Universidade Federal do Ceará
Campus Jardins de Anita
Ciência de Dados

Monitoria Integrada de Estatística para Ciência de Dados

Professora: Elisângela Rodrigues
Monitoras: Bruna Barreto e Larissa Sousa

24 de outubro de 2023

Conteúdo

1 Introdução

Seja bem-vindo(a) à *Monitoria Integrada de Estatística para Ciência de Dados* do Campus de Itapajé da Universidade Federal do Ceará (UFC). Esta apostila foi cuidadosamente elaborada para auxiliá-lo(a) no estudo da Inferência Estatística, um campo fundamental da Ciência de Dados e da Estatística que permite tirar conclusões e fazer previsões com base em dados amostrais.

1.1 Objetivo da Apostila

Nosso principal objetivo é proporcionar uma abordagem abrangente e prática sobre a teoria e aplicação dessa área crucial da Estatística. Com enfoque na consolidação dos conceitos básicos principais, nossa intenção é tornar o aprendizado claro e acessível. Ao longo desta apostila, você terá a oportunidade de explorar de forma detalhada tópicos essenciais, como conceitos básicos, técnicas de amostragem, estimadores eficientes e estatísticas suficientes, métodos de estimação, testes de hipóteses e muito mais. Nosso objetivo é transmitir esses conceitos de maneira didática, para que você possa compreender seu significado e aplicação.

Além disso, esta apostila tem como objetivo explorar a aplicação desses conceitos por meio de exemplos práticos utilizando as linguagens de programação R e Python. Acreditamos que a combinação de teoria e prática será essencial para consolidar seu aprendizado e prepará-lo(a) para enfrentar desafios reais no campo da Ciência de Dados.

1.2 Importância das linguagens R e Python

Figura 1: R

R e Python são duas das linguagens de programação mais populares e poderosas no campo da Ciência de Dados e Estatística. Ambas possuem uma vasta coleção de pacotes e bibliotecas que facilitam a análise e manipulação de dados, permitindo a implementação de diversos métodos estatísticos de forma eficiente.

Figura 2: Python

Nesta apostila, você encontrará exemplos práticos em ambas as linguagens, visando proporcionar uma experiência completa e versátil para o aprendizado. Isso permitirá que você desenvolva habilidades práticas na aplicação da Inferência Estatística utilizando R e Python, tornando-se um profissional mais capacitado para enfrentar os desafios do mercado de trabalho e da pesquisa científica.

1.3 Instruções para os exemplos práticos

Os exemplos práticos nesta apostila foram projetados para serem executados em ambientes que possuam as versões mais recentes do R e do Python, juntamente com as bibliotecas estatísticas relevantes instaladas. Para cada exemplo, forneceremos os códigos necessários e explicações detalhadas sobre como interpretar os resultados.

Recomendamos que você execute e experimente os códigos em seu próprio ambiente, ajustando e explorando os resultados para aprofundar seu entendimento dos conceitos apresentados.

Estamos ansiosas para compartilhar com você todo o conhecimento desta apostila e acreditamos que o aprendizado da Inferência Estatística aliado às habilidades em R e Python abrirá um mundo de possibilidades na sua jornada acadêmica e profissional.

Desejamos a você uma excelente jornada de estudo e descobertas nesta emocionante área da Ciência de Dados e Estatística!

2 Inferência Estatística

De acordo com Morettin e Bussab (2017, p.276), o objetivo da Inferência Estatística é produzir afirmações sobre determinada característica da população a partir de informações colhidas de uma parte dessa população.

Ou seja, é como uma espécie de "ponte"que nos permite usar as informações obtidas da amostra para fazer afirmações ou estimativas sobre a população (maior). É como se a amostra representasse a população e, com base nessa representação, podemos fazer previsões, tirar conclusões e tomar decisões sobre a característica que estamos interessados em estudar.

2.1 Conceitos básicos

Esta seção será dedicada aos principais fundamentos necessários para o estudo da Inferência Estatística. Aqui, nos aprofundaremos nos conceitos essenciais que formam a base dessa área da Estatística.

2.1.1 População

Segundo Werkema (2014, p.10), população é a totalidade dos elementos de um universo sobre o qual desejamos estabelecer conclusões ou exercer ações. A população de interesse pode ser finita ou infinita. Uma população finita possui um número limitado de elementos. Já a população infinita possui um número não limitado de elementos."

Exemplos:

1. Pesquisa de opinião em Itapajé.

   Objetivo: Conhecer a proporção de moradores favoráveis a um projeto de construção de uma nova escola.

   População: Todos os modadores civis.

   Exemplo no Python

```python
import random

# Simulando a populacao com 490 moradores n o favoraveis e 510 favoraveis
populacao_itapaje = ['nao_favoravel'] * 490 + ['favoravel'] * 510

tamanho_populacao = 1000
tamanho_amostra = 100

amostra_itapaje = random.sample(populacao_itapaje, tamanho_amostra)

# Contando a proporcao de moradores favoraveis a nova escola
proporcao_favoravel = amostra_itapaje.count('favoravel') / tamanho_amostra

print("Proporcao de moradores favoraveis a nova escola:", proporcao_favoravel * 100,
    "%")
```

   Exemplo no R

```r
populacao_itapaje <- rep(c('nao_favoravel', 'favoravel'), c(490, 510))
tamanho_amostra <- 100

amostra_itapaje <- sample(populacao_itapaje, tamanho_amostra, replace = TRUE)

# Contando a proporcao de moradores favoraveis a nova escola
proporcao_favoravel <- sum(amostra_itapaje == 'favoravel') / tamanho_amostra

print(paste("Proporcao de moradores favoraveis a nova escola:", proporcao_favoravel *
    100, "%"))
```

2. Pesquisa sócio-econômica na Universidade Federal do Ceára - Campus Jardim de Anita.

   Objetivo: Estimar a renda média das famílias dos estudantes da UFC.

População: Todos os estudantes do campus.

Exemplo no Python

```python
import numpy as np

renda_familias_ufc = np.random.normal(loc=2000, scale=500, size=100)

media_renda = np.mean(renda_familias_ufc) # Estimando a renda media

media_formatada = "R$ {:.2f}".format(media_renda)

print("Estimativa da renda media das familias dos estudantes da UFC:", media_formatada)
```

Exemplo no R

```r
renda_familias_ufc <- rnorm(100, mean = 2000, sd = 500)

media_renda <- mean(renda_familias_ufc)

media_formatada <- paste("R$", format(media_renda, digits = 2, nsmall = 2))

print(paste("Estimativa da renda media das familias dos estudantes da UFC:",
    media_formatada))
```

Exercícios propostos

1) Em uma pequena cidade de 5000 eleitores vai haver uma eleição com os candidatos Tita e Niki. É feita uma prévia em que os 5000 eleitores são entrevistados, sendo que 2501 já se decidiram, definitivamente, por Tita. Determinar a probabilidade de que Niki ganhe a eleição.

2) A probabilidade de três motoristas serem capazes de dirigir até suas casas com segurança depois de ingerirem bebidas alcoólicas é 1/5, 1/6 e 1/2, respectivamente. Se decidirem dirigir até suas casas depois de beberem em uma festa, qual a probabilidade de que:

a) todos os três motoristas sofram acidentes;

b) ao menos um dos motoristas dirija até sua casa a salvo.

3) Renata pode ir para a esquerda, para a direita ou em frente, ao chegar a cada um dos cinco cruzamentos de um labirinto. Determine a probabilidade de Renata atravessar o labirinto corretamente, havendo somente um caminho correto possível.

4) Quantas vezes deve uma moeda honesta ser jogada para que a probabilidade de dar cara pelo menos uma vez seja igual ou maior que 0,9?

2.1.2 Amostra

Morettin e Bussab (2017, p.276) dizem que, uma amostra e um subconjunto de uma populacao.

Exemplos:

1. Pesquisa de opnião em Itapajé.

   Amostra: Conjunto de 100 a 500 moradores que serão entrevistados pelos pesquisadores.

   Exemplo no Python

```python
import random

populacao_itapaje = ['nao_favoravel'] * 490 + ['favoravel'] * 510 # Simulando a
    popula  o

tamanho_amostra = random.randint(100, 500)
```

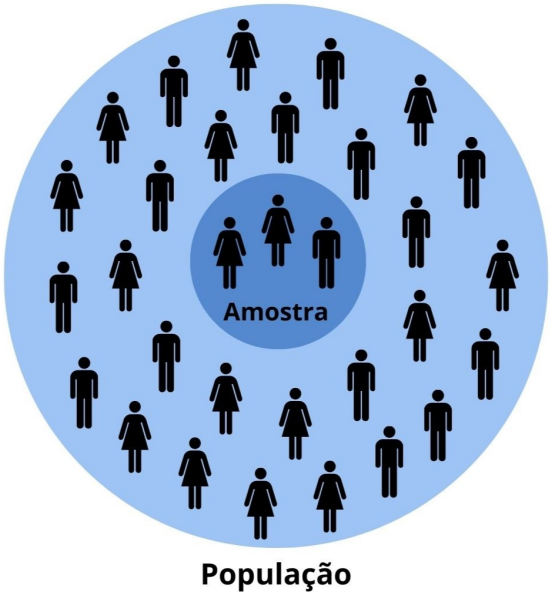


Figura 3: População e amostra

```
6
7 amostra_itapaje = random.sample(populacao_itapaje, tamanho_amostra) # Amostra
      aleat ria
8
9 proporcao_favoravel = amostra_itapaje.count('favoravel') / tamanho_amostra
10
11 print("Amostra de moradores entrevistados:", tamanho_amostra)
```

Exemplo no R

```
1 populacao_itapaje <- rep(c('nao_favoravel', 'favoravel'), c(490, 510))
2
3 tamanho_amostra <- sample(100:500, 1)
4
5 amostra_itapaje <- sample(populacao_itapaje, tamanho_amostra, replace = FALSE)
6
7 proporcao_favoravel <- sum(amostra_itapaje == 'favoravel') / tamanho_amostra
8
9 cat("Amostra de moradores entrevistados:", tamanho_amostra, "\n")
```

2. Pesquisa sócio-econômica na Universidade Federal do Ceára - Campus Jardim de Anita.

   Objetivo: Conjunto de 50 estudantes do campus que serão entrevistados pelos pesquisadores.

   Exemplo no Python

```
1 import numpy as np
2
3 tamanho_amostra = 50
4 renda_familias_ufc = np.random.normal(loc=2000, scale=500, size=tamanho_amostra)
5
6 media_renda = np.mean(renda_familias_ufc) # Estimando a renda m dia
7
8 print("Estimativa da renda media das familias dos estudantes: R$", media_renda)
```

   Exemplo no R

```
1 tamanho_amostra <- 50
2 renda_familias_ufc <- rnorm(tamanho_amostra, mean = 2000, sd = 50)
3
4 media_renda <- mean(renda_familias_ufc)
```

```
5
6 print(paste("Estimativa da renda media das familias dos estudantes: R$", media_renda))
```

Na tabela a seguir, apresentaremos os principais parâmetros da população e estatísticas da amostra. Esses parâmetros e estatísticas são valores numéricos que nos permitem resumir e descrever características importantes da população e da amostra.

Tabela 1: Parâmetros e estatísticas

	Parâmetros (população)	**Estatísticas (amostra)**
Média	μ	$\overline{x}$
Desvio padrão	σ	s
Variância	σ^2	s^2
Proporção	p	$\widehat{p}$

Os parâmetros populacionais, como a média μ e o desvio padrão σ, são medidas que resumem características da população como um todo. Por outro lado, as estatísticas amostrais, como a média amostral $\overline{x}$ e o desvio padrão amostral s, são calculadas a partir dos dados coletados na amostra e servem como estimativas dos parâmetros populacionais.

Exercícios propostos

5) Uma vereadora queria saber o que seus eleitores pensavam de um novo plano diretor. Ela selecionou aleatoriamente 75 nomes da lista telefônica da cidade e realizou uma entrevista pelo telefone. Identifique a população e a amostra neste cenário.

a) A população é todo mundo que está na lista telefônica da cidade; a amostra é as 75 pessoas selecionadas.

b) A população é os habitantes da cidade; a amostra é os eleitores registrados na cidade.

c) A população é os eleitores registrados na cidade; a amostra é todo mundo que está na lista telefônica da cidade.

6) Lúcio quer saber se a comida que ele serve em seu restaurante está dentro de uma faixa segura de temperaturas. Ele aleatoriamente seleciona 70 pratos e mede a temperatura antes de servir ao cliente. Identifique a população e amostra nesse cenário.

a) A população é todos os pratos quentes que Lúcio serve; a amostra é os pratos que estão em uma temperatura segura.

b) A população é os 70 pratos selecionados; a amostra é os pratos que estão em uma temperatura segura.

c) A população é todos os pratos que Lúcio serve; a amostra é os 70 pratos selecionados.

7) Um grupo de bibliotecários está interessado no número de livros e outras mídias que seus clientes retiram da biblioteca. Eles examinam os registros de retirada de 150 clientes adultos selecionados aleatoriamente. Identifique a população e a amostra nesse cenário.

a) A população é todos os clientes adultos da biblioteca; a amostra é os 150 clientes selecionados.

b) A população é todos os clientes da biblioteca; a amostra é os clientes adultos da biblioteca.

c) A população é todos os clientes que retiraram ao menos 1 livro da biblioteca; a amostra é os 150 clientes selecionados.

8) Uma supervisora de fábrica seleciona aleatoriamente 40 hastes rosqueadas dentre aquelas produzidas aquela semana na fábrica. Em seguida, ela testa sua resistência à tração. Identifique a população e a amostra nesse cenário.

a) A população é todas as hastes rosqueadas já produzidas na fábrica; a amostra é as hastes rosqueadas produzidas aquela semana.

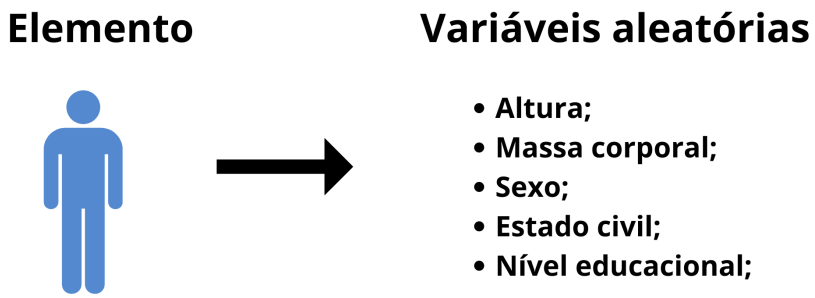
b) A população é as hastes rosqueadas produzidas na fábrica aquela semana; a amostra é as 40 hastes rosqueadas selecionadas.

c) A população é todas as hastes rosqueadas do mundo; a amostra são todas as hastes rosqueadas produzidas na fábrica.

2.1.3 Variável aleatória (v.a.)

Gupta e Irwin Guttman (2016, p.91) afirmam que uma variável aleatória é uma função real de valores individuais X(e) definida para cada elemento e no espaço amostral S.

Figura 4: Variável aleatória (v.a.)



Variáveis aleatórias são fundamentais, sendo uma ferramenta que nos ajuda a entender e modelar o comportamento de eventos incertos. Existem dois tipos de variáveis aleatórias: discretas e contínuas.

1. Variável Aleatória Discreta:

   - Uma variável aleatória discreta é aquela que assume um conjunto finito ou infinito enumerável;
   - Ela é usada para descrever situações em que as observações podem ser quantificadas em termos de contagens ou números inteiros;
   - Exemplos de variáveis aleatórias discretas incluem o número de caras ao lançar uma moeda várias vezes, o número de alunos em uma sala de aula, ou o número de peças defeituosas em um lote de produção;
   - A função de probabilidade de uma variável aleatória discreta associa a cada valor possível dessa variável uma probabilidade correspondente.

2. Variável Aleatória Contínua:

   - Uma variável aleatória contínua é aquela que pode assumir qualquer valor em um intervalo contínuo de números reais;
   - Ela é usada para modelar situações em que as observações podem ser medidas com precisão infinita e não estão restritas a valores inteiros;
   - Exemplos de variáveis aleatórias contínuas incluem a altura de uma pessoa, a temperatura ambiente, ou o tempo necessário para completar uma tarefa;
   - A função de densidade de probabilidade de uma variável aleatória contínua é usada para descrever a probabilidade de que a variável assuma valores em intervalos específicos.

Exemplo:
Considere um questionário com uma série de perguntas, em que cada pergunta possui uma resposta correta. Suponha que você aplique esse questionário a várias pessoas e, para cada participante, conte o número de respostas corretas que ele fornece. Nesse caso, a variável aleatória X pode ser definida como o número de

respostas corretas em um questionário. Seja x um valor numérico possível da variável aleatória X, como 0, 1, 2, 3, etc. A variável aleatória X mapeia os resultados do experimento (número de respostas corretas) em valores numéricos x.

Exemplo no Python

```
import random

respostas_corretas = ['A', 'B', 'C', 'D', 'B', 'A', 'C', 'D', 'B', 'A'] # Definindo as
    respostas corretas

# Simulando as respostas dos participantes do question rio
tamanho_amostra = 100
amostra_participantes = []

for _ in range(tamanho_amostra):
    respostas_participante = random.choices(['A', 'B', 'C', 'D'], k=10)
    amostra_participantes.append(respostas_participante)

num_respostas_corretas = [] # Contando o n mero de respostas corretas de cada participante

for respostas_participante in amostra_participantes:
    num_corretas = sum(respostas_participante[i] == respostas_corretas[i] for i in range(10)
    )
    num_respostas_corretas.append(num_corretas)

print("N mero de respostas corretas de cada participante:", num_respostas_corretas)
```

Exemplo no R

```
respostas_corretas <- c('A', 'B', 'C', 'D', 'B', 'A', 'C', 'D', 'B', 'A')

tamanho_amostra <- 100
amostra_participantes <- matrix(sample(c('A', 'B', 'C', 'D'), size=10*tamanho_amostra,
    replace=TRUE), ncol=10)

num_respostas_corretas <- apply(amostra_participantes, 1, function(respostas_participante) {
    sum(respostas_participante == respostas_corretas)})

print(paste("Numero de respostas corretas de cada participante:", num_respostas_corretas))
```

Exercícios propostos

9) Se as probabilidades de uma criança da faixa etária de 6 a 16 anos consultar um dentista 0, 1, 2, 3, 4, 5 ou 6 vezes por ano são 0,09, 0,25, 0,29, 0,18, 0,14, 0,03 e 0,02, quantas vezes podemos esperar que uma criança daquela faixa etária consulte um dentista em um ano?

10) Uma variável aleatória discreta x tem função de probabilidade dada por:

Valores de X	0	2	6	8
Probabilidades	0,2	0,3	0,3	0,2

A variância de x é igual a

a) 7,2
b) 7,6
c) 8,0
d) 8,4
e) 8,8

11) Os pais de uma estudante prometeram-lhe uma recompensa de R$100,00 se ela obtiver A em estatística, R$50,00 se obtiver B, mas nenhuma recompensa nos demais casos. Qual é o valor esperado se as probabilidades dela obter conceitos A e B são 0,32 e 0,40, respectivamente?

12) Uma variável aleatória discreta x tem função de probabilidade dada por:
A soma dos valores da média e da mediana de x é igual a

X	0	2	4	6
$P(X = x)$	0,3	0,4	0,2	0,1

a) 2,2

b) 3,2

c) 4,2

d) 5,2

e) 6,2

2.1.4 Parâmetros

Para Morettin e Bussab (2017, p.284), um parametro e uma medida usada para descrever uma característica da populacao.

Exemplo:
Imagine que temos uma caixa contendo 100 bolas coloridas. Nessa caixa, 60 bolas são vermelhas e 40 são azuis. O parâmetro de interesse nesse caso seria a proporção de bolas vermelhas em toda a caixa, que é de 60%.

Exemplo no Python

```python
import random

# Simulando a caixa com as bolas coloridas
total_bolas = 100
bolas_vermelhas = 60
bolas_azuis = 40

caixa_bolas = ['vermelha'] * bolas_vermelhas + ['azul'] * bolas_azuis # Criando uma lista com todas as bolas

proporcao_vermelhas = caixa_bolas.count('vermelha') / total_bolas

print("Proporcao de bolas vermelhas na caixa:", proporcao_vermelhas)
```

Exemplo no R

```r
total_bolas <- 100
bolas_vermelhas <- 60
bolas_azuis <- 40

caixa_bolas <- c(rep('vermelha', bolas_vermelhas), rep('azul',
     bolas_azuis))

proporcao_vermelhas <- sum(caixa_bolas == 'vermelha') / total_bolas

print(paste("Proporcao de bolas vermelhas na caixa:",
     proporcao_vermelhas))
```

Exercícios propostos

13) Um fabricante de smartphones deseja verificar a durabilidade da bateria de seus dispositivos. Eles selecionam aleatoriamente 100 smartphones de um lote de produção recente e os submetem a testes rigorosos de bateria. Identifique o parametro neste cenário.

a) O número total de smartphones no lote de produção recente.

b) A durabilidade média da bateria em todos os smartphones do lote de produção recente.

c) A cor da capa dos smartphones no lote.

d) O custo de produção de um único smartphone no lote.

14) Uma empresa de pesquisa quer estudar a opinião dos habitantes de uma cidade sobre a construção de um novo parque público. Eles realizam entrevistas com 500 pessoas em um shopping center da cidade. Identifique o parametro neste cenário.

a) O número total de lojas no shopping center.

b) O preço médio dos itens vendidos no shopping center.

c) A altura média das pessoas entrevistadas no shopping center.

d) A proporção de habitantes da cidade que apoiam a construção do novo parque público.

15) Uma empresa de streaming de música deseja entender as preferências musicais de seus usuários. Eles analisam aleatoriamente as escolhas musicais de 1000 usuários premium. Identifique o parametro neste cenário.

a) O número total de músicas disponíveis no catálogo da empresa de streaming.

b) A porcentagem de usuários premium que preferem um gênero musical específico.

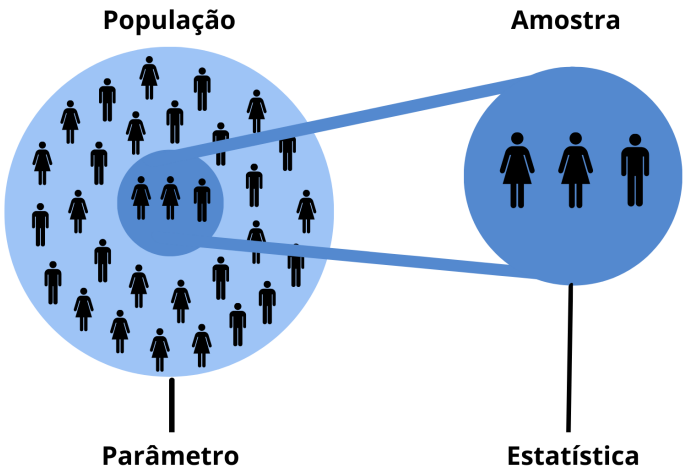
c) A receita total da empresa de streaming no último trimestre.

d) A média de idade dos usuários premium.

