um determinado valor.

i	Tempos (min)	Frequência f_i
1	$85 \vdash 100$	5
2	$100 \vdash 115$	5
3	$115 \vdash 130$	12
4	$130 \vdash 145$	10
5	$145 \vdash 160$	7
6	$160 \vdash 175$	9

Tabela 1.17: Frequência Absoluta Simples

No exemplo, temos: $f_1 = 5$, $f_2 = 5$, $f_3 = 12$, $f_4 = 10$, $f_5 = 7$, $f_6 = 9$, $f_7 = 2$.
A soma de todas as frequências é igual ao número total de dados analisados:

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

em que a notação $\sum_{i=1}^{k} f_i$ representa o somatório das frequências de cada uma das k classes.

1.7.9 FREQUÊNCIA ABSOLUTA ACUMULADA

A frequência absoluta acumulada crescente f_{ac} é a soma das frequências de todos os valores inferiores ao limite superior do intervalo de uma determinada classe. Assim,

$$f_{ac,i} = f_1 + f_2 + \cdots + f_i.$$

A frequência absoluta acumulada decrescente f_{ad} é a soma das frequências de todos os valores superiores ao limite inferior do intervalo de uma determinada classe. Assim,

$$f_{ad,i} = f_i + f_{i+1} + \cdots + f_k.$$

A Tabela 1.18 apresenta a frequência absoluta simples e as frequências absolutas acumuladas (crescente e decrescente) para o exemplo.

Tabela 1.18: Frequência absoluta simples e acumuladas (crescente e decrescente).

i	**Tempos (min.)**	**Frequência (f_i)**	**Frequência Acum. crescente (f_{ac})**	**Frequência Acum. decrescente (f_{ad})**
1	$85 \leq x < 100$	5	5	50
2	$100 \leq x < 115$	5	10	45
3	$115 \leq x < 130$	12	22	40
4	$130 \leq x < 145$	10	32	28
5	$145 \leq x < 160$	7	39	18
6	$160 \leq x < 175$	9	48	11
7	$175 \leq x < 190$	2	50	2
	n	**50**	**50**	**50**

1.7.10 FREQUÊNCIA RELATIVA SIMPLES

A frequência relativa simples corresponde à proporção de dados existentes em uma determinada classe. Para calcular a frequência relativa de uma classe, dividimos a frequência absoluta simples f_i pela frequência total:

$$F_i = \frac{f_i}{n}.$$

Para representar esses valores em termos de porcentagem, basta multiplicarmos por 100%.

Repare que a soma de todas as frequências relativas deve ser igual a 100%:

$$\sum_{i=1}^{k} F_i = 100\%.$$

1.7.11 FREQUÊNCIA RELATIVA ACUMULADA

A frequência relativa acumulada crescente F_{ac} é a proporção de valores inferiores ao limite superior do intervalo de uma dada classe:

$$F_{ac,i} = F_1 + F_2 + \cdots + F_i.$$

A frequência relativa acumulada decrescente F_{ad} é a proporção de valores superiores ao limite inferior do intervalo de uma dada classe:

$$F_{ad,i} = F_i + F_{i+1} + \cdots + F_k.$$

A Tabela apresenta as frequências relativas simples e acumuladas (crescente e decrescente) para o exemplo.

Tabela 1.19: Frequências relativas simples e acumuladas (em %).

i	Tempos (min.)	Frequência Relativa (F_i)	Frequência Rel. Acum. crescente (F_{ac})	Frequência Rel. Acum. decrescente (F_{ad})
1	$85 \leq x < 100$	10%	10%	100%
2	$100 \leq x < 115$	10%	20%	90%
3	$115 \leq x < 130$	24%	44%	80%
4	$130 \leq x < 145$	20%	64%	56%
5	$145 \leq x < 160$	14%	78%	36%
6	$160 \leq x < 175$	18%	96%	22%
7	$175 \leq x < 190$	4%	100%	4%
	$\sum f_i = 50$	$\sum F_i = 100\%$		

1.8 DENSIDADE DE FREQUÊNCIA

A densidade de frequência de uma classe consiste no quociente entre a frequência da classe (absoluta ou relativa) e sua amplitude:

$$\text{densidade} = \frac{\text{frequência}}{\text{amplitude}}.$$

Usando a frequência absoluta e a amplitude do intervalo h_i, temos:

$$d_i = \frac{f_i}{h_i}.$$

Tabela 1.20: Densidade de frequência (para $h_i = 15$).

i	Tempos (min.)	Frequência (f_i)	Densidade de Frequência (d_i)
1	$85 \leq x < 100$	5	$d_1 = \frac{5}{15} = 0{,}33$
2	$100 \leq x < 115$	5	$d_2 = \frac{5}{15} = 0{,}33$
3	$115 \leq x < 130$	12	$d_3 = \frac{12}{15} = 0{,}80$
4	$130 \leq x < 145$	10	$d_4 = \frac{10}{15} = 0{,}67$
5	$145 \leq x < 160$	7	$d_5 = \frac{7}{15} = 0{,}47$
6	$160 \leq x < 175$	9	$d_6 = \frac{9}{15} = 0{,}60$
7	$175 \leq x < 190$	2	$d_7 = \frac{2}{15} = 0{,}13$

Tabela 1.21: Principais conceitos e notações da distribuição de frequências.

Item	**Definição**	**Símbolos e fórmulas**
Número de classes	As classes são os intervalos nos quais o fenômeno é subdividido.	$k = 1 + 3{,}3 \log n$ ou $k = \sqrt{n}$
Limites de classe	Correspondem aos valores extremos de cada classe.	l_{inf} e l_{sup}
Amplitude de um intervalo de classe	Distância entre os limites inferiores (ou superiores) de classes consecutivas.	$h = l_{\text{sup}} - l_{\text{inf}}$
Amplitude total	Diferença entre o limite superior máximo e o limite inferior mínimo.	$AT = l_{\text{max}} - l_{\text{min}}$ ou $AT = h \times k$
Ponto médio	Média aritmética simples dos valores extremos de uma classe.	$PM = \dfrac{l_{\text{inf}} + l_{\text{sup}}}{2}$ $PM = l_{\text{inf}} + \dfrac{h}{2}$, $PM = l_{\text{sup}} - \dfrac{h}{2}$
Frequência absoluta simples	Número de observações correspondentes a uma classe.	f_i
Frequência absoluta acumulada	Soma das frequências até a classe considerada.	$f_{ac,i} = f_1 + f_2 + \cdots + f_i$
Frequência relativa simples	Proporção de dados existentes em uma classe.	$F_i = \dfrac{f_i}{n}$
Frequência relativa acumulada	Proporção acumulada até a classe considerada.	$F_{ac,i} = F_1 + F_2 + \cdots + F_i$

Quartil para dados agrupados em classes

O cálculo do quartil para dados agrupados em classes será realizado por meio das seguintes etapas:

1.9 REPRESENTAÇÃO GRÁFICA DAS DISTRIBUIÇÕES DE FREQUÊNCIAS

1.9.1 GRÁFICO DE HASTES OU BASTÕES

O gráfico de hastes ou bastões é muito utilizado para representar dados não agrupados em classes, o que normalmente ocorre com dados discretos. Nesse caso, não há perda de informação pois os valores da variável aparecem individualmente, conforme constam da amostra. Com relação a sua construção, basta representarmos as frequências simples absolutas ou relativas de cada elemento do conjunto de dados.

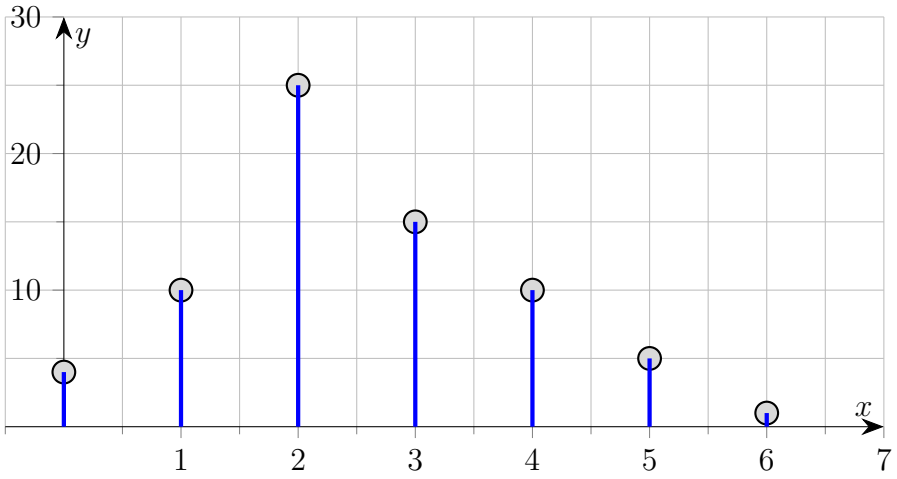


Figura 1.3: Gráfico de Hastes

O histograma é um gráfico destinado a representar dados agrupados em classe, sendo composto por um conjunto de retângulos contíguos (justapostos) cujas bases ficam localizadas sobre o eixo horizontal (eixo x), de forma que os seus pontos médios devem coincidir com os pontos médios dos intervalos de classe e seus limites devem coincidir com os limites da classe.

A quantidade de retângulos em um histograma é equivalente ao número de intervalos de classe. A largura de cada retângulo deve ser igual à amplitude do intervalo de classe, enquanto a altura precisa ser proporcional à frequência do intervalo de classe. Além disso, a área do histograma é proporcional ao somatório das frequências.

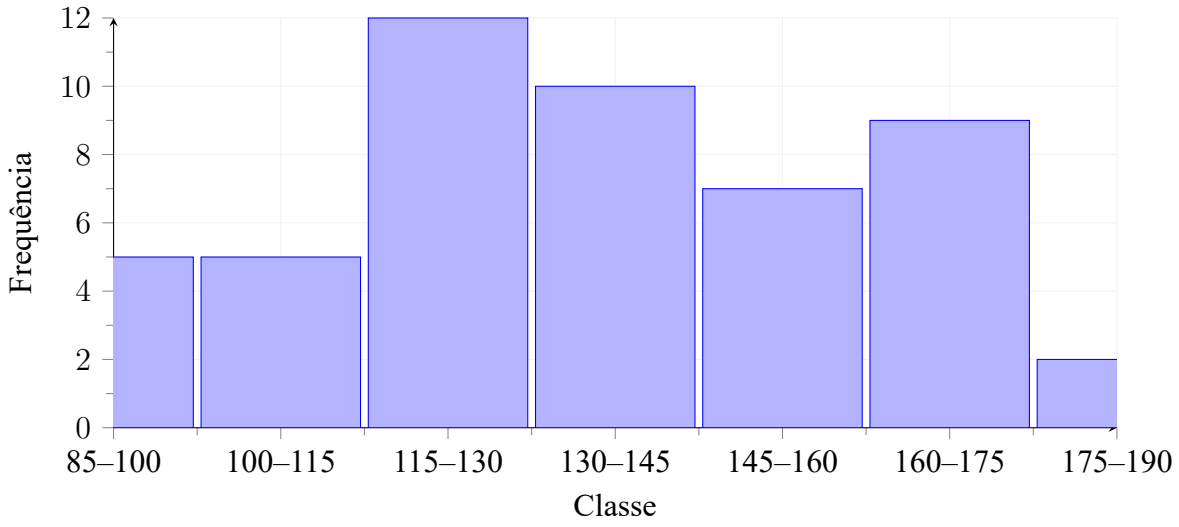


Figura 1.4: Histograma (classes e frequências).

A diferença básica entre um histograma e um gráfico de colunas (estudaremos na próxima seção) é a separação entre os retângulos adjacentes. Veja que não existe separação entre os retângulos no caso do histograma.

Dito isso, é importante mencionarmos a existência do gráfico denominado de poligonal característica, que construímos utilizando somente os contornos do histograma.

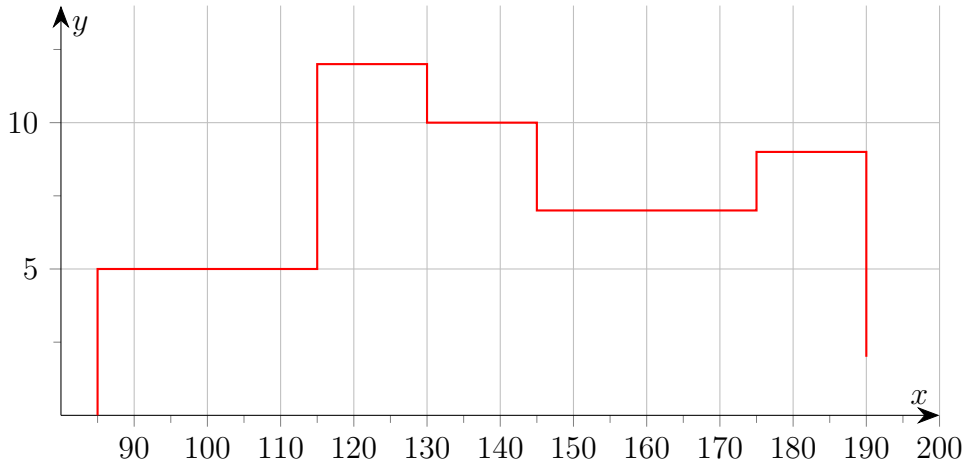


Figura 1.5: Contorno do Histograma.

1.9.2 POLÍGONO DE FREQUÊNCIAS

O polígono de frequências é um gráfico em linha obtido por meio da ligação, por segmentos de reta, dos pontos médios das bases superiores dos retângulos de um histograma. Também é necessário considerar a existência de uma classe anterior à primeira e outra posterior à última, ambas com a frequência nula.

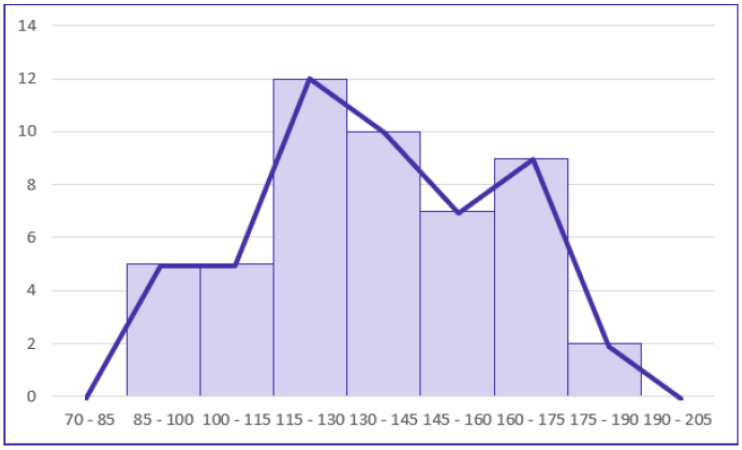


Figura 1.6: Histograma com polígono

1.9.3 CURVA DE FREQUÊNCIAS

A curva de frequências é obtida a partir do polimento de um polígono de frequências. Em sentido geométrico, o polimento corresponde à eliminação dos vértices (cantos) da linha poligonal. Esse processo suaviza os contornos do polígono de frequências, o que evidencia a verdadeira natureza dos dados em análise.

O polígono de frequências fornece a imagem real do fenômeno investigado, enquanto a curva de frequência mostra sua tendência. Naturalmente, quando o conjunto de dados é grande, a linha poligonal se torna curva. Por isso, podemos afirmar que a curva de frequência antecipa o comportamento da distribuição para um número maior de dados.

O processo de polimento é realizado por meio da seguinte fórmula:

$$fc_i = \frac{fant + 2 \times f_i + f_{post}}{4}$$

Tempo Médio (X_i)	Ponto Médio (PM_i)	Frequência(f_i)	Frequência Calculada
$70 \leq x < 85$	77,5	0	$fc_0 = \frac{0+2\times0+5}{4} = 1,25$
$85 \leq x < 100$	92,2	5	$fc_1 = \frac{0+2\times0+5}{4} = 3,75$
$100 \leq x < 115$	107,5	5	$fc_2 = \frac{0+2\times0+5}{4} = 6,75$
$115 \leq x < 130$	122,5	12	$fc_3 = \frac{0+2\times0+5}{4} = 9,75$
$130 \leq x < 145$	137,5	10	$fc_4 = \frac{0+2\times0+5}{4} = 9,75$
$145 \leq x < 160$	152,5	7	$fc_5 = \frac{0+2\times0+5}{4} = 8,75$
$160 \leq x < 175$	167,5	9	$fc_6 = \frac{0+2\times0+5}{4} = 6,75$
$175 \leq x < 190$	192,5	2	$fc_7 = \frac{0+2\times0+5}{4} = 3,25$
$190 \leq x < 205$	197,5	0	$fc_8 = \frac{0+2\times0+5}{4} = 0,50$

Tabela 1.22: Frequências calculadas

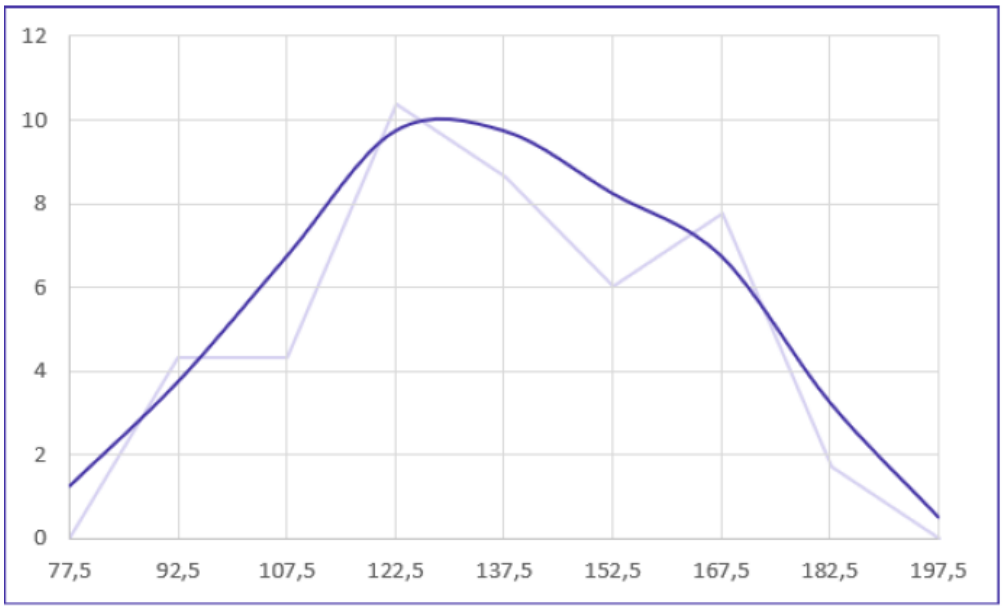


Figura 1.7: Curva de Frequência

Capítulo 2

Medidas de Tendência Central

2.1 Noções Iniciais sobre Médias

A média é um número que, de algum modo, resume as características de um grupo. Nosso contato com a média surge ainda na escola, quando os professores calculam a média das nossas avaliações. Suponha que um aluno tenha obtido as seguintes notas em determinada disciplina: 4; 10; 8; 10; 7; 9. Com base nesses dados, podemos concluir que a média aritmética desse aluno nessa disciplina será 8. Assim, a média consegue representar o conjunto dos valores observados.

