Matheus Muniz
Id: 1104005
MM

mostragem, Estimação e Inferência

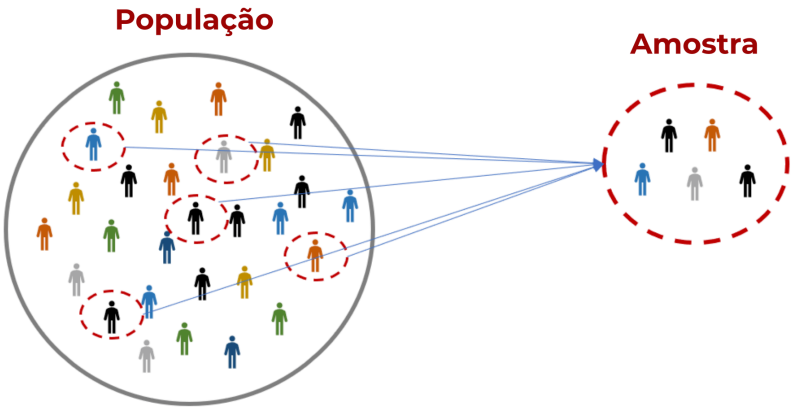
1. Introdução

Neste tópico serão discutidos alguns pontos a respeito de amostragem, sua importância e os tipos de métodos de amostragem que podem ser realizados para experimentações a respeito de uma população. Também serão discutidos pontos sobre a estimação e inferência, como podem ser utilizados nível e intervalo de confiança, além de cálculos para o tamanho amostral.

2. Noções sobre Amostragem

Aplicar **amostragem** faz parte do cotidiano das pessoas sem mesmo elas saberem, como por exemplo na culinária. Ao cozinhar um feijão para saber se está com suficiente sal e temperos, é retirado uma pequena porção para experimentar e, a partir desta porção, concluir algo a respeito do feijão como um todo. Tem exemplos também como em análises clínicas, ao retirar amostras de sangue e outros fluidos para a análise e identificação de eventuais alterações. Vale ressaltar que ambos os exemplos, não seria viável analisar o todo para ter conclusões satisfatórias.

Em estátistica, **amostragem** representa este procedimento que visa obter informações sobre o todo baseando-se no resultado de uma amostra.



Fonte Solvis

2.1 Terminologia sobre Amostragem

Será necessário definir algumas terminologias normalmente utilizadas quando se fala a respeito de amostragem. Estes parâmetros estão descritos a seguir:

- **População**: ou Universo é o conjunto de todas as unidades elementares de interesse. A população deve ser definida claramente e em termos da informação que se pretende conhecer;
- **Unidade**: trata-se de qualquer elemento da população;
- **Amostra**: uma parte ou subconjunto da população;
- **Censo**: observação de todos os elementos da população;
- **Parâmetro Populacional**: é o vetor correspondente a todos os valores de uma variável de interesse. Pode ser qualitativa (gosto musical, opinião sobre o governo, etc) ou quantitativa (média, proporção, quantidade, etc).
- **Função Paramétrica Populacional**: é uma característica numérica da população, ou seja, uma expressão numérica que condensa os valores do vetor de parâmetro populacional. Por exemplo, média, total, proporção, dentre outros.

Utilizando como exemplo, considere uma população formada por 4 alunos de uma escola. Com as seguintes características:

Variável	Valores			
Aluno	1	2	3	4
Nome	Luiz	Marcela	Pedro	Julia
Idade	15	14	13	16
Sexo	M	F	M	F

Neste exemplo, cada aluno é um elemento da população. Com relação à amostragem os subconjuntos (Luiz, Marcela), (Pedro, Julia), (Marcela) são **exemplos de amostra**. **Parâmetros populacionais**: $idade = (15, 14, 13, 16)$ e $sexo = (M, F, M, F)$. Com relação às **funções paramétricas**, poderíamos definir:

- Idade média: fazendo idade = I: $\mu = \overline{I} = \frac{\sum_{i=1}^{4} I_i}{4} = \frac{15 + 14 + 13 + 16}{4} = 14,5$
- Idade máxima: max(Y)= max(15, 14, 13, 16) = 16
- Porporção de meninas: sexo = Y = (M, F, M, F)

$$p(F) = \frac{1}{2} = 0,5$$

2.2 Tipos de Amostragem

Sobre a amostragem, pode-se classificá-la em dois tipos a respeito sobre as probabilidade destas amostra:

- **Amostra Probabilística**: todos os elementos da população apresentam probabilidade maior que zero de serem selecionados;
- **Amostra Não-Probabilística**: quando não há probabilidade clara/conhecida de seleção dos elementos. Os elementos são escolhidos de forma julgamental.

