Estatística para Ciências de Dados

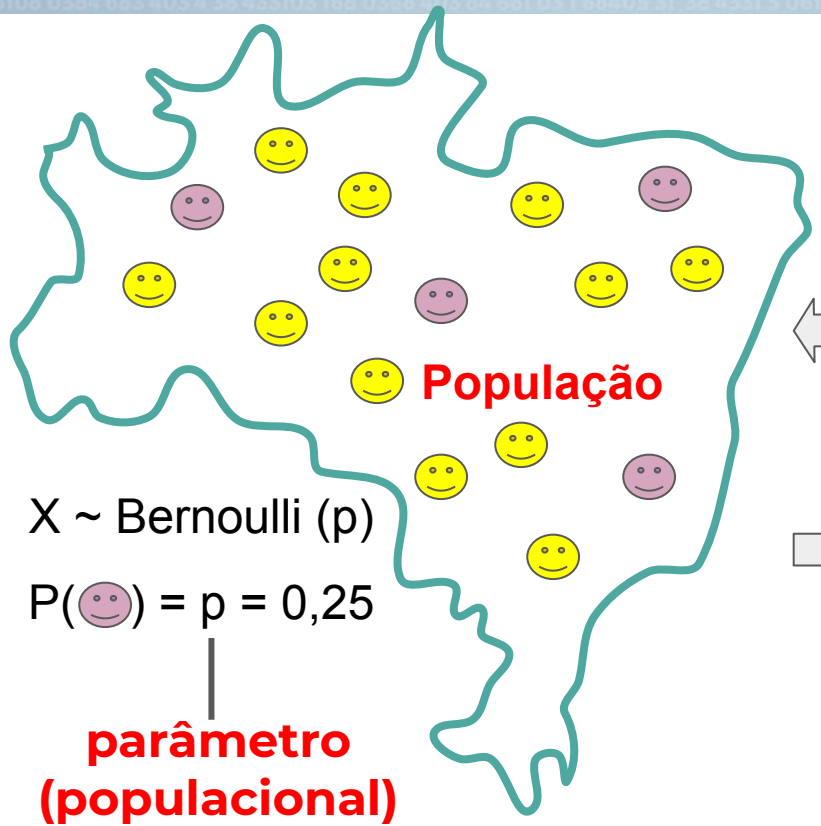
Aula 4: Estimação pontual e intervalar

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

Conteúdo

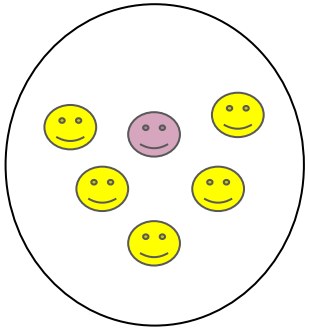
1. Conceitos básicos
    a. Nomenclatura
    b. Inferência estatística clássica
    c. Inferência estatística bayesiana
    d. Amostra aleatória simples
2. Estimação pontual
    a. Método de Máxima Verossimilhança (MV)
    b. Qualidade dos estimadores
3. Estimação intervalar
    a. Método da quantidade pivotal

Conceitos básicos: nomenclatura



Pergunta de interesse sobre uma variável X

Inferência estatística



Amostra

X ~ Bernoulli (p)

P( ) = p = 0,25

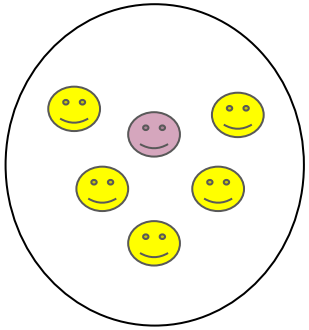
parâmetro (populacional)

$$X_1 = 0,\ X_2 = 0,\ X_3 = 1,\ X_4 = 0,\ X_5 = 0,\ X_6 = 0 \quad \Rightarrow \quad \hat{p} = \sum_{i=1}^{n} X_i / n$$

$$\hat{p} = 0,167$$

estimador (amostral)

