ESTATÍSTICA DESCRITIVA

O principal objectivo da ESTATÍSTICA DESCRITIVA é a redução de dados.

A importância de que se revestem os métodos que visam exprimir a informação relevante contida numa grande massa de dados através de um número muito menor de valores ou medidas características ou através de gráficos simples, é tal que a estatística descritiva se debruça a estudar os métodos que o permitam.

Pelo progresso da ciência que exige que se atenda mais profundamente à aquisição, qualidade e tratamento de dados, J. Tuckey introduziu um conjunto de técnicas estatísticas a que chamou "*Data Analisys*". Na análise de dados reconhecem-se duas componentes: uma mais próxima da estatística descritiva e outra da estatística indutiva.

A estatística descritiva – análise exploratória de dados – pretende isolar as estructuras e padrões mais relevantes e estáveis patenteados pelo conjunto de dados objectos do estudo.

A estatística indutiva – análise confirmatória de dados – pretende avaliar, nomeadamente, através da recolha e análise de novas observações, a reprodutividade ou permanência das estruturas e padrões detectados.

A análise de dados, fortalecida pela quantidade e variedade dos dados disponíveis e pelo poder de computação acessível, que não se pode dissociar das grandes capacidade gráficas, contribui para estabelecer uma maior ligação entre os aspectos descritivos e inferências da estatística.

Os dados estatísticos resultam de experiências ou inquéritos conduzidos sobre um conjunto restrito – a *amostra* – e as conclusões procuram alargar-se a um conjunto mais vasto – a *população*.

O principal objectivo da análise estatística consiste em determinar que generalizações sobre a população podem fazer-se a partir da amostra que da mesma foi recolhida. A designação de "amostra" é tomada correntemente num sentido mais amplo como sinónimo de dados ou observações enquanto a "população" é a totalidade; ou seja, o conjunto de todas as possíveis observações feitas em condições semelhantes.

O *processo de amostragem* é o processo seguido para escolher os elementos da população a incluir na amostra, condicionada, logicamente, a inferências ou conclusões permitidas pela amostra.
Diz-se que se trata de *amostragem casual* quando as *N* variáveis aleatórias observadas – X_i (i=1,2,3,..N) – são independentes e identicamente distribuídas.

Nos estudos por amostragem convém sempre distinguir entre população objectivo e população inquirida.
A *população objectivo* é a totalidade dos elementos que estão sobre estudo e em relação aos quais e se deseja obter certo tipo de informação. Muitas vezes não se consegue uma amostra da população objectivo, podendo-se efectuar a amostragem a partir de outra população relacionada de algum modo com a primeira e que constitui a *população inquirida*.

MEDIDAS DE LOCALIZAÇÃO

Média

A média amostral ou simplesmente média, que se representa por $\overline{x}$ é uma medida de localização do centro da amostra, e obtém-se a partir da seguinte expressão:

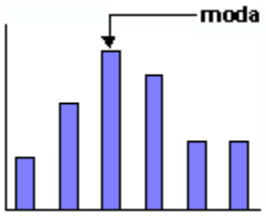
$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$$

onde x_1, x_2, ..., x_n representam os elementos da amostra e **n** a sua dimensão.

Moda

Para um conjunto de dados, define-se moda como sendo: ***o valor que surge com mais frequência*** *se os dados são discretos, ou,* ***o intervalo de classe com maior frequência*** *se os dados são contínuos.*

Assim, da representação gráfica dos dados, obtém-se imediatamente o valor que representa a moda ou a classe modal



Esta medida é especialmente útil para reduzir a informação de um conjunto de dados qualitativos, apresentados sob a forma de nomes ou categorias, para os quais não se pode calcular a média e por vezes a mediana (se não forem susceptíveis de ordenação).

Mediana

A mediana, **m**, é uma medida de localização do centro da distribuição dos dados, definida do seguinte modo: *Ordenados os elementos da amostra, a mediana é o valor (pertencente ou não à amostra) que a divide ao meio, isto é, 50% dos elementos da amostra são menores ou iguais à mediana e os outros 50% são maiores ou iguais à mediana*

Para a sua determinação utiliza-se a seguinte regra, depois de ordenada a amostra de n elementos:

Se n é ímpar, a mediana é o elemento médio.

Se n é par, a mediana é a semi-soma dos dois elementos médios.

