Universidade Federal Fluminense

Instituto de Matemática e Estatística

Departamento de Estatística uff

Estatística Descritiva

Ana Maria Lima de Farias
Departamento de Estatística

Conteúdo

Capítulo 1

Descrição de dados: tabelas e gráficos

De posse de um conjunto de dados, o primeiro passo em sua análise é descobrir o que eles nos dizem. A análise de dados será o objeto de estudo na primeira parte do nosso curso e começamos com gráficos e tabelas, que são ferramentas estatísticas importantes na visualização dos dados.

1.1 Pesquisa estatística – conceitos básicos

1.1.1 População e amostra

Estatística é a ciência da aprendizagem a partir dos dados. Em geral, fazemos levantamentos de dados para estudar e compreender características de uma população. Por exemplo, um grande banco, querendo lançar um novo produto, precisa conhecer o perfil socioeconômico dos seus clientes e, neste caso, a população de interesse é formada pelos clientes de todas as agências do banco. A Federação das Indústrias do Estado do Rio de Janeiro – FIRJAN – mede o grau de confiança dos empresários industriais através de uma pesquisa junto às indústrias, sendo a população de interesse, aqui, o conjunto das empresas industriais do estado do Rio de Janeiro.

Com esses dois exemplos apenas, já podemos ver que o conceito de *população de uma pesquisa estatística* é mais amplo, não se restringindo a seres humanos; ela é definida exatamente a partir dos objetivos da pesquisa.

Embora tenham populações bastante distintas, essas duas pesquisas têm em comum o fato de os resultados desejados serem obtidos a partir de dados levantados em um subconjunto da população – uma *amostra*. Há várias razões para se trabalhar com *pesquisas por amostragem* – custo e tempo, em geral, são as mais comuns. Mas, além de serem mais baratas e rápidas, as pesquisas por amostragem, se bem planejadas, podem fornecer resultados quase tão precisos quanto aqueles fornecidos por *pesquisas censitárias*, em que todos os elementos da população são investigados.

DEFINIÇÃO **População**

População é o conjunto de elementos para os quais se deseja estudar determinada(s) característica(s).

Amostra é um subconjunto da população.

Exemplos clássicos de pesquisa censitária são os Censos Demográficos realizados a cada dez anos no Brasil e em outros países. O objetivo desses censos é levantar informações sobre toda a população do país, de modo a fornecer subsídios para os governantes definirem as políticas públicas. Como exemplos de pesquisa por amostragem, podemos citar também as pesquisas de intenção de voto em eleições, a Pesquisa Nacional por Amostra de Domicílios - PNAD - realizada pelo IBGE, dentre muitas outras.

1.1.2 Alguns tipos de amostragem

Nas pesquisas por amostragem, em particular, o método de seleção da amostra é uma peça fundamental, pois os elementos da amostra têm que ser *representativos* da população à qual os resultados da pesquisa serão estendidos. Por exemplo, numa pesquisa de intenção de voto para prefeito de um município, a amostra tem que ser representativa de todas as regiões do município; não podemos concentrar a pesquisa em um bairro específico, por exemplo, pois o comportamento do eleitorado desse bairro pode ser diferente do comportamento dos eleitores de outros bairros. Na pesquisa de preços para elaboração do Índice Nacional de Preços ao Consumidor – INPC – temos que ter um levantamento em todas as regiões do país para que o índice resultante possa ser representativo do movimento de preços em todo o país.

Um método básico de seleção de amostras é a *amostragem aleatória simples*. Por esse método, todo subconjunto de tamanho n tem a mesma chance de se tornar a amostra selecionada. O processo de amostragem aleatória simples pode ser *com* ou *sem reposição*. Um procedimento comum para se selecionar uma amostra aleatória simples de uma população de tamanho N consiste em numerar os itens da população de 1 a N, escrever esses números em cartões iguais, colocar esses cartões em uma urna bem misturados e daí tirar os n cartões correspondentes à amostra. A amostragem será *com* reposição se cada cartão selecionado for colocado na urna antes da próxima extração; neste caso, há sempre N cartões na urna e cada um deles tem a mesma chance de ser selecionado. Se os cartões selecionados não são colocados de volta na urna antes da próxima extração, então temos amostragem *sem* reposição, que é o método prático mais usual. O número de cartões na urna a cada extração é diferente – para a primeira extração temos N, para a segunda temos $N-1$, para a terceira temos $N-2$ e assim por diante – mas todos eles têm a mesma chance de seleção em cada extração, garantida pelo sorteio aleatório. Na prática, usamos programas computacionais para efetuar o processo de amostragem; já imaginou escrever cartões para representar toda a população brasileira?

Um outro método bastante utilizado é o de *amostragem aleatória estratificada*. Nesse método, a população é dividida em *estratos*, que são subconjuntos da população, mutuamente exclusivos (os estratos não têm elementos em comum) e exaustivos (todo elemento da população pertence a um único estrato), e de cada estrato extrai-se uma amostra aleatória simples. A formação dos estratos deve ser feita de modo que tenhamos máxima homogeneidade dentro de cada estrato e máxima heterogeneidade entre os estratos. Considere, por exemplo, uma pesquisa por amostragem que deve dar resultados para o Brasil. Em vez de se trabalhar com uma amostra aleatória simples de todo o país, podemos estratificar por estado ou por região geográfica, por exemplo. A estratificação tem vantagens administrativas e também estatísticas: com estratos bem definidos, podemos ter resultados precisos com amostras menores e com a vantagem adicional de podermos dar resultados individuais para cada estrato.

Os dois métodos acima descritos são métodos de *amostragem probabilística*, assim chamados porque a aleatoriedade na seleção dos elementos permite que se atribua, a cada elemento da população, uma probabilidade de inclusão na amostra e com essa probabilidade teremos condições de generalizar os resultados da amostra para a população inteira, quantificando a margem de erro.

Considere, agora, que você esteja interessado em avaliar a opinião dos alunos da UFF sobre o serviço de transporte entre os diversos *campi*, oferecido pela administração da universidade. Como você não tem condições nem tempo de selecionar uma amostra de todos os alunos da UFF, você decide entrevistar seus colegas de turma. Essa é uma *amostra de conveniência* e o grande problema é que os resultados obtidos não poderão ser generalizados para uma população maior. Nem mesmo para o seu curso podemos generalizar, porque sua turma pode pode não ser representativa de todas as turmas do seu curso.

Métodos de seleção de amostra mais sofisticados são empregados em diversas pesquisas com o objetivo de se obter uma "boa amostra", ou seja, uma amostra pequena e que forneça resultados precisos sobre a população de interesse.

1.2 Níveis de mensuração

Nas pesquisas estatísticas, as características sobre as quais queremos obter informação são chamadas *variáveis* e uma informação importante sobre essas variáveis é o seu *nível de mensuração*. Isto porque a aplicabilidade ou não de modelos e métodos estatísticos a serem utilizados posteriormente na análise dos dados vai depender em grande parte desse aspecto.

O nível mais elementar de mensuração consiste na classificação dos indivíduos ou objetos de uma população de acordo com uma certa característica, isto é, separam-se os elementos em grupos, conforme possuam essa ou aquela característica em questão. É o que sucede, por exemplo, quando a característica estudada é sexo, religião, estado civil, etc. Nesses casos, as categorias se expressam nominalmente e para a aplicação de métodos estatísticos adequados, é necessário que as categorias sejam *exaustivas* (isto é, cubram todos os elementos da população) e *mutuamente exclusivas* (isto é, um elemento pertence a uma única categoria). Nesses casos, diz-se que a variável em estudo é expressa segundo uma *escala nominal*. Assim, as operações usuais de aritmética não podem ser realizadas sobre esse tipo de escala, mesmo que as categorias estejam expressas em números. No processamento de dados, é bastante comum representar as categorias de sexo Feminino e Masculino por números, como 1 e 2. Naturalmente, não faz sentido dizer que o Masculino é duas vezes o Feminino; o 1 e o 2 são apenas substitutos dos nomes das categorias.

Num nível de mensuração seguinte, podemos ordenar as categorias de uma determinada variável. É o que ocorre com o nível de escolaridade, quando uma população pode ser classificada, por exemplo, em 4 categorias: analfabeto, 1º grau, 2º grau, 3º grau. Aqui podemos dizer que o nível de escolaridade de um indivíduo da categoria 2º grau é maior que o de um indivíduo da categoria 1º grau, mas não podemos dizer que é duas vezes maior. Nesta escala, chamada *escala ordinal*, valem apenas as operações de ordenação, maior do que ou menor do que.

Passa-se deste tipo de escala para um nível de mensuração propriamente dito quando, além da ordenação das categorias, pode-se dizer quanto valem exatamente as diferenças entre essas categorias. Um exemplo típico dessa situação é a medição de temperatura: a diferença entre 90ºC e 70ºC é 20ºC e é igual à diferença entre 30ºC e 10ºC. No entanto, como o zero (0ºC) nesta escala é definido arbitrariamente (não existe naturalmente), não podemos dizer que 90ºC é três vezes mais quente que 30ºC. Dizemos, então, que a temperatura está medida em uma *escala intervalar.*

Quando o zero na escala puder ser estabelecido de forma não arbitrária, todas as operações aritméticas poderão ser realizadas sobre os valores tomados pela variável em estudo. Nesse caso, dizemos que a variável está medida em uma *escala de razão* ou *proporcional*. É o caso da idade, que é contada a partir da data de nascimento do indivíduo.

1.2.1 Variáveis qualitativas e quantitativas

É comum denominar de *variável qualitativa* as características medidas em escala nominal ou ordinal. Já as variáveis medidas em escala intervalar ou proporcional são chamadas *variáveis quantitativas*.

DEFINIÇÃO **Variáveis qualitativas e quantitativas**

Variáveis qualitativas *descrevem* características de elementos de uma população e podem ser medidas em escala nominal ou ordinal.

Variáveis quantitativas *medem* características de elementos de uma população e podem ser expressas em escala de razão ou intervalar.

As variáveis quantitativas, por sua vez, podem ser discretas ou contínuas. Quando a variável puder assumir qualquer valor numérico em um determinado intervalo de variação, ela será uma variável *contínua*. Essas variáveis resultam normalmente de medições, como peso, altura, dosagem de hemoglobina, renda etc. A interpretação desse tipo de variável leva à noção de valor aproximado, pois não existe instrumento de medição capaz de fornecer precisão absoluta na informação. Assim, quando uma balança mostra o peso de uma pessoa como 65,5 kg, esse valor, na verdade, é uma aproximação para qualquer valor entre, digamos, 65,495 kg e 65,505 kg.

Por outro lado, a variável quantitativa *discreta* só poderá assumir valores pertencentes a um conjunto enumerável (pense nos números naturais!); os valores normalmente são obtidos através de algum processo de contagem. Alguns exemplos são o número de filhos de um casal, número de empregados de uma firma de contabilidade, etc.

DEFINIÇÃO **Variáveis discretas e contínuas**

Variáveis quantitativas discretas assumem valores pertencentes a um conjunto enumerável; em geral, resultam de processos de contagem.

Variáveis quantitativas contínuas assumem valores pertencentes a um intervalo de números reais; em geral resultam de processos de medição.

EXEMPLO 1.1 *População e Amostra*

Para cada uma das situações listadas a seguir, identifique a população de interesse e a amostra, se for o caso.

(a) A Pró-Reitoria de Assuntos Estudantis da UFF deseja saber a opinião dos calouros sobre o programa de Acolhimento Estudantil. Sorteia, então, uma amostra de 200 calouros de todos os cursos da UFF, que são entrevistados pelos funcionários.

(b) Uma grande empresa deseja saber a opinião de seus gerentes sobre uma nova proposta de plano de carreira. Para isso, envia um questionário para todos os seus 450 gerentes.

(c) Uma loja de vestuário pretende enviar um questionário de uma pesquisa de satisfação para seus clientes. A partir de seus registros, o gerente de marketing constata que 4345 pessoas fizeram compras com cartão de crédito na loja no último semestre. Ele sorteia uma amostra de 200 desses clientes para os quais envia um questionário.

Solução

(a) A população de interesse é formada por todos os calouros da UFF no ano em questão e a amostra é o conjunto dos 200 alunos entrevistados.

(b) A população é o conjunto dos gerentes da empresa. Como foram entrevistados todos os gerentes, essa é uma pesquisa censitária e não uma pesquisa por amostragem.

(c) A população de interesse é formada por todos os clientes da loja, mas a população de referência, ou seja, a população de onde foi retirada a amostra, é formada pelos clientes que compraram com cartão de crédito. Note que aí não estão incluídos os clientes que pagaram com dinheiro ou cheque.

♦♦

EXEMPLO 1.2 *Classificação de variáveis*

Classifique as variáveis abaixo como qualitativa ou quantitativa (discreta ou contínua).

(a) Altura dos alunos da UFF.

(b) Opinião de consumidores sobre determinado produto (Ruim, Bom ou Excelente).

(c) Número de sanduíches Big Mac vendidos nos estados do Brasil pela rede McDonalds no McDia Feliz.

(d) Temperatura máxima diária na cidade de Niterói no mês de agosto de 2012.

(e) Opinião dos empregados de uma empresa sobre obrigatoriedade do uso do crachá (a favor ou contra).

Solução

(a) Altura é uma variável quantitativa contínua.

(b) A opinião é uma variável qualitativa. Como há uma ordem nas respostas, essa é uma variável qualitativa *ordinal*.

(c) Número de sanduíches é uma variável quantitativa discreta.

(d) Temperatura máxima é uma variável quantitativa contínua.

(e) A opinião, neste caso, é uma variável qualitativa *nominal* - não há qualquer ordem nas respostas possíveis.

♦♦

1.3 Apresentação de dados qualitativos

Vamos considerar o seguinte exemplo fictício, mas verossímil. A direção de uma empresa está estudando a possibilidade de fazer um seguro saúde para seus funcionários e respectivos familiares. Para isso, ela faz um levantamento de seus 500 funcionários, obtendo informação sobre sexo, estado civil, idade, número de dependentes e salário. Como são 500 funcionários, temos que achar uma forma de resumir os dados. Nesta seção, você irá aprender a resumir dados qualitativos em forma de uma distribuição (ou tabela) de frequência e, também, em forma gráfica. Você verá que os gráficos complementam a apresentação tabular.

1.3.1 Distribuições de frequência

Consideremos, inicialmente, a variável qualitativa gênero. O que nos interessa saber sobre essa variável não é que João seja do sexo masculino e Maria do sexo feminino, mas sim quantos funcionários e quantas funcionárias há na empresa. Esse resultado pode ser resumido em uma tabela ou distribuição de frequências da seguinte forma:

Gênero	Número de funcionários
Masculino	270
Feminino	230
Total	500

Os números 270 e 230 resultaram da contagem das frequências de ocorrência de cada uma das categorias da variável sexo. Essa contagem é também chamada de *frequência simples absoluta* ou simplesmente *frequência*. O total de 500 é obtido somando-se o número de homens e de mulheres.

É interessante também expressar esses resultados em forma relativa, isto é, considerar, para cada classe, a *frequência relativa* ao total:

$$\frac{270}{500} = 0,54$$

ou seja, 54% dos funcionários da empresa são do sexo masculino.

É comum apresentar as frequências relativas em forma percentual. Note que:

$$\frac{270}{500} = 0,54 = \frac{54}{100} = 54\%$$

Na **Tabela 1.1**, apresenta-se a versão completa da distribuição dos funcionários por gênero e por estado civil. Note que a soma das frequências absolutas deve ser igual ao número total de elementos sendo pesquisados, enquanto a soma das frequências relativas é sempre 1 ou 100%.

Tabela 1.1 – Número de funcionários por gênero e por estado civil

Gênero	Frequência simples	
	absoluta	relativa
Masculino	270	0,54
Feminino	230	0,46
Total	500	1,00

Estado civil	Frequência simples	
	absoluta	relativa %
Solteiro	125	25,0
Casado	280	56,0
Divorciado	85	17,0
Viúvo	10	2,0
Total	500	100,0

EXEMPLO 1.3 *Dados dos funcionários do Departamento de RH*

Consideremos que, na situação descrita anteriormente, os dados tenham sido levantados por departamento, para depois serem totalizados. Para o Departamento de Recursos Humanos, foram obtidas as seguintes informações:

Nome	Sexo	Estado civil	Número de dependentes
João da Silva	M	Casado	3
Pedro Fernandes	M	Viúvo	1
Maria Freitas	F	Casada	0
Paula Gonçalves	F	Solteira	0
Ana Freitas	F	Solteira	1
Luiz Costa	M	Casado	3
André Souza	M	Casado	4
Patrícia Silva	F	Divorciada	2
Regina Lima	F	Casada	2
Alfredo Souza	M	Casado	3
Margarete Cunha	F	Solteira	0
Pedro Barbosa	M	Divorciado	2
Ricardo Alves	M	Solteiro	0
Márcio Rezende	M	Solteiro	1
Ana Carolina Chaves	F	Solteira	0

Para pequenos conjuntos de dados, podemos construir a tabela à mão e, para isso, precisamos contar o número de ocorrências de cada categoria de cada uma das variáveis. Varrendo o conjunto de dados a partir da primeira linha, podemos marcar as ocorrências da seguinte forma:

Masculino	||||||||	Solteiro	||||||
Feminino	|||||||	Casado	||||||
		Divorciado	||
		Viúvo	|

Obtemos, então, as seguintes distribuições de frequência:

Gênero	Frequência simples	
	absoluta	relativa %
Masculino	8	53,33
Feminino	7	46,67
Total	15	100,0

Estado civil	Frequência simples	
	absoluta	relativa %
Solteiro	6	40,00
Casado	6	40,00
Divorciado	2	13,33
Viúvo	1	6,67
Total	15	100,00

♦♦

1.3.2 Arredondamento de números

No Exemplo 1.3, a divisão de algumas frequências absolutas pelo total de 15 resultou em dízimas. Nesses casos, torna-se necessário arredondar os resultados, mas esse arredondamento deve ser feito com cautela para se evitar que a soma não seja igual a 1 ou 100%.

A primeira etapa no processo de arredondamento consiste em decidir o número de casas decimais desejado. Em geral, frequências relativas percentuais são apresentadas com, no máximo, 2 casas decimais. Isso significa que temos de descartar as demais casas decimais. Existe a seguinte regra de arredondamento:

! Arredondamento de números

Quando o primeiro algarismo a ser suprimido for menor ou igual a 4 (ou seja, for igual a 0,1, 2, 3 ou 4), o último algarismo a ser mantido permanece inalterado. Quando o primeiro algarismo a ser suprimido for igual a 5, 6, 7, 8 ou 9, o último algarismo a ser mantido é acrescido de 1.

Na distribuição de frequências da variável gênero, temos os seguintes resultados:

$$\frac{8}{15} \times 100 = 53,33333\ldots$$

$$\frac{7}{15} \times 100 = 46,66666\ldots$$

No primeiro caso, o primeiro algarismo a ser suprimido é 3; logo, o último algarismo a ser mantido, (3), não se altera e o resultado é 53,33. No segundo caso, o primeiro algarismo a ser suprimido é 6. Logo, o último algarismo a ser mantido, (6), deve ser acrescido de 1 e o resultado é 46,67. Tente sempre usar essa regra em seus arredondamentos; com ela, você evitará erros grosseiros.

Na apresentação de tabelas de frequências relativas, é possível que essas frequências não somem 100%, ou seja, é possível que, ao somarmos as frequências relativas, obtenhamos resultados como 99,9% ou 100,01%. Esses pequenos erros são devidos a arredondamentos e nem sempre é possível evitá-los; no entanto, aceita-se implicitamente que a soma das frequências seja 100%.

1.3.3 Gráficos

As distribuições de frequência para dados qualitativos também podem ser ilustradas graficamente através de gráficos de colunas ou gráficos de setores, também conhecidos como gráficos de pizza. Na Figura 1.1, temos os gráficos de coluna e de setores para os dados da **Tabela 1.1**, referentes ao estado civil dos funcionários.

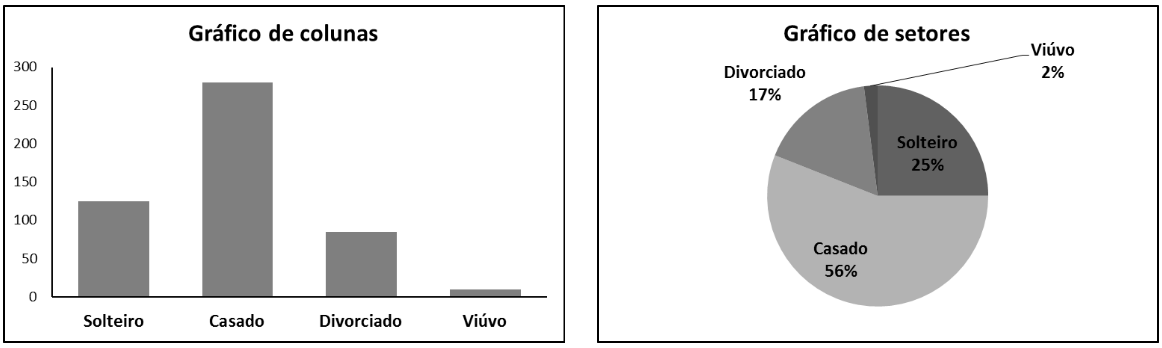


Figura 1.1 – Distribuição do número de funcionários por estado civil

No **gráfico de colunas**, a altura de cada coluna representa a frequência da respectiva classe e o gráfico pode ser construído com base nas frequências absolutas ou relativas. Para diferenciar um do outro, coloca-se no título do eixo o tipo de frequência utilizada. Note que, no eixo horizontal, não há escala, uma vez que aí se representam as categorias da variável, que devem ser igualmente espaçadas.

