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

Análise Descritiva de Dados - Tabelas e Gráficos

Edna A. Reis e Ilka A. Reis

Relatório Técnico
RTE-04/2001

Relatório Técnico
Série Ensino

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

Análise Descritiva de Dados

Tabelas e Gráficos

Edna Afonso Reis
Ilka Afonso Reis

Primeira Edição – Outubro/2001

ÍNDICE

1. Introdução

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

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

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 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 uma 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 uma outra medida mais fácil de se 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.

A planilha de dados dos ursos tem 97 linhas e 10 colunas (veja Anexo). Alguns ursos não tiveram sua idade determinada. Esses dados são chamados *dados faltantes* e é comum representá-los por asteriscos (na verdade, cada *software* tem sua convenção para representar *missing data*).

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

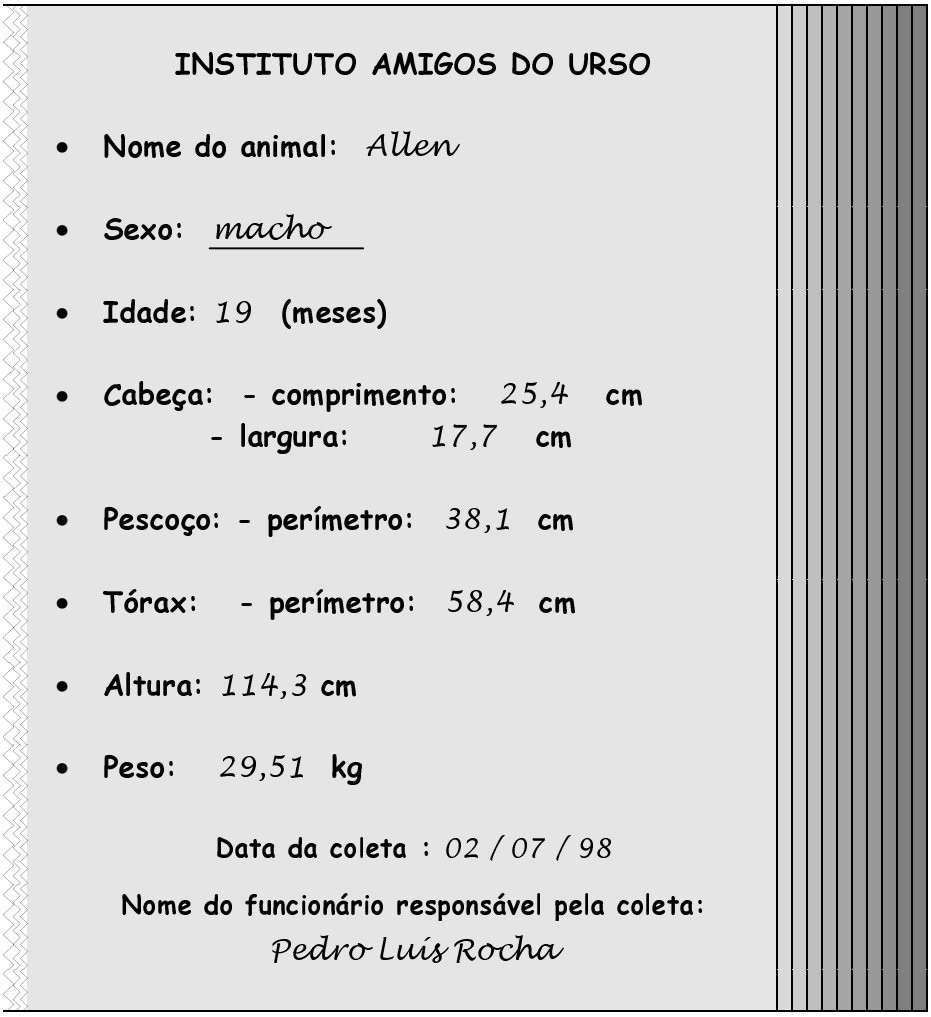


Figura 2.2 – Representação parcial da planilha de dados do exemplo dos ursos.

V A RI Á V E I S →

E L E M E N T O S ↓		**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 elemento da amostra ou população. Como o nome diz, seus valores *variam* de elemento para elemento. 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 fracionais 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 dentre 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 do 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 ocorrem 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, isto é, não há uma ordem natural em suas categorias, a ordem das linhas da tabela pode ser qualquer uma.

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	100,0

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.

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

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	100,0
Total	97	100,0	-----	-----

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.

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*.

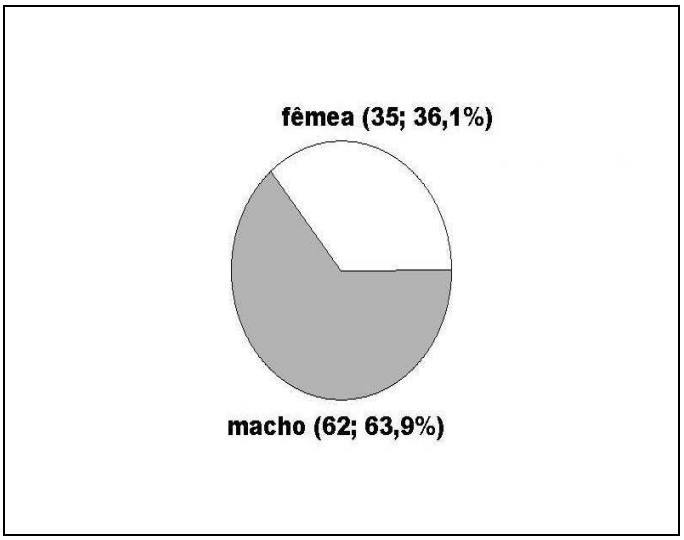
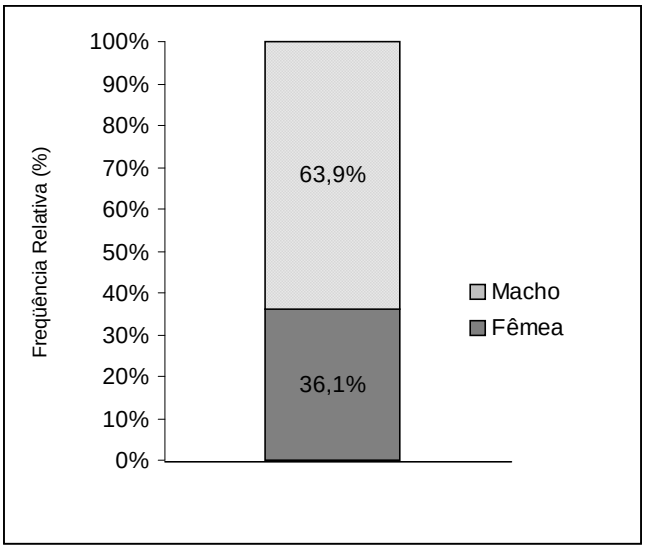


Figura 4.2 – Gráfico de barras para a variável *sexo*.



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 tratam-se 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.

