

Estimação pontual e distribuições amostrais

Fernando de Pol Mayer

Laboratório de Estatística e Geoinformação (LEG)
Departamento de Estatística (DEST)
Universidade Federal do Paraná (UFPR)



Este conteúdo está disponível por meio da Licença Creative Commons 4.0
(Atribuição/NãoComercial/PartilhaIgual)

- 1 Introdução
- 2 Estimação pontual
 - Propriedades dos estimadores
- 3 Erros amostrais
- 4 Distribuições amostrais
 - Distribuição amostral da média
 - Distribuição amostral da proporção
- 5 Referências

1 Introdução

2 Estimação pontual

- Propriedades dos estimadores

3 Erros amostrais

4 Distribuições amostrais

- Distribuição amostral da média
- Distribuição amostral da proporção

5 Referências

Definição (Inferência estatística)

Seja X uma variável aleatória com função densidade (ou de probabilidade) denotada por $f(x, \theta)$, em que θ é um parâmetro desconhecido. Chamamos de **inferência estatística** o problema que consiste em especificar um ou mais valores para θ , baseado em um conjunto de valores X .

A inferência pode ser feita através de duas formas:

- estimativa pontual
- estimativa intervalar

Redução de dados

Um experimentador usa as informações em uma amostra aleatória X_1, \dots, X_n para se fazer inferências sobre θ .

Normalmente n é grande e fica inviável tirar conclusões baseadas em uma longa **lista** de números.

Por isso, um dos objetivos da inferência estatística é **resumir** as informações de uma amostra, da maneira mais **compacta** possível, mas que ao mesmo tempo seja também **informativa**.

Normalmente esse resumo é feito por meio de **estatísticas**, por exemplo, a média amostral e a variância amostral.

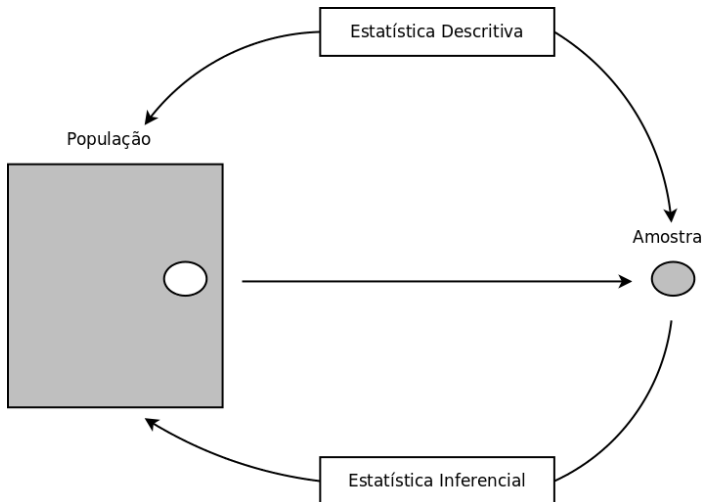
Definição (População)

O conjunto de valores de uma característica associada a uma coleção de indivíduos ou objetos de interesse é dito ser uma população.

Definição (Amostra)

Uma sequência X_1, \dots, X_n de n variáveis aleatórias independentes e identicamente distribuídas (iid) com função densidade (ou de probabilidade) $f(x, \theta)$ é dita ser uma amostra aleatória de tamanho n da distribuição de X . Como normalmente $n > 1$, então temos que a fdp ou fp conjunta será

$$f(\mathbf{x}, \theta) = f(x_1, \dots, x_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$$



População → **censo** → **parâmetro**

Uma medida numérica que descreve alguma característica da população, usualmente representada por letras gregas: θ , μ , σ , ...

Exemplo: média populacional = μ

População → **amostra** → **estatística**

Uma medida numérica que descreve alguma característica da amostra, usualmente denotada pela letra grega do respectivo parâmetro com um acento circunflexo: $\hat{\theta}$, $\hat{\mu}$, $\hat{\sigma}$, ... , ou por letras do alfabeto comum: \bar{x} , s , ...

Exemplo: média amostral = \bar{x}

É importante notar que um parâmetro não é restrito aos modelos de probabilidade. Por exemplo:

$$X \sim N(\mu, \sigma^2) \Rightarrow \text{parâmetros: } \mu, \sigma^2$$

$$Y \sim \text{Poisson}(\lambda) \Rightarrow \text{parâmetro: } \lambda$$

$$Y = \beta_0 + \beta_1 X \Rightarrow \text{parâmetros: } \beta_0, \beta_1$$

$$L_t = L_\infty[1 - e^{-k(t-t_0)}] \Rightarrow \text{parâmetros: } L_\infty, k, t_0$$

Definição (Estatística)

Qualquer função da amostra que não depende de parâmetros desconhecidos é denominada uma estatística, denotada por $T(\mathbf{X}) = T(X_1, X_2, \dots, X_n)$

Exemplos:

- $T_1(\mathbf{X}) = \sum_{i=1}^n X_i = X_1 + X_2 + \dots + X_n$
- $T_2(\mathbf{X}) = \prod_{i=1}^n X_i = X_1 \cdot X_2 \cdots X_n$
- $T_3(\mathbf{X}) = X_{(1)}$
- $T_4(\mathbf{X}) = \sum_{i=1}^n (X_i - \mu)^2$

Verificamos que T_1 , T_2 , T_3 são estatísticas, mas T_4 não.

