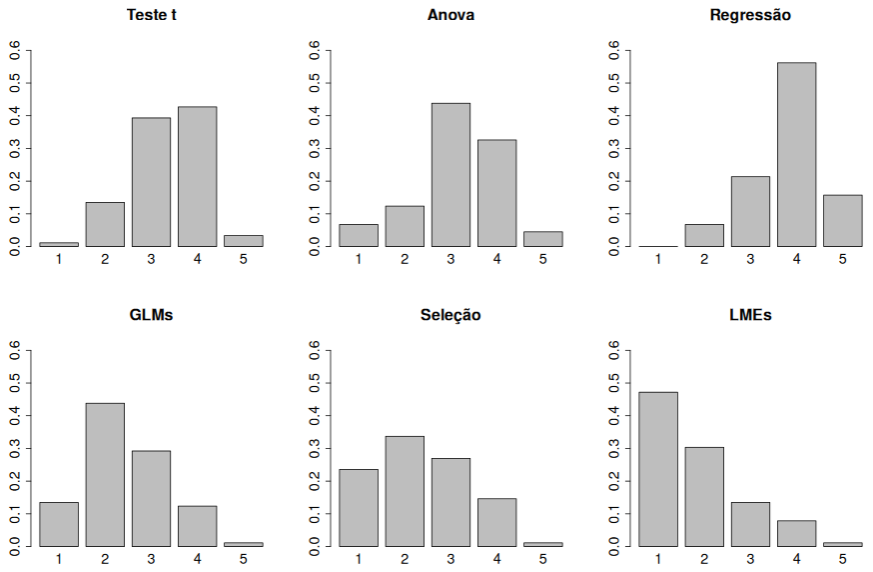
BIE5781
Aula 1

Uma Introdução à Lógica da Modelagem Estatística

Inferência clássica x modelagem



N = 89

Testes de Significância

EXAMPLE 8.1 A two-sample t test for the two-tailed hypotheses, H_0: $\mu_1 = \mu_2$ and H_A: $\mu_1 \neq \mu_2$ (which could also be stated as H_0: $\mu_1 - \mu_2 = 0$ and H_A: $\mu_1 - \mu_2 \neq 0$). The data are human blood-clotting times (in minutes) of individuals given one of two different drugs.

H_0: $\mu_1 = \mu_2$
H_A: $\mu_1 \neq \mu 2$

Given drug B	*Given drug G*
8.8	9.9
8.4	9.0
7.9	11.1
8.7	9.6
9.1	8.7
9.6	10.4
	9.5

$n_1 = 6$	$n_2 = 7$
$\nu_1 = 5$	$\nu_2 = 6$
$\bar{X}_1 = 8.75$ min	$\bar{X}_2 = 9.74$ min
$SS_1 = 1.6950$ min^2	$SS_2 = 4.0171$ min^2

Zar (1999)
Biostatistical Analysis

Testes de Significância

Given drug B	*Given drug G*
8.8	9.9
8.4	9.0
7.9	11.1
8.7	9.6
9.1	8.7
9.6	10.4
	9.5

$n_1 = 6$	$n_2 = 7$
$\bar{X}_1 = 8.75$ min	$\bar{X}_2 = 9.74$ min

Zar (1999)
Biostatistical Analysis

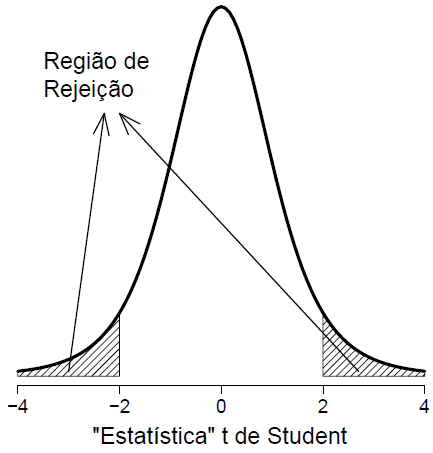
t de Student

William Gosset (1876-1937)

$$t = \frac{\bar{X} - \bar{Y}}{s_{e_{XY}}}$$

$$s_{e_{XY}} = s_{XY}\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}} \qquad s_{XY} = \sqrt{\frac{(n_x - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_x + n_Y - 2}}$$

