Inferência estatística para ciência de dados

Aula 01

Paulo Justiniano Ribeiro Jr

Curso de Especialização em
Data Science & Big Data
Universidade Federal do Paraná

10 de março de 2018

Começando …do começo

Estatística?

OK,
estatística é parte fundamental de DSBD …mas …
Que palavras/idéias *voce* associa com estatística?

Definição

Estatística é a ciência de coletar e interpretar dados …
Dealing with uncertainty is the cornerstone of statistical method.

Diggle & Chetwynd

Definição

Estatística é a ciência de coletar e interpretar dados …
Dealing with uncertainty is the cornerstone of statistical method.

Diggle & Chetwynd

Relevante em praticamente todas áreas do conhecimento.

Papel o processo dedutivo/indutivo do conhecimento

Maths *vs* Stats

Técnicas e idéias: o propósito da aula/curso

More important than learning a few methods and techniques is to understand the statistical thinking.

Box, Hunter & Hunter

Técnicas e idéias: o propósito da aula/curso

More important than learning a few methods and techniques is to understand the statistical thinking.

Box, Hunter & Hunter

Relevante em praticamente todas áreas do conhecimento.

Papel no processo dedutivo/indutivo do conhecimento

Maths *vs* Stats

A necessidade de estatística

- Dois pontos determinam uma reta! …mesmo?
- e se adicionarmos um terceiro?
- no mundo real pode não seguir padrão: imprevisibilidade ou erro?
- erro ou desvio?
- a natureza do erro