2.1.5 Estatísticas

Werkema (2014, p.13) afirma que uma estatística é uma função das observações amostrais, que não depende de parâmetros desconhecidos.

Figura 5: Parâmetro e estatística



Exemplo:
Suponha que, em vez de considerar todas as 100 bolas da caixa, decidimos retirar uma amostra de 20 bolas aleatoriamente (com reposição) e contar quantas são vermelhas. Após realizar várias amostragens, encontramos que, em média, 14 bolas são vermelhas em cada amostra. Nesse caso, a média amostral de 14 bolas vermelhas é a estatística.

Exemplo no Python

```
import random

# Simulando a caixa com as bolas coloridas
total_bolas = 100
bolas_vermelhas = 60
bolas_azuis = 40

caixa_bolas = ['vermelha'] * bolas_vermelhas + ['azul'] * bolas_azuis

amostra = random.choices(caixa_bolas, k=20) # retirada de uma amostra aleatoria
```

```

bolas_vermelhas_amostra = amostra.count('vermelha')

media_amostral = bolas_vermelhas_amostra / 20

print("Media amostral de bolas vermelhas:", media_amostral)
```

Exemplo no R

```
total_bolas <- 100
bolas_vermelhas <- 60
bolas_azuis <- 40

caixa_bolas <- c(rep('vermelha', bolas_vermelhas), rep('azul',
bolas_azuis))

amostra <- sample(caixa_bolas, 20, replace = TRUE) # retirada de uma amostra aleat ria de
    20 bolas (com reposi  o)

bolas_vermelhas_amostra <- sum(amostra == 'vermelha')

media_amostral <- bolas_vermelhas_amostra / 20

print(paste("M dia amostral de bolas vermelhas:", media_amostral))
```

Exercícios propostos

16) Em um estudo sobre o desempenho dos alunos em matemática, uma estatística utilizada foi a média das notas dos alunos em uma amostra de 50 estudantes. Identifique a estatística neste cenário.

a) A média das idades dos alunos.

b) A variância das notas dos alunos.

c) O número total de alunos na escola.

d) A altura média dos alunos.

17) Um fabricante de chips de computador testou a velocidade de processamento de 1000 chips de uma determinada linha de produção. Eles calcularam a média da velocidade de processamento dos chips. Identifique a estatística neste cenário.

a) O desvio padrão da velocidade de processamento dos chips.

b) A velocidade de processamento de um chip específico.

c) A média da velocidade de processamento dos chips.

d) O custo total de produção dos chips.

18) Uma empresa de pesquisa de mercado coletou dados sobre as preferências de sabor de 300 consumidores de sorvetes e calculou a porcentagem de pessoas que preferem o sabor de chocolate. Identifique a estatística neste cenário.

a) A porcentagem de pessoas que preferem o sabor de chocolate.

b) O lucro da empresa no último trimestre.

c) O preço médio de um sorvete.

d) A quantidade total de sorvetes vendidos pela empresa.

19) Um cientista coletou dados sobre a temperatura do oceano em diferentes profundidades e calculou a média das temperaturas em sua amostra. Identifique a estatística neste cenário.

a) O desvio padrão das temperaturas do oceano.

b) O número total de espécies marinhas na região.

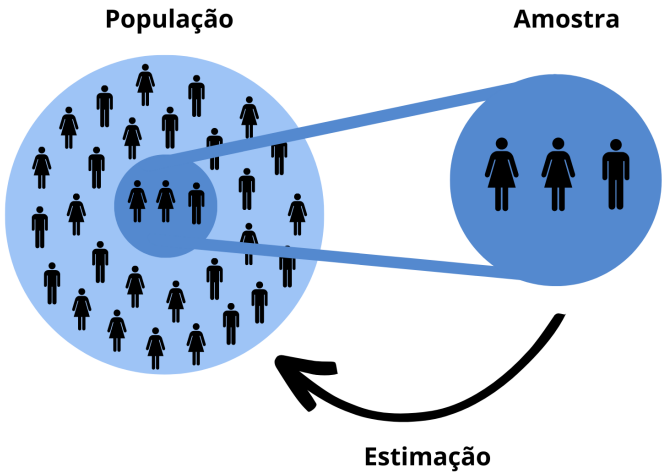
c) A média das temperaturas do oceano em diferentes profundidades.

d) A temperatura do oceano em uma profundidade específica.

2.1.6 Estimativas

Morettin e Bussab (2017, p.276) dizem que um estimador T do parametro e qualquer funcao das observacoes da amostra, ou seja, $T = g(X_1, X_2, ..., X_n)$.

Figura 6: Estimação



Exemplo:

Suponha que um pesquisador deseja estimar a altura média dos estudantes de uma universidade. A população de interesse é composta por todos os estudantes matriculados na universidade.

Para obter uma estimativa da altura média, o pesquisador seleciona uma amostra aleatória de 100 estudantes da universidade e mede suas alturas. Suponha que a média das alturas na amostra seja 170 centímetros.

Nesse caso, a média amostral de 170 centímetros é uma estimativa da altura média dos estudantes na universidade. O estimador, representado por $\overline{x}$, é a média amostral obtida através da fórmula:

$$\overline{x} = \frac{\text{Soma das alturas da amostra}}{\text{Tamanho da amostra}} = \frac{170 \text{ cm}}{100} = 1,70 \text{ m} \tag{1}$$

O parâmetro, representado por μ, é a altura média na população de todos os estudantes da universidade. Como é difícil medir a altura de todos os estudantes, o pesquisador usa a amostra para obter uma estimativa da altura média.

Assim, a estimativa $\overline{x} = 1,70$ metros é usada para fazer inferências sobre o parâmetro μ, a altura média na população de todos os estudantes da universidade.

Exemplo no Python

```python
import numpy as np

altura_estudantes = np.random.normal(loc=170, scale=5, size=100) # Simulando as alturas

media_amostral = np.mean(altura_estudantes)

print("Media amostral das alturas:", media_amostral, "cm")
```

Exemplo no R

```r
altura_estudantes <- rnorm(100, mean = 170, sd = 5)

media_amostral <- mean(altura_estudantes)

print(paste("Media amostral das alturas:", media_amostral, "cm"))
```

Com esses conceitos fundamentais, estaremos preparados para avançar na compreensão dos modelos probabilísticos e aplicá-los em diferentes contextos estatísticos.

Exercícios propostos

20) Em um estudo sobre o tempo médio que os alunos levam para concluir um exame, um pesquisador calcula a média do tempo gasto por uma amostra de 50 alunos. Qual dos seguintes é um exemplo de estimador do parâmetro real, neste caso, o tempo médio gasto por todos os alunos?

a) O tempo gasto pelo primeiro aluno da amostra.

b) A média do tempo gasto pela amostra de 50 alunos.

c) A média do tempo gasto por todos os alunos da escola.

d) O tempo gasto pelo aluno do meio da amostra.

21) Um fabricante de refrigerantes deseja estimar a média de açúcar em todos os refrigerantes de uma linha de produção. Eles coletam uma amostra de 100 garrafas e calculam a média do teor de açúcar dessas garrafas. Qual dos seguintes é um exemplo de estimador do parâmetro real, neste caso, a média real de açúcar em todas as garrafas?

a) O teor de açúcar na primeira garrafa da amostra.

b) A média do teor de açúcar em todas as garrafas da linha de produção.

c) A diferença entre o teor de açúcar na primeira e na última garrafa da amostra.

d) A média do teor de açúcar na amostra de 100 garrafas.

22) Um pesquisador deseja estimar a proporção de pessoas em uma cidade que apoiam uma determinada política. Ele conduz uma pesquisa com uma amostra de 500 pessoas e registra quantas delas apoiam a política. Qual dos seguintes é um exemplo de estimador do parâmetro real, neste caso, a proporção real de apoiadores da política em toda a cidade?

a) O número de pessoas na cidade que não apoiam a política.

b) A diferença entre o número de apoiadores na amostra e o número de não apoiadores na cidade.

c) A proporção de apoiadores na amostra de 500 pessoas.

d) A média das idades das pessoas na cidade.

23) Um agricultor quer estimar a produção média de maçãs por árvore em seu pomar. Ele escolhe aleatoriamente 50 árvores, colhe todas as maçãs delas e calcula a média da produção. Qual dos seguintes é um exemplo de estimador do parâmetro real, neste caso, a produção média de maçãs por árvore em todo o pomar?

a) A média da produção das 50 árvores amostradas.

b) O número total de maçãs colhidas em todo o pomar.

c) A diferença entre a produção da primeira e da última árvore da amostra.

d) A altura média das árvores no pomar.

2.2 Modelos probabilísticos (ou distribuições de probabilidades)

Existem vários modelos probabilísticos utilizados em diferentes contextos estatísticos, como o modelo Normal (Gaussiana), Binomial, Poisson, Uniforme, Exponencial, Gamma, Bernoulli, Qui-quadrado, t de Student, F de Fisher, entre outras. Esses modelos podem ser agrupados em dois tipos: os destinados a variáveis contínuas e, outros a variáveis discretas. Cada um possui suas próprias características e propriedades matemáticas. Alguns deles são amplamente conhecidos e utilizados, enquanto outros têm aplicações mais específicas em campos como ciência, engenharia e economia.

Nesta apostila, nosso foco será no estudo de quatro modelos principais: o modelo Normal (ou Gaussiana), o modelo Exponencial, o modelo Binomial e o modelo de Poisson. Dois deles são discretos, ou seja, seus valores possíveis são um conjunto discreto de números inteiros ou binários, enquanto os outros dois são contínuos, assumindo um intervalo contínuo de números reais. Ao compreender esses quatro modelos, você estará preparado(a) para analisar e interpretar uma variedade de dados em diferentes contextos estatísticos.

2.2.1 Modelo Normal (Gaussiana)

A função de densidade de probabilidade (pdf) de uma variável aleatória normal X com média μ e desvio padrão σ é dada por:

$$f\left(x|\mu,\sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{\left(x-\mu^2\right)}{2\sigma^2}\right], -\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0 \tag{2}$$

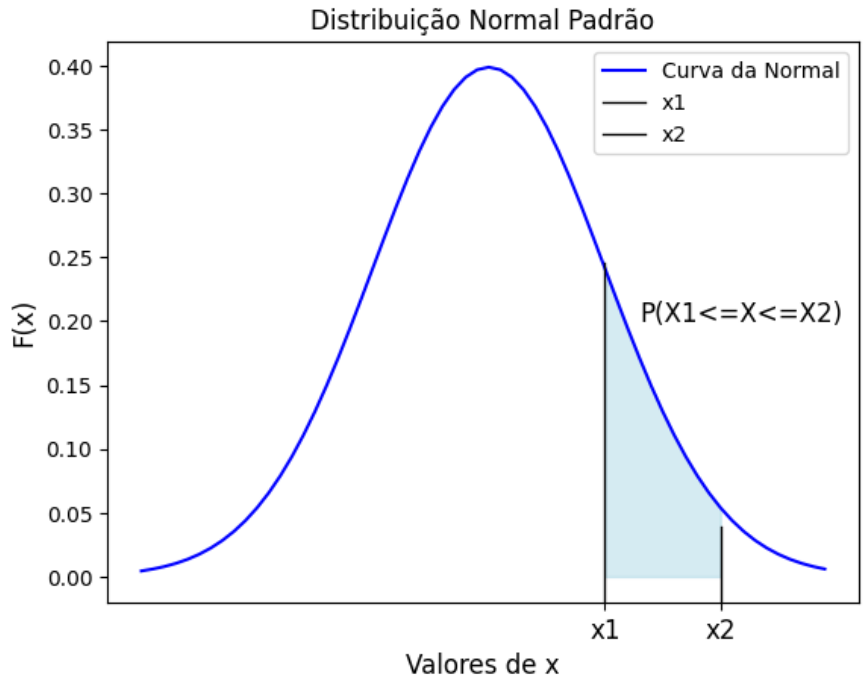
Uma aplicação interessante da distribuição normal é no estudo da inteligência humana, representada pelo Quociente de Inteligência (QI). O QI é uma medida que busca quantificar a inteligência de uma pessoa em relação à média da população. Na distribuição normal, é esperado que o QI das pessoas siga uma distribuição aproximadamente normal.

Quando medimos o QI de uma grande amostra de indivíduos em uma população, é comum observar que a distribuição dos QIs se assemelha a uma curva normal. O ponto mais alto da curva representa a média dos QIs da população, que geralmente é definida como 100. A partir da média, a curva se espalha simetricamente para os dois lados, com o desvio padrão controlando a dispersão dos QIs em relação à média.

Vale salientar que em uma distribuição normal:

- a probabilidade de um valor singular é zero;
- só há sentido em determinar probabilidade de intervalos.

Figura 7: Gráfico da Distribuição Normal



A probabilidade de que a variável esteja dentro do intervalo [x1, x2] é igual à região sob a curva que está delimitada por x1 e x2.

Essa área pode ser determinada através de cálculos de integração, mas, na prática, pode ser facilmente encontrada utilizando uma tabela que fornece diretamente a área entre a média e um valor específico da variável. Em resumo, estamos lidando com uma curva normal padrão, na qual a variável X é substituída por seu escore padronizado Z:

$$Z = \frac{X - \mu}{\sigma} \tag{3}$$

em que é a média e o desvio padrão. Pode ser verificado que a variável reduzida Z possui média 0 e desvio padrão 1.

Figura 8: By http://www.gutenberg.org/ebooks/20662 - Terman, L. (1916). The Measurement of Intelligence. Boston. Houghton Mifflin. p. 66 (Figure 2)

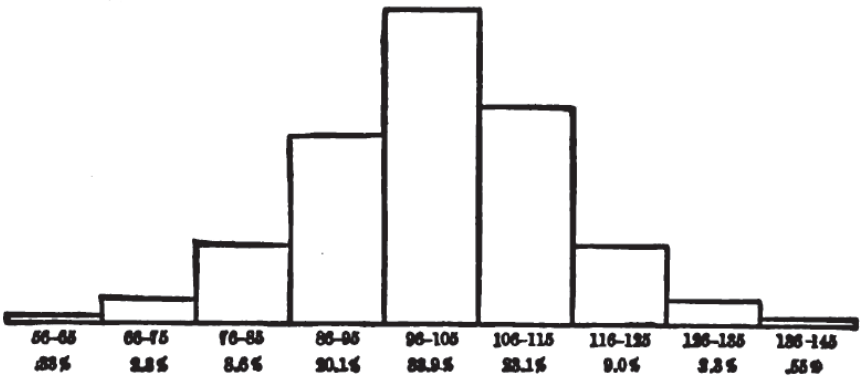


Fig. 2. DISTRIBUTION OF I Q'S OF 905 UNSELECTED CHILDREN, 5–14 YEARS OF AGE

Exemplo:

Imagine que o Quociente de Inteligência (QI) de uma determinada população siga uma distribuição normal, com média $\mu = 100$ e desvio padrão $\sigma = 15$. Nós podemos usar a função de densidade de probabilidade (pdf) da distribuição normal para calcular a probabilidade de encontrar uma pessoa com um QI específico ou em um determinado intervalo.

Exemplo no Python

Vamos usar a biblioteca *scipy.stats* para trabalhar com a distribuição normal e calcular algumas probabilidades.

```
import scipy.stats as stats

# Definindo os par metros da distribui  o normal
media = 100
desvio_padrao = 15

probabilidade_intervalo = stats.norm.cdf(115, media, desvio_padrao) - stats.norm.cdf(85,
    media, desvio_padrao) # Calculando a probabilidade de encontrar uma pessoa com QI entre
    85 e 115

print(f"\nA probabilidade de encontrar uma pessoa com QI entre 85 e 115    de {
    probabilidade_intervalo: .2f} %")

qi_percentil_90 = stats.norm.ppf(0.9, media, desvio_padrao) # Calculando o QI correspondente
     ao percentil 90

print(f"\nO QI correspondente ao percentil 90    de {qi_percentil_90:.2f}")
```

Exemplo no R

Vamos usar a função *pnorm* para calcular a probabilidade acumulada e *qnorm* para calcular o percentil.

```
media <- 100
desvio_padrao <- 15

probabilidade_intervalo <- pnorm(115, media, desvio_padrao) - pnorm(85, media, desvio_padrao
    )

```

```r
cat("A probabilidade de encontrar uma pessoa com QI entre 85 e 115    de", round(
    probabilidade_intervalo, 2), "\n")

qi_percentil_90 <- qnorm(0.9, media, desvio_padrao)

cat("O QI correspondente ao percentil 90    de", round(qi_percentil_90, 2), "\n")
```

Exercícios propostos

24) A Empresa Mandacaru S.A. produz televisores e garante a restituição da quantia paga se qualquer televisor apresentar algum defeito grave no prazo de seis meses. Ela produz televisores do tipo A (comum) e do tipo B (luxo), com um lucro respectivo de $ 100 e $ 200 caso não haja restituição, e com um prejuízo de $ 300 e $ 800 caso haja restituição. Suponha que o tempo para a ocorrência de algum defeito grave seja, em ambos os casos, uma variável aleatória com distribuição normal, respectivamente, com médias de nove e doze meses, e desvios padrões de dois e três meses. Se tivesse de planejar uma estratégia de marketing para a empresa, você incentivaria as vendas dos aparelhos do tipo A ou do tipo B?

25) Em uma distribuição normal, 30% dos elementos são menores que 45 e 10% são maiores que 64. Calcular os parâmetros que definem a distribuição (média e desvio padrão).

26) O consumo de gasolina por km rodado, para certo tipo de carro, em determinadas condições de teste, tem uma distribuição normal média de 100 ml e desvio padrão de 5 ml. Pede-se calcular a probabilidade de:

a) um carro gastar de 95 a 110 ml;

b) em um grupo de seis carros, tomados ao acaso, encontrarmos três carros que gastaram menos de 95 ml;

c) idem, todos terem gasto menos que 110 ml.

27) O tempo de vida de transistores produzidos pela Indústria Zeppelin Ltda. tem distribuição aproximadamente normal, com valor esperado e desvio padrão igual a 500 horas e 50 horas, respectivamente. Se o consumidor exige que pelo menos 95% dos transistores fornecidos tenham vida superior a 400 horas, pergunta-se se tal especificação é atendida. Justifique.

2.2.2 Modelo Exponencial

Em uma situação clássica em que uma distribuição exponencial surge, considere um processo de Poisson com média λ, onde contamos os eventos que ocorrem em um determinado intervalo de tempo ou espaço. Seja X o tempo de espera até o primeiro evento ocorrer. Nesse contexto, para um valor específico $x0$

- Um processo de Poisson é usado para contar eventos que ocorrem em um intervalo de tempo ou espaço, com uma taxa média de ocorrência de λ.
- A distribuição exponencial descreve a probabilidade do tempo de espera ser menor ou igual a um valor específico x.

$$P(X > x) = e^{-x\lambda} \text{ da mesma forma } P(X \leq x) = 1 - e^{-x\lambda} \tag{4}$$

Exemplo:
Suponha que estamos estudando o tempo de espera até um ônibus passar em um ponto de ônibus em uma determinada rua. Sabemos que, em média, um ônibus passa a cada 15 minutos. Nesse caso, podemos modelar o tempo de espera até o próximo ônibus como uma distribuição exponencial com parâmetro $\lambda = \frac{1}{15}$, já que a taxa média de ocorrência de ônibus é de um ônibus a cada 15 minutos.

Exemplo no Python
Vamos usar a biblioteca *scipy.stats* para trabalhar com a distribuição exponencial e calcular algumas probabilidades.

```python
import scipy.stats as stats

# Definindo o par metro lambda
lambda_parametro = 1 / 15

# Calculando a probabilidade de esperar mais de 30 minutos pelo pr ximo   nibus
probabilidade_maior_30_min = 1 - stats.expon.cdf(30, scale=1
    / lambda_parametro)

print(f"A probabilidade de esperar mais de 30 minutos pelo pr ximo   nibus
       de {probabilidade_maior_30_min:.2f}.")

# Calculando a probabilidade de esperar menos de 10 minutos pelo pr ximo   nibus
probabilidade_menor_10_min = stats.expon.cdf(10, scale=1 / lambda_parametro)

print(f"A probabilidade de esperar menos de 10 minutos pelo pr ximo   nibus
       de {probabilidade_menor_10_min:.2f}.")
```

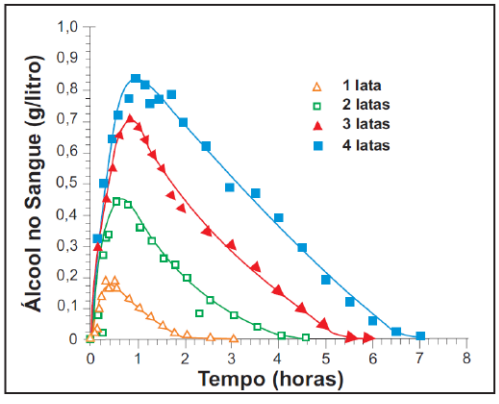
Exemplo no R

Vamos usar a função *pexp* para calcular a probabilidade acumulada.

```r
# Definindo o par metro lambda
lambda_parametro <- 1 / 15

# Calculando a probabilidade de esperar mais de 30 minutos pelo pr ximo   nibus
probabilidade_maior_30_min <- 1 - pexp(30, rate=lambda_parametro)

cat("A probabilidade de esperar mais de 30 minutos pelo pr ximo   nibus
        de", round(probabilidade_maior_30_min, 2), "\n")

# Calculando a probabilidade de esperar menos de 10 minutos pelo pr ximo   nibus
probabilidade_menor_10_min <- pexp(10, rate=lambda_parametro)
cat("A probabilidade de esperar menos de 10 minutos pelo pr ximo   nibus
       de", round(probabilidade_menor_10_min, 2), "\n")
```

Exercícios propostos

28) (ENADE – 2006 – ADMINISTRAÇÃO) A legislação de trânsito brasileira considera que o condutor de um veículo está dirigindo alcoolizado quando o teor alcoólico de seu sangue excede 0,6 gramas de álcool por litro de sangue. O gráfico abaixo mostra o processo de absorção e eliminação do álcool quando um indivíduo bebe, em um curto espaço de tempo, de 1 a 4 latas de cerveja.



(Fonte: National Health Institute, Estados Unidos)

Considere as alternativas a seguir:

I. O álcool é absorvido pelo organismo muito mais rapidamente do que é eliminado.

II. Uma pessoa que vá dirigir imediatamente após a ingestão da bebida pode consumir, no máximo, uma lata de cerveja.

III. Se uma pessoa toma rapidamente três latas de cerveja, o álcool contido na bebida só é completamente eliminado após se passarem cerca de 6 horas da ingestão.

Estão corretas apenas as alternativas:

a) II, apenas.

b) I e II, apenas.

c) I e III, apenas.

d) II e III, apenas.

e) I, II e III.

29) Uma peça cromada resiste a um ensaio de corrosão por três dias, em média, com desvio padrão de cinco horas.

Pede-se calcular:

a) a probabilidade de uma peça resistir menos de 3,5 dias;

b) a probabilidade de uma peça resistir de 60 a 70 horas;

c) sabendo-se que 10 % das peças resistem menos que certo valor, determiná-lo.

