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

Análise Descritiva de Dados

Edna Afonso Reis
Ilka Afonso Reis

Primeira Edição – Junho/2002

ÍNDICE

I- Tabelas e Gráficos

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.

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

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

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.

Ao mesmo tempo em que o uso das ferramentas estatísticas vem crescendo, aumenta também o abuso de tais ferramentas. É muito comum vermos em jornais e revistas, até mesmo em periódicos científicos, gráficos – voluntariamente ou intencionalmente – enganosos e estatísticas obscuras para justificar argumentos polêmicos.

2. Coleta e Armazenamento de Dados

Exemplo Inicial: Ursos Marrons

Pesquisadores do Instituto Amigos do Urso têm estudado o desenvolvimento dos ursos marrons selvagens que vivem em certa floresta do Canadá. O objetivo do projeto é estudar algumas características dos ursos, tais como seu peso e altura, ao longo da vida desses animais. A ficha de coleta de dados, representada na Figura 2.1, mostra as características que serão estudadas na primeira fase do projeto. Na primeira parte do estudo, 97 ursos foram identificados (por nome), pesados e medidos. Os dados foram coletados através do preenchimento da ficha de coleta mostrada na Figura 2.1.

Para que os ursos possam ser identificados, medidos e avaliados, os pesquisadores precisam anestesiá-los. Mesmo assim, medidas como a do peso são difíceis de serem feitas (qual será o tamanho de uma balança para pesar ursos?). Desse modo, os pesquisadores gostariam também de encontrar uma maneira de estimar o peso do urso através de outra medida mais fácil de obter, como uma medida de comprimento, por exemplo, (altura, circunferência do tórax, etc.). Nesse caso, só seria necessária uma grande fita métrica, o que facilitaria muito a coleta de dados das próximas fases do projeto.

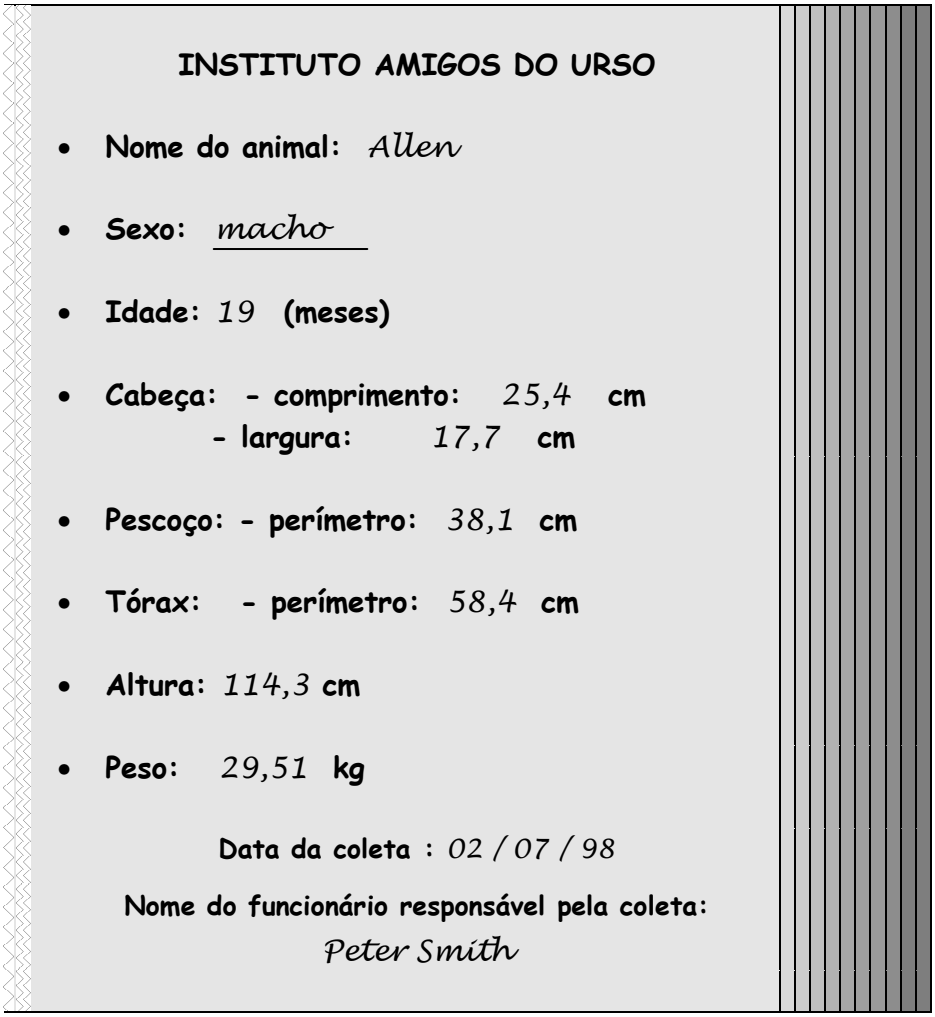
Geralmente, as coletas de dados são feitas através do preenchimento de fichas pelo pesquisador e/ou através de resposta a questionários (o que não foi o caso dos ursos ☺). Alguns dados são coletados através de medições (altura, peso, pressão sangüínea, etc.), enquanto outros são coletados através de avaliações (sexo, cor, raça, espécie, etc.).

Depois de coletados, os dados devem ser armazenados e sistematizados numa planilha de dados, como mostra a Figura 2.2. Hoje em dia, essas planilhas são digitais e essa é a maneira de realizar a entrada dos dados num programa de computador.

A planilha de dados é composta por linhas e colunas. Cada linha contém os dados de um

urso (elemento), ou seja de uma ficha de coleta. As características (variáveis) são dispostos em colunas. Assim, a planilha de dados contém um número de linhas igual a número de participantes do estudo e um número de colunas igual ao número de variáveis sendo estudadas.

Figura 2.1 – Ficha de coleta de dados dos ursos marrons.



INSTITUTO AMIGOS DO URSO

- **Nome do animal:** *Allen*
- **Sexo:** *macho*
- **Idade:** *19* **(meses)**
- **Cabeça: - comprimento:** *25,4* **cm**
  - largura: *17,7* **cm**
- **Pescoço: - perímetro:** *38,1* **cm**
- **Tórax: - perímetro:** *58,4* **cm**
- **Altura:** *114,3* **cm**
- **Peso:** *29,51* **kg**

Data da coleta : *02 / 07 / 98*

Nome do funcionário responsável pela coleta:
Peter Smith

Figura 2.2 – Representação parcial da planilha de dados do exemplo dos ursos.

V A RI Á V E I S →

ELEMENTOS ↓	**Nome**	**Mês Obs.**	**Idade**	**Sexo**	**Cabeça Comp.**	**Cabeça Larg.**	**Pescoço Peri.**	**Altura**	**Tórax Peri.**	**Peso**
1	Allen	jul	19	macho	25,4	12,7	38,1	114,3	58,4	29,5
2	Berta	jul	19	fêmea	27,9	16,5	50,8	120,7	61,0	31,8
3	Clyde	jul	19	macho	27,9	14,0	40,6	134,6	66,0	36,3
4	Doc	jul	55	macho	41,9	22,9	71,1	171,5	114,3	156,2
5	Quincy	set	81	macho	39,4	20,3	78,7	182,9	137,2	188,9
6	Kooch	out	*	macho	40,6	20,3	81,3	195,6	132,1	196,1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
93	Sara	ago	*	fêmea	30,5	12,7	45,7	142,2	82,6	51,8
94	Lou	ago	*	macho	30,5	14,0	38,1	129,5	61,0	37,2
95	Molly	ago	*	fêmea	33,0	15,2	55,9	154,9	101,6	104,4
96	Graham	jul	*	macho	30,5	10,2	44,5	149,9	72,4	58,1
97	Jeffrey	jul	*	macho	34,3	15,2	50,8	157,5	82,6	70,8

3. Tipos de Variáveis

Variável é a característica de interesse que é medida em cada indivíduo da amostra ou população. Como o nome diz, seus valores *variam* de indivíduo para indivíduo. As variáveis podem ter valores numéricos ou não numéricos.

- **Variáveis Quantitativas:** são as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser *contínuas* ou *discretas*.

  Variáveis contínuas: características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores não-inteiros (com casas decimais) fazem sentido. Usualmente devem ser medidas através de algum instrumento. Exemplos: peso (balança), altura (régua), tempo (relógio), pressão arterial, idade.

  Variáveis discretas: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente, são o resultado de contagens. Exemplos: número de filhos, número de bactérias por litro de leite, número de cigarros fumados por dia.

- **Variáveis Qualitativas (ou categóricas):** são as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser *nominais* ou *ordinais*.

  Variável nominais: não existe ordenação entre as categorias.
  Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.

  Variáveis ordinais: existe uma ordenação entre as categorias.
  Exemplos: escolaridade (1º, 2º, 3º graus), estágio da doença (inicial, intermediário, terminal), mês de observação (janeiro, fevereiro,..., dezembro).

Uma variável originalmente quantitativa pode ser coletada de forma qualitativa. Por exemplo, a variável *idade*, medida em anos completos, é quantitativa (contínua); mas, se for informada apenas a faixa etária (0 a 5 anos, 6 a 10 anos, etc...), é qualitativa (ordinal). Outro exemplo é o *peso* dos lutadores de boxe, uma variável quantitativa (contínua) se trabalhamos com o valor obtido na balança, mas qualitativa (ordinal) se o classificarmos nas categorias do boxe (peso-pena, peso-leve, peso-pesado, etc.).

Outro ponto importante é que nem sempre uma variável representada por números é quantitativa. O número do telefone de uma pessoa, o número da casa, o número de sua identidade. Às vezes o sexo do indivíduo é registrado na planilha de dados como 1 se macho e 2 se fêmea, por exemplo. Isto não significa que a variável *sexo* passou a ser quantitativa !

Exemplo dos ursos marrons (continuação).

No conjunto de dados ursos marrons, são qualitativas as variáveis *sexo* (nominal) e *mês da observação* (ordinal); são quantitativas contínuas as demais: *idade, comprimento da cabeça, largura da cabeça, perímetro do pescoço, perímetro do tórax, altura* e *peso*.

4. Estudando a Distribuição de Freqüências de uma Variável

Como já sabemos, as variáveis de um estudo dividem-se em quatro tipos: qualitativas (nominais e ordinais) e quantitativas (discretas e contínuas). Os dados gerados por esses tipos de variáveis são de naturezas diferentes e devem receber tratamentos diferentes. Portanto, vamos estudar as ferramentas - tabelas e gráficos - mais adequados para cada tipo de dados, separadamente.

4.1. Variáveis Qualitativas – Nominais e Ordinais

Iniciaremos essa apresentação com os dados de natureza qualitativa, que são os mais fáceis de tratar do ponto de vista da análise descritiva.

No exemplo dos ursos, uma das duas variáveis qualitativas presentes é o *sexo* dos animais. Para organizar os dados provenientes de uma variável qualitativa, é usual fazer uma **tabela de freqüências**, como a Tabela 4.1, onde estão apresentadas as freqüências com que ocorre cada um dos sexos no total dos 97 ursos observados. Cada categoria da variável *sexo* (feminino, masculino) é representada numa linha da tabela. Há uma coluna com as contagens de ursos em cada categoria (freqüência absoluta) e outra com os percentuais que essas contagens representam no total de ursos (freqüência relativa). Esse tipo de tabela representa a *distribuição de freqüências dos ursos* segundo a variável *sexo*.

Como a variável *sexo* é qualitativa nominal, ou seja, não há uma ordem natural em suas categorias, a ordem das linhas da tabela pode ser qualquer uma. É comum a disposição das linhas pela ordem decrescente das freqüências das classes.

Tabela 4.1: Distribuição de freqüências dos ursos segundo sexo.

Sexo	Freqüência Absoluta	Freqüência Relativa (%)
Feminino	35	36,1
Masculino	62	63,9
Total	97	100

Quando a variável tabelada for do tipo qualitativa ordinal, as linhas da tabela de freqüências devem ser dispostas na ordem existente para as categorias. A Tabela 4.2 mostra a distribuição de freqüências dos ursos segundo o mês de observação, que é uma variável qualitativa ordinal. Nesse caso, podemos acrescentar mais duas colunas com *as freqüências acumuladas* (absoluta e relativa), que mostram, para cada mês, a freqüência de ursos observados *até* aquele mês. Por exemplo, até o mês de julho, foram observados 31 ursos, o que representa 32,0% do total de ursos estudados.

Tabela 4.2: Distribuição de freqüências dos ursos segundo mês de observação.

	Freqüências Simples		Freqüências Acumuladas	
Mês de Observação	Freqüência Absoluta	Freqüência Relativa (%)	Freqüência Absoluta Acumulada	Freqüência Relativa Acumulada (%)
Abril	8	8,3	8	8,3
Maio	6	6,2	14	14,5
Junho	6	6,2	20	20,7
Julho	11	11,3	31	32,0
Agosto	23	23,7	54	55,7
Setembro	20	20,6	74	76,3
Outubro	14	14,4	88	90,7
Novembro	9	9,3	97	100
Total	97	100	-----	-----

Note que as freqüências acumuladas *não fazem sentido* em distribuição de freqüências de variáveis para as quais não existe uma ordem natural nas categorias, as *qualitativas nominais*.

A visualização da distribuição de freqüências de uma variável fica mais fácil se fizermos um gráfico a partir da tabela de freqüências. Existem vários tipos de gráficos, dependendo do tipo de variável a ser representada. Para as variáveis do tipo qualitativas, abordaremos dois tipos de gráficos: os de setores e os de barras.

Os **gráficos de setores**, mais conhecidos como gráficos de pizza ou torta, são construídos dividindo-se um círculo (pizza) em setores (fatias), um para cada categoria, que serão proporcionais à freqüência daquela categoria.

A Figura 4.1 mostra um gráfico de setores para a variável *sexo*, construído a partir da Tabela 4.1. Através desse gráfico, fica mais fácil perceber que os ursos machos são a grande maioria dos ursos estudados. Como esse gráfico contém todas as informações da Tabela 4.1, pode substituí-la com a vantagem de tornar análise dessa variável mais agradável.

Quando houver mais de duas categorias de uma variável nominal, a disposição no gráfico de setores deve ser pela ordem decrescente das freqüências, no sentido horário. A categoria "outros", quando existir, deve ser sempre a última, mesmo não seja a de menor freqüência.

As vantagens da representação gráfica das distribuições de freqüências ficam ainda mais evidentes quando há a necessidade de comparar vários grupos com relação à variáveis que possuem muitas categorias, como veremos mais adiante.

Uma alternativa ao gráfico de setores é o **gráfico de barras** (colunas) como o da Figura 4.2. Ao invés de dividirmos um círculo, dividimos uma barra. Note que, em ambos os gráficos, as freqüências relativas das categorias devem somar 100%. Aliás, esse é a idéia dos gráficos: mostrar como se dá a divisão (distribuição) do total de elementos (100%) em partes (fatias).

Figura 4.1 – Gráfico de setores para a variável *Sexo*.

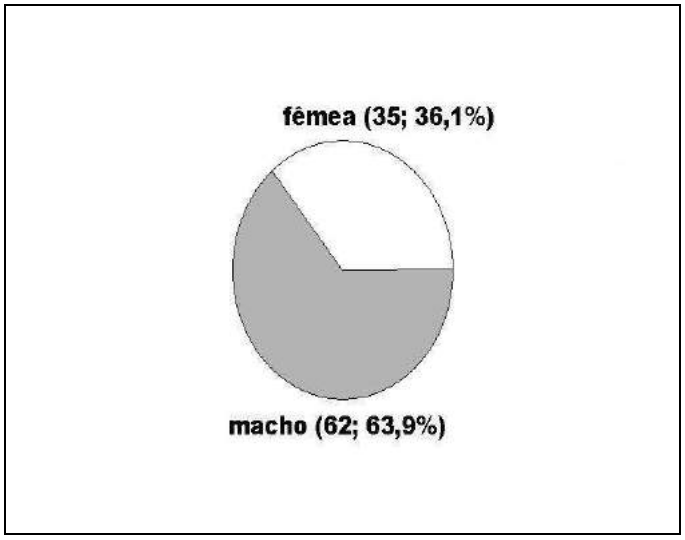
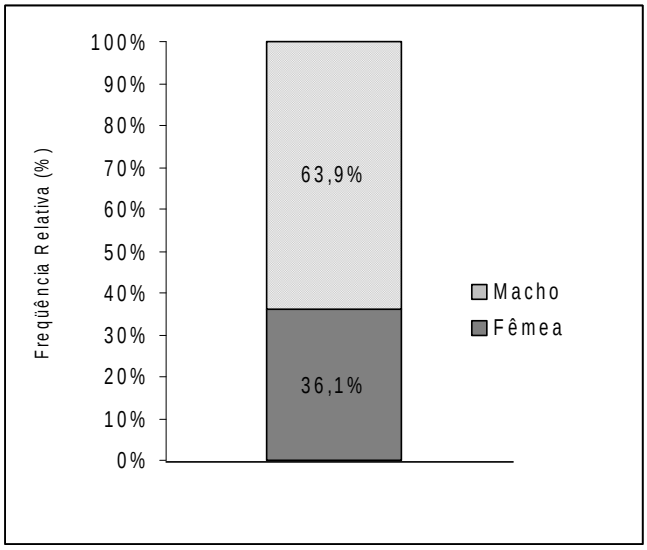


Figura 4.2 – Gráfico de barras para a variável *Sexo*.



Uma situação diferente ocorre quando desejamos comparar a distribuição de freqüências de uma mesma variável em vários grupos, como por exemplo, a freqüência de ursos marrons em quatro regiões de um país. Se quisermos usar o gráfico de setores para fazer essa comparação, devemos fazer quatro gráficos, um para cada região, com duas fatias cada um (ursos marrons e ursos não marrons). Uma alternativa é a construção de um gráfico de colunas (barras) como os gráficos das figuras 4.3 e 4.4, onde há uma barra para cada região representando a freqüência de ursos marrons naquela região. Além de economizar espaço na apresentação, permite que as comparações sejam feitas de maneira mais rápida (tente fazer essa comparação usando quatro "pizzas" e comprove!!)

Note que a soma das freqüências relativas de ursos marrons em cada região não é 100% e nem deve ser, pois se tratam de freqüências calculadas em grupos (regiões) diferentes. A ordem dos grupos pode ser qualquer, ou aquela mais adequada para a presente análise. Freqüentemente, encontramos as barras em ordem decrescente, já antecipando nossa intuição de ordenar os grupos de acordo com sua freqüência para facilitar as comparações. Caso a variável fosse do tipo *ordinal*, a ordem das barras seria a *ordem natural* das categorias, como na tabela de freqüências.

Figura 4.3 – Gráfico de barras horizontais para a freqüência de ursos marrons em quatro regiões.

