CAPÍTULO 7

ANÁLISE EXPLORATÓRIA DE DADOS AMOSTRAIS

"A educação não é o empilhamento de aprendizagem, informações, dados, fatos, habilidades ou capacidades – isso é formação ou instrução – mas sim tornar visivel o que está oculto como uma semente." Thomas Moore, poeta

Nos capítulos anteriores foram resolvidos problemas envolvendo variáveis aleatórias que seguem modelos probabilísticos conhecidos. Por exemplo, desejávamos determinar o valor, em kg, acima do qual encontram-se 5% das cargas de ruptura de um certo tipo de cabo de aço, sendo que, pelo padrão de fabricação desse tipo de cabos, a distribuição das cargas de ruptura é Normal com média e variância conhecidas, μ e σ^2, respectivamente.

- **Na prática, nem o modelo nem os parâmetros são conhecidos de antemão e há a necessidade de determiná-los. Isso pode ser feito a partir da coleta e análise de dados. Uma coleta de dados envolve dois conceitos: população e amostra.**

- **Em Estatística o termo "população" é usado para representar o conjunto de todos os elementos (pessoas ou objetos) cujas propriedades o pesquisador está interessado em estudar.**

- **Essas propriedades podem ser: o resultado de uma medição, um atributo qualitativo, um índice, etc. Por exemplo: diâmetro de um rolamento, duração de um componente eletrônico, caracterização de um produto como perfeito ou defeituoso, etc.**

- **Quando é feito um levantamento completo sobre uma determinada população, ou seja, contemplando cada um dos seus elementos, temos o que se chama de um censo.**

- **Em termos do número de elementos que compõem a população, ela pode ser classificada como finita ou infinita. Os empregados de uma empresa, as agências de um banco, as ruas de uma cidade, o número de carros produzidos mensalmente por uma montadora, etc. são exemplos de populações finitas.**

- **Já os pontos de uma linha, o conjunto dos números reais, etc. constituem populações infinitas.**

- Nas situações mais concretas do mundo real, estamos sempre lidando com **populações finitas**. Por outro lado, as populações ditas infinitas resultam de uma abordagem mais abstrata da realidade.

- Quando uma população, embora finita, é muito grande, ela é tratada, na prática, como se fosse infinita.

- Se uma população é **infinita**, ou **finita mas muito grande**, torna-se impossível ou impraticável a **realização do censo**. Em tais casos, ao invés disso, examina-se somente uma pequena parte da população que chamamos de **amostra**.

- Uma amostra é dita **representativa** da população se a partir da análise dessa amostra podem ser obtidas conclusões passíveis de serem **expandidas** para a população.

- É necessário que a amostra seja extraída de acordo com regras bem definidas. É claro que, se a população da qual a amostra é retirada é muito **homogênea**, essa preocupação não é tão importante.

- É o que ocorre, por exemplo, quando se extrai uma **amostra de um fluido**. Umas poucas gotas podem ser suficientes para se obter a informação desejada, como acontece em um **exame de sangue**.

- Entretanto, quando o material de que está composta a população é muito **heterogêneo** é muito importante o uso de técnicas que nos garantam a obtenção de amostras dignas de confiança.

▪Nas situações mais concretas do mundo real, estamos sempre lidando com **populações finitas**. Por outro lado, as populações ditas infinitas resultam de uma abordagem mais abstrata da realidade.

▪Quando uma população, embora finita, é muito grande, ela é tratada, na prática, como se fosse infinita.

▪Se uma p -se impossível ou imprat disso, examina-se somen e **amostra**.

▪Uma amo análise dessa amostra p andidas para a populaçã

▪É necess s bem definidas. É claro qu o homogênea, essa preo

▪É o que ocorre, por exemplo, quando se extrai uma **amostra de um fluido**. Umas poucas gotas podem ser suficientes para se obter a informação desejada, como acontece em um **exame de sangue**.

▪Entretanto, quando o material de que está composta a população é muito **heterogêneo** é muito importante o uso de técnicas que nos garantam a obtenção de amostras dignas de confiança.

As Técnicas de Amostragem são um tópico importante da Estatística, que trata da obtenção de amostras representativas da população de interesse com um tamanho amostral o menor possível. Quanto maior for o tamanho da amostra, mais cara, demorada e trabalhosa é a pesquisa.

Outro capítulo da Estatística que também se preocupa com a geração de dados é o Planejamento de Experimentos. Os dados são gerados de forma controlada pelo pesquisador, diferentemente da Amostragem onde os dados são em geral selecionados por um procedimento de aleatorização.

Geralmente nas ciências sociais os dados são obtidos por Amostragem, mas nas ciências exatas podem ser realizados experimentos controlados que dependem, por exemplo, da escolha de níveis dos fatores relacionados ao planejamento experimental.

O objetivo do presente Capítulo é apresentar noções de Análise Exploratória, independentemente do processo de obtenção dos dados.

Estimaremos os parâmetros populacionais a partir de dados amostrais, para extrair conclusões sobre populações a partir de amostras.

Em determinadas situações concretas é perfeitamente válido encararmos a população disponível como uma amostra representativa de uma população maior.

Por exemplo:

- **A população de todas as geladeiras de um determinado tipo estocadas no depósito de uma casa comercial pode ser considerada como uma amostra representativa da população de todas as geladeiras desse tipo produzidas ao longo do ano.**

- **A população dos alunos inscritos na disciplina Estatística de um determinado curso de graduação, para efeitos práticos, pode ser vista como uma amostra representativa da população de todos os alunos desse curso inscritos nessa disciplina ao longo de um período de, digamos, 5 anos.**

Sendo assim, se os seus dados a princípio lhe parecem representar toda uma população, talvez valha a pena perguntar a si mesmo se, na verdade, seu propósito não seria extrapolar as conclusões a que você eventualmente chegar para uma realidade maior. Se for esse o caso, os dados devem ser encarados como uma amostra e não como uma população.

7.1 Analisando dados

"Dados! Dados! Dados! Eu não posso fabricar tijolos sem argila.
Sherlock Holmes

"A educação não é o empilhamento de aprendizagem, informações, dados, fatos, habilidades ou capacidades – isso é formação ou instrução – mas sim tornar visível o que está oculto como uma semente." Thomas Moore, poeta

Fazendo uma analogia com a frase do poeta Thomas Moore, os dados não devem ser simplesmente um empilhamento de informações; eles devem ser um instrumento para tornar visível o que está oculto.

Isso é possível quando os analisamos estatisticamente.

O que é analisar dados?

➢...identificar comportamentos médios, comportamentos discrepantes, comparar comportamentos, investigar a interdependência entre variáveis, revelar tendências, etc.

➢... a partir de uma massa de dados, e com o auxílio dos recursos computacionais, separar o que é essencial (estrutura) do que é eventual (ruído).

➢... resumir, de forma inteligente, a informação contida nos dados e assim, permitir que, através desse conhecimento, as decisões sejam tomadas de forma mais consciente.

O que é Análise Exploratória?

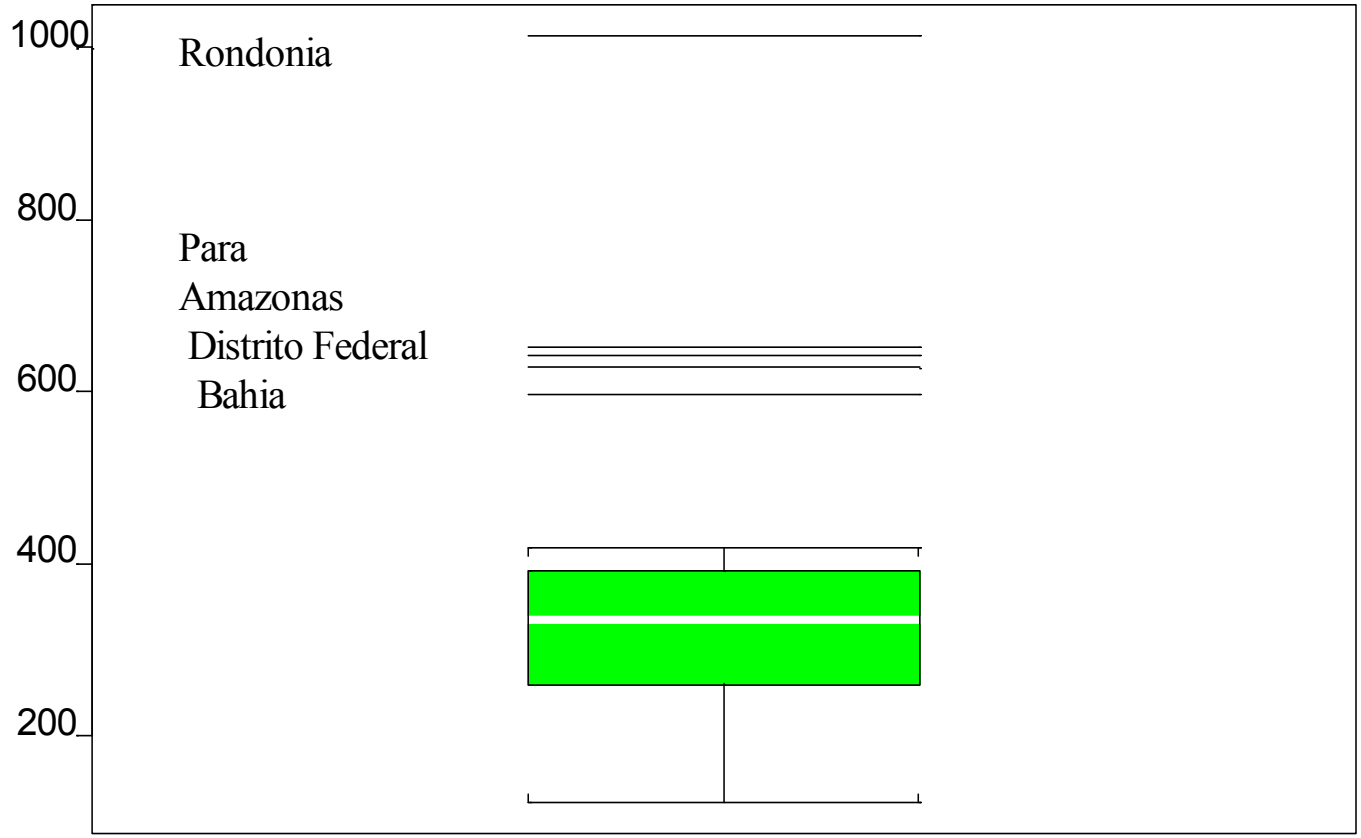
- **Trata-se de um conjunto de técnicas de tratamento de dados que, sem implicar em uma fundamentação matemática mais rigorosa, nos ajuda a fazer uma sondagem do terreno, ou seja, tomar um primeiro contato com a informação disponível.**
- **Supostamente os dados "estão tentando nos dizer algo" a respeito do tema que estamos investigando.**
- **Como extrair e resumir a informação que está contida nos dados? Como devemos usar essa informação para obter mais familiaridade com o problema a ser abordado?**
- **Essas técnicas freqüentemente nos levarão à construção de tabelas e, sobretudo, de gráficos que pretendem facilitar a nossa compreensão do fenômeno em estudo apelando para o poder de visualização do ser humano. Elas também poderão nos guiar na escolha do modelo probabilístico adequado.**

Identificar comportamentos médios

Procurar o centro da informação. Ex:

- Uma turma com 300 alunos gostaríamos de saber o desempenho geral, não olhando individualmente cada aluno e sim a média que é um número que resume o desempenho da turma.

