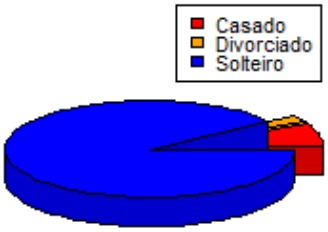
Universidade Federal do Ceará
Campus de Russas

Probabilidade e Estatística



Prof^{a}.: Rosineide F. da Paz

Russas - Ce
Fevereiro–2019

Sumário

Capítulo 1

Introdução

A estatística consiste numa metodologia científica para obtenção, organização, redução, análise e modelagem de dados oriundos das mais variadas áreas das ciências experimentais, cujo objetivo principal é auxiliar a tomada de decisão em situações de incerteza.

Embora não se trate de ramos isolados, basicamente, podemos dividir a estatística em duas áreas:

- Estatística Descritiva: Conjunto de técnicas que objetivam, organizar, resumir, analisar e interpretar os dados experimentais sob consideração.
- Estatística Inferencial: Processo de obter informações sobre uma população a partir de resultados observados na amostra.

> Se o interesse é tirar conclusões sobre um conjunto maior que os dados a serem observados, devemos aplicar técnicas de amostragem para obter uma amostra que seja representativa desse conjunto maior.

A estatística é de grande utilidade quando o método científico é utilizado para testar teoria ou hipóteses em muitas áreas do conhecimento. Esse método pode ser resumido nos seguintes passos.

1. Um problema é formulado em que, muitas vezes, uma hipótese precisa ser testada.
2. Para solucionar o problema, deve-se coletar informações que sejam relevantes, para isso pode-se formular um experimento. Em muitas áreas do conhecimento o planejamento do experimento não é simples, ou até mesmo não é possível, e uma estratégia pode ser a observação de algum fenômeno (variável) de interesse.
3. Os resultados do experimento podem ser utilizados para se obter conclusões, definitivas ou não.
4. os passo 2 e 3 podem ser repetidos quanta vezes forem necessárias.

É notável que nos passos descritos acima a estatística seja uma ferramenta indispensável, podendo ser requerida em todas as etapas.

1.1 Conceitos Básicos

- **População:** consiste em um conjunto de elementos que compartilham de pelo menos uma característica comum.

- **Amostra:** conjunto de elementos extraídos da população.

- **Censo:** É o processo utilizado para levantar as características observáveis, abordando todos os elementos de uma população.

  Exemplo 1.1.1. *Como exemplos de população, podemos citar:*

  - *Pesquisa de opinião pública:*
    - *a população é o total de habitantes de um local;*
    - *a amostra é uma parte dessa população.*
  - *Em sucessivos lançamentos de uma moeda:*
    - *a população é formada por todos os resultados possíveis, cara ou coroa em cada lançamento, aqui a população e infinita;*
    - *a amostra é formada pelos resultados obtidos em uma sequência finita de lançamentos.*
  - *Investigar a porcentagem de lajotas defeituosas fabricadas em uma indústria, durante 6 dias, examinando 20 peças por dia.*
    - *População: todos as lajotas fabricadas durante 6 dias.*
    - *Amostra: o subconjunto de* $6 \times 20 = 120$ *peças, selecionadas para estudo.*

  Na prática não podemos utilizar qualquer amostra para o propósito de inferência. A extração de uma amostra pode ser feita de várias maneiras e deve seguir algumas regras, dependendo do problema que se deseja tratar.

  A Figura 1.1 mostra de forma esquemática e resumida as possíveis etapas de uma análise estatística. Note que existem casos em que apenas uma análise descritiva (exploratória) dos dados é suficiente para tiramos conclusões a respeito da população de interesse. No entanto, se a população é maior do que os dados analisados, sob determinadas condições, podemos fazer uso das teorias das probabilidades para fazer inferência sobre as características (parâmetros) da população de interesse.

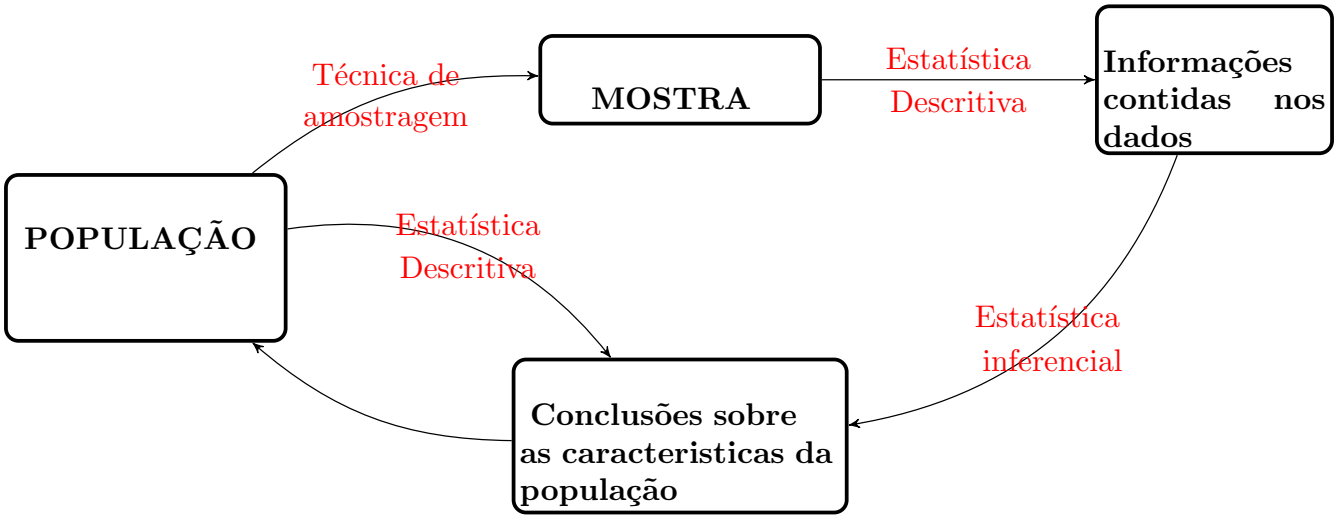


Figura 1.1: Análise estatística a partir da amostra ou da população.

Independentemente de estarmos diante de uma amostra ou de uma população, ao analisar um conjunto de dados devemos sempre fazer uma analise exploratória fazendo uso de ferramentas da **estatística descritiva**. Embora essa exploração ocorra de forma semelhante para população ou amostra, vamos utilizar notações diferentes para indicar se estamos diante de uma população ou de uma amostra.

Capítulo 2

Estatística Descritiva

Antes de qualquer análise estatística mais sofisticada, devemos realizar uma análise exploratória dos dados fazendo uso das ferramentas da **Estatística Descritiva**. Nessa etapa da análise, procuramos obter a maior quantidade possível de informações dos dados observados. Se estamos diante de uma amostra, é nessa fase que devemos obter informações sobre qual modelo probabilístico descreve o fenômeno investigado e pode ser utilizado em uma fase posterior denominada "Inerência Estatística".

2.1 Variável

Os dados são observações de variáveis ou fenômeno de interesse. Uma variável é uma quantidade ou atributo cujo valor pode variar de uma unidade de investigação para outra. Por exemplo, as unidades podem ser pessoas portadoras de dor de cabeça e a variável o tempo entre tomar um remédio e sessar a dor. Uma observação, ou resposta, é o valor assumido por uma variável em uma das unidades investigadas. A observação da variável em várias unidades dá origem aos dados observados.

Exemplo 2.1.1. *Alguns exemplos de variáveis são:*

- *tempo de execução de um algoritmo;*
- *rendimento das famílias de uma grande cidade;*
- *número de erros em pacotes de dados enviados por um servidor;*
- *número de clientes com a mesma dúvida em um site de suporte durante um período de tempo;*
- *opinião dos consumidores de um determinado produto (péssimo, regular, ótimo) etc.*

Existem vários tipos de variáveis, sendo que inicialmente podemos dividi-las em qualitativas e quantitativas. As variáveis são **qualitativa** quando seus valores forem expressos por atributos (não numéricas). Este grupo pode ser subdividido em: qualitativa nominal e ordinal. A variável é nominal se os atributos que ela representa não têm uma ordenação, por exemplo, cor de cabelo, sexo de indivíduos etc, enquanto que as ordinais exprimem alguma ordenação, como por exemplo, opinião sobre a qualidade de um produto (péssimo, regular, ótimo). As variáveis **quantitativas:** assumem valores numéricos. Essas variáveis também podem ser classificadas em dois grupos: contínua e discreta. As variáveis são discretas se assumem valores em um conjunto enumerável (contável), como por exemplo, número de carros que passam por um posto de pedágio em um intervalo de tempo. As variáveis contínuas assumem valores em um conjunto não-enumerável, ou seja, em um intervalo da reta, como por exemplo, alturas de pessoas de um determinado povoado. Veja um resumo das variáveis no esquema a seguir.

$$\text{Variável} \longmapsto \begin{cases} \text{Qualitativa} \longmapsto \begin{cases} \text{nominal (Ex. Região de procedencia)} \\ \text{ordinal (Ex. Grau de instrução)} \end{cases} \\ \text{Quantitativa} \longmapsto \begin{cases} \text{discreta (Ex. número de filhos)} \\ \text{contínua (Ex. Altura)} \end{cases} \end{cases}$$

Em geral, as medições dão origem às variáveis contínuas e as contagens ou enumerações às variáveis discretas.

2.2 Exercícios para a Seção 2.1

Exercício 2.2.1. *Classifique as seguintes variáveis em qualitativas (nominal/ordinal) ou quantitativa (discreta/contínua).*

a) *Classe social.*

b) *Número de clientes em um estabelecimento.*

c) *Salário mensal.*

d) *Cidade de nascimento.*

e) *Departamento que trabalha.*

f) *Número de filho.*

g) *Nível de escolaridade.*

h) *Número de processos analisados.*

i) *Opinião sobre a reforma agrária.*

j) *Opinião sobre atendimento de um estabelecimento.*

k) *Número de telefonemas recebidos.*

l) *Estado Civil.*

m) *Idade (anos).*

n) *Distância de sua casa na faculdade (marcado no velocímetro).*

o) *Número de idas ao cinema por semana.*

Exercício 2.2.2. *Declare se cada uma das seguintes variáveis é do tipo discreta ou contínua:*

1. *O número anual de suicídios no Brasil;*
2. *A concentração de chumbo em uma amostra de água;*
3. *A duração de tempo que um paciente sobrevive depois do diagnóstico de uma doença fatal;*
4. *O número de abortos prévios que uma mãe grávida teve.*

2.3 Estrutura dos Dados e Notação

Os dados, em geral, são dispostos em tabelas, ou planilhas, de modo que em cada coluna podem ser observados os valores observados de uma única variável. Assim, nas linhas da tabela estão os valores observados para cada variável. Essa estrutura simples permite que os dados possam ser analisados por meio de diversos software, tais como library calc, Rstudio, entre outros.