Como é uma função da amostra, então uma estatística também é uma **variável aleatória** → distribuições amostrais

Se podemos utilizar $T(\mathbf{X})$ para extrair toda a informação da amostra, então dizemos que ela é **suficiente** para θ .

Definição (Estatística suficiente)

Seja X_1, \dots, X_n uma amostra aleatória da variável aleatória X , com fdp ou fp $f(x, \theta)$ com $\theta \in \Theta$, dizemos que uma estatística $T(\mathbf{X})$ é suficiente para θ , se a distribuição condicional de \mathbf{X} dado $T(\mathbf{X}) = t$ for independente de θ

$$f_{\mathbf{X}|T(\mathbf{X})}(\mathbf{x}|t) \rightarrow \text{independe de } \theta$$

A definição acima permite verificar se uma estatística é suficiente, mas não como encontrá-la. Dois conceitos fundamentais para encontrar estatísticas (conjuntamente) suficientes são:

- o **critério da fatoração de Neyman**
- o **critério da família exponencial**

Definição (Espaço paramétrico)

O conjunto Θ em que θ pode assumir seus valores é chamado de **espaço paramétrico**

Definição (Estimador)

Qualquer estatística que assume valores em Θ é um estimador para θ .

Dessa forma, um **estimador pontual** para θ é qualquer estatística que possa ser usada para estimar esse parâmetro, ou seja,

$$\hat{\theta} = T(\mathbf{X})$$

Observações:

- 1 Todo estimador é uma estatística, mas nem toda estatística é um estimador.
- 2 O valor assumido pelo estimador pontual é chamado de **estimativa pontual**,

$$T(\mathbf{X}) = T(X_1, \dots, X_n) = t$$

ou seja, o estimador é uma **função** da amostra, e a estimativa é o **valor observado** de um estimador (um número) de uma amostra particular.

1 Introdução

2 Estimação pontual

- Propriedades dos estimadores

3 Erros amostrais

4 Distribuições amostrais

- Distribuição amostral da média
- Distribuição amostral da proporção

5 Referências

A ideia geral por trás da estimação pontual é muito simples:

Quando a amostragem é feita a partir de uma população descrita por uma função $f(x, \theta)$, o conhecimento de θ a partir da amostra, gera todo o conhecimento para a população.

Dessa forma, é natural que se procure um método para se achar um **bom** estimador para θ .

Existem algumas propriedades que definem o que é um bom estimador, ou o “**melhor**” estimador entre uma série de candidatos.

Localização do problema: Considere X_1, \dots, X_n uma amostra aleatória de uma variável aleatória X com fdp ou fp $f(x, \theta)$, $\theta \in \Theta$.
Sejam:

$$\hat{\theta}_1 = T_1(X_1, \dots, X_n) \quad \hat{\theta}_2 = T_2(X_1, \dots, X_n)$$

Qual dos dois estimadores pontuais é **melhor** para θ ?

Como não conhecemos θ , não podemos afirmar que $\hat{\theta}_1$ é melhor do que $\hat{\theta}_2$ e vice-versa.

O problema da estimação pontual é então escolher um estimador $\hat{\theta}$ que se aproxime de θ segundo algumas **propriedades**.

Exemplo 1: Considere uma amostra aleatória (X_1, \dots, X_n) de uma variável aleatória $X \sim N(\mu = 3, \sigma^2 = 1)$ e os estimadores pontuais para μ

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{e} \quad \hat{\theta}_2 = \frac{X_{(1)} + X_{(n)}}{2}$$

Qual dos dois estimadores pode ser considerado como o **melhor** para estimar o verdadeiro valor de μ ?

Considere os seguintes pseudo-códigos para um estudo de simulação do comportamento destes dois estimadores:

Pseudo-código 1

- 1 Simule uma amostra de tamanho $n = 10$ da distribuição considerada
- 2 Para essa amostra, calcule a média ($\hat{\theta}_1$) e o ponto médio ($\hat{\theta}_2$)
- 3 Repita os passos (1) e (2) acima $m = 1000$ vezes
- 4 Faça um gráfico da densidade das $m = 1000$ estimativas de $\hat{\theta}_1$ e $\hat{\theta}_2$ e verifique seu comportamento

Pseudo-código 2

- 1 Simule amostras de tamanhos (n) 2, 3, 5, 10, 20, 50, 100, 500, 1000 da distribuição considerada
- 2 Para cada amostra de tamanho n , calcule a média ($\hat{\theta}_1$) e o ponto médio ($\hat{\theta}_2$)
- 3 Repita os passos (1) e (2) acima $m = 100$ vezes
- 4 Faça um gráfico das $m = 100$ estimativas de $\hat{\theta}_1$ e $\hat{\theta}_2$ para cada tamanho de amostra n e verifique seu comportamento

