1. INTRODUÇÃO

Inicialmente, a actividade estatística surgiu como um ramo da Matemática. Limitava-se ao estudo de medições e técnicas de contagem de fenómenos naturais e ao cálculo de probabilidades de acontecimentos que se podiam repetir indefinidamente. Actualmente, os métodos estatísticos são utilizados em muitos sectores de actividade, tendo como algumas aplicações estudos de fiabilidade, pesquisas de mercado, testes de controle de qualidade, tratamento de inquéritos, sondagens, modelos econométricos, previsões, etc.

Exemplo de uma estatística: os valores da inflação entre 1980 e 1990 constituem uma estatística. Fazer estatística sobre estes dados poderia consistir, por exemplo, em traçar gráficos, calcular a inflação média trimestral ou prever a inflação para 1991.

A análise de um problema estatístico desenvolve-se ao longo de várias fases distintas:

(i) Definição do Problema

Saber exactamente aquilo que se pretende pesquisar; estabelecer o objectivo de análise e definição da <u>população</u>

(ii) Amostragem e Recolha de Dados

Fase operacional. É o processo de selecção e registo sistemático de dados, com um objectivo determinado. Os dados podem ser primários (publicados pela própria pessoa ou organização) ou secundários (quando são publicados por outra organização).

(iii) Tratamento e Apresentação dos Dados

Resumo dos dados através da sua contagem e agrupamento. É a classificação de dados, recorrendo a tabelas ou gráficos.

(iv) Análise e Interpretação dos Dados

A última fase do trabalho estatístico é a mais importante e delicada. Está ligada essencialmente ao cálculo de medidas e coeficientes, cuja finalidade principal é descrever o comportamento do fenómeno em estudo (estatística descritiva). Na estatística indutiva a interpretação dos dados se fundamentam na teoria da probabilidade.

1.1. Definições Gerais

1.1.1. População

Fazer estatística pressupõe o estudo de um conjunto de objectos bem delimitado com alguma característica em comum sobre os quais observamos um certo número de atributos designados por variáveis.
Exemplo: Empresas existentes em Portugal

1.1.2. Variáveis ou atributos

As propriedades de uma população são estudadas observando um certo número de variáveis ou atributos. As variáveis podem ser de natureza qualitativa ou quantitativa. As variáveis quantitativas podem ainda dividir-se entre discretas e contínuas. As variáveis discretas assumem apenas um número finito numerável de valores. As variáveis contínuas podem assumir um número finito não numerável ou um número infinito de valores.
Exemplo: um conjunto de empresas pode ser analisado em termos de sector de actividade (atributo qualitativo), número de trabalhadores (atributo quantitativo discreto), rácio de autonomia financeira (atributo quantitativo contínuo), etc

1.1.3. Processo de amostragem

Para conhecer de forma completa a população, podem efectuar-se:

- recenseamentos (indagação completa de todos os elementos da população); este processo é, no entanto, tipicamente moroso e dispendioso, sendo esses os motivos porque os Censos são realizados apenas em cada 10 anos.
- estudos por amostragem (observação de apenas um subconjunto, tido como representativo do universo). As técnicas de recolha de amostras garantem a sua representatividade e aleatoriedade.