Exemplo 2.3.1. *Suponha, por exemplo, que um questionário foi aplicado aos alunos de um curso da UFC, fornecendo as seguintes informações: Idade em anos; Altura em metros; Peso em quilogramas; Estado Civil: Solteiro, casado, divorciado e viúvo, como mostra o Quadro a seguir.*

Estado civil	*Idade*	*Peso*	*Altura*
solteiro	*20*	*74*	*1,68*
solteiro	*18*	*46*	*1,6*
solteiro	*19*	*62*	*1,6*
solteiro	*19*	*64*	*1,7*
solteiro	*25*	*98*	*1,9*
solteiro	*24*	*68*	*1,72*
solteiro	*20*	*60*	*1,7*
solteiro	*35*	*71*	*1,68*
solteiro	*19*	*67*	*1,62*
solteiro	*20*	*79*	*1,87*
solteiro	*19*	*80*	*1,75*
solteiro	*20*	*65*	*1,74*
solteiro	*20*	*74*	*1,6*
solteiro	*20*	*65*	*1,7*
solteiro	*19*	*53*	*1,63*
solteiro	*19*	*60*	*1,67*
solteiro	*23*	*45*	*1,6*
divorciado	*26*	*70*	*1,7*
solteiro	*20*	*75*	*1,7*
solteiro	*21*	*75*	*1,7*
solteiro	*19*	*73*	*1,76*
casado	*46*	*70*	*1,7*
solteiro	*19*	*70*	*1,78*
solteiro	*28*	*58*	*1,75*
solteiro	*21*	*68*	*1,6*
solteiro	*23*	*62*	*1,7*
solteiro	*19*	*66*	*1,74*
solteiro	*20*	*74*	*1,8*
solteiro	*22*	*90*	*1,86*
casado	*58*	*98*	*1,8*
solteiro	*24*	*74*	*1,73*
solteiro	*20*	*70*	*1,7*
casado	*26*	*95*	*1,6*

...

solteiro	*20*	*46*	*1,54*
solteiro	*21*	*69*	*1,57*
solteiro	*19*	*57*	*1,57*
solteiro	*19*	*59*	*1,61*
solteiro	*17*	*58*	*1,49*
solteiro	*20*	*62*	*1,7*
solteiro	*20*	*60*	*1,65*
solteiro	*22*	*49*	*1,6*

O conjunto de informações disponíveis, após a tabulação do questionário ou pesquisa de campo, é denominado tabela de dados brutos e contém os dados da maneira que foram coletados inicialmente, após a crítica dos valores. Em nosso caso temos quatro variáveis envolvidas sendo uma qualitativa (estado civil) e as restantes quantitativas (idade, peso, altura).

2.3.1 Notação

Uma variável qualquer será denotada por uma letra maiúscula, como por exemplo X e uma sequência de valores observados dessa variável será denotada por letras minúscula, de modo que:

- $x_1, \cdots, x_n$ representa uma sequência de valores observados da variável X em uma amostra e
- $x_1, \cdots, x_N$ representa uma sequência de valores observados da variável X em uma população inteira,

em que n denota o tamanho da amostra e N denota o tamanho da população.

2.4 Distribuição de Frequência

O objetivo da Estatística Descritiva é resumir as principais características dos dados observados fazendo uso de tabelas, gráficos e resumos numéricos, para se ter uma ideia do comportamento da variável estudada. Pois, quando se estuda uma variável, o maior interesse é conhecer o seu comportamento, sua frequência, ou sua **distribuição de frequência**.

2.4.1 Tabelas de Frequência

Para se ter uma ideia dessa distribuição, podemos construir uma tabela de frequência para o conjunto de observações dessa variável.

Exemplo 2.4.1. *Suponha que observamos as notas finais (por conceito) de 30 alunos de um determinado curso e obtivemos os seguintes valores:*

C B B B B A C B B A D D B C B B C B C B C C C B B C B B A C

A variável de interesse é o conceito (A, B, C ou D, em que D significa a reprova do aluno). Será que, de um modo geral, a turma teve um bom desempenho?

Para responder essa questão, devemos observar a frequência da variável "conceito". Essa frequência fica evidente se os dados forem dispostos em uma tabela apropriada. Em particular, para esse conjunto de dados, podemos utilizar uma tabela de frequência simples. A Tabela 2.2 apresenta a frequência absoluta (n_i) e a frequência relativa (f_i) da variável conceito. A frequência relativa indica a proporção de vezes que um determinado valor da variável aparece, por exemplo, note que o conceito

B corresponde a 50% dos valores observados da variável. Em outras palavras, 50% dos estudantes obtiveram conceito B neste curso.

Tabela 2.2: Distribuição de frequência para a variável conceito.

Conceito	Frequência absoluta (n_i)	Frequência relativa (f_i)
A	3	$(3/30) = 0,1$
B	15	$(15/30) = 0,5$
C	10	$(10/30) \cong 0,33$
D	2	$(2/30) \cong 0,07$
Total (n)	30	1

Existem duas possibilidades para a construção de tabelas de frequência:

(i) tabelas de frequências simples;

(ii) e tabelas de frequências em intervalos de classes.

A tabela de frequência simples é usada para variáveis qualitativas e quantitativas discretas com poucos valores possíveis. A tabela em intervalos de classe é apropriada para variáveis quantitativas contínuas (ou discretas com muitos valores possíveis).

Tabela de Frequência Simples

Essa tabela é apropriada para para variáveis qualitativas ou quantitativas discretas com poucos valores possíveis. O formato geral para esse tipo de tabela pode ser visto na Tabela 2.3, em que:

- $x_1, \cdots, x_k$ representam os valores distintos e ordenados que podem ser encontrados no conjunto de dados.
- $n_1, \cdots, n_k$ representam a contagem das repetições de cada valor distinto, e são denominadas frequências absolutas, com $n_1 + n_2 + \cdots + n_k = n$.
- $f_1, \cdots, f_k$ são as frequências relativas (ou proporções).
- f_{ac} representa a frequência relativa acumulada, também podemos incluir uma coluna contendo os valores de frequências absolutas acumuladas (n_{ac}).
- n: número de observações (ou tamanho da amostra, caso sejam dados de uma população usamos N).
- k: número de classes na tabela de frequência.

Tabela 2.3: Formato geral para uma tabela de frequência simples.

Variável	(n_i)	(f_i)	f_{ac}
x_1	n_1	$f_1 = n_1/n$	n_1/n
x_2	n_2	$f_2 = n_2/n$	$(n_1 + n_2)/n$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_k	n_k	$f_k = n_k/n$	$(n_1 + n_2 + \cdots + n_k)/n.$
Total	n	1	

Exemplo 2.4.2. *: Considerando os dados da variável Estado Civil do Exemplo 2.3.1, temos uma tabela de frequência simples com $k = 4$ classes, como pode ser visto na Tabela 2.4. Nesta tabela, acrescentarmos a coluna de frequência relativa, definida por $f_i = n_i/n$, com isso podemos ver rapidamente que 90% dos estudantes entrevistados são solteiros.*

Tabela 2.4: Tabela de frequência para a variável Estado Civil.

Estado Civil	n_i	f_i
Solteiro	37	0,90
Casado	03	0,07
Divorciado	01	0,02
Viúvo	00	0,00

Considerando, novamente, os dados de estudantes apresentados na Tabela 2.3.1, vamos averiguar em torno de que valor tendem a se concentrar as estaturas obtidas, qual a menor ou qual a maior estatura, ou ainda, quantos alunos se acham abaixo ou acima de uma dada estatura. Observe que olhar diretamente para a tabela de dados brutos é difícil se ter uma ideia do comportamento da variável altura considerando o grupo inteiro de estudantes. Novamente podemos interpretar mais facilmente esse conjunto de dados fazendo uso de uma tabela de frequência. Considerando a Tabela de Frequência Simples 2.5, percebemos que esta se mostra muito longa e difere pouco da tabela de dados brutos. Assim, é conveniente montar outra estratégia para resumir esses dados em uma tabela de frequência.

Tabela 2.5: Tabela de frequência simples para a variável Altura.

Variável Altura	**Frequência absoluta** (n_i)
1,49	1
1,54	1
1,57	2
1,60	7
1,61	1
1,62	1
1,63	1
1,65	2
1,67	1
1,68	2
1,70	10
1,72	1
1,73	1
1,74	2
1,75	1
1,76	1
1,78	1
1,80	2
1,86	1
1,87	1
1,90	1

Observe que o processo empregado para variável Altura na Tabela 2.5 é inconveniente, pois exige muito espaço, mesmo quando o número de valores da variável (n) não é muito grande, e não nos esclarece muita coisa. Desta forma, o melhor seria formar agrupamentos. Assim, em vez de trabalharmos com os valores observados da variável, podemos formar intervalos que possam conter esses valores.

Tabelas de Frequências em Intervalos de Classes

Agora vamos construir uma tabela de frequência apropriada para a variável altura dos estudantes. Neste caso temos uma variável que não tem uma natureza discreta, então, é natural não haver muitas repetições no conjunto de dados. Assim, devemos construir uma tabela em intervalos de classes da seguinte procedendo da seguinte forma:

- Determina-se o número de classes, k, fazendo $k \approx \sqrt{n}$, se n é grande, podemos utilizar a regra de Sturges em que
$$k \approx 1 + 3,3 \times \log n,$$
ou ainda determinar esse valor conforme seja mais apropriado.
- definimos L_{inf} um valor menor ou igual ao valor mínimo ($L_{inf} \leq$ **valor mínimo**);
- definimos L_{sup} um valor maior o igual ao valor máximo ($L_{sup} \geq$ **valor máximo**);
- Obtemos a amplitude das classes $AT = L_{sup} - L_{inf}$;
- finalmente, obtemos $h = AT/k$, amplitude de cada classe.

