

AULA 4: DISTRIBUIÇÕES DE PROBABILIDADES AMOSTRAIS

Gleici Castro Perdoná

pgleici@fmrp.usp.br

Exemplo

2. Sabe-se que o tempo gasto no exame de um paciente tem distribuição aproximadamente Normal, com média 30 min e desvio padrão de 5 min.
- Sorteando-se um médico residente ao acaso, qual é a probabilidade dele terminar o exame antes de 24 minutos?
 - Qual deve ser o tempo de exame, de modo a permitir que 95% dos residentes terminem no prazo estipulado?
 - Qual é o intervalo de tempo, simétrico em torno da média tal que 80% dos residentes gastam para completar o exame?

Exercício (10 min)

- Uma população X tem uma distribuição normal de média 100 e desvio padrão 10. Qual $P(95 < X < 105)$?
- Uma população X tem uma distribuição normal de média 10 e desvio padrão 1. Qual $P(X > 7)$?

TÉCNICAS DE AMOSTRAGEM E SEUS USOS

I) Aleatória simples

Populações ricamente homogêneas.

II) Sistemática

Populações ordenadas.

III) Estratificada

Populações heterogêneas.

IV) Conglomerado

Subgrupos de populações.

V) Não-probabilística

Amostragem acidental, intencional ou por quotas.

AMOSTRA ALEATÓRIA SIMPLES

População finita ($n:N$) ou população infinita ($n:\infty$).

COMPUTADOR

Dados gerados num sistema de referência.

SORTEIO

Dados coletados numa tabela de números aleatórios.

COM REPOSIÇÃO

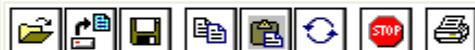
Um elemento pode ser retirado mais de uma vez.

N^n

SEM REPOSIÇÃO

Cada elemento só pode ser retirado uma vez.

C_n^N



```
R version 2.13.0 (2011-04-13)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: i386-pc-mingw32/i386 (32-bit)
```

R é um software livre e vem sem GARANTIA ALGUMA.
Você pode redistribuí-lo sob certas circunstâncias.
Digite 'license()' ou 'licence()' para detalhes de distr:

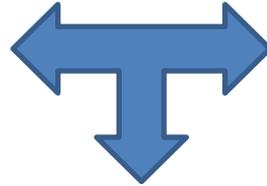
R é um projeto colaborativo com muitos contribuidores.
Digite 'contributors()' para obter mais informações e
'citation()' para saber como citar o R ou pacotes do R e

Digite 'demo()' para demonstrações, 'help()' para o sistema
ou 'help.start()' para abrir o sistema de ajuda em HTML.
Digite 'q()' para sair do R.

```
> x=seq(1:50)
> sample(x,10,replace=T)
 [1] 27 16 38 26 34 33 46 50 25 12
> sample(x,10,replace=F)
 [1]  4 46  7  9 25 18 42 44 23  8
> |
```


AMOSTRA SISTEMÁTICA

Lista de N elementos da população (p.ex., lista telefônica).



Amostra de tamanho n.

Unidade amostral inicial selecionada das primeiras k unidades da lista ($k = N/n$).



