Estatística com Apoio Computacional

Inferência Estatística

Universidade Estadual Vale do Acaraú – UVA

Paulo Regis Menezes Sousa

paulo_regis@uvanet.br

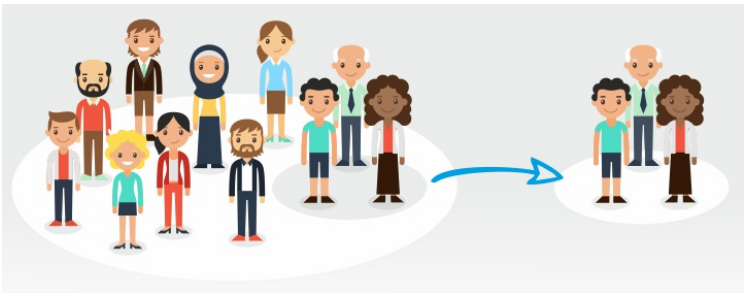
Inferência Estatística

- Amostragem
- Distribuições Amostrais
- Distribuição Amostral da Média
- Determinação do Tamanho de uma Amostra
- Distribuição Amostral de uma Proporção

- O uso de informações de uma amostra para concluir sobre o todo faz parte da atividade diária da maioria das pessoas.
  - *Uma cozinheira verifica se o prato que ela está preparando tem ou não a quantidade adequada de sal.*
  - *Um comprador, após experimentar um pedaço de laranja numa banca de feira, decide se vai comprar ou não as laranjas.*

- Modelos probabilísticos procuram representar a variabilidade de fenômenos casuais de acordo com suas ocorrências.
- Na prática, frequentemente o pesquisador tem alguma ideia sobre a forma da distribuição estudada, mas não dos valores exatos dos parâmetros que a especificam.
- Raramente se consegue obter a distribuição exata de alguma variável, ou porque isso é muito **dispendioso**, ou muito **demorado** ou às vezes porque consiste num **processo destrutivo**.

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



- Isso acontece porque na prática, ou não temos qualquer informação a respeito da variável, ou ela é apenas parcial.

- A maneira de se obter a amostra é tão importante, e existem tantos modos de fazê-lo, que esses procedimentos constituem especialidades dentro da Estatística.
- Poderíamos dividir os procedimentos científicos de obtenção de dados amostrais em três grandes grupos.

Levantamentos Amostrais a amostra é obtida de uma população bem definida, por meio de processos bem protocolados e controlados pelo pesquisador, estes levantamentos podem ser: **probabilísticos** e/ou **não-probabilísticos**.

Planejamento de Experimentos analisa-se o efeito de uma variável sobre outra, com interferência do pesquisador sobre o ambiente em estudo (população), bem como o controle de fatores externos.

Levantamentos Observacionais os dados são coletados sem que o pesquisador tenha controle sobre as informações obtida.

- Utilizando-se um procedimento aleatório, sorteia-se um elemento da população, sendo que todos os elementos têm a mesma probabilidade de ser selecionados. Repete-se o procedimento até que sejam sorteadas as n unidades da amostra.
- Podemos ter uma AAS COM, e SEM, reposição.
  - *Do ponto de vista da quantidade de informação contida na amostra, amostrar sem reposição é mais adequado. Contudo, a amostragem com reposição conduz a um tratamento teórico mais simples.*
- No R podemos utilizar a função `sample` para realizar uma amostragem aleatória simples.

```
> x <- c(1,2,3,4,5,6)
> sample(x, size = 2, replace = TRUE)
[1] 2 5
>
```

Uma **estatística** é uma característica da amostra, ou seja, uma estatística T é uma função de $X_1, X_2, \ldots, X_n$.