Se se representarem os elementos da amostra ordenada com a seguinte notação:

$\mathbf{X_{1:n}}$, $\mathbf{X_{2:n}}$, ... , $\mathbf{X_{n:n}}$ então uma expressão para o cálculo da mediana será:

$$m = \begin{cases} x_{\frac{n+1}{2}:n} & \text{se n é ímpar} \\ \frac{1}{2}\left(x_{\frac{n}{2}:n} + x_{\frac{n}{2}+1:n}\right) & \text{se n é par} \end{cases}$$

Como medida de localização, a mediana é mais robusta do que a média, pois não é tão sensível aos dados!

Quantis

Quantis de ordem p

Generalizando a noção de mediana **m**, que como vimos anteriormente é a medida de localização, tal que 50% dos elementos da amostra são menores ou iguais a **m**, e os outros 50% são maiores ou iguais a **m**, temos a noção de quantil de ordem **p**, com 0<**p**<1, como sendo o valor $\mathbf{Q_p}$ tal que 100p% dos elementos da amostra são menores ou iguais a $\mathbf{Q_p}$ e os restantes 100 (1-p)% dos elementos da amostra são maiores ou iguais a $\mathbf{Q_p}$.

Tal como a mediana, é uma medida que se calcula a partir da amostra ordenada.

Quartis

Os quartis **inferior** e **superior**, Q1 e Q3, são definidos como os valores abaixo dos quais estão um quarto e três quartos, respectivamente, dos dados. Estes dois valores são frequentemente usados para resumir os dados juntamente com o mínimo e o máximo. Eles são obtidos ordenando os dados do menor para o maior, e então conta-se o número apropriado de observações: ou seja é e $\frac{n+1}{4}$ $\frac{3(n+1)}{4}$ para o quartil inferior e quartil superior, respectivamente.

A medidade de dispersão é a **amplitude inter-quartis**, IQR =Q3 Q1, i.e. é a diferença entre o quartil superior e o inferior.

MEDIDAS DE DISPERSÃO

Permitem medir a variabilidade presente num conjunto de dados.

Variância

Define-se a variância, e representa-se por $\mathbf{s^2}$, como sendo a medida que se obtém somando os quadrados dos desvios das observações da amostra, relativamente à sua média, e dividindo pelo número de observações da amostra menos um:

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{(n - 1)} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{(n - 1)}$$

Desvio Padrão

Uma vez que a variância envolve a soma de quadrados, a unidade em que se exprime não é a mesma que a dos dados. Assim, para obter uma medida da variabilidade ou dispersão com as mesmas unidades que os dados, tomamos a raiz quadrada da variância e **obtemos o desvio padrão**:

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

O desvio padrão é uma medida que **só pode assumir valores não negativos** e quanto maior for, maior será a dispersão dos dados.
Algumas propriedades do desvio padrão, que resultam imediatamente da definição, são:

- o desvio padrão é sempre não negativo e será tanto maior, quanta mais variabilidade houver entre os dados.
- se s = 0, então não existe variabilidade, isto é, os dados são todos iguais.

Amplitude

Uma medida de dispersão que se utiliza por vezes, é a **amplitude amostral r**, definida como sendo a diferença entre a maior e a menor das observações: $\mathbf{r = x_{n:n} - x_{1:n}}$

onde representamos por $x_{1:n}$ e $x_{n:n}$, respectivamente o menor e o maior valor da amostra (x_1, x_2, ..., x_n), de acordo com a notação introduzida anteriormente, para a amostra ordenada.

Amplitude Inter-Quartil

A medida anterior tem a grande desvantagem de ser muito sensível à existência, na amostra, de uma observação muito grande ou muito pequena. Assim, define-se uma outra medida, a **amplitude inter-quartil**, que é, em certa medida, uma solução de compromisso, pois não é afectada, de um modo geral, pela existência de um número pequeno de observações demasiado grandes ou demasiado pequenas. *Esta medida é definida como sendo a diferença entre os 1º e 3º quartis*

$$\textbf{Amplitude inter-quartil} = Q_{3/4} - Q_{1/4}$$

Do modo como se define a amplitude inter-quartil, concluímos que 50% dos elementos do meio da amostra, estão contidos num intervalo com aquela amplitude.

Esta medida é não negativa e será tanto maior quanto maior for a variabilidade nos dados.

Atenção: Mas, ao contrário do que acontece com o desvio padrão, uma amplitude inter-quartil nula, não significa necessariamente, que os dados não apresentem variabilidade.

