Introdução à Probabilidade e à Estatística II

Introdução à Inferência Estatística

Capítulo 10, Estatística Básica (Bussab&Morettin, 7a Edição)

Lígia Henriques-Rodrigues

MAE0229 – 1º semestre 2018

Contexto

A Teoria de Probabilidade e a Inferência Estatística são processos ‘complementares”:

- **Teoria da probabilidade**: parte-se de um modelo totalmente especificado, que se assume como correto e se calcula, e.g., as probabilidades de certos acontecimentos;
- **Inferência estatística**: Observam-se certos acontecimentos, e procura inferir-se sobre o modelo probabilístico pelo qual se regirá o experimento aleatório.

Exemplo: Considere-se um grupo numeroso de pessoas entre as quais há uma proporção θ de fumadores.

- Se θ conhecido e estivermos interessados em conhecer a probabilidade de encontrar x fumadores num grupo de 10 pessoas escolhidas ao acaso

  Teoria da Probabilidade

- Na prática, sucede quase sempre que θ é desconhecido. A partir da observação do número de fumadores na amostra de 10 pessoas, pretende-se tirar conclusões sobre a proporção de fumadores na população, θ

  Inferência Estatística

População e Amostra

População: É o conjunto de todos os elementos ou resultados sobre investigação.

Amostra: É qualquer subconjunto da população.

Exemplos:

- Consideramos uma pesquisa para estudar os salários dos 500 funcionários da Companhia MB. Seleciona-se uma amostra de 36 indivíduos, e anotam-se os seus salários.
- Consideramos uma pesquisa para estudar a duração de vida útil de um novo tipo de lâmpadas, pois acredita-se que a duração desse novo tipo é maior. Então 100 lâmpadas do novo tipo são deixadas acesas até queimarem.

Parâmetros

Definição: Um parâmetro é uma medida usada para descrever uma característica da população.

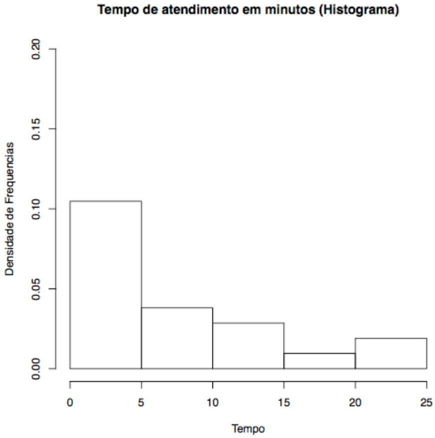
Para que servem os parâmetros?

Como encontrar valores para os parâmetros?

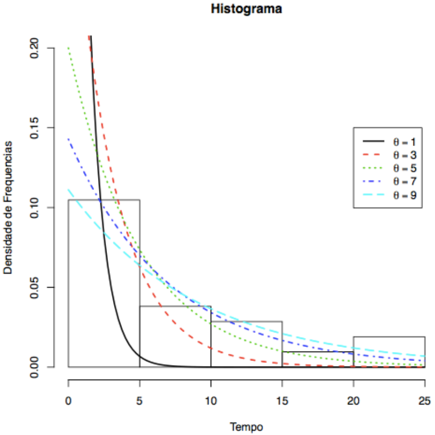
Exemplo: Considere o tempo de atendimento X (em minutos) de um cliente em um supermercado.

Assuma que foi retirada uma amostra de tamanho 20 e foram observados os seguintes valores:

$$\{7, 5, 6, 21, 1, 3, 7, 4, 5, 18, 24, 4, 5, 2, 1, 2, 6, 5, 14, 11, 12\}$$



Pelo histograma, parece que a distribuição subjacente aos dados é exponencial, i.e., $f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}$, $\theta > 0$. Mas qual será o valor de θ?



Qual o valor do parâmetro θ para o qual a curva estimada fica mais próxima do histograma?

Podemos observar o comportamento da amostra e comparar com o que deveríamos esperar na população.