30) O tempo de vida (em horas) de um transistor pode ser considerado uma v.a. com distribuição Exponencial com $\beta = 500$.
A vida média do transistor é, portanto

$$E(T) = 500 \text{ horas}$$

E a probabilidade de que ele dure mais do que a média é?

31) Quanto à distribuição exponencial, julgue o item: "Se a média em uma distribuição exponencial é igual a $\frac{1}{\lambda}$, então a sua variância é igual a $\frac{1}{\lambda^2}$."

a) Certo

b) Errado

2.2.3 Modelo Binomial

A distribuição binomial é frequentemente usada para estimar ou determinar a proporção de indivíduos com um atributo específico em uma grande população. Suponha que uma amostra aleatória de n unidades seja selecionada por amostragem com reposição de uma população finita ou por amostragem sem reposição de uma grande população.

1. O experimento consiste em n tentativas independentes.
2. Cada tentativa tem dois resultados possíveis, usualmente chamados de sucesso e fracasso.
3. A probabilidade de sucesso, p, para cada tentativa é constante durante todo o experimento e, consequentemente, a probabilidade 1 – p de fracasso é constante durante todo o experimento.

As probabilidades dadas pela distribuição binomial podem surgir das seguintes maneiras:

1. Amostragem com reposição de uma população finita.
2. Amostragem de uma população infinita (à qual, em geral, se refere como população indefinidamente grande), com ou sem reposição."

Segundo Morettin e Bussab (2017, p.156), as probabilidades serao indicadas por

$$b(k; n, p) \tag{5}$$

e, quando a v.a. X tiver distribuicao binomial com parametros n e p, escreveremos (notação)

$$X \sim b(n, p). \tag{6}$$

Ou seja, para fazer o cálculo das possibilidades de certo evento ter determinado sucesso seguindo a distribuição binomial precisamos utilizar uma equação (função densidade de probabilidade).

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{7}$$

Sendo:

- k = número de sucessos da amostra
- n = total de ensaios
- p = probabilidade de sucesso
- (1 – p) = probabilidade de fracasso

É importante lembrar que estamos falando da combinação de N elementos escolhidos entre K e K.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{8}$$

Isso significa que estamos interessados em calcular todas as possíveis maneiras de selecionar K elementos de um total de N elementos, onde a ordem dos elementos escolhidos não importa. Essa combinação é representada pela fórmula "N escolha K".

A distribuição binomial possui dois parâmetros importantes, o valor esperado (média) e a variância, que são calculados da seguinte forma:

$$E(X) = np \tag{9}$$

$$Var(X) = np(1-p). \tag{10}$$

Exemplo:

Vamos criar um exemplo simples usando o R para modelar a distribuição binomial. Suponha que queremos estimar a proporção de cães em uma amostra de 50 animais selecionados aleatoriamente em um abrigo em Itapajé. Suponha também que a proporção real de cães na população seja de 0.3.
No R Temos quatro funções para lidar com a distribuição binomial, são elas:

- dbinom()
- pbinom() usada para encontrar a probabilidade cumulativa de dados seguindo a distribuição binomial até um determinado valor $P\left(X <= k\right)$
- qbinom() usada para calcular o quantil (valor que divide a distribuição em uma proporção específica)
- rbinom() usada para gerar amostras aleatórias de uma distribuição binomial.

Neste caso utilizaremos a função dbinom usada para encontrar a probabilidade de um determinado valor para dados que seguem a distribuição: $P(X = k)$ binomial.

Exemplo no R

```
n <- 50  # Tamanho da amostra
p <- 0.3  # Probabilidade de sucesso (propor  o real de c es na popula  o)

# Calculando a probabilidade de obter exatamente k c es na amostra, para k variando de 0 a
    50
valor_k <- 0:n
probabilidade <- dbinom(valor_k, n, p)
print(probabilidade)
```

Agora, vamos fazer o mesmo exemplo em Python. Neste exemplo, vamos usar a função *scipy.stats.binom* para trabalhar com a distribuição binomial e calcular algumas probabilidades.

Exemplo no Python

```
import numpy as np
from scipy.stats import binom

n = 50  # Tamanho da amostra
p = 0.3  # Probabilidade de sucesso (propor  o real de c es na popula  o)

# Calculando a probabilidade de obter exatamente k c es na amostra, para k variando de 0 a
    50
```

```
valor_k = np.arange(0, n+1)
probabilidade = binom.pmf(valor_k, n, p)
print(probabilidade)
```

Exercícios propostos

32) O submarino Malik I dispara cinco torpedos em cadência rápida contra o navio Pégaso. Cada torpedo tem probabilidade igual a 75 % de atingir o alvo. Qual a probabilidade de o navio receber pelo menos um torpedo?

33) A probabilidade de recuperação de uma cápsula registradora de dados, montada em um balão meteorológico, é igual a 90 %. Lançados sete balões, qual a probabilidade de serem recuperadas exatamente cinco cápsulas?

34) Acredita-se que 20% dos moradores das proximidades de uma grande indústria siderúrgica têm alergia aos poluentes lançados ao ar. Admitindo que este percentual de alérgicos é real (correto), calcule a probabilidade de que pelo menos 4 moradores tenham alergia entre 13 selecionados ao acaso.

35) Uma distribuição binomial possui média igual a 3 e variância 2. Calcule $P\left(X \geq 2\right)$.

36) Uma remessa de 800 estabilizadores de tensão é recebida pelo controle de qualidade de uma empresa. São inspecionados 20 aparelhos da remessa, que será aceita se ocorrer no máximo um defeituoso. Há 80 defeituosos no lote. Qual a probabilidade de o lote ser aceito?

$$P(k) = \frac{(n-k+1)\,p}{kq}\,P\left(k-1\right) \forall\; k \geq 1 \tag{11}$$

37) Uma pesquisa de opinião pública revelou que $1/4$ da população de determinada cidade assiste regularmente à televisão. Colocando 300 pesquisadores, cada um entrevistando 10 pessoas diariamente, fazer uma estimativa de quantos desses pesquisadores informarão que até 50 % das pessoas entrevistadas são realmente telespectadoras habituais.

2.2.4 Modelo Poisson

De acordo com Oliveira (2017, p.183), a distribuição de Poisson é considerada o caso limite da distribuição binomial, quando o número de provas n tende para o infinito e a probabilidade p do evento em cada prova é vizinha de zero. Em essência, a distribuição de Poisson é a distribuição binomial adequada para eventos independentes e raros, ocorrendo em um período praticamente infinito de intervalos. Cumpre destacar que a unidade de medida é contínua (em geral, tempo ou espaço), mas a variável aleatória (número de ocorrências) é discreta.

Função densidade de probabilidade:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{12}$$

em que:

- X = número de ocorrências ou sucessos
- e = base neperiana
- λ = valor esperado

Valor esperado:

$$\lambda = np \tag{13}$$

Variância:

$$\sigma^2 = npq \tag{14}$$

Exemplo:

Vamos supor que queremos simular o número de casos de dengue para cada mês em um ano fictício. Utilizaremos uma distribuição de Poisson para simular esses dados, pois a distribuição de Poisson é frequentemente usada para modelar a contagem de eventos raros em um intervalo de tempo.

Exemplo no R

```
set.seed(123) # Defina uma semente para tornar a simula  o reproduz vel
lambda <- 15  # Taxa m dia de casos de dengue por m s (valor fict cio)

# Simulando os dados de casos de dengue para 12 meses em Itapaj
numero_de_meses <- 12
casos_dengue <- rpois(numero_de_meses, lambda)

# Visualizando os dados simulados
meses <- month.abb[1:numero_de_meses]  # Abrevia  es dos meses de janeiro a dezembro
dados_simulados <- data.frame(M s = meses, Casos_Dengue = casos_dengue)
print(dados_simulados)
```

No exemplo a seguir, em Python, utilizaremos a função *numpy.random.poisson* para simular os dados de casos de dengue para 12 meses em Itapajé. Em seguida, mapeamos os números dos meses para suas respectivas abreviações e armazenamos os dados simulados em um DataFrame do pandas.

Exemplo no Python

```
import numpy as np
import pandas as pd

np.random.seed(123)  # Defina uma semente para tornar a simula  o reproduz vel
lambd = 15  # Taxa m dia de casos de dengue por m s (valor fict cio)

numero_de_meses = 12
casos_dengue = np.random.poisson(lambd, size=numero_de_meses)

meses = ['Jan', 'Fev', 'Mar', 'Abr', 'Mai', 'Jun', 'Jul', 'Ago', 'Set', 'Out', 'Nov', 'Dez']

dados_simulados = pd.DataFrame({'M s': meses, 'Casos_Dengue': casos_dengue})

print(dados_simulados)
```

Exercícios propostos

38) Suponha que X_t, o número de partículas emitidas em t horas por uma fonte radioativa, tenha uma distribuição de Poisson com parâmetro 20t. Qual será a probabilidade de que exatamente 5 partículas sejam emitidas durante um período de 15 min?

39) Os clientes chegam a uma loja a uma razão de cinco por hora. Admitindo que esse processo possa ser aproximado por um modelo de Poisson, determine a probabilidade de que durante qualquer hora:

a) não chegue nenhum cliente;

b) chegue mais de um cliente.

40) Um distribuidor de gasolina tem capacidade de receber, nas condições atuais, no máximo três caminhões por dia. Se chegarem mais de três caminhões, o excesso deve ser enviado a outro distribuidor, e, nesse caso, há uma perda média de $ 800 por dia em que não se podem aceitar todos os caminhões. Sabendo-se que o número de caminhões que chegam diariamente obedece à distribuição de Poisson de média 2, calcular:

a) a probabilidade de chegarem de três a cinco caminhões no total de dois dias;

b) a probabilidade de, em certo dia, ser necessário enviar caminhões para outro distribuidor;

c) a perda média mensal (30 dias) por causa de caminhões que não puderam ser aceitos.

41) As chegadas de carros a um posto de gasolina para abastecimento entre as 10h00min e as 16h00min do dia ocorrem de acordo com os postulados de Poisson. Se no transcurso de tal período apresentam-se por hora uma média de 30 carros, qual a probabilidade de nenhum se apresentar em certo intervalo de cinco minutos?

2.2.5 Teorema do Limite Central

(i) Teorema Central de Limite de DeMoivre (1733) – Laplace (1812)

Se $X \sim B(n,p)$ e $Z_n = \frac{X-np}{\sqrt{npq}}$, então $Z_n \xrightarrow{D} N(0,1)$. Isto é,

$\lim_{n\to\infty} P(Z_n \leq z) = \lim_{n\to\infty} F_n(z) = F_{N(0,1)}(z)$,

o que significa que Z_n converge em distribuição (D) para uma N(0,1).

Este teorema mostra que uma $B(n,p)$ pode ser aproximada por uma $N(0,1)$. Ross (1993) sugere que a aproximação seja usada quando $npq10$.

(ii) Teorema Central do Limite – TCL

Sejam $X_1, X_2, ..., X_n$, independentes tais que, para todo $i = 1, 2, ...$, $E(X_i) = \mu_i$ e $V(X_i) = \sigma_2^i$. Seja $S_n = X_i + ... + X_n$. Então, se a condição $Lindeberg^2$ for satisfeita

$$Z_n = \frac{S_n - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \sim N(0,1).$$

Isto é,

$\lim_{n\to\infty} P(Z_n \leq z) = \lim_{n\to\infty} F_n(z) = F_{N(0,1)}(z)$,

ou Z_n converge em distribuição para uma $N(0,1)$,

$Z_n \xrightarrow{D} N(0,1)$.

O teorema de DeMoivre-Laplace mostra que probabilidades envolvendo binomiais podem ser calculadas por meio de aproximação pela $N(0,1)$. Note que a convergência deste último resultado também é convergência em distribuição. O TCL fornece um método efetivo para se calcular probabilidades quando se tem somas de variáveis aleatórias independentes. Isto significa que se um fenômeno do mundo real puder ser modelado por uma soma (S_n) de n fatores independentes, mesmo não sendo possível encontrar uma fórmula para a distribuição de S_n, calcula-se qualquer probabilidade envolvendo S_n pela aproximação com a $N(0,1)$.

Exercícios propostos

42) Uma instituição de caridade deseja realizar uma obra que custa R$3500,00 em sua sede. Entre os contribuintes habituais dessa instituição, cada um pode contribuir com algo em torno de R$120,00 ± um desvio padrão de R$50,00. Se 30 dessas pessoas se quotizarem para levantar fundos com essa finalidade, qual a probabilidade de que eles consigam o montante necessário?

43) Suponha que $X_i, i = 1, 2, ...$, 50 sejam variáveis aleatórias independentes, cada uma com distribuição de Poisson de parâmetro $\beta = 0,03$. Faça $S = X_1 + ... + X_{50}$.

a) Empregando o Teorema Central de Limite, calcule $P(S \geq 3)$.

b) Compare a resposta do item anterior com o valor exato dessa probabilidade.

44) A distribuição dos comprimentos dos elos da corrente de uma bicicleta tem distribuição Normal com média 2 cm e variância 0,01 cm2. Para que uma corrente se ajuste à bicicleta, deve ter comprimento total entre 58 e 61 cm. Qual a probabilidade de que uma corrente com 30 elos não se ajuste à bicicleta? E com 29 elos?

3 Estimadores eficientes e Estatísticas Suficientes

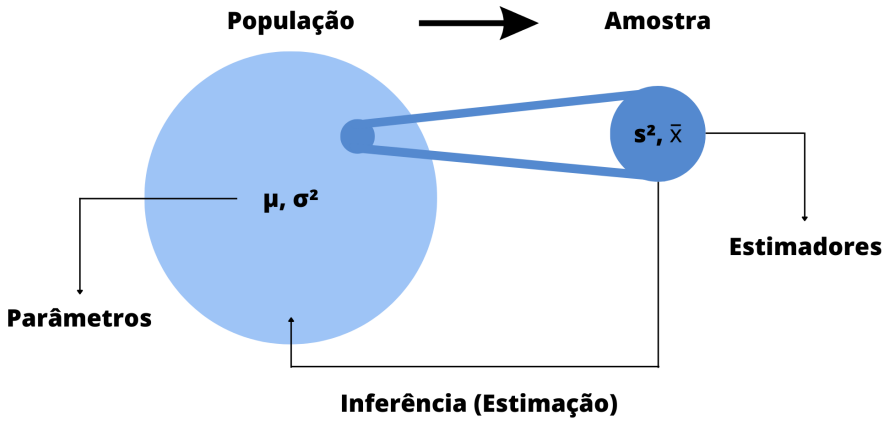
Vamos agora abordar o estudo das Estatísticas. Neste tópico, entenderemos a diferença crucial entre parâmetro, estimador e estimativa.

Parâmetros são características ou medidas associadas a dados de uma população ou de uma amostra. São números que descrevem certas propriedades do fenômeno em estudo, como a média (μ), a variância (σ^2) ou proporções (p) da população. Os parâmetros geralmente são valores desconhecidos, sempre são constantes, e são representados, genericamente, pela letra grega teta (θ).

Já os estimadores estão diretamente relacionados aos parâmetros amostrais. São fórmulas ou procedimentos matemáticos que utilizamos para calcular ou estimar um parâmetro populacional com base nas informações fornecidas por uma amostra. Os estimadores são como ferramentas que nos permitem aproximar ou inferir os valores dos parâmetros desconhecidos a partir dos dados amostrais. Os estimadores são representados, genericamente, pela letra teta com um acento circunflexo ($\hat{\theta}$). Dentre os exemplos de estimadores podemos citar a média da amostra ($\bar{X}$) e a variância da amostra (S^2).

Por fim, as estimativas são os valores numéricos que resultam da aplicação de um estimador específico em uma determinada amostra. Em outras palavras, são os valores assumidos pelo estimador ao utilizar os dados amostrais. As estimativas fornecem uma ideia sobre o valor provável do parâmetro na população, com base nas informações limitadas que temos disponíveis na amostra.

Figura 9: Estimadores eficientes e Estatísticas Suficientes



Exemplo: Estimativa do gasto médio em supermercados.

Suponha que desejamos estimar o gasto médio dos clientes em supermercados de Itapajé, considerando apenas os supermercados localizados na região central da cidade, onde a maioria dos estabelecimentos está concentrada.

- *População*: Clientes que realizaram compras em supermercados na região central do município.
- *Parâmetro*: O gasto médio dos clientes em supermercados na região central, representado por μ. Esse valor é desconhecido, pois não podemos verificar todas as compras realizadas pelos clientes na região.
- *Amostra*: Selecionamos aleatoriamente 50 clientes que realizaram compras nos supermercados.
- *Estimador*: A média amostral ($\overline{x}$) é o estimador utilizado para estimar o gasto médio na população. É calculado pela fórmula:

$$\overline{x} = \frac{\text{Soma dos gastos dos 50 clientes na amostra}}{\text{Tamanho da amostra}} \tag{15}$$

- *Estimativa*: Suponha que após coletar os dados dos 50 clientes, calculamos que a soma total dos gastos foi de R5.000.Então, a média amostral é:

$$\overline{x} = \frac{5000}{50} = 100 \tag{16}$$

  Essa é a estimativa do gasto médio dos clientes em supermercados na região central do município, com base na amostra de 50 clientes.

Nesse exemplo, o parâmetro é o gasto médio dos clientes em todos os supermercados da região central, o estimador é a média amostral que usamos para calcular uma estimativa do gasto médio, e a estimativa é o valor numérico obtido após a aplicação do estimador na amostra específica. Essa estimativa nos fornece uma ideia sobre o gasto médio provável dos clientes em supermercados na região central do município, com base nas informações limitadas que temos disponíveis na amostra de 50 clientes.

Exemplo no Python

```python
import numpy as np

# Dados da amostra
gastos = [120, 90, 150, 80, 200, 130, 110, 100, 140, 160,
          180, 95, 105, 170, 125, 115, 90, 120, 110, 105,
          135, 125, 155, 165, 200, 90, 150, 175, 185, 100,
          190, 115, 130, 110, 160, 140, 105, 175, 120, 200,
          130, 110, 140, 170, 150, 90, 115, 95, 125, 135]

media_amostral = np.mean(gastos) # Estimador da m dia amostral

print(media_amostral)
```

Exemplo no R

```r
# Dados da amostra
gastos <- c(120, 90, 150, 80, 200, 130, 110, 100, 140, 160,
            180, 95, 105, 170, 125, 115, 90, 120, 110, 105,
            135, 125, 155, 165, 200, 90, 150, 175, 185, 100,
            190, 115, 130, 110, 160, 140, 105, 175, 120, 200,
            130, 110, 140, 170, 150, 90, 115, 95, 125, 135)

media_amostral <- mean(gastos)

print(media_amostral)
```

Exercícios propostos

45) Seja $[X_1, X_2, X_3]$ uma a.a. de uma população de Bernoulli com parâmetro θ. A estatística $T = \sum_{X_i}$ é suficiente para (estimar) θ?

46) Seja $T_0 = X_1 + X_3$. A estatística T é suficiente?

3.1 Propriedades dos estimadores

- **Não tendenciosidade (Não viesado/não viciado)**

O estimador $\hat{\theta}$ é chamado não viciado, não viesado, não tendencioso ou imparcial se seu valor esperado ou médio for igual ao verdadeiro valor do parâmetro θ.

$$E(\hat{\theta}) = \theta \tag{17}$$

O viés ou tendência de um estimador $\hat{\theta}$ para um parâmetro θ é igual a $E(\hat{\theta}) = \theta$. Logo, um estimador $\hat{\theta}$ é não viesado para θ, se o seu viés for igual a zero.

Exemplo:
A média amostral X é um estimador não viesado para média populacional μ, pois

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

- **Consistência**

Um estimador é consistente se à medida que o tamanho da amostra aumenta o valor do estimador se aproxima do parâmetro, ou seja, em outras palavras, consistência é uma propriedade por meio da qual a acurácia de uma estimativa aumenta quando o tamanho da amostra aumenta, seu valor esperado converge para o parâmetro de interesse e sua variância converge para zero:

i) $\lim_{n\to\infty} E\left(\hat{\theta}\right) = \theta$;

ii) $\lim_{n\to\infty} V\left(\hat{\theta}\right) = 0$.

Observe que para ser consistente o estimador depende de n, o tamanho da amostra, e somente será não viciado se n for grande. Na definição de estimador não viciado, a propriedade deve valer para qualquer n.

$$\lim_{n\to+\infty} \sigma^2(\hat{\theta}) = 0 \tag{18}$$

Exemplo:
Vamos ver se $\bar{x}$ é um estimador consistente para μ:

$$\text{sendo } \frac{\sum_i^n x_i}{n} \text{ e, portanto, } \bar{x} = \frac{x_1+x_2+x_3...x_n}{n}$$

Assim:

$$\sigma_2(x) = \sigma_2\left[\frac{x_1 + x_2 + x_3... + x_n}{n}\right] \tag{19}$$

$$\sigma_2(x) = \frac{1}{n_2}\left[\sigma_2 x_1 + \sigma_2 x_2 \sigma_2 + x_3...\sigma_2 + x_n\right] \tag{20}$$

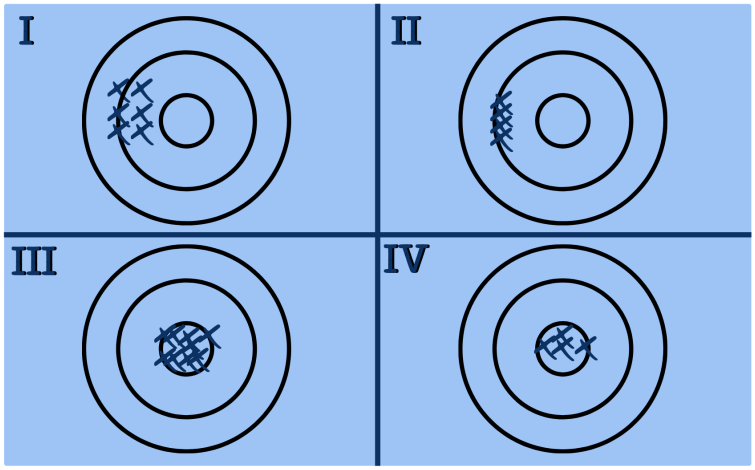
$$\sigma_2 = \frac{1}{n}\sigma_2(x) \tag{21}$$

Dessa forma:

$$\lim_{n\to+\infty} \sigma^2(\hat{x}) = 0 \tag{22}$$

Veja a imagem e a tabela que explicita essa relação entre dois estimadores:

Figura 10: Propriedades do estimadores



		Menos eficiente	Mais eficiente
Estimadores	Enviesados	I	II
	Justos	III	IV

Esta tabela classifica estimadores em quatro categorias com base em duas características principais: viés e eficiência. Os estimadores podem ser classificados como enviesados ou justos e menos eficientes ou mais eficientes. A combinação desses critérios resulta em quatro categorias: menos eficiente e enviesado, mais eficiente e enviesado, menos eficiente e justo, mais eficiente e justo.

- **Eficiência**

Um estimador é eficiente no sentido absoluto quando a variância do estimador é mínima.