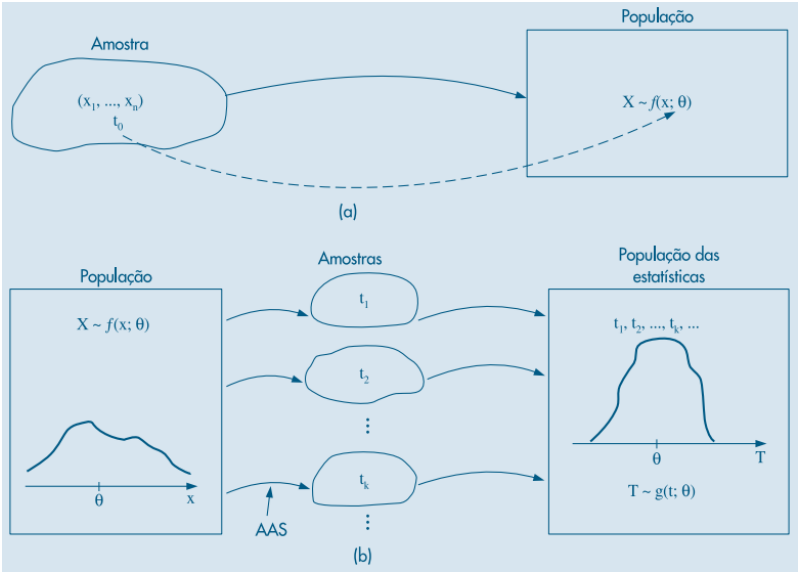
- As estatísticas mais comuns são:
  (a) $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ média da amostra,
  (b) $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ variância da amostra,
  (c) $X_{(1)} = \min(X_1, X_2, \ldots, X_n)$ o menor valor da amostra,
  (d) $X_{(n)} = \min(X_1, X_2, \ldots, X_n)$ o maior valor da amostra.

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

Denominação	População	Amostra
Média	$\mu = E(X)$	$\overline{X} = \sum X_i / n$
Mediana	$\mathrm{Md} = Q_2$	$\mathrm{md} = q_2$
Variância	$\sigma^2 = \mathrm{Var}(X)$	$S^2 = \sum (X_i - \overline{X})^2/(n-1)$
Nº de elementos	N	n
Proporção	p	$\hat{p}$
Quantil	$Q(p)$	$q(p)$
Quartis	$Q_1,\ Q_2,\ Q_3$	$q_1,\ q_2,\ q_3$
Intervalo inter-quartil	$d_Q = Q_3 - Q_1$	$d_q = q_3 - q_1$
Função densidade	$f(x)$	histograma
Função de distribuição	$F(x)$	$F_n(x)$

Figura: Símbolos mais comuns.

- O problema da *inferência estatística* é fazer uma afirmação sobre os parâmetros da população através da amostra.
- Por exemplo:
  1. Queremos fazer uma afirmação sobre um parâmetro θ.
  2. Realizamos uma amostragem (por exemplo uma AAS de n elementos).
  3. Calculamos a estatística T, uma função da amostra $(X_1, X_2, \ldots, X_n)$, ou seja, $T = f(X_1, X_2, \ldots, X_n)$.
  4. Teremos observado um valor particular de T, digamos t_0, e baseados nesse valor faremos uma afirmação sobre θ, o parâmetro populacional.


Amostra
$(x_1, ..., x_n)$
t_0
População
$X \sim f(x; \theta)$
(a)
População
$X \sim f(x; \theta)$
θ
x
Amostras
t_1
t_2
t_k
AAS
População das estatísticas
$t_1, t_2, ..., t_k, ...$
θ
T
$T \sim g(t; \theta)$
(b)

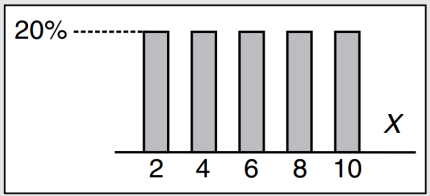
Exercício 1

Suponha que uma urna contém cinco bolas numeradas com os números 2, 4, 6, 8 e 10. Considerando que da urna serão retiradas amostras com reposição de tamanho $n = 2$, o objetivo é determinar o valor esperado da média de todas as combinações possíveis de serem formadas.

Solução

Analisemos a população formada pelas cinco bolas {2, 4, 6, 8, 10} dentro da urna.

1. Cada uma das cinco bolas tem a mesma probabilidade de ser escolhida, sendo 20% a probabilidade de uma bola ser escolhida.
2. A distribuição de frequências relativas da população {2, 4, 6, 8, 10} é uma distribuição discreta e uniforme com média igual a seis, $\mu = 6$, como mostra o histograma da figura seguinte.



$$E(\bar{X}) = 2(0,2) + 4(0,2) + 6(0,2) \ldots$$
$$\cdots + 8(0,2) + 10(0,2)$$
$$E(\bar{X}) = 6$$

Solução

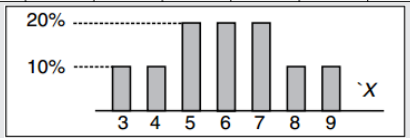
3. Analisemos o **experimento** de retirar amostras de tamanho $n = 2$ com reposição. O número total de amostras diferentes é dez, resultado obtido da contagem das combinações de cinco bolas tomadas duas a duas.

