Estatística I

Prof. Fernando de Souza Bastos
fernando.bastos@ufv.br

Departamento de Estatística
Universidade Federal de Viçosa
Campus UFV - Viçosa

Sumário

Introdução

Por que não confiar só na média?

Imagine que você está com muita fome e abre o aplicativo de delivery. A informação aparece assim:

"O tempo médio de entrega é de 30 minutos."

Você pensa:

"Perfeito! Em meia horinha eu tô comendo!"

Por que não confiar só na média?

Mas... será que é tão simples assim?

Pense nos dados que o aplicativo usa para calcular essa média:

- Algumas entregas foram muito rápidas (15, 20 minutos).
- Outras demoraram bastante (60, 70, até 90 minutos).

A **média** de 30 minutos parece bonita... mas esconde toda essa variabilidade!

A verdade por trás da média

Se o aplicativo dissesse:

"Com 95% de confiança, seu pedido chegará entre 20 minutos e 1 hora e 10 minutos."

Agora sim, você entende o jogo!

Isso significa que:

- É possível que chegue rápido (20 min).
- Mas também existe uma chance real de demorar mais de uma hora.

Percebe a diferença?

"A média é uma informação solitária. O intervalo de confiança é uma informação honesta."

"A média é uma informação solitária. O intervalo de confiança é uma informação honesta."

Confiar só na média é como dirigir olhando apenas pelo retrovisor... Parece informação, mas não te mostra o que vem pela frente.

Dois tipos de estimativas

Estimativa Pontual
É quando usamos um único número, calculado a partir da amostra, para estimar um parâmetro populacional.

Exemplos:

- Média amostral ($\bar{x}$) para estimar a média populacional (μ).
- Proporção amostral ($\hat{p}$) para estimar a proporção populacional (p).

Limitação: Fornece apenas um valor. Não diz nada sobre a incerteza ou confiabilidade desse valor.

Dois tipos de estimativas

Estimativa Intervalar (Intervalo de Confiança)
Em vez de fornecer um único número, fornece um **intervalo de valores plausíveis** para o parâmetro populacional.

Exemplo:

"Com 95% de confiança, a média populacional está entre 25 e 35."

Vantagem: Expressa não só a estimativa, mas também a incerteza associada a ela.

Por que precisamos de um intervalo de confiança?

Todo estimador (como a média amostral) é uma **variável aleatória**.

- Se coletarmos outra amostra, vamos obter outro valor.
- A cada amostra possível, temos uma média diferente.

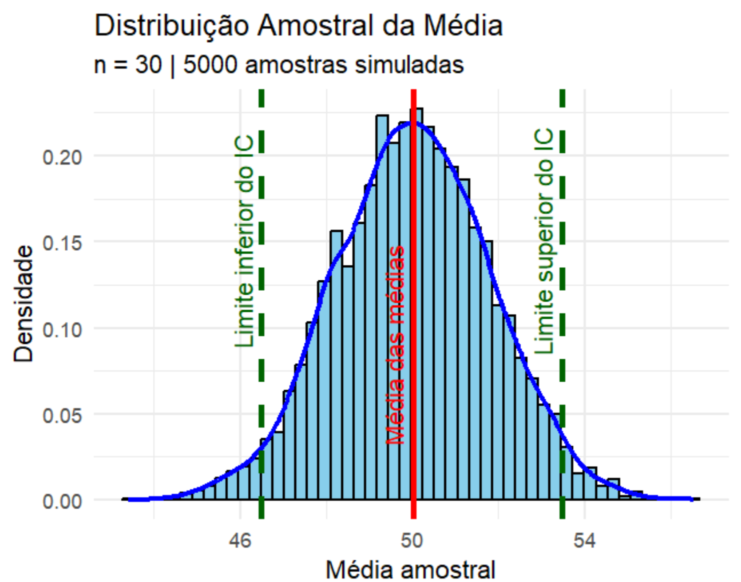
Por isso, o estimador possui uma **distribuição de probabilidade**, chamada de **distribuição amostral**.

E é exatamente a partir dessa distribuição que construímos o intervalo de confiança.

