Estatística Descritiva

EFT

índice

Início de tudo

A palavra Estatística procede do neolatim **statisticum collegium** (*conselho de Estado*) e do Italiano **statista** (*estadista ou político*), pois desde que se estabeleceram as sociedades organizadas, uma das funções principais dos governos era o de estabelecer registros da população, tais como nascimentos, mortes, impostos, colheitas, etc.
Alguns historiadores sinalam a **ilha de Sardenha**, em italiano: Sardegna, em sardo: Sardigna, e em catalão: Sardenya) como o local mais antigo de registros estatísticos.

Início de tudo

"Antes da chegada dos fenícios a partir do século X a.C., houve três civilizações autóctones que prosperaram na ilha: a de Bonuighinu, que surgiu no Quarto milénio a.C., a misteriosa população Shardana, e a mais célebre cultura nurágica, que se desenvolve a partir do século XVI a.C., senão antes, da qual os vestígios mais monumentais são os mais de 7000 nuragues (chamados localmente nuraghes no norte e nuraxis no sul), torres defensivas em forma de tronco cónico construídas com grandes blocos de pedra talhada e trabalhada, que se encontram espalhados por toda a ilha".
Existem monumentos da época dos Nuragues em cujas paredes estão gravados alguns sinais que eram a contagem do gado e da caça. (extraído do WIKIPEDIA)

Início de tudo

3000 anos a.C, os habitantes de babilonia usavam pequenas tábuas de argila para coletar dados sobre produção agrícola e os produtos trocados ou vendidos.
No Egito já eram analisados os dados da população, e a renda antes das pirâmides. A realização de censos eram comum assim como o registro de todos os movimentos populacionais (antes do ano 3050 a.C).
Na Biblia observamos num dos libros do Pentateuco, sob o nome de Números, o censo que realizou Moisés depois da saida de Egito. Na época do nascimento de Jesús, também é possível saber do censo realizado pelo imperador César Augusto.
Na China existiam os censos chineses ordenados pel emperador Tao perto do ano 2.200 a.C

Início de tudo

"No século XIX, a estadística entra numa nova fase com a generalização do método para estudar fenómenos das ciências naturais e sociais. Galton $(1.822 - 1.911)$ e Pearson $(1.857 - 1936)$ podem ser considerados como os pais da estatística moderna, pois a eles deve-se o passo da estatística dedutiva à estatística indutiva"
$(http://www.estadisticaparatodos.es/historia/histo_esta.html)$
Os fundamentos da estadtística atual e muitos dos métodos de inferência são devidos a R. A. Fisher.

Introdução

- uma população é um conjunto bem definido de objetos
- Quando se dispõe da informação de toda a população, teremos um censo.
- Um subconjunto qualquer da população é exemplo de uma amostra

Tipos de dados

- Dados univariados consistem de observações de uma variável particular
- Dados multivariados, para mais de duas variáveis.

Estatística Descritiva e Inferencial

- Estatística Descritiva: sumarização (resumo) e descrição de dados coletados
- Estatística Inferencial: processo de generalização a partir de amostras para as populações.

