

# ANADI

## Análise de Dados em Informática

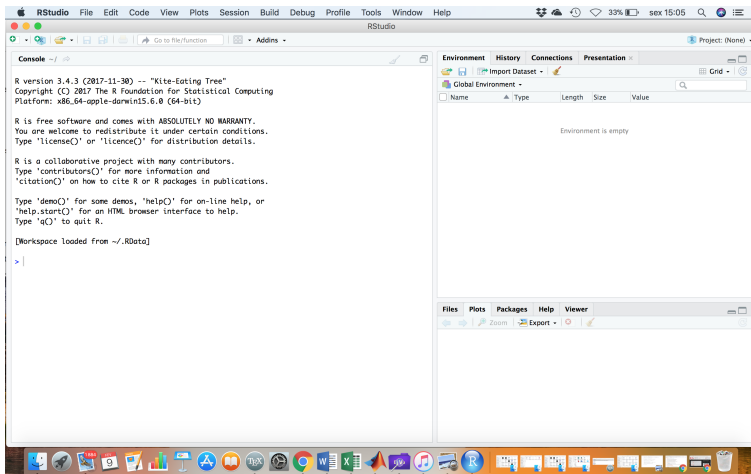
Aulas TP (revisões)

Instituto Superior de Engenharia do Porto

Ano letivo 2017/2018



- Para instalar o R: <http://www.r-project.org/>
- Para instalar o RStudio: <http://www.rstudio.com/>



- Existem três sinais de atribuição em R: (deve-se optar por usar apenas um)

```
> a <- 2
> a
[1] 2
> b=3.8
> b
[1] 3.8
> 9.55 -> c
```

- Podemos definir diferentes tipos de variáveis:

```
> x = 'Palmira'
> y = TRUE
> z=2+6i
> x1 = pi
> mode(x); mode(y); mode(z); mode(x1)
[1] "character"
[1] "logical"
[1] "complex"
[1] "numeric"
```

- Exemplos de operações aritméticas:

```
> 3*6
[1] 18
> 12/3
[1] 4
> 2^3+6
[1] 14
```

# Objectos em R

Objectos	Tipos	Diferentes tipos?
vector	numérico, caractere, complexo ou lógico	Não
factor	numérico ou caractere	Não
matriz	numérico, caractere, complexo ou lógico	Não
array	numérico, caractere, complexo ou lógico	Não
data.frame	numérico, caractere, complexo ou lógico	Sim
ts	numérico, caractere, complexo ou lógico	Sim
lista	numérico, caractere, complexo, lógico, função, expressão, etc.	Sim

# Vectores

```
> x=c(2,3,5,7,11,13,17) # criar um vector com os primeiros números primos
> x
[1] 2 3 5 7 11 13 17
> y=c(x,13,14,17) # concatenação de vectores
> y
[1] 2 3 5 7 11 13 17 13 14 17
> k=c('a','b') # vector de caracteres
> k
[1] "a" "b"
> kk=1:19 # criar vector com o comando :
> kk
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
#print(kk) # em alternativa kk (isto é um comentário)
> x^2 # as operações são efectuadas elemento a elemento
[1] 4 9 25 49 121 169 289
> x-y
[1] 0 0 0 0 0 0 0 -11 -11 -12
Warning message:
In x - y : longer object length is not a multiple of shorter object length
> length(x); length(y) # número de entradas dos vectores x e y
[1] 7
[1] 10
> x[4] # 4º número primo
[1] 7
> x[c(2,4,6)] # 2º, 4º e 6º número primo
[1] 3 7 13
> seq(1,13,2) # vector gerado pelo comando seq() (passo positivo)
[1] 1 3 5 7 9 11 13
> seq(33,1,-3) # vector gerado pelo comando seq() (passo negativo)
[1] 33 30 27 24 21 18 15 12 9 6 3
```

# Listas

```
> Mlista=list(aluno=c('jose','joao','antonio','maria'),nota=c(12,15,9,13))
> Mlista$aluno
[1] "jose"      "joao"      "antonio"   "maria"
> Mlista$aluno[3]
[1] "antonio"
> Mlista[[2]]
[1] 12 15 9 13
> Mlista[[2]][3]
[1] 9
# Algumas funções retornam listas p. ex.
> tt<-t.test(rnorm(500),rnorm(500),var.equal=T)
> tt

      Two Sample t-test
data:  rnorm(500) and rnorm(500)
t = 0.037737, df = 998, p-value = 0.9699
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.1211556  0.1259068
sample estimates:
 mean of x  mean of y
-0.0128063 -0.0151819

> is.list(tt)
[1] TRUE
> names(tt)
[1] "statistic" "parameter" "p.value"    "conf.int"   "estimate"   "null.value"
[7] "alternative" "method"     "data.name"

> tt$conf.int
[1] -0.1211556  0.1259068
attr(,"conf.level")
[1] 0.95
```

# Matrizes

```
> z=1:24 # criar vector
> zmatc=matrix(z,ncol=3) # criar matriz com 3 colunas com entradas de z
> zmatc
      [,1] [,2] [,3]
[1,]     1     9    17
[2,]     2    10    18
[3,]     3    11    19
[4,]     4    12    20
[5,]     5    13    21
[6,]     6    14    22
[7,]     7    15    23
[8,]     8    16    24
#
#---- outras formas de construir matrizes
# usar comandos cbind e rbind (usar o help...)
> ?cbind
> ?rbind
> zmatc[5,2] # extrair entrada da 5ª linha e 2ª coluna
[1] 13
> zmatc[,3] # extrair 3ª coluna
[1] 17 18 19 20 21 22 23 24
> zmatc[c(1,4,7),] # extrair 1ª, 4ª e 7ª linha
      [,1] [,2] [,3]
[1,]     1     9    17
[2,]     4    12    20
[3,]     7    15    23
```