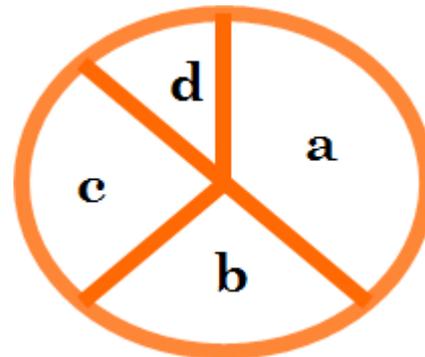
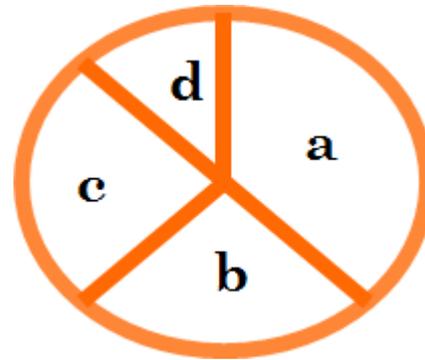
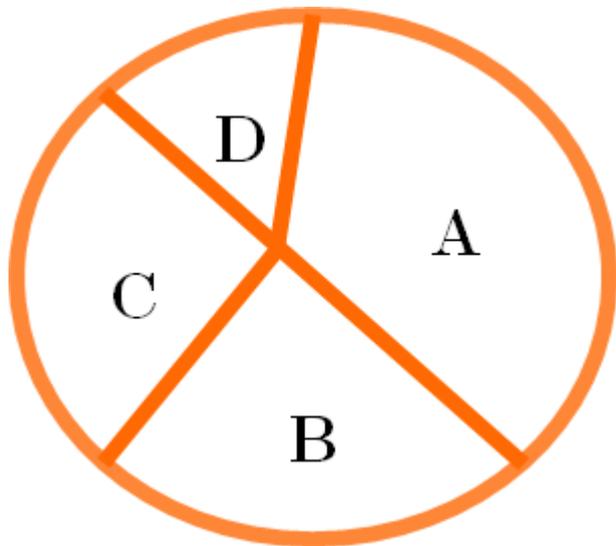
Ex.: seleciona-se, aleatoriamente, a 4ª pessoa da lista; a amostra segue, então, com os elementos $4+k, 4+2k\dots$