Estatística e o método científico

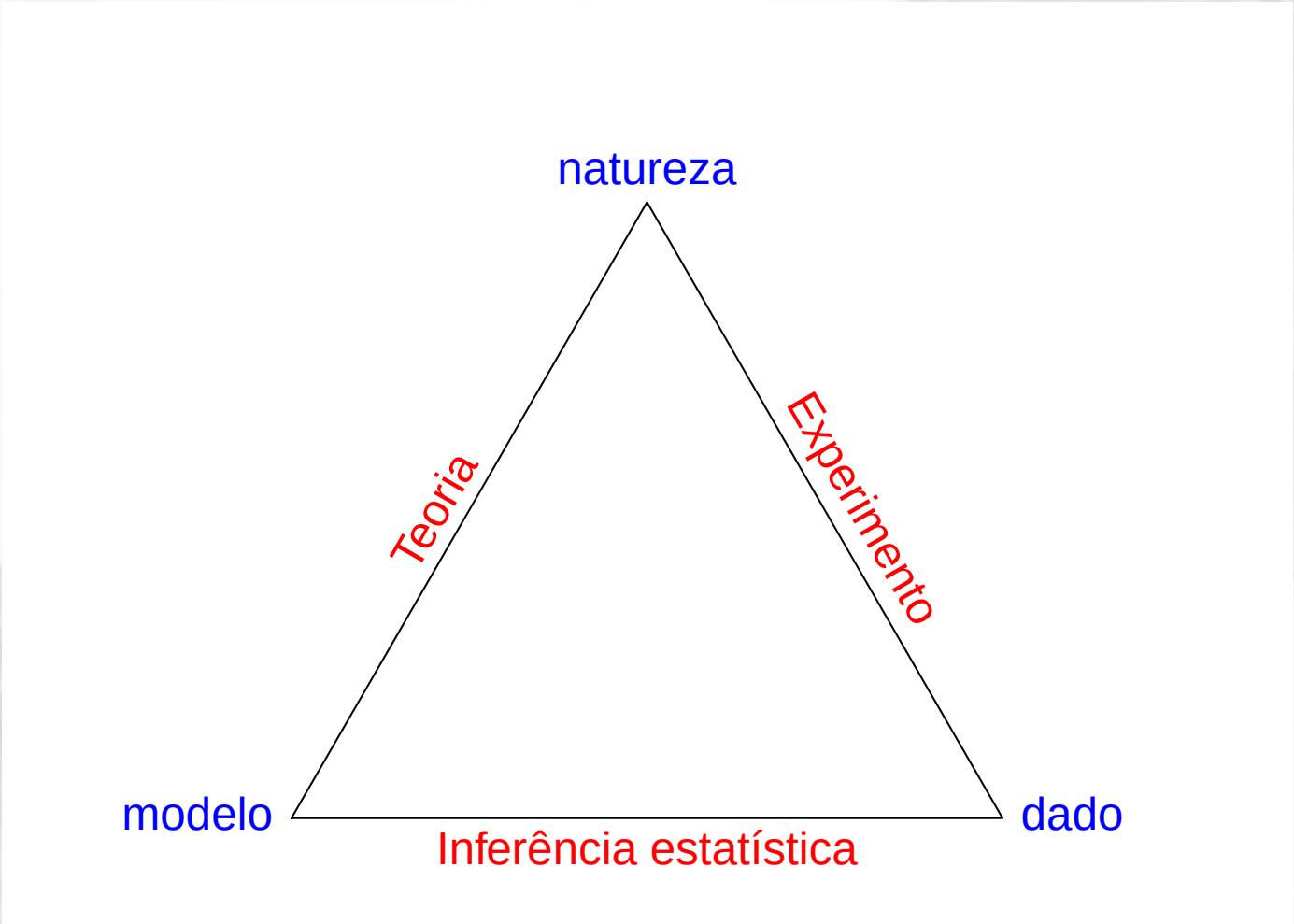


Figura 1. Estatística e o método científico.

Princípios e idéias

- Qual a questão sobre a natureza que queremos investigar?
- Pode ser respondida por dados?
- Que experimento possível pode fornecer os dados?
- Como podemos aprender com os dados sobre a questão de interesse?
- Existe incerteza? qual sua origem? como mensurar?
- Como avaliar o que foi feito?

Conceitos básicos em um exemplo *simples*

Dados...

```
##    d     t
## 1 10 0.241
## 2 40 0.358
## 3 70 0.460
## 4 10 0.249
## 5 45 0.395
## 6 75 0.485
```

Vamos (como?) analisar!!

Mas espere um pouco …

Antes de cofar no método para analisar os dados …

- Por que queremos analisar os dados?