Comportamentos discrepantes



Comparar comportamentos

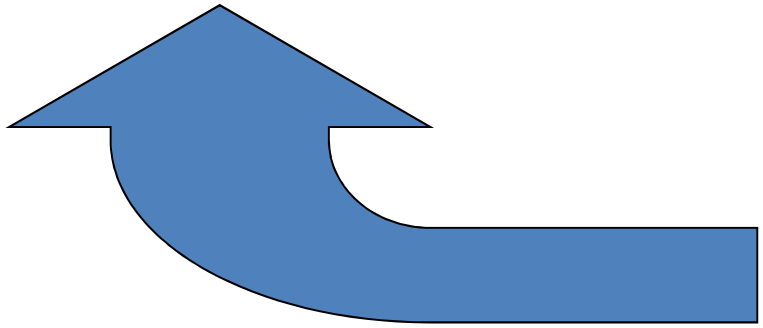
Comparação de dois grupos:

- Placebo
- Remédio

Verificar, através de duas amostras, uma de cada grupo tomando remédio e outra não, se existe diferença no aumento médio da pressão sistólica.

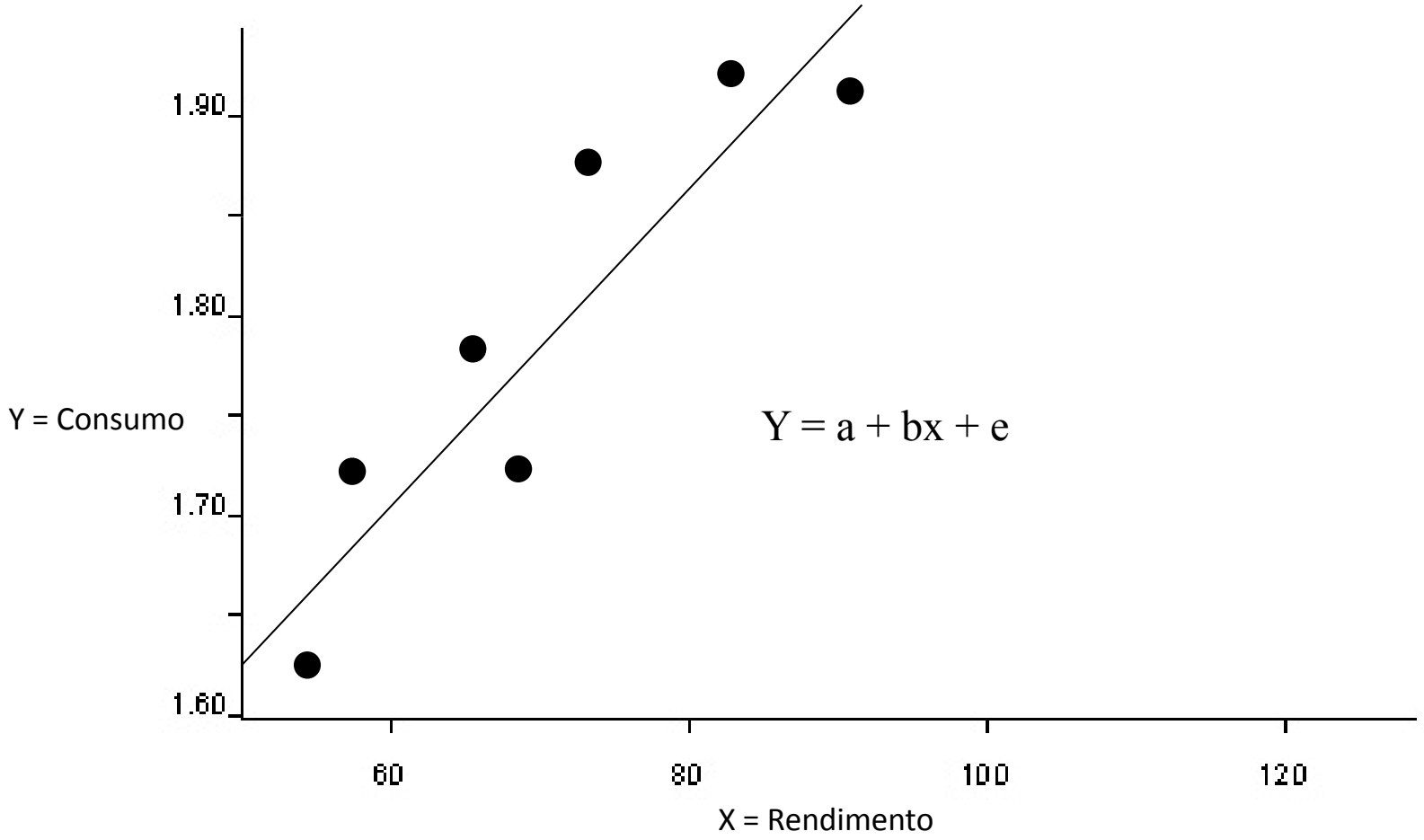
Investigar a Interdependência entre Variáveis

Curso \ Sexo	M	F	
M - Matemática	40	60	100
E - Estatística	30	20	50
I - Informática	30	70	100
	100	150	**250**



	Curso	Sexo
1	I	M
2	I	M
3	E	F
4	E	M
5	I	F
6	E	F
7	I	M
.	.	.
.	.	.
.	.	.
250	I	M

Revelar Tendências



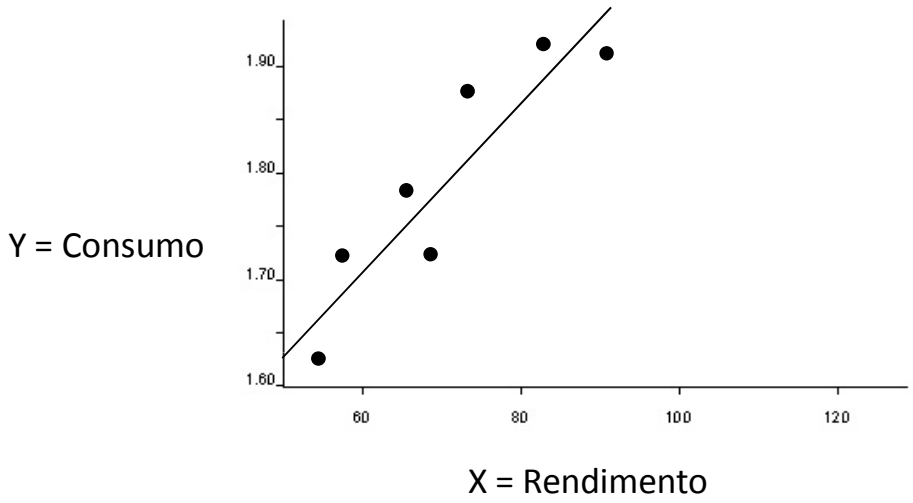
Recursos Computacionais

- R - www.r-project.org (Splus)
- SPSS
- SAS - Statistical Analysis System
- Statistica
- Minitab
- Systat
- Microsoft Exel

Estrutura & Ruído

$$Y = a + bx + e$$

Estrutura Ruído



Nem toda família com a mesma renda reage igual com relação ao consumo:

- Logo incorpora-se o Ruído (**e**) ao modelo.
- Sendo **a + bx** a Estrutura.

O que vem depois da Análise Exploratória?

Uma vez de posse das "pistas" a respeito do tema em estudo, que nos foram fornecidas pela Análise Exploratória, podemos partir para a chamada Inferência, onde serão aplicados aos dados métodos estatísticos mais sofisticados, cuja fundamentação matemática está no Cálculo de Probabilidades.

7.2 - Tipologia das variáveis

Quando é feito um levantamento de dados a respeito de um determinado assunto, eles costumam ser representados em uma tabela como abaixo, onde cada linha corresponde a uma observação e cada coluna corresponde a uma variável.

As observações também são às vezes chamadas de indivíduos, objetos, casos, unidades amostrais, etc. As variáveis também costumam ser referidas como atributos, características, propriedades, etc.

Exemplo 7.1: Imóveis à venda

A tabela 7.1 abaixo mostra os dados brutos de uma amostra de 27 imóveis anunciados para venda nos anúncios de um *site* especializado.

Tabela 7.1 - Amostra sistemática, de 20 em 20, dos imóveis anunciados para venda nos anúncios de um *site* especializado

Nº da Obs.	Bairro	Tipo	Nº de quartos	Preço (*)
1	Barra	Apto.	2	165
2	Barra	Apto.	3	240
3	Barra	Cobt.	-	158
4	Barra	Sala	-	150
5	Botafogo	Apto.	2	59
6	Catete	Apto.	1	54
7	Centro	Sala	-	35
8	Copacabana	Apto.	2	83
9	Copacabana	Apto.	3	180
10	Copacabana	Apto.	4+	85
11	Flamengo	Apto.	1	58
12	Flamengo	Cobt.	-	120
13	Gávea	Apto.	4+	250
14	Ipanema	Apto.	3	130
15	Jacarepaguá	Apto.	3	90
16	Lagoa	Apto.	2	130
17	Laranjeiras	Apto.	2	68
18	Laranjeiras	Apto.	4+	360
19	Leblon	Apto.	3	300
20	Leblon	Apto.	4+	600
21	Maracanã	Apto.	3	137
22	Recreio	Cobt.	-	240
23	São Conrado	Casa	4+	650
24	Tijuca	Apto.	2	49
25	Tijuca	Apto.	2	95
26	Tijuca	Casa	4+	170
27	Vila Isabel	Apto.	2	57

(*) em unidades monetárias (u.m.)