Para os dados da variável Altura, vamos fixar $L_{sup} = 1,98$ e $L_{inf} = 1,48$ para determinar

$$AT = L_{sup} - L_{inf} = 1,98 - 1,48 = 0,50$$

Como $n = 41$, temos $\sqrt{41} \approx 6,4$, mas aqui assumiremos $k = 5$ que fornece $h = 0,50/5 = 0,1$. Logo, obtemos os seguintes **limites das classes**:

- $1,48 + 0,1 = 1,58 \quad \Rightarrow \textbf{classe}_1 : 1,48 \dashv 1,58 = (1,48; 1,58]$
- $1,58 + 0,1 = 1,68 \quad \Rightarrow \textbf{classe}_2 : 1,58 \dashv 1,68 = (1,58; 1,68]$
- $1,68 + 0,1 = 1,78 \quad \Rightarrow \textbf{classe}_3 : 1,68 \dashv 1,78 = (1,68; 1,78]$
- $1,78 + 0,1 = 1,88 \quad \Rightarrow \textbf{classe}_4 : 1,78 \dashv 1,88 = (1,78; 1,88]$
- $1,88 + 0,1 = 1,98 \quad \Rightarrow \textbf{classe}_5 : 1,88 \dashv 1,98 = (1,88; 1,98]$

Observe que usamos duas notações para intervalos, sendo $1,48 \dashv 1,58$ um intervalo que vai desde 1,48 até 1,58 aberto em 1,48 e fechado em 1,58, assim como $(1,48; 1,58]$. Aqui vamos usar a notação com $\dashv$.

Para construir a coluna das frequências absolutas, a parti do hol (dados ordenados), contamos quantos elementos pertencem a cada intervalo de classe obtido. A distribuição de frequência para a variável Altura pode ser vista na Tabela 2.6 e os dados ordenados são:
(1,49; 1,54; 1,57; 1,57; 1,60; 1,60; 1,60; 1,60; 1,60; 1,60; 1,60; 1,61; 1,62; 1,63; 1,65; 1,67; 1,68; 1,68; 1,70; 1,70; 1,70; 1,70; 1,70; 1,70; 1,70; 1,70; 1,70; 1,70; 1,72; 1,73; 1,74; 1,74; 1,75; 1,75; 1,76; 1,78; 1,80; 1,80; 1,86; 1,87; 1,90).

Tabela 2.6: Frequência dos alunos segundo sua altura.

X	n_i	f_i	f_{ac}
$1,48 \dashv 1,58$	4	$4/41 \approx 0,10$	$4/41 \approx 0,10$
$1,58 \dashv 1,68$	14	$14/41 \approx 0,34$	$18/41 \approx 0,44$
$1,68 \dashv 1,78$	18	$18/41 \approx 0,44$	$36/41 \approx 0,88$
$1,78 \dashv 1,88$	4	$4/41 \approx 0,10$	$40/41 \approx 0,98$
$1,88 \dashv 1,98$	1	$1/41 \approx 0,02$	$1/41 = 1$
Total	41	1	

Note que, na Tabela 2.6, foi incluída uma coluna extra para uma sequência denominada **frequência relativa acumulada** (f_{ac}), essas frequências são importantes para qualquer tipo de variável que contém uma certa ordenação, pois a partir dessas frequências podemos tirar conclusões sobre quantos por cento dos valores estão acima ou abaixo de um determinado valor observado da variável. Como exemplo, podemos notar nesta tabela que aproximadamente 88% dos estudantes tem altura até 1,78 m, isso quer dizer que apenas 12% tem mais de 1,78 m.

Exceto pealas classes, a tabela de frequência em intervalos de classes é construída da mesma maneira da tabela de frequência simples. Sua interpretação também é similar, apenas devemos levar em consideração que os valores possíveis não são precisos.

Exemplo 2.4.3. *A Tabela 2.7 categoriza visitas ao consultório de doenças cardiovasculares por duração de cada visita (em minutos). Uma duração de 0 (zero) minuto implica que o paciente não teve contato direto com o especialista.*

Tabela 2.7: Tempo (minutos) de duração de visita ao cardiologista de um grupo de 10614 pessoas.

Duração (minutos)	Número de visitas (n_i)	f_i	f_{ac}
0	390	0,036	0,036
$1 \vdash 6$	227	0,02	0,056
$6 \vdash 11$	1023	0,09	0,146
$11 \vdash 16$	3390	0,31	0,456
$16 \vdash 21$	4431	0,41	0,866
$21 \vdash 26$	968	0,09	0,956
$26 \vdash 31$	390	0,03	0,986
Duração≥ 31	185	0,015	1
Total	10614		

Observe que essa é uma situação em que o limite superior da última classe não é determinado na tabela. Observando a Tabela 2.7 podemos procurar responder as seguintes questões:

1. *Qual o tamanho da amostra?*
2. *Pode-se fazer a afirmação de que as visitas a consultórios de especialistas de doenças cardiovasculares têm duração mais frequente entre 16 e 21 minutos?*
3. *É razoável que a secretária agende uma nova consulta a cada 10 minutos?*
4. *Se não for satisfatório agendar uma nova consulta a cada 10 minutos, qual seria o tempo entre o agendamento de duas consultas que você julga satisfatório?*

Como uma tabela deve aparecer em um texto?

1. Uma tabela deve sempre ser apresentada com suas laterais abertas, se as laterais são fechadas trata-se de um quadro.

2. O título da tabela deve sempre aparecer no topo, não em baixo.

3. Na parte de baixo, pode-se apresentar a fonte e outras informações.

4. Tabelas são elementos flutuantes em um texto, ela pode ser apresentada em qualquer lugar com sua devida numeração.

5. Ao se fazer referencia a uma tabela em um texto, deve-se usar sua numeração (ex. na Tabela 1 pode ser visto...).

2.4.2 Exercícios para a Seção 2.4.1

Exercício 2.4.1. *A massa (em quilogramas) de 17 trabalhadores de uma empresa com 100 funcionários está registrada a seguir:*

52 73 80 65 50 70 80 65 70 77 82 91 52 68 86 70 80

Com base nos dados obtidos, responda:

(a) *Qual é a variável nessa pesquisa? Ela é discreta ou contínua?*

(b) *Que frequências absolutas têm os valores 65 kg, 75 kg, 80 kg e 90 kg?*

Exercício 2.4.2. *A cantina de uma escola selecionou 50 alunos ao acaso e verificou o número de vezes por semana que eles compravam lanche., obtendo os seguintes resultados:*
0; 2; 2; 4; 3; 2; 2; 1; 1; 2; 1; 1; 0; 1; 1 ; 1; 1; 1; 1; 2; 2; 2; 3; 2; 2; 2; 0; 2; 2; 1; 1; 0; 2; 0; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 1; 2; 5; 4.

(a) *Construa uma tabela de distribuição de frequências absolutas, relativas e frequências relativas acumuladas com esses dados.*

(b) *Qual é a proporção de alunos que compram pelo menos 2 lanches por semana?*

Exercício 2.4.3. *Um hospital tem o interesse em determinar a altura média dos pacientes de uma determinada área e relacioná-la com a incidência de determinada anomalia ortopédica. Foram selecionados 80 pacientes e as alturas (em m) podem ser encontradas abaixo.*

1,72	1,78	1,87	1,86	1,79	1,79	1,83	1,74	1,64	1,62
1,75	1,65	1,75	1,58	1,63	1,77	1,64	1,68	1,66	1,82
1,68	1,80	1,74	1,76	1,74	1,72	1,75	1,89	1,73	1,76
1,72	1,71	1,63	1,81	1,65	1,58	1,63	1,70	1,73	1,57
1,75	1,64	1,73	1,70	1,75	1,56	1,70	1,68	1,68	1,79
1,75	1,71	1,62	1,83	1,72	1,76	1,67	1,82	1,67	1,60
1,67	1,61	1,61	1,67	1,75	1,80	1,70	1,77	1,73	1,77
1,64	1,66	1,74	1,66	1,66	1,79	1,68	1,79	1,69	1,80

Construa a tabela de distribuição de frequências por intervalos de classes, apresentando também a coluna da distribuição de frequências acumulada.

2.4.3 Gráficos de Frequência

Vimos que a distribuição de frequências pode ser representada em tabelas de frequências. Outra opção para representar essas frequências é o uso de gráficos. Os principais gráficos para representação de distribuição de frequências são:

1. Gráficos em barras (apropriado para variáveis qualitativas);
2. Gráficos em setores (apropriado para variáveis qualitativas);
3. Histograma (apropriado para variáveis quantitativas) e
4. Polígono de frequências absolutas.

Muitos outros tipos de gráficos são encontrados, no entanto vários são versões diferentes dos tipos citados acima. Aqui, vamos nos limitar a discutir cada um desses.

Gráficos em barras

Os gráficos em barras são comumente usados para exibir distribuição de frequências para as variáveis qualitativas, como por exemplo a variável Estado Civil do Exemplo 2.3.1, como mostra a Figura 2.4.3. Podemos também representar a tabela de frequência simples apresentada na Figura 2.4.3 pelo Gráfico em Setores, como mostra a Figura 2.2 e 2.3. O gráfico em setores é comumente utilizado para representar parte de um todo, geralmente em porcentagens, e é, também, apropriado para variáveis qualitativas.

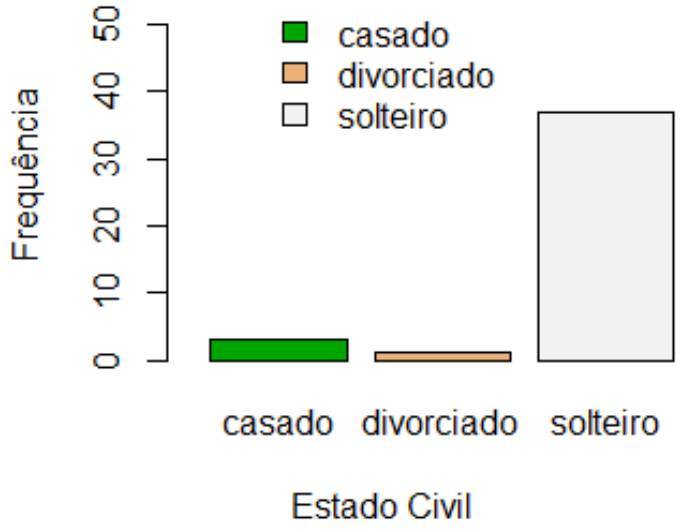


Tabela de frequência simples.