A média da amostra (tempo de atendimento em minutos) abaixo $\{7, 5, 6, 21, 1, 3, 7, 4, 5, 18, 24, 4, 5, 2, 1, 2, 6, 5, 14, 11, 12\}$ é

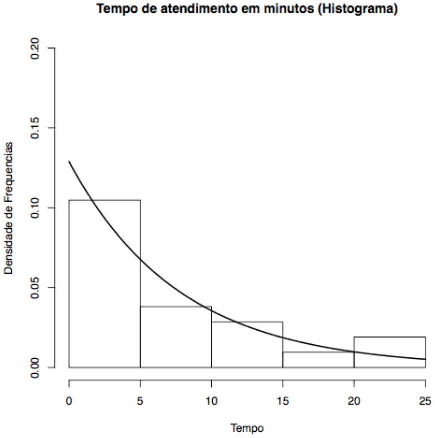
$$\overline{x} = 7,76.$$

Com a parametrização utilizada, sabemos que, se X segue distribuição exponencial, então

$$E(X) = \theta,$$

pelo que esperamos que a média amostral seja próxima da média populacional!!

Usando como valor aproximado de θ a quantidade $\overline{x} = 7,76$, ajustamos a curva $f(x) = 0,13e^{-0,13x}$



Inferência Estatística

O objetivo da **Inferência Estatística** é produzir afirmações sobre dada característica da população, na qual estamos interessados, a partir de informações recolhidas de uma parte dessa população.

A inferência clássica procura responder a questões do tipo:

- Admitindo a validade do modelo, como escolher um ou mais elementos do modelo que representem adequadamente os parâmetros desconhecidos à custa da informação contida nos dados? Neste ponto insere-se a estimação pontual e intervalar.

- Os dados x são compatíveis com o modelo teórico?
  Neste ponto inserem-se os chamados testes de hipóteses.

Amostragem Aleatória Simples (AAS)

Aleatoriamente sorteia-se um elemento da população, sendo que todos os elementos têm a mesma chance de ser escolhidos. Repete-se o procedimento até que sejam sorteadas as n unidades da amostra.

AAS com/sem reposição: com reposição implica a propriedade de independência entre unidades selecionadas. Isso facilita o tratamento matemático de propriedades de estimadores que vamos construir em cima da amostra.

Definição: Uma amostra aleatória simples (AAS) de tamanho n de uma variável aleatória X, com dada distribuição, é o conjunto de n variáveis aleatórias independentes $X_1,\ X_2, \ldots,\ X_n$, cada uma com a mesma distribuição de X, ou seja, identicamente distribuídas,

$$(X_1, \ldots, X_n)\text{: Amostra aleatória simples}$$
$$(x_1, \ldots, x_n)\text{: Amostra observada}$$

Estatísticas

Definição: Uma estatística é uma característica da amostra, ou seja, é qualquer função da amostra aleatória que não depende de parâmetros desconhecidos.

Exemplos:

- A média amostral $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ é uma estatística;
- A variância amostral $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ é uma estatística;
- $X_{(1)} = \min(X_1,\ X_2, \ldots,\ X_n)$ é uma estatística;
- $X_{(n)} = \max(X_1,\ X_2, \ldots,\ X_n)$ é uma estatística.

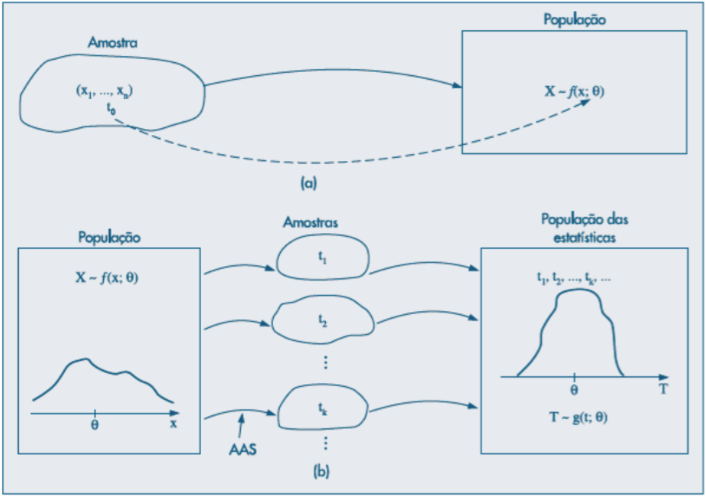
Nota: As estatísticas são v.a.'s!!!