Nesse exemplo cada observação é um imóvel e cada variável é um atributo dos imóveis (bairro, tipo, n° de quartos, preço).

As variáveis, de uma forma geral, podem ser classificadas em tipos conforme abaixo:

- Variável
  - Qualitativa
    - Nominal
    - Ordinal
  - Quantitativa
    - Discreta
    - Contínua

Variável Qualitativa nominal ou categórica - seus valores possíveis são diferentes categorias não ordenadas, em que cada observação pode ser classificada. Exemplos: Área de Atividade, marca de um produto, qualidade do produto (perfeito ou defeituoso).

Variável Qualitativa ordinal - seus valores possíveis são diferentes categorias ordenadas, em que cada observação pode ser classificada. Exemplos: Resposta a uma pesquisa sobre a qualidade de um serviço (bom, regular, ruim), Nível de Instrução, Classe social.

Variável Quantitativa discreta - seus valores possíveis são em geral resultantes de um processo de contagem. Exemplos: Número de empregados de uma empresa, Número de peças defeituosas num lote.

Variável Quantitativa contínua - seus valores possíveis podem ser expressos através de números reais e varrem uma escala contínua de medição. Exemplos: Diâmetro da seção circular de um pistão, Duração da carga de uma bateria.

7.3 - Distribuições de Freqüências. Tabelas e Gráficos.

Para melhor descrever o comportamento de uma variável é comum apresentar os valores que ela assume organizados sob a forma de tabelas de freqüências e gráficos.

na construção das tabelas e gráficos, o tipo de cada variável é o que vai determinar a forma pela qual ela será tratada.

7.3.1 – Tabelas de Freqüências para Variáveis Qualitativas

Em uma Tabela de Freqüências para uma variável qualitativa:

- **Cada linha corresponde a uma categoria possível da variável.**
- **Através de um processo de contagem são obtidos os valores que constam na coluna de Freqüências da tabela. O resultado dessa contagem é a chamada freqüência absoluta.**
- **A partir das freqüências absolutas podem ser também calculadas freqüências relativas, usualmente apresentadas sob a forma de percentuais em relação à freqüência absoluta.**

Exemplo 7.2: Tipo de imóvel

A tabela de freqüências para a variável Tipo de Imóvel do Exemplo 7.1 é:

Tabela 7.2 - Freqüência e Percentual dos 27 imóveis segundo o Tipo

CATEGORIA	Freqüência Absoluta	Percentual
Apartamento	20	74,07
Cobertura	3	11,11
Casa	2	7,41
Sala	2	7,41
Total	27	100,00

Note que a ordem na qual as categorias são dispostas na tabela é irrelevante. Neste caso optou-se por organizá-las em ordem decrescente no que se refere às suas freqüências.

Observação:

Quando a variável é Qualitativa Ordinal, as linhas devem seguir a ordem relativa das possíveis categorias da variável.

7.3.2 – Gráficos de barras e Gráficos de setores para Variáveis Qualitativas

Com base em uma tabela de freqüências podem ser construídos gráficos da distribuição de freqüências, entre os quais os mais comuns são o gráfico de barras e o gráfico de setores

No Gráfico de Barras as categorias são representadas por retângulos dispostos ao longo de um eixo (em geral o horizontal) e as freqüências ou percentagens, correspondentes a cada categoria, são as alturas desses retângulos com relação ao outro eixo (em geral o vertical).

Já no Gráfico de Setores, os 360º do círculo são divididos em setores (fatias) proporcionalmente ao percentual de cada categoria.

Exemplo 7.3: Tipo de Imóvel (Cont.)

Foram construídos o gráfico de barras na Fig.7.1 e o gráfico de setores na Fig. 7.2.

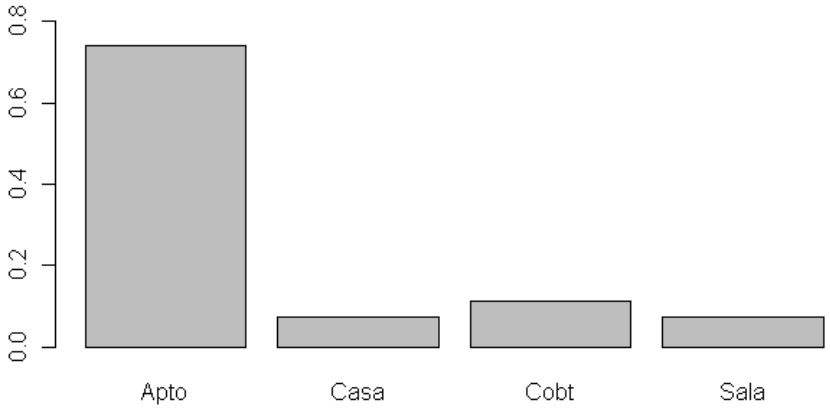


Fig. 7.1: Gráfico de barras exemplo 7.1

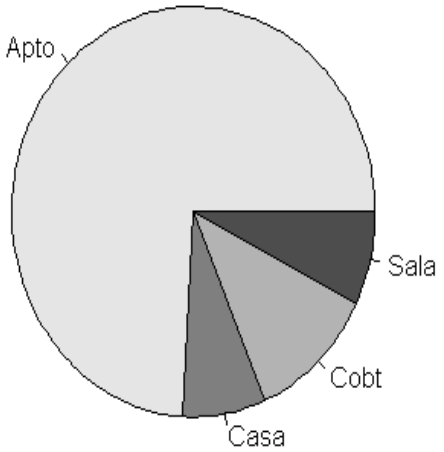


Fig. 7.2: Gráfico de setores exemplo 7.1

Observações sobre cada tipo de gráfico:

- O gráfico de setores, por não implicar em uma ordenação das categorias é mais apropriado para as variáveis qualitativas nominais.

- Enquanto isso, o gráfico de barras, onde as categorias estão naturalmente ordenadas, é mais apropriado para as variáveis qualitativas ordinais.

- Para representar a distribuição de freqüências de uma variável através de um gráfico de setores é importante que a variável não possua muitas categorias, pois isto dificulta a visualização das proporções.

7.3.3 – Tabelas de Freqüências para Variáveis Quantitativas

No caso de variável quantitativa discreta com um pequeno número de valores possíveis (por exemplo: Número de Quartos, no Exemplo 7.1), a construção de uma Tabela de Freqüência segue os mesmos moldes do que foi visto para variáveis qualitativas. Nesse caso cada classe ou categoria é representada por um valor da variável.

Quando trabalhamos com uma variável quantitativa discreta com um grande número de valores possíveis ou com uma variável quantitativa contínua, para avaliarmos sua distribuição através de uma Tabela de Freqüências, antes de mais nada, é preciso dividir o seu intervalo de variação em sub-intervalos (de preferência todos eles com a mesma amplitude). Ao adotar esse procedimento o problema se torna muito semelhante ao caso de variáveis qualitativas.

Exemplo 7.4: Carga de ruptura de cabos de aço

A medição da carga de ruptura, em kg, para 30 espécimes de cabos de aço da mesma espessura resultou nas observações abaixo:

83 96 73 102 93 94 99 85 91 118 93 103 87 95 102

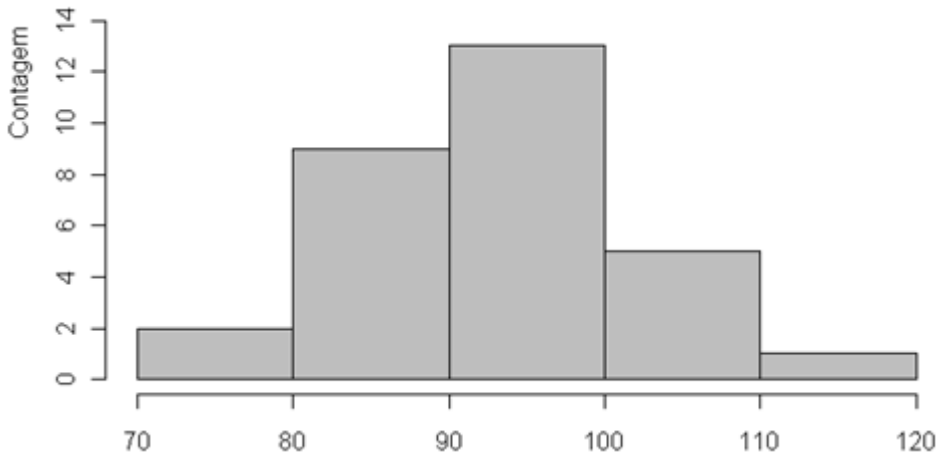
84 100 95 90 81 102 98 94 89 91 78 85 83 105 96

Organize os dados numa tabela de freqüências com intervalos de igual amplitude.

Solução : A menor das observações é 73 kg e a maior, 108 kg . Assim, por conveniência, podemos considerar o intervalo [70, 110] e subdividi-lo em 5 intervalos de amplitude 10. Cada observação é alocada no correspondente sub-intervalo resultando assim a Tabela 7.3 abaixo :

Tabela 7.3 – Freqüências e percentuais das cargas de ruptura

Carga (kg)	Freqüência	Percentuais
70 ├ 80	2	6,67
80 ├ 90	8	26,67
90 ├ 100	13	43,33
100 ├ 110	6	20,00
110 ├ 120	1	3,33
Total	30	100,00



7.3.4 – Histogramas e Diagramas Ramo-Folha para Variáveis Quantitativas

•De forma similar ao gráfico de barras, no Histograma os intervalos de classe da variável considerada são marcados em um eixo e

•as freqüências (ou percentuais) no outro eixo.

•Se os intervalos estiverem no eixo horizontal e as freqüências no eixo vertical dizemos que é um histograma vertical, caso contrário o histograma será denominado horizontal.

•No caso, por exemplo, de um histograma vertical, a largura das barras corresponde à amplitude do intervalo e a altura é proporcional à freqüência (ou ao percentual).

•Qualquer que seja o histograma, vertical ou horizontal, não existe espaço entre as barras.