Distribuição de t sob hipótese nula



Cálculo

```
> X <- c(8.8, 8.4, 7.9, 8.7, 9.1, 9.6)
> Y <- c(9.9, 9, 11.1, 9.6, 8.7, 10.4, 9.5)
> ## sample sizes
> n.X <- length(X)
> n.Y <- length(Y)
> ## Pooled sd
> s.XY <- sqrt( ((n.X-1)*var(X) +
+ (n.Y-1)*var(Y)) / (n.X+n.Y-2) )
> ## Standard error of differences
> se.XY <- s.XY * sqrt(1/n.X + 1/n.Y)
```


Cálculo

```
> #t
> (t.XY <- (mean(X) - mean(Y)) / se.XY)
[1] -2.47649
> ## Degrees of freedom
> (df.XY <- n.X + n.Y - 2)
[1] 11
> ## Bicaudal test
> pt(q = t.XY, df = df.XY) * 2
[1] 0.0307649
```

Cálculo

```
> t.test(X, Y, var.equal=TRUE)

    Two Sample t-test

data:  X and Y
t = -2.4765, df = 11, p-value = 0.03076
alternative hypothesis: true difference in means is not equal
to 0
95 percent confidence interval:
 -1.8752609 -0.1104534
sample estimates:
mean of x mean of y
```


Quais os modelos implícitos no Teste t ?

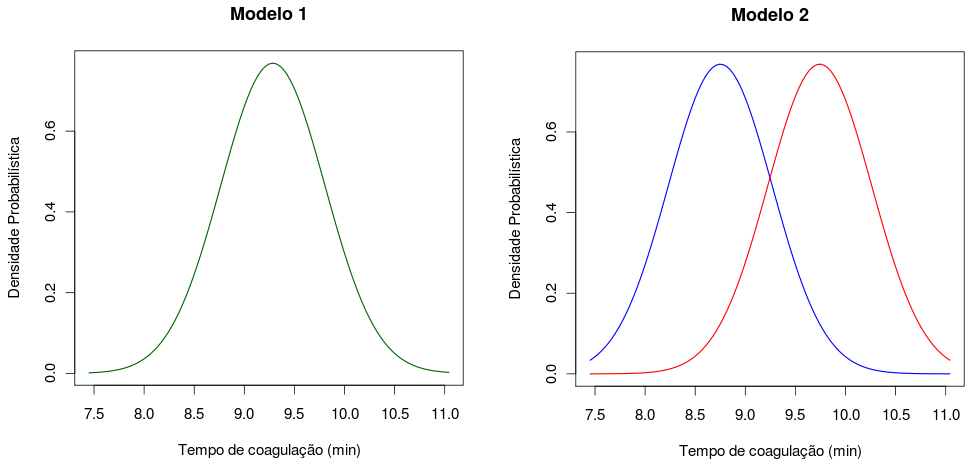
A distribuição de t Student é a distribuição de probabilidade da estatística t sob a hipótese nula de que as duas médias amostrais vêm da mesma distribuição se:

1 - A observações que compõem as amostras são independentes;

2 - As amostras foram tomadas de distribuições Gaussianas com igual variância*.