1.2. A Estatística Descritiva e a Estatística Indutiva

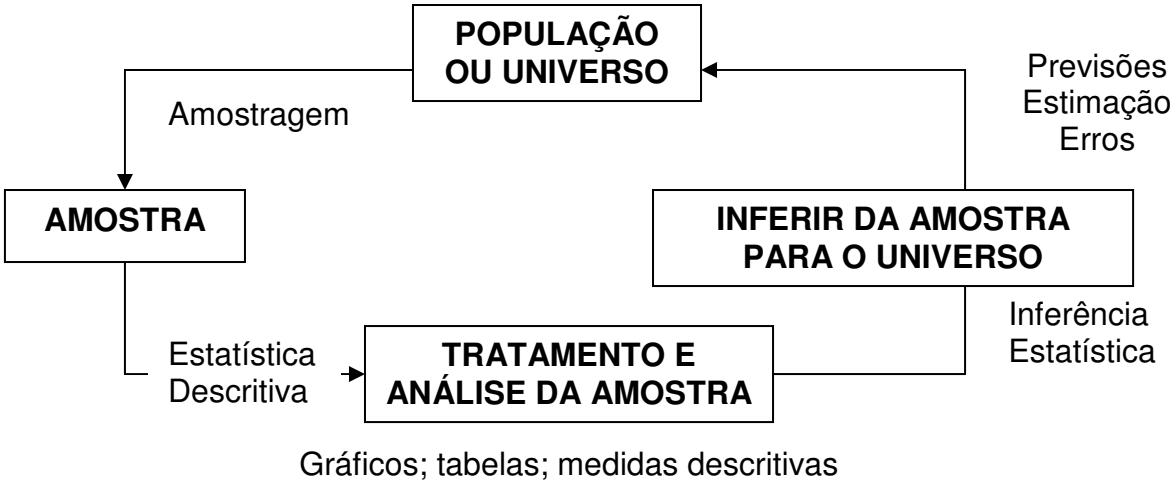
Para além do ramo de amostragem, a estatística compreende dois grandes ramos: a estatística descritiva e a estatística indutiva.

A estatística descritiva é o ramo da estatística que se encarrega do tratamento e análise de dados amostrais. Assim, depois de recolhida a amostra de acordo com técnicas que garantem a sua representatividade e aleatoriedade, fica disponível um conjunto de dados sobre o universo "em bruto" ou não classificados. Para que seja possível retirar qualquer tipo de conclusões, torna-se necessário classificar os dados, recorrendo a tabelas de frequências e a representações gráficas, isto é, é preciso tratar os dados. Depois de tratados, será possível proceder à análise dos dados através de várias medidas que descrevem o seu comportamento: localização, dispersão, simetria dos dados, concentração, etc. São disso exemplo indicadores numéricos bem conhecidos como a média ou a variância.

A estatística indutiva é o ramo da estatística que se ocupa em inferir das conclusões retiradas sobre a amostra para a população. De facto, a amostra não é mais do que um passo intermédio e exequível de obter informações sobre o verdadeiro objecto de estudo, que é o universo. A estatística indutiva (ou inferência estatística) garante a ligação entre amostra e universo: se algo se concluiu acerca da amostra, até que ponto é possível afirmar algo semelhante para o universo? É nesta fase que se procuram validar as hipóteses formuladas numa fase prévia exploratória. Claro que o processo de

indução implica um certo grau de incerteza associado à tentativa de generalização de conclusões da “parte” (amostra) para o “todo” (universo). O conceito de probabilidade vai ter aqui, então, um papel fundamental. Isto é, não vai ser possível afirmar com toda a certeza que o comportamento da amostra ilustra perfeitamente o comportamento do universo, mas apenas que o faz com forte probabilidade. As inferências indutivas são assim elaboradas medindo, ao mesmo tempo, o respectivo grau de incerteza. Daí que, na ficha das técnicas das sondagens eleitorais, por exemplo, apareçam referências ao “nível de confiança” associado aos resultados e ao “erro” cometido.

O esquema seguinte ilustra a “roda” da disciplina de estatística, relacionando os seus diferentes ramos:



2. ESTATÍSTICA DESCRITIVA

Os resultados da observação de um atributo sobre os elementos do conjunto a analisar constituem os dados estatísticos. O ramo da estatística que se ocupa do tratamento, apresentação e análise de dados amostrais denomina-se de estatística descritiva.

2.1. Variáveis Qualitativas

Os dados qualitativos são organizados na forma de uma tabela de frequências, que representa o número n_i de elementos de cada uma das categorias ou classes e que é chamado de frequência absoluta. A soma de todas as frequências é igual à dimensão da amostra (n).