O intervalo de confiança nos permite afirmar algo do tipo:

"Se eu repetir esse processo muitas vezes, 95% dos intervalos conterão o verdadeiro valor do parâmetro."

Visualizando a incerteza



Intervalos de Confiança para a Média: σ conhecido

Suposições Necessárias

Para construirmos um intervalo de confiança para a média (com σ conhecido), precisamos garantir:

- A amostra é uma **amostra aleatória simples (AAS)**.
- O desvio padrão da população (σ) é conhecido.
- E uma das seguintes condições:
  - A população tem distribuição normal;
  - ou o tamanho da amostra é suficientemente grande ($n > 30$).

Erro Amostral: Sempre Existe!

Ao coletar uma amostra, a média amostral ($\bar{X}$) dificilmente será exatamente igual à média populacional (μ).

Essa diferença é chamada de erro amostral:

$$e = \bar{X} - \mu \quad \Leftrightarrow \quad \bar{X} = \mu + e$$

Sabemos que a média amostral segue uma distribuição:

$$\bar{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Ou seja, as médias amostrais **variam** de amostra para amostra!

O que é a Margem de Erro?

Se padronizarmos a média amostral, obtemos:

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

Z responde: Quantos desvios padrão minha média amostral está distante da média populacional.

A **margem de erro** (e) representa o erro máximo aceitável, dentro de um grau de confiança (γ):

$$e = z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Onde $z_{\gamma/2}$ é o **valor crítico da normal padrão**.

Construindo o Intervalo de Confiança

Raciocínio
Queremos capturar o valor de μ dentro de um intervalo simétrico ao redor da média amostral $\bar{x}$.

$$P\left(-z_{\gamma/2} < Z < z_{\gamma/2}\right) = \gamma$$

Substituindo Z pela padronização da média:

$$P\left(\bar{x} - z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = \gamma$$

Pronto! Este é o intervalo de confiança para μ com confiança γ.

Valor Crítico $z_{\gamma/2}$

O valor $z_{\gamma/2}$ é o ponto da distribuição normal padrão que deixa uma área de $\gamma/2$ em cada cauda.

Por exemplo, para $\gamma = 0,95$:

- A área central é 95%.
- Sobra 5% para as caudas $\rightarrow$ 2,5% em cada lado.
- Buscamos na tabela da normal padrão a área acumulada até 0,975.
- Resultado: $z_{0,025} = 1,96$.

A área central corresponde ao nível de confiança γ.

Fórmula do Intervalo de Confiança

O intervalo de confiança para μ, com nível de confiança γ, é dado por:

$$\left[\bar{X} - z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}}\,;\ \bar{X} + z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Interpretação: Uma faixa de valores plausíveis para a média populacional, considerando a variabilidade natural das amostras.

Passos para construir o Intervalo de Confiança

1. Verificar as suposições:
   - AAS
   - σ conhecido
   - População normal ou $n > 30$
2. Escolher o nível de confiança γ e determinar $z_{\gamma/2}$.
3. Calcular a margem de erro:

$$e = z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}}$$

4. Construir o intervalo:

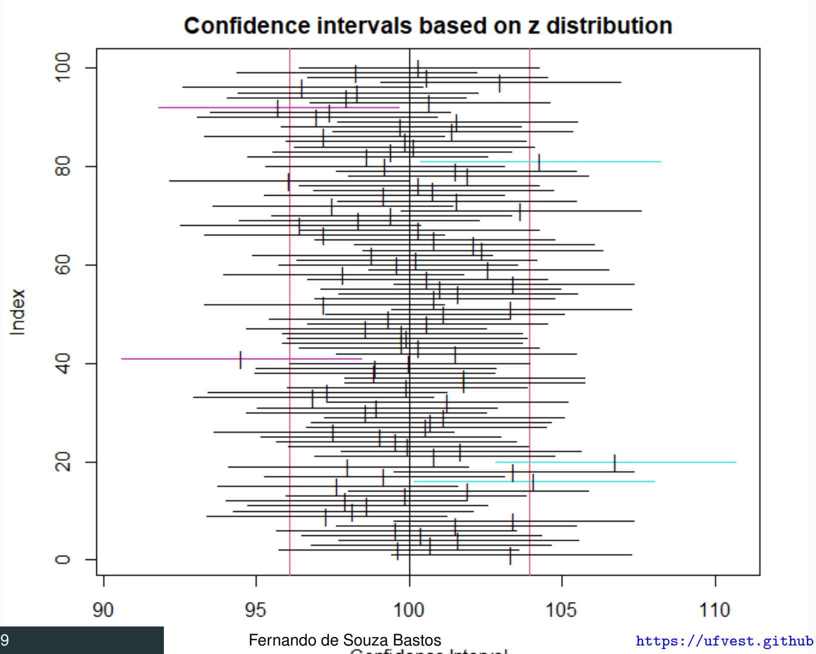
$$\bar{X} \pm e$$

Interpretação do Intervalo de Confiança

Atenção: O que significa γ?
- O **parâmetro** (μ) é fixo. - O que varia é o **intervalo**, porque ele depende da amostra.

Se construirmos 100 intervalos de confiança de 95%, usando 100 amostras diferentes, **esperamos que 95 deles contenham μ**, e 5 não contenham.

Importante: *Não dizemos que "a probabilidade de μ estar no intervalo é 95%". μ não é aleatório. O que é aleatório é o intervalo.*



Confidence intervals based on z distribution
Index
0
20
40
60
80
100
90
95
100
105
110
Confidence Interval

Exemplo

Uma empresa de marketing deseja estimar quanto tempo, em média, as pessoas passam no WhatsApp por semana.

Eles selecionaram, aleatoriamente, uma amostra de 25 pessoas. O tempo médio de uso foi de **22,4 horas por semana** – isso mesmo, quase um emprego de meio período só no zap!

Com base em estudos anteriores, eles assumem que o desvio padrão populacional é $\sigma = 5,2$ horas e que o tempo de uso tem distribuição normal.

Construa um intervalo de confiança de 95% para a média populacional μ.

Exemplo

Solução

$$\bar{X} = 22,4, \quad n = 25, \quad \sigma = 5,2$$

$$z_{0,025} = 1,96$$

$$e = 1,96 \cdot \frac{5,2}{\sqrt{25}} = 2,038$$

$$IC = [22,4 - 2,038\ ;\ 22,4 + 2,038]$$

$$(20,362 \leq \mu \leq 24,438)$$

Interpretação: Com 95% de confiança, o tempo médio que as pessoas passam no WhatsApp por semana está entre aproximadamente **20,4 e 24,4 horas**.

Intervalos de Confiança para a Média: σ desconhecido

Quando σ é Desconhecido

Estimativa da Variância Amostral

Na maioria das situações práticas, não sabemos o verdadeiro valor do desvio padrão populacional (σ). Se σ é desconhecido, ele precisa ser estimado a partir da amostra.

Sendo $(X_1, \ldots, X_n)$ uma amostra aleatória de uma variável aleatória $X \sim \mathrm{N}(\mu, \sigma^2)$, o estimador da variância populacional é a variância amostral:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Esse estimador é **não viciado** e **consistente** para σ^2.

A Distribuição t de Student

Quando substituímos σ pela estimativa S, a variável padronizada passa a ser:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Essa variável não segue mais a distribuição normal. Ela segue uma **distribuição t de Student** com $n-1$ **graus de liberdade**:

$$T \sim t(n-1)$$

A distribuição t é mais "espalhada"que a normal, pois leva em conta a incerteza adicional de estimar σ.

Valores Críticos da Distribuição t

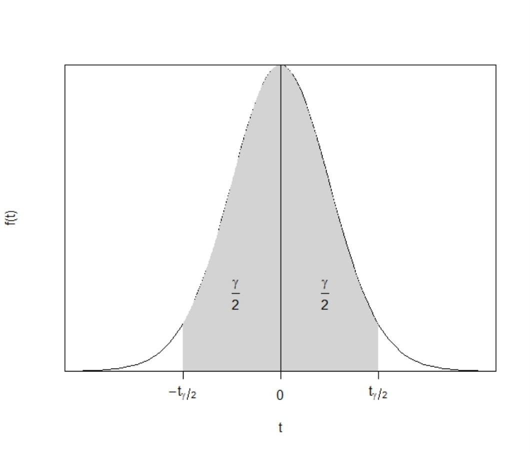
Para determinar o valor crítico $t_{\gamma/2}$, usamos:

- O nível de confiança (γ);
- O número de graus de liberdade ($gl = n - 1$).