AMOSTRA ESTRATIFICADA

População heterogênea

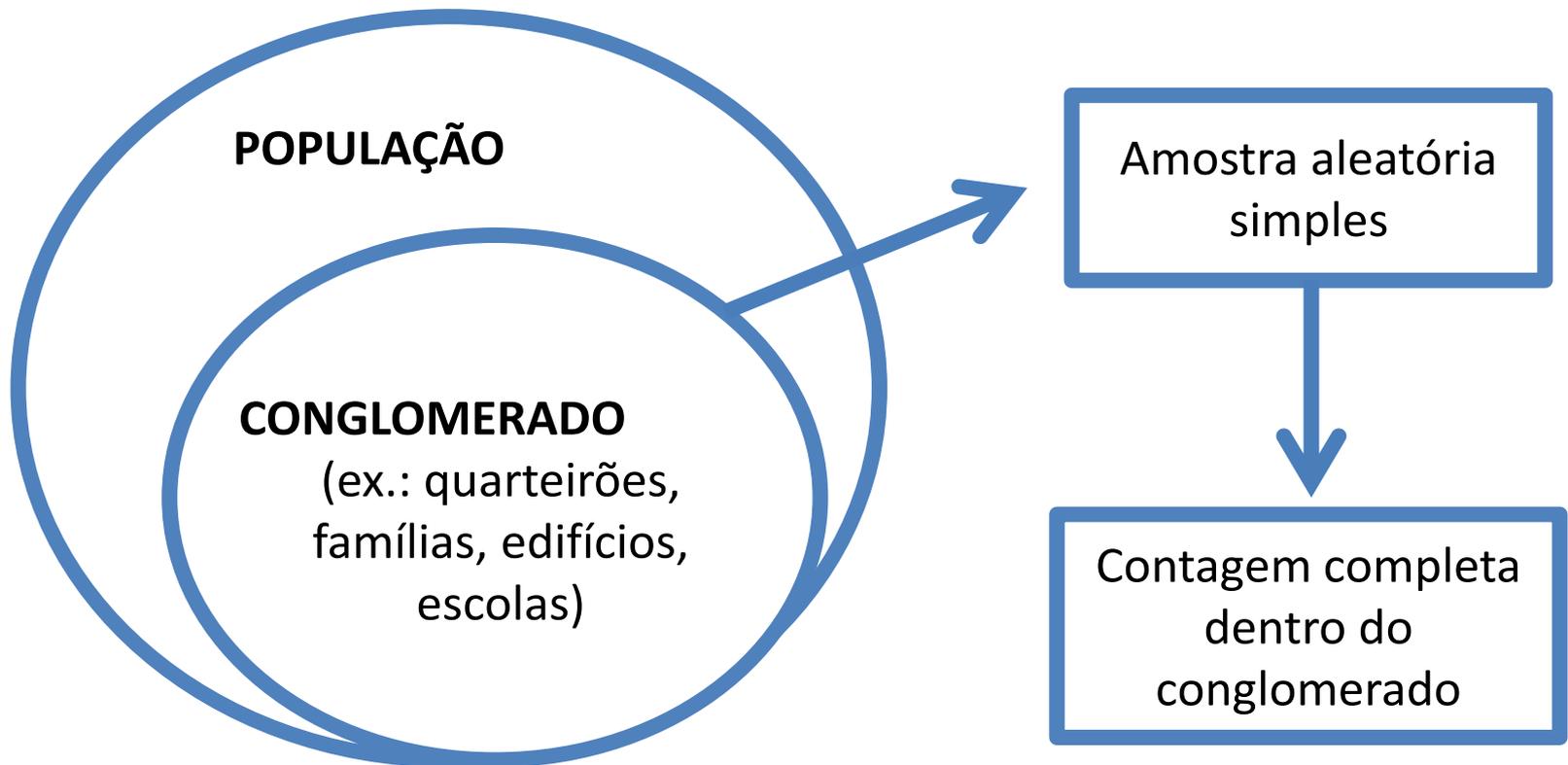


Estratos mais ou menos homogêneos



Amostragem simples ao acaso

AMOSTRA POR CONGLOMERADOS



AMOSTRAGEM NÃO-PROBABILÍSTICA

(Não-inferência – não se conhece a probabilidade de um indivíduo ser incluído na amostra)

AMOSTRAGEM
ACIDENTAL

AMOSTRAGEM
INTENCIONAL

AMOSTRAGEM POR
QUOTAS

Coletam-se elementos até se atingir o número desejado (ex.: pesquisas de opinião).

Os elementos são coletados dentro do grupo de interesse (ex.: sala de espera da clínica X).

A amostra recebe quotas proporcionais ao total da população (ex.: pesquisas de mercado).

EXERCÍCIO Exercício (10min)

Use o rol de dados abaixo (pesos, em kg, de 50 frequentadores do NSF X de Ribeirão Preto) para responder às questões.

45,2	53,1	57,7	58,4	61,4	61,5	62,3	63,5	63,6	64,3
64,8	65,7	66,7	66,7	67,5	67,8	67,8	68,0	68,0	68,9
68,9	71,3	71,3	71,5	71,6	72,5	73,1	74,0	74,1	74,1
74,2	76,1	76,1	76,5	76,7	77,5	77,7	77,7	79,1	79,4
79,5	79,9	81,9	82,2	82,3	84,9	85,0	87,7	89,8	94,1

- 1) Obtenha uma amostra aleatória simples de 10 elementos e calcule sua média.
- 2) Compare as duas estimativas encontradas com a média populacional.

Exemplo

Intervalo (10 minutos)



Estimador, estimativas e parâmetros

Uma característica da população é denominada parâmetro.

Um parâmetro é um valor, um número que representa uma característica única da população.

Se X uma variável de uma população, os principais parâmetros seriam:

- A média de X , anotada por μ
- A variância de X , anotada por σ^2
- O desvio padrão de X , anotado por σ
- A proporção de elementos de P que apresentam determinada característica, anotada por: p , entre outros.

Exemplo

- $X = \{ 1, 3, 5, 6 \}$ é amigos na república
- $\mu = (1 + 3 + 5 + 6) / 4 = 15 / 4 = 3,75$
- $\sigma^2 = (1 + 9 + 25 + 36) / 4 - 3,75^2 = 71/4 - 3,75^2$
 $= 17,75 - 14,0625 = 3,6875 = 3,69.$
- $\sigma = 1,9203 = 1,92$
- $p = 3 / 4 = 75\%$, exemplo para a proporção de numero ímpar.