Exemplos:

1. Verifique o melhor estimador de $\mu(x)$:

$$x_i = x_1, x_2, x_3, x_4, x_5 \tag{23}$$

Justeza:

$$\theta_1 = x_1 \;\; II.\theta_2 = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) \;\; III.\theta_3 = \frac{1}{2}(x_1 + 2x_5) \tag{24}$$

Solução:

I. $\mu(\theta_1) = \mu(x_1) = \mu(x)$

II. $\mu(\theta_2) = \mu(\frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) = \frac{1}{5}\mu(x_1 + x_2 + x_3 + x_4 + x_5) = \frac{1}{5}5\mu(x)$

III. $\mu(\theta_3) = \mu(\frac{1}{2}(x_1 + 2x_5)) = \frac{1}{2}\mu(x_1 + 2x_5 = \frac{1}{2}\left[\mu(x) + 2\mu(x)\right] = ... = \frac{2}{3}\mu(x)$

Dessa forma apenas o III não satisfaz a exigência de justeza ou imparcialidade.

2. Vamos considerar um exemplo prático para ilustrar a propriedade de justeza (imparcialidade) de um estimador. Imagine que estamos conduzindo um estudo sobre a altura média das crianças em uma escola primária. Queremos estimar a altura média da população de alunos dessa escola usando uma amostra de cinco alunos selecionados aleatoriamente.

Parâmetro: A altura média de todos os alunos da escola (μ).
Estimador: Três possíveis estimadores da altura média são:

1. $\hat{\theta}_1$: Altura do primeiro aluno selecionado (x_1).
2. $\hat{\theta}_2$: Média das alturas dos cinco alunos selecionados ($\frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5)$).
3. $\hat{\theta}_3$: Média ponderada das alturas do primeiro e último alunos ($\frac{1}{2}(x_1 + 2x_5)$).

Vamos verificar se esses estimadores são imparciais (justos):

Justeza (Imparcialidade): Um estimador $\hat{\theta}$ é imparcial se o valor esperado de $\hat{\theta}$ for igual ao verdadeiro valor do parâmetro ($E(\hat{\theta}) = \theta$).

Cálculos:

$$E(\hat{\theta}_1) = E(x_1) = \mu \text{ (Imparcial)}$$
$$E(\hat{\theta}_2) = \frac{1}{5}E(x_1 + x_2 + x_3 + x_4 + x_5) = \frac{1}{5} \cdot 5\mu = \mu \text{ (Imparcial)}$$
$$E(\hat{\theta}_3) = \frac{1}{2}E(x_1 + 2x_5) = \frac{1}{2}(\mu + 2\mu) = \frac{3}{2}\mu \text{ (Não é imparcial)}$$

Concluímos que apenas os estimadores $\hat{\theta}_1$ e $\hat{\theta}_2$ são imparciais, o que significa que eles são justos em relação à estimativa da altura média da população. O estimador $\hat{\theta}_3$ não é imparcial e, portanto, não satisfaz a propriedade de justeza.

Isso demonstra a importância da propriedade de justeza (imparcialidade) em estimadores, pois garante que a estimativa média seja correta em relação ao parâmetro populacional que estamos tentando estimar.

Solução:

$$\text{I. } \sigma_2(\hat{\theta}_1) = \sigma_2(x_1) = \sigma_2(x)$$
$$\text{II. } \sigma_2(\hat{\theta}_2) = (\frac{1}{5}x_1 + x_2 + x_3 + x_4 + x_5) = \frac{1}{5}\sigma_2(x)$$

O estimador, dentre os justos, com menor desvio padrão foi $\hat{\theta}_2$ sendo ele o melhor estimador de $\mu(x)$, dentre os fornecidos.

Exemplo no Python

```python
import numpy as np

# Alturas dos cinco alunos selecionados
alturas = np.array([130, 140, 125, 135, 150])

# Estimadores
theta_1 = alturas[0]  # Altura do primeiro aluno
theta_2 = np.mean(alturas)  # Media das alturas
theta_3 = (alturas[0] + 2 * alturas[-1]) / 2  # Media ponderada das alturas do primeiro e ultimo aluno

# Funcao para calcular o valor esperado
def valor_esperado(estimador):
    return np.mean(estimador)

# Calcular os valores esperados
esperado_theta_1 = valor_esperado(theta_1)
esperado_theta_2 = valor_esperado(theta_2)
esperado_theta_3 = valor_esperado(theta_3)

# Visualizar os resultados
print("Valor esperado de theta_1:", esperado_theta_1)
print("Valor esperado de theta_2:", esperado_theta_2)
print("Valor esperado de theta_3:", esperado_theta_3)
```

Exemplo no R

```r
# Alturas dos cinco alunos selecionados
alturas <- c(130, 140, 125, 135, 150)

# Estimadores
theta_1 <- alturas[1]  # Altura do primeiro aluno
theta_2 <- mean(alturas)  # Media das alturas
theta_3 <- (alturas[1] + 2 * alturas[5]) / 2  # Media ponderada das alturas do 1   e  ltimo aluno

# Funcao para calcular o valor esperado
valor_esperado <- function(estimador) {
  return(mean(estimador))
}

# Calcular os valores esperados
esperado_theta_1 <- valor_esperado(theta_1)
esperado_theta_2 <- valor_esperado(theta_2)
esperado_theta_3 <- valor_esperado(theta_3)

```

```
# Visualizar os resultados
cat("Valor esperado de theta_1:", esperado_theta_1, "\n")
cat("Valor esperado de theta_2:", esperado_theta_2, "\n")
cat("Valor esperado de theta_3:", esperado_theta_3, "\n")
```

Exercícios propostos

47) Se $\hat{\theta}$ for um estimador de um parâmetro populacional θ tal que $E[\hat{\theta}] = \theta$, então se diz que $\hat{\theta}$ é um estimador

a) consistente.

b) suficiente.

c) não viciado.

d) intervalar.

e) de máxima verossimilhança.

48) Foram sorteadas 15 famílias com filhos num certo bairro e observado o número de crianças de cada família, matrícula das na escola. Os dados foram 1, 1, 2, 0, 2, 0, 2, 3, 4, 1, 1, 2, 0, 0, e 2.

Obtenha as estimativas correspondentes aos seguintes estimadores da média de crianças na escola nesse bairro,

$$\widehat{\mu}_1 = \frac{(\text{mínimo} + \text{máximo})}{2} \tag{25}$$

$$\widehat{\mu}_2 = \frac{(X_1 + X_2)}{2} \tag{26}$$

$$\widehat{\mu}_3 = \overline{X} \tag{27}$$

Qual deles é o melhor estimador da média e por que?

49) Verifique se o estimador para a variância a seguir é não-viciado e consistente.

$$S_2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \tag{28}$$

50) Sobre as principais propriedades dos estimadores pontuais, para pequenas e grandes amostras, é correto afirmar que:

a) se $\widehat{\theta}$ é tendencioso, só poderá ser mais eficiente do que $\check{\theta}$ caso $EQM\left(\widehat{\theta}\right) < Var\left(\check{\theta}\right)$;

b) uma condição necessária para que um estimador $\widehat{\theta}$ de θ seja assintoticamente eficiente é que ele seja não tendencioso, também em termos assintóticos;

c) se $\lim_{n\to\infty} EQM\left(\widehat{\theta}\right) = +\infty$, então $\lim_{n\to\infty} P\left(\left|\widehat{\theta} - \theta\right| < \varepsilon\right) \neq 0$;

d) se $f_{X_1,\ldots,X_n} = \theta.\lambda\left(X_1, X_2, ..., X_n\right)$ é a densidade conjunta da amostra e $f_{\widehat{\theta}} = 0.k\left(X_1, X_2, ..., X_n\right) - \theta$ é a função densidade do estimador $\widehat{\theta}$, então $\widehat{\theta}$ é um estimador suficiente de θ;

e) um estimador que seja assintoticamente tendencioso não poderá ser consistente.

51) Faça $X_1, X_2, ..., X_n$ denotar uma amostra aleatória, proveniente de uma população tendo média μ e variância σ^2. Considere os seguintes estimadores de μ:

$$\hat{\theta}_1 = \frac{X_1 + X_2 + ... + X_7}{7} \quad \text{e} \quad \hat{\theta}_2 = \frac{2X_1 - X_6 + X_4}{2} \tag{29}$$

a) Os dois estimadores são não tendenciosos?

b) Qual é o melhor estimador? Em que sentido ele é melhor?

3.2 Como usar a tabela normal

A tabela normal, também conhecida como tabela Z, é uma ferramenta importante em estatística para calcular probabilidades associadas à distribuição normal padrão, que é uma distribuição contínua com média (μ) igual a 0 e desvio padrão (σ) igual a 1. A distribuição normal padrão é representada como $N(0, 1)$.

Para usar a tabela normal, siga os seguintes passos:

1. Padronize a variável aleatória: Antes de usar a tabela normal, é necessário converter a variável aleatória em uma variável com média 0 e desvio padrão 1, ou seja, padronizá-la. Isso é feito subtraindo a média da variável original e dividindo pelo desvio padrão. Onde:

$$Z = \frac{X - \mu}{\sigma} \tag{30}$$

   Onde:

   - X é o valor da variável;
   - μ é a média da distribuição;
   - σ é o desvio padrão da distribuição;
   - z é o valor padronizado.

2. Encontre o valor Z: Após a padronização, você obtém um valor z, que representa o número de desvios padrão pelo qual a observação está afastada da média da distribuição normal padrão ($\mu = 0$ e $\sigma = 1$).

3. Consulte a tabela: Com o valor z em mãos, consulte a tabela normal padrão, que fornece valores da função de distribuição acumulada (CDF) da distribuição normal padrão para diferentes valores de z. A tabela normal fornece as probabilidades $P(Z \leq z)$.

4. Interprete o resultado: Ao encontrar o valor Z na tabela normal, você terá a probabilidade associada à sua observação. Isso pode ser usado para responder a perguntas sobre a probabilidade de uma variável aleatória normal cair em um intervalo específico ou ser maior/menor que um determinado valor.

Lembre-se de que a tabela normal fornece valores acumulados da distribuição normal padrão. Portanto, para calcular probabilidades específicas, pode ser necessário fazer cálculos adicionais, como subtrair probabilidades acumuladas ou usar a simetria da distribuição normal.

Usar a tabela normal é uma habilidade fundamental em estatística e é amplamente aplicada em análises estatísticas e testes de hipóteses!

Exercícios propostos

52) Numa escola os registos indicam que os exames finais têm uma classificação com média de 510 e um desvio padrão de 90. Sabendo que 100 estudantes fazem o teste, qual a probabilidade da sua classificação média ser de:

a) Mais de 530?

b) Menos de 500?

c) Entre 495 e 515?

53) Um catálogo de um fabricante indica para um determinado produto uma vida média de 1.200 horas. Assuma o desvio padrão igual à 120 horas. Um cliente decide selecionar aleatoriamente 35 itens do referido produto e rejeitar a amostra $\overline{X} < 1.160$ horas. Se a indicação do fabricante for verdadeira, qual a probabilidade de rejeitar a amostra?

54) Uma fábrica de sapatos tem uma máquina que corta peças de borracha comprimida para serem usadas em solas. A espessura dessas solas é uma variável aleatória normalmente distribuída com desvio padrão igual a 2 mm, com valor médio μ. Para se tentar corrigir estas medidas, reajustando a máquina, é conveniente verificar a qualidade do produto, medindo espessura das solas de uma amostra aleatória retirada periodicamente da máquina. De uma amostra de 5 elementos foram registadas as espessuras respectivas e calculada a média aritmética.
Se $\overline{X} < 24,8$ ou $\overline{X} > 25,2$ diz-se que a máquina não está controlada, pelo que é parada e reajustada. Com a média $\mu = 25$ mm, qual a probabilidade de a amostra indicar que a máquina não está controlada?

55) Suponha que as medidas da corrente elétrica em pedaço de fio sigam a distribuição Normal, com uma média de 10 miliamperes e uma variância de 4 miliamperes.

a) Qual a probabilidade de a medida exceder 13 miliamperes?

b) Qual a probabilidade de a medida da corrente estar entre 9 e 11 miliamperes?

c) Determine o valor para o qual a probabilidade de uma medida da corrente estar abaixo desse valor seja 0,98.

56) Uma fábrica de carros sabe que os motores de sua fabricação têm duração normal com média 150000 km e desvio-padrão de 5000 km. Qual a probabilidade de que um carro, escolhido ao acaso, dos fabricados por essa firma, tenha um motor que dure:

a) Menos de 170000 km?

b) Entre 140000 km e 165000 km?

c) Se a fábrica substitui o motor que apresenta duração inferior à garantia, qual deve ser esta garantia para que a porcentagem de motores substituídos seja inferior a 0,2%?

57) Um criador possui 5.000 cabeças de vaca leiteira. Sabendo-se que cada vaca produz em média 3 litros por dia, obedecendo a uma distribuição normal, com desvio padrão de 0,5 litros, calcular a probabilidade de produzir, diariamente:

a) Mais de 15.110 litros;

b) Entre 14.910 e 14.960 litros.

58) $X : B(n; p)$, onde $n = 100$ e $p = \frac{1}{2}$. Calcular usando a aproximação pela normal:

a) $P(X \geq 25)$;

b) $P(X \leq 70)$;

c) $P(X > 57)$;

d) $P(X = 52)$;

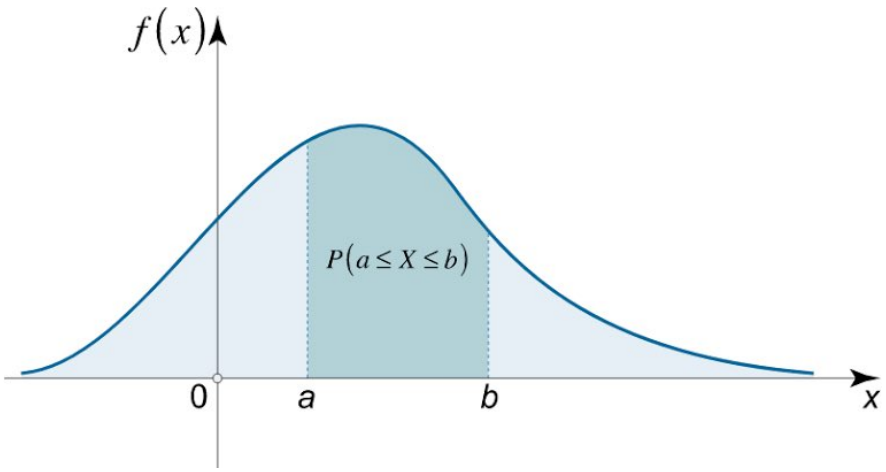
e) $P(25 < X < 57)$.

3.3 Distribuições amostrais

Para Gupta e Irwin Guttman (2016, p.195), "As distribuições de probabilidade das várias estatísticas são chamadas distribuições amostrais.".

Vamos nos aprofundar agora em como as estatísticas se distribuem, ou seja, qual é a cara da função densidade de probabilidade dos principais estimadores, assim como suas respectivas médias e variâncias.

Figura 11: Ilustração de probabilidade



3.3.1 Distribuição Amostral da Média para populações normais ($\overline{X}$)

Seja $\overline{X}$ uma variável aleatória $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$.

Propriedades:

$$Média = E(\overline{x}) = \mu_{\overline{x}} = \mu \tag{31}$$

$$Variância = VAR(\overline{x}) = \sigma\frac{2}{x} = \frac{\sigma^2}{x} \tag{32}$$

$$Desvio\ padrão = DP(\overline{x}) = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} \tag{33}$$

Demonstração:

$$E(\bar{x}) = \left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \frac{1}{n}E(X_1 + X_2 + ... + X_n) \tag{34}$$

$$= \frac{1}{n}\left[E(X_1 + E(X_2 + ... + E(X_n)))\right] \tag{35}$$

$$= \frac{1}{n}(\mu + \mu + ... + \mu) = \frac{1}{n}n\mu = \mu \tag{36}$$

Sabendo que, sendo k constante e x e y variáveis temos que:

$$Var(k * x) = k^2 * var(x) \tag{37}$$

$$Var(y + x) = Var(x) + Var(y) \tag{38}$$

Então:

$$Var(\bar{X}) = \left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = \left(\left(\frac{1}{n}\right)^2 X_1 + X_2 + ... + X_n\right) = \frac{1}{n_2} Var\left(X_1 + X_2 + ... + X_n\right) \quad (39)$$

$$= \frac{1}{n_2}\left[Var(X_1) + Var(X_2) + ... + Var(X_n + \sum_{i \neq j} Cov\left(X_i, X_j\right))\right] \quad (40)$$

$$= \frac{1}{n_2}\left(\sigma_2 + \sigma_2 + ... + 0\right) = \frac{1}{n_2} n\sigma_2 = \frac{\sigma_2}{n} \quad (41)$$

Exercícios propostos

59) Uma população tem uma distribuição com média de 325 unidades e variância 144 unidades. Considere uma amostra de tamanho n=36. Determine:

a) A média amostral;

b) O desvio padrão amostral;

c) $\left[320 < \overline{X} < 322\right]$;

d) $\left[321 < \overline{X} < 327\right]$;

e) $\left[\overline{X} < 323\right]$;

f) $\left[\overline{X} > 328\right]$.

60) Um catálogo de um fabricante indica para um determinado produto uma vida média de 1.200 horas. Assuma o desvio padrão igual à 120 horas. Um cliente decide selecionar aleatoriamente 35 itens do referido produto e rejeitar a amostra se $\overline{X} < 1.160$ horas. Se a indicação do fabricante for verdadeira, qual a probabilidade de rejeitar a amostra?

61) Uma fábrica de sapatos tem uma máquina que corta peças de borracha comprimida para serem usadas em solas. A espessura dessas solas é uma variável aleatória normalmente distribuída com desvio padrão igual a 2 mm, com valor médio $\bar{X}$. Para se tentar corrigir estas medidas, reajustando a máquina, é conveniente verificar a qualidade do produto, medindo espessura das solas de uma amostra aleatória retirada periodicamente da máquina. De uma amostra de 5 elementos foram registadas as espessuras respectivas e calculada a média aritmética. Se $\overline{X} < 24$ e $\overline{X} > 25,2$ diz-se que a máquina não está controlada, pelo que é parada e reajustada. Com a média $\mu = 25$ mm, qual a probabilidade de a amostra indicar que a máquina não está controlada?
62) A idade dos assinantes de um jornal é uma variável aleatória com densidade Normal com média 35,5 anos e desvio padrão 4,8 anos. Calcule as seguintes probabilidades:

a) De que um assinante escolhido aleatoriamente tenha entre 30 e 40 anos;

b) De que um assinante escolhido aleatoriamente tenha mais que 40 anos;

c) Suponha que você toma uma amostra de 16 assinantes. Qual a probabilidade de que a idade média na amostra exceda 40 anos?

63) Em um teste de fadiga à tração de certo material, o número esperado de ciclos para a o início de uma trinca é $\mu = 28.000$ e o desvio padrão do número de ciclos é $\sigma = 5000$. Assuma que $X_1, X_2, ..., X_{25}$ são itens de uma amostra aleatória, em que cada X_i é o número de ciclos para um corpo de prova diferente.

a) Qual é o valor esperado do número de ciclos médio da amostra?

b) E o desvio padrão do número de ciclos médio?

3.3.2 Distribuição Amostral da Proporção

A proporção possui distribuição binomial. Sua média e variância, calculadas a partir de sua definição, são dadas por:

- $$\mu(\hat{p}) = p \tag{42}$$

- $$\sigma_2(\hat{p}) = \frac{p(1-p)}{n} \tag{43}$$

Quando $np > 5$ e $n(1-p) > 5$, a Distribuição Binomial se aproxima de uma Distribuição Normal. Validada essa demonstração a partir do *Teorema do Limite Central*.

A distribuição binomial é definida como "número de sucessos em n repetições independentes de um experimento de Bernoulli com parâmetro p". No entanto, em algumas situações, quando o número de tentativas n se torna grande, pelo teorema citado acima e usando o fato de que se $X \sim Bern(p)$, logo $E(X) = p$ e $Var(X) = p(1-p)$, podemos dizer que a distribuição binomial com parâmetros n e p se aproxima de uma distribuição normal com média np e variância $np(1-p)$ quando $n \to \infty$.

Alguns cuidados devem ser tomados na aproximação da binomial pela normal. Um fato importante a observar é que a distribuição binomial é discreta, enquanto a variável normal é contínua. Veja a figura a seguir.

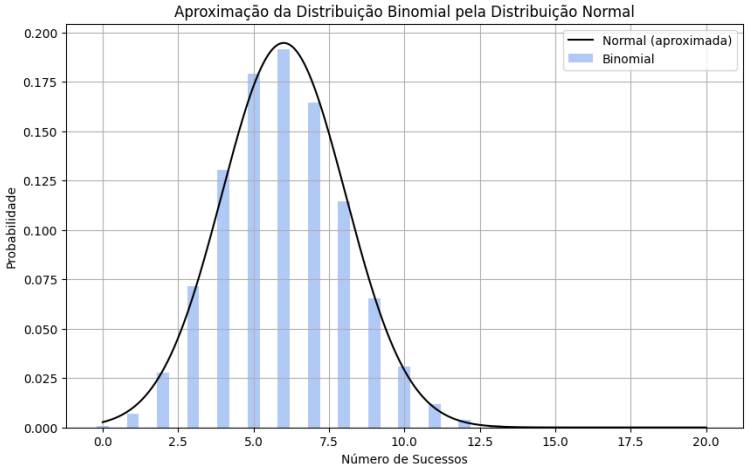


Figura 12: Aproximação da distribuição binomial pela normal

Código do gráfico passo a passo (Python)

```
import numpy as np # biblioteca para lidar com fuco es matematicas
import matplotlib.pyplot as plt # para criar graficos, visualizacao de dados
from scipy.stats import binom, norm # estatisticas e distribuicoes de probabilidade

# Parametros da distribui  o binomial:
n = 20  # Numero de tentativas
p = 0.3  # Probabilidade de sucesso

# Criando um array de valores de x (numero de sucessos)
x = np.arange(0, n + 1)

# Calculando as probabilidades da distribuicao binomial
binom_probs = binom.pmf(x, n, p)

# Parametros da distribuicao normal aproximada:
mu = n * p # M dia
sigma = np.sqrt(n * p * (1 - p)) # Desvio padr o

# Calcula as probabilidades da distribuicao normal:
```

```
norm_probs = norm.pdf(x, mu, sigma)

# Criando o grafico:
plt.figure(figsize=(10, 6))

# Histograma da distribuicao binomial:
plt.bar(x, binom_probs, label='Binomial', color='CornflowerBlue', alpha=0.5, width=0.4)

# Curva da distribuicao normal (aproximada):
x_normal = np.linspace(0, n, 1000)
y_normal = norm.pdf(x_normal, mu, sigma)
plt.plot(x_normal, y_normal, color='black', label='Normal (aproximada)')

plt.title('Aproximacao da Distribuicao Binomial pela Distribuicao Normal')
plt.xlabel('Numero de Sucessos')
plt.ylabel('Probabilidade')
plt.legend()
plt.grid(True)
plt.show()
```

Exercícios propostos

64) Suponha que a proporção de usuários frequentes de drogas em uma população seja de 0,50, enquanto que a proporção em outra população seja de 0,33. Qual é a probabilidade de que amostras de tamanho 100 das duas populações tenham um valor $\widehat{p}_1 - \widehat{p}_2$ maior que 0,30?