Exemplo:

- $\gamma = 95\%$ e $n = 7 \Rightarrow gl = 6$
- Buscamos na tabela t a linha dos $gl = 6$ e a coluna correspondente a 5% (2,5% em cada cauda)
- Encontramos: $t_{0,025} = 2,447$

Distribuição *t* de Student



Distribuição t de Student: mais "espalhada"que a normal, especialmente para amostras pequenas.

Fórmula do Intervalo de Confiança (σ desconhecido)

O intervalo de confiança para μ, quando σ é desconhecido, é:

$$\left[\bar{X} - t_{\gamma/2} \cdot \frac{S}{\sqrt{n}}\,;\ \bar{X} + t_{\gamma/2} \cdot \frac{S}{\sqrt{n}}\right]$$

Interpretação: Intervalo de valores plausíveis para μ, levando em conta a incerteza tanto da variabilidade amostral quanto da estimativa do desvio padrão.

Procedimentos para a construção de intervalos de confiança

1. Verifique se as suposições necessárias estão satisfeitas
    - Temos uma AAS
    - Temos uma estimativa de s
    - A população tem distribuição normal ou $n > 30$
2. Determine o nível de confiança γ, e encontre o valor crítico $t_{\gamma/2}$
3. Calcule a margem de erro $e = t_{\gamma/2} \cdot (s/\sqrt{n})$
4. Calcule IC(μ, γ)

Exemplo

Uma turma de Estatística, conhecida por conversar bastante durante as aulas, fez uma prova com nota máxima 100.

Foram selecionadas, aleatoriamente, as notas de **15 alunos**. A média da amostra foi **62,4** e o desvio padrão amostral foi **18,5**.

Segundo relatos, há alguns poucos alunos que prestam muita atenção e puxam as notas para cima, enquanto o resto... conversa.

Construa um intervalo de confiança de 95% para a média verdadeira das notas dessa turma.

Exemplo

Solução

$$\bar{X} = 62,4, \quad S = 18,5, \quad n = 15$$

$$gl = n - 1 = 14$$

$$t_{0,025,14} = 2,145$$

$$e = 2,145 \cdot \frac{18,5}{\sqrt{15}} = 10,24$$

$$IC = [62,4 - 10,24 \; ; \; 62,4 + 10,24]$$

$$(52,16 \leq \mu \leq 72,64)$$

Interpretação: Com 95% de confiança, a média real das notas da turma está entre aproximadamente **52,2 e 72,6**.

Intervalo de Confiança para Proporção

Intervalo de Confiança para Proporção

Distribuição da proporção amostral

Seja uma variável aleatória binária com:

- Sucesso: probabilidade p;
- Fracasso: probabilidade $1 - p$.

A proporção amostral é:

$$\hat{p} = \frac{x}{n} = \frac{\text{número de sucessos}}{\text{tamanho da amostra}}$$

Intervalo de Confiança para Proporção

Propriedades da distribuição amostral de $\hat{p}$:

- $\hat{p}$ é um estimador não-viesado de p.
- Se n é suficientemente grande, $\hat{p}$ segue aproximadamente uma distribuição normal:
$$\hat{p} \sim \mathrm{N}\left(p, \frac{p(1-p)}{n}\right)$$
- Esperança: $\mathrm{E}(\hat{p}) = p$;
- Variância: $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$.

Distribuição amostral da proporção $\hat{p}$

Se a condição de normalidade é satisfeita, podemos padronizar a proporção amostral:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathrm{N}(0, 1)$$

Isso permite construir intervalos de confiança para p usando a distribuição normal padrão.

Fórmula do Intervalo de Confiança

O intervalo de confiança para a proporção populacional p, com nível de confiança γ, é dado por:

$$\text{IC}(p, \gamma) = \left[\hat{p} - z_{\gamma/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\ ;\ \hat{p} + z_{\gamma/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

Interpretação: Fornece uma faixa de valores plausíveis para a verdadeira proporção da população.

