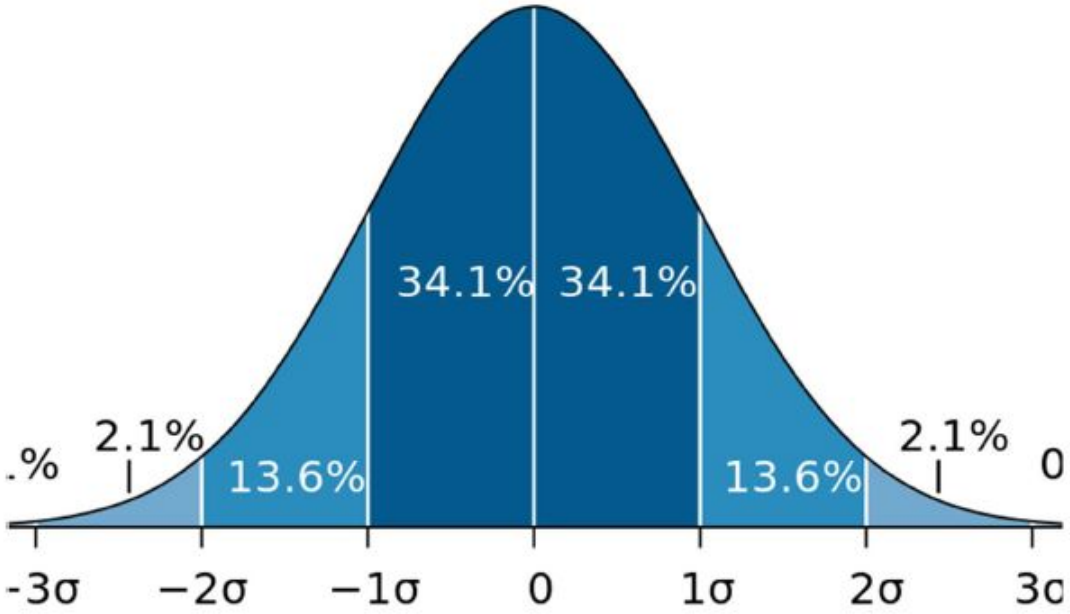
# DATA SCIENCE

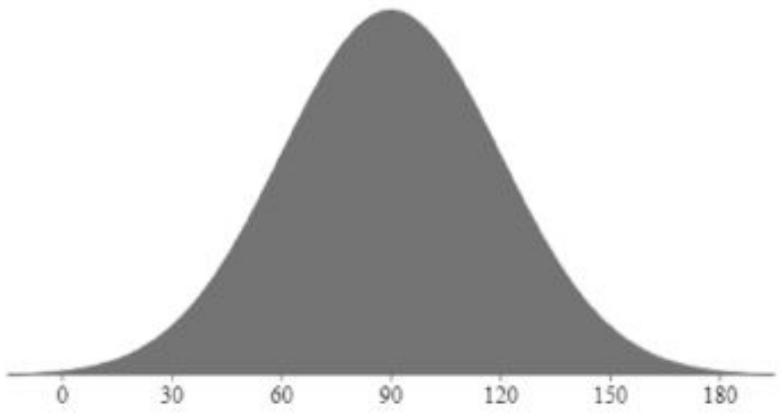
## AULA 4 - ESTATÍSTICA

PROFª. ANA CAROLINA B. ALBERTON

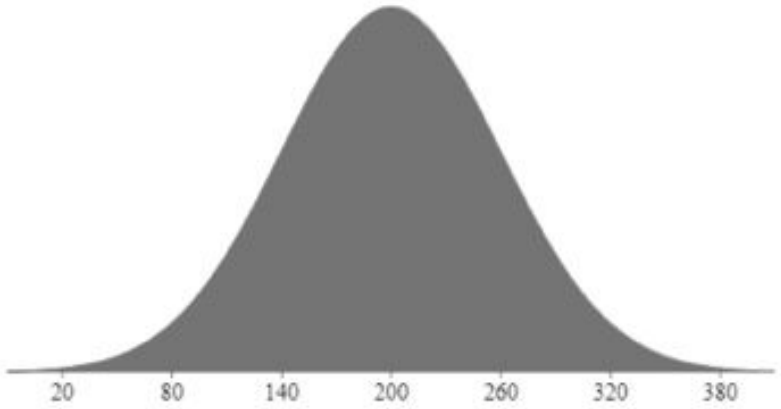
# AULA PASSADA - DISTRIBUIÇÃO NORMAL



# AULA PASSADA - DISTRIBUIÇÃO NORMAL

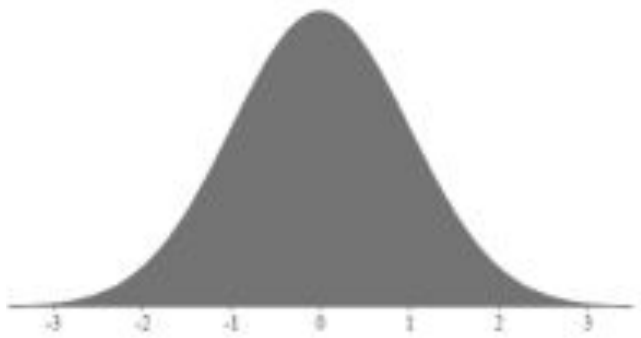



$\mu = 90$
$\sigma = 30$




$\mu = 200$
$\sigma = 60$

# AULA PASSADA - DISTRIBUIÇÃO NORMAL PADRÃO (Z)




- Distribuição de Referência para outras Distribuições Normais
- Média Zero e Desvio Padrão =1

# AULA PASSADA - DISTRIBUIÇÃO NORMAL PADRÃO (Z) E PROBABILIDADE

Para encontrar a probabilidade, utiliza-se a Tabela Z para facilitar.

Com a fórmula abaixo você transforma a probabilidade da sua distribuição na probabilidade da tabela Z

Então você olha a probabilidade na tabela!

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

onde X = seu valor

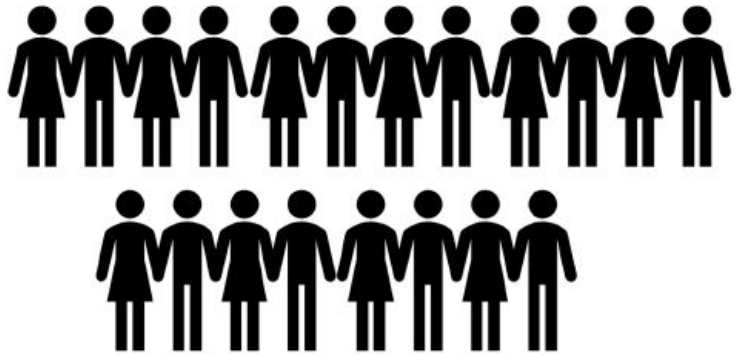
μ = média

σ = desvio padrão

# INTERVALOS DE CONFIANÇA

Uso de amostras:

- Possui um 'custo': erro padrão / nível de confiança
- Possui Riscos: dados ruins, enviesamento



POPULAÇÃO

AMOSTRA

# INTERVALOS DE CONFIANÇA

Como estamos utilizando amostra..

....devemos esperar variação.

- A primeira amostra pode variar com relação a segunda...
- A segunda com relação a terceira etc..

Mas devemos poder "medir" o quanto pode ser esta variação

# INTERVALOS DE CONFIANÇA

- Intervalo de Confiança: O parâmetro mais ou menos a margem de erro estimada
- Parâmetro: valor a ser estimado
- Margem de erro: variabilidade, para mais ou para menos
- Nível de Confiança: de 80 a 99%
- