

Inferência Estatística

Inferência sobre uma proporção

sob diferentes paradigmas

Paulo Justiniano Ribeiro Jr

Curso de Estatística e Ciência de Dados
Departamento de Estatística
Universidade Federal do Paraná

17 de setembro de 2024

Como proceder os passos de inferência?

Como *aprender* com os dados?

- ▶ Como estimar quantidade(s) de interesse?
- ▶ Como expressão incerteza, por exemplo com uma estimativa intervalar?
- ▶ Como utilizar resultados para testar hipóteses sobre quantidades de interesse?
- ▶ Como obter a predição/classificação e avaliar a incerteza associada?

Exemplo: estimando uma proporção

Em uma população (considerada *infinita*) uma proporção θ de indivíduos apresenta determinada característica.

Inferências sobre θ :

- ▶ estimar θ ,
- ▶ expressar a incerteza sobre esta estimativa,
- ▶ verificar se θ (e portanto a população) está fora da norma/referência (proporção max. de 20%), se há evidências de um desvio “relevante” (*significativo*).

Dados de *uma* amostra (considerada aleatória):

$n = 80$ e $y = 19$.

Como proceder?

Questões:

- ▶ O que devo fazer?
- ▶ O que os dados dizem?
- ▶ Em que devo acreditar?

Paradigmas e métodos de inferência

Objetivos:

Estimativa de θ ,
expressão da incerteza,
opinião em relação a valor de interesse $\theta_0 = 0,20$

Abordagens (paradigmas):

- ▶ O que devo fazer?
O que aconteceria em outras amostras? – frequentista
- ▶ O que os dados dizem?
- ▶ Em que devo acreditar?

Abordagem frequentista

Objeto para inferência é a **distribuição amostral** (textos/procedimentos “usuais”)

Modelo : $Y \sim B(n, \theta)$

Parâmetro : θ (**de interesse**)

Estimador : $p = \hat{\theta} = \frac{Y}{n}$

$p = \hat{\theta} \sim N(\mu = \theta, \sigma^2 = \frac{\theta(1 - \theta)}{n})$ (**distribuição amostral**)

IC : $p \pm z_{1-\alpha/2} \sqrt{\frac{\theta(1 - \theta)}{n}}$ (**margem de erro**)

adota-se $\theta = \hat{\theta}$ (assintótico) ou $\theta = 0,5$ (conservador)

TH : $(\theta > \theta_0) : z = \frac{p - \theta_0}{\sqrt{\frac{\theta_0(1 - \theta_0)}{n}}} \sim N(0, 1)$ (**estatística de teste**)

Exemplo da inferência sobre proporção

Utilizando $\theta = \hat{\theta}$ na distribuição amostral:

Modelo : $Y \sim B(n = 80, \theta)$ **Parâmetro** : θ (de interesse)

Estimador : $\hat{\theta} = \frac{Y}{n}$ **Estimativa** : $\hat{\theta} = \frac{19}{80} = 0,2375$

$\hat{\theta} \approx N(\mu = \theta, \sigma^2 = \frac{\hat{\theta}(1 - \hat{\theta})}{n})$ (distribuição amostral "aproximada")

IC(95%) : $0,2375 \pm 1,96 \sqrt{\frac{0,2375(1 - 0,2375)}{80}}$ (margem de erro)

utilizando $\theta = \hat{\theta}$ o intervalo é dito ser *assintótico*

TH : $(\theta \leq 0,20 \text{ vs } \theta > 0,20)$: $z = \frac{0,2375 - 0,20}{\sqrt{\frac{0,20(1-0,20)}{80}}} = 0.839$ (estatística de teste)

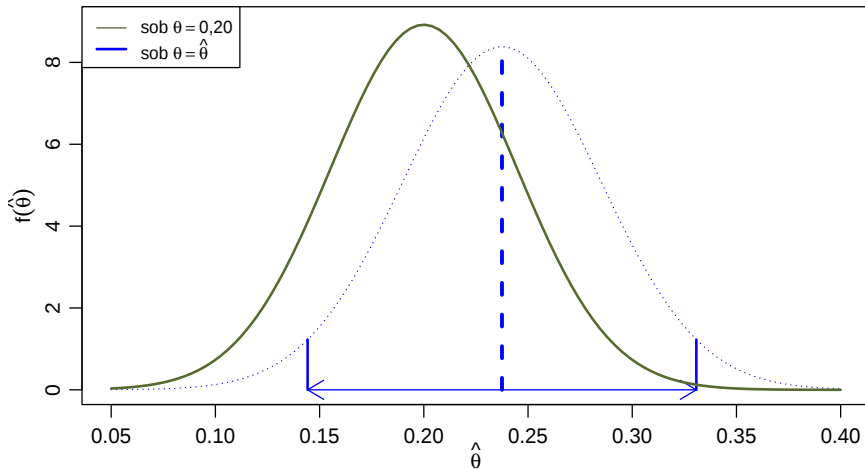
p-valor = $P(\hat{\theta} \geq 0,2375) = P(Z \geq 0.839) = 0.201$

(não rejeita a hipótese nula que $\theta \leq 0,20$)