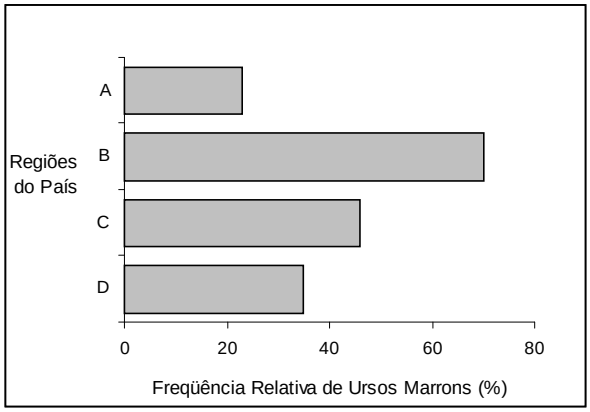
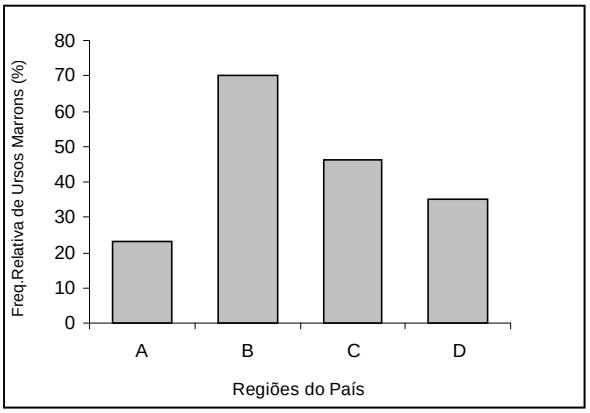


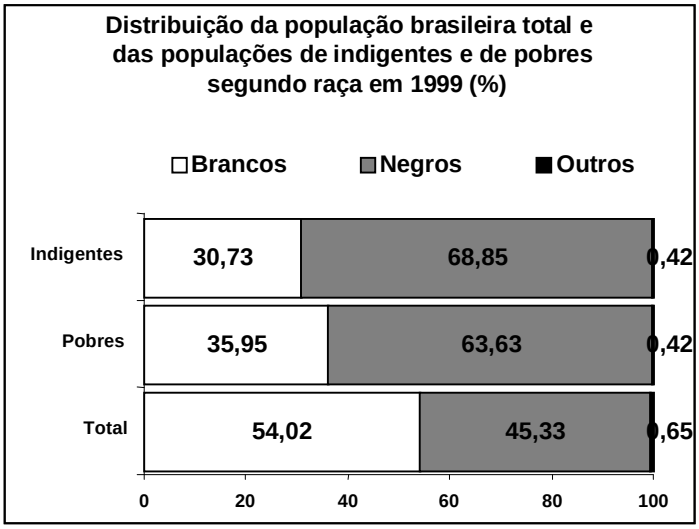
Figura 4.4 – Gráfico de barras horizontais para a freqüência de ursos marrons em quatro regiões.



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.

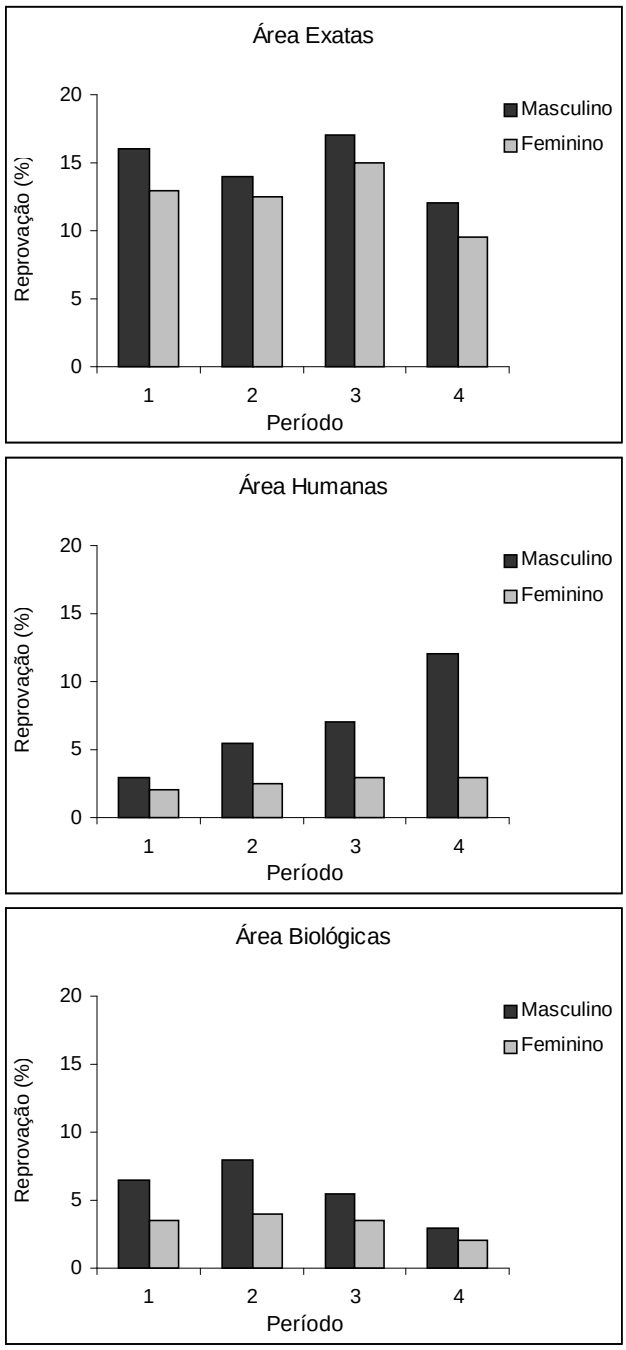
Analisando os três gráficos da Figura 4.6, podemos notar que o percentual de reprovação entre os alunos do sexo masculino é sempre maior do que o percentual de reprovação entre os alunos do sexo feminino, em todas as áreas, durante todos os períodos. A área de ciências exatas é a que possui os maiores 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 tipo.

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	100,0
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%) têm um ou nenhum filho.

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

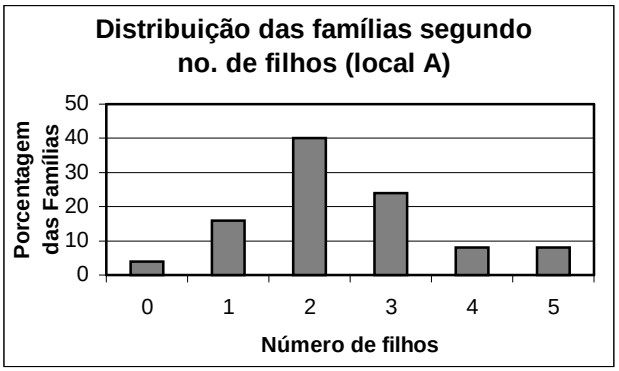
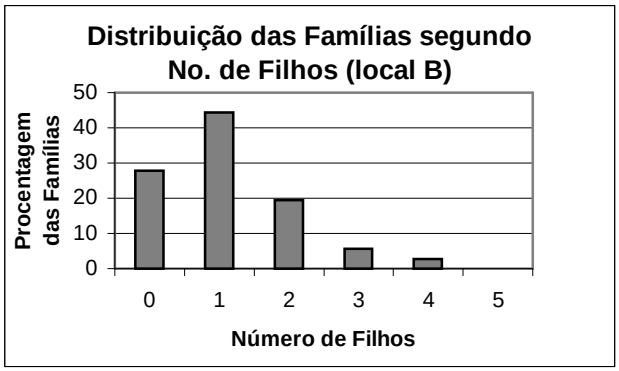


Figura 4. 8: Distribuição de freqüências do número de filhos por família na localidade B (36 lares).



