Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Análise descritiva de dados
Síntese numérica

Edna Afonso Reis
Ilka Afonso Reis

Relatório Técnico
RTP-02/2002
Série Ensino

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Departamento de Estatística

Análise Descritiva de Dados

Síntese Numérica

Edna Afonso Reis
Ilka Afonso Reis

Primeira Edição – Julho/2002

ÍNDICE

1. Introdução

A coleta de dados estatísticos tem crescido muito nos últimos anos em todas as áreas de pesquisa, especialmente com o advento dos computadores e surgimento de *softwares* cada vez mais sofisticados. Ao mesmo tempo, olhar uma extensa listagem de dados coletados não permite obter praticamente nenhuma conclusão, especialmente para grandes conjuntos de dados, com muitas características sendo investigadas.

A Análise Descritiva é a fase inicial deste processo de estudo dos dados coletados. Utilizamos métodos de Estatística Descritiva para organizar, resumir e descrever os aspectos importantes de um conjunto de características observadas ou comparar tais características entre dois ou mais conjuntos de dados.

A descrição dos dados também tem como objetivo identificar anomalias, até mesmo resultante do registro incorreto de valores, e dados dispersos, aqueles que não seguem a tendência geral do restante do conjunto.

Não só nos artigos técnicos direcionados para pesquisadores, mas também nos artigos de jornais e revistas escritos para o público leigo, é cada vez mais freqüente a utilização destes recursos de descrição para complementar a apresentação de um fato, justificar ou referendar um argumento.

As ferramentas descritivas são os muitos tipos de gráficos e tabelas e também medidas de síntese como porcentagens, índices e médias.

As ferramentas gráficas e o uso de tabelas são abordados no Relatório Técnico RTE04-2001 (*Análise Descritiva de Dados - Tabelas e Gráficos*). Neste texto, abordaremos as medidas de síntese numérica, usadas quando a variável em questão é do tipo quantitativa. Serão discutidas as medidas de tendência central, variabilidade e ainda as medidas de posição.

Ao sintetizarmos os dados, perdemos informação, pois não se têm as observações originais. Entretanto, esta perda de informação é pequena se comparada ao ganho que temos com a clareza da interpretação proporcionada.

2. Medidas de Tendência Central

A *tendência central* da distribuição de freqüências de uma variável em um conjunto de dados é caracterizada pelo *valor típico* dessa variável. Essa é uma maneira de resumir a informação contida nos dados, pois escolheremos um valor para representar todos os outros.

Assim, poderíamos perguntar, por exemplo, qual é a altura típica dos brasileiros adultos no final da década de 90 e compará-la com o valor típico da altura dos brasileiros no final da década de 80, a fim de verificar se os brasileiros estão se tornando, em geral, mais altos, mais baixos ou não sofreram nenhuma alteração em sua altura típica. Fazer essa comparação utilizando medidas-resumo (as alturas típicas em cada período) é bem mais sensato do que comparar os dois conjuntos de dados valor a valor, o que seria inviável.

Mas, como identificar o valor típico de um conjunto de dados? Existem três medidas que podem ser utilizadas para descrever a tendência central de um conjunto de dados: a média, a mediana e a moda. Apresentaremos essas três medidas e discutiremos suas vantagens e desvantagens.

2.1. Média Aritmética Simples

A média aritmética simples (que chamaremos apenas de média) é a medida de tendência central mais conhecida e usada para o resumo de dados. Essa popularidade pode ser devida à facilidade de cálculo e à idéia simples que ela nos sugere. De fato, se queremos um valor que represente a altura dos brasileiros adultos, por que não medir as alturas de uma amostra de brasileiros adultos, somar os valores e dividir esse "bolo" igualmente entre os participantes? Essa é a idéia da média aritmética.

Para apresentar a média, primeiramente vamos definir alguma notação. A princípio, essa notação pode parecer desnecessária, mas facilitará bastante nosso trabalho futuro.

Notação	
$\mathbf{n}$	número de indivíduos no conjunto de dados
$\mathbf{x_i}$	valor da i-ésima observação do conjunto de dados, i = 1, 2, 3,..., n
$\sum \mathbf{x_i}$	soma de todas as observações da amostra (a letra grega Σ é o símbolo que indica soma).
$\overline{X}$	é o símbolo usado para representar a média aritmética simples.

Assim,

$$\overline{X} = \frac{\text{Soma de todas as observações do conjunto de dados}}{\text{tamanho do conjunto de dados}} = \frac{\sum \mathbf{x_i}}{\mathbf{n}}$$

Exemplo 2.1: No conjunto de dados (3,0 ; 4,5 ; 5,5 ; 2,5 ; 1,3 ; 6,0), temos n = 6,
x1 = 3,0 x2 = 4,5 x3 = 5,5 x4 = 2,5 x5 = 1,3 X6 = 6,0

Σx_i = 3,0 + 4,5 + 5,5 + 2,5 + 1,3 + 6,0 = 22,8 e $\overline{X} = \frac{22,8}{6} = 3,8$

Se esses seis valores representassem, por exemplo, as quantidades de peixe pescado (em toneladas) durante seis dias da semana, a quantidade típica pescada por dia, naquela semana, seria 3,8 toneladas. Como estamos representando o valor típico pela média aritmética, podemos falar em quantidade média diária naquela semana.

Exemplo 2.2: No conjunto de dados (2,0 ; 3,3 ; 2,5 ; 5,6 ; 5,0 ; 4,3 ; 3,2), temos

$$\overline{X} = \frac{2,0 + 3,3 + 2,5 + 5,6 + 5,0 + 4,3 + 3,2}{7} = \frac{25,9}{7} = 3,7$$

Novamente, vamos imaginar que esses sete valores representam as quantidades de peixe pescado por dia (em toneladas) durante uma semana inteira. Se quisermos representar a quantidade típica pescada por dia nessa semana usando a média, então poderemos dizer que a média diária de peixe pescado nessa semana foi de 3,7 toneladas. Essa média pode ser comparada com a média diária da semana anterior, mesmo que se tenha trabalhado um dia a mais nessa semana. Assim, concluímos que a produção diária média da semana anterior foi ligeiramente superior à produção diária média da semana deste exemplo.

2.2. Mediana

A mediana de um conjunto ordenado de dados é definida como sendo o "valor do meio" desse conjunto de dados, dispostos em ordem crescente, deixando metade dos valores acima dela e metade dos valores abaixo dela.
Como calcular a mediana ? Basta seguir sua definição. Vejamos:

n é ímpar: Existe apenas um "valor do meio", que é a mediana
Seja o conjunto de dados (2,0 ; 3,3 ; 2,5 ; 5,6 ; 5,0 ; 4,3 ; 3,2).
Ordenando os valores (2,0 ; 2,5 ; 3,2 ; 3,3 ; 4,3 ; 5,0 ; 5,6).
O valor do meio é o 3,3 . A mediana é o valor 3,3.

n é par: Existem dois "valores do meio". A mediana é média aritmética simples deles.
Seja o conjunto de dados (3,0 ; 4,5 ; 5,5 ; 2,5 ; 1,3 ; 6,0).
Ordenando os valores (1,3 ; 2,5 ; 3,0 ; 4,5 ; 5,5 ; 6,0)
Os valores do meio são 3,0 e 4,5. A mediana é (3,0 + 4,5)/2 = 3,75.

Como medida de tendência central, a mediana é até mais intuitiva do que a média, pois representa, de fato, o centro (meio) do conjunto de valores ordenados. Assim como a média, o valor da mediana não precisa coincidir com algum dos valores do conjunto de dados. Em particular, quando os dados forem de natureza contínua, essa coincidência dificilmente ocorrerá. Nos exemplos 2.1 e 2.2, podemos dizer que a produção diária mediana foi de 3,75 toneladas de peixe na primeira semana e de 3,3 toneladas na segunda semana.

2.3. Moda:

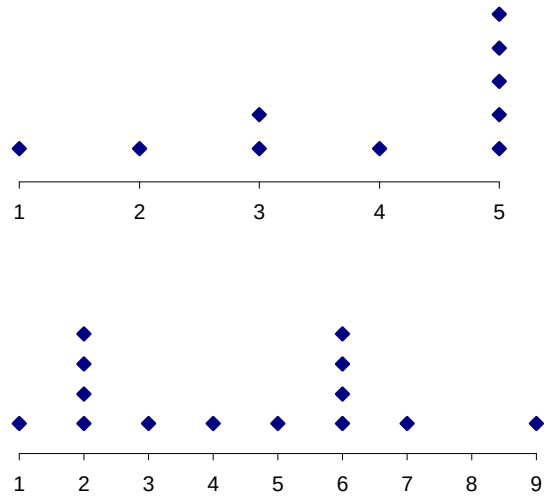
Uma maneira alternativa de representar o que é "típico" é através do valor mais freqüente da variável, chamado de moda.

Exemplo 2.3:

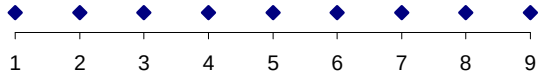
No conjunto de dados
(1; 2 ; 3 ; 3 ; 4 ; 5 ; 5 ; 5 ; 5 ; 5),
há apenas uma moda, o valor 5.
O conjunto de dados é **unimodal.**

No conjunto de dados
(1 ; 2 ; 2 ; 2 ; 2 ; 3 ; 4 ; 5 ; 6 ; 6 ; 6 ; 6 ; 7 ; 9),
existem duas modas, os valores 2 e 6.
O conjunto de dados é **bimodal**

Figura 2.1: Diagrama de pontos para dados (a) unimodal, (b) bimodal e (c) amodal.



Nem sempre a moda existe ou faz sentido.
No conjunto de dados
(1 ; 2 ; 3 ; 4 ; 5 ; 6 ; 7; 8 ; 9)
não existe um valor mais freqüente que os demais.
Portanto, o conjunto de dados é **amodal**.



No caso de uma tabela de freqüências, a classe de maior freqüência é chamada *classe modal*. A moda é também a única das medidas de tendência central que faz sentido no caso de variáveis qualitativas. Assim, a categoria dessas variáveis que aparecer com maior freqüência é chamada de *categoria modal*. No exemplo dos ursos marrons (RTE04-2001), a categoria modal para a variável *sexo* é a categoria *macho*, com 64% dos ursos.

2.4. Moda, Mediana ou Média: Como Escolher ?

Devemos sempre apresentar os valores de todas as medidas de tendência central. Nesta seção, apenas fazemos uma comparação entre elas em situações onde a diferença entre seus valores poderá levar a conclusões diversas sobre os dados.

Mediana *versus* Média

A média é uma medida-resumo muito mais usada na prática do que a mediana. Existem várias razões para essa popularidade da média, entre elas, a facilidade de tratamento estatístico e algumas propriedades interessantes que a média apresenta, o que ficará mais claro quando estudarmos os métodos de estimação.

No entanto, a média é uma medida muito influenciada pela presença de valores extremos em um conjunto de dados (valores muito grandes ou muito pequenos em relação aos demais). Como a média usa os valores de cada observação em seu cálculo, esses valores extremos "puxam" o valor da média em direção a si, deslocando também a representação do centro, que já não será tão central como deveria ser.

A mediana, por sua vez, não é tão influenciada por valores extremos, pois o que utilizamos para calculá-la é a ordem dos elementos e não diretamente seus valores. Assim, se um elemento do conjunto de dados tem o seu valor alterado (um erro, por exemplo), mas sua ordem continua a mesma, a mediana não sofre influência nenhuma.

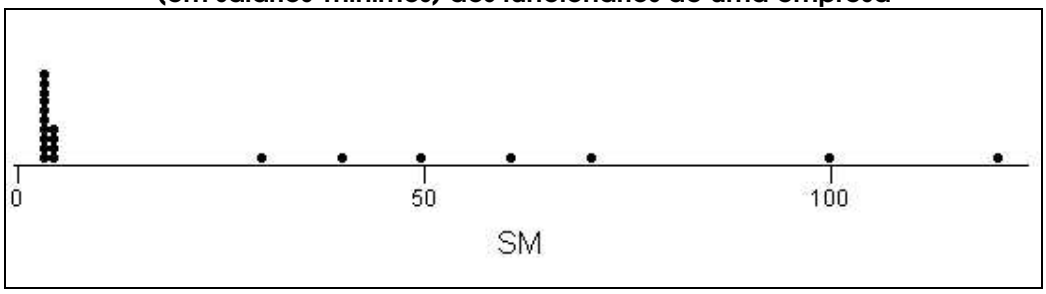
Vejamos um exemplo bastante esclarecedor dessas idéias.

Exemplo 2.4: Salário pago (em salários-mínimos) aos 21 funcionários de uma empresa.