As estatísticas são v.a's e a sua distribuição amostral desempenha um papel fundamental na teoria da inferência estatística.

Esquema:

- uma população X, com determinado parâmetro de interesse θ;
- todas as amostras retiradas da população, via AAS;
- para cada amostra, calculamos o valor observado, t, da estatística T;
- os novos valores t formam uma nova população, cuja distribuição recebe o nome de distribuição amostral de T.

Figura 10.1: (a) Esquema de inferência sobre θ.
(b) Distribuição amostral da estatística T.



[Fonte: Bussab & Morettin]

Exemplo 1: Variáveis Bernoulli

Seja X a v.a. que vale 1 se o indivíduo é portador da característica A e zero caso contrário.

Usamos o modelo estatístico Bernoulli, com $\theta \in \Theta = (0, 1)$, com:

$$P_\theta(X = 1) = \theta \qquad \text{e} \qquad P_\theta(X = 0) = 1 - \theta,$$

sendo θ a probabilidade associada à característica A.

Para estimar θ criamos um experimento que produza uma amostra $X_1, \ldots, X_n$, sendo $X_i = 1$ se o indivíduo é portador da característica A e zero caso contrário.

Considere-se a estatística $\overline{X}$, a proporção (média) amostral e vejamos para $n = 3$ qual a distribuição de $\overline{X}$.

(X_1, X_2, X_3)	$\overline{X}$	$P(X_1, X_2, X_3)$
$(0,0,0)$	0	$(1-\theta)^3$
$(1,0,0)$	$1/3$	$\theta(1-\theta)^2$
$(0,1,0)$	$1/3$	$\theta(1-\theta)^2$
$(0,0,1)$	$1/3$	$\theta(1-\theta)^2$
$(1,1,0)$	$2/3$	$\theta^2(1-\theta)$
$(1,0,1)$	$2/3$	$\theta^2(1-\theta)$
$(0,1,1)$	$2/3$	$\theta^2(1-\theta)$
$(1,1,1)$	1	θ^3

Na prática podemos observar uma das possibilidades acima!

Neste caso, a identificação da distribuição associada à proporção amostral é fácil. Assim, e para um n geral,

$$P\left(\overline{X} = \frac{i}{n}\right) = \binom{n}{i}\theta^i(1-\theta)^{n-i}, \quad i = 1, 2, \ldots, n,$$

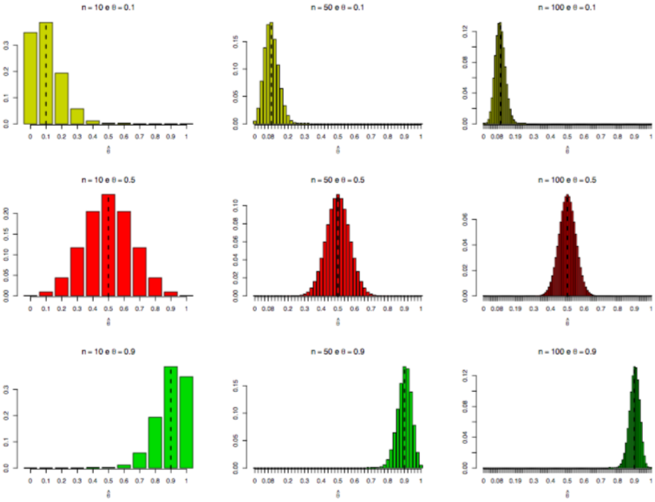
ou seja

$$n\overline{X} \sim Bin(n, \theta).$$

Conseguimos derivar a distribuição da média amostral para variáveis de Bernoulli.

Porém, observamos certa dificuldade em encontrar a distribuição da média amostral para outros tipos de variáveis (algumas variáveis discretas e contínuas).

Veremos que quando o tamanho amostral n é grande, poderemos aproximar a distribuição da média amostral pela distribuição normal.


n = 10 e θ = 0.1
n = 50 e θ = 0.1
n = 100 e θ = 0.1
n = 10 e θ = 0.5
n = 50 e θ = 0.5
n = 100 e θ = 0.5
n = 10 e θ = 0.9
n = 50 e θ = 0.9
n = 100 e θ = 0.9
$\hat{\theta}$

Exemplo 2 (Exemplo 10.7 de Bussabb & Morettin)

Considere uma população em que a variável X assume um dos valores do conjunto $\{1, 3, 5, 7\}$. A f.p da v.a. X é

x	1	3	5	7
$P(X = x)$	$1/5$	$1/5$	$2/5$	$1/5$

É fácil ver que $\mu_X = E(X) = 4,2$ e que $\sigma_X^2 = Var(X) = 4,16$.

Considere todas as amostras possíveis de tamanho 2 recolhidas com reposição. Obtenha a distribuição da média amostral,

$$\overline{X} = \frac{X_1 + X_2}{2},$$

sendo

X_1 : valor selecionado na 1º extração

X_2 : valor selecionado na 2º extração

Amostra (X_1, X_2)	Probabilidade	Média amostral
(1,1)	1/25	1
(1,3)	1/25	2
(1,5)	2/25	3
(1,7)	1/25	4
(3,1)	1/25	2
(3,3)	1/25	3
(3,5)	2/25	4
(3,7)	1/25	5
(5,1)	2/25	3
(5,3)	2/25	4
(5,5)	4/25	5
(5,7)	2/25	6
(7,1)	1/25	4
(7,3)	1/25	5
(7,5)	2/25	6
(7,7)	1/25	7

A distribuição de probabilidade de $\overline{X}$ para $n = 2$ é

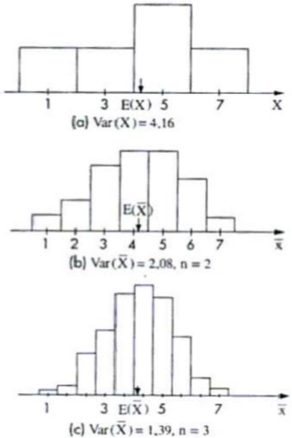
$\overline{x}$	1	2	3	4	5	6	7
$P(\overline{X} = \overline{x})$	1/25	2/25	5/25	6/25	6/25	4/25	1/25

Neste caso,

$$E(\overline{X}) = 4,2 = \mu_X \qquad \text{e} \qquad Var(\overline{X}) = 2,08 = \frac{\sigma_X^2}{2}.$$

Poderíamos repetir o mesmo procedimento para amostras de tamanho $n = 3$, ou $n = 9$ ou $n = 100$

Figura: Histogramas das distribuições de X e de $\overline{X}$ para amostras de dimensão $n = 2$ e $n = 3$ da população $\{1, 3, 5, 7\}$.

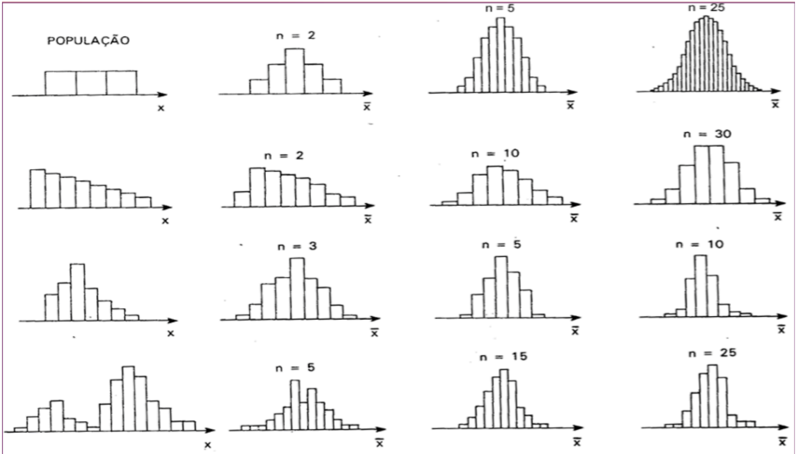


[Fonte: Bussab & Morettin]