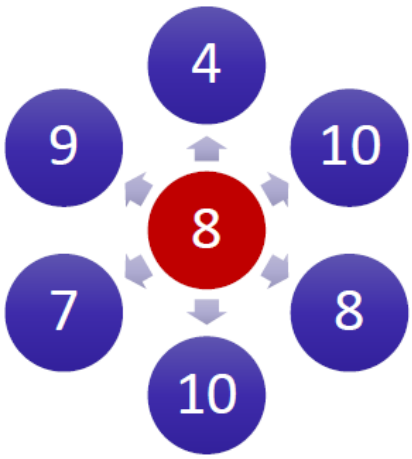


Figura 2.1: Média

A média está presente em nosso cotidiano. Com certa frequência, as notícias abordam conceitos que estão relacionados com a média: expectativa de vida dos brasileiros; idade média de uma população; renda domiciliar per capita brasileira; consumo médio de combustível; tempo médio de deslocamento em um trajeto.

Vamos analisar o primeiro exemplo. Quando o noticiário diz que expectativa de vida no Japão teve um aumento recorde na última década, chegando a 84,2 anos. O que você entende diante dessa informação?

Essa informação reflete a qualidade de vida da população japonesa. Ela nos mostra que a população japonesa está envelhecendo de forma saudável e que o sistema de saúde está sendo eficaz.

Se o noticiário também disser que a expectativa de vida em Angola gira em torno de 60 anos, você será capaz comparar essas duas populações, certo? A resposta é sim. Quando comparados, os

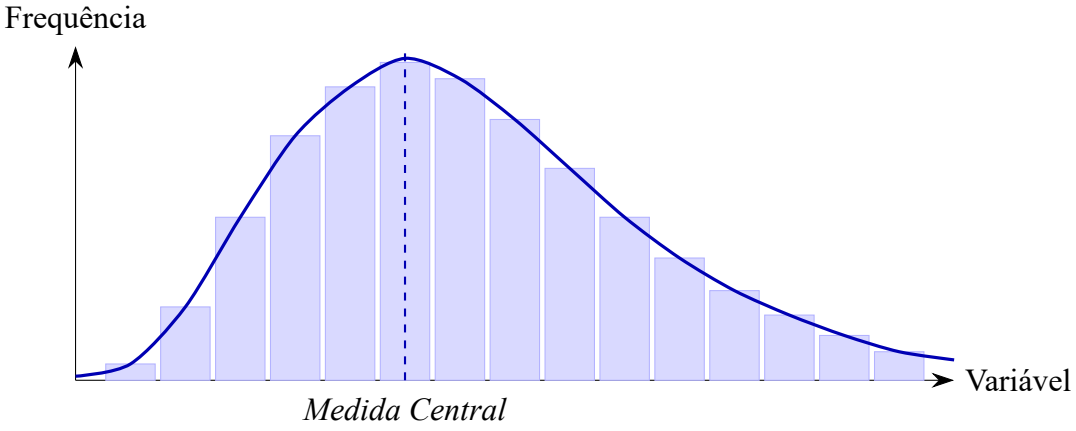
números mostram que a qualidade de vida em Angola não é tão boa quanto a do Japão. Também nos dizem que os angolanos tendem a viver, em média, 24 anos a menos que os japoneses.

Ao longo dessa aula, aprenderemos a calcular a média em diversas situações. Vamos ver que, a depender de como os dados nos forem apresentados, o cálculo será feito de uma forma diferente. Além disso, conheceremos algumas propriedades da média que facilitam a resolução de questões.

2.2 MEDIDAS DE POSIÇÃO

Muitas vezes, queremos resumir um conjunto de dados apresentando um ou alguns valores que sejam representativos de uma série toda. As medidas de posição são estatísticas que caracterizam o comportamento dos elementos de uma série de dados, orientando quanto à posição da distribuição em relação ao eixo horizontal do gráfico da curva de frequência.

Em outras palavras, podemos dizer que as medidas de posição indicam a tendência de concentração dos elementos de uma série, apontando o valor que melhor representa o conjunto de dados. Por exemplo, podemos ter uma medida para representar a posição de maior frequência de uma distribuição:

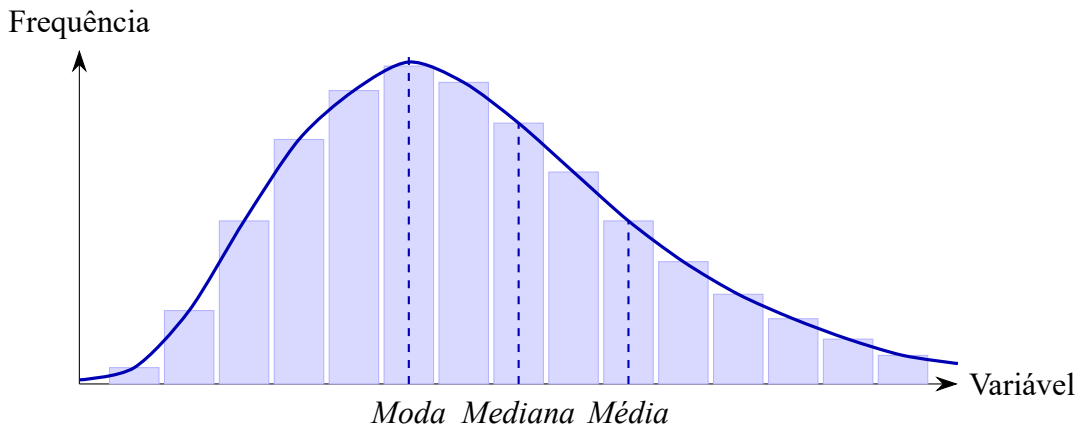


As medidas de posição podem ser divididas em:

a. medidas de tendência central: representam o ponto central ou o valor típico de um conjunto de dados, indicando onde está localizada a maioria dos valores de uma distribuição.

   - média aritmética: é a medida de posição mais utilizada, sendo o valor resultante da divisão entre a soma de todos os valores de uma série de observações e o número de observações;
   - mediana: valor que ocupa a posição central de uma série de observações, quando organizadas em ordem crescente ou decrescente; e
   - moda: valor mais frequente em uma série de observações,

Somente a título de exemplo, vejamos como as medidas de tendência central se posicionam em relação a uma distribuição de frequências. Notem que essas medidas tendem a ocupar as posições centrais da distribuição, sendo denominadas de medidas de tendência central.



b. medidas separatrizes: dividem (ou separam) uma série em duas ou mais partes, cada uma contendo a mesma quantidade de elementos. As medidas mais utilizadas são:

- mediana: divide uma série em duas partes iguais. Reparem que, além de ser uma medida separatriz, a mediana também é uma medida de tendência central;
- quartis: dividem uma série em quatro partes iguais;
- decis: dividem uma série em dez partes iguais; e
- percentis: dividem uma série em cem partes iguais.

2.3 NOTAÇÃO DE SOMATÓRIO

Com frequência, as fórmulas matemáticas exigem a adição de muitas variáveis, como a média aritmética. O somatório ou notação sigma é uma forma simples e conveniente de abreviação, usada para fornecer uma expressão concisa para a soma dos valores de uma variável. Por exemplo, se quisermos representar a soma de um número de termos tais como:

$$1 + 2 + 3 + 4 + 5$$

ou,

$$1^2 + 2^2 + 3^2 + 5^2 + 5^2$$

em que há um padrão evidente para os números envolvidos.

De modo geral, se tomarmos uma sequência de números $x_1, x_2, x_3, \ldots, x_n$ então podemos escrever a soma desses números como $x_1, x_2, x_3, \ldots, x_n$. Nesse conjunto, x_1 representa o primeiro termo; x_2 representa o segundo; x_3, o terceiro; e x_i o i-ésimo termo da soma.

Essa soma pode ser representada de uma forma mais simples e concisa, deixando que x_i represente o termo geral da sequência. Para isso, empregamos a seguinte notação:

$$\sum_{i=1}^{n} x_i$$

Então, em vez de usarmos vários elementos para determinarmos o somatório, utilizamos somente o símbolo do somatório:

Essa notação envolve um símbolo de somatório, $\sum$, sendo a letra grega maiúscula Sigma Σ. Basicamente, esse símbolo está nos instruindo a somar determinados elementos de uma sequência. Os elementos típicos da sequência que está sendo somada aparecem à direita do símbolo de somatório:

$$\sum x_i$$

Observe que essa notação também requer a definição de um índice, que fica localizado abaixo do símbolo de somatório. Esse índice é frequentemente representado por i, embora também seja comum encontrarmos questões adotando j ou n. Esse índice normalmente aparece como uma expressão, por exemplo, $i = a$, em que o índice assume um valor inicial atribuído no lado direito da equação, conhecido como limite inferior a. Se considerarmos que $i = 1$, estamos dizendo que o primeiro elemento da sequência a ser considerado é o de índice igual a 1, isto é, o primeiro elemento da sequência. A condição de parada ou limite superior do somatório é o valor localizado acima do símbolo, no caso b. A condição de parada indica o último elemento da sequência a ser considerado no somatório.

$$\sum_{i=a}^{b} x_i$$

Então, se tivermos uma sequência de 10 valores, devemos interpretar a notação a seguir como a soma dos valores da sequência x_i, com i variando de 1 a 10:

$$\sum_{i=1}^{10} x_i$$

Vejamos alguns exemplos típicos de operações envolvendo somatórios. Para isso, tomaremos como exemplo a sequência x_i = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10:

Notação	Interpretação e exemplo
$\sum_{i=3}^{10} x_i$	Representa a soma dos valores de x, começando em x_3 e terminando em x_{10}. $$\sum_{i=3}^{10} x_i = x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$$ $$= 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 = 55$$
$\sum x$	Quando os limites do somatório são omitidos, entende-se que a soma é feita de x_1 até x_n. $$\sum x = x_1 + x_2 + x_3 + \cdots + x_n$$ $$= 1 + 2 + 3 + \cdots + 10 = 55$$
$\sum_{i=1}^{n} x_i^2$	Representa a soma dos quadrados dos valores de x, de x_1 até x_n. $$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2$$ $$= 1^2 + 2^2 + 3^2 + \cdots + 10^2 = 385$$
$\sum_{k=1}^{4}(2x_k + 1)$	Representa a soma dos termos da sequência $(2n + 1)$, com k variando de 1 a 4. $$= (2 \cdot 1 + 1) + (2 \cdot 2 + 1) + (2 \cdot 3 + 1) + (2 \cdot 4 + 1)$$ $$= 3 + 5 + 7 + 9 = 24$$
$\sum_{i=3}^{5}\left(\frac{x_i}{x_i + 1}\right)$	Representa a soma dos termos $\frac{x_i}{x_i + 1}$, com i começando em 3 e terminando em 5. $$\sum_{i=3}^{5}\left(\frac{x_i}{x_i + 1}\right) = \frac{x_3}{x_3 + 1} + \frac{x_4}{x_4 + 1} + \frac{x_5}{x_5 + 1}$$ $$= \frac{3}{3+1} + \frac{4}{4+1} + \frac{5}{5+1} = \frac{3}{4} + \frac{4}{5} + \frac{5}{6}$$ $$= \frac{45 + 48 + 50}{60} = \frac{143}{60}$$

Tabela 2.1: Interpretação de diferentes notações de somatório.

Agora que já entendemos o funcionamento básico dessa notação, precisamos analisar outras operações aritméticas que também podem ser realizadas com as variáveis num somatório. Para tanto, vamos tomar como base as sequências x_i = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 e y_i = 3, 6, 9, 12, 15, 18, 21, 24, 27, 30.

Por exemplo, na SOMA DOS PRODUTOS, multiplicamos x_1 por y_1; x_2 por y_2; e assim por diante, até x_n por y_n. Em seguida, somamos os resultados de cada multiplicação. A SOMA DOS PRODUTOS da variável x pela variável y, com i variando de 1 a 10, pode ser representada por meio da seguinte expressão:

$$\sum_{i=1}^{n} x_i \times y_i = x_1 \times y_1 + x_2 \times y_2 + x_3 \times y_3 + \cdots + x_n \times y_n$$

$$\sum_{i=1}^{n} x_i \times y_i = 1 \times 3 + 2 \times 6 + 3 \times 9 + \cdots + 10 \times 30 = 1155$$

Observe que essa expressão é diferente de $\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i \times$, que representa o **PRODUTO DAS SOMAS** dessas duas variáveis. No **PRODUTO DAS SOMAS**, primeiro somamos toda a sequência x, depois toda a sequência y e, em seguida, multiplicamos o resultado das somas:

$$\sum_{i=1}^{10} x_i \times \sum_{i=1}^{10} y_i = (1 + 2 + 3 + \cdots + 10) \times (3 + 6 + 9 + \cdots + 30)$$

$$\sum_{i=1}^{10} x_i \times \sum_{i=1}^{10} y_i = 55 \times 165 = 9075$$

Dessa forma, temos que:

SOMA DOS PRODUTOS $\neq$ PRODUTO DAS SOMAS

Também podemos utilizar a notação para representar o QUADRADO DA SOMA dos valores de x, com i iniciando em 1 e terminando em 10. No QUADRADO DA SOMA, somamos toda a sequência e elevamos o resultado ao quadrado:

$$\left(\sum_{i=1}^{n} x_i \right)^2 = (x_1 + x_2 + x_3 + \cdots + x_n)^2$$

$$\left(\sum_{i=1}^{n} x_i \right)^2 = (1 + 2 + 3 + \cdots + 10)^2$$

$$\left(\sum_{i=1}^{n} x_i \right)^2 = (55)^2 = 3025$$

Veja que essa expressão é diferente de $\sum_{i=n}^{n} x_i^2$, que representa a **SOMA DOS QUADRADOS**. Na **SOMA DOS QUADRADOS**, cada elemento da sequência é elevado ao quadrado e depois os resultados são somados:

$$\sum_{i=1}^{n} x_i^2 = 1^2 + 2^2 + 3^2 + \cdots + 10^2 = 385$$

Por fim, ainda podemos representar o somatório de uma constante k. Digamos que essa constante tenha valor igual a 3:

$$\sum_{i=1}^{n} k = k + k + k + \cdots + k = k \times n$$
$$\sum_{i=1}^{n} 3 = 3 + 3 + 3 + \cdots + 3 = 3 \times n$$

2.3.1 PROPRIEDADES DO SOMATÓRIO

As propriedades apresentadas nesta seção facilitam o desenvolvimento de expressões algébricas com a notação de somatório.

1. O somatório de uma constante k é igual ao produto do número de termos pela constante.

$$\sum_{i=1}^{n} k = k + k + k + \cdots + k = k \times n$$

Para demonstrar essa propriedade, consideraremos que cada constante está multiplicada pelo valor um, isto é:

$$\sum_{i=1}^{n} k = k + k \times 1 + k \times 1 + \cdots + k \times 1$$

Agora, colocaremos os novos valores em evidência:

$$\sum_{i=1}^{n} k = k \underbrace{(1 + 1 + 1 + \cdots + 1)}_{n \text{ termos}} = k \times n$$

Considere uma lista composta por noventa e nove elementos repetidos e iguais a 9:

$$\underbrace{\{9, 9, 9, 9, \ldots, 9\}}_{\text{99 termos repetidos}}$$

O somatório dos elementos dessa lista será:

$$\sum_{i=1}^{99} 9 = \underbrace{(9 + 9 + 9 + \cdots + 9)}_{n \text{ termos repetidos}} = 9 \times \underbrace{(9 + 9 + 9 + \cdots + 9)}_{n \text{ termos repetidos}} = 9 \times 99 = 891$$

2. O somatório do produto de uma constante por uma variável é igual ao produto da constante pelo somatório da variável.

Para demonstrarmos essa propriedade, colocaremos em evidência cada constante k:

$$\sum_{i=1}^{n} k \times x_i = k \times x_1 + k \times x_2 + k \times x_3 + \cdots + k \times x_n = k \times \sum_{n=1}^{n} x_i$$

Para demonstrarmos essa propriedade, colocaremos em evidência cada constante k:

$$\sum_{i=1}^{n} k \times x_i = k \times x_1 + k \times x_2 + k \times x_3 + \cdots + k \times x_n = k \times (x_1 + x_2 + x_3 + \cdots + x_n)$$

Já sabemos que $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$. Logo:

$$\sum_{i=1}^{n} k \times x_i = k \times \sum_{i=1}^{n} x_i$$

Portanto, a constante pode sair de dentro do somatório, passando a multiplicá-lo.

2.4 MÉDIA ARITMÉTICA SIMPLES

A média aritmética simples está muito presente em nosso cotidiano, seja no consumo médio de combustível, na temperatura média ou na renda per capita. Essa medida é definida como o **QUOCIENTE** entre a **SOMA DE TODOS OS ELEMENTOS** e o **NÚMERO DELES**. A propriedade principal da média é preservar a soma dos elementos de um conjunto de dados.

Podemos adotar o seguinte raciocínio para encontrarmos a fórmula da média aritmética. Dada uma lista de n números, $x_1, x_2, \ldots, x_n$ a soma de seus termos é igual a:

$$\overbrace{x_1 + x_2 + x_3 + \cdots + x_n}^{\text{n fatores}}$$

A média aritmética dessa lista é um número , tal que, se todos os elementos forem substituídos por $\bar{x}$, a soma da lista permanecerá preservada. Assim, substituindo todos os elementos por $\bar{x}, \bar{x}, \ldots, \bar{x}$, teremos uma nova lista, cuja soma é:

$$\overbrace{\bar{x} + \bar{x} + \cdots + \bar{x}}^{\text{n fatores}} = n \times \bar{x}$$

Como as somas das duas listas são iguais, temos:

$$n \times \bar{x} = \bar{x_1} + \bar{x_2} + \cdots + \bar{x_n}$$

Portanto, a média aritmética é:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

2.4.1 PROPRIEDADES DA MÉDIA ARITMÉTICA

Nessa seção, vamos estudar algumas propriedades importantes sobre a média aritmética.

Propriedade 2.4.1. Dado um conjunto com $n \geq 1$ elementos, a média aritmética sempre existirá e será única.

Desde que o conjunto tenha pelo menos um elemento, podemos afirmar que a média aritmética sempre existe, pois sempre conseguiremos calcular o quociente entre a soma dos elementos e o número deles. Além disso, como o somatório dos elementos resulta em um único número, o valor da média também sempre será único.

Propriedade 2.4.2. A média aritmética $\bar{x}$ de um conjunto de dados satisfaz a expressão $m \leq \bar{x} \leq M$, em que m e M são, respectivamente, os elementos que representam o valor mínimo e o valor máximo desse conjunto.

$$\textbf{Mínimo} \leq \bar{x} \leq \textbf{Máximo}$$

Essa propriedade diz respeito ao fato da média aritmética sempre se encontrar entre os números mínimo e máximo de um conjunto.