Relação entre E.D. e E.I

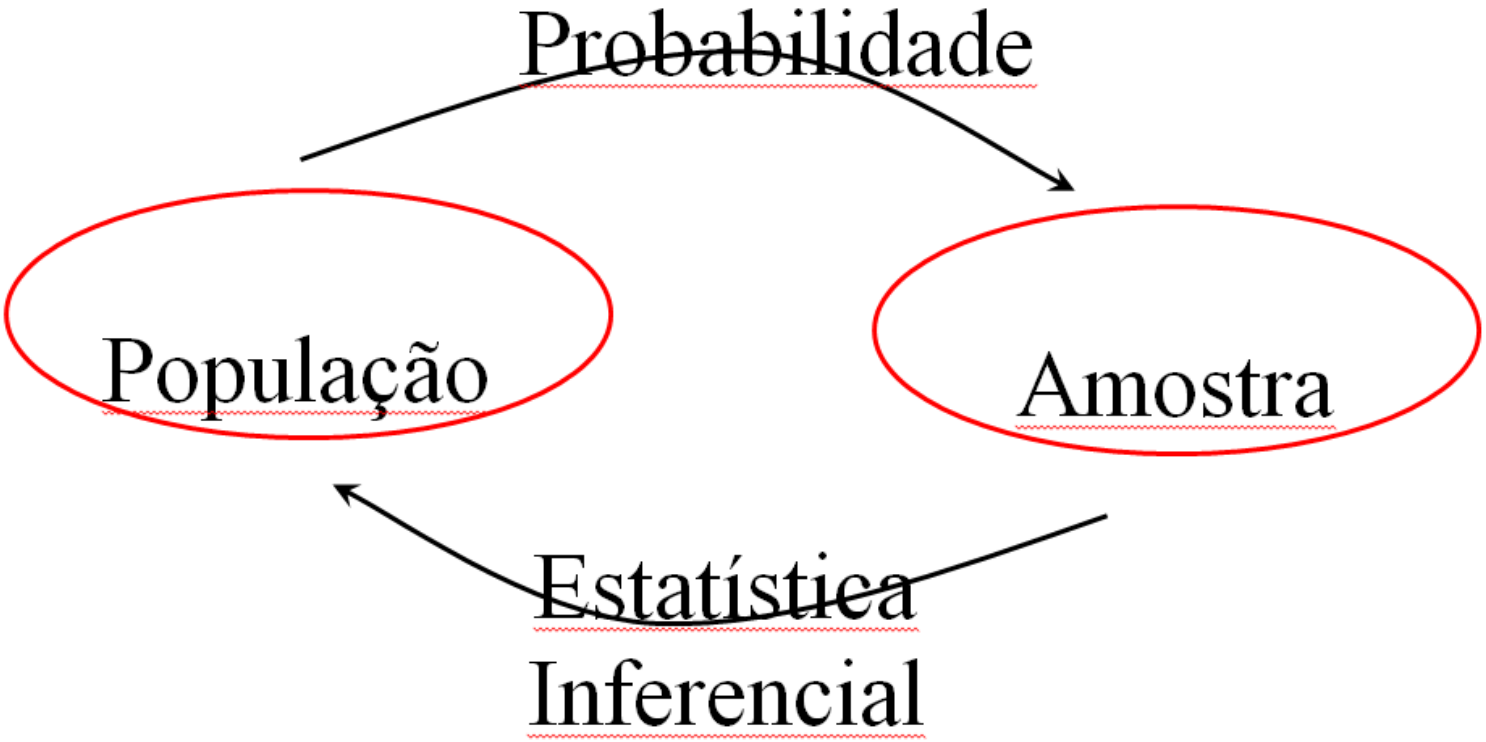

Figura: Relação entre E.D. e E.I.

Ramos e folhas

- selecione um ou mais valores iniciais para os ramos. Os dígitos sobrantes serão as folhas,
- disponha os ramos numa coluna vertical,
- anote a folha para cada observação,
- indique as unidades para os ramos e as folhas.