A Figura 2.4 mostra o diagrama de pontos para os salários pagos aos 21 funcionários de uma empresa e, logo abaixo, estão calculados os valores da média e mediana em duas situações: uma, considerando todos os funcionários e outra, sem os quatro salários mais altos. Na primeira situação, a média está muito deslocada em direção aos mais altos salários e não seria uma medida-resumo adequada para representar os salários pagos nessa empresa. A mediana representa bem melhor a realidade, informando que pelo menos metade dos empregados ganham até 4 salários-mínimos.

É claro que, dependendo do uso que se queira fazer da informação sobre os salários dessa empresa, emprega-se a média ou a mediana como medida de centro. Por exemplo, se os administradores quisessem pintar um bom quadro para os salários, escolheriam a média. Já o sindicato, em suas negociações salariais, escolheria a mediana. Nenhum dos dois lados estaria errado do ponto de vista estritamente técnico, estariam apenas usando o que mais lhes convém. Por esta e outras razões, é que devemos estar atentos às estatísticas divulgadas, no sentido de saber como foram calculadas e se, pela natureza dos dados, seriam a forma mais adequada para a descrição do fenômeno estudado.

Figura 2.4 – Diagrama de pontos para os salários pagos (em salários-mínimos) aos funcionários de uma empresa



Situação I: dados completos:
Média = 24,6 SM
Mediana = 4 SM

Situação II: sem os quatro valores mais altos:
Média = 9,8 SM
Mediana = 3 SM

Na segunda situação, quando são removidos os quatro maiores salários, o valor da média muda drasticamente, enquanto o da mediana muda muito pouco, mostrando como ela é pouco influenciada por valores extremos. Na verdade, se esses quatro valores tivessem sido apenas modificados, mas continuassem a ser os quatro maiores, a mediana não mudaria. De fato, a alteração de valor dos maiores salários, desde que eles continuem a ser os maiores, não altera a realidade salarial da empresa e a mediana consegue captar isso, mas a média, não.

De modo geral, o uso da mediana é indicado quando :

- Os valores para a variável em estudo têm distribuição de freqüências assimétrica (verificada através das ferramentas gráficas);
- O conjunto de dados possui algumas poucas observações extremas (valores muito mais altos ou muito mais baixos que os outros);
- Não conhecemos exatamente o valor de algum elemento, mas temos alguma informação sobre a ordem que ele ocupa no conjunto de dados. Por exemplo, no caso dos salários, se alguém não quisesse informar o quanto recebe, mas apenas dissesse que ganha mais (ou menos) do que um certo valor, de modo que conseguíssemos determinar uma ordem para essa pessoa, poderíamos calcular a mediana, mas não a média.

Uma alternativa para a análise de conjunto de dados com observações extremas usando a média como medida de tendência central é analisar separadamente os dois grupos (valores extremos e não extremos) que, quase sempre, podem ser criados a partir de informações externas. No exemplo dos salários, poderíamos criar dois grupos: o das pessoas que recebem 3 e 4 salários-mínimos, que são provavelmente os operários da empresa, e um outro grupo com os demais empregados, que, a julgar pelos salários, devem exercer funções administrativas ou de gerenciamento. Em cada um dos grupos, não haveria mais valores extremos e a média pode ser usada sem problemas.

No caso de poucos valores extremos, pode-se fazer a análise do grupo sem esses valores e justificar a retirada deles, porém, sem os esquecer.

Moda *versus* Média e Mediana

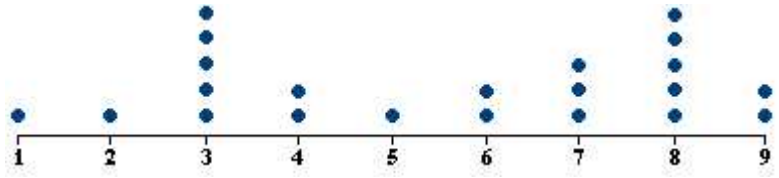
A moda não é uma medida de tendência central muito conhecida, mas tem suas vantagens em relação à média e à mediana, especialmente quando estamos lidando com variáveis que possuem distribuição de freqüências bimodais ou multimodais.

Exemplo 2.5: Considere uma pesquisa de opinião na qual foi perguntado a 26 pessoas de baixa renda: "*Incluindo crianças e adultos, que tamanho de família você acha ideal?*". As respostas são apresentadas na Tabela 2.1 e representadas graficamente na Figura 2.5.

Tabela 2.1 – Distribuição de freqüências das respostas sobre o tamanho ideal de família.

Tamanho ideal da família	1	2	3	4	5	6	7	8	9	10
Freqüência da resposta	1	2	6	2	1	2	3	6	2	1

Figura 2.5 – Diagrama de pontos para as respostas sobre o tamanho ideal de família.



De acordo com essa pesquisa de opinião, o tamanho médio para a família ideal é de 6 pessoas, assim como o tamanho mediano. No entanto, pela Figura 2.5, podemos notar claramente que há dois grupos de pessoas: as que preferem famílias pequenas (até 5 pessoas) e as que gostariam de famílias maiores (6 ou mais pessoas). A distribuição de freqüências das respostas é bimodal. A média nem mediana seriam boas medidas-resumo para a opinião dessas pessoas.

A alternativa é usar a moda e, nesse caso, há duas: os valores 3 e 8. Esses dois valores representam melhor os valores típicos desse conjunto de dados, além de evidenciar uma divisão na opinião desse grupo acerca do tamanho de família ideal.

Nesse exemplo, os valores 3 e 8 têm a mesma freqüência (6 respostas), caracterizando-se como as duas modas. No entanto, mesmo que um dos valores tivesse uma freqüência um pouco inferior (cinco ou quatro respostas), ainda assim poderia haver uma divisão de opinião dessas pessoas. Nesse caso, seria melhor descrever esses dois grupos separadamente, fazendo essa separação dos grupos, se possível, por uma variável externa, como, por exemplo, tamanho da família do respondente. Curiosamente, poderíamos encontrar que pessoas nascidas numa família pequena prefeririam famílias grandes e vice-versa (mas essa é só uma especulação).

2.5. A Forma da Distribuição de Freqüências e as Medidas de Tendência Central

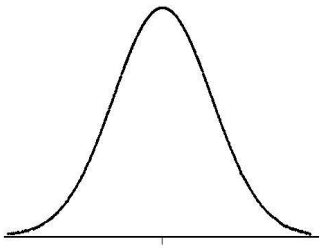
Como já sabemos, a distribuição de freqüências de uma variável pode ter várias formas, mas existem três formas básicas, representadas esquematicamente pelos histogramas da Figura 2.5. Nesta figura, também está a posição de cada uma das medidas de tendência central apresentadas neste texto.

Quando uma distribuição é **simétrica** em torno de um valor (o mais freqüente, isto é, a moda), significa que as observações estão igualmente distribuídas em torno desse valor (metade acima e metade abaixo) Ou seja, esse valor também é a mediana. A média dessas observações também coincidirá com a moda, que coincide com a mediana, pois, se as observações valores estão simetricamente distribuídas em torno de um valor, a média delas será esse valor. Assim, quando a distribuição de freqüências uma variável é simétrica, as três medidas de tendência central têm o mesmo valor.

Se a distribuição é **assimétrica com concentração à esquerda** (ou cauda à direita), a mediana é menor do que a média. Isto acontece porque a mediana é "puxada" em direção à concentração dos valores, enquanto a média é "puxada" em direção à cauda (valores extremos).

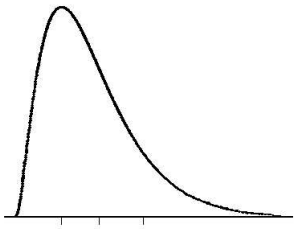
Figura 2.5 – Representação esquemática da forma da distribuição de freqüências e as posições relativas das medidas de tendência central.

Simétrica



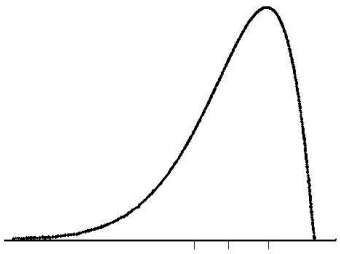
moda = mediana = média

Assimétrica (concentração à esquerda) ou (cauda à direita)



moda < mediana < média

Assimétrica (concentração à direita) ou (cauda à esquerda)



média < mediana < moda

Desse modo, é fácil deduzir o que ocorre quando a distribuição é **assimétrica com concentração à direita** (ou cauda à esquerda): a mediana, "puxada" em direção à concentração dos valores, é maior do que a média, que é influenciada pelos valores pequenos da cauda.

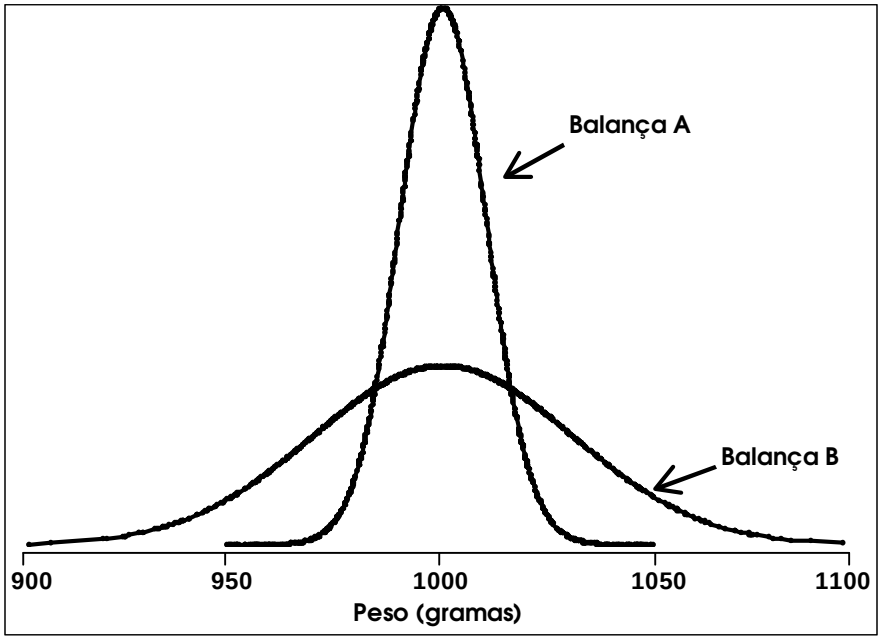
3. Medidas de Variabilidade

As medidas de tendência central (média, mediana, moda) conseguem resumir em um único número, o valor que é "típico" no conjunto de dados. Mas, somente com essas medidas, conseguimos descrever adequadamente o que ocorre em um conjunto de dados?

Vejamos um exemplo: quando pesamos algo em uma balança, esperamos que ela nos dê o verdadeiro peso daquilo que estamos pesando. No entanto, se fizermos várias medições do peso de um mesmo objeto em uma mesma balança, teremos diferentes valores para o peso deste objeto. Ou seja, existe variabilidade na medições de peso fornecidas pela balança. Neste caso, quanto menor a variabilidade desses valores, mais precisa é a balança (considerando que a média dos medidas de peso coincida como seu valor real). Observe a Figura 3.1, onde estão representada as distribuições de freqüências das medições do peso de uma esfera de 1000 g, feitas por duas balanças (A e B). As duas balanças registram o mesmo peso médio de 1000 g (média dos pesos de todas as medições feitas). Isto é, as duas balanças tipicamente acertam o verdadeiro peso da esfera. Porém, pela Figura 3.1, podemos notar que

- As medições da balança A variam pouco em torno de 1000g: oscilam basicamente entre cerca de 950g e 1050g (uma "imprecisão" de 50 g);
- As medições da balança B variam muito em torno de 1000 g: oscilam basicamente entre 900 g e 1100 g, (uma "imprecisão" de 100 g).

Figura 3.1 – Representação esquemática da distribuição de freqüências das medições do peso da esfera de 1000 gramas feitas nas balanças A e B.



Dois conjuntos de dados podem ter a mesma medida de centro (valor típico), porém com uma dispersão diferente em torno desse valor. Desse modo, além de uma medida que nos diga qual é o valor "típico" do conjunto de dados, precisamos de uma medida do *grau de dispersão (variabilidade)* dos dados em torno do valor típico.

O objetivo das medidas de variabilidade é quantificar esse grau de dispersão. Nesta seção, apresentaremos três dessas medidas (amplitude total, desvio-padrão e coeficiente de variação), discutindo suas vantagens e desvantagens. Em discussões posteriores, apresentaremos medidas de variabilidade alternativas.

3.1. Amplitude Total

A medida de variabilidade mais simples é a chamada amplitude total (AT), que é a diferença entre o valor máximo e o valor mínimo de um conjunto de dados

AT = Máximo – Mínimo

Exemplo 3.1: Medições do peso de uma esfera de 1000 g em duas balanças (A e B).