* Há uma aproximação para variâncias diferentes

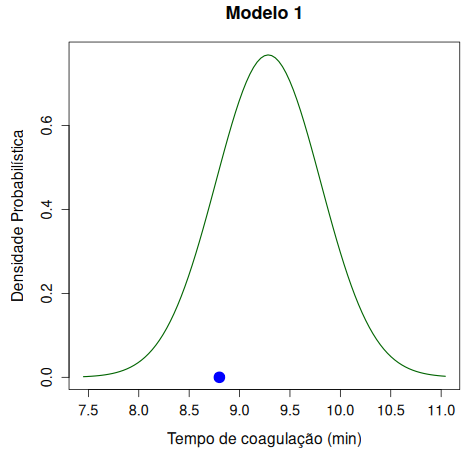
Modelos implícitos no Teste t

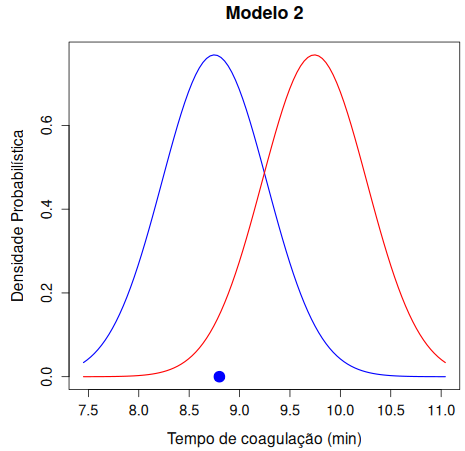


Parâmetros: μ, σ

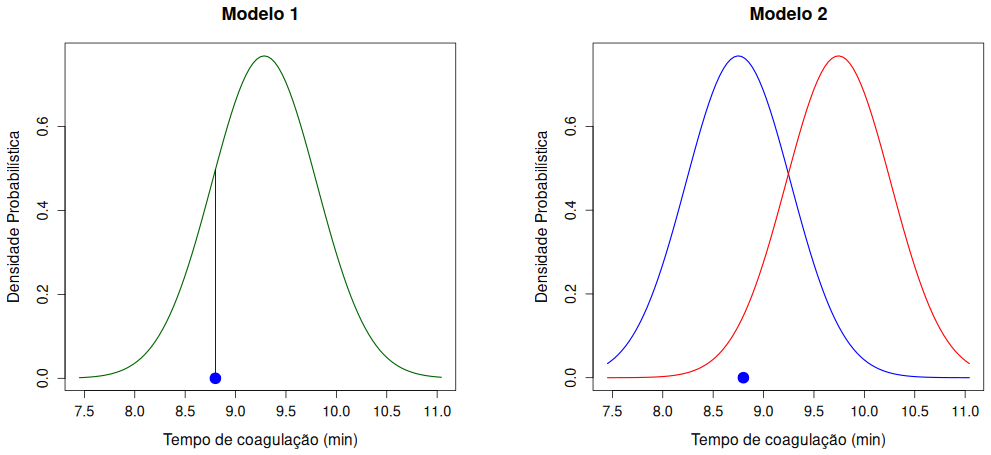
Parâmetros: μ_1, μ_2, σ

Modelos estatísticos atribuem probabilidades a observações



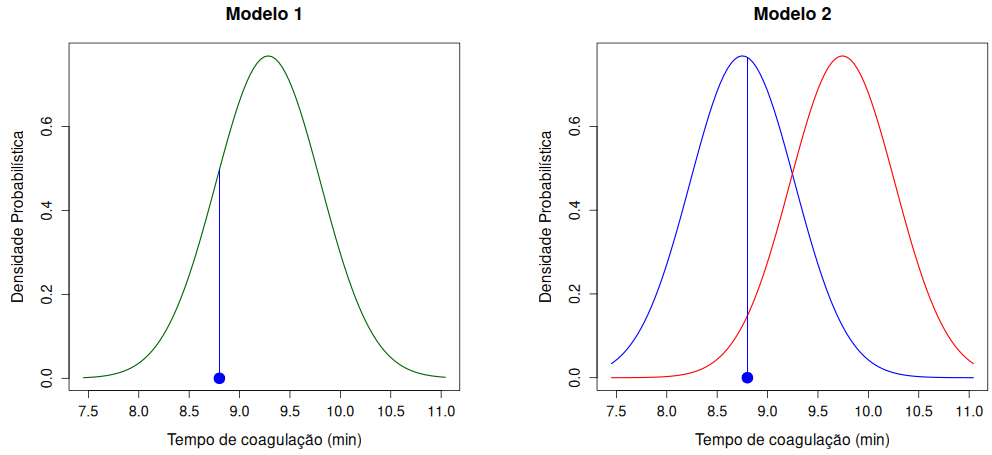


Probabilidade atribuída a uma observação por cada modelo



Modelos estatísticos em ecologia e recursos naturais

Probabilidade atribuída a uma observação por cada modelo



PG-Ecologia-USP Paulo Inácio Prado & João L. F. Batista

Modelos estatísticos em ecologia e recursos naturais

Lei da Verossimilhança (Um Enunciado Informal)

Dado que:

- Há mais de um modelo para um conjunto de dados.
- Cada modelo atribui uma probabilidade diferente aos dados.

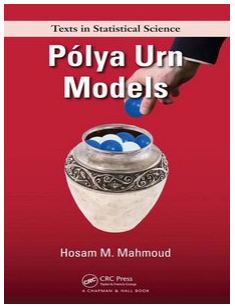
Então:

O MODELOS MAIS PLAUSÍVEL SERÁ AQUELE QUE ATRIBUIR A MAIOR PROBABILIDADE AOS DADOS.

PG-Ecologia-USP Paulo Inácio Prado & João L. F. Batista

Modelos estatísticos em ecologia e recursos naturais

O que é Verossimilhança?