Um estimador é uma característica da amostra.

Veja que se a amostra é aleatória o estimador é uma variável aleatória.

Então tudo de distribuição de probabilidade para variáveis aleatórias, aplica-se aos estimadores. A distribuição de probabilidade de um estimador é denominada de distribuição amostral.

Entendendo....

A média da amostra, \bar{X} que é um estimador da média da população: μ

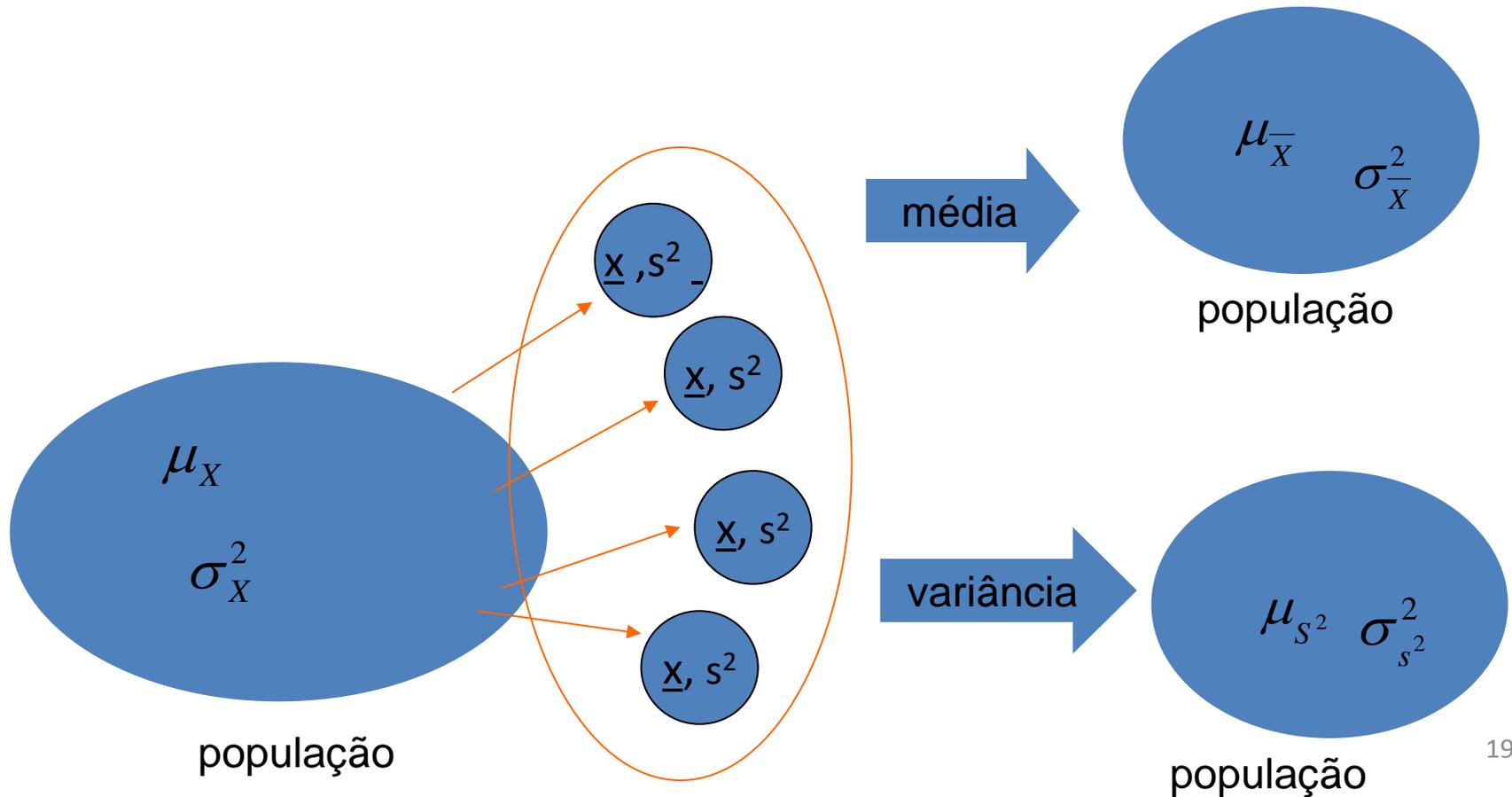
A variância amostral, S^2 que é um estimador da variância populacional: σ^2