Balança A:	Min = 945 g	Balança B:	Min = 895 g
	Max = 1040 g		Max = 1095 g
	AT = 1040 - 945 = 95 g		AT = 1095 - 895 = 200 g

A variabilidade das medições de peso da balança B é maior que a variabilidade das medições de peso da balança A (apesar do valor médio ser igual)

Embora seja uma medida simples de variabilidade, a amplitude total é um tanto grosseira, pois depende somente de dois valores do conjunto de dados (máximo e mínimo), não captando o que ocorre com os outros valores. Vejamos um exemplo.

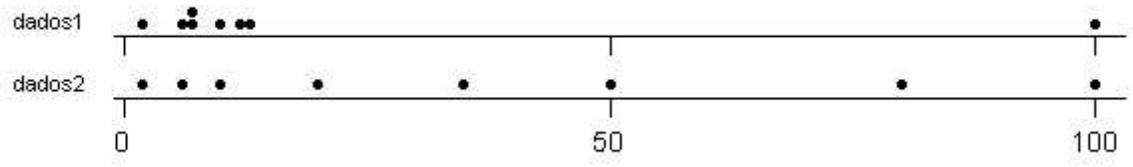
Exemplo 3.2: A Figura 3.2 mostra o diagrama de pontos para os conjuntos de dados I e II:

Dados I: {2, 6, 7, 7, 10, 12, 13, 100}, $AT_1 = 100 - 2 = 98$

Dados II: {2, 6, 10, 20, 35, 50, 80, 100}, $AT_2 = 100 - 2 = 98$

As variabilidades desses dois conjuntos de dados são claramente diferentes. No entanto, apenas usando a amplitude total para medi-las, concluiríamos que os dois conjuntos de dados são igualmente dispersos.

Figura 3.2 – Diagrama de pontos para os conjuntos de dados I e II



Embora a amplitude total sozinha não seja uma boa medida de variabilidade, pode ser usada como uma medida auxiliar na análise da dispersão de um conjunto de dados. Pode também medir o quanto do eixo de valores possíveis para a variável é ocupado pelo conjunto de dados observado.

3.2. Desvio Padrão

Uma boa medida de dispersão deve considerar todos os valores do conjunto de dados e resumir o grau de dispersão desses valores em torno do valor típico.

Considerando a média como a medida de tendência central, podemos pensar em medir a dispersão (desvio) de cada valor do conjunto de dados em relação à ela. A medida mais simples de desvio entre duas quantidades é a diferença entre elas. Assim, para cada valor $\mathbf{X_i}$, teremos o seu desvio em relação à $\overline{X}$ representado por $\mathbf{(X_i - \overline{X})}$.

Exemplo 3.3: no conjunto de dados (1; 1; 2; 3; 4; 4; 5; 6; 7; 7), relativo ao número de filhos de 10 mulheres, temos $\overline{X}$ = 4 filhos. A coluna 1 da Tabela 3.1 mostra esses 10 valores e a coluna 2 mostra o desvio de cada deles até a média.

Tabela 3.1 – Exemplo de cálculo do desvio-padrão.

Coluna 1 X_i	Coluna 2 $X_i - \overline{X}$	Coluna 3 $(X_i - \overline{X})^2$
1	-3	9
1	-3	9
2	-2	4
3	-1	1
4	0	0
4	0	0
5	1	1
6	2	4
7	3	9
7	3	9
Σ	-	46

A idéia do desvio-padrão

Como temos um desvio para cada elemento, poderíamos pensar em resumi-los em um desvio típico, a exemplo do que fizemos com a média. Porém, quando somarmos esses desvios para o cálculo do desvio médio, a soma dará sempre zero, como mostrado na última linha da coluna 2. Isto ocorre com qualquer conjunto de dados, pois os desvios negativos sempre compensam os positivos.

No entanto, os sinais dos desvios não são importantes para nossa medida de dispersão, já que estamos interessados na quantidade de dispersão em torno da média, mas não na direção dela. Portanto, eliminaremos os sinais elevando os desvios ao quadrado[1], como mostrado na coluna 3. A soma desses desvios ao quadrado pode ser, então, dividida entre os participantes do "bolo". Na verdade, por razões absolutamente teóricas, dividiremos essa soma pelo total de participantes menos 1 (n-1). Assim, usando a notação definida anteriormente, teremos

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

Para os dados da Tabela 3.1, teremos 46/(10-1) = 5,11. Esse valor pode ser visto como uma quase-média dos desvios ao quadrado e é chamado de **variância**.

A variância seria nossa medida de variabilidade se não fosse o fato de que ela está expressa em uma unidade diferente da unidade dos dados, pois, ao elevarmos os desvios ao quadrado, elevamos também as unidades de medida em que eles estão expressos. No caso dos dados da Tabela 3.1, medidos em número de filhos, a variância vale 5,11 "filhos ao quadrado", algo que não faz nenhum sentido.

Para eliminar esse problema, extraímos a raiz quadrada da variância e, finalmente, temos a nossa medida de variabilidade, que chamaremos **desvio-padrão (DP)** .

$$DP = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

O desvio-padrão, como o nome já diz, representa o desvio típico dos dados em relação à média, escolhida como medida de tendência central. No exemplo 3.3, temos que o desvio-padrão vale 2,26. Isto significa que a distância típica (padrão) de cada mãe até o número médio de filhos (4 filhos) é de 2,26 filhos. Quanto maior o desvio-padrão, mais diferentes entre si serão as quantidades de filhos de cada mãe.

[1] A função módulo | | também resolve o problema dos sinais (ex: |-3|=3). No entanto, do ponto de vista matemático, a função quadrado é mais fácil de ser tratada.

O desvio-padrão, em alguns livros chamado de *s*, é uma medida sempre positiva. Se observarmos a maneira como ele é calculado, veremos que não há como obter um valor negativo.

Exemplo 3.4: Os agentes de fiscalização de certo município realizam, periodicamente, uma vistoria nos bares e restaurantes para apurar possíveis irregularidades na venda de seus produtos. A seguir, são apresentados dados de uma vistoria sobre os pesos (em gramas) de uma amostra de 10 bifes, constantes de um cardápio de um restaurante como "bife de 200 gramas".

170 175 180 185 190 195 200 200 200 205

Como podemos notar, nem todos os "bifes de 200 gramas" pesam realmente 200 gramas. Esta variação é natural e é devida ao processo de produção dos bifes. No entanto, esses bifes deveriam pesar cerca de 200 gramas e com pouca variação em torno desse valor. Com o auxílio da Tabela 3.2, calcularemos a média e o desvio-padrão.

Tabela 3.2 – Cálculo do desvio-padrão para o exemplo dos "bifes de 200 gramas".

i	Xi	Xi - $\overline{X}$	(Xi - $\overline{X}$)^2
1	170	170 – 190 = -20	400
2	175	175 – 190 = -15	225
3	180	180 – 190 = -10	100
4	185	185 – 190 = -05	25
5	190	190 – 190 = 0	0
6	195	195 – 190 = 05	25
7	200	200 – 190 = 10	100
8	200	200 – 190 = 10	100
9	200	200 – 190 = 10	100
10	205	205 – 190 = 15	225
Soma	1900	0	1300

Média: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1900}{10} = 190\,gramas$

Desvio Padrão: $s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1300}{9}} = 12\,gramas$

Os bifes desse restaurante pesam, em média, 190 gramas, com um desvio-padrão de 12 gramas. Ou seja, os pesos dos "bifes de 200 gramas" variam tipicamente entre 178 e 202 gramas ($\bar{x} \pm s$). Analisando esses valores, concluímos que esse restaurante pode estar lesando a maior parte de seus clientes.

Para casos como esse, os agentes fiscalizadores podem estabelecer parâmetros (valores) para saber até quanto a média pode se desviar do valor correto e o quanto de variação eles podem permitir numa amostra para concluir que o processo de produção de bifes não possui problemas. Por exemplo, a média da amostra não poderia ser inferior a 200 gramas, com um desvio-padrão que não seja superior a 5% dessa média.

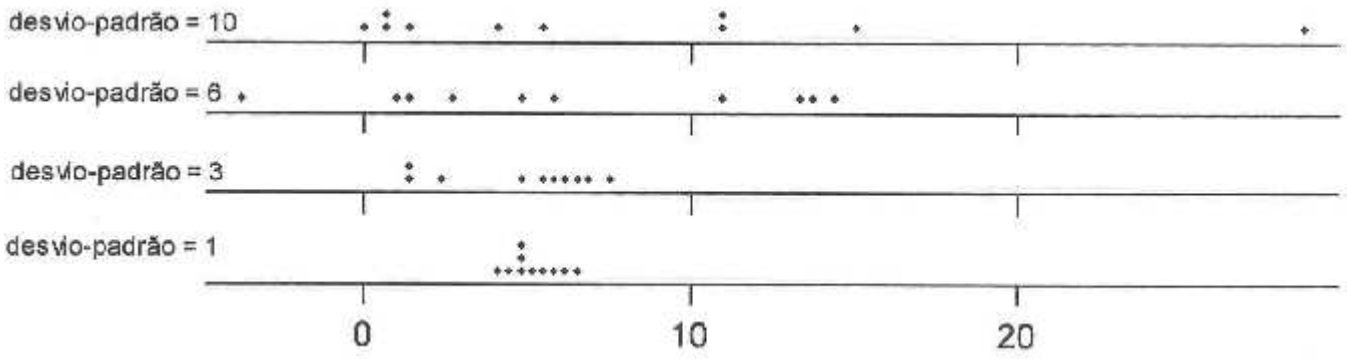
Essas idéias são utilizadas no controle do processo de produção das indústrias, onde já se espera alguma variação entre as unidades produzidas. Porém, essa variação deve estar sob controle[2]. Numa indústria farmacêutica, por exemplo, espera-se que os comprimidos de um certo medicamento sejam produzidos com uma certa variação em sua composição (maior ou menor quantidade do princípio ativo), devido à própria maneira como os comprimidos são produzidos (máquinas, pessoas, etc.). No entanto, esta variação deve ser pequena, para que não sejam produzidos comprimidos com doses sub-clínicas (com pouco do princípio ativo) ou com superdosagem do princípio ativo, o que, em ambos os casos, pode causar sérias complicações à saúde do paciente.

[2] O chamado Controle Estatístico de Processos é um conjunto de ferramentas estatísticas (gráficos, medidas de centro e de variabilidade) bastante usado no Controle da Qualidade Total em uma grande parte das indústrias nacionais e internacionais.

O desvio-padrão nos permite distinguir numericamente conjuntos de dados de mesmo tamanho, mesma média, mas que são visivelmente diferentes, como mostra a ilustração da Figura 3.3. Usando o desvio-padrão, também conseguimos representar numericamente a variabilidade das medições das balanças A e B (exemplo 3.1), que, apesar de possuírem a mesma média, possuem variabilidades bastante diferentes.

Quando os conjuntos de dados a serem comparados possuem médias diferentes, a comparação da variabilidade desses conjuntos deve levar em conta essa diferença. Por esta e outras razões, definiremos uma terceira medida de variabilidade, o *coeficiente de variação*.

Figura 3.3 – Conjuntos de dados de mesmo tamanho, mesma média e variabilidades diferentes.



Tamanho de amostra: n = 10
Média = 5

3.3. Coeficiente de Variação

Ao analisarmos o grau de dispersão de um conjunto de dados, poderemos nos deparar com uma questão do tipo: um desvio-padrão de 10 unidades é pequeno ou grande?

Vejamos:

- Se estivermos trabalhando com um conjunto de dados cuja a média é 10.000, um desvio típico de 10 unidades em torno dessa média significa pouca dispersão;
- Mas, se a média for igual a 100, um desvio típico de 10 unidades em torno dessa média significa muita dispersão.

Assim, antes de responder se um desvio-padrão de 10 unidades é grande ou pequeno, devemos avaliar sua magnitude em relação à média:

- No primeiro caso, o desvio-padrão corresponde a 0,1% da média: $\frac{10}{10.000} = 0{,}001$ ou 0,1%;
- No segundo caso, o desvio-padrão corresponde a 10% da média: $\frac{10}{100} = 0{,}1$ ou 10%.

À essa razão entre o desvio-padrão e a média damos o nome de **Coeficiente de Variação:**

$$CV = \frac{\text{desvio padrão}}{\text{média}}$$

Quanto menor o Coeficiente de Variação de um conjunto de dados, menor é a sua variabilidade. O Coeficiente de Variação expressa o quanto da escala de medida, representada pela média, é ocupada pelo desvio-padrão.

O Coeficiente de Variação é uma medida adimensional, isto é, não depende da unidade de medida. Essa característica nos permite usá-lo para comparar a variabilidade de conjuntos de dados medidos em unidades diferentes, o que seria impossível usando o desvio-padrão.

a) Comparação da homogeneidade de uma mesma variável entre grupos diferentes.

Exemplo 3.5: Numa pesquisa em saúde Saúde Ocupacional, deseja-se comparar a idade de motoristas e cobradores de ônibus da região metropolitana de Belo Horizonte. Algumas estatísticas descritivas são apresentadas na Tabela 3.3 .

Tabela 3.3 - Estatísticas descritivas para idade de motoristas e cobradores.

Grupo	n	$\overline{X}$ (anos)	DP (anos)	CV
Motoristas	150	35,6	7,08	0,199 (19,9%)
Cobradores	50	22,6	3,11	0,137 (13,7%)

Fonte: dados fictícios

Os motoristas são, em média, 13 anos mais velhos do que os cobradores. Ao compararmos o grau de dispersão dos dois grupos usando o desvio-padrão, concluiríamos que os motoristas são menos homogêneos quanto à idade do que os cobradores. Ao fazermos isso, estamos esquecendo que, apesar de estarem em unidades iguais, as medidas de idade nos dois grupos variam em escalas diferentes. As idades dos motoristas variam em torno dos 35 anos e podem chegar até a 18 anos (idade mínima para se conseguir a habilitação), numa amplitude de 17 unidades. Enquanto isso, as idades dos cobradores variam em torno de 22 anos e também só podem chegar até a 18 anos, uma amplitude de apenas 4 anos. Assim, os motoristas têm a possibilidade de ter um desvio-padrão maior do que o dos cobradores. Se levarmos em conta a escala de medida, usando o coeficiente de variação, veremos que os motoristas são somente um pouco mais heterogêneos (dispersos) quanto à idade do que os cobradores.

b) Comparação da homogeneidade entre variáveis diferentes em um mesmo grupo.

Na mesma pesquisa do item (a), foram coletados dados para outras variáveis, como tempo de profissão e salário, cujas as estatísticas descritivas para o grupo de motoristas são apresentadas na Tabela 3.4.

Tabela 3.4 - Estatísticas descritivas para idade, tempo de profissão salário de motoristas.

Variável	$\overline{X}$	DP	CV
Idade	35,6 anos	5,08 anos	0,143 (14,3%)
Tempo de profissão	6,5 anos	2,98 anos	0,458 (45,8%)
Salário	537,52 reais	25,34 reais	0,047 (4,7%)

Gostaríamos de saber em qual das variáveis os motoristas são mais parecidos entre si. Essa informação é conseguida através da análise da variabilidade, procurando-se a variável mais homogênea. Se usarmos o desvio-padrão nessa análise, além de não considerarmos as diferentes escalas de medidas, estaremos fazendo comparações um tanto estranhas, pois, como comparar anos com reais? Desse modo, o coeficiente de variação é a única opção nesse caso, pois ele é adimensional. Ao analisarmos os CV´s das três variáveis, concluiremos que os motoristas são mais homogêneos quanto ao salário. Uma pessoa desatenta, usando o desvio-padrão, escolheria o tempo de profissão como a mais homogênea, que, na verdade, é a mais heterogênea.

Classificação do grau de homogeneidade da variável através do coeficiente de variação

Quanto menor o coeficiente de variação, mais homogênea é a variável naquele conjunto de dados. A definição do que pode ser considerado pouco ou muito homogêneo segundo o coeficiente de variação varia de acordo com a área de estudo de onde provêm os dados.

Em geral, um coeficiente de variação menor de que 0,25 indica uma variável homogênea. Em populações onde já se espera uma variabilidade maior entre os indivíduos, essas faixas de homogeneidade devem ser redefinidas. Espera-se, por exemplo, que as notas de um processo seletivo como o vestibular tenham uma alta variabilidade, devido aos diferentes níveis de preparação das pessoas que prestam os exames. Nesse caso, se uma determinada prova consegue um coeficiente de variação de 0,27, por exemplo, isto pode significar um grau de homogeneidade razoável. Depois de alguma experiência analisando esse tipo de dado, consegue-se definir faixas de homogeneidade.

Exemplo 3.6: Três grupos de cães machos de raças diferentes, porém criados para a mesma utilidade (cães de guarda), foram submetidos a um teste de força da mordida, medida em toneladas por cm^3. As medidas-síntese de tendência central e variabilidade são apresentadas na Tabela 3.5.

Tabela 3.5 - Estatísticas descritivas para força de mordida (ton/cm^3) de três raças de cães.

Raça	$\overline{x}$	DP	CV
I	1,5	0,6	0,40
II	2,4	0,6	0,25
III	2,2	0,3	0,14

Fonte: dados fictícios

Os cães da raça II são, em média, os mais fortes e os da raça III são os mais homogêneos (menor CV). Se uma pessoa está interessada em criar cães de guarda, deveria optar pelos cães da raça III, que, embora um pouco mais fracos do que os da raça II, são mais parecidos entre si. Isto é, as crias de cães dessa raça têm uma força que varia pouco em torno de 2,2 ton/cm^3, enquanto os da raça II são, em média, um pouco mais fortes, porém podem variar mais em torno desse valor médio, tanto para valores mais altos como para valores mais baixos.

Embora seja uma medida de variabilidade bastante usada, o desvio-padrão nem sempre é a medida mais adequada. Quando a média não é a medida de tendência central mais indicada (por causa da presença de valores extremos, por exemplo), o desvio-padrão e, consequentemente, o coeficiente de variação, também não são indicados para medir a variabilidade, pois ambos dependem da média e sofrem dos mesmos problemas que ela. Desse modo, existem medidas de variabilidade alternativas, que serão apresentadas oportunamente neste texto, pois dependem de conceitos ainda não trabalhados até o momento.

3.4. Regra do Desvio-Padrão para Dados com Distribuição Simétrica

Quando a distribuição dos dados é simétrica, existe uma regra que nos permite determinar a freqüência de dados contidos em certos intervalos construídos a partir do conhecimento da média e do desvio-padrão. Os intervalos mais comuns são aqueles simétricos em torno da média e que se afastam dela por um, dois ou três desvios-padrão, para a direita e para a esquerda, como ilustra a Figura 3.4.

Essa regra pode ter os mais variados usos como, por exemplo, na classificação de um valor como extremo. Pela regra, 99,7% dos dados estão contidos no intervalo de três desvios-padrão a partir da média. Se um valor está fora desse intervalo, pode ser considerado extremo por essa regra e merecer uma atenção especial[3].

Exemplo 3.7: Numa certa população, os recém-nascidos (de gestações únicas de 37 semanas sem intercorrências) têm peso médio de 3000g e um desvio-padrão de 250g, sendo a distribuição dos pesos simétrica em torno dessa média. Assim, podemos concluir que 68,3% dos recém-nascidos pesam entre 2750g e 3250g, 95,4% deles pesam de 2500g a 3500g e 99,7% desses bebês pesam de 2250g a 3750g.

Os intervalos deste exemplo podem servir como **faixas de referência** para o peso de bebês nascidos de gravidez normal. Uma faixa de referência para uma determinada característica é um intervalo de valores considerados típicos para uma certa porcentagem da população considerada "sadia". O intervalo de 2500g a 3500g pode ser visto como uma faixa de referência de 95,4% para o peso de bebês vindos de gravidez normal, pois 95,4% desses bebês nascem com pesos nesse intervalo.

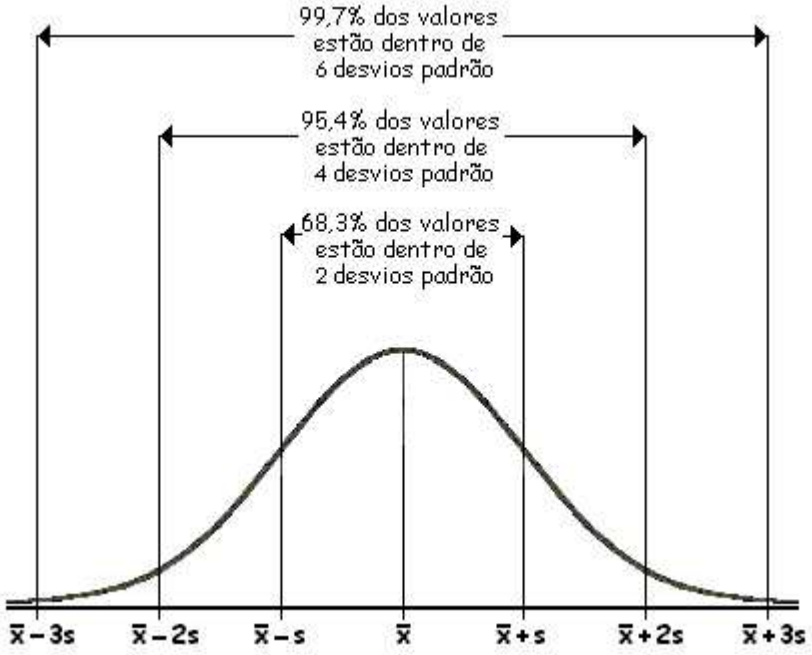
Outros exemplos de faixas de referência podem ser encontrados em resultados de exames clínicos, onde é apresentada, juntamente com o resultado do indivíduo, a faixa de referência (ou normalidade) para a característica em exame. Um indivíduo com resultado fora da faixa de referência deve ser investigado. No exemplo dos recém-nascidos, aqueles que nascem com peso

[3] Existem outros métodos para detecção de valores extremos (ou discrepantes), que serão apresentados mais adiante.

fora da faixa de referência de 95,4% estão entre os 2,3% mais (ou menos) pesados. Descartada a possibilidade de erro, podem pertencer a população "sadia", porém, é mais provável que tenham vindo de uma população centrada em outro valor médio, não considerada "sadia", e devem ser investigados.

A regra para dados de distribuição simétrica nos fornece faixas de referência de 68,3%, 95,4% e 99,7%. Podemos construir faixas de referência com outros percentuais e este assunto será abordado mais adiante.

Figura 3.4 – Ilustração da regra do desvio padrão para dados com distribuição simétrica.



4. Medidas de Posição

- Então, qual foi sua posição final na corrida ?
- Ah, eu fiquei em 3º lugar!
- Puxa...Foi mesmo ? E quantos estavam correndo ?
- Três.

Quando falamos de posição ou colocação de um indivíduo em uma corrida ou em um teste como o Vestibular, freqüentemente nos referimos ao seu posto, como 1°, 2° , 3°, 29° ou último lugar. Mas como vimos na estória acima, para sabermos se uma dada colocação é ou não um bom resultado, precisamos informar quantos indivíduos participaram da corrida ou do Vestibular.

As duas medidas de posição que veremos aqui, os percentis e os escores padronizados, solucionam este e outros problemas de posicionamento (*ranking)*. A posição de um indivíduo no conjunto de dados é mostrada, pelo percentil, contando-se (em porcentagem) quantos indivíduos do conjunto têm valores menores que o deste indivíduo. O escore padronizado mostra a posição do indivíduo em relação à média de todos os indivíduos do conjunto de dados, considerando também a variabilidade de tais medidas.

Como veremos, estas duas medidas de posição podem ser usadas para comparar a posição do indivíduo em diferentes conjuntos de dados, nos quais foram medidas as mesmas variáveis ou variáveis diferentes. Veremos também que os escores padronizados de um indivíduo calculados de diferentes variáveis podem ser combinados para gerar a posição *global* do indivíduo.

4.1. Percentis

Quando dizemos que certo aluno está entre os 5% melhores do colégio ou que um país está entre os 10% mais pobres, não precisamos nem saber quantos alunos tem o colégio ou em quantos países estão sendo consideradas as rendas. Aqui já houve uma padronização da posição usando-se a ***porcentagem*** de alunos ou países com desempenho ou renda ***abaixo*** do valor considerado. É este raciocínio que define os percentis.

Definição:

O percentil de ordem K (onde k é qualquer valor entre 0 e 100), denotado por P_k, é o valor tal que K% dos valores do conjunto de dados são menores ou iguais a ele.