Dos histogramas, observamos que

- à medida que n aumenta, os valores de $\overline{X}$ tendem a se concentrar cada vez mais em torno de
$$E(\overline{X}) = \mu_X = 4,2$$
uma vez que a variância vai diminuindo;
- os casos extremos passam a ter pequena probabilidade de ocorrência;
- para n suficientemente grande, a forma do histograma aproxima-se de uma distribuição normal.

Figura: Histogramas das distribuições de $\overline{X}$ para amostras de algumas populações.



Fonte: BUSSAB & MORETTIN. ***Estatística Básica***. São Paulo, Atual, 4ª edição, 2002, pp. 273.

Esses gráficos sugerem que,

quando n aumenta, independentemente da forma da distribuição de X, a distribuição de probabilidade da média amostral aproxima-se de uma distribuição normal.

Propriedades da média e da variância

Seja X uma variável aleatória e $X_1, \ldots, X_n$ uma amostra aleatória, i.e., $X_1, \ldots, X_n$ são v.a.'s independentes e identicamente distribuídas de X.

Então:

$$E(\overline{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = E(X)$$

e

$$Var(\overline{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}Var\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{Var(X)}{n}$$

Teorema do Limite Central

Seja X uma v. a. que tem média μ e variância $\sigma^2 < \infty$. Para uma amostra aleatória $X_1, X_2, \ldots, X_n$, retirada ao acaso e com reposição de X, temos:

- Definindo $S_n = \sum_{i=1}^{n} X_i$, então a distribuição de probabilidade de S_n aproxima-se, para n grande, de uma distribuição normal, com média $n\mu$ e variância $n\sigma^2$, ou seja,

$$S_n \approx N\left(n\mu, n\sigma^2\right) \qquad \text{ou seja} \qquad \frac{S_n - n\mu}{\sqrt{n}\sigma} \approx N(0,1).$$

- A distribuição de probabilidade da média amostral $\overline{X} = S_n/n$ aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n, ou seja,

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \qquad \text{ou seja} \qquad \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

Nota: A quantidade $\dfrac{\sigma}{\sqrt{n}} = \sigma_{\overline{X}}$ é designada de erro padrão.

Teorma do Limite Central - caso Bernoulli

Sendo X uma v.a. Bernoulli de parâmetro p, i.e., $X \sim Ber(p)$ então

$$E(X) = p \qquad \text{e} \qquad Var(X) = p(1-p).$$

Logo, e denotando por $\overline{X} = \hat{p}$, a proporção amostral

$$E(\hat{p}) = p \qquad \text{e} \qquad Var(\hat{p}) = \frac{p(1-p)}{n}.$$

Assim, quando n é grande:

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right) \qquad \text{ou seja} \qquad \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \approx N(0,1).$$

Caso especial

Um caso especial ocorre para o modelo estatístico Normal, ou seja, quando $X \sim N(\mu, \sigma^2)$ e $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^+$. Neste caso temos que:

Seja $X_1, \ldots, X_n$ uma amostra aleatória de X e $\overline{X}$ a média amostral associada. Sabemos que, para qualquer $n \geq 1$,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Ou seja, quando assumimos o modelo normal, a distribuição de $\overline{X}$ será normal EXATA. Para outros casos temos apenas uma aproximação.

Dimensionamento da amostra

Chamamos de **erro amostral** (da média ou da proporção) e designamos por e a v.a. que mede a diferença entre a estatística e o parâmetro. Assim,

$$e = \overline{X} - \mu \qquad \text{ou} \qquad e = \hat{p} - p.$$

Tendo-se

$$e \approx N\left(0, \frac{\sigma^2}{n}\right) \qquad \text{ou} \qquad e \approx N\left(0, \frac{p(1-p)}{n}\right).$$

Suponhamos que pretendemos determinar o valor de n tal que a probabilidade da média amostral $\overline{X}$ estar a uma distância de, no máximo, ε da média populacional μ (desconhecida),

$$P(|\overline{X} - \mu| \leq \varepsilon) = \gamma$$

com $0 < \gamma < 1$ e ε é o erro amostral (observado) máximo que podemos suportar.

