

# Estatística aplicada

Lino Costa

Departamento de Produção e Sistemas  
Escola de Engenharia  
lac@dps.uminho.pt

Ano letivo 2015/2016

# Sumário

## População e amostra

### 1. Amostragem aleatória

- População e amostra
- Amostra representativa
- Planos de amostragem
- Aleatoriedade
- Tipos de dados e escalas de medida

# População e amostra

## População (ou Universo)

conjunto de indivíduos ou objetos que apresentam uma ou mais características em comum que se pretende analisar

## Amostra

subconjunto da população, que se observa com o objetivo de se obter informação sobre características desconhecidas da população de onde foi retirada

## Variável

qualquer característica (populacional) da unidade que constitui a população

# População e amostra

## Exemplo

Foi feito um inquérito a um grupo de 40 compradores de carros novos de uma certa marca para determinar quantas reparações ou substituições de peças foram feitas durante o primeiro ano de utilização dos carros.

- população: todos os compradores de carros novos da marca
- amostra: os 40 compradores de carros novos da marca
- variável: o número de reparações ou substituições de peças feitas durante o primeiro ano de utilização dos carros

# Parâmetro e estatística

## Parâmetro

características numéricas que descrevem a população, em geral, desconhecidas

## Estatística

característica numérica que descreve a amostra, calculada a partir dos valores observados na amostra

## Estimação

utiliza-se a estatística para estimar o parâmetro desconhecido da população

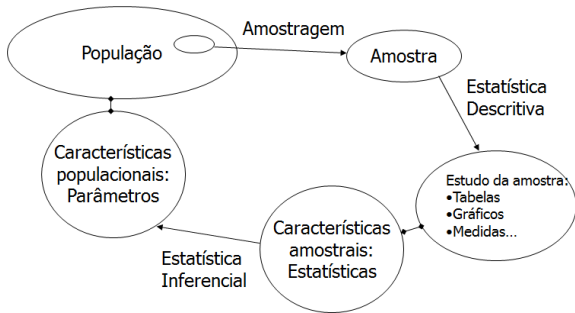
# Estatística descritiva e estatística inferencial

## Estatística descritiva

resumir a informação contida na amostra em tabelas, gráficos e estatísticas

## Estatística inferencial

tirar conclusões acerca da população com base na amostra, utilizando técnicas estatísticas convenientes



## Recenseamento ou censo

Estudo científico de um universo de pessoas, instituições ou objetos físicos com o propósito de adquirir conhecimentos, observando todos os seus elementos, e fazer juízos quantitativos acerca das características importantes desse universo.

### Exemplo

XV Recenseamento Geral da População Portuguesa (2011)  
(<http://www.ine.pt>)

# Sondagem

Estudo científico de uma parte de uma população com o objectivo de estudar atitudes, hábitos e preferências da população relativamente a acontecimentos, circunstâncias e assuntos de interesse comum.

## Fases de uma sondagem

- Amostragem
- Obtenção de informação
- Análise dos dados
- Relatório final

## Exemplo

- Sondagens para obter informação acerca da atitude dos eleitores
- Sondagens para testar as preferências dos consumidores



# Representatividade da amostra

A amostra deve ser tão representativa quanto possível da população que se está a estudar, evitando o enviesamento.

## Exemplo

Amostras enviesadas ou tendenciosas

- Amostragem por conveniência
  - utilizar uma amostra de sócios do FCP para estudar a qualidade da arbitragem do último Benfica-Porto
  - utilizar uma amostra de alunos de um curso para tirar conclusões acerca do aproveitamento dos alunos universitários
- Amostragem por resposta voluntária
  - “sondagens” realizadas pelas estações de televisão utilizando respostas voluntárias, por exemplo, por SMS

# Amostra aleatória simples

Dada uma população, uma amostra aleatória simples de dimensão  $n$  é um conjunto de  $n$  unidades da população, tal que qualquer outro conjunto de  $n$  unidades teria igual probabilidade de ser selecionado.

## Características a.a.s.

- minimiza o enviesamento
- tem em conta o princípio da aleatoriedade
- corresponde a uma recolha sem reposição
- os  $n$  elementos da amostra são independentes entre si
- todos os  $n$  elementos da amostra têm igual distribuição de probabilidade

## Amostra aleatória simples

### Exemplo

Discuta o seguinte procedimento de amostragem: "A opinião pública acerca da contracepção é estudada telefonando a lares selecionados aleatoriamente durante os dias da semana entre as 8h e as 17h."