Passos para Construção do Intervalo de Confiança

1. Verifique as suposições:
   - A amostra é uma AAS;
   - As condições da binomial são satisfeitas:
     - Tentativas independentes;
     - Dois resultados possíveis (sucesso/fracasso);
     - Probabilidade de sucesso constante.
   - Condição para aproximação normal:

$$n\hat{p} \geq 5 \text{ e } n(1-\hat{p}) \geq 5$$

2. Defina o nível de confiança γ e encontre $z_{\gamma/2}$.
3. Calcule a margem de erro e o intervalo:

$$e = z_{\gamma/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow \hat{p} \pm e$$

Exemplo - Proporção

Durante uma aula de Estatística, o professor percebe que a turma está mais no WhatsApp do que prestando atenção. Para investigar, ele sorteia uma amostra de 40 alunos e verifica que 28 estavam usando o celular na hora da explicação.

A proporção amostral é:

$$\hat{p} = \frac{28}{40} = 0,70$$

Construa um intervalo de confiança de 95% para a proporção de alunos que usam o celular durante a aula.

Exemplo - Proporção

Solução

$$\hat{p} = 0,70, \;\; n = 40, \;\; z_{0,025} = 1,96$$

$$e = 1,96 \cdot \sqrt{\frac{0,70 \cdot 0,30}{40}} = 0,144$$

$$IC = [0,70 - 0,144 \; ; \; 0,70 + 0,144] = [0,556 \; ; \; 0,844]$$

Interpretação: Com 95% de confiança, entre 55,6% e 84,4% dos alunos usam o celular durante a aula.

Determinação do Tamanho Amostral (σ conhecido)

Determinação do Tamanho Amostral

Nosso objetivo é coletar dados para estimar a **média populacional** μ.

A grande pergunta é:

Quantos elementos (pessoas, objetos, itens...) devemos amostrar?

Sabemos que, de forma geral, $n > 30$ costuma ser suficiente para muitas situações.

Mas... podemos ser mais inteligentes que isso!

ula que permite calcular exatamente o ta-

Fórmula para o Tamanho Amostral

Derivação
Sabemos que a **margem de erro** para a média é dada por:

$$e = z_{\gamma/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Isolando *n*:

$$n = \left(\frac{z_{\gamma/2} \cdot \sigma}{e}\right)^2$$

Esta é a fórmula que usamos para calcular o tamanho amostral necessário.

Interpretando a Fórmula

O que influencia o tamanho amostral?

$$n = \left(\frac{z_{\gamma/2} \cdot \sigma}{e}\right)^2$$

O tamanho amostral depende de:

- O **nível de confiança** desejado ($z_{\gamma/2}$);
- O **erro máximo admissível** (e) – *precisão desejada*;
- O **desvio padrão populacional** (σ).

Não depende do tamanho da população (se for muito grande).

Importante: Sempre arredondamos n para o **próximo superior**

Exemplo - Determinação do Tamanho Amostral

Queremos estimar o tempo médio que alunos passam no WhatsApp por semana.

Desejamos um intervalo de confiança de 95%, com **margem de erro de no máximo 2 horas**. Estudos anteriores indicam que o desvio padrão é **5,2 horas**.

Qual deve ser o tamanho da amostra?

Exemplo - Determinação do Tamanho Amostral

Solução

$$\sigma = 5,2, \;\; e = 2, \;\; z_{0,025} = 1,96$$

$$n = \left(\frac{1,96 \cdot 5,2}{2}\right)^2 = (5,096)^2 = 25,97$$

$$n = 26 \text{ (arredondado para o inteiro superior)}$$

Conclusão: Precisamos de uma amostra com pelo menos 26 alunos.

