Estatística para Ciências de Dados

Aula 2: Probabilidade

Mariana Cúri
ICMC/USP
mcuri@icmc.usp.br

CEPID - Centro de Ciências Matemáticas Aplicadas à Indústria

Conteúdo

Conceitos básicos

São Paulo, sábado, 14 de fevereiro de 2009 FOLHA DE S.PAULO ilustrada

Texto Anterior | Próximo Texto | Índice

Livros/Crítica/"Uma Senhora Toma Chá"

Professor conta bons casos sobre paradoxos da ciência

"... A tal senhora do título dizia que **o gosto do chá fica diferente se alguém põe antes o leite na xícara e depois derrama o chá, ou se alguém põe antes o chá e depois derrama o leite.** Ouvindo isso, naquela tarde de verão em Cambridge, Ronald Aymler Fisher propôs que se testasse a proposição: <u>oferecer diferentes xícaras de chá com leite àquela senhora, convenientemente vendados os seus olhos, e verificar se ela era capaz de acertar a ordem da mistura</u>."

Conceitos básicos

Experimento aleatório

Um experimento cujo resultado não se conhece com certeza, mesmo se repetido exatamente da mesma maneira.

Espaço amostral (Ω)

Conjunto cujos elementos são todos os possíveis resultados do experimento. Pode ser <u>discreto</u> (finito ou infinito enumerável) ou <u>contínuo</u>.

Evento (A, B, ...)

Qualquer subconjunto de Ω.

Oferecer uma xícara de chá com leite à senhora e verificar se ela acerta ou não a ordem da preparação.

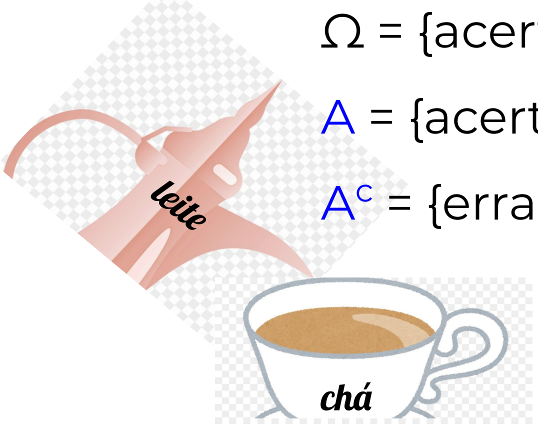
Desenho do Experimento

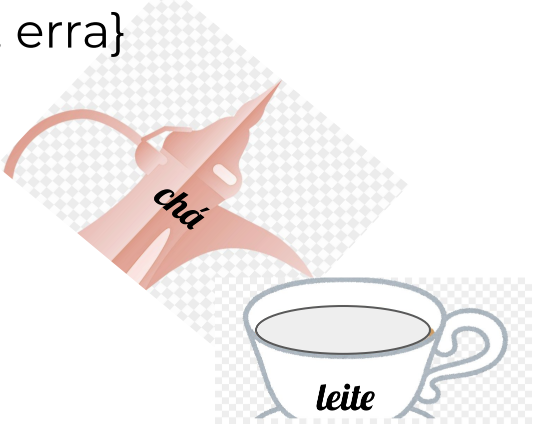
- sorteia-se um dos 2 tratamentos (tipos de preparo)
- olhos vendados
- repete-se 4 vezes o experimento

Ω = {acerta, erra}

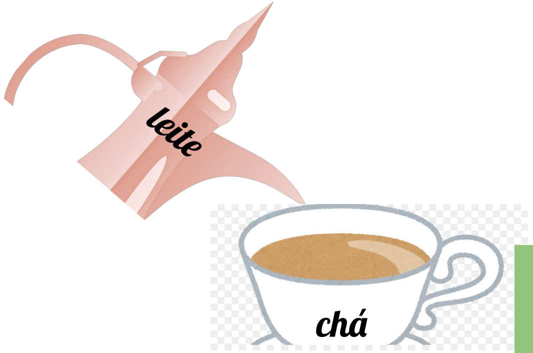
A = {acerta}

A^c = {erra}





Conceitos básicos



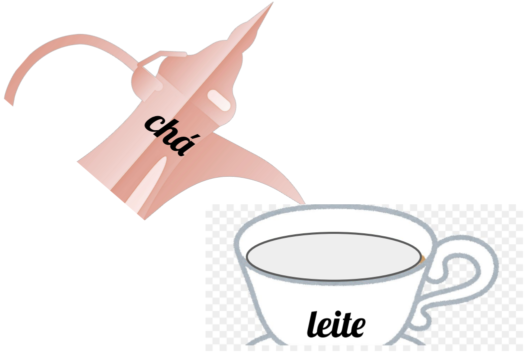
acerta acerta

erra erra

- mesma quantidade de chá e de leite nos 2 tratamentos
- xícara com camada dupla para isolamento térmico

QUANTO SE ESPERA DE ACERTO AO ACASO?

ao acaso: **P(acerto)=0,5**



acerta acerta

erra erra

Conceitos básicos: espaço amostral contínuo

Encontrar os valores de referência de normalidade para exames laboratoriais de hemograma da população brasileira.