No **gráfico de setores**, a frequência de cada categoria é representada pelo tamanho (ângulo) do setor (ou fatia da pizza). Para construir um gráfico de setores à mão, você precisará de um compasso para fazer um círculo de raio arbitrário e, em seguida, traçar um raio qualquer no círculo. A partir daí, você marcará os raios de acordo com os ângulos de cada setor, utilizando um transferidor. Para determinar o ângulo de cada setor, você deverá usar a seguinte regra de proporcionalidade: o ângulo total – 360º– corresponde ao número total de observações; o ângulo de cada setor corresponde à frequência da respectiva classe. Dessa forma, você obtém a seguinte regra de três para os solteiros:

$$\frac{360^{\circ}}{500} = \frac{x}{125} \Rightarrow x = 90^{\circ}$$

Esses gráficos podem ser construídos facilmente com auxílio de programas de computador, como o programa de planilhas Excel da Microsoft ®.

1.4 Apresentação de dados quantitativos discretos

1.4.1 Distribuições de frequências

Quando uma variável quantitativa discreta assume poucos valores distintos, é possível construir uma distribuição de frequências da mesma forma que fizemos para as variáveis qualitativas. A diferença é que, em vez de termos categorias nas linhas da tabela, teremos os distintos valores da variável. Continuando com o nosso exemplo, vamos trabalhar agora com a variável número de dependentes. Suponha que alguns funcionários não tenham dependentes e que o número máximo de dependentes seja 7. Obteríamos, então, a seguinte distribuição de frequências:

Número de	Frequência simples	
dependentes	absoluta	relativa %
0	120	24,0
1	95	19,0
2	90	18,0
3	95	19,0
4	35	7,0
5	30	6,0
6	20	4,0
7	15	3,0
Total	500	100,0

O processo de construção é absolutamente o mesmo, mas, dada a natureza quantitativa da variável, é possível acrescentar mais uma informação à tabela.

Suponha, por exemplo, que a empresa esteja pensando em limitar o seu projeto a 4 dependentes, de modo que funcionários com mais de 4 dependentes terão que arcar com as despesas extras. Quantos funcionários estão nessa situação?

Para responder a perguntas desse tipo, é costume acrescentar à tabela de frequências uma coluna com as *frequências acumuladas*. Essas frequências são calculadas da seguinte forma: para cada valor da variável (número de dependentes), contamos quantas ocorrências correspondem a valores menores ou iguais a esse valor.

Por exemplo, valores da variável menores ou iguais a 0 correspondem aos funcionários sem dependentes. Logo, a frequência acumulada para o valor 0 é igual à frequência simples: 120. Analogamente, valores da variável menores ou iguais a 1 correspondem aos funcionários sem dependentes *mais* os funcionários com 1 dependente. Logo, a frequência acumulada para o valor 1 é igual a $120 + 95 = 215$. Para o valor 2, a frequência acumulada é igual a $120 + 95 + 90 = 215 + 90 = 305$. Repetindo esse procedimento, obtemos a **Tabela 1.2**.

Note que aí acrescentamos também as frequências acumuladas em forma percentual. Essas frequências são calculadas como a proporção da frequência acumulada em relação ao total; por exemplo,

$$87,0 = \frac{435}{500} \times 100$$

Suponhamos, agora, que se pergunte para cada um dos 500 funcionários a sua idade, em anos completos. Essa é, também, uma variável discreta, mas a diferença é que a idade

Tabela 1.2 – Distribuição de frequências para o número de dependentes

Número de	Frequência simples		Frequência acumulada	
dependentes	absoluta	relativa %	absoluta	relativa %
0	120	24,0	120	24,0
1	95	19,0	215	43,0
2	90	18,0	305	61,0
3	95	19,0	400	80,0
4	35	7,0	435	87,0
5	30	6,0	465	93,0
6	20	4,0	485	97,0
7	15	3,0	500	100,0
Total	500	100,0		

pode assumir um número maior de valores, o que resultaria em uma tabela grande, caso decidíssemos relacionar todos os valores, da mesma forma que fizemos para o número de dependentes. Além disso, em geral não é necessário apresentar a informação em tal nível de detalhamento.

Por exemplo, para as seguradoras de planos de saúde, as faixas etárias importantes – aquelas em que há reajuste por idade – são 0 a 18; 19 a 23; 24 a 28; 29 a 33; 34 a 38; 39 a 43; 44 a 48; 49 a 53; 54 a 58 e 59 ou mais. Sendo assim, podemos agrupar os funcionários segundo essas faixas etárias e construir uma **tabela de frequências agrupadas** em que cada frequência corresponde ao número de funcionários na respectiva faixa etária, tal como a **Tabela 1.3**:

Tabela 1.3 – Distribuição de frequência das idades de 500 funcionários

Faixa	Frequência Simples		Frequência Acumulada	
Etária	Absoluta	Relativa %	Absoluta	Relativa %
19 — 23	1	0,2	1	0,2
24 — 28	23	4,6	24	4,8
29 — 33	103	20,6	127	25,4
34 — 38	246	49,2	373	74,6
39 — 43	52	10,4	425	85,0
44 — 48	50	10,0	475	95,0
49 — 53	25	5,0	500	100,0
Total	500	100,0		

1.4.2 Gráfico da distribuição de frequências simples

A representação gráfica da distribuição de frequências de uma variável quantitativa discreta pode ser feita através de um gráfico de colunas, desde que o número de valores seja pequeno. A diferença, neste caso, quando comparamos com as variáveis qualitativas, é que, no eixo horizontal do gráfico, é representada a escala da variável quantitativa, que deve ser definida cuidadosamente de modo a representar corretamente os valores.

Na Figura 1.2, temos o gráfico de colunas para o número de dependentes dos 500 funcionários.

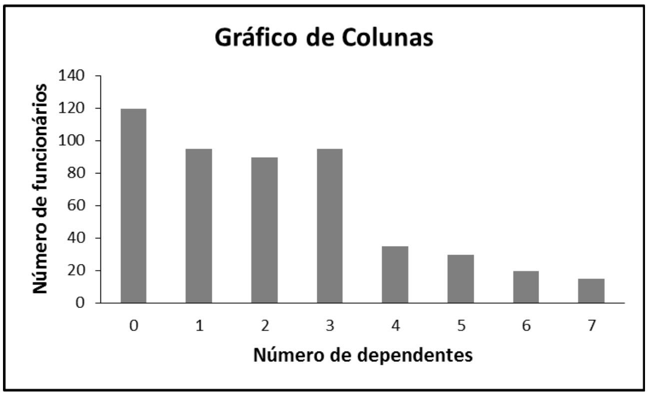


Figura 1.2 – Distribuição do número de dependentes por funcionário

> **! Gráfico de setores para dados quantitativos**
>
> Embora nem sempre incorreto, não é apropriado representar dados quantitativos discretos em um gráfico de setores, uma vez que, neste gráfico, não é possível representar a escala dos dados.

1.5 Apresentação de dados quantitativos contínuos

1.5.1 Distribuições de frequência

Para as variáveis quantitativas contínuas, devemos também trabalhar com distribuições de frequências agrupadas. O processo de construção é idêntico ao visto para as variáveis discretas, mas aqui devemos tomar um cuidado especial na construção das classes. A escolha dos limites das classes deve ser feita com base na natureza, valores e unidade de medida dos dados. As regras que deverão ser seguidas são as seguintes:

> **! Classes em uma distribuição de frequências agrupadas**
>
> 1. As classes têm que ser exaustivas, isto é, todos os elementos devem pertencer a alguma classe.
> 2. As classes têm que ser mutuamente exclusivas, isto é, cada elemento tem que pertencer a uma única classe.

O primeiro passo é definir o número de classes desejado; esse número, de preferência, deve estar entre 5 e 25. Em seguida, devemos determinar a **amplitude** dos dados, ou seja, o intervalo de variação dos valores observados da variável em estudo.

DEFINIÇÃO Amplitude

A **amplitude** de um conjunto de dados, representada por Δ_{total}, é definida como a diferença entre os valores máximo e mínimo:

$$\Delta_{total} = V_{\text{Máx}} - V_{\text{Mín}} \tag{1.1}$$

Se quisermos trabalhar com classes de mesmo comprimento (e essa é uma opção bastante comum), para determinar esse comprimento, é necessário dividir a amplitude total pelo número de classes desejado. No entanto, para garantir a inclusão dos valores mínimo e máximo, podemos, como regra geral, usar o seguinte procedimento: considere o primeiro múltiplo do número de classes maior que o valor da amplitude e use esse número como a nova amplitude.

Por exemplo, se a amplitude for 28 e quisermos trabalhar com cinco classes, vamos considerar 30 como a nova amplitude. Dividindo esse valor pelo número de classes, obtemos o comprimento de cada classe. Os limites de classe podem ser obtidos somando-se o comprimento de classe a partir do valor mínimo dos dados.

Continuando com o nosso exemplo, o comprimento de classe é $30 \div 5 = 6$; se o valor mínimo dos dados for 4, então os limites de classe serão:

$$\begin{aligned} & & 4 \\ 4+6 &=& 10 \\ 10+6 &=& 16 \\ 16+6 &=& 22 \\ 22+6 &=& 28 \\ 28+6 &=& 34 \end{aligned}$$

e as classes serão:

$$[4, 10) \;\; [10, 16) \;\; [16, 22) \;\; [22, 28) \;\; [28, 34)$$

Note o tipo de intervalo utilizado: para incluir o valor mínimo, 4, na primeira classe, o intervalo deve ser fechado no extremo inferior: [4,.

Se fechássemos o intervalo no limite superior, o 10 estaria incluído na primeira classe e, portanto, não poderia estar na segunda classe. Isso resultaria em $[4, 10]$ como a primeira classe e $(10, 16)$ como a segunda classe. Assim, as duas primeiras classes estariam definidas de forma diferente, o que não é conveniente, pois dificultaria a leitura da tabela. É preferível incluir o 10 na segunda classe, o que resulta nas classes apresentadas anteriormente.

EXEMPLO 1.4 *Salários de 500 funcionários*

Suponha que, dentre os 500 funcionários da nossa empresa, o menor salário seja de 2800 e o maior salário seja de 12400. Para agrupar os dados em cinco classes, devemos fazer o seguinte:

$$\Delta_{total} = V_{\text{Máx}} - V_{\text{Mín}} = 12400 - 2800 = 9600$$

$$\text{Próximo múltiplo de } 5 = 9605$$

$$\text{Comprimento de classe} = \frac{9605}{5} = 1921$$

Os limites de classe, então, são:

$$\begin{aligned}
& & 2800 \\
2800 + 1921 &=& 4721 \\
4721 + 1921 &=& 6642 \\
6642 + 1921 &=& 8563 \\
8563 + 1921 &=& 10484 \\
10484 - 1921 &=& 12405
\end{aligned}$$

e as classes podem ser definidas como:

$$\begin{array}{ll}
[2800, 4721) & \text{(2800 incluído; 4721 excluído)} \\
[4721, 6642) & \text{(4721 incluído; 6642 excluído)} \\
[6642, 8563) & \text{(6642 incluído; 8563 excluído)} \\
[8563, 10484) & \text{(8563 incluído; 10484 excluído)} \\
[10484, 12405) & \text{(10484 incluído; 12405 excluído)}
\end{array}$$

Essa é uma regra que resulta em classes corretamente definidas, mas nem sempre as classes resultantes são apropriadas ou convenientes. Neste exemplo, seria preferível trabalhar com classes de comprimento 2000, o que resultaria nas classes

$$[2800, 4800) \;\; [4800, 6800) \;\; [6800, 8800) \;\; [8800, 10800) \;\; [10800, 12800)$$

que são corretas e mais fáceis de ler.

Fazendo a contagem do número de funcionários em cada classe, a distribuição resultante seria:

Tabela 1.4 – Distribuição de frequência dos salários de 500 funcionários

Salário (reais)	Frequência Simples		Frequência Acumulada	
	Absoluta	Relativa %	Absoluta	Relativa %
2800 $\vdash$ 4800	87	17,4	87	17,4
4800 $\vdash$ 6800	203	40,6	290	58,0
6800 $\vdash$ 8800	170	34,0	460	92,0
8800 $\vdash$ 10800	30	6,0	490	98,0
10800 $\vdash$ 12800	10	2,0	500	100,0

♦♦

1.5.2 Histogramas e polígonos de frequência

O histograma e o polígono de frequências são gráficos usados para representar uma distribuição de frequências simples de uma variável quantitativa contínua. A ogiva de frequência representa graficamente a distribuição das frequências acumuladas.

DEFINIÇÃO **Histograma**

Um **histograma** é um gráfico formado por um conjunto de retângulos contíguos, com bases sobre um eixo horizontal, cuja escala é definida de acordo com as classes da distribuição da variável de interesse. As bases desses retângulos, construídas sobre o eixo horizontal, representam as classes e as **áreas são proporcionais ou iguais às frequências**.

Vamos ilustrar a construção de um histograma usando como exemplo a distribuição de frequência dos dados sobre salários dada na **Tabela 1.4**.

Começamos construindo os eixos: no eixo horizontal, representamos os limites das classes e, no eixo vertical, construímos a escala apropriada para representar as frequências absolutas. Veja a Figura 1.3. Poderíamos, também, trabalhar com as frequências relativas.

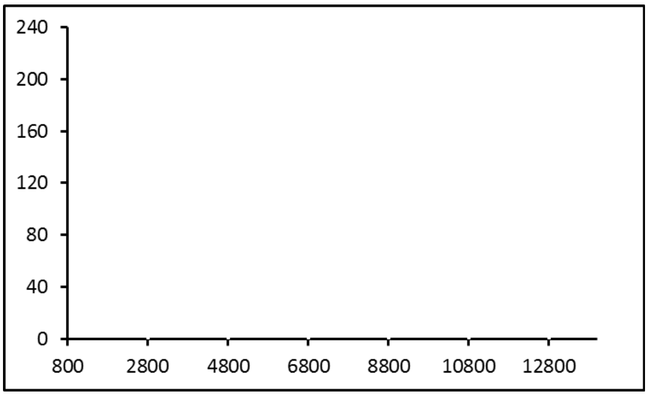


Figura 1.3 – Construção do Histograma da Distribuição dos Salários - Passo 1

Passamos, agora, a construir os retângulos, tendo em mente que a área de cada um representa a frequência da respectiva classe. Como neste exemplo as classes têm o mesmo comprimento, o histograma pode ser construído de tal modo que as alturas dos retângulos sejam *iguais* às frequências das classes. Dessa forma, as áreas serão *proporcionais* (e não iguais) às frequências, conforme ilustrado no histograma da Figura 1.4. Note que cada área é igual à frequência da classe multiplicada por 2000, o comprimento de cada classe.

Para construir o histograma baseado em retângulos com áreas exatamente iguais às frequências das classes, usa-se a fórmula da área de um retângulo com base igual ao comprimento de classe e área igual à frequência da classe. Por exemplo, para a classe [2800, 4800), a frequência (área) é 87 e a base do retângulo (comprimento de classe) é 2000. Logo, a altura h do retângulo correspondente é encontrada da seguinte forma:

$$87 = h * 2000 \Longrightarrow h = \frac{87}{2000} = 0,0435$$

O resultado dessa divisão é denominado **densidade**, uma vez que dá a frequência em cada classe por unidade da variável. Na Figura 1.5, temos o histograma em que a área de cada retângulo é exatamente *igual* à frequência absoluta da classe.

Observe as Figuras 1.4 e 1.5. Em ambos os gráficos, a forma dos retângulos é a mesma; o que muda é a escala no eixo vertical.

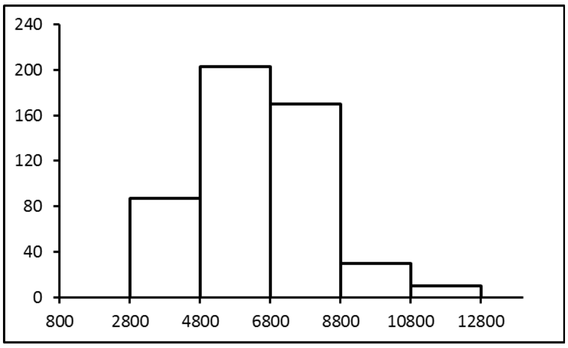


Figura 1.4 – Histograma dos salários - Altura = Frequência

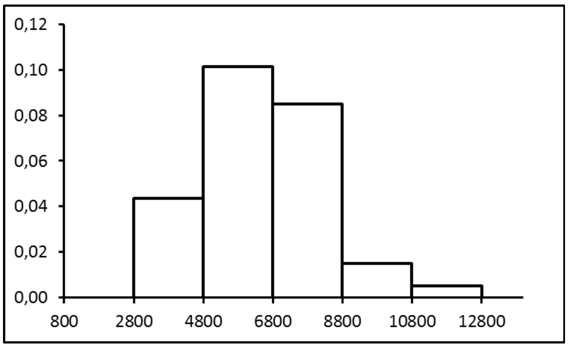


Figura 1.5 – Histograma dos salários - Área = Frequência

De modo geral, quando as classes têm o mesmo comprimento – e essa é a situação mais comum –, podemos representar as alturas dos retângulos pelas frequências das classes, o que facilita a interpretação do gráfico.

DEFINIÇÃO **Polígono de frequência**

Um **polígono de frequências** é um *gráfico de linha* obtido quando são unidos, por uma poligonal, os pontos correspondentes às frequências das diversas classes, centrados nos respectivos pontos médios. Mais precisamente, são plotados os pontos com coordenadas (ponto médio, frequência simples).

Para obter as interseções da poligonal com o eixo, cria-se em cada extremo uma classe com frequência nula.

Na Figura 1.6, temos o polígono de frequências para a distribuição dos salários dos 500 funcionários. É comum apresentar-se o polígono de frequências junto com o histograma, o que facilita a visualização dos resultados. Note que o polígono de frequência dá uma ideia da forma da distribuição dos dados.

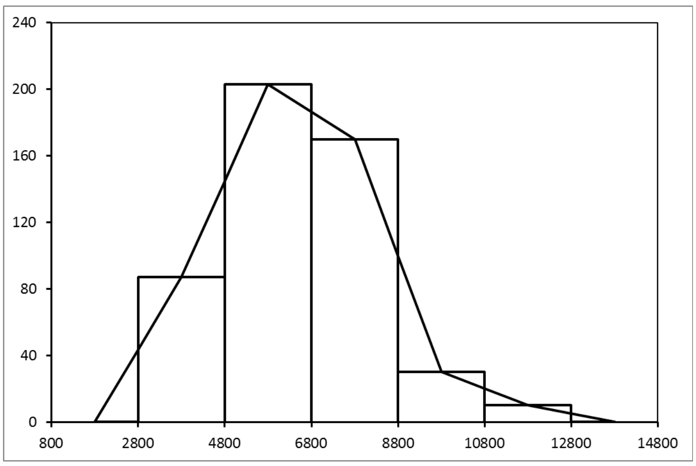


Figura 1.6 – Histograma e Polígono de Frequências para a Distribuição dos Salários

1.5.3 Diagrama de ramo-e-folhas

Um outro gráfico usado para mostrar a forma da distribuição de um conjunto de dados quantitativos é o **diagrama de ramo-e-folhas**, desenvolvido pelo estatístico John Tukey. Para a construção desse gráfico, cada observação do conjunto de dados é "quebrada" em duas partes. Uma dessas partes é a folha, que deve ser formada por apenas um algarismo, e os algarismos restantes formam o galho. Como numa árvore, as folhas são "penduradas" no galho apropriado.

Para construir o diagrama, traça-se uma linha vertical para separar os galhos das folhas. À esquerda dessa linha escrevem-se os diferentes ramos, um em cada linha horizontal, e escrevem-se as folhas no respectivo galho.

EXEMPLO 1.5 *Notas de 50 alunos*

Considerando as notas dos 50 alunos, vamos construir o diagrama de ramo-e-folhas com esses dados.