Propriedade 2.4.3. Somando-se (ou subtraindo-se) uma constante c de todos os valores de uma variável, a média do conjunto fica aumentada (ou diminuída) dessa constante.

Propriedade 2.4.4. Multiplicando-se (ou dividindo-se) uma constante c de todos os valores de uma variável, a média do conjunto fica multiplicada (ou dividida) por esta constante.

$$\bar{y} = \bar{x} \times c \quad \text{ou} \quad \bar{y} = \bar{x} \div c$$

Portanto, não importa qual a sequência de números, a soma dos desvios em relação à média é sempre igual a zero.

Propriedade 2.4.5. A soma algébrica dos desvios em relação à média é nula.

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Propriedade 2.4.6. A soma dos quadrados dos desvios da sequência de números x_1, em relação a um número a, é mínima se a for a média aritmética dos números.

$$\sum_{i=1}^{n} (x_i - a)^2 \geq \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Essa propriedade afirma que, caso os desvios sejam calculados com relação a um número diferente da média, e os resultados de tais desvios sejam elevados ao quadrado e somados, teremos um número necessariamente maior do que obteríamos caso a mesma operação fosse realizada utilizando-se a média.

2.5 MÉDIA PONDERADA

Muitas vezes, certos elementos de um conjunto de dados possuem relevância maior que os demais. Nessa situação, para calcular a média de tais conjuntos, devemos encontrar uma média ponderada. Uma média ponderada é a média de um conjunto de dados cujos valores possuem pesos variados.

Ela é calculada pela igualdade a seguir, em que p é o peso de cada valor de x:

$$\bar{x} = \frac{\sum_{i=1}^{n} (x_i \times p_1)}{\sum_{i=1}^{n} p_i}$$

Observe que no numerador cada valor será multiplicado pelo seu respectivo peso, enquanto no denominador teremos a soma de todos os pesos.

Suponha que um candidato tenha prestado um concurso público para o cargo de Auditor Fiscal, alcançando as seguintes notas:

Disciplina	Nota x_i
Língua Portuguesa	4,0
Direito Administrativo	4,0
Direito Constitucional	4,0
Direito Tributário	7,0
Legislação Tributária	7,0
Contabilidade	8,0
Auditoria	8,0

Tabela 2.2: Notas.

Considere, também, que o edital desse concurso previa que algumas disciplinas teriam importância maior do que outras, por isso foram atribuídos pesos diferentes às várias disciplinas. Digamos que os pesos tenham sido distribuídos da seguinte forma:

Disciplina	Peso p_i
Língua Portuguesa	1
Direito Administrativo	2
Direito Constitucional	2
Direito Tributário	3
Legislação Tributária	3
Contabilidade	3
Auditoria	4

Tabela 2.3: Pesos.

Agora, admita que o candidato deveria alcançar uma nota 7,0 ou superior na prova objetiva para que fosse convocado para a etapa discursiva. Se você fosse um dos avaliadores desse concurso, você consideraria o candidato aprovado na prova objetiva?

Para responder a esse questionamento, devemos calcular a média aritmética ponderada desse candidato, considerando os pesos de cada disciplina. Dessa forma, devemos multiplicar cada nota pelo seu respectivo peso, somar esses produtos e dividir pela soma dos pesos.

Disciplina	Notas x_i	Pesos p_i	$x_i \times p_i$
Língua Portuguesa	4,0	1	$4,0 \times 1$=4,0
Direito Administrativo	4,0	2	$4,0 \times 2$=8,0
Direito Constitucional	4,0	2	$4,0 \times 2$=8,0
Direito Tributário	6,0	3	$7,0 \times 3$=18,0
Legislação Tributária	6,0	3	$7,0 \times 3$=18,0
Contabilidade	7,0	3	$8,0 \times 3$=21,0
Auditoria	7,0	3	$8,0 \times 3$=21,0

Tabela 2.4: Notas.

Nesse ponto, temos uma lista contendo todos os produtos de notas e pesos. Então, a média aritmética ponderada é dada por:

$$\bar{x} = \frac{\sum_{i=1}^{n}(x_i \times p_1)}{\sum_{i=1}^{n}} = \frac{4,0+8,0+8,0+18,0+18,0+21,0+21,0}{1+2+2+3+3+3+3} = \frac{98}{17} = 5,76$$

2.6 MÉDIA PARA DADOS AGRUPADOS

Em estatística, os dados podem ser definidos como informações que representam os atributos qualitativos ou quantitativos de uma variável, ou de um conjunto de variáveis. Esses dados podem ser classificados em agrupados e não-agrupados. Normalmente, logo após a etapa de coleta, temos dados não-agrupados ou dados brutos.

Por exemplo, suponha que o Estratégia Concursos esteja realizando um experimento com um grupo de dez alunos, para mensurar o tempo médio de resposta a uma questão de estatística. Logo após a coleta, os dados continuam brutos, pois não passaram por nenhuma análise nem foram agrupados de alguma forma. Então, teríamos uma tabela similar a seguinte:

Aluno	1	2	3	4	5	6	7	8	9	10
Tempo Médio	2	3	5	7	9	6	6	5	3	9

Tabela 2.5: Tempos.

Por sua vez, os dados agrupados são aqueles que passaram por algum nível de análise, o que significa que já não são brutos. Os dados agrupados podem ser organizados por frequência de um determinado valor ou por intervalos de classes. Quando por frequência de valor, os dados são organizados de forma ascendente e suas ocorrências são contabilizadas:

Tempo Médio X_i	Frequência f_i
1	1
3	2
5	2
6	2
7	1
9	2

Tabela 2.6: Frequências.

Quando por intervalos de classes, os dados também são organizados de forma ascendente, porém, em classes preestabelecidas, e as ocorrências de cada classe são contabilizadas:

Tempo Médio X_i	Frequência f_i
$0 \leq x \leq 2$	1
$2 \leq x \leq 4$	2
$4 \leq x \leq 6$	2
$6 \leq x \leq 8$	3
$8 \leq x \leq 10$	2

Tabela 2.7: Frequências.

Para dados agrupados e apresentados como diagramas ou tabelas, a definição da média permanece inalterada, então, tudo o que estudamos até o momento permanece válido, mas teremos métodos específicos para obtenção da média. A seguir, veremos como proceder em cada caso.

2.6.1 MÉDIA PARA DADOS AGRUPADOS POR VALOR

Dando continuidade ao nosso exemplo, vamos a calcular a média aritmética de dados agrupados por valor. Os dados foram organizados na tabela a seguir:

Tempo Médio X_i	Frequência f_i
1	1
3	2
5	2
6	2
7	1
9	2

Tabela 2.8: Frequências.

Como podemos interpretar essa tabela? Basta você saber que as frequências refletem o número de repetições de cada valor da nossa variável tempo médio. Isto é, um aluno conseguiu responder à questão em 1 minuto, dois alunos conseguiram em 3 minutos, dois alunos conseguiram em 5 minutos, e assim sucessivamente.

Para calcularmos a média a partir de uma tabela de frequências como esta, devemos utilizar a seguinte fórmula:

$$\bar{x} = \frac{\sum_{i=1}^{n}(X_i \times f_i)}{\sum_{i=1}^{n} f_i}$$

A aplicação dessa fórmula é bem simples. O raciocínio é o mesmo adotado para a média ponderada, sendo que, agora, o peso é representado pela frequência. Desse modo, vamos multiplicar cada valor por sua respectiva frequência, somar tudo e dividir pela soma das frequências:

Tempo Médio X_i	Frequência f_i	$f_i \times X_i$
1	1	$1 \times 1 = 1$
3	2	$2 \times 3 = 2$
5	2	$2 \times 5 = 10$
6	2	$2 \times 6 = 12$
7	1	$1 \times 7 = 7$
9	2	$2 \times 9 = 18$

Tabela 2.9: $f_i \times X_i$

Após isso, somaremos todos os valores da coluna $X_i \times f_i$, obtendo o termo $\sum_{i=1}^{n} X_i \times f_i$, e também somaremos os termos da coluna f_1 , obtendo o termo $\sum_{i=1}^{n} f_i$. Veja a última linha da tabela:

Tempo Médio X_i	Frequência f_i	$f_i \times X_i$
1	1	$1 \times 1 = 1$
3	2	$2 \times 3 = 2$
5	2	$2 \times 5 = 10$
6	2	$2 \times 6 = 12$
7	1	$1 \times 7 = 7$
9	2	$2 \times 9 = 18$
Totais	10	54

Tabela 2.10: $f_i \times X_i$

Agora, basta dividirmos um valor pelo outro, obtendo:

$$\bar{x} = \frac{\sum_{i=1}^{n}(X_i \times f_i)}{\sum_{i=1}^{n} f_i} = \frac{54}{10} = 5,4$$

Portanto, a média dos dados apresentados na tabela é 5,4.

2.6.2 MÉDIA PARA DADOS AGRUPADOS POR CLASSE

Retomando nosso exemplo, vamos calcular a média aritmética de dados agrupados por classe. Os dados foram organizados na tabela a seguir:

Tempo Médio X_i	Frequência f_i
$0 \leq x \leq 2$	1
$2 \leq x \leq 4$	2
$4 \leq x \leq 6$	2
$6 \leq x \leq 8$	3
$8 \leq x \leq 10$	2

Tabela 2.11: Separação por classe.

Como podemos interpretar essa tabela? Basta sabermos que as frequências refletem o número de ocorrências em cada um dos intervalos definidos para a variável tempo médio. Isto é, um aluno respondeu à questão com tempo médio abaixo de 2 minutos, dois responderam com tempo médio entre 2 e 4 minutos, dois com tempo médio entre 4 e 6 minutos, e assim sucessivamente.

Ao agruparmos os dados em classes, precisaremos fazer uma modificação em relação ao cálculo anterior: substituir os intervalos pelos seus respectivos pontos médios. Como assim? Ao invés de considerarmos o intervalo de 0 a 2 minutos, por exemplo, substituiremos pelo valor de 1 minuto.

Em nosso exemplo, a identificação dos pontos médios é relativamente fácil. Mas é possível que você encontre situações em que isso não seja tão trivial. Como fazer nesses casos? Devemos calcular a média dos dois extremos do intervalo. Assim, o ponto médio PM é calculado pela seguinte expressão:

$$PM = \frac{l_{sup} + l_{inf}}{2}$$

em que l_inf e l_sup são, respectivamente, os limites inferior e superior do intervalo considerado.

Na tabela abaixo, repare que foi incluída uma nova coluna para o cálculo dos pontos médios:

Tempo Médio X_i	Ponto Médio PM_i	Frequência f_i
$0 \leq x \leq 2$	1	1
$2 \leq x \leq 4$	3	2
$4 \leq x \leq 6$	5	2
$6 \leq x \leq 8$	7	3
$8 \leq x \leq 10$	9	2

Tabela 2.12: Separação por classe.

O próximo passo consiste em calcular os valores das multiplicações $PM_i \times f_i$ multiplicando essas duas colunas. Vamos ver:

Tempo Médio X_i	Ponto Médio PM_i	Frequência f_i	$PM_i \times f_i$
$0 \leq x \leq 2$	1	1	$1 \times 1 = 1$
$2 \leq x \leq 4$	3	2	$3 \times 2 = 6$
$4 \leq x \leq 6$	5	2	$5 \times 2 = 10$
$6 \leq x \leq 8$	7	3	$7 \times 3 = 21$
$8 \leq x \leq 10$	9	2	$9 \times 2 = 18$

Tabela 2.13: Separação por classe.

Após isso, somaremos todos os valores da coluna $PM_i \times f_i$, obtendo o termo $\sum_{i=1}^{n} PM_i \times f_i$, e também somaremos os termos da coluna f_i, obtendo o termo $\sum_{i=1}^{n} f_i$. Veja a última linha da tabela:

Tempo Médio X_i	Ponto Médio PM_i	Frequência f_i	$PM_i \times f_i$
$0 \leq x \leq 2$	1	1	$1 \times 1 = 1$
$2 \leq x \leq 4$	3	2	$3 \times 2 = 6$
$4 \leq x \leq 6$	5	2	$5 \times 2 = 10$
$6 \leq x \leq 8$	7	3	$7 \times 3 = 21$
$8 \leq x \leq 10$	9	2	$9 \times 2 = 18$
Total		10	56

Tabela 2.14: Separação por classe.

Agora, basta dividirmos um valor pelo outro, obtendo:

$$\bar{x} = \frac{\sum_{i=1}^{n}(PM_i \times f_i)}{\sum_{i=1}^{n} f_i} = \frac{56}{10} = 5,6$$

Finalmente, note que a média de dados agrupados por classe (5,6) foi diferente da média de dados agrupados por valor (5,4). Por que isso ocorreu? Isso ocorreu porque, ao agruparmos os valores da variável em classes, perdemos detalhes que eram relevantes para o cálculo exato da média, embora a forma de apresentação tenha sido simplificada.

2.7 MÉDIA GEOMÉTRICA

2.8 MÉDIA HARMÔNICA

A média harmônica é muito utilizada quando precisamos trabalhar com grandezas inversamente proporcionais. É o caso de problemas clássicos, como o cálculo da velocidade média de um automóvel ou da vazão de torneiras (quanto tempo duas ou mais torneiras levam para encher um tanque).

Essa medida é definida, para o conjunto de números positivos, como o inverso da média aritmética dos inversos. A propriedade principal dessa média é preservar a soma dos inversos dos elementos de um conjunto de números.

O raciocínio para encontrarmos a fórmula da média harmônica é similar ao adotado para as médias aritmética e geométrica. Dada uma lista de *x* números, $x_1, x_2, \ldots, x_n$, a soma dos inversos de seus termos é igual a:

$$\underbrace{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}_{\text{n fatores}}$$

A média harmônica dessa lista é um número *H*, tal que, se todos os elementos forem substituídos por *H*, a soma dos inversos permanecerá preservada. Assim, substituindo todos os elementos por H, teremos uma lista, $H, H, \ldots, H$, cuja soma dos inversos é:

$$\underbrace{\frac{1}{H} + \frac{1}{H} + \cdots + \frac{1}{H}}_{\text{n fatores}} = \frac{n}{H}$$

Como as somas dos inversos das duas listas são iguais, temos:

$$\frac{n}{H} = \frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{n}$$

$$n = H \times \left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{n} \right)$$

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{n}}$$

Como vimos no início, muitas vezes, a média harmônica é descrita como o inverso da média aritmética dos inversos. Isso porque a fórmula acima também pode ser escrita na forma mostrada a seguir, em que $(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{n})/n$ corresponde à média aritmética dos inversos.

2.9 DESIGUALDADE DAS MÉDIAS

Dada uma lista de .. números positivos, $x_1, x_2, \ldots, x_n$, podemos afirmar que:

$$\boxed{\bar{x} \geq G \geq H}$$

em que $\bar{x}$ é a média aritmética; G é a média geométrica e H é a média harmônica. Significa dizer que a média aritmética será sempre maior ou igual à média geométrica que, por sua vez, será sempre maior ou igual à média harmônica. A igualdade ocorrerá quando os números da lista forem todos iguais.

Tomemos como exemplo os números 4, 12 e 20. Como sabemos, a média aritmética será:

$$\bar{x} = \frac{4 + 12 + 20}{3} = 12$$

A média geométrica é

$$G = \sqrt[3]{4 \times 12 \times 20} \approx 9,86$$

E a média harmônica é:

$$H = \frac{3}{\frac{1}{4} + \frac{1}{12} + \frac{1}{20}} \approx 2,61$$

Obtivemos, portanto, uma média aritmética $\bar{x} = 12$ maior que a média geométrica $G = 9,86$ que, por sua vez, é maior que a média harmônica $H = 2,61$.

Agora, analisaremos um caso em que as três médias são iguais: considere uma lista composta pelos números 5, 5 e 5. Nesse caso, temos que $\bar{x}, GeH$ são, respectivamente:

$$\bar{x} = \frac{5 + 5 + 5}{3} = 5$$

A média geométrica é

$$G = \sqrt[3]{5 \times 5 \times 5} = 5$$

E a média harmônica é:

$$H = \frac{3}{\frac{1}{5} + \frac{1}{5} + \frac{1}{5}} = 5$$

Portanto, quando todos os números da lista são iguais, as média aritmética $\bar{x}$, geométrica G e harmônica H também são iguais.

2.10 MEDIANA

2.10.1 MEDIDAS SEPARATRIZES

As separatrizes são medidas que dividem (ou separam) uma série ordenada em duas ou mais partes, cada uma contendo a mesma quantidade de elementos. Nesse caso, o nome da medida separatriz é definido conforme a quantidade de partes onde a série é dividida:

- mediana: divide uma série ordenada em duas partes iguais, cada uma contendo 50% dos valores da sequência;
- quartis: dividem uma série ordenada em quatro partes iguais, cada uma contendo 25% dos valores da sequência;
- quintis: dividem uma série ordenada em cinco partes iguais, cada uma contendo 20% dos valores da sequência;
- decis: dividem uma série ordenada em dez partes iguais, cada uma contendo 10% dos valores da sequência; e
- percentis: dividem uma série ordenada em cem partes iguais, cada uma contendo 1% dos valores da sequência.

Ao longo da aula, vamos estudar a mediana, os quartis, os decis e os percentis. Os quintis, por não serem tão explorados em provas de concurso, não serão abordados.

2.10.2 MEDIANA

A mediana é, simultaneamente, uma MEDIDA DE POSIÇÃO, de TENDÊNCIA CENTRAL e SEPARATRIZ. Ela caracteriza a posição central de uma série de valores. Além disso, também separa uma série de valores em duas partes de tamanhos iguais, cada uma contendo o mesmo número de elementos. Muitas vezes, a mediana é designada como valor mediano, sendo representada pelos símbolos M_d ou, em menor frequência, $\tilde{x}$

Mediana para Dados não Agrupados

O método para determinação da mediana envolve a realização de uma etapa anterior, que consiste na ordenação do conjunto de dados. Feito isso, a mediana é o elemento que ocupa a POSIÇÃO CENTRAL de uma série de observações ORDENADA segundo suas grandezas (isto é, dados brutos organizados em rol crescente ou decrescente).