Amostras	2, 4	2, 6	2, 8	2, 10	4, 6	4, 8	4, 10	6, 8	6, 10	8, 10
Média*X*	3	4	5	6	5	6	7	7	8	9

Solução

4. A tabela seguinte registra a distribuição de frequências das médias das amostras X, tendo presente que todas as amostras são igualmente prováveis.

Média $\overline{X}$	3	4	5	6	7	8	9
$P(\overline{X})$	0,1	0,1	0,2	0,2	0,2	0,1	0,1



$$E(\bar{X}) = 3(0,1) + 4(0,1) + 5(0,2) + 6(0,2) + 7(0,2) + 8(0,1) + 9(0,1)$$
$$E(\bar{X}) = 6$$

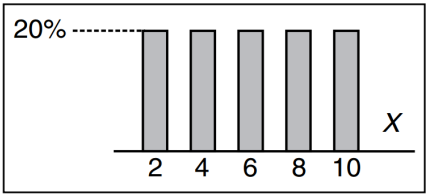


Figura: Dist. de freq. da população

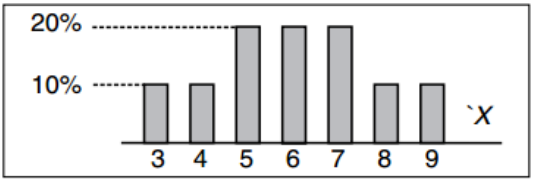


Figura: Dist. amostral

```r
# Exercício 1
urna = c(2,4,6,8,10)
tam  = 2                          # tamanho da amostra
num  = 1000                       # numero de amostras
amostras = matrix(0, num, tam) # cria matriz 1000x2

for (i in 1:num) {
   amostras[i,] <-sample(urna,           # amostra aleatória
                         size= tam,      # tamanho da amostra
                         replace= TRUE) # com, ou sem, reposição
}
medias <- apply(amostras,  # aplica função nas margens da matriz
                MARGIN= 1, # margem (linhas=1 ou colunas=2)
                FUN= mean) # função que deve ser aplicada

par(mfrow=c(1,2))     # parâmetro gráfico: 1 linha e 2 colunas
b <- c(1,3,5,7,9,11) # pontos de quebra das células do histograma
hist(amostras[,1], probability= TRUE, col= 3, breaks= b)
hist(medias,       probability= TRUE, col= 4, breaks= b)
```

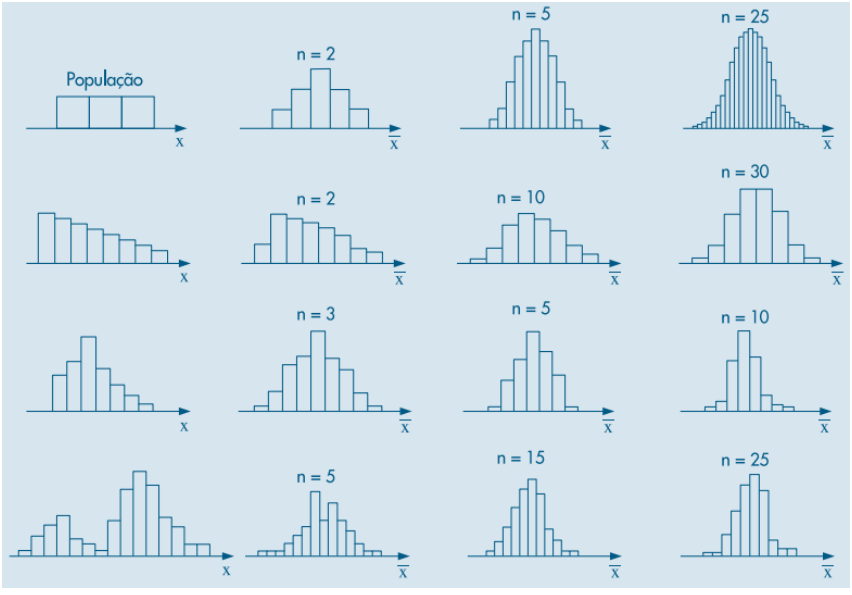
Teorema 1

Seja X uma variável aleatória com média μ e variância σ^2, e seja $(X_1, X_2, \ldots, X_n)$ uma AAS de X. Então,

$$E(\bar{X}) = \mu \quad \text{e} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

Teorema 2 (TLC)

Para amostras aleatórias simples $(X_1, X_2, \ldots, X_n)$, retiradas de uma população com média μ e variância σ^2 finita, a distribuição amostral da média $\bar{X}$ aproxima-se, para n grande, de uma distribuição normal, com média μ e variância $\frac{\sigma^2}{n}$.


