

APONTAMENTOS  
DE  
ESTATÍSTICA APLICADA

## Índice

1. Conceitos Estatísticos Básicos.....	3
1.1.Introdução.....	3
1.2 Objectivo da Estatística.....	3
1.3. O Método Científico e a Análise Estatística .....	3
1.4.O Processo de Pesquisa e o seu Desenho .....	4
1.4.1. Definição do problema, objectivos e hipóteses .....	4
1.4.2. Tipos de Pesquisa e Métodos de Recolha de Dados .....	4
1.4.3. População, Amostra e Métodos de Amostragem .....	5
1.5.Tipos de Dados.....	6
1.5.1.Variáveis Qualitativas e Quantitativas .....	6
1.5.2.Escalas de Medida .....	7
1.6. Estatística Descritiva, Inferência Estatística, Estatísticas e Parâmetros.....	8
2. Estatística Descritiva .....	10
2.1. Introdução.....	10
2.2. Tabelas de Distribuição de Frequências.....	10
2.2.1. Variáveis Qualitativas .....	10
2.2.2. Variáveis Quantitativas .....	11
2.3. Representação Gráfica de Distribuições de Frequência.....	13
2.4. Indicadores Numéricos.....	15
2.4.1. Medidas de Tendência Central: Média, moda e mediana .....	15
2.4.2. Medidas de Partição: Quartis, Decis e Percentis.....	20
2.4.3. Medidas de Dispersão .....	22
2.5. Assimetria e Curtose. Diagramas de Caixa e Bigodes (“Box Plot”).....	26
2.5.1. Assimetria.....	26
2.5.2. Curtose .....	28
2.5.2. Diagramas de Caixa e Bigodes.....	30
2.6. Tabelas de contingência .....	31
2.7. Diagramas de dispersão e coeficientes de correlação .....	33

## **1. Conceitos Estatísticos Básicos**

### **1.1. Introdução**

Qualquer um de nos, enquanto cidadãos do mundo moderno, está exposto a um enorme conjunto de informação resultante de estudos sociológicos, económicos e de mercado, de sondagens políticas e de estudos científicos. Muita desta informação baseia-se em resultados colhidos junto de alguns elementos (amostra) da população. Para que a informação seja relevante a amostra deve ser representativa, ter uma dimensão adequada e ser seleccionada aleatoriamente. Caso estes pressupostos não se verifiquem não se pode fazer extrapolação dos resultados obtidos na amostra para a população. Assim, o conhecimento da estatística permite que se avaliem os métodos de recolha de dados, os resultados e as conclusões definidos num dado estudo permitindo que se detectem falsas conclusões.

Para além da sua utilidade ao nível cívico a necessidade de saber estatística encontra-se intimamente ligada ao exercício de profissões no domínio da engenharia, da economia, da psicologia e da sociologia, sendo ainda uma ferramenta indispensável à investigação científica.

### **1.2 Objectivo da Estatística**

O objectivo fundamental deste ramo do conhecimento consiste na recolha, compilação, análise e interpretação de dados, havendo a necessidade de se distinguir entre estatística descritiva e inferência estatística.

No âmbito da estatística descritiva procura-se sintetizar e representar de forma compreensível e sintetizada a informação contida num conjunto de dados. Esta tarefa concretiza-se na construção de tabelas e gráficos e no cálculo de valores que representem a informação contida nos dados.

O objectivo da inferência estatística consiste, em última análise, em fazer previsões a partir da parte para o todo, ou seja, com base na análise de um conjunto limitado de alguns dados (amostra) recolhidos junto de um conjunto total de indivíduos (população), pretendemos caracterizar a população.

### **1.3. O Método Científico e a Análise Estatística**

A metodologia utilizada na análise estatística tem um paralelismo evidente com o método científico. Esta metodologia consiste em cinco passos fundamentais:

- i) Estabelecer o objectivo da análise e definir a população
- ii) Conceber o procedimento mais adequado para a recolha de dados
- iii) Proceder à recolha de dados
- iv) Analisar os dados
- v) Inferir acerca da população

#### **1.4.O Processo de Pesquisa e o seu Desenho**

##### **1.4.1. Definição do problema, objectivos e hipóteses**

De modo a se proceder a qualquer análise ou a estabelecer uma hipótese sobre um determinado conjunto de dados é sempre necessário identificar correctamente o problema. Isto é, saber qual a informação relevante para o problema que se pretende estudar<sup>1</sup>.

##### **1.4.2. Tipos de Pesquisa e Métodos de Recolha de Dados**

Para além de se dever recolher informação relevante a recolha deve ser realizada em tempo útil podendo ser obtida a partir de diversas fontes e de formas diversas.

Se os dados forem recolhidos directamente pelo analista através de inquéritos ou determinações laboratoriais os dados dizem-se primários. Se os dados forem recolhidos e publicados por pessoas ou instituições das quais o analista não depende dizem-se secundários. São deste tipo os dados que podemos obter junto de institutos governamentais ou associações empresariais

A recolha de dados primários pode ainda ser efectuada recorrendo a dois tipos de processos: processos experimentais e processos observacionais.

Nos processos experimentais exerce-se um controle directo sobre os factores que potencialmente afectam a característica ou o conjunto de características em análise

##### Exemplo:

Para estudar o efeito poluente de uma fábrica sobre a água de um rio efectuaram-se medições da concentração de um determinado contaminante sobre um conjunto de amostras colhidas em vários pontos do rio a jusante da fábrica. Metade das amostras foram

---

<sup>1</sup> Um estudante de biologia interessado no comportamento das rãs pretendia saber quais os factores que influenciavam a audição destes animais. Para tal começou por pegar numa rã e cortar-lhe uma perna, pousando-a em seguida e sussurrando-lhe ao ouvido :- salta rã!, ordem que a rã obedeceu imediatamente. Repetiu a experiência com várias rãs e obteve os mesmos resultados. No estágio seguinte da experiência o estudante recorreu às rãs previamente amputadas e cortou-lhes uma segunda perna mas, desta vez, ao pousa-las sussurrando – salta rã!, os animais mantiveram-se imóveis. Conclusão do estudante: As rãs sem pernas não ouvem...

colhidas no final de dias úteis e a outra metade durante o fim-de-semana. É de esperar que as concentrações medidas na primeira metade das amostras sejam superiores às medidas na segunda metade. Neste caso o momento em que se efectua a recolha da amostra influencia os resultados da análise.

Nos processos observacionais os factores que potencialmente afectam as características a analisar não são controlados.

Exemplo:

No âmbito de um estudo de tráfego num túnel rodoviário procurou-se analisar a relação entre a densidade do tráfego e a velocidade média de circulação. Para se proceder a esse estudo efectuaram-se medições em simultâneo da velocidade e da densidade do tráfego ao longo de um mês. Neste caso verifica-se que o momento em que é efectuada a medição não tem influência sobre os dados já que em cada momento existirá sempre uma relação única entre os valores observados.

### **1.4.3. População, Amostra e Métodos de Amostragem**

Designa-se por Universo ou População o conjunto de dados que expressam a característica que se pretende medir para a totalidade dos indivíduos que constituem o objecto da análise. Designa-se por amostra um subconjunto dos dados pertencentes à população:

Exemplos:

1) População: intenção de voto dos eleitores de uma cidade

Amostra: intenção de voto de alguns dos eleitores dessa cidade seleccionados a partir da lista telefónica.

2) População: Consumo de um novo detergente pelos clientes de um supermercado

Amostra: Consumo do produto recolhido por entrevista à porta do supermercado

A dimensão da população pode ser finita ou infinita. Frequentemente as populações, apesar de finitas, têm uma dimensão tão elevada que se torna mais simples tratá-las como infinitas. Assim, a dimensão da amostra pode ser um dos factores para que se proceda à análise por amostragem.

Exemplos

- População Finita (susceptível de ser tratada como tal): conjunto das intenções de voto dos eleitores de uma freguesia;
- População Infinita (susceptível de ser tratada como infinita): conjunto das alturas dos portugueses com mais de 18 anos;

- População Infinita: conjunto das pressões atmosféricas que se verificam num determinado instante à superfície terrestre.

Para além da dimensão da população existem outras razões que podem contribuir para não se analisarem todos os elementos de uma população, entre elas distinguem-se as seguintes:

- 1- Custo excessivo do processo de recolha e tratamento de dados, como resultado da sua elevada dimensão ou da sua complexidade de caracterização
- 2- Tempo excessivo de recolha e tratamento dos dados, que pode conduzir à obtenção de informação desactualizada (por alteração da população) ou obsoleta (por exceder o prazo de utilidade da informação)
- 3- Destruição da população provocada pelos métodos de recolha de informação
- 4- Inacessibilidade a alguns elementos da população (por exemplo, por razões de carácter legal)

De modo que a análise feita sobre a amostra possa ser extrapolada para a população a amostragem deve seguir algumas regras. O processo de amostragem deve recorrer a métodos probabilísticos nos quais cada um dos elementos da população tem uma certa probabilidade (conhecida) de ser incluído na amostra. Dentre estes métodos o mais utilizado é o de amostragem aleatória. Este processo garante que todos os elementos da população têm a mesma probabilidade de serem incluídos na amostra e, através dele, consegue-se evitar qualquer enviesamento no processo de selecção, ou seja, é afastada qualquer tendência sistemática para sub representar ou sobre representar na amostra alguns elementos da população.

## **1.5. Tipos de Dados**

Os dados podem ser qualitativos ou quantitativos dependendo do acontecimento a medir. A intenção de voto dos eleitores de uma freguesia constitui um conjunto de dados qualitativos. A duração em horas de um lote de lâmpadas constitui um conjunto de dados quantitativos, ou seja, os dados quantitativos são expressos por um valor numérico.

### **1.5.1. Variáveis Qualitativas e Quantitativas**

Constituem variáveis qualitativas a profissão, o sexo, a raça, a localização geográfica, o sector de actividade económica (por exemplo).

Como exemplos de variáveis quantitativas podemos referir, a idade, o peso, a distância, a temperatura, a altitude, o nº de trabalhadores nos diferentes departamentos de uma empresa...

### **1.5.2. Escalas de Medida**

Dependendo do tipo de variáveis que constituem os dados estes podem ser expressos em quatro escalas distintas: nominal, ordinal, por intervalo e por rácios.

Os dados qualitativos exprimem-se nas duas primeiras escalas e os dados quantitativos nas duas últimas. Em relação a este tipo de dados devemos distinguir os que constituem variáveis discretas dos que constituem variáveis contínuas. Considerem-se os valores que se podem obter dos seguintes acontecimentos:

- resultados de 150 lançamentos de um dado.
- distância diária em km a percorrer por um vendedor no próximo mês.