A proporção amostral, P , que é um estimador amostral da proporção populacional p .

Estimativa é um valor particular de um estimador

- O estimador é a expressão (fórmula) enquanto que a estimativa é o valor particular que ele assume (número).

MÉDIA E VARIÂNCIA: POPULAÇÃO X AMOSTRA



Teorema do Limite Central

Seja X uma v. a. que tem média μ e variância σ^2 . Para uma amostra X_1, X_2, \dots, X_n , retirada ao acaso e com reposição de X , a distribuição de probabilidade da média amostral \bar{X} aproxima-se, para n grande, de uma distribuição normal, com média μ e variância σ^2/n , ou seja,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \text{ para } n \text{ grande, aproximadamente.}$$

• Se a distribuição de X é normal, então \bar{X} tem distribuição normal exata, para todo n .

• O desvio padrão $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$, que é

o desvio padrão da média amostral, também é denominado erro padrão.

DISTRIBUIÇÃO AMOSTRAL DAS MÉDIAS

Se n suficientemente grande, distribuição amostral \sim normal.

Média da distribuição amostral das médias = média da população

$$E(\bar{X}) = \mu_{\bar{X}} = \mu_X$$

População infinita, amostra com reposição, variância da distribuição amostral das médias:

$$E((\bar{X} - \mu)^2) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

População finita, amostra sem reposição, variância da distribuição amostral das médias:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

DISTRIBUIÇÃO AMOSTRAL DAS f_r fr

$$\text{Média: } E(\hat{fr}) = p$$

$$\text{Variância: } \downarrow \text{Var}(fr) = \frac{pq}{n}$$

DISTRIBUIÇÃO AMOSTRAL DE s^2

$$\text{Média: } E(S^2) = \mu_{s^2} = \sigma^2$$

$$\text{Variância: } \downarrow \text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

Voltando ao caso dos amigos da república

- $X:\{ 1, 3, 5, 6 \}$, todas as amostras possíveis de tamanho $n = 2$ extraídas com reposição. Para cada amostra vai-se calcular a média. Ter-se-á assim um conjunto de 16 valores que serão dispostos em uma tabela, com as respectivas probabilidades, e que constituirá então a distribuição amostral da média da amostra.
- As possíveis amostras com as respectivas médias são:
- $(1,1) = 1$
- $(1,3)=2$
- $(1,5)=3$, etc então temos:

Distribuição amostral da média é

\bar{x}	$f(\bar{x}) = P(\bar{X} = \bar{x})$	$\bar{x} f(\bar{x})$	$\bar{x}^2 f(\bar{x})$
1,0	1/16	1/16	1,0/16
2,0	2/16	4/16	8,0/16
3,0	3/16	9/16	27,0/16
3,5	2/16	7/16	24,5/16
4,0	2/16	8/16	32,0/16
4,5	2/16	9/16	40,5/16
5,0	1/16	5/16	25,0/16
5,5	2/16	11/16	60,5/16
6,0	1/16	6/16	36,0/16
Σ	1	60/16	254,5/16

$$\mu = 3,75 \quad \text{e} \quad \sigma^2 = 254,5/16 - 3,75^2 = 1,84 = 3,69/2$$

E sem reposição?