Estado Civil	n_i
Casado	3
Divorciado	1
Solteiro	37
Total	41

Figura 2.1: Gráfico de frequência de estudantes por estado civil.

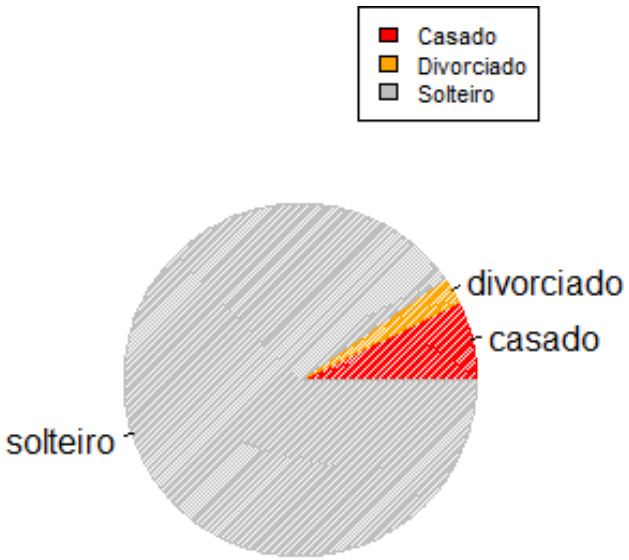


Figura 2.2: Frequência de estudantes por estado civil.

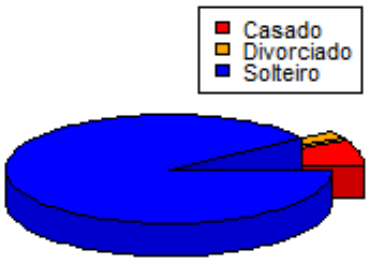


Figura 2.3: Frequência de estudantes por estado civil.

Histograma e o Polígono de Frequência

O histograma é um gráfico de barras contíguas apropriado para representar distribuições de frequências de variáveis quantitativas (contínuas ou discretas com muitos valores possíveis) e é um

dos mais importantes gráficos no estudo de frequência de variáveis quantitativa. Sua construção é similar ao procedimento empregado para construir uma tabela de frequências em intervalos de classes.

As barras contíguas do histograma têm bases de mesma amplitude. Caso a tabela de frequência em intervalos de classes já tenha sido construída, as barras do histograma podem ser construídas de modo a serem proporcionais aos intervalos das classes dessa tabela. É importante que as áreas das barras sejam proporcionais as suas correspondentes frequência absoluta (n_i) ou relativa (f_i), sendo mais usual utilizar as frequências relativas (ou as porcentagens, $P_i = 100 \times f_i$).

Para construir o histograma de modo que as áreas das barras sejam dadas pelas frequências relativas, devemos obter a altura (h_i) de cada retângulo (ou barra) de modo que $h_i = f_i/\Delta$ (ou $h_i = n_i/\Delta$ caso seja utilizada a frequência absoluta), em que Δ é a amplitude das bases das barras, para $i = 1, \cdots, k$ com k sendo o número de barras (ou classes na tabela). Deste modo, quanto mais dados contiver a i-ésima classe, mais alto será o i-ésimo retângulo.

Como exemplo, vamos construir um histograma a partir da tabela de frequência para variável altura de um estudante, Tabela 2.7. Para isso, seguiremos os seguintes passos.

- vamos considerar k=5 classes com $L_{inf} = 1,48$ e $L_{sup} = 1,98$;
- obtemos $AT = 1,98 - 1,48 = 0,5$ é a amplitude total considerada;
- em seguida obtemos a largura das barras, ou amplitude, $\Delta = \frac{AT}{k} = \frac{0,5}{5} = 0,1$;
- com essa amplitude obtemos os limites das bases de cada barra de modo semelhante ao que foi feito na tabela de frequência;
- para cada amplitude, obtemos a altura da base de modo que $h_i = f_i/\Delta$.

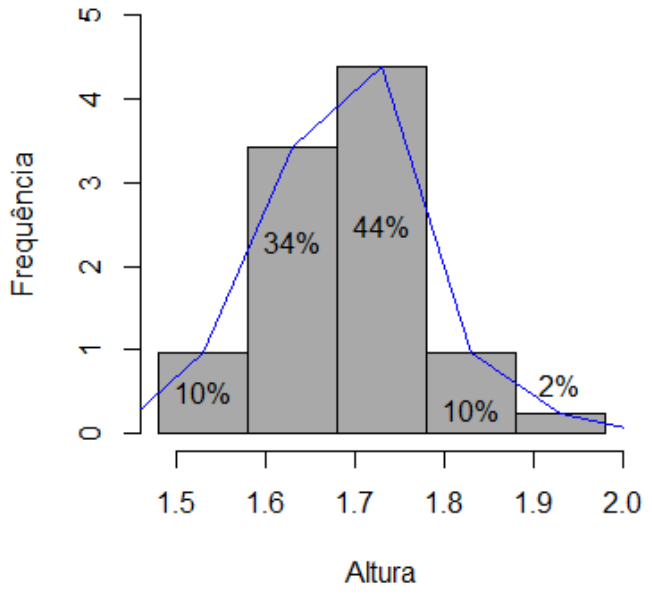


Figura 2.4: Frequência dos estudantes segundo altura.

Frequência dos alunos segundo altura.

X	n_i	f_i
$1,48 \dashv 1,58$	4	$4/41 \approx 0,10$
$1,58 \dashv 1,68$	14	$14/41 \approx 0,34$
$1,68 \dashv 1,78$	18	$18/41 \approx 0,44$
$1,78 \dashv 1,88$	4	$4/41 \approx 0,10$
$1,88 \dashv 1,98$	1	$1/41 \approx 0,02$
Total	41	1

A Figura 2.4.3 mostra o histograma construído a partir da tabela de frequência, de modo que a área de cada barra é igual a porcentagem de dados no intervalo de sua base, para isso basta multiplicar a altura de cada barra por 100, ou seja $h_i = 100 \times f_i/\Delta$. A parti deste histograma, podemos perceber que a grande maioria dos estudantes tem entre 1,60 e 1,80 metros de altura, correspondendo à 78% dos indivíduos desse grupo. Iss quer dizer que se sorteássemos ao acaso

(aleatoriamente) um indivíduo desse grupo, teríamos 78% de chance de escolhermos alguém com a altura neste intervalo. Note que, pelo ponto médio do topo de cada barra, foi traçado uma linha (curva). Essa curva é denominada **polígono de frequência** e é uma ferramenta bastante útil para para dar uma ideia do comportamento da frequência de ocorrência da variável em questão. O histograma, assim como o polígono de frequência, é denominado densidade empírica da variável.

2.4.4 Exercícios para a Seção 2.4.3

Exercício 2.4.4. *Uma distribuição de frequências para os níveis séricos (em microgramas por decilitro) de zinco de 462 homens entre as idades de 15 a 17 anos é exibida na Tabela 2.8.*

Tabela 2.8: Nível sérico de zinco (mg/dl).

Nível	Número de homens
$50 \vdash 60$	6
$60 \vdash 70$	35
$70 \vdash 80$	110
$80 \vdash 90$	116
$90 \vdash 100$	91
$100 \vdash 110$	63
$110 \vdash 120$	30
$120 \vdash 130$	5
$130 \vdash 140$	2
$140 \vdash 150$	2
$150 \vdash 160$	2
Total	462

1. *Complete a tabela de frequências. O que você pode concluir sobre essa distribuição de níveis séricos de zinco?*

2. *Elabore um histograma dos dados. Descreva a forma do histograma.*

Exercício 2.4.5. *A distribuição de frequência na Tabela 2.9 exibe os números de casos pediátricos de Aids registrados nos EUA entre 1983 e 1989.*

Tabela 2.9: Casos pediátricos de Aids registrados nos EUA.

Ano	Número de casos
1983	122
1984	250
1985	455
1986	848
1987	1412
1988	2811
1989	3098

Construa um gráfico de barras que mostre o número de casos por ano. O que o gráfico lhe conta sobre a Aids pediátrica nesse período?

Exercício 2.4.6. *Contou-se o número de erros de impressão da primeira página de um jornal durante 50 dias, obtendo-se os resultados: 8, 11, 8, 12, 14, 13, 11, 14, 14, 5, 6, 10, 14, 19, 6, 12, 7, 5, 8, 8, 10, 16, 10, 12, 12, 8, 11, 6, 7, 12, 7, 10, 14, 5, 12, 7, 9, 12, 11, 9, 14, 8, 14, 8, 12, 10, 12, 12, 7, 15.*

Represente os dados graficamente.

2.5 Medidas de Resumo

Frequentemente é útil descrever numericamente as características dos dados observados a partir de variáveis quantitativas, uma vez que sua natureza permite o calculo de algumas medidas que podem ser uteis para dar uma ideia do comportamento dessas variáveis.

- Dentre as medidas de resumo existentes, destacam-se:
  - as **medidas de posição,**
    - ∗ média,
    - ∗ moda e
    - ∗ separatrizes;
  - as **medidas de dispersão,**
    - ∗ variância,
    - ∗ desvio-padrão,
    - ∗ distância interquartílica e
    - ∗ coeficiente de variação.

As medidas de posição localizam a distribuição de frequência da variável no eixo das abcissas, enquanto as medidas de dispersão fornecem informações sobre o "espalhamento" dessa distribuição.

Vamos iniciar nossos estudos sobre as medidas de resumo pelas medidas de posição, as quais podem apresentar-se de várias formas, sendo que as mais importantes são:

- **medidas de tendência central**, são assim denominadas devido a tendência dos dados observados se agruparem em torno desses valores (média, moda e mediana);

- **separatrizes**, são assim chamadas porque separam, dividem um conjunto de dados ordenado em partes percentuais iguais (quartil, decil e percentil, genericamente quantís).

2.5.1 Medidas de Tendencia Central

A Figura 2.5 mostra o histograma para a variável altura dos estudantes construído a partir dos dados brutos, apresentado no Exemplo 2.3.1. Neste gráfico podemos observar geometricamente a média desse conjunto de dados, que se localiza próxima ao centro da distribuição de frequência representada pelo histograma.