Por exemplo, vamos determinar a mediana da seguinte série de valores:

$$\{3, 17, 13, 19, 2, 5, 7, 1, 8, 21, 9\}$$

Conforme a definição da mediana, a primeira etapa consiste na ordenação (crescente ou decrescente) da série de valores. Desse modo, obtemos:

$$\{1, 2, 3, 5, 7, 8, 9, 13, 17, 19, 21\}$$

Agora, determinaremos o elemento que ocupa a posição central desse conjunto de dados. Para isso, devemos encontrar o termo que possui o mesmo número de elementos tanto à sua esquerda quanto à sua direita. Em nosso exemplo, esse valor é o 8, pois existem cinco elementos antes dele e cinco após ele.

$$\underbrace{1, 2, 3, 5, 7}_{\text{5 elementos antes}}, \quad \underbrace{8}_{\text{Elemento central}} \quad , \underbrace{9, 13, 17, 19, 21}_{\text{5 elementos depois}}$$

É importante notarmos que essa série possui um número ímpar de elementos. Quando isso acontece, isto é, quando uma série possui um NÚMERO ÍMPAR de elementos, a MEDIANA SEMPRE COINCIDE com o ELEMENTO CENTRAL do conjunto de dados. Portanto, temos:

$$M_d = 8$$

Contudo, se porventura a série tivesse um número par de elementos, POR CONVENÇÃO, a MEDIANA seria a MÉDIA ARITMÉTICA dos dois termos centrais. Assim, caso dicionássemos o número 23 ao conjunto de dados apresentado anteriormente, teríamos a seguinte situação:

$$\underbrace{1, 2, 3, 5, 7}_{\text{5 elementos antes}}, \quad \underbrace{8, 9}_{\text{Elemento central}} \quad , \underbrace{13, 17, 19, 21, 23}_{\text{5 elementos depois}}$$

Nesse caso, em que temos um número par de elementos, a mediana é definida como a média aritmética dos termos centrais, sendo os números 8 e 9. Assim,

$$M_d = \frac{8 + 9}{2} = 8,5$$

Note que, quando o número é ímpar, o termo central sempre ocupa a posição $\frac{n+1}{2}$. Por outro lado, quando o número de termos é par, existem dois termos centrais, sendo que o primeiro ocupa a posição $\frac{n}{2}$; e o segundo ocupa a posição imediatamente seguinte, ou seja, $\frac{n}{2} + 1$.

Essas relações são importantes porque nem sempre conseguiremos identificar a posição central tão facilmente. Por exemplo, se tivermos uma série composta por 501 elementos, podemos afirmar que o termo central será o elemento ocupando a posição $\frac{n+1}{2} = \frac{501+1}{2} = 251$, sem precisar recorrer a qualquer outro método. Logo, a mediana terá o mesmo valor do termo central:

$$M_d = x_{251}$$

Vejamos a disposição do termo central em relação aos demais elementos da série:

$$\underbrace{x_1, x_2, \ldots, x_{250}}_{\text{250 elementos antes}}, \quad \underbrace{x_{251}}_{\text{Termo central}} \quad , \underbrace{x_{252}, x_{253}, \ldots, x_{501}}_{\text{250 elementos depois}}$$

Agora, se tivermos uma série composta por 500 elementos, os termos centrais serão os elementos ocupando as posições:

$$\frac{n}{2} = \frac{500}{2} = 250 \text{ e } \frac{n}{2} + 1 = \frac{500}{2} + 1 = 251$$

Vejamos a disposição dos termos centrais em relação aos demais elementos da série:

$$\underbrace{x_1, x_2, \ldots, x_{249}}_{\text{250 elementos antes}}, \underbrace{x_{250}, x_{251}}_{\text{Termo central}}, \underbrace{x_{252}, x_{253}, \ldots, x_{500}}_{\text{250 elementos depois}}$$

Nessa situação, por convenção, a mediana será a média aritmética entre os termos centrais,

$$M_d = \frac{x_{250} + x_{251}}{2},$$

Portanto, podemos estabelecer que a mediana de um conjunto composto por $n$elementos ordenados de forma crescente ou decrescente será:

a. se n for impar, o termo de ordem $\frac{n+1}{2}$, isto é:

$$M_d = \frac{x_n + 1}{2}$$

b. se n for par, a média aritmética dos termos de ordem $\frac{n}{2}$ e $\frac{n}{2} + 1$, isto é:

$$M_d = \frac{\frac{x_n}{2} + \frac{x_n}{2} + 1}{2}$$

Como vimos, a mediana depende somente do termo que ocupa a posição central em um conjunto de dados, e não dos valores de todos os elementos que compõem a série. Por isso, dizemos que a mediana não sofre tanta influência pela presença de valores extremos/discrepantes quanto à média. Essa é, inclusive, uma das principais diferenças entre essas duas medidas.

Podemos constatar essa propriedade da mediana por meio de um exemplo. Considere que tenhamos inicialmente a seguinte série:

$$\{1, 2, 4, 6, 7, 9, 10, 11, 13\}$$

A média aritmética desses valores é:

$$\bar{x} = \frac{1 + 2 + 4 + 6 + 7 + 9 + 10 + 11 - 13}{9} = \frac{63}{9} = 7$$

Como o número de elementos é ímpar, n= 9, a mediana será o elemento que ocupa a posição:

$$\frac{n}{2} + 1 = \frac{9}{2} + 1 = \frac{10}{2} = 5$$

$$M_d = x_5 \Rightarrow M_d = 7$$

Agora, considere que o elemento de valor 13 tenha sido alterado para 130.000. Veja o que acontece com a média aritmética desse conjunto:

$$\bar{x} = \frac{1 + 2 + 4 + 6 + 7 + 9 + 10 + 11 - 130.000}{9} = 14.450$$

Mas como o número de elementos é o mesmo assim como o elemento central, a mediada continua sendo 7.

Portanto, a alteração no valor de um único elemento do conjunto de dados causou um impacto significativo na média, ao passo que a mediana permaneceu inalterada. Por isso, dizemos que a média é mais influenciada pela presença de valores extremos que a mediana.

Mediana para dados Agrupados sem Intervalos de Classe

O raciocínio adotado no cálculo da mediana para dados agrupados por valor (sem intervalos de classe) é similar ao empregado no caso dos dados não-agrupados. Basicamente, teremos que encontrar um valor que dividirá a distribuição de frequências em duas partes contendo o mesmo número de elementos.

Considere a seguinte situação hipotética: uma empresa realizou uma pesquisa para medir o nível de satisfação dos clientes com relação ao seu atendimento. Os clientes puderam atribuir notas de 0 a 5 no que diz respeito ao nível de satisfação, resultando na seguinte distribuição de frequências:

Nível de Satisfação X_i	Frequência f_i
0	5
1	5
2	8
3	10
4	13
5	10

Tabela 2.15: Frequência de níveis de Satisfação

O total de clientes entrevistados foi de:

$$5 + 5 + 8 + 10 + 13 + 10 = 49$$

Como o número de entrevistados é ímpar, n = 49, a mediana será o termo que ocupa a posição de ordem:

$$\frac{n}{2} + 1 = \frac{49}{2} + 1 = \frac{50}{2} = 25$$

Em outros termos, a mediana será o elemento que ocupa a vigésima quinta posição. Para chegarmos a esse elemento, precisamos percorrer cada um dos níveis de satisfação. Reparem que três clientes atribuíram a nota 0 (zero); cinco atribuíram a nota 1 (um); e oito atribuíram a nota 2 (dois). Portanto, até esse ponto, temos um total de 16 avaliações:

$$3 + 5 + 8 = 16$$

Nivel de Satisfação X_i	Frequênia f_i	Frequência Acumulada	Memória de Cálculo
0	3	3	0+3=3
1	5	8	3+5=8
2	8	16	8+8=16
3	10	26	16+10=26
4	13	39	26+13=39
5	10	49	39+10=49

Tabela 2.16: Frequência Acumulada

Nivel de Satisfação X_i	Frequênia f_i	Frequência Acumulada
0	3	3
1	5	8
2	8	16
3	10	26
4	13	39
5	10	49

Tabela 2.17: Frequência Acumulada

Vejam que ainda não chegamos na posição desejada, isto é, na vigésima quinta. Contudo, sabemos que o próximo nível de satisfação, referente à nota 3 (três), teve frequência absoluta igual a 10. Se somarmos essas dez novas avaliações com o total obtido anteriormente, chegaremos a um valor que ultrapassa a posição procurada (16 + 10 = 26). Assim, descobrimos que a mediana está localizada nessa faixa de avaliação. Portanto,

$$M_d = x_2 5 = 3$$

Esse procedimento pode ser simplificado com a introdução de uma coluna adicional para armazenar as frequências acumuladas. Já vimos que, para calcularmos a frequência acumulada, devemos repetir a primeira frequência e somar as frequências subsequentes, exibindo os resultados a cada linha. Observem:

Vamos remover a memória de cálculo para simplificar a tabela:

Reparem que o número 16, na terceira linha da coluna de frequências acumuladas, representa a soma das frequências absolutas simples das três primeiras linhas, isto é, 3 + 5 + 8. Assim, concluímos que 16 clientes avaliaram o atendimento da empresa com nota igual ou inferior a 2. De forma análoga, como 49 clientes participaram da pesquisa, podemos afirmar que 33 avaliaram o atendimento com nota igual ou superior a 3.

Observem que a introdução da coluna de frequências acumuladas permite calcularmos a mediana de forma praticamente imediata. Nesse sentido, se n for ímpar, basta identificarmos o valor da variável correspondente à primeira frequência acumulada imediatamente igual ou superior à posição de ordem $\frac{n+1}{2}$; e, se n for par, basta identificarmos os dois valores correspondentes às frequências acumuladas imediatamente iguais ou superiores às posições de ordens $\frac{n}{2}$ e $\frac{n+1}{2}$, respectivamente, e tirarmos a média aritmética desses dois valores.

Em nosso exemplo, como a frequência total é ímpar, teremos que buscar pela posição $\frac{n+1}{2} = \frac{49+1}{2} = 25$A mediana será o valor da variável correspondente à primeira frequência acumulada

Nivel de Satisfação X_i	Frequênia f_i	Frequência Acumulada
0	3	3
1	5	8
2	8	16
3	10	$26 \geq 25$
4	13	39
5	10	49

Tabela 2.18: Frequência Acumulada

maior ou igual a essa posição, portanto, $M_d = 3$. Vejamos:

Assim, podemos estabelecer que a mediana de uma tabela de frequências composta por um total de n elementos será:

b. se n for ímpar, o valor identificado na tabela correspondente à frequência acumulada imediatamente igual ou superior à posição de ordem $\frac{n+1}{2}$, isto é,

$$M - d = \frac{X_n + 1}{2}$$

b. se n for par, a média aritmética dos valores identificados na tabela correspondentes às frequências acumuladas imediatamente iguais ou superiores às posições de ordens $\frac{n}{2}$ e $\frac{n}{2} + 1$, isto é,

$$M_d = \frac{\frac{x_n}{2} + \frac{x_n}{2} + 1}{2}$$

Mediana para dados agrupados em classes

O raciocínio adotado no cálculo da mediana para dados agrupados em classes é muito similar ao empregado no tópico anterior. Agora, contudo, não nos importaremos com o número de elementos da série. Adotaremos um único procedimento de cálculo, independentemente de termos um número par ou ímpar de elementos.

Considere a distribuição de frequências descrita a seguir, que resume as idades de um grupo de 50 pessoas:

Idade	Frequência f_i
$23 \vdash 16$	3
$26 \vdash 29$	4
$29 \vdash 32$	10
$32 \vdash 35$	13
$35 \vdash 38$	10
$38 \vdash 41$	6
$41 \vdash 44$	4
Total	50

Tabela 2.19: Distribuição de Frequência de Idades

A etapa inicial do cálculo da mediana consiste na construção da coluna de frequências a partir da tabela das frequências acumuladas:

Idade	Frequência f_i	Frequência Acumulada f_{ac}
$23 \vdash 16$	3	3
$26 \vdash 29$	4	7
$29 \vdash 32$	10	17
$32 \vdash 35$	13	30
$35 \vdash 38$	10	40
$38 \vdash 41$	6	46
$41 \vdash 44$	4	50
Total	50	

Tabela 2.20: Cáculo da mediana

Para calcular a mediana de dados agrupados por intervalo de classes, precisamos identificar a lasse em que se encontra a mediana, a chamada classe mediana, que corresponde à frequência acumulada imediatamente igual ou superior à metade da frequência total, ou seja, metade da soma das frequências simples, $\sum \frac{f_i}{2}$. Em nosso exemplo, temos:

$$\frac{\sum f_i}{2} = \frac{50}{2} = 25$$

Agora, devemos comparar o valor encontrado com os valores presentes na coluna de frequências acumuladas, percorrendo-os de cima para baixo. A classe mediana será a primeira classe onde a frequência acumulada for igual ou superior a 25. Assim, teremos que analisar o seguinte:

- a primeira frequência acumulada (3) é maior ou igual a 25? Não;
- a segunda frequência acumulada (7) é maior ou igual a 25? Não;
- a terceira frequência acumulada (17) é maior ou igual a 25? Não;
- a quarta frequência acumulada (30) é maior ou igual a 25? Sim.

Pronto, encontramos a classe mediana. Nesse ponto, paramos a comparação e verificamos que a mediana se encontra na quarta classe, isto é, no intervalo entre 32 e 35.

Conhecendo a classe mediana, podemos aplicar a fórmula da mediana, a seguir:

$$M_d = l_{inf} + \left[\frac{\left(\frac{\sum f_i}{2} \right) - f_{ac_{ant}}}{f_i} \right] \times h$$

onde:

- l_{inf} é o limite inferior da classe mediana;
- $l_{fac_{ant}}$ é a frequência acumulada da classe anterior à classe mediana;
- l_i é a frequência simples da classe mediana; e
- h é a amplitude do intervalo da classe mediana.

Já sabemos que a amplitude é a diferença entre os limites da classe. Assim, temos:

$$h = 35 - 32 = 3$$

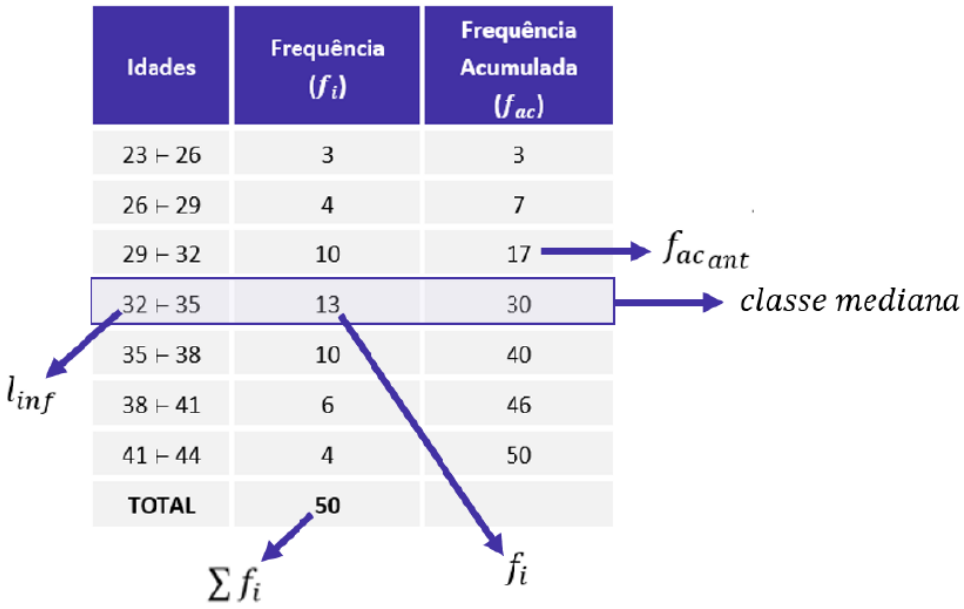


Figura 2.2: classe mediana

Após identificarmos os elementos, precisamos aplicá-los na fórmula evidenciada anteriormente:

$$M_d = l_{inf} + \left[\frac{\left(\frac{\sum f_i}{2}\right) - }{f_i}\right] \times h$$

$$M_d = 32 + \left[\frac{\left(\frac{50}{2}\right) - 17}{13}\right] \times 3$$

$$M_d = 32 + \frac{8}{13} \times 3 \approx 33,85$$

Propriedades da Mediana

Propriedade 2.10.1. Somando-se (ou subtraindo-se) uma constante *c* a todos os valores de uma variável, a mediana do conjunto fica aumentada (ou diminuída) dessa constante.

Propriedade 2.10.2. Multiplicando-se (ou dividindo-se) todos os valores de uma variável por uma constante *c*, a mediana do conjunto fica multiplicada (ou dividida) por esta constante.

Propriedade 2.10.3. A soma dos desvios absolutos de uma sequência de números, em relação a um número *a*, é mínima quando *a* é a mediana dos números.

2.11 QUARTIS, DECIS E PERCENTIS

Já vimos que a mediana separa uma série em duas partes iguais, cada uma contendo o mesmo número de elementos. Contudo, uma série também pode ser dividida em um número maior de partes, todas compostas por quantidades iguais de elementos. Nesse caso, o nome da medida separatriz é atribuído conforme a quantidade de partes onde a série é dividida:

- quartil: divide uma série em quatro partes iguais Q_1, Q_2, Q_3;

- quintil: divide uma série em cinco partes iguais $Qt_1, Qt_2, \ldots, Qt_5$;
- decil: divide uma série em dez partes iguais $D_1, D_2, \ldots, D_9$;
- percentil: divide uma série em cem partes iguais $P_1, P_2, \ldots, P_{99}$.

2.11.1 QUARTIS

Denominamos de quartis os valores de uma série que a dividem em quatro partes iguais, isto é, quatro partes contendo o mesmo número de elementos (25%). A imagem a seguir mostra os quartis de uma distribuição hipotética:

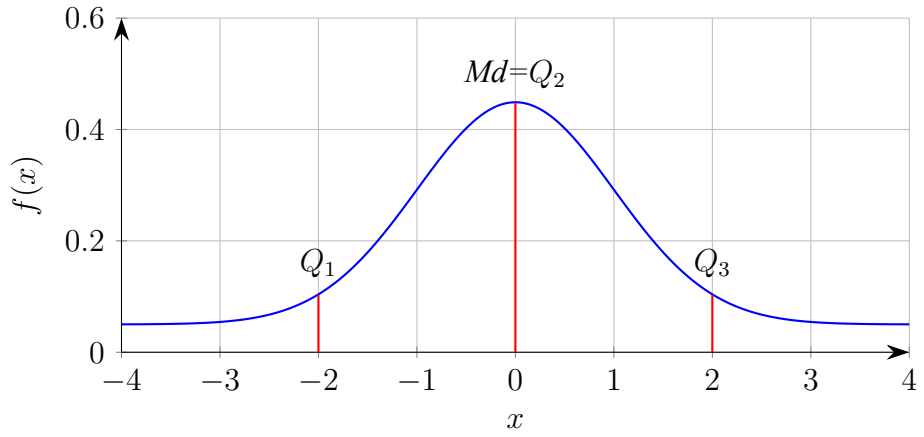


Figura 2.3: Quartis

Temos, então, 3 quartis Q_1 , Q_2 e Q_3 para dividir uma série em quatro partes iguais:

- ◇ o primeiro quartil corresponde à separação dos primeiros 25% de elementos da série;
- ◇ o segundo quartil corresponde à separação de metade dos elementos da série, coincidindo com a mediana $Q_2 = M_d$;
- ◇ o terceiro quartil corresponde à separação dos primeiros 75% de elementos da série, ou dos últimos 25% de elementos da série.