- Que dados queremos ou podemos coletar?

Mas espere um pouco …

Antes de cofar no método para analisar os dados …

- Por que queremos analisar os dados?
- Que dados queremos ou podemos coletar?

O contexto dos dados acima: estimar g em:

$$d = \frac{1}{2}gt^2$$

Delineamento

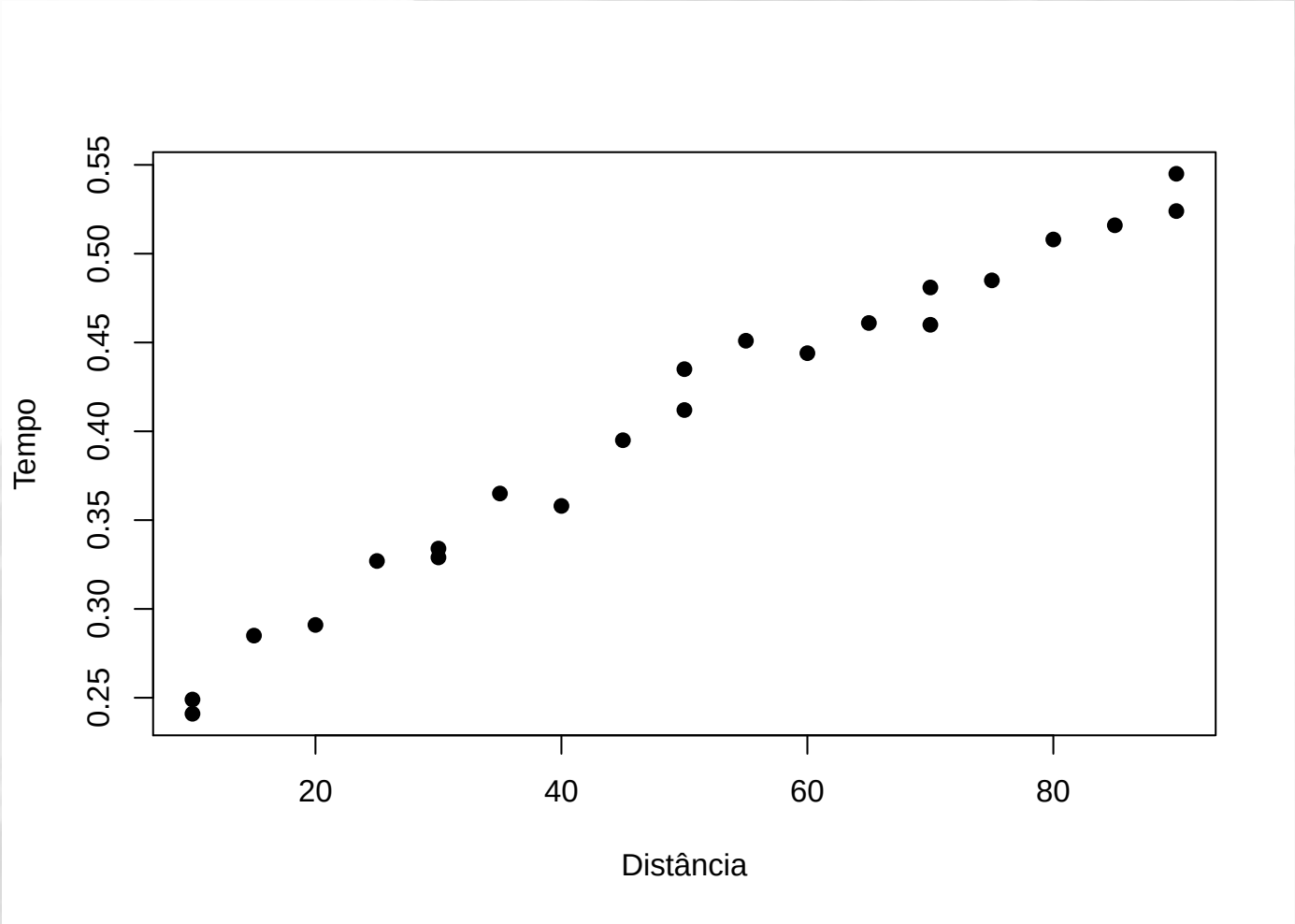
- Garantir validade e depois eficiência
- o que fixar e o que medir (input/output)?
- pontos para *d* e variação sistemática
- tempo de reação e variação imprevisível ou aleatória
- variação de interesse ou exógena (dependendo do contexto)
- variação aleatória:
  - eliminar ou reduzir
  - aleatorização para proteger validade (mas não eficiência …)
  - réplicas: permitem separar aleatório e sistemático