Tabela 1.5 – Notas de 50 alunos

2,9	3,8	3,7	4,9	4,7	5,6	7,3	8,3	5,5	7,7	8,9	8,7	7,6
8,3	7,3	6,9	6,8	7,0	5,4	6,5	7,6	5,2	9,0	7,4	8,4	6,8
7,5	8,7	9,7	7,9	7,2	8,1	9,4	6,6	7,0	8,0	9,2	8,8	
6,3	6,5	5,8	6,9	6,9	8,2	7,0	6,0	6,2	7,1	7,5	8,2	

A quebra de cada observação em duas partes aqui é bastante natural: a folha será o algarismo decimal, enquanto o ramo será a parte inteira. As duas primeiras observações são quebradas da seguinte forma: Por outro lado, a menor observação é 2,9 e a maior é 9,7; assim,

2	9
3	7

os galhos vão de 2 a 9, e organizamos a nossa escala da seguinte forma:

2	
3	
4	
5	
6	
7	
8	
9	

Continuando o processo, penduramos as folhas no respectivo galho, obtendo o Diagrama 1.1:

Diagrama 1.1 – *Notas de 50 alunos*

2	9												
3	8	7											
4	9	7											
5	6	5	4	2	8								
6	9	8	5	8	6	3	5	9	9	0	2		
7	3	7	6	3	0	6	4	5	9	2	0	0	1	5
8	3	9	7	3	4	7	1	0	8	2	2		
9	0	7	4	2									

Para facilitar a leitura, as folhas em cada ramo são ordenadas. É importante também definir corretamente a escala. Como indicar no diagrama que a primeira observação é 2,9 e não 29? Veja uma forma de fazer isso no Diagrama 1.2:

Diagrama 1.2 – *Notas de 50 alunos - versão final*

Escala

1	0	1,0

2	9													
3	7	8												
4	7	9												
5	2	4	5	6	8									
6	0	2	3	5	5	6	8	8	9	9	9			
7	0	0	0	1	2	3	3	4	5	5	6	6	7	9
8	0	1	2	2	3	3	4	7	7	8	9			
9	0	2	4	7										

♦♦

EXEMPLO 1.6 *Notas de duas turmas*

Suponha que, no Exemplo 1.5, a mesma prova tenha sido aplicada a duas turmas diferentes. Para comparar os resultados, podemos construir o *diagrama de ramo-e-folhas lado a lado*. Um conjunto é representado no lado direito da escala e, o outro, no lado esquerdo. Em ambas as partes, as folhas crescem da escala para as margens. Veja o Diagrama 1.3.

♦♦

Diagrama 1.3 – *Notas dos alunos das turmas A (lado esquerdo) e B (lado direito)*

Escala

1 | 0 | 1,0

8	1	
3 2 2	2	9
	3	7 8
7 5 0	4	7 9
2 1	5	2 4 5 6 8
6 5 4 3 3 2 0 0 0 0	6	0 2 3 5 5 6 8 8 9 9 9
2 2 2 0 0	7	0 0 0 1 2 3 3 4 5 5 6 6 7 9
4 3 3 2 1 0 0	8	0 1 2 2 3 3 4 7 7 8 9
5	9	0 2 4 7

1.5.4 Gráficos temporais

O **gráfico temporal** é um gráfico de linha, usado para representar observações feitas ao longo do tempo, isto é, observações de uma **série de tempo**.

No eixo horizontal, colocam-se as datas em que foram realizadas as observações e, no eixo vertical, os valores observados. Os pontos assim obtidos são unidos por segmentos de reta para facilitar a visualização do comportamento dos dados ao longo do tempo.

Para efeitos de comparação, é possível também construir um gráfico temporal em que duas séries são representadas conjuntamente. Use símbolos ou cores diferentes para identificar cada uma das séries.

EXEMPLO 1.7 *Homicídios - RJ e SP*

Na Tabela 1.6, temos dados sobre o número de homicídios e a taxa de homicídios por 100.000 habitantes nos estados do Rio de Janeiro e São Paulo no período de 1980 a 2009. Nas Figuras 1.7 e 1.8, apresentamos os gráficos. Observe a diferença entre eles. Quando trabalhamos com números absolutos, São Paulo tem mais homicídios que o Rio de Janeiro. Mas São Paulo tem uma população bem maior que a do Rio de Janeiro; assim, é razoável que ocorra um número maior de homicidios. Apresentar as taxas por 100.000 habitantes elimina esse problema e nos permite ver mais claramente a real situação.

Tabela 1.6 – Número e taxa de homicídios por 100.000 habitantes

Ano	Homicídios				Ano	Homicídios			
	Número		Taxa (100.000 hab)			Número		Taxa (100.000 hab)	
	RJ	SP	RJ	SP		RJ	SP	RJ	SP
1980	2.946	3.452	26,09	13,78	1995	8.183	11.566	61,54	34,32
1981	2.508	4.187	21,98	16,39	1996	8.049	12.350	60,04	36,20
1982	2.170	4.183	18,79	15,99	1997	7.966	12.552	58,77	36,12
1983	1.861	5.836	15,91	21,79	1998	7.569	14.001	55,32	39,68
1984	2.463	7.063	20,81	25,78	1999	7.249	15.810	52,50	44,14
1985	2.550	7.015	21,29	25,04	2000	7.337	15.631	50,98	42,21
1986	2.441	7.195	20,14	25,14	2001	7.352	15.745	50,50	41,84
1987	3.785	7.918	30,87	27,09	2002	8.321	14.494	56,51	37,96
1988	3.054	7.502	24,64	25,16	2003	7.840	13.903	52,69	35,92
1989	4.287	9.180	34,22	30,21	2004	7.391	11.216	49,16	28,58
1990	7.095	9.496	56,05	30,69	2005	7.098	8.727	46,14	21,58
1991	5.039	9.671	39,34	30,62	2006	7.122	8.166	45,77	19,89
1992	4.516	9.022	34,96	28,15	2007	6.313	6.234	40,11	14,96
1993	5.362	9.219	41,04	28,19	2008	5.395	6.117	33,99	14,92
1994	6.414	9.990	78,66	30,08	2009	4.198	6.319	26,22	15,27

Fonte: IPEADATA

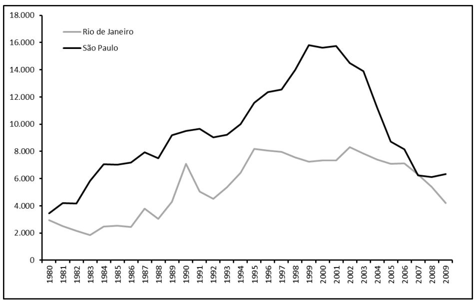


Figura 1.7 – Número de Homicídios - RJ e SP - 1980-2009

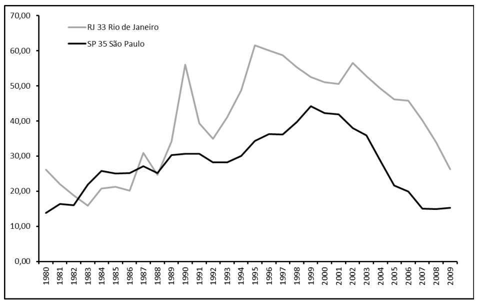


Figura 1.8 – Taxa de Homicídios (100.000 habitantes) - RJ e SP - 1980-2009

♦♦

Capítulo 2

Descrição de dados: resumos numéricos

A redução dos dados através de tabelas de frequências ou gráficos é um dos procedimentos disponíveis para se ilustrar o comportamento de um conjunto de dados. No entanto, muitas vezes, queremos resumir ainda mais esses dados, apresentando valores únicos que descrevam suas principais características. Estudaremos, neste capítulo, medidas que descrevem a tendência central, a dispersão e a assimetria das distribuições de dados.

2.1 Medidas de posição

As medidas de posição ou tendência central, como o próprio nome indica, são medidas que informam sobre a posição típica dos dados.

Na Figura 2.1, podemos notar os seguintes fatos: em (a) e (b), as distribuições são idênticas, exceto pelo fato de a segunda estar deslocada à direita. Em (c), podemos ver que há duas classes com a frequência máxima e, em (d), há uma grande concentração na cauda inferior e alguns poucos valores na cauda superior. As medidas de posição que apresentaremos a seguir irão evidenciar essas diferenças.

2.1.1 Média aritmética simples

No nosso dia a dia, o conceito de média é bastante comum, quando nos referimos, por exemplo, à altura média dos brasileiros, à temperatura média dos últimos anos etc.

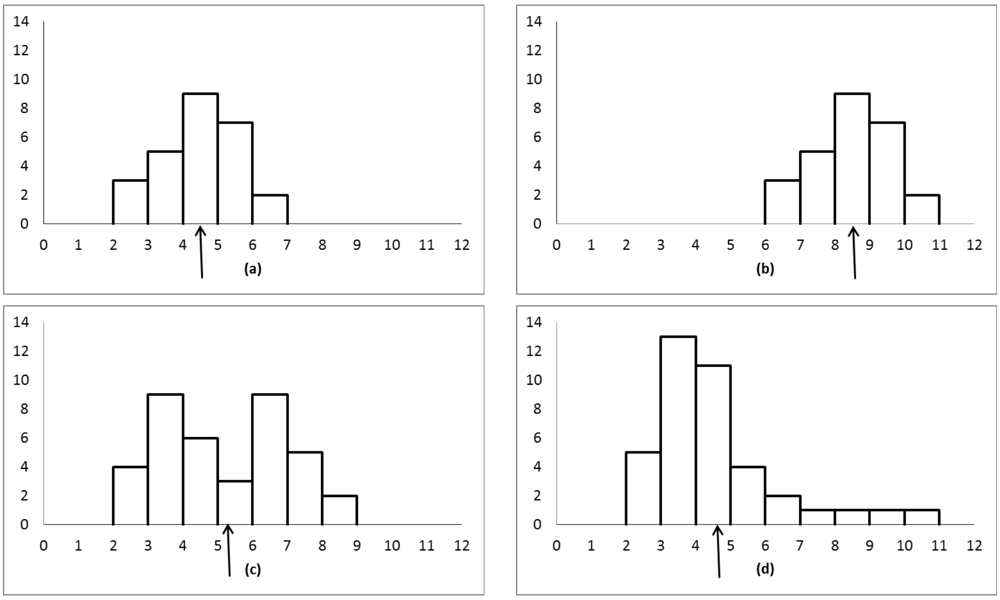


Figura 2.1 – Exemplos ilustrativos do conceito de medidas de posição

> **DEFINIÇÃO Média aritmética simples**
>
> Dado um conjunto de n observações $x_1, x_2, \ldots, x_n$, a **média aritmética simples** é definida como
>
> $$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

A notação $\overline{x}$ (lê-se x barra), usada para indicar a média, é bastante comum; em geral, usa-se a mesma letra adotada para indicar os dados com a barra em cima.

Na definição anterior, fazemos uso do símbolo de somatório, representado pela letra grega sigma maiúscula, Σ. Mais adiante, você aprenderá mais sobre essa notação e suas propriedades. Por enquanto, entenda como a média aritmética de um conjunto de dados é calculada. Observe, inicialmente, que ela só pode ser calculada para dados quantitativos. (Não faz sentido somar masculino + feminino!) O seu cálculo é feito somando-se todos os valores e dividindo-se pelo número total de observações.

Considere as idades dos funcionários do Departamento de Recursos Humanos, apresentadas no diagrama de ramo-e-folhas a seguir.

Diagrama 2.1 – *Idades de 15 Funcionários do Departamento de Recursos Humanos*

Escala
1 | 0 10

2	4	5	6	6	9	9
3	1	5	6	7	8	
4	2	5				
5	1	3				

A idade média é

$$
\begin{aligned}
\overline{x} &= \frac{24+25+26+26+29+29+31+35+36+37+38+42+45+51+53}{15} \\
&= \frac{527}{15} = 35,13
\end{aligned}
$$

Como as idades estão em anos, a idade média também é dada nessa unidade, ou seja, a idade média é 35,13 *anos*. Isso é regra geral: *a média de um conjunto de dados tem a mesma unidade dos dados originais*.

Como interpretação física da média aritmética, temos que ela representa o centro de gravidade da distribuição. Nos quatro histogramas da Figura 2.1, ela é o ponto de equilíbrio, indicado pela seta.

Note que o valor da média aritmética é um valor tal que, se substituíssemos todos os dados por ela, isto é, se todas as observações fossem iguais à média aritmética, a soma total seria igual à soma dos dados originais. Então, a média aritmética é uma forma de se distribuir o total observado por n elementos, de modo que todos tenham o mesmo valor.

Considere os seguintes dados fictícios referentes aos salários de cinco funcionários de uma firma: $136, 210, 350, 360, 2500$. O total da folha de pagamentos é 3236, havendo um salário bastante alto, discrepante dos demais. A média para esses dados é 647,20. Se todos os cinco funcionários ganhassem esse salário, a folha de pagamentos seria a mesma, e todos teriam o mesmo salário.

2.1.2 Moda

No histograma (c) da Figura 2.1, duas classes apresentam a mesma frequência máxima. Esse é o conceito de *moda*.

DEFINIÇÃO **Moda**

A **moda** de uma distribuição ou conjunto de dados, que representaremos por x^*, é o valor que mais se repete, ou seja, o valor mais frequente.

Podemos ter distribuições amodais (todos os valores ocorrem o mesmo número de vezes), unimodais (uma moda), bimodais (duas modas), etc. Para os dados do Diagrama 2.1, temos as seguintes modas: $x^* = 26$ e $x^* = 29$ anos e, portanto, essa é uma distribuição bimodal. Assim como a média, *a moda sempre tem a mesma unidade dos dados originais*. Mas note que a moda é sempre igual a um dos dados originais, o que não ocorre com a média.

2.1.3 Mediana

Vamos analisar, novamente, os seguintes dados referentes aos salários (em R$) de cinco funcionários de uma firma: 136, 210, 350, 360, 2500. Como visto, o salário médio é R$ 647,20. No entanto, esse valor não representa não representa, de forma adequada, os salários mais baixos e o salário mais alto, isso porque o mais alto é muito diferente dos demais.

Esse exemplo ilustra um fato geral sobre a média aritmética: ela é muito influenciada por *valores discrepantes* (em inglês, *outliers*), isto é, valores muito grandes (ou muito pequenos) que sejam distintos da maior parte dos dados. Nesses casos, é necessário utilizar outra medida de posição para representar o conjunto. Uma medida possível de ser utilizada é a *mediana*.

DEFINIÇÃO Mediana

Seja $x_1, x_2, \dots, x_n$ um conjunto de n observações, e seja $x_{(i)},\ i = 1, \dots, n$ o conjunto das observações ordenadas, de modo que $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$. Então, a **mediana** Q_2 é definida como o valor tal que 50% das observações são menores e 50% são maiores que ela. Para efeito de cálculo, valem as seguintes regras:

$$\begin{array}{ll} n \text{ ímpar:} & Q_2 = x_{\left(\frac{n+1}{2}\right)} \\ n \text{ par:} & Q_2 = \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \end{array} \tag{2.2}$$

Dessa definição, podemos ver que a mediana é o valor central dos dados e, para calculá-la, é necessário ordenar os dados. Para as idades no Diagrama 2.1, o número total de observações é $n = 15$. A mediana é o valor central, que deixa sete observações abaixo e sete observações acima. Logo, a mediana é a oitava observação, uma vez que

$$\frac{n+1}{2} = \frac{15+1}{2} = 8.$$

Sendo assim, a idade mediana é $Q_2 = 35$ anos. *A unidade de medida da mediana é a mesma dos dados.*

Note que, da definição de mediana, tem-se que sua posição é sempre dada por $\frac{n+1}{2}$. Quando esse cálculo resultar em um número inteiro, a mediana será a observação nessa posição. Caso contrário, a mediana será a média dos dois valores centrais. Por exemplo, se o resultado for 20,5, então a mediana será a média da vigésima e da vigésima primeira observações na lista ordenada. Já se o resultado for 7,5, a mediana será a média da sétima e da oitava observações na lista ordenada. Se o resultado for 9, a mediana será a nona observação na lista ordenada dos dados.

EXEMPLO 2.1 *Número de dependentes dos funcionários do departamento de RH*

Vamos calcular as medidas de posição para os dados referentes ao número de dependentes dos funcionários do Departamento de Recursos Humanos, apresentados na tabela abaixo.

Nome	Dependentes	Nome	Dependentes
João da Silva	3	Ana Freitas	1
Patrícia Silva	2	Pedro Barbosa	2
Pedro Fernandes	1	Luiz Costa	3
Regina Lima	2	Ricardo Alves	0
Maria Freitas	0	André Souza	4
Alfredo Souza	3	Márcio Rezende	1
Paula Gonçalves	0	Ana Carolina Chaves	0
Margarete Cunha	0		

Os dados ordenados são

$$0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4$$

e a média é

$$\overline{x} = \frac{5 \times 0 + 3 \times 1 + 3 \times 2 + 3 \times 3 + 1 \times 4}{15} = \frac{22}{15} = 1,47$$

Em média, temos 1,47 dependentes por funcionário do Departamento de RH. A moda é 0 dependente e a mediana é ($n = 15$)

$$Q_2 = x_{\left(\frac{15+1}{2}\right)} = x_{(8)} = 1 \text{ dependente.}$$

♦♦

EXEMPLO 2.2 *Notas de 50 alunos*

No capítulo anterior, obtivemos o diagrama de ramo-e-folhas a seguir para as notas de 50 alunos.

Diagrama 2.2 – *Notas de 50 alunos*

Escala

1 | 0   1,0

2	9													
3	7	8												
4	7	9												
5	2	4	5	6	8									
6	0	2	3	5	5	6	8	8	9	9	9			
7	0	0	0	1	2	3	3	4	5	5	6	6	7	9
8	0	1	2	2	3	3	4	7	7	8	9			
9	0	2	4	7										

Com $n = 50$, a posição da mediana é

$$\frac{n+1}{2} = \frac{51}{2} = 25,5 \tag{2.3}$$

e, assim, a mediana é a média das observações nas posições 25 e 26, ou seja,

$$Q_2 = \frac{71+72}{2} = 71,5 \tag{2.4}$$

Essa é uma distribuição bimodal, com modas $x^* = 69$ e $x^* = 70$. A média é

$$\overline{x} = \frac{3529}{50} = 70,58 \tag{2.5}$$

♦♦

2.1.4 Separatrizes

A mediana é um caso particular de um conjunto mais amplo de medidas estatísticas, chamadas *separatrizes*.

DEFINIÇÃO Separatriz

A **separatriz de ordem** p é um valor tal que pelo menos p% dos dados são menores do que ele e pelo menos $(1-p)$% são maiores.

As separatrizes mais comuns são os *quartis*, *decis* e *percentis*, cujos fatores de divisão são 4, 10 e 100. Mais precisamente, existem 3 quartis, 9 decis e 99 percentis. Os quartis serão representados pela letra Q e são eles:

- primeiro quartil Q_1 : deixa pelo menos 25% das observações abaixo dele e pelo menos 75% acima;
- segundo quartil Q_2 : deixa pelo menos 50% das observações abaixo dele e pelo menos 50% acima; Q_2 é a mediana;
- terceiro quartil Q_3 : deixa pelo menos 75% das observações abaixo dele e pelo menos 25% acima.

Os decis serão representados pela letra D e os percentis pela letra P; assim, por exemplo:

- o terceiro decil D_3 deixa pelo menos 30% das observações abaixo e pelo menos 70% acima;
- o quinto decil e o 50º percentil são a mediana;
- o octagésimo percentil deixa pelo menos 80% das observações abaixo e pelo menos 20% acima.

No cálculo das separatrizes quase sempre será necessário algum procedimento de arredondamento e aproximação. Para os quartis, podemos adotar o seguinte procedimento: depois de calculada a mediana, considere as duas partes dos dados, a parte abaixo da mediana e a parte acima da mediana, em ambos os casos excluindo a mediana. O primeiro quartil pode ser calculado como a mediana da parte abaixo da mediana original e o terceiro quartil como a mediana da parte acima da mediana original.

EXEMPLO 2.3 *Notas de duas turmas*

Consideremos, novamente, os dados do Exemplo 1.6, cujo diagrama de ramo-e-folhas apresentamos a seguir.

Diagrama 2.3 – *Notas dos alunos das turmas A (lado esquerdo) e B (lado direito)*

Escala

1	0	1,0
8	1	
3 2 2	2	9
	3	7 8
7 5 0	4	7 9
2 1	5	2 4 5 6 8
6 5 4 3 3 2 0 0 0 0	6	0 2 3 5 5 6 8 8 9 9 9
2 2 2 0 0	7	0 0 0 1 2 3 3 4 5 5 6 6 7 9
4 3 3 2 1 0 0	8	0 1 2 2 3 3 4 7 7 8 9
5	9	0 2 4 7

Consideremos as notas da turma A: temos 32 observações e a mediana é a média dos valores centrais (16ª e 17ª observações). Então, as duas partes consistem nas 16 observações inferiores e nas 16 observações superiores, respectivamente. Como 16 é um número par, a mediana é a média dos valores centrais, ou seja, o primeiro quartil é a média da oitava e da nona observações. Analogamente, o terceiro quartil é a média da oitava e da nona observações da metade superior; como na metade inferior já temos 16 observações, o terceiro quartil será a média da $(16+8)^{a}$ e da $(16+9)^{a}$, ou seja, Q_3 é a média da $(24)^{a}$ e da $(25)^{a}$ observação. Veja a Figura 2.2.