## Data.frames/ arrays

- As data.frames são objectos semelhantes às matrizes.
- Ao contrário das matrizes podemos colocar elementos de diferentes tipos em cada coluna.

```
> alphab=LETTERS[1:26] # vector com as letras do abecedário!  
> amostra = sample(alphab, 5, replace = TRUE) # amostra aleatória de 5 letras com reposição  
mydata = data.frame(x = 1, y = 1:5, amostra = amostra) # Criada data.frame 'mydata'  
> mydata  
> x y amostra  
1 1 1      A  
2 1 2      L  
3 1 3      H  
4 1 4      S  
5 1 5      B
```

- As data.frames podem ser guardadas em arquivos em vários formatos p. ex. : .txt, .csv, etc...
- Para importar arquivos de dados no RStudio usar, **File - import data - "tipo de arquivo"** .
- **Arrays** são objectos idênticos às **matrizes** mas de dimensão superior a dois.



# Fenómenos probabilísticos vs determinísticos

Quando estudamos um fenómeno, o seu comportamento pode ser:

- **determinístico**: existe uma relação funcional bem definida entre as variáveis independentes e dependentes, e em que é possível prever com exatidão o seu comportamento.

**Explo:** No sistema de faturação de eletricidade, o valor da fatura é uma função bem definida entre a energia consumida e um perfil de consumo.

- **probabilístico**: não se consegue definir completamente uma relação funcional entre as variáveis independentes e dependentes, atribuindo-se importância decisiva ao fator sorte.

**Explo:** As preferências partidárias dos eleitores; o resultado do próximo jogo Sporting - Porto; número de clientes que chegam a uma agência bancária entre as 10h e as 11h.

A **Probabilidade** proporciona as ferramentas para se desenvolver modelos matemáticos que nos ajudam a prever o comportamento deste tipo de fenómenos, estimando o fator sorte.

Uma **experiência aleatória** é um qualquer processo que gera um resultado que pode ser diferente de cada vez que o processo é executado em iguais condições.

Uma **observação** é o resultado associado a uma realização de uma experiência aleatória.

O **espaço amostral**,  $S$  ou  $\Omega$ , é o conjunto de todos os resultados possíveis de uma experiência aleatória.

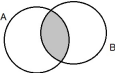
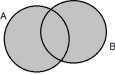
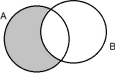

Um **acontecimento** (ou evento) é um subconjunto do espaço amostral.  
Diz-se:

- **simples** se contém exatamente um dos resultados possíveis de uma experiência aleatória (e.a.);
- **composto** se contém mais do que um dos resultados possíveis de uma e.a..

## Exemplo

- e.a.: Lançar um dado numerado até ao aparecimento de face "5".  
espaço amostral:  $S = \{5, \bar{5}5, \bar{5}\bar{5}5, \bar{5}\bar{5}\bar{5}5, \dots\}$  é um conjunto discreto (infinito numerável).
- e.a.: Contagem do número anual de acidentes rodoviários numa dada autoestrada.  
espaço amostral:  $S = \{0, 1, 2, 3, \dots\} = \mathbb{N}_0$  é um conjunto discreto (infinito numerável).
- e.a.: Medição do tempo de duração de uma lâmpada.  
espaço amostral:  $S = \mathbb{R}_0^+$  é um conjunto contínuo (infinito não numerável).

Sejam  $S$  o espaço amostral de uma e.a., e  $A$  e  $B$  dois acontecimentos.

operação	notação	descrição verbal	diagrama de Venn
interseção	$A \cap B$	realização simultânea de $A$ e $B$	
reunião	$A \cup B$	realização de pelo menos um dos dois eventos	
diferença	$A - B$	realização de $A$ sem que se realize $B$	
complementar	$\bar{A}$	não realização de $A$	

Dois acontecimentos dizem-se **mutuamente exclusivos, disjuntos ou incompatíveis** se e só se  $A \cap B = \emptyset$ .

## Leis de De Morgan

$$\overline{A \cap B} = \overline{A} \cup \overline{B} \quad \text{e} \quad \overline{A \cup B} = \overline{A} \cap \overline{B}$$

### Definição clássica de probabilidade:

Dada uma experiência aleatória com  $N$  resultados possíveis e diferentes ( $\#S = N$ ), igualmente prováveis e em número finito, se um acontecimento  $A$  tiver  $n$  ocorrências, então

$$P(A) = \frac{n}{N} = \frac{\text{n}^\circ \text{ resultados favoráveis}}{\text{n}^\circ \text{ resultados possíveis}}$$

## Propriedades:

- ①  $0 \leq P(A) \leq 1$
- ②  $P(S) = 1$
- ③  $P(\emptyset) = 0$
- ④ Se  $A_1, A_2, \dots, A_n$  forem acontecimentos mutuamente exclusivos, i.e.,  $A_i \cap A_j = \emptyset$ ,  $i \neq j$ , então
$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n)$$
- ⑤  $P(\bar{A}) = 1 - P(A)$
- ⑥  $A \subseteq B \Leftrightarrow P(A) \leq P(B)$

### Exemplo

Consideremos a e.a. que consiste no lançamento de um dado numerado até ao aparecimento de face "5". Verifiquemos a veracidade da Propriedade (2).

## Exercício

Consideremos uma e.a. onde estão associados dois acontecimentos  $A$  e  $B$ . A probabilidade de que só ocorra  $A$  é de 0,3, a probabilidade de que ambos ocorram é de 0,15 e a probabilidade de ocorrência de  $B$  é de 0,5. Determine a probabilidade de:

- 1 que pelo menos um dos acontecimentos ocorra;
- 2 que nenhum dos acontecimentos ocorra.