2.3 Quando utilizar amostras?

Como no exemplo do feijão e do paciente, existem alguns casos que são ideais para a utilização de amostras em análises. A seguir, serão listados alguns casos de aplicação de amostragem:

- **Populações infinitas**: Quando é impossível investigar todos os elementos de uma população;
- **Teste Destrutivos**: Estudos onde os elementos avaliados passaram por algum processo de transformação, sendo este processo destrutivo para a amostra. Exemplo: Ensaio sobre fadiga em asas de avião;
- **Resultados Rápidos**: Pesquisas que precisam de mais agilidade na divulgação. Exemplo: pesquisas de opinião, pesquisas que envolvam problemas de saúde pública;
- **Custos Elevados**: Quando a população é finita mas muito numerosa, o custo de um censo pode tornar o processo inviável. Exemplo: modelar uma base de 10 milhões de clientes em uma máquina local não adequada.

3. Métodos de Amostragem

Neste material, será abordado apenas os métodos relacionados à amostragem probabilística, que tem como objetivo de obter uma **amostra representativa**. Uma amostra é considerada representativa quando consegue **refletir as características da população**.

3.1 Amostra Aleatória Simples

Este é o método mais simples e mais importante de seleção de uma amostra, pois pode ser usada em combinação com outros métodos. A premissa assumida é que a população é homogênea com relação à característica de interesse.

A amostra aleatória simples (AAS) pode ser realizada com ou sem reposição. No caso em que há reposição, cada elemento pode ser sorteado mais de uma vez. Para exemplificar, suponha que se queira sortear um número aleatório de uma urna, se for uma AAS com reposição, este número voltará para urna para participar do próximo sorteio. Se não houver reposição, cada elemento só poderá ser selecionado uma vez para compor a amostra.

Considere uma população formada por N elementos (conhecido e finito). Este método consiste em selecionar n elementos, sendo que cada elemento tem a mesma probabilidade de ser selecionado.

Exemplo de Aplicação: Considere uma população formada por 50 alunos. Selecionar de forma aleatória 10 alunos, sem reposição.

```
# Carrega a função sample do random
from random import sample

# Gera uma lista de número identificando os alunos
alunos = list(range(1, 51))

# Gera uma amostra com tamanho 10
sample(alunos, k = 10)
```

3.2 Amostra Sistemática

Usada quando os elementos população estão ordenados (população de lista telefônica, casas em uma rua). Considere uma população de tamanho N e que se queira uma amostra de tamanho n. O processo de amostragem deste método consiste em:

- Dividir o tamanho populacional em k partes: $k = \frac{N}{n}$
- Definir a posição de início da amostragem (que também será o primeiro elemento da amostra). Para tal fim, é sorteado i com o uso da amostra aleatória simples no intervalo, em que $i \in [1, k]$;

- A partir do elemento selecionado aleatoriamente, é realizada sucessão aritmética para selecionar os $n-1$ indivíduos restantes: $i, i+k, i+2k, i+3k, ..., i+(n-1)k$

Segue um exemplo de como implementar uma amostragem sistemática utilizando o *Python*:

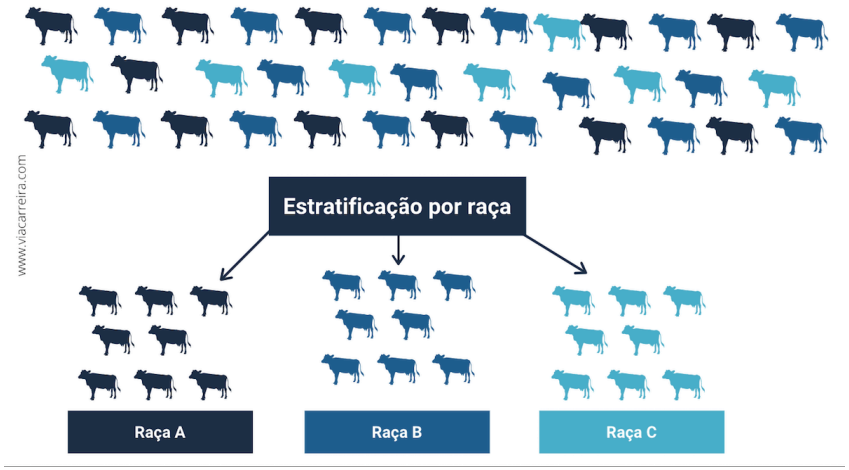
```
# Carrega a biblioteca random
import random

# Define a função para amostragem sistemática
def amostra_sistematica(populacao, n):
    N = len(populacao) # define o tamanho da população
    k = N // n         # define o tamanho das partes

    # Escolhe a posição inicial de amostragem
    first_elem = np.random.randint(0, k)

    # Cria uma lista vazia para armazenar os elementos
    amostras = []

    # Escolhe os elementos seguinte a frequência dada pela amostra sistemática
    for i in range(first_elem, N, k):
        amostras.append(populacao[i])

    # Retorna a amostra gerada
    return amostras

# Cria uma amostra sistemática com 10 alunos
```

```
print(amostra_sistematica(alunos, 10))
```

3.3 Amostra Estratificada

Trata-se do método em que a população é dividida em grupos (estratos) segundo alguma(s) característica(s) conhecida(s) na população sob estudo. São exemplos de estrato o gênero, faixa etária, região geográfica, profissão. No geral, é usada quando a população é heterogênea sob a ótica das características analisadas. Procedimento de amostragem:

- Dividir as N unidades da população em $N_1, N_2, \cdots, N_j$ estratos distintos e homogêneos;
- Selecionar, ao acaso, uma amostra de tamanhos $n_1, n_2, \cdots, n_j$, de modo que o tamanho da amostra seja $n = n_1 + n_2 + \cdots + n_j$. O tamanho amostral pode ser proporcional à representatividade do estrato.



Fonte: Via Carreira

Exemplo de Aplicação: Considere a população formada pelos integrantes de uma escola. Dependendo do objetivo do estudo, esta população poderia ser dividida em alunos, professores, e demais funcionários (grupos mais homogêneos com relação à função na escola). Agora considere que a proporção de cada estrato seja: 60% alunos, 30% professores e 10% servidores. A amostragem

poderia ser realizada dentro de cada estrato de forma que o tamanho amostral preserve esta característica. Sendo assim, se a amostra total é n, a composição será $0,6 \times n$ de alunos, $0,3 \times n$ de professores e $0,10 \times n$ de servidores.

3.4 Amostra por conglomerados

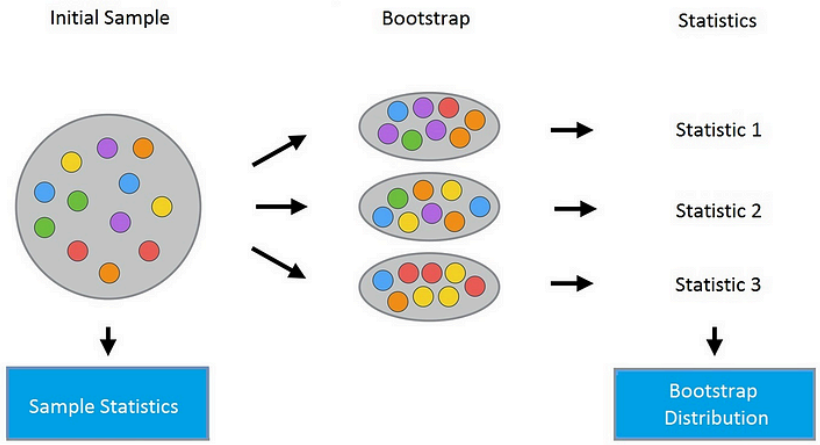
A população é dividida em subpopulações (conglomerados) heterogêneas distintas (quarteirões, residências, famílias, bairros, etc.). Alguns dos conglomerados são selecionados segundo amostra aleatória simples e **todos os elementos** nos conglomerados selecionados são observados. Note que a amostragem é feita sobre os conglomerados, e não mais sobre os indivíduos da população. Este procedimento amostral é adequado quando é possível dividir a população em um grande número de pequenas subpopulações.