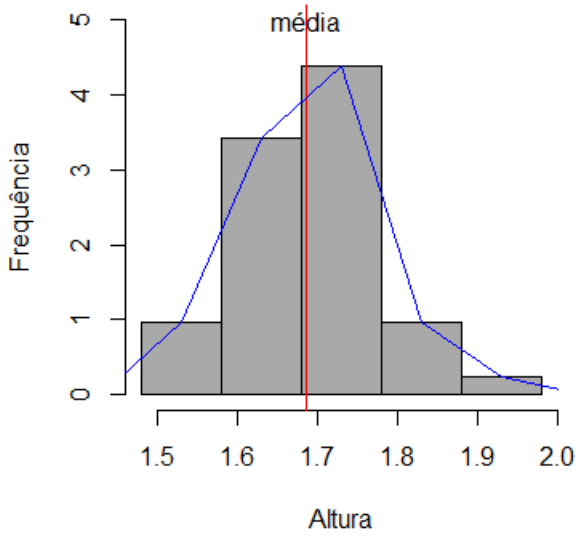


Figura 2.5: Frequência dos estudantes segundo a altura.

Média a partir da série de dados

Vamos considerar uma série de dados $(x_1, \ldots, x_n)$, ou $(x_1, \ldots, x_N)$, a partir desse conjunto podemos obter a média como:

- $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$, para uma amostra
- $\mu = \dfrac{\sum_{i=1}^{n} x_i}{N}$, para uma população inteira, sendo:
  - $\overline{x}$ a notação para média amostral,
  - μ a notação para média populacional,
  - n a quantidade de elementos na amostral e
  - N a quantidade de elementos na população.

Média a partir da tabela de frequência simples

Para exemplificar a obtenção da média a partir de uma tabela de frequência simples, consideremos a distribuição de frequência de uma amostra de 34 famílias de quatro filhos quanto ao número de filhos do sexo masculino apresentada na Tabela 2.10. Neste caso temos que o número médio de filhos homens por família é dado por:

Tabela 2.10: Frequência de famílias com quatro filhos segundo número de filhos do sexo masculino.

Número de meninos	n_j
0	2
1	6
2	10
3	12
4	4
Total	34

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i n_i}{n} = \frac{(0 \times 2) + (1 \times 6) + (2 \times 10) + (3 \times 12) + (4 \times 4)}{34} = \frac{78}{34} \cong 2,3$$

Assim a média desta amostra é de 2,3 meninos por família.

Genericamente, temos

- $\overline{x} = \dfrac{\sum_{j=1}^{k} n_j x_j}{n}$, para uma amostra,
- $\mu = \dfrac{\sum_{i=1}^{k} n_j x_j}{N}$, para uma população inteira,

  em que
  - x_j é o j-ésimo valor possível da variável (valores sem as repetições e ordenados),
  - n_j é a j-ésima frequência absoluta e
  - k é o número de classes da tabela de frequência.

Média a partir da tabela de frequência em intervalos de classe

A tabela a seguir representa a idade dos estudantes do curso de medicina veterinária da UFBA, ano/1993. A partir dessa tabela, vamos calcular a idade média desses alunos. Para isso, consideremos duas colunas extras na tabela: a coluna dos pontos médios das classes e a coluna dos pontos médios multiplicados pelas frequências absolutas ($x_j.n_j$).

Classe de Idade	n_j	Ponto médio das classes (x_j)	$x_j.n_j$
$21 \vdash 24$	7	$\frac{21+24}{2}$ =22,5	157,5
$24 \vdash 27$	8	$\frac{24+27}{2}$ = 25,5	204
$27 \vdash 30$	1	$\frac{27+30}{2}$=28,5	28,5
$30 \vdash 33$	5	$\frac{30+33}{2}$ =31,5	157,5
$33 \vdash 36$	7	$\frac{33+36}{2}$=34,5	241,5
Total	28	142,5	789

- Neste caso, temos:

$$\overline{x} = \frac{\sum_{j=1}^{k} x_j.n_j}{n} = \frac{789}{28} \cong 28,18$$

  Logo a idade média dos alunos é de aproximadamente 28,2 anos.

Então, pra o cálculo da média, convencionamos que todos os valores incluídos em um determinado intervalo da classe coincidem com seu ponto médio, assim determinamos a média aritmética ponderada aproximada por

- $\overline{x} = \dfrac{\sum_{j=1}^{k} x_j n_j}{n}$ (para uma amostra),

- $\mu = \dfrac{\sum_{j=1}^{k} x_j n_j}{N}$ (para uma população),

  em que:

  - x_j representa o ponto médio da j-ésima classe,
  - n_j é a j-ésima frequência absoluta e
  - k é o número de classes da tabela de frequência.

Moda de uma série de dados

A moda é o valor que ocorre com maior frequência em uma série de dados.

Exemplo 2.5.1. *Vamos obter a moda nos seguintes casos:*

- *considerando a série: (7 , 8 , 9 , 10 , 10 , 10 , 11 , 12),*
  neste caso a moda é igual a 10;
- *considerando a série: (3 , 5 , 8 , 10 , 12), neste caso a série é* ***amodal;***
- *considerando a série: (2 , 3 , 4 , 4 , 4 , 5 , 6 , 7 , 7 , 7 , 8 , 9),*
  neste caso apresentam-se duas modas: 4 *e* 7*, a série é* ***bimodal.***

Em situações em que a série apresenta mais de duas modas, dizemos que a série é **multimodal**.

Moda a partir de uma tabela de frequência simples

Uma vez que os dados encontram-se agrupados em uma tabela de frequência simples, é possível obter imediatamente a moda: basta observar o valor da variável de maior frequência.

Exemplo 2.5.2. *Consideremos a distribuição relativa a 34 famílias de quatro filhos tomando para a variável o número de filhos do sexo masculino apresentada na Tabela 2.10 e repetida abaixo:*

Número de meninos	n_j
0	2
1	6
2	10
3	**12**
4	4
Total	34

Observe que a moda da variável número de meninos é $M_o = 3$*, conforme destacado acima. A classe destacada é denominada* ***classe modal.***

Moda a partir de uma tabela de frequência simples em intervalos de classe

Se os dados estão agrupados em uma tabela de frequência em intervalos de classe, devemos aproximar a moda dentro da classe modal. Aqui faremos essa aproximação de modo similar ao que foi feito no caso da média, ou seja, obter o ponto médio da classe.

Exemplo 2.5.3. *Veja a seguir os dados de idade dos alunos do curso de medicina veterinária da UFBA, ano/1993.*

Classe de Idade	n_i	x_i
$21 \vdash 24$	7	–
$\mathbf{24 \vdash 27}$	**8**	**25,5**
$27 \vdash 30$	1	–
$30 \vdash 33$	5	–
$33 \vdash 36$	7	–

Para esse exemplo, a moda é $M_o = 25,5$, ou seja, há uma maior quantidade de alunos com idade de 25,5 anos, aproximadamente.

Mediana de uma série de dados

A média, embora seja uma medida de tendência central muito utilizada, muitas vezes não descreve de maneira adequada um conjunto de dados, pois essa é uma medida que pode ser afetada por algumas características que os dados pode conter, como por exemplo a presença de assimetria acentuada na distribuição dos dados, ou presença de pontos que destoam dos demais, seja para cima ou para baixo. Nessas situações é necessário procurar medidas que não sejam afetadas por essas características. Uma medida que pode ser empregada nessas situações é a mediana, pois esta não é afetada por assimetria ou por pontos atípicos.

A mediana de um conjunto de valores ordenados $(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$, é o valor situado de tal forma no conjunto que o separa em dois subconjuntos, de mesmo número de elementos. Aqui, $x_{(1)}$ corresponde ao **valor mínimo** da série e $x_{(n)}$ corresponde ao **valor máximo** da série de dados.

A mediana é considerada uma separatriz, por dividir a distribuição ou o conjunto de dados em duas partes iguais.

Para a obtenção da mediana de uma variável X, devemos considerar:

$$Med(X) = \begin{cases} x_{(\frac{n+1}{2})}, & \text{se n ímpar;} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & \text{se n par.} \end{cases}$$

Exemplo 2.5.4. *Vamos considerar os exemplos a seguir.*

Para a série $X = (5, 2, 6, 13, 9, 15, 10)$, temos a seguinte ordenação:

$$(\underbrace{2, 5, 6,}_{3\ elementos}\ \boxed{9,}\ \underbrace{10, 13, 15}_{3\ elementos}) \Rightarrow Med(X) = 9$$

Para a série $Y = (1, 3, 0, 0, 2, 4, 1, 3, 5, 6)$, temos a ordenação:

$$(\underbrace{0, 0, 1, 1,}_{4\ elementos}\ \boxed{2,3,}\ \underbrace{3, 4, 5, 6}_{4\ elementos}) \Rightarrow Med(Y) = \frac{2+3}{2} = 2,5$$

Mediana a partir da tabela de frequência simples

Para obtenção da mediana a partir de uma tabela de frequência, vamos olhar, inicialmente para a coluna das frequências relativas acumuladas, pois a mediana é um valor que contém abaixo dele 50% dos dados, com isso podemos encontrar facilmente a classe mediana diretamente da tabela.

Exemplo 2.5.5. *Consideremos a distribuição relativa a 34 famílias de quatro filhos tomando para a variável o número de filhos do sexo masculino:*

Número de meninos	n_i	f_{ac}
0	*2*	*2/34 ≈ 0,06*
1	*6*	*8/34≈0,24*
2	**10**	**18/34≈0,53**
3	*12*	*30/34≈0,88*
4	*4*	*34/34=1*
Total	*34*	*1*

- *Para se obter a mediana, observe que a até a terceira classe acumulam-se mais de 50% dos dados (53% dos dados), sendo assim, esta é a classe que contém a mediana com sobra, deste modo não importa se o total de elementos na série é par ou ímpar, a mediana é o valor que está nessa classe.*

- *Assim, a mediana é dada por* $Med = 2$ *meninos.*

Se n é par e existe uma classe que concentra até ela exatamente 50% dos dados, devemos obter a média aritmética simples entre o valor da classe mediana e o valor imediatamente posterior.