R: (1) 0,8; (2) 0,2.

## Definição frequencista de probabilidade:

Dada uma experiência aleatória, a probabilidade do acontecimento  $A$  é igual ao limite da frequência relativa da ocorrência desse evento

$$P(A) = \lim_{N \rightarrow +\infty} \frac{n(A)}{N},$$

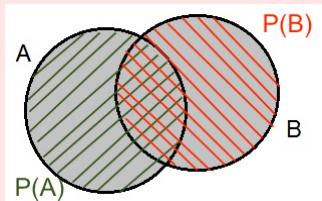
onde  $N$  é o número de realizações da e.a. e  $n(A)$  representa o número de vezes que o evento  $A$  ocorre nas  $N$  realizações da e.a..

# Reunião de eventos e regras aditivas

## Teorema

- 1 Sejam  $A$  e  $B$  dois eventos quaisquer de um espaço amostral  $S$ . Então

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$



- 2 Sejam  $A$ ,  $B$  e  $C$  eventos quaisquer de um espaço amostral  $S$ . Então
- $$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$



# Probabilidade condicionada

A probabilidade de um acontecimento  $A$  ocorrer dado que o evento  $B$  ocorreu denota-se por  $P(A|B)$ .

O efeito de se saber que  $B$  ocorreu é  $B$  tornar-se o espaço amostral.

Tem-se

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ se } P(B) \neq 0.$$

## Exercício

A tabela seguinte refere uma situação de emprego dos habitantes (adultos) de uma comunidade e está organizado em função do género sexual.

	nº empregados	nº desempregados	
homens	940	110	1050
mulheres	860	90	950
	1800	200	2000

- 1 Selecciona-se, ao acaso, um dos habitantes.
  - 1 Qual a probabilidade de ser mulher?
  - 2 Qual a probabilidade de estar desempregado?
  - 3 Qual a probabilidade de ser mulher e estar desempregada?
- 2 Selecciona-se, ao acaso, um dos habitantes e verifica-se que é mulher. Qual a probabilidade de estar desempregada?
- 3 Selecciona-se, ao acaso, um dos habitantes e verifica-se que está desempregado. Qual a probabilidade de ser mulher?

R: (1.1) 47,5%; (1.2) 10%; (1.3) 4,5%; (2) 9,5%; (3) 45%.

# Independência de eventos

Notemos que, se  $A$  e  $B$  são acontecimentos mutuamente exclusivos, então  $P(A|B) = 0 = P(B|A)$ , pois a ocorrência de um deles impede a ocorrência do outro. Consequentemente,  $A$  e  $B$  **não são independentes**.

Por outro lado, se  $B \subseteq A$ , então  $P(A|B) = 1$  e, consequentemente,  $A$  e  $B$  também **não são independentes**.

Ou seja, saber que  $B$  ocorreu dá-nos alguma informação acerca da ocorrência de  $A$ .

Quando saber que  $B$  ocorreu não traz qualquer informação sobre a ocorrência de  $A$ , dizemos que  $A$  e  $B$  **são acontecimentos independentes**.

$A$  e  $B$  são acontecimentos independentes se e só se

$$P(A \cap B) = P(A)P(B).$$

### Teorema

Se  $A$  e  $B$  são acontecimentos independentes, com  $P(A) > 0$  e  $P(B) > 0$ , então

$$P(A|B) = P(A) \quad \text{e} \quad P(B|A) = P(B).$$

# Interseção de eventos e regras multiplicativas

## Teorema

- ❶ A probabilidade conjunta dos eventos  $A$  e  $B$  é

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B),$$

com  $P(A) > 0$  e  $P(B) > 0$ .

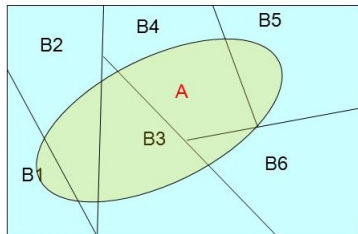
- ❷ A probabilidade conjunta dos eventos  $A_1, A_2, \dots, A_n$  é

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1)P(A_2|A_1) \cdots P(A_n|A_1 \cap A_2 \cap \dots \cap A_{n-1}),$$

com  $P(A_i) > 0$  para todo o  $i = 1, \dots, n$ .

# Teorema de Bayes

Considere uma partição  $B_1, B_2, \dots, B_n$  do espaço amostral  $S$  e seja  $A$  um evento de  $S$ , como mostra a figura:



Temos  $A = (A \cap B_1) \cup (A \cap B_2) \cup \dots \cup (A \cap B_n)$ .

Como os acontecimentos são mutuamente exclusivos, então

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \dots + P(A \cap B_n).$$

e, pela probabilidade conjunta de eventos, resulta que

$$P(A) = \sum_{i=1}^n P(B_i)P(A|B_i).$$

## Teorema da probabilidade total

Se os eventos  $B_1, B_2, \dots, B_n$  constituem uma partição do espaço amostral  $S$ , então, para qualquer evento  $A$ , temos

$$P(A) = \sum_{i=1}^n P(B_i)P(A|B_i).$$

## Exercício

Uma peça é manufaturada por 3 fábricas: 1, 2, 3. Sabe-se que a fábrica 1 produz o dobro das peças de cada uma das fábricas 2 e 3. Além disso, 2% das peças produzidas pelas fábricas 1 e 2 são defeituosas, enquanto 4% das produzidas pela fábrica 3 são defeituosas. Todas as peças produzidas são colocadas no mesmo armazém. Tira-se uma peça ao acaso:

- 1 Qual é a probabilidade dessa peça ser defeituosa?
- 2 Se a peça retirada for defeituosa, qual a probabilidade de ter sido produzida na fábrica 1?