65) Em uma certa população de adolescentes, sabe-se que 10% dos rapazes são obesos. Se a mesma proporção de garotas da população for obesa, qual a probabilidade de que uma amostra aleatória de 250 rapazes e 200 garotas tenha $\widehat{p}_1 - \widehat{p}_2 \geq 0,06$?

66) As proporções dos carros produzidos pelas fábricas A e B que apresentam defeito durante o primeiro ano de uso são, respectivamente, iguais a 8% e a 5%. Se forem selecionadas amostras aleatórias de carros produzidos pelas duas fábricas com tamanhos iguais a 230 e 250 respectivamente, qual a probabilidade de que a diferença entre as proporções amostrais de carros com defeito no primeiro ano de uso seja maior do que 1%?

67) Use a aproximação normal para calcular as probabilidades pedidas, tendo o cuidado de verificar se as condições para essa aproximação são realmente satisfeitas.

a) $Pr\,(X \leq 25)$ se $X \sim bin\,(50; 0,7)$

b) $Pr(42 < X \leq 56)$ se $X \sim bin\,(100; 0,5)$

c) $Pr\,(X > 60)$ se $X \sim bin\,(100; 0,5)$

d) $Pr\,(X = 5)$ se $X \sim bin\,(20; 0,4)$

e) $Pr\,(X \geq 12)$ se $X \sim bin\,(30; 0,3)$

f) $Pr(9 < X < 11)$ se $X \sim bin(80; 0,1)$

g) $Pr(12 \leq X \leq 16)$ se $X \sim bin(30; 0,2)$

h) $Pr(X > 18)$ se $X \sim bin(50; 0,3)$

i) $Pr(X = 6)$ se $X \sim bin(28; 0,2)$

j) $Pr(30 \leq X < 48)$ se $X \sim bin(95; 0,4)$

68) Com base em dados históricos, uma companhia aérea estima em 15% a taxa de desistência entre seus clientes, isto é, 15% dos passageiros com reserva não aparecem na hora do vôo. Para otimizar a ocupação de suas aeronaves, essa companhia decide aceitar 400 reservas para os vôos em aeronaves que comportam apenas 350 passageiros. Calcule a probabilidade de que essa companhia não tenha assentos suficientes em

um desses vôos. Essa probabilidade é alta o suficiente para a companhia rever sua política de reserva?

69) Supondo que meninos e meninas sejam igualmente prováveis, qual é a probabilidade de nascerem 36 meninas em 64 partos? Em geral, um resultado é considerado não-usual se a sua probabilidade de ocorrência é pequena, digamos, menor que 0,05 é não-usual nascerem 36 meninas em 64 partos?

3.3.3 Distribuição Amostral da Variância

Vamos reescrever S_2 para enxergar como a sua distribuição se comporta:

$$S_2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1} \rightarrow S_2 = \frac{\sigma_2}{\sigma_2} \frac{(x_i - \bar{x})}{n-1} \tag{44}$$

$$S_2 = \frac{\sigma_2}{n-1} \sum_i^n \left[\frac{x_i - \bar{x}}{\sigma} \right]^2 \rightarrow S_2 = \frac{\sigma_2}{n-1} \sum_i^n z_i^2 \tag{45}$$

Exercícios propostos

70) Definimos a variável $e = \overline{X} - \mu$ como sendo o erro amostral da média, onde $\overline{X}$ é a média de uma aas de tamanho n de uma população com média μ e desvio padrão σ.

a) Determine $E(e)$ e $Var(e)$.

b) Se a população é normal com $\sigma = 20$, que proporção das amostras de tamanho 100 terá erro amostral absoluto maior do que 2 unidades?

c) Neste caso, qual deve ser o valor de δ para que $Pr(|e| > \delta) = 0,01$?

d) Qual deve ser o tamanho da amostra para que 95% dos erros amostrais absolutos sejam inferiores a 1 unidade?

71) A máquina de empacotar um determinado produto o faz segundo o faz segundo uma distribução normal, com média μ e desvio padrão 10g.

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

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

72) No exemplo anterior, e após a máquina estar regulada, programou-se uma carta conrole de qualidade. De hora em hora, será retirada uma amostra de quatro pacotes e esses pesados. Se a média da amostra for inferior a 495g ou superior a 520g, encerra-se a produção para reajustar a máquina, isto é, reajustar o peso médio.

a) Qual é a probabilidade de ser feita uma parada desnecessária?

b) Se o peso médio da máquina desregulou-se para 500g, qual é a probabilidade de continuar a prdução fora dos padrões desejados?

4 Métodos de Estimação

Nessa seção veremos duas maneiras de estimar parâmetros a partir dos estimadores selecionados: a estimação por ponto e a estimação por intervalo de confiança.

4.1 Estimação por ponto

É o processo através do qual obtemos um único ponto, ou seja, um único valor para estimar o parâmetro
Exemplo: Amostra (1,3,5)

$$\bar{x} = \sum x_i = \frac{1+3+5}{3} = 3 \ \leftarrow \textit{estimativa pontual de } \mu \tag{46}$$

$$S_2 = \frac{\sum(x_i - \bar{x})}{n-1} = \frac{(1-3)^2 + (3-3)^2 + (5-3)^2}{3-1} = 4 \ \leftarrow \textit{estimativa pontual de } \sigma_2 \tag{47}$$

Exercícios propostos

73) (Montgomery; Runger) Dados sobre a força (libra-força) de remoção de conectores usados em um motor de automóveis são os seguintes: 79,3; 75,1; 78,2; 74,1; 73,9; 75,0; 77,6; 77,3; 73,8; 74,6; 75,5; 74,0; 74,7; 75,9; 72,9; 73,8; 74,2; 78,1; 75,4; 76,3; 75,3; 76,2; 74,9; 78,0; 75,1; 76,8.

a) Calcule a estimativa pontual da força média de remoção de todos os conectores na população. Que estimador você usou e por que?

b) Calcule as estimativas pontuais da variância e do desvio-padrão da população.

c) Calcule o erro-padrão da estatística (estimador pontual) usado no item a). Forneça uma interpretação do erro-padrão.

d) Calcule uma estimativa pontual da proporção de todos os conectores na população cuja força de remoção é menor do que 73 libras-força.

74) Suponha que X seja o número de "sucessos" observados em uma amostra de n observações, em que p é a probabilidade de sucesso em cada observação. a) Mostre que $\hat{p} = \frac{X}{n}$ é um estimador não-tendencioso de p. b) Mostre que o erro-padrão de $\hat{p}$ é $\sqrt{p\left(\frac{1-p}{n}\right)}$. Como você estimaria o erro-padrão?

4.2 Estimação intervalar

A estimação por ponto é uma técnica prática e comum para obter uma estimativa do valor de um parâmetro desconhecido em uma população, com base em informações amostrais. No entanto, é importante ressaltar que essas estimativas estão sujeitas a um certo grau de incerteza, uma vez que as estimativas pontuais podem diferir do valor real do parâmetro. Para obter uma melhor noção da incerteza associada à estimação, construímos o que chamamos de "intervalo de confiança"em torno da estimativa pontual.
Todo intervalo de confiança conta com dois parâmetros importantes, são eles:

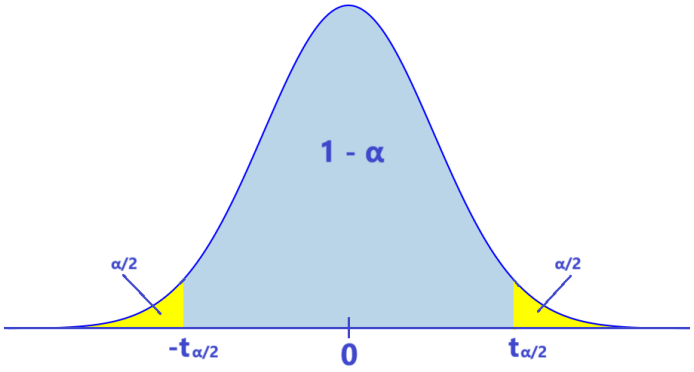
- α = significância
- $1 - \alpha$ = confiança

Por serem variáveis aleatórias, os estimadores pontuais possuem uma distribuição de probabilidade (distribuições amostrais). Com isso, podemos apresentar uma estimativa mais informativa para o parâmetro de interesse, que inclua uma medida de precisão do valor obtido → estimativa intervalar ou intervalo de confiança. Os intervalos de confiança são obtidos a partir da distribuição amostral de seus estimadores.

Exercícios propostos

75) Para uma população normal com variância conhecida σ^2:

Figura 13: Estimação intervalar



a) Qual valor de $Z_{\alpha/2}$ fornece 95% de confiança?

b) Qual valor de $Z_{\alpha/2}$ fornece 99% de confiança?

c) Qual valor de $Z_{\alpha/2}$ fornece 90% de confiança?

d) Qual valor de $Z_{\alpha/2}$ fornece 98% de confiança?

76) Suponha que $n = 100$ amostras aleatórias de água proveniente de um lago com água fresca foram retiradas, sendo medida a concentração (miligramas por litro) de cálcio. Um IC de 95% para a concentração média de cálcio é $(0,49; 0,82)$.

a) Um IC de 99% calculado a partir dos dados da amostra seria maior ou menor?

b) Considere a seguinte afirmação: há uma chance de 95% de μ estar entre 0,49 e 0,82. Esta afirmação é correta? Explique sua resposta.

c) Considere a seguinte afirmação: se $n = 100$ amostras aleatórias de água proveniente do lago forem tomadas e o IC de 95% para μ for calculado e esse processo for repetido 1000 vezes, 950 dos ICs conterão o valor verdadeiro de μ. Esta afirmação está correta? Explique sua resposta.

77) O rendimento de um processo químico está sendo estudado. De experiências prévias com esse processo, sabe-se que o rendimento tem distribuição normal com $\sigma = 3$. Os últimos cinco dias de operação da planta resultaram nos seguintes rendimentos percentuais: 91,6; 88,75; 90,8; 89,95 e 91,3. Encontre um intervalo com 95% de confiança para o rendimento médio verdadeiro.

78) Um fabricante produz anéis para pistões de um motor de um carro. Sabese que o diâmetro do anel tem distribuição normal com $\sigma = 0,001$ milímetro. Em uma amostra aleatória de 15 anéis foi obtido um diâmetro médio de $\bar{x} = 74,036$ milímetros. Construa um intervalo 99% de confiança para o diâmetro médio do anel do pistão.

79) Um engenheiro do setor de pesquisa de uma fábrica de pneus está investigando o tempo de vida útil do pneu em relação a um novo componente de borracha. Ele fabricou 16 pneus e testou-os até o final da vida em um teste na estrada. A vida média e o desvio padrão da amostra foram 60139,7 e 3645,94km. Encontre um intervalo de 95% de confiança para a vida média do pneu.

5 Técnicas de amostragem

De acordo com Silva et al. (2018, p.202), a amostragem é o conjunto de técnicas utilizadas para a seleção de uma amostra. Essas técnicas podem ser subdivididas em dois grupos principais: amostragem aleatória e amostragem não aleatória. Na amostragem não aleatória, estão incluídas técnicas como a amostragem intencional, na qual o pesquisador seleciona intencionalmente os componentes da amostra, e a amostragem voluntária, onde os componentes da população se oferecem voluntariamente para participar da amostra, independentemente do julgamento do pesquisador.

O processo de amostragem desempenha um papel crucial na obtenção de resultados precisos e representativos, permitindo-nos tirar conclusões válidas sobre toda a população com base nas informações contidas na amostra.

Existem várias técnicas de amostragem que podem ser utilizadas, e a escolha da técnica adequada depende do objetivo do estudo, das características da população e das restrições de recursos disponíveis.

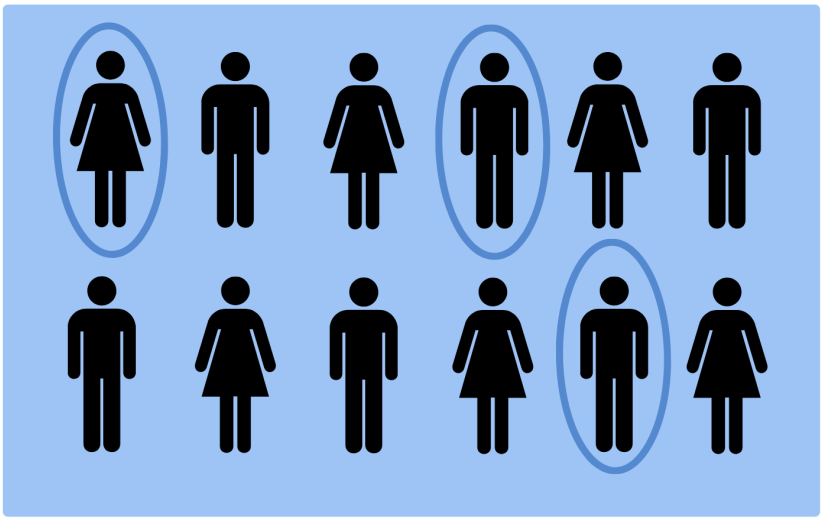
A seguir, apresentaremos algumas das principais técnicas de amostragem.

5.1 Amostragem Aleatória Simples

Para Silva et al. (2018, p. 202), a Amostra Aleatória Simples (AAS) é um método de seleção em que todos os elementos da população têm a mesma chance de serem escolhidos para fazer parte da amostra. Em outras palavras, cada elemento tem uma probabilidade igual de participar da amostra, o que significa que todos os grupos ou indivíduos têm uma oportunidade igualitária de serem selecionados.

Morettin e Bussab (2017, p. 276) dizem que "pode-se ter uma AAS com reposicao, se for permitido que uma unidade possa ser sorteada mais de uma vez, e sem reposicao, se a unidade sorteada for removida da populacao."

Figura 14: Amostra aleatória simples (AAS)



Ou seja, nesse método, cada elemento da população tem a mesma probabilidade de ser selecionado na amostra. É como realizar um sorteio justo, onde todos os indivíduos têm uma chance igual de serem escolhidos. Esse tipo de amostragem é útil quando a população é homogênea - todos os itens da amostra são escolhidos porque têm características semelhantes ou idênticas - e não existe uma estrutura específica.

Exemplo: Imagine que você é um pesquisador interessado em investigar o nível de satisfação dos clientes de uma loja de eletrônicos. A população de interesse é composta por todos os clientes que fizeram compras na loja nos últimos 6 meses, e você deseja selecionar uma amostra para coletar suas opiniões.

Exemplo no Python

```
1   import random
```

```
# Populacao de clientes que fizeram compras na loja de eletronicos
nos ultimos 6 meses
populacao_clientes = ['Cliente 1', 'Cliente 2', 'Cliente 3', 'Cliente 4',
      'Cliente 5', 'Cliente 6', 'Cliente 7', 'Cliente 8', 'Cliente 9',
      'Cliente 10']

# Tamanho da amostra desejada
tamanho_amostra = 5

# Realizando a Amostragem Aleatoria Simples (AAS) sem reposicao
amostra_aas = random.sample(populacao_clientes, tamanho_amostra)

print("Amostra Aleat ria Simples:", amostra_aas)
```

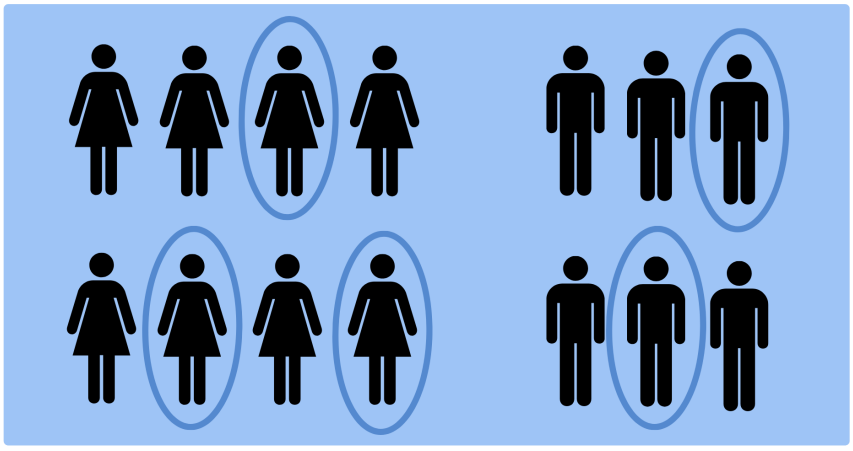
Exemplo no R

```
    populacao_clientes <- c('Cliente 1', 'Cliente 2', 'Cliente 3', 'Cliente 4',
     'Cliente 5', 'Cliente 6', 'Cliente 7', 'Cliente 8', 'Cliente 9',
     'Cliente 10')

tamanho_amostra <- 5  # Tamanho da amostra desejada

# Realizando a Amostragem Aleat ria Simples (AAS) sem reposi  o
amostra_aas <- sample(populacao_clientes, tamanho_amostra,
     replace = FALSE)

print(paste("Amostra Aleatoria Simples:", amostra_aas))
```

5.2 Amostragem Estratificada

Para Bolfarine (2005, p. 110), "Amostragem estratificada consiste na divisão de uma população em grupos (chamados estratos) segundo alguma(s) característica(s) conhecida(s) na população sob estudo, e de cada um desses estratos são selecionadas amostras em proporções convenientes."

Figura 15: Amostra Estratificada



Nessa abordagem, a população é dividida em estratos ou subgrupos com características semelhantes. Em seguida, uma amostra aleatória simples é coletada em cada estrato. Isso garante que cada estrato esteja representado na amostra, o que é útil quando diferentes grupos da população têm características distintas.

5.3 Amostragem Sistemática

Digamos que queremos fazer uma pesquisa sobre a satisfação dos moradores da cidade de Itapajé, no Ceará, em relação aos serviços de transporte público. Para realizar uma amostragem sistemática, devemos primeiro obter uma lista de todos os habitantes elegíveis.

Suponha que a cidade de Itapajé tenha 50.000 habitantes e cada habitante seja identificado por um número único de registro. Para selecionar uma amostra sistemática de 5.000 habitantes, fazemos o seguinte:

1. Calcular o intervalo de amostragem:

$$\text{Intervalo} = \frac{\text{População total}}{\text{Tamanho da amostra desejada}} = \frac{50.000}{5.000} = 10 \tag{48}$$

2. Em seguida, um número aleatório entre 1 e 10 é selecionado. Suponha que o número aleatório escolhido seja 8.

3. Os moradores cujo número de registro é 8 módulo 10 (8, 28, 48, 68, ...) são selecionados a partir do número aleatório escolhido (8) até que o tamanho amostral desejado (5.000) seja atingido. Assim, uma amostra sistemática incluiria moradores identificados com os números de registro 8, 18, 28, 38, 48, ... até o 5.000º morador selecionado.

Exemplo no Python

```
    import random

tamanho_populacao = 50000
tamanho_amostra = 5000

intervalo = tamanho_populacao // tamanho_amostra

numero_aleatorio = random.randint(1, intervalo)

amostra_sistematica = list(range(numero_aleatorio, tamanho_populacao + 1,
      intervalo))

print(amostra_sistematica)
```

Exemplo no R

```
tamanho_populacao <- 50000  # Tamanho da popula  o de Itapaj

tamanho_amostra <- 5000  # Tamanho da amostra desejada

intervalo <- ceiling(tamanho_populacao / tamanho_amostra)

numero_aleatorio <- sample(1:intervalo, 1)

# Selecionando a amostra sistem tica
amostra_sistematica <- seq(from = numero_aleatorio, to = tamanho_populacao,
     by = intervalo)

print(amostra_sistematica)
```

Nesse método, selecionamos um ponto de partida aleatório e, em seguida, escolhemos periodicamente os elementos subsequentes até atingir o tamanho desejado da amostra. Essa técnica é útil quando os elementos estão organizados em uma ordem específica, como uma lista ou arquivo.

5.4 Amostragem por Conglomerados

Aqui, a população é dividida em grupos ou conglomerados, e uma seleção aleatória de conglomerados é feita. Em seguida, todos os elementos dos conglomerados selecionados são incluídos na amostra.

Em outras palavras, a amostragem por conglomerados consiste em dividir a população em grupos menores e heterogêneos, conhecidos como conglomerados, e, em seguida, realizar um sorteio aleatório para selecionar alguns desses grupos, que serão considerados como a amostra completa.

Figura 16: Amostra Sistemática

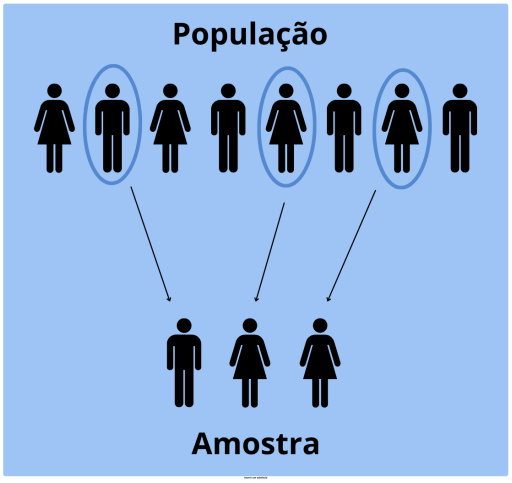
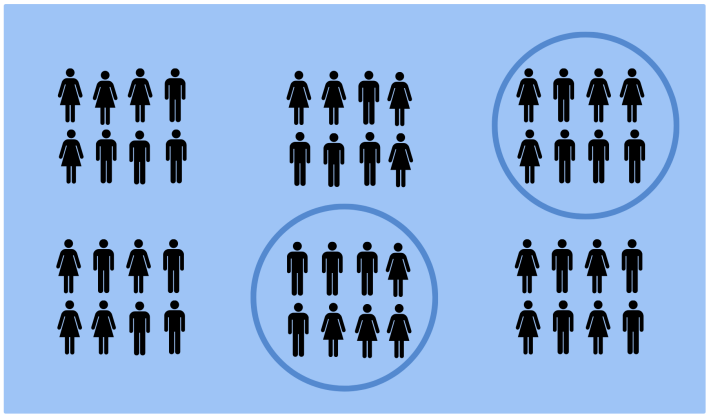


Figura 17: Amostragem por Conglomerados



Exemplo: Vamos supor que o município de Itapajé seja dividido em quatro bairros: Centro, Bairro A, Bairro B e Bairro C. Queremos realizar uma amostragem por conglomerados para estudar o nível de satisfação dos moradores em relação aos serviços públicos oferecidos em cada bairro.

Para isso, faremos o seguinte:

1. Dividiremos a população do município em quatro conglomerados: Centro, Bairro A, Bairro B e Bairro C.
2. Faremos um sorteio aleatório e selecionaremos dois dos quatro conglomerados para fazer parte da amostra. Suponha que os conglomerados sorteados sejam Centro e Bairro B.
3. Entrevistaremos todos os moradores do Centro e do Bairro B selecionados, pois a amostra será composta por todos os elementos dos conglomerados sorteados.
4. Com as informações coletadas dos moradores do Centro e do Bairro B, poderemos analisar o nível de satisfação em relação aos serviços públicos nesses dois bairros específicos.