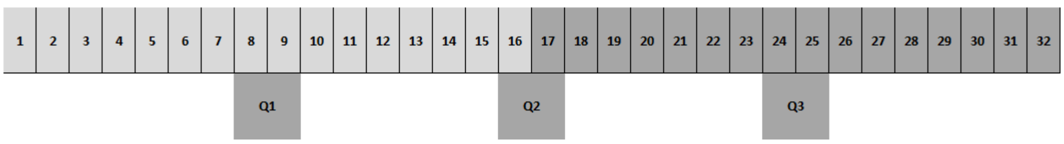


Figura 2.2 – Cálculo dos quartis de 32 notas

Analogamente, para a turma B, que tem 50 notas, a mediana é a média da 25ª e da 26ª observações. Em cada metade, ficam 25 observações e o primeiro quartil é a observação central da metade inferior, ou seja, Q_1 é a (13ª observação. O terceiro quartil será a observação central da metade superior, ou seja, Q_3 é a $(25+13)^{a}$ observação. Veja a 2.3.

O primeiro decil para as notas da turma A pode ser calculado como (note que $\frac{32}{10} = 3,2)$:

$$D_{1,B} = x_{(4)} = 2,3$$

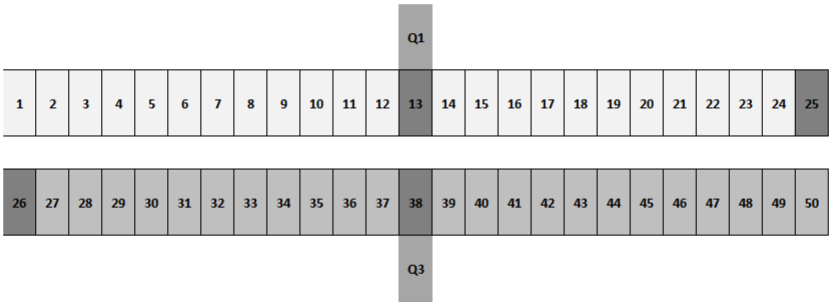


Figura 2.3 – Cálculo dos quartis de 50 notas

e o quarto decil como (note que $4 \times \dfrac{32}{10} = 12,8 \simeq 13)$:

$$D_{4,B} = x_{(13)} = 6,0$$

Todos esses arredondamentos são necessários mas um pouco arbitrários; não existe uma regra definida para tratar as diversas situações e diferentes programas podem dar resultados diferentes.

2.1.5 Média aritmética ponderada

Vimos que a média aritmética simples equivale a dividir o "todo" (soma dos valores) em partes iguais, ou seja, estamos supondo que os números que desejamos sintetizar têm o mesmo grau de importância. Entretanto, em algumas situações não é razoável atribuir a mesma importância a todos os dados.

Por exemplo, o Índice Nacional de Preços ao Consumidor (INPC) é calculado com uma média dos Índices de Preço ao Consumidor (IPC) de diversas regiões metropolitanas do Brasil, mas a importância dessas regiões é diferente. Uma das variáveis que as diferencia é a população residente. Nesse tipo de situação, em vez de se usar a média aritmética simples, adota-se a *média aritmética ponderada*, que será representada por $\overline{x}_p$.

DEFINIÇÃO Média aritmética ponderada

A **média aritmética ponderada** de números $x_1, x_2, \dots, x_n$ com pesos $\omega_1, \omega_2, \dots, \omega_n$ é definida como

$$\overline{x}_p = \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n = \sum_{i=1}^{n} \omega_i x_i, \qquad (2.6)$$

em que $\omega_1 + \omega_2 + \cdots + \omega_n = \sum_{i=1}^{n} \omega_i = 1$

Note que a média aritmética simples é um caso particular da média aritmética ponderada, onde todas as observações têm o mesmo peso $\omega_i = \dfrac{1}{n}$.

EXEMPLO 2.4 *INPC*

Para a construção do Índice Nacional de Preços ao Consumidor (INPC), o peso de cada índice regional é definido pela população residente urbana, conforme dados da Tabela 2.1. Os pesos, apresentados em porcentagem, representam a participação da população residente urbana da região metropolitana no total da população residente urbana das 11 regiões metropolitanas pesquisadas.

Tabela 2.1 – Estrutura básica de ponderação regional para cálculo do INPC - Agosto 2012

Área Geográfica	Peso (%)	IPC - Ago/12
Belém	6,9	0,74
Fortaleza	6,4	0,83
Recife	7,1	0,45
Salvador	10,6	0,29
Belo Horizonte	11,1	0,48
Rio de Janeiro	10,2	0,59
São Paulo	25,6	0,27
Curitiba	7,2	0,44
Porto Alegre	7,5	0,57
Goiânia	5,1	0,36
Distrito Federal	2,2	0,31
INPC - Geral		0,45

Fonte: IBGE

O índice geral, dado pela média ponderada, é calculado como

$$\begin{aligned}
\text{INPC}_{08/12} &= 0,069 \times 0,74 + 0,064 \times 0,83 + 0,071 \times 0,45 + \\
&\quad 0,106 \times 0,29 + 0,111 \times 0,48 + 0,102 \times 0,59 + \\
&\quad 0,256 \times 0,27 + 0,072 \times 0,44 + 0,075 \times 0,57 + \\
&\quad 0,051 \times 0,36 + 0,022 \times 0,31 = 0,44906 \simeq 0,45
\end{aligned}$$

♦♦

EXEMPLO 2.5 *Nota Média*

Segundo o critério de avaliação adotado pelo Departamento de Estatística, cada aluno será submetido a duas provas, a primeira tendo peso 2 e a segunda tendo peso 3. Para ser aprovado sem precisar fazer prova final, a média obtida nas duas provas deve ser, no mínimo, 6. Se um aluno tirar 5,5 na primeira prova, quanto deverá tirar na segunda prova para não precisar fazer prova final?

Solução

A média nas duas provas é calculada como

$$\overline{x}_p = \frac{2 \times N_1 + 3 \times N_2}{2+3} = \frac{2 \times N_1 + 3 \times N_2}{5}$$

O problema pede que $\overline{x}_p \geq 6$. Então é necessário ter

$$\frac{2 \times 5,5 + 3 \times N_2}{5} \geq 6 \Rightarrow N_2 \geq 6,33$$

O aluno deve tirar nota maior que 6,3 para que não precise fazer prova final.

♦♦

2.1.6 Propriedades das medidas de posição

Da interpretação física da média como centro de gravidade da distribuição, fica claro que seu valor está sempre entre os valores mínimo e máximo dos dados. O mesmo resultado vale para a mediana e a moda, o que é imediato a partir das respectivas definições. Resumindo, temos:

Propriedade 1

$$\begin{aligned} x_{\min} &\leq \overline{x} \leq x_{\max} \\ x_{\min} &\leq Q_2 \leq x_{\max} \\ x_{\min} &\leq x^* \leq x_{\max} \end{aligned} \tag{2.7}$$

Iremos apresentar as outras duas propriedades através do seguinte exemplo:

Em uma turma de estatística, os resultados de uma prova ficaram abaixo do que a professora esperava. Como todos os alunos participavam ativamente de todas as atividades, demonstrando interesse especial pela matéria, a professora resolveu dar um ponto a mais na prova para todos os alunos. Além disso, ela deu os resultados com as notas variando de 0 a 10, mas a secretaria da faculdade exige que as notas sejam dadas em uma escala de 0 a 100. Sendo assim, a professora precisa multiplicar todas as notas por 10. O que acontecerá com a média, a moda e a mediana depois dessas alterações?

Vamos ver o que ocorre, selecionando como exemplo o seguinte conjunto de cinco notas: $5, 4, 2, 3, 4$.

As notas ordenadas são $2, 3, 4, 4, 5$ e temos as seguintes medidas de posição:

$$\begin{aligned} \overline{x} &= \frac{5+4+2+3+4}{5} = \frac{18}{5} = 3,6 \\ Q_2 &= x^* = 4 \end{aligned}$$

Somando 1 ponto, as notas passam a ser $3, 4, 5, 5, 6$ com as seguintes medidas de posição:

$$\begin{aligned} \overline{y} &= \frac{3+4+5+5+6}{5} = \frac{23}{5} = 4,6 = 3,6+1 \\ Q_{2,y} &= y^* = 5 = 4+1 \end{aligned}$$

Ao somar 1 ponto em todas as notas, o conjunto sofre um deslocamento (uma translação), o que faz com que o seu centro também fique deslocado 1 ponto. Sendo assim, todas as três medidas de posição ficam acrescidas de 1 ponto.

Multiplicando as novas notas por 10, obtemos 30, 40, 50, 50, 60 e

$$\begin{aligned} \overline{z} &= \frac{30+40+50+50+60}{5} = \frac{230}{5} = 46,0 = 4,6 \times 10 \\ Q_{2,z} &= z^* = 50 = 5 \times 10, \end{aligned}$$

ou seja, todas as medidas de posição ficam multiplicadas por 10.

Esse exemplo ilustra as propriedades a seguir.

Propriedade 2

Somando-se um mesmo valor a cada observação x_i, obtemos um novo conjunto de dados $y_i = x_i + k$, para o qual temos as seguintes medidas de posição:

$$y_i = x_i + k \Rightarrow \begin{cases} \overline{y} = \overline{x} + k \\ Q_{2,y} = Q_{2,x} + k \\ y^* = x^* + k \end{cases} \tag{2.8}$$

Propriedade 3

Multiplicando cada observação x_i por uma mesma constante não nula k, obtemos um novo conjunto de dados $y_i = kx_i$, para o qual temos as seguintes medidas de posição:

$$y_i = kx_i \Rightarrow \begin{cases} \overline{y} = k\overline{x} \\ Q_{2,y} = kQ_{2,x} \\ y^* = kx^* \end{cases} \tag{2.9}$$

EXEMPLO 2.6 *Temperaturas*

A relação entre as escalas Celsius e Fahrenheit é a seguinte:

$$C = \frac{5}{9}(F - 32)$$

Se a temperatura média em determinada localidade for de $45°F$, qual será a temperatura média em graus Celsius?

Solução

Se cada observação for transformada de graus Fahrenheit para Celsius, a média sofrerá a mesma mudança, ou seja,

$$\overline{x} = 45°F \Rightarrow \overline{y} = \frac{5}{9}(45 - 32) = 7,2°C$$

♦♦

2.2 Medidas de dispersão

Considere os conjuntos de dados representados por *diagramas de pontos* na Figura 2.4. Nesses gráficos, as "pilhas" de pontos representam as frequências de cada valor. Podemos ver facilmente que os três conjuntos têm a mesma média (o centro de gravidade ou ponto de equilíbrio é o mesmo), a mesma mediana e a mesma moda. No entanto, esses conjuntos têm características diferentes, e ao sintetizá-los com base em apenas uma medida de posição essas características se perderão. Tal característica é a *dispersão* dos dados e iremos estudar algumas medidas de dispersão que nos permitirão diferenciar entre essas três distribuições.

2.2.1 Amplitude

Analisando os diagramas da Figura 2.4, vemos que os valores se distribuem entre 4 e 8 na distribuição (a) ao passo que, nas distribuições (b) e (c), eles se encontram mais dispersos,

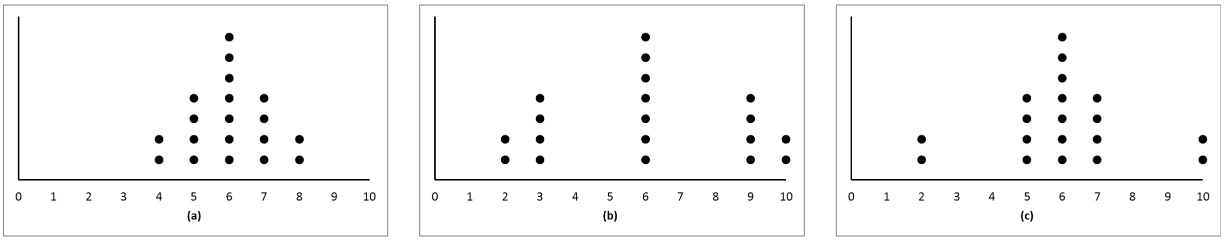


Figura 2.4 – Exemplos ilustrativos do conceito de medidas de dispersão

variando de 2 a 10. Considerar, então, a distância entre o mínimo e o máximo nos permite quantificar diferenças nas dispersões. Como já visto, esse é o conceito de *amplitude*.

DEFINIÇÃO **Amplitude**

A **amplitude** de um conjunto de dados é a distância entre o maior valor e o menor valor.

$$\Delta_{total} = V_{\max} - V_{\min}. \tag{2.10}$$

A amplitude tem a mesma unidade dos dados, mas, como medida de dispersão, ela tem algumas limitações, conforme ilustrado nas distribuições (b) e (c) da Figura 2.4, que possuem a mesma média, a mesma mediana e a mesma amplitude. No entanto, essas medidas não conseguem caracterizar o fato de a distribuição dos valores entre o mínimo e o máximo ser diferente nos dois conjuntos. A limitação da amplitude também fica patente pelo fato de ela se basear em apenas duas observações, independentemente do número total de observações.

2.2.2 Desvio médio absoluto

Uma maneira de se medir a dispersão dos dados é considerar os tamanhos dos *desvios* $x_i - \overline{x}$ de cada observação em relação à média. Observe, nos exemplos da Figura 2.4, que quanto mais disperso for o conjunto de dados, maiores serão os desvios. Para obtermos uma medida-resumo, isto é, um único número, poderíamos somar esses desvios, considerando a seguinte medida:

$$D = \sum_{i=1}^{n} (x_i - \overline{x}) = (x_1 - \overline{x}) + (x_2 - \overline{x} + \cdots + (x_n - \overline{x}). \tag{2.11}$$

No entanto, essa medida, que representa a soma dos desvios em relação à média, é sempre nula, não importa o conjunto de dados! Logo, ela não serve para diferenciar quaisquer conjuntos!

Daremos uma explicação intuitiva para esse fato, que nos permitirá obter correções para tal fórmula. Pela definição de média, sempre há valores menores e maiores que a média, que resultam, respectivamente, em desvios negativos e positivos. Esses desvios positivos e negativos, ao serem somados, se anulam.

Pois bem, se o problema está no fato de termos desvios positivos e negativos, por que não trabalhar com o valor absoluto das diferenças? De fato, esse procedimento nos leva à definição de *desvio médio absoluto*.

DEFINIÇÃO **Desvio médio absoluto**

O **desvio médio absoluto** de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido por

$$DMA = \frac{|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \overline{x}| \tag{2.12}$$

onde as barras verticais representam o valor absoluto ou módulo.

Note que, nessa definição, estamos trabalhando com o desvio médio, isto é, tomamos a média dos desvios absolutos. Isso evita interpretações equivocadas, pois, se trabalhássemos apenas com a soma dos desvios absolutos, um conjunto com um número maior de observações tenderia a apresentar um resultado maior para a soma, devido apenas ao fato de ter mais observações. Esta situação é ilustrada com os seguintes conjuntos de dados:

- Conjunto 1: $\{1, 3, 5\}$
- Conjunto 2: $\left\{1, \frac{5}{3}, 3, \frac{13}{3}, 5\right\}$

Para os dois conjuntos, $\overline{x} = 3$, e para o conjunto 1,

$$\sum_{i=1}^{3}|x_i - \overline{x}| = |1-3| + |3-3| + |5-3| = 4$$

Já para o conjunto 2,

$$\begin{aligned}\sum_{i=1}^{5}|x_i - \overline{x}| &= |1-3| + \left|\frac{5}{3} - 3\right| + |3-3| + \left|\frac{13}{3} - 3\right| + |5-3| \\ &= \frac{20}{3} = 6,667.\end{aligned}$$

Então, o somatório para o segundo conjunto é maior, mas o desvio médio absoluto é o mesmo para ambos. De fato, para o primeiro conjunto, temos

$$DMA = \frac{4}{3}$$

e, para o segundo conjunto,

$$DMA = \frac{\frac{20}{3}}{5} = \frac{4}{3}$$

Ao dividirmos o somatório pelo número de observações, compensamos o fato de o segundo conjunto ter mais observações do que o primeiro.

O desvio médio absoluto tem a mesma unidade dos dados.

2.2.3 Variância e desvio-padrão

Considerar o valor absoluto das diferenças $(x_i - \overline{x})$ é uma das maneiras de se contornar o fato de a soma dos desvios em torno da média ser zero. Mas há uma outra possibilidade de correção, com propriedades matemáticas e estatísticas mais adequadas, que consiste em trabalhar com o quadrado dos desvios. Isso nos leva à definição de *variância*.

DEFINIÇÃO Variância

A **variância** de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definida por

$$\sigma^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{2.13}$$

Essa definição nos diz que *a variância é a média dos desvios quadráticos*. Uma expressão alternativa mais simples de ser usada em cálculos manuais é dada por

$$\sigma^2 = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2 \tag{2.14}$$

Essa forma de escrever a variância facilita os cálculos feitos à mão ou em calculadoras menos sofisticadas, pois o número de cálculos envolvidos é menor. Podemos ler essa fórmula como *a variância é a média dos quadrados menos o quadrado da média*.

Suponhamos que os valores x_i representem os pesos, em quilogramas, de um conjunto de pessoas. Então, o valor médio $\overline{x}$ representa o peso médio dessas pessoas e sua unidade também é quilogramas, o mesmo acontecendo com as diferenças $(x_i - \overline{x})$. Ao elevarmos essas diferenças ao quadrado, passamos a ter a variância medida em quilogramas ao quadrado, uma unidade que não tem interpretação física. Uma forma de se obter uma medida de dispersão, com a mesma unidade dos dados, consiste em tomar a raiz quadrada da variância.

DEFINIÇÃO **Desvio-padrão**

O **desvio-padrão** de um conjunto de dados $x_1, x_2, \ldots, x_n$ é definido como a raiz quadrada da variância:

$$\sigma = \sqrt{\text{Variância}} = \sqrt{\sigma^2} \tag{2.15}$$

EXEMPLO 2.7 *Idades de funcionários*

Novamente, vamos considerar os dados referentes às idades dos funcionários do Departamento de Recursos Humanos. Essas idades são:

24 25 26 26 29 29 31 35 36 37 38 42 45 51 53

e sua média é $\frac{527}{15} = 35,1\overline{3}$. Assim, a variância, em anos^2, é

$$\begin{aligned}
\sigma^2 &= \frac{1}{15}\left[\begin{array}{c}(24-35,13)^2+(25-35,13)^2+2\times(26-35,13)^2+\\ 2\times(29-35,13)^2+(31-35,13)^2+(35-35,13)^2+\\ (36-35,13)^2+(37-35,13)^2+(38-35,13)^2+\\ (42-35,13)^2+(42-35,13)^2+(45-35,13)^2+\\ (51-35,13)^2+(53-35,13)^2\end{array}\right] = \\
&= \frac{1213,73}{15} = 80,92
\end{aligned}$$

e o desvio-padrão, em anos, é

$$\sigma = \sqrt{80,92} = 8,995$$

Usando a fórmula 2.14, temos:

$$\begin{aligned}
\sigma^2 &= \frac{1}{15}\left[24^2+25^2+25^2+2\times 26^2+2\times 29^2+31^2+35^2+36^2\right]+\\
&+ \frac{1}{15}\left[37^2+38^2+39^2+42^2+45^2+51^2+53^2\right]-\left(\frac{527}{15}\right)^2 = \\
&= \frac{19729}{15}-\left(\frac{527}{15}\right)^2 = \\
&= \frac{19729\times 15-527^2}{15^2} = \frac{295935-277729}{225} = \frac{18206}{225} = 80,916
\end{aligned}$$

Na comparação dos resultados obtidos pelas duas fórmulas, pode haver alguma diferença por causa dos arredondamentos, uma vez que a média é uma dízima. Em geral, a fórmula 2.14 fornece resultados mais precisos e certamente requer menos cálculos.

♦♦

EXEMPLO 2.8 *Número de dependentes dos funcionários do departamento de RH*

Consideremos, novamente, o número de dependentes dos funcionários do Departamento de Recursos Humanos, apresentados no Exemplo 2.1. Os dados são

$$3 \quad 2 \quad 1 \quad 2 \quad 0 \quad 3 \quad 0 \quad 0 \quad 1 \quad 2 \quad 3 \quad 0 \quad 4 \quad 1 \quad 0$$

Como o menor valor é 0 e o maior é 4, temos que a amplitude dos dados é de 4 dependentes. A média calculada para esses dados foi $\overline{x} = \frac{22}{15} = 1,467$. Vamos calcular a soma dos desvios em torno da média, usando o fato de termos observações repetidas.

$$\begin{aligned}
\sum(x_i - \overline{x}) &= 5 \times \left(0 - \frac{22}{15}\right) + 3 \times \left(1 - \frac{22}{15}\right) + 3 \times \left(2 - \frac{22}{15}\right) + \\
&+ 3 \times \left(3 - \frac{22}{15}\right) + \left(4 - \frac{22}{15}\right) = \\
&= -\frac{110}{15} - \frac{21}{15} + \frac{24}{15} + \frac{69}{15} + \frac{38}{15} = -\frac{131}{15} + \frac{131}{15} = 0
\end{aligned}$$

Caso trabalhássemos com o valor aproximado 1,467, o resultado aproximado seria −0,005.