Coeficiente de variação

O **coeficiente de variação** é uma medida de dispersão que se presta para a comparação de distribuições diferentes. O desvio-padrão, uma medida de dispersão, é relativo à média e como duas distribuições podem ter médias diferentes, o desvio dessas duas distribuições não é comparável. A solução é usar o coeficiente de variação, que é igual ao desvio-padrão dividido pela média:

$$cv = \frac{s}{\overline{x}} (\times 100\%)$$

DADOS, TABELAS E GRÁFICOS

Representação Gráfica de Dados

HISTOGRAMA

Um HISTOGRAMA é uma de duas representações gráficas de distribuição de frequências.
Consiste num conjunto de rectângulos que têm:

- a base num eixo horizontal (eixo dos xx) com centro no ponto médio e a largura igual à amplitude dos intervalos das classes.
- as áreas proporcionais às frequências das classes.

Se todos os intervalos tiverem a mesma amplitude, as alturas dos rectângulos serão proporcionais às frequências das classes; costuma-se tomar as alturas numéricas iguais a essas frequências.

Um POLÍGONO de frequência é um gráfico de linha em que as frequências são colocadas sobre perpendiculares, levantadas nos pontos médios. Pode-se também obtê-los ligando-se os pontos médios dos topos dos rectângulos de um histograma.

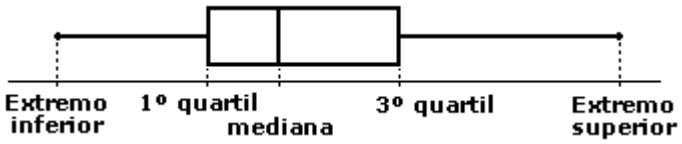
Diagrama de caule-e-folhas

É um tipo de representação que se pode considerar entre a tabela e o gráfico, uma vez que são apresentados os verdadeiros valores da amostra, mas numa apresentação sugestiva, que faz lembrar um histograma.

Consiste em escrever do lado esquerdo de uma linha vertical o dígito (ou dígitos) da classe de maior grandeza, seguidos dos restantes. A representação obtida terá o seguinte aspecto:

Diagrama de extremos e quartis e Caixa dos bigodes

É um tipo de representação gráfica, em que se realçam algumas características da amostra. O conjunto dos valores da amostra compreendidos entre o 1º e o 3º QUARTIS, que vamos representar por Q_1 e Q_3 é representado por um rectângulo (caixa) com a MEDIANA indicada por uma barra. A largura do rectângulo não dá qualquer informação, pelo que pode ser qualquer. Consideram-se seguidamente duas linhas que unem os meios dos lados do rectângulo com os extremos da amostra. Para obter esta representação, começa por se recolher da amostra, informação sobre 5 números, que são: os 2 extremos (mínimo e máximo), a mediana e o 1º e 3º quartis. A representação do diagrama de extremos e quartis tem o seguinte aspecto:



O extremo inferior é o mínimo da amostra, enquanto que o extremo superior é o máximo da amostra.

MEDIDAS DA FORMA DA DISTRIBUIÇÃO

Simetria

Uma distribuição diz-se simétrica se a média divide o histograma em duas metades iguais, uma constituindo uma imagem em espelho da outra.

Se a distribuição não é simétrica, diz-se assimétrica. O que quer dizer que umdos lados do gráfico da distribuição é mais alongado do que o outro. Adistribuição é assimétrica positiva se o alongamento tende a ocorrer no lado direito e é assimétrica negativa se o alongamento ocorrer predominantemente do lado esquerdo.

A curva normal, por exemplo, tem uma assimetria de 0. Se a assimetria é maior do que ±1, a forma da distribuição começa a afastar-se significativamente da curva normal.

Curtose

Esta é uma medida do grau de achatamento e afunilamento da curva que descreve a distribuição. O seu valor diz-nos se a curva tende a ser muito afunilada, com uma

elevada proporção dos dados aglomerados junto do centro, ou achatada, com os dados espalhando-se ao longo de uma grande amplitude.

A distribuição normal tem uma *curtose* igual a 0. Um valor positivo indica que os dados estão concentrados no centro e que a distribuição apresenta um forte pico/elevação nesse lugar (neste caso, diz-se que a distribuição é *leptocúrtica*). Um valor negativo indica que os dados estão dispersos e que a distribuição é mais achatada do que a curva normal (diz-se que a distribuição é *platicúrtica*). A curva normal diz-se *mesocúrtica*.

Valores de *curtose* superiores a ±1 indicam que a curva não é *mesocúrtica*.