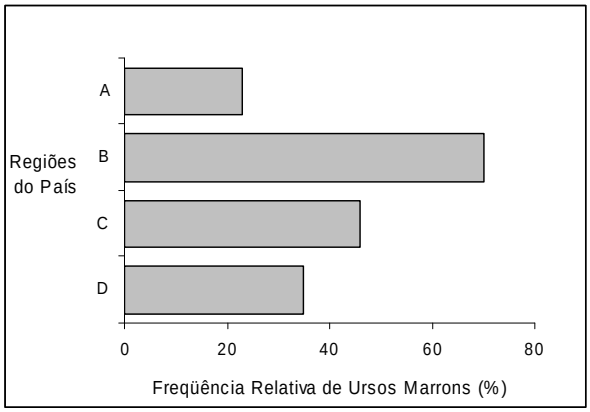
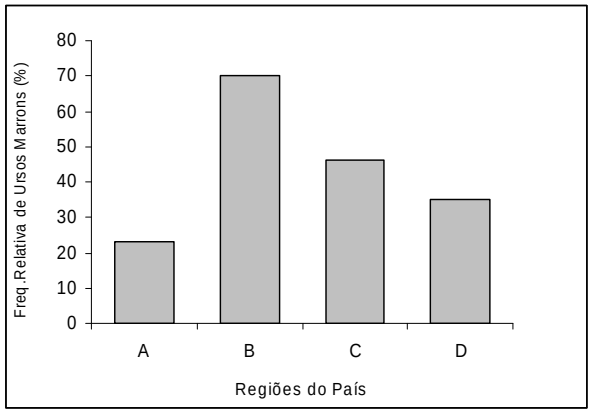


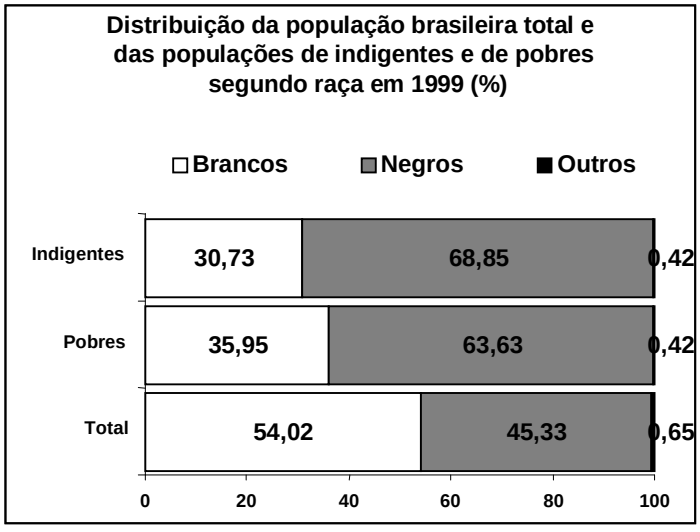
Figura 4.4 – Gráfico de barras verticais para a freqüência de ursos marrons em quatro regiões.



A Figura 4.5 mostra um gráfico de barras que pode ser usado da comparação da distribuição de freqüências de uma mesma variável em vários grupos. È também uma alternativa ao uso de vários gráficos de setores, sendo, na verdade, a junção de três gráficos com os da Figura 4.2 num só gráfico. Porém, esse tipo de gráfico só deve ser usado quando não houver muitos grupos a serem comparados e a variável em estudo não tiver muitas categorias (de preferência, só duas). No exemplo da Figura 4.5, a variável *raça* tem três categorias, mas uma delas é muito menos freqüente do que as outras duas.

Através desse gráfico, podemos observar que a população brasileira total, em 1999, dividia-se quase que igualmente entre brancos e negros, com uma pequena predominância de brancos. Porém, quando nos restringimos às classes menos favorecidas economicamente, essa situação se inverte, com uma considerável predominância de negros, principalmente na classe da população considerada indigente, indicando que a classe sócio-econômica influencia a distribuição de negros e brancos na população brasileira de 1999.

Figura 4.5 – Gráfico de barras para comparação da distribuição de freqüências de uma variável (raça) em vários grupos (indigentes, pobres e população total).



Freqüentemente, é necessário fazer comparações da distribuição de freqüências de uma variável em vários grupos simultaneamente. Nesse caso, o uso de gráficos bem escolhidos e construídos torna a tarefa muito mais fácil. Na Figura 4.6, está representada a distribuição de freqüências da reprovação segundo as variáveis sexo do aluno, período e área de estudo.

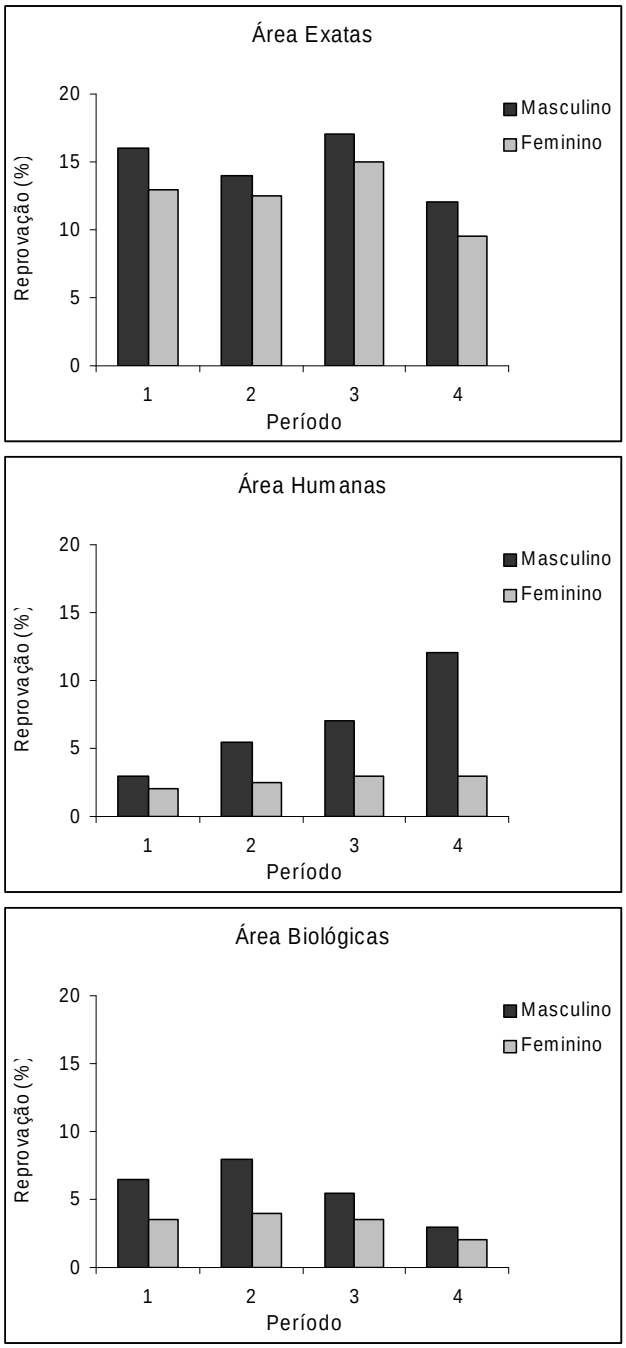
Analisando os três gráficos da Figura 4.6, podemos notar que o percentual de reprovação entre os alunos do sexo masculino é sempre maior do que o percentual de reprovação entre os alunos do sexo feminino, em todas as áreas, durante todos os períodos. A área de ciências exatas é a que possui os maiores percentuais de reprovação, em todos os períodos, nos dois sexos. Na área de ciências humanas, o percentual de reprovação entre os alunos do sexo masculino cresce com os períodos, enquanto esse percentual entre as alunas se mantém praticamente constante

durante os períodos. Na área de ciências biológicas, há uma diminuição do percentual de reprovação, a partir do segundo período, entre os alunos do dos sexos, sendo mais acentuado entre os estudantes do sexo masculino.

Chegar às conclusões colocadas no parágrafo anterior através de comparação numérica de tabelas de freqüências seria muito mais árduo do que através da comparação visual possibilitada pelo uso dos gráficos. Os gráficos são ferramentas poderosas e devem ser usadas sempre que possível.

É importante observar que a *comparação* dos três gráficos da Figura 4.6 só foi possível porque eles usam a *mesma escala*, tanto no eixo dos períodos (mesma ordem) quanto no eixo dos percentuais de reprovação (mais importante). Essa observação é válida para toda comparação entre gráficos de quaisquer tipos.

Figura 4.6: Distribuição de freqüências de reprovação segundo área, período e sexo do aluno.



Fonte: *A Evasão no Ciclo Básico da UFMG*, em Cadernos de Avaliação 3, 2000.

4.2. Variáveis Quantitativas Discretas

Quando estamos trabalhando com uma **variável discreta que assume poucos valores**, podemos dar a ela o mesmo tratamento dado às variáveis qualitativas ordinais, assumindo que cada valor é uma classe e que existe uma ordem natural nessas classes.

A Tabela 4.3 apresenta a distribuição de freqüências do número de filhos por família em uma localidade, que, nesse caso, assumiu apenas seis valores distintos.

Tabela 4.3 – Distribuição de freqüências do número de filhos por família em uma localidade (25 lares).

Número de filhos	Freqüência Absoluta	Freqüência Relativa (%)	Freqüência Relativa Acumulada (%)
0	1	4,0	4,0
1	4	16,0	20,0
2	10	40,0	60,0
3	6	24,0	84,0
4	2	8,0	92,0
5	2	8,0	100
Total	25	100	------

Analisando a Tabela 4.3, podemos perceber que as famílias mais freqüentes são as de dois filhos (40%), seguida pelas famílias de três filhos. Apenas 16% das famílias têm mais de três filhos, mas são ainda mais comuns do que famílias sem filhos.

A Figura 4.7 mostra a representação gráfica da Tabela 4.3 e a Figura 4.8 mostra a distribuição de freqüências do número de filhos por família na localidade B. Como o número de famílias estudadas em cada localidade é diferente, a freqüência utilizada em ambos os gráficos foi a relativa (em porcentagem), tornando os dois gráficos comparáveis. Comparando os dois gráficos, notamos que a localidade B tende a ter famílias menos numerosas do que a localidade A . A maior parte das famílias da localidade B (cerca de 70%) tem um ou nenhum filho.

Figura 4. 7: Distribuição de freqüências do número de filhos por família na localidade A (25 famílias).

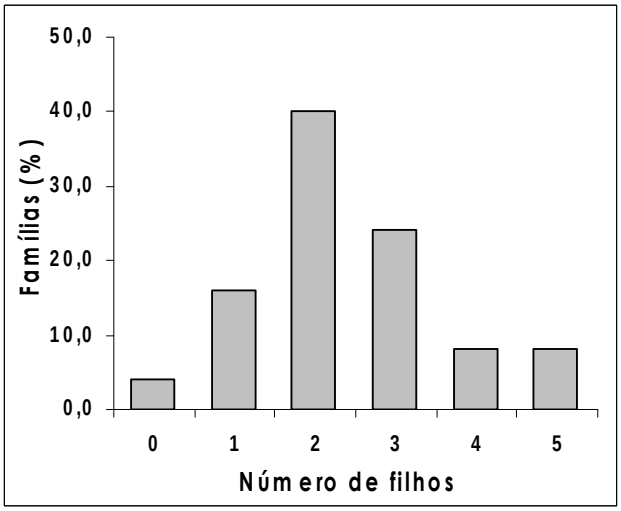
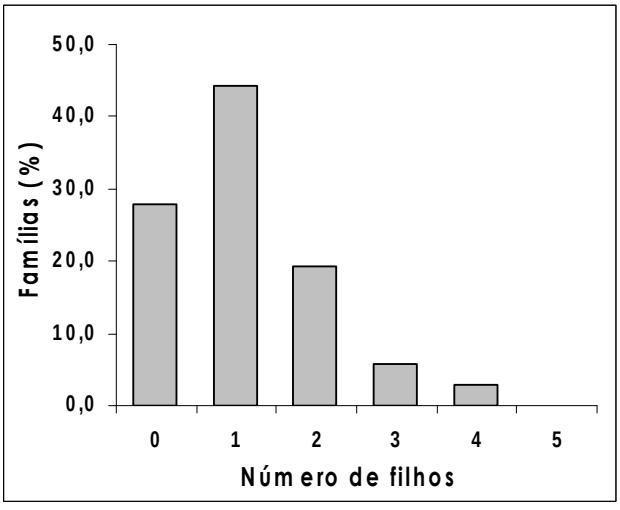


Figura 4. 8: Distribuição de freqüências do número de filhos por família na localidade B (36 famílias).



ⓘ **Importante:** Na comparação da distribuição de freqüências de uma variável entre dois ou mais grupos de tamanhos (número de observações) diferentes, devemos usar as freqüências relativas na construção do gráfico. Deve-se, também usar a mesma escala em todos os gráficos, tanto na escala vertical quanto na horizontal.

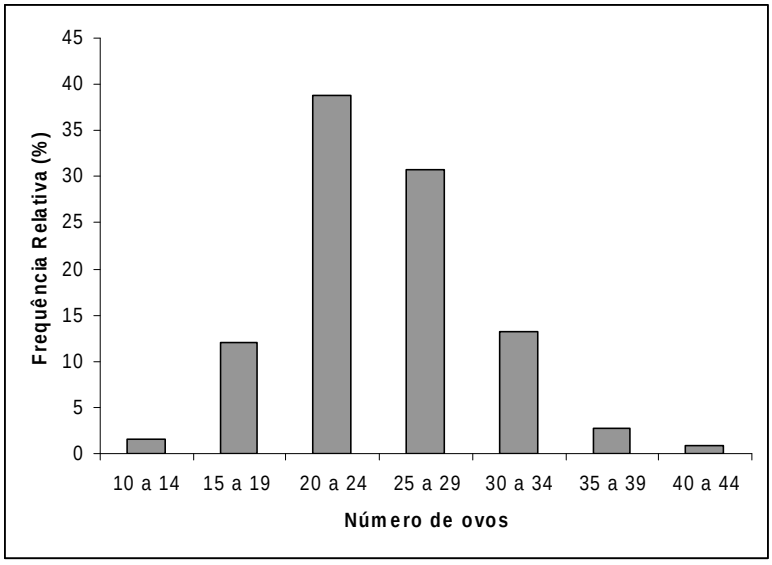
Quando trabalhamos com uma **variável discreta que pode assumir um grande número de valores distintos** como, por exemplo, o número de ovos que um inseto põe durante sua vida, a construção da tabela de freqüências e de gráficos considerando cada valor como uma categoria fica inviável. A solução é agrupar os valores em classes ao montar a tabela, como mostra a Tabela 4.4.

Tabela 4.4: Distribuição de freqüências do número de ovos postos por 250 insetos.

Número de ovos	Freqüências Simples		Freqüências Acumuladas	
	Freqüência Absoluta	Freqüência Relativa (%)	Freq. Abs. Acumulada	Freq. Rel. Acumulada (%)
10 a 14	4	1,6	4	1,6
15 a 19	30	12,0	34	13,6
20 a 24	97	38,8	131	52,4
25 a 29	77	30,8	208	83,2
30 a 34	33	13,2	241	96,4
35 a 39	7	2,8	248	99,2
40 a 44	2	0,8	250	100
Total	250	100	---	---

A Figura 4.9 mostra o gráfico da distribuição de freqüências do número de ovos postos por 250 insetos ao longo de suas vidas. Podemos perceber que o número de ovos está concentrado em torno de 20 a 24 ovos com um ligeiro deslocamento para os valores maiores.

Figura 4. 9: Distribuição de freqüências do número de ovos postos por 250 insetos.



A escolha do *número de classes* e do *tamanho das classes* depende da amplitude dos valores a serem representados (no exemplo, de 10 a 44) e da quantidade de observações no conjunto de dados. Classes muito grandes resumem demais a informação contida nos dados, pois forçam a construção de poucas classes. No exemplo dos insetos, seria como, por exemplo, construir classes do tamanho 10, o que reduziria o número de classes para quatro (Figura 4.10). Por outro lado, classes muito pequenas nos levaria a construir muitas classes, o que poderia não resumir a informação como gostaríamos. Além disso, para conjuntos de dados pequenos, podem ocorrer classes com muito poucas observações ou mesmo sem observações. Na Figura 4.11, há classes sem observações, mesmo o conjunto de dados sendo grande. Alguns autores recomendam que tabelas de freqüências (e gráficos) possuam de 5 a 15 classes, dependendo do tamanho do conjunto de dados e levando-se em consideração o que foi exposto anteriormente.

Figura 4.10: Distribuição de freqüências do número de ovos postos por 250 insetos (classes de tamanho 10).

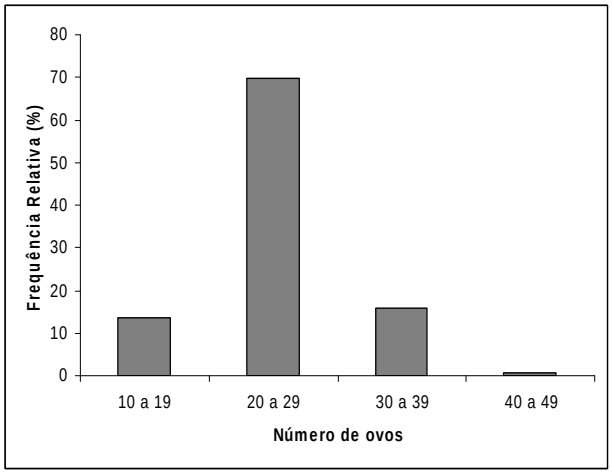
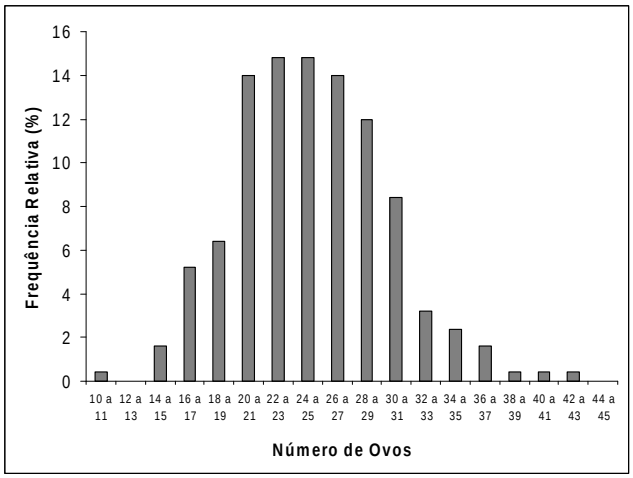


Figura 4. 11: Distribuição de freqüências do número de ovos postos por 250 insetos (classes de tamanho 2).

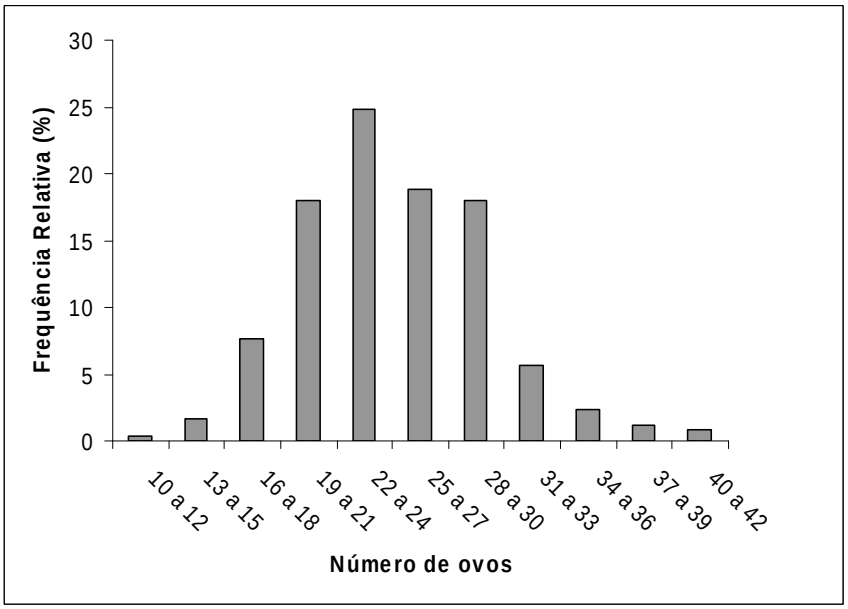


Os limites inferiores e superiores de cada classe dependem do tamanho (amplitude) de classe escolhido, que deve ser, na medida do possível, igual para todas as classes. Isso facilita a interpretação da distribuição de freqüências da variável em estudo.