No primeiro caso os dados podem tomar valores pertencentes a um conjunto finito: {1,2,..., 6} dizendo-se nesse caso que os dados são discretos ou que são realizações de uma variável aleatória discreta. No segundo caso, se admitirmos que as distâncias podem ser medidas com precisão absoluta, existem um número infinito de distâncias diárias, entre um valor mínimo e um valor máximo, que o vendedor pode percorrer diariamente. Nesta situação os dizem-se contínuos ou realizações de uma variável aleatória contínua.

Escala Nominal: Suponha-se que para elaborar um estudo de mercado se pretendia conhecer a profissão de uma população constituída pelos consumidores de um determinado produto. Esse tipo de estudo conduziria a uma lista onde, por exemplo, se incluiria:

- Trabalhador não qualificado
- Trabalhador qualificado
- Professor
- Engenheiro
- Médico
- Advogado
- Etc.

Suponha-se que para efeitos de processamento dos dados se atribuía um código a cada um dos elementos da lista anterior: 1 para trabalhador não qualificado; 2 para trabalhador qualificado, 3 para professor, etc. Apesar desta codificação não podemos considerar os dados como quantitativos uma vez que não é possível estabelecer uma ordem entre eles, isto é, a profissão codificada com 1 não é menor nem maior que a profissão codificada com 2.

Um caso particular deste tipo de escala de medida ocorre quando a característica em estudo envolve apenas duas categorias (por exemplo o sexo, ou questões que apenas podem ser respondidas com “sim” ou “não”). Essas características são denominadas binárias ou dictômicas.

Escala Ordinal: esta escala de medida pode ser construída a partir de escalas nominais quando existe paralelismo evidente entre a escala nominal e uma sequência crescente ou decrescente com significado. Por exemplo, pode-se perguntar a um consumidor qual é a sua opinião sobre um determinado produto alimentar de acordo com a seguinte lista:

- detesta
- gosta pouco
- indiferente
- gosta
- adora

Sendo evidente que esta lista corresponde a uma sequência ordenada com cinco categorias.

Escala por Intervalos: este tipo de escala é usada com dados quantitativos tanto discretos como contínuos sendo que a distância entre os valores que constituem os intervalos deve ser igual. O número de automóveis que atravessa a ponte da Arrábida em cada hora pode ser definido numa escala por intervalos de valores discretos, por exemplo, entre 0 e 150; entre 150 e 300; entre 300 e 450, etc. A temperatura mínima diária do ar em °C numa estação meteorológica num determinado ano pode ser definido numa escala por intervalos de valores contínuos, por exemplo, [-5, 0]; [0, 5]; [5, 10]; [10, 15]; etc.

Escala por Rácios: As escalas deste tipo têm as mesmas propriedades que as escalas por intervalos para variáveis contínuas e, adicionalmente, apresentam a característica de possuírem um zero absoluto como valor mínimo de modo que as razões entre duas medidas têm sempre o mesmo valor qualquer que seja a unidade utilizada. Por exemplo o peso pode constituir uma escala por rácios (a razão entre os pesos de dois pacotes de açúcar, por exemplo, é sempre o mesmo qualquer que seja a unidade de medida: g, kg, ton., etc.) mas a temperatura não ( $10^{\circ}\text{C} = 50^{\circ}\text{F}$ ;  $30^{\circ}\text{C} = 86^{\circ}\text{F}$  porém  $10/30 \neq 50/86$ )

## **1.6. Estatística Descritiva, Inferência Estatística, Estatísticas e Parâmetros**

A estatística descritiva ocupa-se da recolha, classificação e organização de dados permitindo elaborar conclusões apenas para o conjunto limitado de indivíduos que serviram de base à recolha desses dados. Pelo contrário. A inferência estatística permite estimar as características desconhecidas de uma população, mesmo que a população não tenha sido analisada na totalidade, e testar se são plausíveis determinadas hipóteses formuladas sobre

essas características (por exemplo, permite testar se é verdadeira ou falsa a afirmação de um vendedor de detergentes para automóveis quando diz que os resultados da lavagem da marca que vende são superiores aos da concorrência)

Quando calculamos determinados valores sobre o conjunto de dados que constituem a amostra designamos esses valores por “estatísticas”. Quando esses mesmos valores dizem respeito à população designam-se por “parâmetros”.

Os métodos de inferência estatística envolvem o cálculo de estatísticas a partir dos quais se inferem os parâmetros da população, ou seja, permitem, com determinado grau de probabilidade, generalizar para a população determinadas características dos resultados amostrais.

Exemplo:

Um fabricante de máquinas de lavar pretende determinar o número médio de lavagens efectuado por um determinado modelo de máquina até necessitar de reparação. O responsável pela qualidade selecciona aleatoriamente algumas máquinas produzidas mensalmente e regista para cada uma o número de lavagens efectuadas até ocorrer uma avaria, calculando em seguida o nº médio de lavagens em cada amostra (média amostral – estatística) concluindo que os valores obtidos podem ser extrapolados para o nº médio de lavagens dos lotes mensais (média populacional – parâmetro cujo valor real é desconhecido).

## 2. Estatística Descritiva

### 2.1. Introdução

A estatística descritiva pretende organizar, sintetizar e analisar os dados obtidos no estudo de variáveis relativas a uma população de modo a permitir caracterizar a população e conhecer o seu comportamento.

A informação fornecida pelos dados é compilada e sintetizada em tabelas e gráficos e através do cálculo de indicadores numéricos. O desafio da Estatística Descritiva consiste não na própria construção das tabelas ou dos gráficos mas na escolha mais adequada destas ferramentas de modo a caracterizar correctamente as variáveis em estudo

### 2.2. Tabelas de Distribuição de Frequências

Nas tabelas de distribuição de frequências representa-se a forma como uma dada variável se encontra distribuída pelo conjunto dos indivíduos em que essa variável foi analisada, tendo aplicação tanto em variáveis qualitativas como quantitativas.

#### 2.2.1. Variáveis Qualitativas

Suponha-se que se pretende estudar a marca de computadores portáteis preferida pelos estudantes do ensino superior. Tendo-se questionado 50 estudantes obtiveram-se os dados representados na Tabela 2.1.

Tabela 2.1. Marca de Computadores Portáteis preferida por 50 estudantes do Ensino Superior

COMPAQ	HP	COMPAQ	COMPAQ	COMPAQ	ACER	ACER	ACER	FUJITSU	TOSHIBA
TOSHIBA	COMPAQ	COMPAQ	TOSHIBA	TOSHIBA	ACER	ACER	COMPAQ	FUJITSU	COMPAQ
FUJITSU	IBM	FUJITSU	IBM	COMPAQ	TOSHIBA	FUJITSU	FUJITSU	IBM	TOSHIBA
ACER	TOSHIBA	IBM	IBM	IBM	IBM	FUJITSU	COMPAQ	TOSHIBA	ACER
FUJITSU	COMPAQ	ACER	IBM	IBM	TOSHIBA	COMPAQ	TOSHIBA	COMPAQ	TOSHIBA

Para construir a tabela de frequências deve-se proceder à contagem do número de vezes que cada marca é referida. Verifica-se que a marca COMPAQ foi referida 13 vezes pelo que este valor será a frequência para essa marca (neste caso, a frequência absoluta, em contraste com a frequência relativa que é dada em percentagem por  $\frac{13}{50} \times 100 = 26\%$  )

Na tabela de frequências é usual representar a frequência absoluta,  $n_i$ , a frequência relativa,  $f_i$ , e a frequência relativa acumulada,  $F_i$ .

A tabela de frequências para os dados apresentados na Tabela 2.1 é dada pela Tabela 2.2. e permitem saber como se distribuem as preferências dos 50 elementos da amostra relativamente às marcas de computadores portáteis.

Tabela 2.2. Distribuição de frequências para a marca de computadores portáteis preferidos

Marca de PC	Freq. Abs $n_i$	Freq. Relat. (%) $f_i$	Freq. Relat. Acumulada (%) $F_i$
COMPAQ	13	26	26
HP	1	2	26+2=28
TOSHIBA	11	22	28+22=50
ACER	8	16	50+16=66
IBM	12	24	66+24=90
FUJITSU	5	10	90+10=100
$\Sigma$	50	100	

A marca mais frequente é a COMPAC, com 26 % das preferências, seguida pela IBM, com 24 %, podendo verificar-se as preferências relativas para as 6 marcas analisadas.

A frequência relativa acumulada permite verificar que metade dos indivíduos da amostra prefere três marcas e a outra metade prefere as marcas restantes (neste caso, também três).  $F_i$  obtém-se somando ordenadamente as frequências relativas de cada variável.

## 2.2.2. Variáveis Quantitativas

No caso de variáveis quantitativas discretas, desde que o espaço amostral não tenha dimensões muito elevadas, o procedimento utilizado para construir tabelas de frequências é idêntico ao que foi apresentado para as variáveis qualitativas.

Suponha-se uma experiência concebida para verificar se um dado é ou não viciado. A experiência consistiu em lançar os dados 112 vezes e registar o resultado obtido em cada lançamento. O espaço amostral desta experiência é um conjunto discreto, limitado e de dimensão reduzida, correspondendo a  $A = \{1, 2, 3, 4, 5, 6\}$ .

Os resultados obtidos na experiência encontram-se na Tabela 2.3. A Tabela 2.4 apresenta a respectiva distribuição de frequências

Tabela 2.3. Resultados obtidos em 112 lançamentos de um dado

2	5	4	2	2	4	5	6	4	5	5	3	6	3
4	4	2	5	6	5	2	5	4	3	5	2	3	5
3	6	3	5	5	3	2	1	4	2	4	2	4	3
3	2	3	4	5	5	5	5	1	2	4	5	5	3
3	5	1	6	5	3	1	5	1	4	5	5	4	1
2	2	3	2	1	4	6	3	5	4	3	1	5	5

3	4	2	4	6	3	4	4	2	5	1	3	4	5
6	5	3	5	3	2	3	2	5	3	5	2	5	5

Tabela 2.4. Distribuição de frequências para os resultados do lançamento de um dado

Pontos	$n_i$	$f_i$	$F_i$
1	9	8,035714	8,035714
2	19	16,96429	25
3	23	20,53571	45,53571
4	20	17,85714	63,39286
5	33	29,46429	92,85714
6	8	7,142857	100
$\Sigma$	112	100	

Vamos agora estudar o caso de uma variável quantitativa contínua. Suponhamos que se registou a duração, em horas, de uma amostra de 40 pilhas extraídas ao acaso de um lote de produção diário. Os valores observados constam da Tabela 2.5.