5.5 Amostragem por Conveniência

Essa técnica envolve a seleção de elementos da população com base em sua disponibilidade e acessibilidade. Embora seja uma forma rápida e conveniente de obter dados, pode levar a um viés de seleção, pois os

elementos mais acessíveis podem não ser representativos da população em geral.

Exemplo: Suponha que um pesquisador queira coletar informações sobre o nível de satisfação dos estudantes do campus Jardins de Anita da Universidade Federal do Ceará (UFC) em relação aos serviços do centro de convivência. O pesquisador decide entrevistar os estudantes que estão próximos à entrada do centro de convivência e que têm tempo disponível naquele momento.

Nesse exemplo, a amostragem por conveniência torna a coleta de dados rápida e conveniente, pois os entrevistados estão facilmente acessíveis. No entanto, é importante lembrar que essa abordagem pode introduzir viés de seleção, uma vez que estudantes que frequentam essa área podem não ser representativos de toda a população de estudantes da UFC Campus Jardins de Anita. Portanto, os resultados obtidos por essa técnica podem não refletir a opinião geral de todos os estudantes do campus.

5.6 Amostragem por Julgamento

Nesse método, os pesquisadores escolhem deliberadamente elementos específicos da população que consideram mais relevantes ou representativos para a pesquisa. Embora possa ser útil em certos cenários, essa abordagem também pode levar a um viés de seleção e não é recomendada para estudos que exijam alta representatividade.

Exemplo: Vamos supor que os estudantes do curso de Ciência de Dados da UFC Campus Jardins de Anitta desejam realizar um estudo sobre as habilidades de programação dos alunos do curso. Eles decidem utilizar a amostragem por julgamento para selecionar os participantes da pesquisa.

Nesse método, os pesquisadores analisam cuidadosamente os alunos do curso de Ciência de Dados e selecionam deliberadamente aqueles que consideram mais relevantes para a pesquisa, com base em critérios específicos, como desempenho acadêmico, experiência em programação ou participação em projetos de tecnologia.

Imaginemos que a equipe de pesquisadores escolheu 30 alunos que são considerados como os melhores programadores do curso. Eles conduzem entrevistas detalhadas e aplicam testes específicos para avaliar as habilidades de programação desses alunos selecionados.

Embora a amostragem por julgamento possa fornecer informações detalhadas sobre um grupo específico de alunos altamente habilidosos em programação, ela não garante representatividade para toda a população de estudantes de Ciência de Dados da UFC. Outros alunos com habilidades diferentes podem não ser incluídos na pesquisa, e, portanto, os resultados podem não refletir precisamente a média das habilidades de programação de todos os alunos do curso.

Exercícios propostos

80) Um estudo sobre o desempenho dos vendedores de uma grande cadeia de lojas de varejo está sendo planejado. Para tanto, deve ser colhida uma amostra probabilística dos vendedores. Classifique cada uma das amostras abaixo conforme a seguinte tipologia:

(A) Amostragem aleatória simples
(B) Amostragem sistemática
(C) Amostragem estratificada
(D) Amostragem por conglomerados

( ) Lista de todos os vendedores (que atuam em todas as lojas da rede). Selecionei todos vendedores que ocupavam posições múltiplas de 15 (15ª posição, 30ª posição, 45ª posição, 60ª posição, 75ª posição, 90ª posição, 105ª posição, etc) A amostra é sistemática, já que a retirada dos elementos da amostra é periódica, ou seja, um vendedor é retirado a cada 15 presentes na lista.
( ) Escolhi casualmente 3 lojas da rede. A amostra foi composta de todos os vendedores que atualmente em cada uma destas 3 lojas. A amostragem é por meio de conglomerados, já que a população (vendedores) apresenta uma subdivisão em grupos (lojas), e 3 grupos foram casualmente escolhidos (sorteados) para compor a amostra.
( ) Em cada uma das lojas, identifiquei todos os vendedores (lista de vendedores por loja). Selecionei aleatoriamente k vendedores da loja. Onde k é um número inteiro proporcional à quantidade de vendedores da loja A amostragem é estratificada, já que a amostra leva em consideração a presença de estratos (lojas) na composição da população (vendedores). Neste caso, a amostragem estratificada é proporcional, pois o número de elementos selecionados em cada estrato é proporcional à quantidade de elementos existentes no

estrato.
( ) Lista de todos os vendedores (que atuam em todas as lojas da rede). Selecionei aleatoriamente N vendedores
A amostragem é casual simples, pois todos os elementos da população (vendedores) tem igual probabilidade de pertencer à amostra.

Marque a opção que corresponde à classificação correta das amostras:

a) A, B, C, D

b) B, D, C, A

c) D, C, B, A

d) B, A, C, D

81) Considere as técnicas de amostragem listadas a seguir:

I. Amostragem por conglomerados.

II. Amostragem por conveniência.

III. Amostragem sistemática.

IV. Amostragem aleatória simples sem reposição.

É incorreto afirmar que:

a) Todas são probabilísticas.

b) Somente I é probabilística.

c) Apenas I e IV são probabilístias.

d) Somente II é não probabilística.

82) Caso uma amostra aleatória simples, sem reposição, de tamanho $n = 3$, seja retirada de uma população constituída por $N = 10$ elementos e representada pelo conjunto $Y_1, Y_2, ..., Y_{10}$, o total de configurações nesse tipo de levantamento será igual a

a) 1.000.

b) 720.

c) 120.

d) 240.

83) Sabemos que um estudo estatístico pode ser feito com todos os elementos da população ou com uma parte desta população (amostra). Para que possamos usar os resultados obtidos na amostra para fazer inferências sobre a população de interesse, precisamos garantir que a amostra seja representativa desta população. (FERREIRA, Valéria. Estatística Básica. Rio de Janeiro: SESES, 2015. Adaptado.)

Para obter uma amostra, pode-se utilizar diferentes técnicas de amostragem; analise-as.

I. Na amostragem aleatória simples, todos os elementos da população têm igual probabilidade de pertencer à amostra, assemelhando-se a um sorteio.

II. Na amostragem aleatória estratificada, os elementos da população são divididos em subgrupos (estratos) e é possível selecionar quantidades proporcionais de elementos de cada subgrupo.

III. Na amostragem sistemática, os elementos da população são organizados e ordenados; seleciona-se um número inicial aleatório, em seguida, os demais elementos são selecionados mantendo-se os intervalos regulares, a partir do número inicial.

Está correto o que se afirma em

a) I, II e III.

b) I, apenas.

c) I e II, apenas.

d) I e III, apenas.

e) II e III, apenas.

84) Uma escola com 500 alunos está fazendo uma pesquisa. O diretor utilizou um computador para gerar 50 números de identificação aleatórios e estes alunos foram convidados a responder a pesquisa. Assinale a alternativa que apresenta que tipo de amostragem é essa.

a) Amostragem aleatória simples

b) Amostragem aleatória estratificada

c) Amostragem aleatória por agrupamento

d) Amostragem aleatória sistemática

85) Suponha que um estatístico tenha feito uma amostragem da intenção de votos dos professores, servidores e alunos de uma universidade, em uma disputa eleitoral entre duas chapas para o cargo de reitor dessa universidade. Assinale a alternativa que apresenta para que pelo menos 1 professor, 1 servidor e 1 aluno estejam na amostra, qual o plano de amostragem o estatístico deverá aplicar?

a) Plano de amostragem de conglomerados em 1 estágio

b) Plano de amostragem aleatória simples

c) Plano de amostragem sistemática

d) Plano de amostragem estratificada proporcional

6 Intervalo de Confiança (IC)

6.1 Conceito

Segundo Silva et al. (2018, p.209), diz que o intervalo de confiança é uma faixa de valores reais, centrada na estimativa pontual que deverá conter o parâmetro com determinada probabilidade. A probabilidade de o intervalo conter o parâmetro estimado é denominado nível de confiança associado ao intervalo.

O Intervalo de Confiança é uma ferramenta estatística que nos permite estimar um valor desconhecido, como a média ou proporção de uma população, com base em uma amostra aleatória dos dados. O IC é construído de tal forma que fornece uma faixa de valores onde a verdadeira medida populacional provavelmente se encontra, com uma determinada probabilidade.

6.2 Como calcular o IC?

A construção de um Intervalo de Confiança depende do parâmetro que estamos estimando (por exemplo, média, proporção) e da distribuição amostral adequada. A ideia geral é que, ao coletar várias amostras aleatórias da mesma população, os valores estimados variarão. O IC é calculado de forma a abranger esses diferentes valores estimados com uma probabilidade pré-definida, conhecida como nível de confiança.

Por exemplo, se quisermos calcular o IC para a média de uma população com um nível de confiança de 95%, usamos a média amostral calculada a partir da amostra e o desvio padrão amostral para determinar a extensão do intervalo. Quanto maior o tamanho da amostra, menor será o intervalo de confiança, o que significa que estamos mais certos de que a estimativa está próxima do valor populacional real.

O cálculo do intervalo de confiança envolve o uso de uma fórmula estatística que leva em conta a estimativa pontual do parâmetro, o tamanho da amostra, o desvio padrão e o nível de confiança desejado.

Sendo assim, existem diferentes fórmulas para calcular o intervalo de confiança dependendo do tipo de parâmetro e da distribuição da amostra.

Confira a seguir as etapas necessárias para realizar o cálculo do intervalo de confiança:

1. Calcule a média amostral (x) e o desvio padrão amostral (s).
2. Determine o nível de confiança desejado (por exemplo, 95%, 99%, etc.). O nível de confiança é a probabilidade de que o verdadeiro valor do parâmetro esteja contido no intervalo de confiança.
3. Determine o valor crítico da distribuição t-Student correspondente ao nível de confiança e ao tamanho da amostra.
4. Calcule o erro padrão (SE) da média, que é dado por SE = s / sqrt(n), onde n é o tamanho da amostra.
5. Calcule o intervalo de confiança, que é dado por x ± (valor crítico da distribuição t-Student) * SE.

6.3 Intervalo de Confiança para a Média (Distribuição Normal)

Quando estamos estimando a média populacional usando uma amostra grande (geralmente com tamanho amostral maior que 30) e a variância populacional é desconhecida, podemos utilizar a distribuição normal para calcular o Intervalo de Confiança.

A fórmula para o Intervalo de Confiança para a média é dada por:

$$\text{Intervalo de Confiança para a Média} = \bar{X} \pm Z \times \frac{S}{\sqrt{n}} \tag{49}$$

Onde:

- $\bar{X}$ é a média amostral.
- Z é o valor crítico da distribuição normal padrão associado ao nível de confiança desejado. Por exemplo, para um nível de confiança de 95%, $Z \approx 1.96$.
- S é o desvio padrão amostral.

- n é o tamanho da amostra.

Exercícios propostos

86) Suponha que os comprimentos de jacarés adultos de certa raça siga o modelo Normal com média µ desconhecida e variância igual a $0,01m^2$.Uma amostra de 10 animais foi sorteada e forneceu média 1,69 m. Encontre uma estimativa intervalar de 95% para o parâmetro desconhecido μ.

87) Uma nova ligação metálica é planejada para ser usada em aviões. As medidas de durabilidade num laboratório experimental são feitas com 15 peças da liga metálica. Dessa amostra, foram obtidas a média amostral x = 39 e o desvio padrão amostral s = 2,6. Achar o intervalo de confiança 90% para a média populacional.

88) Um provedor de acesso à Internet esta monitorando a duração do tempo das conexões de seus clientes, com o objetivo de dimensionar seus equipamentos. São desconhecidas a média e distribuição de probabilidades desse tempo, mas o desvio padrão, por analogia a outros serviços, é considerado igual a $\sqrt{50}$ minutos. Uma amostra de 500 conexões resultou num valor médio observado de 25 minutos. O que dizer da verdadeira média com confiança de 92%.

89) Obtenha o intervalo de confiança para média de uma população normal μ a partir da afirmativa probabilística dada por:

$$P\left(-t_{\alpha/2} \leq \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \leq t_{\alpha/2}\right) = 1-\alpha \tag{50}$$

6.4 Intervalo de Confiança para a Proporção (Distribuição Binomial)

Quando estamos estimando a proporção populacional usando uma amostra grande (geralmente com tamanho amostral maior que 30) e a proporção populacional é desconhecida, podemos utilizar a distribuição binomial para calcular o Intervalo de Confiança.

A fórmula para o Intervalo de Confiança para a proporção é dada por:

$$\text{Intervalo de Confiança para a Proporção} = \hat{p} \pm Z \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{51}$$

Onde:

- $\hat{p}$ é a proporção amostral (número de sucessos dividido pelo tamanho da amostra).
- Z é o valor crítico da distribuição normal padrão associado ao nível de confiança desejado. Por exemplo, para um nível de confiança de 95%, $Z \approx 1.96$.
- n é o tamanho da amostra.

Exemplo

Suponha que você deseja estimar a altura média de todos os funcionários de uma fábrica calçadista. Como medir a altura de todos os funcionários é inviável devido ao grande número de pessoas, você decide utilizar uma amostra aleatória para obter essa informação.

Passo 1: Coleta dos dados e cálculo das estatísticas

Você coleta uma amostra aleatória de 50 funcionários e registra a altura de cada um deles. Após coletar os dados, você calcula a média amostral ($\bar{x}$) e o desvio padrão amostral (s) das alturas. Suponha que a média amostral seja de 165 cm e o desvio padrão amostral seja de 6 cm.

$$\text{Tamanho da amostra (n)} = 50 \text{ funcionários} \tag{52}$$

Passo 2: Determinação do nível de confiança e valor crítico

Você decide calcular um intervalo de confiança de 95% para a verdadeira média populacional de altura dos funcionários. Portanto, o nível de confiança é de 95%.

Para calcular o valor crítico, você precisa consultar a tabela t-Student para a distribuição t-Student com 49 graus de liberdade (n - 1), onde n é o tamanho da amostra (50 - 1 = 49). Para um nível de confiança de 95%, o valor crítico é aproximadamente 2,009.

Passo 3: Cálculo do erro padrão (SE)

O erro padrão (SE) é calculado pela fórmula:

$$SE = \frac{s}{\sqrt{n}} \tag{53}$$

Substituindo os valores:

$$SE = \frac{6}{\sqrt{50}} \approx 0,848 \tag{54}$$

Passo 4: Cálculo do intervalo de confiança

O intervalo de confiança é calculado utilizando a fórmula:

$$\text{Intervalo de Confiança} = \bar{x} \pm (\text{valor crítico} \times SE) \tag{55}$$

Substituindo os valores:

$$\text{Intervalo de Confiança} = 165 \pm (2,009 \times 0,848) = 165 \pm 1,702 \tag{56}$$

Portanto, o intervalo de confiança de 95% para a altura média de todos os funcionários da fábrica é de 163,298 cm a 166,702 cm. Isso significa que com 95% de confiança, a verdadeira média de altura de todos os funcionários da fábrica está contida nesse intervalo de valores.

Exemplo no Python

```
import numpy as np
from scipy.stats import t

# Dados da amostra (altura dos funcion rios em cent metros)
amostra_altura = np.array([167, 172, 165, 160, 174, 169, 171, 168, 175, 166,
                           170, 172, 163, 168, 169, 172, 167, 170, 173, 168,
                           171, 169, 172, 170, 166, 169, 174, 167, 170, 172,
                           175, 170, 168, 172, 171, 169, 166, 174, 167, 172,
                           171, 168, 169, 172, 165, 170, 173, 168, 171, 169])

# Tamanho da amostra
n = len(amostra_altura)

# M dia amostral
media_amostral = np.mean(amostra_altura)

# Desvio padr o amostral
desvio_padrao_amostral = np.std(amostra_altura, ddof=1)

# N vel de confian a desejado (95%)
nivel_confianca = 0.95

# Graus de liberdade
graus_liberdade = n - 1

# Valor cr tico da distribui  o t-Student
valor_critico = t.ppf((1 + nivel_confianca) / 2, graus_liberdade)

# Erro padr o
erro_padrao = desvio_padrao_amostral / np.sqrt(n)

```

```
# Intervalo de confian a
intervalo_confianca = (media_amostral - valor_critico * erro_padrao,
                       media_amostral + valor_critico * erro_padrao)

print("Intervalo de confian a de 95% para a altura m dia dos funcion rios:")
print(f"{intervalo_confianca[0]:.2f} cm a {intervalo_confianca[1]:.2f} cm")
```

Exemplo no R

```
    # Dados da amostra (altura dos funcion rios em cent metros)
amostra_altura <- c(167, 172, 165, 160, 174, 169, 171, 168, 175, 166,
                    170, 172, 163, 168, 169, 172, 167, 170, 173, 168,
                    171, 169, 172, 170, 166, 169, 174, 167, 170, 172,
                    175, 170, 168, 172, 171, 169, 166, 174, 167, 172,
                    171, 168, 169, 172, 165, 170, 173, 168, 171, 169)

# Tamanho da amostra
n <- length(amostra_altura)

# M dia amostral
media_amostral <- mean(amostra_altura)

# Desvio padr o amostral
desvio_padrao_amostral <- sd(amostra_altura)

# N vel de confian a desejado (95%)
nivel_confianca <- 0.95

# Graus de liberdade
graus_liberdade <- n - 1

# Valor cr tico da distribui  o t-Student
valor_critico <- qt((1 + nivel_confianca) / 2, df = graus_liberdade)

# Erro padr o
erro_padrao <- desvio_padrao_amostral / sqrt(n)

# Intervalo de confian a
intervalo_confianca <- c(media_amostral - valor_critico * erro_padrao,
                         media_amostral + valor_critico * erro_padrao)

cat("Intervalo de confian a de 95% para a altura m dia dos funcion rios:\n")
cat(sprintf("%.2f cm a %.2f cm", intervalo_confianca[1], intervalo_confianca[2]))
```

Exercícios propostos

90) Em uma amostra aleatória de 85 mancais de eixos de manivelas de motores de automóveis, 10 têm um acabamento superficial mais rugoso do que as especificações permitidas. Calcule um intervalo de confiança para o 95% da proporção.

91) Sabe-se que a proporção de animais contaminados com uma determinada doença não é superior a 10%. Qual deve o tamanho da amostra para que a semi amplitude do intervalo com 92% de confiança para a fração populacional não seja superior a 2%?

92) De 1.000 casos selecionados de aleatoriamente de câncer de pulmão, 823 resultaram em morte. Construa um intervalo de confiança de 95% para a taxa de morte de câncer de pulmão.

93) Suponha que $p = 30\%$ dos estudantes de uma escola sejam mulheres. Colhemos uma amostra aleatória simples 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?

94) Supondo que a produção do exemplo anterior esteja sob controle, isto é, $p = 10\%$, e que os itens sejam vendidos em caixas com 100 unidades, qual é a probabilidade que uma caixa:

a) tenha mais do que 10% de itens defeituosos?

b) não tenha itens defeituosos?

95) Um distribuidor de sementes determina, por meio de testes, que 5% das sementes não germinam. Ele vende pacotes com 200 sementes com garantia de 90% de germinação. Qual é a probabilidade de que um pacote não satisfaça à garantia?

96) Uma amostra aleatória de 625 donas de casa revela que 70% delas preferem a marca A de detergente. Construir um intervalo de confiança para p = proporção das donas de casa que preferem A com coeficiente de confiança $\gamma = 90\%$.

7 Testes de Hipóteses

7.1 Conceito

Os testes de hipóteses são procedimentos estatísticos utilizados para tomar decisões sobre afirmações ou suposições feitas a respeito de uma população com base em informações obtidas a partir de uma amostra. O objetivo principal de um teste de hipóteses é avaliar se os dados fornecem evidências suficientes para rejeitar ou não uma afirmação sobre o parâmetro populacional.

As afirmações que são testadas em um teste de hipóteses são chamadas de hipóteses estatísticas. Uma hipótese é geralmente dividida em duas partes: a hipótese nula (H_0) e a hipótese alternativa (H_1). A hipótese nula é a afirmação inicial que assume que não há efeito ou diferença real, enquanto a hipótese alternativa é a afirmação que o pesquisador pretende testar e mostra a direção da diferença ou efeito.

7.2 Hipóteses estatísticas

A partir de amostras, realizamos inferências - suposições - sobre os parâmetros de uma população, o que implica em formular suposições que podem ser verídicas ou não. São exemplos de hipóteses estatísticas:

- A média populacional de renda dos residentes de um determinado bairro é de R$ 3.500, ou seja, $\mu = R\$3.500$.
- A proporção de estudantes aprovados em um exame é de 75%, o que equivale a $p = 0,75$.

7.3 Tipos de hipóteses

No procedimento de teste de hipóteses, são elaboradas duas proposições, conhecidas como Hipótese Nula (H_0) e Hipótese Alternativa (H_1). Ambas estão relacionadas, principalmente, a um parâmetro que representa um valor na população e à sua estimativa correspondente a partir da amostra.

A Hipótese Nula (H_0) representa a suposição que está sujeita a teste. Em outras palavras, é a hipótese que se busca verificar por meio do teste estatístico.

A Hipótese Alternativa (H_1) engloba todas as outras possíveis proposições que são distintas da Hipótese Nula. É uma afirmação que se coloca como alternativa à Hipótese Nula, refletindo diferentes cenários ou resultados que podem ocorrer.

Recaptulando:

- A **hipótese nula (H_0)** sustenta que não existe qualquer discrepância entre o parâmetro e seu estimador.
- A **hipótese alternativa (H_1)** tem a obrigação de refutar a hipótese nula em todos os casos.

7.4 Tipos de teste

Um teste estatístico de hipótese é um procedimento que, quando aplicado a valores amostrais, orienta a escolha entre aceitar ou rejeitar uma hipótese específica. Em geral, quando o parâmetro de interesse é a média (μ) de uma população, as hipóteses nula e alternativa são formuladas da seguinte maneira:

1. Teste bicaudal ou bilateral: $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$
2. Teste unicaudal ou unilateral à direita: $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$
3. Teste unicaudal ou unilateral à esquerda: $\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$

7.5 Nível de significância

Ao tomar uma decisão com base na análise de hipóteses estatísticas, existe sempre a possibilidade de cometer um erro, o que significa aceitar erroneamente uma hipótese. Essa probabilidade é denominada nível de significância e é representada por α.

7.6 Tipos de Erros

Em um teste de hipótese, existem duas categorias de erros possíveis. Um erro de decisão ocorre quando uma hipótese que é comprovadamente falsa é aceita como verdadeira, ou quando uma hipótese que é verdadeira é erroneamente considerada falsa.

- Erro do Tipo 1: ocorre quando a hipótese H_0 é rejeitada, mesmo que seja verdadeira. O nível de significância do teste, representado por α, indica a probabilidade de cometer esse tipo de erro.
- Erro do Tipo 2: ocorre quando a hipótese H_0 é aceita, embora seja falsa. Rejeitar H_0 implica a aceitação de H_1, e vice-versa. A probabilidade de cometer um erro do Tipo 2 é igual a 1 menos o nível de significância $(1 - \alpha)$.

7.7 Regiões de Aceitação e Rejeição

- Região de Aceitação É a região onde a hipótese nula (H_0) é considerada válida.
- Região de Rejeição É a região onde a hipótese nula (H_0) é descartada, e é o oposto da região de aceitação. Também é conhecida como região crítica.