Para o cálculo dos quartis, empregaremos a mesma fórmula adotada no cálculo da mediana, somente substituindo a expressão $\frac{\sum f_i}{2}$ por $\frac{k \times \sum f_i}{4}$ em que k indica a ordem do quartil e assume valores inteiros no intervalo de 1 a 3.

Quartil para dados Não Agrupados

O cálculo do quartil para dados não-agrupados é realizado, aproximadamente, por meio das etapas descritas a seguir:

- **primeira etapa**: determinamos a posição do quartil, por meio da expressão:

$$P_{Q_k} = \frac{k \times n}{4} \quad (k = 1,\ 2,\ 3)$$

- **segunda etapa**: identificamos a posição do quartil na coluna de frequências acumuladas, isto é, a frequência acumulada imediatamente igual ou superior à posição do quartil;
- **terceira etapa**: verificamos o valor da variável que corresponde a essa posição.

Sempre que houver necessidade, teremos que incluir uma coluna de frequências acumuladas.

Embora fórmula anterior possa ser utilizada para o cálculo da posição de Q_2, por depender de uma aproximação, nem sempre o valor do segundo quartil resultará no valor convencionado para a mediana. Por isso, para o cálculo de Q_2, vamos adotar o procedimento utilizado para encontrar a mediana. Isto é, se o número de elementos for ímpar, Q_2 será representado pelo elemento que ocupar a posição central, $\frac{n}{2}$. Se o número de elementos do conjunto for par, Q_2 será representado pela média aritmética entre os elementos que ocuparem as posições centrais $\frac{n}{2}$ e $\frac{n}{2}+1$

$$Q_k = l_{inf_{Q_k}} + \left[\frac{\frac{k \times \sum f_i}{4} - f_{ac_{ant}}}{f_{Q_k}} \right] \times h_{Q_k}$$

onde:

- $l_{inf_{Q_k}}$ = limite inferior da classe do quartil considerado;
- $f_{ac_{ant}}$ = frequência acumulada da classe anterior à classe do quartil considerado;
- Q_k = amplitude do intervalo de classe do quartil considerado;
- f_{Q_k} = frequência simples da classe do quartil considerado.

2.11.2 DECIS

Denominamos de decis os valores de uma série que a dividem em dez partes iguais, isto é, dez partes contendo o mesmo número de elementos (10%). A imagem a seguir mostra os decis de uma distribuição hipotética:

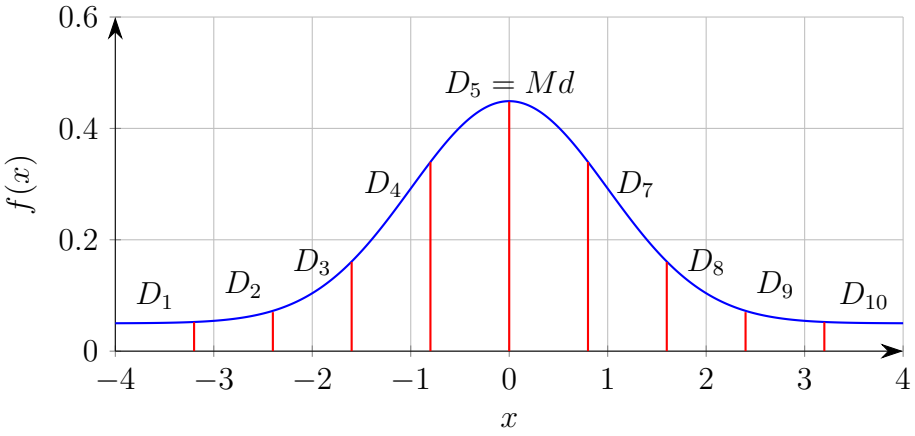


Figura 2.4: Decis

Temos, então, 9 decis $D_1, D_2, \ldots, D_9$ para dividir uma série em dez partes iguais:

- D_1: o primeiro decil corresponde à separação dos primeiros 10% de elementos da série;
- D_5: o quinto decil corresponde à separação de metade dos elementos da série, coincidindo com a mediana $D_5 = Md$;
- D_9: o nono decil corresponde à separação dos primeiros 90% de elementos da série, ou dos últimos 10% de elementos da série.

Para o cálculo dos decis, empregaremos a mesma fórmula adotada no cálculo da mediana, somente substituindo a expressão $\sum \frac{\Sigma f_i}{2}$ por $k \times \sum \frac{\Sigma f_i}{10}$, em que k indica a ordem do decil e assume valores inteiros no intervalo de 1 a 9.

Decis para dados Não Agrupados

O cálculo do decil segue o mesmo raciocínio empregado no cálculo do quartil para dados não-agrupados. A primeira tarefa que devemos realizar, se houver necessidade, é organizar o conjunto de valores por ordem de magnitude. Depois disso, procedemos conforme as seguintes etapas:

- **primeira etapa**: determinados a posição do decil conforme a seguinte fórmula:

$$P_{D_k} = \frac{k \times n}{10} \, (k = 1, 2, \ldots, 9);$$

- **segunda etapa**: identificamos a posição mais próxima ao rol;
- **terceira etapa**: verificamoso o valor que está ocupando essa posição.

Exemplo
Calcular os decís D_1 e D_8 para o seguinte conjunto de valores:

(5, 12, 15, 20, 2, 3, 4, 18, 10, 22)

Note que os valores não estão organizados. Portanto, nossa primeira tarefa será colocá-los em ordem de magnitude (rol):

(**2**, 3, 4, 5, 10, 12, 15, 18, 20, 22)

a. Cálculo de D_1:

$$P_{D_1} = \frac{1 \times 10}{10} = 1$$

Depois, identificamos a posição mais próxima no rol. Como o resultado foi um número inteiro, a posição mais próxima coincidirá com o valor encontrado, não havendo necessidade de aproximação.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
2	3	4	5	10	12	15	18	20	22

Tabela 2.21: Rol em ordem crescente

Portanto, o valor 2 corresponde a 10% do rol.

b. Calcular o decil D_8 O rol é o mesmo:

(2, 3, 4, 5, 10, 12, 15, 18, 20, 22)

Cálculo de D_1:

$$P_{D_8} = \frac{8 \times 10}{10} = 8$$

Depois, identificamos a posição mais próxima no rol. Como o resultado foi um número inteiro, a posição mais próxima coincidirá com o valor encontrado, não havendo necessidade de aproximação.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
2	3	4	5	10	12	15	**18**	20	22

Tabela 2.22: Rol em ordem crescente

Portanto, o valor 8 corresponde a 80% do rol.

Decil para dados agrupados sem intervalos de classe.

O cálculo do decil para dados agrupados sem intervalos de classe será realizado por meio das etapas descritas a seguir:

em que $\sum f_i$ é a soma das frequências simples;

- **primeira etapa**:etapa: determinamos a posição do decil, por meio da expressão:

$$P_{D_k} = \frac{k \times n}{10}\,(k = 1, 2, \ldots, 9);$$

- **segunda etapa**: identificamos a posição do decil na coluna de frequências acumuladas, isto é, a frequência acumulada imediatamente igual ou superior à posição do decil;
- **terceira etapa**: verificamos o valor da variável que corresponde a essa posição.

Sempre que houver necessidade, teremos que incluir uma coluna de frequências acumuladas.

Vamos calcular D_3 e D_8 da tabela de frequências a seguir, que representa a quantidade de filhos de um grupo de pessoas:

Filhos	Frequência f_i	Frequência Acumulada f_{ac}
0	18	18
1	35	$53(\geq 51)$
2	46	99
3	28	127
4	25	152
5	10	162
6	5	167
7	3	170
Total		170

Tabela 2.23: Cálculo de decis

a. Determinar a posição de D_3;

$$P_{D_3} = \frac{3 \times 170}{10} = 51$$

Portanto, a quantidade de 1 filho corresponde a 30% do rol.

b. Cálculo de D_8:

$$P_{D_8} = \frac{8 \times 170}{10} = 136$$

Filhos	Frequência f_i	Frequência Acumulada f_{ac}
0	18	18
1	35	53
2	46	99
3	28	127
4	25	$152 \geq 136$
5	10	162
6	5	167
7	3	170
Total		170

Tabela 2.24: Cálculo de decis

Decil para dados agrupados em classes

O cálculo do decil para dados agrupados em classes será realizado por meio das seguintes etapas:

- **primeira etapa**: determinamos a posição do decil, por meio da expressão:

$$P_{D_k} = \frac{k \times n}{10} \, (k = 1, 2, \ldots, 9);$$

em que:
$\sum f_i =$ somatório das frequências simples.
$k =$ índice do decil.

- **segunda etapa**: identificamos a posição do decil na coluna de frequências acumuladas, isto é, a frequência acumulada imediatamente igual ou superior à posição do decil;
- **terceira etapa** verificamos as informações referentes à classe correspondente a essa posição; e
- **quarta etapa** calculamos o valor do decil por meio da fórmula apresentada a seguir, que consiste em uma variação da fórmula da mediana para dados agrupados em classes, mudando-se somente o $\frac{k \times \sum f_i}{10}$

$$Q_k = l_{inf_{Q_k}} + \left[\frac{\frac{k \times \sum f_i}{4} - f_{ac_{ant}}}{f_{Q_k}} \right] \times h_{Q_k}$$

em que:

- $l_{inf_{Q_k}}$ = limite inferior da classe do quartil considerado;
- $f_{ac_{ant}}$ = frequência acumulada da classe anterior à classe do quartil;
- h_{Q_k} = amplitude do intervalo de classe do quartil considerado;
- f_{Q_k} = frequência simples da classe do quartil considerado.

Sempre que houver necessidade, teremos que incluir uma coluna de frequências acumuladas.

a. Calcular Q_1 Calcular a posição de Q_1 por:

$$P_{D_1} = \frac{1 \times 54}{10} = 5,4$$

i	Altura (cm)	Frequência f_i	Frequência Acumulada f_{ac}	Comentários
1	120 ⊢128	6	$6 \geq 5,4$	Classe considerada
2	128 ⊢136	12	18	
3	136 ⊢144	16	34	
4	144 ⊢152	13	47	
5	152 ⊢160	7	54	
Total			54	

onde o limite inferior é 120 cm e a frequência é 6.

$$D_1 = l_{inf_{Q_1}} + \left[\frac{\frac{k \times \sum f_i}{10} - f_{ac_{ant}}}{f_{Q_1}} \right] \times h_{Q_1}$$

$$D_1 = 120 + \frac{5,4 - 0}{6} \times 8 = 127.5\,\text{cm}$$

b. Calcular D_2:

Para começar vamos achar a posição de D_2

$$P_{D_2} = \frac{2 \times 54}{10} = 10,5$$

i	Altura (cm)	Frequência f_i	Frequência Acumulada f_{ac}	Comentários
1	120 ⊢128	6	6	F. acumulada anterior
2	128 ⊢136	12	$18 \geq 10,5$	Classe considerada
3	136 ⊢144	16	34	
4	144 ⊢152	13	47	
5	152 ⊢160	7	54	
Total		$\longrightarrow$	54	

$$D_2 = l_{inf_{Q_2}} + \left[\frac{\frac{k \times \sum f_2}{10} - f_{ac_{ant}}}{f_{Q_2}} \right] \times h_{Q_2}$$

$$D_1 = 128 + \frac{10,8 - 6}{12} \times 8 = 131.2\,\text{cm}$$

c. Cálculo para D_7:

Para começar vamos achar a posição de D_7

$$P_{D_7} = \frac{7 \times 54}{10} = 37,8$$

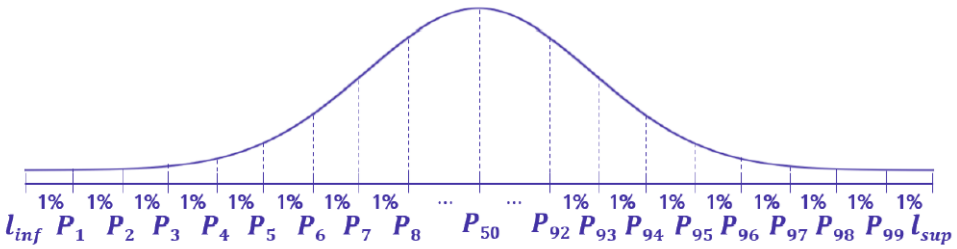
i	Altura (cm)	Frequência f_i	Frequência Acumulada f_{ac}	Comentários
1	120 ⊢128	6	6	
2	128 ⊢136	12	18	
3	136 ⊢144	16	34	F. acumulada anterior
4	144 ⊢152	13	47	Classe considerada
5	152 ⊢160	7	54	
Total		$\longrightarrow$	54	

$$D_7 = l_{inf_{D_7}} + \left[\frac{\frac{k \times \sum f_7}{10} - f_{ac_{ant}}}{f_{D_7}} \right] \times h_{D_7}$$

$$D_1 = 144 + \frac{37,8 - 34}{13} \times 8 = 146.3\,\text{cm}$$

2.11.3 PERCENTIS

Denominamos de percentis os valores de uma série que a dividem em cem partes iguais, isto é, cem partes contendo o mesmo número de elementos (1%). A imagem a seguir mostra os percentis de uma distribuição hipotética:



Temos, então, 99 percentis $P_1, P_2, \ldots, P_{99}$ para dividir uma série em cem partes iguais:

- P_1: o primeiro percentil corresponde à separação do primeiro 1% de elementos da série;
- P_50: o quinquagésimo percentil corresponde à separação de metade dos elementos da série, coincidindo com a mediana $P_{=Md}$;
- P_99: o nonagésimo nono percentil corresponde à separação dos primeiros 99% de elementos da série, ou do último 1% de elementos da série.

Para o cálculo dos percentis, empregaremos a mesma fórmula adotada no cálculo da mediana, somente substituindo a expressão $\frac{sumf_i}{2}$ por $\frac{k \times sumf_i}{100}$, em que □ indica a ordem do percentil e assume valores inteiros no intervalo de 1 a 99.

Percentil para dados não-agrupados

O cálculo do percentil segue o mesmo raciocínio empregado nos cálculos do quartil e do decil para dados não-agrupados. A primeira tarefa que devemos realizar, se houver necessidade, é organizar o conjunto de valores por ordem de magnitude. Depois disso, colocamos em prática as seguintes etapas:

- **primeira etapa**: determinamos a posição do percentil, por meio da expressão:

$$PP_k = \frac{k \times n}{100} \, (k = 1, \, 2, \, \ldots . 99);$$

- **segunda etapa**: identificamos a posição mais próxima do rol;
- **terceira etapa**: verificamos o valor que está ocupando essa posição.

Exemplo Calcularemos os percentis P_{27}eP_{83} para o seguinte conjunto de valores:

$$\{15, 2, 4, 6, 10, 12, 13, 7, 21, 18, 20\}$$

Reparem que os valores não estão organizados. Portanto, nossa primeira tarefa será colocá-los em ordem de magnitude (rol):

$$\{2, 4, 6, 7, 10, 12, 13, 15, 18, 20, 21\}$$

a. Cálculo de P_{27}

Primeiramente determina-se a posição de P_27

$$P_27 = \frac{27 \times 11}{100} = 2,97$$

Depois, identificamos a posição mais próxima no rol:

$$P_{27} = 3$$

Em seguida, verificamos o valor que está ocupando essa posição.

Em seguida, seguida, verificamos o valor que está ocupando essa posição.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_11
2	4	**6**	7	10	12	13	15	18	20	21

Tabela 2.25: Rol em ordem crescente

Portanto, o valor 6 corresponde a 27% do rol.

b. Determinar a posição P_{83}

Determina-se Primeiramente a posição de P_83

$$P_83 = \frac{83 \times 11}{100} = 9,13$$

Depois, identificamos a posição mais próxima no rol:

$$P_{83} = 9$$

Em seguida, verificamos o valor que está ocupando essa posição.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_11
2	4	6	7	10	12	13	15	**18**	20	21

Tabela 2.26: Rol em ordem crescente

Percentil para dados agrupados em classes

O cálculo do percentil para dados agrupados em classes será realizado por meio das seguintes etapas:

- **primeira etapa**: determinamos a posição do percentil, por meio da expressão:

$$PP_k = \frac{k \times n}{100} \, (k = 1, \, 2, \, \ldots . 99);$$

em que:
$k =$ índice do percentil
$\sum f_i =$ somatório das frequencias simples
- **segunda etapa**: identificamos a posição do percentil na coluna de frequências acumuladas, isto é, a frequência acumulada imediatamente igual ou superior à posição do percentil;
- **terceira etapa**: verificamos as informações referentes à classe correspondente a essa posição; e
- **quarta etapa**: calculamos o valor do percentil por meio da fórmula apresentada a seguir, que consiste em uma variação da fórmula da mediana para dados agrupados em classes, mudando-se somente o $\frac{k \times \sum f_i}{100}$

$$P_k = l_{inf_{Pk}} + \left[\frac{\frac{k \times \sum f_i}{100} - f_{ac_{ant}}}{fP_k} \right] \times h_{P_k}$$

em que:

- $l_{inf_{Pk}}$ = limite inferior da classe do percentil considerado;
- $f_{ac_{ant}}$ = frequência acumulada;
- fP_k = amplitude do intervalo de classe do percentil considerado;
- h_{P_k} = frequência simples da classe do percentil considerado.

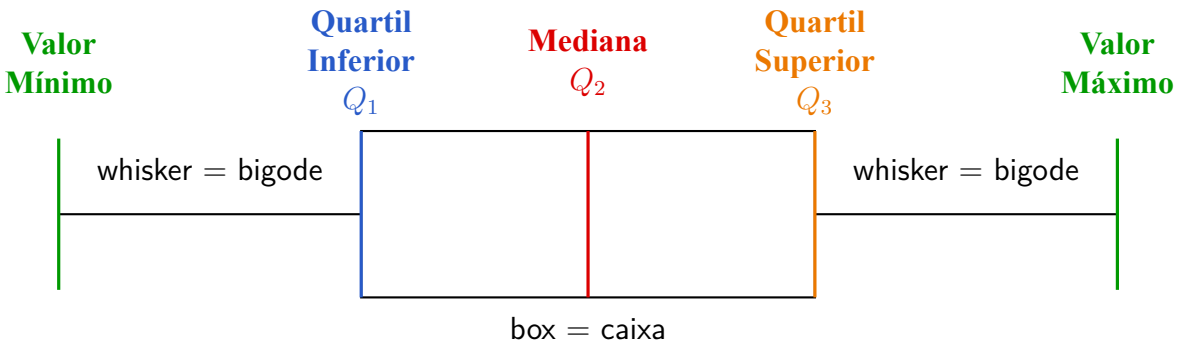
Sempre que houver necessidade, teremos que incluir uma coluna de frequências acumuladas.