Com o uso do computador na análise estatística de dados, a tarefa de construção de tabelas e gráficos ficou menos trabalhosa e menos dependente de regras rígidas. Se determinado agrupamento de classes não nos pareceu muito bom, podemos construir vários outros quase que instantaneamente e a escolha da melhor representação para a distribuição de freqüências para aquela variável fica muito mais tranqüila[1].

O gráfico da Figura 4.12, com classes de tamanho três, é uma alternativa ao gráfico da Figura 4.9.

Figura 4.12: Alternativa à distribuição de freqüências do número de ovos da Figura 4.9.



[1] Em publicações mais antigas sobre construção de tabelas de freqüências, há fórmulas para determinação do número de classes de acordo com o número de dados. Essas fórmulas eram úteis, pois a construção dos gráficos era muito custosa sem o auxílio do computador. Hoje em dia, essas fórmulas só são (ou deveriam ser) usadas pelos programas de computador, que precisam de fórmulas na geração de tabelas e gráficos no modo automático. Esse procedimento é aconselhável como uma primeira visualização da distribuição de freqüências de uma variável.

4.3. Variáveis Quantitativas Contínuas

Quando a variável em estudo é do tipo contínua, que assume muitos valores distintos, o agrupamento dos dados em classes será sempre necessário na construção das tabelas de freqüências. A Tabela 4.5 apresenta a distribuição de freqüências para o peso dos ursos machos.

Tabela 4.5: Distribuição de freqüências dos ursos machos segundo peso.

Peso (kg)	Freqüência Absoluta	Freqüência Relativa (%)	Freq. Abs. Acumulada	Freq. Rel. Acumulada (%)
0 |- 25	3	4,8	3	4,8
25 |- 50	11	17,7	14	22,6
50 |- 75	15	24,2	29	46,8
75 |- 100	11	17,7	40	64,5
100 |- 125	3	4,8	43	69,4
125 |- 150	4	6,5	47	75,8
150 |- 175	8	12,9	55	88,7
175 |- 200	5	8,1	60	96,8
200 |- 225	1	1,6	61	98,4
225 |- 250	1	1,6	62	100
Total	62	100	-	-

Os limites das classes são representados de modo diferente daquele usado nas tabelas para variáveis discretas: o limite superior de uma classe é igual ao limite inferior da classe seguinte. Mas, afinal, onde ele está incluído? O símbolo |- resolve essa questão. Na segunda classe (25 |- 50), por exemplo, estão incluídos todos os ursos com peso de 25,0 a 49,9 kg. Os ursos que porventura pesarem exatos 50,0 kg serão incluídos na classe seguinte. Ou seja, ursos com pesos maiores ou iguais a 25 kg e menores do que 50 kg.

A construção das classes da tabela de freqüências é feita de modo a facilitar a interpretação da distribuição de freqüências, como discutido anteriormente. Geralmente, usamos tamanhos e limites de classe múltiplos de 5 ou 10. Isso ocorre porque estamos acostumados a pensar no nosso sistema numérico, que é o decimal. Porém, nada nos impede de construirmos classes de outros tamanhos (inteiros ou fracionários) desde que isso facilite nossa visualização e interpretação da distribuição de freqüências da variável em estudo. Mesmo assim, para os que não se sentem à vontade com tamanha liberdade, disponibilizamos, no Anexo 2, os passos a serem seguidos na construção de uma tabela de freqüências para variáveis contínuas.

A representação gráfica da distribuição de freqüências de uma variável contínua é feita através de um gráfico chamado **histograma**, mostrado nas figuras 4.13 e 4.14. O histograma nada mais é do que o gráfico de barras verticais, porém construído com as barras unidas, devido ao caráter contínuo dos valores da variável.

Figura 4.13: Histograma para a distribuição de freqüências (absolutas) do peso dos ursos machos.

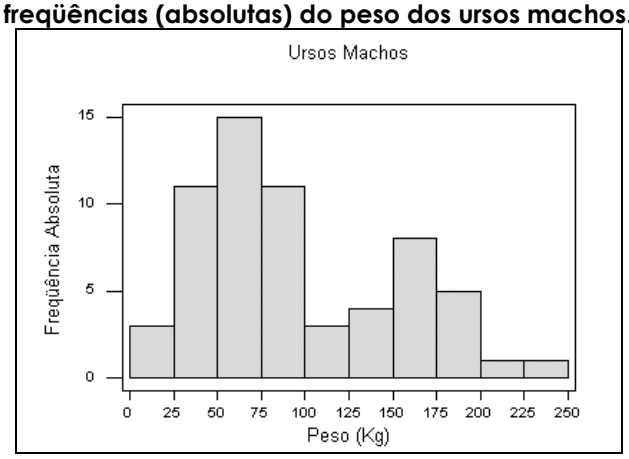
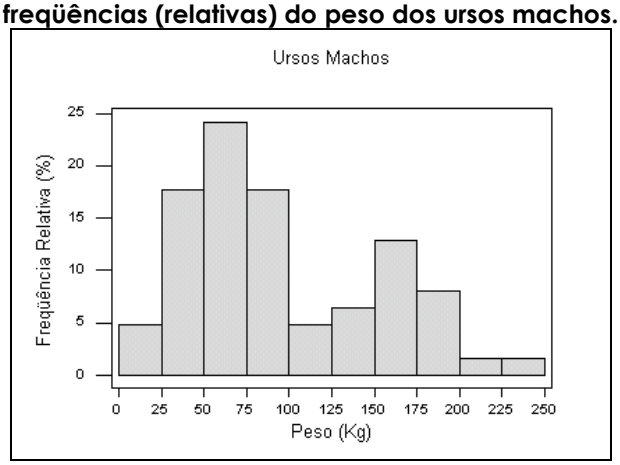


Figura 4.14: Histograma para a distribuição de freqüências (relativas) do peso dos ursos machos.



Os histogramas das figuras 4.13 e 4.14 têm a mesma forma, apesar de serem construídos usando as freqüências absolutas e relativas, respectivamente. O objetivo dessas figuras é mostrar que a escolha do tipo de freqüência a ser usada não muda a forma da distribuição. Entretanto, o uso da freqüência relativa torna o histograma comparável a outros histogramas, mesmo que os conjuntos de dados tenham tamanhos diferentes (desde a mesma escala seja usada!)

Analisando o histograma para o peso dos ursos machos, podemos perceber que há dois grupos de ursos: os mais leves, com pesos em torno de 50 a 75 Kg, e os mais pesados, com pesos em torno de 150 a 175 Kg. Essa divisão pode ser devida a uma outra característica dos ursos, como idades ou hábitos alimentares diferentes, por exemplo.

A Tabela 4.6 apresenta a distribuição de freqüências para o peso dos ursos fêmeas, representada graficamente pelo histograma da Figura 4.15. Apesar de não haver, neste conjunto de dados, fêmeas com peso maior de que 175 Kg, as três últimas classes foram mantidas para que pudéssemos comparar machos e fêmeas quanto ao peso.

Tabela 4.6: Distribuição de freqüências dos ursos fêmeas segundo peso.

Peso (kg)	Freqüência Absoluta	Freqüência Relativa (%)	Freq. Abs. Acumulada	Freq. Rel. Acumulada
0 |- 25	3	8,6	3	8,6
25 |- 50	5	14,3	8	22,9
50 |- 75	18	51,4	26	74,3
75 |- 100	5	14,3	31	88,6
100 |- 125	2	5,7	33	94,3
125 |- 150	1	2,9	34	97,1
150 |- 175	1	2,9	35	100,0
175 |- 200	0	0	35	100,0
200 |- 225	0	0	35	100,0
225 |- 250	0	0	35	100,0
Total	35	100	-	-

A Figura 4.16 mostra o histograma para o peso dos ursos machos. Note que ele tem a mesma forma dos histogramas das figuras 4.13 e 4.14, porém com as barras mais "achatadas", devido à mudança de escala no eixo vertical para torná-lo comparável ao histograma das fêmeas.

Comparando as distribuições dos pesos dos ursos machos e fêmeas, podemos concluir que as fêmeas são, em geral, menos pesadas do que os machos, distribuindo-se quase simetricamente em torno da classe de 50 a 75 Kg . O peso das fêmeas é mais homogêneo (valores mais próximos entre si) do que o peso dos ursos machos.

Figura 4.15: Histograma para a distribuição de freqüências do peso dos ursos fêmeas.

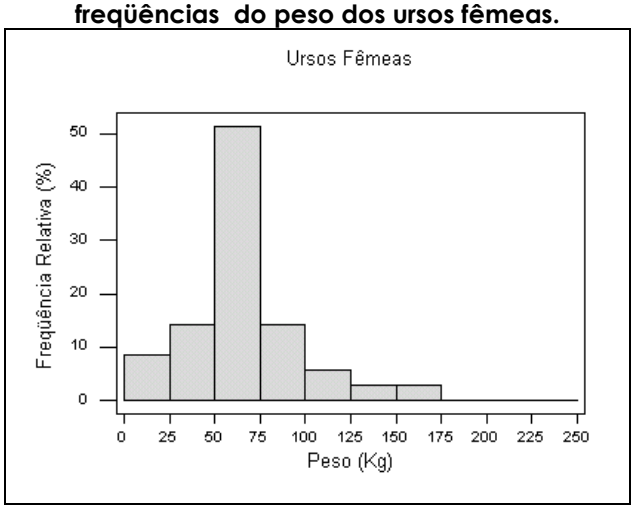
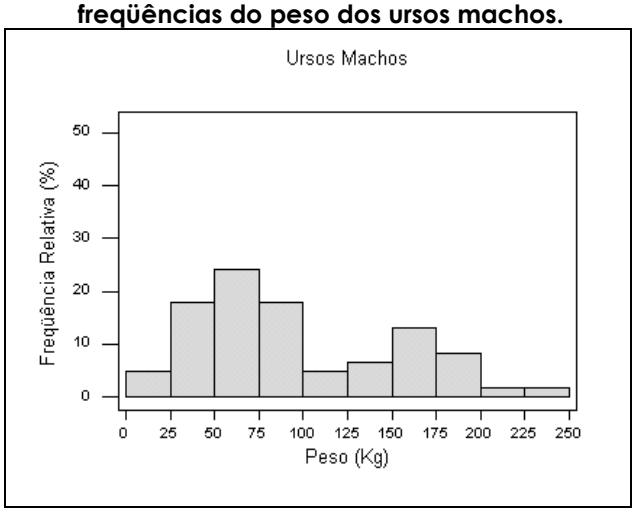
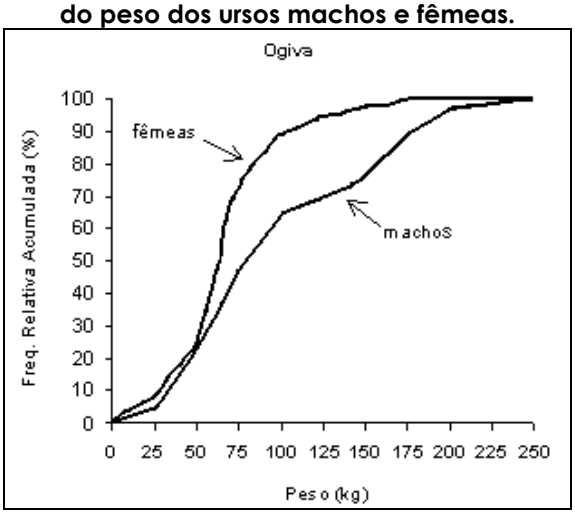


Figura 4.16: Histograma para a distribuição de freqüências do peso dos ursos machos.



Muitas vezes, a análise da distribuição de freqüências acumuladas é mais interessante do que a de freqüências simples, representada pelo histograma. O gráfico usado na representação gráfica da distribuição de freqüências acumuladas de uma variável contínua é a **ogiva**, apresentada na Figura 4.17. Para a construção da ogiva, são usadas as *freqüências acumuladas* (absolutas ou relativas) no eixo vertical e os limites *superiores* de classe no eixo horizontal.

Figura 4.17: Ogivas para as distribuições de freqüências do peso dos ursos machos e fêmeas.



O primeiro ponto da ogiva é formado pelo limite inferior da primeira classe e o valor zero, indicando que abaixo do limite inferior da primeira classe não existem observações. Daí por diante, são usados os limites superiores das classes e suas respectivas freqüências acumuladas, até a última classe, que acumula todas as observações. Assim, uma ogiva deve começar no valor zero e, se for construída com as freqüências relativas acumuladas, terminar com o valor 100%.

A ogiva permite que sejam respondidas perguntas do tipo:

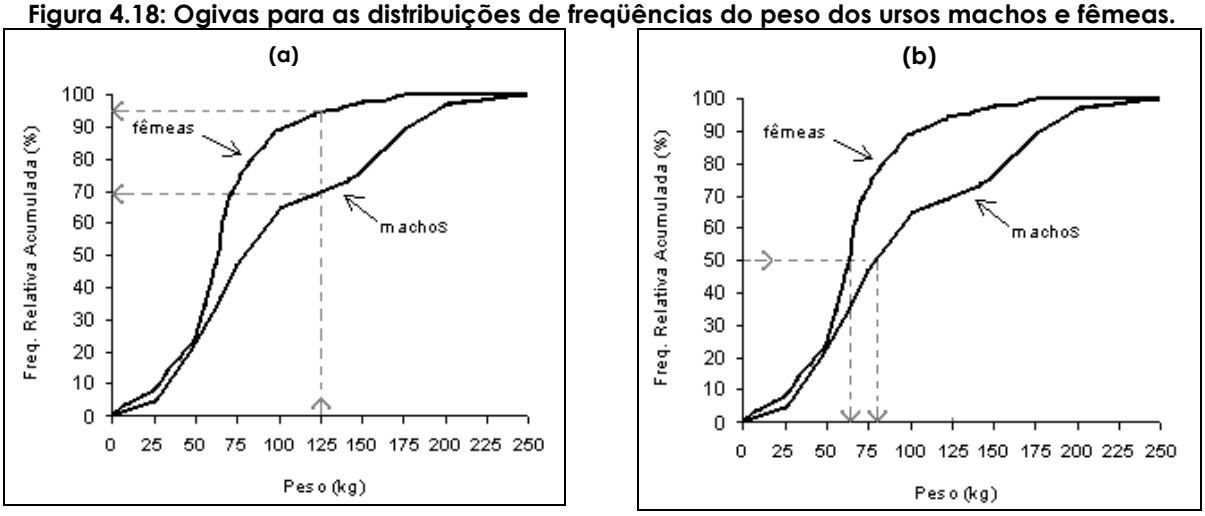
a) Qual o percentual de ursos tem peso de até 125 Kg?

Na Figura 4.18(a), traçamos uma linha vertical partindo do ponto 125 kg até cruzar com cada ogiva (fêmeas e machos). A partir deste ponto de cruzamento, traçamos uma linha horizontal até o eixo das freqüências acumuladas, encontrando o valor de 70% para os machos e 95% para as fêmeas. Assim, 95% das fêmeas têm até 125 kg, enquanto 70% dos machos têm até 125 kg. È o mesmo que dizer que apenas 5% das fêmeas pesam mais que 125 kg, enquanto 30% dos machos pesam mais que 125 kg.

b) Qual o valor do peso que deixa abaixo (e acima) dele 50% dos ursos?

Na Figura 4.18(b), traçamos uma linha horizontal partindo da freqüência acumulada de 50% até encontrar as duas ogivas. A partir destes pontos de encontro, traçamos uma linha vertical até o eixo dos valores de peso, encontrando o valor de 80 kg para os machos e 65 kg para as fêmeas. Assim, metade dos machos pesa até 80 kg (e metade pesam mais que 80 kg), enquanto metade das fêmeas pesam até 65 kg.

Figura 4.18: Ogivas para as distribuições de freqüências do peso dos ursos machos e fêmeas.



4.4. Outros Gráficos para Variáveis Quantitativas

Quando construímos uma tabela de freqüências para uma variável quantitativa utilizando agrupamento de valores em classes, estamos resumindo a informação contida nos dados. Isto é desejável quando o número de dados é grande e, sem um algum tipo de resumo, ficaria difícil tirar conclusões sobre o comportamento da variável em estudo.

Porém, quando a quantidade de dados disponíveis não é tão grande, o resumo promovido pelo histograma não é aconselhável.

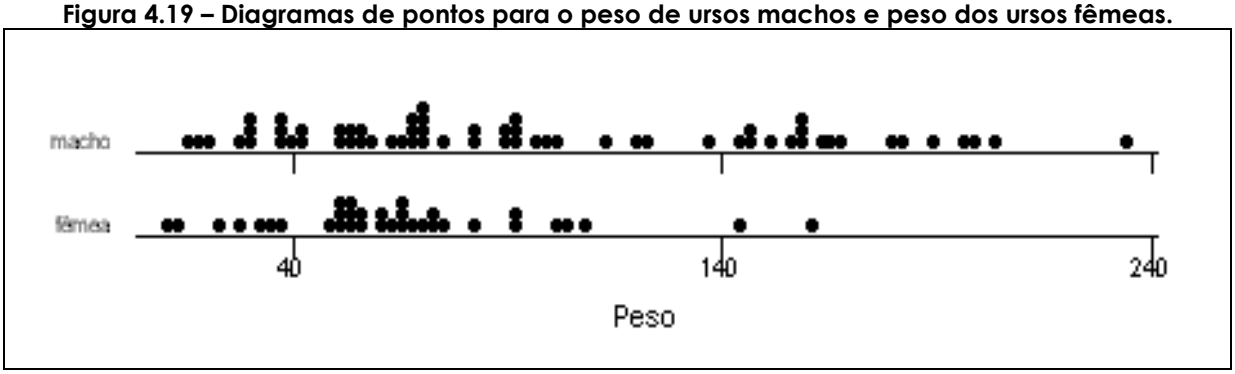
Para os casos em que o número de dados é pequeno, uma alternativa para a visualização da distribuição desses dados são os gráficos denominados **diagrama de pontos** e **diagrama de ramo-e-folhas**.

O Diagrama de Pontos

Uma representação alternativa ao histograma para a distribuição de freqüências de uma variável quantitativa é o *diagrama de pontos*, como aqueles mostrados na Figura 4.19.

Neste gráfico, cada ponto representa uma observação com determinado valor da variável. Observações com mesmo valor são representadas com pontos empilhados neste valor.

Figura 4.19 – Diagramas de pontos para o peso de ursos machos e peso dos ursos fêmeas.



Através da comparação dos diagramas de pontos da Figura 4.19, podemos ver que os ursos machos possuem pesos menos homogêneos (mais dispersos) do que as fêmeas, que estão concentradas na parte esquerda do eixo de valores de peso.

O Diagrama de Ramo-e-Folhas

Outro gráfico útil e simples para representar a distribuição de freqüências de uma variável quantitativa com poucas observações é o *diagrama de ramo-e-folhas*. Sua maior vantagem sobre os demais é que ele explicita os valores dos dados, como veremos.