Desenho do Experimento

- amostra de brasileiros sem doenças prévias
- limites estratificados por sexo, faixa etária
- 24h sem exercício físico e 48h sem álcool

Espaço amostral

(Hemoglobina)

$\Omega = \mathfrak{R}_+$

A = [12; 16,5]

B = [11; 13]

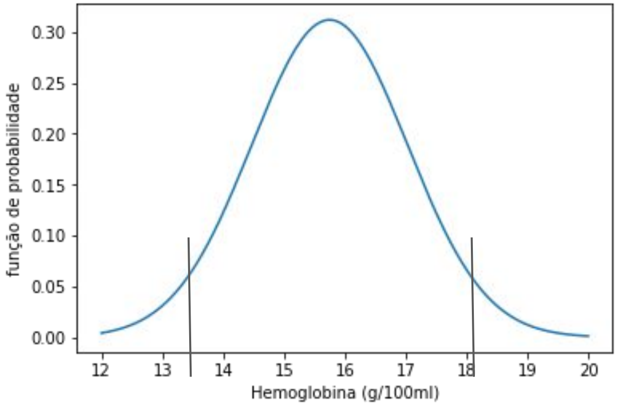
C = [13,5; 18]

eventos mutuamente exclusivos

$B \cap C = \varnothing$

Valores normais para eritrócitos, hemoglobina, hematócrito[3]

Tipo de indivíduo	Eritrócitos (x 10^6/mm³)	Hemoglobina (g/100mL)	Hematócrito (%)
Recém nascidos (a termo)	4 - 5,6	13,5 - 19,6	44 - 62
Crianças (3 meses)	4,5 - 4,7	9,5 - 12,5	32 - 44
Crianças (1 ano)	4,0 - 4,7	11,0 - 13	36 - 44
Crianças (10 a 12 anos)	4,5 - 4,7	11,5 - 14,8	37 - 44
Mulheres (gestantes)	3,9 - 5,6	11,5 - 16,0	34 - 47
Mulheres	4,0 - 5,6	12 - 16,5	35 - 47
Homens	4,5 - 6,5	13,5 - 18	40 - 54



Fonte: Wikipedia: https://pt.wikipedia.org/wiki/Hemograma

Conceitos básicos

Variável aleatória (X)

Função que associa um valor real a cada elemento de Ω. Também podem ser discretas ou contínuas.

$$X : \omega \rightarrow \Re$$
$$X(\omega) = x$$

$X:$ resposta da senhora sobre a ordem de preparo da bebida

$x = 0, 1$

se acerta ao acaso: p=0,5

$$X = \begin{cases} 0, \text{ se erra} & P(X=0) = (1 - p) \\ 1, \text{ se acerta} & P(X=1) = p \end{cases}$$

função de probabilidade
(ou função massa de probabilidade)

4 repetições		Probabilidade
$X_1 = x_1$	= 0	(1-p)
$X_2 = x_2$	= 1	p
$X_3 = x_3$	= 1	p
$X_4 = x_4$	= 1	p

$P(X_1=0, \dots, X_4= 1) = p^3 . (1-p)^1$

suposições?

Probabilidade

Probabilidade: definição 1

Se os elementos de Ω são equiprováveis, a probabilidade de um evento A (subconjunto de Ω) é:

$$P(A) = \frac{\text{nº de elementos em A}}{\text{nº de elementos em } \Omega}$$

Ω = {acertar, errar}

A = {acertar}

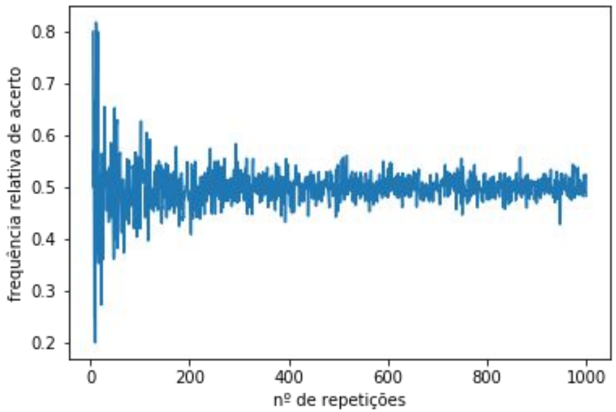
P(A) = ½ ?

Apenas se P(acertar)=P(errar)

Probabilidade: definição 2

Frequência relativa de vezes que ocorre o evento A em infinitas repetições do experimento.

$$P(A) = \lim_{n \to \infty} \frac{\text{nº de vezes que ocorre A}}{n}$$



Independência

Eventos independentes

A_1 e A_2 são eventos independentes se:

$$P(A_1 \cap A_2) = P(A_1).\ P(A_2)$$

Consequência:

A_1^c e A_2 são independentes

A_1^c e A_2^c são independentes

A_1 e A_2^c são independentes

Sob a condição de independência:

$$P(X_1=0, X_2=1, X_3=1, X_4=1) = p^3 .(1-p)^1$$

$$P(A_1^c \cap A_2 \cap A_3 \cap A_4) = P(A_1^c) . P(A_2) . P(A_3) . P(A_4)$$

função de distribuição conjunta

função de distribuição acumulada

Y: número de acertos

y = 0, 1, 2, 3, 4

$$P(Y=3) + P(Y=4) = 1-P(Y\leq 2)$$

Qual a probabilidade de ela acertar 3 ou mais ao acaso?