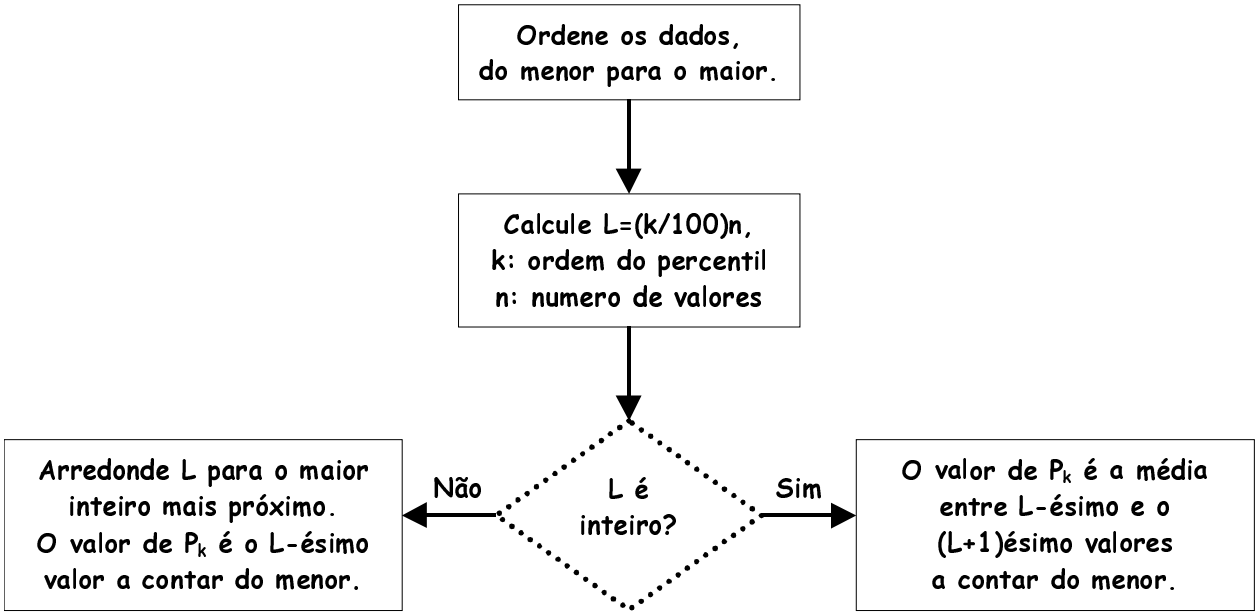
Assim, o percentil de ordem 10, o P_{10}, é o valor da variável tal que 10% dos valores são menores ou iguais a ele; o percentil de ordem 65 deixa 65% dos dados menores ou iguais a ele, etc.

Os percentil de ordem 10, 20, 30, ... 90 dividem o conjunto de dados em dez partes com mesmo número de observações e são chamados de *decis*.

Os percentis de ordem 25, 50 e 75 dividem o conjunto de dados em quatro partes com o mesmo número de observações. Assim, estes três percentis recebem o nome de quartis – **primeiro quartil (Q_1), segundo quartil (Q_2) e terceiro quartil (Q_3)**, respectivamente. O segundo quartil é a já conhecida mediana.

Existem vários processos para calcular os percentis, usando interpolação. Vamos ficar com um método mais simples, mostrado na Figura 3.5 a seguir. As diferenças serão muito pequenas e desaparecerão à medida que aumenta o número de dados.

Figura 3.5 - Determinação do Percentil de ordem K (Triola, 1996).



Exemplo 4.1: Considere as notas finais dos 40 candidatos ao curso de Direito no Vestibular de certa faculdade, já colocadas em ordem crescente:

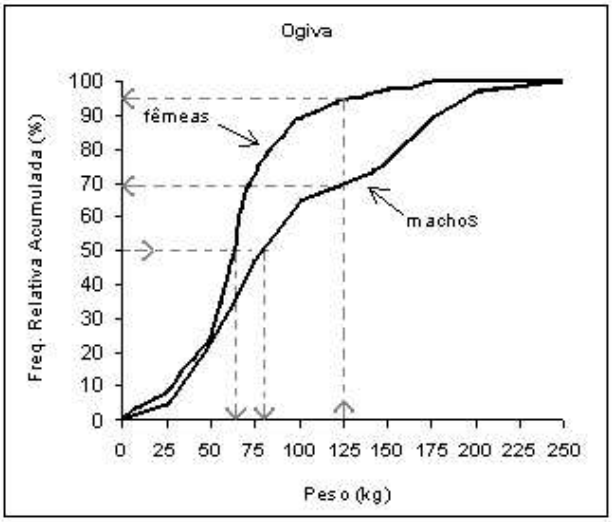
40 41 42 42 44 47 48 48 49 49 51 52 53 58 59 62 63 64 65 66
67 68 69 70 75 76 83 83 85 86 86 87 87 88 92 93 94 95 97 98

Vamos calcular alguns percentis.

- Percentil de ordem 10: 10% de 40 = 4. Então o P_{10} = média(4° e 5° valores)=(42+44)/2 = 43.
- Percentil de ordem 95: 95% de 40 = 38. Então o P_{95} = média(38° e 39° valores)=(95+97)/2 = 96.
- Primeiro Quartil: 25% de 40 = 10. Então o Q_1 = média(10° e 11° valores)=(49+51)/2 = 50.
- Terceiro Quartil: 75% de 40 = 30. Então o Q_3 = média(30° e 31° valores)=(86+86)/2 = 86.
- Mediana: 50% de 40 = 20. Então mediana = média(20° e 21° valores)=(66+67)/2 = 66,5.

A **ogiva** é o gráfico representativo dos percentis. Dada a ogiva, podemos determinar quaisquer percentis. Na Figura 3.6, temos as ogivas para peso de ursos marrons machos e fêmeas do Exemplo Inicial (RTE04-2001). Partindo do valor 125kg, podemos ver que corresponde ao percentil de ordem 70 para os machos e de ordem 95 para as fêmeas. Partindo da freqüência acumulada de 50%, podemos ver que 80kg é a mediana do peso dos machos e 65kg é a mediana do peso para as fêmeas.

Figura 3.6: Ogiva para peso de urso marrons, segundo sexo.



4.2. Escores Padronizados

Os escores padronizados são medidas que, calculadas para cada observação do conjunto de dados, nos permitem fazer comparações entre valores de variáveis medidas em escalas diferentes. Vejamos um exemplo.

Exemplo 4.2: (Dados fictícios) Os 20 alunos da oitava série de uma escola foram submetidos, pelo seu professor de Educação Física, a cinco testes de aptidão física e a um teste de conhecimento desportivo:

1- Abdominal: número de abdominais realizados em 2 minutos;
2- Salto em extensão: comprimento do salto (centímetros);
3- Suspensão de braços flexionados: tempo em suspensão (segundos);
4- Corrida: distância (em metros) percorrida em 12 minutos ;
5- Natação: tempo (em segundos) para nadar 50 metros;
6- Conhecimento desportivo: prova escrita (0 a 100 pontos).

Os resultados dos seis testes para os 20 alunos da turma são mostrados no Quadro 4.1 a seguir.

Quadro 4.1: Resultados (escores originais) dos 20 alunos[4] nos seis testes.

Aluno	Abdominal	Salto	Suspensão	Corrida	Natação	Conhecimento
Pedro	34	108	64	1989	34	64
João	30	88	33	1461	32	82
Manuel	27	87	23	1333	27	66
Maria	25	94	12	1858	29	78
Vinícius	26	102	10	1986	30	68
Luiza	27	80	16	1267	32	84
Marina	28	90	20	1743	33	76
Camila	28	92	27	1833	31	71
Guido	29	71	30	1255	29	72
Bárbara	29	88	36	1503	35	75
Luiz	30	89	42	1600	28	77
Gabriela	30	90	39	1747	31	76
Antônio	30	98	45	1930	33	74
Daniele	31	84	48	1276	30	73
Marcelo	31	91	51	1716	25	81
Rodrigo	32	70	57	1054	27	69
Luciana	32	89	54	1535	28	74
Rafael	33	74	60	1084	30	86
Flávia	33	106	67	1968	26	79
Ana	35	69	67	1019	30	75

Vamos tentar resolver algumas questões propostas pelo professor sobre o desempenho dos alunos nos testes:

Questão nº 1: Em um dado teste, qual foi o aluno de melhor desempenho ? E de pior desempenho? Como se poderia classificar os alunos de acordo com seu desempenho em um dado teste ?

[4] Por razões didáticas, os alunos estão identificados pelo nome. No entanto, numa análise de dados, a identidade das unidades de análise (neste caso, os alunos) deve ser mantida em sigilo, por questões éticas.

Esta é fácil. No teste de abdominal, por exemplo, a Ana se saiu melhor, pois fez mais abdominais no tempo marcado, enquanto a Maria, que fez o menor número de abdominais, teve o pior desempenho da turma neste teste.

Assim, o aluno com o melhor desempenho em um teste é aquele que obteve o maior escore no teste, exceto para o de natação, onde o melhor desempenho é do aluno que fez o menor tempo. Para cada teste, os alunos poderiam ser classificados de acordo com seu desempenho naquele teste (do melhor para o pior) simplesmente ordenando os valores de seus escores naquele teste (do menor para o maior, no caso da natação, ou o contrário, para os outros testes). O Quadro 4.1 mostra os alunos com o melhor e pior desempenho da turma em cada teste.

Quadro 4.1: Alunos com o melhor e o pior desempenho em cada teste.

Teste	**Pior Desempenho**	**Melhor desempenho**
Número de abdominais em 2 minutos	Maria	Ana
Salto em extensão (centímetros)	Ana	Pedro
Suspensão de braços flexionados (segundos)	Vinícius	Flávia e Ana
Distância percorrida em 12 minutos (metros)	Ana	Pedro
Natação de 50 metros (segundos)	Bárbara	Marcelo
Conhecimento desportivo (0 a 100 pontos)	Pedro	Rafael

Questão nº 2 Para um dado aluno, em qual teste onde ele se saiu melhor (ou pior) em relação à turma ? Como se poderia classificar os testes de acordo com o desempenho de um dado aluno ?

Vamos definir o *desempenho geral da turma em um teste* como sendo o escore médio de todos os alunos naquele teste:

Teste	**Média da turma**
Abdominais em 2 minutos	30 abdominais
Salto em extensão	88 centímetros
Suspensão de braços flexionados	40 segundos
Corrida em 12 minutos	1558 metros
Natação de 50 metros	30 segundos
Conhecimento desportivo	75 pontos

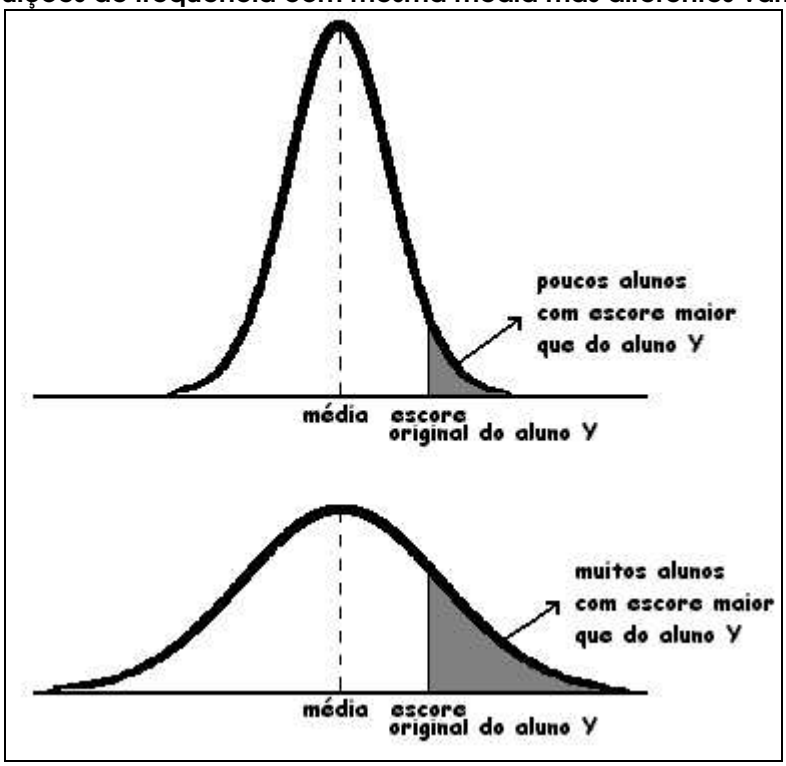
Inicialmente, podemos pensar em classificar o desempenho de um aluno em um teste como *bom* se seu escore ficou acima da média da turma (abaixo, para o teste de natação), e como *ruim* se seu escore ficou abaixo da média da turma (acima, para o teste de natação).

	Teste	O resultado está	...	da média da turma	Desempenho
Para o Pedro:	Abdominal:	34 – 30 =	4 abdominais	Acima	bom
	Salto:	108 – 88 =	20 centímetros	Acima	bom
	Suspensão:	64 – 40 =	24 segundos	Acima	bom
	Corrida:	1989 – 1558 =	431 metros	Acima	bom
	Natação:	34 – 30 =	4 segundos	Acima	ruim
	Conhecimento:	64 – 75 =	-11 pontos	Abaixo	ruim

Agora sabemos que há quatro testes em que o Pedro foi bem em relação à turma, mas como saber em qual destes ele testes teve o melhor desempenho ? Não podemos comparar diretamente seus escores em cada teste, pois estão em unidades de medida diferentes.