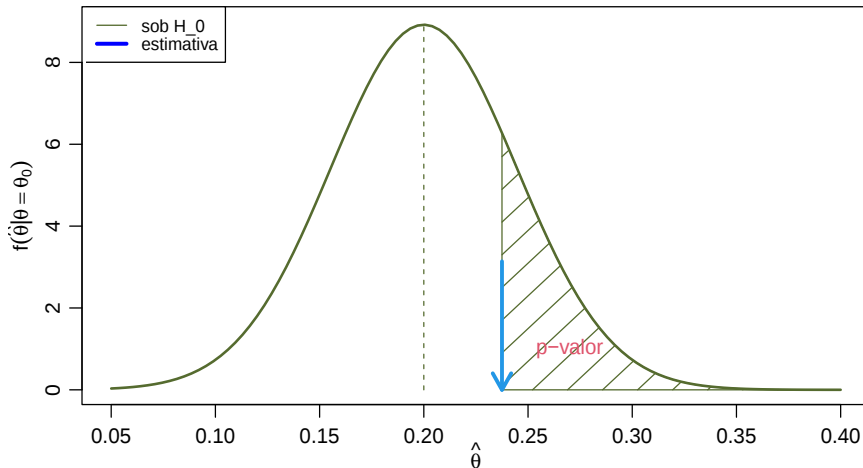
Abordagem Frequentista

- ▶ Baseia-se em considerar o comportamento das quantidades de interesse medidas na amostra, supondo que diversas amostras fossem tomadas da população.
- ▶ Tais quantidades, por serem baseadas em amostras aleatórias são portanto aleatórias, e possuem alguma distribuição de probabilidades.
- ▶ Tal distribuição de probabilidades é chamada de distribuição amostral.
- ▶ As inferências frequentistas são baseadas em probabilidades medidas nestas distribuições.
- ▶ Usual nos métodos, técnicas e procedimentos de estatística, especialmente os ligados a cursos e textos básicos e aplicados a diversas áreas.
- ▶ As distribuições amostrais podem ser obtidas analiticamente em alguns casos (e.g teste- t), aproximadas por distribuições conhecidas, ou obtidas por procedimentos computacionais intensivos (e.g. testes aleatorizados e bootstrap).

Frequentista: Distribuição amostral



Distribuição amostral e teste de hipótese

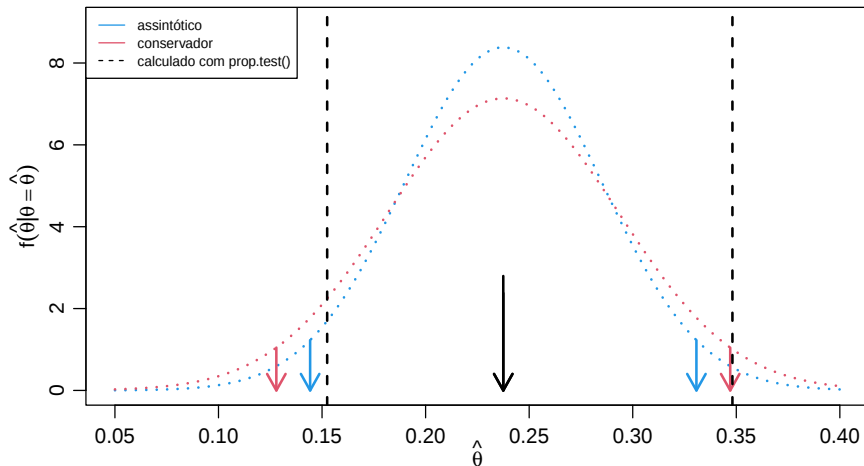


Inferência para proporção - frequentista

Na prática, com recursos computacionais

```
prop.test(19, 80)$conf
## [1] 0.1524765 0.3481396
## attr(,"conf.level")
## [1] 0.95
prop.test(19, 80, p=0.20, alt="greater")
##
## 1-sample proportions test with continuity correction
##
## data: 19 out of 80, null probability 0.2
## X-squared = 0.48828, df = 1, p-value = 0.2423
## alternative hypothesis: true p is greater than 0.2
## 95 percent confidence interval:
## 0.1632771 1.0000000
## sample estimates:
##      p
## 0.2375
```

Métodos alternativos



Métodos alternativos

Resultados diferentes

```
binom::binom.confint(19, 80)
```

##	method	x	n	mean	lower	upper
## 1	agresti-coull	19	80	0.2375000	0.1568987	0.3421559
## 2	asymptotic	19	80	0.2375000	0.1442487	0.3307513
## 3	bayes	19	80	0.2407407	0.1507750	0.3342939
## 4	cloglog	19	80	0.2375000	0.1512112	0.3348730
## 5	exact	19	80	0.2375000	0.1494538	0.3457770
## 6	logit	19	80	0.2375000	0.1569143	0.3426513
## 7	probit	19	80	0.2375000	0.1548012	0.3399211
## 8	profile	19	80	0.2375000	0.1534490	0.3381093
## 9	lrt	19	80	0.2375000	0.1534455	0.3381075
## 10	prop.test	19	80	0.2375000	0.1524765	0.3481396
## 11	wilson	19	80	0.2375000	0.1576467	0.3414078

Inferência frequentista

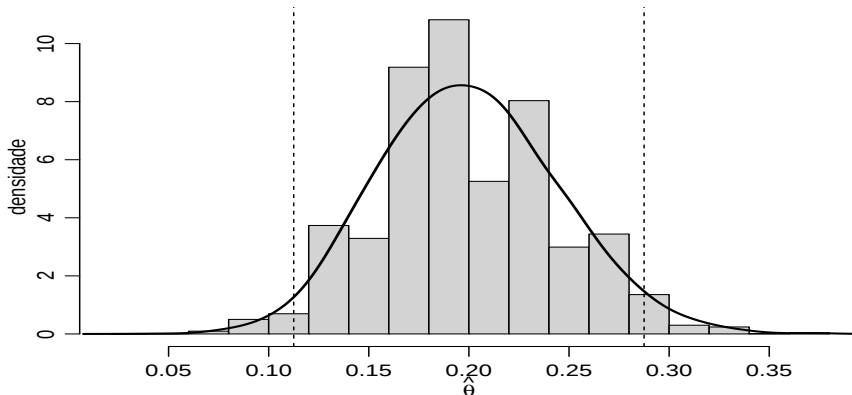
Um passo atrás: Revisando os fundamentos, entendendo e generalizando.
Experimento computacional "emulando" a realidade.