Determinação do Tamanho Amostral (σ desconhecido)

Determinação do Tamanho Amostral (σ desconhecido)

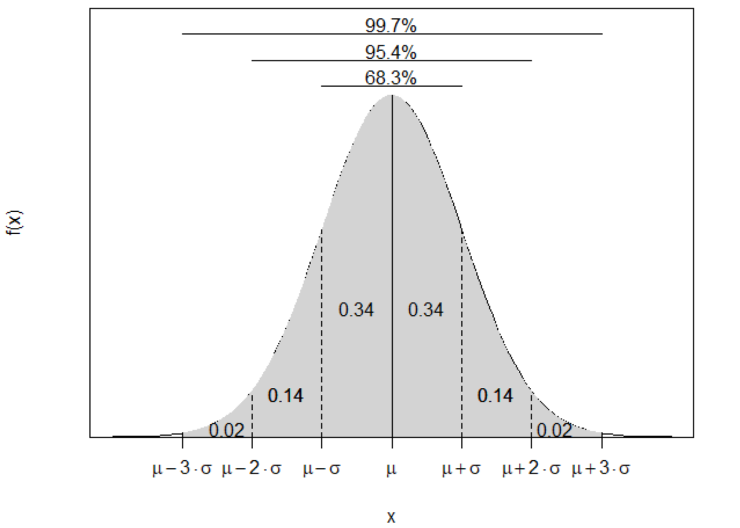
Se σ for desconhecido, temos algumas estratégias práticas:

- Utilizar uma **estimativa de** σ baseada em estudos anteriores ou literatura;
- Realizar uma **amostra piloto** e utilizar o desvio padrão amostral (s) como aproximação de σ;
- Usar a **regra empírica da amplitude**, válida para dados aproximadamente normais:

$$\sigma \approx \frac{\text{amplitude}}{4}$$

  onde amplitude é $\max - \min$ de valores típicos (sem outliers extremos).

Regra Empírica para Estimar σ



Aplicando a Regra Empírica

Definimos como **valores usuais** aqueles que não são extremos.

Para uma distribuição aproximadamente normal, sabemos que:

$$4\sigma \approx \text{amplitude} = \max - \min$$

Portanto, uma estimativa prática para σ é:

$$\sigma \approx \frac{\max - \min}{4}$$

Essa técnica é muito útil quando não temos dados prévios ou quando queremos uma estimativa rápida e prática.

Exemplo - Determinação do Tamanho Amostral

Uma clínica deseja estimar o **tempo médio de espera dos pacientes para atendimento**.

Eles querem garantir que a média amostral esteja, no máximo, **5 minutos distante da média real**, com **90% de confiança**.

Sabe-se que os tempos de espera costumam variar entre **10 e 50 minutos**.

Pergunta: Qual deve ser o tamanho da amostra?

Exemplo - Determinação do Tamanho Amostral

Estimando o desvio padrão usando a regra empírica:

$$\sigma = \frac{50 - 10}{4} = 10$$

Para 90% de confiança, $z_{0,05} = 1,645$.

Aplicando a fórmula:

$$n = \left(\frac{1,645 \cdot 10}{5}\right)^2 = (3,29)^2 = 10,82$$

$$n = 11 \text{ (arredondado para o inteiro superior)}$$

Conclusão: É necessário amostrar pelo menos 11 pacientes para estimar a média do tempo de espera com a precisão desejada.

Referências

- Bastos, Fernando de Souza (2025). ***Apostila Interativa.*** Disponível online: https://ufvest.shinyapps.io/ApostilaInterativa/.
- Ferreira, Eric Batista e Marcelo Silva de Oliveira (2020). ***Introdução à Estatística com R.*** Editora Universidade Federal de Alfenas. URL: https://www.unifal-mg.edu.br/bibliotecas/wp-content/uploads/sites/125/2021/12/32-EBR_Unifal.pdf.
- Meyer, Paul L (1982). ***Probabilidade: aplicações à estatística.*** Livros Técnicos e Científicos.

Referências ii

- Montgomery, D. C. e G. C Runger (2016). ***Estatística Aplicada E Probabilidade Para Engenheiros.*** 6ª ed. São Paulo: Grupo Gen-LTC.
- Morettin, P.A. e W.O Bussab (2023). ***Estatística básica.*** 10ª ed. São Paulo: Editora Saraiva.
- Peternelli, Luiz Alexandre (s.d.). ***Apostila (EST 106).*** Formato slide Disponível no PVANet - Moodle.