Tabela 2.5. Resultados obtidos em 40 observações efectuadas à duração em horas de lote de pilhas

45,7	44,7	44,5	44,1	45,3	45,6	45,2	43,8	44,2	45,2
45,0	44,8	43,7	44,4	43,8	43,9	43,9	44,2	44,7	45,5
45,8	44,5	45,0	44,7	44,2	44,9	43,7	45,5	44,9	45,7
45,0	43,8	44,9	44,5	45,3	44,0	43,6	43,9	44,5	44,4

Conforme se pode verificar, entre o valor mínimo (43,6 horas) e o valor máximo (45,8 horas) existe um número infinito de valores, alguns observados, outros não. Nesse caso recorre-se à distribuição dos dados por classes ou células. A regra para se saber qual o número de classes a considerar consiste em determinar a raiz quadrada do número de observações, N:

$$K = N^{\circ} \text{ de Classes} = \sqrt{N}$$

Esta regra não é estrita, fornecendo unicamente uma ordem de grandeza. No caso que exemplificamos  $\sqrt{40} = 6,6(6)$  porém, por razões de natureza prática e de melhor compreensão dos dados só foram consideradas 5 classes com amplitude de 0,5 horas. Para determinar a amplitude de cada classe efectua-se a razão entre a diferença do valor máximo e mínimo observados pelo número de classes.

A rotina mais adequada para construir uma tabela de frequências com dados agrupados em classes é:

- 1º) Calcular K e arredondar convenientemente (no exemplo anterior K=7)
- 2º) Considerar  $K_1=K$ ;  $K_2=K+1$  e  $K_3=K-1$  (No exemplo anterior  $K_1=7$ ;  $K_2=8$ ;  $K_3=6$ )
- 3º) Determinar as respectivas amplitudes,  $a_i$ , para os diferentes valores de  $K_i$  ( $i=1,2,3$ ). Ou seja:

$$a1 = \frac{45,8 - 43,6}{7} = 0,3143; a2 = \frac{45,8 - 43,6}{8} = 0,275; a3 = \frac{45,8 - 43,6}{6} = 0,36(6)$$

4º) Seleccionar a amplitude mais consistente com os dados (neste caso, a2)

Voltando à decisão inicial (meramente prática) de representar 5 classes com amplitude 0,5h, constrói-se a tabela de frequências representada na Tabela 2.6.

Tabela 2.6. Distribuição de frequências para os resultados da duração em horas de uma amostra de pilhas

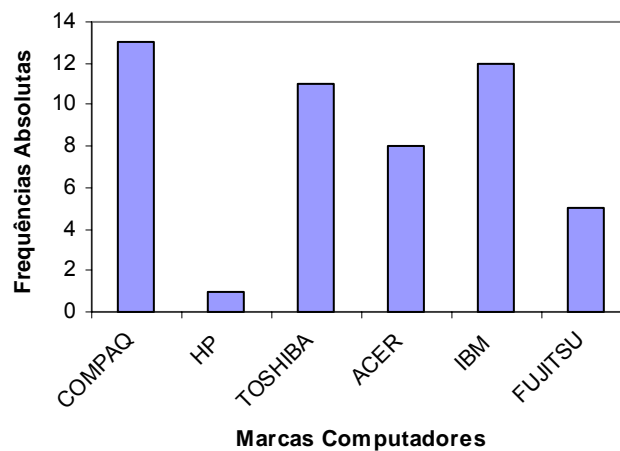
Duração (h)	ni	fi (%)	Fi (%)
[43,5 a 44[	9	22,5	22,5
[44 a 44,5[	7,0	17,5	40,0
[44,5 a 45[	11,0	27,5	67,5
[45 a 45,5[	7,0	17,5	85,0
[45,5 a 46[	6,0	15	100,0
$\Sigma$	40	100	

Na Tabela 2.6. podemos concluir 67,5% da amostra tem duração inferior a 45 horas, enquanto que 32,5% tem duração superior ou igual a 45 hs.

### 2.3. Representação Gráfica de Distribuições de Frequência

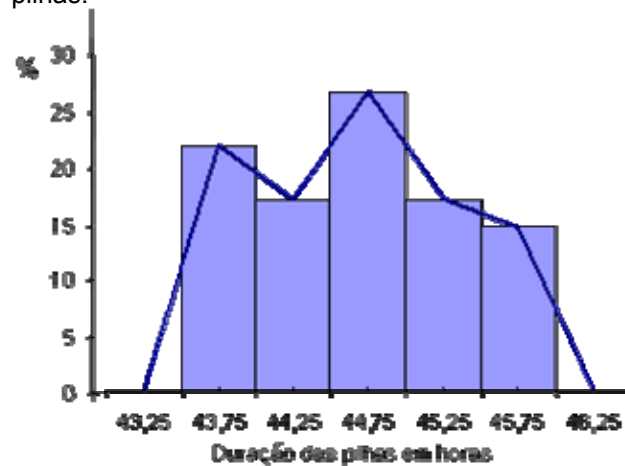
As frequências absolutas e relativas para variáveis quantitativas discretas ou qualitativas podem ser representadas em gráficos de barras, conforme se ilustra na Figura 2.1.

Figura 2.1. Gráfico de colunas para as marcas preferidas de PC portáteis



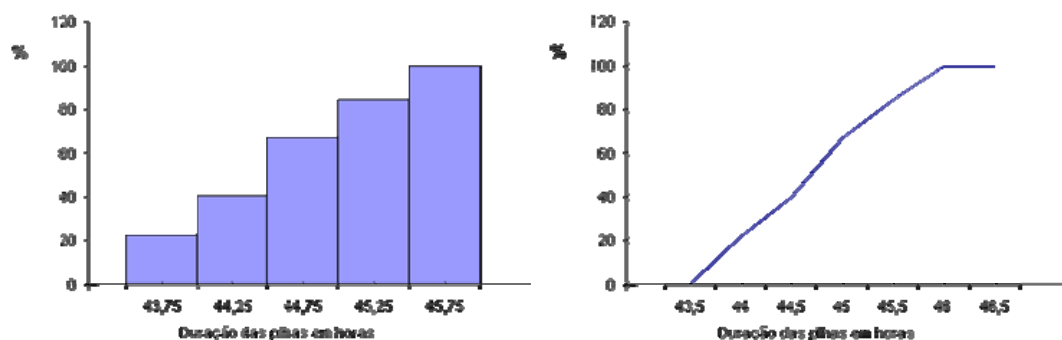
As frequências absolutas e relativas para variáveis quantitativas contínuas podem ser representadas por histogramas (colunas contíguas) ou polígonos de frequência (linha poligonal), conforme se ilustra na Figura 2.2.

Figura 2.2. Histograma e polígono de frequências relativas para a duração em horas de uma amostra de pilhas.



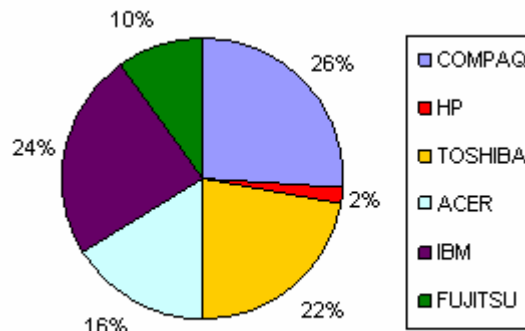
Podemos igualmente construir um histograma de frequências relativas (ou absolutas) acumuladas e o respectivo polígono de frequências acumuladas (também denominado Ogiva). Na construção do polígono de frequências acumuladas considera-se o limite inferior da primeira classe como tendo frequência acumulada nula – temos assim o par ordenado (43,5; 0) para o exemplo da Tabela 2.6 – a frequência acumulada da primeira classe será considerada no limite inferior da classe – e temos o par ordenado (44; 22,5) para o exemplo da Tabela 2.6. Deste modo, a frequência acumulada da última classe é marcada para o limite superior dessa classe. Na Figura 2.3 apresenta-se o histograma de frequências relativas acumuladas e a respectiva ogiva para os dados da Tabela 2.6.

Figura 2.3. Frequências relativas acumuladas e Ogiva para a duração em horas de uma amostra de pilhas.



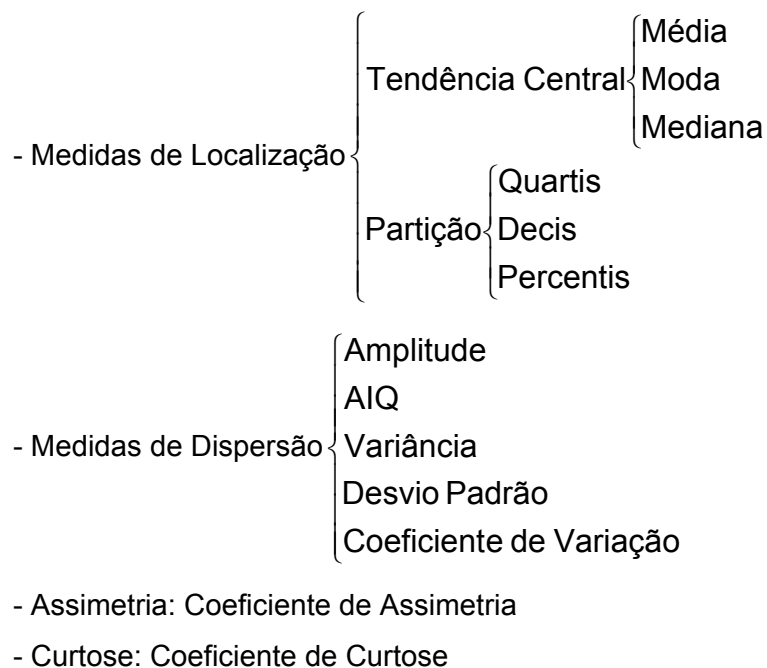
Variáveis qualitativas ou quantitativas, tanto discretas como contínuas, podem ser representadas em diagramas circulares (Figura 2.4.)

Figura 2.4. Diagrama circular para as marcas preferidas de PC portáteis (frequências relativas)



## 2.4. Indicadores Numéricos

Os indicadores numéricos resumem a informação contida nos dados e, quando calculados para uma amostra, denominam-se estatísticas. Estes indicadores podem ser classificados em:



### 2.4.1. Medidas de Tendência Central: Média, moda e mediana

#### 2.4.1.1. Média

A média é uma medida de localização de tendência central, sendo representada por  $\bar{X}$  ou por  $\mu$  conforme se trate, respectivamente, da média amostral (estatística) ou da

média populacional (parâmetro). A média de um conjunto de dados quantitativos obtém-se somando todos os valores e dividindo o resultado pelo nº total de valores.

Sejam as seguintes idades, em anos, dum grupo de 15 pessoas:

39, 43, 41, 43, 39, 45, 39, 39, 39, 43, 41, 43, 39, 45, 39,

A idade média do grupo é dada por:

$$\bar{X} = \frac{7 \times 39 + 2 \times 41 + 4 \times 43 + 2 \times 45}{15} = \frac{617}{15} = 41,1(3)$$

Ou seja:

$$\bar{X} = \frac{\sum niXi}{N} = \sum fiXi$$

Com :  $Xi$  – valor das variáveis observadas

$ni$  – Frequência absoluta

$N$  – Nº total de dados

$fi$  – Frequência relativa

No caso dos dados estarem agrupados em classes a média será dada por:

$$\bar{X} = \frac{\sum niCi}{N} = \sum fiCi$$

sendo  $Ci$  o valor central de cada classe.