R: (1) 0,025; (2) 0,40.

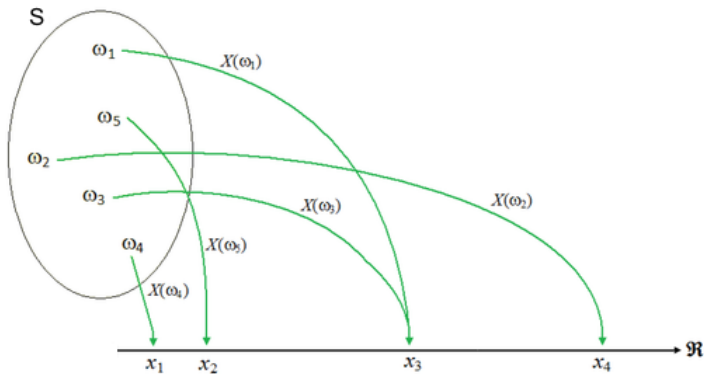
## Teorema de Bayes

Se os eventos  $B_1, B_2, \dots, B_n$  constituem uma partição do espaço amostral  $S$ , então, para qualquer evento  $A$  de  $S$ , temos

$$P(B_k|A) = \frac{P(B_k \cap A)}{P(A)} = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^n P(B_i)P(A|B_i)}, \quad k = 1, \dots, n.$$



Uma **variável aleatória (v.a.)**  $X$  é uma função que associa um número real  $x$  a cada resultado  $s$  do espaço amostral  $S$ .



As variáveis aleatórias classificam-se em:

- **discretas**: se o conjunto dos valores possíveis de  $X$  for finito ou for infinito numerável;
- **contínuas**: se o conjunto dos valores possíveis de  $X$  for não numerável (intervalo ou reunião de intervalos).

## Exemplo

- Discretas:
  - 1 Analisam-se livros com 250 páginas para se determinar número de páginas com erros:  $S' = \{0, 1, 2, \dots, 250\}$ ;
  - 2 Contam-se as partículas emitidas por uma fonte radioativa durante um intervalo de tempo:  $S' = \{0, 1, 2, \dots\} = \mathbb{N}$ .
- Contínuas:
  - 1 Estima-se o consumo de gasolina de um determinado modelo de automóvel. Para isso, regista-se a distância percorrida com um litro de gasolina (admite-se que não excede 50 km):  $S' = [0, 50]$ ;
  - 2 Observa-se o tempo entre avarias de uma máquina em funcionamento numa fábrica:  $S' = [0, +\infty[$ .

## V.A. Discretas

### Função de probabilidade

A **função de probabilidade** de uma v. a. discreta  $X$  é uma função  $f$  que associa a cada valor possível  $x$  de  $X$  a sua probabilidade

$$f(x) = P(X = x),$$

ou seja,

$$f(x) = \begin{cases} P(X = x), & \text{se } x \in S', \\ 0, & \text{se } x \notin S'. \end{cases}$$

#### Propriedades:

- 1  $f(x) \geq 0$ , para todo o  $x \in \mathbb{R}$ ;
- 2  $\sum_x f(x) = P(S) = 1$ .

# Função de distribuição acumulada

A **função de distribuição acumulada**  $F$  de uma v. a. discreta  $X$  com função de probabilidade  $f$  está definida para qualquer valor real  $x$  e é dada por

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i).$$

## Propriedades:

- 1  $F$  é uma função com descontinuidades e em escada;
- 2  $F$  é constante para todo o  $x \in [x_i, x_{i+1}[$ ;
- 3  $F(x) = 0$  para todo o  $x < x_{min}$ ;
- 4  $F(x) = 1$  para todo o  $x \geq x_{max}$ .

# Parâmetros de uma v.a. discreta

Média ou valor esperado:

$$\mu = E(X) = \sum_i x_i f(x_i)$$

Propriedades:

- ①  $E(k) = k$ , onde  $k = \text{constante}$ ;
- ②  $E(kX) = kE(X)$ ;
- ③  $E(X \pm Y) = E(X) \pm E(Y)$ ;
- ④  $E(XY) = E(X)E(Y)$ , se  $X$  e  $Y$  forem v.a. independentes.

## Variância:

$$\sigma^2 = V(X) = E[(X - \mu)^2] = E(X^2) - \mu^2 = \sum_i x_i^2 f(x_i) - \mu^2$$

## Desvio padrão:

$$\sigma = +\sqrt{V(X)}$$

## Propriedades:

- 1  $V(k) = 0$ , onde  $k = \text{constante}$ ;
- 2  $V(kX) = k^2 V(X)$ ;
- 3  $V(X + Y) = V(X) + V(Y)$ , se  $X$  e  $Y$  forem v.a. independentes.

## V.A. Contínuas

Seja  $X$  uma v.a. contínua que representa, por exemplo:

- ① a temperatura na cidade do Porto, amanhã, às 12h;
- ② o peso de um estudante do ISEP.

Então, a probabilidade de  $X$  ser, exatamente,

- ①  $16^{\circ}\text{C}$ ,
- ②  $70\text{ kg}$ ,

é tão pequena que é considerada nula.

Uma v.a. contínua tem **probabilidade nula de tomar exatamente qualquer um dos seus valores**.

Isto porque, como a v.a. contínua é definida num intervalo não numerável de valores, então antes da experiência ser efetuada, a probabilidade de ocorrer exatamente um desses valores é nula.