O desvio médio absoluto é

$$\begin{aligned}
DMA &= \frac{1}{n}\sum |x_i - \overline{x}| = \\
&= \frac{1}{15} \times \left[5 \times \left|0 - \frac{22}{15}\right| + 3 \times \left|1 - \frac{22}{15}\right| + 3 \times \left|2 - \frac{22}{15}\right|\right] + \\
&+ \left[3 \times \left|3 - \frac{22}{15}\right| + \left|4 - \frac{22}{15}\right|\right] = \\
&= \frac{1}{15} \times \left[\frac{110}{15} + \frac{21}{15} + \frac{24}{15} + \frac{69}{15} + \frac{38}{15}\right] = \\
&= \frac{1}{15} \times \left[\frac{131}{15} + \frac{131}{15}\right] = \frac{262}{225} = 1,1644
\end{aligned}$$

A variância é

$$\begin{aligned}
\sigma^2 &= \frac{1}{n}\sum(x_i-\overline{x})^2 = \\
&= \frac{1}{15}\times\left[5\times\left(0-\frac{22}{15}\right)^2+3\times\left(1-\frac{22}{15}\right)^2+3\times\left(2-\frac{22}{15}\right)^2\right]+ \\
&+ \frac{1}{15}\times\left[3\times\left(3-\frac{22}{15}\right)^2+\left(4-\frac{22}{15}\right)^2\right]= \\
&= \frac{1}{15}\times\left[\frac{2420}{225}+\frac{147}{225}+\frac{192}{225}+\frac{1587}{225}+\frac{1444}{225}\right]= \\
&= \frac{5790}{15\times 225}=1,715556
\end{aligned}$$

e

$$\sigma=\sqrt{\frac{5790}{15\times 225}}=1,3098$$

Vamos agora calcular a variância usando a fórmula alternativa:

$$\begin{aligned}
\sigma^2 &= \frac{1}{15}\times\left(5\times 0^2+3\times 1^2+3\times 2^2+3\times 3^2+4^2\right)-\left(\frac{22}{15}\right)^2= \\
&= \frac{3+12+27+16}{15}-\frac{484}{225}=\frac{58}{15}-\frac{484}{225}=\frac{58\times 15-484}{225}= \\
&= \frac{386}{225}=1,715556
\end{aligned}$$

Com essa fórmula, os cálculos ficam bem mais simples, uma vez que é necessário fazer menos conta!

♦♦

2.2.4 Amplitude interquartil

Assim como a média, a variância e o desvio-padrão são muito afetados por valores discrepantes. Vamos, então, apresentar uma outra medida de dispersão que não se altera tanto na presença de tais valores atípicos. Essa medida se baseia nos *quartis*, definidos anteriormente.

Vimos que os três quartis dividem o conjunto de dados em 4 partes com o mesmo número de observações. Da definição dos quartis, resulta que, entre Q_1 e Q_3, há sempre 50% dos dados, qualquer que seja a distribuição. Assim, quanto maior for a distância entre Q_1 e Q_3, mais dispersos serão os dados. Temos, assim, uma nova medida de dispersão, a *amplitude interquartil*.

DEFINIÇÃO Amplitude interquartil

A **amplitude interquartil**, que denotaremos por AIQ, é definida como a distância entre o primeiro e o terceiro quartis, isto é:

$$AIQ = Q_3 - Q_1 \tag{2.16}$$

A amplitude interquartil tem a mesma unidade dos dados. A vantagem da amplitude interquartil sobre o desvio-padrão é que, assim como a mediana, a AIQ não é muito influenciada por poucos valores discrepantes.

EXEMPLO 2.9 *Número de dependentes dos funcionários*

Vamos calcular os quartis e a amplitude interquartil para o número de dependentes dos funcionários do Departamento de Recursos Humanos, cujos valores já ordenados são:

$$0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad 1 \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4$$

Como há 15 observações, a mediana é a oitava observação:

$$0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1 \quad \boxed{1} \quad 2 \quad 2 \quad 2 \quad 3 \quad 3 \quad 3 \quad 4$$

isto é,

$$Q_2 = x_{\left(\frac{n+1}{2}\right)} = x_{(8)} = 1$$

Excluída a oitava observação, a parte inferior dos dados, com 7 observações, é

$$0 \quad 0 \quad 0 \quad \boxed{0} \quad 0 \quad 1 \quad 1$$

cuja mediana é a observação marcada, ou seja:

$$Q_1 = x_{\left(\frac{7+1}{2}\right)} = x_{(4)} = 0$$

A parte superior dos dados, excluída a mediana, é

$$2 \quad 2 \quad 2 \quad \boxed{3} \quad 3 \quad 3 \quad 4$$

e, portanto,

$$Q_3 = x_{(4+8)} = x_{(12)} = 3$$

A amplitude interquartil é calculada como

$$AIQ = Q_3 - Q_1 = 3 - 0 = 3.$$

♦♦

2.2.5 Propriedades das medidas de dispersão

Como visto para as medidas de posição, vamos estudar as principais propriedades das medidas de dispersão.

Propriedade 1

Todas as medidas de dispersão são não negativas:

$$\begin{aligned} &\Delta \geq 0 \\ &DMA \geq 0 \\ &\sigma^2 \geq 0 \\ &\sigma \geq 0 \\ &AIQ \geq 0 \end{aligned} \tag{2.17}$$

Propriedade 2

Somando-se uma mesma constante a todas as observações, as medidas de dispersão não se alteram. Essa propriedade é bastante intuitiva: note que, ao somar uma constante aos dados, estamos simplesmente fazendo uma translação dos mesmos, sem alterar a dispersão.

$$y_i = x_i + k \Rightarrow \begin{cases} \Delta_y = \Delta_x \\ DMA_y = DMA_x \\ \sigma_y^2 = \sigma_x^2 \\ \sigma_y = \sigma_x \\ AIQ_y = AIQ_x \end{cases} \tag{2.18}$$

Propriedade 3

Ao multiplicarmos todos os dados por uma constante não nula, temos:

$$y_i = kx_i \Rightarrow \begin{cases} \Delta_y = |k| \ \Delta_x \\ DMA_y = |k| \ DMA_x \\ \sigma_y^2 = k^2 \sigma_x^2 \\ \sigma_y = |k| \ \sigma_x \\ AIQ_y = |k| \ AIQ_x \end{cases} \tag{2.19}$$

Note que é razoável aparecer o módulo da constante, já que as medidas de dispersão são não negativas.

EXEMPLO 2.10 *Temperaturas*

Se o desvio-padrão das temperaturas diárias de uma determinada localidade for de $5,2°F$, qual será o desvio-padrão em graus Celsius? Lembre-se de que a relação entre as duas escalas é

$$C = \frac{5}{9}(F - 32)$$

Solução

Se cada observação for transformada de graus Fahrenheit para Celsius, a única operação que afetará o desvio-padrão será a multiplicação pelo fator 5/9, ou seja,

$$\sigma_C = \frac{5}{9} \times \sigma_F \tag{2.20}$$

♦♦

2.3 Medidas de assimetria

Considere os diagramas de pontos da Figura 2.5, onde a seta indica a média dos dados. Analisando-os, podemos ver que a principal e mais marcante diferença entre eles diz respeito à simetria da distribuição. A distribuição do centro é simétrica, enquanto as outras duas são assimétricas.

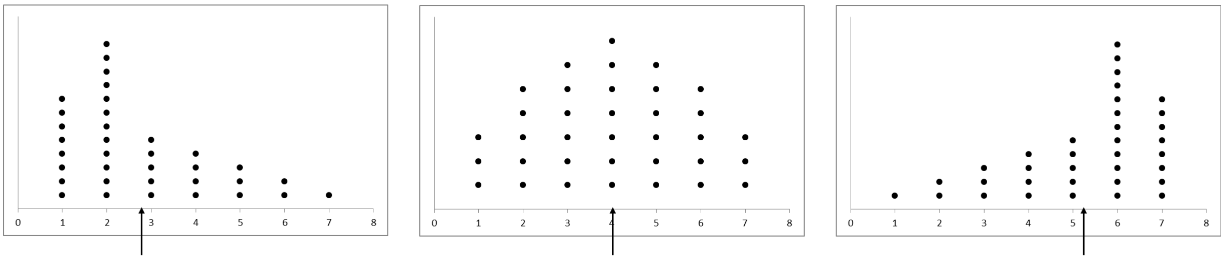


Figura 2.5 – Distribuições com diferentes tipos de assimetria

No diagrama à esquerda, a assimetria é tal que há maior concentração na cauda inferior, enquanto no diagrama à direita, a concentração é maior na cauda superior. Visto de outra maneira, no diagrama à direita, os dados se estendem para o lado positivo da escala, enquanto no diagrama à esquerda, os dados se estendem para o lado negativo da escala. Dizemos que a distribuição ilustrada no diagrama à esquerda apresenta uma *assimetria à direita*, ao passo que a do diagrama à direita apresenta uma *assimetria à esquerda*. No diagrama do centro, temos uma *simetria* perfeita ou *assimetria nula*.

DEFINIÇÃO Simetria e assimetria

Uma distribuição é simétrica se os lados direito e esquerdo do histograma (ou diagrama de pontos) são, aproximadamente, a imagem espelhada um do outro.
Uma distribuição é assimétrica à direita se a cauda direita do histograma se estende muito mais do que a cauda esquerda. Ela é assimétrica à esquerda se a cauda esquerda do histograma se estende muito mais do que a cauda direita.

2.3.1 O coeficiente de assimetria de Pearson

Esses três tipos de assimetria podem ser caracterizados pela posição da moda com relação à média dos dados. No primeiro tipo, a moda tende a estar à esquerda da média, enquanto no terceiro tipo, a moda tende a estar à direita da média. (Lembre-se de que a média é o centro de gravidade ou ponto de equilíbrio da distribuição). Para distribuições simétricas, a moda coincide com a média. Temos, assim, a seguinte caracterização dos três tipos de assimetria:

- se a média é maior que a moda ($\overline{x} > x^*$), dizemos que a distribuição é *assimétrica à direita* ou tem *assimetria positiva* [diagrama à esquerda na Figura 2.5];

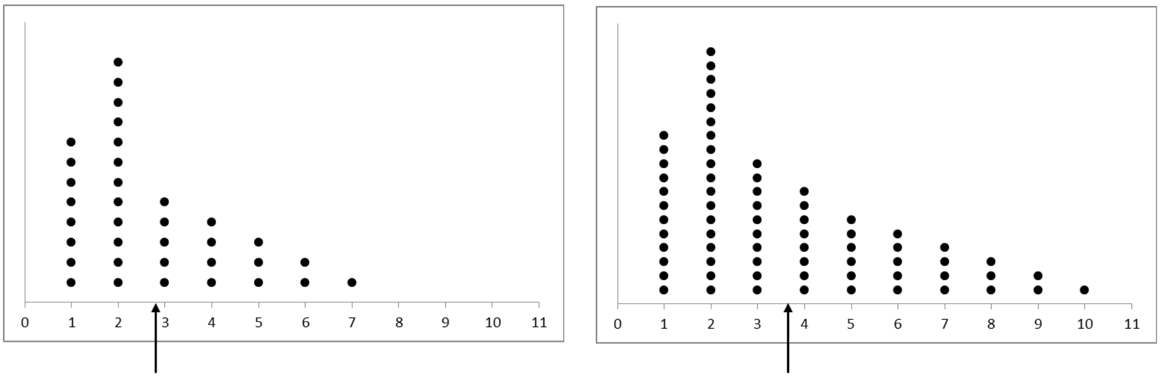


Figura 2.6 – Distribuições assimétricas à direita

- se a média é igual à moda ($\overline{x} = x^*$), dizemos que a distribuição é simétrica ou tem assimetria nula [diagrama central na Figura 2.5];
- se a média é menor que a moda ($\overline{x} < x^*$), dizemos que a distribuição é assimétrica à esquerda ou tem assimetria negativa [diagrama à direita na Figura 2.5].

Essas definições, no entanto, não permitem "medir" diferentes graus de assimetria. Por exemplo, considere os diagramas de pontos da Figura 2.6, ambos assimétricos à direita. Uma forma de medirmos essas diferentes assimetrias é através do desvio $\overline{x} - x^*$ entre a média e a moda. Mas como as distribuições podem ter graus de dispersão diferentes, é importante considerarmos a diferença acima na mesma escala. Como visto na definição dos escores padronizados, a forma de se fazer isso é dividindo o desvio pelo desvio-padrão, o que nos leva ao *coeficiente de assimetria de Pearson*.

DEFINIÇÃO Coeficiente de assimetria de Pearson

O **coeficiente de assimetria de Pearson** é definido como

$$e = \frac{\overline{x} - x^*}{\sigma}. \tag{2.21}$$

Se o coeficiente for negativo, a distribuição terá assimetria negativa; se for positivo, assimetria positiva, e se for nulo, a distribuição será simétrica.

Note que aqui, assim como nos escores padronizados, tiramos o efeito de escalas diferentes ao dividirmos pelo desvio-padrão, o que resulta na adimensionalidade do coeficiente.

Para os dados do diagrama à esquerda da Figura 2.6, temos $x^* = 2$, $\overline{x} = 2,7714$ e $\sigma = 1,6228$, logo,

$$e = \frac{2,7714 - 2}{1,6228} = 0,475351$$

Para o diagrama à direita, $x^* = 2$, $\overline{x} = 3,6232$ e $\sigma = 2,3350$, logo,

$$e = \frac{3,6232 - 2}{2,3350} = 0,6952$$

o que indica uma assimetria mais acentuada.

2.3.2 O coeficiente de assimetria de Bowley

Da definição dos quartis, sabemos que entre Q_1 e Q_2 e entre Q_2 e Q_3 há sempre 25% dos dados. Então, a diferença entre as distâncias $Q_2 - Q_1$ e $Q_3 - Q_2$ nos dá informação sobre a assimetria da distribuição.

Se $Q_2 - Q_1 < Q_3 - Q_2$, isso significa que "andamos mais rápido" para cobrir os 25% inferiores do que os 25% superiores, ou seja, a distribuição "se arrasta" para a direita.

Analogamente, se $Q_2 - Q_1 > Q_3 - Q_2$, isso significa que "andamos mais devagar" para cobrir os 25% inferiores do que os 25% superiores, ou seja, a distribuição "se arrasta" para a esquerda. De forma mais precisa, temos o seguinte resultado:

$$Q_2 - Q_1 < Q_3 - Q_2 \Longrightarrow \text{assimetria positiva}$$

$$Q_2 - Q_1 > Q_3 - Q_2 \Longrightarrow \text{assimetria negativa}$$

$$Q_2 - Q_1 = Q_3 - Q_2 \Longrightarrow \text{simetria ou assimetria nula}$$

Podemos, então, usar a diferença $(Q_3 - Q_2) - Q_2 - Q_1$ como uma medida de assimetria. Mas, aqui, também é necessário tirar o efeito de escala e, para isso, temos de dividir por uma medida de dispersão – lembre-se de que dividimos pelo desvio-padrão quando trabalhamos com as diferenças $\overline{x} - x^*$. Para não termos efeito dos valores discrepantes, usaremos a amplitude interquartil para gerar a seguinte medida de assimetria, que é chamada *coeficiente de assimetria de Bowley*.

DEFINIÇÃO Coeficiente de assimetria de Bowley

O coeficiente de assimetria de Bowley é definido como

$$B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} \tag{2.22}$$

que pode ser reescrito como

$$B = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_2) + (Q_2 - Q_1)} \tag{2.23}$$

Analisando a expressão (2.23), percebemos que, quanto mais assimétrica à direita for uma distribuição, mais próximos serão Q_1 e Q_2 e, portanto, B se aproximará de $+1$. Analogamente, quanto mais assimétrica à esquerda, mais próximos serão Q_2 e Q_3 e, portanto, B irá se aproximar de -1.

2.4 O boxplot

A partir dos quartis constrói-se um gráfico chamado *boxplot* ou *diagrama em caixa*, que ilustra os principais aspectos da distribuição e é também muito útil na comparação de distribuições.

O boxplot é formado basicamente por um retângulo vertical (ou horizontal). O comprimento do lado vertical (ou horizontal) é dado pela amplitude interquartil. Veja a Figura 2.7-(a), onde estamos trabalhando com um retângulo vertical. O tamanho do outro lado é indiferente, sugerindo-se apenas uma escala razoável. Na altura da mediana, traça-se uma linha, dividindo o retângulo em duas partes. Veja a Figura 2.7-(b).

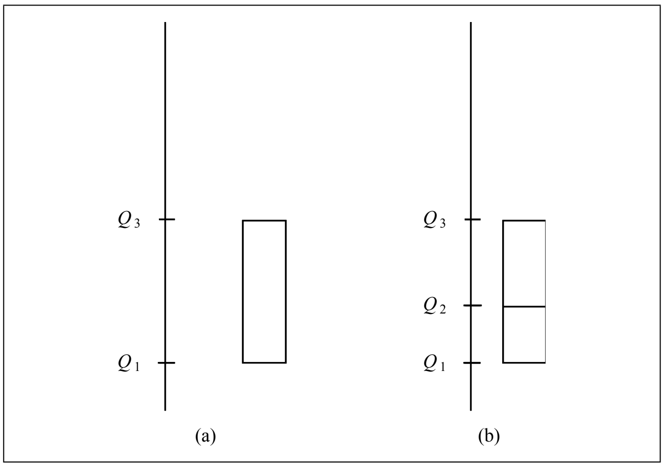


Figura 2.7 – Construção do boxplot - Parte 1

Observe que, nesse momento, não só temos representados 50% da distribuição, como também temos ideia da assimetria da mesma -? nessa figura, percebemos uma leve assimetria à direita, já que $Q_2 - Q_1 < Q_3 - Q_2$. Para representar os 25% restantes em cada cauda da distribuição, temos de cuidar, primeiro, da presença de possíveis *outliers* ou valores discrepantes, que, como já dito, são valores que se distanciam dos demais.

! Regra de valores discrepantes

Um dado x será considerado valor discrepante ou *outlier* se

$$x < Q_1 - 1,5\ AIQ$$

ou

$$x > Q_3 + 1,5\ AIQ$$

Veja a Figura 2.8-(a). Qualquer valor para fora das linhas pontilhadas é considerado um valor discrepante.

Para representar o domínio de variação dos dados na cauda inferior que não são *outliers*, traça-se, a partir do lado do retângulo definido por Q_1, uma linha para baixo até o menor valor que não seja *outlier*. Da mesma forma, na cauda superior, traça-se, a partir do lado do

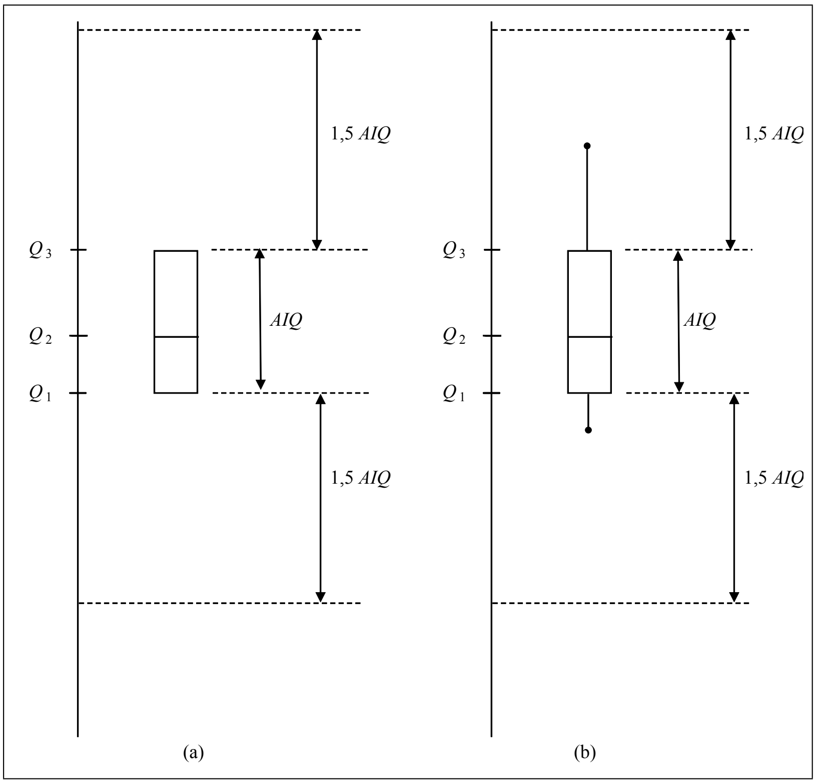


Figura 2.8 – Construção do boxplot - Parte 2

retângulo definido por Q_3, uma linha para cima até o maior valor que não seja *outlier* (veja a Figura 2.8-(b)). Esses pontos são chamados *juntas*. Dito de outra forma, as juntas são os valores mínimo e máximo do conjunto de dados formado pelos valores não discrepantes.