$$\begin{cases} P(\mathbf{X}=(0,1,1,1)) + P(\mathbf{X}=(1,0,1,1)) + P(\mathbf{X}=(1,1,0,1)) + P(\mathbf{X}=(1,1,1,0)) + P(\mathbf{X}=(1,1,1,1)) \\ = 4 . (0,5)^3 . (0,5)^1 + (0,5)^4 = 0,25 + 0,0625 = 31,25\% \end{cases}$$

Funções de distribuição

X é uma variável aleatória discreta, com possíveis valores em $R_X = \{x_1, x_2, \ldots\}$

f(x) é sua função de probabilidade se:

- $0 \leq f(x_i) \leq 1$, para todo i
- $P(X = x_i) = f(x_i)$, para $x_i \in R_X$
- $\sum_{x_i : f(x_i) > 0} f(x_i) = 1$

A função de distribuição conjunta de uma amostra **X**=(X_1, X_2, ..., X_n) é:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x})$$

para $\boldsymbol{x}$ em $(R_x)^n$

F(x) é sua função de distribuição acumulada é:

- $F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$, em que $x_i \in R_X$

Probabilidade condicional

Se A_1 e $A_2 \subseteq \Omega$, eventos, a probabilidade condicional de A_2 dado que ocorreu A_1 é:

$$P(A_2|A_1) = \frac{P(A_2 \cap A_1)}{P(A_1)}\text{ , se } P(A_1)>0. \Longrightarrow P(A_2 \cap A_1) = P(A_2|A_1).P(A_1)$$

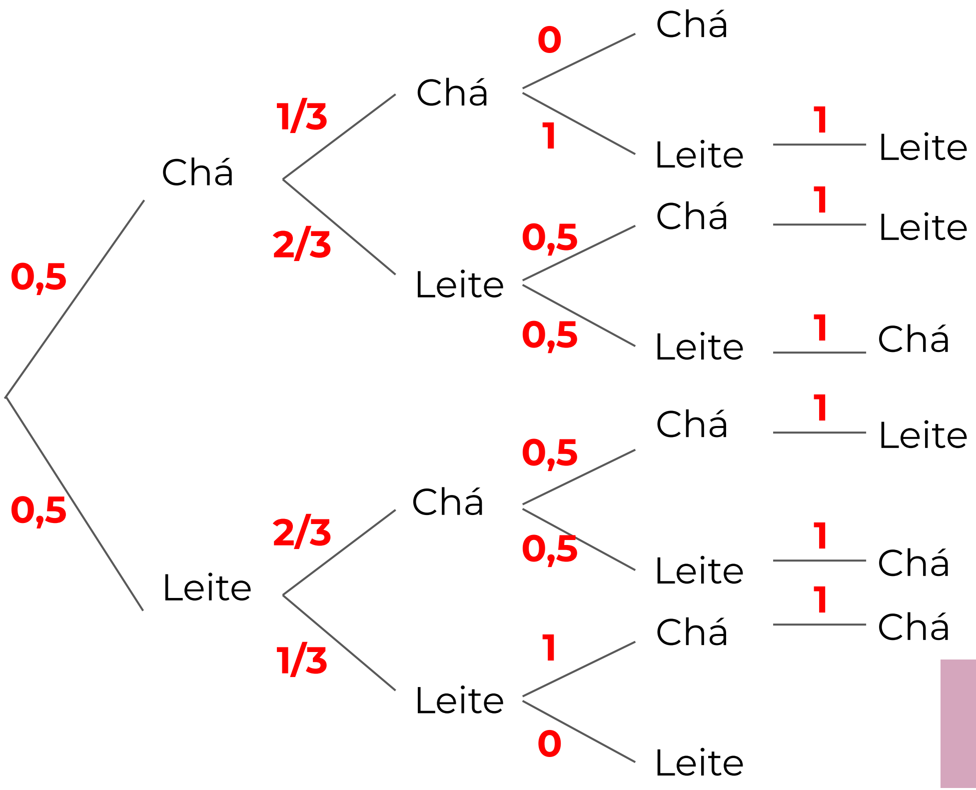
NOTE QUE SE A_1 e A_2 SÃO INDEPENDENTES, ENTÃO:

$P(A_1|A_2)=P(A_1)$ e $P(A_2|A_1)=P(A_2)$

Se 2 xícaras de cada tratamento fossem aleatoriamente servidas à senhora (esta sabendo deste número), comunicando, após cada degustação, qual era a ordem de preparo da bebida.