7.8 Construção do Teste

A construção de um teste de hipóteses envolve os seguintes passos:

1. Formular as hipóteses e o tipo de teste:

    - A hipótese nula (H_0) sustenta que não existe qualquer discrepância entre o parâmetro e seu estimador.
    - A hipótese alternativa (H_1) tem a obrigação de refutar a hipótese nula em todos os casos.
    - Teste bicaudal ($\mu \neq \mu_0$), teste unilateral à direita ($\mu > \mu_0$) ou teste unilateral à esquerda ($\mu < \mu_0$).

2. Escolher o nível de significância (α):

    O nível de significância é a probabilidade de cometer um erro do tipo I, ou seja, rejeitar a hipótese nula quando ela é verdadeira. É geralmente definido comumente em 5% (0,05) ou 1% (0,01).

3. Selecionar a estatística de teste apropriada:

    A escolha da estatística de teste depende da natureza dos dados e do parâmetro que está sendo testado. Algumas delas são:

    - Teste *t de Student*: Usado para testar diferenças entre médias de duas amostras ou uma amostra em relação a um valor teórico.
    - Teste *Z*: Similar ao teste t de Student, mas é usado quando a população tem uma distribuição normal e o desvio padrão populacional é conhecido.
    - Teste *Qui-quadrado*: Utilizado para testar a independência entre duas variáveis categóricas.
    - Teste de *ANOVA* (Análise de Variância): Usado para testar a igualdade de médias entre três ou mais grupos.

- Teste de *Wilcoxon-Mann-Whitney*: Usado para testar diferenças entre medianas de duas amostras independentes.
- Teste de *Kolmogorov-Smirnov*: Utilizado para testar se uma amostra segue uma determinada distribuição teórica.

4. Calcular o valor da estatística de teste:

   A estatística de teste é calculada a partir dos dados da amostra e é usada para comparar com um valor crítico.

5. Comparar o valor da estatística de teste com o valor crítico:

   Você compara o valor da estatística de teste com o valor crítico adequado, determinado pelo nível de significância e pelo tipo de teste (bicaudal ou unilateral). A decisão de rejeitar ou não rejeitar a hipótese nula é tomada com base nessa comparação. Se o valor da estatística de teste estiver na região de rejeição, rejeitamos a hipótese nula e aceitamos a hipótese alternativa. Caso contrário, não rejeitamos a hipótese nula.

6. Interpretar os resultados:

   Com base na decisão tomada, interpretamos os resultados do teste e tiramos conclusões sobre a hipótese nula e a hipótese alternativa.

7.9 Teste de Hipótese para a Média Populacional (μ)

1. Determinar a hipótese nula $H_0 : \mu = \mu_0$;
2. Determinar a hipótese alternativa;
   $H_1 : \begin{cases} \mu \neq \mu_0 \\ \mu > \mu_0 \\ \mu < \mu_0 \end{cases}$
3. Fixar o nível de significância α;
4. Determinar a região de rejeição (bilateral, unilateral à direita ou unilateral à esquerda);
5. Calcular a estatística amostral;
   $$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \tag{57}$$

   onde:

   - $\bar{x}$ é a média da amostra;
   - μ_0 é a média da população (hipótese a ser testada);
   - σ é o desvio padrão da população;
   - n é o número de elementos da amostra.

   Para grandes amostras ($n > 30$) pode-se utilizar o desvio padrão da amostra (s). Quando a amostra for pequena ($n \leq 30$) devemos utilizar a distribuição de Student.

   $$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \tag{58}$$

   onde:

   - $\bar{x}$ é a média da amostra;
   - μ_0 é a média da população (hipótese a ser testada);

- s é o desvio padrão da amostra;
- n é o número de elementos da amostra

6. Conclusões:

   (a) Se $\left|z\right| > Z_{\frac{\alpha}{2}}$, rejeita-se H_0. (Teste bilateral)

   (b) Se $Z > Z_{\alpha}$, rejeita-se H_0. (Teste unilateral)

   (c) Se $Z < -Z_{\alpha}$, rejeita-se H_0. (Teste unilateral)

Os valores críticos de z associados aos níveis de significância mais utilizados estão apresentados na tabela a seguir.

	Nível de significância α				
	0,002	**0,005**	**0,01**	**0,05**	**0,10**
Valores z para testes unilaterais	2,88	2,58	2,33	1,65	1,28
Valores z para testes bilaterais	3,08	2,81	2,58	1,96	1,65

OBS.: Os valores críticos de t devem ser retirados da tabela da Distribuição de Student, visto que dependem do grau de liberdade ($v = n - 1$). Observação. Nos testes bilaterais, para determinar o t crítico, utilizar como parâmetro na tabela de Student a metade do valor do nível de significância.

7.10 Teste de Hipótese para a Proporção Populacional (p)

1. Estabelecer a hipótese nula $H_0 : p = p_0$;
2. Estabelecer a hipótese alternativa;

   - $H_1 : p \neq p_0$
   - $H_1 : p > p_0$
   - $H_1 : p < p_0$

3. Fixar o nível de significância α;
4. Determinar a região de rejeição;
5. Calcular a estatística Z_calc:

$$Z_{calc} = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \tag{59}$$

   onde:

   - $\hat{p}$ é a proporção ou percentagem da amostra
   - p é a proporção ou percentagem da população
   - q é $(1 - p)$
   - n é o número de elementos da amostra

6. Conclusões:

   (a) Se $|Z_{calc}| > Z_{\frac{\alpha}{2}}$, rejeita-se H_0, onde $Z_{\frac{\alpha}{2}} = Z_{critico} = Z(tabelado)$

   (b) Se $Z > Z_{\alpha}$, rejeita-se H_0

(c) Se $Z < -Z_\alpha$, rejeita-se H_0

Exemplo
Teste de Hipóteses para Médias

Suponha que uma empresa de e-commerce deseje avaliar se o tempo médio de entrega de seus pedidos é menor do que 4 dias. Para isso, eles coletam uma amostra aleatória de 50 pedidos e registram o tempo de entrega de cada um deles. O tempo médio de entrega na amostra é de 3,8 dias, e o desvio padrão é de 0,6 dias.

A hipótese nula (H_0) é que o tempo médio de entrega é igual a 4 dias, e a hipótese alternativa (H_1) é que o tempo médio de entrega é menor do que 4 dias.

Agora, vamos realizar o Teste de Hipóteses para Médias no Python e R:

Exemplo no Python

```
   import numpy as np
from scipy.stats import ttest_1samp

# Dados da amostra
amostra = np.array([3.2, 3.9, 3.7, 4.1, 3.6, 3.9, 4.2, 3.8, 3.5, 3.9,
                    4.0, 3.6, 3.9, 3.7, 3.8, 3.5, 3.9, 4.0, 3.8, 4.1,
                    3.7, 3.9, 4.3, 3.6, 3.9, 3.7, 4.0, 3.8, 4.2, 3.9,
                    4.1, 3.6, 4.0, 3.8, 3.9, 3.7, 3.5, 3.9, 3.6, 4.1,
                    3.7, 3.8, 3.5, 3.9, 3.6, 3.8, 4.0, 3.9, 3.7, 3.5])

# Realizando o teste de hip teses
resultado_teste = ttest_1samp(amostra, popmean=4, alternative='less')

# Extraindo os resultados do teste
valor_p = resultado_teste.pvalue
estatistica_teste = resultado_teste.statistic

# Exibindo os resultados
print("Resultado do Teste de Hip teses para M dias no Python:")
print("Valor-p:", valor_p)
print("Estat stica do Teste:", estatistica_teste)
```

Exemplo no R

```
    # Dados da amostra
amostra <- c(3.2, 3.9, 3.7, 4.1, 3.6, 3.9, 4.2, 3.8, 3.5, 3.9,
             4.0, 3.6, 3.9, 3.7, 3.8, 3.5, 3.9, 4.0, 3.8, 4.1,
             3.7, 3.9, 4.3, 3.6, 3.9, 3.7, 4.0, 3.8, 4.2, 3.9,
             4.1, 3.6, 4.0, 3.8, 3.9, 3.7, 3.5, 3.9, 3.6, 4.1,
             3.7, 3.8, 3.5, 3.9, 3.6, 3.8, 4.0, 3.9, 3.7, 3.5)

# Realizando o teste de hip teses
resultado_teste <- t.test(amostra, mu = 4, alternative = "less")

# Extraindo os resultados do teste
valor_p <- resultado_teste$p.value
estatistica_teste <- resultado_teste$statistic

# Exibindo os resultados
cat("Resultado do Teste de Hip teses para M dias no R:\n")
cat("Valor-p:", valor_p, "\n")
cat("Estat stica do Teste:", estatistica_teste, "\n")



# Dados da amostra
amostra <- c(3.2, 3.9, 3.7, 4.1, 3.6, 3.9, 4.2, 3.8, 3.5, 3.9,
             4.0, 3.6, 3.9, 3.7, 3.8, 3.5, 3.9, 4.0, 3.8, 4.1,
             3.7, 3.9, 4.3, 3.6, 3.9, 3.7, 4.0, 3.8, 4.2, 3.9,
             4.1, 3.6, 4.0, 3.8, 3.9, 3.7, 3.5, 3.9, 3.6, 4.1,
             3.7, 3.8, 3.5, 3.9, 3.6, 3.8, 4.0, 3.9, 3.7, 3.5)

```

```r
# Realizando o teste de hip teses
resultado_teste <- t.test(amostra, mu = 4, alternative = "less")

# Extraindo os resultados do teste
valor_p <- resultado_teste$p.value
estatistica_teste <- resultado_teste$statistic

# Exibindo os resultados
cat("Resultado do Teste de Hip teses para M dias no R:\n")
cat("Valor-p:", valor_p, "\n")
cat("Estat stica do Teste:", estatistica_teste, "\n")
```

Exercícios propostos

97) Um pesquisador deseja estudar o efeito de certa substância no tempo de reação de seres vivos a um certo tipo de estímulo. Um experimento é desenvolvido com cobaias, que são inoculadas com a substância e submetidas a um estímulo elétrico, com seus tempos de reação (em segundos) anotados. Os seguintes valores foram obtidos 9,1; 9,3; 7,2; 7,5; 13,3; 10,9; 7,2; 9,9; 8,0; 8,6. Admite-se que, em geral, o tempo de reação tem distribuição Normal com média 8 segundos e desvio padrão 2 segundos. Entretanto, o pesquisador desconfia que o tempo médio aumenta por influência da substância.

a) Construa o intervalo de confiança para o tempo médio de reação das cobaias que usam tal substância considerando o coeficiente de confiança de 95%.

b) Formule as hipóteses adequadas para verificar a opinião do pesquisador;

c) Interprete os erros tipo I e tipo II;

d) Apresente a região crítica do teste e conclua ao nível de significância de 5%;

e) A partir da região crítica do item c, conclua o teste.

98) O teste de hipótese é um procedimento estatístico que auxilia na tomada de decisões. A respeito desse, julgue o item.

"A dita hipótese nula é a tomada como verdadeira para a construção do teste de hipótese."

a) Certo

b) Errado

99) Nas sentenças a seguir, marque V se verdadeiro e F se falso.
( ) O teste de hipóteses é um procedimento estatístico onde se busca verificar uma hipótese a respeito da população, tendo por base dados amostrais.
( ) A hipótese estatística é uma suposição feita a respeito de uma ou mais estatísticas, como, por exemplo a média amostral e a proporção amostral.
( ) A hipótese nula, supõe a igualdade dos parâmetros que estão sendo comparados.
( ) Quando não temos motivos suficientes para supor que uma das médias será maior que a outra, formulamos uma hipótese alternativa unilateral (mais genérica).
( ) Joãozinho Mão Leve foi levado a julgamento. O juiz inocentou Joãozinho, porém ele era culpado. Pensando estatisticamente (H0: culpado), o juiz cometeu, neste caso, o ERRO TIPO I.
( ) José Inocêncio foi absolvido no julgamento. Ele não havia cometido o delito. Pensando estatisticamente, (H0: inocente), evitou-se o ERRO TIPO II.
( ) José Inocêncio foi absolvido no julgamento. Ele não havia cometido o delito. Pensando estatisticamente, (H0: culpado), evitou-se o ERRO TIPO II.
( ) Joãozinho Mão Leve foi levado a julgamento. O juiz inocentou Joãozinho, porém ele era culpado. Pensando estatisticamente (H0: inocente),o juiz cometeu, neste caso, o ERRO TIPO I.

Marque a opção que corresponde corretamente aos itens:

a) V, V, V, V, F, F, F, F

b) V, F, V, F, V, F, V, F

c) F, V, V, V, F, F, V, F

d) F, V, F, F, F, V, V, F

100) Um fabricante de material desportivo desenvolve uma nova linha de pesca sintética sobre a qual ele afirma que tem resistência média à ruptura de 8 kg com desvio padrão de 0,5 kg. Teste a hipótese de que $\mu = 8$ kg, contra a hipótese de que $\mu 8$ kg, se uma amostra de 50 linhas foi testada e apresentou uma média de resistência a ruptura de 7,8 kg. Use um nível de 0,01 de significância.

101) Para reduzir ambos os tipos de erro, devemos:

a) acrescentar uma terceira hipótese ao teste;

b) aumentar o tamanho da amostra;

c) aumentar somente o nível de significância (α);

d) diminuir o tamanho da amostra;

e) aumentar o nível de confiança, ou seja, β.

102) A vida média de uma amostra de 100 lâmpadas, produzidas por uma fábrica, foi calculada em 1570 horas, o desvio padrão indicado é de 120 horas. Se μ é a vida média de todas as lâmpadas produzidas pela companhia, testar a hipótese $\mu = 1600$ horas, em face da hipótese alternativa $\mu \neq 1600$ horas, adotando o nível de significância:

a) 0,05;

b) 0,01.

103) De 50.000 válvulas fabricadas por uma companhia retira-se uma amostra de 400 válvulas, e obtém-se a vida média de 800 horas e desvio padrão de 100 horas.

a) Qual o intervalo de confiança de 99% para a vida média da população de válvulas?

b) Com que confiança dir-se-ia que a vida média é $800 \pm 0{,}98$?

104) O salário médio dos empregados das indústrias da construção civil é de 2,5 salários mínimos, com um desvio padrão de 0,5 salários mínimos. Se uma firma particular emprega 49 empregados com um salário médio de 2,3 salários mínimos, podemos afirmar que esta indústria paga salários inferiores, ao nível de 5%?

105) Uma companhia fabrica cabos cujas tensões de ruptura têm a média de 300 kg e o desvio padrão de 24 kg. Acredita-se que, mediante um processo recentemente aperfeiçoado, a tensão média de ruptura pode ser aumentada. Foram ensaiados 64 cabos feitos através do novo processo obtendo-se uma média de 310 kg para a ruptura. Ao nível de significância de 1% é possível afirmar que houve melhoria no processo?

106) Um fabricante garante que 90% dos equipamentos que fornece a uma fábrica estão de acordo com as especificações exigidas. O exame de uma amostra de 200 peças desse equipamento revelou 25 defeituosas. Teste a afirmativa do fabricante, aos níveis de 5% e 1%.

107) Um industrial deseja certificar-se de que a fração do mercado que prefere seu produto ao de seu concorrente é superior a 70%. Para tanto, colheu uma amostra aleatória de 165 opiniões, das quais 122 lhe foram favoráveis. Pode o industrial ficar satisfeito com esse resultado, adotado o nível de 5% de significância?

8 Considerações Finais

É com imensa satisfação que concluímos esta apostila de Inferência eEstatística, fruto do nosso compromisso como monitoras da "Monitoria Integrada de Estatística para Ciência de Dados"na Universidade Federal do Ceará, sob a orientação e inspiração constante da Profa Dra Elisângela Rodrigues. Esta jornada de aprendizado e criação deste material enriquecedor proporcionou não apenas um aprofundamento nos princípios fundamentais da disciplina, mas também uma oportunidade única de compartilhar esse conhecimento com outros alunos do curso. Os exercícios junto do gabarito fornecidos ao longo desta apostila visam aprimorar a compreensão e a aplicação prática desses conceitos tão essenciais para a Ciência de Dados. Esperamos que esta apostila seja um recurso valioso não apenas para meus colegas, mas para todos que buscam dominar a arte da Inferência Estatística usando as poderosas linguagens de programação R e Python. Agradecemos à Profa Dra Elisângela Rodrigues por sua orientação dedicada e ao programa de monitoria por proporcionar essa incrível oportunidade de contribuir para o nosso aprendizado. Que este material inspire futuras gerações de cientistas de dados na busca por insights valiosos a partir dos dados.

A Apêndice A: Exercícios - Respostas

1) 0%
2) a) 33,33%
b) 66,66%
3) 0,41%
4) 4 vezes
5) A
6) C
7) A
8) B
9) $E(X) = 2,2$ vezes.
10) E
11) $E(X) = 52,00$
12) E
13) B
14) D
15) B
16) B
17) C
18) A
19) C
20) B
21) D
22) C
23) A
24) Tipo B
25) 50.517095
26) a) 81,86%
b) 4,76%
c) 0,02%
27) Como aproximadamente 97,70% têm vida média superior a 40 horas, concluiu-se que a especificação foi satisfeita.
28) D
29) a) 99,18%
b) 33,64%
c) 65,5 horas
30) $\cong 0,3678$
31) A
32) 99,90%
33) 12,40%
34) 0,2526
35) 85,7%
36) 0,3917
37) 290 pesquisadores.
38) 0.1754
39) a) $\cong 0,0067$
b) $\cong 0,9595$
40) a) 54,70%
b) 14,28% c) $ 3427,20
41) 8,21%
42) 0,64, ou seja, a chance da obra ser realizada com base nesses 30 donativos seria de aproximadamente 64%.
43) a)0,1112
b)0,193, probabilidade obtida de uma Poisson de parâmetro 1,5.
44) 0,0336; 0,7946
45) Sim
46) Não
47) C
48) O estimador $\widehat{\mu}_3$ é melhor por usar todas as observações disponíveis, além de ser não viciado e consistente. As estimativas são: $\widehat{\mu}_1 = 2$, $\widehat{\mu}_2 = 1$ e $\widehat{\mu}_3 = 1,4$.
49) FALTA RESPOSTA
50) D
51) a) FALTA RESPOSTA
b) FALTA RESPOSTA
52) a) 0,0132
b) 0,1335
c) 0,6613
53) 0,0244
54) 0,8258
55) a) 0,06681
b) 0,38292
c) 14.1 miliamperes.
56) a) 4
b) 3
c) A garantia deve ser de 135650 km.
57) a) 0,000935
b) 0,123852
58) a) 0,15386
b)0,29389
c) 0,39743
d) 0,01591
e) 0,03595
59) a) 325 unidades
b) 2 unidades
c) 0,0606
d) 0,8185
e) 0,1587
f) 0,668
60) 0,0244
61) 0,8258
62) a) 0,7013
b) 0,1736
c) 0,0001
63) a)28.000
b)1.000
64) 0,0294
65) 0,0174
66) FALTA RESPOSTA
67) a) 0,00169
b) 0,83639
c) 0,01786
d) 0,07234

e) 0,15866
f) 0,11155
g) 0,05475
h) 0,14007
i) 0,18634
j) 0,93916
68) 0,07078. Essa é uma probabilidade um pouco alta, talvez valha a pena a companhia rever a política de reservas e aceitar menos que 400 reservas.
69) 0,12924. Esse é um resultado que pode ocorrer por mero acaso, ou seja, não é um resultado não-usual.
70) a) $E(e) = 0$ e $Var(e) = \frac{\sigma^2}{n}$
b) 0,31732
c) $\delta = 5,16$
d) $n \approx 1537$
71) a) $\mu = 512,89g$
b) 5.23×10^{-3}
72)a)0,07512
b)84.13%
73)a)$\overline{X} = 75,6154$ libras-força. O estimador utilizado foi $\overline{X}$, pois a média amostral é um estimador não tendencioso e de mínima variância de μ.
b) E.P. $s^2 = 2,7382$ libras-força^2 e D.P. $s = 1,6547$ libras-força.
c) O erro padrão de $\overline{X}$ é dado por $\sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$. Uma estimativa do erro padrão de $\bar{X}$ é $\frac{s}{\sqrt{n}} = 0,3245$. O erro padrão de $\bar{X}$ é o desvio padrão das médias das possíveis amostras de tamanho n extraídas da população.
d) $\hat{p} = \frac{1}{26} = 0,0385$
74) FALTA RESPOSTA
75) a) 1,96
b) 2,575
c) 1,645
d) 2,325
76) a) O intervalo de confiança $(1-\alpha) \times 100\%$ para μ é dado por $\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Aumentando a confiança para 99% a amplitude do intervalo também aumenta, pois $Z_{\frac{0,05}{2}} = 1,96$ e $Z_{\frac{0,01}{2}} = 2,575$.
b) A afirmação está incorreta, pois μ não é uma variável aleatória e sim uma constante desconhecida. Nesse caso μ ou estará ou não estará no intervalo.
c) Sim, a afirmação está correta, o método é baseado na construção de intervalos de confiança para cada uma das possíveis amostras de mesmo tamanho, obtidas a partir do mesmo procedimento de seleção. Estabelecendo-se um erro de 5% tem-se que 95% dos intervalos conterão μ e o restante dos intervalos 5% não conterão μ.
77) IC 95% para μ : $(87,8504; 93,1096)$
78) IC 95% para μ : $(74,0353; 74,0367)$
79) $(58197,33; 62082,07)$
80) B
81) D
82) C
83) A
84) A
85) D
86) (1,63;1,75)
87) (38,12;40,48)
88)(24,45;25,55)
89)FALTA RESPOSTA
90) $I.C. = [0,05; 0,19]$ ou $0,05 \leq p \leq 0,19$
91) $n = 689,1 \approx 690$ elementos
92) $I.C. = [0,799; 0,847]$ ou $0,799 \leq p \leq 0,847$
93) $P(-0.07 < Z < 0.07) = 0.056$
94) a) $P(Z > 0) = 0.5$
b) $P(0$ defeituosos em $100) = 0.9^{100} \approx 2.65 \times 10^{-5}$, ou quase zero.
95) 0.9994 e portanto a garantia do vendedor é falsa em somente 0.06% das vezes.
96) $(0.6692; 0.7308)$
97)FALTA RESPOSTA
98) A
99) B
100) $Z_{calc} = -2,828$ (menor do que $-2,58$ (Z_{crtico})), rejeitamos H_0 e concluímos que a resistência média à ruptura é diferente de $8kg$ (é menor que $8kg$).
101) B
102) a) $Z_{calc} = -2,5$ e $Z_{crtico} = -1,96$, rejeita-se H_0, ou seja, o resultado amostral afirma que a média é diferente de 1600h (na realidade é menor do que 1600h), ao nível de 95% de confiança, $1570 \notin (1576,48; 1623,52)$.
b) $Z_{calc} = -2,5$ e $Z_{crtico} = -2,58$, aceita-se H_0, ou seja, o resultado afirma que a média está compreendida no intervalo que satisfaz ao nível de 99% de confiança, $1570 \in (1569,04; 1630,96)$.
103) a) $787,1 < \mu < 812,9$
b)15,86%
104) Sim
105) Sim. Rejeita-se $H_0 : \mu = 300$ e aceita-se $H_1 : \mu > 300$ logo houve melhoria no processo.
106) Aceita-se a hipótese $H_0 : p = 0,9$ para ambos os níveis.
107) Aceita-se a hipótese $H_0 : p = 0,7$, logo o empresário não pode ficar satisfeito.