Exemplo dos ursos marrons (continuação)

Dos 35 ursos fêmeas observados, somente 20 puderam ter sua idade estimada. Para visualizar a distribuição dos valores de idade dessas fêmeas, usaremos um diagrama de ramo-e-folhas, já que um histograma resumiria mais ainda algo que já está resumido.

Os 20 valores de idade (em meses) disponíveis, já ordenados são:

```
8  9  11  17  17  19  20  44  45  53  57  57  57  58  70  81  82  83  100  104
```

Podemos organizar os dados, separando-os pelas dezenas, uma em cada linha:

```
  8  9
 11  17  17  19
 20
 44  45
 53  57  57  57  58
 70
 81  82  83
100 104
```

Como muitos valores em cada linha tem as dezenas em comum, podemos colocar as dezenas em "evidência", separando-as das unidades por um traço. Ao dispor os dados dessa maneira, estamos construindo um diagrama de ramo-e-folhas (Figura 4.20). O lado com as dezenas é chamado de *ramo*, no qual estão "dependuradas" as unidades, chamadas *folhas*.

Figura 4.20 - Ramo-e-folhas da idade (meses) dos ursos fêmeas.

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

Legenda: 5 | 3 = 53 meses
10| 0 = 100 meses

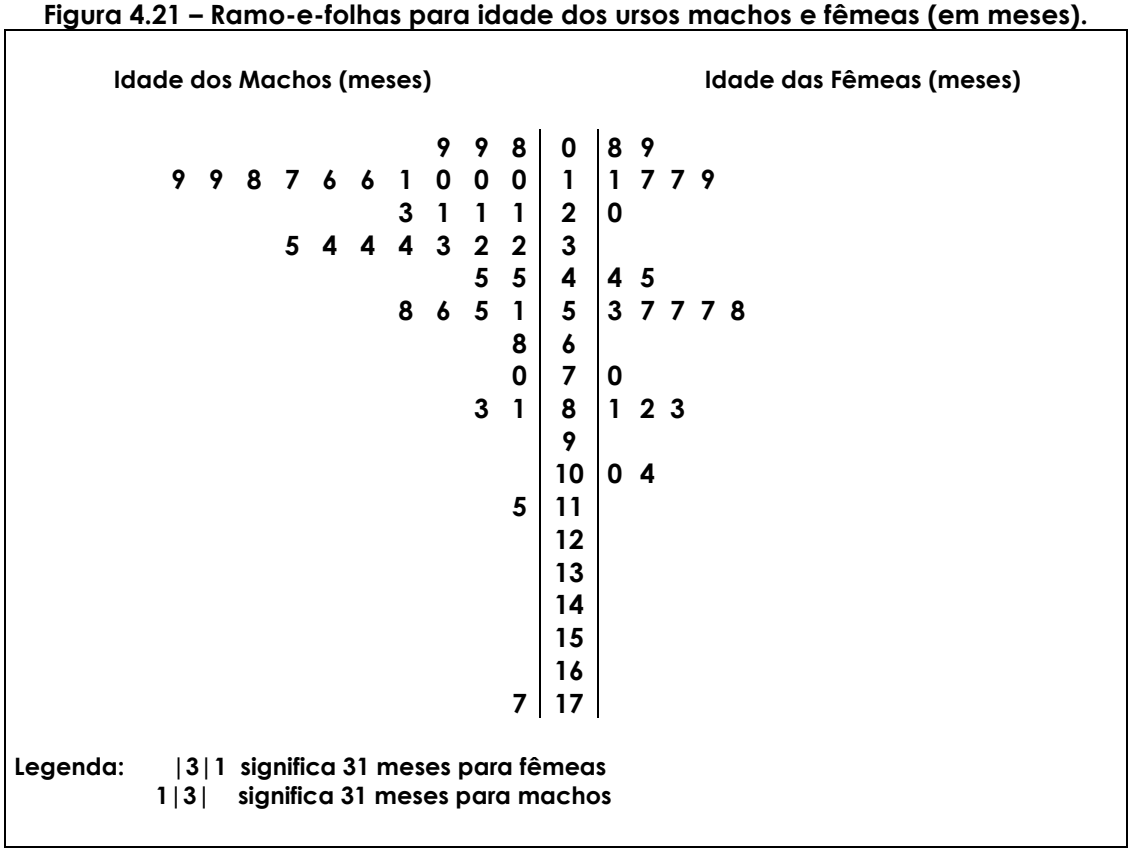
Os ramos e as folhas podem representar quaisquer unidades de grandeza (dezenas e unidades, centenas e dezenas, milhares e centenas, etc). Para sabermos o que está sendo representado, um ramo-e-folhas deve ter sempre uma legenda, indicando o que significam os ramos e as folhas. Se a idade estivesse medida em dias, por exemplo, usando esse mesmo ramo-e-folhas, poderíamos estabelecer que o ramo representaria as centenas e as folhas, as dezenas. Assim, `0|8` seria igual a 80 dias e `10|4` seria igual a 1040 dias.

Analisando o ramo-e-folhas para a idade dos ursos fêmeas, percebemos a existência de três grupos: fêmeas mais jovens (até 20 meses), fêmeas mais crescidas (de 44 a 58 meses) e um grupo mais velho (mais de 70 meses), com destaque para duas fêmeas bem mais velhas.

O ramo-e-folhas também pode ser usado para comparar duas distribuições de valores, como mostra a Figura 4.21. Aproveitando o mesmo ramo do diagrama das fêmeas, podemos fazer o diagrama dos machos, utilizando o lado esquerdo. Observe que as folhas dos ursos machos são dependuradas de modo espelhado, assim como explica a legenda, que agora deve ser dupla.

Observando a Figura 4.21, notamos que os ursos machos são, em geral, mais jovens do que os ursos fêmeas, embora possuam dois ursos bem "idosos" em comparação com os demais.

Figura 4.21 – Ramo-e-folhas para idade dos ursos machos e fêmeas (em meses).



ⓘ **Importante**: No ramo-e-folhas, estamos trabalhando, implicitamente, com freqüências absolutas. Assim, ao comparar dois grupos de tamanhos diferentes, devemos levar isso em conta. Caso os tamanhos dos grupos sejam muito diferentes, não se deve adotar o ramo-e-folhas como gráfico para comparação de distribuições.

4.5. Aspectos Gerais da Distribuição de Freqüências

Ao estudarmos a distribuição de freqüências de uma variável quantitativa, seja em um grupo apenas ou comparando vários grupos, devemos verificar basicamente três características:

- Tendência Central;
- Variabilidade;
- Forma.

O histograma (ou o diagrama de pontos, ou o ramo-e-folhas) permite a visualização destas características da distribuição de freqüências, como veremos a seguir. Além disso, elas podem ser quantificadas através das *medidas de síntese numérica* (não discutidas aqui).

Tendência Central

A tendência central da distribuição de freqüências de uma variável é caracterizada pelo valor (ou faixa de valores) "típico" da variável.

Uma das maneiras de representar o que é "típico" é através do valor mais freqüente da variável, chamado de **moda**. Ou, no caso da tabela de freqüências, a classe de maior freqüência, chamada de **classe modal**. No histograma, esta classe corresponde àquela com barra mais alta ("pico").

No exemplo dos ursos marrons, a classe modal do peso dos ursos fêmeas é claramente a terceira, de 50 a 75 kg (Figura 4.16). Assim, os ursos fêmeas pesam, tipicamente, de 50 a 75 kg. Entretanto, para os ursos machos, temos dois picos: de 50 a 75 kg e de 150 a 175 kg (Figura 4.17). Ou seja, temos um grupo de machos com peso típico como o das fêmeas e outro grupo, menor, formado por ursos tipicamente maiores.

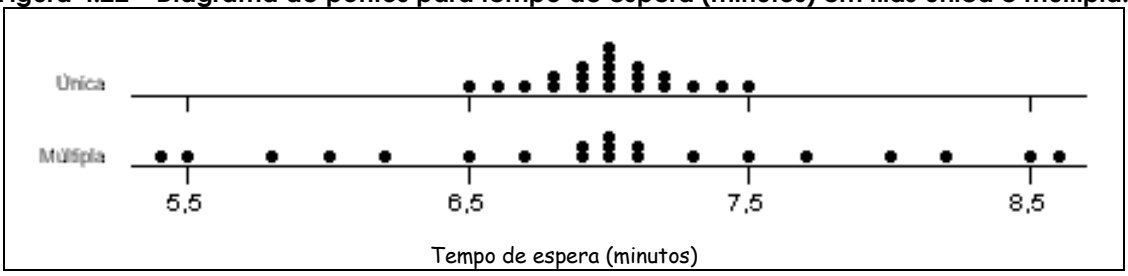
Dizemos que a distribuição de freqüências do peso dos ursos fêmeas é **unimodal** (apenas uma moda) e dos ursos machos é **bimodal** (duas modas). Geralmente, um histograma bimodal indica a existência de dois grupos, com valores centrados em dois pontos diferentes do eixo de valores. Uma distribuição de freqüências pode também ser **amodal**, ou seja, todos os valores são igualmente freqüentes.

Variabilidade

Para descrever adequadamente a distribuição de freqüências de uma variável quantitativa, além da informação do valor representativo da variável (tendência central), é necessário dizer também o quanto estes valores variam, ou seja, o quão dispersos eles são.

De fato, somente a informação sobre a tendência central de um conjunto de dados não consegue representá-lo adequadamente. A Figura 4.22 mostra um diagrama de pontos para os tempos de espera de 21 clientes de dois bancos, um com fila única e outro com fila múltipla, com o mesmo número de atendentes. Os tempos de espera nos dois bancos têm a mesma tendência central de 7 minutos. Entretanto, os dois conjuntos de dados são claramente diferentes, pois os valores são muito mais **dispersos** no banco com fila múltipla. Assim, quando entramos num fila única, esperamos ser atendidos em cerca de 7 minutos, com uma **variação** de, no máximo, meio minuto a mais ou a menos. Na fila múltipla, a variação é maior, indicando-se que tanto pode-se esperar muito mais ou muito menos que o valor típico de 7 minutos.

Figura 4.22 – Diagrama de pontos para tempo de espera (minutos) em filas única e múltipla.



Forma

A distribuição de freqüências de uma variável pode ter várias *formas*, mas existem três formas básicas, apresentadas na Figura 4.23 através de histogramas e suas respectivas ogivas.

Quando uma distribuição é **simétrica** em torno de um valor (o mais freqüente), significa que as observações estão igualmente distribuídas em torno desse valor (metade acima e metade abaixo).

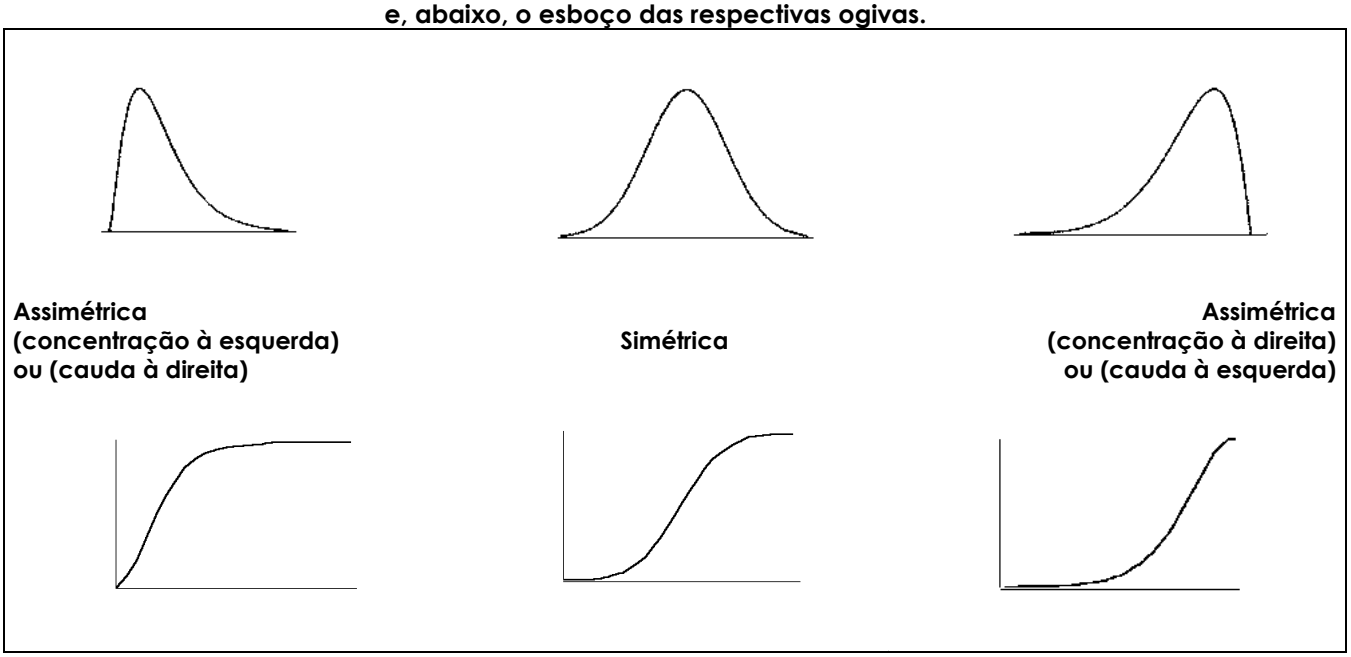
A **assimetria** de uma distribuição pode ocorrer de duas formas:

- quando os valores concentram-se à esquerda (assimetria com concentração à esquerda ou assimetria com cauda à direita);
- quando os valores concentram-se à direita (assimetria com concentração à direita ou com assimetria cauda à esquerda);

Ao definir a assimetria de uma distribuição, algumas pessoas preferem se referir ao lado onde está a concentração dos dados. Porém, outras pessoas preferem se referir ao lado onde está "faltando" dados (cauda). As duas denominações são alternativas.

Em alguns casos, apenas o conhecimento da forma da distribuição de freqüências de uma variável já nos fornece uma boa informação sobre o comportamento dessa variável. Por exemplo, o que você acharia se soubesse que a distribuição de freqüências das notas da primeira prova da disciplina de Estatística que você está cursando é, geralmente, assimétrica com concentração à direita ? Como você acha que é a forma da distribuição de freqüências da renda no Brasil ?

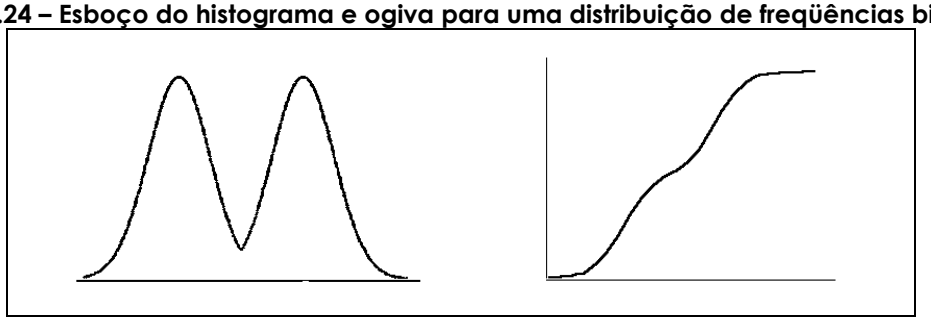
Figura 4.23 – Formas básicas para a distribuição de freqüências de uma variável quantitativa e, abaixo, o esboço das respectivas ogivas.



Note que, quando a distribuição é assimétrica com concentração à esquerda, a ogiva cresce bem rápido, por causa do acúmulo de valores do lado esquerdo do eixo. Por outro lado, quando a distribuição é assimétrica com concentração à direita, a ogiva cresce lentamente no começo e bem rápido na parte direita do eixo, por causa do acúmulo de valores desse lado. Quando a distribuição é simétrica, a ogiva tem a forma de um "S" suave e simétrico.

A ogiva para uma distribuição de freqüências bimodal (Figura 4.24) mostra essa característica da distribuição através de um platô ("barriga") no meio da ogiva. A ogiva para o peso dos ursos machos (Figura 4.18) também mostra essa "barriga".

Figura 4.24 – Esboço do histograma e ogiva para uma distribuição de freqüências bimodal.



5. Séries Temporais

Séries temporais (ou séries históricas) são um conjunto de observações de uma mesma variável quantitativa (discreta ou contínua) feitas ao longo do tempo.

O conjunto de todas as temperaturas medidas diariamente numa região é um exemplo de série temporal.

Um dos objetivos do estudo de séries temporais é conhecer o comportamento da série ao longo do tempo (aumento, estabilidade ou declínio dos valores). Em alguns estudos, esse conhecimento pode ser usado para se fazer previsões de valores futuros com base no comportamento dos valores passados.

A representação gráfica de uma série temporal é feita através do **gráfico de linha**, como exemplificado nas figuras 5.1 e 5.2. No eixo horizontal do gráfico de linha, está o indicador de tempo e, no eixo vertical, a variável a ser representada. As linhas horizontais pontilhadas são opcionais e só devem ser colocadas quando ajudarem na interpretação do gráfico. Caso contrário, devem ser descartadas, pois, como já enfatizamos antes, um gráfico deve ser o mais "limpo" possível.

Figura 5.1 – Gráfico de linha para o número de ursos machos e fêmeas observados ao longo dos meses de pesquisa.

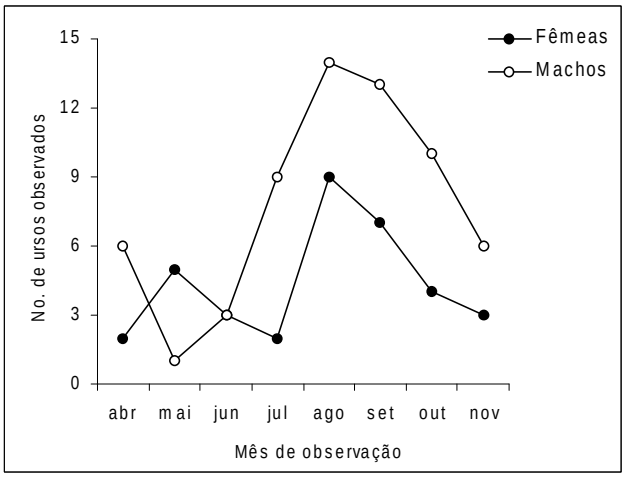
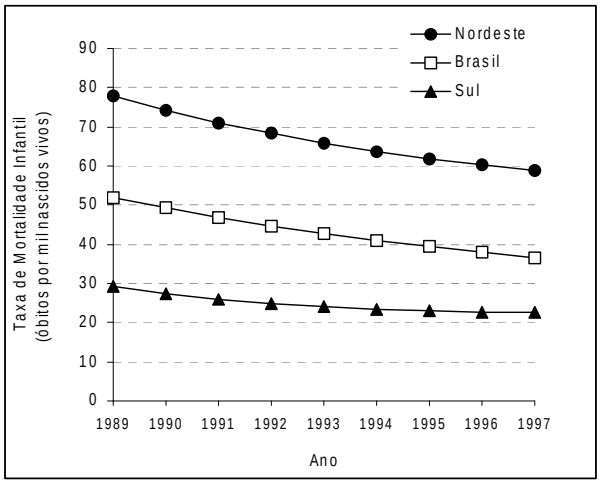


Figura 5.2 – Gráfico de linha para a taxa de mortalidade infantil de 1989 a 1997 nas Regiões Nordeste e Sul e no Brasil.



No gráfico da Figura 5.2, podemos notar que a taxa de mortalidade infantil na região Nordeste esteve sempre acima da taxa da região Sudeste durante todo o período considerado, com um declínio das taxas nas duas regiões e também no Brasil como um todo ao longo do período. Embora o declínio absoluto na taxa da região Nordeste tenha sido maior (aproximadamente 20 casos em mil nascidos vivos), a redução percentual na taxa da região Sudeste foi maior (cerca de 8 casos a menos nos 30 iniciais, ou seja, 27% a menos, enquanto 20 casos a menos nos 80 iniciais na região Nordeste representam uma redução de 25%. Podemos observar ainda uma tendência à estabilização da taxa de mortalidade infantil da região Sudeste a partir do ano de 1994, enquanto a tendência de declínio permanece na região Nordeste e no Brasil.

Ao analisar e construir um gráfico de linhas, devemos estar atentos a certos detalhes que podem mascarar o verdadeiro comportamento dos dados. A Figura 5.3(a) apresenta um gráfico de linhas para o preço médio do litro de leite entre os meses de maio e agosto de 2001. Apesar de colocar os valores para cada mês, o gráfico não mostra a escala de valores e não representa a série desde o começo da escala, o valor zero. Essa concentração da visualização da linha somente na parte do gráfico onde os dados estão situados distorce a verdadeira de dimensão da queda do preço, acentuando-a. Ao compararmos com o gráfico da Figura 5.3(b), cujo escala vertical começa no zero, percebemos que houve mesmo uma queda, mas não tão acentuada quanto aquela mostrada no gráfico divulgado no jornal.

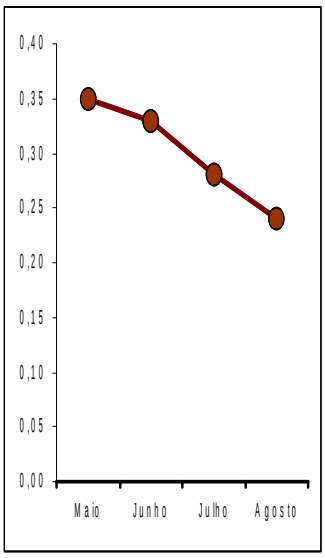
Outro aspecto mascarado pela falta da escala é que as diferenças entre os valores numéricos não correspondem às distâncias representadas no gráfico. Por exemplo, no gráfico de linha divulgado para a série do preço do leite, vemos que a queda no preço de maio para junho foi de R$0,02 e, de julho para agosto, foi de R$0,04, duas vezes maior. No entanto, a distância

(vertical) entre os pontos de maio e julho é *maior* do que a distância (vertical) entre os pontos de julho e agosto!! E mais, a queda de junho para junho foi de R$0,05, pouco mais do que a queda de R$0,04 de junho a agosto. Porém, a distância (vertical) no gráfico entre os pontos de junho e julho é cerca de quatro vezes maior do que a distância (vertical) dos pontos de julho e agosto!! Examinando o gráfico apenas visualmente, sem nos atentar para os números, tenderemos a pensar que as grandes quedas no preço do leite ocorreram no começo do período de observação (de maio a julho), enquanto, na verdade, as quedas se deram quase da mesma forma mês a mês, sendo um pouco maiores no final do período (de julho a agosto). Além disso, a palavra "despenca" nos faz pensar numa queda abrupta, que é o que o gráfico divulgado parece querer mostrar. No entanto, analisando o gráfico da Figura 5.3(a), que corrige essas distorções, notamos que houve sim uma queda, mas não tão abrupta quanto colocada na Figura 5.3(b).

Figura 5.3 – Gráfico de linhas para o preço médio do litro de leite: (a) original (jornal Folha de São Paulo, set/2001), (b) modificado, com a escala de valores mostrada e iniciando-se no zero.

a)

b)



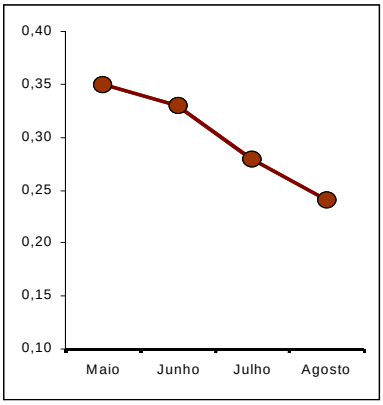
A Figura 5.4 mostra os efeitos na representação de uma série temporal quando mudamos o começo da escala de valores do eixo vertical. À medida que aproximamos o começo da escala do valor mínimo da série, a queda nos parece mais abrupta. A mesma observação vale para o caso em que o gráfico mostrar um aumento dos valores da série: quanto mais o início da escala se aproxima do valor mínimo da série, mais acentuado parecerá o aumento.

De maneira geral, um gráfico de linhas deve ser construído de modo que:

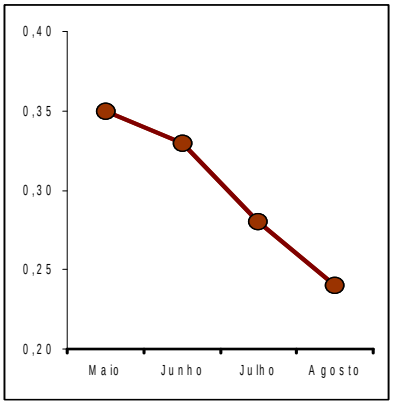
- O início do eixo vertical seja o valor mínimo possível para a variável que está sendo representada (para o caso do preço de leite, o valor zero, leite de graça), para evitar as distorções ilustradas na Figura 5.4;
- O final do eixo vertical seja tal que a série fica centrada em relação ao eixo vertical, como mostrado na Figura 5.5(a);
- Os tamanhos dos eixos sejam o mais parecidos possível, para que não ocorra a distorção mostrada nos gráficos (b) e (c)) da Figura 5.5.

Figura 5.5 – Efeitos da mudança no início e/ou final da escala do gráfico em linhas da série temporal do preço do leite.

a)



b)



c)

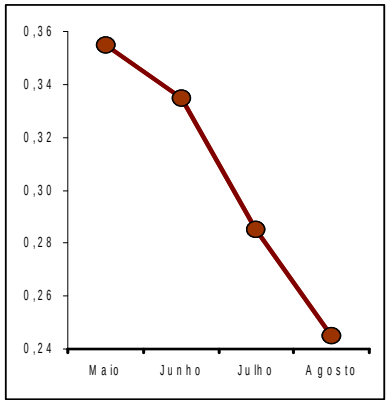
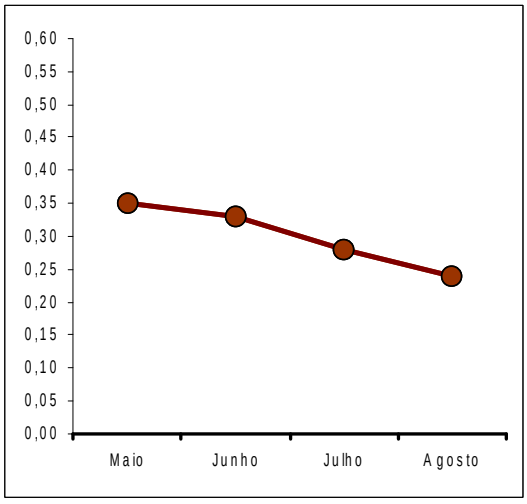
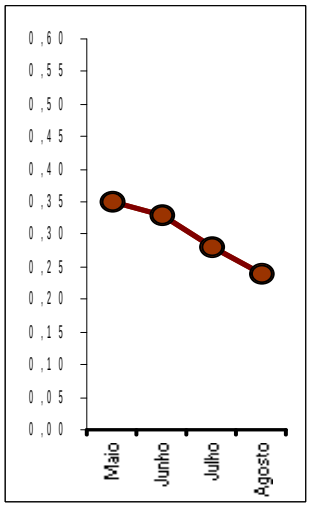


Figura 5.6 - Efeitos de alterações na dimensão horizontal do gráfico de linhas da série do preço do leite

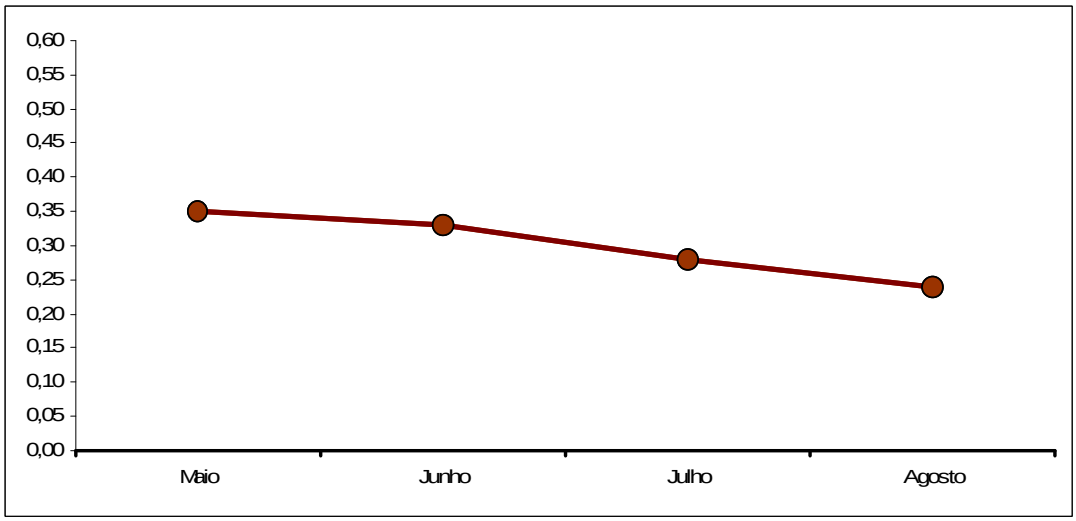
a)



b)



c)



6. O Diagrama de Dispersão

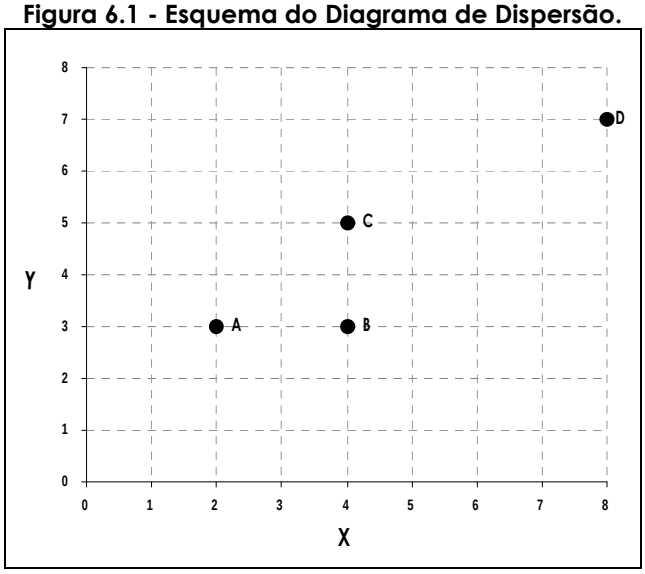
O diagrama de dispersão é um gráfico onde pontos no espaço cartesiano XY são usados para representar simultaneamente os valores de duas variáveis quantitativas medidas em cada indivíduo do conjunto de dados.

O Quadro 6.1 e a Figura 6.1 mostram um esquema do desenho do diagrama de dispersão. Neste exemplo, foram medidos os valores de duas variáveis quantitativas, X e Y, em quatro indivíduos. O eixo horizontal do gráfico representa a variável X e o eixo vertical representa a variável Y.

Quadro 6.1 - Dados esquemáticos.

Indivíduos	Variável X	Variável Y
A	2	3
B	4	3
C	4	5
D	8	7

Figura 6.1 - Esquema do Diagrama de Dispersão.



O diagrama de dispersão é usado principalmente para visualizar a relação/associação entre duas variáveis, mas também para é muito útil para:

- Comparar o efeito de dois tratamentos no mesmo indivíduo.
- Verificar o efeito tipo antes/depois de um tratamento;

A seguir, veremos quatro exemplos da utilização do diagrama de dispersão. Os dois primeiros referem-se ao estudo da associação entre duas variáveis. O terceiro utiliza o diagrama de dispersão para comparar o efeito de duas condições no mesmo indivíduo. O último exemplo, similar ao terceiro, verifica o efeito da aplicação de um tratamento, comparando as medidas antes e depois da medicação.

Exemplo dos ursos marrons (continuação).

Recorde que um dos objetivos dos pesquisadores neste estudo é encontrar uma maneira de conhecer o peso do urso através de uma medida mais fácil de obter do que a direta (carregar uma balança para o meio da selva e colocar os ursos em cima dela) como, por exemplo, uma medida de comprimento (altura, perímetro do tórax, etc.).

O problema estatístico aqui é encontrar uma variável que tenha uma *relação forte* com o peso, de modo que, a partir de seu valor medido, possa ser calculado (*estimado*, na verdade) o valor peso indiretamente, através de uma equação matemática.

O primeiro passo para encontrar esta variável é fazer o diagrama de dispersão das variáveis candidatas (eixo horizontal) versus o peso (eixo vertical), usando os pares de informações de todos os ursos. Você pode tentar as variáveis: idade, altura, comprimento da cabeça, largura da cabeça, perímetro do pescoço e perímetro do tórax.

Nas figuras 6.2 e 6.3, mostramos a relação entre peso e altura e entre peso e perímetro do tórax. Respectivamente.

Figura 6.2 - Diagrama de dispersão da altura *versus* o peso dos ursos marrons.

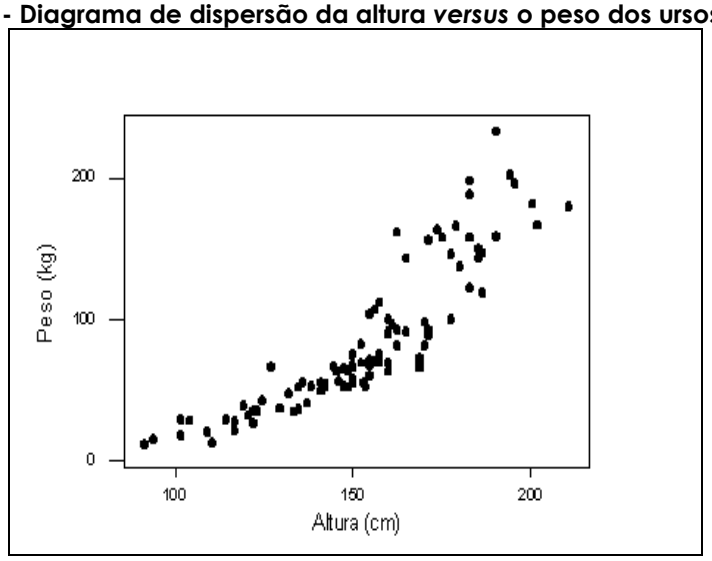
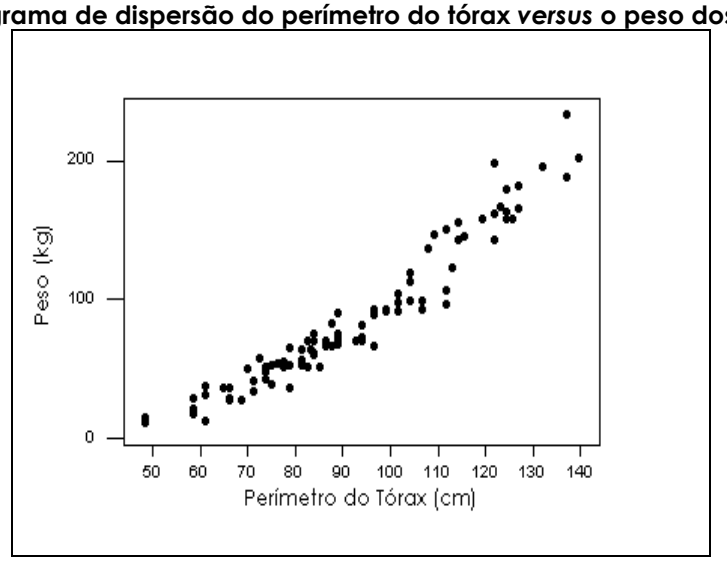


Figura 6.3 - Diagrama de dispersão do perímetro do tórax *versus* o peso dos ursos marrons.



Podemos ver que, tanto a altura quanto o perímetro do tórax são fortemente associados ao peso do urso, no sentido de que quanto mais alto o urso ou quanto maior a medida de seu tórax, mais pesado ele será. Mas note que este crescimento é linear para o perímetro do tórax e não-linear para a altura. Além disso, com os pontos estão mais dispersos no gráfico da altura, a variável mais adequada para estimar, sozinha, o peso é o perímetro do tórax (a técnica estatística adequada aqui se chama Regressão Linear Simples).

Exemplo dos morangos.

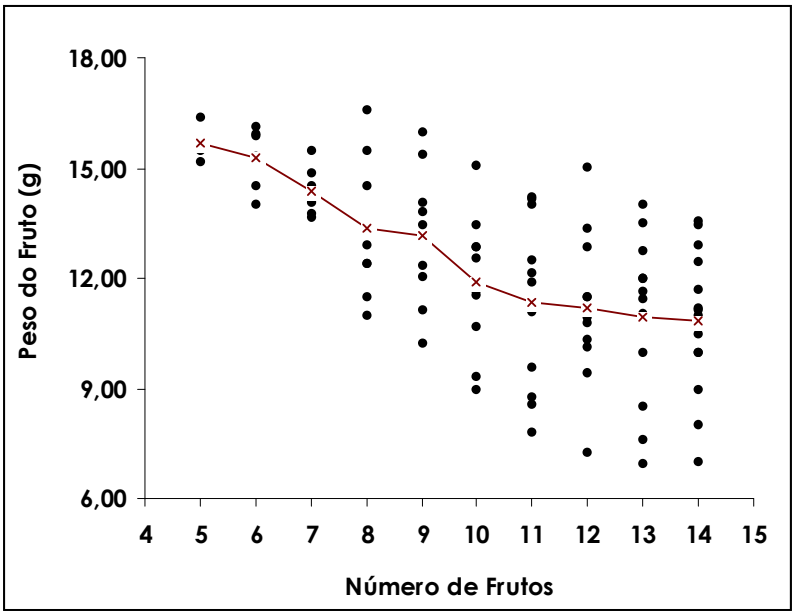
Um produtor de morangos para exportação deseja produzir frutos grandes, pois frutos pequenos têm pouco valor mesmo no mercado interno. Além disso, os frutos, mesmo grandes, não devem ter tamanhos muito diferentes entre si. O produtor suspeita que uma dos fatores que altera o tamanho dos frutos é o número de frutos por árvore.

Para investigar a relação entre o número de frutos que uma planta produz e o peso destes frutos, ele observou dados de 10 morangueiros na primeira safra (Quadro 6.2). O diagrama de dispersão é mostrado na Figura 6.4.

Quadro 6.2 – Peso dos frutos e número de frutos por planta em 10 morangueiros na primeira safra.