Probabilidade condicional: árvore de probabilidades

Experimento: Seleção das 4 xícaras de chá



$P(C_1C_2L_3L_4) = P(C_1).P(C_2|C_1).P(L_3|C_1 \cap C_2).P(L_4|C_1 \cap C_2 \cap L_3)$

$P(C_1L_2C_3L_4) = 0{,}5 \,.\, 0{,}67 \,.\, 0{,}5.\ 1 = 0{,}167$

$P(C_1L_2L_3C_4) = 0{,}5 \,.\, 0{,}67 \,.\, 0{,}5 \,.\, 1 = 0{,}167$

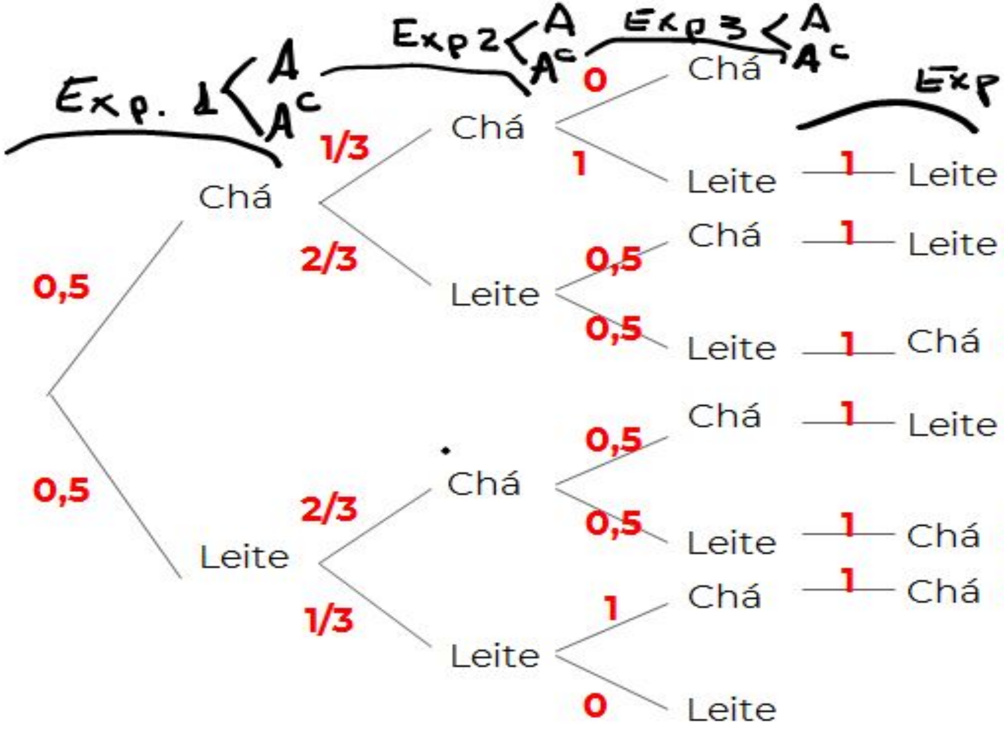
$P(L_1C_2C_3L_4) = 0{,}5 \,.\, 0{,}67 \,.\, 0{,}5 \,.\, 1 = 0{,}167$

$P(L_1C_2L_3C_4) = 0{,}5 \,.\, 0{,}67 \,.\, 0{,}5 \,.\, 1 = 0{,}167$

$P(L_1L_2C_3C_4) = 0{,}5 \,.\, 0{,}33 \,.\, 1 \,.\, 1 = 0{,}167$

Total = 1: soma de eventos mutuamente exclusivos que formam uma partição de Ω

Probabilidade condicional



$P(C_1C_2L_3L_4) = P(C_1).P(C_2|C_1).P(L_3|C_1 \cap C_2).P(L_4|C_1 \cap C_2 \cap L_3)$

$P(C_1L_2C_3L_4) = 0,5 \,.\, 0,67 \,.\, 0,5.\, 1 = 0,167$

$P(C_1L_2L_3C_4) = 0,5 \,.\, 0,67 \,.\, 0,5 \,.\, 1 = 0,167$

$P(L_1C_2C_3L_4) = 0,5 \,.\, 0,67 \,.\, 0,5 \,.\, 1 = 0,167$

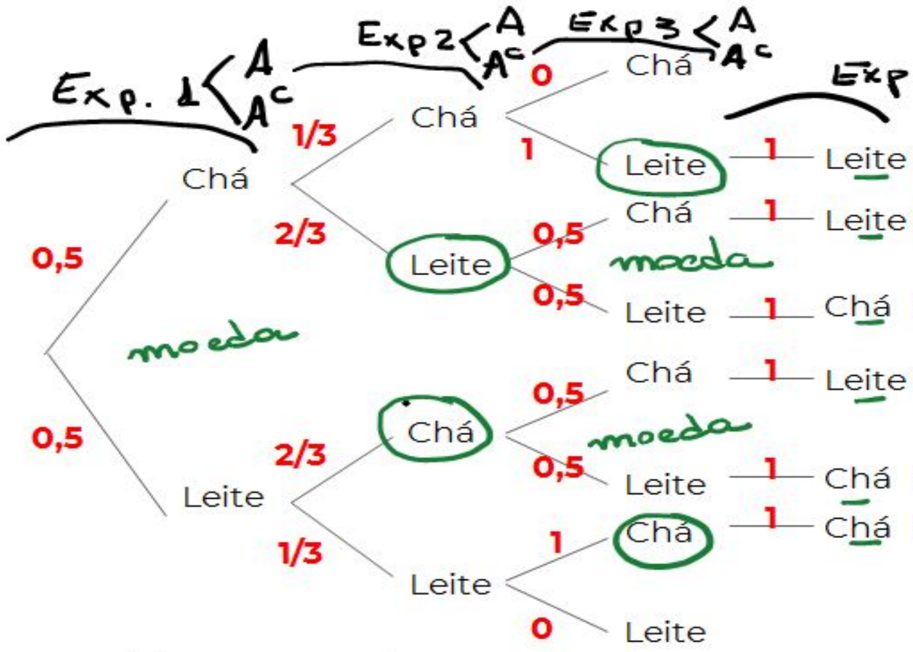
$P(L_1C_2L_3C_4) = 0,5 \,.\, 0,67 \,.\, 0,5 \,.\, 1 = 0,167$