Ramos e folhas

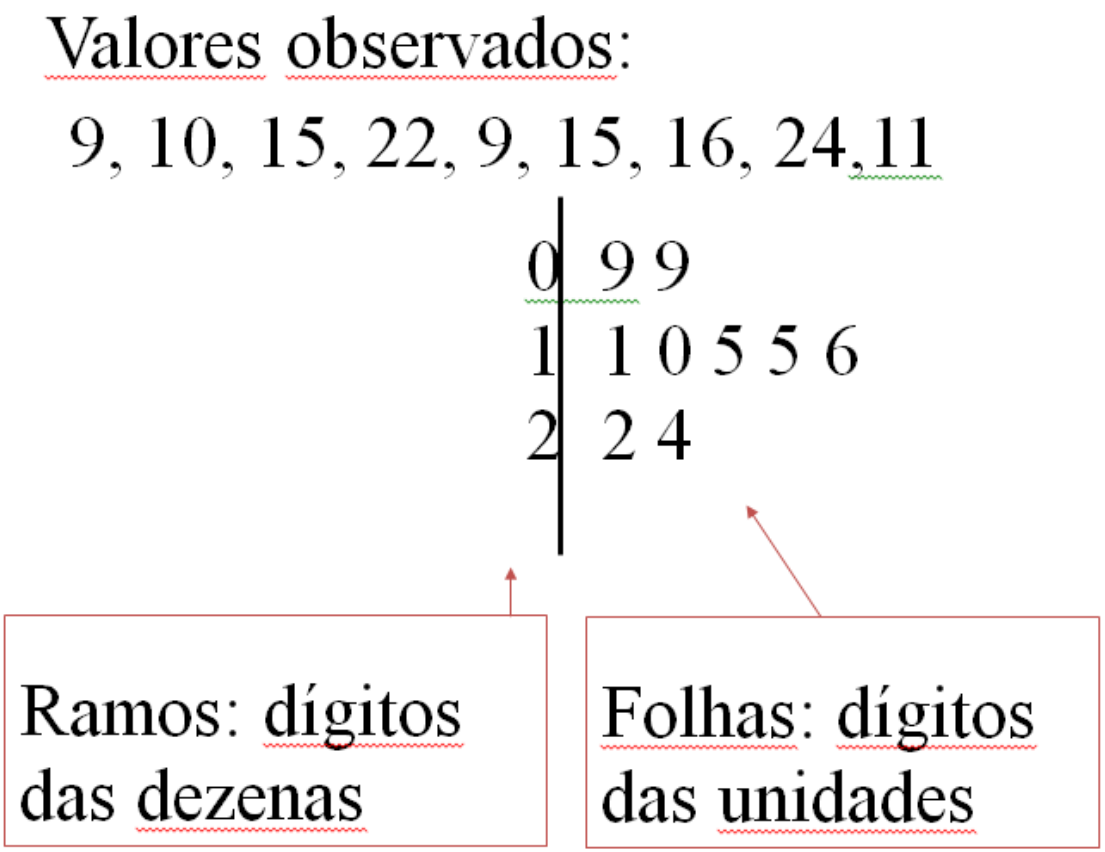

Figura: Exemplo de uma representação em Ramo e folhas

O que é mostrado num gráfico de ramo e folhas

- Identificação de valores típicos
- Como um valor é diseminado
- Presença de lacunas
- Observação da simetría
- Número e localização de picos
- Presença de outliers

Diagrama de pontos

- Representa os dados por meio de pontos
- O exemplo seguinte, conserva os mesmos valores do anterior: $9, 10, 15, 22, 9, 15, 16, 24, 11$

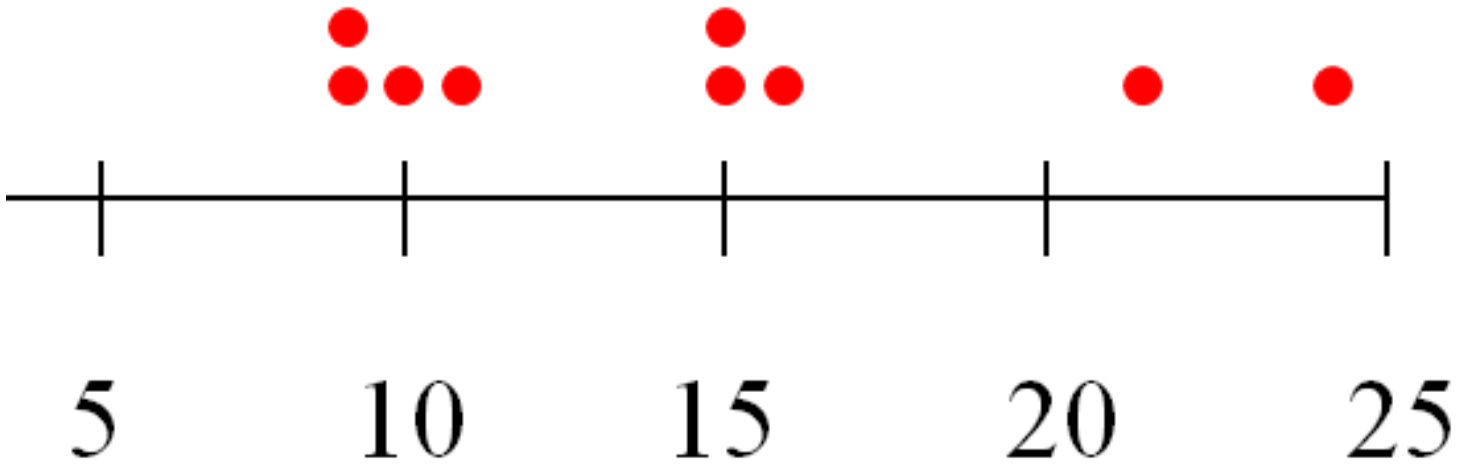

Figura: Exemplo de uma representação do diagrama de pontos

Variáveis discretas e contínuas

- Variável discreta: uma variável é discreta se o seu conjunto de valores possíveis é um conjunto finito ou uma sequência infinita.
- Variável contínua: uma variável e contínua se o seu conjunto de valores possíveis consiste de um intervalo inteiro numa linha numérica.