- ▶ Vamos simular uma população com proporção verdadeira ou sob hipótese $\theta = 0,20$ (já que na prática é desconhecida).
- ▶ Podemos emular computacionalmente a obtenção de amostras e respectivas estimativas desta população.
- ▶ Com isto podemos obter a distribuição (amostral) dos valores de $p = \hat{\theta}$ que expressa a variabilidade que pode ocorrer "naturalmente nas amostras".
- ▶ Esta distribuição contém os elementos para inferência, por exemplo, limites que contém 95% dos valores (IC) ou probabilidade acima de certo valor (teste de hipótese/p-valor).
- ▶ ...

Simulação da distribuição amostral

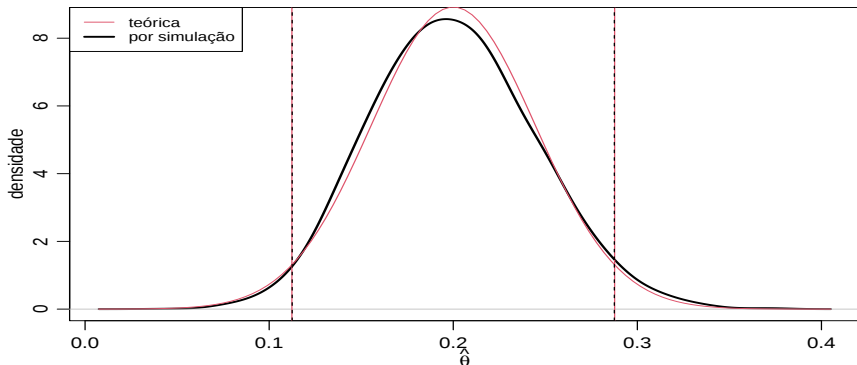
Tomar (no computador) diversas amostras da população e obter estimativa em cada.
 (código em **inf-prop.R**)

```
##      0%   2.5%   50%  97.5%  100%
## 0.0375 0.1125 0.2000 0.2875 0.3750
```



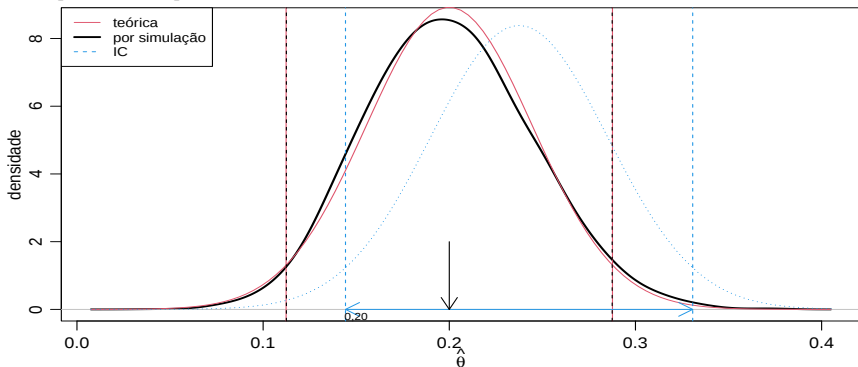
Inferência frequentista - distribuição amostral

- ▶ Pode ter uma "aparência" de alguma distribuição conhecida.
- ▶ Pode ser deduzida em alguns casos chegando à alguma distribuição "conhecida".
- ▶ Se identificada, pode-se fazer inferências sem as diversas amostras da população.
- ▶ Caso contrário utilizam-se procedimentos computacionais.



Inferência frequentista

- ▶ Mas ainda temos um problema: não conhecemos θ .
- ▶ Usamos o equivalente a uma distribuição com $\theta = \hat{\theta}$.
- ▶ Obtemos o IC que tem uma certa probabilidade (nível de confiança) de conter o valor verdadeiro do parâmetro.
- ▶ Notar $P[\hat{\theta} > 0,20]$ nas diferentes distribuições.



Inferência para proporção - frequentista

Resumindo:

- ▶ Se baseia no comportamento das *possíveis amostras* que poderiam ser retiradas da população
- ▶ Interpretação de intervalo de confiança: *o calculado a partir da amostra é um entre os possíveis, sendo que uma proporção dos possíveis (nível de confiança) conteria o verdadeiro valor*
- ▶ Dedos cruzados ...fé! (ou seja, não se tomou uma amostra atípica por mera chance.)
- ▶ Interpretação do Teste de Hipótese e **valor-p**: *mesmo sob H_0 uma proporção das possíveis amostras produziria valores tão ou mais extremos que o visto na amostra. Se esta proporção é **baixa** (nível de significância) a amostra é considerada incompatível com a hipótese nula.*

Método fortemente baseado em suposições e aproximações!

Uma alternativa (ainda) frequentista: Teste aleatorizado

Ideia básica:

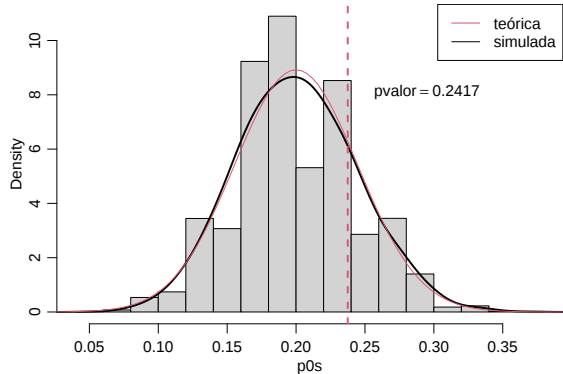
Reproduzir a essência da ideia frequentista porém obtendo a **distribuição amostral** por simulação sob H_0

Algoritmo:

- ▶ Simular amostras da população sob H_0
- ▶ Calcular o valor de interesse ou estatística de teste para cada amostra simulada
- ▶ **valor-p** proporção destes que são mais “extremos” do que o valor observado na amostra

Teste aleatorizado

```
quantile(p0s, prob=c(0,0.025, 0.5, 0.95, 1))
##      0%   2.5%   50%   95%  100%
## 0.0500 0.1125 0.2000 0.2750 0.3750
(pvalor <- mean(p0s >= 19/80))
## [1] 0.2417
```



Características do testes aleatorizados

- ▶ Implementação computacional simples em diversos problemas.
- ▶ Contorna/dispensa aproximações teóricas para distribuição amostral.
- ▶ Implementação segue a essência do pensamento frequentista.
- ▶ Foco em teste de hipóteses na formulação original.