População
n = 2
n = 5
n = 25
n = 2
n = 10
n = 30
n = 3
n = 5
n = 10
n = 5
n = 15
n = 25
x
$\bar{x}$

Corolário

Se $(X_1, X_2, \ldots, X_n)$ for uma amostra aleatória simples da população X, com média μ e variância σ^2 finita, e $\bar{X} = (X_1 + \cdots + X_n)/n$, então

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

Exercício 2

Uma v.a. X tem distribuição normal, com média 100 e desvio padrão 10.

a) Qual a $P(90 < X < 110)$?

b) Se $\bar{X}$ for a média de uma amostra de 16 elementos retirados dessa população, calcule $P(90 < \bar{X} < 110)$.

c) Represente, num único gráfico, as distribuições de X e $\bar{X}$.

d) Que tamanho deveria ter a amostra para que $P(90 < \bar{X} < 110) = 0,95$?

```r
# a) p(90<X<110)
p90  <- pnorm(90, 100, 10)
p110 <- pnorm(110, 100, 10)
p    <- p110 - p90
p

# b) p(90<Xamostra<110), n = 16
p90  <- pnorm(90, 100, 10/sqrt(16))
p110 <- pnorm(110, 100, 10/sqrt(16))
p    <- p110 - p90
p
```

```r
# c) Graficos de X e Xamostra
curve(dnorm(x,100,10/sqrt(16)),
      from = 50,
      to   = 150,
      xlab = "X", ylab = "Probabilidade",
      main = "Distribuições",
      col  = "blue")
curve(dnorm(x,100,10),
      from = 50,
      to   = 150,
      add  = TRUE,
      col="orange")
legend("topleft",
       legend = c("Distribuição original", "Distribuição amostral"),
       col = c("orange","blue"),
       lty = 1)
```

$$P\left(\frac{\sqrt{n}(90-100)}{10} < Z < \frac{\sqrt{n}(110-100)}{10}\right) = 0,95$$

$$P\left(-\sqrt{n} < Z < \sqrt{n}\right) = 0,95$$

$$P(Z < -\sqrt{n}) = 0,025$$

```r
# d)
p <- qnorm(0.025)
p
```

$$-\sqrt{n} = -1.95996$$

$$n = (1.95996)^2 \cong 4$$

```r
n <- ceiling(p^2)
n
```

Exercício 3

A máquina de empacotar lembas o faz segundo uma distribuição normal, com média μ e desvio padrão 10 g.

a) Em quanto deve ser regulado o peso médio μ para que apenas 10% dos pacotes tenham menos do que 500 g?

b) Com a máquina assim regulada, qual a probabilidade de que o peso total de 4 pacotes escolhidos ao acaso seja inferior a 2 kg?

$$X \sim N(\mu, 10)$$

$$P(X < 500) = 0,1$$

$$P\left(Z < \frac{500 - \mu}{10}\right) = 0,1$$

$$\frac{500 - \mu}{10} = -1,28$$

$$-\mu = -1,28 \times 10 - 500$$

$$-\mu = -512.81$$

$$\mu = 512.81$$

```r
# a)
q = qnorm(0.1)
q
m = abs(q * 10 - 500)
m
```

$$P\left(\sum_{i=1}^{4} X_i < 2000\right) = P\left(\frac{1}{4}\sum_{i=1}^{4} X_i < \frac{2000}{4}\right) = P\left(\bar{X} < 500\right) = 0,005203566$$

```
# b)
# prob. de 1 pacote ter menos de 500g
p1pac = pnorm(500, m, 10)
p1pac

# prob. de uma amostra de 4 pacotes ter média menor que 500g
p4pac = pnorm(500, m, 10/sqrt(4))
p4pac
```

Exercício 4

A capacidade máxima de um elevador é de 500 kg. Se a distribuição X dos pesos dos usuários for suposta $N(70, 100)$:

a) Qual é a probabilidade de sete passageiros ultrapassarem esse limite?

b) E seis passageiros?