Voltando ao experimento

Como descrever a relação entre distância e tempo?
Que gráfico fazer?

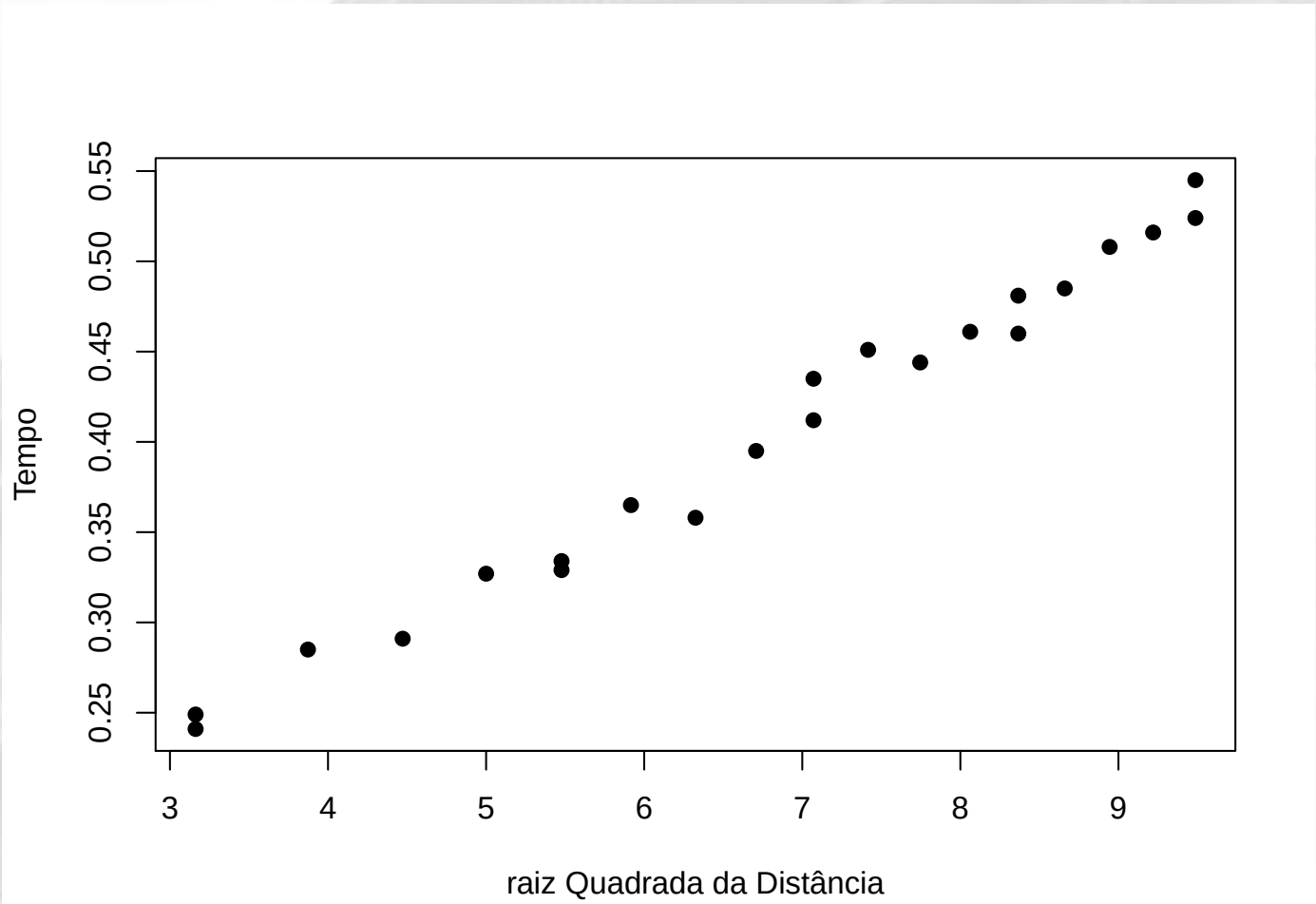
Voltando ao experimento

Como descrever a relação entre distância e tempo?
Que gráfico fazer?

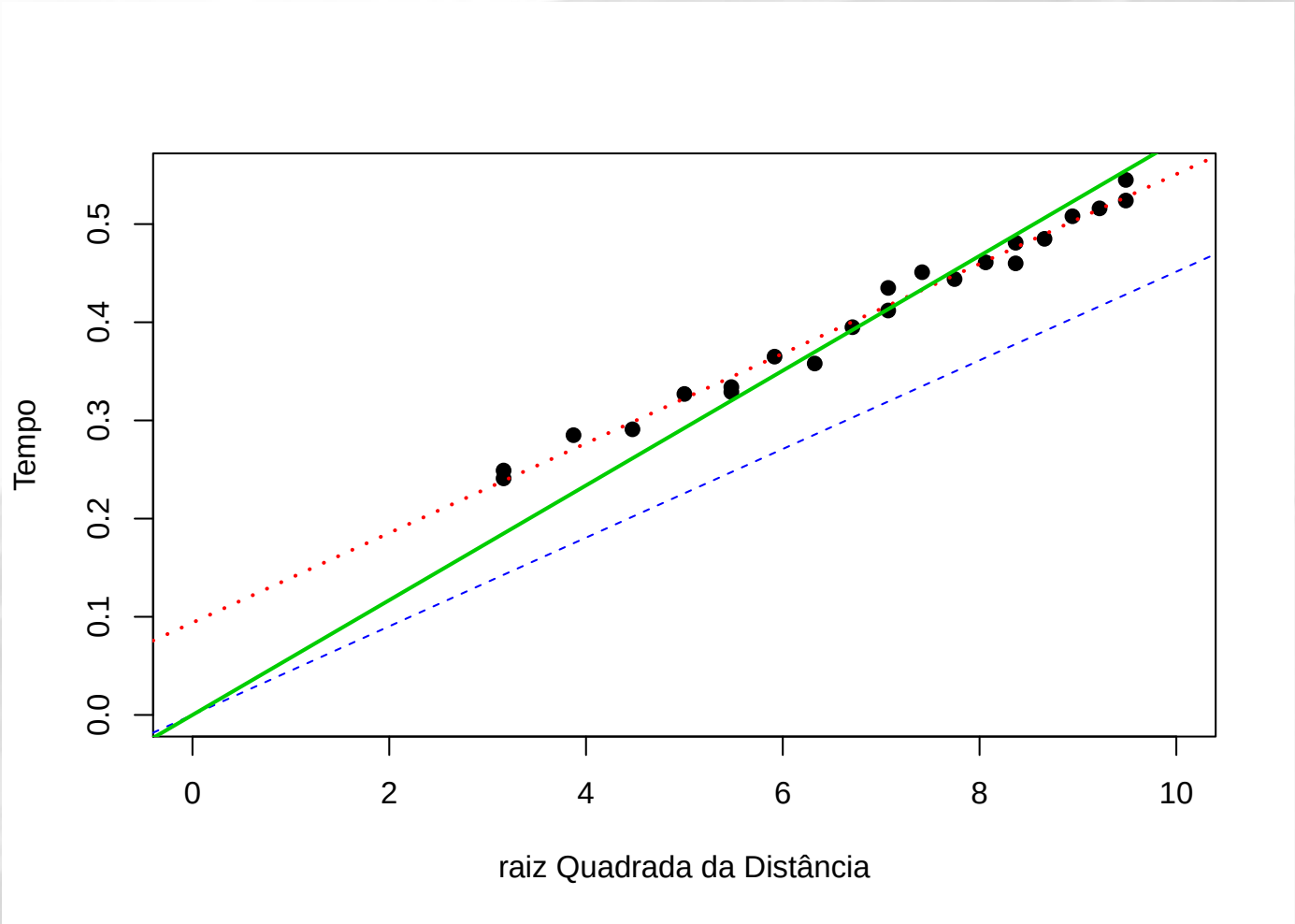


Um outro gráfico

$$t = \sqrt{\frac{2}{g}}\sqrt{d} \longrightarrow y = \beta x$$



$y = 0$ quando $x = 0$?



E aí Newton???

Modelo estatístico para o experimento

Começamos com um modelo mecanístico

$$d = \frac{1}{2} g t^2$$

Modelo estatístico para o experimento

Começamos com um modelo mecanístico

$$d = \frac{1}{2} g t^2$$

Denotamos inputs por x e outputs por Y.
Temos no caso que $x = \sqrt{d}$ e $Y = t$.
Denotamos o parâmetro desconhecido por $\beta = \sqrt{2/g}$.

$$Y = \beta x$$

Modelo estatístico para o experimento

Começamos com um modelo mecanístico

$$d = \frac{1}{2} g t^2$$

Denotamos inputs por x e outputs por Y.
Temos no caso que $x = \sqrt{d}$ e $Y = t$.
Denotamos o parâmetro desconhecido por $\beta = \sqrt{2/g}$.

$$Y = \beta x$$

Adicionamos a variação exógena α e a aleatória ϵ.
Incluímos índices para apontar o que varia e o que é constante.

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

Modelo estatístico para o experimento

Começamos com um modelo mecanístico

$$d = \frac{1}{2} g t^2$$

Denotamos inputs por x e outputs por Y.
Temos no caso que $x = \sqrt{d}$ e $Y = t$.
Denotamos o parâmetro desconhecido por $\beta = \sqrt{2/g}$.

$$Y = \beta x$$

Adicionamos a variação exógena α e a aleatória ϵ.
Incluímos índices para apontar o que varia e o que é constante.

$$Y_i = \alpha + \beta x_i + \epsilon_i$$

Cada termo pode ser interpretado.

O que temos até aqui?

- Gráficos, em geral, são úteis!
- Modelos
  - respeitando dados
  - respeitando o conhecimento contextual
- Dados podem/devem ser transformados em certos casos
- Resultado deve ser interpretável em relação à questão original

Método estatístico

Delineamento/aquisição de dados:
como o dado deve ser obtido para tratar a questão de interesse

Modelagem:
como a variação nos dados pode ser descrita matematicamente de tal forma que:

não seja inconsistente com os dados
incorpore o conhecimento contextual ao ponto que é disponpivel
seja o mais simples possível,
respeitando as condições acima

Inferência:
O que se pode dizer e concluir a partir dos dados e modelos acima?

Inferência

Parâmetro: constante desconhecida e de interesse

$$\beta \text{ e } g = 2/\beta^2$$

Estimativa pontual:

$$\hat{\beta} = 0.0457 \text{ e } \hat{g} = 958,32$$

Intervalo de confiança: conjunto de valores razoáveis sobre parãmetro, expressando incerteza

$$\beta : (0,0431; 0,0482) \text{ e } g : (860,9; 1076.7)$$

Teste de hipótese:

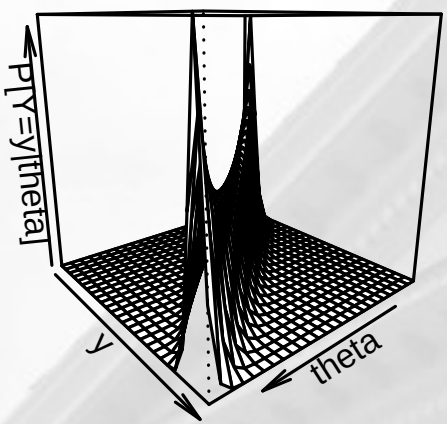
Hipótese: declaração sobre parâmetro
teste estatístico: confronto dos dados com a declaração na hipótese

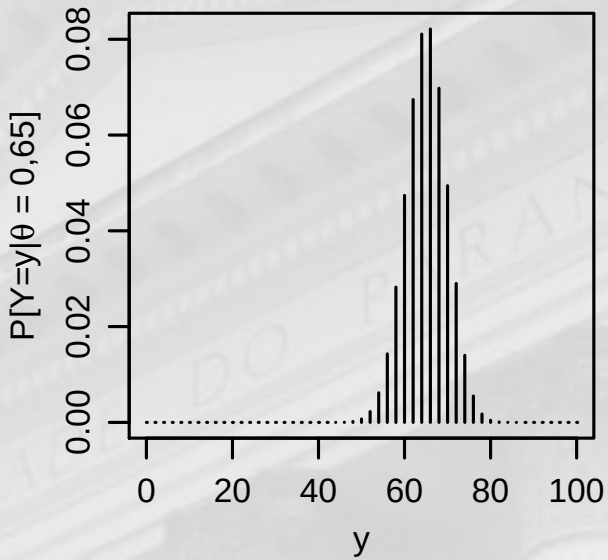
Predição: uso do modelo além dos dados

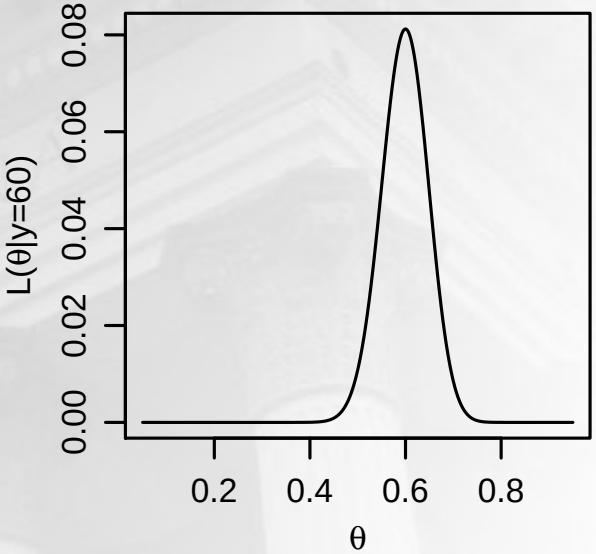
Como proceder inferência?

Como aprender com os dados?