Paradigmas e métodos de inferência

Objetivos:

Estimativa de θ ,
expressão da incerteza,
opinião em relação a valor de interesse $\theta_0 = 0, 20$

Abordagens (paradigmas):

- ▶ O que devo fazer? O que aconteceria em outra amostras? – frequentista
- ▶ O que os dados dizem? – verossimilhança
- ▶ Atualizei o que eu pensava? Em que acredito?

Compatibilidade com os dados

Se a proporção é θ , podemos avaliar a chance (probabilidade) de obter um certo número Y de indivíduos com a característica em uma amostra de n indivíduos. Sob certas suposições é razoável adotar:

$$P[Y = y|\theta] = \binom{n}{y} \theta^y (1 - \theta)^{n-y}.$$

Quando obtemos a amostra temos $n = 80$ e $y = 19$. Para cada θ temos então a probabilidade de obter esta particular amostra:

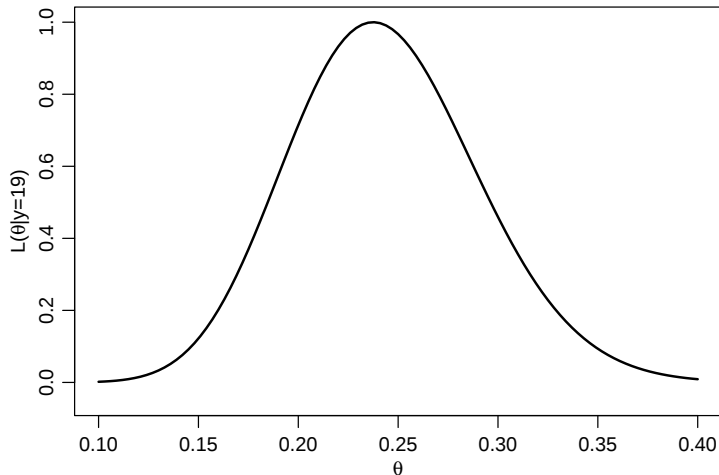
$$P[Y = y|\theta] = \binom{80}{19} \theta^{19} (1 - \theta)^{80-19}.$$

Como esta chance muda para cada valor de θ ?

Considerando todos os θ possíveis temos uma função de θ :

$$L[\theta|y] = \binom{80}{19} \theta^{19} (1 - \theta)^{80-19}.$$

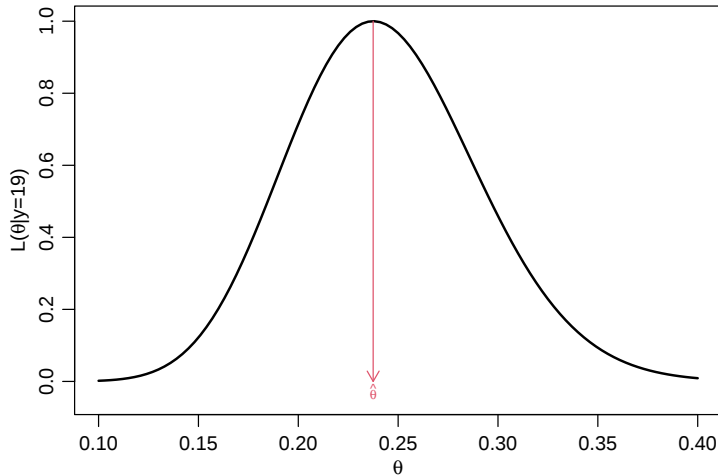
Uma função representando ideias



Objetivos:

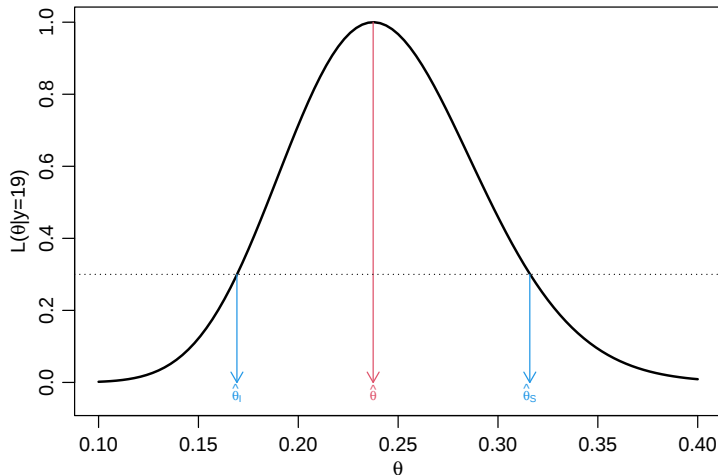
- Estimativa de θ ,
- expressão da incerteza,
- opinião em relação a valor de interesse $\theta_0 = 0,20$