# Função densidade de probabilidade

A **função densidade de probabilidade** de uma v.a. contínua  $X$  é uma função  $f$  tal que

$$P(a < X < b) = \int_a^b f(x)dx, \forall a, b \in \mathbb{R} \text{ t.q. } a \leq b$$

e que satisfaz as **propriedades** seguintes:

- 1  $f(x) \geq 0$ , para todo o  $x \in \mathbb{R}$ ;
- 2  $\int_{-\infty}^{+\infty} f(x)dx = 1$ .

**Nota:** Para qualquer  $x \in \mathbb{R}$ ,  $P(X = x) = 0$ .



## Função de distribuição acumulada

Seja  $X$  uma v.a. contínua com função densidade de probabilidade  $f$ . A sua **função de distribuição acumulada**  $F$  está definida para qualquer valor real  $x$  e é dada por

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du.$$

### Propriedades:

- 1  $F$  é uma função contínua;
- 2 Se  $x \leq y$ , então  $F(x) \leq F(y)$ ;
- 3  $F(-\infty) = \lim_{x \rightarrow -\infty} F(x) = 0$ ;
- 4  $F(+\infty) = \lim_{x \rightarrow +\infty} F(x) = 1$ ;
- 5  $f(x) = \frac{dF(x)}{dx}$ , se a função de distribuição acumulada for derivável;
- 6  $P(a \leq X \leq b) = F(b) - F(a)$ .

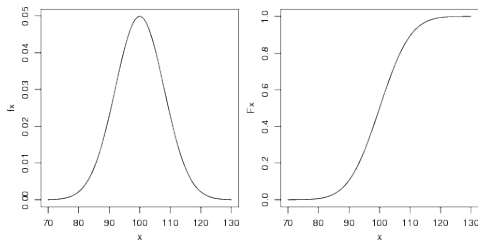


Figure: Função densidade de probabilidade vs. função de distribuição acumulada.

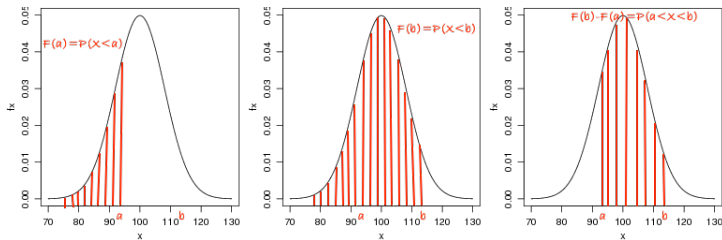


Figure: Ilustração da Propriedade (6).

# Parâmetros de uma v.a. contínua

Média ou valor esperado:

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx$$

Propriedades: goza das mesmas das v.a. discretas.

Variância:

$$\sigma^2 = V(X) = E[(X - \mu)^2] = E(X^2) - \mu^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x)dx$$

Desvio padrão:

$$\sigma = +\sqrt{V(X)}$$

Propriedades: goza das mesmas das v.a. discretas.

# Distribuições de probabilidade teóricas

## Distribuições Discretas: Binomial

Seja  $X$  uma v.a. que representa o número de sucessos em  $n$  experiências independentes com apenas dois resultados possíveis, sendo  $p$  a probabilidade de sucesso.

Então, diz-se que  $X$  tem *distribuição binomial* com parâmetros  $n$  e  $p$  e denota-se por

$$X \sim \text{Bin}(n, p)$$

A função de probabilidade da v.a.  $X \sim \text{Bin}(n, p)$  é

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{se } x \in \{0, 1, 2, \dots, n\}, \\ 0, & \text{se } x \notin \{0, 1, 2, \dots, n\}. \end{cases}$$

A função de distribuição acumulada da v.a.  $X \sim \text{Bin}(n, p)$  é

$$F(x) = \begin{cases} 0, & \text{se } x < 0 \\ \sum_{i=0}^{\lfloor x \rfloor} \binom{n}{i} p^i (1-p)^{n-i}, & \text{se } 0 \leq x < n \\ 1, & \text{se } x \geq n. \end{cases}$$

## Exercício

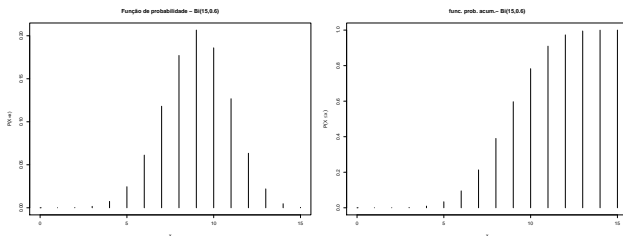
A Sara e o Francisco gostam muito de jogar xadrez, e a Sara ganha 60% dos jogos. Resolveram efetuar um campeonato de 15 jogos.

- a) Qual é a probabilidade da Sara ganhar exatamente 10 jogos?
- b) Qual é a probabilidade da Sara ganhar pelo menos 10 jogos?
- c) Qual é a probabilidade da Sara ganhar entre 4 e 8 jogos?

```
# resolução no R
> # alínea a)
> dbinom(10,size=15,prob=0.6) # dbinom: função de probabilidade
[1] 0.1859378
> # alínea b)
> 1-pbinom(9,size=15,prob=0.6) # pbinom: função de distribuição acumulada
[1] 0.4032156
> # alínea c)
> pbinom(8,size=15,prob=0.6)-pbinom(3,size=15,prob=0.6)
[1] 0.3882591
> # resolução alternativa
> sum(dbinom(10:15, 15, 0.6)) # alínea b)
[1] 0.4032156
> sum(dbinom(4:8, 15, 0.6)) # alínea c)
[1] 0.3882591
```