$P(L_1L_2C_3C_4) = 0,5 \,.\, 0,33 \,.\, 1 \,.\, 1 = 0,167$

Estratégia: escolhe a preparação com maior probabilidade de ocorrer.

Probabilidade condicional

escolhas da senhora



$P(C_1C_2L_3L_4) = P(C_1).P(C_2|C_1).P(L_3|C_1 \cap C_2).P(L_4|C_1 \cap C_2 \cap L_3)$

$P(C_1L_2C_3L_4) = 0{,}5 \,.\, 0{,}67 \,.\, 0{,}5 \,.\, 1 = 0{,}167$

$P(C_1L_2L_3C_4) = 0{,}5 \,.\, 0{,}67 \,.\, 0{,}5 \,.\, 1 = 0{,}167$

$P(L_1C_2C_3L_4) = 0{,}5 \,.\, 0{,}67 \,.\, 0{,}5 \,.\, 1 = 0{,}167$

$P(L_1C_2L_3C_4) = 0{,}5 \,.\, 0{,}67 \,.\, 0{,}5 \,.\, 1 = 0{,}167$

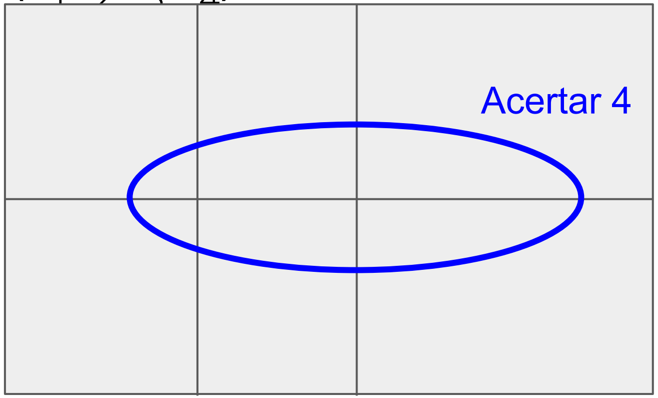
$P(L_1L_2C_3C_4) = 0{,}5 \,.\, 0{,}33 \,.\, 1 \,.\, 1 = 0{,}167$

Estratégia: escolhe a preparação com maior probabilidade de ocorrer.

Decisão: acertar 3 ou mais?

Partição: $(C_1C_2L_3L_4)$, $(C_1L_2C_3L_4)$, $(C_1L_2L_3C_4)$, $(L_1C_2C_3L_4)$, $(L_1C_2L_3C_4)$, $(L_1L_2C_3C_4)$

Ω



Fórmula da Probabilidade Total

Se $(B_1, B_2, B_3, \ldots, B_k)$ uma partição de Ω e $A \subseteq \Omega$, então:

$P(A) = P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + \ldots + P(B_k) \cdot P(A|B_k)$

Fórmula da Bayes

$$P(B_i \mid A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}$$

$$\begin{aligned} P(Y=4) = {} & P((Y=4) \cap (C_1C_2L_3L_4)) + P((Y=4) \cap (C_1L_2C_3L_4)) + P((Y=4) \cap (C_1L_2L_3C_4)) + \\ & P((Y=4) \cap (L_1C_2C_3L_4)) + P((Y=4) \cap (L_1C_2L_3C_4)) + P((Y=4) \cap (L_1L_2C_3C_4)) \\ & = 0 + 0{,}0417 + 0{,}0417 + 0{,}0417 + 0{,}0417 + 0 = 16{,}68\% \end{aligned}$$

“Árvore” de probabilidades

ChaLeite conhecendo resultado - Excel

ARQUIVO PÁGINA INICIAL INSERIR LAYOUT DA PÁGINA FÓRMULAS DADOS REVISÃO EXIBIÇÃO PDF Architect 4 Creator Mariana Curi