Exemplo: Pretende-se calcular o preço médio das habitações T2 numa cidade em 1993. Os dados são apresentados na Tabela 2.7.

Tabela 2.7. Distribuição de frequências para os resultados da duração em horas de uma amostra de pilhas

Preço (contos)	f abs. (fi)	Ci	Cixfi
[13600, 14800[	7	14200	99400
]14800, 16000[	15	15400	231000
[16000, 17200[	24	16600	398400
]17200, 18400[	27	17800	480600
]18400, 19600]	17	19000	323000
$\Sigma=$	90		1532400

$$\bar{X} = \frac{1532400}{90} = 17026,67 \text{ contos}$$

#### 2.4.1.2. Mediana

A mediana ( $\tilde{X}$ ) é o valor central de um conjunto de dados. No caso de dados brutos discretos, obtém-se ordenando os dados de forma crescente ou decrescente e identificando

o valor central, caso o nº de dados seja ímpar, ou a média entre os dois dados centrais, caso o nº de dados seja par.

Exemplo : Pretende-se determinar a mediana para os pontos obtidos em 25 lançamentos de um dado com os seguintes resultados:

1 1 1 2 2 2 3 4 5 5 4 4 5 6 6 6 1 3 3 3 4 4 4 5 5

1º) 25 é ímpar logo o local do valor mediano obtém-se pela expressão  $\frac{N+1}{2}$ , ou seja

$$Lme = \frac{25+1}{2} = 13$$

2º) Ordenando os dados:

1 1 1 1 2 2 2 3 3 3 3 4 4 4 4 4 5 5 5 5 6 6 6

Logo  $\tilde{X} = 4$

No caso dos dados estarem representados numa tabela de frequências temos que localizar a mediana através das frequências absolutas ou relativas acumuladas. Considerando o exemplo anterior temos:

Pontos	Freq abs.	Freq. Abs. Acum.
1	4	4
2	3	7
3	4	11
4	6	17
5	5	22
6	3	25

17 contém 13 logo  $\tilde{X} = 4$

No caso de frequências relativas acumuladas devemos determinar o valor em percentagem ao qual corresponde a posição da mediana. No exemplo apresentado a

posição em percentagem é  $\frac{13}{25} \times 100 = 52\%$

Pontos	Freq rel.	Freq. rel. Acum.
1	16	16
2	12	28
3	16	44
4	24	68
5	20	88
6	12	100

68% contém 52% logo  $\tilde{X} = 4$

O modo de proceder no caso do n° de dados ser par é análogo porém, neste caso, a

posição do valor mediano é dado pela expressão  $\left( \frac{\frac{N}{2} + \frac{N+2}{2}}{2} \right)$ . Para exemplificar

calculemos a mediana na seguinte distribuição de frequências:

**Idades (anos) de ingresso no Infântário numa amostra de 42 indivíduos:**

Idades (anos)	fi (%)	fi Acum. (%)
1	11,9	11,9
2	23,8	35,7
3	35,7	71,4
4	19,1	90,5
5	9,5	100
	100	

A posição da mediana é dada por  $\left( \frac{\frac{42}{2} + \frac{42+2}{2}}{2} \right) = \frac{43}{2} = 21,5$ , correspondendo em

percentagem a  $\frac{21,5}{42} \times 100 = 51,19\%$ . 71,4% contém 51,19% logo  $\tilde{X} = 3$

No caso de dados agrupados em classe o primeiro passo consiste em identificar a classe mediana pelos processos já descritos. Em seguida calcula-se o valor da mediana através da seguinte expressão:

$$\tilde{X} = li(me) + \frac{\frac{N}{2} - CumFabs(me-1)}{Fabs(me)} \times a(me)$$

Com:

$li(me)$ : limite inferior da classe mediana

$N$ : dimensão da amostra

$CumFabs(me-1)$ : cumulante das frequências absolutas na classe anterior à classe mediana

$Fabs(me)$ : frequência absoluta da classe mediana

$a(me)$ : amplitude da classe mediana

A expressão para frequências relativas é semelhante:

$$\tilde{X} = li(me) + \frac{0,5 - CumFrel(me-1)}{Frel(me)} \times a(me)$$

Com:

$li(me)$ : limite inferior da classe mediana

$CumFrel(me-1)$ : cumulante das frequências relativas na classe anterior à classe mediana

$Fabs(me)$ : frequência relativa da classe mediana

$a(me)$ : amplitude da classe mediana

Exemplo: Calcular o valor da mediana na seguinte distribuição de frequências:

classes	Fabs	FabsAcum
[35, 45[	5	5
[45, 55[	12	17
[55, 65[ ←	18 ←	35 ←
[65, 75[	14	49
[75, 85[	6	55
[85, 95]	3	58

O primeiro passo será localizar a classe mediana: Localização =  $\left( \frac{\frac{58}{2} + \frac{58+2}{2}}{2} \right) = 29,5$ , logo,

a classe mediana é [55, 65[ e  $\tilde{X} = 55 + \frac{29-17}{18} \times 10 = 61,67$  unidades de medida dos dados.

### 2.4.1.3. Moda

A moda ( $\hat{X}$ ) pode ser determinada para dados quantitativos e qualitativos. Para variáveis quantitativas discretas ou qualitativas é simplesmente a variável mais frequente (ou mais observada). Para variáveis quantitativas agrupadas em classes é necessário:

1º Identificar a classe modal

2º Determinar o valor da moda dentro da classe através das expressões:

$$\hat{X} = li(mo) + \frac{F(mo+1)}{F(mo-1) + F(mo+1)} \times a(mo)$$

ou

$$\hat{X} = li(mo) + \frac{f(mo+1)}{f(mo-1) + f(mo+1)} \times a(mo)$$

Com:

$li(mo)$ : limite inferior da classe modal

$F(mo+1)$ : frequência absoluta da classe seguinte à modal

$f(mo+1)$ : frequência relativa da classe seguinte à modal

$F(mo-1)$ : frequência absoluta da classe anterior à modal

$f(mo-1)$ : frequência relativa da classe anterior à modal

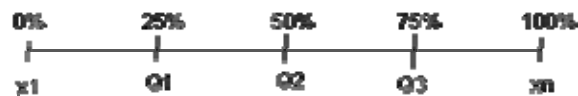
$a(mo)$ : amplitude da classe modal

Assim, no último exemplo tem-se:  $\hat{X} = 55 + \frac{14}{12+14} \times 10 = 60,4$  unidades de medida dos dados

## 2.4.2. Medidas de Partição: Quartis, Decis e Percentis

### 2.4.2.1. Quartis

Os quartis são os valores da variável observada que dividem a distribuição de frequências em 4 partes iguais.



Q1 – Primeiro Quartil – é o valor da variável observada tal que o nº de observações para valores inferiores a Q1 será de 25% e o nº de observações para valores superiores a Q1 será de 75%.

Q2 – Segundo Quartil – é o valor da variável tal que metade das observações encontram-se à sua esquerda e a outra metade à sua direita, logo, coincide com a mediana.

Q3 – Terceiro Quartil – é o valor da variável observada tal que o nº de observações para valores inferiores a Q3 será de 75% (3/4) e o nº de observações para valores superiores a Q3 será de 25% (1/4).

Para determinar os quartis, tal como ocorre na determinação da mediana, é necessário ordenar os dados.

Exemplo: O editor de uma obra literária pretende estudar as idades dos leitores da obra tendo obtido, numa amostra de 15 leitores, as seguintes respostas:

15 15 15 16 17 17 17 17 18 19 19 20 20 20 20

O primeiro passo para determinar os quartis é localizá-los:

$$Q1: \text{localização } \frac{1}{4} \times 15 = 3,75 \rightarrow 4 \Rightarrow Q1 = 16$$

$$Q2: \text{localização } \frac{2}{4} \times 15 = 7,5 \rightarrow 8 \Rightarrow Q2 = 17$$

$$Q3: \text{localização } \frac{3}{4} \times 15 = 11,25 \rightarrow 12 \Rightarrow Q3 = 20$$

Nas situações em que os dados se encontram agrupados em classes, após identificar a classe a que corresponde o quartil que pretendemos determinar, aplica-se a expressão seguinte:

$$Q_i = li(Q_i) + \frac{\frac{iN}{4} - CumFabsAnterior}{Fabs} \times a(Q_i)$$

Esta expressão também pode ser aplicada às frequências relativas.

Exemplo. Considere-se que se pretende determinar os quartis na distribuição de frequências para as classificações obtidas num teste de estatística. Os dados constam da tabela seguinte:

classes	Fabs	FrelAcum(%)
[0, 4[	27	24,5
[4, 8[	16	39
[8, 12[	34	70
[12, 16[	17	85,5
[16, 20[	16	100

Posição de Q1:  $110/4=27,5 \rightarrow 28$ . Em percentagem:  $\frac{28}{110} \times 100 = 25,5\%$

Posição de Q2:  $110/2=55 \rightarrow 55,5$ . Em percentagem:  $\frac{55,5}{110} \times 100 = 50,5\%$

Posição de Q3:  $(3 \times 110)/4=82,5 \rightarrow 83$ . Em percentagem:  $\frac{83}{110} \times 100 = 75,5\%$

Então:  $Q_1 = 4 + \frac{28 - 27}{16} \times 4 = 4,125$  valores

$$Q_2 = 8 + \frac{55 - 43}{34} \times 4 = 9,4 \text{ valores}$$

$$Q_3 = 12 + \frac{83 - 77}{17} \times 4 = 13,3 \text{ valores}$$

#### 2.4.2.2. Decis e Percentis

Os decis são os valores da variável que dividem a distribuição em 10 partes iguais. Os percentis dividem a distribuição em 100 partes iguais. O número de decis é 9 (do D1 até o D9) e o de percentis é 99 (do P1 ao P99).

À semelhança do que se fez no cálculo da mediana e dos quartis o primeiro passo consiste em determinar a posição destes indicadores. A posição do 1º decil é efectuada dividindo o nº de dados por 10, a do 2º decil obtém-se multiplicando o nº de dados por 2/10, etc. As posições dos percentis obtém-se multiplicando o nº de dados por (ordem do percentil)/100. Por exemplo, para calcular a posição do P72 é suficiente fazer  $\frac{72}{100} \times N$ .