Numa tabela de frequências, além das frequências absolutas, também se apresentam as frequências relativas (f_i), obtida dividindo a frequência absoluta pelo número total de observações.

Modalidades	Frequências absolutas	Frequências relativas
Mod. 1	n1	f1
Mod. j	nj	fj
Mod. n	nn	fn
Total	n: dimensão da amostra	1

$fi = \frac{ni}{n}$; n_i: nº de vezes que cada modalidade da variável foi observada.

Estes dados podem também ser representados graficamente através de:

Diagrama de barras

Para cada modalidade, desenha-se uma barra de altura igual à frequência absoluta ou relativa (as frequências relativas são de preferir, pois permitem a comparação de amostras de diferentes dimensões).

Diagrama sectorial ou circular

Esta representação é constituída por um círculo, em que se apresentam tantas "fatias" quantas as modalidades em estudo. O ângulo correspondente a cada modalidade é proporcional às frequências das classes, fazendo corresponder o total da amostra (n) a 360º Geralmente, juntamente com a identificação da modalidade, indica-se a frequência relativa respectiva.

2.2. Variáveis Quantitativas Discretas

São variáveis que assumem um número finito ou infinito numerável de valores. A apresentação destas amostras é semelhante às variáveis qualitativas, fazendo-se uma tabela de frequências e uma representação gráfica recorrendo ao diagrama de barras.

Valores da variável	Frequências absolutas	Frequências relativas
X1	n1	f1
Xj	nj	fj
Xn	nn	fn
Total	n: dimensão da amostra	1

Também é possível calcular as frequências (absolutas – Ni - e relativas - Fi) acumuladas, como se pode ver no exemplo:

Nº defeituosos (X)	Nº embalagens (ni)	% embalagens (fi)	Ni	Fi
0	80	40%	80	40%
1	60	30%	80+60	40%+30%
2	30	15%	170	85%
3	20	10%	190	95%
4	10	5%	200	100%
Total	200	1		

2.3. Variáveis Quantitativas Contínuas

Como foi dito anteriormente, uma variável (ou atributo) é contínua quando assume um número infinito não numerável de valores, isto é, podem assumir qualquer valor dentro de um intervalo.

Neste caso, a construção da tabela compreende duas etapas:

(i) Definição de classes de valores disjuntas, correspondentes a intervalos de números reais fechados à esquerda e abertos à direita, cuja constituição obedece a certas regras

(ii) Contagem das observações pertencentes a cada classe

<u>Regra de construção de classes</u>
(pressupõe a formação de classes de igual amplitude)

- Número de classes a constituir
  Depende de n = dimensão da amostra
  Se $n \geq 25$, o número de classes a constituir deve ser 5
  Se $n < 25$, o número de classes a constituir deve ser $\sqrt{n}$
- Amplitude comum a todas as classes
  Sendo a amplitude total dos dados dada pela diferença entre o valor máximo e o valor mínimo observados, então a amplitude de cada classe será:

$$\frac{\text{Valor máximo da variável observado} - \text{Valor mínimo da variável observado}}{\text{Nº de classes a constituir}}$$

Classes de valores da variável	Frequências absolutas	Frequências relativas
$[x_1; x_2[$	n1	f1
$[x_2; x_3[$		
$[x_3; x_4[$	nj	fj
$[x_{n-1}; x_n]$	n	fn
Total	n: dimensão da amostra	1

A distribuição de frequências representa-se através de um **histograma**.

Um histograma é uma sucessão de rectângulos adjacentes, em que a base é uma classe e a altura a frequência (relativa ou absoluta) por unidade de amplitude (n_i/a_i ou f_i/a_i), sendo a amplitude de cada classe $a_i=e_i-e_{i-1}$. A área total do histograma é a soma das frequências relativas, isto é, 1.