Observação:

N de bolas brancas em um sorteio

Hipóteses (ou modelos)

H_1: Há apenas bolas brancas na urna.

H_2: Metade das bolas da urna são brancas e metade são azuis.

Problema:

Identificar a hipótese mais plausível, **dada a observação**.

PG-Ecologia-USP Paulo Inácio Prado & João L. F. Batista

Força da Evidência de uma observação

- Um sorteio de uma bola.
- A bola sorteada é branca.

$$P(x=1|H_1)=1,0$$
$$P(x=1|H_2)=0,5$$

H_1 é $\frac{1,0}{0,5}=2$ vezes mais plausível que H_2

E para observações múltiplas?

- Dois sorteios de uma bola cada.
- Em ambos tivemos uma bola branca.

$$P(x_1=1, x_2=1|H_1)=1,0\times1,0=1,0$$
$$P(x_1=1, x_2=1|H_2)=0,5\times0,5=0,25$$

H_1 é $\frac{1,0}{0,25}=4$ vezes mais plausível que H_2

Função de Verossimilhança

Qualquer função proporcional ao produto das probabilidades que um modelo atribui a cada valor dos dados*

$$L \propto P(x_1|H)\times P(x_2|H)\times \ldots P(x_n|H)$$

* Sob a premissa de que os dados são realizações independentes de um mesmo processo.

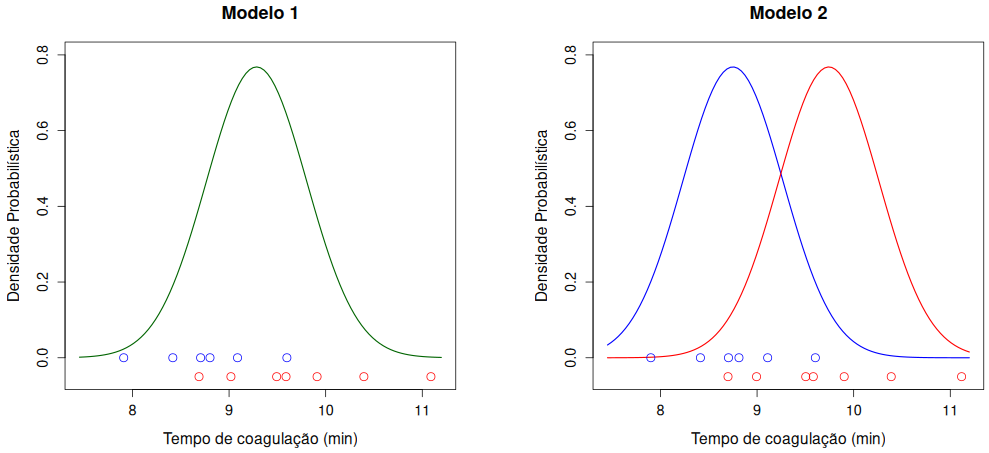
Função de Log-Verossimilhança

Logaritmo de uma função de verossimilhança, ou seja:
qualquer função proporcional à soma dos logaritmos das probabilidades que um modelo atribui a cada valor dos dados*

$$\mathrm{LL} \propto \ln P(x_1|H) + \ln P(x_2|H) + \ldots \ln P(x_n|H)$$

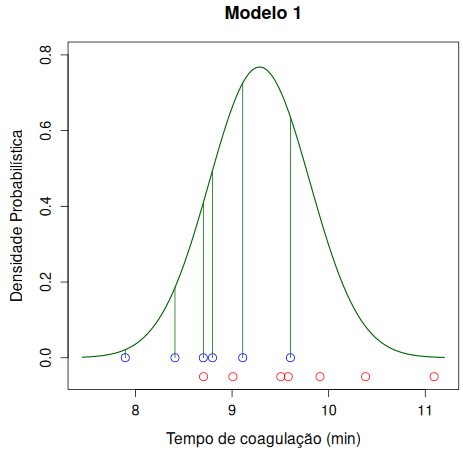
* Sob a premissa de que os dados são realizações independentes de um mesmo processo.