Uma função representando ideias



Objetivos:
 → Estimativa de θ ,

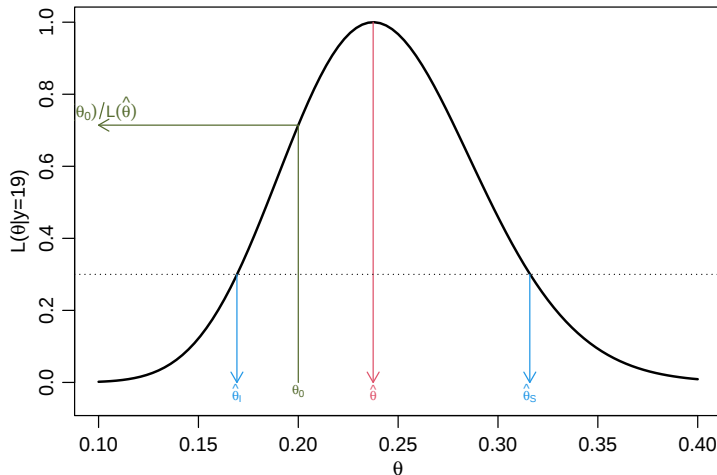
Uma função representando ideias



Objetivos:

- Estimativa de θ ,
- expressão da incerteza,

Uma função representando ideias



Objetivos:

- Estimativa de θ ,
- expressão da incerteza,
- **opinião em relação a valor de interesse**
 $\theta_0 = 0, 20$.

Representações alternativas

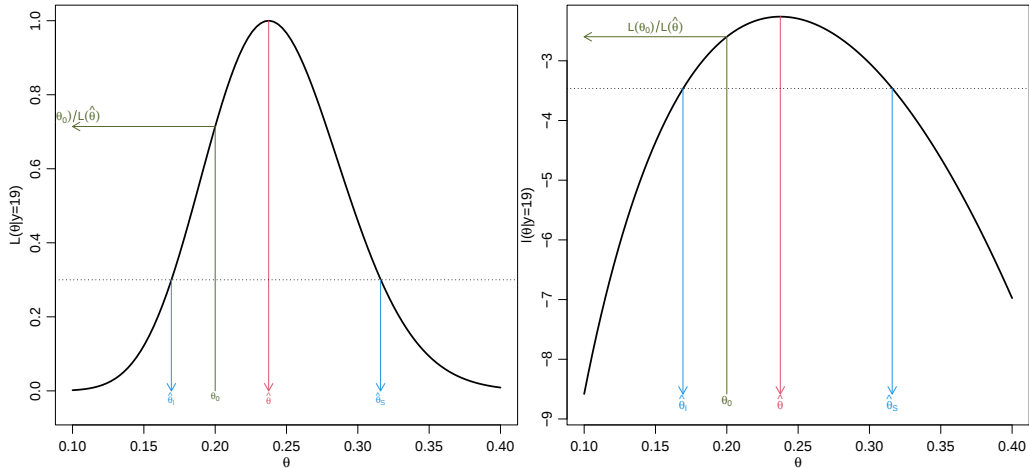


Figura 1. Funções de verossimilhança relativa (esquerda) e log-verossimilhança (direita).

Suposições e considerações adicionais

- ▶ Suposições:
 - ▶ amostra aleatória (independência),
 - ▶ população *infinita*,
 - ▶ modelo binomial (critério de parada: n fixo) e invariância.

- ▶ Critérios são necessários.
 - ▶ Onde efetuar o corte da função para determinar faixas de incerteza?
 - ▶ Como avaliar o valor de verossimilhança (relativa) para θ_0 ?

Observação extrema: acaso ou "fora do padrão"?

- ▶ Joga-se um dado 3 vezes e se obtém 3 caras.
- ▶ Joga-se um dado 5 vezes e se obtém 5 caras.
- ▶ Joga-se um dado 8 vezes e se obtém 8 caras.
- ▶ Joga-se um dado 10 vezes e se obtém 10 caras.

Quantas caras seguidas te fazem crer que o dado não é *honesto*?

Observação extrema: acaso ou "fora do padrão"?



Inferência pela verossimilhança

Necessidade de critérios.

- ▶ Definir o valor para corte da função para obter intervalos de confiança (IC's).
- ▶ Definir limiar para o valor de verossimilhança (relativa ao máximo) para θ_0 .

Possíveis soluções.

- ▶ Critérios de razoabilidade e comparação (e.g. moedas ou lembre da família da foto).
- ▶ Argumento frequentista (comportamento “médio” da verossimilhança) estabelece relações:

r	“Caras”	$P[Z < \sqrt{c^*}]$
50%	1,00	0,761
26%	1,94	0,899
15%	2,74	0,942
3,6%	4,80	0,990

Suposições e considerações adicionais

- ▶ Suposições:
 - ▶ amostra aleatória (independência),
 - ▶ população *infinita*,
 - ▶ modelo binomial (critério de parada: n fixo) e invariância.

- ▶ Critérios são necessários.
 - ▶ Onde efetuar o corte da função para determinar faixas de incerteza.
 - ▶ Como avaliar o valor de verossimilhança (relativa) para θ_0 .

- ▶ Indo além dos dados.
 - ▶ Não há ou não se usa nenhuma informação “acessória/preliminar” sobre θ ?
 - ▶ Como se comportariam outras amostras que fossem eventualmente tomadas?
 - ▶ Questões motivam e caracterizam diferentes abordagens!

Paradigmas e métodos de inferência

Objetivos:

Estimativa de θ ,
expressão da incerteza,
opinião em relação a valor de interesse $\theta_0 = 0, 20$

Abordagens (paradigmas):

- ▶ O que devo fazer? O que aconteceria em outra amostras? – frequentista
- ▶ O que os dados dizem? – verossimilhança
- ▶ **Atualizei o que eu pensava? Em que acredito?** – bayesiana

Estados da natureza e atualização

O problema do teste de diagnóstico (e análogos).