1. Esta distribuição permite visualizar o tipo de distribuição e deve salientar alguns aspectos mais relevantes desta (moda, classe modal, ...). Como as classes podem ter amplitudes diferentes, para que todos os rectângulos (colunas) sejam comparáveis é necessário corrigir as frequências das classes (calculando as frequências que se teria se a amplitude de todas as classes fosse igual e igual a 1)
2. É preferível representar o histograma com f_i/h_i do que com n_i/h_i uma vez que deste modo é possível comparar distribuições com diferente número de observações amostrais.

Também é possível calcular as frequências (absolutas – Ni - e relativas - Fi) acumuladas.

2.4. Medidas de localização

2.4.1. Média ($\overline{X}$)

É a medida de localização mais usada, sobretudo pela sua facilidade de cálculo.

Dados não-classificados (não agrupados numa tabela de frequências)

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$ Média aritmética simples

Dados classificados (isto é, agrupados numa tabela de frequências)
Variáveis discretas

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} n_i \, x_i = \sum_{i=1}^{n} f_i x_i$$ Média ponderada dos valores de X

Dados classificados (isto é, agrupados numa tabela de frequências)
Variáveis contínuas

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} n_i \, c_i = \sum_{i=1}^{n} f_i c_i$$ Média ponderada dos pontos médios das classes

onde c_i é o ponto médio de cada classe ($\frac{\lim.\inf. + \lim.\sup.}{2}$)

A média é uma medida de localização que, geralmente, indica o valor central da distribuição, entendido como o valor em torno do qual se distribuem os valores observados. Desta forma, a média é muitas vezes utilizada como valor representativo da amostra.

No entanto, a média tem o grande inconveniente de ser sensível a valores muito extremados ou aberrantes da distribuição (*outliers*). Em casos desses, a média deixa de ser um valor que aparece na parte central da distribuição para ser "empurrada" para os extremos. Nestes casos, é preferível recorrer à informação complementar fornecida por outras medidas de localização, como a moda e a mediana, que se definem a seguir.

2.4.2. Mediana (Me)

A mediana não se calcula a partir do valor de todas as observações, mas a partir da posição dessas observações.

Dados não-classificados

Se tivermos n valores x_1, x_2, ... x_n

Se n fôr ímpar,

$$Me = x_{\frac{n+1}{2}}$$

Se n fôr par,

$$Me = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

Dados classificados

A mediana é o valor tal que $F_i = 0,5$

Variáveis discretas

Se existe um valor de x_i para o qual $F_i = 0,5$, então fala-se em intervalo mediano.

Se não existe nenhum valor de x_i para o qual $F_i = 0,5$, então a mediana é o primeiro valor para o qual $F_i > 0,5$.

Variáveis contínuas

Em geral, determina-se o valor para o qual $F_i = 0,5$ através de uma regra de três simples, atendendo a que as frequências acumuladas variam uniformemente dentro de cada classe.

De uma forma geral:

$$Me = L\inf + \frac{0.5 - FL\inf}{FL\sup - FL\inf} xamp.classe\, mediana$$

2.4.3. Moda (Mo)

Variáveis discretas

A moda é valor de X para o qual fi é máximo, isto é, é o valor mais frequente da distribuição.

Variáveis contínuas

A classe modal é a classe de valores de X para o qual fi/hi é máximo, isto é, é a classe a que corresponde maior frequência por unidade de amplitude.

2.5. Medidas de ordem

Tal como se definiu para a mediana, é possível definir outros valores de posição ou valores separadores da distribuição em partes iguais.

Chama-se quantil de ordem p ao valor de x a que corresponde $F_i = p$.

- Se p=0,01; 0,02;.....0,99, chama-se ao quantil **percentil**
- Se p=0,1; 0,2;...0,9, chama-se ao quantil **decil**
- Se p=0,25, 0,5, 0,75, chama-se ao quantil **QUARTIL** (Q1, Q2 e Q3). A mediana é uma caso particular dos quartis (coincide com Q2)

Variável discreta

O quantil de ordem p é o primeiro valor de x para o qual $_i$>p.