Exemplo 2.5.6. *Como exemplo, veja a distribuição de frequência abaixo:*

x_i	n_i	f_{ac}
12	1	1/8
14	2	3/8
15	1	**4/8=0,5**
16	**2**	**6/8=0,75**
17	1	7/8
20	1	8/8
Total	8	1

Na tabela acima, tem-se $f_{ac_3} = 0,5$ *com n par, assim sabemos que existem dois valores diferente ocupando a posição central dos dados ordenados, logo:*

$$Med = \frac{15 + 16}{2} = 15,5$$

Mediana a partir da tabela de frequência em intervalos de classe

A obtenção da mediana de dados agrupados em uma tabela de frequência em intervalos de classes é feira, primeiramente, localizando a classe que contém a mediana, assim como no caso de uma tabela de frequência simples. No entanto, o valor da mediana não pode ser obtido exatamente, exigindo, assim, uma aproximação dentro do intervalo que contém esse valor. Essa aproximação será feita aqui de modo a levar em consideração a distribuição de frequência por meio da relação:

$$Med = L_i + \left[\frac{n\left(0,5 - f_{ac(ant)}\right)}{n_i}\right] \times \Delta$$

em que,

- L_i: limite inferior da classe mediana,
- $f_{ac(ant)}$: frequência relativa acumulada da classe anterior à classe mediana,
- Δ: amplitude da classe e
- n_i é a frequência absoluta (contagem) da classe mediana.

Exemplo 2.5.7. *A tabela a seguir representa a idade dos alunos do curso de medicina veterinária da UFBA, ano/1993.*

Classe de Idade	n_i	f_{ac}
$21 \vdash 24$	7	$7/28$
$\boldsymbol{24 \vdash 27}$	$\boldsymbol{8}$	$\boldsymbol{15/28 \approx 0,54}$
$27 \vdash 30$	1	$16/28$
$30 \vdash 33$	5	$21/28$
$33 \vdash 36$	7	$28/28$
Total	28	1

Aqui, a classe que contém a mediana é a segunda, pois mais de 50% dos valores estão acumulados até essa classe. Para este exemplo, temos:

- $L_i = 24$,
- $f_{ac(ant)} = \frac{7}{28} = 0,25$,
- $\Delta = 27 - 24 = 3$,
- $n_i = 8$ *e*
- $n = 28$

Portanto,

$$\begin{aligned} Med &= L_i + \left[\frac{n\left(0,5 - f_{ac(ant)}\right)}{n_i}\right] \times \Delta \\ &= 24 + \frac{28(0,5 - 0,25)}{8} \times 3 \\ &= 24 + \frac{21}{8} = \boldsymbol{26,63} \end{aligned}$$

2.5.2 Separatrizes

Separatrizes (ou **quantis**) são os valores que dividem a série de dados ordenados $(x_{(1)}, x_{(2)}, \cdots, x_{(n)})$ em partes iguais. Vimos que a mediana divide a série dados ordenados em duas partes iguais, então a mediana é uma separatriz. No entanto, essa não é a única medida relevante que divide uma série de valores. Temos outras separatrizes também importantes. Como os **quartis**, os **decis** e os **percentis**.

- Os **quartis** são os valores da variável que dividem uma série de dados ordenados em quatro partes iguais, portanto são três medidas que são denotadas por Q_1 (primeiro quartil), Q_2(segundo quartil) e Q_3(terceiro quartil).
  - o primeiro quartil, Q_1, é o valor que divide aproximadamente, a quarta parte (25%) das observações abaixo dele, e os 75% restantes, acima dele.
  - O segundo quartil é exatamente a mediana ($Q_2 = Med$).
  - O terceiro quartil, Q_3, tem aproximadamente os três quartos (75%) das observações abaixo dele e os demais 25% acima.

A estratégia para a obtenção dos quartis é semelhante aquela empregada para se obter a mediana, ou seja, primeiramente temos que encontrar a classe que contém o quartil desejado. Para isso, basta observar as frequências relativas acumuladas. Com isso, podemos usar as equações:

$$p_j = j \cdot \frac{1}{4}, \text{ para } j = 1, 2, 3 \tag{2.1}$$

$$Q_j = L_i + \left[\frac{n\left(p_j - f_{ac(ant)}\right)}{n_i} \right] \times \Delta \tag{2.2}$$

com,

- L_i é o limite inferior da classe definida por p_j;
- $f_{ac(ant)}$ é a frequência absoluta acumulada da classe anterior à que contém o j-ésimo quartil;
- Δ é a amplitude da classe e
- n_i é a frequência da classe definida por p_j.

Obtenção dos quartis a partir do histograma

Exemplo 2.5.8. *Observe a Figura 2.6, em que é mostrada uma série de dados agrupados em intervalos de classe. Vamos determinar os quartis geometricamente, utilizando essa figura.*

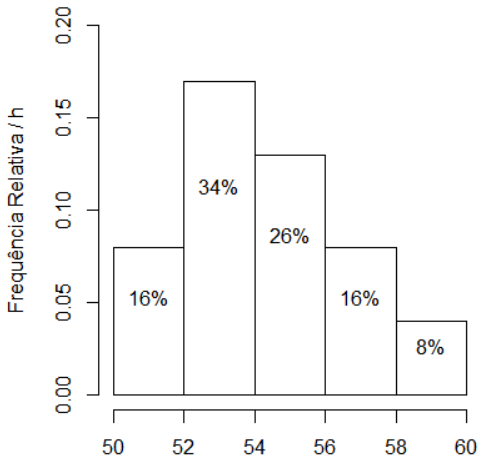


Figura 2.6: Intervalos de tempo (em minutos) de montagem de 50 equipamentos.

Em cada barra podemos observar a porcentagem de dados em cada base dos retângulos, então, é fácil concluir que o primeiro quartil (Q_1) está contido no intervalo $52 \vdash 55$*, pois até esse intervalos acumulam-se 25% dos dados. A aproximação desse quartil dentro do intervalo pode, também, ser aproximado considerando que os dados estão dispostos uniformemente dentro do intervalo. Com isso, e levando em consideração que a área de cada retângulo corresponde a porcentagem de dos dados na classe, podemos aproximar facilmente o primeiro quartil da seguinte forma:*

$$\begin{aligned} 34 - 25 &= 9\% \\ \textit{altura da barra} &= 34/2 = 17 \\ \delta &= 9/17 \approx 0,53 \\ Q_1 = 52 + \delta &= 52 + 0,53 = 52,53 \end{aligned}$$

A obtenção dos demais quartis será deixada como exercício.

Decis e Percentis

- Se dividimos o conjunto em dez partes iguais temos os **decis**.
- Quando dividimos em cem partes, temos os **percentis**,

Para a obtenção dos decis(D_j) e percentis (P_j) podemos utilizar as seguintes expressões:

$$p_j = \frac{j}{10}, \text{ para } j = 1, 2, 3, .., 9 \text{ (para os decis)}$$

$$p_j = \frac{j}{100}, \text{ para } j = 1, 2, 3, .., 99 \text{ (para os percentis)}$$

$$D_j = P_j = L_i + \left[\frac{n\left(p_j - f_{ac(ant)}\right)}{n_i} \right] \times \Delta$$

Observe que existem relações entre quartis, decis e percentis. $Q_1 = P_{25}$, $Q_2 = D_5 = P_{50}$, $Q_3 = P_{75}$, por exemplo.

2.5.3 Boxplot

O Boxplot, também conhecido como Desenho Esquemático, é um gráfico bastante útil na análise do comportamento de uma variável a partir de um conjunto de valores observados. Dentre as vantagens do boxplot, podemos destacar:

- a detecção rápida de uma possível assimetria na distribuição de frequência dos dados;
- a capacidade de fornecer uma ideia sobre a existência de possíveis pontos atípicos (muito além ou muito aquém dos demais pontos);
- a exibição dos quartis.

Para sua construção, vamos obter mais duas medidas para decidir quais são os pontos atípicos da série de dados. Vamos chamar essas medidas de **limite superior** (l_{sup}) e **limite inferior** (l_{inf}). Para obtê-los, fazemos:

$$l_{inf} = Q_1 - \frac{3}{2}(Q_3 - Q_1)$$
$$l_{sup} = Q_3 + \frac{3}{2}(Q_3 - Q_1).$$

Com essas medidas, podemos obter os valores que estão muito aquém de Q_1 ou muito além de Q_3. Tais pontos são chamados de **pontos discrepantes (ou aberrantes, ou ainda outliers)**.

Após a obtenção dos limites (l_{inf} e l_{sup}), podemos construir o boxplot da seguindo os seguintes passos:

1. No eixo cartesiano, constrói-se um retângulo na vertical de modo que:
   - A base no retângulo corresponde ao primeiro quartil (Q_1) e
   - o topo (lado superior) corresponde ao terceiro quartil (Q_3);
2. divide-se o retângulo em duas partes por meio de um segmento de reta orientado pela mediana;
3. Acima do retângulo traça-se um segmento orientado por l_{sup}
4. Abaixo do retângulo também é apresentado um traço orientado por l_{inf}
5. acima de l_{sup} e abaixo de l_{inf}, marca-se, geralmente com os pontos discrepantes.

A Figura 2.5.3, mostra o histograma para os dados de médias de notas da prova Brasil do ano de 2015 dos municípios do estado do Ceará. Neste gráfico podemos perceber que não existem assimetria nos dados, pois a mediana se encontra no centro do retângulo. Além disso, observamos que existem pontos acima e abaixo dos limites traçados, indicando presença de pontos atípicos, ou seja, existem municípios cujas médias estão muito acima, ou muito abaixo, das notas dos demais municípios.