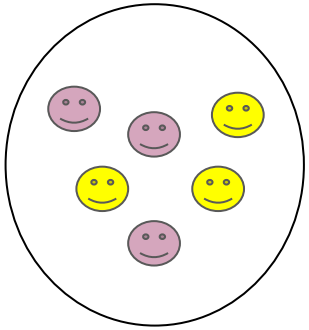
Conceitos básicos: nomenclatura



Amostra 1

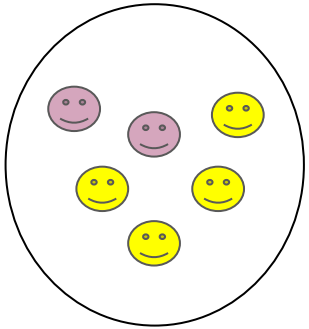
$\hat{p}_1 = 0,167$

estimador pontual



Amostra 3

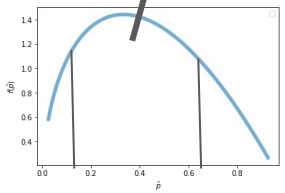
$\hat{p}_3 = 0,5$

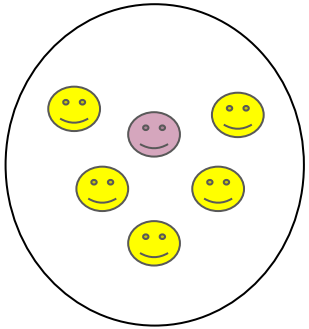


Amostra 2

$\hat{p}_2 = 0,333$

estimador intervalar





Amostra 4

$\hat{p}_4 = 0,167$

distribuição amostral:
distribuição de probabilidade do estimador

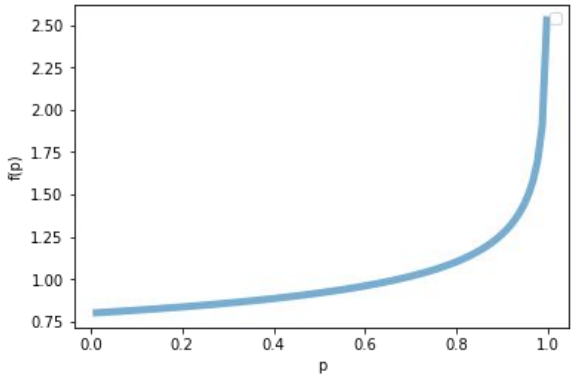
Conceitos básicos: inferência clássica x bayesiana

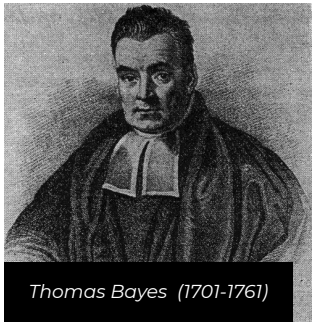
Inferência clássica ou frequentista

- parâmetros são nºs reais
- utiliza apenas a informação da amostra (e modelo probabilístico)
- Ex: 1) no lançamento de uma moeda, a probabilidade de sair cara é: igual a um número real (entre 0 e 1, no caso de probabilidade) que deve ser 0,5, se a moeda é "honesta".
  2) prob. de ter COVID-19, se chegou da Itália: $p \in (0,1)$

Inferência bayesiana

- parâmetros são aleatórios, que variam conforme uma distribuição de probabilidades
- utiliza também informação prévia (*a priori*), além da amostra





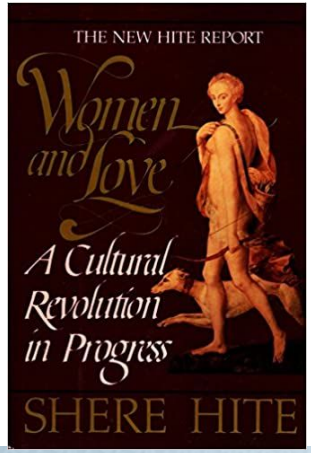
Thomas Bayes (1701-1761)