Várias observações, dois modelos



Modelos estatísticos em ecologia e recursos naturais

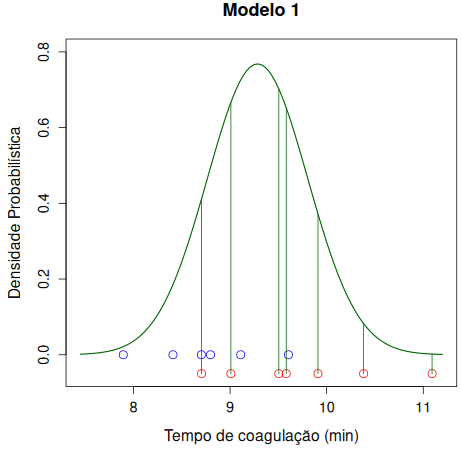
Várias observações, modelo 1



PG-Ecologia-USP Paulo Inácio Prado & João L. F. Batista

Modelos estatísticos em ecologia e recursos naturais

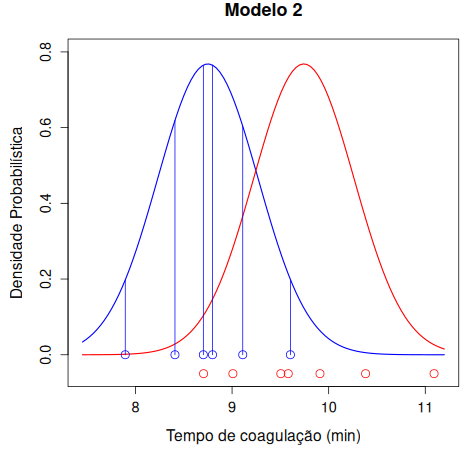
Várias observações, modelo 1



PG-Ecologia-USP Paulo Inácio Prado & João L. F. Batista

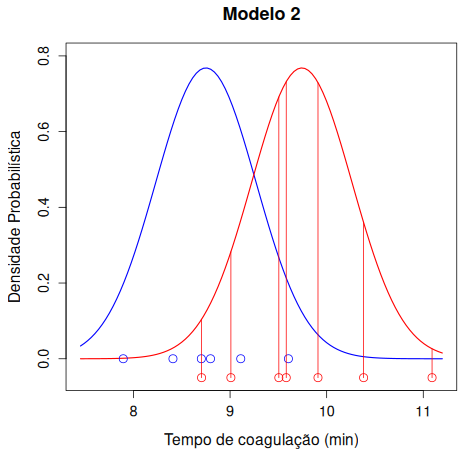
Modelos estatísticos em ecologia e recursos naturais

Várias observações, modelo 2

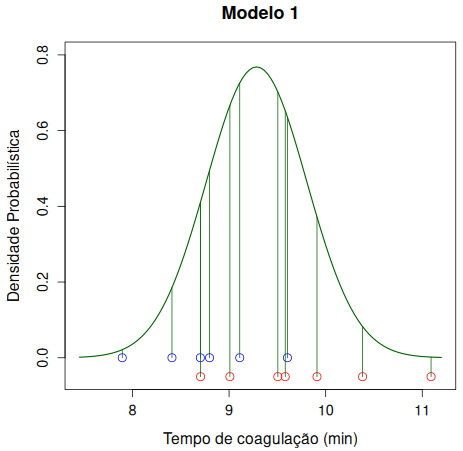


PG-Ecologia-USP Paulo Inácio Prado & João L. F. Batista

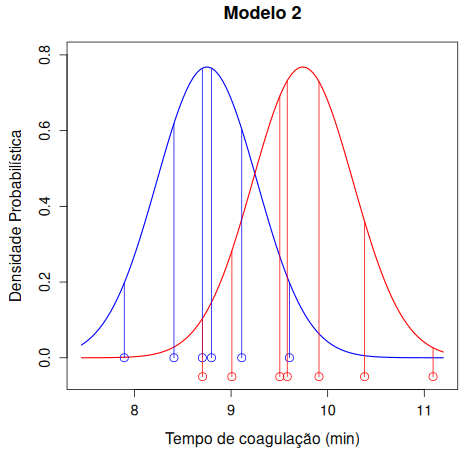
Várias observações, modelo 2



Várias observações, dois modelos



Log-Verossimilhança = -19,9



Log-Verossimilhança = -14,1

Uma medida de plausibilidade

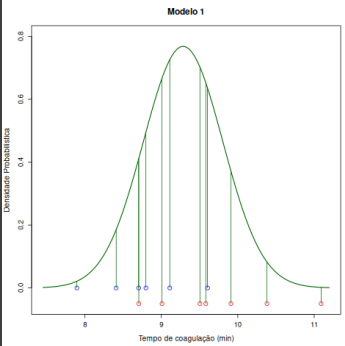
SE:

1. Temos dados que podem ser explicados por mais de uma hipótese, e
2. Cada hipótese é um modelo que atribui alguma probabilidade aos dados

ENTÃO:

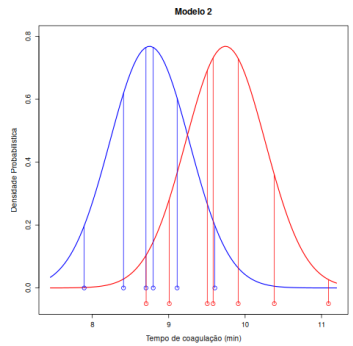
Podemos expressar o quão plausível uma hipótese é em relação às outras por meio de uma função, chamada verossimilhança (ou pelo seu logaritmo, chamada função de log-verossimilhança)..

Várias observações, três modelos



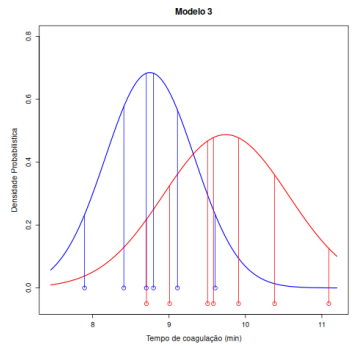
LL = -19,9

μ, σ



LL = -14,1

μ_1, μ_2, σ



LL = -12,8

μ_1, μ_2, σ_1, σ_2

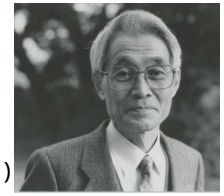
Parcimônia!

MODELO	Parâmetros	LL
H1	μ, σ	-19,9
H2	μ_1, μ_2, σ	-14,1
H3	μ_1, μ_2, σ_1, σ_2	-12,8

AIC

AIC = - 2 x Log-Verossimilhança + 2 x n de parâmetros

MODELO	Parâmetros	LL	AIC
H1	μ, σ	-19,9	43,8
H2	μ_1, μ_2, σ	-14,1	34,2
H3	μ_1, μ_2, σ_1, σ_2	-12,8	33,6



Hirotsugu Akaike (1927-2011)

Os 3 passos da Inferência Baseada em Modelos

1. **ESPECIFICAÇÃO**: defina os modelos concorrentes.
2. **ESTIMAÇÃO**: busque o melhor ajuste de cada modelo (combinação de parâmetros que maximiza a verossimilhança).
3. **SELEÇÃO**: Fique com o melhor modelo (com maior verossimilhança).

RECAPITULANDO

- Modelos probabilísticos descrevem a probabilidade de que seu ensaio tenha um certo resultado.
- Uma vez observado o resultado de um ensaio, encontramos a verossimilhança máxima de cada modelo proposto.
- Usamos a verossimilhança, ou uma função dela, para identificar o(s) modelo(s) mais plausível(is).

Para saber mais

Burnham, K. P., & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical-Theoretic Approach, 2nd ed. New York, Springer-Verlag.

Bolker, B. (2008). Ecological Models and Data in R. Princeton, Princeton University Press.

Hilborn, R. & Mangel, M. (1997). The Ecological Detective – Confronting Models with Data. Princeton, Princeton University Press.

Royall, R. M. (2000). Statistical Evidence: A Likelihood Paradigm. London, Chapman and Hall.

Nossa cadeia de dependências

- Seleção de modelos
  - Definição dos modelos concorrentes
    - Conhecer diferentes classes de modelos
  - Ajuste de cada modelo aos dados
    - Função de verossimilhança de cada modelo
      - Distribuições de probabilidade assumidas para cada modelo

Nosso roteiro

1) Introdução (esta aula)
2) Distribuições de probabilidade
   a) Contínuas
   b) Discretas
3) Verossimilhança
4) Modelos estatísticos
   a) Parâmetros constantes
   b) Gaussianos
   c) Não Gaussianos
5) Seleção de modelos

CONCEITOS-CHAVE

- **Ensaios**: procedimentos de geração de dados (e.g. experimentos ou amostragens).
- **Cenário**: situação na qual podem ser conduzidos ensaios.
- **Cenário estocástico**: aquele em que ensaios têm mais de um resultado possível, cada um com uma certa chance de ocorrer.
- **Modelo estocástico**: construto matemático que simula os resultados possíveis de um ensaio em um cenário estocástico.