Planta	Nº de frutos	Peso dos Frutos (gramas)
1	5	15,15 15,45 15,63 15,65 16,38
2	6	14,00 14,50 15,35 15,86 15,94 16,13
3	7	13,67 13,76 14,06 14,11 14,54 14,89 15,50
4	8	11,00 11,50 12,39 12,39 12,90 14,50 15,50 16,56
5	9	10,24 11,12 12,05 12,37 13,48 13,80 14,04 15,39 16,00
6	10	9,00 9,32 10,67 11,56 11,67 12,56 12,83 12,84 13,43 15,09
7	11	7,82 8,56 8,74 9,57 11,08 11,92 12,13 12,50 14,14 14,20 14,00
8	12	7,25 9,41 10,15 10,33 10,80 10,95 11,13 11,48 11,49 12,86 13,37 15,04
9	13	6,95 7,61 8,53 10,00 10,94 11,04 11,43 11,63 11,97 12,02 12,74 13,53 14,00
10	14	7,00 8,00 9,00 10,00 10,00 10,50 11,00 11,16 11,17 11,70 12,45 12,89 13,47 13,54

Figura 6.4 - Diagrama de dispersão do número de frutos por árvore *versus* o peso do fruto e linha unindo os pesos médios dos frutos.



O diagrama de dispersão mostra-nos dois fatos. O primeiro, que há um decréscimo no valor médio do peso do fruto por árvore à medida que cresce o número de frutos na árvore. Ou seja, não é vantagem uma árvore produzir muitos frutos, pois eles tenderão a ser muito pequenos.

O segundo fato que percebemos é que, com o aumento no número de frutos na árvore, cresce também a variabilidade no peso, gerando tanto frutos muito grandes, como muito pequenos.

Assim, conclui-se que não é vantagem ter poucas plantas produzindo muito frutos, mas sim muitas plantas produzindo poucos frutos, mas grandes e uniformes. Uma análise mais detalhada poderá determinar o número ideal de frutos por árvore, aquele que maximiza o peso médio e, ao mesmo tempo, minimiza a variabilidade do peso.

Exemplo da Capacidade Pulmonar.

Em um estudo[2] sobre técnicas usadas para medir a capacidade pulmonar, coletaram-se dados fisiológicos de 10 indivíduos. Os valores constantes no Quadro 6.3 a seguir representam a capacidade vital forçada (CVF) dos indivíduos em posição sentada e em posição deitada. Deseja-se verificar se a posição (sentada/deitada) influi ou não na medição da capacidade vital forçada.

Quadro 6.3 - Capacidade Vital Forçada (litros) medida em 10 indivíduos nas posições sentada e deitada.

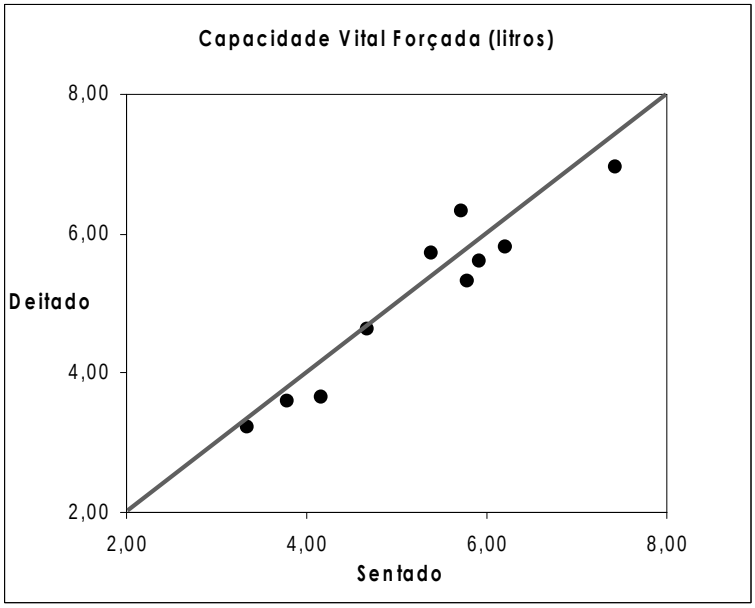
Indivíduo	A	B	C	D	E	F	G	H	I	J
Sentado	4,66	5,70	5,37	3,34	3,77	7,43	4,15	6,21	5,90	5,77
Deitado	4,63	6,34	5,72	3,23	3,60	6,96	3,66	5,81	5,61	5,33

As amostras de cada posição (sentada/deitada) são do tipo *emparelhadas*, pois os mesmos indivíduos foram utilizados nas duas amostras. Assim, é natural compararmos a CVF em cada posição para cada indivíduo, tomando a diferença na CVF *deitada – sentada* (ou o contrário):

Deitado - Sentado: -0,03 0,64 0,35 -0,11 -0,17 -0,47 -0,49 -0,40 -0,29 -0,44

Para grande maioria dos indivíduos, a CVF na posição sentada é maior do que na posição deitado. Mas, como podemos visualizar isto e, ainda, ver se estas diferenças são grandes?Com a ajuda do diagrama de dispersão mostrado na Figura 6.5.

Figura 6.5 - Diagrama de dispersão da capacidade vital forçada nas posições sentada e deitada e linha correspondendo à igualdade das posições.



Cada ponto no diagrama de dispersão corresponde às medidas de CVF de um indivíduo, medida com o indivíduo sentado e deitado. A linha marcada no diagrama corresponde à situação onde a CVF do indivíduo é a mesma nas duas posições. Os pontos acima desta linha são os indivíduos cuja CVF é maior quando deitado; os pontos abaixo da linha são os indivíduos cuja CVF é menor quando deitados. Quanto maior a distância dos pontos à linha, maior é a diferença na CVF entre as duas posições.

Podemos ver que, embora a maior parte dos pontos esteja abaixo da linha, eles estão bem próximos a ela, mostrando que a diferença não é significativa.

[2] Dados de "Validation of Esophageal Balloon Technique at Difefrent Lung Volumes and Postures", de Baydur et al,. Journal of Applied Physiology, v. 62, n. 1.

Exemplo do Captopril.

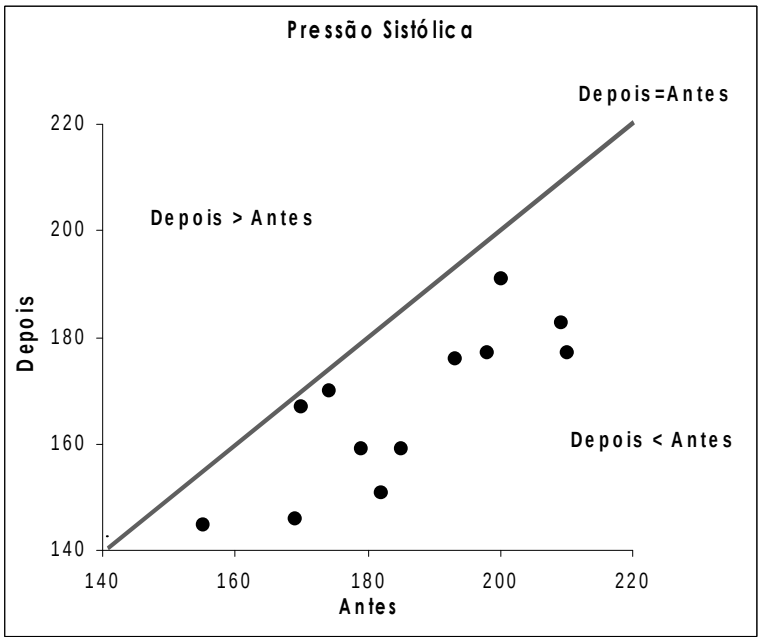
Captopril é um remédio destinado a baixar a pressão sistólica. Para testar seu efeito, ele foi ministrado a 12 pacientes, tendo sido medida a pressão sistólica antes e depois da medicação (Quadro 6.4).

Quadro 6.4 - Pressão sistólica (mmHg) medida em 12 pacientes antes e depois do Captopril.

Paciente	A	B	C	D	E	F	G	H	I	J	K	L
Antes	200	174	198	170	179	182	193	209	185	155	169	210
Depois	191	170	177	167	159	151	176	183	159	145	146	177

Os mesmos indivíduos foram utilizados nas duas amostras (Antes/depois). Assim, é natural compararmos a pressão sistólica para cada indivíduo, comparando a pressão sistólica *depois* e *antes*. Para todos os pacientes, a pressão sistólica depois do Captopril é menor do que antes da medicação. Mas como podemos "ver" se estas diferenças são grandes? Com a ajuda do diagrama de dispersão mostrado na Figura 6.6.

Figura 6.6 - Diagrama de dispersão da pressão sistólica antes X depois da medicação e linha correspondendo ao não efeito individual da medicação.



Cada ponto no diagrama de dispersão corresponde às medidas de pressão sistólica de um paciente, medida antes e depois da medicação. A linha marcada no diagrama corresponde à situação onde a pressão sistólica não se alterou depois do paciente tomar o Captopril. Veja que todos os pontos estão abaixo desta linha, ou seja para todos os pacientes o Captopril fez efeito. Grande parte destes pontos está bem distante da linha, mostrando que a redução na pressão sistólica depois do uso do medicamento não foi pequena.

II- Síntese Numérica

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 na primeira parte deste texto. Nesta segunda parte, 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 (da mesma)tens 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,

$$\overline{X} = \frac{\textit{Soma de todas as observações da amostra}}{\textit{tamanho da amostra}} = \frac{\sum \mathbf{x_i}}{\mathbf{n}}$$

Exemplo 2.1: No conjunto de dados (3 ; 4,5 ; 5,5 ; 2,5 ; 1,3 ; 6), temos n = 6,

$x_1 = 3 \quad x_2 = 4,5 \quad x_3 = 5,5 \quad x_4 = 2,5 \quad x_5 = 1,3 \quad x_6 = 6$

$\sum x_i = 3 + 4,5 + 5,5 + 2,5 + 1,3 + 6 = 22,8$ 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 ; 3,3 ; 2,5 ; 5,6 ; 5 ; 4,3 ; 3,2), temos

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

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 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 ; 3,3 ; 2,5 ; 5,6 ; 5 ; 4,3 ; 3,2).
  Ordenando os valores (2 ; 2,5 ; 3,2 ; 3,3 ; 4,3 ; 5 ; 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 ; 4,5 ; 5,5 ; 2,5 ; 1,3 ; 6).
  Ordenando os valores (1,3 ; 2,5 ; 3 ; 4,5 ; 5,5 ; 6)
  Os valores do meio são 3 e 4,5. A mediana é (3 + 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.

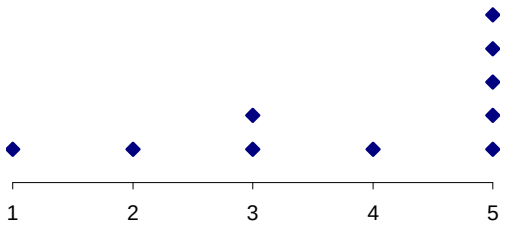
2.3. Moda:

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

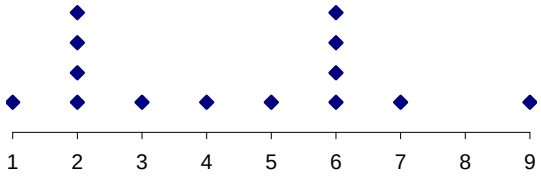
Exemplo 2.3:

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

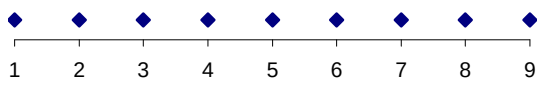
No conjunto de dados
(1; 2 ; 3 ; 3 ; 4 ; 5 ; 5 ; 5 ; 5 ; 5),
há apenas uma moda, o valor 5.
O conjunto de dados é **unimodal.**



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**



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, a categoria modal para a variável *sexo* é a categoria *macho*, pois é a mais frequente.

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 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.

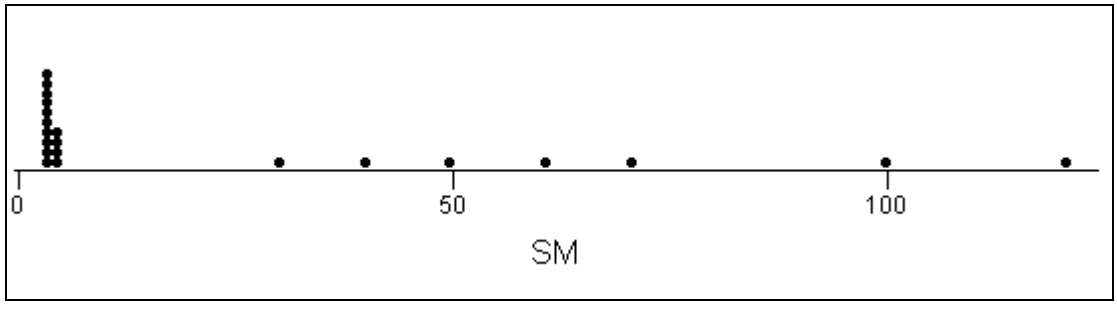
Vejamos um exemplo bastante esclarecedor dessas idéias.

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 aos funcionários de uma empresa



Situação I: dados completos:	Situação II: sem os quatro valores mais altos:
Média = 24,6 SM	Média = 9,8 SM
Mediana = 4 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, sem os esquecer.

Moda *versus* Média e Mediana

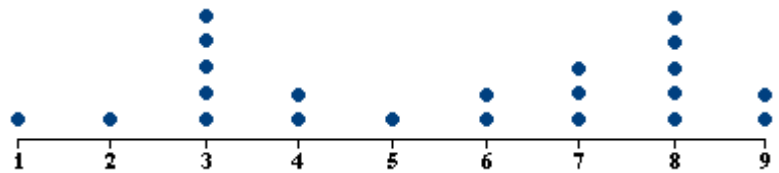
A moda não é uma medida de tendência central muito conhecida, mas tem suas vantagens em relação à média e à mediana, especialmente quando estamos lidando com variáveis que possuem distribuição de freqüências bimodais ou multimodais.

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).

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.

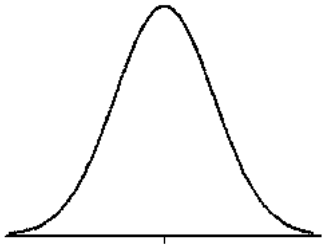
Existem medidas numéricas para quantificar o grau e a direção da assimetria de uma distribuição de freqüências, baseadas na posição relativa entre a média e a mediana. Uma destas **medidas de assimetria** é o **coeficiente de assimetria de Pearson**, dado por :

$$IAP = \frac{3(\text{média} - \text{mediana})}{\text{desvio padrão}}$$

onde desvio-padrão é uma medida de variabilidade que veremos a seguir. Se IAP > 1 ou IAP < -1, a distribuição da variável pode ser considerada significativamente assimétrica com cauda à direita ou significativamente assimétrica com cauda à esquerda, respectivamente.

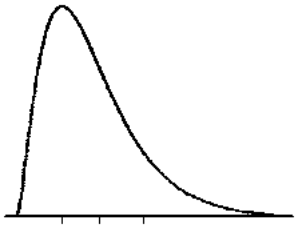
Figura 2.5 – Representação esquemática da forma da distribuição de freqüências e as posições relativas das medidas de tendência central.

Simétrica



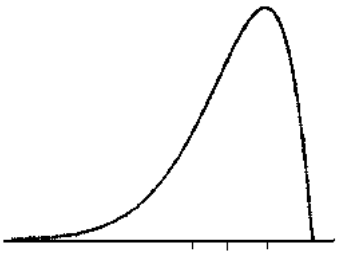
moda = mediana = média

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



moda < mediana < média

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



média < mediana < moda

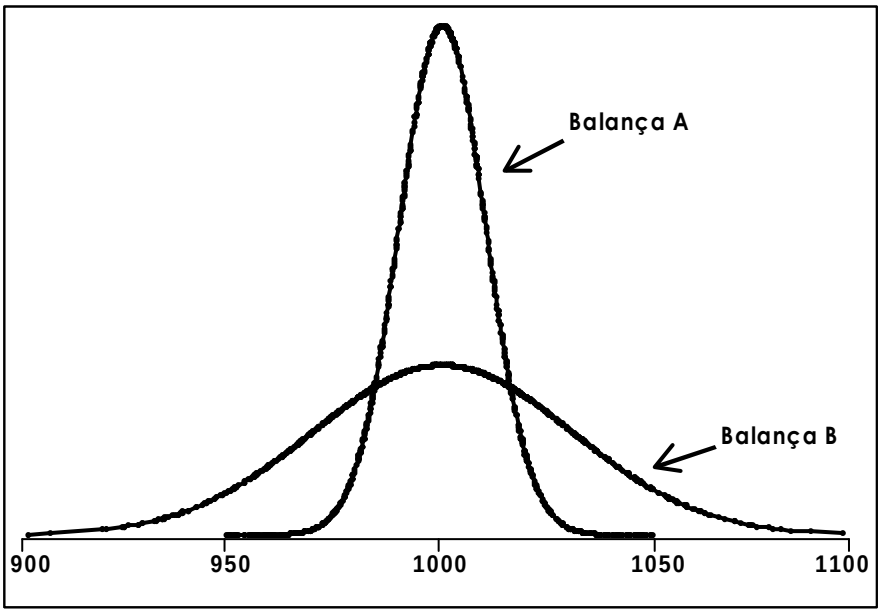
3. Medidas de Variabilidade

As medidas de tendência central (média, mediana, moda) conseguem resumir em um único número, o valor que é "típico" no conjunto de dados. Mas, será que, 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 nas medições de peso fornecidas pela balança. Neste caso, quanto menor a variabilidade desses valores, mais precisa é a balança (considerando que a média das medidas de peso coincida como seu valor real). Observe a Figura 3.1, onde estão representadas 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.

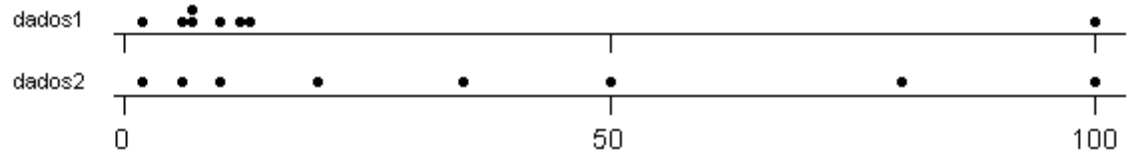
Exemplo 3.2: A Figura 3.2 mostra o diagrama de pontos para os conjuntos de dados I e II:

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 compensaram os positivos.

No entanto, os sinais dos desvios não são importantes para nossa medida de dispersão, já que estamos interessados na quantidade de dispersão e não na direção dela. Portanto, eliminaremos os sinais elevando os desvios ao quadrado[3], 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

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

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)** .

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

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.

[3] 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 e, assim, optou-se por utilizá-la.