Conceitos básicos: amostra aleatória simples

- plano amostral: define como se dá a seleção da amostra
- define-se a probabilidade de cada unidade ser selecionada (em amostragens probabilísticas)
- ignorar o plano amostral pode subestimar (ou superestimar) a variância do estimador
- Amostra Aleatória Simples: cada unidade da população tem mesma probabilidade de ser selecionada para a amostra, sem estratificação e em um único estágio com seleção aleatória.
- Outras: estratificada, sistemática, múltiplos estágios, etc
- amostragens não probabilísticas: de conveniência, voluntários

Conceitos básicos: amostra

- 84% insatisfeitas em seus relacionamentos
- 70% das casadas por 5+ anos têm relações fora do casamento
- 95% reportaram assédio emocional dos parceiros



- 100.000 questionários distribuídos, mas 4,5% retornaram
- 127 questões (longo!)

The New York Times

By Arlie Russell Hochschild
Nov. 15, 1987

This ''Hite Report'' is not a report on average women. Its methodology is flawed. The author tells us she distributed 100,000 questionnaires to various organizations, church groups and a ''wide range'' of others. But wait. Even if every organization distributed every questionnaire to every woman and every woman answered it, Ms. Hite could only talk about women in these organizations, unless these members ''represent'' all women. How many women join feminist organizations? Or any organizations, for that matter? To accept this study as ''science'' would be wrong. But eventually the statisticians will come out with their brooms to sweep up the mess. Meanwhile, fishy statistics don't necessarily

Estimação pontual: MV

- Há vários métodos para encontrar estimadores de parâmetros
- Método de MV é um dos mais usados

Função de verossimilhança

Se $f(\boldsymbol{x})$ é a função de distribuição de probabilidade conjunta de uma amostra $X_1, X_2, ..., X_n$, que depende de um parâmetro $\boldsymbol{\theta}$, então $L(\boldsymbol{\theta}) = f(\boldsymbol{x})$ é chamada de função de verossimilhança.

ou $L(\theta \mid \boldsymbol{x})$

ou $f(\boldsymbol{x} \mid \theta)$

Se $X_1, X_2, ..., X_n$ são variáveis independentes, então:

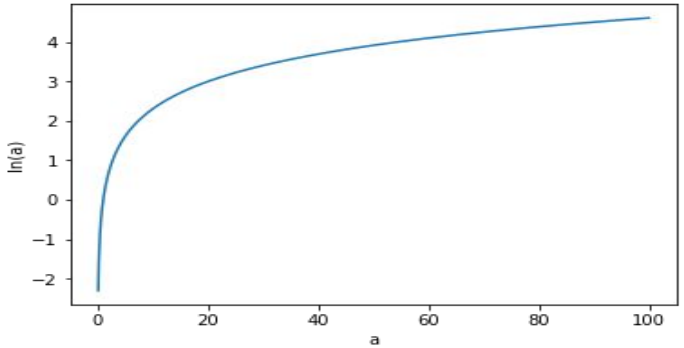
$$L(\theta) = \prod_{i=1}^{n} f(x_i)$$

Estimação pontual: MV

Ex: $X_1, X_2, ..., X_n$ são variáveis independentes com distribuição Bernoulli(p)

$$L(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

$$= p^{\sum_{i=1}^{n} x_i} (1-p)^{n-\sum_{i=1}^{n} x_i}$$



Estimador de Máxima de Verossimilhança (EMV)

É o valor do parâmetro que maximiza a função de verossimilhança quando os valores de ***x*** são iguais aos que observamos na amostra.