Comparando com o primeiro exercício

- Uma população X tem uma distribuição normal de média 100 e desvio padrão 10. Se \bar{X} é a média de 16 elementos extraída desta população, qual a $P(95 < \bar{X} < 105)$?
- A renda de um conjunto de pessoas de uma certa região tem média 6 salários mínimos e desvio padrão de 2. Extraída uma amostra de $n = 100$ pessoas, qual a probabilidade de a média desta amostra ter um valor superior a 6,3 sal.min?

DISTRIBUIÇÃO AMOSTRAL DAS MÉDIAS

Se n suficientemente grande, distribuição amostral \sim normal.

Média da distribuição amostral das médias = média da população

$$E(\bar{X}) = \mu_{\bar{X}} = \mu_X$$

População infinita, amostra com reposição, variância da distribuição amostral das médias:

$$E((\bar{X} - \mu)^2) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$$

População finita, amostra sem reposição, variância da distribuição amostral das médias:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \left(\frac{N-n}{N-1} \right)$$

DISTRIBUIÇÃO AMOSTRAL DAS f_r fr

$$\text{Média: } E(\hat{f}_r) = p$$

$$\text{Variância: } \downarrow \text{Var}(f_r) = \frac{pq}{n}$$

DISTRIBUIÇÃO AMOSTRAL DE s^2

$$\text{Média: } E(S^2) = \mu_{s^2} = \sigma^2$$

$$\text{Variância: } \downarrow \text{Var}(S^2) = \frac{2\sigma^4}{n-1}$$

APLICAÇÃO DO TCL

Tem-se a distribuição dos níveis séricos de colesterol de todos os homens de 20 a 74 anos (EUA), $\mu = 211$ mg/100 mL e $\sigma = 46$ mg/100 mL. Seleccionando amostras repetidas de tamanho 25 da população, que proporção de amostras terá um valor médio de 230 mg/100 mL ou acima?

- Pelo TCL, a distribuição de médias de amostras de tamanho 25 é aproximadamente normal com média $\mu = 211$ e desvio-padrão $\sigma/\sqrt{n} = 9,2$ mg/100 mL. Como $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ é uma variável aleatória normal padrão, $z = \frac{230 - 211}{9,2} = 2,07$. Acima desse valor, encontra-se 0,019 da área sob a curva normal padrão, logo 1,9% das amostras terá um valor médio acima de ou igual a 230 mg/100 mL.
- É possível obter, por exemplo, os limites superior e inferior que incluem 95% das médias das amostras de tamanho 25 extraídas da população, e conforme o tamanho das amostras aumenta, a quantidade de variabilidade entre as médias diminui; consequentemente, os limites que englobam 95% dessas médias se aproximam.

Podemos pensar em um intervalo para a média μ

Vimos até o momento estimador e estimativas pontuais. O estimador por intervalo para a média μ tem a forma

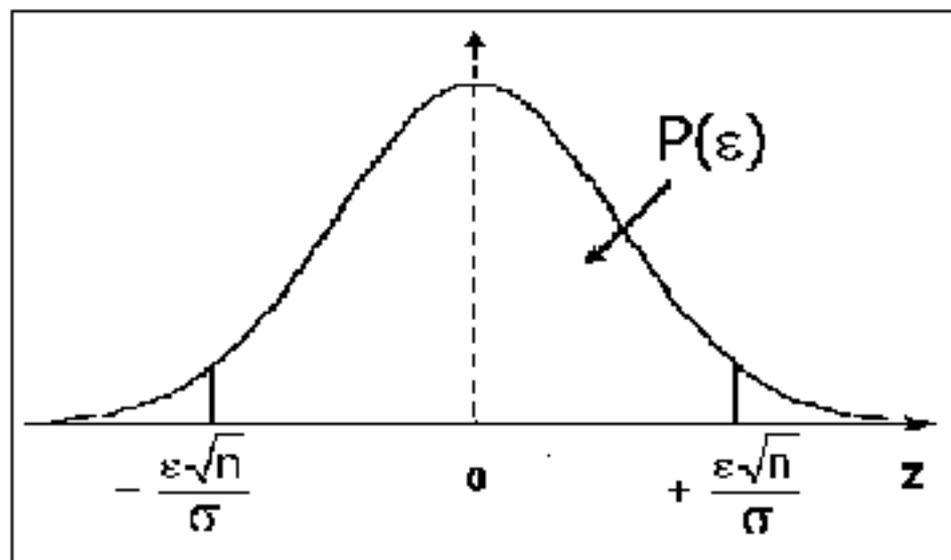
$$[\bar{X} - \varepsilon ; \bar{X} + \varepsilon]$$