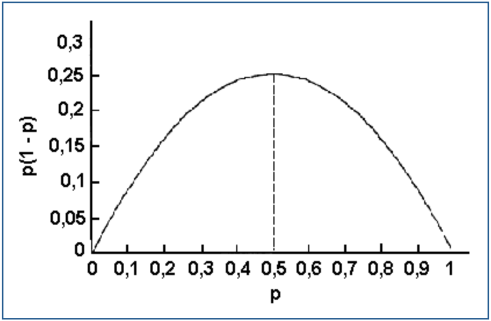
Definindo $\gamma = P(-z_\gamma < Z < z_\gamma)$, com $z_\gamma = \frac{\sqrt{n}\,\varepsilon}{\sigma}$, obtemos

$$n = \frac{\sigma^2\, z_\gamma^2}{\varepsilon^2}.$$

No caso de proporções,

$$n = \frac{p(1-p)\, z_\gamma^2}{\varepsilon^2}.$$

Observando o gráfico da função $p(1-p)$, para $0 \leq p \leq 1$



vemos que:

- a função $p(1-p)$ é uma parábola simétrica em torno de $p = 0,5$;
- a função atinge o seu máximo (0,25) quando $p = 0,5$.

Deste modo, na prática, **se nada sabemos sobre o valor de** p, substituimos $p(1-p)$ por 0,25 obtendo-se

$$n = 0,25 \left(\frac{z_\gamma}{\varepsilon}\right)^2,$$

que pode fornecer um valor de n maior do que o necessário.

- Se soubermos que $p \leq 0,20$, então o valor máximo de $p(1-p)$ é $0,2 \times 0,8 = 0,16$ e reduzimos o valor de n para $n = 0,16 \left(\frac{z_\gamma}{\varepsilon}\right)^2$.
- Se soubermos que $p \geq 0,70$, então o valor máximo de $p(1-p)$ é $0,7 \times 0,3 = 0,21$ e reduzimos o valor de n para $n = 0,21 \left(\frac{z_\gamma}{\varepsilon}\right)^2$.
- Se soubermos que $0,40 \leq p \leq 0,70$, então o valor máximo de $p(1-p)$ é $0,5 \times 0,5 = 0,25$ e, neste caso, não há redução, ou seja, $n = 0,25 \left(\frac{z_\gamma}{\varepsilon}\right)^2$.

Exercício 1:
Uma máquina está regulada para encher pacotes de café automaticamente segundo uma distribuição normal com média de 10 gramas e desvio padrão $\sigma = \sqrt{0.2}$ gramas. Qual é a probabilidade de encontrarmos a média $\overline{X}$ diferindo de 10 gramas de menos de 0.32 gramas.

Exercício 2: Suponha que 30% dos estudantes de uma escola sejam mulheres. Colhemos uma AAS de $n = 100$ estudantes e calculamos $\hat{p}$ = proporção de mulheres na amostra. Qual probabilidade de que $\hat{p}$ difira de p em menos de 0,01?

Exercício 3: A renda per-capita domiciliar numa certa região tem distribuição normal com desvio padrão $\sigma = 250$ reais e média μ desconhecida. Se desejamos estimar a renda média μ com erro $\varepsilon = 50$ reais e com uma confiança $\gamma = 95\%$, quantos domicílios devemos consultar?

Exercício 4: Suponha que se pretende estimar, com base em uma amostra, a proporção **p** de alunos da USP que foram ao teatro no último mês.

- Qual deve ser o tamanho da amostra para que o erro cometido ao se estimar p seja no máximo $0,02$ com um coeficiente de confiança de 95%?
- Suponha que temos informação de que no máximo 30% dos alunos da USP foram ao teatro no último mês. Qual a dimensão da amostra que devemos recolher? Houve redução?