Além disso, não basta saber que, em um dado teste, ele ficou acima da média, mas também o quão distante da média em relação à variabilidade dos resultados da turma. No esquema da Figura 3.7, temos um aluno com o mesmo escore original em um teste feito em duas turmas onde a média dos resultados foi a mesma, mas com variabilidades diferentes. Assim, para este aluno, embora a distância, em valores absolutos, de seu resultado à média tenha sido a mesma nas duas turmas, ele se saiu melhor na primeira, pois há menos alunos com escore tão alto quanto o seu. Em outras palavras, como há menos variabilidade na primeira turma, um escore alto provoca maior "efeito" em termos de posição.

Figura 3.7: Esquema comparativo da posição relativa de um valor de escore original em duas distribuições de freqüência com mesma média mas diferentes variabilidades.



Uma solução para estes dois problemas – **unidades de medida diferentes** e **necessidade de se levar em conta a variabilidade dos resultados da turma** – quando da comparação dos resultados de um aluno entre os vários teste é dividir a distância entre o escore original e a média da turma pelo desvio-padrão da turma no dado teste. Assim, temos:

Teste		**Média**	**Desvio-Padrão**	
Abdominais em 2 minutos	30	abdominais	3	Abdominais
Salto em extensão	88	centímetros	11	Centímetros
Suspensão de braços flexionados	40	segundos	18	Segundos
Corrida em 12 minutos	1558	metros	327	Metros
Natação de 50 metros	30	segundos	3	Segundos
Conhecimento desportivo	75	pontos	6	Pontos

	Teste	(escore original – média)/desvio padrão
Para o Pedro:	Abdominal:	(34 – 30)/3 = 1,33
	Salto:	(108 – 88)/11 = 1,82
	Suspensão:	(64 – 40)/18 = 1,33
	Corrida:	(1989 – 1558)/327 = 1,32
	Natação:	(34 – 30)/3 = 1,33
	Conhecimento:	(64 – 75)/6 = -1,83

Agora temos medidas de desempenho do Pedro em cada teste, que são *adimensionais* e, assim, podemos fazer um *ranking* entre os testes, comparando diretamente estas medidas (lembrando que, no teste de natação, quanto menor o valor menor, melhor é o desempenho do aluno):

Ordenação dos testes, do melhor para o pior, segundo o desempenho do aluno Pedro				
Salto	Abdominais/Suspensão	Corrida	Natação	Conhecimento

Esta conta que fizemos para o Pedro chama-se ***escore padronizado***.

Definição:

O escore padronizado de um aluno em um teste é calculado como:

$$\textbf{EscorePadronizado} = \frac{\textbf{EscoreOriginal} - \textbf{Média}}{\textbf{DesvioPadrão}}$$

onde *média* e *desvio padrão* são calculados para cada teste usando os escores originais de todos os 20 alunos naquele teste.

O Quadro 4.2 mostra os escores padronizados dos 20 alunos nos seis testes.

Quadro 4.2: Escores padronizados dos 20 alunos nos seis testes.

Aluno	Abdominal	Salto	Suspensão	Corrida	Natação	Conhecimento
Pedro	1,33	1,82	1,33	1,32	1,33	-1,83
João	0,00	0,00	-0,39	-0,30	0,67	1,17
Manuel	-1,00	-0,09	-0,94	-0,69	-1,00	-1,50
Maria	-1,67	0,55	-1,56	0,92	-0,33	0,50
Vinícius	-1,33	1,27	-1,67	1,31	0,00	-1,17
Luiza	-1,00	-0,73	-1,33	-0,89	0,67	1,50
Marina	-0,67	0,18	-1,11	0,57	1,00	0,17
Camila	-0,67	0,36	-0,72	0,84	0,33	-0,67
Guido	-0,33	-1,55	-0,56	-0,93	-0,33	-0,50
Bárbara	-0,33	0,00	-0,22	-0,17	1,67	0,00
Luiz	0,00	0,09	0,11	0,13	-0,67	0,33
Gabriela	0,00	0,18	-0,06	0,58	0,33	0,17
Antônio	0,00	0,91	0,28	1,14	1,00	-0,17
Daniele	0,33	-0,36	0,44	-0,86	0,00	-0,33
Marcelo	0,33	0,27	0,61	0,48	-1,67	1,00
Rodrigo	0,67	-1,64	0,94	-1,54	-1,00	-1,00
Luciana	0,67	0,09	0,78	-0,07	-0,67	-0,17
Rafael	1,00	-1,27	1,11	-1,45	0,00	1,83
Flávia	1,00	1,64	1,50	1,25	-1,33	0,67
Ana	1,67	-1,73	1,50	-1,65	0,00	0,00

O escore padronizado mede a distância do escore original à média em ***número de desvios-padrão***. Assim, nos testes de abdominais e natação, o Pedro está 1,33 desvios-padrão acima da média de sua turma, o que significa um bom resultado em abdominais, mas um resultado ruim em natação. Note que o conjunto de escores padronizados em um teste têm média igual a zero e desvio padrão-igual a 1.

Questão nº 3: Como se poderia resumir, em um único número, o desempenho do aluno em todos os testes (desempenho global) ? Como se poderia classificar os alunos de acordo com seu desempenho global ?

Se o professor desejar calcular, para cada aluno, um único número, um índice, que expresse o desempenho global do aluno em todos os testes, ele deve usar os escores padronizados, pois já sabemos que são medidas adimensionais. Ele poderia calcular, por exemplo, a *média dos escores*

padronizados, lembrando-se de inverter o sinal do escore padronizado do teste de natação. Para o Pedro, temos:

$$DG_{Pedro} = \frac{(1{,}33) + (1{,}82) + (1{,}33) + (1{,}32) + (-1{,}33) + (-1{,}83)}{6} = \frac{2{,}64}{6} = 0{,}44$$

Fazendo o mesmo cálculo para os demais alunos, temos o seguinte *ranking* dos alunos segundo seu desempenho global, do melhor para o pior:

> Flávia (1,23), Marcelo (0,73), Pedro (0,44), Luciana (0,33), Luiz (0,22), Rafael (0,20), Antônio (0,19), Gabriela (0,09), João e Ana (-0,03), Daniele (-0,13), Maria (-0,15), Camila (-0,20), Vinícius e Rodrigo (-0,26), Marina (-0,31), Bárbara (-0,40), Luiza (-0,52), Manuel (-0,54), Guido (-0,59).

Suponha que, na definição da medida para o desempenho global de cada aluno, o professor queira dar mais peso àqueles testes com maior poder de discriminação entre os alunos, ou seja, dar mais peso aos testes com resultados mais heterogêneos entre os 20 alunos. Uma sugestão é utilizar uma média *ponderada* dos escores padronizados, usando como *peso* de cada teste seu coeficiente de variação (CV), pois, sabemos que esta é uma medida de variabilidade adimensional.

Assim, temos:

Teste	**CV da turma**
Abdominais em 2 minutos	0,10
Salto em extensão	0,13
Suspensão de braços flexionados	0,45
Corrida em 12 minutos	0,21
Natação de 50 metros	0,10
Conhecimento desportivo	0,08

Chamando de Z_1, Z_2, Z_3, Z_4, Z_5 e Z_6 o escore padronizado de cada teste, a fórmula do índice de Desempenho Global Ponderado (DGP) pelo coeficiente de variação é descrita abaixo:

$$\mathbf{DGP} = \frac{\mathbf{(0{,}10 \times Z_1) + (0{,}13 \times Z_2) + (0{,}45 \times Z_3) + (0{,}21 \times Z_4) + (0{,}10 \times (-Z_5)) + (0{,}08 \times Z_6)}}{\mathbf{0{,}10 + 0{,}13 + 0{,}45 + 0{,}21 + 0{,}10 + 0{,}08}}.$$

Como exemplo, para o aluno Pedro, temos:

$$DGP_{Pedro=} \frac{(0{,}10 \times 1{,}33) + (0{,}13 \times 1{,}82) + (0{,}45 \times 1{,}33) + (0{,}21 \times 1{,}32) + (0{,}10 \times -1{,}33) + (0{,}08 \times -1{,}83)}{1{,}07} = \frac{0{,}96}{1{,}07} = 0{,}90$$

Fazendo o mesmo cálculo para os demais alunos, temos o seguinte *ranking* dos alunos segundo seu desempenho global ponderado, do melhor para o pior:

> Flávia (1,34), Pedro (0,90), Marcelo (0,65), Luciana (0,44), Antonio (0,34), Rafael (0,26), Ana (0,25), Luiz (0,17), Gabriela (0,09), Daniele e Rodrigo (-0,02), João (-0,20), Camila (-0,24), Bárbara (-0,31), Marina (-0,48), Maria e Vinícius (-0,50), Guido (-0,64), Manuel (-0,66), Luiza (-0,87).

5. O *Boxplot*

O *Boxplot* é um gráfico proposto para a detecção de valores discrepantes (*outliers*), que são aqueles valores muito diferentes do restante do conjunto de dados.

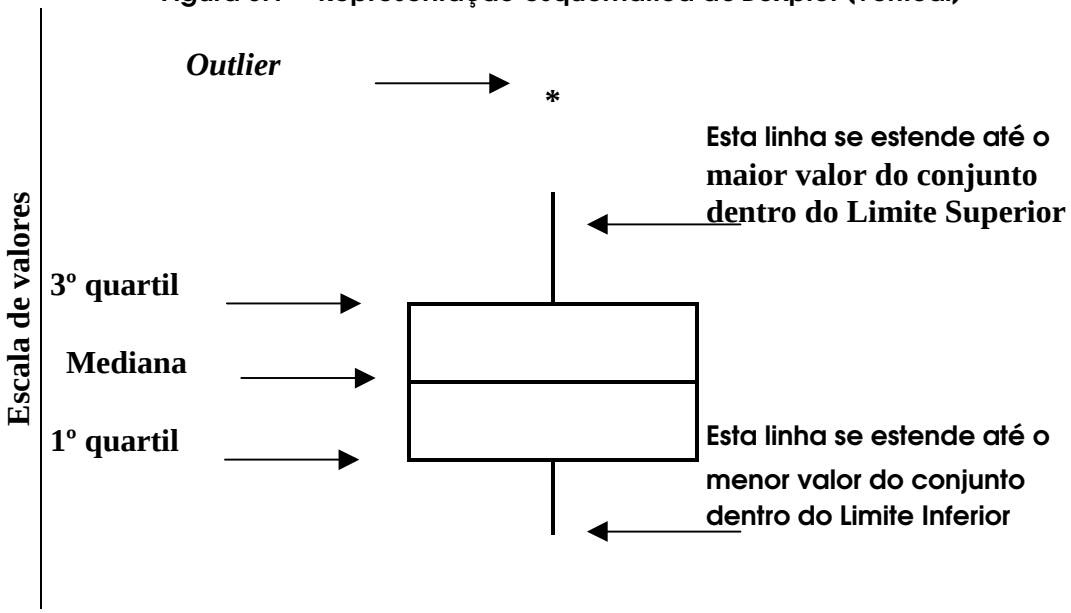
Esses valores discrepantes podem representar erros no processo de coleta ou de processamento dos dados, e, nesse caso, devem ser corrigidos ou excluídos do banco de dados.

No entanto, os *outliers* podem ser valores corretos, que, por alguma razão, são muito diferentes dos demais valores. Nesse caso, a análise desses dados deve ser cuidadosa, pois, como sabemos, algumas estatísticas descritivas, como a média e o desvio-padrão, são influenciadas por valores extremos.

Na construção do *Boxplot*, utilizamos alguns percentis (mediana, primeiro e terceiro quartis), que são pouco influenciados por valores extremos. Além disso, precisamos saber quais são os valores mínimo e máximo do conjunto de dados.

O *Boxplot* é constituído por uma caixa atravessada por uma linha, construído usando um eixo com uma escala de valores, como mostra a Figura 5.1. O fundo da caixa é marcado na escala de valores na altura do primeiro quartil (Q1). O topo da caixa é marcado na altura do terceiro quartil (Q3). Uma linha é traçada dentro da caixa na altura da mediana, que não precisa estar necessariamente no meio da caixa. Como sabemos, entre o primeiro e o terceiro quartis, temos 50% dos dados. Podemos pensar, então, que essa caixa contém metade dos dados do conjunto. A altura da caixa é dada por (Q3 - Q1), que é denominada distância *interquartílica* (DQ).

Figura 5.1 – Representação esquemática do Boxplot (vertical)



Como um gráfico tem que representar todos os valores do conjunto de dados, precisamos representar os outros 50%, sendo 25% abaixo do Q1 e 25% acima do Q3. Esses valores serão representados pelas duas *linhas* que saem das extremidades da caixa. Cada uma das linhas é traçada, a partir das extremidades da caixa, até que:

- Encontre o valor máximo (linha superior) ou mínimo (linha inferior) ou
- Atinja o maior valor dentro do *limite superior* (Q3 + 1,5 x DQ), no caso da linha superior, e atinja o menor valor dentro do *limite inferior* (Q1 – 1,5 x DQ), no caso da linha inferior.