# Gráfico da f.p. e da f.d.a. - Bin(15, 0.6)

```
> x=c(0:15)
> plot(x,dbinom(x,size=15,prob=0.6),xlab='x',ylab='P(X=x)',
+      main='Função de probabilidade - Bi(15,0.6)',type='h')
> plot(x,dbinom(x,size=15,prob=0.6),xlab='x',
+      ylab=expression(P(X <= x)),main='func. prob. acum. - Bi(15,0.6)', type='h')
```



# Propriedades

- Parâmetros:  $n \in \mathbb{N}$ ,  $p \in ]0, 1[$ .
- Gama de valores:  $\{0, 1, 2, \dots, n\}$ .
- Média ou valor esperado:

$$\mu = E(X) = \sum_{x=0}^n xP(X=x) = \sum_{x=0}^n x \binom{n}{x} p^x (1-p)^{n-x} = np.$$

- Variância:

$$\sigma^2 = V(X) = E(X^2) - [E(X)]^2 = \sum_{x=0}^n x^2 \binom{n}{x} p^x (1-p)^{n-x} - (np)^2 = np(1-p).$$

- Moda:

$$\begin{cases} \lfloor (n+1)p \rfloor, & \text{se } (n+1)p \notin \mathbb{Z}, \\ (n+1)p \text{ e } (n+1)p - 1, & \text{se } (n+1)p \in \mathbb{Z}. \end{cases}$$

# Distribuições Discretas: Poisson

Na distribuição binomial:

- sabemos o tamanho da amostra;
- sabemos quantas vezes o acontecimento ocorreu e quantas vezes não ocorreu.

Nem sempre isto é possível:

- observar uma tempestade durante uma hora e contar o número de relâmpagos (não faz sentido contar quantas vezes não relampejou);
- número de golos num desafio de futebol;
- número de defeitos num tapete de arraiolos.

Assim, estamos interessados em contar o número de eventos que ocorrem durante um dado intervalo de tempo ou numa dada região espacial.



Seja  $\lambda > 0$  o número médio de eventos que ocorrem num dado intervalo de tempo (ou numa dada região espacial) e seja

$X$  uma v.a. que representa o número de eventos (independentes) que ocorrem nesse intervalo de tempo (ou nessa região espacial).

Então, diz-se que  $X$  segue *distribuição de Poisson* com parâmetro  $\lambda$  e denota-se por

$$X \sim Po(\lambda).$$

A função de probabilidade da v.a.  $X \sim Po(\lambda)$  é

$$f(x) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!}, & \text{se } x \in \mathbb{N}_0 = \{0, 1, 2, 3, \dots\}, \\ 0, & \text{se } x \notin \mathbb{N}_0. \end{cases}$$

A função de distribuição acumulada da v.a.  $X \sim Po(\lambda)$  é

$$F(x) = \begin{cases} 0, & \text{se } x < 0 \\ \sum_{i=0}^{\lfloor x \rfloor} \frac{e^{-\lambda} \lambda^i}{i!}, & \text{se } x \geq 0. \end{cases}$$

# Propriedades

- Parâmetro:  $\lambda \in ]0, +\infty[$ .
- Gama de valores:  $\{0, 1, 2, \dots\} = \mathbb{N}_0$ .
- Média ou valor esperado:

$$\mu = E(X) = \sum_{x=0}^{+\infty} xP(X=x) = \sum_{x=0}^{+\infty} x \frac{e^{-\lambda} \lambda^x}{x!} = \lambda.$$

- Variância:

$$\sigma^2 = V(X) = \lambda.$$

- Moda:

$$\begin{cases} \lfloor \lambda \rfloor, & \text{se } \lambda \notin \mathbb{Z}, \\ \lambda - 1 \text{ e } \lambda, & \text{se } \lambda \in \mathbb{Z}. \end{cases}$$

- Se  $X_1, X_2, \dots, X_n$  são v.a. independentes e se  $X_i \sim Po(\lambda_i)$ , então

$$X_1 + X_2 + \dots + X_n \sim Po(\lambda_1 + \lambda_2 + \dots + \lambda_n).$$

- Se  $\lambda$  for o número médio de eventos num dado intervalo de tempo (região espacial) e se pretendermos calcular probabilidades para um intervalo de tempo de amplitude  $t$  vezes a inicial, então o novo parâmetro deixará de ser  $\lambda$  e passará a ser  $\lambda t$ . Assim

$$f(x) = \begin{cases} \frac{e^{-\lambda t}(\lambda t)^x}{x!}, & \text{se } x \in \mathbb{N}_0, \\ 0, & \text{se } x \notin \mathbb{N}_0. \end{cases}$$

$$E(X) = \lambda t \quad \text{e} \quad V(X) = \lambda t$$

## Exercício

O número de chamadas que chegam à central telefónica de uma associação de defesa do consumidor é uma v.a. de Poisson com média 1,5 chamadas em cada 10 minutos.

- 1 Considere o período entre as 9:00 e as 9:10. Determine a probabilidade da associação:
  - a) Não receber qualquer chamada.
  - b) Receber mais de duas chamadas.
- 2 Considere o período entre as 11:00 e as 11:30. Determine a probabilidade da associação:
  - a) Não receber qualquer chamada.
  - b) Receber mais de duas chamadas.

```
> #1 alínea a) e b)
> dpois(0,lambda=1.5)
[1] 0.2231302
> 1-ppois(2,lambda=1.5)
[1] 0.1911532
> #2 alíneas a) e b)
> dpois(0,lambda=3*1.5)
[1] 0.011109
> 1-ppois(2,lambda=3*1.5)
[1] 0.8264219
```

## Distribuições Contínuas: Uniforme

A v.a.  $X$  tem *distribuição uniforme contínua* no intervalo  $[a, b]$  e denota-se por

$$X \sim U(a, b),$$

se a probabilidade de  $X$  tomar um valor num subintervalo de  $[a, b]$  for proporcional ao comprimento desse subintervalo (e, portanto, a função densidade de probabilidade é constante em  $[a, b]$ ).