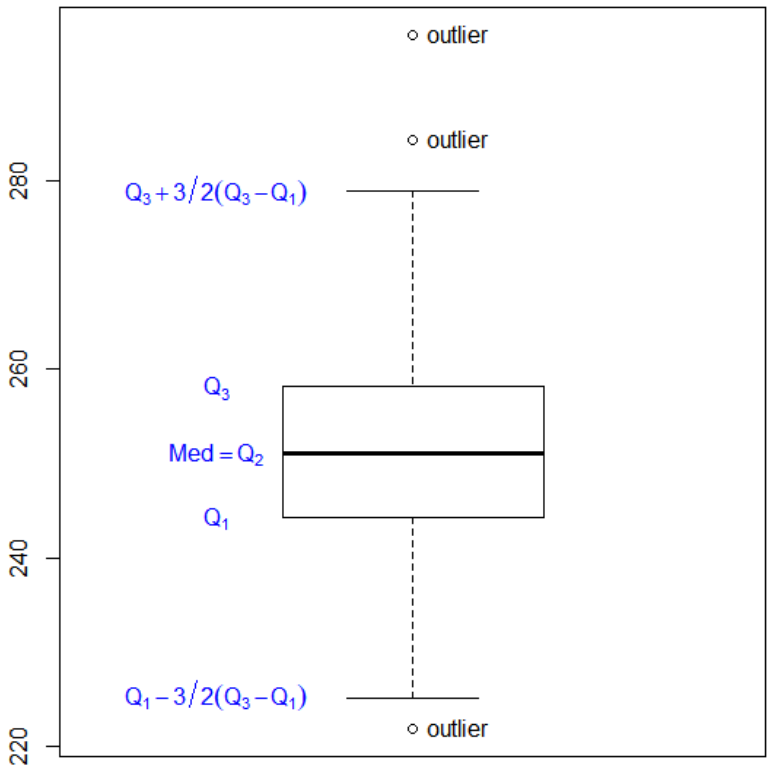


Figura 2.7: Boxplot para os dados de desempenho na prova de português dos estudantes do estado do Ceará na prova Brasil de 2015.

2.5.4 Exercícios para a Seção 2.5

Exercício 2.5.1. *Num determinado país a população feminina representa 53% da população total. Sabendo-se que a idade média (média aritmética das idades) da população feminina é de 38 anos e a da masculina é de 35 anos. Qual a idade média da população?*

Exercício 2.5.2. *Obtenha o segundo e o terceiro quartil dos dados de montagem de equipamentos representados pelo histograma apresentado na Figura 2.6.*

Exercício 2.5.3. *Os dados do quadro a seguir mostram os diâmetros abdominais em centímetros de 36 indivíduos adultos.*

59	60	60	60	62	63	63	63	63	64	66	66
66	68	69	69	69	70	71	74	75	75	77	78
81	85	86	86	87	88	88	91	95	101	107	120

Fonte: Academia Aquarius

a) Construa a tabela de distribuição de frequência em intervalos de classes, apresentando as frequências absolutas, as frequências relativas, as frequências acumuladas e o ponto médio de cada classe.

b) Obtenha a média a partir da tabela de frequência.

c) A média é uma boa medida de tendência central para este conjunto de dados? Justifique.

Exercício 2.5.4. *O gráfico apresentado na Figura 2.8 representa a variável nível de potássio no plasma em miliequivalentes (mEq), foram examinadas 20 pessoas.*

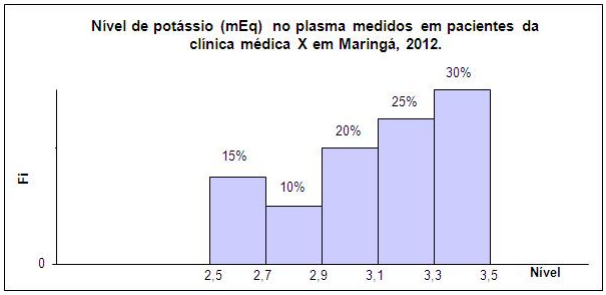


Figura 2.8: Histograma para a variável nível de potássio no plasma em miliequivalentes (mEq).

a) Calcule os quartis (Q_1, Q_2 e Q_3). Interprete-os.

b) Qual a porcentagem exata de nível de potássio acima de 3,15 mEq?

c) Faça o boxplot para esse conjunto de dados e interprete-o.

Exercício 2.5.5. *Os dados a seguir representam a fração de colesterol em miligramas por decilitro (mg/dl) fornecida pelo laboratório Vida e Saúde, de 9 indivíduos do sexo feminino.*

27,0	22,0	24,0	30,2	14,5	30,0	43,0	29,0	53,0

a) Qual o nível médio de colesterol deste grupo de indivíduos?

b) Qual o nível mediano e qual a moda?

Exercício 2.5.6. *Um hospital tem o interesse em determinar a altura média dos pacientes de uma determinada área e relacioná-la com a incidência de determinada anomalia ortopédica. Foram selecionados 80 pacientes e as alturas (em m) podem ser encontradas na tabela abaixo.*

Altura dos pacientes

1,72 1,78 1,87 1,86 1,79 1,79 1,83 1,74 1,64 1,62 1,75 1,65 1,75 1,58 1,63 1,77 1,64 1,68 1,66 1,82 1,68 1,80 1,74 1,76 1,74 1,72 1,75 1,89 1,73 1,76 1,72 1,71 1,63 1,81 1,65 1,58 1,63 1,70 1,73 1,57 1,75 1,64 1,73 1,70 1,75 1,56 1,70 1,68 1,68 1,79 1,75 1,71 1,62 1,83 1,72 1,76 1,67 1,82 1,67 1,60 1,67 1,61 1,61 1,67 1,75 1,80 1,70 1,77 1,73 1,77 1,64 1,66 1,74 1,66 1,66 1,79 1,68 1,79 1,69 1,80

Use o computador para resolver os itens abaixo.

a) (Feito na lista 1) Construa uma tabela de frequência agrupando os dados em intervalos de classe.

b) Construa o histograma.

c) Calcule o primeiro, segundo e o terceiro quartil e interprete-os.

c) Construa o boxplot.

2.6 Medidas de Dispersão

O resumo de um conjunto de dados por uma única medida de tendência central esconde toda a informação sobre a variabilidade do conjunto de observações. Por essa razão, precisamos empregar outro tipo de medida que informe sobre quão dispersos os dados estão.

Por exemplo, suponhamos que se deseja comparar a performance de dois empregados, com base na seguinte produção diária de determinada peça:

Funcionário	Variáveis	Total
A	70; 71; 69; 70; 70	350
B	60; 80; 70; 59; 81	350

Observe que $\overline{x}_A = 70$ e $\overline{x}_B = 70$, de acordo com as médias diríamos que a performance de B é igual a de A, no entanto se observarmos a variabilidade nos dados, observamos que a performance de A é bem mais uniforme.

Dependendo da medida de tendência central empregada, devemos adotar uma medida de dispersão apropriada. Aqui abordaremos as seguintes medidas de dispersão:

- variância e desvio-padrão,
- distância interquartil e
- coeficiente de variação.

2.6.1 Variância

A medida de tendência central mais comumente utilizada é a média. Uma vez que essa medida é adotada para descrever a posição da distribuição dos dados, faz-se necessário a escolha de uma medida da variabilidade em torno dessa média. Neste caso, a variância e o desvio padrão podem ser adotados.

Variância a partir de uma série de dados

A variância é a medida que fornece o grau de dispersão, ou variabilidade dos valores do conjunto de observações em torno da média. Ela é calculada tomando-se a média dos quadrados dos desvios em relação à média. Ou seja,

para uma população:

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

Em que,

- μ é a média da população.
- N é a quantidade de elementos na população;

no caso de uma amostra temos a variância amostral:

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

Em que

- n_j é a frequência da j-ésima classe.
- $\bar{x}$ é a **média da amostra** ou **média amostral**.

Variância a partir de uma tabela de frequência

Se os dados estão apresentados em uma tabela de frequência, a variância é obtida tomando-se a **média ponderada** dos quadrados dos desvios dos valores possíveis da variável (ou dos ponto médio das classes).

$$\sigma^2 = \frac{\sum_{j=1}^{K}(x_j - \mu)^2.n_j}{N} \qquad \text{ou} \qquad S^2 = \frac{\sum_{j=1}^{K}(x_j - \mu)^2.n_j}{n-1}$$

Em que

- n_j é a frequência da j-ésima classe.
- K é o número de classes na tabela.
- x_j é o j-ésimo valor possível da variável (ou ponto médio da classe).

2.6.2 Desvio-padrão

Como a variância é uma medida de dimensão igual ao quadrado da dimensão dos dados, pode-se causar problemas de interpretação. Então costuma-se usar o desvio padrão, que é definido como a raiz quadrada da variância, como segue:

$$\sigma = \sqrt{\sigma^2} \text{ (para uma população) ou } S = \sqrt{S^2} \text{ (para uma amostra).}$$

Propriedades do desvio padrão e da variância:

1. Somando (ou subtraindo) um valor constante e arbitrário, C a cada elemento de um conjunto de números, o desvio padrão desse conjunto não se altera, essa propriedade também vale para variância.

2. Multiplicando (ou dividindo) por um valor constante C, cada elemento de um conjunto de números, o desvio padrão fica multiplicado (ou dividido) pela constante C,

3. no caso da variância ela fica multiplicada pela constante elevada ao quadrado.

2.6.3 Distância interquartil

Se a mediana é usada como medida de tendencia central, a distância entre o primeiro e o terceiro quartil pode ser usada como uma medida da variabilidade dos dados em torno da mediana. Essa medida é chamada de distância "interquartil" e é dada por:

$$D = Q_3 - Q_1$$

Também é muito utilizado a "amplitude ou desvio semi-quartil", que seria o interquartil dividido por 2. Neste caso, essa é uma boa medida de dispersão, pois em um intervalo igual ao interquartil em torno da mediana estão 50% dos dados. Neste caso, o boxplot pode ser utilizado para visualizar o comportamento da variável que gerou os dos dados.

2.6.4 Amplitude total

Também podem ser usadas outras medidas para se ter uma ideia da dispersão dos dados. Um exemplo é a Amplitude Total (AT) que é a diferença entre o maior e o menor valor observado.

$$AT = x_{(\text{máx})} - x_{(\text{ mín})}$$

Dados agrupados em classes: Neste caso a AT é dada pela diferença entre o limite superior da última classe e o limite inferior da primeira classe.

$$AT = L_s - L_i$$

Obs: essa não é muito utilizada devido ser altamente afetada por pontos discrepantes, além de ser pouco informativa.

2.6.5 Coeficiente de Variação

O coeficiente de variação é uma medida de variabilidade relativa, que é definida como a razão entre o desvio padrão e a média. Assim, essa é uma medida expressa em percentual e é dada por:

$$CV\% = \frac{\sigma}{\mu} \times 100.$$

Note que o coeficiente de variação não tem unidade, pois o desvio-padrão e a média estão na mesma unidade, fazendo com que estas se cancelem.

Com o coeficiente de variação, podemos avaliar se a média é ou não uma boa medida de tendencia central.