Exemplo 7.5: Novamente a Carga de Ruptura

Consideremos a variável Carga de Ruptura de cabos de aço , do Exemplo 7.4. A partir da Tabela 7.3 construímos o Histograma abaixo :

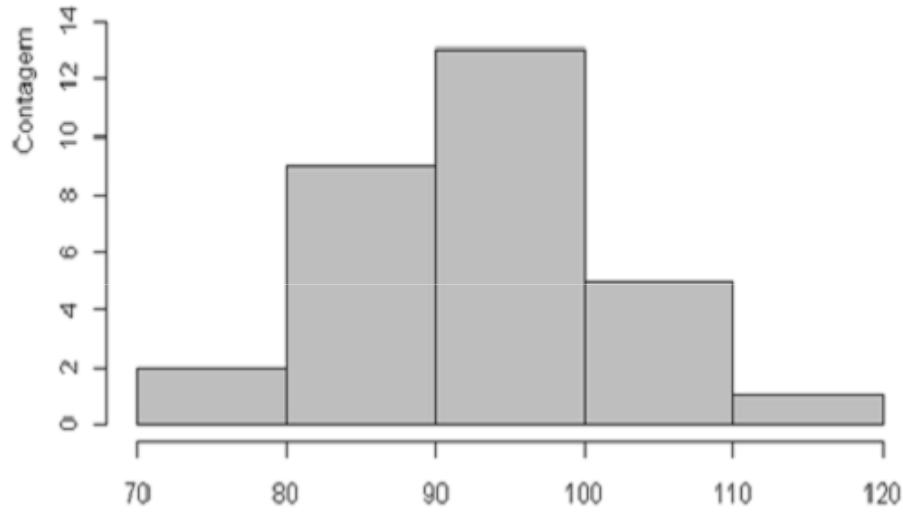


Figura 7.3 - Gráfico de Barras, correspondentes as Cargas de Ruptura

Um outro gráfico que também costuma ser utilizado para analisar uma variável quantitativa é o gráfico ramo-folha, cuja construção é feita através de uma seqüência de passos a serem percorridos, como no exemplo a seguir.

<u>Exemplo 7.6</u>: Mais uma vez a Carga de Ruptura de cabos de aço

Para obter o gráfico ramo-folha:

O primeiro passo é escolher os ramos a partir dos quais serão colocadas as folhas. O primeiro ramo corresponderá a todos os valores entre 70 e 79, o segundo a todos os valores entre 80 e 89, o terceiro a todos os valores entre 90 e 99, e assim por diante.

Em seguida localizaremos cada observação como uma folha (no caso o número de unidades) no ramo correspondente

7	3
8	3
9	6
10	
11	

Figura 7.4 - Ramos (dezenas)

83 96 73 102 93 94 99 85 91 118 93 103 87 95 102
84 100 95 90 81 102 98 94 89 91 78 85 83 105 96

Resultado, depois de **ordenar crescentemente as folhas dentro de cada Ramo**:

Ramos	Folhas	Freqüências
7	38	2
8	13345579	8
9	0113344556689	13
10	022235	6
11	8	1

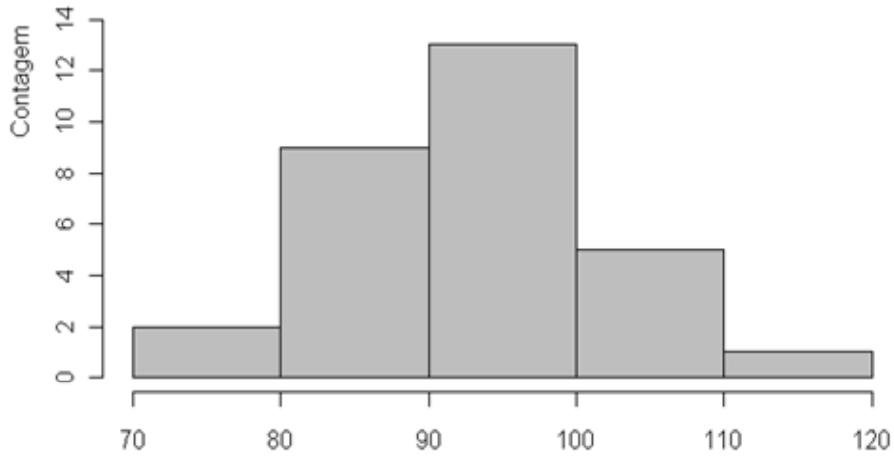


Figura 7.5 - Gráfico ramo-folha para a Carga de Ruptura

Note que o gráfico ramo-folha é muito semelhante ao histograma. Contudo ele é mais informativo porque apresenta os valores de todas as observações da variável, o que não ocorre nem na Tabela de Freqüências nem no Histograma.

7.4 - Medidas de Centralidade para dados amostrais quantitativos

Para uma dada variável quantitativa, uma <u>medida de centralidade</u> é um "valor típico" em torno do qual se situam os valores daquela variável.

Há várias formas de se definir uma medida de centralidade: a média aritmética, a mediana e a moda são as mais conhecidas entre elas.

Sejam $x_1, x_2, ..., x_n$ os valores observados da variável considerada.

A <u>média aritmética</u> dos dados é definida por

$$\overline{X} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Sejam $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$ os mesmos valores que compõem o conjunto de dados, porém dispostos em ordem crescente.

A <u>mediana</u> dos dados é

$$Q2 = \begin{cases} \text{valor da observação de posição central, se n é ímpar} = x_{\left(\frac{n+1}{2}\right)} \\ \text{média dos valores das 2 observações de posição central, se n é par} = \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \end{cases}$$

Exemplo 7.7: Média aritmética e mediana para os dados de Carga de Ruptura.

Para os dados do Exemplo 7.4, relativos à Carga de Ruptura em kg de 30 espécimes de cabos de aço, temos:

n = 30 ; $\sum_{i=1}^{30} x_i$ = 83 + 96 + ... + 105 +96 = 2785

Logo, $\overline{x} = \frac{\sum_{i=1}^{30} x_i}{30} = \frac{2785}{30} = 92{,}83$

Portanto, a média aritmética das observações das cargas de ruptura para os 30 cabos de aço é 92,83 kg.

Para a obtenção da mediana podemos usar o gráfico ramo-folha da Figura 7.5

Como n = 30 (número par) a mediana é a média aritmética das observações $x_{(15)}$ e $x_{(16)}$ que são 93 e 94 , respectivamente. Assim,

Mediana (x) = $\frac{93+94}{2}$ = 93,5

Ou seja, metade dos valores das 30 observações fica abaixo do valor 93,5 enquanto que a outra metade fica acima deste valor.

Note que, neste exemplo, a média aritmética e a mediana têm valores relativamente próximos entre si. Isto se deve ao fato de que os valores distribuem-se de forma aproximadamente simétrica em torno do valor central, conforme é mostrado pelo histograma e pelo gráfico ramo-folha.

7.5 Medidas de Dispersão para dados amostrais quantitativos

Como no caso de variáveis aleatórias podemos definir medidas de dispersão para dados amostrais quantitativos, isto é, indicadores do grau de espalhamento dos valores da amostra em torno da medida de centralidade.

Há diferentes formas de se medir a dispersão de uma variável quantitativa. Entre estas serão vistas aqui a variância amostral, o desvio padrão amostral, o coeficiente de variação amostral e a distância interquartil amostral.

A variância amostral é calculada por

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

O desvio padrão amostral é a raiz quadrada não negativa da variância, ou seja,

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

O coeficiente de variação amostral é o quociente entre o desvio padrão e a média amostrais.

$$cv = \frac{s}{\bar{x}}$$

Exemplo 7.8: Continuando com os dados de Carga de Ruptura

a) Cálculo da variância amostral

Temos : n = 30 ; $\sum_{i=1}^{30} x_i$ = 2785 ; $\sum_{i=1}^{30} x_i^2$ = 261 037 ; $\bar{x}$ = 92,8333

$$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1} = \frac{261037 - 30(92,8333)^2}{29} = 86,08 \text{ kg}^2$$

b) Desvio-padrão amostral :

$s = \sqrt{86,08}$ = 9,3 kg

c) Coeficiente de variação amostral :

$cv(x) = \frac{s}{\bar{x}} = \frac{9,3}{92,83} = 0,10$ (10%)

Embora o coeficiente de variação amostral seja relativamente pequeno, em termos de qualidade do produto pode representar uma variabilidade excessiva em relação à media. Na prática, todos os esforços devem ser feitos para se fabricar cabos de aço com uma carga de ruptura de menor variabilidade.

Sejam $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ os dados dispostos em ordem crescente.

Já vimos que a mediana é um valor tal que metade dos dados é menor que ele e metade dos dados é maior que ele.

Analogamente, os 3 quartis são valores que dividem os dados em 4 grupos, cada um deles contendo 1/4 do tamanho total da amostra.

O primeiro quartil Q1 tem 1/4 dos dados abaixo dele e 3/4 dos dados acima dele.

O terceiro quartil Q3 tem 3/4 dos dados abaixo dele e 1/4 dos dados acima dele.

O segundo quartil Q2 é a própria mediana.

A distância interquartil é dada por DIQ = Q3 – Q1.

Para o cálculo dos quartis devemos determinar a posição que eles ocupam quando os dados são dispostos em ordem crescente.

Pela definição, a mediana ocupa a posição $(n+1)/2$. Se n é ímpar a posição da mediana é um número inteiro e corresponde exatamente ao valor central do conjunto ordenado de observações. Se n for par, a mediana é calculada como a média das observações de ordens $n/2$ e $n/2+1$. Isto é,

$$Q2 = \text{Mediana}(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ é ímpar} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ é par} \end{cases}$$

A posição do primeiro quartil, Q1, é uma posição intermediária entre a posição 1 e a posição da mediana, $(n+1)/2$. Assim, o primeiro quartil ocupa a posição de ordem $[1+(n+1)/2]/2 = (n+3)/4$. Se o número da posição não é inteiro, Q1 deve ser calculado por interpolação entre os valores cujas posições são vizinhas a $(n+3)/4$.

Analogamente, a posição do terceiro quartil, Q3, é dada por $[(n+1)/2 + n]/2 = (3n+1)/4$.

Se o número da posição não é inteiro procedemos como no caso de Q1. .

Exemplo 7.9: Continuando com os dados de Carga de Ruptura (cont)

Cálculo da mediana:

Temos n= 30 observações.

Para obtermos a mediana (segundo quartil), calculamos primeiramente (n+1)/2 = 15,5. Ou seja, Q2 está entre $x_{(15)}$ e $x_{(16)}$. A saber,

$$Q2 = 0{,}5\, x_{(15)} + 0{,}5\, x_{(16)} = 0{,}5 \times 93 + 0{,}5 \times 94 = 93{,}5 \text{ kg}$$

Isto quer dizer que metade dos cabos de aço têm suas cargas de ruptura inferiores a 93,5 kg e a outra metade têm suas cargas de ruptura superiores a 93,5 kg.