### Função densidade de probabilidade

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{se } x \in [a, b], \\ 0, & \text{se } x \notin [a, b]. \end{cases}$$

### Função de distribuição acumulada

$$F(x) = \begin{cases} 0, & \text{se } x < a \\ \frac{x-a}{b-a}, & \text{se } x \in [a, b] \\ 1, & \text{se } x > b. \end{cases}$$

# Propriedades

- Parâmetros:  $a, b$ , com  $a < b$ .
- Gama de valores:  $[a, b]$ .
- Média ou valor esperado:

$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_a^b \frac{x}{b-a} dx = \frac{1}{b-a} \left[ \frac{x^2}{2} \right]_a^b = \frac{a+b}{2}.$$

- Variância:

$$\sigma^2 = V(X) = \frac{(b-a)^2}{12}.$$

## Exercício

Encomendaram-se peças de um determinado tipo sujeitas à especificação seguinte (em mm):

Diâmetro (D):  $9.60 \leq D \leq 10.40$

Comprimento (C):  $99 \leq C \leq 101$ .

Sabe-se que o diâmetro (em mm) tem distribuição uniforme no intervalo  $[9.5, 10.4]$  e o comprimento tem distribuição uniforme no intervalo  $[98, 102]$ .

- Calcule a probabilidade de uma dada peça ser rejeitada.
- De entre as peças rejeitadas, qual é a percentagem das que têm diâmetro que satisfaz a especificação?

```
> PA_D=1-punif(9.6,min=9.5,max=10.4) # prob de o diametro ser aceite
> PA_C=punif(101,min=98,max=102)-
+   punif(99,min=98,max=102) # prob de o comprimento ser aceite
> 1-(PA_D*PA_C) # resp. alínea a)
[1] 0.5555556
> PA_D*(punif(99,min=98,max=102)
+   +1-punif(101,min=98,max=102))/(1-(PA_D*PA_C)) # resp. alínea b)
[1] 0.8
```

## Distribuições Contínuas: Exponencial

É usada em situações em que se pode identificar um processo de Poisson, i.e., de ocorrência de eventos, a uma taxa constante, num intervalo de tempo ou numa região do espaço. Enquanto a distribuição de Poisson é usada para contar o número de eventos, a distribuição exponencial usa-se para determinar o intervalo de tempo entre dois eventos.

É usada, em particular, para representar intervalos de tempo entre eventos independentes (exemplo: tempo entre a chegada de clientes a um estabelecimento, tempo entre a ocorrência de avarias, duração de máquinas, etc.).

**Função densidade de probabilidade** Diz-se que a v.a. contínua  $X$  tem *distribuição exponencial* com parâmetro  $\lambda > 0$  e denota-se por

$$X \sim \text{Ex}(\lambda),$$

se a sua função densidade de probabilidade for dada por

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{se } x \geq 0, \\ 0, & \text{se } x < 0. \end{cases}$$



## Função de distribuição acumulada

$$F(x) = \begin{cases} 0, & \text{se } x < 0, \\ 1 - e^{-\lambda x}, & \text{se } x \geq 0. \end{cases}$$

### Propriedades:

- Parâmetro:  $\lambda > 0$ .
- Gama de valores:  $[0, +\infty[$ .
- Média ou valor esperado:  
$$\mu = E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \int_0^{+\infty} x\lambda e^{-\lambda x}dx = \dots = \frac{1}{\lambda}.$$
- Variância:  
$$\sigma^2 = V(X) = E(X^2) - \mu^2 = \dots = \frac{1}{\lambda^2}.$$
- É a única distribuição contínua que não tem memória, i.e.,  
$$P(X > k + c | X > k) = P(X > c).$$
- Se  $X$  é uma v.a. discreta que representa o número de ocorrências de um determinado evento num dado intervalo de tempo e  $Y$  é uma v.a. contínua que representa o tempo entre ocorrências sucessivas desse mesmo evento, então

$$X \sim Po(\lambda) \Leftrightarrow Y \sim Ex(\lambda).$$

## Exercício

Os acidentes que ocorrem num estaleiro de construção naval são estatisticamente independentes e o número de acidentes tem uma distribuição de Poisson com média de 6 por mês. Considere a variável aleatória  $T$  que representa o intervalo de tempo entre acidentes consecutivos. Suponha que o estaleiro está em laboração contínua.

- Identifique a distribuição de  $T$ .  $R: T \sim Ex(6)$
- Qual a probabilidade de não ocorrerem acidentes na próxima semana?
- Qual a probabilidade de ocorrerem acidentes em, pelo menos, 4 das próximas 5 semanas?

```
> # resolução usando as distribuições Po e Ex
> # alínea b)
> dpois(0,1/4*6) # Poisson
[1] 0.2231302
> 1-pexp(1/4,rate=6) # Exponencial
[1] 0.2231302
> pexp(1/4,rate=6,lower.tail=F) # Exponencial (alternativa)
[1] 0.2231302
> # alínea c)
> P1s=1-dpois(0,1/4*6) # Prob. de haver acidentes numa semana
> pbinom(3,5,P1s,lower.tail=F)
[1] 0.6893403
```

## Distribuições Contínuas: Normal

É a distribuição contínua mais importante, pois, entre outros:

- É um modelo adequado para representar muitos fenómenos da vida real (altura e o peso humano, etc.).
- É muito usada na inferência estatística (adiante).