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. 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 198 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[4]. Numa indústria farmacêutica, por exemplo, espera-se que os comprimidos de uma certa droga 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 inócuos (com pouco do princípio ativo) ou com extra-dosagem do

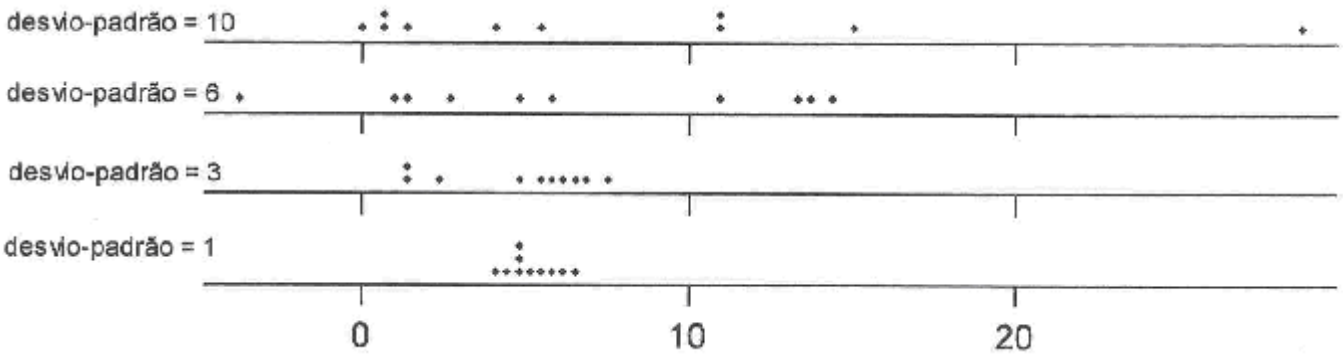
[4] O chamado Controle Estatístico de Processos é um conjunto de ferramentas estatísticas (gráficos, medidas de centro e de variabilidade) bastante usado no Controle da Qualidade Total em uma grande parte das indústrias nacionais e internacionais.

princípio ativo, o que, em ambos os casos, pode causar sérias complicações à saúde do paciente.

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 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 na área de 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	5,08	0,143 (14,3%)
Cobradores	50	22,6	3,11	0,137 (13,7%)

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 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 tem 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 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 naquela população. 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 de raças diferentes 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 duas 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

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 para competição, 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 tem uma força que varia pouco em torno de 2,2 ton/cm^3, enquanto os da raça II são, em média, 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[5].

Exemplo 3.7: Numa certa população, os recém-nascidos de gravidez considerada normal 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 do exemplo 3.7 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 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.

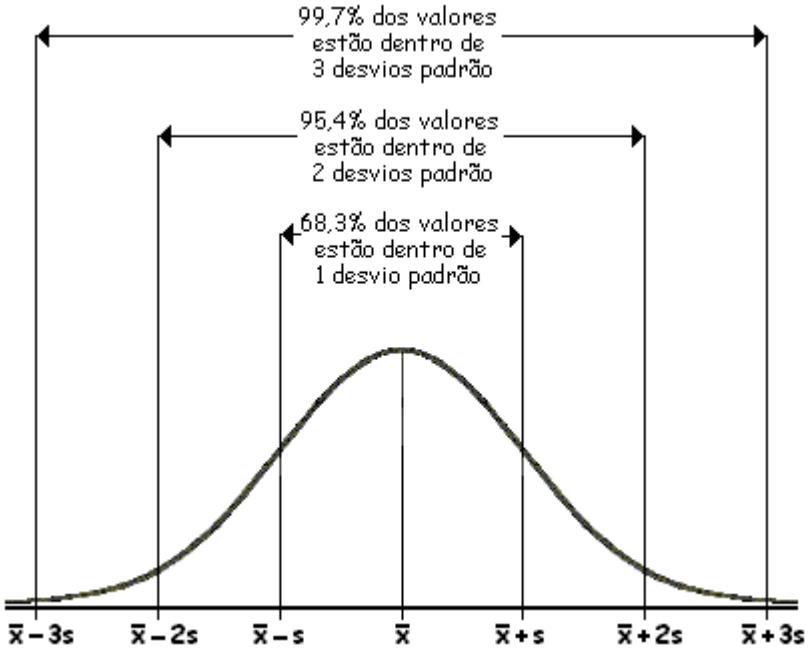
Outros exemplos de faixas de referência podem ser encontrados em resultados de exames clínicos, onde é apresentada, juntamente com o resultado do indivíduo, a faixa de referência (ou normalidade) para a característica em exame. Um indivíduo com resultado fora da faixa de

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

referência deve ser investigado. No exemplo dos recém-nascidos, aqueles que nascem com peso 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 1º, 2º , 3º, 29º 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

Considere o trecho a seguir, sobre a posição do Brasil, entre os países do mundo, quanto à renda per capita:

> *"O Brasil obviamente não é país rico, mas também não está entre os mais pobres. ...Mais de três quartos da população mundial vivem em países de renda per capita menor."*
> (Folha de São Paulo (12/12/2001), Editorial: Pobres Argumentos).

Neste caso, o posição do Brasil é dada pela quantidade de países que têm renda per capita menores que o Brasil, a saber três quartos ou 75%. O mesmo tipo de raciocínio fazemos quando dizemos que certo aluno está entre os 5% melhores do colégio. 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.

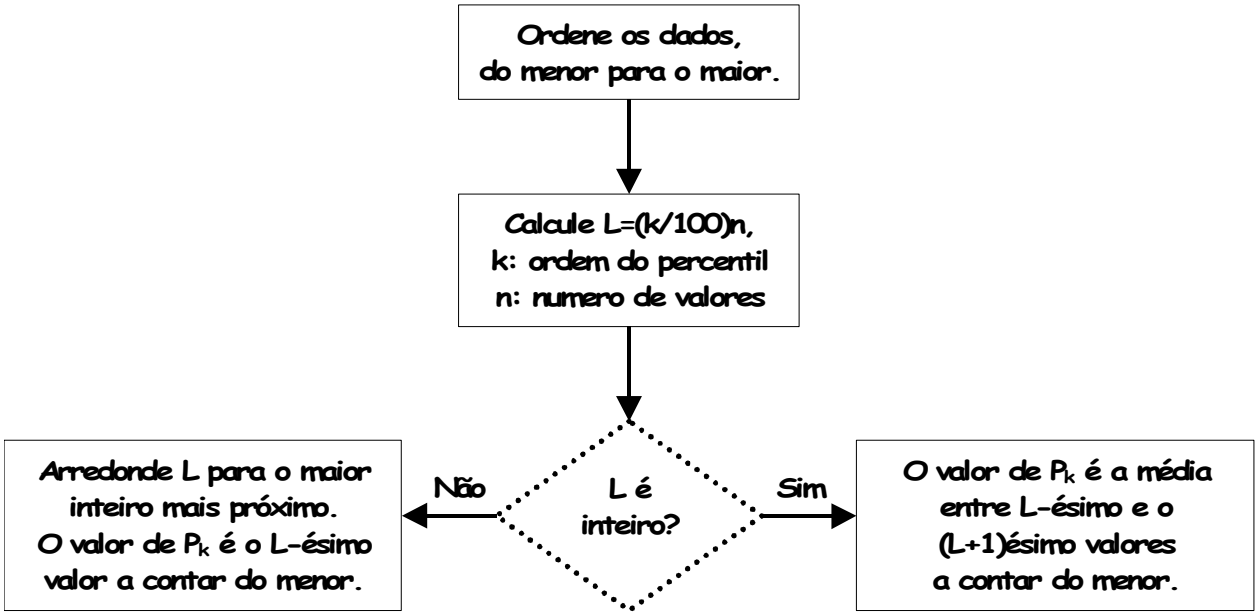
Assim, o percentil de ordem 10, o P_{10}, é o valor da variável tal que 10% dos valores são menores ou iguais a ele; o percentil de ordem 65 deixa 65% dos dados menores ou iguais a ele, etc.

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:

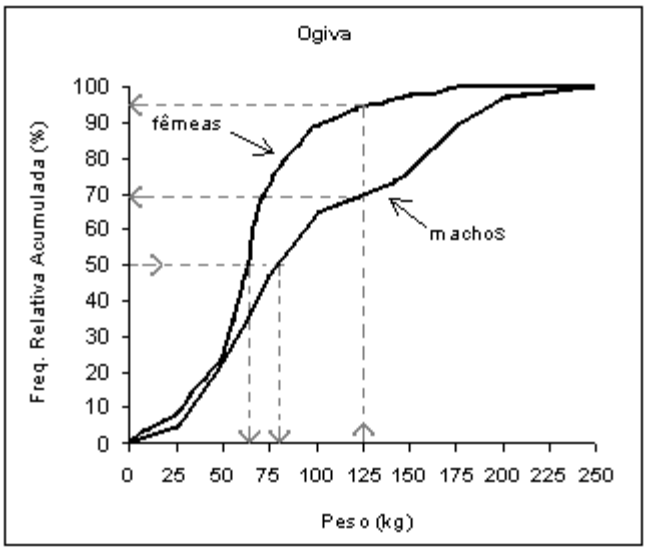
40	41	42	42	44	47	48	48	49	49	51	52	53	58	59	62	63	64	65	66
67	68	69	70	75	76	83	83	85	86	86	87	87	88	92	93	94	95	97	98

Vamos calcular alguns percentis.

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

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

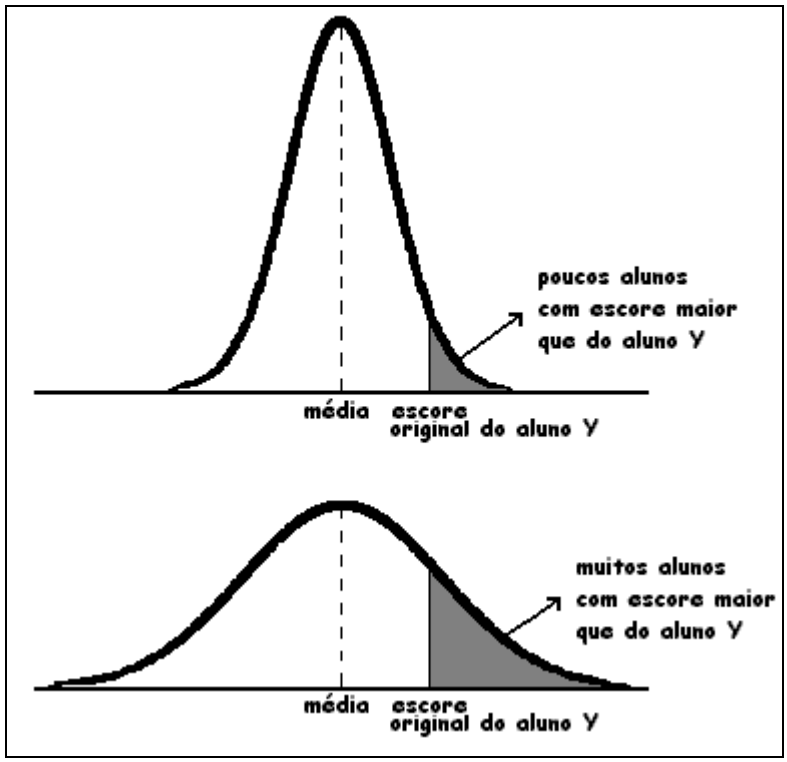
Inicialmente, podemos pensar em classificar o desempenho de um aluno em um teste como *bom* se ele se seu escore ficou acima da média da turma (abaixo, para o teste de natação), e como *ruim* se seu escore ficou abaixo da média da turma (acima, para o teste de natação).

	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.

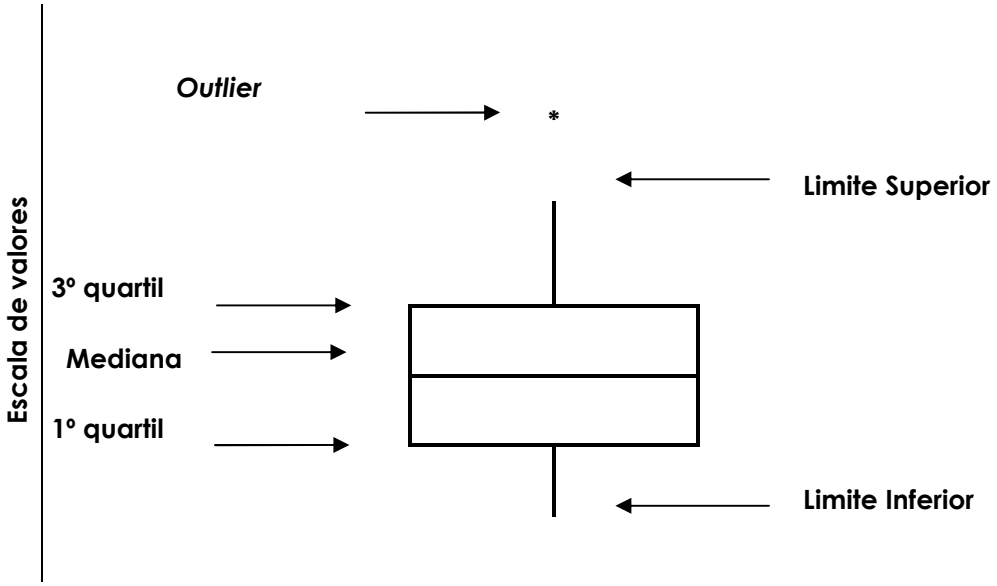
Esses valores discrepantes podem representar erros no processo de coleta ou de processamento dos dados, e, nesse caso, devem ser corrigidos ou excluídos do banco de dados.

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 comprimento máximo de 1,5 vezes a altura da caixa (DQ).

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.

Como foram calculado anteriormente, 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).

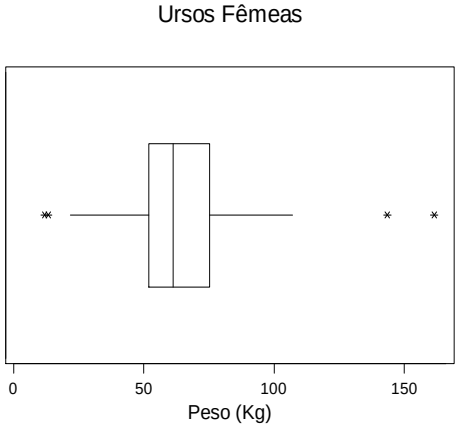
Para traçarmos as linhas laterais do boxplot, precisamos calcular a distância interquartílica (DQ), que é (75,4 – 51,8) = 23,6 . Desse modo, as linhas laterais podem ter comprimento máximo de 1,5 x 23,6 = 35,4. Vejamos até onde vão as linhas laterais:

- 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.

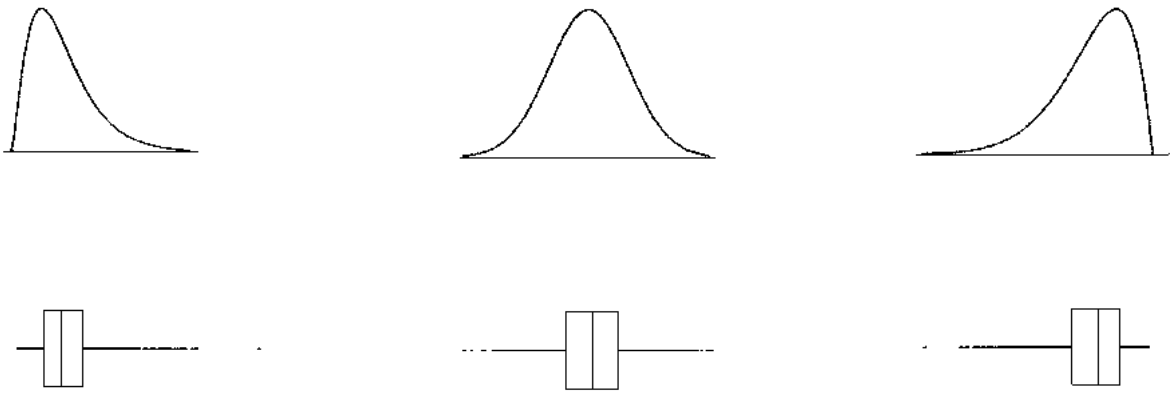
Através do boxplot, podemos:

- 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)**

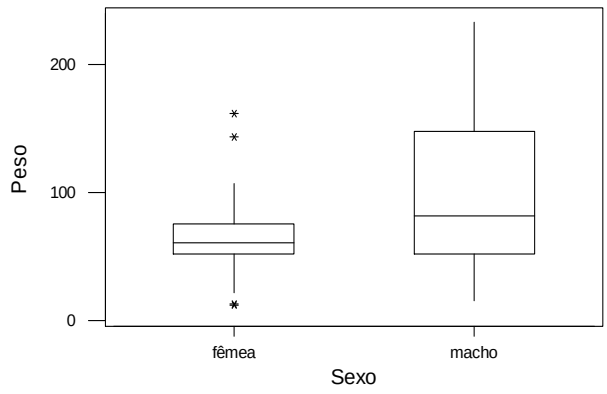
Observando a Figura 5.4, notamos que a distribuição do peso dos ursos fêmeas pode ser considerada simétrica em torno do valor mediano do peso (61,1Kg), enquanto a distribuição do peso dos ursos machos é assimétrica com concentração à esquerda (valores menores).

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 dos ursos fêmeas e machos



5.3. Vantagens e Desvantagem do Boxplot

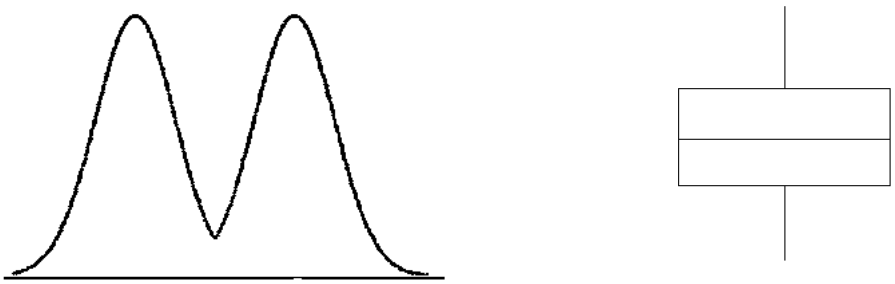
Além de permitir analisar a forma da distribuição de freqüências de um conjunto de valores, assim como a sua variabilidade e tendência central, o boxplot é uma forma mais prática de comparação entre dois ou mais grupos. Podemos representar vários boxplots numa mesma figura, enquanto isso não é possível 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[6].

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



[6] Na seção sobre histogramas (Parte I), 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:

			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	23
(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

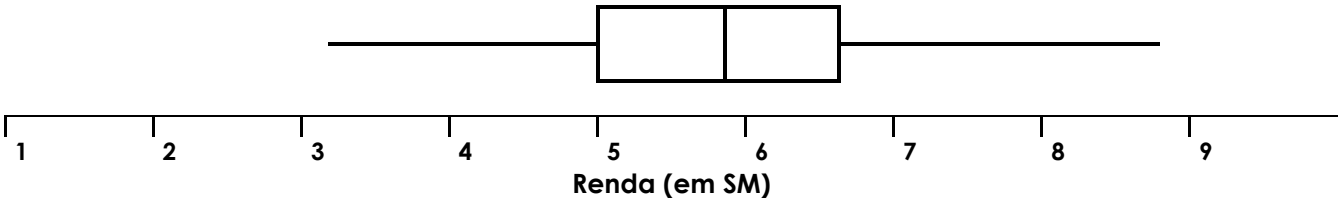
Dados necessários à construção do gráfico Boxplot:

Min = 3,1
Q_1 = 5,0
Mediana = 5,9
Q_3 = 6,5
Max = 8,8

Distância Interquartílica $DQ = Q_3 - Q_1 = 6,5 - 5,0 = 1,5$
$1,5DQ = 1,5(1,5) =$ **2,3**

Final da linha da esquerda $Q_1 - 1,5DQ = 5,0 - 2,3 =$ **2,7** < Min (o Min será o final)

Final da linha da direita $Q_3 + 1,5DQ = 6,6 + 2,3 =$ **8,9** > Max (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. O boxplot também reflete essa característica, posicionando sua caixa no meio da linha.