Critérios para interpretação.

Quanto menor for o coeficiente de variação, mais representativa dos dados será a média. Coeficiente de variação acima de 50%, a média não é representativa e outra medida de tendencia central deve ser utilizada, como a mediana por exemplo.

Uma interpretação para o coeficiente de variação

- Se $0\% \leq CV\% < 30\%$, conclui-se pela baixa variabilidade dos dados e a média é uma ótima medida para representar os dados;
- Se $30\% \leq CV\% < 50\%$, conclui-se pela média variabilidade dos dados e a média é uma boa medida para representar os dados;
- Se $CV\% \geq 50\%$, conclui-se pela alta variabilidade dos dados e a média não é uma medida apropriada para representar os dados. Neste caso, deve-se pensar na mediana ou moda.

Por ser adimensional, o coeficiente de variação é uma boa opção para se comparar o grau de concentração dos dados em torno da média de séries de dados distintas, mesmo que tenham unidades diferentes.

Interpretação em Controle Estatístico de Processo

- Vários autores indicam diferentes métodos para se classificar o coeficiente de variação.
- Além disso, essa medida é intrínseca a cada processo, sendo muito utilizado na área agrícola, mais especificamente na experimentação agronômica.
- Em controle estatístico de processo, também pode-se utilizar a interpretação mostrada no quadro abaixo:

Faixa	CV %	Dispersão
menor ou igual a 15%	baixo	baixa dispersão dos dados
entre 15% e 30%	médio	média dispersão dos dados
maior que 30%	alto	alta dispersão dos dados

Exemplo 2.6.1. *Voltando ao exemplo da performance dos dois empregados, vamos calcular a variância, o desvio padrão e o coeficiente de variação dos dois conjuntos de valores de produção diaria dos empregados A e B:*

Empregado	*Variáveis*	Σ
A	*70; 71; 69; 70; 70*	*350*
B	*60; 80; 70; 59; 81*	*350*

Vimos que: $\bar{X_A} = 70$ *e* $\bar{X_B} = 70$. *Vamos agora obter medidas de dispersão para esse conjunto de valores.*

- *Variância de A.*

$$S_A^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{(70-70)^2+(70-71)^2+(70-69)^2+(70-70)^2+(70-70)^2}{5-1} = \frac{1+1}{4} = \frac{2}{4} = 0,5$$

- *Variância de B.*

$$S_B^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1} = \frac{(70-60)^2+(70-80)^2+(70-70)^2+(70-59)^2+(70-81)^2}{5-1}$$
$$= \frac{100+100+121+121}{4} = \frac{442}{4} = 110,5$$

- *Desvio padrão e coeficiente de variação de A.*

$$S_A = \sqrt{0,5} = 0,7 \qquad e \qquad CV_A\% = \frac{0,7}{70}.100 = 1\%$$

- *Desvio padrão e coeficiente de variação de B.*

$$S_B = \sqrt{110,5} = 10,51 \qquad e \qquad CV_B\% = \frac{10,51}{70}.100 = 15,01\%$$

Conclusão: *as duas médias representam muito bem os dados, no entanto é fácil verificar que a dispersão dos valores de B é muito maior que a de A.*

Exemplo 2.6.2. *Considere a seguinte distribuição de frequências correspondente aos diferentes preços de um determinado produto em vinte lojas pesquisadas.*

Preços (R$)	*Nº de lojas*
50	*2*
51	*5*
52	*6*
53	*6*
54	*1*
Total	*20*

Vamos determinar a média, a variância, o desvio padrão e o coeficiente de variação dos preços. Adicionando as colunas complementares, a tabela completa fica:

Preços (R$)	*Nº de lojas*	$x_i.n_i$	$(x_i - \overline{x})$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2.n_i$
50	*2*	*100*	*-1,95*	*3,8025*	*7,605*
51	*5*	*255*	*-0,95*	*0,9025*	*4,5125*
52	*6*	*312*	*0,05*	*0,0025*	*0,015*
53	*6*	*318*	*1,05*	*1,1025*	*6,615*
54	*1*	*54*	*2,05*	*4,2025*	*4,2025*
Total	*20*	*1039*			*22,95*

A partir dessa tabela, obtemos os valores desejados como segue:

- $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i.n_i}{n} = \frac{1039}{20} = 51,95(R\$)$

- $S^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2.n_i}{n-1} = \frac{22,95}{19} = 1,21(R\$)^2$

- $S = \sqrt{1,21} = 1,1(R\$) \qquad e \qquad CV\% = \frac{1,1}{51,95}.100 = 2,12\%$

A média, nesse caso, é uma ótima medida para representar os dados, pois existe uma baixa variabilidade em torno desse valor.

Exemplo 2.6.3. *Um comerciante atacadista vende determinado produto em sacas que deveriam conter 16,50 kg. A pesagem de 40 sacas revelou os resultados representado na tabela:*

Classes de peso	n_i
$14,55 \vdash 15,05$	*1*
$15,05 \vdash 15,55$	*3*
$15,55 \vdash 16,05$	*8*
$16,05 \vdash 16,55$	*9*
$16,55 \vdash 17,05$	*10*
$17,05 \vdash 17,55$	*6*
$17,55 \vdash 18,05$	*3*
Total	*40*

Determinar a média, a variância, o desvio padrão e o coeficiente de variação dos pesos.

Tabela completada.

Classe de peso	n_i	x_i	$x_i n_i$	$(x_i - \overline{X})$	$(x_i - \overline{X})^2$	$(x_i - \overline{X})^2.n_i$
14,55 ⊢ 15,05	*1*	*14,8*	*14,8*	*-1,68*	*2,8224*	*2,8224*
15,05 ⊢ 15,55	*3*	*15,3*	*45,9*	*-1,18*	*1,3924*	*4,1772*
15,55 ⊢ 16,05	*8*	*15,8*	*126,4*	*-0,68*	*0,4624*	*3,6992*
16,05 ⊢ 16,55	*9*	*16,3*	*146,7*	*-0,18*	*0,0324*	*0,2916*
16,55 ⊢ 17,05	*10*	*16,8*	*168*	*0,32*	*0,1024*	*1,024*
17,05 ⊢ 17,55	*6*	*17,3*	*103,8*	*0,82*	*0,6724*	*4,0344*
17,55 ⊢ 18,05	*3*	*17,8*	*53,4*	*1,32*	*1,7424*	*5,2272*
Total	*40*		*659*			*21,276*

A partir da última tabela, obtemos os valores desejados como segue:

- $\bar{x} = \frac{\sum_{i=1}^{n} x_i.n_i}{n} = \frac{659}{40} = 16,475kg$

- $S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2.n_i}{n-1} = \frac{21,276}{39} = 0,55kg^2$

- $S = \sqrt{0,55} = 0,74Kg \qquad e \qquad CV\% = \frac{0,74}{16,475}.100 = 4,48\%$

O coeficiente de variação mostra a baixa variabilidade dos dados em torno da média, o que faz com que essa medida seja uma ótima medida para representar os dados.

Exercício 2.6.1. *Uma distribuição de frequências para os níveis séricos (em microgramas por decilitro) de zinco de 462 homens entre as idades de 15 a 17 anos é exibida na Tabela 2.11.*

Tabela 2.11: Nível sérico de zinco (mg/dl).

Nível	Número de homens
50 ⊢ 60	6
60 ⊢ 70	35
70 ⊢ 80	110
80 ⊢ 90	116
90 ⊢ 100	91
100 ⊢ 110	63
110 ⊢ 120	30
120 ⊢ 130	5
130 ⊢ 140	2
140 ⊢ 150	2
150 ⊢ 160	2
Total	462

1. *Obtenha a média, a moda e a mediana dos níveis de zinco.*

2. *Obtenha o desvio padrão e o coeficiente de variação. O que você pode concluir?*

3. *Apresente os quartins.*

4. *Exiba as separatrizes em um diagrama de caixa (boxplot).*

Exercício 2.6.2. *A seguir temos a distribuição de frequência dos pesos de uma amostra de 45 alunos:*

Pesos(Kg)	$40 \vdash 45$	$45 \vdash 50$	$50 \vdash 55$	$55 \vdash 60$	$60 \vdash 65$	$65 \vdash 70$
Número de alunos	04	10	15	08	05	03

a) Determinar a média, a moda e a mediana.
R: 53,5
b) Determinar o desvio padrão e o coeficiente de variação. O que você pode concluir? R: var= 45; d.p.= 6,708; C.V.=0,125

Exercício 2.6.3. *Entre os formados de 2012 de uma certa universidade no Paraná, 382 formados em ciências humanas receberam ofertas de emprego com salários anuais médios de 24 373 reais, 450 formados em ciências sociais receberam ofertas de emprego com salários anuais médios de 22 684 reais e 113 formados em ciência da computação receberam ofertas de emprego com salários anuais médios de de 31 329 reais. Qual foi a média de salário anual oferecida a esses 945 formados?*

Exercício 2.6.4. *Os dados a seguir referem-se aos rendimentos médios, em kg/ha, de 32 híbridos de milho recomendados para a Região Oeste catarinense*

Tabela 2.12: Rendimento médios, em kg/ha, de 32 híbridos de milho, Região Oeste, 1987/88.

3.973	4.550	4.770	4.980	5.117	5.403	6.166
4.500	4.680	4.778	4.993	5.166	5.513	6.388
4.550	4.685	4.849	5.056	5.172	5.823	
4.552	4.760	4.960	5.063	5.202	5.889	
4.614	4.769	4.975	5.110	5.230	6.047	

Fonte: Andrade e Ogliari

Use o computador para resolver os itens abaixo.
a) Construa uma tabela de frequência agrupando os dados em intervalos de classe.
b) Construa o histograma.
c) Calcule o primeiro, segundo e o terceiro quartil e interprete-os.
c) Construa o boxplot.

Referêcias

Barbetta, P. A., Reis, M. M. & Bornia, A. C. (2004). *Estatística: para cursos de engenharia e informática*, volume 3. Atlas São Paulo.

Hazzan, S. (2013). *Fundamentos de matemática elementar, 5, Combinatória, Probabilidade*. Atual, São Paulo - SP.

Morettin, L. (2010). *Estatistica basica: probabilidade e inferência : volume único*. MAKRON.

WILLIAM, J. S. (1981). Estatística aplicada à administração.