Para dados agrupados em classes rcorre-se à seguinte expressão:

$$C_i \text{ ou } D_i = li + \frac{ni - CumFabsAnterior}{Fabs} \times a$$

Com:

$li$ : limite inferior da classe decil ou percentil

$ni$ : nº de observações até ao decil (percentil)

$CumFabAnteriors$ : cumulante das frequências absolutas na classe anterior à classe do decil (percentil)

$Fabs$ : frequência absoluta da classe decil (percentil)

$a$ : amplitude da classe do decil (percentil)

Exemplo: Calculem-se o D4 e o P72 da seguinte distribuição:

classes	F abs	Cum F
[4, 9[	8	8
[9, 14[	12	20
[14, 19[	17	37
[19, 24]	3	40

Posição de D4 =  $\frac{4}{10} \times 40 = 16 \rightarrow$  D4 pertence à classe [9, 14[

Posição de P72 =  $\frac{72}{100} \times 40 = 28,8 \rightarrow$  P72 pertence à classe [14, 19[

$$D4 = 9 + \frac{16 - 8}{12} \times 5 = 12,3$$

$$P72 = 14 + \frac{29 - 20}{17} \times 5 = 16,9$$

### 2.4.3. Medidas de Dispersão

As medidas de dispersão têm por finalidade verificar a representatividade das medidas de localização.

#### 2.4.3.1. Exactidão e Dispersão

Sejam as observações seguintes relativas às variáveis x e y

X	20	20	20	20	20
y	15	10	20	25	30

Verifica-se que  $\bar{X}(x) = 20$  e que  $\bar{X}(y) = 20$  mas, enquanto os valores de x não apresentam variações em relação à média, os valores de y apresentam variações significativas em torno do seu valor médio. Ou seja, x não apresenta qualquer dispersão em torno de  $\bar{X}(x)$  enquanto que y tem os seus valores dispersos em torno de  $\bar{X}(y)$ . A dispersão faz com que a medida considerada para a média possa não ser representativa por pouca exactidão. A exactidão é uma medida do desvio entre o valor obtido e o verdadeiro valor.

### 2.4.3.2. Erros Fortuitos e Erros Sistemáticos

As medidas efectuadas e os valores observados podem ser afectados por erros de medição ou de observação. Os erros são sistemáticos quando ocorrem sempre e da mesma maneira, podendo ser evitados. Os erros fortuitos são acidentais, acontecendo esporadicamente e não podendo ser evitados.

### 2.4.3.3. AIQ, Variância, Desvio-Padrão e Coeficiente de Variação

As medidas de dispersão dividem-se em três categorias principais:

- Medidas de distância – cujos valores estão representados nas mesmas unidades que os dados e onde não é necessário o cálculo de uma medida de localização, por exemplo o AIQ (Amplitude Inter Quartis);
- Medidas de desvio em relação a uma medida de localização que é utilizada como termo de comparação, por exemplo a variância e o desvio padrão;
- Medidas de dispersão relativa entre 2 ou mais conjuntos de dados, por exemplo o CV (Coeficiente de Variação)

#### AIQ

Esta medida de dispersão define-se pelo valor da diferença entre o 3º e o primeiro quartis:

$$AIQ = Q3 - Q1$$

medindo a amplitude que existe entre 50% das observações centrais.

Considerem-se os preços dos vários modelos de um determinado produto dados na tabela seguinte:

Modelo	A	B	C	D	E	F	G	H	I	J	K
Preço(€)	133	135	146	175	179	188	195	204	219	233	254

A amplitude total dos dados é  $254-133=121$  € e  $AIQ=219-146=73$  €, concluindo-se que das 11 observações registadas para os preços, as que correspondem a 50% dos valores centrais (entre o 3º e o 9º registo) têm uma variação de preço de 73 €. Este valor representa mais de metade da amplitude total dos dados pelo que a distribuição pode ser considerada dispersa.

#### Variância e Desvio Padrão

A variância e o desvio padrão fornecem uma medida da variabilidade dos dados em torno do seu valor médio.

A variância amostral ( $s^2$ ) é dada pelas seguintes expressões:

$$s^2 = \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1} \text{ para dados não agrupados (brutos)}$$

$$s^2 = \frac{N \sum_{k=1}^K f_k (x_k - \bar{x})^2}{N-1} \text{ com } f_k \text{ frequências relativas, e para dados discretos agrupados}$$

$$s^2 = \frac{N \left[ \sum_{k=1}^K f_k (C_k - \bar{x})^2 - \frac{a^2}{12} \right]}{N-1} \text{ com } f_k \text{ frequências relativas e para dados contínuos agrupados em classes, onde } C_k \text{ é o centro de cada classe e } a \text{ a amplitude das classes.}$$

Considere-se a tabela de frequências seguinte:

Pesos em gramas dos conteúdos de 100 garrafas	Freq. Relativa (%)	Freq. Relativa Acumulada (%)
[297, 298[	8	8
[298, 299[	21	29
[299, 300[	28	57
[300, 301[	15	72
[301, 302[	11	83
[302, 303[	10	93
[303, 304[	5	98
[304, 305[	1	99
[305, 306]	1	100
Totais	100	

De modo a calcular a variância amostral e o desvio padrão deste conjunto de dados contínuos agrupados em classes, teremos primeiro que fazer os cálculos que constam da tabela seguinte. Nessa tabela devemos ter o cuidado de reparar que, como  $N=100$ , os valores das frequências relativas são iguais aos valores das frequências absolutas, pelo que se no cálculo da média são assumidos como absolutos (por exemplo, a frequência absoluta da classe [302, 303[ é 10), no cálculo da variância devem ser assumidos como relativos (por exemplo, a frequência relativa da classe [302, 303[ é 0,1)

Pesos em gramas dos conteúdos de 100 garrafas	Freq. Relativa (%)	Freq. Relativa Acumulada (%)	ci	fi*ci	(ci-med(x))^2	fi/100*(ci-med(x))^2
[297, 298[	8	8	297,5	2380	6,8121	0,544968
[298, 299[	21	29	298,5	6268,5	2,5921	0,544341
[299, 300[	28	57	299,5	8386	0,3721	0,104188
[300, 301[	15	72	300,5	4507,5	0,1521	0,022815
[301, 302[	11	83	301,5	3316,5	1,9321	0,212531
[302, 303[	10	93	302,5	3025	5,7121	0,57121
[303, 304[	5	98	303,5	1517,5	11,4921	0,574605
[304, 305[	1	99	304,5	304,5	19,2721	0,192721
[305, 306]	1	100	305,5	305,5	29,0521	0,290521
Totais	100			30011		3,0579

Aplicando as fórmulas teremos então:

$$\text{Média } \bar{x} = \frac{\sum f_i c_i}{N} = \frac{30011}{100} = 300,11g$$

Amplitude das células  $a=1g$

$$\text{Variância } s^2 = \frac{N \sum \frac{f_i}{100} (c_i - \bar{x})^2}{N - 1} = \frac{100 \times 3,0579}{99} = 3,005 \text{ g}^2$$

$$\text{Desvio padrão } s = \sqrt{s^2} = \sqrt{3,005} = 1,733 \text{ g}$$

O desvio padrão, dado pela raiz quadrada da variância, tem a vantagem de ser expresso nas mesmas unidades dos dados a partir dos quais foi calculado, podendo ser interpretado como o valor absoluto de um desvio “típico” dos dados em relação à média amostral.

### **Coeficiente de Variação**

O coeficiente de variação (CV) é particularmente útil quando se pretende tirar conclusões acerca da representatividade da média como medida estatística. A expressão do CV é dada por:

$$CV = \frac{s}{\bar{x}} \times 100\%$$

E mede o grau de dispersão relativa. De modo geral considera-se que:

- $CV \leq 10\% \rightarrow$  Dispersão reduzida
- $CV \leq 30\% \rightarrow$  Dispersão moderada
- $CV > 30\% \rightarrow$  Dispersão elevada

Para melhor interpretar o significado deste coeficiente, consideremos o exemplo seguinte:

Uma mesma peça é fornecida por dois fornecedores, A e B. A peça destina-se à indústria automóvel e o seu diâmetro deve ser de 1,3 cm. Tanto o fornecedor A como o B garantem estas dimensões no diâmetro médio das peças e estas são vendidas por ambos ao mesmo preço. De modo a decidir qual o fornecedor a escolher, com base nas garantias de qualidade oferecidas, o comprador recolheu uma amostra de 6 peças junto de cada fornecedor, tendo medido o diâmetro de cada uma. Os resultados obtidos encontram-se na tabela seguinte.

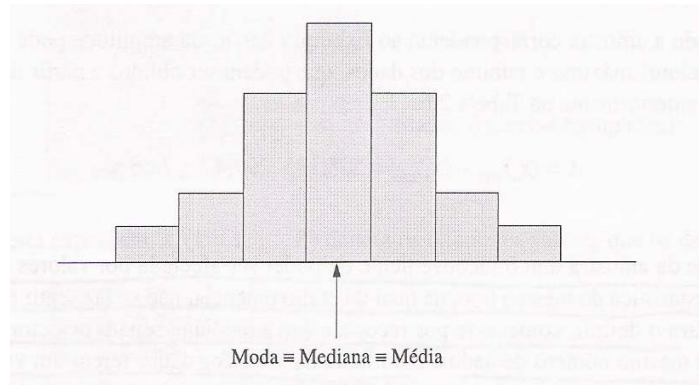
Fornecedor A Diâmetros (cm)	Fornecedor B Diâmetros (cm)
1,5	2,0
1,4	1,2
1,2	1,0
1,0	1,5
1,3	1,2
1,3	0,9

Pode-se verificar a partir da tabela que  $\bar{x}_A = \bar{x}_B = 1,3cm$ . Porém,  $s_A=0,14$  cm e  $s_B=18,5$  cm. Sendo assim  $CV_A=7\%$  e  $CV_B=18,5\%$  pelo que se pode concluir que as peças do fornecedor A terão diâmetros mais uniformes, optando-se então pelo fornecedor A.

## **2.5. Assimetria e Curtose. Diagramas de Caixa e Bigodes (“Box Plot”)**

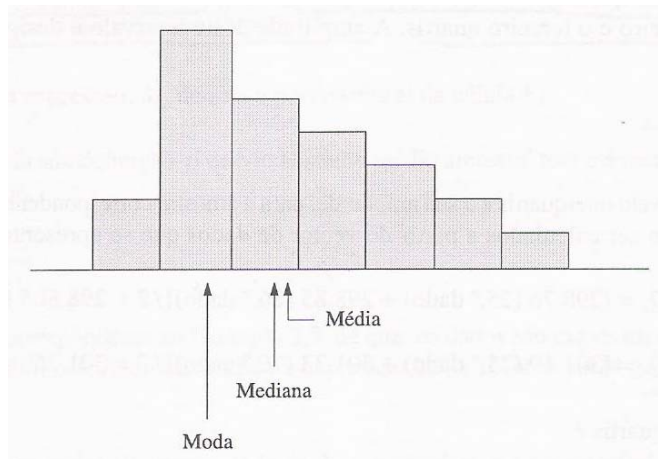
### **2.5.1. Assimetria**

Uma distribuição de frequências diz-se simétrica quando os valores da moda, de média e da mediana coincidem entre si. O histograma da Figura seguinte representa uma distribuição simétrica.