2.12 BOX PLOT

Um boxplot (também chamado de box-and-whisker plot) é uma ferramenta gráfica frequentemente utilizada na análise exploratória de dados que permite visualizar a distribuição dos dados e os valores discrepantes (outliers), assim como a distância dos valores extremos em relação à maioria dos dados. Essa ferr amenta resume cinco medidas descritivas de um conjunto de dados, incluindo: o valor mínimo, o primeiro quartil, a mediana, o terceiro quartil e o valor máximo.

Para construir um gráfico de boxplot, usamos uma haste horizontal ou vertical e uma caixa retangular (box). O local onde a haste começa (da esquerda para a direita) indica o valor mínimo e o ponto em que a haste termina indica o valor máximo.

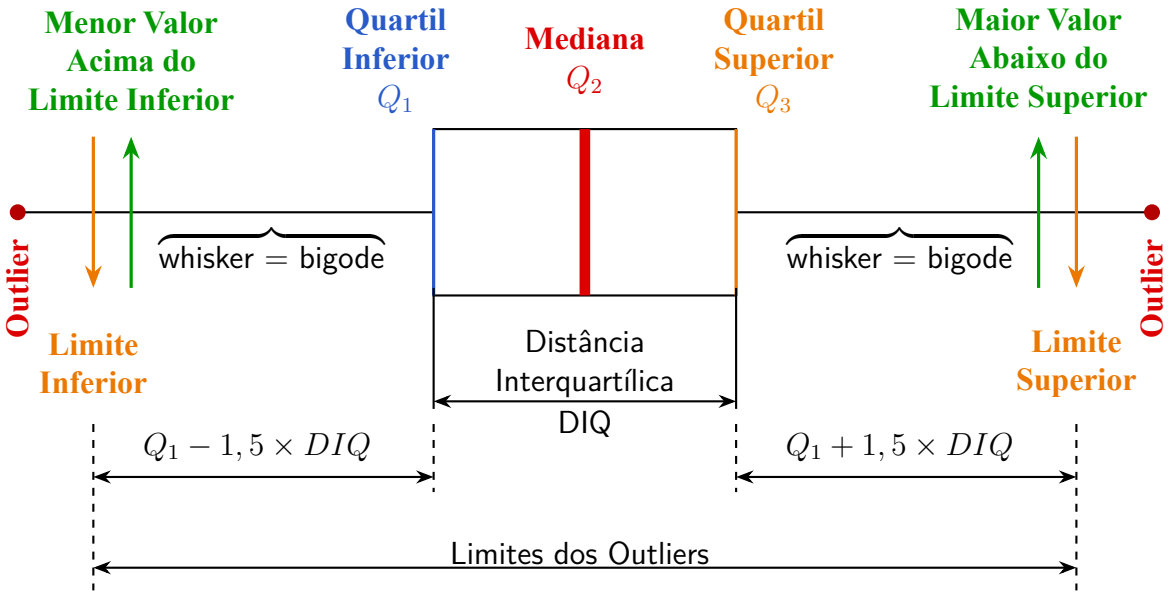
A caixa retangular, localizada no meio da haste, em geral, possui três linhas. A primeira linha, na extremidade esquerda da caixa, indica o primeiro quartil. A terceira linha, na extremidade direita, indica o terceiro quartil. A linha do meio, no interior da caixa, indica o segundo quartil ou a mediana. O segundo quartil pode estar entre o primeiro e o terceiro quartis, ou pode coincidir com um, ou outro, ou ambos.



Além disso, há dois traços, chamados de whiskers(ou bigodes), ligando o valor mínimo à extremidade esquerda da caixa e o valor máximo à extremidade direita da caixa. Cada um desses traços comporta, aproximadamente, 25% dos dados. O restante, cerca de 50%, está distribuído no interior da caixa.

Também podemos encontrar gráficos de box plot com pontos ou asteriscos marcando valores discrepantes(outliers). Nesses casos, os whiskers não se estendem aos valores mínimo e máximo do conjunto de dados, mas ficam limitados a um comprimento máximo de 1,5 $\times$, em que DIQ é a distância interquartílica. A distância interquartílica (ou amplitude interquartílica, ou intervalo interquartílico) é calculada pela fórmula:

$$DIQ = Q3 - Q1$$



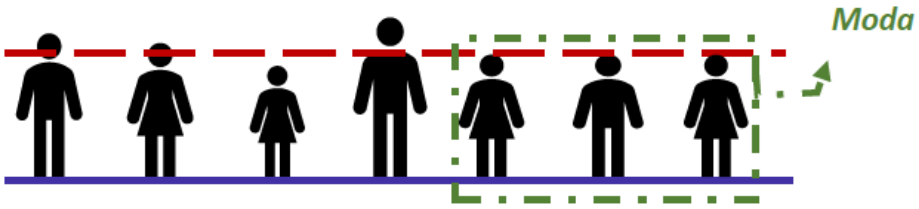
Dessa forma, valores menores que $Q_1 - 1,5 \times DIQ$ ou maiores que $Q_3 + 1,5 \times DIQ$ são considerados VALORES DISCREPANTES (OUTLIERS) e representados por PONTOS ou ASTERISCOS.

É importante salientarmos que a fórmula da distância interquartílica se parece muito com a do desvio quartílico (ou amplitude semi-interquartílica), podendo ser facilmente confundida. O desvio quartílico é calculado pela expressão apresentada a seguir:

$$\boxed{D_q = \frac{Q_3 - Q_1}{2}}$$

2.13 MODA

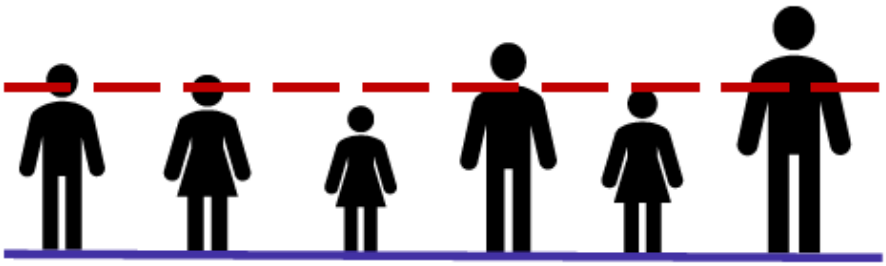
Nessa aula, aprenderemos outra importante medida descritiva: a moda estatística. A moda é uma medida de posição e de tendência central que descreve o valor mais frequente de um conjunto de observações, ou seja, o valor de maior ocorrência dentre os valores observados.



Moda referente ao tamanho mais frequente da população.

Na estatística, o termo moda foi introduzido por Karl Pearson, em 1895, influenciado, muito provavelmente, pela forma com que as pessoas se referiam àquilo que estava em destaque, em evidência, com o significado de coisa mais frequente.

A definição evidencia que um conjunto de valores pode possuir uma ou mais modas, ou não possuir nenhuma. Assim, dizemos que um conjunto é unimodal, bimodal, trimodal ou plurimodal, conforme o número de modas que apresenta. A ausência de uma moda caracteriza o conjunto como amodal.



Conjunto amodal.

Em geral, a moda é utilizada em distribuições nas quais o valor mais frequente é o mais importante da distribuição. A moda também é útil para a determinação da medida de posição de variáveis qualitativas nominais, ou seja, variáveis não-numéricas que não podem ser ordenadas.

O cálculo da moda ocorre de diferentes formas, a depender de como os dados estão organizados. Nesse contexto, aprenderemos a calcular a moda para as seguintes situações:

- dados não-agrupados;
- dados agrupados sem intervalos de classe (ou por valores); e
- dados agrupados em classes.

2.13.1 MODA PARA DADOS NÃO-AGRUPADOS

Para determinarmos a moda de um conjunto ordenado de valores não agrupados em classes, basta identificarmos o elemento (ou elementos) de maior frequência no conjunto.

Com relação ao número de modas, o conjunto de pode ser classificado como:

- amodal: quando todos os elementos apresentam a mesma frequência, isto é, quando todosaparecem o mesmo número de vezes:

$$(1, 2, 3, 5, 6, 8, 10)$$

- unimodal: quando a frequência de um elemento é maior que as frequências dos demais elementos. Assim, um único elemento se destaca entre os demais. Isto é, quando o conjunto tem uma única moda. No conjunto a seguir, o elemento 2 repete-se cinco vezes enquanto o elemento 3 aparece duas vezes. Logo, $Mo = 2$.

$$\left(\underbrace{2, 2, 2, 2, 2,}_{5\ elementos} \underbrace{3, 3}_{2\ elementos} \right)$$

- bimodal: quando as frequências de dois elementos são iguais e maiores que as frequências dos demais elementos. Isto é, quando o conjunto tem duas modas. No conjunto a seguir, os elementos 2 e 3 repetem-se cinco vezes, enquanto o elemento 4 aparece duas vezes. Logo, $Mo = 2$ e $Mo = 3$.

$$\left(\underbrace{2, 2, 2, 2, 2,}_{5\ elementos} \underbrace{3, 3, 3, 3, 3}_{5\ elementos} \underbrace{4, 4}_{2\ elementos} \right)$$

- multimodal ou plurimodal: quando as frequências de três ou mais elementos são iguais e maiores que as frequências dos demais elementos. Isto é, quando o conjunto tem três ou mais modas. No conjunto a seguir, os elementos 2, 3 e 4 repetem-se cinco vezes, enquanto o elemento 5 aparece duas vezes. Logo, $Mo = 2$ $Mo = 3$ e $Mo = 4$.

$$\left(\underbrace{2, 2, 2, 2, 2,}_{5\ elementos} \underbrace{3, 3, 3, 3, 3}_{5\ elementos} \underbrace{4, 4, 4, 4, 4}_{5\ elementos} \underbrace{5, 5}_{2\ elementos} \right)$$

2.13.2 MODA PARA DADOS AGRUPADOS SEM INTERVALOS DE CLASSE

Quando os dados estão agrupados por valores, isto é, quando não agrupados em intervalos de classe, o cálculo da moda também é realizado de maneira simples e rápida. Para tanto, devemos identificar o valor que apresenta a maior frequência absoluta. Vejamos um exemplo.

Exemplo 2.13.1. *Considere que o Estratégia Concursos tenha realizado um simulado, contendo 50 questões, com 100 estudantes da área fiscal, obtendo a seguinte distribuição de acertos:*

Nº de Acertos (X_i)	Frequência Absoluta (f_i)
45	8
46	12
47	28
48	32
49	17
50	3

Tabela 2.27: Número de Acertos

Para calcular a moda dessa distribuição, devemos identificar o maior valor existente na coluna de frequências. Veja novamente a tabela 2.27, o valor da maior frequência é 32 que está marcado em vermelho. Logo, a moda da distribuição é o resultado de 48 questões corretas, pois corresponde a maior frequência. Portanto, podemos concluir que a maior concentração dos participantes errou somente duas questões:

$$Mo = 48$$

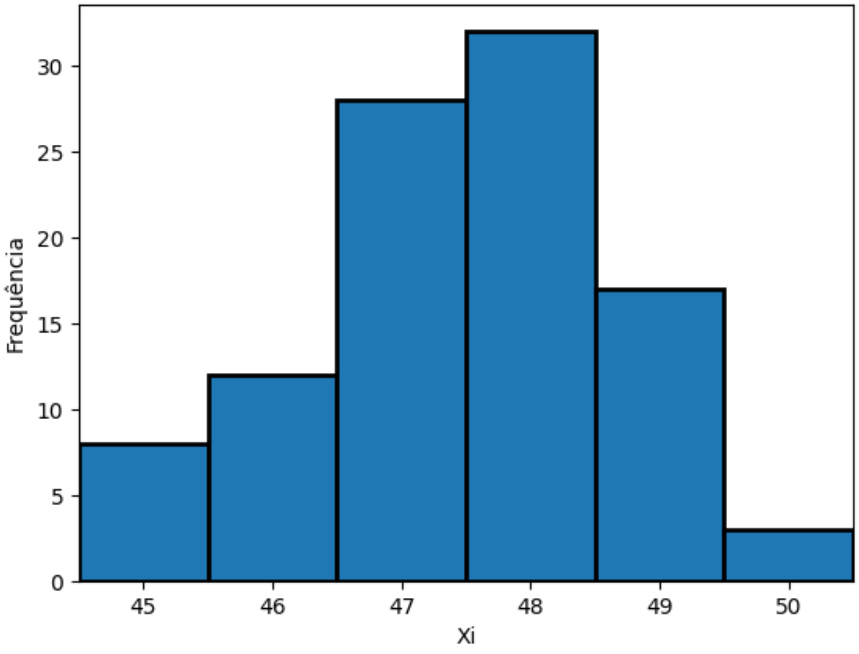


Figura 2.5: Histograma do número de acertos.

Perceba que a maior barra do gráfico, referente à frequência 32, corresponde à moda da distribuição, isto é, um total de 48 questões corretas.

2.13.3 MODA PARA DADOS AGRUPADOS EM CLASSES

Quando os dados estão agrupados em classes de mesma amplitude, a moda será o valor dominante da classe que apresenta a maior frequência, a qual é denominada classe modal. Como já vimos, a

amplitude de classe é a diferença entre os limites superior e inferior de uma determinada classe. Assim, quando as amplitudes são todas iguais, a moda estará contida na classe de maior frequência. A seguir, veremos os principais métodos empregados no cálculo da moda de distribuições agrupadas por intervalos de classe: moda bruta, moda de Pearson, moda de Czuber e moda de King.

Moda Bruta

A maneira mais simples de calcular a moda é tomar o ponto médio da classe modal. Esse valor, ao qual denominamos de moda bruta, é determinado pela seguinte fórmula:

$$\boxed{M_o = \frac{l_{inf} + l_{sup}}{2}}$$

em que l_{inf} é o limite inferior da classe modal; e l_{sup} é o limite superior da classe modal.

Exemplo 2.13.2. *Assim podemos exemplificar:*

Faixa etária $(X_i)	Frequência (f_i)
10 ⊢20	30
20 ⊢30	50
30 ⊢40	70
40 ⊢50	60
50 ⊢60	10
Total 220	

Tabela 2.28: Moda agrupada por classes

Como todas as classes possuem a mesma amplitude, a classe modal é aquela com maior frequência simples. No caso, trata-se da terceira classe marcada em vermelho na tabela 2.28. Temos, então, as seguintes informações:

- *limite inferior da classe modal:* $l_{inf} = 30$*; e*
- *limite superior da classe modal:* $l_{sup} = 40$.

Aplicando a fórmula da moda bruta, temos:

$$M_o = \frac{l_{inf} + l_{sup}}{2}$$
$$M_o = \frac{30 + 40}{2} = 35$$

Moda de Pearson

O matemático Karl Pearson observou a existência de uma relação empírica que permite calcular a moda quando são conhecidas a média $(\bar{x})$ e a mediana (M_d) de uma distribuição moderadamente assimétrica. Quando essas condições são satisfeitas, podemos aplicar a relação denominada de moda de Pearson:

$$M_o = 3 \times M_d - 2 \times \bar{x}$$

Onde: $\bar{x}$ é a média e M_d é a mediana da distribuição.

Moda de Czuber

O matemático Emanuel Czuber elaborou um processo gráfico capaz de aproximar o cálculo da moda. Para determinar graficamente a moda, Czuber partiu de um histograma, utilizando os três retângulos correspondentes à classe modal e às classes adjacentes (anterior e posterior).

$$M_o = l_{inf} + x$$

Nesse caso, o valor de x é determinado pela intersecção dos segmentos $\overline{AB}$ (que une o limite superior da classe que antecede a classe modal ao limite superior da classe modal) e $\overline{CD}$ (que une o limite inferior da classe modal ao inferior da classe posterior).

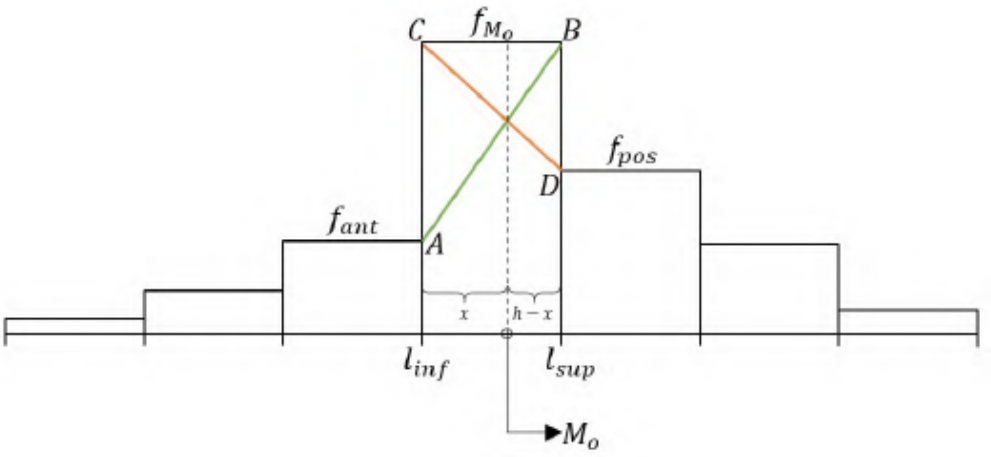


Figura 2.6: Moda de Czber

Note os triângulos (I) e (II), indicados na figura a seguir:

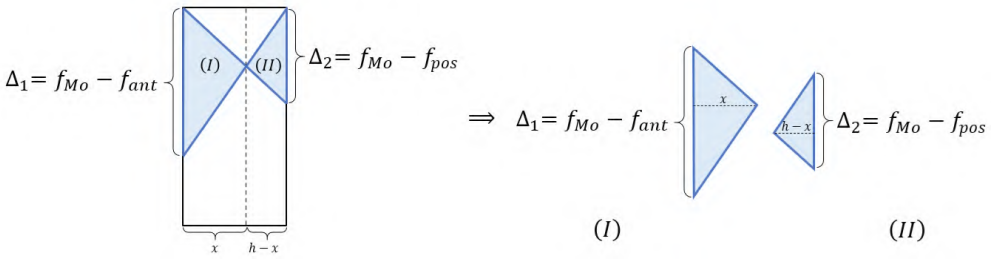


Figura 2.7: Moda de Czber (geometria)

Por semelhança de triângulos temos:

$$\frac{\Delta_1}{x} = \frac{\Delta_2}{h - x}$$

Fazendo a multiplicação cruzada das frações, temos que:

$$\begin{aligned}
\Delta_1(h - x) &= \Delta_2(x) \\
h\Delta_1 - x\Delta_1 &= x\Delta_2 \\
h\Delta_1 &= x\Delta_2 + x\Delta_1 \\
h\Delta_1 &= x(\Delta_2 + x\Delta_1)
\end{aligned}$$

Então temos:

$$x = \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] \times h$$

Como a moda é igual ao limite inferior da classe modal adicionado de □, temos que:

$$M_o = l_{inf} + x$$

e portanto,

$$M_o = l_{inf} + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2}\right] \times h$$

ou

$$M_o = l_{inf} + \left[\frac{f_{Mo} - f_{ant}}{2 \times fM_o - f_{post}}\right] \times h$$

2.13.4 PROPRIEDADES DA MODA

São propriedades da Moda:

- Somando-se (ou subtraindo-se) uma constante c a todos os valores de uma variável, a moda do conjunto fica aumentada (ou diminuída) dessa constante.
- Multiplicando-se (ou dividindo-se) uma constante c a todos os valores de uma variável, a moda do conjunto fica aumentadamultiplicada (ou dividida) dessa constante.