Estimação pontual

Pseudo-código 1 - $X \sim N(3, 1)$

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual

Propriedades

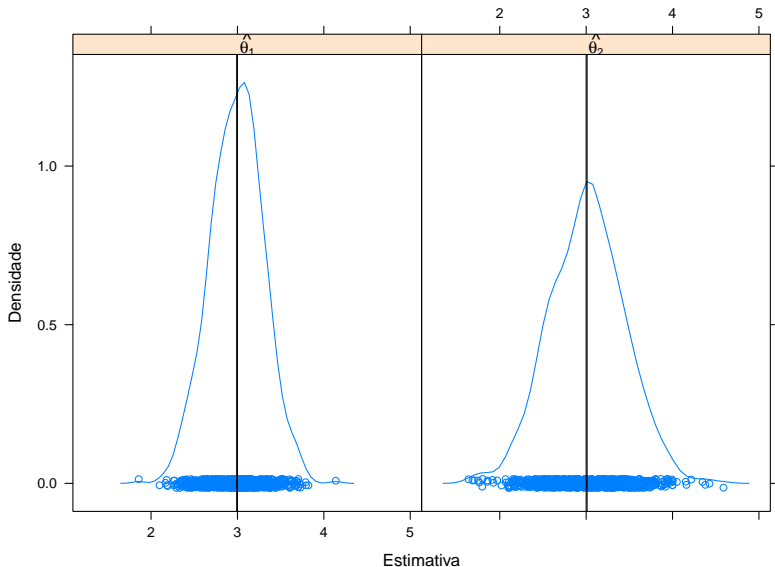
Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências



Estimação pontual

Pseudo-código 2 - $X \sim N(3, 1)$

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual

Propriedades

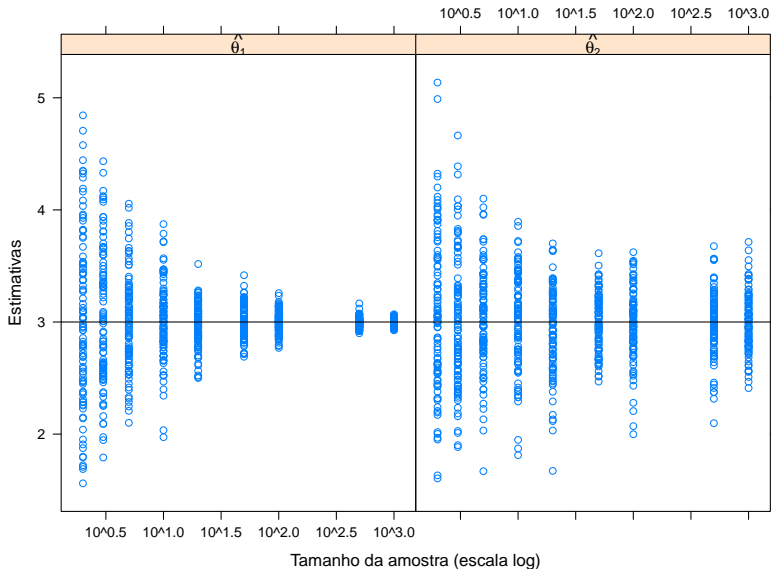
Erros amostrais

Distribuições amostrais

Distribuição amostral da média

Distribuição amostral da proporção

Referências



Exemplo 2: Considere uma amostra aleatória (X_1, \dots, X_n) de uma variável aleatória $Y \sim U(\min = 2, \max = 4)$ (distribuição uniforme no intervalo $[2,4]$) e os estimadores pontuais para μ

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{e} \quad \hat{\theta}_2 = \frac{X_{(1)} + X_{(n)}}{2}$$

Qual dos dois estimadores pode ser considerado como o **melhor** para estimar a média de Y ?