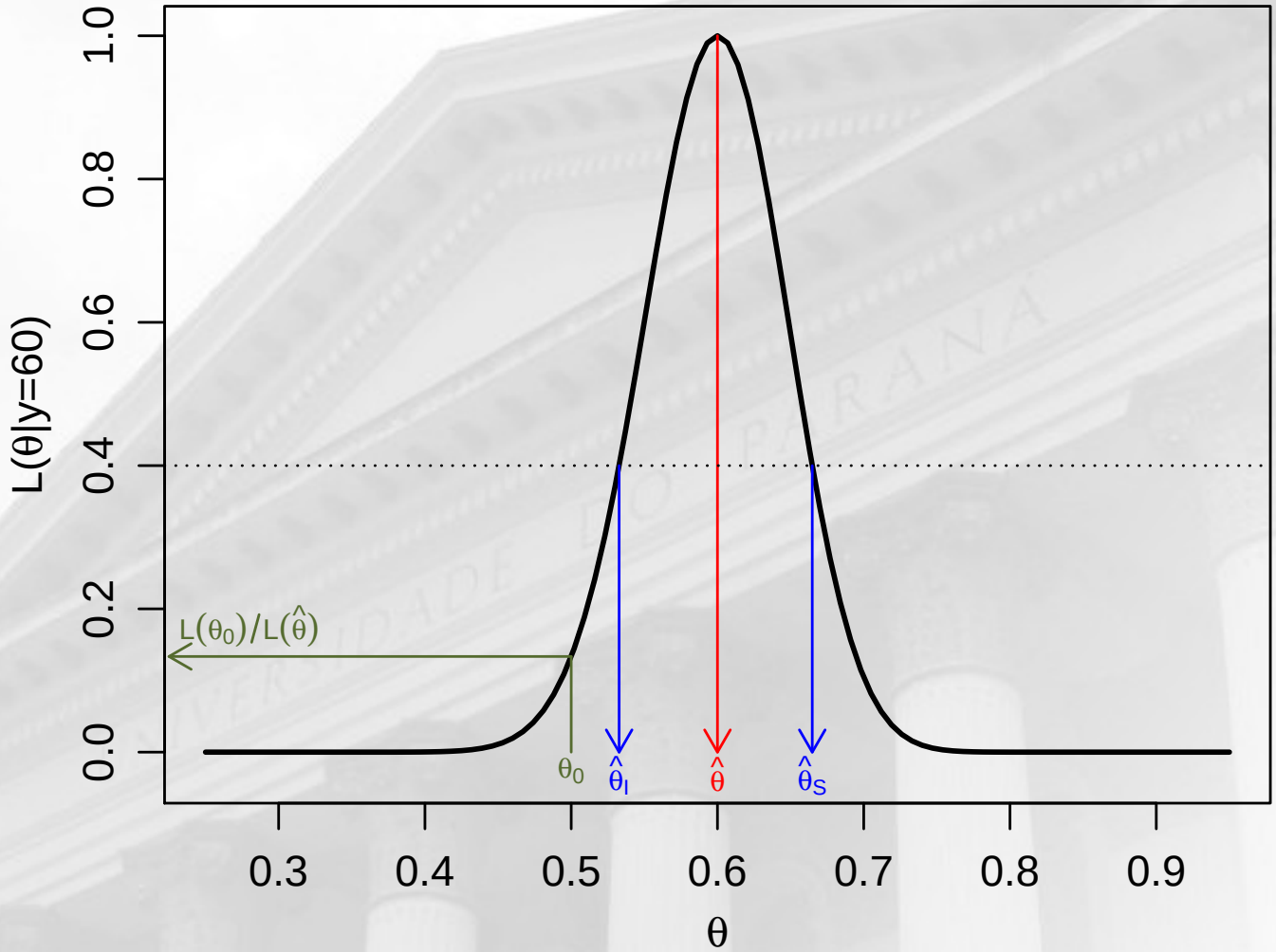
O mundo estatístico



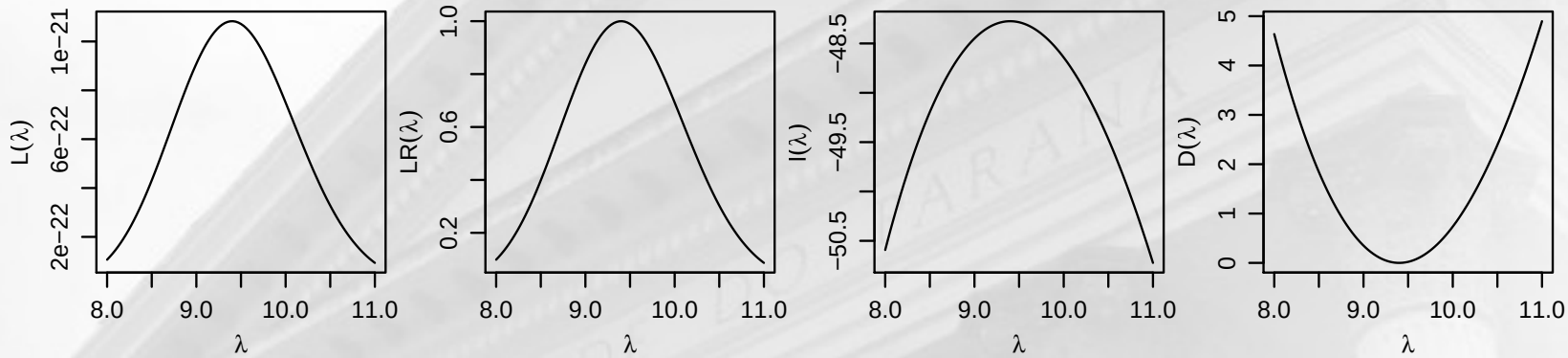




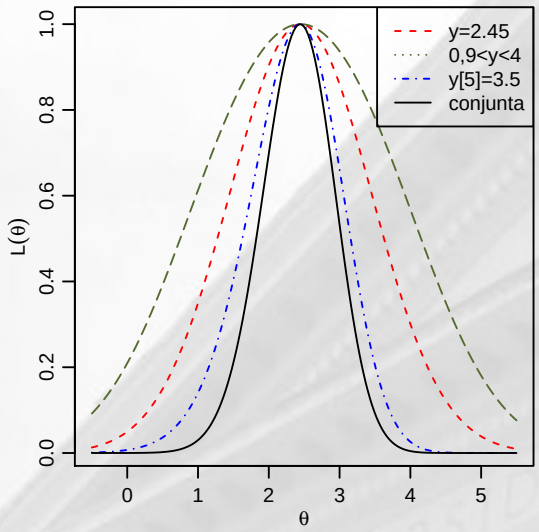
All we need is …likelihood