Quanto aos *outliers*, eles são representados individualmente por um X (ou algum outro tipo de carácter), explicitando-se, de preferência, os seus valores, mas com uma possível quebra de escala no eixo Figura 2.9).

Note que a construção do boxplot é toda baseada nos quartis, que são medidas resistentes contra valores discrepantes.

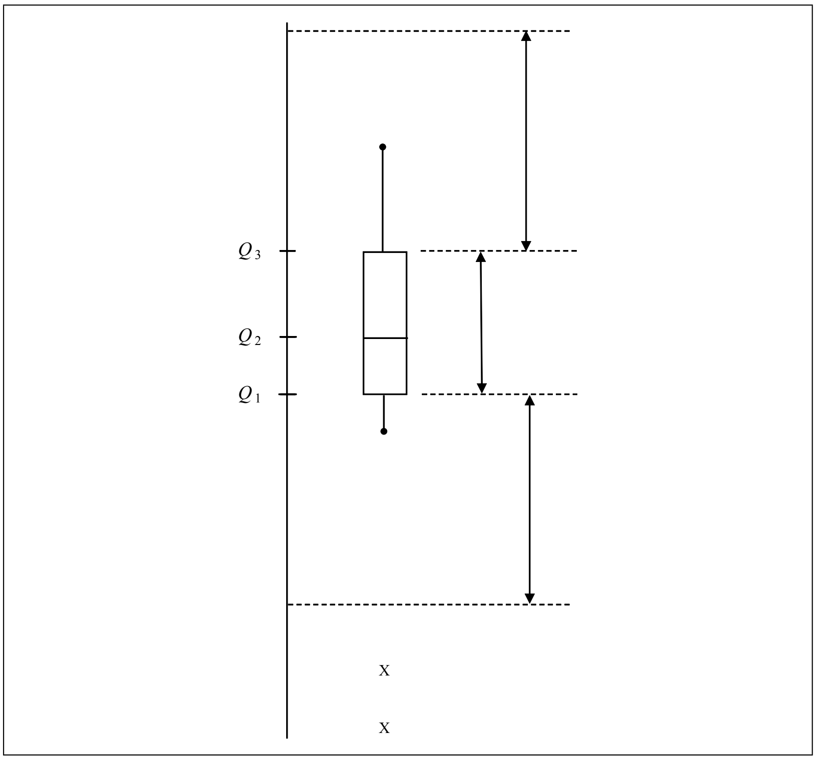


Figura 2.9 – Construção do boxplot - Parte 3

EXEMPLO 2.11 *Comprimento de flores tropicais*

Na Tabela 2.2, temos dados referentes ao comprimento das flores de três variedades da *heliconia* e, na Figura 2.10, apresenta-se o diagrama em caixa ou boxplot para esses dados. Pode-se ver que os comprimentos das três variedades são bem diferentes, com a *H. bihai* apresentando os maiores comprimentos. A variedade *H. caribaea* amarela apresenta os menores comprimentos, enquanto a dispersão dos comprimentos da *H. caribaea* vermelha é a maior de todas.

Tabela 2.2 – Comprimento das flores de três variedades da *Heliconia*

H.bihai							
47,12	46,75	46,81	47,12	46,67	47,43	46,44	46,64
48,07	48,34	48,15	50,26	50,12	46,34	46,94	48,36
H.caribaea vermelha							
41,90	42,01	41,93	43,09	41,47	41,69	39,78	40,57
39,63	42,18	40,66	37,87	39,16	37,40	38,20	38,07
38,10	37,97	38,79	38,23	38,87	37,78	38,01	
H.caribaea amarela							
36,78	37,02	36,52	36,11	36,03	35,45	38,13	37,10
35,17	36,82	36,66	35,68	36,03	34,57	34,63	

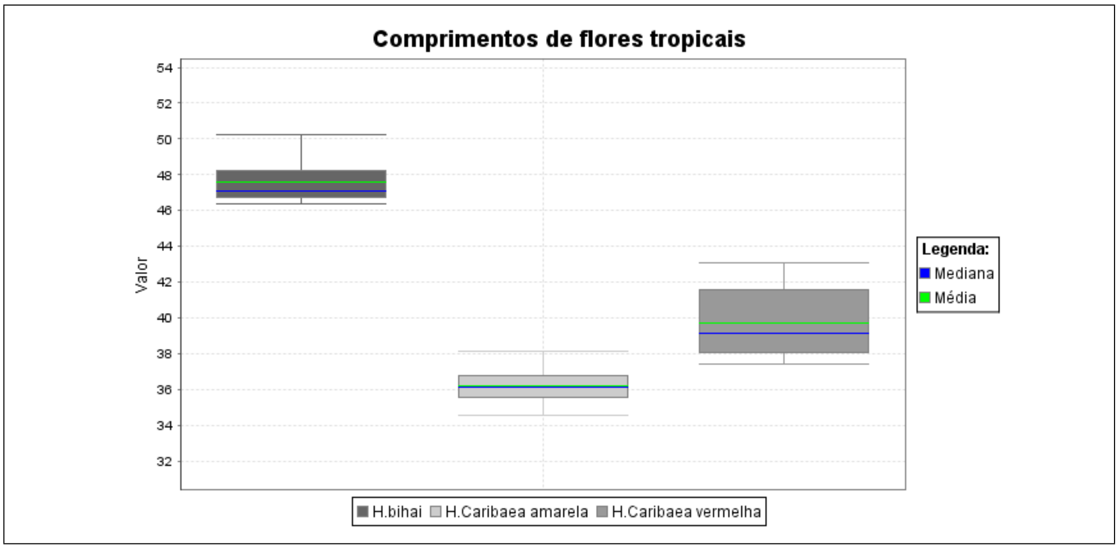


Figura 2.10 – Comprimentos de flores tropicais

2.5 Medidas de posição para distribuições de frequências agrupadas

Considere a distribuição de frequências do salário dos 500 funcionários reproduzida na **Tabela 2.3**. Essa tabela foi construída a partir dos dados individuais dos funcionários da nossa empresa fictícia. Essas informações estão disponíveis para a empresa, mas, em geral, não são divulgadas nesse nível de detalhamento. Imagine, então, que não dispomos dos dados individuais (também chamados *dados brutos*) e temos acesso, somente, às informações da Tabela 2.3. Como poderíamos calcular a média, a moda e a mediana? Isso é o que você aprenderá nesta seção.

Tabela 2.3 – Distribuição de frequência dos salários de 500 funcionários

Salário (reais)	Frequência Simples		Frequência Acumulada	
	Absoluta	Relativa %	Absoluta	Relativa %
2800 ⊢ 4800	87	17,4	87	17,4
4800 ⊢ 6800	203	40,6	290	58,0
6800 ⊢ 8800	170	34,0	460	92,0
8800 ⊢ 10800	30	6,0	490	98,0
10800 ⊢ 12800	10	2,0	500	100,0

2.5.1 Média aritmética simples

Quando agrupamos os dados em uma distribuição de frequências, estamos perdendo informação, uma vez que não apresentamos os valores individuais. Informar apenas que há 87 valores na classe 2800 ⊢ 4800 nos obriga a escolher um valor típico, representante de tal classe. Esse valor será sempre o *ponto médio* da classe.

DEFINIÇÃO Ponto médio

Numa distribuição de frequências agrupadas, o ponto médio de cada classe é escolhido como o valor representativo de todas as observações agrupadas na classe.
O ponto médio é o ponto do meio do intervalo de classe. Se a classe tiver limites inferior e superior representados por l e L respectivamente, então o ponto médio x será calculado como

$$x = \frac{l + L}{2} \tag{2.24}$$

Com essa convenção, o fato de haver 87 observações na primeira classe é interpretado como a existência de 87 valores iguais a 3800, que é o ponto médio dessa classe. Esta é a interpretação básica da tabela de frequências: *todos os valores de uma classe são considerados iguais ao ponto médio da classe*. Na **Tabela 2.4**, acrescentamos uma coluna para informar o ponto médio de cada classe.

Tabela 2.4 – Distribuição de frequência dos salários de 500 funcionários

Salário (reais)	Ponto médio	Frequência Simples		Frequência Acumulada	
		Absoluta	Relativa %	Absoluta	Relativa %
2800 ⊢ 4800	3800	87	17,4	87	17,4
4800 ⊢ 6800	5800	203	40,6	290	58,0
6800 ⊢ 8800	7800	170	34,0	460	92,0
8800 ⊢ 10800	9800	30	6,0	490	98,0
10800 ⊢ 12800	11800	10	2,0	500	100,0

A interpretação da tabela de frequências nos diz que há 87 observações iguais a 3800, 203 observações iguais a 5800, e assim por diante. Então, esses dados podem ser vistos como o seguinte conjunto de observações:

$$\left.\begin{array}{r} 3800 \\ \\ \vdots \\ 3800 \end{array}\right\} \text{87 ocorrências do 3800}$$

$$\left.\begin{array}{r} 5800 \\ \\ \vdots \\ 5800 \end{array}\right\} \text{203 ocorrências do 5800}$$

$$\left.\begin{array}{r} 7800 \\ \\ \vdots \\ 7800 \end{array}\right\} \text{170 ocorrências do 7800}$$

$$\left.\begin{array}{r} 9800 \\ \\ \vdots \\ 9800 \end{array}\right\} \text{30 ocorrências do 9800}$$

$$\left.\begin{array}{r} 11800 \\ \\ \vdots \\ 11800 \end{array}\right\} \text{10 ocorrências do 11800}$$

Para calcular a média desse novo conjunto de dados, temos de fazer:

$$\begin{aligned}
\overline{x} &= \frac{87 \times 3800 + 203 \times 5800 + 170 \times 7800 + 30 \times 9800 + 10 \times 11800}{500} \\
&= \frac{87}{500} \times 3800 + \frac{203}{500} \times 5800 + \frac{170}{500} \times 7800 + \frac{30}{500} \times 9800 + \frac{10}{500} \times 11800 \\
&= 0,174 \times 3800 + 0,406 \times 5800 + 0,340 \times 7800 + 0,06 \times 9800 + 0,02 \times 11800 \\
&= 6492
\end{aligned}$$

Note, na penúltima linha da equação anterior, que os pontos médios de cada classe são multiplicados pela frequência relativa da mesma. Dessa forma, a média dos dados agrupados é uma média ponderada dos pontos médios, onde os pesos são definidos pelas frequências das classes.

Representando o ponto médio da classe por x_i e a frequência relativa (não multiplicada por 100) por f_i, temos que

$$\overline{x} = \sum_{i=1}^{k} f_i x_i \tag{2.25}$$

Os pesos (frequências) aparecem exatamente para compensar o fato de as classes possuírem números diferentes de observações.

2.5.2 Moda

Embora haja métodos geométricos para se calcular a moda de dados agrupados, tais métodos não são muito utilizados na prática. Sendo assim, estimaremos a moda de uma distribuição de frequências agrupadas pelo ponto médio da *classe modal*, que é a classe de maior frequência.

No exemplo anterior, $4800 \vdash 6800$ é a classe modal, de modo que a moda é estimada como $x^* = 5800$.

2.5.3 Quartis

Estando os dados agrupados em classes, há um método geométrico que produz uma estimativa dos quartis. As ideias subjacentes a esse método são a própria definição dos quartis e o fato de que, no histograma da distribuição, as áreas dos retângulos são iguais (proporcionais) às frequências relativas.

Considere o histograma da Figura 2.11, referente aos salários dos 500 funcionários da Tabela 2.3. Na primeira classe, temos $17,4\%$ das observações e, nas duas primeiras classes, temos $58,0\%$. Logo, a mediana é algum ponto da *classe mediana* $4800 \vdash 6800$ e, abaixo desse ponto, devemos ter 50% da distribuição, ou seja, a soma da área do primeiro retângulo com a área do retângulo sombreado representa 50% da frequência total.

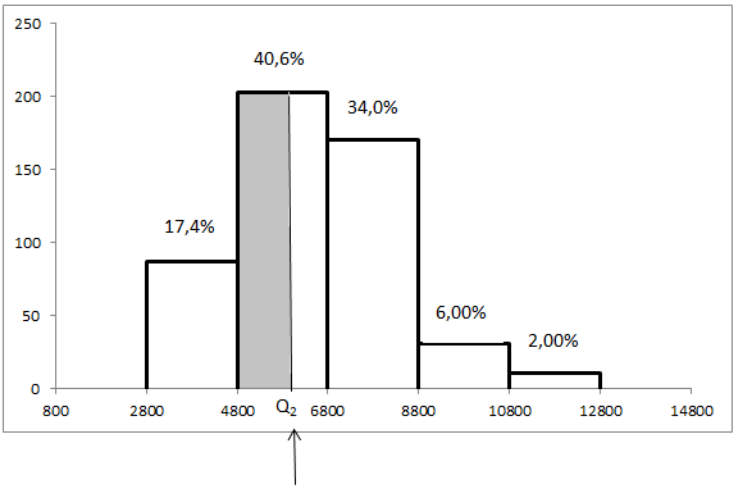


Figura 2.11 – Cálculo da mediana da distribuição dos salários

Então, para identificar a mediana, devemos notar que, na classe mediana, faltam $32,6\% = 50\% - 17,4\%$ da distribuição para completar 50%. Então, a área A_1 do retângulo sombreado deve ser igual a $32,6\%$, enquanto o retângulo da classe mediana tem área $A_m = 40,6\%$. Note que o retângulo sombreado e o retângulo da classe mediana têm a mesma altura. Usando a fórmula da área de um retângulo, obtém-se:

$$\begin{aligned} A_1 &= 32,6 = (Q_2 - 4800) \times h \\ A_m &= 40,6 = (6800 - 4800) \times h \end{aligned}$$

em que h é a altura comum dos dois retângulos. Dividindo as duas igualdades, termo a termo, obtém-se a seguinte regra de proporcionalidade:

$$\frac{32,6}{40,6} = \frac{Q_2 - 4800}{6800 - 4800} \Rightarrow Q_2 = 4800 + 2000 \times \frac{32,6}{40,6} \Rightarrow Q_2 = 6405,91$$

Seguindo o mesmo raciocínio, vemos que o primeiro quartil também está na segunda classe $4800 \vdash 6800$. Como na primeira classe a frequência é 17,4%, faltam $7,6\% = 25\% - 17,4\%$ para completar os 25%. A regra de três que fornece o primeiro quartil é

$$\frac{7,6}{40,6} = \frac{Q_1 - 4800}{6800 - 4800} \Rightarrow Q_1 = 4800 + 2000 \times \frac{7,6}{40,6} \Rightarrow Q_1 = 5174,38$$

O terceiro quartil está na terceira classe $6800 \vdash 8800$. Como nas duas primeiras classes a frequência acumulada é de $17,4\% + 40,6\% = 58\%$, faltam $17\% = 75\% - 58\%$ para completar os 75%. A regra de três que fornece o terceiro quartil é

$$\frac{17}{34} = \frac{Q_3 - 6800}{8800 - 6800} \Rightarrow Q_3 = 6800 + 2000 \times \frac{17}{34} \Rightarrow Q_3 = 7800$$

EXEMPLO 2.12 *Medidas de posição de dados agrupados*

Vamos calcular a média, a moda e a mediana da seguinte distribuição:

Classes	Frequência Simples		Frequência Acumulada	
	Absoluta	Relativa %	Absoluta	Relativa %
0 ⊢ 5	5	6,25	5	6,25
5 ⊢ 10	21	26,25	20	32,50
10 ⊢ 15	28	35,00	42	67,50
15 ⊢ 20	18	22,50	60	90,00
20 ⊢ 25	8	10,00	80	100,00

Os pontos médios das classes são

$$\frac{0+5}{2} = 2,5 \qquad \frac{5+10}{2} = 7,5 \qquad \cdots \qquad \frac{20+25}{2} = 22,5$$

e a média é calculada como

$$\begin{aligned} \overline{x} &= 0,0625 \times 2,5 + 0,2625 \times 7,5 + 0,3500 \times 12,5 + 0,2250 \times 17,5 + \\ &\quad +0,10 \times 22,5 = 12,6875 \end{aligned}$$

Note que é preferível trabalhar com as frequências relativas em forma decimal, pois, se trabalhássemos com as frequências relativas em forma percentual, teríamos de dividir o resultado por 100. *Lembre-se de que a média tem de estar entre o valor mínimo* 0 *e o valor máximo* 25.

A classe modal é 10 ⊢ 15 e, portanto, estimamos a moda como $x^* = 12,5$.

Da coluna de frequências relativas acumuladas, vemos que a mediana está na terceira classe, ou seja, 10 ⊢ 15 é a classe mediana. Nas duas primeiras classes, temos 32,50% dos dados, e faltam 17, 50% para completar 50% (veja a 2.12).

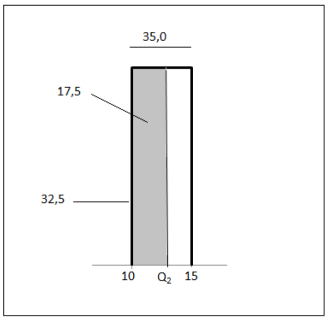


Figura 2.12 – Cálculo da mediana da distribuição do Exemplo 2.12

A regra de três resultante é

$$\frac{Q_2 - 10}{17,5} = \frac{15 - 10}{35,0} \Rightarrow Q_2 = 12,5$$

O primeiro quartil está na segunda classe $5 \vdash 10$. Como, na primeira classe, temos $6,25\%$, faltam $25\% - 6,25\% = 18,75\%$ para completar 25%. A regra de três que define o primeiro quartil é

$$\frac{Q_1 - 5}{10 - 5} = \frac{18,75}{26,25} \Rightarrow Q_1 = 5 + 5 \times \frac{18,75}{26,25} = 8,57$$

O terceiro quartil está na quarta classe $15 \vdash 20$. Como, nas três primeiras classes, temos $67,50\%$, faltam $75\% - 67,5\% = 7,5\%$ para completar 75%. A regra de três que define o terceiro quartil é

$$\frac{Q_3 - 15}{20 - 15} = \frac{7,5}{22,5} \Rightarrow Q_3 = 15 + 5 \times \frac{7,5}{22,5} = 16,67$$

♦♦

Capítulo 3

Análise bidimensional

Até o momento, vimos como organizar e resumir informações referentes a uma única variável. No entanto, é bastante frequente depararmos com situações onde há interesse em estudar conjuntamente duas ou mais variáveis. Para os dados da Tabela 3.1, por exemplo, podemos estudar se há alguma relação entre sexo e a matéria predileta no Ensino Médio. Num estudo sobre mortalidade infantil, é importante acompanhar também o tratamento pré-natal da mãe; espera-se, neste caso, que haja uma diminuição da taxa de mortalidade infantil com o aumento dos cuidados durante a gravidez.

3.1 Variáveis qualitativas

3.1.1 Representação tabular: Distribuição bivariada de frequências

Nesta seção iremos considerar o caso de duas variáveis *qualitativas*. Como exemplo, vamos trabalhar com os dados apresentados na Tabela 3.1, onde temos a matéria predileta no Ensino Médio (C = Ciências; G = Geografia; H = História; T = Matemática; P = Português) e o sexo (M = Masculino; F = Feminino) de 40 alunos. Podemos ver que, mesmo para um pequeno número de observações (40), essa tabela é de difícil leitura. Além disso, para cada aluno, independente de seu nome, a informação que realmente queremos é sobre o sexo e a matéria predileta. Note que cada aluno dá origem a um par de valores (x_i, y_i); na Tabela 3.1, por exemplo, para o aluno Daniel temos o par (M, H).

Uma forma de representar conjuntamente as informações referentes a essas duas variáveis é através de uma *distribuição ou tabela conjunta de frequências*, muitas vezes chamada de *tabela de contingência*. Como temos duas variáveis de interesse, precisamos de duas dimensões, linha e coluna, para representar as informações disponíveis. Assim, o primeiro passo é escolher qual variável será representada em cada uma dessas duas dimensões.

A escolha da variável linha e da variável coluna depende do objetivo do estudo. Em geral, coloca-se na coluna a variável que define os grupos que desejamos comparar. No exemplo, poderíamos estar interessados em comparar homens e mulheres com relação à matéria predileta e nesse caso, sexo seria a variável coluna. Se nosso interesse fosse comparar a preferência pelas diferentes matérias, matéria preferida seria a variável coluna e sexo seria a variável linha.