Variável contínua

Calcula-se por uma regra de três simples, como a mediana.

De uma forma geral:

25% maiores

$$Q1 = L\inf + \frac{0.25 - FL\inf}{FL\sup - FL\inf} xamp.classe\, Q1$$

$$Q3 = L\inf + \frac{0.75 - FL\inf}{FL\sup - FL\inf} xamp.classe\, Q3$$

A representação gráfica destas medidas designa-se de **diagrama de extremos e quartis** e serve para realçar algumas características da amostra. Os valores da amostra compreendidos entre os 1º e 3º quartis são representados por um rectângulo (caixa) com a mediana indicada por uma barra. Seguidamente, consideram-se duas linhas que unem os meios dos lados do rectângulo com os extremos da amostra.

A partir deste diagrama, pode reconhecer-se a simetria ou enviesamento dos dados e a sua maior ou menor concentração:

2.6. Medidas de assimetria

A assimetria é tanto maior quanto mais afastados estiverem os valores da média, mediana e moda. Concretamente, se:

- $\overline{X}$ = Me = Mo, a distribuição diz-se simétrica
- $\overline{X}$ > Me > Mo, a distribuição diz-se assimétrica positiva (ou enviesada à esquerda)
- $\overline{X}$ < Me < Mo, a distribuição diz-se assimétrica negativa (ou enviesada à direita)

Coeficiente de assimetria de Bowley (g'): $\frac{(Q3 - Q2) - (Q2 - Q1)}{Q3 - Q1}$

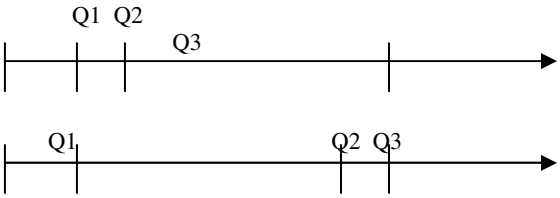
Se g' = 0 .....a distribuição é simétrica positiva ou equilibrada

Os quartis estão à mesma distância da mediana.

Se g' > 0 .....a distribuição é assimétrica positiva ou "puxada" para

a esquerda (se fôr = 1, assimetria é máxima)
A mediana desliza para o lado do Q1,
logo Q3-Q2 > Q2-Q1

Se g' < 0 .....a distribuição é assimétrica negativa ou "puxada" para a direita (se fôr = -1, assimetria é máxima)
A mediana desliza para o lado do Q3,
logo Q2-Q1 > Q3-Q2



Assimétrica positiva

Assimétrica negativa

2.7. Medidas de dispersão

Duas distribuições podem distinguir-se na medida em que os valores da variável se dispersam relativamente ao ponto de localização (média, mediana, moda). Apresentam-se de seguida algumas das mais utilizadas, classificadas consoante a medida de localização usada para referenciar a dispersão das observações:

2.7.1 Medidas de dispersão absoluta

(i) **Em relação à mediana**

Amplitude inter-quartis = Q = Q3 – Q1

Significa que 50% das observações se situam num intervalo de amplitude Q. Quanto maior (menor) a amplitude do intervalo, maior (menor) a dispersão em torno da mediana.

(ii) **Em relação à média**

Variância amostral: mede os desvios quadráticos de cada valor observado em relação à média, havendo pouca dispersão se os desvios forem globalmente pequenos, e havendo muita dispersão se os desvios forem globalmente grandes.

Dados não-classificados

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}\left(xi - \bar{x}\right)^2$$

Dados classificados

Variáveis discretas

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} ni\left(xi - \bar{x}\right)^2 = \sum_{i=1}^{n} fi\left(xi - \bar{x}\right)^2$$

Dados classificados

Variáveis contínuas

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} ni\left(ci - \bar{x}\right)^2 = \sum_{i=1}^{n} fi\left(ci - \bar{x}\right)^2$$

onde c_i é o ponto médio de cada classe i.