$$\bar{X} \sim N\left(70, \frac{100}{7}\right)$$

$$P\left(\sum_{i=1}^{7} X_i > 500\right) = P\left(\frac{1}{7}\sum_{i=1}^{7} X_i > \frac{500}{7}\right) = P\left(\bar{X} > 71,42857\right) = 35,27\%$$

```r
# a)
p = pnorm(500/7, 70, 10/sqrt(7), lower.tail = F)
p

# b)
p = pnorm(500/6, 70, 10/sqrt(6), lower.tail = F)
p
```

- Há uma observação importante a ser feita: se a população for finita e de tamanho N conhecido, e se a amostra de tamanho n dela retirada for sem reposição, então:

$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n} \underbrace{\left(\frac{N-n}{N-1}\right)}_{\text{fator de correção}^*}. \tag{2}$$

- Quando tiramos uma amostra em que $\frac{n}{N} \leq 0,05$, é indiferente usar o fator de correção para populações finitas, porque o erro é muito pequeno.

Exercício 5

Em uma turma de 20 alunos de uma faculdade observou-se que a altura média dos alunos é de 175 cm e o desvio padrão, 5 cm. Retiramos uma amostra sem reposição, de tamanho $n = 9$.

$$X \sim N(175 \text{ cm}, 5 \text{ cm}).$$

Dessa forma temos:

$$\frac{n}{N} = \frac{9}{20} = 0,45 > 0,05$$

com correção:

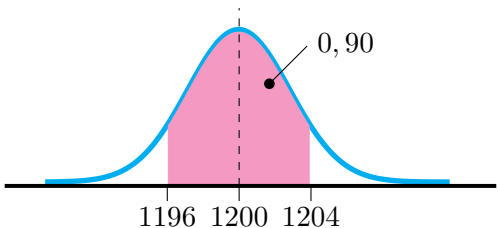
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} = \frac{5}{\sqrt{9}}\sqrt{\frac{20-9}{20-1}} = 1,268143$$

Logo, a média das médias amostrais é 175 cm e o desvio padrão da média amostral é aproximadamente 1,27 cm.

- Não há dúvida de que uma amostra não representa perfeitamente uma população. Ou seja, a utilização de uma amostra implica na aceitação de uma margem de erro que denominaremos erro amostral.
- Não podemos evitar a ocorrência do erro amostral, porém podemos limitar seu valor através da escolha de uma amostra de tamanho adequado.
- Quanto maior o tamanho da amostra, menor o erro cometido e vice-versa.

Exercício 6

Seja $X \sim N(1200, 28.982)$. Qual deverá ser o tamanho de uma amostra, de tal forna que $P(1196 < \bar{X} < 1204) = 0.90$?



$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - 1200}{\frac{28,982}{n}}$$

$$Z_{0.95} = 1,64$$

$$1,64 = \frac{1204 - 1200}{\frac{28,982}{\sqrt{n}}}$$

$$\therefore n = \left(\frac{1,64 \times 28,982}{4}\right)^2 \cong 141$$

- Considerando uma população em que a proporção de elementos portadores de certa característica é p. Logo, podemos definir uma v.a. X, da seguinte maneira:

$$X = \begin{cases} 0, & \text{se o indivíduo for portador da característica} \\ 1, & \text{se o indivíduo não for portador da característica,} \end{cases}$$

logo,

$$\mu = E(X) = p, \quad \sigma^2 = \mathrm{Var}(X) = \frac{p(1-p)}{n}.$$

- Para n grande podemos considerar a distribuição amostral de p como aproximadamente normal:

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

onde, $\hat{p}$ é a distribuição amostral e n o número de amostras.

Exemplo 5

Suponha que $p = 30\%$ dos estudantes de uma escola sejam mulheres. Colhemos uma AAS de $n = 10$ estudantes e calculamos $\hat{p} =$ proporção de mulheres na amostra. Qual a probabilidade de que $\hat{p}$ difira de p em menos de 0,01?

Solução

Temos que essa probabilidade é dada por

$$P(|\hat{p} - p| < 0,01) = P(-0,01 < \hat{p} - p < 0,01).$$

Para a distribuição de $\hat{p} - p$ temos que, $\hat{p} - p \sim N\left(0, \frac{p(1-p)}{n}\right)$, e como $p = 0,3$, podemos calcular a variância como

$$\text{Var}(\hat{p}) = \frac{0,3(1-0,3)}{10} = \frac{0,3 \times 0,7}{10} = 0,021,$$

e, portanto, a probabilidade pedida é igual a

$$P\left(\frac{-0,01}{\sqrt{0,021}} < Z < \frac{0,01}{\sqrt{0,021}}\right) = P(-0,07 < Z < 0,07) = 0,05580634.$$