Tabela 3.1 – Sexo e matéria predileta no Ensino Médio

Aluno	Sexo	Matéria	Aluno	Sexo	Matéria	Aluno	Sexo	Matéria
Alice	F	T	Jeferson	M	T	Marina	F	P
Ana Luiza	F	T	Jessica	F	P	Mateus	M	H
André	M	G	Julia	F	C	Miguel	M	P
Andrea	F	T	Juliana	F	P	Paula	F	C
Beatriz	F	H	Letícia	F	G	Paulo	M	C
Camila	F	T	Luana	F	T	Pedro	M	H
Carolina	F	G	Lucas	M	H	Rafael	M	G
Cristina	F	T	Luiz	M	G	Renato	M	C
Daniel	M	H	Luiza	F	P	Ricardo	M	C
Daniela	F	P	Luna	F	C	Tatiana	F	T
Fernando	M	P	Marcela	F	T	Thais	F	P
Gabriel	M	C	Maria Luiza	F	P	Tiago	M	P
Gabriela	F	G	Marília	F	C	Vitor	M	C

Tomada a decisão sobre as variáveis linha e coluna, podemos começar a construção da tabela. Vamos construir inicialmente a tabela para o caso em que sexo é a variável coluna. Na Tabela 3.2 ilustra-se o primeiro passo do procedimento, que consiste em rotular as linhas e colunas da tabela.

Tabela 3.2 – Distribuição conjunta de sexo e matéria predileta no Ensino Médio - Passo 1

Matéria predileta	Sexo	
no segundo grau	Masculino (M)	Feminino (F)
Ciências (C)		
Geografia (G)		
História (H)		
Matemática (T)		
Português (P)		

Note que cada célula na tabela corresponde a um par e temos ao todo 10 pares: (C,M), (C,f), (G,M), (G,F), (H,M), (H,F), (T,M), (T,F), (P,M), (P,F). A forma de se preencher a tabela é registrando em cada célula a frequência observada do par correspondente. Da Tabela 3.1, podemos ver que há 5 ocorrências do par (C,M), ou seja, 5 alunos do sexo Masculino preferiam Ciências no Ensino Médio. Assim, na primeira célula registramos a frequência 5. Continuando com esse raciocínio, obtemos a Tabela 3.3:

Tabela 3.3 – Distribuição conjunta de sexo e matéria predileta no Ensino Médio - Passo 2

Matéria predileta	Sexo	
no segundo grau	Masculino	Feminino
Ciências	5	4
Geografia	3	3
História	4	2
Matemática	1	8
Português	3	7

O último passo consiste em registrar os totais de linha (número de alunos que preferiam cada uma das cinco matérias) e os totais de coluna (número de alunos por sexo). Esses totais são obtidos somando-se os elementos de cada linha e cada coluna. Por exemplo, para a primeira linha, $5 + 4 = 9$ alunos preferiam Ciências; para a primeira coluna, $5 + 3 + 4 + 1 + 3 = 16$ alunos eram do sexo Masculino. Na Tabela 3.4 apresentamos a distribuição conjunta final das duas variáveis. Como já visto no caso univariado, essa forma de apresentação é mais interessante, uma vez que não estamos interessados na observação individual e, sim, no comportamento dos grupos.

Tabela 3.4 – Distribuição conjunta de sexo e matéria predileta no Ensino Médio

Matéria predileta no Ensino Médio	Sexo		Total
	Masculino	Feminino	
Ciências	5	4	9
Geografia	3	3	6
História	4	2	6
Matemática	1	8	9
Português	3	7	10
Total	16	24	40

A título de ilustração, apresenta-se na Tabela 3.5 a tabela com Matéria Predileta como a variável coluna.

Tabela 3.5 – Distribuição conjunta de sexo e matéria predileta no Ensino Médio

Sexo	Matéria predileta no Ensino Médio					Total
	Ciências	Geografia	História	Matemática	Português	
Masculino	5	3	4	1	3	16
Feminino	4	3	2	8	7	24
Total	9	6	6	9	10	40

3.1.2 Frequências relativas

Na construção de distribuições de frequências univariadas, foi acrescentada à tabela a coluna de frequências relativas, que davam a proporção de elementos em cada classe com relação ao número total de elementos. Um procedimento análogo pode ser feito para as tabelas bidimensionais; a diferença é que, neste caso, existem três possibilidades para expressarmos as proporções de cada cela: (i) com relação ao total geral; (ii) com relação ao total de cada linha e (iii) com relação ao total de cada coluna. A escolha entre essas três possibilidades deverá ser feita de acordo com o objetivo da análise.

Continuando com nosso exemplo, vamos construir as distribuições de frequências relativas com a variável Sexo na coluna.

Em relação ao total geral

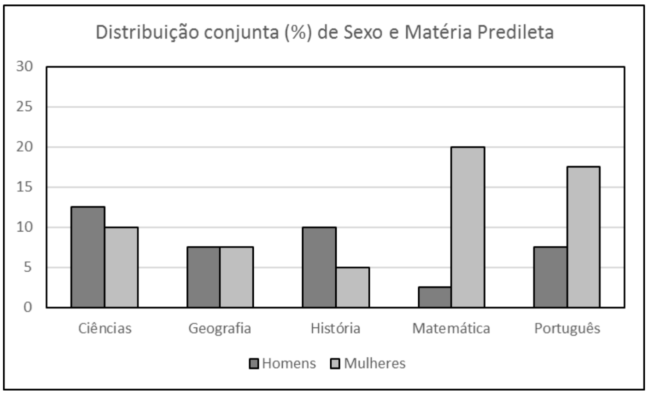
Para cada célula, calcula-se a frequência relativa ao total geral (40). Para a primeira célula, isso nos dá $5/40 = 0,125$ ou 12,5%, o que significa que 12,5% de todos os alunos são do sexo Masculino e preferiam Ciências no Ensino Médio. Completando os cálculos, obtemos a Tabela 3.6

Na Figura 3.1 apresentam-se duas formas de representar essa distribuição graficamente: gráfico de colunas ou gráfico de colunas empilhadas.

A distribuição de frequências relativas ao total geral não são muito usadas na prática, uma vez que, na maioria das análises bidimensionais, o objetivo está em comparar grupos.

Tabela 3.6 – Distribuição conjunta (%) de sexo e matéria predileta no Ensino Médio

Matéria predileta no Ensino Médio	Sexo		Total
	Masculino	Feminino	
Ciências	12,5	10,0	22,5
Geografia	7,5	7,5	15,0
História	10,0	5,0	15,0
Matemática	2,5	20,0	22,5
Português	7,5	17,5	25,0
Total	40,0	60,0	100,0



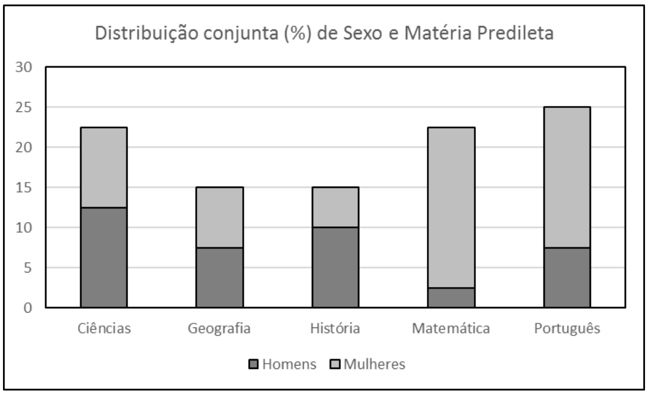


Figura 3.1 – Distribuição conjunta (%) de sexo e matéria predileta no Ensino Médio

Assim, é mais comum vermos as distribuições de frequências relativas aos totais de coluna ou linha.

Em relação ao total das colunas

O objetivo agora é comparar os "grupos" definidos pelas categorias da variável coluna. Sendo assim, temos que uniformizar os totais de todas as colunas, ou seja, para cada coluna a frequência relativa total será 100%, o que permite a comparação. Para construir essa distribuição, temos que trabalhar em cada coluna separadamente, calculando para cada célula a frequência relativa ao total da respectiva coluna. Na Tabela 3.7 mostram-se os cálculos e na Tabela 3.8 temos a distribuição completa. Essa distribuição é chamada *distribuição condicional da Matéria Predileta dado o Sexo do aluno*

Tabela 3.7 – Construção da distribuição condicional (%) de Matéria Predileta no Ensino Médio dado o Sexo do aluno

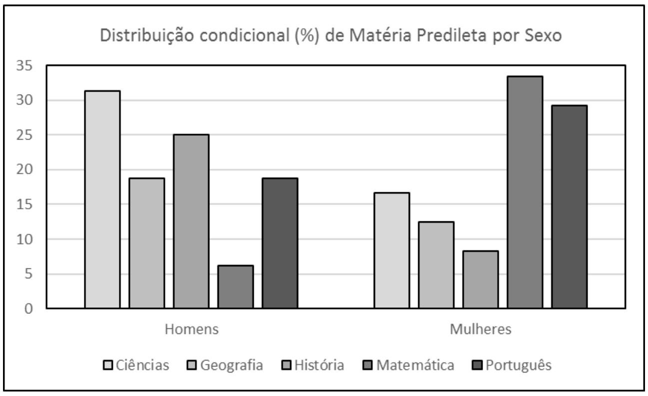
Matéria predileta no Ensino Médio	Sexo		Total
	Masculino	Feminino	
Ciências	100×5/16	100×4/24	100×9/40
Geografia	100×3/16	100×3/24	100×6/40
História	100×4/16	100×2/24	100×6/40
Matemática	100×1/16	100×8/24	100×9/40
Português	100×3/16	100×7/24	100×10/40
Total	100×16/16	100×24/24	100×40/40

Tabela 3.8 – Distribuição condicional (%) de Matéria Predileta no Ensino Médio dado o Sexo do aluno

Matéria predileta no Ensino Médio	Sexo		Total
	Masculino	Feminino	
Ciências	31,25	16,67	22,50
Geografia	18,75	12,50	15,00
História	25,00	8,33	15,00
Matemática	6,25	33,33	22,50
Português	18,75	29,17	25,00
Total	100,00	100,00	100,00

Da Tabela 3.8 podemos concluir que, entre os homens, 31,25% preferiam Ciências, enquanto essa porcentagem cai para 16,67% entre as mulheres. Vemos, também, que, embora os números absolutos de homens e mulheres que preferiam Geografia sejam iguais, os percentuais por sexo são diferentes: 3 em 16 (0,1875) é maior que 3 em 24 (0,125).

Na Figura 3.2 apresentam-se os gráficos apropriados para ilustrar essa distribuição.



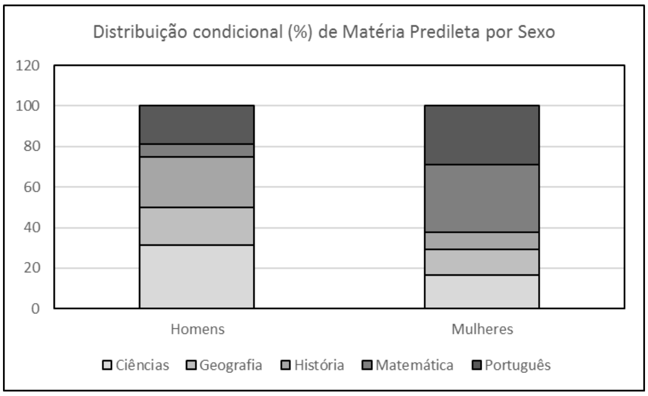


Figura 3.2 – Distribuição condicional (%) de Matéria Predileta por Sexo

Em relação ao total das linhas

A seguir são dadas as representações tabular e gráfica da distribuição condicional de Sexo por Matéria Predileta.

Tabela 3.9 – Distribuição condicional de Sexo dada a Matéria Predileta no Ensino Médio

Matéria predileta no Ensino Médio	Sexo		Total
	Masculino	Feminino	
Ciências	55,56	44,44	100,00
Geografia	50,00	50,00	100,00
História	66,67	33,33	100,00
Matemática	11,11	88,89	100,00
Português	30,00	70,00	100,00
Total	40,00	60,00	100,00

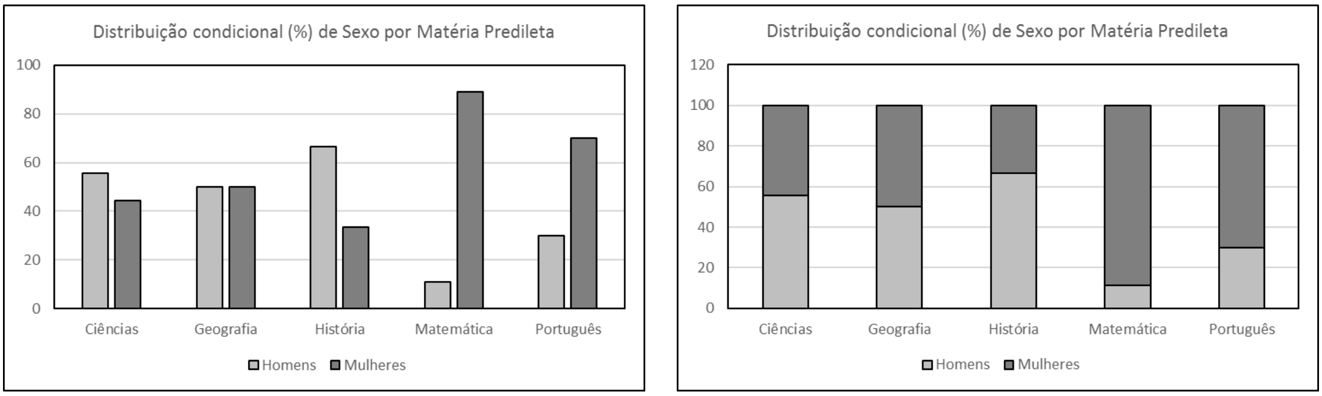


Figura 3.3 – Distribuição condicional (%) de Sexo por Matéria Predileta

3.2 Variáveis quantitativas

No caso de variáveis quantitativas discretas com poucos valores, a construção de tabelas bivariadas pode ser feita de maneira análoga às variáveis qualitativas. Para variáveis quantitativas contínuas ou discretas com muitos valores, a construção é possível, mas não muito usual, uma vez que há muita perda de informação pois, assim como no caso univariado, é preciso agrupar os dados em classes.

3.2.1 Diagramas de dispersão

Quando as variáveis envolvidas em uma análise bidimensional são do tipo *quantitativo* (salário, idade, altura etc.), um instrumento de análise bastante útil é o *diagrama de dispersão*.

DEFINIÇÃO **Diagrama de dispersão**

O diagrama de dispersão é um gráfico bidimensional, em que os valores das variáveis envolvidas são representados como pares ordenados no plano cartesiano. Essas variáveis são variáveis *quantitativas*, medidas sobre os mesmos indivíduos.

Para ilustrar a construção de um diagrama de dispersão, vamos considerar uma amostra de 10 alunas do curso de Hotelaria da UFF (dados fictícios) para as quais foram medidos seu peso (em kg) e sua altura (em cm). Na Tabela 3.10, apresentam-se os dados obtidos.

O primeiro passo consiste em desenhar os eixos cartesianos e definir as escalas de forma apropriada. Não é necessário começar da origem, ou seja, pode-se fazer uma quebra de escala. Na Figura 3.4 ilustra-se o sistema de eixos cartesianos, com a variável Peso representada no eixo vertical e a variável altura no eixo horizontal. O motivo para essa escolha é que nosso interesse está em estudar o efeito da altura sobre o peso das meninas. Note que a escala no eixo horizontal começa em 150 (a menor altura é 155) e termina em 180 (a maior altura é 177). Analogamente, a escala no eixo vertical começa em 30 e termina em 90.

O próximo passo consiste em marcar os pontos nesse sistema de eixos. Vamos ilustrar o procedimento para a primeira aluna Ana, que tem altura 155 e peso 50. Na escala horizontal

Tabela 3.10 – Peso e altura de uma amostra de 10 alunas

Aluno	Altura	Peso
Ana	155	50
Ludmilla	158	61
Cristina	162	65
Tereza	168	68
Patrícia	170	69
Mariana	170	65
Ana Paula	172	82
Dirce	173	79
Fabiana	173	75
Tatiane	177	80

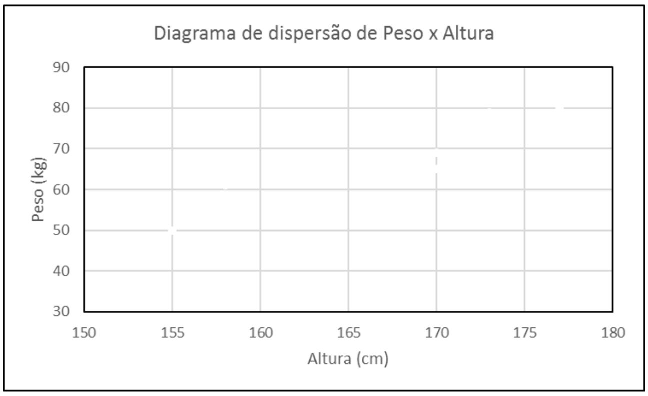


Figura 3.4 – Diagrama de dispersão para peso e altura - os eixos

procuramos o valor 155 e a partir desse valor traçamos uma linha auxiliar paralela ao eixo vertical; na escala vertical procuramos o valor 50 e a partir dele traçamos uma linha auxiliar paralela ao eixo horizontal. O ponto de interseção dessas duas linhas auxiliares representa o ponto (155,50) referente à aluna Ana. Veja a Figura 3.5.

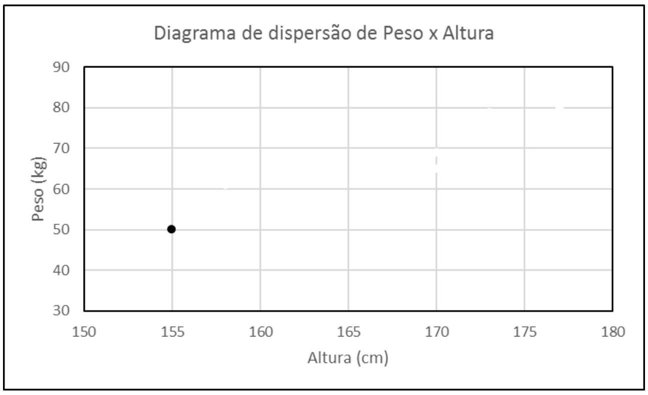


Figura 3.5 – Diagrama de dispersão para peso e altura - marcação do ponto para Ana

Repetindo o processo para cada uma das alunas obtém-se o diagrama de dispersão dado na Figura 3.6.

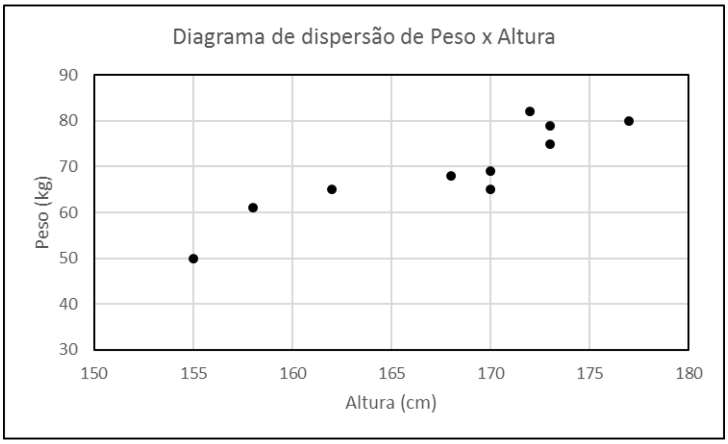
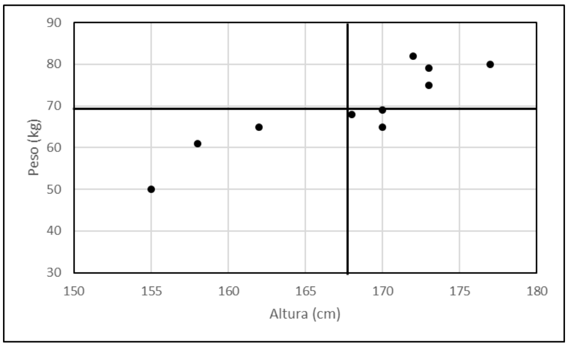


Figura 3.6 – Diagrama de dispersão para peso e altura de 10 alunas

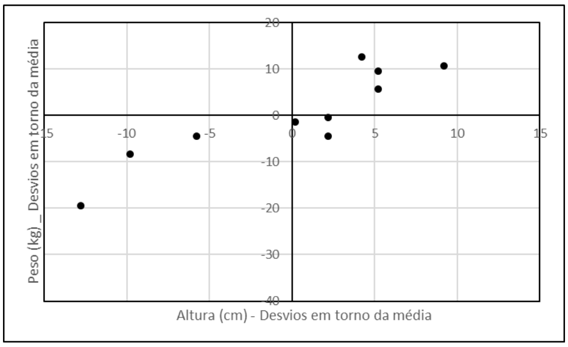
3.2.2 Covariância

Analisando o diagrama de dispersão de peso e altura, podemos ver que há uma tendência de crescimento do peso à medida que a altura aumenta e essa tendência parece linear, ou seja, seria razoável traçar uma reta através dos pontos. Com essa reta, poderíamos "estimar" o peso de alguma aluna conhecendo sua altura. Mas análises visuais são sempre subjetivas. Assim, vamos estudar, agora, uma medida de associação entre variáveis quantitativas que medirá o grau de associação linear. Então, tal medida irá representar o quanto a "nuvem" de pontos em um diagrama de dispersão se aproxima de uma reta.