Estados (iniciais) da natureza:

Estado (θ)	$\theta = 0(\bar{D})$	$\theta = 1(D)$
Probabilidade	0,980	0,020

Estados da natureza após primeiro exame positivo:

Estado ($\theta y_1 = 1$)	$\theta = 0(\bar{D})$	$\theta = 1(D)$
Probabilidade	0,916	0,084

Estados da natureza após segundo exame positivo:

Estado ($\theta y_1 = 1, y_2 = 1$)	$\theta = 0(\bar{D})$	$\theta = 1(D)$
Probabilidade	0,708	0,292

Estados da natureza após terceiro exame positivo:

Estado ($\theta y_1 = 1, y_2 = 1, y_3 = 1$)	$\theta = 0(\bar{D})$	$\theta = 1(D)$
Probabilidade	0,350	0,650

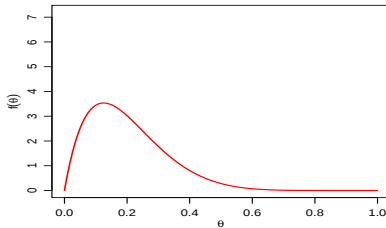
Estados da natureza e atualização

O problema de estimar proporção (e análogos).

A proporção θ é uma variável aleatória contínua.

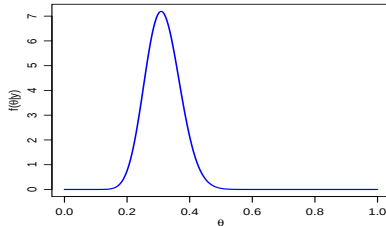
Estados (iniciais) da natureza:

θ tem uma distribuição de probabilidades $f(\theta)$.



Estados da natureza após observar os dados y :

θ tem uma distribuição de probabilidades revisada $f(\theta|y)$.



Inferência Bayesiana

O objeto de inferência é a **distribuição à posteriori**

- ▶ A incerteza inicial sobre θ é expressa na forma de uma distribuição **priori** para θ
- ▶ Com amostra **atualizamos** opinião θ com a informação contida na **verossimilhança**
- ▶ O conhecimento/incerteza atualizados sobre θ é expresso pela distribuição **posteriori**

Formalmente:

$$f(\theta|y) \propto f(\theta) \cdot L(\theta|y)$$

ou, usando jargão técnico:

$$\text{posteriori} \propto \text{priori} \cdot \text{verossimilhança}$$

A essência de Bayes ilustrada (I)

Exemplo I : estimação da proporção de atributo (θ) na população

- **Priori**: Acredita-se que o atributo ocorre em 40% da população com 70% de chance de estar entre 30 e 50%.

Informação expressa como distribuição de probabilidades para θ :

$$[\theta] \sim \text{Beta}(10; 15) \longrightarrow f(\theta) = C \theta^{10-1} (1 - \theta)^{15-1}$$

- **Verossimilhança**: Modelo Binomial, amostra $n=80$ e $y=19$

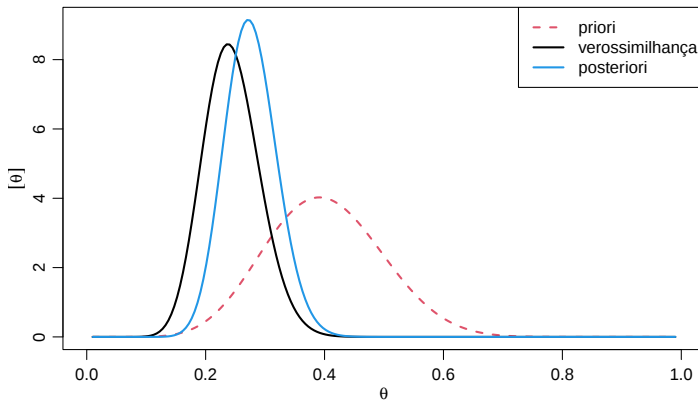
$$L[\theta] = \binom{80}{19} \theta^{19} (1 - \theta)^{80-19}$$

- **Posteriori**: a distribuição de probabilidades para θ após observar os dados:

$$[\theta|y] \sim \text{Beta}(29; 76) \longrightarrow f(\theta|y) = C \theta^{29-1} (1 - \theta)^{76-1}$$

C é um valor constante (não depende θ) e conhecido.

A essência de Bayes ilustrada (I)



A essência de Bayes ilustrada (I)

```
(prI <- prioriBeta(0.4, c(0.30, 0.50), 0.70))  
##      alpha      beta  
##  9.912277 14.868415  
postBinom(19, 80, prI, plot=FALSE)  
## $pars  
##           alpha      beta  
## priori      9.912277 14.86842  
## posteriori 28.912277 75.86842  
##  
## $summary  
##           moda      media  variancia  
## priori      0.3912206 0.4000000 0.009309292  
## posteriori 0.2715712 0.2759313 0.001888750  
##  
## $EMV  
## [1] 0.2375
```

A essência de Bayes ilustrada (II)

Uma priori bem diferente:

- **Priori:** Acredita-se que o atributo ocorre em 8% da população com 90% de chance de estar entre 3 e 20%.

$$[\theta] \sim \text{Beta}(2; 24) \longrightarrow f(\theta) = C \theta^{2-1} (1 - \theta)^{24-1}$$

- **Verossimilhança:** Modelo Binomial, amostra $n=80$ e $y=19$

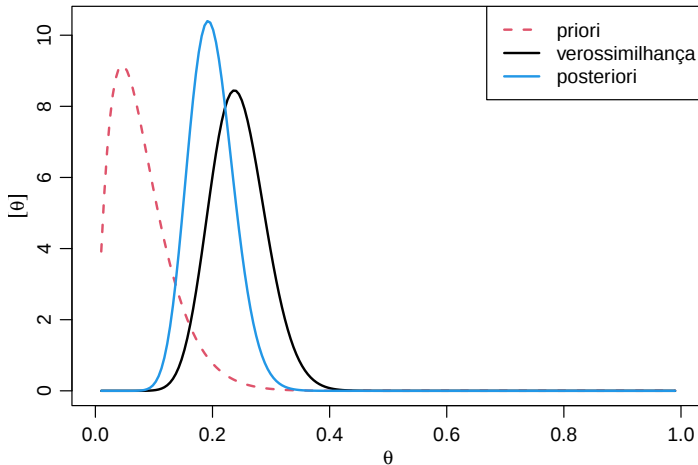
$$L[\theta] = \binom{80}{19} \theta^{19} (1 - \theta)^{80-19}$$

- **Posteriori:** após observar os dados:

$$[\theta|y] \sim \text{Beta}(21; 85) \longrightarrow f(\theta|y) = C \theta^{21-1} (1 - \theta)^{85-1}$$

C é um valor constante conhecido.