Histograma de dados discretos

- determine a freqüência e a freqüência relativa para cada valor de x
- marque os valores possíveis de x numa escala horizontal
- sobre cada valor, desenhe um retângulo cuja altura é a freqüência relativa desse valor.

Exemplo de dados discretos

Dados sobre o número de cartões de crédito que os alunos de uma pequena faculdade dizeram ter, onde x é a variável que indica o número de cartões dos estudantes.

x	Frequência	Frequência relativa
0	12	0,08
1	42	0,28
2	57	0,38
3	24	0,16
4	9	0,06
5	4	0,03
6	2	0,01

Histograma

x	Freq.relativa
0	0.08
1	0.28
2	0.38
3	0.16
4	0.06
5	0.03
6	0.01

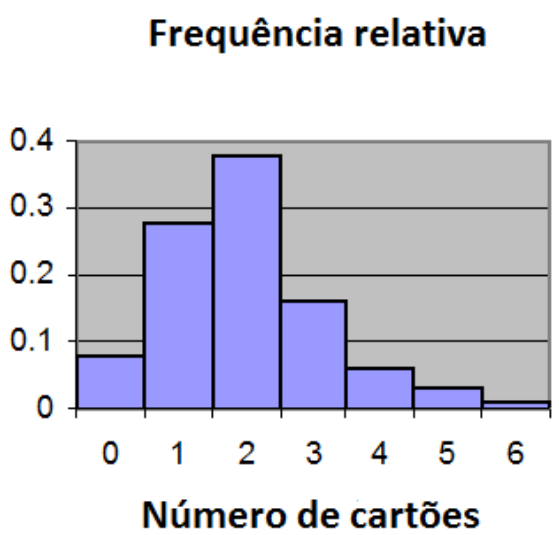

Figura: Histograma de frequências

Histograma de dados contínuos: classes de tamanhos iguais

- determine a freqüência e a freqüência relativa para cada classe
- marque os limites das classes numa escala horizontal
- sobre cada intervalo de classe, desenhe um retângulo cuja altura é a freqüência relativa dessa classe.

Histograma de dados contínuos: classes de tamanhos diferentes

- Após determinar as frequências e frequências relativas, calcule a altura de cada retângulo usando a seguinte fórmula:

$$\text{altura do retângulo} = \frac{\text{frequência relativa da classe}}{\text{amplitude da classe}}$$

- as alturas resultantes são chamadas de densidades
- a escala vertical é a escala de densidades