Geralmente este método é usado quando os sistemas de referência da população não são adequados e o custo para atualização é alto, ou quando identificar os elementos da população em campo é cara e consome muito tempo.

Exemplo de Aplicação: Suponha que o objetivo de uma pesquisa seja determinar a renda média familiar de moradores de uma cidade. Dificilmente dispõe-se de uma lista de famílias, a unidade elementar da população de interesse. Pode-se usar como sistema de referência a lista de setores censitários do IBGE. Neste caso, os setores censitários seriam os conglomerados.

3.5 *Bootstrapping*

O **Bootstrapping** é uma interessante técnica de reamostragem, que consiste em gerar amostras aleatórias a partir de uma amostra de uma população finita:



Fonte: Complex systems and AI

Esta é uma técnica fácil de aplicar utilizando programação devido aos milhares de processos de reamostragem e vem ganhando espaço dentro de *Machine Learning* em modelos baseados em testes massivos, servindo como base para o modelo de floresta aleatória (*Random Forest*).

Técnicas de testagem massivas utilizando amostras aleatórias podem ser conhecidas também como **Métodos de Monte Carlo.**

Exemplo de Aplicação: Seja a variável aleatória com distribuição de probabilidade: P(X=3)=0,4; P(X=6)=0,3; P(X=8)=0,3.

```
# Função para a esperança
def esperanca(X, P):
    E = 0
    for i in range(0, len(X)):
        E = E + X[i]*P[i]
    return E

# Função para a variância
def variancia(X ,P):
    E = 0; E2 = 0
    for i in range(0, len(X)):
```

```
        E = E + X[i]*P[i]
        E2 = E2 + (X[i]**2)*P[i]
    V = E2-E**2
    return V

# Vetor de Eventos
X = [3,6,8]

# Vetor de Probabilidades
P = [0.4,0.3,0.3]

# Cálculo da Esperança
E = esperanca(X,P)

# Cálculo da Variância
V = variancia(X,P)

# Print das métricas
print("Esperança: ", E)
print("Variância: ", V)

# Tamanho de amostras
n = 40

# Número de simulações
ns = 1000

# Vetor vazio para armazenar a média amostral
```

```
vx = [] # armazena a média amostral

# Laço para as simulações
for s in range(0, ns):
    A = np.random.choice(X, n, p = P)
    vx.append(np.mean(A))

# Mostra o gráfico da distribuição criada
plt.figure(figsize=(8,6))
plt.hist(x=vx, bins='auto',color='#0504aa', alpha=0.7, rwidth=0.85, density = True)
plt.xlabel(r'$\bar{X}$', fontsize=20)
plt.ylabel(r'$P(\bar{X})$', fontsize=20)
plt.show()

# Compara as métricas da amostra e da população
print("Média das amostras: ", np.mean(vx))
print("Média da população: ", E)
```

4. Tamanho Amostral

Ao se realizar uma amostra para inferir uma determinada função paramétrica (média, máximo ou outra função de um parâmetro), há um erro associado ao planejamento amostral. À medida que o tamanho da amostra aumenta, o erro do estimador decresce. Vale ressaltar que uma amostra muito grande pode implicar custos desnecessários, enquanto que uma amostra pequena pode tornar a pesquisa inconclusiva. Deste modo, o ponto chave de um levantamento amostral é determinar o tamanho da amostra. Uma forma de garantir que o tamanho amostral seja significativo em relação à população, é ultilizado do Teorema Central do Limite.

Recapitulando, seja uma amostra aleatória $(x_1, x_2, \cdots, x_n)$ de uma variável aleatória X com qualquer distribuição, média μ e desvio padrão σ. A medida que n cresce, a distribuição de probabilidade da média amostral, $\overline{X}$, se aproxima de uma Normal com

média μ e desvio padrão $\frac{\sigma}{\sqrt{n}}$. Isto é $\overline{X} \sim N(\mu, \sigma^2/n)$. Se a transformação a baixo for realizada, então $Z \sim N(0,1)$.

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

O Teorema do Limite Central afirma que, com o aumento do tamanho da amostra, a distribuição das médias amostrais se aproxima de uma distribuição normal com média igual à média da população e desvio padrão igual ao desvio padrão da variável original dividido pela raiz quadrada do tamanho da amostra. Este fato é assegurado para n maior ou igual a 30.