ⓘ **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 histograma. Deve-se, também usar a mesma escala em todos os histogramas, tanto na escala vertical quanto na horizontal.

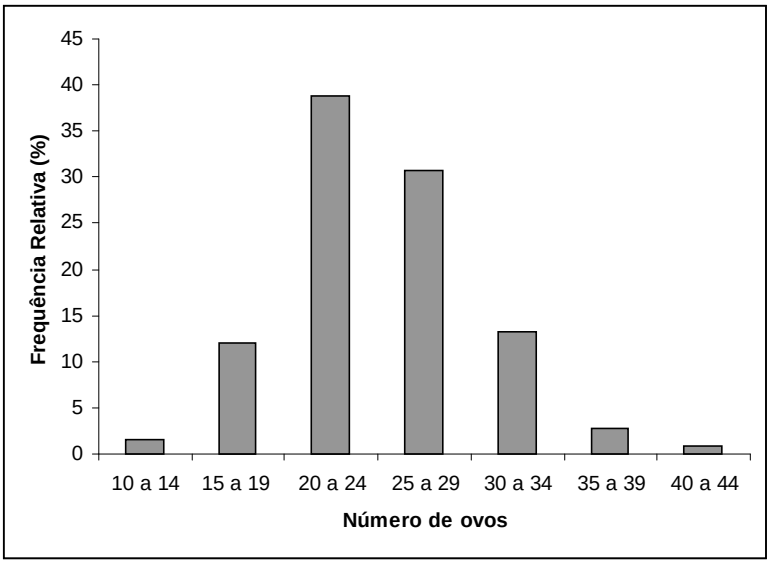
Quando trabalhamos com uma **variável discreta que pode assumir um grande número de valores distintos** como, por exemplo, o número de ovos que um inseto põe durante sua vida, a construção da tabela de freqüências e de gráficos considerando cada valor como uma categoria fica inviável. A solução é agrupar 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	100,0
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 da tamanho 10, o que reduziria para quatro o número de classes (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, pode 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).

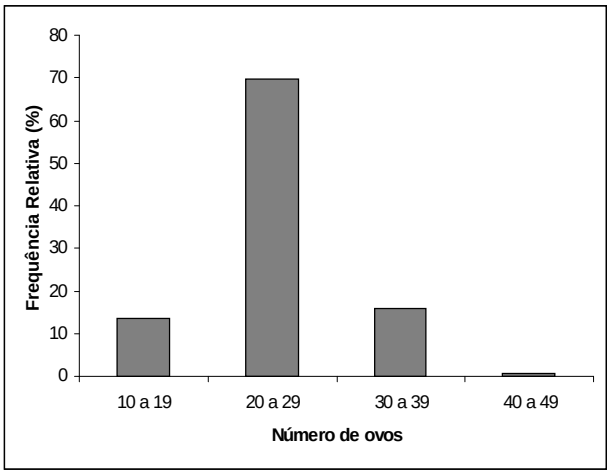
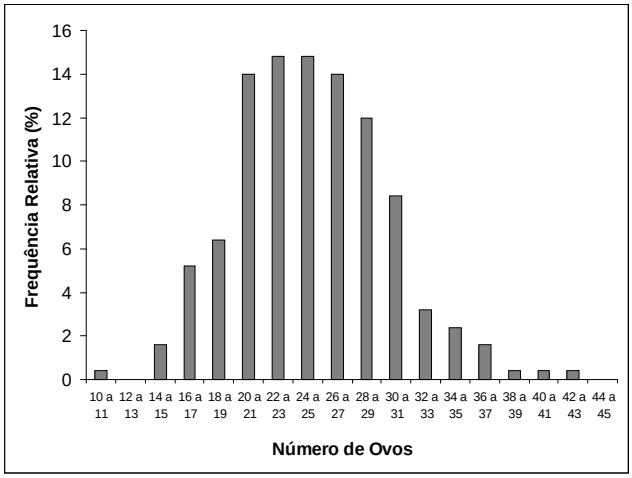


Figura 4. 11: Distribuição de freqüências do número de ovos postos por 250 insetos (classes de tamanho 2).

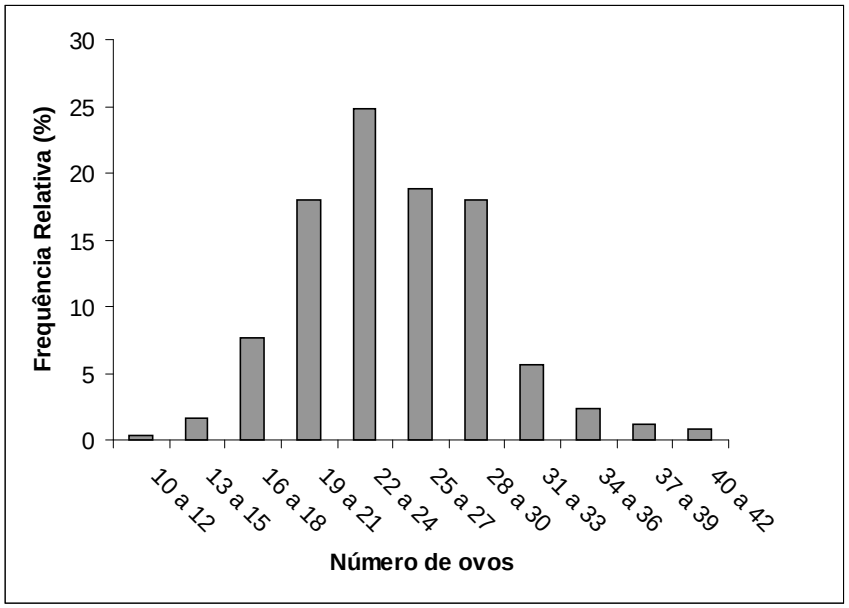


Os limites inferiores e superiores de cada classe dependem do tamanho (amplitude) de classe escolhido, que deve ser, na medida do possível, igual para todas as classes. Isso facilita a interpretação da distribuição de freqüências da variável em estudo.

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	100,0
Total	62	100,0	-	-

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.