Desvio-padrão: Medida de dispersão com significado real, mas que só é possível calcular indirectamente, através da raiz quadrada da variância. Está expressa nas mesmas unidades da variável.

2.7.2 Medidas de dispersão relativa

Muitas vezes, avaliar a dispersão através de um indicador de dispersão absoluta não é conveniente, assim como comparara a dispersão de duas distribuições, uma vez que estas medidas vêm expressas na mesma unidade da variável – como é o caso, por exemplo, da variância. Assim, é de esperar que os valores da variância sejam mais elevados quando os valores da variável são maiores, o que não significa que a distribuição seja muito dispersa. Para comparar diferentes distribuições de frequência são precisas medidas de dispersão relativa:

$$Dispersão\ relativa = \frac{Dispersão\ absoluta}{Medida\ de\ localização\ em\ relação\ à\ qual\ está\ definida}$$

Coeficiente de variação

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

Outras medidas

$$\frac{Q3 - Q1}{Q2}$$

Estas medidas não estão expressas em nenhuma unidade, e permitem comparar dispersões entre duas amostras, pois não são sensíveis à escala (eventualmente diferente) em que as variáveis estejam expressas.

2.8. Análise da concentração

A noção de concentração apareceu associada ao estudo de desigualdades económicas, como a repartição do rendimento ou a distribuição de salários. O fenómeno de concentração está relacionado com a variabilidade ou dispersão dos valores observados, apesar de não poder ser analisado através das medidas de dispersão atrás descritas, que apenas medem a dispersão dos valores em relação a um ponto. O objectivo é determinar como o atributo (rendimento, salários, número de empresas) se distribui (se de forma mais ou menos uniforme) pelos diferentes indivíduos da amostra (que devem ser susceptíveis de serem adicionados, isto é, a análise de concentração não se aplica a idade, altura, peso, etc).

Se o atributo estiver igualmente repartido pelos indivíduos, temos uma situação extrema de igual distribuição; e vice-versa de o atributo estiver concentrado num só indivíduo, temos uma situação extrema de máxima concentração. Em geral, interessa medir o grau de concentração em situações intermédias.

Para analisar a concentração, existem dois instrumentos: a Curva de Lorenz e o Índice de Gini.

2.8.1 Curva de Lorenz

O objectivo é comparar a evolução das frequências acumuladas ($F_i = p_i$) com a evolução da soma dos valores da variável (q_i)

Quadro de dados

Classes de valores da variável	ni	Quantidade atributo	Freq.relativa acumuladas	Proporção atrib.acumul,
$[x_1; x_2[$	n1	yi	p1	q1
$[x_2; x_3[$				
$[x_3; x_4[$	nj	yj	pj	qj
$[x_{n-1}; x_n[$	nn	yn	pn=1	qn=1
Total	n			

Os pontos (pi;qi) pertencem ao quadrado (0,1) por (0,1). A curva que os une é a **curva de Lorenz**. Se houver igual distribuição, a frequência das observações deve ter uma evolução igual à proporção do atributo correspondente, isto é, pi=qi. Nesse caso, a curva de Lorenz coincide com a diagonal do quadrado, que é designada de **recta de igual repartição**. Quanto mais a curva se afastar da recta, maior é a concentração. A zona entre a diagonal e acurva de Lorenz designa-se, por isso, de **zona de concentração**.

2.8.2 Índice de Gini

O índice de Gini é calculado pela seguinte expressão

$$G = \frac{\sum_{i=1}^{n-1}(pi - qi)}{\sum_{i=1}^{n-1} pi}$$

Quando G = 0, a concentração é nula, havendo igual repartição. Caso o valor de G seja 1, a concentração será máxima. O valor de G varia entre 0 e 1, e quanto maior o seu valor, maior a concentração.