Maximizar $L(p)$ é equivalente à maximizar $lnL(p)$

Estimação pontual: MV

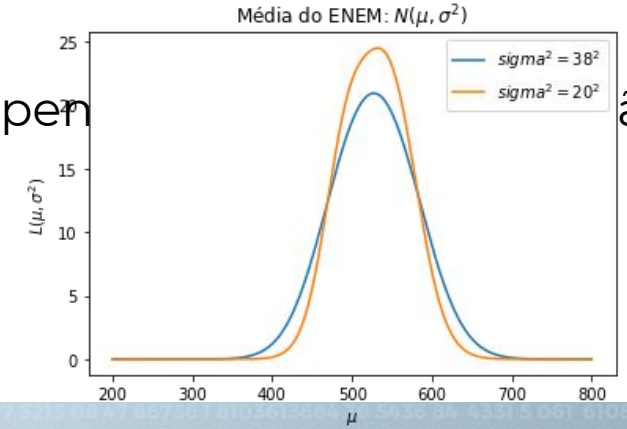
Ex: $X_1, X_2, ..., X_n$ são variáveis independentes com distribuição Bernoulli(p)

$$l(p) = lnL(p) = \sum_{i=1}^{n} x_i ln(p) + (n - \sum_{i=1}^{n} x_i) ln(1-p)$$

$$\frac{dl(p)}{dp} = \frac{\sum_{i=1}^{n} x_i}{p} - \frac{(n - \sum_{i=1}^{n} x_i)}{(1-p)} = 0 \quad \Rightarrow \quad \hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Exemplo: $X_1, X_2, ..., X_n$ são variáveis indepen ão $N(\mu,\sigma^2)$

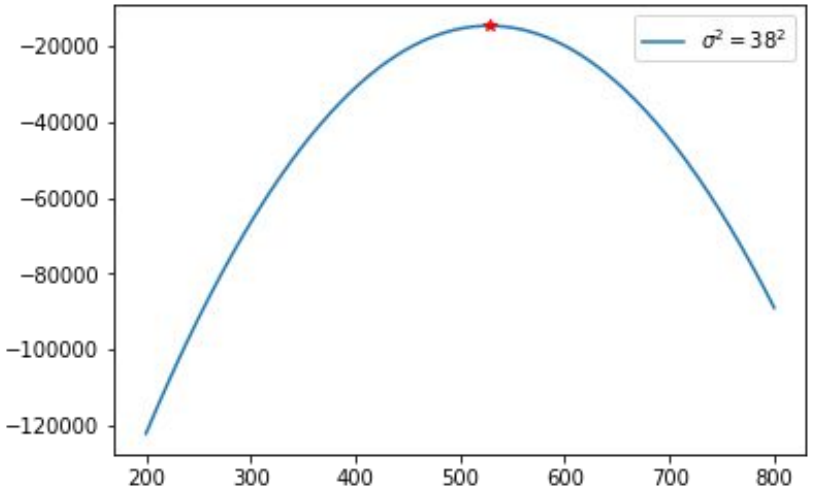
$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$



Estimação pontual: MV

Exemplo: $X_1, X_2, ..., X_n$ são variáveis independentes com distribuição $N(\mu,\sigma^2)$

$$l(\mu,\sigma^2) = ln(\sqrt{2\pi\sigma^2})^n - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$$



equações de estimação

$$\frac{\partial l(\mu,\sigma^2)}{\partial \mu} = 0$$

$$\frac{\partial l(\mu,\sigma^2)}{\partial \sigma^2} = 0$$

$$\hat{\mu} = \bar{X}$$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(X_i-\bar{X})^2}{n}$$

ex. ENEM:
527,6

$38{,}052^2$

Estimação pontual: Qualidade dos estimadores

Estimador não viciado de θ (ou não viesado)