No esquema da Figura 5.1, a linha inferior do boxplot atendeu à primeira condição, encontrando o valor mínimo dos dados antes de atingir o comprimento máximo permitido (1,5 x DQ). Assim, o limite inferior do boxplot coincide com o valor mínimo.

Caso a segunda situação ocorra, os valores que ainda não foram representados devem ser devidamente marcados por um asterisco (*) em suas respectivas posições na escala de valores. Foi o que ocorreu com o valor máximo no esquema da Figura 5.1, que não conseguiu ser incluído na linha superior do gráfico. Esses valores são considerados *outliers* pelo critério do boxplot. Obviamente, o limite superior do boxplot não coincidiu com o valor máximo do conjunto de dados, que foi considerado um valor discrepante (*outlier*).

Existem outros critérios para detecção de *outliers*, dos quais falaremos adiante.

5.1. Exemplo de Construção do *Boxplot*

Para exemplificar a construção do *boxplot*, utilizaremos os dados de peso dos ursos fêmeas do Exemplo Inicial (RTE04-2001, Anexo I).

Como pode ser calculado, os valores do primeiro, segundo e terceiro quartis para o peso dos ursos fêmeas são, respectivamente, 51,8 Kg, 61,1 Kg e 75,4 Kg. Assim, os limites da caixa são marcados nos valores de Q1 e Q3, como mostra a Figura 5.2. A mediana é marcada com uma linha dentro da caixa na sua respectiva posição na escala de valores (61,1 Kg).

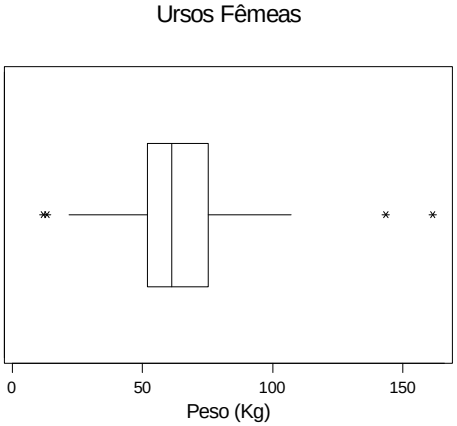
Para traçarmos as linhas laterais do *boxplot*, precisamos calcular a distância interquartílica (DQ), que é (75,4 – 51,8) = 23,6 . Desse modo, as linhas laterais podem ter comprimento máximo de 1,5 x 23,6 = 35,4. Vejamos até onde *podem ir* as linhas laterais:

- Linha da esquerda: Q1 – 1,5 x DQ = 51,8 – 1,5 x 23,6 = 51,8 – 35,4 = 16,4.
  a linha da esquerda pode se estender até o valor 16,4.
- Linha da direita: Q3 – 1,5 x DQ = 75,4 + 1,5 x 23,6 = 75,4 + 35,4 = 110,8.
  a linha da direita pode se estender até o valor 110,8.

Como existem ursos com pesos inferiores a 16,4 Kg (11,8 Kg e 13,2 Kg), eles serão representados por asteriscos colocados em suas respectivas posições. Do mesmo modo, serão representados os ursos com pesos superiores a 110,8 Kg (143,5 Kg e 161,6 Kg).

Os valores de peso para essas quatro fêmeas são considerados discrepantes (*outliers*) em relação ao grupo, pelo critério do boxplot, e devem ser investigados antes de se prosseguir a análise.

Figura 5.2 – *Boxplot* (horizontal) do peso dos ursos fêmeas.



Investigando o conjunto de dados do Exemplo Inicial, descobrimos que as duas fêmeas menos pesadas (Addy e Ness), consideradas *outliers*, são as mais jovens (8 e 9 meses, respectivamente). Já as duas fêmeas mais pesadas (Edith e Diane), também apontadas como *outliers*, não são as mais velhas (70 e 82 meses, respectivamente), mas podem estar grávidas, um dado do qual não dispomos nesse estudo. Caso essa suspeita possa ser confirmada, os dados de peso devem ser analisados sem as fêmeas Edith e Diane, pois elas estão numa situação especial. Quanto aos bebês Addy e Ness, pode-se optar por analisar dos dados de peso desse grupo de ursos separando-se os bebês dos ursos adultos.

Caso a opção seja analisar todos os ursos juntos, deve-se tomar cuidado em relação às medidas descritivas usadas, como já discutimos anteriormente.

Obviamente, se os dados dos *outliers* (ou qualquer outro urso) forem considerados erros, devem ser corrigidos se possível. Se não, devem ser imediatamente excluídos do conjunto de dados.

5.2. Outras Aplicações do Boxplot

Além da detecção de valores discrepantes, o *boxplot* pode ser muito útil na análise da distribuição dos valores de um conjunto de dados.

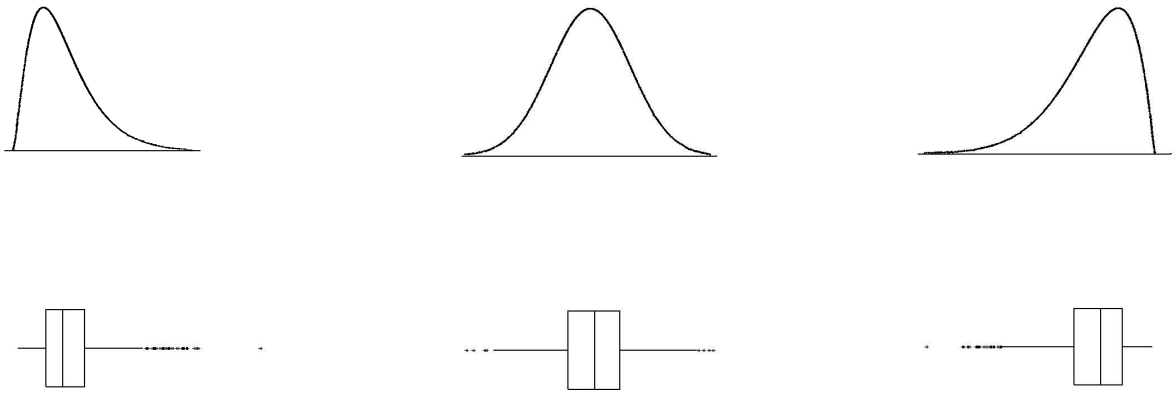
Através do *boxplot*, podemos:

- Identificar a forma da distribuição (simétrica ou assimétrica);
- Avaliar e comparar a tendência central (mediana) de dois ou mais conjuntos de dados;
- Comparar a variabilidade de dois ou mais conjuntos de dados.

Para avaliar a forma da distribuição, devemos observar o deslocamento da caixa em relação a linha do *boxplot* (Figura 5.3). Lembrando que a caixa do boxplot contém 50% dos dados, o seu deslocamento na linha nos informa onde estão concentrados os dados.

Se a caixa está mais deslocada para um dos lados da linha, significa que metade dos dados estão concentrados naquele lado da escala de valores e, assim, a distribuição é assimétrica.
Se a caixa está praticamente no meio da linha, dividindo-a em duas partes iguais, é distribuição será considerada simétrica.

Figura 5.3 – Representação esquemática das formas básicas de uma distribuição de freqüências utilizando-se histogramas e boxplots.



Assimétrica
(concentração à esquerda)

Simétrica

Assimétrica
(concentração à direita)

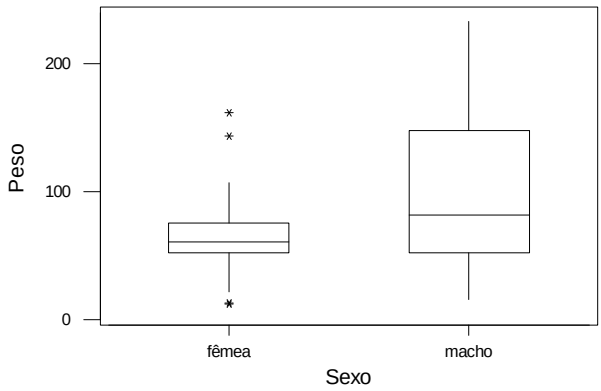
Observando a Figura 5.4, notamos que a distribuição do peso dos ursos fêmeas pode ser considerada simétrica em torno do valor mediano do peso (61,1Kg), enquanto a distribuição do peso dos ursos machos é assimétrica com concentração à esquerda (valores menores).

Ao avaliar e comparar a variabilidade de conjunto de dados através do boxplot, devemos observar a largura (ou altura) das caixas. No caso do peso dos ursos machos e fêmeas, notamos que a altura da caixa das fêmeas é bem menor do que a altura da caixa dos machos, indicando que as fêmeas são mais homogêneas quanto ao peso, apesar de contar com quatro valores atípicos. De fato, metade dos dados do conjunto de fêmeas ocupa um espaço bem menor na escala de valores do que o espaço ocupado por metade do conjunto de machos.

Finalmente, através da comparação dos boxplots para os pesos dos ursos machos e fêmeas, podemos concluir que os ursos fêmeas possuem, em geral, pesos menores e mais homogêneos do que o peso dos ursos machos.

O fato de não termos detectado o aparecimento de *outliers* no grupo dos ursos machos pode ser devido a sua grande variabilidade (altura da caixa), pois sabemos que o comprimento máximo das linhas do boxplot depende do valor dessa altura (DQ). Quanto maior o DQ, maior a possibilidade de que os valores extremos sejam incluídos na linha e, assim, não sejam considerados como *outliers*.

Figura 5.4 – *Boxplot* do peso (em Kg) dos ursos fêmeas e machos



5.3. Vantagens e Desvantagem do Boxplot

Além de permitir analisar a forma da distribuição de freqüências de um conjunto de valores, assim como a sua variabilidade e tendência central, o *boxplot* é uma forma mais prática de **comparação** entre dois ou mais grupos. Podemos representar vários *boxplots* numa mesma figura, enquanto isso não é possível quando utilizamos histogramas, por exemplo.

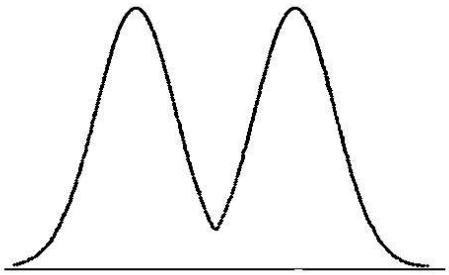
Outra vantagem do *boxplot* é que ele pode ser feito para um número reduzido de dados, enquanto histogramas (ou ogivas) não são recomendados quando o conjunto de dados é pequeno.

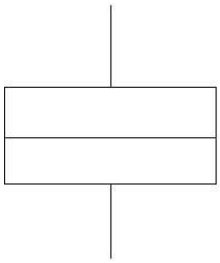
Apesar de suas muitas vantagens, o *boxplot* tem uma restrição: não deve ser usado quando a distribuição de freqüências dos dados tiver mais de uma classe modal, ou seja, mais de um pico. O uso do boxplot nesse caso esconderá essa característica da distribuição (Figura 5.5). Nessa figura, notamos como o *boxplot* mascara o caráter bimodal da distribuição através de uma falsa simetria em torno da mediana dos dados, que, como sabemos, não é melhor medida de tendência central no caso de distribuições multimodais.

Desse modo, é recomendável que se faça uma investigação sobre esse aspecto antes da opção pelo *boxplot*. Um diagrama de ramo-e-folhas ou de pontos podem ser úteis nessa tarefa[5].

Figura 5.5 – Histograma e boxplot para uma distribuição de freqüências bimodal.

[5] Na seção sobre histogramas (RTE04/2001), vimos que a distribuição do peso dos ursos machos é bimodal, com valores mais concentrados em torno da moda menor (Figura 4.13). Isto é mascarado pelo *boxplot* da Figura 5.4 através de uma indicação de assimetria à esquerda. Essa assimetria não é de todo falsa, mas a existência de dois grupos (menos pesados e mais pesados) é escondida pelo boxplot. Nesse caso, o gráfico mais indicado para a comparação dos dois grupos seria o ramo-e-folhas, devido ao pequeno número de fêmeas. A título de exemplo de comparação entre *boxplots*, manteremos o *boxplot* para os ursos machos da Figura 5.4 .