Cálculo da distância interquartil:

Para obtermos o primeiro quartil, calculamos (n+3)/4 = 33/4 = 8,25. Ou seja, Q1 está entre $x_{(8)}$ e $x_{(9)}$. Mais precisamente:

$$Q1 = 0{,}75\, x_{(8)} + 0{,}25\, x_{(9)} = 0{,}75 \times 85 + 0{,}25 \times 87 = 85{,}5 \text{ kg}$$

Para o terceiro quartil, calculamos (3n+1)/4 = 91/4 = 22,75 . Portanto, Q3 está entre $x_{(22)}$ e $x_{(23)}$, ou seja,

$$Q3 = 0{,}25\, x_{(22)} + 0{,}75\, x_{(23)} = 0{,}25 \times 98 + 0{,}75 \times 99 = 98{,}75 \text{ kg}$$

Dessa forma , a distância interquartil é dada por

$$DIQ = Q3 - Q1 = 98{,}75 - 85{,}5 = 13{,}25 \text{ kg}$$

Isto quer dizer que aproximadamente metade dos cabos de aço tem uma carga de ruptura compreendida entre 85,5 kg e 98,75 kg.

7.6 O conceito de resistência de uma medida

Diz-se que uma medida de centralidade ou de dispersão é resistente quando ela é pouco afetada pela presença de observações discrepantes. É claro então que as medidas mais resistentes são mais convenientes que as menos resistentes.

Entre as medidas de centralidade, a média aritmética é bem menos resistente que a mediana.

Por outro lado, entre as medidas de dispersão, o desvio padrão é bem menos resistente que a distância interquartil.

Exemplo 7.10: Os cabos de aço mais uma vez

Admita que nas observações das cargas de ruptura dos 30 cabos de aço, a observação de valor 105 tivesse sido, por engano ou por erro do instrumento de medição, anotada como 150. É fácil ver que nesse caso a média aritmética que era de 92,8 kg passaria a ser, com o novo valor, 94,3 Kg. Analogamente, a variância de 86,08 kg^2 e o erro padrão de 9,3 kg passariam a ter os novos valores de 217,34 kg^2 e 14,7 kg, respectivamente. Há, portanto, uma mudança importante nos valores dessas três medidas.

Contudo, a mediana e os quartis superior e inferior todos mantém o mesmo valor que tinham antes. Conseqüentemente, o mesmo ocorre com a distância interquartil.

7.7 Identificação de Discrepâncias em Variáveis Quantitativas

"Fatos que a princípio parecem improváveis irão – mesmo que aparentemente sem explicação – deixar cair o manto que os mantém escondidos e se apresentarão em sua beleza nua e simples".
Galileu Galilei, cientista

Eventualmente em uma massa de dados há valores que foram coletados em condições anormais (falha de equipamento, queda de energia, erro do operador, erro de leitura, erro de digitação, etc.). Esses valores, principalmente quando estão muito afastados dos demais (para mais ou para menos) infelizmente podem afetar de forma substancial o resultado das análises

Uma vez detectada a presença de uma observação discrepante, poderá ser tomada a decisão de repetir aquele experimento, ou meramente expurgar aquele dado da amostra (ou até mesmo mantê-lo, se for encontrada uma explicação plausível para aquela discrepância...).

Um **critério para a identificação de observações discrepantes**, que se baseia em medidas pouco resistentes, é apontar toda observação que estiver fora do intervalo $(\bar{x}-3\cdot s;\bar{x}+3\cdot s)$.

Um **segundo critério** também muito usado, que se baseia em **medidas mais resistentes** para a identificação de observações discrepantes é apontar qualquer valor abaixo da

Cerca Inferior = Q1 - $\frac{3}{2}$ × DIQ ou acima da **Cerca superior = Q3 + $\frac{3}{2}$ × DIQ**.

Exemplo 7.11: Discrepância nas cargas de ruptura. Consideremos, mais uma vez, os dados do Exemplo 7.4. Temos, para eles : $\bar{x}$ = 92,8 kg; s = 9,3 kg

Usando o critério para identificação de valores discrepantes baseado em medidas pouco resistentes, encontramos : $\bar{x}$ - 3 s = 92,8 – 3x9,3 = 64,9 kg e $\bar{x}$ + 3 s = 92,8 + 3x9,3 = 120,7 kg

Como todos os 30 valores observados estão dentro do intervalo (64,9 ; 120,7) concluímos que não há valores discrepantes.

Para a utilização do 2º critério , temos : Q1 = 85,5 kg; Q3 = 98,75 kg e DIQ = 13,25 kg

As cercas são: Q1 - $\frac{3}{2}$ DIQ = 45,75 kg e Q2 + $\frac{3}{2}$ DIQ = 138,5 kg

Como todos os 30 valores estão entre as duas cercas, não foram identificados valores discrepantes.

Suponha, agora, como no exemplo anterior, que o valor 105 kg foi erroneamente anotado como 150kg. Já vimos que, nesse caso, os quartis, e a distância interquartil não mudam. O mesmo ocorre, é claro, com as cercas. Desta maneira, o valor 150 kg, por ser maior do que a cerca superior, 138,5 kg, pode ser classificado como discrepante.

E no caso do critério com medidas não resistentes?. Neste caso, teríamos

$\bar{x} - 3\,s = 94{,}3 - 3\times 14{,}7 = 50{,}2$ kg e $\bar{x} + 3\,s = 94{,}3 + 3\times 14{,}7 = 138{,}4$ kg

O valor 150 kg é, novamente, identificado como *outlier* ou discrepante, porque ele está fora do intervalo (50,2 ; 138,4)

Considerações sobre a simetria do perfil de freqüências:

(a) Convém observar que ambos os critérios aqui apresentados, pela própria forma como foram propostos, pressupõem que a distribuição de freqüências (representada pelo Histograma ou pelo Gráfico Ramo Folha) dos dados é simétrica com relação à medida de centralidade adotada (média ou mediana).

(b) Se, para a variável em exame, a distribuição de freqüências for muito assimétrica (é mais comum o caso de assimetria para a direita), um expediente útil é aplicar uma transformação à variável original (por exemplo: raiz quadrada, logaritmo,...) e depois usar o critério para detecção de observações discrepantes para a variável já transformada.

7.8 - *Box Plot* para Variáveis Quantitativas

O *Box Plot* ou Desenho Esquemático é um gráfico que se costuma utilizar para sintetizar em uma mesma figura várias informações relativas à distribuição de uma determinada variável quantitativa:

1. **Inicialmente é traçado um eixo vertical onde serão representados os valores da variável considerada.**
2. **Depois se desenha um retângulo cuja posição da base inferior corresponde ao valor do 1º quartil Q1 e cuja posição da base superior corresponde ao valor do 3º quartil Q3. A posição da mediana é indicada por um traço horizontal no interior desse retângulo.**
3. **Em seguida são traçados dois segmentos de reta verticais que vão, um deles desde o ponto médio da base inferior do retângulo até a posição da menor observação não discrepante, e o outro desde o ponto médio da base superior do retângulo até a posição da maior observação não discrepante.**
4. **Cada uma das observações discrepantes é explicitada (e, muitas vezes, devidamente rotulada) no gráfico.**

Observe que nesta figura a dimensão horizontal não tem qualquer significado.

Exemplo 7.12: *Box-Plot* para as Cargas de Ruptura dos 30 cabos de aço.

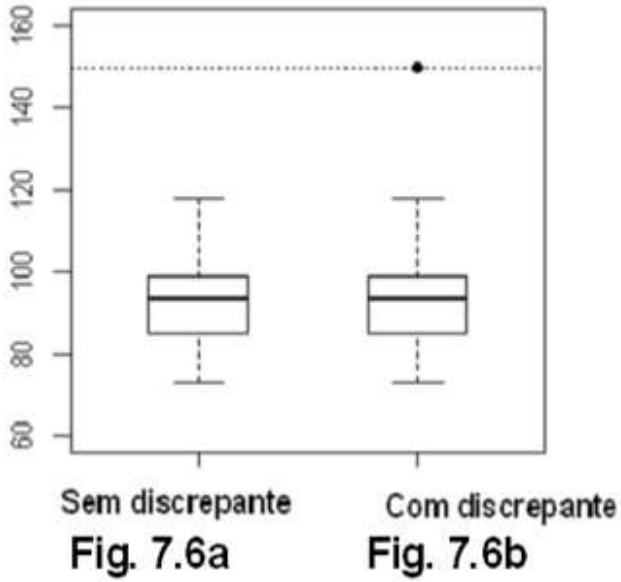


Fig. 7.6a Fig. 7.6b

Figura 7.6- *Box-plot* dos dados de Ruptura dos cabos de aço

A Figura 7.6a é o Box-Plot correspondente aos dados originais do Exemplo 7.7, enquanto que a Figura 7.6b apresenta o Box-Plot para os dados modificados (com a observação 150 kg ao invés da observação 105 kg). Note que no segundo caso, 150 kg é um valor discrepante, identificado por um ponto isolado.

"Quantidades numéricas focam em valores esperados, resumos gráficos focam em valores inesperados."

John Tukey, estatístico

7.9 – Estudando a relação entre duas variáveis

Muitas vezes estamos interessados em duas características dos elementos de uma amostra. Por exemplo, em uma amostra de carros produzidos por uma montadora podemos determinar, para cada carro, o seu modelo e o tipo de combustível usado; numa amostra de fios elétricos as características de interesse podem ser o seu diâmetro e a sua condutividade, etc. As variáveis a serem medidas podem ser qualitativas ou quantitativas. Começaremos considerando o caso de duas variáveis qualitativas.

7.9.1 – Relação entre Variáveis Qualitativas. Tabelas de Contingência

Quando se deseja investigar a relação entre duas variáveis qualitativas, o caminho natural é, a partir de um conjunto de dados montarmos uma tabela de contingência, contendo as freqüências cruzadas relativas a essas duas variáveis.

A montagem da tabela de contingência implica somente em se contar o número de ocorrências em cada cruzamento das 2 variáveis (quadrícula da tabela).

Uma vez obtida a tabela de contingência é importante também que sejam calculados os percentuais de linha e/ou de coluna. Através dessa abordagem, uma das variáveis é usada para dividir a população em estratos e depois determina-se o perfil de freqüências relativas (ou percentuais) da outra variável em cada um desses estratos.