## Função densidade de probabilidade

Diz-se que a v.a. contínua  $X$  tem *distribuição normal* com média  $\mu$  e variância  $\sigma^2$  e denota-se por

$$X \sim N(\mu, \sigma^2),$$

se a sua função densidade de probabilidade for dada por

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad x \in \mathbb{R}.$$

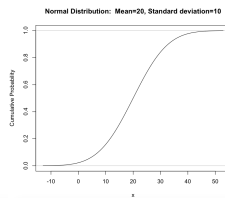
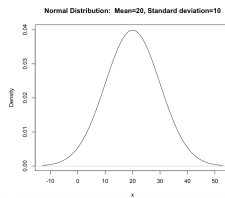
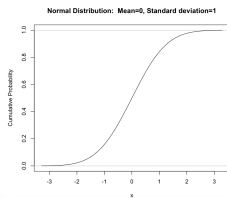
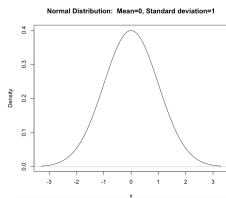
# Características

- Tem a forma de sino e um único máximo (é unimodal) em  $x = \mu$ .
- É simétrica relativamente a um eixo vertical que passa por  $x = \mu$ , pelo que se tem média = mediana = moda.
- $\lim_{x \rightarrow -\infty} = 0 = \lim_{x \rightarrow +\infty}$ .
- Tem pontos de inflexão para  $x = \mu \pm \sigma$ .

# Função de distribuição acumulada

$$F(x) = P(X \leq x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^x e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

A função  $F$  não tem forma fechada, ou seja, não há nenhuma função expressa por um número finito de funções elementares cuja derivada seja a integranda acima.



**Figure:** Função densidade de probabilidade vs função de distribuição acumulada.

# Propriedades

- Parâmetros:  $\mu$  e  $\sigma^2$ .
- Gama de valores:  $] -\infty, +\infty[$ .

- Média ou valor esperado:

$$E(X) = \int_{-\infty}^{+\infty} xf(x)dx = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} xe^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx.$$

Fazendo a mudança de variável  $z = \frac{x-\mu}{\sigma}$ , onde  $dz = \frac{1}{\sigma} dx$ , vem:

$$\begin{aligned} E(X) &= \frac{\sigma}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (\sigma z + \mu) e^{-\frac{1}{2}z^2} dz = \frac{\sigma}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (\sigma z + \mu) e^{-\frac{1}{2}z^2} dz = \\ &= \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} ze^{-\frac{1}{2}z^2} dz + \frac{\sigma\mu}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} e^{-\frac{1}{2}z^2} dz = 0 + \mu 1 = \mu. \end{aligned}$$

- Variância:

$$V(X) = E(X^2) - \mu^2 = \dots = \sigma^2.$$

## Distribuição Normal Standardizada

Diz-se que a v.a. contínua  $Z$  tem *distribuição normal standardizada* se

$$Z \sim N(0, 1).$$

$$\text{Se } X \sim N(\mu, \sigma^2), \text{ então } Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Assim, o cálculo de probabilidades envolvendo a v.a.  $X \sim N(\mu, \sigma^2)$  pode ser reduzido ao cálculo com a v.a.  $Z \sim N(0, 1)$ , da forma seguinte:

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}\right)$$

# Aditividade

Qualquer combinação linear de v.a. normais e independentes é, ainda, uma v.a. normal. Mais concretamente, se as v.a.

$$X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \dots, n,$$

são independentes, então

$$\sum_{i=1}^n a_i X_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right),$$

com  $a_i \in \mathbb{R}$ , para todo o  $i = 1, 2, \dots, n$ .



## Exercício

A duração  $T$  de um relógio de determinado modelo, antes de avariar, é uma v.a. normal, com média de 11 anos e desvio padrão de 1 ano. O fabricante pretende oferecer um período de garantia, dentro do qual os relógios avariados são substituídos por relógios novos. Esse período deverá ser tão grande quanto possível, mas sem que os custos se tornem in comportáveis.

- 1 Qual a probabilidade de um relógio durar mais de 11 anos?
- 2 Qual a probabilidade de um relógio durar menos de 10 anos?
- 3 Qual deverá ser o período de garantia, se a fábrica não pretender substituir mais do que 5% dos relógios?
- 4 Considere um cliente que comprou 5 relógios.
  - a) Qual a probabilidade de pelo menos dois deles durarem mais de 12 anos?
  - b) Qual a probabilidade de pelo menos um deles dura menos de 10 anos?

## Resolução:

```
> pnorm(11,mean=11,sd=1,lower.tail=F) # pergunta 1
[1] 0.5
> pnorm(10,11,1) # pergunta 2
[1] 0.1586553
> qnorm(0.05,11,1,lower.tail=T) # pergunta 3
[1] 9.355146
> # pergunta 4
> p1rM12=pnorm(12,mean=11,sd=1,lower.tail=F) # prob. de um relógio durar mais que 12 anos
> pbinom(1,5,p1r12,lower.tail=F) # alínea a)
[1] 0.1809451
> p1rm10=pnorm(10,mean=11,sd=1,lower.tail=T) # prob. de um relógio durar menos que 12 anos
> pbinom(0,5,p1rm10,lower.tail=F) # alínea b)
[1] 0.5784298
```

**Nota:** Se pretendermos arredondar os dados com n casa decimais pode-se usar a função "**round**". p. ex. se pretendermos os resultados em percentagem com 1 casa decimal

```
> # pergunta 4 respostas em percentagem com uma casa decimal
> round(100*pbinom(1,5,p1r12,lower.tail=F),1) # alínea a)
[1] 18.1
> round(100*pbinom(0,5,p1rm10,lower.tail=F),1) # alínea a)
[1] 57.8
```