Nessa comunidade, metade das famílias recebem de 5 a 6,5 salários-mínimos e apenas 25% recebem mais de 8,8 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:

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

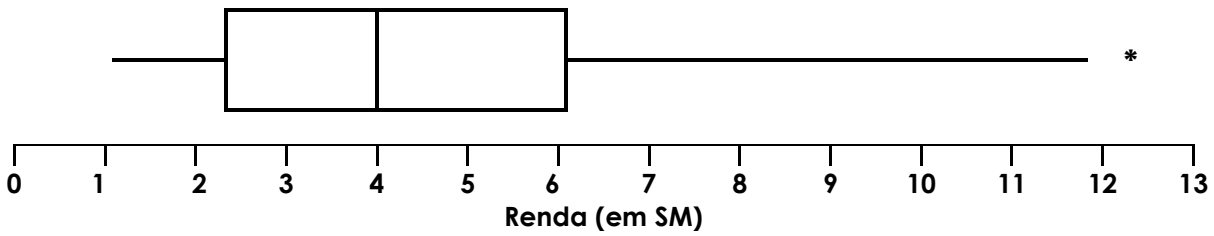
Legenda: 1|1 = 1,1 salários mínimos.

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**

$DQ = Q_3 - Q_1 = 6,1 - 2,3 = 3,8$
$1,5 \times DQ = 1,5 \times (3,8) = 5,7$
$Q_1 - 1,5DQ = 2,3 - 5,7 = -3,4$ (-3,4 é menor do que o Min; o Min=1,1 será o final da linha esquerda)
$Q_3 + 1,5DQ = 6,1 + 5,7 = 11,8$ (11,8 é menor do que o Max; 11,8 será o final da linha direita)

O valor máximo é um *outlier* e será marcado com um asterisco.

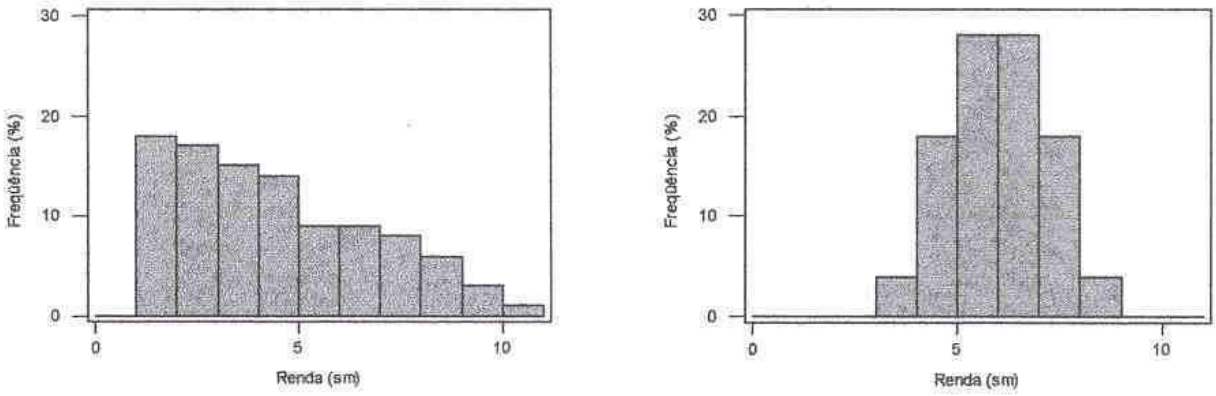


A distribuição de freqüências da renda familiar nessa comunidade é claramente assimétrica com concentração à esquerda, o que pode ser visto pelo ramo-e-folhas e pelo deslocamento da caixa do boxplot em direção ao lado esquerda da escala de valores. Do total de famílias dessa comunidade, metade ganha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

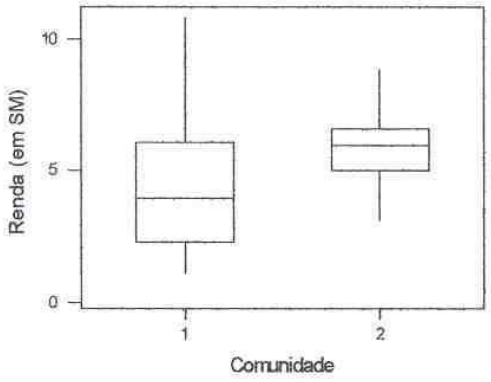
Ramo-e-folhas

	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

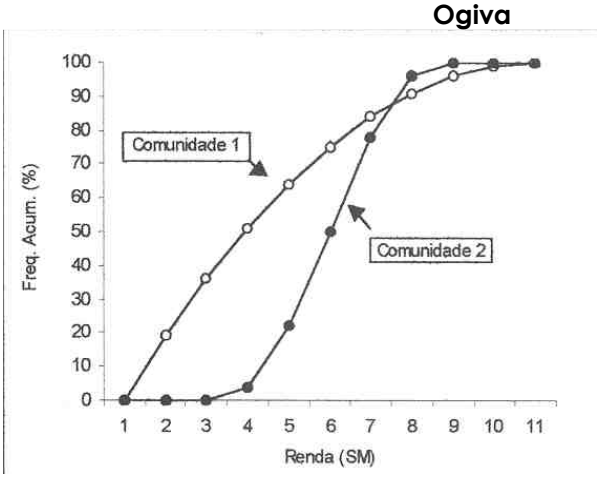
Legenda: | 1 | 2 : 1,2 salários-mínimos
1 | 3 | : 3,1 salários-mínimos

O ramo-e-folhas 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. and Simon, G.A. (2000) *Estatística Aplicada – Economia, Administração e Contabilidade*. 9ª Edição, Bookman, 404 pg, ISBN 85-7307-531-7.

- Huff, D. (1982) *How To Lie With Statistics*. W.W. Norton & Company, 142 pg, ISBN 0-393-31072-8.

- Lopes, P. A. (1999) *Probabilidades e Estatística*. Reichmann & Affonso Editores, 174 pg, ISBN 85-87148-07-9.

- MINITAB – Statistical Software, Release 13.30. Licenciado para Departamento de Estatística – UFMG.

- Peixoto, M.C.L., Braga, M.M. e Bogutchi, T.F. (2000) 'A Evasão no Ciclo Básico da UFMG´. *Cadernos de Avaliação 3*. Avaliação Institucional PAIUB-PROGRAD-UFMG, p. 7-28.

- Triola, M. F. (1999) *Introdução à Estatística* (tradução). 7ª edição, Editora LTC, 410 pg, ISBN 85-216-1154-4.

- Zeisel, H. (1985) *Say It With Figures*. 6ª edição, Harper & Row Publishers,272 pg, ISBN 0-06-181982-4.

Anexo I: Conjunto de Dados do Exemplo dos Ursos Marrons

	Nome	Mês da Obs.	Idade	Sexo	Cabeça Comprimento	Cabeça Largura	Pescoço Perímetro	Altura	Tórax Perímetro	Peso
1	Allen	jul	19	macho	25,4	12,7	38,1	114,3	58,4	29,5
2	Berta	jul	19	fêmea	27,9	16,5	50,8	120,7	61,0	31,8
3	Clyde	jul	19	macho	27,9	14,0	40,6	134,6	66,0	36,3
4	Doc	jul	55	macho	41,9	22,9	71,1	171,5	114,3	156,2
5	Quincy	set	81	macho	39,4	20,3	78,7	182,9	137,2	188,9
6	Kooch	out	*	macho	40,6	20,3	81,3	195,6	132,1	196,1
7	Charlie	jul	115	macho	43,2	25,4	80,0	182,9	124,5	158,0
8	Geraldine	ago	104	fêmea	39,4	16,5	55,9	157,5	88,9	75,4
9	Fannie	abr	100	fêmea	33,0	17,8	53,3	177,8	104,1	99,9
10	Dieter	jul	56	macho	38,1	19,1	67,3	186,7	104,1	118,9
11	John	abr	51	macho	34,3	20,3	68,6	174,0	124,5	163,4
12	Xeronda	set	57	fêmea	34,3	17,8	50,8	162,6	96,5	92,6
13	Clara	mai	53	fêmea	31,8	15,2	45,7	147,3	78,7	65,4
14	Abe	jun	*	macho	30,5	21,1	47,0	153,2	81,3	55,4
15	Eugene	ago	68	macho	40,6	22,9	73,7	185,4	111,8	150,7
16	Floyd	ago	8	macho	22,9	11,4	33,0	94,0	48,3	15,4
17	Kim	ago	44	fêmea	31,8	11,4	26,7	160,0	81,3	63,6
18	Ichabod	ago	32	macho	35,6	12,7	54,6	170,2	94,0	81,7
19	Lorie	ago	20	fêmea	29,2	12,7	44,5	132,1	73,7	47,7
20	Mighty	ago	32	macho	33,0	20,3	54,6	149,9	83,8	75,4
21	Oliver	set	45	macho	34,3	17,8	61,0	162,6	99,1	92,6
22	Ness	set	9	fêmea	22,9	11,4	30,5	91,4	48,3	11,8
23	Pete	set	21	macho	33,0	15,2	48,3	149,9	76,2	54,5
24	Robert	set	177	macho	40,6	24,1	76,2	182,9	121,9	197,9
25	Smokey	set	57	fêmea	31,8	12,7	48,3	146,1	81,3	56,8
26	Tozia	set	81	fêmea	33,0	12,7	50,8	154,9	83,8	59,9
27	Unser	set	21	macho	33,0	12,7	43,2	137,2	71,1	40,9
28	Viking	set	9	macho	25,4	10,2	33,0	101,6	58,4	18,2
29	Walt	set	45	macho	40,6	15,2	61,0	160,0	106,7	99,9
30	Xavier	set	9	macho	25,4	10,2	34,3	109,2	58,4	20,9
31	Yogi	set	33	macho	34,3	15,2	55,9	168,9	86,4	69,9
32	Zelda	set	57	fêmea	33,0	14,0	44,5	153,7	78,7	52,7
33	Allison	set	45	fêmea	33,0	16,5	53,3	152,4	87,6	82,6
34	Buck	set	21	macho	36,8	14,0	50,8	154,9	86,4	68,1
35	Christophe	out	10	macho	24,1	11,4	40,6	101,6	66,0	29,5
36	Diane	out	82	fêmea	34,3	16,5	71,1	162,6	121,9	161,6
37	Edith	out	70	fêmea	36,8	16,5	66,0	165,1	121,9	143,5
38	Gary	out	10	macho	27,9	12,7	43,2	124,5	73,7	42,7
39	Herman	out	10	macho	29,2	12,7	43,2	119,4	74,9	39,0
40	Jim	out	34	macho	33,0	17,8	53,3	149,9	88,9	68,1
41	Ken	out	34	macho	41,9	16,5	68,6	182,9	113,0	122,6
42	Leon	out	34	macho	35,6	14,0	61,0	165,1	99,1	91,7
43	Noreen	out	58	fêmea	34,3	16,5	54,6	160,0	101,6	91,7
44	Orville	out	58	macho	39,4	17,8	71,1	179,1	127,0	165,7
45	Pasquale	nov	11	macho	29,2	15,2	41,9	121,9	78,7	35,9
46	Rich	nov	23	macho	30,5	16,5	48,3	127,0	96,5	67,2
47	Ian	out	70	macho	39,4	17,8	71,1	194,3	139,7	202,5
48	Suzie	nov	11	fêmea	22,9	12,7	38,1	116,8	68,6	28,1

Continua...

....Continuação

	Nome	Mês da Obs.	Idade	Sexo	Cabeça Comprimento	Cabeça Largura	Pescoço Perímetro	Altura	Tórax Perímetro	Peso
49	Thelma	nov	83	fêmea	36,8	17,8	58,4	156,2	111,8	107,1
50	U-Sam	nov	35	macho	34,3	21,6	58,4	161,3	111,8	96,2
51	Bill	abr	*	macho	47,0	21,6	59,7	171,5	106,7	92,6
52	Wille	abr	16	macho	25,4	10,2	39,4	121,9	66,0	27,2
53	XRay	abr	16	macho	25,4	12,7	38,1	104,1	66,0	29,1
54	Vanessa	abr	*	fêmea	33,0	17,8	53,3	149,9	86,4	66,3
55	Zack	abr	*	macho	39,4	22,9	73,7	200,7	127,0	181,6
56	Albert	abr	*	macho	34,3	17,8	62,2	157,5	104,1	112,6
57	*	ago	*	macho	40,6	22,9	80,0	190,5	119,4	158,9
58	*	mai	17	macho	29,2	12,7	43,2	134,6	77,5	51,8
59	Denise	mai	17	fêmea	29,2	12,7	38,1	133,4	71,1	34,5
60	Evelyn	mai	17	fêmea	27,9	11,4	33,0	116,8	58,4	21,8
61	Fran	mai	*	fêmea	30,5	15,2	48,3	144,8	87,6	67,2
62	Gert	mai	*	fêmea	34,3	12,7	43,2	147,3	73,7	51,8
63	Michele	jun	*	fêmea	34,3	12,7	43,2	147,3	74,9	52,7
64	Villager	jun	*	macho	35,6	16,5	53,3	160,0	88,9	89,9
65	Sally	jun	*	fêmea	30,5	12,7	48,3	148,6	85,1	51,8
66	Mary	jun	*	fêmea	33,0	15,2	44,5	154,9	83,8	61,3
67	Sonny	jul	*	macho	36,8	16,5	54,6	162,6	94,0	81,7
68	Davy	jul	*	macho	30,5	16,5	47,0	141,0	69,9	49,9
69	Patty	jul	*	fêmea	27,9	12,7	39,4	123,2	64,8	35,9
70	Friday	ago	*	macho	36,8	15,2	57,2	170,2	101,6	98,1
71	Swartz	ago	*	macho	38,1	20,3	67,3	180,3	108,0	137,1
72	Ann	ago	*	fêmea	30,5	15,2	48,3	135,9	81,3	55,4
73	Tiffy	ago	*	macho	43,2	22,9	74,9	177,8	115,6	146,2
74	Ralph	ago	*	macho	39,4	20,3	50,8	160,0	83,8	69,9
75	Bronson	ago	*	macho	30,5	15,2	45,7	168,9	86,4	66,3
76	Eddie	ago	*	macho	44,5	20,3	76,2	210,8	124,5	179,8
77	Ozz	ago	*	macho	33,0	12,7	45,7	141,0	77,5	55,4
78	Margie	ago	*	fêmea	33,0	14,0	49,5	156,2	94,0	70,8
79	Pam	ago	*	fêmea	31,8	15,2	49,5	148,6	81,3	64,5
80	Addy	ago	8	fêmea	25,4	11,4	25,4	110,5	61,0	13,2
81	Curt	ago	*	macho	41,9	21,6	74,9	175,3	125,7	158,0
82	Kermit	set	*	macho	43,2	21,6	77,5	201,9	123,2	167,1
83	Paul	set	*	macho	30,5	14,0	45,7	138,4	81,3	52,7
84	Frieda	set	*	fêmea	35,6	17,8	53,3	168,9	94,0	72,6
85	Chet	set	*	macho	33,0	16,5	52,1	152,4	92,7	69,9
86	Brander	out	*	macho	40,6	19,1	71,1	185,4	114,3	143,5
87	Louise	out	*	fêmea	34,3	14,0	49,5	154,9	88,9	71,7
88	Nan	nov	*	fêmea	31,8	14,0	48,3	142,2	81,3	54,5
89	Ian	nov	83	macho	39,4	20,3	77,5	190,5	137,2	233,4
90	Larry	nov	*	macho	39,4	19,1	64,8	186,7	109,2	147,1
91	Scott	nov	*	macho	36,8	17,8	55,9	171,5	96,5	89,0
92	Grizz	jun	18	macho	31,8	21,6	45,7	145,5	83,3	63,6
93	Sara	ago	*	fêmea	30,5	12,7	45,7	142,2	82,6	51,8
94	Lou	ago	*	macho	30,5	14,0	38,1	129,5	61,0	37,2
95	Molly	ago	*	fêmea	33,0	15,2	55,9	154,9	101,6	104,4
96	Graham	jul	*	macho	30,5	10,2	44,5	149,9	72,4	58,1
97	Jeffrey	jul	*	macho	34,3	15,2	50,8	157,5	82,6	70,8

Anexo II: Passos para Construção da Tabela de Distribuição de Freqüências de uma Variável Contínua

1- Encontre o **menor** e o **maior valor** das observações;

2- Determine o **tamanho das classes** (geralmente, valores múltiplos de 5 ou 10 facilitam a interpretação dos resultados posteriormente);

3- Construa as classes, lembrando-se de começar antes do valor mínimo e terminar depois do valor máximo;

4- Lembre-se de que o número de classes está associado ao

- Tamanho de classe escolhido. Uma tabela de freqüência não deve ter:
  - menos de 6 classes (muito resumida),
  - mais de 15 classes (muito dispersa);

- Número de observações. Um grande número de observações pode ser distribuído em muitas classes, mas um pequeno número de observações requer poucas classes;

5- Em todas as etapas da construção das classes deve prevalecer o **bom senso*:***
se a primeira distribuição de freqüências construída não ficou boa (muito resumida ou muito dispersa), aumente ou diminua o número de classes, diminuindo ou aumentando o tamanho delas.

Exemplo: Construção de tabela de distribuição de freqüências

Quadro AII.1 - Emissões de Óxido de Enxofre (em toneladas) de uma indústria em 70 dias.

15,8	26,4	17,3	11,2	23,9	24,8	18,7	13,9	9,0	13,2
22,7	9,8	6,2	14,7	17,5	26,1	12,8	28,6	17,6	23,7
26,8	22,7	18,0	20,5	11,0	20,9	15,5	19,4	16,7	10,7
19,1	15,2	22,9	26,6	20,4	21,4	19,2	21,6	16,9	19,0
18,5	23,0	24,6	20,1	16,2	18,0	7,7	13,5	23,5	14,5
8,3	21,9	12,3	22,3	13,3	11,8	19,3	20,0	25,7	31,8
25,9	10,5	15,9	27,5	18,1	17,9	9,4	24,1	20,1	28,5

1 - min = 6,2 max = 31,8

2 - Tamanho de classe: 5 toneladas;

3 - 1ª classe: 5,0 |- 10,0
2ª classe: 10,0 |- 15,0
3ª classe: 15,0 |- 20,0
4ª classe: 20,0 |- 25,0
5ª classe: 25,0 |- 30,0
6ª classe: 30,0 |- 35,0

Tabela AII.1 – Distribuição de Freqüências das Emissões de Óxido de Enxofre (em toneladas) de uma indústria em 70 dias.

Emissão de Óxido de Enxofre (toneladas)	Freqüência Absoluta	Freqüência Relativa (%)	
5,0 |- 10,0	6	8,6	
10,0 |- 15,0	13	18,6	
15,0 |- 20,0	21	30,0	
20,0 |- 25,0	20	28,5	(28,6)*
25,0 |- 30,0	9	12,9	
30,0 |- 35,0	1	1,4	
Total	70	100,0	

* Devido a erros de arredondamento, muitas vezes a soma das freqüências relativas não fecha nos 100% exatos, somando 99,9% ou 100,1%. Quando isso ocorrer, o ajuste (somar ou subtrair 0,1%) deve ser feito na classe de maior freqüência, se possível. Se não, faz-se o ajuste na classe de segunda maior freqüência e assim por diante. Nesse exemplo, a soma das freqüências relativas é igual a 100,1%. Se fizéssemos o ajuste na categoria de maior freqüência (30%), ficaria estranho, pois 21÷70 é exatamente 0,30. Desse modo, preferimos fazer o ajuste na classe com a segunda maior freqüência (28,6%), ajustando-a para 28,5%.