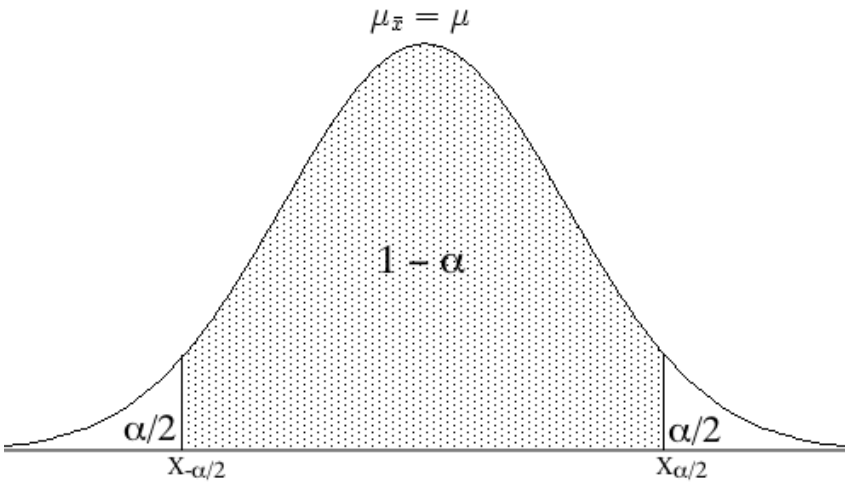
$$\sigma_{\bar{x}=\frac{\sigma}{\sqrt{n}}}$$

O desvio padrão das médias amostrais é conhecido como **erro padrão da média**.

5. Níveis de confiança e significância

O **nível de confiança** ($1-\alpha$) representa a probabilidade de acerto da estimativa. De forma complementar o **nível de significância** (α) expressa a probabilidade de erro da estimativa. O **nível de confiança** representa o grau de confiabilidade do resultado da estimativa estar dentro de determinado intervalo. Quando fixado em uma pesquisa um **nível de confiança** de 95%, por exemplo, é assumindo que existe uma probabilidade de 95% dos resultados da pesquisa representarem bem a realidade, ou seja, estarem corretos.

O **nível de confiança** de uma estimativa pode ser obtido a partir da área sob a curva normal como ilustrado na figura abaixo.



Fonte: Wikimedia

6. Erro inferencial e Intervalo de Confiança

O **erro inferencial** é definido pelo **desvio padrão das médias amostrais** $\sigma_{\bar{x}}$ e pelo **nível de confiança** determinado para o processo.

$$e = z\frac{\sigma}{\sqrt{n}}$$

Já o **intervalo de confiança** pode ser obtido adicionando o erro inferencial aos parâmetros de referência. Levantando os casos de intervalo de confiança para a média da população, pode-se ser aplicados em duas ocasiões:

- **Com desvio padrão populacional conhecido:** $\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}}$

- **Com desvio padrão populacional desconhecido:** $\mu = \bar{x} \pm z\frac{s}{\sqrt{n}}$

Exemplo de Aplicação: Suponha que os pesos dos sacos de arroz de uma indústria alimentícia se distribuem aproximadamente como uma normal de **desvio padrão populacional igual a 150 g**. Selecionada uma **amostra aleatória de 30 sacos** de um lote específico, obteve-se um **peso médio de 5.050 g**. Construa um intervalo de confiança para a **média populacional** assumindo um **nível de significância de 5%**.

Implementando a resolução do exemplo em *Python*, temos que:

```
# Carregando as funções para distribuição normal do SciPy
from scipy.stats import norm

# Parâmetros
X = 5050              # média amostral
desvpad = 150         # desvio padrão populacional
alpha = 0.05          # nível de significância
conf = 1 - alpha      # nível de confiança
n = 30                # tamanho da amostra

# Calculando a probabilidade
prob = conf + (alpha / 2)

# Calculando o Z-score para a probabilidade
z = norm.ppf(prob)
print("Z-score: ", np.round(z, 4))

# Definindo o erro inferencial
```

```
e = (z * desvpad) / np.sqrt(n)
print("Erro inferencial: ", np.round(e,4))


# Definindo o intervalo de confiança de duas formas
print("Intervalo de confiança pela fórmula: ", (X - e, X + e))
print("Intervalo de confiança usando o SciPy: ", norm.interval(conf, loc = X, scale = desvpad / np.sqrt(n)))
```

Na tabela abaixo, segue uma relação de valores que normalmente são utilizados para nível de confiança, com a probabilidade e o *Z-score*:

Nível de confiança	Valor da área sob a curva normal	z
90%	0,90	1,645
95%	0,95	1,96
99%	0,99	2,575

7. Cálculo do tamanho amostral baseado na estimativa da média populacional

7.1 População Infinita

Uma população é considerada infinita quando seu tamanho é muito grande.