A42

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Sai	P(sair)	Fala	P(A)	Sai	P(sair)	Fala	P(A)	Sai	P(sair)	Fala	P(A)	Prob	Y: nº acertos	Y * Prob		Result1	Result2	Result3
2	Chá	0.5	A: fala chá	0.5	Chá	0.33	A: fala chá	0.00	Chá	0			0		0		1	1	
3		0.5		0.5		0.33		0.00		0			0		0		1	1	
4		0.5		0.5		0.33		0.00	Leite	1	A: fala leite	1	0	4	0		1	1	
5		0.5		0.5		0.33		0.00		1	E: fala chá	0	0	3	0		1	1	
6		0.5		0.5		0.33	E: fala leite	1.00	Chá	0			0		0		1	0	
7		0.5		0.5		0.33		1.00		0			0		0		1	0	
8		0.5		0.5		0.33		1.00	Leite	1	A: fala leite	1	0.0833	3	0.2500		1	0	
9		0.5		0.5		0.33		1.00		1	E: fala chá	0	0	2	0		1	0	
10		0.5		0.5	Leite	0.67	A: fala leite	1.00	Chá	0.5	A: fala chá	0.5	0.0417	4	0.1667		1	1	
11		0.5		0.5		0.67		1.00		0.5	E: fala leite	0.5	0.0417	3	0.1250		1	1	
12		0.5		0.5		0.67		1.00	Leite	0.5	A: fala leite	0.5	0.0417	4	0.1667		1	1	
13		0.5		0.5		0.67		1.00		0.5	E: fala chá	0.5	0.0417	3	0.1250		1	1	
14		0.5		0.5		0.67	E: fala chá	0.00	Chá	0.5	A: fala chá	0.5	0	3	0		1	0	
15		0.5		0.5		0.67		0.00		0.5	E: fala leite	0.5	0	2	0		1	0	
16		0.5		0.5		0.67		0.00	Leite	0.5	A: fala leite	0.5	0	3	0		1	0	
17		0.5		0.5		0.67		0.00		0.5	E: fala chá	0.5	0	2	0		1	0	
18		0.5	E: fala leite	0.5	Chá	0.33	A: fala chá	0.00	Chá	0			0		0		0	1	
19		0.5		0.5		0.33		0.00		0			0		0		0	1	
20		0.5		0.5		0.33		0.00	Leite	1	A: fala leite	1	0	3	0		0	1	
21		0.5		0.5		0.33		0.00		1	E: fala chá	0	0	2	0		0	1	
22		0.5		0.5		0.33	E: fala leite	1.00	Chá	0			0		0		0	0	
23		0.5		0.5		0.33		1.00		0			0		0		0	0	

Esperança

Qual o número de acertos esperados, no caso de ela “chutar” a ordem de preparação (p=0,5)?

No caso de independência

(não se revela qual bebida provou)

2?

Y: número de acertos
y = 0, 1, 2, 3, 4

P(Y=3) =0,25

P(Y=4) = 0,0625

P(Y=0) = $(1-0,5)^4$= 0,0625

P(Y=1) = 4. $0,5^1.(1-0,5)^3$ = 0,25

P(Y=2) = 6. $0,5^2.(1-0,5)^2$ = 0,375

Esperança (ou valor esperado)

$$E(Y) = \sum_{i=1}^{n} y_i \cdot f(y_i)$$

= 0 . 0,0625 + 1 . 0,25 + 2 . 0,375 + 3 . 0,25 + 4 . 0,0625 = 2

No caso de dependência

(sabe-se que são 2 de cada tipo e revela-se qual bebida provou, após cada resposta da senhora)

E(Y)=2,8

Esperança

Esperança (ou valor esperado) é a média teórica

$$\overline{Y} = \sum_{i=1}^{n} \frac{Y_i}{n}$$

probabilidades

= 0 . 0,0625 + 1 . 0,25 + 2 . 0,375 + 3 . 0,25 + 4 . 0,0625 = 2

valores que a variável assume

Ex: Uma amostra, $n = 16$

$x_1 = 4$ $x_7 = 1$ $x_{13} = 2$

$x_2 = 1$ $x_8 = 0$ $x_{14} = 3$

$x_3 = 2$ $x_9 = 3$ $x_{15} = 1$

$x_4 = 3$ $x_{10} = 1$ $x_{16} = 2$

$x_5 = 2$ $x_{11} = 2$

$x_6 = 3$ $x_{12} = 2$

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

$$= \frac{0.1 + 1.4 + 2.6 + 3.4 + 4.1}{16}$$

$\frac{1}{16} = 0,625$: freq. relativa de 0 na amostra

Variância

σ^2 ou $V(X) = E[(X-\mu)^2]$, em que $\mu = E[X]$

Se Y = h(X), o valor esperado de Y é: $E[h(X)] = \sum_{x:f(x)>0} h(x).\,f(x)$,
em que f(x) é a função de probabilidade de X.

Assim: $$\begin{aligned} V(X) &= \sum_{x:f(x)>0} (x-\mu)^2.\,f(x) \\ &= \sum_{x:f(x)>0} (x^2 - 2\mu x + \mu^2).\,f(x) \\ &= \sum_x x^2.\,f(x) - 2\mu \sum_x x.\,f(x) + \mu^2 \sum_x f(x) \\ &= E(X^2) - E^2(X) \end{aligned}$$