- números de telefone selecionados aleatoriamente (amostra aleatória simples)
- a amostra pode ser tendenciosa porque a população ativa não está em casa durante os dias da semana entre as 8h e as 17h
- os resultados do estudo podem não refletir a verdadeira opinião da população

## Amostra aleatória estratificada

As unidades, donde se vai retirar a amostra, são divididas por grupos ou estratos, escolhidos pelo interesse especial que esses grupos representam ou pela semelhança das unidades dentro de cada grupo. De seguida, retira-se uma a.a.s. de cada grupo e combinam-se todas as amostras para construir a amostra aleatória estratificada.

### Ficha Técnica

Estudo de opinião efectuado pela Eurosondagem, S.A. para o Expresso, SIC e Rádio Renascença, de 20 a 25 de Outubro de 2011. O universo é a população com 18 anos ou mais, residente em Portugal Continental e habitando em lares com telefone da rede fixa. A amostra foi estratificada por região (Norte – 19,9%; A.M. do Porto – 14,3%; Centro – 29,9%; A.M. de Lisboa – 26,1%; Sul – 9,8%), e aleatória no que concerne ao sexo e faixa etária, donde resultou feminino (51,1%), masculino (48,9%) e 18/30 anos (17,4%), 31/59 anos (47,4%) e 60 anos ou mais (35,2%), num total de 1.032 entrevistas telefónicas validadas, que correspondem a uma taxa de resposta de 77,2%. O objecto da sondagem foi a intenção de voto para eleições legislativas, a actuação de órgãos de soberania e líderes partidários, e questões de âmbito político e social da actualidade. O resultado projectado da intenção de voto é calculado mediante um exercício meramente matemático, presumindo que os 21,5% respondentes “Ns/Nr” se absterem. O erro máximo da amostra é de 3,05%, para um grau de probabilidade de 95,0%.

# Dados estatísticos

## Dados

- **qualitativos** - informação que identifica alguma qualidade, categoria ou característica, não suscetível de medida, mas de classificação, assumindo várias modalidades
- **quantitativos** - informação resultante de características suscetíveis de serem medidas, apresentando-se com diferentes intensidades
  - **natureza discreta** - pode tomar um número finito (ou infinito numerável) de valores distintos
  - **natureza contínua** - pode tomar todos os valores numéricos, compreendidos no seu intervalo de variação

# Dados estatísticos

## Dados

- **qualitativos** - cor dos olhos, desporto preferido, sexo..
- **quantitativos**
  - **natureza discreta** - número de acidentes, resultado do lançamento de um dado, número de irmãos...
  - **natureza contínua** - peso, altura, nível de colesterol no sangue, idade...

# Escalas de medição

A medição de uma propriedade significa atribuir uma quantidade numérica para a representar.

## Tipos de escala

- **nominal** - a medição apenas define a que classe a unidade pertence, em relação àquela propriedade
- **ordinal** - a medição também esclarece quando uma unidade tem mais da propriedade do que outra unidade
- **intervalar** - a medição também diz que uma unidade é diferente por uma certa quantidade da propriedade, de outra unidade
- **proporcional** - a medição diz que uma unidade tem tantas vezes mais da propriedade do que outra unidade

# Escalas de medição

## Tipos de escala

- **nominal** - raça, situação profissional, sexo... Pode-se atribuir um código a cada categoria (por exemplo para o sexo: feminino-0, masculino-1 ou feminino-1, masculino-0; o valor em si não é importante).
- **ordinal** - opinião (má, aceitável, boa), pontuação de um júri (1 - mais fraca a 10 - mais forte)... A ordem dos números atribuídos às categorias tem significado, mas o valor dos números não tem significado (por exemplo, a pontuação 8 não é duas vezes melhor do que a 4).
- **intervalar** - temperatura ( $^{\circ}C$ )... Uma temperatura de  $40^{\circ}C$  não é duas vezes mais quente do que uma de  $20^{\circ}C$ .
- **proporcional** - o tempo de reação ( $s$ ), o comprimento ( $cm$ )... Um comprimento de  $4cm$  é duas vezes maior do que um de  $2cm$  e o 0 tem o seu significado.