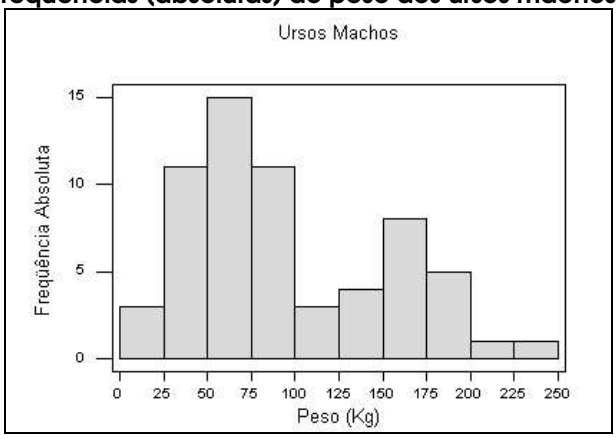
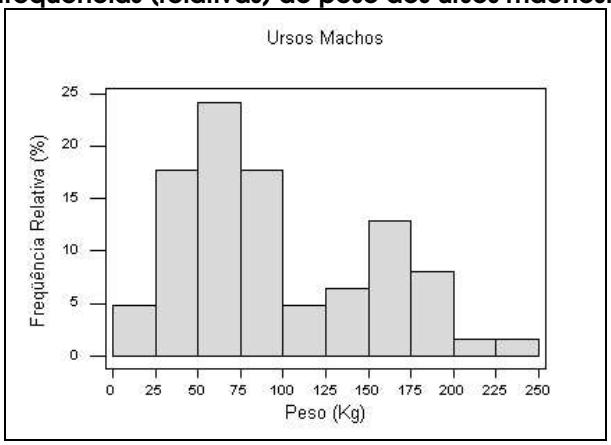


Figura 4.14: Histograma para a distribuição de freqüências (relativas) do peso dos ursos machos.



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	100,0	-	-

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.

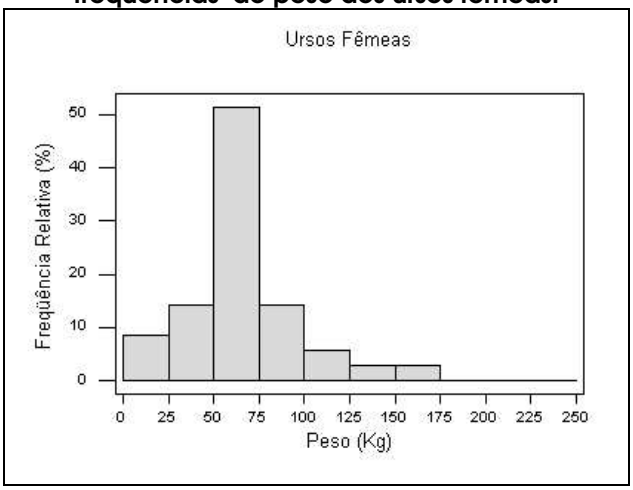
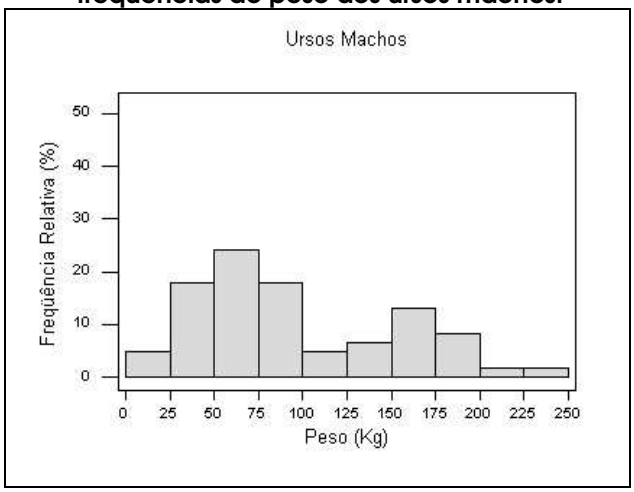
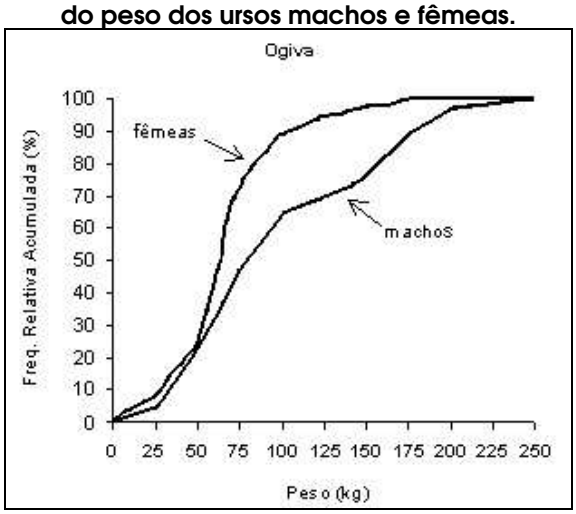


Figura 4.16: Histograma para a distribuição de freqüências do peso dos ursos machos.



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.



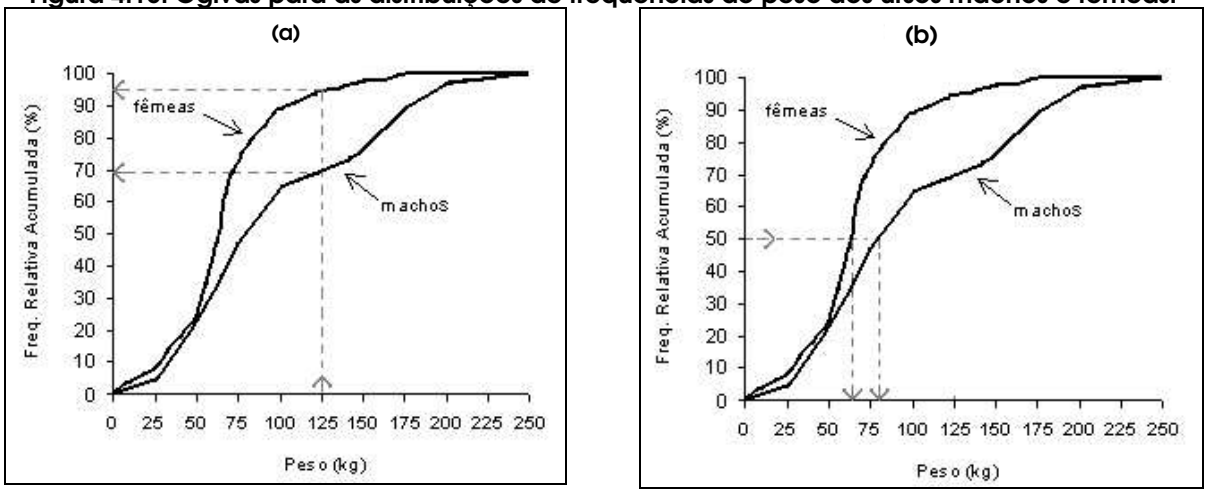
O primeiro ponto da ogiva é formado pelo limite inferior da primeira classe e o valor zero, indicando que abaixo do limite inferior da primeira classe não existem observações. Daí por diante, são usados os limites superiores das classes e suas respectivas freqüências acumuladas, até a última classe, 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 têm peso de até 125 Kg?
   Na Figura 4.18(a), traçamos uma linha vertical partindo do ponto 120 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 do valores de peso, encontrando o valor de 80 kg para os machos e 65 kg para as fêmeas. Assim, metade dos machos pesam 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.