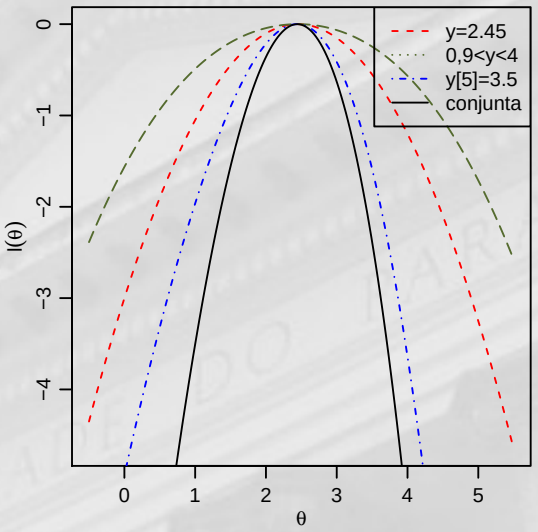


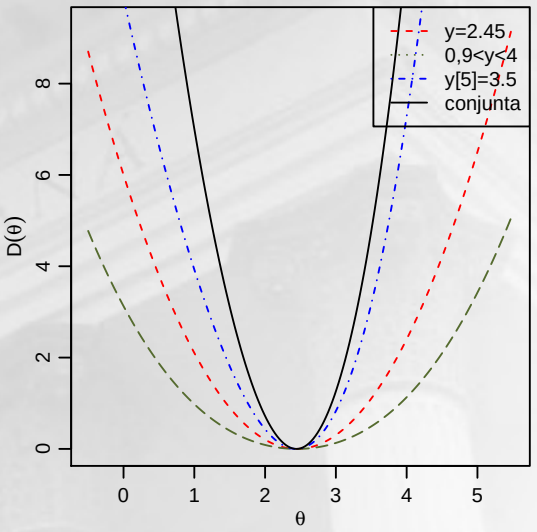
Diferentes roupagens



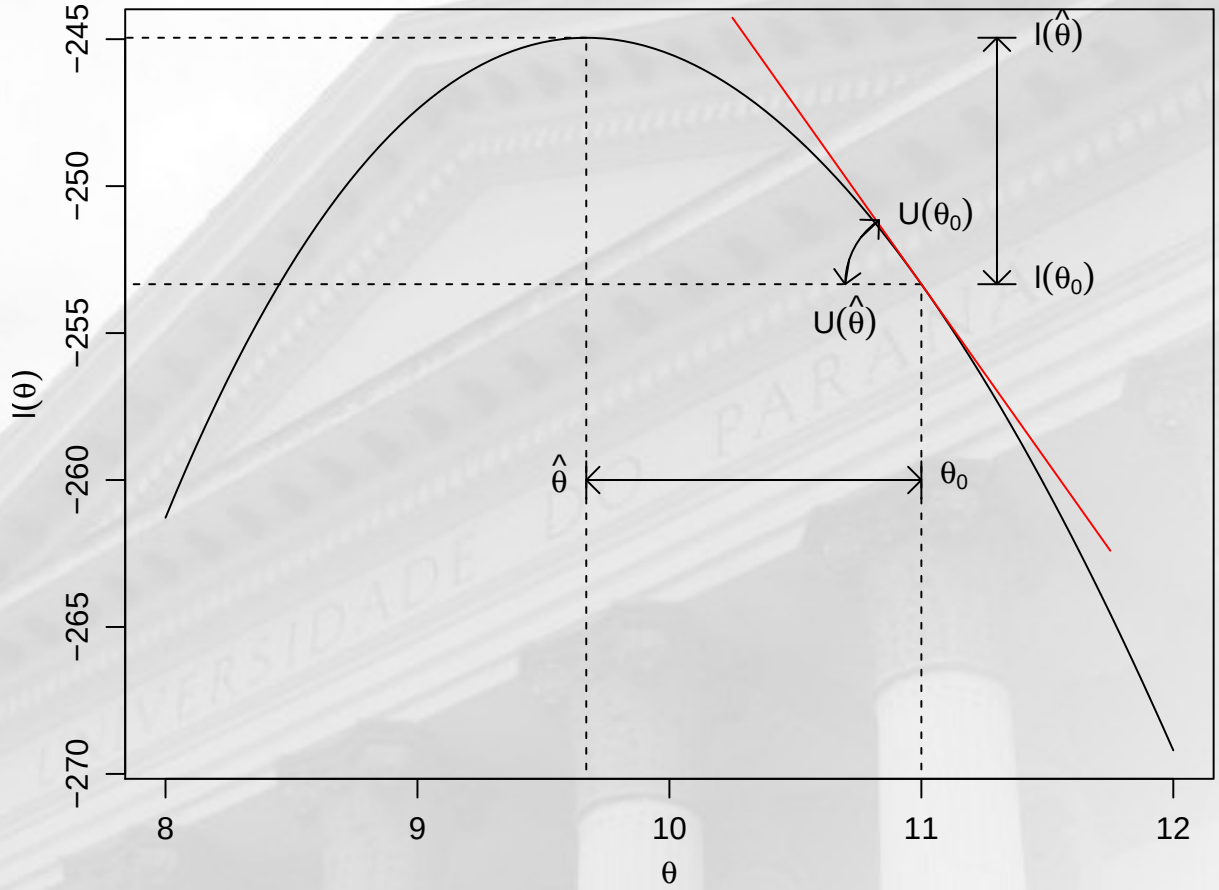
Testes à vontade







Informação de cada dado



Decidindo …