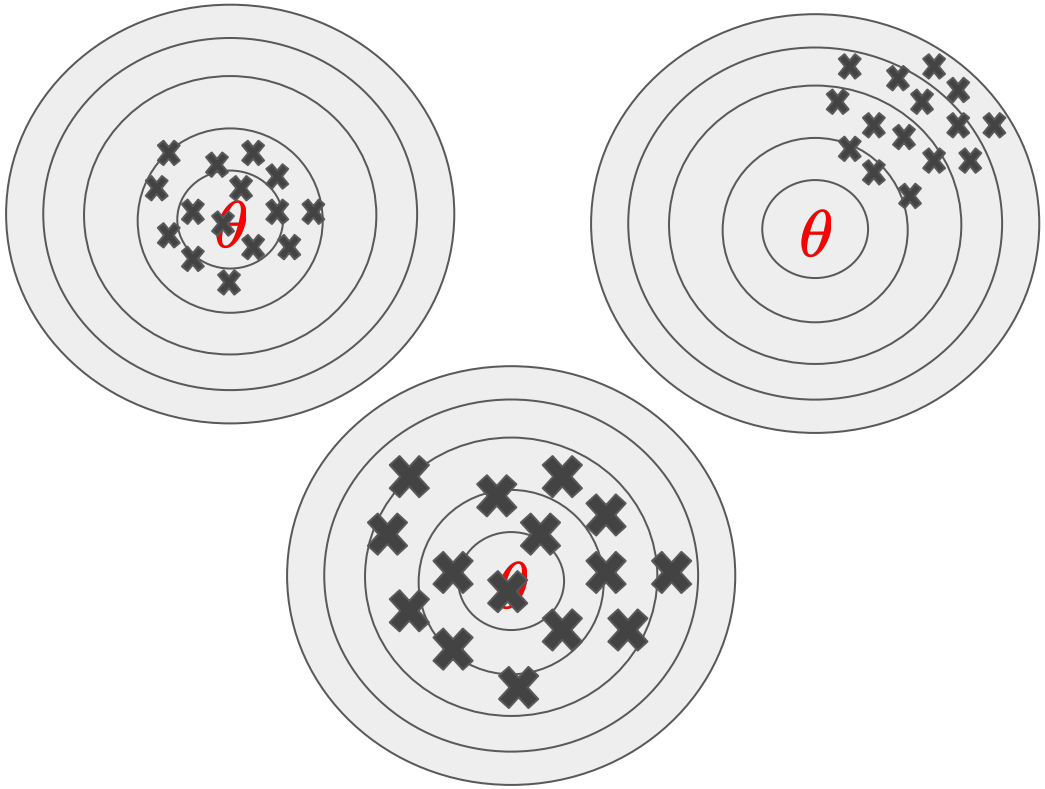
$$E(\hat{\theta}) = \theta$$

Vício ou viés do estimador

$$Bias(\hat{\theta}, \theta) = E(\hat{\theta}) - \theta$$

$\bar{X}$ e $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n-1}$

são não viciados

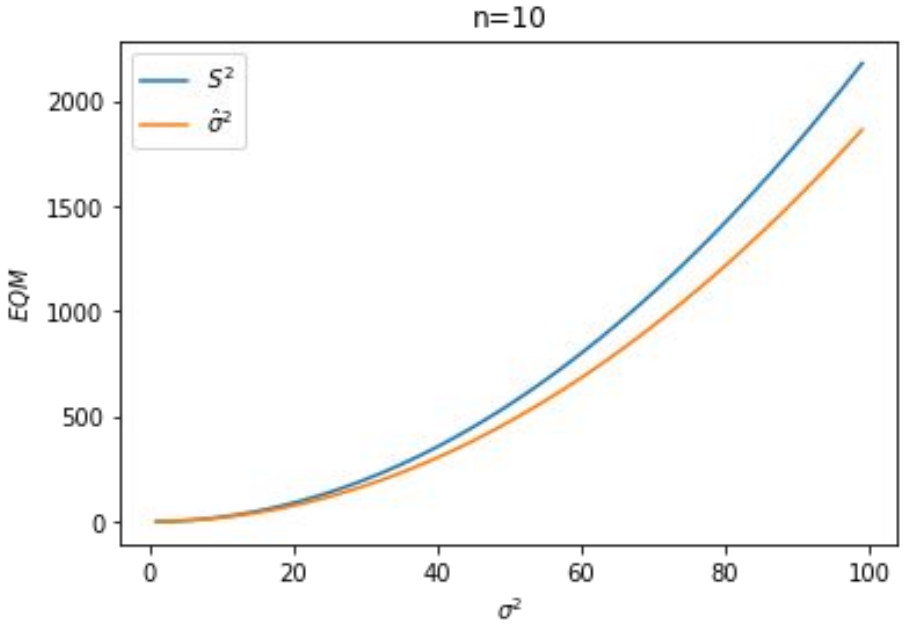


Estimação pontual: Qualidade dos estimadores

Erro Quadrático Médio de um estimador

$$EQM(\hat{\theta}, \theta) = E[(\hat{\theta} - \theta)^2]$$

$$EQM(\hat{\theta}, \theta) = V(\hat{\theta}) + Bias^2(\hat{\theta}, \theta)$$



Estimação intervalar

Se L e U são funções da amostra $X_1, X_2, ..., X_n$,

com distribuição de probabilidade dependente de θ

e $P(L < \theta < U) = 1 - \alpha$, então:

[L,U] é um intervalo de confiança de nível $100(1 - \alpha)\%$ para θ: $IC_\theta(1 - \alpha)$

a amplitude do intervalo reflete a incerteza a respeito de θ

ou γ:
99%, 95%, 90%,
são comuns

Estimação intervalar: quantidade pivotal

Quantidade pivotal

Função da amostra $X_1, X_2, ..., X_n$ e do parâmetro θ

com distribuição de probabilidade independente de parâmetros desconhecidos

Ex: $X_1, X_2, ..., X_n$ são variáveis independentes com $E(X)=\mu$ e $V(X)=\sigma^2$.
Quantidades pivotais para o $IC_\mu(1 - \alpha)$:

Se σ^2 conhecido: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{TCL}{\approx} N(0,1)$

Se σ^2 desconhecido: $\frac{\bar{X}-\mu}{\hat{\sigma}/\sqrt{n}} \approx N(0,1)$