5.4. Outros Exemplos de Construção de *Boxplot*

Exemplo 5.2: Ramo-e-folhas para renda mensal (em salário mínimos) de chefes de famílias em 100 domicílios de uma comunidade (dados fictícios):

			Acumulado
(4)	**3**	1 2 5 9	4
(18)	**4**	0 0 0 3 3 4 4 5 5 5 5 5 6 6 6 7 7 8 8	22
(28)	**5**	0 0 0 1 1 2 2 3 3 3 3 4 4 4 5 5 5 5 5 5 6 6 6 7 8 8 9 9	50
(28)	**6**	0 0 0 1 1 1 1 1 1 2 2 2 2 2 3 3 3 4 4 4 5 5 5 5 6 6 6 7	78
(18)	**7**	1 1 1 1 1 2 2 3 3 4 5 5 5 6 6 6 7 9	96
(4)	**8**	0 1 2 8	100

Legenda: 3 | 1 = 3,1 salários mínimos

Dados necessários à construção do gráfico *Boxplot*:

Min = 3,1

Q_1 = 5,0

Mediana = 5,9

Q_3 = 6,6

Max = 8,8

Distância Interquartílica $DQ = Q_3 - Q_1 = 6,6 - 5,0 = 1,6$

$1,5DQ = 1,5(1,6) =$ **2,4** (comprimento máximo da linhas)

Limite Inferior $Q_1 - 1,5DQ = 5,0 - 2,4 =$ **2,6**

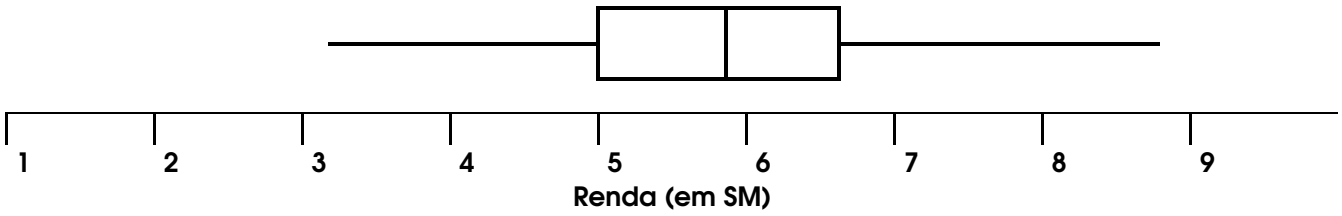
$2,6 < Min \Rightarrow$ o Min está dentro do limite inferior

o Min será o final da linha esquerda

Limite Superior $Q_3 + 1,5DQ = 6,6 + 2,4 =$ **9,0**

$9,0 > Max \Rightarrow$ o Max está dentro do limite superior

o Max será o final



Como podemos notar no ramo-e-folhas, a distribuição de freqüências da renda das famílias dessa comunidade é bastante simétrica em torno do salário mediano. O boxplot também reflete essa característica, posicionando sua caixa no meio da linha.

Nessa comunidade, metade das famílias recebem de 5 a 6,6 salários-mínimos e apenas 25% recebem mais de 6,6 salários-mínimos.

Exemplo 5.3: Boxplot para detecção de valores discrepantes

Ramo-e-folhas para renda mensal (em salário mínimos) de chefes de famílias em 101 domicílios de uma outra comunidade (dados fictícios):

Ramo-e-folhas

(18)	1	1 2 5 5 5 5 6 6 6 6 6 6 6 7 7 8 8 9
(35)	2	0 1 1 1 2 2 3 3 4 4 4 4 6 7 8 8 9
(50)	3	0 0 2 2 2 2 3 3 4 5 6 7 8 9 9
(64)	4	0 1 2 3 4 46 6 6 6 7 7 8 9
(73)	5	0 0 1 2 5 5 6 8 9
(82)	6	0 0 1 2 4 4 5 6 7
(90)	7	0 0 1 2 5 5 8 8
(96)	8	0 1 1 2 4 9
(99)	9	1 1 7
(100)	10	8
(100)	11	
(101)	12	2

Legenda: 1|1 = 1,1 salários mínimos.

Dados necessários à construção do gráfico Boxplot:

Min =	**1,1**
Q_1:	25% de 101 = 25,3 (arredonda para cima: 26). O Q_1 é 26º valor **Q_1 = 2,3**
Mediana	50% de 101 = 50,5 (arredonda para cima: 51) A mediana é o 51º valor **Mediana = 4,0**
Q_3:	75% de 101 = 75,8 (arredonda para cima: 66). O Q_3 é 76º valor **Q_3 = 6,1**
Max =	**12,2**

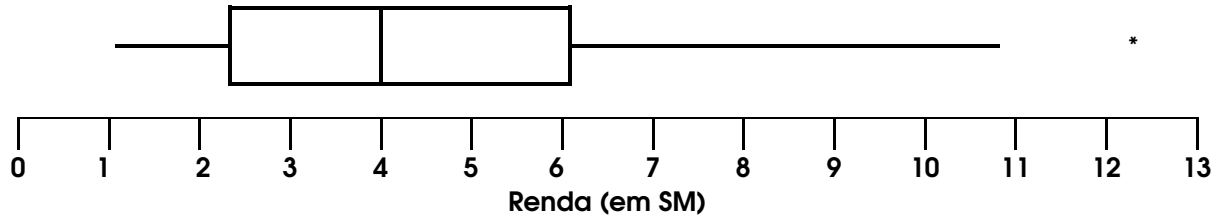
$DQ = Q_3 - Q_1 = 6,1 - 2,3 = 3,8$

Comprimento máximo das linhas : 1,5 x DQ = 1,5 x (3,8) = 5,7

Limite inferior: $Q_1 - 1,5DQ = 2,3 - 5,7 = -3,4$ (-3,4 é menor do que o Min; o Min=1,1 será o final da linha esquerda)

Limite Superior: $Q_3 + 1,5DQ = 6,1 + 5,7 = 11,8$ (o maior valor dentro do limite superior é 10,8, que será o final da linha direita)

O valor máximo está fora do limite superior é um *outlier* e será marcado com um asterisco.

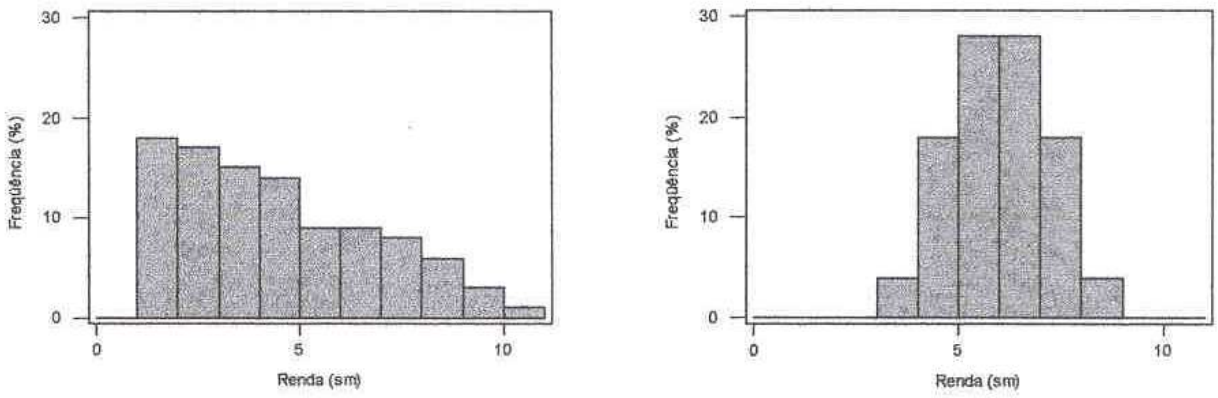


A distribuição de freqüências da renda familiar nessa comunidade é claramente assimétrica com concentração à esquerda, o que pode ser visto pelo ramo-e-folhas e pelo deslocamento da caixa do boxplot em direção ao lado esquerda da escala de valores. Do total de famílias dessa comunidade, metade ganha até 4 salários-mínimos e apenas 25% ganha mais do que 6,1 salários-mínimos.

6. Comparação Gráfica de Conjuntos de Dados

Com a introdução do *boxplot*, somamos cinco alternativas para a representação gráfica de um conjunto de dados quantitativos: histograma, ramo-e-folhas, ogiva, diagrama de pontos e boxplot. Em seções anteriores, já discutimos os usos, as vantagens e desvantagens de cada uma dessas alternativas. Nesta seção, discutiremos a representação gráfica de mais de um conjunto de dados com o objetivo de compará-los. Para isso, usaremos os dados de renda familiar da comunidade do exemplo 5.2 e da comunidade do exemplo 5.3, sem a família de maior renda.

Histograma



Ao escolhermos o histograma, necessitaremos de uma figura para cada conjunto de dados. Quando a quantidade de grupos a serem comparados é maior do que dois, essa alternativa de representação gráfica se torna menos econômica e só deve ser escolhida, se, por outras razões, o histograma for a melhor maneira de representar graficamente os dados.

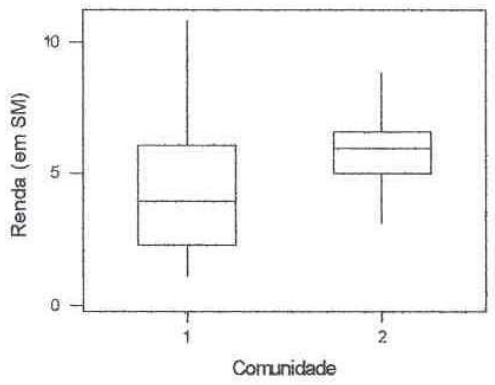
Ramo-e-folhas

	1	1 2 5 5 5 5 6 6 6 6 6 6 7 7 8 8 9
	2	0 1 1 1 2 2 3 3 4 4 4 4 6 7 8 8 9
9 5 2 1	**3**	0 0 2 2 2 2 3 3 4 5 6 7 8 9 9
8 8 7 7 6 6 6 5 5 5 5 5 4 4 3 3 0 0 0	**4**	0 1 2 3 4 4 6 6 6 6 7 7 8 9
9 9 8 8 7 6 6 6 5 5 5 5 5 5 4 4 4 3 3 3 3 2 2 1 1 0 0 0	**5**	0 0 1 2 5 5 6 8 9
7 6 6 6 5 5 5 5 4 4 4 3 3 3 2 2 2 2 2 1 1 1 1 1 1 0 0 0	**6**	0 0 1 2 4 4 5 6 7
9 7 6 6 6 5 5 5 4 3 3 2 2 1 1 1 1 1	**7**	0 0 1 2 5 5 8 8
8 2 1 0	**8**	0 1 1 2 4 9
	9	1 1 7
	10	8

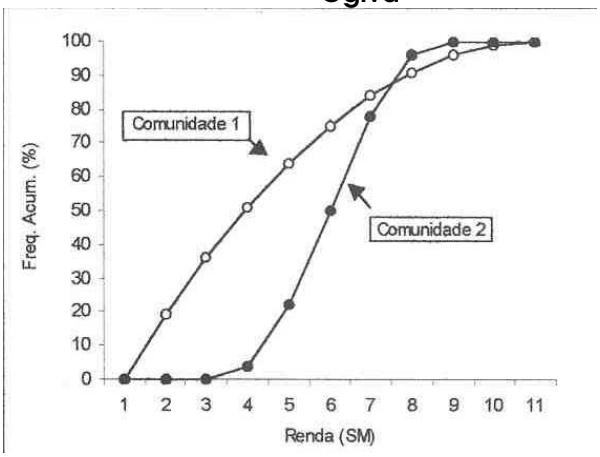
Legenda: 1 1 1 2 : 1,2 salários-mínimos
1 1 3 1 : 3,1 salários-mínimos

O ramo-e-folhas lado-a-lado é uma boa alternativa quando se tem pares de grupos a serem comparados. No entanto, torna-se também pouco econômica quando o número de pares é grande, pois necessitaremos de uma figura para cada par.

Boxplot



Ogiva



O *boxplot* e a ogiva são alternativas que permitem a representação de vários grupos numa mesma figura, facilitando a comparação. À medida que o número de grupos a serem representados aumenta, o *boxplot* torna-se uma alternativa melhor do que a ogiva, onde um grande número de linhas pode dificultar a interpretação do gráfico. No *boxplot*, a adição de grupos significa a adição de caixas numa mesma figura, o que requer apenas a diminuição da largura dessas caixas (no *boxplot* vertical).

Além dessas considerações, a escolha da representação gráfica mais adequada deve considerar os objetivos da análise descritiva, no sentido de facilitar a compreensão e interpretação dos resultados.

Referências Bibliográficas

Freund, J. E., Simon, G. A. (2000). *Estatística Aplicada*, 9ª edição, Editora Bookman.

Reis, E. A., Reis, I. A. (2001). *Análise Descritiva de Dados- Tabelas e Gráficos*. (Relatório Técnico RTE04/2001) Departamento de Estatística - ICEx – Universidade Federal de Minas Gerais.

Triola, M. F. (1996). *Introdução à Estatística*, 7ª edição, Editora LTC.