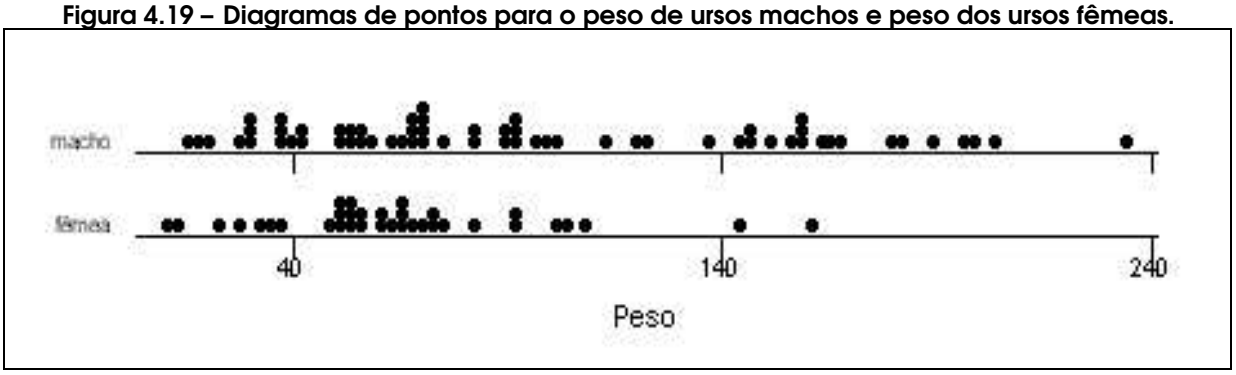
Para os casos em que o número de dados é pequeno, uma alternativa para a visualização da distribuição desses dados são os gráficos denominados **diagrama de pontos** e **diagrama de ramo-e-folhas**.

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 mostrado 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*. A sua 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 pela 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

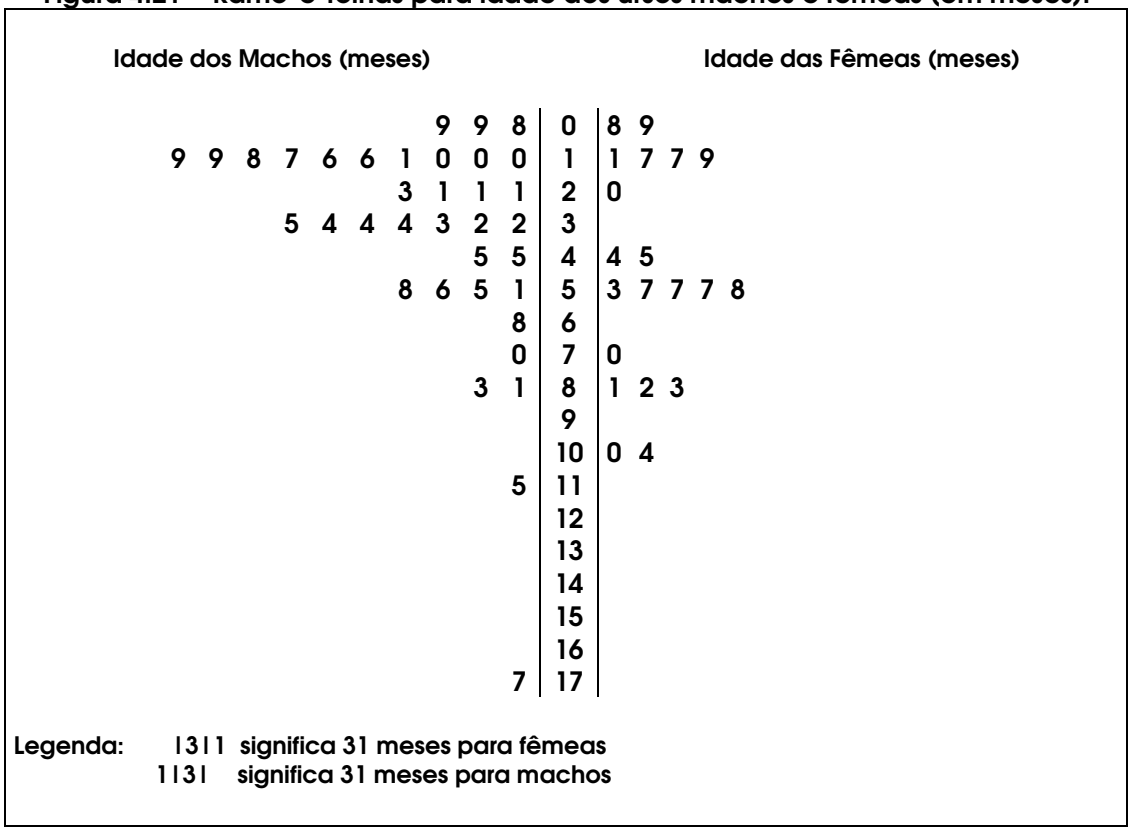
Os ramos e as folhas podem representar quaisquer unidades de grandeza (dezenas e unidades, centenas e dezenas, milhares e centenas, etc). Para sabermos o que está sendo representado, um ramo-e-folhas deve ter sempre uma legenda, indicando o que significam os ramos e as folhas. Se a idade estivesse medida 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.

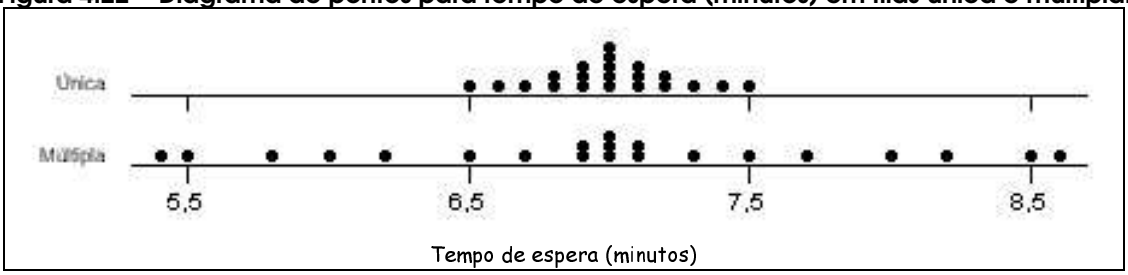
Dizemos que a distribuição de freqüências do peso dos ursos fêmeas é **unimodal** (apenas uma moda) e dos ursos machos é **bimodal** (duas modas). Geralmente, um histograma bimodal indica a existência de dois grupos, com valores centrados em dois pontos diferentes do eixo de valores. Uma distribuição 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:

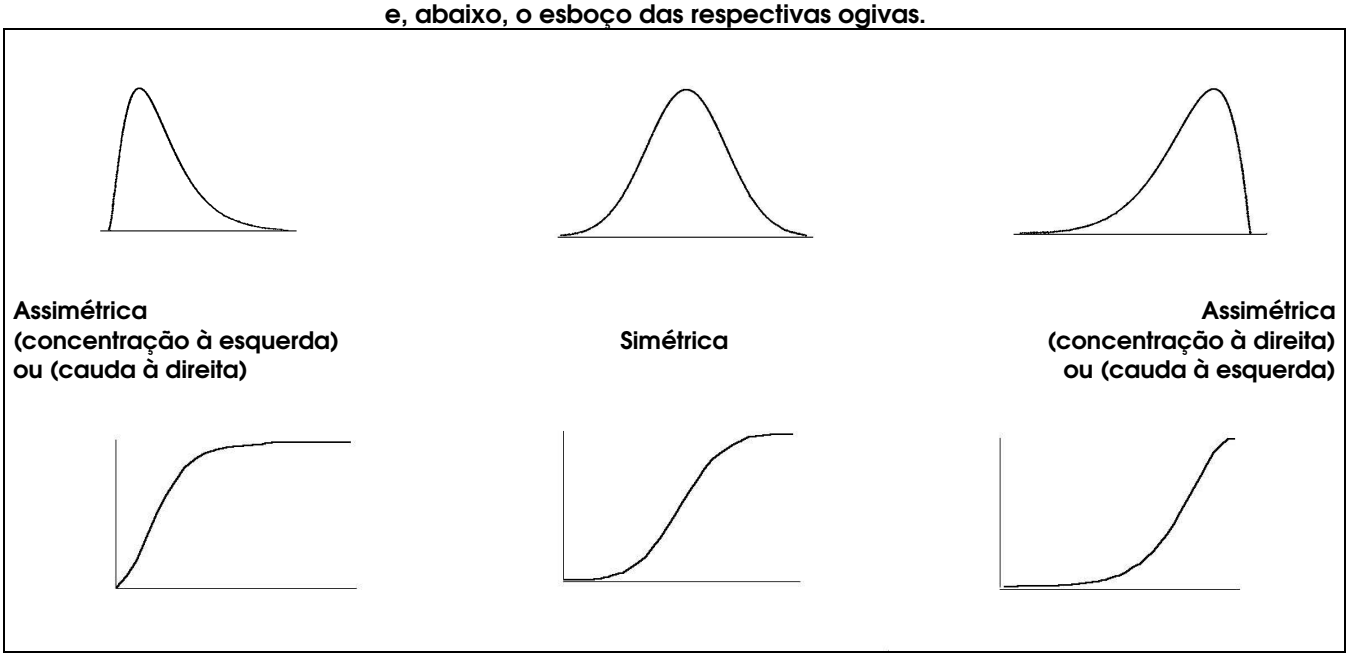
- quando os valores concentram-se à esquerda (assimetria com concentração à esquerda ou assimetria com cauda à direita);
- quando os valores concentram-se à direita (assimetria com concentração à direita ou com assimetria cauda à esquerda);

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 ?

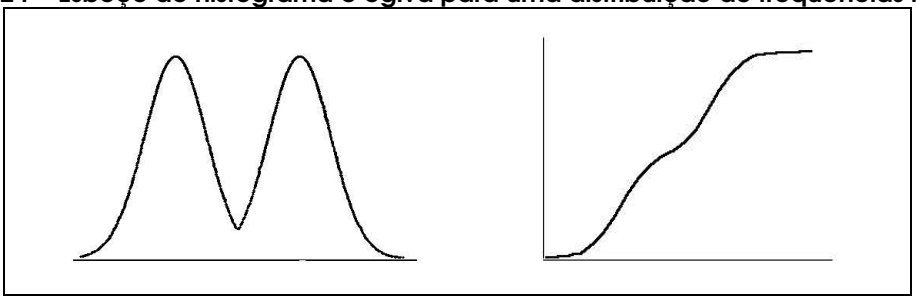
Figura 4.23 – Formas básicas para a distribuição de freqüências de uma variável quantitativa e, abaixo, o esboço das respectivas ogivas.



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, o 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.

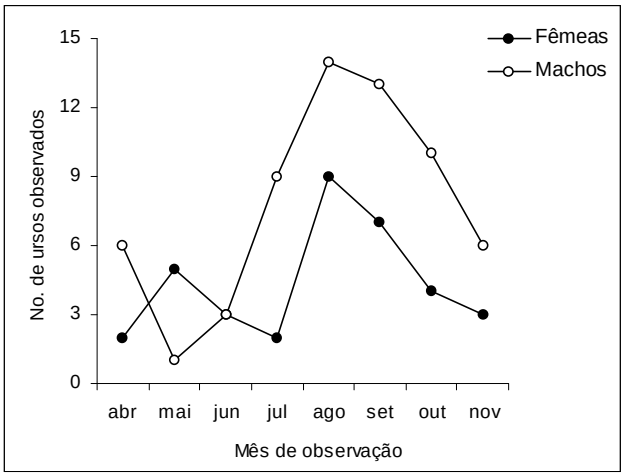
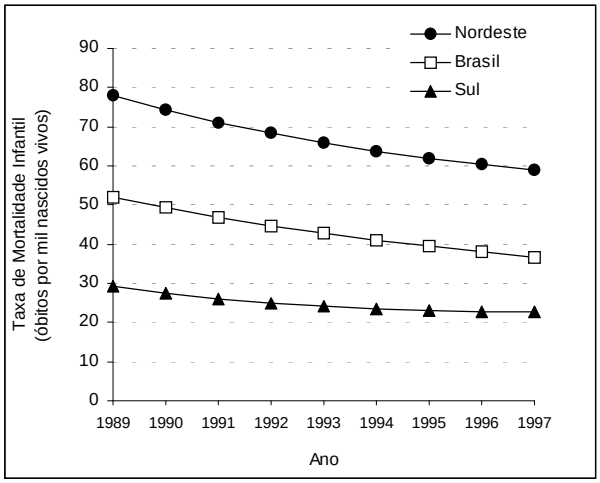


Figura 5.2 – Gráfico de linha para a taxa de mortalidade infantil de 1989 a 1997 nas Regiões Nordeste e Sul e no Brasil.



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.

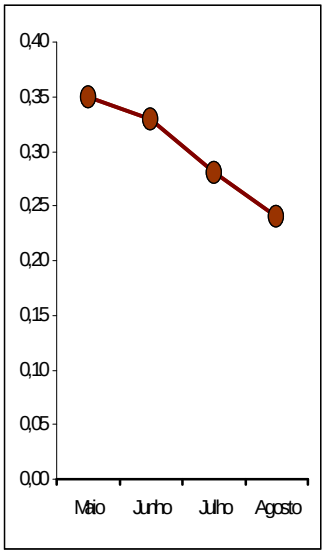
Outro aspecto mascarado pela falta da escala é que as diferenças entre os valores numéricos não correspondem às distâncias representadas no gráfico. Por exemplo, no gráfico de linha divulgado para a série do preço do leite, vemos que a queda no preço de maio para junho foi de R$0,02 e, de julho para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