2.14 MEDIDAS DE DISPERSÃO

2.14.1 MEDIDADES DE VARIABILIDADE

Já estudamos mecanismos para encontrar valores (média, mediana e moda) que sintetizam o comportamento dos elementos de um conjunto de dados. Esses valores fornecem parâmetros significativos para uma análise dos dados, porém, também é importante identificarmos como variam ou como se diferenciam as características dos elementos de um conjunto.

Imagine, por exemplo, que você precise avaliar três grupos de alunos, cada um com cinco elementos, no que diz respeito ao domínio de uma determinada matéria. Os testes mostraram os seguintes resultados:

$$\begin{aligned} A =& 7, 7, 7, 7, 7 \\ B =& 5, 6, 7, 8, 9 \\ C =& 1, 4, 10, 10, 10 \end{aligned}$$

Para analisar esses dados, podemos, inicialmente, calcular a média aritmética dos três grupos. Concluímos, então, que todos possuem a mesma média aritmética $\bar{x} = 7$. Contudo, ao observarmos a variação dos dados, percebemos que os grupos se comportam de maneira diferente, apesar de todos possuírem a mesma média.

A média de A, de B e de C são iguais a 7.

Nesse caso, a média, embora seja uma medida representativa do conjunto, não indica o grau de homogeneidade ou heterogeneidade existente entre os valores que compõem o conjunto. Desse modo,

precisamos recorrer a procedimentos matemáticos que possibilitem a compreensão da discrepância existente entre os valores do conjunto.

As medidas de dispersão (ou variabilidade) são justamente métricas que mostram a variação dos dados de um conjunto. Elas podem ser divididas em dois grupos:

a. medidas de dispersão absoluta:

   – amplitude total;
   – amplitude interquartílica;
   – desvio médio;
   – variância; e
   – desvio-padrão.

b. medidas de variação relativa:

   – coeficiente de variação (de Pearson); e
   – variância relativa.

Agora, aprenderemos a medir o grau de concentração ou dispersão dos dados em torno da média. Para isso, estudaremos as principais medidas de dispersão, sendo: amplitude total, amplitude interquartílica, desvio médio, variância, desvio padrão, coeficiente de variação e variância relativa.

A amplitude total (ou simplesmente amplitude) é a diferença entre os valores extremos de um conjunto de observações, ou seja, a diferença entre o maior e o menor elemento desse conjunto:

$$A_T = \bar{x}_m ax - \bar{x}_m in$$

Essa medida de dispersão chama atenção por ser extremamente simples e muito fácil de se calcular. Contudo, há uma certa restrição quanto ao seu uso por conta de sua grande instabilidade, vez que considera somente os valores extremos da série.

Por exemplo, vamos comparar os conjuntos A e B da tabela a seguir:

Conjunto	Média $\bar{x}$	Amplitude total (AT)
$A = 5, 7, 8, 9, 10, 11, 55$	$\bar{x} = 15$	$55 - 5 = 50$
$B = 12, 13, 14, 15, 16, 17, 18$	$\bar{x} = 15$	$18 - 12 = 6$

Tabela 2.29: Médias e Amplitudea

As médias aritméticas dos dois conjuntos são iguais a 15. Portanto, no que diz respeito a essa medida de posição, podemos considerá-los idênticos. Porém, ao calcularmos a amplitude total, verificamos que os valores do conjunto A apresentam um grau de dispersão bem maior que os do conjunto B.

Isso acontece porque, no cálculo da amplitude total, desconsideramos os valores da série que se encontram entre os extremos, o que pode conduzir a interpretações equivocadas. Com frequência, um valor discrepante pode afetar a medida de maneira acentuada. É o caso, por exemplo, do último valor (55) do conjunto A, sensivelmente maior que seu antecessor (11), que elevou a magnitude da amplitude total para 50.

Além disso, a amplitude total também é sensível ao tamanho de amostra. Normalmente, a amplitude total tende a aumentar com o incremento da dimensão da amostra, ainda que não proporcionalmente. Ainda, a amplitude total pode apresentar muita variação de uma amostra para outra, ainda que extraídas de uma mesma população.

Apesar das limitações dessa medida, há situações em que ela pode ser aplicada de forma satisfatória. É o caso, por exemplo, da variação da temperatura em um dia. Também é o caso de quando uma compreensão rápida dos dados é mais relevante que a exatidão de um procedimento complexo.

2.14.2 AMPLITUDE PARA DADOS NÃO AGRUPADOS

Para dados não agrupados, o cálculo da amplitude total pode ser expresso pela fórmula:

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

em que x_max é o maior elemento; e x_min é o menor elemento do conjunto.

Exemplo 2.14.1. *Calcular a amplitude total dos conjuntos apresentados a seguir:*

$$A = 50, 50, 50, 50, 50, 50, 50$$
$$B = 47, 48, 49, 50, 51, 52, 53$$
$$C = 20, 30, 40, 50, 60, 70, 80$$

Aplicando a fórmula anterior para esses dados, obtemos os seguintes resultados:

$$ATA = x_{max} - x_{min} = 50 - 50 = 0$$
$$ATB = x_{max} - x_{min} = 53 - 47 = 6$$
$$ATC = x_{max} - x_{min} = 80 - 20 = 60$$

Nesse caso, podemos observar que o conjunto A obteve uma amplitude total igual a 0, ou seja, uma dispersão nula. Então, significa que os valores não variam entre si. O conjunto B, por sua vez, obteve uma amplitude igual a 6. Já a variável C teve uma amplitude total igual a 60. Embora o valor da amplitude total seja diferente para os conjuntos A, B e C, todos possuem a mesma média aritmética (50). Independentemente da média, verificamos que o conjunto A possui elementos mais homogêneos do que os conjuntos B e C. E, também, que os elementos do conjunto B são mais homogêneos do que os do conjunto C.

2.14.3 AMPLITUDE TOTAL PARA DADOS AGRUPADOS SEM INTERVALOS DE CLASSES

Para dados agrupados SEM intervalos de classe, a fórmula usada para a identificação da amplitude total é similar à adotada para dados não-agrupados. A única diferença consiste na identificação dos valores mínimo e máximo, que agora ocorre por meio de uma tabela de frequências.

Exemplo 2.14.2. *Calcular a amplitude total da tabela de frequências apresentada a seguir.*

Nesse caso, como 1 e 9 são os valores mínimo e máximo da variável x_i, temos o seguinte resultado:

$$AT = x_{max} - x_{min}$$
$$AT = 9 - 1 = 8$$

É importante ressaltar que esses valores foram selecionados independentemente da frequência associada a eles.

x_i	f_i
1	10
3	15
5	10
7	8
9	7

Tabela 2.30: Frequências

2.14.4 AMPLITUDE TOTAL PARA DADOS AGRUPADOS EM CLASSES

Para dados agrupados em intervalos de classe, podemos definir a amplitude total de duas formas:

1. pela diferença entre o limite superior da última classe l_{inf} e o limite inferior da primeira classe l_{sup}, conforme expresso na fórmula a seguir:

$$A = (l_{sup}) - (l_{inf})$$

2. pela diferença entre o ponto médio da última classe PM_{Utl} e o ponto médio da primeira classe PM_{pri}, conforme expresso na fórmula a seguir:

$$A = PM_{ult} - PM_{pri}$$

Exemplo 2.14.3. *Calcular a amplitude total da distribuição de frequências apresentada a seguir:*

Classes	PM_i	f_i
1 ⊢5	3	5
5 ⊢9	7	10
9 ⊢13	11	15
13 ⊢17	15	10
17 ⊢21	19	5
Total		45

Tabela 2.31: Amplitudes agrupadas por classes

Pelo primeiro método, temos que o limite superior da última classe é 21, enquanto o limite inferior da primeira classe é 1. Portanto, temos a seguinte amplitude:

$$A = L_{sup} - L_{inf} A = 21 - 1 = 20$$

Pelo segundo método, temos que o ponto médio da última classe é 19, enquanto o ponto médio da primeira classe é 3. Portanto, temos a seguinte amplitude:

$$A = PM_{ult} - PM_{pri} A = 19 - 3 = 16$$

Observe que a amplitude é menor pelo segundo método, porque os extremos da distribuição são desconsiderados.

2.14.5 PROPRIEDADES DA AMPLITUDE TOTAL

Nesse tópico, estudaremos as principais propriedades da amplitude total:

1. Somando-se (ou subtraindo-se) uma constante □ a todos os valores de uma variável, a amplitude do conjunto não é alterada.

2. Multiplicando-se (ou dividindo-se) todos os valores de uma variável por uma constante c, a amplitude do conjunto fica multiplicada (ou dividida) por essa constante.

Como já sabemos, denominamos de quartis os valores de uma série que a dividem em quatro partes iguais, isto é, quatro partes contendo o mesmo número de elementos (25%). A imagem a seguir mostra os quartis de uma distribuição hipotética:

2.14.6 AMPLITUDE INTERQUARTÍLICA

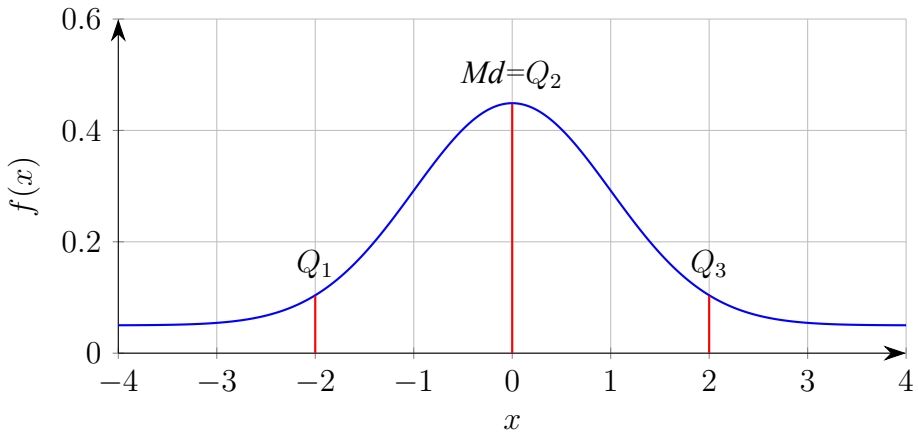


Figura 2.8: Amplitude interquartílica

Temos, então, 3 quartis ($Q1, Q2, Q3$) para dividir uma série em quatro partes iguais:

– Q1: o primeiro quartil corresponde à separação dos primeiros 25% de elementos da série;
– Q2: o segundo quartil corresponde à separação de metade dos elementos da série, coincidindo com a mediana $Q2 = Md$;
– Q3: o terceiro quartil corresponde à separação dos primeiros 75% de elementos da série, ou dos últimos 25% de elementos da série.

A amplitude interquartílica (ou distância interquartílica, ou intervalo interquartílico) é o resultado da subtração entre o terceiro quartil e o primeiro quartil:

$$A_{IQ} = Q_1 - Q_3$$

A amplitude semi-interquartílica (ou desvio quartílico) é definida como a metade desse valor, sendo calculada pela expressão apresentada a seguir:

$$D_Q = \frac{Q_3 - Q_1}{2}$$

A fórmula da amplitude interquartílica (ou distância interquartílica) é muito parecida com a fórmula da amplitude semi-interquartílico (ou desvio quartílico), podendo ser facilmente confundida.

2.14.7 PROPRIEDADES DA AMPLITUDE INTERQUARTÍLICA

1. Somando-se (ou subtraindo-se) uma constante □ a todos os valores de uma variável, a amplitude interquartílica (e o desvio quartílico) do conjunto não é alterada.

2. Multiplicando-se (ou dividindo-se) todos os valores de uma variável por uma constante c, a amplitude interquartílica (e o desvio quartílico) do conjunto fica multiplicada (ou dividida) por essa constante.

2.14.8 DESVIOS EM RELAÇÃO À MÉDIA ARITMÉTICA E MEDIANA

Antes de apresentarmos as fórmulas para o cálculo do desvio médio e da variância, precisamos compreender qual o conceito de desvio em estatística. Um desvio é a distância entre qualquer observação do conjunto de dados e uma medida descritiva desse conjunto:

devio = observação - medida

Em especial, os desvios em relação à média aritmética e em relação à mediana:

$$di = x - \bar{x} \text{ (média)}$$

ou

$$di = x - Md \quad \text{(mediana)}$$

Quando os desvios em relação a uma medida descritiva são pequenos, as observações estão concentradas em torno dessa medida e, portanto, a variabilidade dos dados é pequena. Agora, quando os desvios são maiores, significa que as observações estão dispersas e, portanto, a variabilidade dos dados é grande.

2.14.9 PROPRIEDADES DOS DESVIOS EM RELAÇÃO À MÉDIA ARITMÉTICA E MEDIANA

1. A soma algébrica dos desvios em relação à média é nula.

2. A soma dos quadrados dos desvios da sequência de números x_i, em relação a um número a, é mínima se a for a média aritmética dos números.

3. A soma dos desvios absolutos de uma sequência de números, em relação a um número a, é mínima quando a é a mediana dos números.

2.14.10 DESVIO ABSOLUTO MÉDIO

O desvio absoluto médio, ou simplesmente desvio médio, mede a dispersão entre os valores da distribuição e a média dos dados coletados. Para compreender essa medida, vamos supor que o Estratégia Concursos tenha realizado uma semana de revisão para estudantes da área fiscal, obtendo os seguintes números de visualizações:

Dia da semana	Número de visualizações
Domingo	2.000
Segunda	4.000
Terça	5.200
Quarta	6.300
Quinta	5.400
Sexta	4.100
Sábado	2.400
Total	$\sum f_i = 29.400$

Tabela 2.32: Desvio Absoluto médio

Isso significa que a semana de revisão teve uma média diária de 4.200 visualizações. Esse resultado, porém, não retrata a realidade com fidedignidade, pois alguns dias tiveram mais visualizações do que a média; enquanto outros não. Por isso, é importante sabermos o quão distante a média está em relação aos valores reais por ela representados.

Para calculá-los, basta subtrairmos o valor da média de cada observação, conforme mostrado a seguir:

Dia da semana	Número de visualizações	$x_i - \bar{x}$	
Domingo	2.000	2.000 - 4.200 =	-2200
Segunda	4.000	4.000 - 4.200 =	-200
Terça	5.200	5.200 - 4.200 =	1000
Quarta	6.300	6.300 - 4.200 =	2100
Quinta	5.400	5.400 - 4.200 =	1200
Sexta	4.100	4.100 - 4.200 =	-100
Sábado	2.400	2.400 - 4.200 =	-1800

Tabela 2.33: Cálculo do desvio

Ao calcularmos o desvio médio, obtemos resultados positivos e negativos, que se anulam ao serem somados. Percebam que existem valores de observações que estão muito próximos da média, enquanto outros estão mais distantes.

Como a soma de todos os desvios médios é sempre igual a zero para qualquer conjunto de dados (Primeira propriedade dos desvios), sabemos que $\sum(x - \bar{x})n_i$ não nos fornecerá nenhuma informação relevante nem nos ajudará a compreender o que está acontecendo com essa variável.

Para superar essa dificuldade, podemos utilizar somente os resultados positivos dos desvios calculados. A fórmula do cálculo do desvio médio se apresenta da seguinte maneira:

$$D_m = \frac{\sum_{n=1}^{n} |x_i - \bar{x}|}{n}$$

em que D_m representa o desvio médio, $|x_i - \bar{x}|$ representa o módulo da diferença entre uma determinada observação e a média calculada, f_i representa a frequência de um determinado valor para a variável da distribuição, e n representa o total de elementos formados pela distribuição.

O desvio médio é uma medida de dispersão mais robusta do que a amplitude total e a amplitude interquartílica, pois considera todos os valores do conjunto. O inconveniente dessa medida é a operação de módulo, que, por conta de suas características matemáticas, dificulta o estudo de suas propriedades.

2.14.11 DESVIO MÉDIO PARA DADOS NÃO-AGRUPADOS

O desvio absoluto médio (D_m), de um conjunto de n observações $x_1 1, \ldots, x_n$, é a média dos valores absolutos das diferenças entre as observações e a média. Isto é,

$$D_m = \frac{\sum_{n=1}^{n} |x_i - \bar{x}|}{n}$$

As barras verticais indicam a operação de módulo, responsável por transformar qualquer número negativo em um número positivo, isto é, retornar o valor absoluto.

Exemplo 2.14.4. *Calcular o desvio médio do conjunto mostrado a seguir:*

$$1, 2, 3, 5, 9$$

Iniciaremos pelo cálculo da média aritmética:

$$\bar{x} = \frac{1 + 2 + 3 + 5 + 9}{5} = 4$$

x_i	$x_i - \bar{x}$	$\lvert x_i - \bar{x} \rvert$
1	$(1 - 4) = -3$	3
2	$(2 - 4) = -2$	2
3	$(3 - 4) = -1$	1
5	$(5 - 4) = 1$	1
9	$(9 - 4) = 5$	5

Tabela 2.34: Desvio médio

Aplicando a fórmula do desvio médio, temos:

$$D_m = \frac{\sum_{n=1}^{n} |x_i - 4|}{n} = \frac{\sum_{n=1}^{n} |x_i - \bar{x}|}{n} = \frac{12}{5} = 2,4$$

2.14.12 DESVIO MÉDIO PARA DADOS AGRUPADOS SEM INTERVALO DE CLASSE

Quando os valores vierem dispostos em uma tabela de frequências, o desvio médio será calculado por meio da seguinte fórmula:

$$D_m = \frac{\sum_{i=1}^{m} |x_i - \bar{x}| \times f_i}{\sum f_i}$$

Em que m indica o número de grupos que os dados estão organizados; e $|x_1 - \bar{x}|$ representa o módulo da diferença entre uma determinada observação e a média calculada.

Durante uma pesquisa, o Estratégia Concursos registrou a quantidade de filhos de seus professores, obtendo a tabela de frequências apresentada a seguir. Vamos calcular o desvio médio dessa distribuição.