Estimação pontual

Pseudo-código 1 - $Y \sim U(2, 4)$

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual

Propriedades

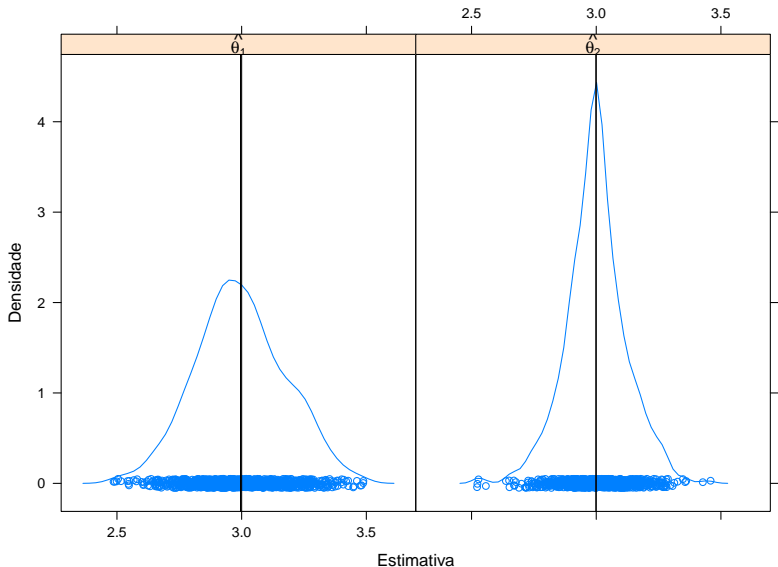
Erros amostrais

Distribuições amostrais

Distribuição amostral da média

Distribuição amostral da proporção

Referências



Estimação pontual

Pseudo-código 2 - $Y \sim U(2, 4)$

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual

Propriedades

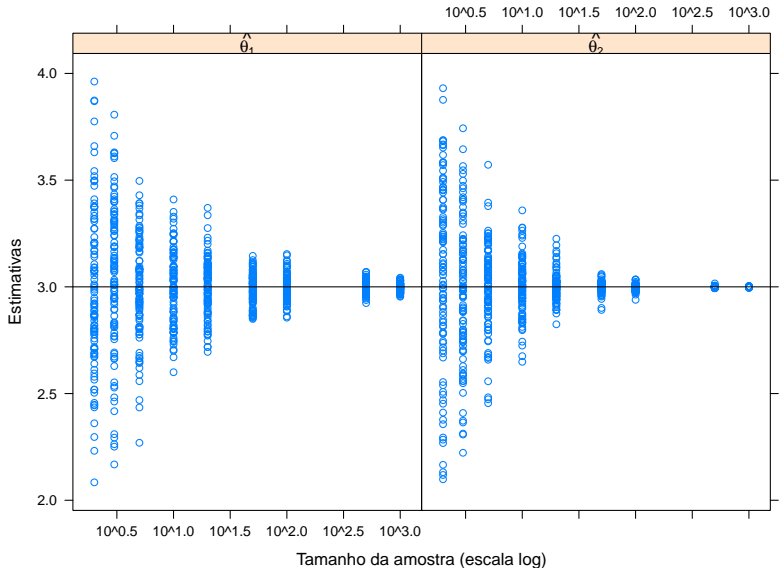
Erros amostrais

Distribuições amostrais

Distribuição amostral da média

Distribuição amostral da proporção

Referências



1 Introdução

2 Estimação pontual

- Propriedades dos estimadores

3 Erros amostrais

4 Distribuições amostrais

- Distribuição amostral da média
- Distribuição amostral da proporção

5 Referências

De modo geral, um “**bom**” estimador deve ser

- 1 Não viciado
- 2 Consistente
- 3 Eficiente

Propriedades dos estimadores

1) Não viciado

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual

Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Definição (Erro quadrático médio (EQM))

O Erro Quadrático Médio (EQM) de um estimador $\hat{\theta}$ de θ é dados por

$$\begin{aligned}\text{EQM}[\hat{\theta}] &= E[(\hat{\theta} - \theta)^2] \\ &= \text{Var}[\hat{\theta}] + B[\hat{\theta}]^2\end{aligned}$$

onde

$$B[\hat{\theta}] = E[\hat{\theta}] - \theta$$

é denominado de **vício** do estimador $\hat{\theta}$. Portanto, dizemos que um estimador é **não viciado** para θ quando

$$B[\hat{\theta}] = 0 \quad \Rightarrow \quad E[\hat{\theta}] = \theta$$

Propriedades dos estimadores

1) Não viciado

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual

Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média
Distribuição
amostral da
proporção

Referências

Definição (Estimador não viciado)

Seja (X_1, \dots, X_n) , uma amostra aleatória de uma variável aleatória com fdp ou fp $f(x, \theta)$, $\theta \in \Theta$, dizemos que o estimador $\hat{\theta} = T(\mathbf{X})$ é não viciado para θ se

$$E[\hat{\theta}] = E[T(\mathbf{X})] = \theta \quad \forall \theta \in \Theta$$

Um estimador $\hat{\theta}$ é dito **assintoticamente não viciado** se

$$\lim_{n \rightarrow \infty} E[\hat{\theta}] = \theta$$

Ou seja, para grandes amostras, $\hat{\theta}$ passa a ser imparcial.

Definição (Estimador consistente)

Seja (X_1, \dots, X_n) , uma amostra aleatória de uma variável aleatória com fdp ou fp $f(x, \theta)$, $\theta \in \Theta$, o estimador $\hat{\theta} = T(\mathbf{X})$ é consistente para θ se satisfaz simultaneamente

$$\lim_{n \rightarrow \infty} E[\hat{\theta}] = \theta$$

e

$$\lim_{n \rightarrow \infty} \text{Var}[\hat{\theta}] = 0$$

Definição (Eficiência relativa)

Sejam $\hat{\theta}_1 = T_1(\mathbf{X})$ e $\hat{\theta}_2 = T_2(\mathbf{X})$ dois estimadores pontuais **não viciados** para θ . A eficiência relativa de $\hat{\theta}_1$ em relação a $\hat{\theta}_2$ é

$$ER[\hat{\theta}_1, \hat{\theta}_2] = \frac{\text{Var}[\hat{\theta}_1]}{\text{Var}[\hat{\theta}_2]}$$

Se:

- $ER[\hat{\theta}_1, \hat{\theta}_2] > 1 \Rightarrow \hat{\theta}_2$ é mais eficiente
- $ER[\hat{\theta}_1, \hat{\theta}_2] < 1 \Rightarrow \hat{\theta}_1$ é mais eficiente

Exemplo: média amostral $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ como estimador da média populacional μ :

$$E(\bar{X}) = E\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \mu$$

$$\text{Var}(\bar{X}) = \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{\sigma^2}{n}$$

Portanto \bar{X} é um estimador **não viciado** e **consistente** para μ .