Formas comúns de histogramas

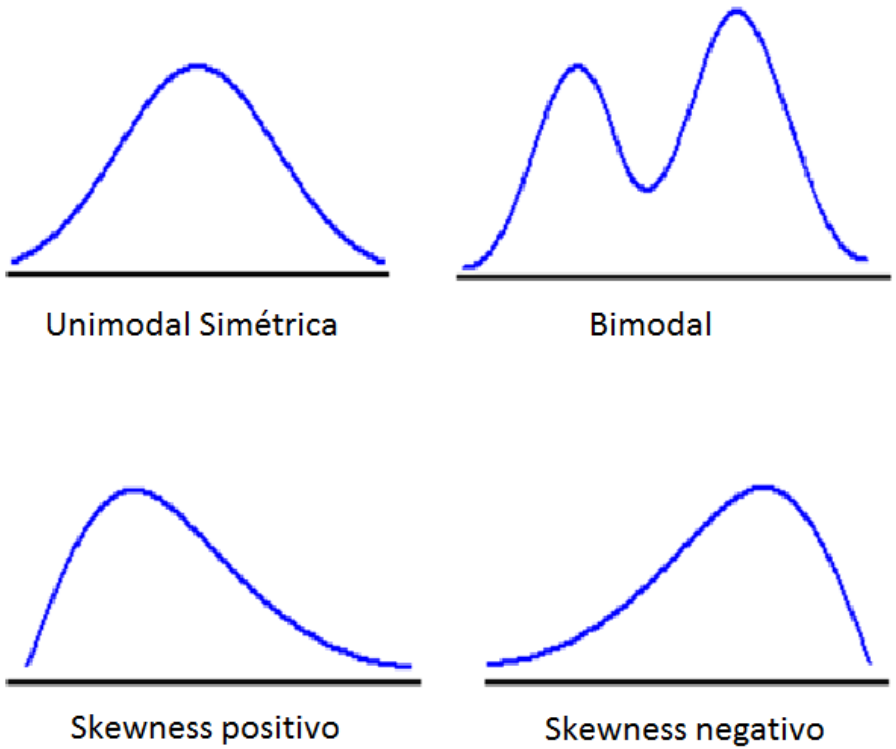

Figura: Representação das formas de histogramas

A média

A média de n números $x_1, x_2, ..., x_n$ é $\bar{x}$, onde

$$\bar{x} = \frac{\sum_i^n x_i}{n} \tag{1}$$

A média da população é denotada por μ

A mediana

- A mediana amostral ($med(x)$ ou $\tilde{x}$), é o valor central em um conjunto de dados ordenado de forma ascendente.
- Para um número par de dados, a médiana é a média dos dois pontos centrais

A mediana da população é denotada por $\tilde{\mu}$

Formas de distribuições populacionais

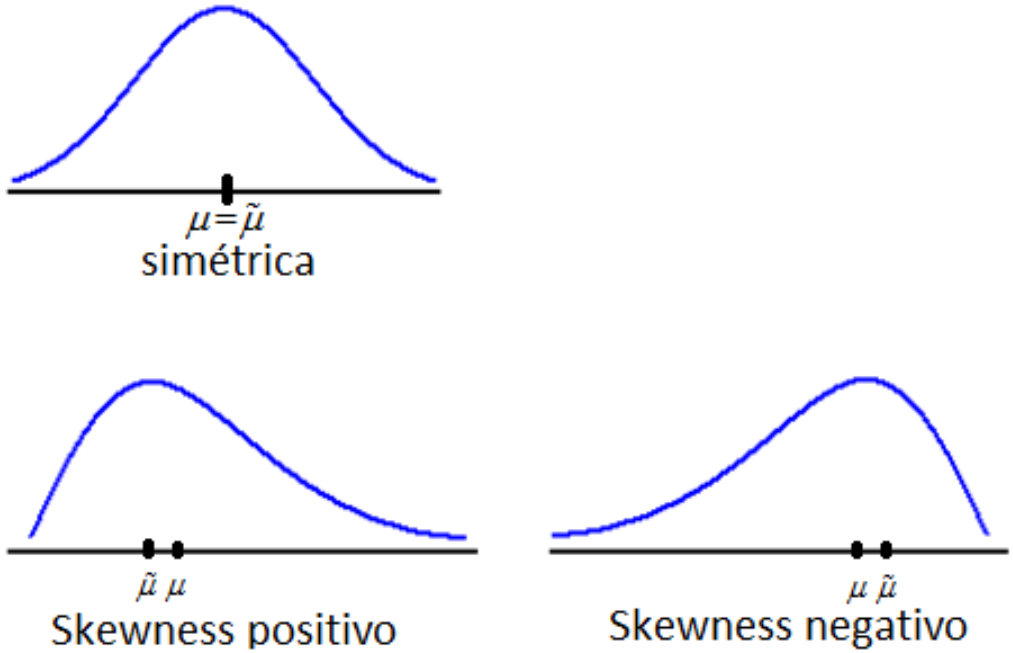

Figura: Três formas diferentes de distribuições populacionais

os Percentis

- Fornece a informação sobre como os dados estão dispersos no intervalo do menor ao maior valor.
- o p-ésimo percentil de um conjunto de dados é um valor tal que ao menos $p\%$ dos itens tomam este valor ou algum valor menor e ao menos $(100 - p)\%$ dos itens toma este valor ou algum valor maior.