Exemplo 7.13: Condições de trabalho vistas por empregados de diferentes setores de uma empresa.

- Uma Pesquisa de Clima Organizacional é uma pesquisa de opinião através da qual se pretende investigar o nível de satisfação e motivação dos empregados de uma determinada empresa.
- Suponha que foi feita uma Pesquisa de Clima Organizacional na Empresa X. Os empregados a serem entrevistados foram selecionados através de um processo de Amostragem dentro de cada um dos três Departamentos da empresa. O relatório correspondente indica que, com relação às Condições de Trabalho oferecidas aos empregados, dentro de cada um dos três Departamentos dessa empresa as opiniões se dividem conforme indica a tabela abaixo.

Departamento	Avaliação das Condições de Trabalho			Número de Entrevistas
	Insatisfeitos	Parcialmente satisfeitos	Satisfeitos	
Comercial	63	8	4	75
Pessoal	40	30	5	75
Produção	84	72	44	200
Total	187	110	53	350

Esta tabela é dita uma tabela de contingência de 3×3, porque tem três linhas e três colunas. Em geral, uma tabela com h linhas e k colunas é dita uma tabela de contingência de h×k.

Departamento	Avaliação das Condições de Trabalho			Número de Entrevistas
	Insatisfeitos	Parcialmente satisfeitos	Satisfeitos	
Comercial	63	8	4	75
Pessoal	40	30	5	75
Produção	84	72	44	200
Total	187	110	53	350

- **Por exemplo, na tabela 7.4 verifica-se que o número total de empregados entrevistados do Departamento Comercial é igual a 75. Entre eles 63 estão na Classe Insatisfeitos, portanto concluí-se que**

$\left(\frac{63}{75}\right) \times 100$ **= 84,00% dos entrevistados do DP estão Insatisfeitos.**

Tabela 7.5.- Percentuais (de linha) correspondentes às Classes de Avaliação das Condições de Trabalho para cada um dos Departamentos aos quais pertencem os entrevistados

Departamento	Avaliação das Condições de Trabalho			Total
	Insatisfeitos	Parcialmente satisfeitos	Satisfeitos	
Comercial	84,00%	10,67%	5,33%	100%
Pessoal	53,33%	40,00%	6,67%	100%
Produção	42,00%	36,00%	22,00%	100%
Total	53,43%	31,43%	15,14%	100%

Note que os percentuais somam 100% ao longo de cada uma das linhas.

A tabela 7.5 parece sugerir, por exemplo, que o percentual de Insatisfeitos é duas vezes maior entre os empregados do Departamento Comercial que entre os do Departamento de Produção. Por outro lado, entre os Satisfeitos, percentualmente os empregados do Departamento de Produção são aproximadamente 4 vezes mais numerosos que os do Departamento Comercial.

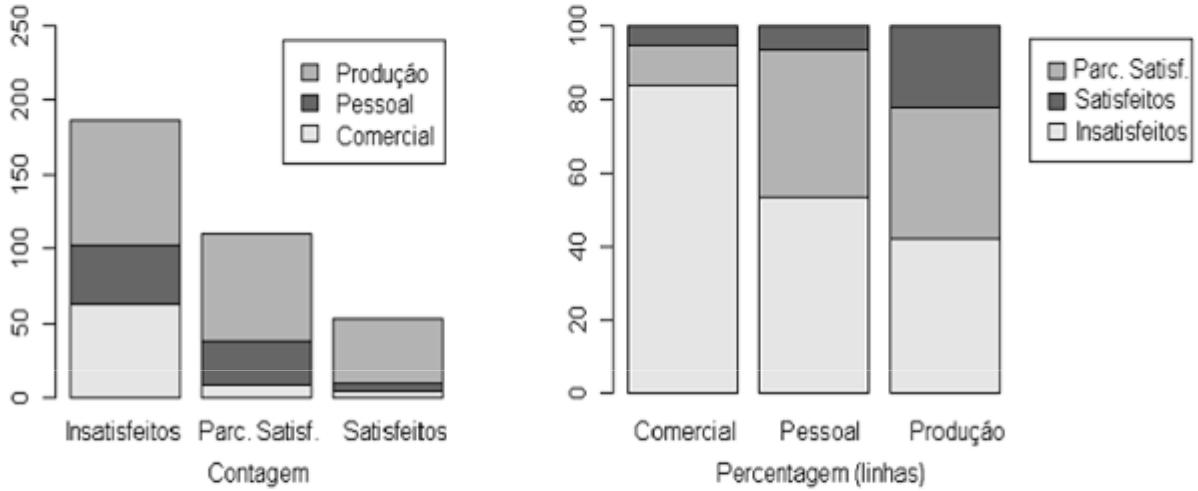


Figura 7.7 – Distribuição das opiniões sobre as Condições de Trabalho dentro de cada Departamento

Observemos agora o que acontece quando calculamos percentuais de coluna ao invés de percentuais de linha:

Na montagem da Tabela 7.6 a partir da Tabela 7.4, uma vez fixada uma coluna da tabela (Classe de Avaliação das Condições de Trabalho), foi calculado o percentual correspondente a cada linha (Departamento) com respeito ao total de coluna.

Na tabela 7.4 notamos, por exemplo, que existe um total de 187 empregados que se declararam Insatisfeitos. Entre eles 163 são do Departamento Comercial, portanto $\left(\frac{63}{187}\right)\times 100=$ 33,69 dos empregados Insatisfeitos são do Departamento Comercial.

Note que os percentuais somam 100% ao longo de cada uma das colunas.

Tabela 7.6 - Percentuais (de coluna) correspondentes aos Departamentos a que pertencem os empregados para cada uma das Classes de Avaliação das Condições de Trabalho

Departamento	Avaliação das Condições de Trabalho			Total
	Insatisfeitos	Parcialmente satisfeitos	Satisfeitos	
Comercial	33,69%	7,27%	7,55%	21,43%
Pessoal	21,39%	27,27%	9,43%	21,43%
Produção	44,92%	65,46%	83,02%	57,14%
Total	100,00%	100,00%	100,00%	100%

Vemos aqui, entre outras coisas, que o percentual relativo aos Empregados do Departamento Comercial é mais de quatro vezes maior entre Insatisfeitos do que entre os Parcialmente Satisfeitos ou entre os Satisfeitos. Por outro lado, o percentual dos empregados Satisfeitos do Departamento de Produção é quase duas vezes maior que o de Insatisfeitos, no mesmo Departamento.

7.9.2 - Covariância e Correlação entre Variáveis Quantitativas

- **A estratégia *que acabamos de apresentar (Tabelas de Contingência) pode também ser utilizada* para se analisar a relação entre duas variáveis quantitativas discretas.**

- **Ainda mais, poderia ser utilizada no caso de variáveis contínuas se, antes de mais nada, for feita uma "discretização". Neste último caso basta dividirmos em subintervalos e montarmos a tabela de contingência correspondente.**

- **O mesmo procedimento se aplicaria ao caso de variáveis quantitativas discretas com muitos valores.**

- **Contudo, no caso específico de variáveis quantitativas contínuas há medidas, semelhantes às apresentadas, no Capítulo 3, para variáveis aleatórias contínuas que permitem analisar de forma mais precisa a relação entre aquelas variáveis.**

- **Consideremos, então, duas variáveis quantitativas contínuas, X e Y. Cada dado a partir de uma amostra de tamanho n será representado por um par ordenado (x_i , y_i) para i = 1,2,...,n , onde x_i e y_i são, respectivamente, a i-ésima observação de X e de Y.**

- **O problema é determinar se X e Y estão relacionadas e de que forma. Uma primeira tentativa de descobrir a forma aproximada de relacionamento entre as duas variáveis é através de uma representação gráfica dos dados como pontos no plano xy. Um gráfico dessa natureza chama-se diagrama de dispersão.**

Exemplo 7.14: Difusividade Térmica e Temperatura

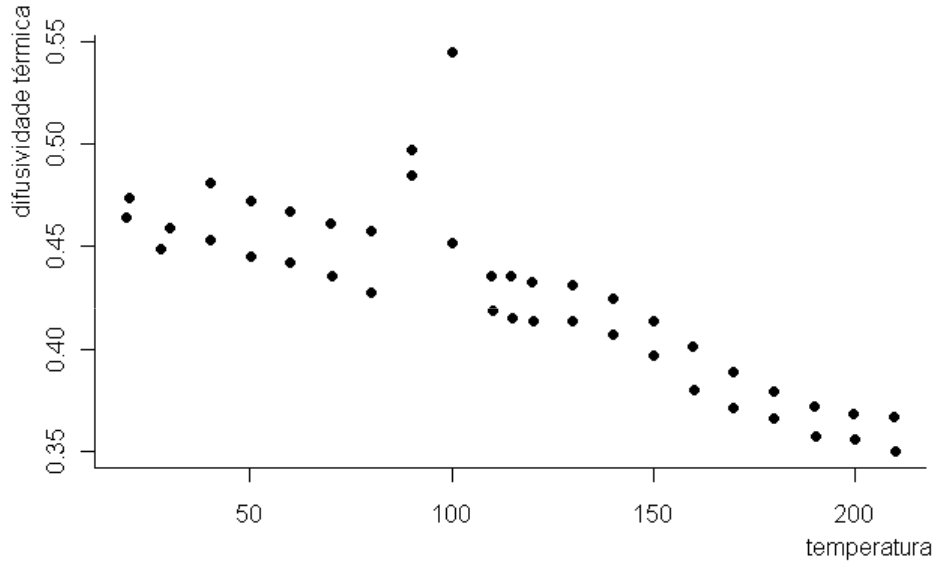
O dados a seguir mostram como a Difusividade Térmica de uma fibra de carbono, sem envelhecimento, varia em função da temperatura.