Propriedades dos estimadores

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual

Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Exemplo: variância amostral $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ como estimador da variância populacional σ^2 :

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

Portanto $\hat{\sigma}^2$ é um estimador **viciado** para σ^2 . Para aliminar esse vício, definimos então um novo estimador: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$, e

$$E(s^2) = E \left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right] = \sigma^2$$

que é então um estimador **não viciado** para σ^2 .

O **erro padrão** de um estimador dá uma ideia da **precisão** da estimativa.

Definição (Erro padrão de um estimador)

O erro padrão (EP) de um estimador é o seu desvio-padrão, ou seja,

$$EP(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}$$

Exemplo: Sabemos que a distribuição de \bar{X} é normal com média μ e variância σ^2/n . Então o erro padrão de \bar{X} é

$$EP(\bar{X}) = \sqrt{\text{Var}(\bar{X})} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual Propriedades

Erros amostrais

Distribuições amostrais

Distribuição amostral da média Distribuição amostral da proporção

Referências

- 1 Introdução
- 2 Estimação pontual
 - Propriedades dos estimadores
- 3 Erros amostrais
- 4 Distribuições amostrais
 - Distribuição amostral da média
 - Distribuição amostral da proporção
- 5 Referências

Erros amostrais

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média
Distribuição
amostral da
proporção

Referências

Erros amostrais

Diferença entre o resultado da amostra e o verdadeiro valor da população. Ocorre pois as amostras são **aleatórias**!

Exemplo: a diferença entre a média amostral \bar{X} e a média populacional μ

$$e = \bar{X} - \mu$$

é chamada de *erro amostral da média*.

Erros não amostrais

Ocorre quando os dados amostrais são coletados **incorretamente**, devido a uma *amostra tendenciosa*, instrumento de medida defeituoso, anotações erradas, ...

Erros amostrais

Erros amostrais

Diferença entre o resultado da amostra e o verdadeiro valor da população. Ocorre pois as amostras são **aleatórias**!

Exemplo: a diferença entre a média amostral \bar{X} e a média populacional μ

$$e = \bar{X} - \mu$$

é chamada de *erro amostral da média*.

Erros não amostrais

Ocorre quando os dados amostrais são coletados **incorretamente**, devido a uma *amostra tendenciosa*, instrumento de medida defeituoso, anotações erradas, ...

Atenção!

Os erros não amostrais não devem existir, ou devem ser minimizados

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média
Distribuição
amostral da
proporção

Referências

Não importa quão bem a amostra seja coletada, os **erros amostrais** sempre irão ocorrer

Cada vez que uma amostra aleatória for retirada de uma população, um resultado diferente será observado

Selecione uma amostra de tamanho $n = 5$ das idades dos estudantes de uma sala: 22 21 24 23 20 22 21 25 24 24 23 19 25 24
23 23 20 21 23 20 23 22 23 23 25 25 20 23 24 20

Repita 5 vezes (tente ser o mais aleatório possível!), calcule a média de cada amostra, e compare com a média populacional $\mu = 22,5$

Um exemplo

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual Propriedades

Erros amostrais

Distribuições amostrais

Distribuição amostral da média Distribuição amostral da proporção

Referências

Amostra	\bar{x}	$e = \bar{x} - \mu$
23 23 23 24 23	23.2	0.7
24 22 20 20 20	21.2	-1.3
21 20 19 22 25	21.4	-1.1
22 23 25 20 22	22.4	-0.1
21 20 22 24 20	21.4	-1.1

- O que isso nos diz a respeito das médias amostrais?
- O que isso nos diz a respeito da variabilidade das médias amostrais?
- E se fizemos uma “média das médias” de todas as amostras?

1 Introdução

2 Estimação pontual

- Propriedades dos estimadores

3 Erros amostrais

4 Distribuições amostrais

- Distribuição amostral da média
- Distribuição amostral da proporção

5 Referências

Suponha que vamos retirar uma amostra de $n = 100$ indivíduos de uma população

Se selecionarmos aleatoriamente um indivíduo desta população, ele terá apenas um valor, x_1 , de todos os possíveis valores da variável aleatória X_1

Da mesma forma, um segundo indivíduo amostrado aleatoriamente terá o valor x_2 da variável aleatória X_2 , e assim sucessivamente até o centésimo indivíduo amostrado com valor x_{100} da variável aleatória X_{100}

De maneira geral, uma amostra de tamanho n será descrita pelos valores x_1, x_2, \dots, x_n das variáveis aleatórias $X_1, X_2, \dots, X_n \Rightarrow$

Amostra Aleatória

No caso de uma Amostragem Aleatória Simples (AAS) **com reposição**, X_1, X_2, \dots, X_n serão variáveis aleatórias **independentes e identicamente distribuídas** (iid) com função de probabilidade (fp) ou função densidade de probabilidade (fdp) $f(x)$

Isto significa que quando observamos cada amostra x_i de uma população indexada por um parâmetro θ (um escalar ou um vetor), então cada observação possui fp ou fdp dada por $f(x, \theta)$

Se somente uma observação X é feita, então as probabilidades referentes a X podem ser calculadas diretamente utilizando $f(x, \theta)$

No entanto, na maioria das vezes temos $n > 1$ observações de X . Como vimos que as variáveis X_i são iid, temos que a fp ou fdp conjunta será

$$f(x_1, x_2, \dots, x_n, \theta) = f(x_1, \theta) \cdot f(x_2, \theta) \cdots f(x_n, \theta) = \prod_{i=1}^n f(x_i, \theta)$$