Como calcular os percentis

- Ordenar os dados de forma crescente.
- calcule o indice i: a posição do p-ésimo percentil:

$$i = (\frac{p}{100}) \times n$$

- se i não é inteiro, arredonde-o. O p-ésimo percentil é o valor na posição i.
- se i é inteiro, o p-ésimo percentil é a média dos valores nas posições i e $i + 1$.

os quartis

Quartis são percentis específicos:

- O primeiro quartil (Q1) = 25 percentil
- O segundo quartil (Q2 ou Mediana) = 50 percentil
- O terceiro quartil (Q3) = 75 percentil

A variância amostral

- Variância é a medida de dispersão dos dados
- a variância amostral de n dados: $x_1, x_2, ..., x_n$ é dado por:

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

- consideramos s^2 como sendo baseado em $n-1$ graus de liberdade

o desvio padrão

- O desvio padrão é a medida de dispersão usando as mesmas unidades dos dados.
- o desvio padrão amostral de n dados é definido como a raíz quadrada da variância amostral:

$$s = \sqrt{s^2}$$

fórmulas para s^2

Uma expressão alternativa para o numerados de s^2 é:

$$s^2 = S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

propriedades

Seja $x_1, x_2, ..., x_n$ uma amostra quaisquer e seja c uma constante diferente de zero:

- Se $y_1 = x_1 + c$, $y_2 = x_2 + c$,..., $y_n = x_n + c$, então:

$$s_y^2 = s_x^2$$

- Se $y_1 = cx_1$, $y_2 = cx_2$,..., $y_n = cx_n$, então:

$$s_y^2 = c^2 s_x^2$$

Rank interquartil

- é a distância entre o terceiro e o primeiro quartil
- representa a dispersão dos 50% dados centrais
- é robusto a valores extremos

Box-plots

- Box-plot é um desenho de uma caixa com os extremos localizados nos quartis Q1 e Q3.
- Uma linha vertical é desenhada na posição da mediana ou Q2.
- limites são calculados usando a distância interquartil ($DIQ = Q3 - Q1$).
- O limite inferior é: $Q1 - 1,5 \times DIQ$
- O limite superior é: $Q3 + 1,5 \times DIQ$
- linhas tracejadas são desenhadas (dentro dos limites inferior e superior) entre o menor e o maior valor.
- Dados acima ou abaixo dos limites são considerados outliers ou valores discordantes.

desenho de box-plot

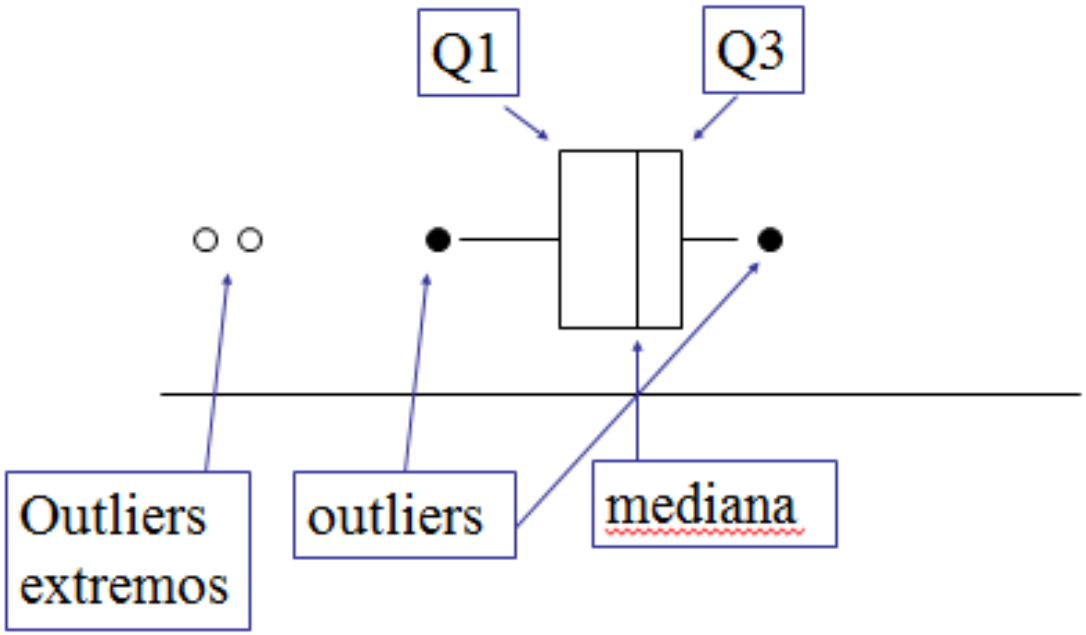

Figura: Desenho de um box-plot