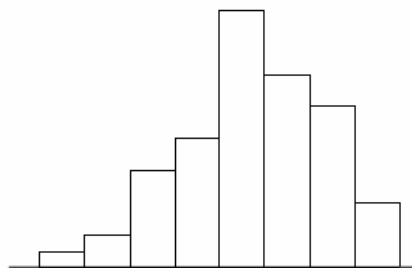
Histograma Simétrico

Quando o valor da moda é inferior ao da mediana que, por sua vez, possui um valor menor que a média, a distribuição diz-se assimétrica positiva ou assimétrica à direita. No histograma da figura seguinte representa-se uma distribuição assimétrica positiva.



Histograma Assimétrico Positivo

Quando o valor da moda é superior ao da mediana que, por sua vez, possui um valor superior ao da média, a distribuição diz-se assimétrica negativa ou assimétrica à esquerda. No histograma da figura seguinte representa-se uma distribuição assimétrica negativa



Histograma Assimétrico Negativo

A assimetria é fácil de determinar graficamente, podendo dizer-se se uma distribuição é simétrica ou assimétrica (positiva ou negativa) pelo aspecto do seu histograma. Quando não se dispõem de meios gráficos o grau de assimetria de uma distribuição pode ser medido utilizando um indicador: o coeficiente de assimetria.

Para dados não agrupados em classe o coeficiente de assimetria é dado por:

$$g_1 = \frac{N}{(N-1)(N-2)} \times \frac{\sum_{i=1}^N (x_i - \bar{x})^2}{s^2}$$

Dizendo-se que a distribuição é:

- Simétrica se  $g_1 = 0$
- Assimétrica à direita se  $g_1 > 0$
- Assimétrica à esquerda se  $g_1 < 0$

No caso de dados agrupados em classes recorreremos ao coeficiente de assimetria de Pearson, dado por:

$$G_1 = \frac{\bar{x} - \hat{x}}{s}$$

Dizendo-se que a distribuição é:

- Simétrica se  $G_1 = 0$
- Assimétrica à direita se  $G_1 > 0$
- Assimétrica à esquerda se  $G_1 < 0$

## 2.5.2. Curtose

As medidas de curtose dão uma indicação da intensidade das frequências na vizinhança dos valores de tendência central, por comparação com a distribuição Normal.

A distribuição Normal (ou de Gauss – em homenagem ao matemático alemão Carl F. Gauss), cujo estudo será efectuado mais a frente, é a que mais frequentemente se utiliza para descrever fenómenos que são traduzidos por variáveis aleatórias contínuas que resultam da soma de um grande número de efeitos provocados por causas independentes, em que o efeito de cada causa é desprezável em relação à soma de todos os outros efeitos. Esta distribuição pode ser caracterizada por uma função  $f(x)$  simétrica em torno de  $x = \mu$  ( $\mu$  é a média populacional), que tem um máximo em  $x = \mu$  e pontos de inflexão em  $x = \pm\sigma$ , sendo  $\sigma$  o desvio padrão populacional. Na Figura seguinte representa-se uma curva típica de Gauss.

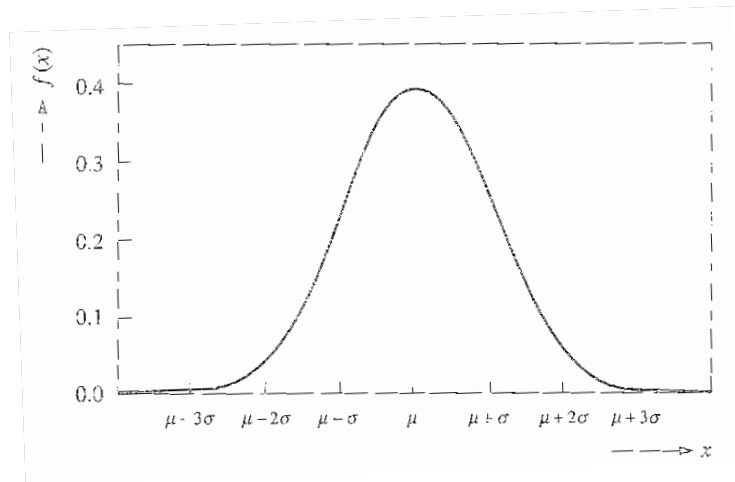


Gráfico da Distribuição Normal ou Curva de Gauss

O grau de curtose de uma qualquer distribuição fica assim definido pelo seu “achatamento” por comparação com a Normal.

Para variáveis não agrupadas pode-se medir a curtose através do coeficiente de curtose,  $g_2$ , dado por:

$$g_2 = \frac{(N+1)N}{(N-1)(N-2)(N-3)} \frac{\sum_{i=1}^N (x_i - \bar{x})^4}{s^4} - 3 \frac{(N-1)^2}{(N-2)(N-3)}$$

Dizendo-se que a distribuição é:

- Tão achatada quanto a Normal se  $g_2 = 0$
- Menos achatada do que a Normal se  $g_2 > 0$
- Mais achatada do que a Normal se  $g_2 < 0$

Para dados agrupados em classes o coeficiente de curtose,  $K$ , é dado por:

$$K = \frac{Q3 - Q1}{2(P90 - P10)}$$

Dizendo-se que a distribuição é:

- Tão achatada quanto a Normal se  $K = 0,263$
- Menos achatada do que a Normal se  $K < 0,263$
- Mais achatada do que a Normal se  $K > 0,263$

Exemplo: O gerente de um supermercado registou as chegadas dos clientes numa terça-feira entre as 15 h. e as 17 h., procedendo ao registo de chegadas em 100 períodos com a duração de um minuto seleccionados ao acaso, obtendo a tabela seguinte, a partir da qual se verifica que a distribuição é assimétrica positiva e menos achatada do que a normal.

$$\begin{aligned} g_1 &= 1,020408 \\ G_1 &= 0,699968 \end{aligned}$$

xi	fi	xi*fi	(xi-média)^2
0	1	0	14,3641
1	8	8	7,7841
2	19	38	3,2041
3	23	69	0,6241
4	17	68	0,0441
5	15	75	1,4641
6	8	48	4,8841
7	3	21	10,3041
8	3	24	17,7241
9	2	18	27,1441
10	1	10	38,5641
Totais	100	379	126,1051

$$g^2 = 16,40553$$

### 2.5.2. Diagramas de Caixa e Bigodes

Os diagramas deste tipo designam-se em inglês por “box-plot” ou “box and whisker plot”. Nestes diagramas representam-se graficamente o 1º e 3º quartis (que delimitam a caixa), representando a mediana no interior da caixa, e unindo por duas linhas à caixa tanto o valor mínimo da maior concentração de dados como o valor máximo da maior concentração de dados. Para melhor compreender este tipo de diagramas veja-se a Figura seguinte:

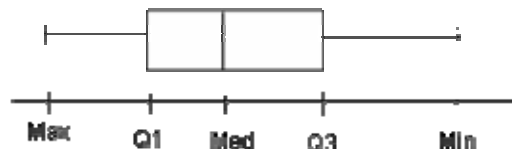


Diagrama de caixa e bigodes representando uma distribuição assimétrica à direita

Para representar diagramas deste tipo é necessário, numa primeira fase, identificar se existem valores aberrantes, que não pertencem ao intervalo no interior do qual se encontram contidos a maior parte dos dados, os denominados “outliers”. De modo a detectar a existência destes valores deveremos calcular 4 limites:

- $LL1 = Q1 - 3AIQ$
- $LL4 = Q3 + 3AIQ$
- $LL2 = Q1 - 1,5AIQ$
- $LL3 = Q3 + 1,5AIQ$

Os valores da variável menores que LL1 e maiores que LL4 são outliers severos. Os valores compreendidos entre LL1 e LL2 e entre LL3 e LL4 são outliers moderados.

Os outliers são representados no diagrama de caixa como pontos isolados, adoptando-se normalmente um asterisco (\*) para os severos e um ponto aberto (o) para os moderados.

Para melhor compreensão deste assunto, consideremos o exemplo seguinte:

Uma dada variável apresenta  $Q1=50$ ,  $Q3=60$  e  $\tilde{X}=55$ . Os valores mais pequenos medidos para a variável foram: 18, 25 e 40. Os maiores valores medidos foram: 70 e 85. Os restantes valores situavam-se 40 e 70.

Para construir o correspondente “box-plot” investiguemos a existência de outliers:

$$AIQ=60-50=10$$

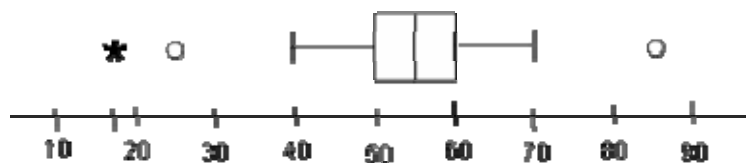
$$LL1=50-30=20$$

$$LL4=60+30=90$$

$$LL2=35$$

$$LL3=75$$

Concluimos que existem outliers severos ( $x=18$ ) e moderados ( $x=25$  e  $x=85$ ) devendo assinalá-los no “box-plot” pelos pontos que lhe correspondem. Obtem-se então o diagrama da Figura seguinte:



## 2.6. Tabelas de contingência

Existe um grande número de estudos estatísticos que não é dedicado apenas a analisar o valor de uma única variável mas de várias variáveis em simultâneo. Neste tipo de estudos, a cada elemento analisado corresponde um conjunto de valores, sendo os dados resultantes designados por multivariados (podemos para cada indivíduo analisar em simultâneo, por exemplo a idade, a altura, o peso, o estado civil, o distrito onde reside, etc.). No caso particular de serem medidas duas variáveis para cada um dos elementos que constituem a amostra obtém-se um conjunto de dados bivariado.

Nestes conjuntos de dados interessa desenvolver instrumentos que meçam o grau de associação ou a existência de alguma relação entre as variáveis.

Para dados qualitativos recorre-se à elaboração de tabelas de contingência e para dados quantitativos utilizam-se os diagramas de dispersão e os coeficientes de correlação.

As tabelas de contingência são semelhantes às tabelas de distribuição de frequências e permitem analisar a associação entre variáveis quantitativas, apresentando os dados numa forma sumária. Para as elaborar contam-se o número de observações que são comuns a cada par de categorias. Por exemplo, quando se pretende relacionar o género sexual com o estado civil de um conjunto de indivíduos é necessário contabilizar quantos são homens casados, homens divorciados, homens viúvos, homens solteiros, mulheres casadas, etc.