Onde o mesmo valor do parâmetro θ é utilizado em cada um dos termos no produto

Exemplo: distribuição conjunta da Bernoulli(π)

Para uma observação, temos que a fp da Bernoulli(π) é

$$f(x, \pi) = \pi^x (1 - \pi)^{1-x} \mathbb{I}_{\{0,1\}}(x)$$

Para uma amostra aleatória X_1, X_2, \dots, X_n

$$\begin{aligned} f(\mathbf{x}, \pi) &= \prod_{i=1}^n \pi^{x_i} (1 - \pi)^{1-x_i} \mathbb{I}_{\{0,1\}}(x_i) \\ &= \pi^{\sum_{i=1}^n x_i} (1 - \pi)^{n - \sum_{i=1}^n x_i} \prod_{i=1}^n \mathbb{I}_{\{0,1\}}(x_i) \end{aligned}$$

Quando uma amostra X_1, X_2, \dots, X_n é obtida, geralmente estamos interessados em um resumo destes valores, que pode ser expresso matematicamente pela estatística $T(x_1, x_2, \dots, x_n)$

A função $T(\cdot)$ pode ser um valor real ou um vetor. Dessa forma, $Y = T(x_1, x_2, \dots, x_n)$ **é também uma variável aleatória** (ou vetor aleatório). Se Y é uma VA, *então ela possui uma distribuição de probabilidade*.

Uma vez que a amostra aleatória X_1, X_2, \dots, X_n tem uma estrutura probabilística simples (porque X_i são iid), Y é particularmente tratável. Uma vez que a distribuição de Y é derivada desta estrutura, vamos denominá-la de **distribuição amostral** de Y .

Definição (Distribuição amostral)

A distribuição de probabilidade de uma estatística $Y = T(x_1, x_2, \dots, x_n)$ é denominada de **distribuição amostral** de Y . Assim, uma estatística também é uma variável aleatória, pois seus valores mudam conforme a amostra aleatória

Exemplo: duas estatísticas comumente utilizadas para o resumo de uma amostra aleatória são a **média amostral**

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

e a **proporção amostral**

$$\hat{p} = \frac{x}{n}$$

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual Propriedades

Erros amostrais

Distribuições amostrais

Distribuição amostral da média

Distribuição amostral da proporção

Referências

- 1 Introdução
- 2 Estimação pontual
 - Propriedades dos estimadores
- 3 Erros amostrais
- 4 Distribuições amostrais
 - Distribuição amostral da média
 - Distribuição amostral da proporção
- 5 Referências

Distribuição amostral da média

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Para estudarmos a distribuição amostral da estatística \bar{X} , considere uma população identificada pela VA X , com parâmetros

$$E(X) = \mu = \text{média}$$

$$\text{Var}(X) = \sigma^2 = \text{variância}$$

supostamente conhecidos. Em seguida, realizamos os seguintes passos:

- 1 Retiramos m amostras aleatórias (AAS com reposição) de tamanho n dessa população
- 2 Para cada uma das m amostras, calculamos a média amostral \bar{x}
- 3 Verificamos a distribuição das m médias amostrais e estudamos suas propriedades

Voltando ao exemplo ...

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual

Propriedades

Erros amostrais

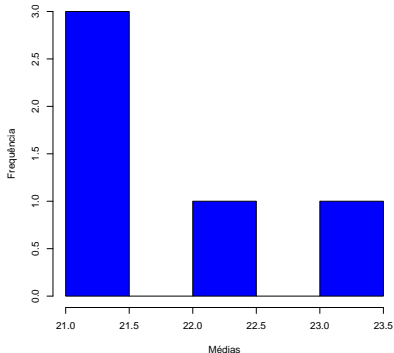
Distribuições amostrais

Distribuição amostral da média

Distribuição amostral da proporção

Referências

Amostra	\bar{x}	$\epsilon = \bar{x} - \mu$
23 23 23 24 23	23.2	0.7
24 22 20 20 20	21.2	-1.3
21 20 19 22 25	21.4	-1.1
22 23 25 20 22	22.4	-0.1
21 20 22 24 20	21.4	-1.1



Do exemplo anterior, temos que $\mu = 22,5$, e $\sigma^2 = 3,09$

Para esta tabela, com $m = 5$ e $n = 5$:

- A média das médias é $\mu_{\bar{X}} = 21,9$
- A variância das médias é $\sigma_{\bar{X}}^2 = 0,732$

Distribuição amostral da média

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual
Propriedades

Erros amostrais

Distribuições amostrais

Distribuição amostral da média

Distribuição amostral da proporção

Referências

E se pudéssemos retirar **todas** as amostras **com reposição** de tamanho $n = 5$ dessa população???

Teríamos que fazer $N^n = 20^5 = 3.200.000$ amostragens!

Para $n = 10 \Rightarrow N^n = 20^{10} = 1,024 \times 10^{13}$

Para $n = 15 \Rightarrow N^n = 20^{15} = 3,2768 \times 10^{19}$

O computador pode fazer isso, e o resultado é (para $n = 15$)

- $\mu_{\bar{X}} = 22,5$
- $\sigma_{\bar{X}}^2 \approx 0,2 = \sigma^2/n \approx 3,09/15$

Conclusão:

- A média de **todas** as médias é igual à média da população!
- A variância das médias é menor porque a variabilidade entre as médias é menor!