Nossa questão é determinar ε

Seja $P(\varepsilon) = \gamma$, a probabilidade da média amostral \bar{X} estar a uma distância de, no máximo ε , da média populacional μ (desconhecida), ou seja,

$$\begin{aligned}\gamma &= \mathbf{P}\left(\left|\bar{X}-\mu\right| \leq \varepsilon\right) = \mathbf{P}\left(\mu-\varepsilon \leq \bar{X} \leq \mu+\varepsilon\right) \\ &= \mathbf{P}\left(\frac{-\varepsilon}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq \frac{\varepsilon}{\frac{\sigma}{\sqrt{n}}}\right) \equiv \mathbf{P}\left(\frac{-\varepsilon\sqrt{n}}{\sigma} \leq \mathbf{Z} \leq \frac{\varepsilon\sqrt{n}}{\sigma}\right),\end{aligned}$$

sendo $\mathbf{Z} \sim \mathbf{N}(0,1)$.



Denotando $\frac{\varepsilon\sqrt{n}}{\sigma} = z$, temos que

$$\gamma = P(-z \leq Z \leq z).$$

Assim, conhecendo-se o coeficiente de confiança γ obtemos z .

Erro na estimativa intervalar

Da igualdade $z = \frac{\varepsilon \sqrt{n}}{\sigma}$, segue que

o erro amostral ε é dado por

$$\varepsilon = z \frac{\sigma}{\sqrt{n}},$$

sendo z tal que $\gamma = P(-z \leq Z \leq z)$, com $Z \sim N(0,1)$.

O intervalo de confiança para a média μ , com coeficiente de confiança γ fica, então, dado por

$$\left[\bar{X} - z \frac{\sigma}{\sqrt{n}} ; \bar{X} + z \frac{\sigma}{\sqrt{n}} \right],$$

sendo σ o desvio padrão de X .

A partir da relação $\varepsilon = z \frac{\sigma}{\sqrt{n}}$,

o tamanho da amostra n é determinado por

$$n = \left(\frac{z}{\varepsilon} \right)^2 \sigma^2,$$

conhecendo-se o desvio padrão σ de X , o erro ε da estimativa e o coeficiente de confiança γ do intervalo, sendo z tal que

$$\gamma = P(-z \leq Z \leq z) \text{ e } Z \sim N(0,1).$$

Exemplo

- O colesterol das mulheres universitárias tem uma distribuição de probabilidades com $\sigma=50\text{mg/dl}$ e média desconhecida. Desejamos estimar a média μ com erro de 20 mg/dl e confiança de 90% , quantas alunas precisamos na amostra?

BIBLIOGRAFIA RECOMENDADA

BUSSAB, W.O.; MORETTIN, P. **Estatística básica**. 4 ed. São Paulo, Atual, 1987.

PAGANO, M. e GAUVREAU, K. Princípios de Bioestatística - Tradução da 2ª Edição Norte Americana, Pioneira Thonpson Learning, São Paulo, SP,2004.

MEDRONHO R; CARVALHO DM; BLOCH KV; LUIZ RR; WERNECK GL. Epidemiologia. Atheneu, 2 ed. São Paulo, 2008

ROSNER, B. Fundamentos de bioestatística. 8ª Edição Norte Americana, Cengage Learning, 2016.

OBRIGADA

- EXERCÍCIOS estarão no stoa para entregar como tarefa para próximas aulas.