# Sumário

## Estatística descritiva

1. Métodos numéricos ou quantitativos
  - medidas de localização
  - medidas de dispersão
2. Dados agrupados em classes
3. Métodos gráficos
  - gráficos de barra
  - gráficos circulares
  - histograma de frequências
  - gráfico de extremos e quartis

# Estatística descritiva

Amostra de  $n$  observações:  $x_1, x_2, \dots, x_n$

## Medidas de localização

- **média**

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

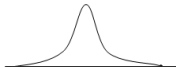
- **mediana** - valor que divide as observações ordenadas em duas metades iguais

$$Med = \begin{cases} \frac{x(\frac{n}{2}) + x(\frac{n}{2}+1)}{2} & \text{se } n \text{ par} \\ x(\frac{n+1}{2}) & \text{se } n \text{ ímpar} \end{cases}$$

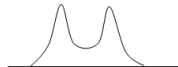
- **moda** - valor mais frequente



moda inexistente



uma moda (unimodal)



múltiplas modas (multimodal)

# Estatística descritiva

Amostra de  $n$  observações:  $x_1, x_2, \dots, x_n$

## Medidas de dispersão

- **variância** - indica a dispersão dos dados em relação à média

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

- **desvio padrão**

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

- **amplitude**

$$R = \max_{i=1, \dots, n} x_i - \min_{i=1, \dots, n} x_i$$

# Estatística descritiva

Amostra de  $n$  observações:  $x_1, x_2, \dots, x_n$

## Medidas de dispersão

- **quartis** ( $Q_1, Q_2, Q_3$ ) - dividem os dados ordenados por ordem crescente em quatro partes iguais

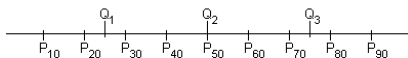
$$Q_i = x_{((n+1) \times \frac{i}{4})}$$

- **amplitude inter-quartil** ( $AIQ$ )

$$AIQ = Q_3 - Q_1$$

- **percentis** ( $P_i$  com  $0 < i < 100$ ) - indica o valor abaixo do qual estão  $i\%$  dos dados

$$P_i = x_{((n+1) \times \frac{i}{100})}$$



# Estatística descritiva

Dados discretos classificados de acordo com  $k$  valores da variável  $x_j$ , frequência absoluta  $f_j$ , frequência relativa  $fr_j$  com  $j = 1, \dots, k$ , sendo  $n = \sum_{j=1}^k f_j$

## Medidas de localização

- **média**

$$\bar{x} = \sum_{j=1}^k fr_j x_j = \frac{1}{n} \sum_{j=1}^k f_j x_j$$

## Medidas de dispersão

- **variância**

$$s^2 = \frac{n}{n-1} \sum_{j=1}^k fr_j (x_j - \bar{x})^2 = \frac{1}{n-1} \sum_{j=1}^k f_j (x_j - \bar{x})^2$$

# Estatística descritiva

Dados contínuos agrupados em  $k$  classes com valor médio  $M_j$ , frequência absoluta  $f_j$ , frequência relativa  $fr_j$  com  $j = 1, \dots, k$ , sendo  $n = \sum_{j=1}^k f_j$

## Medidas de localização

- **média**

$$\bar{x} \approx \sum_{j=1}^k fr_j M_j = \frac{1}{n} \sum_{j=1}^k f_j M_j$$

## Medidas de dispersão

- **variância**

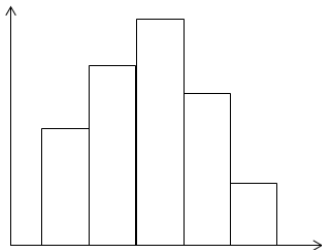
$$s^2 \approx \frac{n}{n-1} \sum_{j=1}^k fr_j (M_j - \bar{x})^2 = \frac{1}{n-1} \sum_{j=1}^k f_j (M_j - \bar{x})^2$$

# Estatística descritiva

## Organização dos dados

- Tabelas de frequências (absolutas e relativas)
- Representações gráficas
  - gráfico de barras (dados qualitativos ou quantitativos discretos)
  - gráfico circulares (dados qualitativos ou quantitativos discretos)
  - diagrama de caule-e-folhas (dados quantitativos discretos ou contínuos)
  - histograma (dados quantitativos contínuos)
  - gráficos de extremos e quartis ou “Boxplot” (dados quantitativos discretos ou contínuos)

# Construção de um histograma

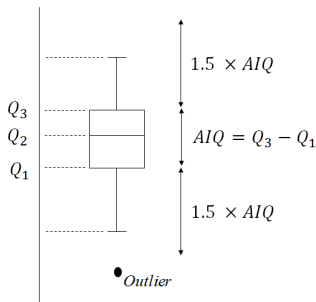


Histograma é uma representação gráfica dos dados formado por uma sucessão de rectângulos adjacentes, tendo cada um por base um intervalo de classe e por área a frequência relativa (ou a frequência absoluta). Deste modo a área total será igual a 1 (ou igual a  $n$ , a dimensão da amostra).

- calcular o número de classes ( $k$ ):  $k = 1 + 3.3 \log(n)$  (regra de Sturges)
- calcular a amplitude ( $R$ ):  $R = \max x_i - \min x_i$  para  $i = 1, \dots, n$
- calcular o intervalo de classe ( $w$ ):  $w > \frac{R}{k}$
- calcular o limite inferior da primeira classe ( $l$ ) e o limite superior da última classe ( $u$ ):  $l = \min x_i - \frac{kw - R}{2}$  para  $i = 1, \dots, n$  e  $u = l + kw$
- construir tabela de frequências



# Construção de um “Boxplot”

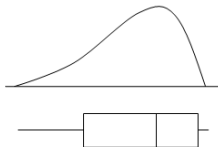


“Boxplot” é uma representação gráfica dos dados que representa, simultaneamente, localização central, dispersão, simetria e “outliers”.

- Caixa: primeiro quartil ( $Q_1$ ), segundo quartil ( $Q_2 = Med$ ) e terceiro quartil ( $Q_3$ )
- Fios: menor observação dentro de  $1.5 \times AIQ$  a partir de  $Q_1$  e maior observação dentro de  $1.5 \times AIQ$  a partir de  $Q_3$
- “Outliers”: observações menores do que  $Q_1 - 1.5 \times AIQ$  e observações maiores do que  $Q_3 + 1.5 \times AIQ$

# Simetria da distribuição

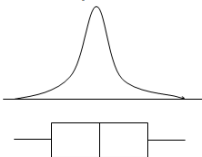
Distribuição assimétrica à esquerda



$\text{moda} > \text{mediana} > \text{média}$

$\text{Skewness} < 0$

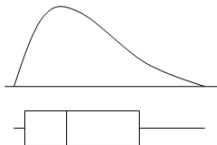
Distribuição simétrica



$\text{moda} = \text{mediana} = \text{média}$

$\text{Skewness} = 0$

Distribuição assimétrica à direita



$\text{moda} < \text{mediana} < \text{média}$

$\text{Skewness} > 0$

- analisada através da comparação das medidas de localização moda, mediana e média
- observável em representações gráficas como o histograma e o diagrama de extremos e quartis (“Boxplot”)
- medidas do grau de simetria (“skewness”)

# Construção de histograma e “Boxplot”

## Exemplo

Os seguintes dados são os tempos de espera numa fila de supermercado de 60 sujeitos selecionados aleatoriamente:

4	18	8	25	5.5	7	7	26	8	16	2	1
12	3	2	9	16	4	21	7	13	27	8	8
34.5	4	27	19	7	5	18	9	12	16	2	6
12	10	7	21	3	1	0.5	11	10	13	4	5
20	1.5	5	7	12	2	8.5	12	5	10	18	0.5

- o tempo de espera é uma variável ( $x$ ) quantitativa contínua de escala proporcional
- tamanho da amostra  $n = 60$
- número de classes ( $k$ ):  $k = 1 + 3.3 \log(60) = 6.87 \approx 7$
- amplitude ( $R$ ):  $R = 34.5 - 0.5 = 34$
- intervalo de classe ( $w$ ):  $w > \frac{34}{7} = 4.86 \approx 5$  logo  $w = 5$
- limite inferior da primeira classe ( $l$ ):  $l = 0.5 - \frac{7 \times 5 - 34}{2} = 0$ ; limite superior da última classe ( $u$ ):  $u = 0 + 7 \times 5 = 35$

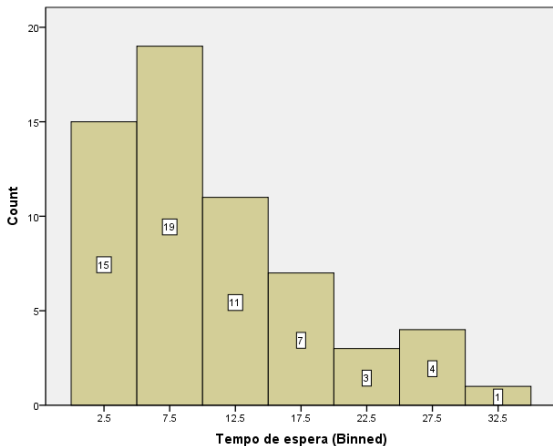
# Construção de histograma e “Boxplot”