Veja a figura `dist_amostrais_idades.pdf`

- O primeiro gráfico é a distribuição da população original
- O segundo gráfico é a distribuição de 1000 médias, calculadas a partir de 1000 amostras de tamanho 5 ($m = 1000$ e $n = 5$)
- Os demais gráficos mostram a distribuição amostral de 1000 médias calculadas com amostras de tamanho $n = 10$ e $n = 15$
- Repare que:
 - A distribuição das 1000 médias se torna cada vez mais próxima de uma normal, conforme o tamanho da amostra aumenta
 - A variabilidade da distribuição amostral das médias diminui conforme o tamanho da amostra aumenta
 - A distribuição amostral tende a se concentrar cada vez mais em torno da média populacional verdadeira

Distribuição amostral da média

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Através do estudo da distribuição da média amostral chegamos em um dos resultados mais importantes da inferência estatística

Teorema (Distribuição amostral da média)

- $E(\bar{X}) = \mu_{\bar{X}} = \mu$
- $\text{Var}(\bar{X}) = \sigma_{\bar{X}}^2 = \sigma^2/n$

Portanto, se

$$X \sim N(\mu, \sigma^2) \quad \text{então} \quad \bar{X} \sim N(\mu_{\bar{X}}, \sigma_{\bar{X}}^2)$$

mas, como

$$\mu_{\bar{X}} = \mu \quad \text{e} \quad \sigma_{\bar{X}}^2 = \sigma^2/n$$

então, a **distribuição amostral** da média amostral \bar{X} é

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Teorema (Teorema Central do Limite (TCL))

Para amostras aleatórias simples (X_1, X_2, \dots, X_n) , retiradas de uma população normal com média μ e variância σ^2 , a distribuição amostral da média \bar{X} , terá forma dada por

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

no limite quando $n \rightarrow \infty$, que é a distribuição normal padrão:
 $Z \sim N(0, 1)$.

- Se a população for normal, então \bar{X} terá distribuição *exata* normal.
- A rapidez da convergência para a normal depende da distribuição da população da qual as amostras foram geradas

Distribuição amostral da média

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Este teorema nos mostra que, para amostras suficientemente grandes ($n > 30$), **a média amostral \bar{X} converge para o verdadeiro valor da média populacional μ (é um estimador não viesado de μ)**

Além disso, a variância das médias amostrais $\sigma_{\bar{X}}^2$ tende a diminuir conforme $n \rightarrow \infty$ (é um estimador **consistente**)

Estes resultados sugerem que, quando o tamanho da amostra aumenta,

independente do formato da distribuição da população original,

a distribuição amostral de \bar{X} aproxima-se cada vez mais de uma distribuição normal, um resultado fundamental na teoria de probabilidade conhecido como **Teorema Central do Limite**

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Exemplo computacional → veja a figura `dist_amostrais.pdf`

Em palavras, o teorema garante que para n grande, a distribuição da média amostral, devidamente padronizada, **se comporta segundo um modelo normal** com média 0 e variância 1.

Pelo teorema, temos que quanto maior o tamanho da amostra, **melhor é a aproximação**.

Estudos envolvendo simulações mostram que, em muitos casos, **valores de n ao redor de 30** fornecem aproximações bastante boas para as aplicações práticas.

Quando calculamos a probabilidade de um valor estar em um determinado intervalo de valores, podemos usar o modelo Normal, como vimos anteriormente.

No entanto, quando temos uma **amostra**, e queremos calcular probabilidades associadas à **média amostral** (a probabilidade da média amostral estar em um determinado intervalo de valores), precisamos necessariamente usar os resultados do TCL.

Já vimos que o **erro amostral da média** é dado pela diferença entre \bar{X} e μ , ou seja,

$$e = \bar{X} - \mu$$

Dessa forma, se

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

então a distribuição de e também será normal padrão, pois

$$\frac{e\sqrt{n}}{\sigma} \sim N(0, 1)$$

Esse resultado será fundamental na construção de estimativas intervalares.

Usando o TCL

Exemplo: Uma máquina de empacotamento que abastece pacotes de feijão apresenta distribuição normal com média de 500 g e desvio-padrão de 22 g. De acordo com as normas de defesa do consumidor, os pacotes de feijão não podem ter peso inferior a 2% do estabelecido na embalagem.

- Determine a probabilidade de **um pacote** selecionado aleatoriamente ter a peso inferior a 490 g.
- Determine a probabilidade de **20 pacotes** selecionados aleatoriamente terem peso médio inferior a 490 g.
- Como podemos interpretar os resultados dos itens anteriores? O que é mais indicado para se tomar uma decisão sobre o funcionamento da máquina: selecionar um pacote ou uma amostra?

Usando o TCL

Exemplo: Uma pesquisa com 12000 estudantes mostrou que a média de horas de estudo por semana foi de 7,3 horas, com desvio-padrão de 4,2 horas. **O tempo de estudo não apresenta distribuição normal.** Com isso calcule:

- A probabilidade de que **um** estudante exceda 8 horas de estudo por semana.
- Dada uma amostra de 45 estudantes, a probabilidade de que o **tempo médio** de estudo exceda 8 horas por semana.
- Dada uma amostra de 45 estudantes, a probabilidade de que o **tempo médio** de estudo seja igual ou superior a 7 horas por semana.