Se amostra da $N(\mu,\sigma^2)$: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ ou $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{(n-1)}$

Estimação intervalar: IC para a média da Normal

Ex: $X_1, X_2, ..., X_n$ amostra aleatória da $N(\mu,\sigma^2)$

Como: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

ex. ENEM:
IC(95%) para a média da variável "Média": (526.2; 529.0)

usando a t-Student, pois a variância é desconhecida

$$P(-1,96 \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq 1,96) = 0,95$$

multiplicar por: $\sigma/\sqrt{n}$
subtrair: $\bar{X}$
multiplicar por (-1), invertendo o sinal da inequação

$$P\left(\bar{X} - 1,96.\sigma/\sqrt{n} \leq \mu \leq \bar{X} + 1,96.\sigma/\sqrt{n}\right) = 0,95$$

$IC_{\mu}(95\%)$

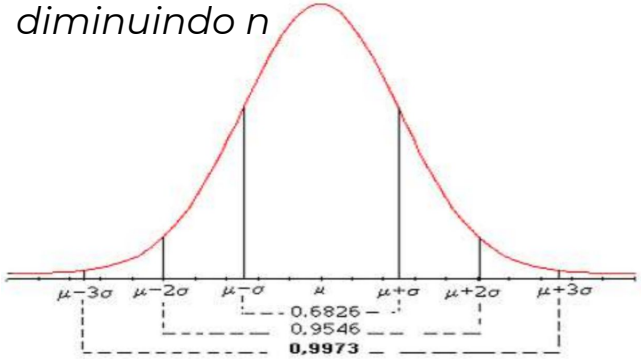
Estimação intervalar: IC para a proporção

Ex: $X_1, X_2, ..., X_n$ amostra aleatória da Bernoulli(p): E(X)=p e V(X)=p(1-p)

Como: $\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \approx N(0,1)$

$$P\left(-1,96 \leq \frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq 1,96\right) = 0,95$$

Aumenta-se a amplitude do IC:
aumentando a confiança (γ=1-α)
diminuindo n



$IC_p(95\%) = (\hat{p} - 1,96\sqrt{\hat{p}(1-\hat{p})/n}; \hat{p} + 1,96\sqrt{\hat{p}(1-\hat{p})/n})$

Se substituído por ¼: intervalo de confiança conservador para p