A tabela seguinte exemplifica uma tabela de contingência para as variáveis: x=idade e y=região do país preferida para férias.

Região preferida	Idade (anos)				Total
	]20, 30]	]30, 40]	]40, 50]	>50	
Norte	84	80	4	2	170
Sul	80	130	90	10	310
Centro	4	10	30	25	69
Total	168	220	124	37	544

Na tabela verifica-se que a faixa etária entre os 30 e os 40 anos prefere o Sul como destino de férias. Observa-se também que apenas 6 dos indivíduos com 40 ou mais anos preferem o Norte.

Como as faixas etárias não foram igualmente amostradas, a informação que se pode obter é ainda mais útil se forem obtidas as percentagens em linha ou em coluna. Para obter as percentagens em linhas dividem-se os valores das frequências em cada linha pelo resultado total da linha e multiplica-se por 100. Assim, temos:

Região preferida	Idade (anos)				Total
	]20, 30]	]30, 40]	]40, 50]	>50	
Norte	49,9%	47,0%	2,4%	1,2%	100%
Sul	25,8%	41,9%	29,0%	3,3%	100%
Centro	5,8%	14,5%	43,5%	36,2%	100%
Total	30,5%	40,1%	22,5%	6,9%	100%

Na tabela anterior podemos verificar que, dentre os indivíduos que preferem o Norte, 49,9 % tem idade compreendida entre 20 e 30 anos e 47,0 % têm idade entre 30 e 40 anos. Apenas 3,6 % das pessoas que preferem o Norte têm idade superior a 40 anos.

Como conclusão geral dos dados apresentados podemos dizer que existe uma tendência para as pessoas mais jovens preferirem o Norte como destino de férias enquanto que nas idades mais avançadas prefere-se o Centro.

## 2.7. Diagramas de dispersão e coeficientes de correlação

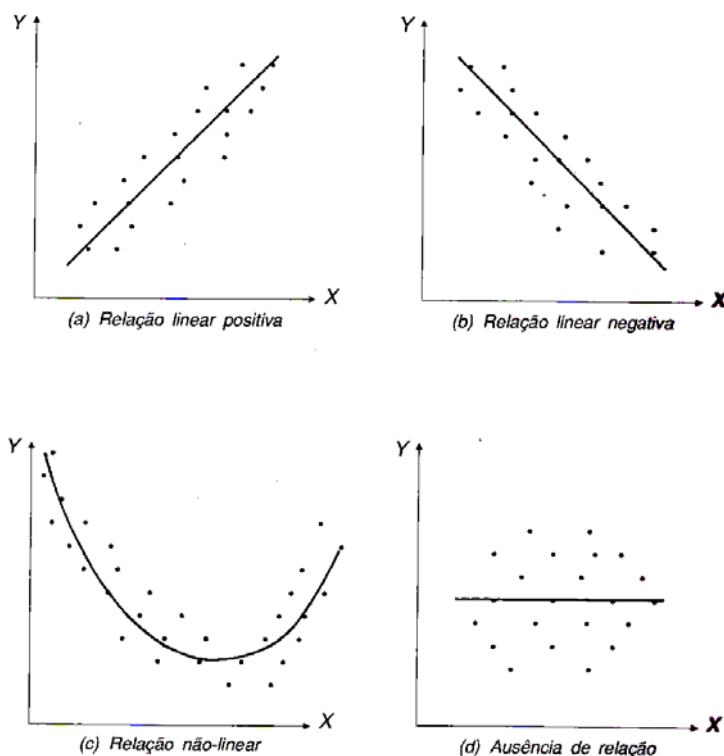
Os diagramas de dispersão e os coeficientes de correlação são utilizados para analisar possíveis associações entre variáveis quantitativas. Um pressuposto indispensável à existência dessas associações que permitirá, numa fase posterior, elaborar previsões, é o facto de se poder estabelecer uma relação do tipo causa-efeito entre as variáveis. Isto é, só é viável fazer previsões com base em relações estatísticas entre variáveis se a variação de uma delas puder ser atribuída à variação da outra.

Após se estabelecer que existe uma relação causal entre as variáveis, o passo seguinte consiste em determinar a forma ou o tipo de relação. Esta determinação pode ser feita mediante a construção de um diagrama de dispersão.

Um diagrama de dispersão consiste num gráfico constituído por pontos discretos onde cada ponto,  $P_i$ , representa um par de valores observados,  $(x_i, y_i)$ .  $x_i$  representa o valor da variável independente observada para o indivíduo  $P_i$  e  $y_i$  representa o valor da variável dependente observada para esse mesmo indivíduo.

O diagrama de dispersão tem uma função dupla: por um lado ajuda a destrinçar se existe alguma associação entre as variáveis, por outro permite identificar qual o modelo matemático (equação) mais apropriado para descrever essa associação.

Nos gráficos da Figura seguinte apresentam-se vários exemplos de diagramas de dispersão e as conclusões que deles se podem tirar acerca da relação entre as variáveis.



No caso de existir uma relação entre as variáveis esta pode ser de vários tipos: linear (casos (a) e (b) da figura), polinomial (caso (c) da figura – polinómio do 2º grau), exponencial, logarítmica, etc.

A relação mais simples é do tipo linear, sendo possível linearizar algumas das relações não lineares exemplificadas no parágrafo anterior.

Uma relação do tipo linear entre as variáveis pode ser descrita matematicamente pela equação:

$$y = b + mx + e$$

Esta equação constitui o modelo de regressão linear simples sendo:

*y*: variável explicada ou dependente

*x*: variável explicativa ou independente

*e*: variável residual que inclui outros factores explicativos de *y* não incluídos em *x* ou erros de medição

*b* e *m*: parâmetros da regressão. *b* é a intersecção da recta com o eixo vertical e *m* é o seu declive.

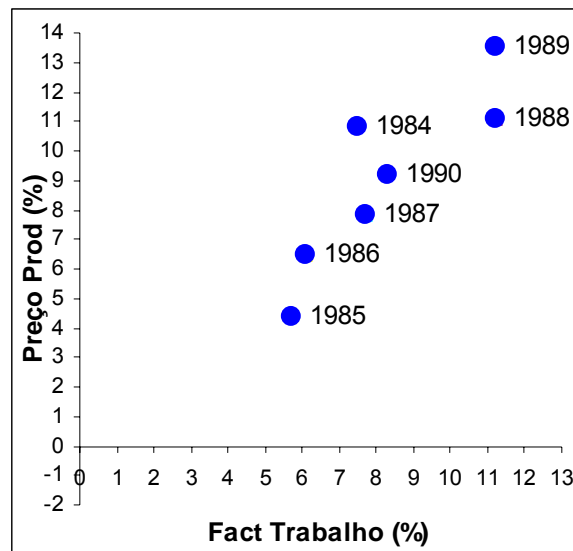
A equação anterior representa pois uma recta que, quando ajustada aos dados do diagrama de dispersão, se chama recta de regressão ou recta ajustada.

Exemplo: pretende-se estudar a relação entre os custos do factor trabalho (em percentagem) e o índice de preços no produtor (em percentagem) com o objectivo de fazer previsões acerca desta última variável a partir de valores conhecidos da primeira. Os valores conhecidos das variáveis para um período de 7 anos constam da tabela seguinte:

<b>Ano</b>	<b>Fact Trab (%)</b>	<b>Preço Prod (%)</b>
1984	7,5	10,8
1985	5,7	4,4
1986	6,1	6,5
1987	7,7	7,8
1988	11,2	11,1
1989	11,2	13,5
1990	8,3	9,2

Recorrendo aos dados dessa Tabela podemos construir o seguinte diagrama de dispersão que se ilustra na Figura seguinte e onde aparentemente existe uma relação linear positiva entre as variáveis, isto é, quando o custo do factor trabalho aumenta, o índice de preços também aumenta.

Podemos agora tentar ajustar uma recta de regressão aos dados, assumindo que existe entre eles uma relação de causa-efeito.



O modelo de regressão linear assume que existe a seguinte relação entre os dados:

$$y = b + mx + e$$

Com o seguinte significado:

- cada valor observado para a variável dependente (y) pode ser encontrado a partir de um efeito constante (b)
- de um efeito que resulta da variável independente (mx)
- de um efeito resultante de uma variável residual (e) que impede a existência de uma relação linear perfeita entre y e x.

Ao ajustar uma recta de regressão aos dados observados anulamos os efeitos da variável residual. A recta ajustada terá então a forma:

$$y_a = b + mx$$

A obtenção da recta ajustada implica o conhecimento dos parâmetros m e b de tal modo que o desvio entre os valores reais e os valores ajustados seja mínimo. Um método que permite minimizar estes desvios é o método dos mínimos quadrados. Neste método é minimizado o somatório dos quadrados das distâncias entre os valores observados e os valores ajustados.

Os valores observados são dados por  $y = b + mx + e$  e os valores ajustados são dados por  $y_a = b + mx$ . O que se pretende através do método dos mínimos quadrados é minimizar o somatório dos quadrados das diferenças,  $e_i = y_i - y_{ia}$ , para cada indivíduo, i, observado. Isto é:

$$y_i = b + mx + e_i \quad \text{e} \quad y_{ia} = b + mx \quad \text{então} \quad y_i - y_{ia} = e_i \quad \text{e} \quad e_i = y_i - b - mx$$

Pretendendo determinar-se m e b de tal modo que:

$$\text{Min}_{m,b} \sum e_i^2 = \text{Min}_{m,b} \sum (y_i - y_{ia})^2 = \text{Min}_{m,b} \sum (y_i - b - mx)^2$$

Para o último somatório, ou para qualquer outro polinómio quadrático, os pontos mínimos encontram-se quando as primeiras derivadas forem nulas e as segundas derivadas forem positivas (concavidade voltada para cima). Assim, a função de minimização estabelecida, pode ser resolvida através dos sistemas:

$$\begin{cases} \frac{\partial \sum e_i^2}{\partial b} = 0 \\ \frac{\partial \sum e_i^2}{\partial m} = 0 \end{cases} \quad \text{e} \quad \begin{cases} \frac{\partial^2 \sum e_i^2}{\partial b^2} > 0 \\ \frac{\partial^2 \sum e_i^2}{\partial m^2} > 0 \end{cases}$$

que conduzem às seguintes expressões para  $m$  (declive da recta ajustada) e para  $b$  (ordenada na origem da recta ajustada):

$$m = \frac{N \sum_{i=1}^N x_i y_i - \sum_{i=1}^N x_i \sum_{i=1}^N y_i}{N \sum_{i=1}^N x_i^2 - \left( \sum_{i=1}^N x_i \right)^2}$$

$$b = \bar{y} - m \bar{x}$$

Exemplo: Para a relação entre o custo do factor trabalho e o índice de preços no produtor dada anteriormente, iremos calcular a recta de regressão pelo método dos mínimos quadrados. Esse cálculo é efectuado através da tabela seguinte:

Ano	X Fact Trab (%)	Y Preço Prod (%)	X*Y	x^2
1984	7,5	10,8	81	56,25
1985	5,7	4,4	25,08	32,49
1986	6,1	6,5	39,65	37,21
1987	7,7	7,8	60,06	59,29
1988	11,2	11,1	124,32	125,44
1989	11,2	13,5	151,2	125,44
1990	8,3	9,2	76,36	68,89
N=7	57,7	63,3	557,67	505,01

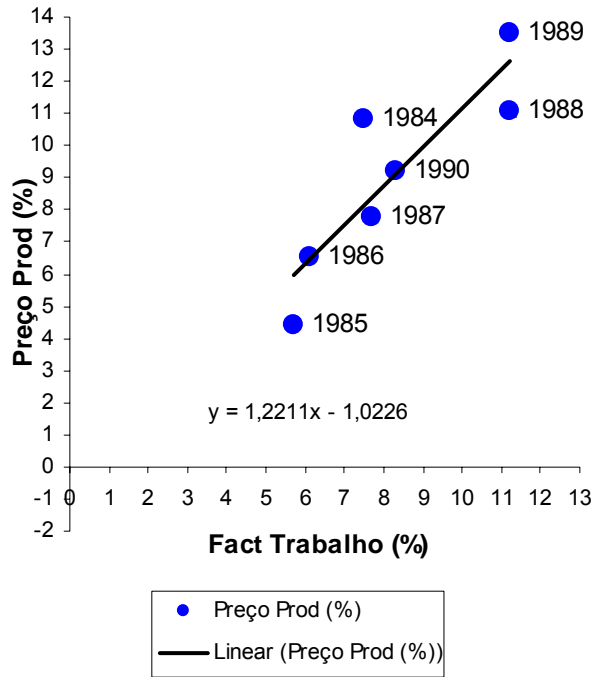
Dessa tabela concluímos que:

$$m = \frac{7 \times 557,67 - 57,7 \times 63,3}{7 \times 505,01 - 57,7^2} = 1,22 \quad \text{e} \quad b = \frac{63,3}{7} - 1,22 \times \frac{57,7}{7} = -1,022$$

Pelo que a recta ajustada tem a seguinte expressão:

$$y_a = 1,22x - 1,022$$

Podendo agora desenhar-se essa recta sobre o diagrama de dispersão, conforme se ilustra na seguinte Figura:



Na Figura anterior vê-se que existe uma relação linear positiva, embora imperfeita, entre as variáveis analisadas. Uma maneira de se analisar a possibilidade de existência de uma associação linear entre um par de variáveis é através do coeficiente de correlação linear.

O coeficiente de correlação linear,  $r$ , é um valor real compreendido entre -1 e 1 que pode ser calculado pela expressão seguinte:

$$r = \frac{\text{COV}(x, y)}{s_x s_y}$$

onde  $\text{COV}(x, y) = \frac{\sum [(x_i - \bar{x})(y_i - \bar{y})]}{N - 1}$  é a covariância entre as variáveis  $x$  e  $y$ ,  $s_x$  é o desvio-padrão de  $x$  e  $s_y$  é o desvio padrão de  $y$ .

No caso de não existir qualquer relação linear entre as variáveis a covariância será nula e quanto maior for o valor absoluto da covariância maior será o grau de associação linear entre as variáveis. Porém as unidades da covariância são difíceis de compreender. Se estivermos a relacionar precipitação (em mm) com pressão atmosférica (em bar) as unidades da covariância respectiva será mmxbar, o que não tem um significado compreensível. Opta-se então pelo coeficiente de correlação para medir o grau de ajuste linear entre um par de variáveis. Este coeficiente, obtido da divisão entre a covariância e o produto dos desvios padrões é adimensional e tem o seguinte significado:

$r=1$ : correlação linear perfeita positiva

$0,7 \leq r < 1$ : correlação linear forte positiva

$0,3 \leq r < 0,7$ : correlação linear moderada positiva

$0 < r \leq 0,3$ : correlação linear fraca positiva

$r = 0$ : não existe correlação linear (podendo ou não existir outro tipo de relação)

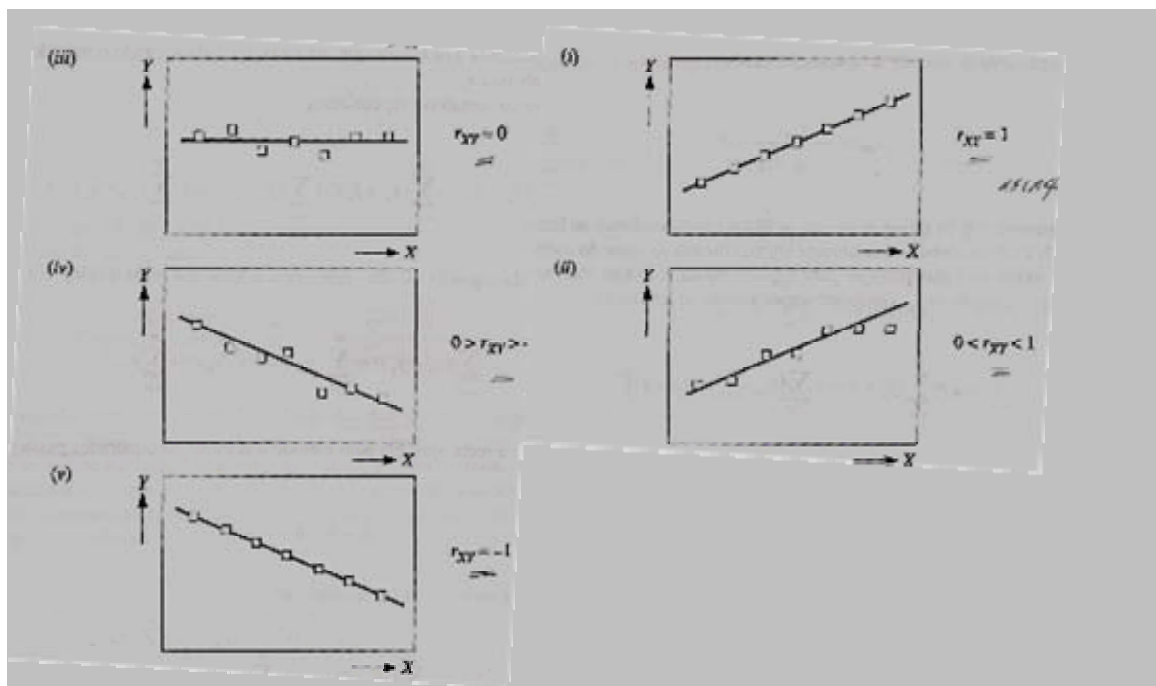
$r = -1$ : correlação linear perfeita negativa

$-1 < r \leq -0,7$ : correlação linear forte negativa

$-0,7 < r \leq -0,3$ : correlação linear moderada negativa

$-0,3 < r < 0$ : correlação linear fraca negativa

Nos gráficos da figura seguinte ilustra-se a relação entre o valor do coeficiente de correlação linear e o ajuste entre os dados observados e a recta de regressão.

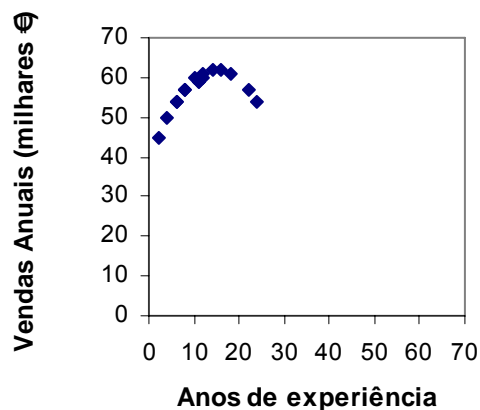


Não devemos tirar conclusões apressadas sobre o relacionamento entre duas variáveis apenas com base no cálculo do coeficiente de correlação linear. Por vezes este coeficiente sugere a existência de uma correlação linear entre os dados (positiva ou negativa) que é refutada pelo diagrama de dispersão. Para melhor compreender estes casos considere-se o seguinte exemplo.

Exemplo: Com o objectivo de estabelecer se existe uma relação linear entre o número de anos de trabalho e o volume de vendas anual para os vendedores de uma empresa analisou-se este par de variáveis na totalidade dos 16 trabalhadores da empresa. Os valores obtidos constam da tabela seguinte:

Anos de Experiência	Vendas anuais (milhares de €)
24	54
8	57
2	45
12	61
8	57
4	50
6	54
6	54
11	59
12	60
11	59
16	62
14	62
10	60
18	61
22	57

O valor obtido para o coeficiente de correlação linear deste conjunto de dados é 0,5, indicando uma correlação linear moderada positiva. Porém, ao construir o diagrama de dispersão, verifica-se que existe uma relação não linear (mas polinomial do 2º grau) entre as variáveis (ver a Figura seguinte).



Quando se procede ao ajuste de uma determinada recta de regressão aos dados observados, podemos ainda tirar conclusões acerca da qualidade do ajuste através do cálculo de outro coeficiente: o coeficiente de determinação.

O coeficiente de determinação, notado por  $R^2$ , mede a qualidade do ajuste entre a recta e os dados e o seu valor é um número real compreendido entre 0 e 1. Se  $R^2$  for 1 a qualidade do ajuste é perfeita (positiva ou negativa), não existindo relação linear se  $R^2 = 0$ .

O coeficiente de determinação representa ainda a proporção (ou percentagem) da variação da variável dependente (y) que é explicada pelas variações da variável independente (x), sendo o seu valor obtido através da seguinte expressão:

$$R^2 = \frac{m \sum [(x_i - \bar{x})(y_i - \bar{y})]}{\sum (y_i - \bar{y})^2}$$

No exemplo da variação dos índices de preço no consumidor com o custo do factor trabalho, o coeficiente de determinação é 0,8 (verifique), o que significa que 80% das variações dos índices de preço no consumidor são devidas às variações verificadas pelo custo do factor trabalho, sendo que os restantes 20% são devidos a outras causas.

Nesse exemplo foi obtida uma recta de regressão cuja equação é  $y_a = 1,22x - 1,022$ , o coeficiente de correlação linear indica uma correlação linear positiva forte (verifique), o coeficiente de determinação é próximo da unidade e no diagrama de dispersão vê-se que existe de facto uma relação linear entre as variáveis. Deste conjunto de observações concluímos que a equação da recta obtida constitui um modelo matemático (embora estatístico e não fenomenológico) que permite prever a evolução dos índices de preço com o custo do factor trabalho. Podemos, por exemplo, prever para o ano 2007, sabendo que o custo do factor trabalho nesse ano será 12 %, o índice de preço no consumidor como sendo

$$y_p = 1,22 \times 12 - 1,022 = 13,618 \%$$