Iniciamos pelo cálculo da média aritmética:

$$D_m = \frac{\sum_{i=1}^{m} |x_i - \bar{x}| \times f_i}{\sum f_i} = \frac{30}{20} = 1,5 \text{ Filhos por professor}$$

Nº de filhos por professor	f_i	$x_i \times f_i$
0	4	0 ×4 = 0
1	8	1 ×8 = 8
2	4	2 ×4 = 8
3	2	3 ×2 = 6
4	2	4 ×2 = 8
*Pesquisa populacional	$\sum f_i = 20$	$\sum x_i \times f_i = 30$

Tabela 2.35: Distribuição de frequências

Em seguida, adicionaremos uma nova coluna à tabela anterior, em que calcularemos os produtos dos desvios absolutos por suas respectivas frequências:

Nº de filhos por professor	f_i	$x_i \times f_i$	$\lvert x_i - \bar{x}\rvert \times xf_i$
0	4	0	|0 - 1,5| ×4 = 6
1	8	8	|1 - 1,5| ×8 = 4
2	4	8	|2 - 1,5| ×4 = 2
3	2	6	|3 - 1,5| ×2 = 3
4	2	8	|4 - 1,5| ×2 = 5
* Pesquisa populacional	$\sum f_i = 20$	$\sum x_i \times f_i = 30$	$\sum \lvert x_i - \bar{x}\rvert \times f_i = 20$

Tabela 2.36

Por fim, aplicando a fórmula do desvio médio, temos:

$$D_m = \frac{\sum_{i=1}^{m} |x_i - \bar{x}| \times f_i}{\sum f_i} = \frac{20}{20} = 1$$

2.14.13 DESVIO MÉDIO PARA DADOS AGRUPADOS EM CLASSES

Se os dados estiverem agrupados em classe, deveremos adotar a mesma convenção que tomamos para o cálculo da média: vamos assumir que todos os valores coincidem com os pontos médios das suas respectivas classes.

Exemplo 2.14.5. *Durante uma pesquisa, uma escola registrou as estaturas de 40 alunos, obtendo a distribuição de frequências apresentada a seguir. Calcule o desvio médio dessa distribuição*

Estaturas	Frequência f_i
150 ⊢154	4
154 ⊢158	9
158 ⊢162	11
162 ⊢166	8
166 ⊢170	5
170 ⊢174	3
*Pesquisa amostral	$\sum f_i = 40$

Tabela 2.37

Inicialmente, construiremos uma tabela como a mostrada a seguir:

Estaturas	Frequência (f_i)	x_i	$x_i \times f_i$	$(x_i - \bar{x})$	$\lvert x_i \bar{x} \rvert$	$\lvert x_i - \bar{x} \rvert \times f_i$
150 ⊢154	4	152	608	-9	9	36
150 ⊢154	9	156	1.404	-5	5	45
150 ⊢154	11	160	1.760	-1	1	11
150 ⊢154	8	164	1.312	3	3	24
150 ⊢154	5	168	840	7	7	35
150 ⊢154	3	172	516	11	11	33
* Pesquisa amostral	$\sum f_i = 40$		$\sum x_i \times f_i = 6440$			$\sum \lvert x_i - \bar{x} \rvert \times f_i = 184$

Tabela 2.38: Desvio Médio

Feito isso, podemos calcular a média da distribuição por meio da seguinte fórmula:

$$\bar{x} = \frac{\sum x_i \times f_i}{\sum f_i} = \frac{6640}{40} = 161$$

Conhecendo a média, completamos a tabela com as diferenças e os produtos necessários para o cálculo do desvio médio. Assim, aplicando a fórmula do desvio médio, temos:

$$D_m = \frac{\sum_{i=1}^{6} |x_i - \bar{x}| \times f_i}{\sum f_i} = \frac{184}{40} = 4,6$$

Portanto, o desvio médio para essa distribuição de estaturas é 4,6 cm.

2.15 VARIÂNCIA

Existem outras formas de se eliminar o problema com os números negativos. Além da operação de módulo, podemos trabalhar com potências pares. A utilização de potências de expoente par, como o número dois, além de transformar números negativos em positivos, simplifica o cálculo.

A variância é determinada pela média dos quadrados dos desvios em relação à média aritmética. Por meio dessa medida de dispersão ou variabilidade, podemos avaliar o quanto os dados estão dispersos em relação à média aritmética. Nesse sentido, quanto maior a variância, maior a dispersão dos dados.

A variância considera a totalidade dos valores da variável em estudo, e não somente os valores extremos, como faz a amplitude total. Por isso, essa medida de variabilidade é considerada muito estável.

Além disso, a variância complementa as informações obtidas pelas medidas de tendência central.

Até o momento, as medidas que estudamos não sofriam nenhuma alteração quando o cálculo era realizado para uma amostra. Contudo, para a variância, devemos considerar essa informação, pois há uma pequena diferença entre o cálculo da variância populacional e da variância amostral.

A variância populacional é simbolizada pela letra grega σ (sigma), sendo calculada usando todos os elementos da população, pela seguinte fórmula:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}$$

em que: x_i é o valor de ordem i assumido pela variável; μ é a média populacional de x; σ^2 é a variância populacional; e z é o número de dados da população. A variância amostral é simbolizada pela letra s, sendo calculada a partir de uma amostra da população, pela seguinte fórmula:

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

em que: x_1 é o valor de ordem i assumido pela variável; $\bar{x}$ é a média amostral de n; s^2 é a variância amostral; e n é o número de dados da amostra.

Normalmente, uma população possui uma grande quantidade de elementos, o que inviabiliza a realização de um estudo de suas medidas, chamadas de parâmetros populacionais. Nesse caso, recorremos ao estudo de amostras representativas dessa população, buscando obter indícios do valor correto do parâmetro populacional desconhecido. Esse valor amostral é denominado de estimador do parâmetro populacional.

Em nosso caso, a variância populacional cumpre o papel de parâmetro populacional, enquanto a variância amostral atua como um estimador. Já vimos a variância populacional e a variância amostral são representadas por símbolos diferentes: σ^2 e s^2. O mesmo acontece com a média populacional e a média amostral, que também possuem símbolos diferentes: μ (parâmetro populacional) e $\bar{x}$ (estimador).

Reparem que, quando a variância representa uma descrição da amostra e não da população, caso mais frequente em estatística, o denominador das expressões deve ser $n-1$, em vez de n. Isso ocorre porque a utilização do divisor $(n-1)$ resulta em uma melhor estimativa do parâmetro populacional.

Além disso, como a soma dos desvios em relação à média aritmética é sempre nula, apenas $((n-1))$ dos desvios $(n-\bar{x})$ são independentes, vez que $(n-1)$ desvios determinam automaticamente o valor desconhecido. Para amostras grandes $(n > 30)$, não há diferença significativa entre os resultados proporcionados pela utilização de qualquer dos dois divisores, n ou $(n-1)$.

Em determinadas situações, a aplicação dessas fórmulas pode requerer um esforço considerável. É o caso do que acontece quando a média não é um número natural, situação em que a obtenção da soma dos quadrados dos desvios se torna muito trabalhosa. Por isso, é importante aprendermos outras fórmulas que podem nos ajudar no cálculo da variância.

Já ouviram dizer que a variância é igual à média dos quadrados menos o quadrado da média? Pois bem, essa é a fórmula que expressa a variância populacional:

$$\sigma^2 = \bar{x}^2 - \bar{x}^2$$

em que $\overline{x^2}$ é a média dos quadrados; e $\bar{x}^2$ é o quadrado da média.

Como vimos, para encontrarmos a fórmula da variância amostral, basta substituirmos n por $(n-1)$. Isso é equivalente a multiplicarmos a variância populacional por

$$\frac{n}{n-1}$$

É exatamente o que faremos agora:

$$s^2 = [\overline{x^2} - \bar{x}^2] \times \frac{n}{n-1}$$

em que $\overline{x^2}$ é a média dos quadrados; $\bar{x}^2$ é o quadrado da média; e n é o tamanho da amostra.

2.15.1 VARIÂNCIA PARA DADOS NÃO AGRUPADOS

Considere um conjunto de dados $x_1, x_2, \ldots, x_n$.

a. **Para populações:**

A variância populacional é dada por

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

ou, na forma computacional,

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n}.$$

b. **Para amostras:**

A variância amostral é dada por

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

ou, na forma computacional,

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}.$$

A relação entre a variância amostral (s^2) e a variância populacional σ^2 é dada por:

$$s^2 = \left(\frac{n}{n-1}\right) \times \sigma^2$$

Exemplo 2.15.1. *Calcular a variância amostral do conjunto de números mostrado a seguir:*

1, 2, 3, 5, 9

Calculando a média aritmética

$$\bar{x} = \frac{1+2+3-5+9}{5} = \frac{20}{5} = 4$$

Agora, vamos montar uma tabela para facilitar o cálculo da variância:

x_i	$(x_i - \bar{x})^2$
1	$(1-4)^2 = 9$
2	$(2-4)^2 = 4$
3	$(3-4)^2 = 1$
5	$(5-4)^2 = 1$
9	$(9-4)^2 = 25$
Total	$\sum(x_i - \bar{x})^2 = 40$

Tabela 2.39

Por fim, aplicando a fórmula da variância amostral, temos:

$$s^2 = \frac{\sum_{i=1}^{n}(x - \bar{x})^2}{n-1} = \frac{40}{5-1} = 10$$

2.15.2 VARIÂNCIA PARA DADOS AGRUPADOS SEM INTERVALOS DE CLASSE

Considere uma distribuição de frequências composta por valores distintos $X_1, X_2, \ldots, X_m$, com respectivas frequências $f_1, f_2, \ldots, f_m$, e tamanho total

$$n = \sum_{i=1}^{m} f_i.$$

a. **Para populações:**

A variância populacional pode ser calculada por:

$$\sigma^2 = \frac{\sum_{i=1}^{m} (X_i - \mu)^2 \, f_i}{n}$$

ou, na forma computacional,

$$\sigma^2 = \frac{\sum_{i=1}^{m} X_i^2 f_i - \frac{\left(\sum_{i=1}^{m} X_i f_i\right)^2}{n}}{n}.$$

b. **Para amostras:**

A variância amostral é dada por

$$s^2 = \frac{\sum_{i=1}^{m} (X_i - \bar{x})^2 \, f_i}{n-1}$$

ou, na forma computacional,

$$s^2 = \frac{\sum_{i=1}^{m} X_i^2 f_i - \frac{\left(\sum_{i=1}^{m} X_i f_i\right)^2}{n}}{n-1}.$$

Em que

$$n = \sum_{i=1}^{m} f_i$$

e

$$\bar{x} = \frac{\sum_{i=1}^{m} X_i f_i}{n}$$

Exemplo 2.15.2. *Durante uma pesquisa, em uma escola, registrou-se a quantidade de filhos por professor, obtendo a tabela de frequências apresentada a seguir. Sendo assim, calcule a variância amostral dessa tabela.*

Nº de filhos por professor	f_i	$x_i \times f_1$
0	4	0 ×4 = 0
1	8	1 ×8 = 8
2	4	2 ×4 = 8
3	2	3 ×2 = 6
4	2	4 ×2 = 8
* Pesquisa populacional	$f_i = 20$	$x_i \times f_i = 30$

Tabela 2.40

começando pela média aritimética temos:

$$\bar{x} = \frac{\sum x_i \times f_i}{\sum f_i} = \frac{30}{20} = 1,5 \textit{ filhos por professor}$$

Em seguida, adicionaremos uma nova coluna à tabela anterior, em que calcularemos os produtos dos quadrados dos desvios por suas respectivas frequências:

Nº de filhos por professor	f_1	$x_i \times f_i$	$(x_i - \bar{x})^2 \times f_i$
0	4	0 ×4 = 0	$(0-1,5)^2 \times 4 = 9$
1	8	1 ×8 = 8	$(1-1,5)^2 \times 8 = 2$
2	4	2 ×4 = 8	$(2-1,5)^2 \times 4 = 1$
3	2	3 ×2 = 6	$(3-1,5)^2 \times 2 = 4,5$
4	2	4 ×2 = 8	$(4-1,5)^2 \times 2 = 12,5$
* Pesquisa populacional	f_i = 20	$x_i \times f_i = 30$	$\left(\sum x_i - \bar{x}\right)^2 \times f_i$=29

Tabela 2.41

Por fim, aplicando a fórmula do desvio padrão amostral, temos:

$$s = \sqrt{\frac{\sum_{i=1}^{m}(X_i - \bar{x})^2 \, f_i}{n-1}}$$

$$s = \sqrt{\frac{29}{19}} = \sqrt{1,53} \cong 1,23$$

2.15.3 DESVIO-PADRÃO PARA DADOS AGRUPADOS EM CLASSES

Quando tivermos que calcular o desvio-padrão para dados agrupados em classes, usaremos as mesmas fórmulas para dados sem intervalos de classes, utilizando para x_i os pontos médios de cada classe, mas adotando os mesmos procedimentos.

Exemplo 2.15.3. *Durante uma pesquisa, feito em um grupo de estudadntes registrou as estaturas de 40 alunos, obtendo a distribuição de frequências apresentada a seguir. Vamos calcular o desvio-padrão amostral dessa distribuição.*

Estaturas	Frequência f_i
150 ⊢154	4
154 ⊢158	9
158 ⊢162	11
162 ⊢166	8
166 ⊢170	5
170 ⊢174	3
*Pesquisa amostral	$\sum f_i = 40$

Tabela 2.42

Inicialmente, construiremos uma tabela como a mostrada a seguir:

Estaturas	Frequência (f_i)	x_i	$x_i \times f_i$	$(x_i - \bar{x})$	$(x_i\bar{x})^2$	$(x_i - \bar{x})^2 \times f_i$
150 ⊢154	4	152	608	-9	81	324
154 ⊢158	9	156	1.404	-5	25	225
158 ⊢162	11	160	1.760	-1	1	11
162 ⊢166	8	164	1.312	3	9	72
166 ⊢170	5	168	840	7	49	245
170 ⊢174	3	172	516	11	121	363
* Pesquisa amostral	$\sum f_i = 40$		$\sum x_i \times f_i = 6440$			$\sum(x_i - \bar{x}) \times f_i = 1240$

Tabela 2.43

Feito isso, podemos calcular a média da distribuição por meio da seguinte fórmula:

$$\bar{x} = \frac{\sum x_i \times f_i}{\sum f_i} = \frac{6440}{40} = 161$$

Conhecendo a média, completamos a tabela com as diferenças e os produtos necessários para o cálculo da variância. Agora, aplicando a fórmula da variância amostral, temos:

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2}{n-1} \times f_i = \frac{\sum_{i=1}^{6}(x_i - 161)^2}{40-1} \times f_i = \frac{1240}{39} = 31.79\,\text{cm}^2$$

2.15.4 PRIOPRIEDADES DA VARIÂNCIA

1. Somando-se (ou subtraindo-se) uma constante c a todos os valores de uma variável, a variância do conjunto não é alterada.

2. Multiplicando-se (ou dividindo-se) todos os valores de uma variável por uma constante c, a variância do conjunto fica multiplicada (ou dividida) pelo QUADRADO dessa constante.

2.16 DESVIO PADRÃO

O desvio padrão $(sou\sigma)$ é definido como sendo a raiz quadrada da média aritmética dos quadrados dos desvios e, dessa forma, é determinado pela raiz quadrada da variância. É uma das medidas de variabilidade mais utilizadas porque consegue apontar de forma mais precisa a dispersão dos valores em relação à média aritmética.

Valores muito próximos da média resultarão em um desvio-padrão pequeno, enquanto valores mais espalhados levarão a desvios maiores. Essa medida será sempre maior ou igual a zero. Ela será igual a zero quando todos os elementos do conjunto forem iguais.

O desvio padrão é utilizado para comparar a variabilidade de dois conjuntos de dados diferentes quando as médias forem aproximadamente iguais e quando as unidades de medidas para os dois conjuntos forem idênticas.

A fórmula para o cálculo do desvio padrão populacional é:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Para o desvio padrão amostral, a fórmula é a seguinte:

$$s = \sqrt{\frac{\sum_{i=1}^{m}(x_i - \bar{x})^2}{n-1}}$$

Como vimos no tópico anterior, a utilização do divisor $(n-1)$ resulta em uma melhor estimativa do parâmetro populacional. Além disso, como a soma dos desvios em relação à média aritmética é sempre nula, somente $(n-1)$ dos desvios $(x_i - \bar{x})$ são independentes, uma vez que esses $(n-1)$ desvios determinam automaticamente o valor desconhecido.

Por fim, o desvio-padrão é expresso nas mesmas unidades dos dados originais. Tanto o desvio padrão como a variância são usados como medidas de dispersão ou variabilidade. O uso de uma medida ou de outra dependerá da finalidade que se tiver em mente.

O desvio-padrão será igual a zero quando todos os elementos forem iguais. Se todos os elementos forem iguais, a média aritmética do conjunto será igual ao valor dos elementos e todos os desvios também serão iguais a zero. Logo, o desvio-padrão também será zero.

O desvio-padrão é sempre maior ou igual a zero, isto é, sempre tem valor positivo.

2.16.1 DESVIO-PADRÃO PARA DADOS NÃO-AGRUPADOS

Para dados não agrupados, o desvio-padrão pode ser expresso por meio das seguintes fórmulas:

a. para populações;

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

b. para amostras;

$$s = \sqrt{\frac{\sum_{i=1}^{n} d_i^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Exemplo 2.16.1. *Vamos calcular o desvio-padrão amostral do conjunto de números mostrado a seguir:*

{1, 2, 3, 5, 9}

Iniciaremos pelo cálculo da média aritmética:

$$\bar{x} = \frac{1+2+3+5+9}{5} = \frac{20}{5} = 4$$

Em seguida, montaremos uma tabela para facilitar o cálculo do desvio padrão:

x_i	$(x_i - \bar{x})$
1	$(1-4)^2 = 9$
3	$(1-4)^2 = 4$
5	$(1-4)^2 = 1$
7	$(1-4)^2 = 1$
9	$(1-4)^2 = 25$
	$\sum(x_i - \bar{x})^2 = 40$

Tabela 2.44

Por fim, aplicando a fórmula do desvio padrão temos:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{40}{5-1}} = \sqrt{10} \cong 3.16$$

2.16.2 DESVIO-PADRÃO PARA DADOS AGRUPADOS SEM INTERVALO DE CLASSE

Quando os valores vierem dispostos em uma tabela de frequências, o desvio-padrão será calculado por meio de uma das seguintes fórmulas:

a. para populações;

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(d_i \times f_i)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}[(X_i - \mu)^2 \times f_i]}{n}}$$

b. para amostras:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(d_i \times f_i)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}[(X_i - \bar{x})^2 \times f_i]}{n-1}}$$

em que

$$n = \sum_{i=1}^{m} \text{ e } \bar{x} = \frac{\sum_{i=1}^{mX_i\bar{x}}}{n}$$

Exemplo 2.16.2. *Durante a mesma pesquisa sobre a quantidade de filhos dos professores de uma escola, produziu-se a tabela de frequências apresentada a seguir. Vamos calcular o desvio-padrão amostral dessa distribuição.*

Nº de filhos por professor	f_i	$x_i \times f_1$
0	4	0 ×4 = 0
1	8	1 ×8 = 8
2	4	2 ×4 = 8
3	2	3 ×2 = 6
4	2	4 ×2 = 8
* Pesquisa populacional	$f_i = 20$	$x_i \times f_i = 30$

Tabela 2.45