Tabela 7.7 – Temperatura (de que?) e Difusividade Térmica de uma fibra de carbono, sem envelhecimento

Temp (°C)	Dif Term (mm^2/s)	Temp (°C)	Dif Term (mm^2/s)	Temp (°C)	Dif Term (mm^2/s)
19,2	0,464	150,1	0,397	90,0	0,497
30,2	0,459	160,0	0,38	100,0	0,545
40,2	0,453	170,0	0,371	110,0	0,436
50,3	0,445	180,0	0,366	114,8	0,436
60,1	0,442	190,2	0,357	120,0	0,433
70,2	0,436	200,0	0,356	130,1	0,431
80,2	0,428	210,0	0,35	139,9	0,425
90,1	0,485	20,0	0,474	150,0	0,414
100,2	0,452	27,8	0,449	159,9	0,401
110,1	0,419	40,1	0,481	170,0	0,389
115,1	0,415	50,2	0,472	179,9	0,379
120,2	0,414	60,0	0,467	190,0	0,372
130,1	0,414	70,1	0,461	199,8	0,368
140,1	0,407	80,1	0,458	209,9	0,367

Sejam X a variável Temperatura e Y Difusividade Térmica.. O diagrama de dispersão correspondente aos dados da Tabela acima é mostrado na Fig. 7.7.

Figura 7.7 – Diagrama de dispersão da Temperatura (°C) versus Difusividade Térmica (mm^2/s) de fibras de carbono, sem envelhecimento



Uma análise do diagrama de dispersão da Figura 7.7 nos revela que há uma tendência de valores pequenos de X estarem associados a valores grandes de Y, ao mesmo tempo em que valores grandes de X estão associados a valores pequenos de Y. Além disso, a natureza da relação entre X e Y parece ser bem expressa por meio de uma reta (embora haja alguns poucos pontos destoantes).

Para variáveis contínuas tais como X e Y acima, seria desejável medir de algum modo o grau de associação entre elas.

Uma primeira medida, que permite determinar o grau de variação conjunta entre duas variáveis contínuas, X e Y, a partir de uma amostra de n elementos , é a <u>covariância amostral</u> , definida a seguir.

A <u>covariância amostral</u> entre X e Y é definida por

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n-1}$$

onde $\bar{x}$ e $\bar{y}$ são as médias aritméticas de x e y X e Y, respectivamente.

Uma fórmula de cálculo muito utilizada para a covariância entre X e Y é a seguinte:

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{n-1}$$

<u>Relação entre o comportamento do gráfico de dispersão e a covariância:</u>

- A covariância é positiva, se o diagrama de dispersão de x e y mostra uma tendência crescente, isto é, se valores pequenos de X estão associados a valores pequenos de Y e valores grandes de X estão associados a valores grandes de Y.

- A covariância é zero se o diagrama de dispersão não mostra qualquer tendência, crescente ou decrescente. Na prática, devido aos possíveis erros de amostragem, neste último caso, de fato, o que ocorre é que, em módulo, a covariância é próxima de zero.

Exemplo 7.15: Covariância entre a Temperatura e a Difusividade Térmica

Para os dados do Exemplo 7.14, representando por X a Temperatura (ºC) e por Y a Difusividade Térmica (mm^2/s), temos :

Tabela 7.7

$n = 42$; $\sum x_i = 4829,2$; $\sum y_i = 17,865$; $\sum x_i^2 = 688553,1$;

$\sum y_i^2 = 7,678455$; $\sum x_i y_i = 1968,807$

Substituindo na terceira fórmula de cálculo da covariância , encontramos:

$$s_{xy} = \frac{1968,807 - \frac{(4829,2)(17,865)}{42}}{41} = -2,0812 \ ºC.mm^2/s.$$

O sinal negativo de s_{xy} está de acordo com a tendência decrescente apresentada pelo Diagrama de Dispersão da Figura 7.7

Além do sinal, indicador de uma tendência, crescente ou decrescente, apresentada pelo diagrama de dispersão, muito pouca informação pode ser extraída da covariância amostral, dada sua difícil interpretação. Isso porque o valor de s_{xy} é fortemente dependente das unidades de medida de X e de Y.

Uma medida de uso mais freqüente para se estabelecer a associação entre duas variáveis quantitativas contínuas, X e Y, é o coeficiente de correlação , também chamado de coeficiente de correlação linear ou coeficiente de correlação de Pearson , que definimos a seguir.

Sejam X e Y duas variáveis quantitativas contínuas e consideremos uma amostra de tamanho n cujos elementos são pares (x_i , y_i) , i = 1, 2,...., n.

O coeficiente de correlação entre X e Y, r_{xy} , é dado por : $r_{xy} = \frac{s_{xy}}{s_x s_y}$,

onde s_x e s_y são, respectivamente, os desvios-padrões amostrais de X e de Y.

A partir das expressões de s_{xy} , s_x e s_y obtemos as seguintes fórmulas para calcular o coeficiente de correlação :

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum x_i^2 - n\bar{x}^2)(\sum y_i^2 - n\bar{y}^2)}} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{\{n\sum x_i^2 - (\sum x_i)^2\}\{n\sum y_i^2 - (\sum y_i)^2\}}}$$

Algumas propriedades do coeficiente de correlação

1. O coeficiente de correlação é adimensional, isto é, o seu valor não é afetado pelas unidades de medida nas quais são expressas as variáveis X e Y. Por esse motivo ele é preferido à covariância para quantificar o grau de associação entre duas variáveis quantitativas contínuas.
2. O valor r_{xy} está sempre compreendido entre -1 e +1.
3. O coeficiente de correlação mede um tipo específico de interdependência, a saber, interdependência linear. Isso quer dizer que mesmo havendo uma forte dependência entre duas variáveis quantitativas, se a relação entre elas for do tipo não linear, o coeficiente de correlação poderá não ser muito alto em módulo.

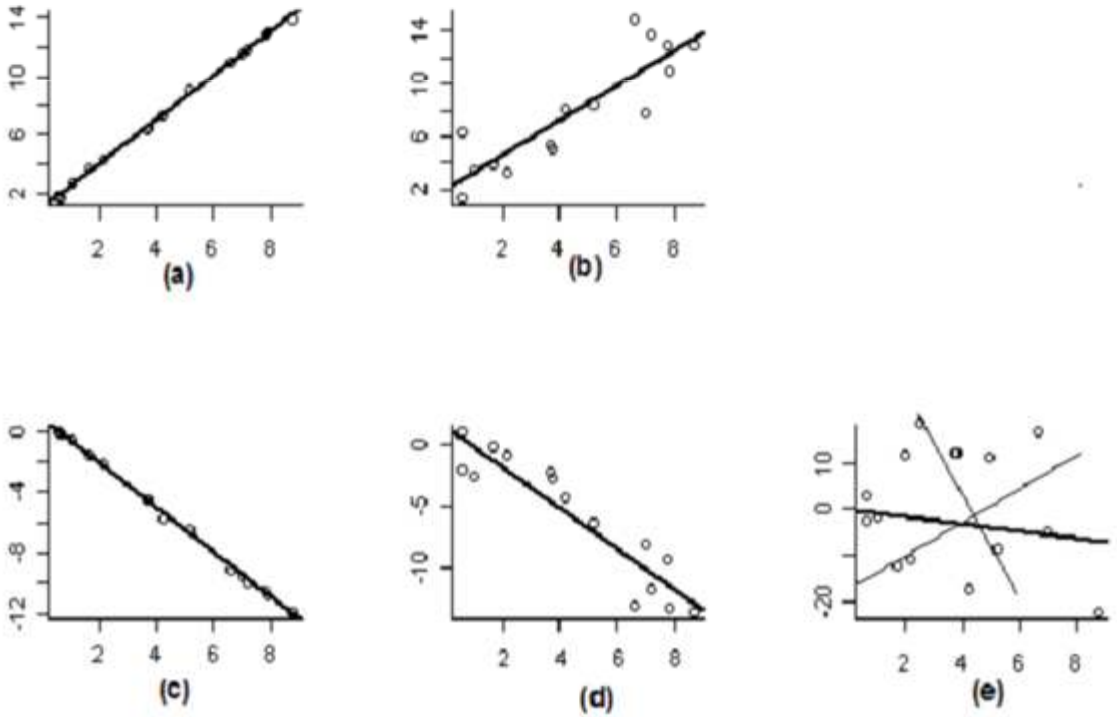


Figura7.8: Relação entre o comportamento do gráfico de dispersão e o coeficiente de correlação

A análise da Figura 7.8 indica que :

- Se os pontos estiverem dispostos em torno de uma reta com inclinação positiva, o valor de r_{xy} estará entre 0 e +1, como na Figura 7.8(b). Quanto mais próximos esses pontos estiverem da reta, mais próximo r_{xy} estará de +1, como na Figura 7.8(a).
- Se os pontos estiverem dispostos em torno de uma reta com inclinação negativa, o valor de r_{xy} estará entre -1 e 0, como na Figura 7.8(d). Quanto mais próximos esses pontos estiverem da reta, mais próximo r_{xy} estará de -1, como na Figura 7.8(c)
- Se os pontos estiverem bastante dispersos, de forma a que não se possa identificar algum tipo de dependência linear entre x e y, r_{xy} estará próximo de zero, como na Figura 7.8(e).

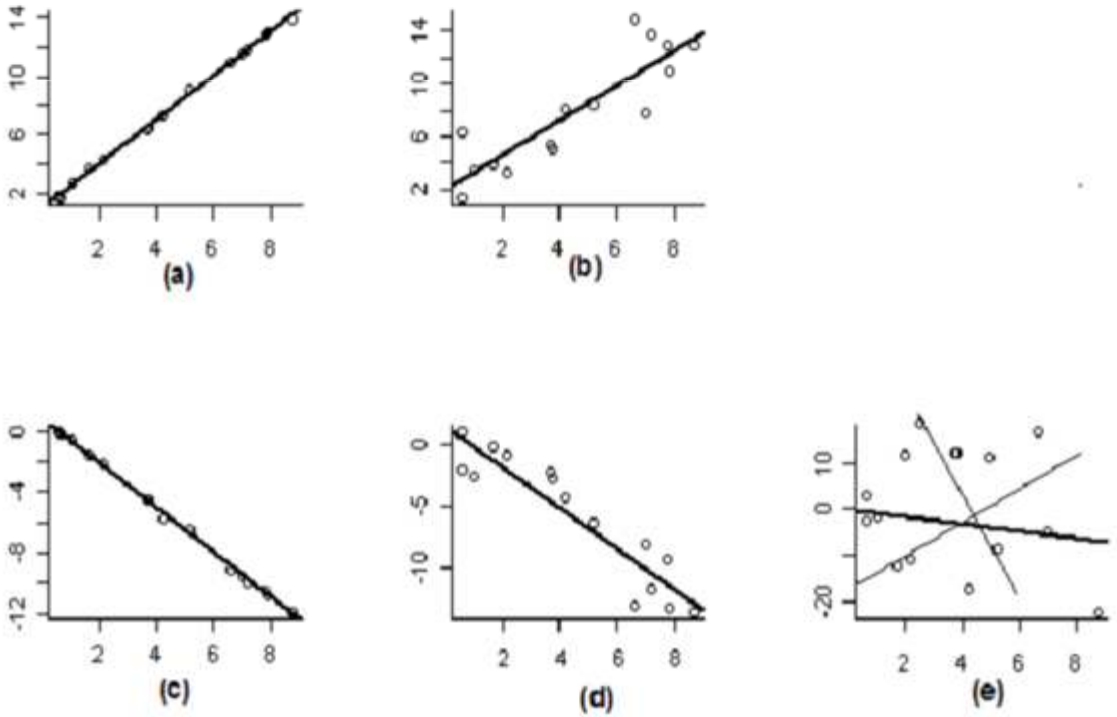


Figura7.8: Relação entre o comportamento do gráfico de dispersão e o coeficiente de correlação

Exemplo 7.16: Novamente a Temperatura e a Difusividade Térmica.