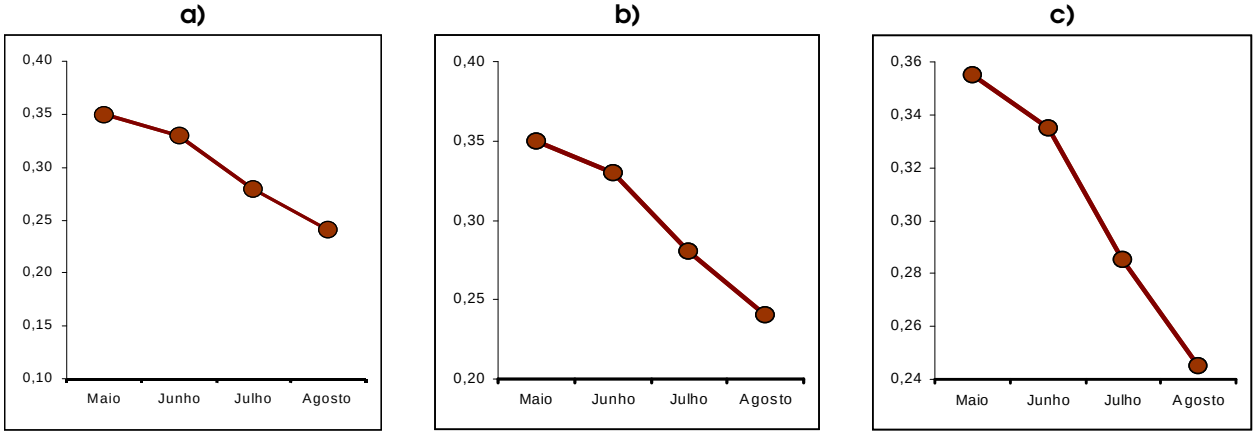
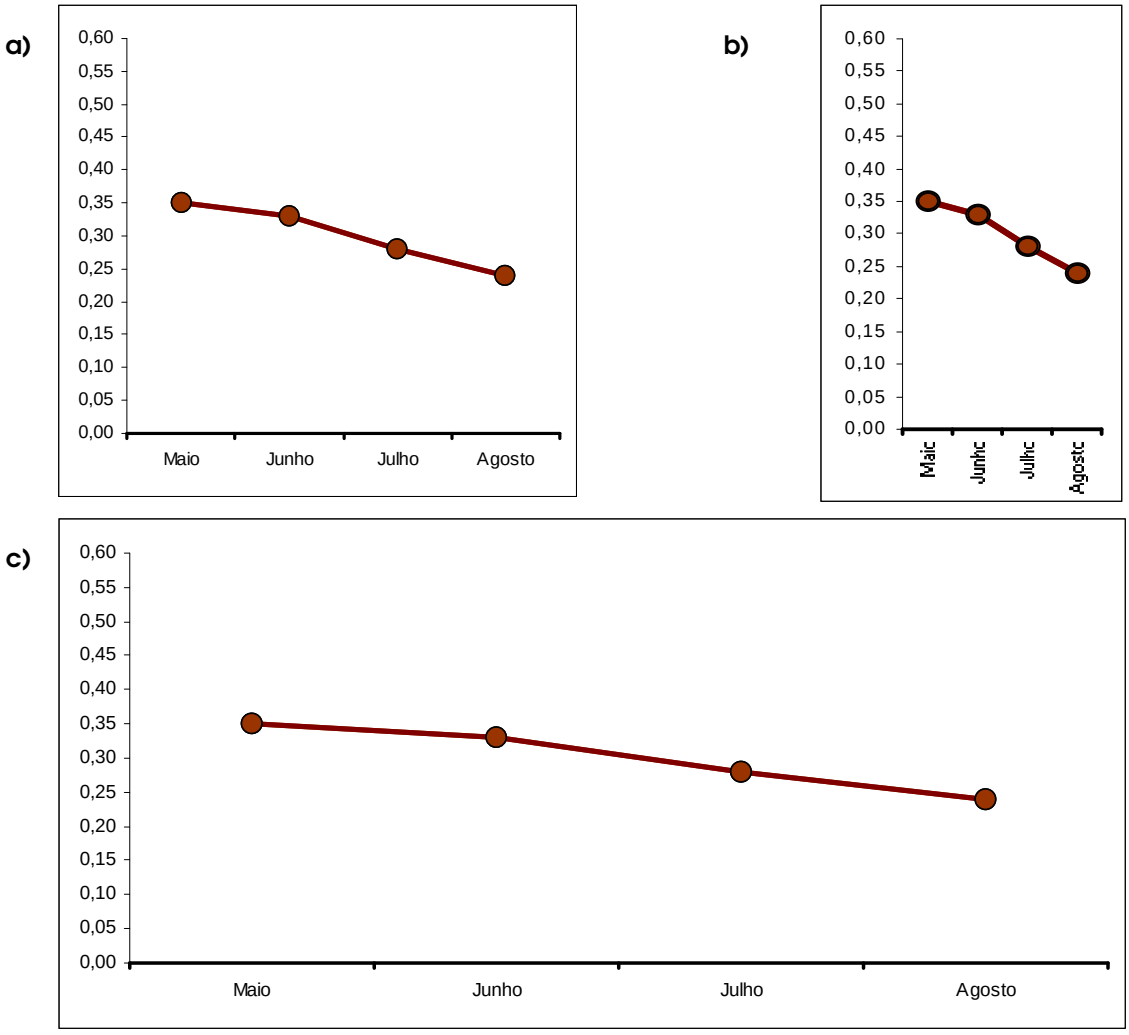


Figura 5.6 - Efeitos de alterações na dimensão horizontal do gráfico de linhas da série do preço do leite



6. O Diagrama de Dispersão

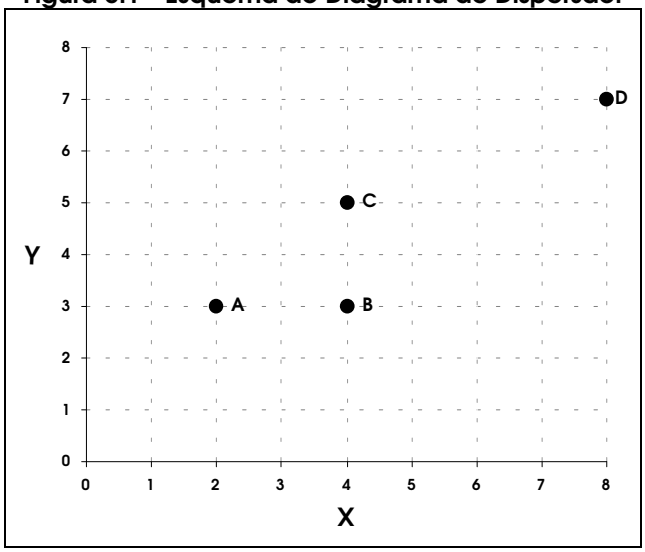
O diagrama de dispersão é um gráfico onde pontos no espaço cartesiano XY são usados para representar simultaneamente os valores de duas variáveis quantitativas medidas em cada elemento do conjunto de dados.

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 se 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.