A essência de Bayes ilustrada (II)



A essência de Bayes ilustrada (II)

```
(prII <- prioriBeta(0.08, c(0.03, 0.20), 0.90))
##      alpha      beta
## 2.124901 24.436358
postBinom(19, 80, prII, plot=FALSE)
## $pars
##           alpha      beta
## priori      2.124901 24.43636
## posteriori 21.124901 85.43636
##
## $summary
##           moda      media  variancia
## priori      0.0457998 0.0800000 0.002670415
## posteriori 0.1924700 0.1982418 0.001477688
##
## $EMV
## [1] 0.2375
```

A essência de Bayes ilustrada (III)

Uma **priori vaga** :

- **Priori**: Não se sabe praticamente nada sobre θ . Expressa-se então que o atributo ocorre em 50% da população mas com 90% de chance de estar entre 5 e 95%.

$$[\theta] \sim \text{Beta}(1.2; 1.2) \longrightarrow f(\theta) = C \theta^{1.2-1} (1 - \theta)^{1.2-1}$$

- **Verossimilhança**: Modelo Binomial, amostra $n=80$ e $y=19$

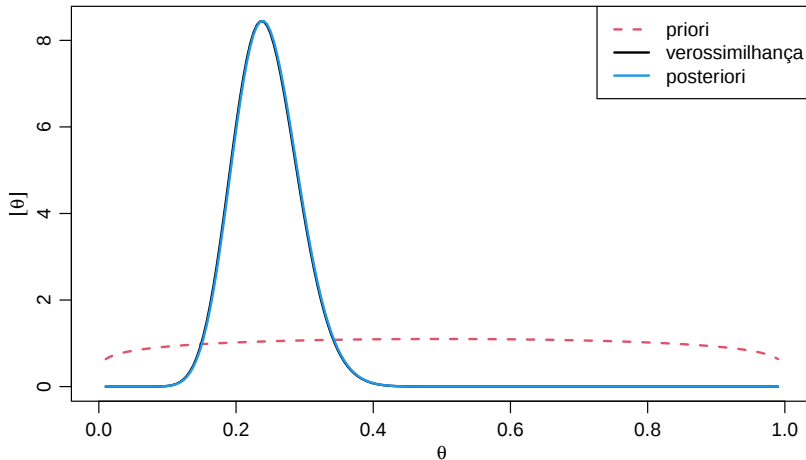
$$L[\theta] = \binom{80}{19} \theta^{19} (1 - \theta)^{80-19}$$

- **Posteriori**: após observar os dados:

$$[\theta|y] \sim \text{Beta}(20, 62) \longrightarrow f(\theta|y) = C \theta^{20-1} (1 - \theta)^{62-1}$$

C é um valor constante conhecido.

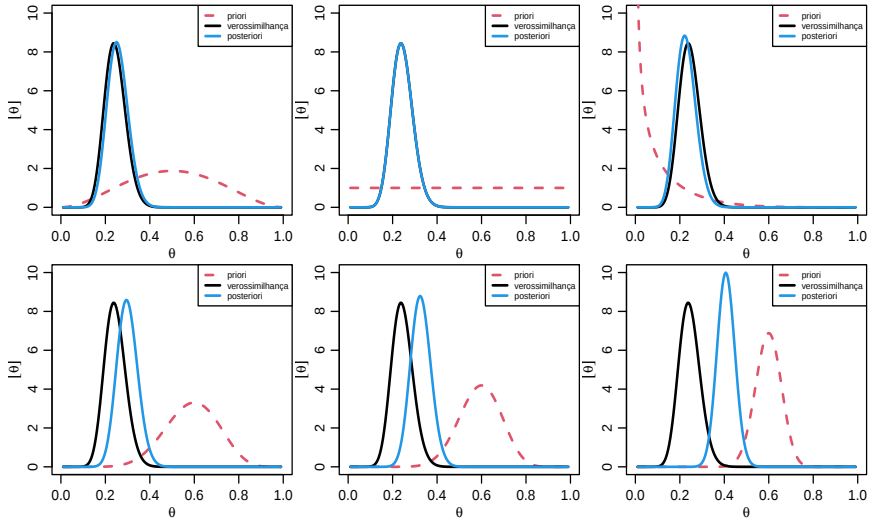
A essência de Bayes ilustrada (III)