Consideremos, mais uma vez, os dados do Exemplo 7.14 para X = Temperatura (ºC) e Y = Difusividade Térmica (mm^2/s)

$$s_x^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{688553{,}1 - \frac{(4829{,}2)^2}{42}}{41} = 3250{,}9045$$

Tabela 7.7

Dai , $s_x = 57{,}0167$

A variância amostral de y é:

$$s_y^2 = \frac{\sum y_i^2 - \frac{(\sum y_i)^2}{n}}{n-1} = = \frac{7{,}678455 - \frac{(17{,}865)^2}{42}}{41} = 0{,}001938$$

Portanto, $s_y = 0{,}0440$

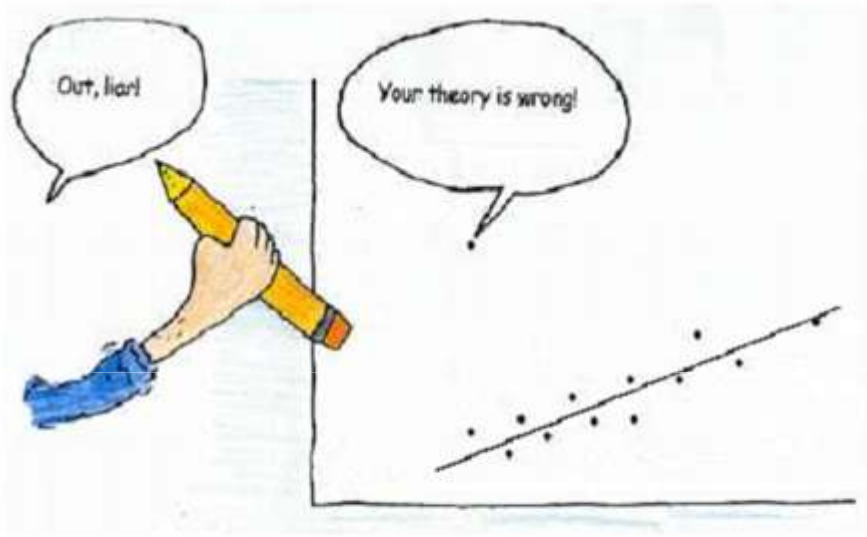
Desta forma, o coeficiente de correlação é:

$$r_{xy} = \frac{-2{,}0812}{(57{,}0167)(0{,}0440)} = -0{,}8296 \approx -0{,}83$$

O valor de r_{xy} mostra que há uma forte correlação linear negativa entre a Temperatura e a Difusividade Térmica, no caso de uma fibra de carbono sem envelhecimento.

Figura 7.7

Assim como a média e o desvio-padrão, o coeficiente de correlação r_{xy} é uma medida pouco resistente à presença de observações discrepantes. Isso quer dizer que se a um conjunto de pontos, todos situados em torno de uma reta, forem acrescentados alguns poucos pontos que estejam bastante afastados dessa reta, o módulo do coeficiente de correlação poderá diminuir substancialmente.



- Sua teoria está errada!

- Para fora, mentiroso!

Do mesmo modo que um coeficiente de correlação nulo, ou muito pequeno, em módulo, não implica a inexistência de algum tipo de relação entre X e Y, um valor relativamente alto de $|r_{xy}|$ não significa que há, necessariamente, uma relação de causa e efeito entre X e Y. Cabe ao ao pesquisador, determinar, conforme o seu conhecimento da natureza do problema, se o valor observado corresponde ou não à existência de uma efetiva relação entre as duas variáveis.

7.9.3 - Reta de Regressão

Quando se verifica através do coeficiente de correlação (e pelo próprio aspecto visual do diagrama de dispersão) que existe uma forte relação linear entre duas variáveis X e Y , pode ser de interesse calcular a equação da reta que representa esta relação: $y = a + b.x$

Y é a variável cujo comportamento se deseja explicar e

X a variável a ser usada para explicar o comportamento de Y.

Por isso Y é denominada variável resposta ou variável dependente e X a variável explicativa ou variável independente .

A equação da reta pode ser usada, por exemplo, para se estimar qual seria o valor y_0 da variável resposta Y correspondente a um determinado valor x_0 da variável preditora X.

Esse procedimento costuma ser utilizado principalmente nos caso em que a medição da variável resposta Y é mais cara, difícil ou demorada, enquanto a medição da variável explicativa X é mais barata, fácil ou rápida.

Suponha que se dispõe de um conjunto de n pares de dados (x_i , y_i) como nos exemplos anteriores. Então pode ser usado o que denominamos **método dos mínimos quadrados** para se obter a equação da reta que melhor se ajusta aos n pontos correspondentes a esses dados no plano bidimensional. Como, em geral, a relação de dependência linear entre X e Y não é perfeita, costuma-se introduzir na equação um termo relativo ao erro do modelo de predição: $y = a + b.x + erro.$

Assim , para cada um dos pontos (x_i , y_i) , mantendo fixos a e b, podemos escrever :

$$y_i = a + b.x_i + (erro)_i$$

O **método dos mínimos quadrados** para determinar os coeficientes a e b, consiste em encontrar aqueles valores de a e b que minimizam:

$$\sum_{i=1}^{n}(y_i - (a + bx_i))^2 ,$$

ou seja, a soma dos quadrados das n diferenças entre os dois valores de Y : o observado (y_i) e o valor ajustado $\hat{y}_i$, calculado através da equação da reta (a +bx_i).

As fórmulas que nos permitem calcular os valores de a e b a partir dos dados são:

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} = \frac{s_{xy}}{s_x^2}$$

$$a = \bar{y} - b\bar{x} = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}$$

A interpretação do significado dos coeficientes a e b é a habitual, ou seja:

- O coeficiente **b** mede a inclinação ou coeficiente angular da reta de regressão. Então, ao passarmos de um ponto a outro sobre a reta, b mede a relação $\frac{\Delta y}{\Delta x}$, onde Δy e Δx representam, respectivamente, as variações de y e de x .

- O coeficiente **a** mede o valor de y quando x é igual a zero, ou seja, o intercepto da reta de regressão.

• Exemplo 7.17: Mais uma vez **Temperatura e Difusividade Térmica**

Usaremos os dados do Exemplo 7.14 para determinar a reta de regressão $y = a + bx$, onde

Y representa a Difusividade Térmica , em mm^2/s, e

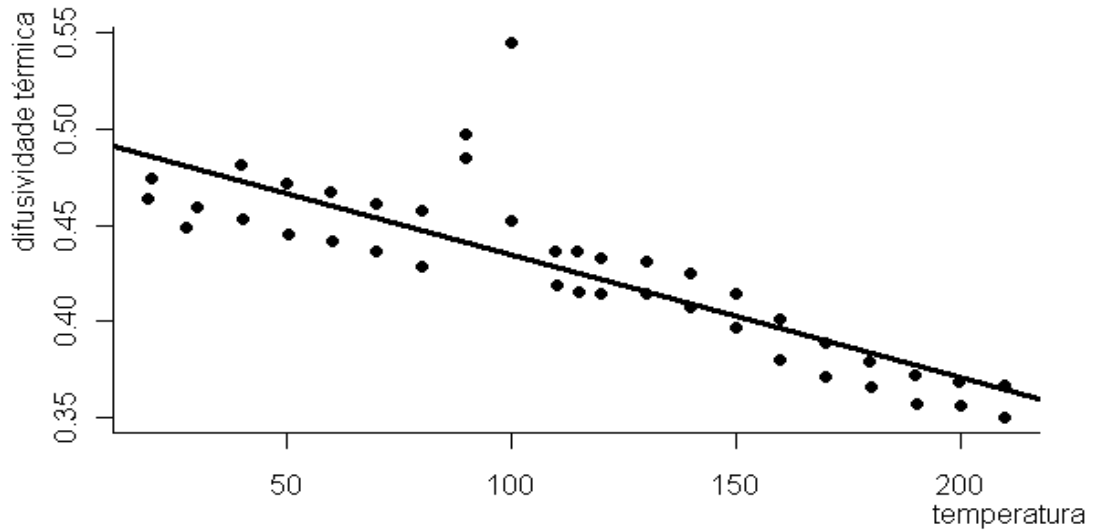
X , a Temperatura, em graus Celsius.

Para o cálculo de a e b temos :

$$b = \frac{s_{xy}}{s_x^2} = \frac{-2{,}0812}{3250{,}9045} = -0{,}00064$$

$$a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n} = \frac{17{,}865-(-0{,}00064)(4829{,}2)}{42} = 0{,}4987$$

A equação da reta de regressão ajustada aos dados é: $y = 0{,}4987 - 0{,}00064x$



A equação de regressão obtida mostra que:

1) Para cada 1° C de aumento da temperatura, a difusividade térmica experimenta uma diminuição de 0,00064 mm^2 / s

2) Para um dado valor de X, digamos $x_0 = 125\ ^oC$, a difusividade térmica esperada é

$$y_0 = 0{,}4987 - 0{,}00064 \times 125 = 0{,}4187 \approx 0{,}419\ mm^2 / s$$

Note que este valor é coerente com os dados da Tabela 7.7

Precaução: Devemos **evitar fazer extrapolações** sobre o comportamento de Y para valores de X fora do intervalo considerado no experimento.

Por exemplo, no caso do exemplo anterior, somente com base nos dados observados, **não podemos inferir** qual seria o verdadeiro comportamento da difusividade térmica **para valores muito altos ou muito baixos** da temperatura.