O primeiro passo para isso é uniformizar o centro da nuvem, de modo que todas as nuvens de pontos (diagramas de dispersão) que venhamos a analisar estejam centradas na origem. Veja as Figuras 3.7a e 3.7b. Na primeira, temos os dados originais e o centro da nuvem está no ponto $(\overline{x}, \overline{y})$. Na segunda, temos os dados deslocados, de modo que o centro da nuvem está no ponto $(0, 0)$. A forma da nuvem é a mesma; apenas "arrastamos" a nuvem para a origem. Note, na Figura 3.7b, que a maioria dos pontos está no primeiro e terceiro quadrantes!



(a) Dados originais



(b) Desvios em torno das médias

Figura 3.7 – Diagrama de dispersão de peso e altura

Para obter essa nova nuvem, temos que subtrair, de cada altura, a média das alturas e de cada peso, a média dos pesos. Na Tabela 3.11 apresentamos os cálculos. Note que a média dos desvios é igual a 0, ou seja, a nuvem está centrada na origem!

Tabela 3.11 – Peso e altura de uma amostra de 10 alunas

Aluno	Dados originais		Desvios em torno da média	
	Altura (X_i)	Peso (Y_i)	Altura ($X_i - \overline{X}$)	Peso ($Y_i - \overline{Y}$)
Ana	155	50	$155 - 167,8 = -12,8$	$50 - 69,4 = -19,4$
Ludmilla	158	61	$158 - 167,8 = -9,8$	$61 - 69,4 = -8,4$
Cristina	162	65	$162 - 167,8 = -5,8$	$65 - 69,4 = -4,4$
Tereza	168	68	$168 - 167,8 = 0,2$	$68 - 69,4 = -1,4$
Patrícia	170	69	$170 - 167,8 = 2,2$	$69 - 69,4 = -0,4$
Mariana	170	65	$170 - 167,8 = 2,2$	$65 - 69,4 = -4,4$
Ana Paula	172	82	$172 - 167,8 = 4,2$	$82 - 69,4 = 12,6$
Dirce	173	79	$173 - 167,8 = 5,2$	$79 - 69,4 = 9,6$
Fabiana	173	75	$173 - 167,8 = 5,2$	$75 - 69,4 = 5,6$
Tatiane	177	80	$177 - 167,8 = 9,2$	$80 - 69,4 = 10,6$
MÉDIA	167,8	69,4	0,0	0,0

Consideremos, agora, os dados da Tabela 3.12, em que temos as temperaturas médias anuais e a latitude de uma amostra de 15 cidades dos Estados Unidos (Dunn e Clark (1974)). Na Figura 3.8 temos o diagrama de dispersão desses dados e podemos ver que há uma relação decrescente entre as variáveis: aumentando a latitude a temperatura decresce. Como antes, é razoável pensar em traçar uma reta por esses dados, ou seja, há uma tendência linear decrescente entre as variáveis.

Latitude	Temperatura (^{o}F)
34	56,4
32	51,0
39	36,7
39	37,8
41	36,7
45	18,2
41	30,1
33	55,9
34	46,6
47	13,3
44	34,0
39	36,3
41	34,0
32	49,1
40	34,5

Tabela 3.12 – Latitude e temperatura

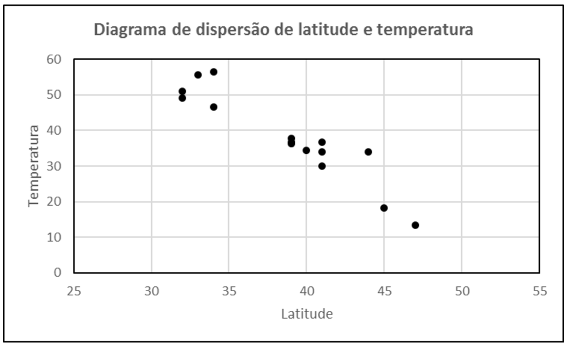
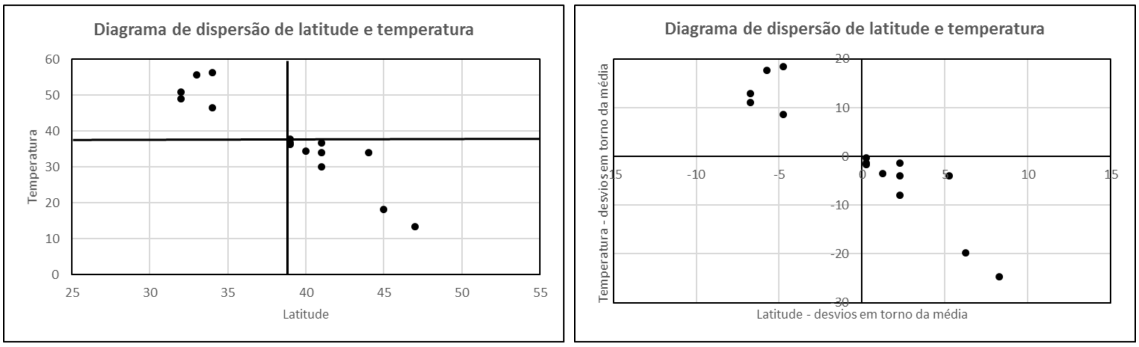


Figura 3.8 – Latitude e temperatura

Como no exemplo anterior, vamos centralizar a nuvem na origem, plotando os desvios em torno da média. Veja as Figuras 3.9a e 3.9b. No diagrama centrado na origem, a maioria dos pontos está no segundo e no quarto quadrantes!

Analisando esses dois exemplos, podemos observar que, para o primeiro conjunto de dados, em que a tendência entre as variáveis é crescente, a maioria dos desvios está no primeiro e terceiro quadrantes, enquanto no segundo exemplo, em que a relação é decrescente, a maioria dos desvios está no segundo e quarto quadrantes.



(a) Dados originais　　**(b)** Desvios em torno das médias

Figura 3.9 – Diagrama de dispersão de latitude e temperatura

O primeiro e terceiro quadrantes se caracterizam pelo fato de as abscissas e ordenadas terem o mesmo sinal e, portanto, seu produto é positivo; já no segundo e quarto quadrantes, as abscissas e ordenadas têm sinais opostos e, portanto, seu produto é negativo. Então, para diferenciar esses gráficos, podemos usar uma medida baseada no produto dos desvios, isto é, $(X_i - \overline{X})(Y_i - \overline{Y})$. Como no caso da variância ou desvio médio absoluto, para considerar todos os pares possíveis e descontar o número de observações, vamos tomar o valor médio desses produtos.

DEFINIÇÃO **Covariância**

A *covariância entre as variáveis X e Y* é definida por

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) \tag{3.1}$$

onde X_i e Y_i são os valores observados.

Uma fórmula alternativa mais simples de se trabalhar é a seguinte:

$$\text{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}\,\overline{Y} \tag{3.2}$$

Analisando a fórmula (3.2) podemos ver que a covariância é a "média dos produtos menos o produto das médias". Resulta também que a covariância entre X e X é a variância de X, isto é: $\text{Cov}(X, X) = \text{Var}(X)$.

É bastante importante salientar a interpretação da covariância: ela mede o grau de *associação linear* entre variáveis. Considere os dados apresentados na Tabela 3.13, cujo diagrama de dispersão é dado na Figura 3.10. Este diagrama exibe uma associação quadrática perfeita entre as variáveis; no entanto, a covariância entre elas é nula. Note que $\overline{X} = 0$, assim como $\sum_{i=1}^{n} X_i Y_i = 0$.

X	Y	X	Y
-3	9,00	0,2	0,04
-2,8	7,84	0,4	0,16
-2,6	6,76	0,6	0,36
-2,4	5,76	0,8	0,64
-2,2	4,84	1,0	1,00
-2,0	4,00	1,2	1,44
-1,8	3,24	1,4	1,96
-1,6	2,56	1,6	2,56
-1,4	1,96	1,8	3,24
-1,2	1,44	2,0	4,00
-1,0	1,00	2,2	4,84
-0,8	0,64	2,4	5,76
-0,6	0,36	2,6	6,76
-0,4	0,16	2,8	7,84
-0,2	0,04	3	9,00
0,0	0,00		

Tabela 3.13 – Covariância nula

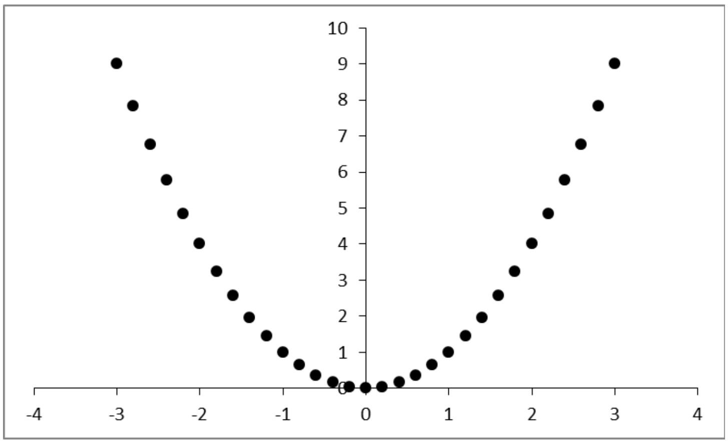


Figura 3.10 – Associação quadrática perfeita, covariância nula

3.2.3 Coeficiente de correlação

Um dos problemas da covariância é a sua dependência da escala dos dados, o que faz com que seus valores possam variar de $-\infty$ a $+\infty$. Observe que sua unidade de medida é dada pelo produto das unidades de medida das variáveis X e Y envolvidas. Isso torna difícil a comparação de situações como as ilustradas nos gráficos das Figuras 3.11 e 3.12. Esses dois diagramas de dispersão representam os dados sobre latitude e temperatura já analisados anteriormente. Na Figura 3.11, as temperaturas estão medidas em graus Fahrenheit e na Figura 3.12, em graus Celsius. Sendo assim, a informação que os dados nos trazem é, basicamente, a mesma. Mas, para o primeiro conjunto, a covariância é $-51,816$ e, para o segundo, $-28,7867$.

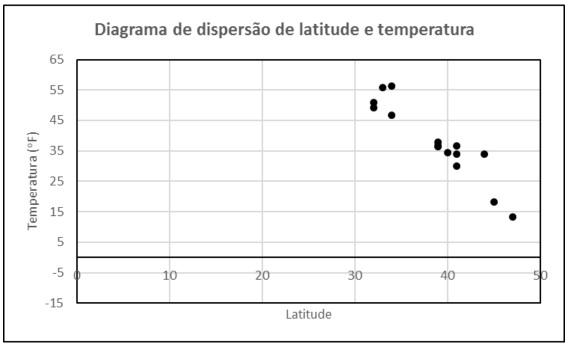


Figura 3.11 – Latitude e temperatura (^{o}F)

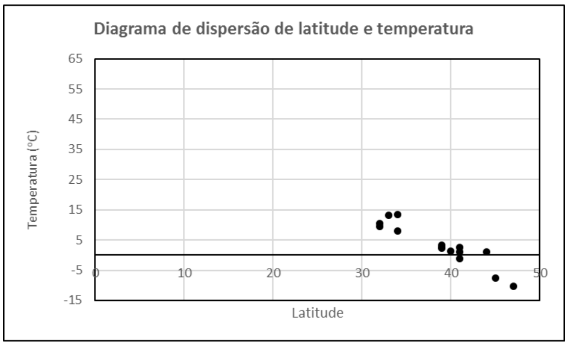


Figura 3.12 – Latitude e temperatura (^{o}C)

Uma maneira de se tirar o efeito da escala é dividir os dados pelo seu desvio padrão, ou seja, trabalhar com os *escores padronizados* $\frac{X_i - \overline{X}}{\sigma_X}$ e $\frac{Y_i - \overline{Y}}{\sigma_Y}$. Nas Figuras 3.13 e 3.14, apresentam-se os diagramas de dispersão para os dados padronizados sobre latitude e temperatura, com a temperatura medida em graus Fahrenheit e Celsius. Note que os diagramas, agora, são idênticos!

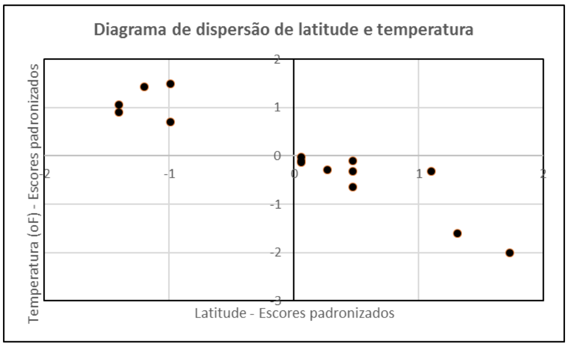


Figura 3.13 – Escores padronizados de latitude e temperatura (^{o}F)

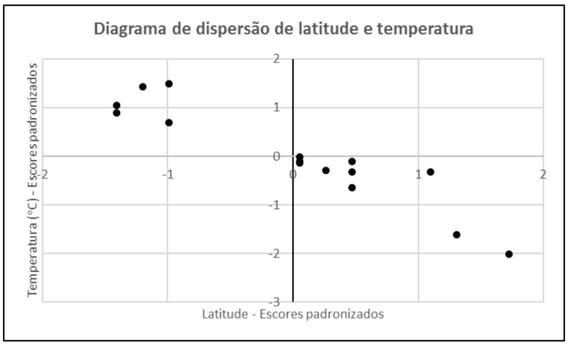


Figura 3.14 – Escores padronizados de latitude e temperatura (^{o}C)

A covariância entre variáveis padronizadas recebe o nome de *coeficiente de correlação*.

DEFINIÇÃO Coeficiente de correlação

O *coeficiente de correlação entre as variáveis X e Y* é definido como

$$\operatorname{Corr}(X,Y) = \rho(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\sigma_x}\right)\left(\frac{y_i - \overline{y}}{\sigma_y}\right) = \frac{\operatorname{Cov}(X,Y)}{\sigma_x\,\sigma_y} \tag{3.3}$$

Os dois conjuntos de dados das Figuras 3.13 e 3.14 têm, ambos, o mesmo coeficiente de correlação, igual a $0,9229$.

Observe que o coeficiente de correlação é *adimensional*. Além disso, ele tem uma propriedade bastante interessante, que é a seguinte:

$$-1 \leq \rho(X,Y) \leq 1 \tag{3.4}$$

Assim, valores do coeficiente de correlação próximos de 1 indicam uma forte associação linear crescente entre as variáveis, enquanto valores próximos de -1 indicam uma forte associação linear decrescente. Já valores próximos de zero indicam fraca associação linear (isso não significa que não exista algum outro tipo de associação; veja o caso da Figura 3.10).

Vamos ver, agora, mais um exemplo para ilustrar todos os passos necessários para o cálculo do coeficiente de correlação entre duas variáveis.

1. O primeiro passo é analisar o diagrama de dispersão para ver se é razoável pensar em uma relação linear entre as variáveis.
2. Calcular a média de cada uma das variáveis. Para isso, precisamos somar os valores de cada uma das variáveis.
3. Calcular a variância de cada uma das variáveis pela fórmula (2.14): para isso precisamos somar os quadrados dos valores de cada uma das variáveis.
4. Calcular a covariância das variáveis usando a fórmula (3.2: para isso precisamos somar os produtos dos valores das variáveis.

5. Calcular a correlação.

A execução desses passos fica facilitada com a construção de uma tabela, conforme ilustraremos agora.

EXEMPLO 3.1 *Barcos registrados e mortes de peixes-bois*

A Tabela 3.14 contém dados sobre o número de barcos registrados na Flórida (em milhares) e o número de peixes-bois mortos por barcos, entre os anos de 1977 e 19962013 (Moore et al. (2017)). Construa o diagrama de dispersão para esses dados e calcule o coeficiente de correlação entre as variáveis, interpretando seu resultado.

Tabela 3.14 – Barcos registrados e mortes de peixes-bois na Flórida

Ano	Barcos	Mortes	Ano	Barcos	Mortes	Ano	Barcos	Mortes	Ano	Barcos	Mortes
1977	447	13	1987	645	39	1997	755	54	2007	1027	73
1978	460	21	1988	675	43	1998	809	66	2008	1010	90
1979	481	24	1989	711	50	1999	830	82	2009	982	97
1980	498	16	1990	719	47	2000	880	78	2010	942	83
1981	513	24	1991	681	53	2001	944	81	2011	922	88
1982	512	20	1992	679	38	2002	962	95	2012	902	81
1983	526	15	1993	678	35	2003	978	73	2013	897	72
1984	559	34	1994	696	49	2004	983	69			
1985	585	33	1995	713	42	2005	1010	79			
1986	614	33	1996	732	60	2006	1024	92			

Solução

Na Figura 3.15, temos o diagrama de dispersão, onde se vê que, à medida que aumenta o número de barcos registrados, há um aumento do número de mortes de peixes-bois na Flórida. A associação entre as variáveis tem um forte padrão linear crescente.

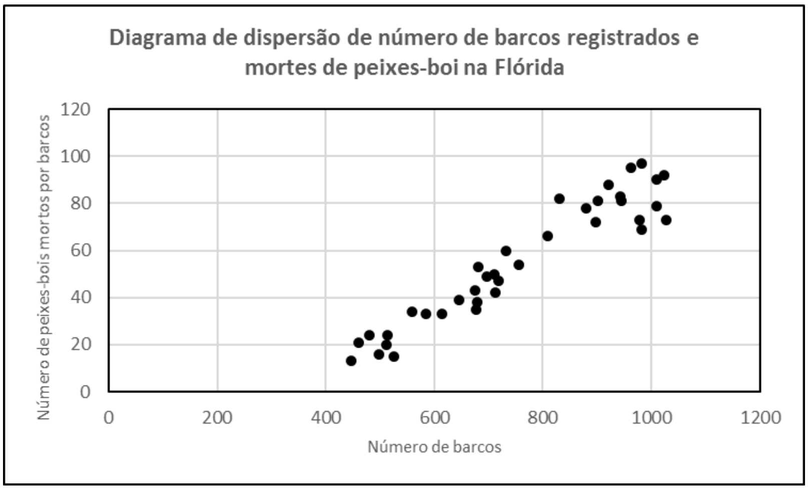


Figura 3.15 – Barcos registrados e mortes de peixes-bois na Flórida

Na tabela a seguir, temos os detalhes dos cálculos a serem feitos, no caso de se estar utilizando uma calculadora mais simples.

	X	Y	X^2	Y^2	XY
	447	13	199809	169	5811
	460	21	211600	441	9660
	481	24	231361	576	11544
	498	16	248004	256	7968
	513	24	263169	576	12312
	512	20	262144	400	10240
	526	15	276676	225	7890
	559	34	312481	1156	19006
	585	33	342225	1089	19305
	614	33	376996	1089	20262
	645	39	416025	1521	25155
	675	43	455625	1849	29025
	711	50	505521	2500	35550
	719	47	516961	2209	33793
	681	53	463761	2809	36093
	679	38	461041	1444	25802
	678	35	459684	1225	23730
	696	49	484416	2401	34104
	713	42	508369	1764	29946
	732	60	535824	3600	43920
	755	54	570025	2916	40770
	809	66	654481	4356	53394
	830	82	688900	6724	68060
	880	78	774400	6084	68640
	944	81	891136	6561	76464
	962	95	925444	9025	91390
	978	73	956484	5329	71394
	983	69	966289	4761	67827
	1010	79	1020100	6241	79790
	1024	92	1048576	8464	94208
	1027	73	1054729	5329	74971
	1010	90	1020100	8100	90900
	982	97	964324	9409	95254
	942	83	887364	6889	78186
	922	88	850084	7744	81136
	902	81	813604	6561	73062
	897	72	804609	5184	64584
Soma	**27981**	**2042**	**22422341**	**136976**	**1711146**

A covariância de X e Y é a "média dos produtos menos o produto das médias", ou seja:

$$\text{Cov}(x, y) = \frac{1711146}{37} - \frac{27981}{37} \times \frac{2042}{37} = 4510,738$$

A variância de cada variável é a "média dos quadrados menos o quadrado da média", ou seja:

$$\text{Var}(X) = \frac{22422341}{37} - \left(\frac{27981}{37}\right)^2 = 34105,37327$$

$$\text{Var}(Y) = \frac{136976}{37} - \left(\frac{2042}{37}\right)^2 = 656,2074$$

O coeficiente de correlação é:

$$\text{Corr}(X, Y) = \frac{4510,738}{\sqrt{34105,3733 \times 656,2074}} = 0,953489$$

Esta alta correlação positiva confirma a forte relação linear crescente entre as variáveis, já vislumbrada no diagrama de dispersão.

♦♦

Bibliografia

Dunn, O. J. e Clark, V. A. (1974) *Applied Statistics: Analysis of Variance and Regression*. Wiley.

Moore, D. S., Notz, W. I. e Fligner, M. A. (2017) *A Estatística Básica e sua Prática*. LTC Editora, 7a. ed.