## Exemplo

Tabela de frequências

Classe	$M_j$	$f_j$
$0 \leq x < 5$	2.5	15
$5 \leq x < 10$	7.5	19
$10 \leq x < 15$	12.5	11
$15 \leq x < 20$	17.5	7
$20 \leq x < 25$	22.5	3
$25 \leq x < 30$	27.5	4
$30 \leq x < 35$	32.5	1

Histograma (SPSS)



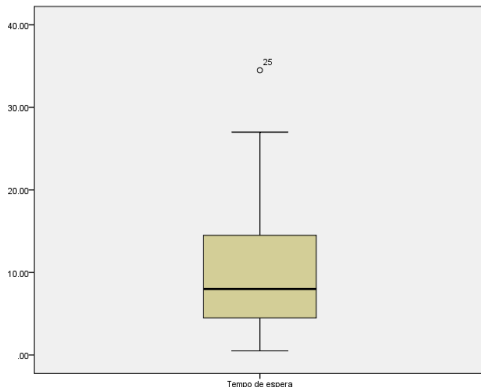
# Construção de histograma e “Boxplot”

## Exemplo

$i$	1	2	...	15	16	...	30	31	...	45	46	...	59	60
$x_{(i)}$	0.5	0.5	...	4	5	...	8	8	...	13	16	...	27	34.5

### “Boxplot” (SPSS)

- $Q_1 = x_{(15.25)} = 4.25$
- $Q_2 = \frac{x_{(30)} + x_{(31)}}{2} = 8$
- $Q_3 = x_{(45.75)} = 15.25$
- $AIQ = Q_3 - Q_1 = 15.25 - 4.25 = 11$
- $Q_1 - 1.5 \times AIQ = -12.25$ ;  
 $Q_3 + 1.5 \times AIQ = 31.75$   
logo  $x_{(60)} = 34.5 > 31.75$  é um “outlier”
- $\min x_i = x_{(1)} = 0.5$
- $\max x_i = x_{(59)} = 27$   
(excluindo  $x_{(60)}$ )



# Medidas descritivas

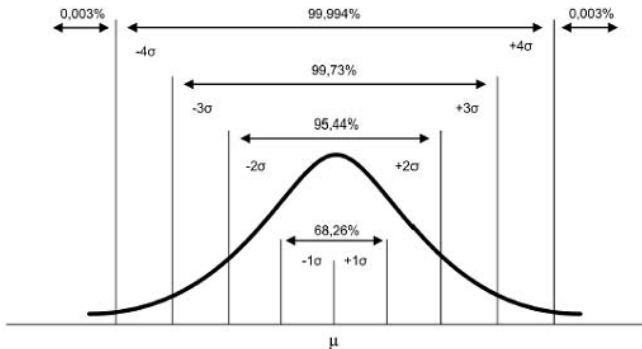
## Exemplo

- $\bar{x} = 10.2667$
- $Med = Q_2 = 8$
- $s^2 = 60.004$
- $s = 7.74622$
- $\min x_i = 0.5$
- $\max x_i = 34.5$
- $R = 34$
- $AIQ = 11$
- Assimetria à direita  
( $Skewness = 1.041 > 0$ )

## Medidas descritivas (SPSS)

Descriptives			Statistic	Std. Error
Tempo de espera	Mean		10.2667	1.00003
	95% Confidence Interval for Mean	Lower Bound	8.2656	
		Upper Bound	12.2677	
	5% Trimmed Mean		9.7315	
	Median		8.0000	
	Variance		60.004	
	Std. Deviation		7.74622	
	Minimum		.50	
	Maximum		34.50	
	Range		34.00	
	Interquartile Range		11.00	
	Skewness		1.041	.309
	Kurtosis		.695	.608

# Distribuição gaussiana ou normal



- tem forma de “sino” e é caracterizada completamente pela média ( $\mu$ ) e desvio padrão ( $\sigma$ )
- a distribuição normal padrão tem média 0 e desvio padrão 1
- é simétrica em relação à média  $\mu$
- unimodal com máximo em  $\mu$
- tende para 0 quando a variável tende para  $+\infty$  ou  $-\infty$
- tem dois pontos de inflexão nas abscissas  $\mu - \sigma$  e  $\mu + \sigma$  (a distribuição normal padrão tem pontos de inflexão em  $-1$  e  $+1$ )