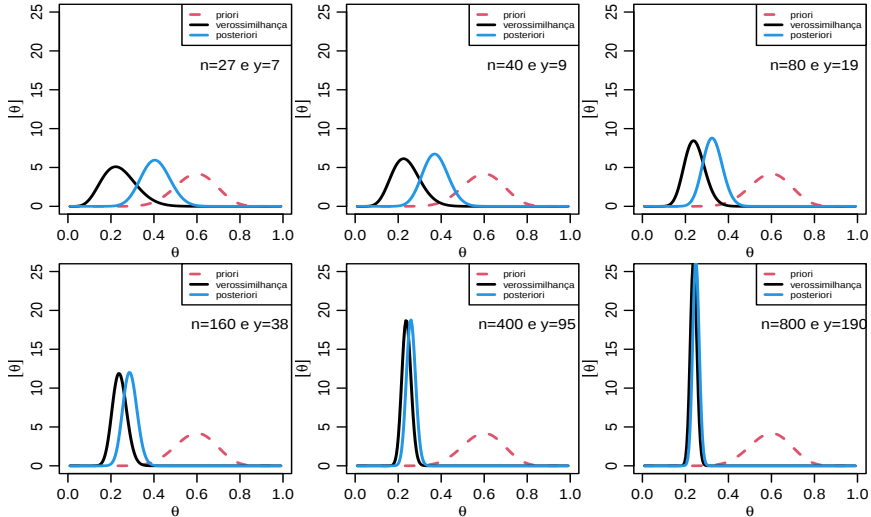
A essência de Bayes ilustrada (III)

```
(prIII <- prioriBeta(0.50, c(0.05, 0.95), 0.90))
##      alpha      beta
## 1.170088 1.170088
postBinom(19, 80, prIII, plot=FALSE)
## $pars
##           alpha      beta
## priori      1.170088 1.170088
## posteriori 20.170088 62.170088
##
## $summary
##           moda      media  variancia
## priori      0.5000000 0.5000000 0.074846333
## posteriori 0.2386115 0.2449605 0.002219276
##
## $EMV
## [1] 0.2375
```

Efeito da priori (fixando amostra)



Efeito do tamanho da amostra (fixando priori)



Comentários

- ▶ Expressão da opinião “a priori” é necessária e sua especificação é um desafio.
- ▶ As interpretações de intervalo de confiança são agora probabilísticas, por exemplo pode-se avaliar intervalo (de credibilidade) para uma certa probabilidade:

$$P[a < \theta < b] = 0,95.$$

- ▶ No contexto do exemplo, pode-se avaliar

$$P[\theta \geq 0,20].$$

Inferência Bayesiana

Em resumo:

- ▶ Estimativa de θ : alguma medida resumo da posteriori (média, moda, mediana, ...)
- ▶ expressão da incerteza: variabilidade da distribuição posteriori
- ▶ opinião em relação a valor de interesse $\theta \geq 0.20$: probabilidade na posteriori

Paradigmas e métodos de inferência

Objetivos:

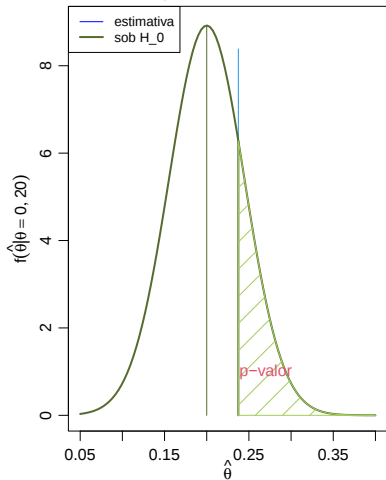
Estimativa de θ ,
expressão da incerteza,
opinião em relação a valor de interesse $\theta_0 = 0, 20$

Abordagens (paradigmas):

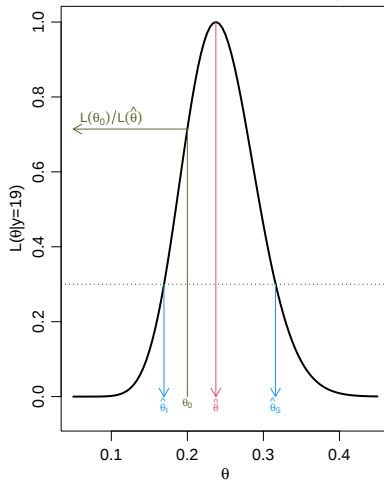
- ▶ O que devo fazer? O que aconteceria em outra amostras? – frequentista
- ▶ O que os dados dizem? – verossimilhança
- ▶ Atualizei o que eu pensava? Em que acredito? – inferência bayesiana

Comparando paradigmas

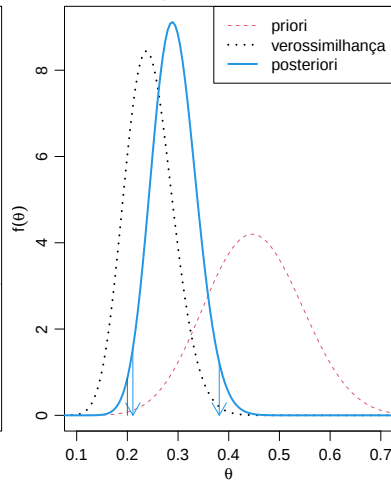
Frequentista
Distribuição amostral



Verossimilhança
Função de Verossimilhança



Bayesiano
Distribuição Posteriori



Comparando paradigmas

- ▶ **Qual o valor de θ ?** (estimação pontual):
 - ▶ Frequentista: fornecido por algum método de estimação
 - ▶ Verossimilhança: máximo (supremo) da função de verossimilhança
 - ▶ Bayesiana: alguma medida resumo da posteriori (média, moda, mediana, ...)
- ▶ **Expressão da incerteza sobre θ** (estimação intervalar):
 - ▶ Frequentista: variabilidade na distribuição amostral (intervalo de confiança)
 - ▶ Verossimilhança: faixa de valores dentro de um limite de compatibilidade com a amostra, curvatura da função
 - ▶ Bayesiana: variabilidade na distribuição posteriori (intervalo de credibilidade)
- ▶ **Opinião em relação a valor de interesse $\theta_0 = 0,20$** (teste de hipótese):
 - ▶ Frequentista: probabilidade na distribuição amostral (p-valor)
 - ▶ Verossimilhança: comparação da verossimilhança deste valor com a do máximo
 - ▶ Bayesiana: probabilidade na posteriori