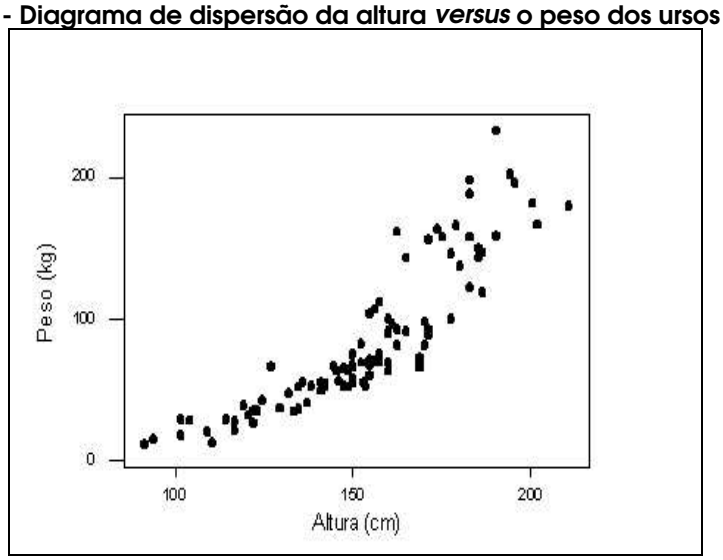
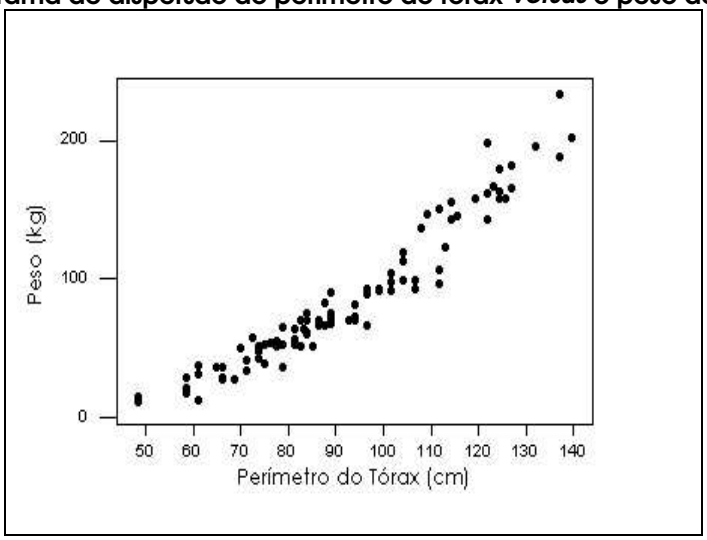


Figura 6.3 - Diagrama de dispersão do perímetro do tórax *versus* o peso dos ursos marrons.



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 chama-se Regressão Linear Simples).

Exemplo dos morangos.

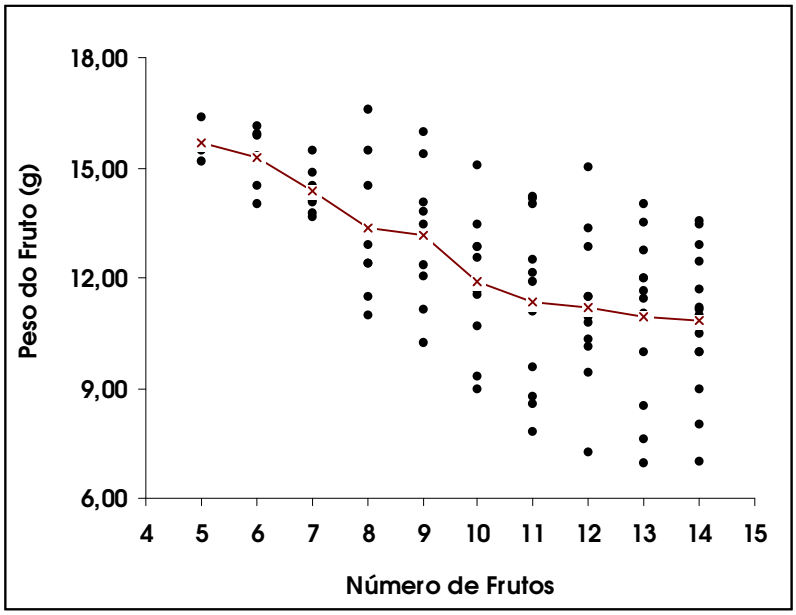
Um produtor de morangos para exportação deseja produzir frutos grandes, pois frutos pequenos têm pouco valor mesmo no mercado interno. Além disso, os frutos, mesmo grandes, não devem ter tamanhos muito diferentes entre si. O produtor suspeita que uma dos fatores que altera o tamanho dos frutos é o número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

O segundo fato que percebemos é que, com o aumento no número de frutos na árvores, 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.

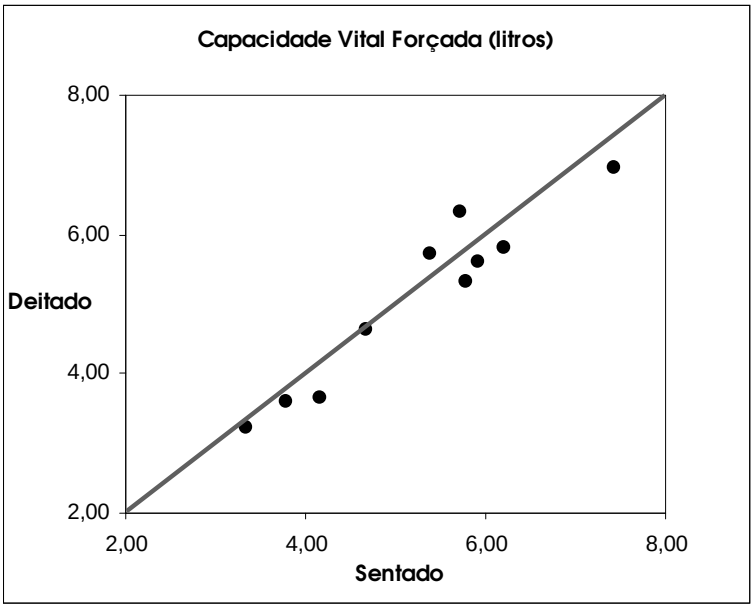
Indivíduo	A	B	C	D	E	F	G	H	I	J
Sentado	4,66	5,70	5,37	3,34	3,77	7,43	4,15	6,21	5,90	5,77
Deitado	4,63	6,34	5,72	3,23	3,60	6,96	3,66	5,81	5,61	5,33

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 ? Através 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.

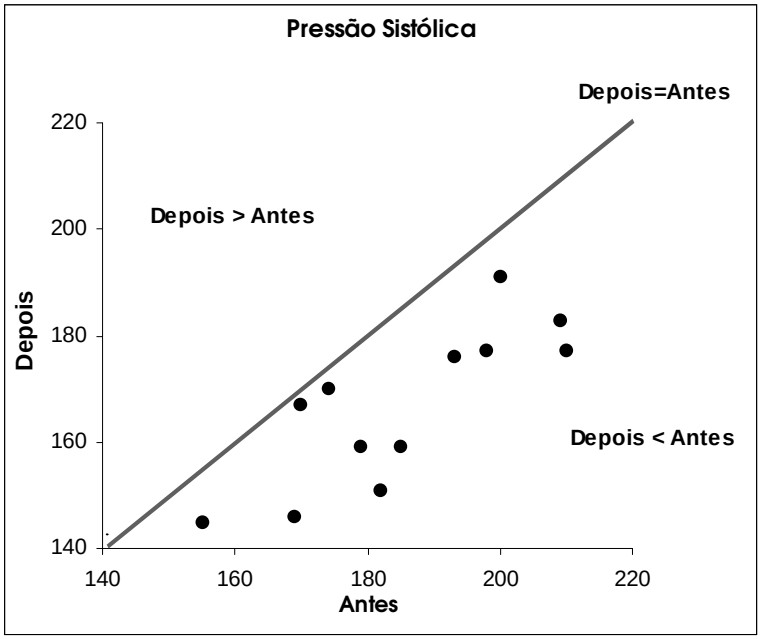
Captopril é um remédio destinado a baixar a pressão sistólica. Para testar seu efeito, ele foi ministrado a 12 pacientes, tendo sido medida a pressão sistólica antes e depois da medicação (Quadro 6.4).

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 ? Através 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.

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%.