X variável aleatória discreta

x	$f(x)$
-1	$0{,}3$
0	$0{,}4$
1	$0{,}3$

$E(X) = (-1).0{,}3 + 0.0{,}4 + 1.0{,}3$
$= 0$

Defino $Y = X^2$

$(-1)^2 = 1$
$0^2 = 0$
$1^2 = 1$

$P(Y=0) = 0{,}4$
$P(Y=1) = P(X=-1) + P(X=1)$
$= 0{,}6$

$E(Y) = 0.0{,}4 + 1.0{,}6 = 0{,}6$

ou

$$E(Y) = E(X^2) = \sum_{x: f(x)>0} x^2 . f(x) = (-1)^2.0{,}3 + 0^2.0{,}4 + 1^2.0{,}3$$
$$= 0{,}3 + 0{,}3 = 0{,}6$$

Propriedades da Esperança e da Variância

Se X e Y são variáveis aleatórias e *a* e *b* são números reais:

- $E(a)=a$
- $E(aX \pm b) = aE(X) \pm b$
- $E(aX \pm bY) = aE(X) \pm bE(Y)$
- $V(a)=0$
- $V(aX)=a^2V(X)$
- Se X e Y são **independentes**, então: $V(aX \pm bY) = a^2V(X) + b^2V(Y)$

Covariância e Correlação

A covariância entre duas variáveis aleatórias X e Y é:

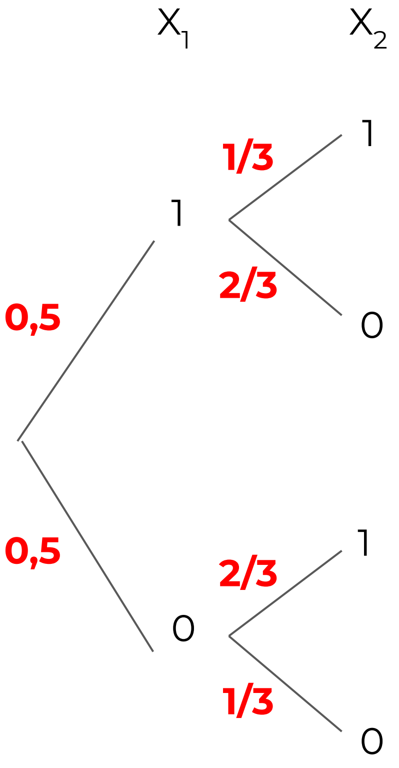
$$\sigma_{XY} \quad \text{ou} \quad Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$$
$$= E(X.Y) - E(X).E(Y)$$

A correlação entre duas variáveis aleatórias X e Y é:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X).V(Y)}} = \frac{\sigma_{XY}}{\sigma_X.\sigma_Y} \qquad -1 \leq \rho_{XY} \leq 1$$

- Se X e Y são **dependentes**, então: $V(X \pm Y) = V(X) + V(Y) \pm 2Cov(X,Y)$

Covariância e Correlação



		X_2		
		0	1	distribuição marginal de X_1
X_1	0	1/6	1/3	1/2
	1	1/3	1/6	1/2
marginal de X_2		1/2	1/2	1

- $E(X_1) = E(X_2) = \frac{1}{2}$
- $VX_1) = V(X_2) = (0-\frac{1}{2})^2 . \frac{1}{2} + (1-\frac{1}{2})^2 . \frac{1}{2} = \frac{1}{4}$
- $Cov\ (X_1,X_2) = E(X_1.X_2) - E(X_1).E(X_2) = \frac{1}{6} - \frac{1}{4} = -1/12$

Covariância e Correlação

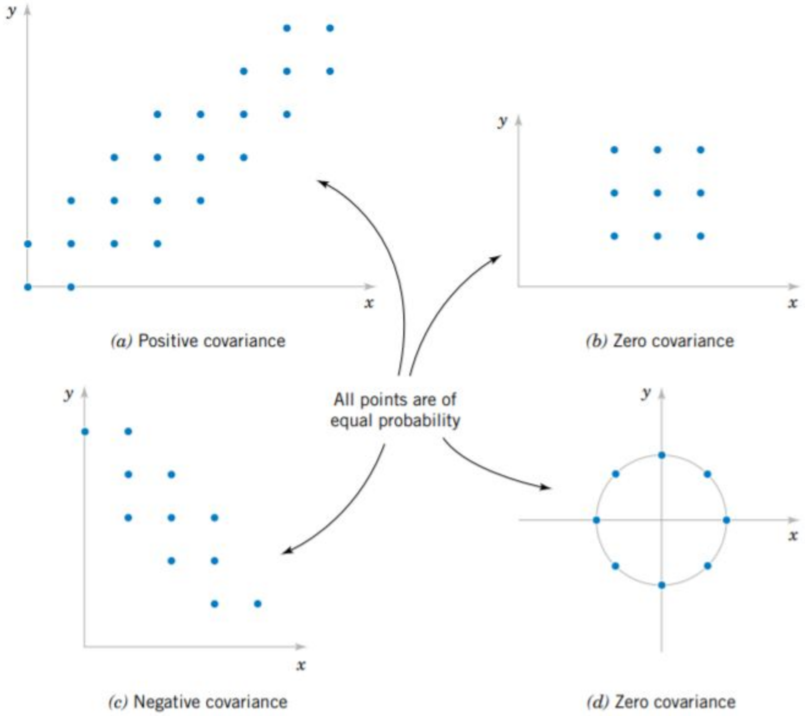


Figure 5-13 Joint probability distributions and the sign of covariance between X and Y.