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual Propriedades

Erros amostrais

Distribuições amostrais

Distribuição amostral da média

Distribuição amostral da proporção

Referências

1 Introdução

2 Estimação pontual

- Propriedades dos estimadores

3 Erros amostrais

4 Distribuições amostrais

- Distribuição amostral da média
- Distribuição amostral da proporção

5 Referências

Distribuição amostral da proporção

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Muitas vezes, o interesse é conhecer uma **proporção**, e não a média de uma população.

Suponha que uma amostra de tamanho n foi obtida de uma população, e que $x \leq n$ observações nessa amostra pertençam a uma classe de interesse (ex.: pessoas do sexo masculino).

Dessa forma, a proporção amostral

$$\hat{p} = \frac{x}{n} = \frac{\text{número de sucessos}}{\text{total de tentativas}}$$

é o “melhor estimador” para a proporção populacional p .

Note que n e p são os parâmetros de uma **distribuição binomial**.

Exemplo: em 5 lançamentos de uma moeda considere que o evento “cara” (C) seja o sucesso (“sucesso” = 1; “fracasso” = 0). Um possível resultado seria o conjunto {C, C, R, R, C}. A proporção amostral seria

$$\hat{p} = \frac{x}{n} = \frac{\text{número de sucessos}}{\text{total de tentativas}} = \frac{3}{5} = 0,6$$

Exemplo: em uma amostra de 2500 eleitores de uma cidade, 1784 deles eram favoráveis à reeleição do atual prefeito. A proporção amostral é então

$$\hat{p} = \frac{x}{n} = \frac{\text{número de sucessos}}{\text{total de tentativas}} = \frac{1784}{2500} = 0,7136$$

Distribuição amostral de uma proporção

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

A distribuição amostral de uma **proporção** é a distribuição das proporções de todas as possíveis amostras de tamanho n retiradas de uma população

Ver figura `dist_amostal_proporcoes.pdf`:

- Uma moeda é lançada $n = 10$ vezes, e a proporção de caras é registrada
- Esse processo é repetido $m = 10, 30, 100, 1000, 10000$ vezes

Com isso, concluímos que:

- A média das proporções para $m \rightarrow \infty$ tende para a verdadeira proporção populacional $p = 0,5$
- A **distribuição amostral** das proporções segue aproximadamente uma **distribuição normal**

Distribuição amostral de uma proporção

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Através do estudo da distribuição amostral da proporção, chegamos aos seguintes resultados

- $E(\hat{p}) = \mu_{\hat{p}} = p$
- $\text{Var}(\hat{p}) = \sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$

Ou seja, \hat{p} é um estimador **não viciado** e **consistente** para p .

Assim, a **distribuição amostral** de \hat{p} será

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

Distribuição amostral de uma proporção

Estimação
pontual e
distribuições
amostrais

Introdução

Estimação
pontual
Propriedades

Erros
amostrais

Distribuições
amostrais

Distribuição
amostral da
média

Distribuição
amostral da
proporção

Referências

Note que o **erro padrão** de \hat{p} será

$$EP(\hat{p}) = \sqrt{\text{Var}(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$$

Assim, usando o TCL, podemos mostrar que a quantidade

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

segue uma distribuição **normal padrão** com média 0 e variância 1.

Quando não conhecemos p , usamos $\hat{p} = x/n$ como estimativa para calcular o erro padrão.

A normal como aproximação da binomial

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual
Propriedades

Erros amostrais

Distribuições amostrais

Distribuição amostral da média

Distribuição amostral da proporção

Referências

Sob determinadas condições, podemos usar a distribuição normal como aproximação da distribuição binomial.

Se X for uma VA binomial com parâmetros n e p , então

$$Z = \frac{X - np}{\sqrt{np(1-p)}}$$

será uma VA **normal padrão**, $Z \sim N(0, 1)$, desde que as seguintes condições sejam satisfeitas:

- $np \geq 5$
- $n(1-p) \geq 5$

Dessa forma, podemos calcular probabilidades para uma VA binomial, aproximadas por uma distribuição normal com média $\mu = np$ e desvio-padrão $\sigma = \sqrt{np(1-p)}$.

Estimação pontual e distribuições amostrais

Introdução

Estimação pontual Propriedades

Erros amostrais

Distribuições amostrais

Distribuição amostral da média Distribuição amostral da proporção

Referências

- 1 Introdução
- 2 Estimação pontual
 - Propriedades dos estimadores
- 3 Erros amostrais
- 4 Distribuições amostrais
 - Distribuição amostral da média
 - Distribuição amostral da proporção
- 5 Referências

- Bussab, WO; Morettin, PA. **Estatística básica**. São Paulo: Saraiva, 2006. [Cap. 10]
- Magalhães, MN; Lima, ACP. **Noções de Probabilidade e Estatística**. São Paulo: EDUSP, 2008. [Cap. 7]
- Montgomery, DC; Runger, GC. **Estatística aplicada e probabilidade para engenheiros**. Rio de Janeiro: LTC Editora, 2012. [Cap. 7]