Ao realizar o cálculo do tamanho da amostra n, deve-se levar em consideração o erro ϵ máximo que deseja-se assumir (ao estimar a função paramétrica) e o nível de confiança do resultado (probabilidade). Sendo assim, o problema consiste em determinar n de forma que: $P(|\ \overline{X} - \mu\ | \leq \epsilon) \simeq 1 - \alpha$

Mas pelo Teorema Central do Limite, a equação acima pode ser reescrita como: $$P\left(\mid \overline{X} - \mu \mid \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \simeq 1 - \alpha$$

Sendo assim, dados um erro máximo e nível de confiança, calcular o tamanho amostral consiste em:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = \epsilon \Longrightarrow n = \left(\frac{z_{\alpha/2}\sigma}{\epsilon}\right)^2$$

Exemplo de Aplicação: Um estudo a respeito do rendimento mensal dos chefes de domicílios no Brasil, determinou que o **erro máximo em relação a média seja de R\$$$ 100,00**. Sabendo que o desvio padrão populacional deste grupo de trabalhadores é de **R$$$ 3.323,39**. Para um **nível de confiança de 95%**, qual deve ser o tamanho da amostra de nosso estudo?

```
# Parâmetros
desvpad = 3323.39  # desvio padrão
e = 100          # erro inferencial
conf = 0.95        # nível de confiança
alpha = 1- conf    # nível de significância

# Calculando o Z-Score
z = norm.ppf(conf + (alpha / 2))

# Determinando o tamanho da amostra
n = ((z * desvpad) / e)**2

# Valor da amostra
print("Tamanho da amostra será: ", np.round(n, 0))
```

7.2 População Finita

No caso em que o tamanho populacional não é tão grande, a consideramos finita. **Caso a amostra tenha um tamanho n maior ou igual a 5% do tamanho da população N, considera-se que a população é finita**. Neste caso, aplica-se um fator de correção à fórmula vista anteriormente:

$$n = \frac{N(z_{\alpha/2}\sigma)^2}{(N-1)\epsilon^2 + (z_{\alpha/2}\sigma)^2}$$

Exemplo de Aplicação: Em um lote de **10.000 latas** de refrigerante foi realizada uma amostragem aleatória simples de **500 latas** e foi obtido o **desvio padrão amostral do conteúdo das latas igual a 12 ml**. O fabricante estipula um **erro máximo sobre a média populacional de apenas 5 ml**. Para garantir um **nível de confiança de 95%** qual o tamanho de amostra deve ser selecionado para este estudo?

```
# Parâmetros
desvpad = 12       # desvio padrão
e = 5              # erro inferencial
N = 10000          # Tamanho da população
conf = 0.95        # nível de confiança
alpha = 1 - conf   # nível de significância

# Cálculo do Z-score
z = norm.ppf(confianca + significancia / 2)

# Cálculo do tamanho da amostra
n = (N * (z * desvpad)**2)/(((N - 1) * e**2) + (z * desvpad)**2)

# Valor da amostra
print("Tamanho da amostra será: ", np.round(n, 0))
```

7.3 Variância populacional desconhecida

No caso em que a variância populacional é desconhecida, pode-se realizar uma amostragem aleatória preliminar (ao menos 30 elementos) para estimar a variância amostral e usá-la na equação acima.

$$\widehat{\sigma^2} = s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{X})^2}{N-1}$$

Materiais Complementares

Documentação do SciPy;

Artigo *Bootstrapping using Python and R* escrito por Michael Grogan;

Referências

Pedro A. Morettin, Wilton O. Bussab, Estatística Básica, 8ª edição

Peter Bruce, Andrew Bruce & Peter Gedeck, Practical Statistics for Data Scientists, 50+ Essential Concepts Using R and Python, 2ª edition

Ron Larson & Betsy Farber, Estatística Aplicada, 6ª edição.

Próximo Tópico