Se X e Y são **independentes**:

$$\sigma_{XY} = \rho_{XY} = 0$$

$\Rightarrow$

$\cancel{\Leftarrow}$

Fonte: Montgomery, Douglas C., Hubele, Norma Francis, Runger, George C. (2004). **Estatística Aplicada à Engenharia**. 2a edição

Variáveis Contínuas

Definições análogas às apresentadas para variáveis discretas,

com ∑ substituído por ∫. Se X é uma variável aleatória contínua:

- função densidade de probabilidade: $\int_{-\infty}^{\infty} f(x)dx = 1$

  $P(a \leq X \leq b) = \int_a^b f(x)dx$

- função de distribuição acumulada: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$

  do teorema fundamental de Cálculo: $f(x) = \frac{d}{dx} F(x)$

- esperança: $E(X) = \int_{-\infty}^{\infty} x.\, f(x)dx$

Variáveis Contínuas

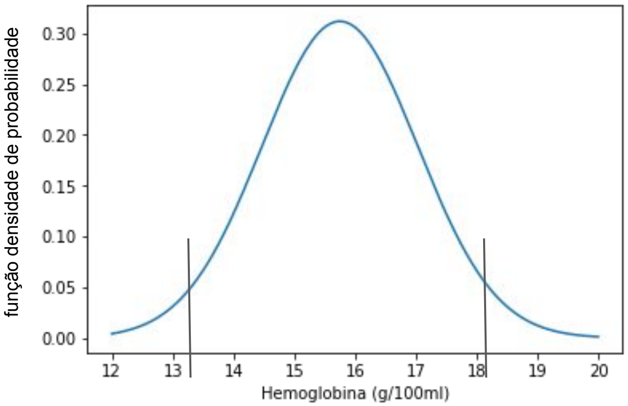
- $E[h(X)] = \int_{-\infty}^{\infty} h(x).\, f(x)dx$
- $V(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2.\, f(x)dx$

$$= E(X^2) - E^2(X)$$

- Cov(X,Y) = E(X.Y)-E(X).E(Y) $= \rho_{XY} \cdot \sqrt{V(X).\,V(Y)}$

Variáveis contínuas: exemplo

Valores de referência para hemoglobina



$$f(x) = \frac{1}{\sqrt{2\pi 1{,}28^2}} exp\{-\frac{(x-15{,}75)^2}{2.\,1{,}28^2}\}dx$$

$$P(13,2 \leq X \leq 18,3) = \int_{13{,}2}^{18{,}3} \frac{1}{\sqrt{2\pi 1{,}28^2}} exp\{-\frac{(x-15{,}75)^2}{8^2}\}dx$$

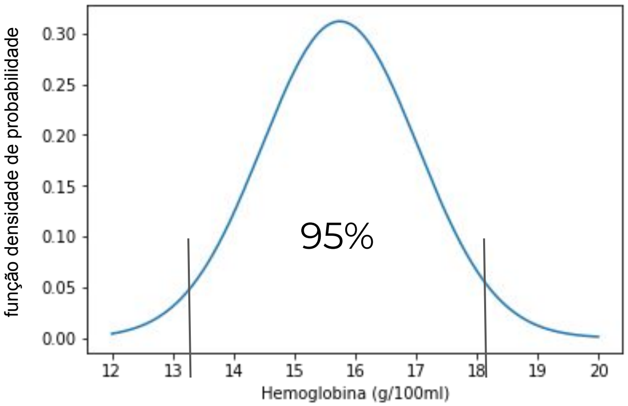
≃ 95%

$$E(X) = \int_{-\infty}^{\infty} x.\, \frac{1}{\sqrt{2\pi 1{,}28^2}} exp\{-\frac{(x-15{,}75)^2}{2.1{,}28^2}\}dx = 15{,}75$$

Intervalo de normalidade:
[13,2; 18,3]

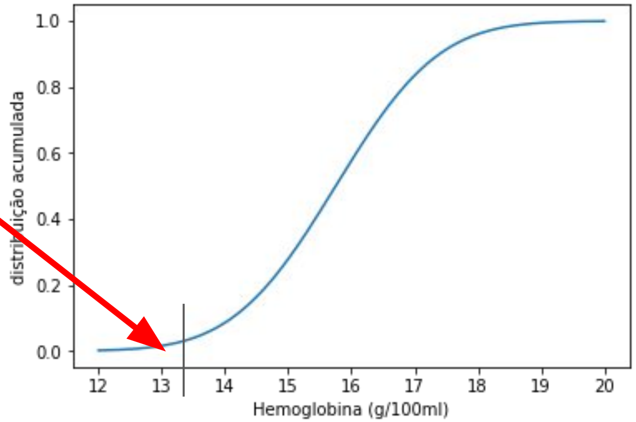
Variáveis contínuas: exemplo

Valores de referência para hemoglobina



$$f(x) = \frac{1}{\sqrt{2\pi 1{,}28^2}} exp\{-\frac{(x-15{,}75)^2}{2.\,1,\,28^2}\}dx$$

2,5%



$$F(13,2) = P(X \leq 13,2) = \int_{-\infty}^{13,2} f(x)dx = 2,5\%$$