B Apêndice B: Como usar a tabela normal

Exploramos a distribuição normal, uma das bases mais sólidas da Estatística, também conhecida como Distribuição Gaussiana. Nesta seção, avançaremos ao próximo nível, compreendendo como aproveitar a tabela de valores padronizados para a distribuição normal padrão. Vamos fazer uma breve recapitulação sobre valores padronizados.

Entendemos que o cálculo da probabilidade entre dois pontos a e b envolve a determinação da área sob a curva, uma tarefa que frequentemente requer cálculos complexos e detalhados como:

$$P\left(a < X < b\right) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e \frac{(x-\mu)^2}{2\sigma^2}. \tag{60}$$

Sabemos que a distribuição normal é definida por dois parâmetros, a média μ e o desvio padrão σ portanto oferece diversas distribuições possíveis. No entanto, uma tabela amplamente empregada é aquela associada à distribuição normal padronizada, onde a média é 0 e o desvio padrão é 1.

Muitos fenômenos aleatórios exibem comportamentos que se assemelham a essa distribuição. Alguns exemplos incluem a altura, pressão sanguínea, peso, entre outros. Tomemos a altura como nosso exemplo. Se considerarmos que a altura segue uma distribuição normal com média de 1,7 metros e um desvio padrão de 0,08 metros, qual é a probabilidade de uma pessoa ter altura entre 1,7 e 1,8 metros?

Para calcular essa probabilidade, recorremos à tabela da distribuição normal padronizada. No entanto, antes de consultar a tabela, é essencial padronizar os valores. Para isso, padronizamos o valor 1,8 metros calculando sua distância em relação à média. Essa distância é determinada pela fórmula:

$$\frac{1,8 - 1,7}{0,08} = 1,25. \tag{61}$$

Portanto, a distância de 1,8 metros em relação à média é 1,25 vezes o desvio padrão $(1,25 \times \sigma)$.

Com esse valor padronizado, podemos inferir que a probabilidade discutida anteriormente é equivalente à probabilidade na distribuição normal padronizada em que o valor de z está entre 0 e 1,25.

Essa abordagem nos permite encontrar rapidamente a área sob a curva da distribuição normal padronizada entre dois valores, refletindo a probabilidade de um evento ocorrer nesse intervalo. A tabela fornece diretamente a probabilidade acumulada para valores de z, simplificando cálculos complexos.

Agora, iremos nos concentrar na compreensão da utilização de duas tabelas: a tabela da **Distribuição Normal Padronizada** e a tabela da **Distribuição Normal Acumulada**.

1. **Tabela da Distribuição Normal Padronizada**: Os valores na tabela representam as probabilidades de um valor aleatório estar entre 0 (média) e um valor positivo de Z, onde Z é calculado como (valor observado - média) / desvio padrão. Cada valor na tabela corresponde à probabilidade de ocorrência de uma determinada faixa de valores. Isso é útil para calcular probabilidades exatas em uma distribuição normal, onde:

$$p = P\left(0 < z_c < Z\right) \tag{62}$$

Código do gráfico da distribuição Normal Padrão

```
library(ggplot2)

x <- seq(-3, 3, 0.01)

pdf <- dnorm(x, mean = 0, sd = 1)

df <- data.frame(x, pdf)

# Criando o grafico
ggplot(data = df, aes(x = x, y = pdf)) +
  geom_line(color = "blue") +
```

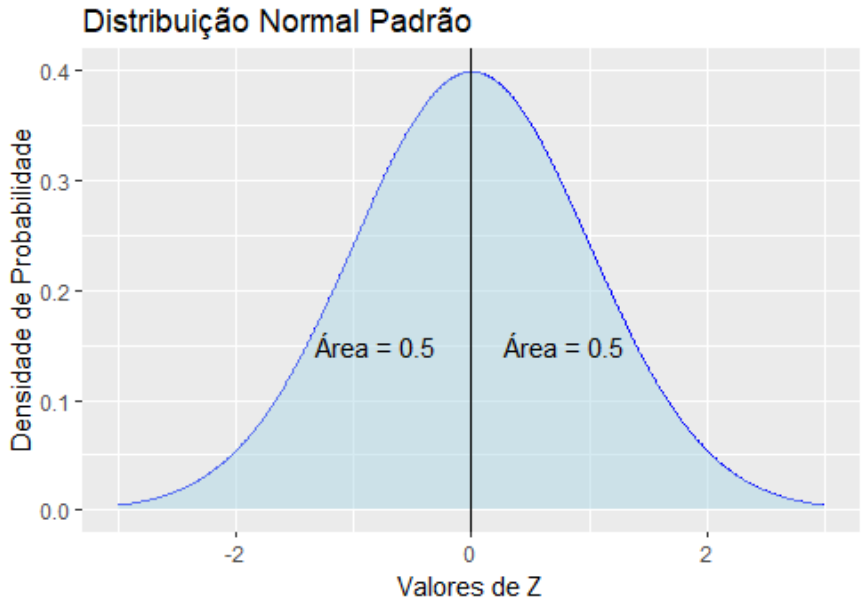


Figura 18: Demonstração da distribuição Normal Padrão no R

```
  geom_area(fill = "lightblue", alpha = 0.5, aes(x = x, y = pdf)) +
  labs(title = "Distribui  o Normal Padr o",
       x = "Valores de Z",
       y = "Densidade de Probabilidade") +
  annotate("text", x = 0.8, y = 0.15, label = paste(" rea  = 0.5"), color = "black") +
  annotate("text", x = -0.8, y = 0.15, label = paste(" rea  = 0.5"), color = "black") +
  geom_vline(xintercept = 0, color = "black")
```

2. **Tabela da Distribuição Normal Acumulada**: Essa tabela fornece os valores da função de distribuição acumulada (CDF) da distribuição normal padrão. Em outras palavras, ela dá a probabilidade acumulada de um valor aleatório ser menor ou igual a um valor específico de Z. Isso é particularmente útil quando você deseja calcular a probabilidade de um valor aleatório estar abaixo ou acima de um certo limite, o que pode ser necessário em problemas estatísticos e de probabilidade, onde:

$$p = P\left(-\infty < Z < z\right) \tag{63}$$

Código do gráfico da distribuição Normal Padrão Acumulada

```
library(ggplot2)

x <- seq(-3, 3, 0.01)

pdf <- dnorm(x, mean = 0, sd = 1)

area_total <- 1

df <- data.frame(x, pdf)

ggplot(data = df, aes(x = x, y = pdf)) +
  geom_line(color = "blue") +
  geom_area(fill = "lightblue", alpha = 0.5, aes(x = x, y = pdf)) +
  labs(title = "Distribui  o Normal Padr o Acumulada",
       x = "Valores de Z",
       y = "Densidade de Probabilidade") +
  annotate("text", x = 0, y = 0.15, label = paste(" rea  Total =", area_total), color =
     "black")
```

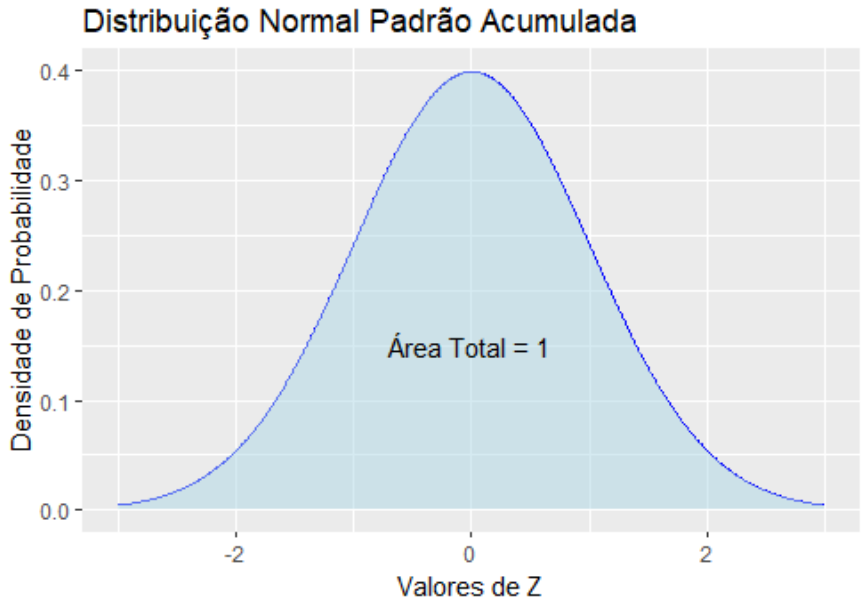


Figura 19: Demonstração da distribuição Normal Padrão Acumulada no R

Vamos ver a seguir como funciona essa diferença na prática:

Soluções:

a) $P\left(Z \leq 0,32\right)$

$$Normal : A(0,32) = 0,12552 \tag{64}$$
$$= 0,12552 + 0,5 \tag{65}$$
$$= 0,62552 \tag{66}$$

$$Acumulada : A(0,32) = 0,62552 \tag{67}$$

Código do gráfico do item A

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

z = 0.32
probabilidade = stats.norm.cdf(z) # Calculando a probabilidade P(Z <= 0,32)

x = np.linspace(-3, 3, 1000) # Criando os valores dos eixos x e y

pdf = stats.norm.pdf(x, loc=0, scale=1) # Calculando a funcao de densidade de
    probabilidade da  normal padrao

plt.figure(figsize=(10, 6))
plt.plot(x, pdf, label='Curva da Distribuicao Normal', color='blue')

plt.fill_between(x, pdf, where=(x <= z), alpha=0.2, color='SkyBlue', label='P(Z <=
    0.32)') # Preenchendo a area sob a curva

plt.scatter([z], [0], color='blue', label=f'P(Z <= {z}) = {probabilidade:.4f}', zorder
    =5) # Destacando o ponto z

plt.title('C lculo de P(Z <= 0.32) na Distribuicao Normal Padrao')
```

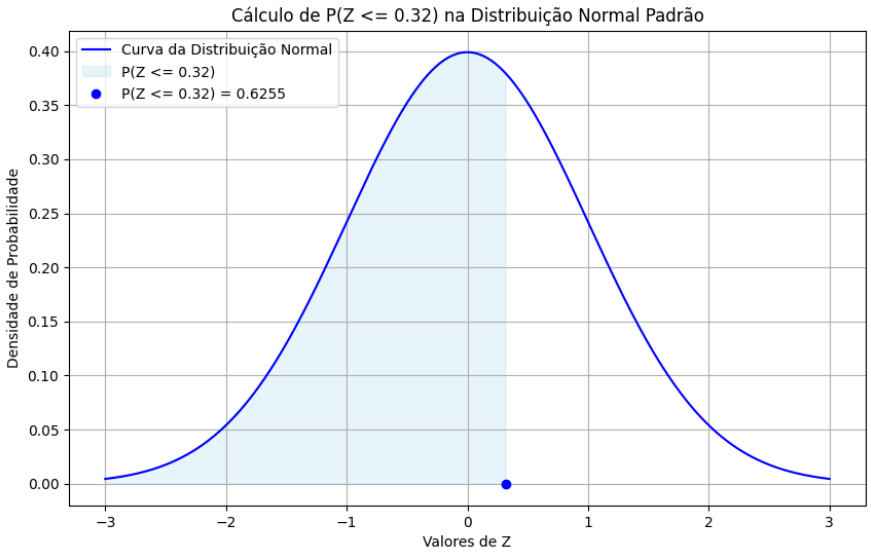


Figura 20: Demonstração do item A usando Python

```
plt.xlabel('Valores de Z')
plt.ylabel('Densidade de Probabilidade')
plt.legend(loc='upper left')
plt.grid()
plt.show()
```

b) $P\left(0 < Z \leq 1,71\right)$

$$Normal : A(1,71) = 0,45637 \tag{68}$$

$$\begin{aligned} Acumulada : A(1,71) &= 0,9564 - 0,5 && (69)\\ &= 0,4664 && (70)\end{aligned}$$

c) $P\left(-1,32 < Z < 0\right)$

$$Normal : A(-1,32) = A(1,32) = 0,40658 \tag{71}$$

$$\begin{aligned} Acumulada : A(-1,32) &<=> A(1,32) && (72)\\ &= 0,9066 - 0,5 && (73)\\ &= 0,4066 && (74)\end{aligned}$$

d) $P\left(1,32 < Z < 1,79\right)$

$$\begin{aligned} Normal : &A(1,79) - A(1,32) && (75)\\ &= 0,46327 - 0,40658 && (76)\\ &= 0,05669 && (77)\end{aligned}$$

$$\begin{aligned} Acumulada : &A(1,79) - A(1,32) && (78)\\ &= 0,9633 - 0,9066 && (79)\\ &= 0,0576 && (80)\end{aligned}$$

Código do gráfico do item D

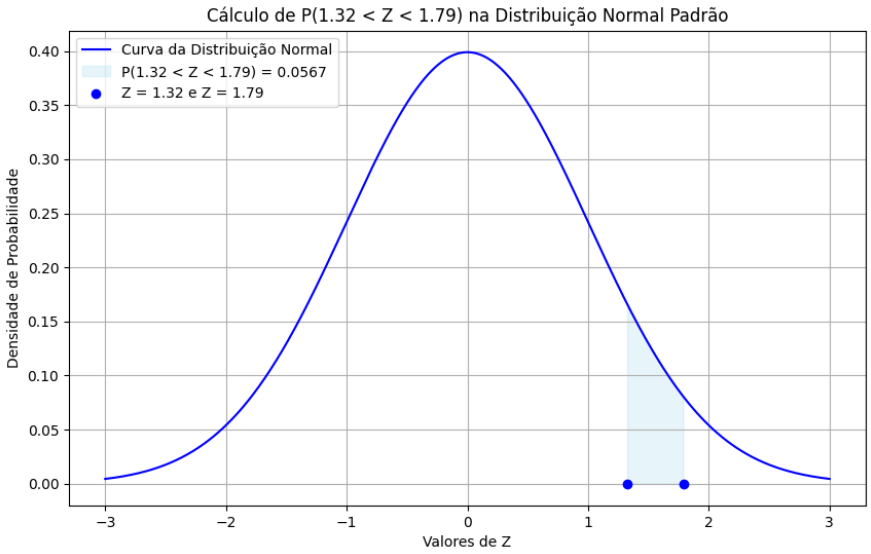


Figura 21: Demonstração do item D usando Python

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

# Valores de Z
z1 = 1.32
z2 = 1.79

# Calculando as probabilidades P(1,32 < Z < 1,79)
prob1 = stats.norm.cdf(z1)
prob2 = stats.norm.cdf(z2)
prob = prob2 - prob1

# Criando um array de valores para o eixo x
x = np.linspace(-3, 3, 1000)

pdf = stats.norm.pdf(x, loc=0, scale=1)

plt.figure(figsize=(10, 6))
plt.plot(x, pdf, label='Curva da Distribuicao Normal', color='blue')

# Preenchendo a area sob a curva entre os valores
plt.fill_between(x, pdf, where=(x >= z1) & (x <= z2), alpha=0.2, color='SkyBlue', label
    =f'P({z1} < Z < {z2}) = {probabilidade:.4f}')

# Destacando os pontos z1 e z2
plt.scatter([z1, z2], [0, 0], color='blue', zorder=5, label=f'Z = {z1} e Z = {z2}')

plt.title(f'C lculo de P({z1} < Z < {z2}) na Distribuicao Normal Padrao')
plt.xlabel('Valores de Z')
plt.ylabel('Densidade de Probabilidade')
plt.legend(loc='upper left')
plt.grid()
plt.show()
```

e) $P\left(-2,3 < Z \leq -1,49\right)$

$$Normal: A(2,3) - A(1,49) \tag{81}$$
$$= 0,48928 - 0,43189 \tag{82}$$
$$= 0,0574 \tag{83}$$

$$Acumulada : A(2,3) - A(1,49) \tag{84}$$
$$= 0,9893 - 0,9319 \tag{85}$$
$$= 0,0574 \tag{86}$$

f) $P\left(Z \geq 1,5\right)$

$$Normal : 0,5 - A(1,5) \tag{87}$$
$$= 0,5 - 0,43319 \tag{88}$$
$$= 0,06681 \tag{89}$$

$$Acumulada : 1 - A(1,5) \tag{90}$$
$$= 1 - 0,9332 \tag{91}$$
$$= 0,06681 \tag{92}$$

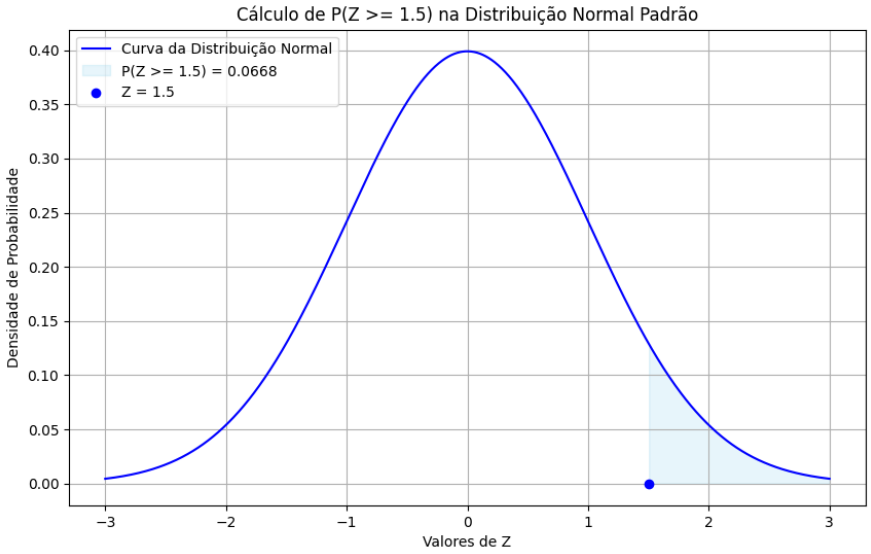


Figura 22: Demonstração do item F usando Python

Código do gráfico do item F

```
import numpy as np
import matplotlib.pyplot as plt
import scipy.stats as stats

z = 1.5

# Calculando a probabilidade
probabilidade = 1 - stats.norm.cdf(z)

x = np.linspace(-3, 3, 1000)

pdf = stats.norm.pdf(x, loc=0, scale=1)

plt.figure(figsize=(10, 6))

plt.plot(x, pdf, label='Curva da Distribui  o Normal', color='blue')

plt.fill_between(x, pdf, where=(x >= z), alpha=0.2, color='SkyBlue', label=f'P(Z >= {z
    }) = {probabilidade:.4f}')

plt.scatter([z], [0], color='blue', zorder=5, label=f'Z = {z}')
```

```
plt.title(f'Calculo de P(Z >= {z}) na Distribuicao Normal Padrao')
plt.xlabel('Valores de Z')
plt.ylabel('Densidade de Probabilidade')
plt.legend(loc='upper left')
plt.grid()
plt.show()
```

g) $P\left(Z \leq -1,3\right)$

$$Normal: 0,5 - A(1,3) \tag{93}$$
$$= 0,5 - 0,40320 \tag{94}$$
$$= 0,0968 \tag{95}$$

$$Acumulada: 1 - A(1,3) \tag{96}$$
$$= 1 - 0,9032 \tag{97}$$
$$= 0,0968 \tag{98}$$

h) $P\left(-1,5 \leq Z \leq 1,5\right)$

$$Normal: 2 \times (1,5) <=> A(-1,5) + A(1,5) \tag{99}$$
$$= 2 \times 0,43319 \tag{100}$$
$$= 0,8664 \tag{101}$$

$$Acumulada: [A(1,5) - 0,5] \times 2 \tag{102}$$
$$= [0,9332 - 0,5] \times 2 \tag{103}$$
$$= 0,4332 \times 2 \tag{104}$$
$$= 0,8664 \tag{105}$$

C Referências

[1] Sobrenome, A. B., Sobrenome, C. D., & Sobrenome, E. F. (Ano). Título do Artigo. *Revista de Estatística*, *Volume*(Número), Páginas.

[2] Morettin, P. A.; Bussab, W. O. Estatística básica. São Paulo: Editora Saraiva, 2017. E-book. ISBN 9788547220228. Disponível em: https://app.minhabiblioteca.com.br//books/9788547220228/. Acesso em: 25 jul. 2023.

[3] Gupta, C B.; Guttman, Irwin. Estatística e Probabilidade com Aplicações para Engenheiros e Cientistas. [Digite o Local da Editora]: Grupo GEN, 2016. E-book. ISBN 9788521632931. Disponível em: https://app.minhabiblioteca.com.br//books/9788521632931/. Acesso em: 27 jul. 2023.

[4] Silva, Ermes Medeiros, D. et al. Estatística, 5ª edição. Disponível em: Minha Biblioteca, Grupo GEN, 2018.

[5] Bolfarine, Heleno. Elementos de amostragem. Disponível em: Minha Biblioteca, Editora Blucher, 2005.

[6] Werkema, Cristina. Inferência Estatística - Como Estabelecer Conclusões com Confiança no Giro do PDCA e DMAIC. Disponível em: Minha Biblioteca, Grupo GEN, 2014.

[7] Oliveira, Francisco Estevam Martins de. Estatistica e Probabilidade - Exercicios Resolvidos e Propostos, 3ª edição. São Paulo: Grupo GEN, 2017. E-book. ISBN 9788521633846. Disponível em: https://app.minhabiblioteca.com.br//books/9788521633846/. Acesso em: 18 ago. 2023.

[8] SILVA, Ermes Medeiros da; SILVA, Elio Medeiros da; GONÇALVES, Valter; MUROLO, Afrânio C. Estatística, 5ª edição. [Digite o Local da Editora]: Grupo GEN, 2018. E-book. ISBN 9788597014273. Disponível em: https://app.minhabiblioteca.com.br//books/9788597014273/. Acesso em: 01 ago. 2023.

[9] CAMPOS, Marcilia A.; RÊGO, Leandro C.; MENDONÇA, André Feitoza de. Métodos Probabilísticos e Estatísticos com Aplicações em Engenharias e Ciências Exatas. [Digite o Local da Editora]: Grupo GEN, 2016. E-book. ISBN 9788521633143. Disponível em: https://app.minhabiblioteca.com.br//books/9788521633143/. Acesso em: 27 ago. 2023.

VER DEPOIS

BARBETTA, Pedro Alberto. Estatística: para cursos de engenharia e informática. São Paulo: Atlas, 2004.

GUIMARÃES, Inácio Andruski. Notas de Aulas. Curitiba, 2005.

MORETTIN, Pedro A. e BUSSAB, Wilton de O. Estatística Básica. São Paulo: Saraiva, 2003.

MARTINS, Gilberto de Andrade. Estatística Geral e Aplicada. São Paulo: Editora Atlas, 2002.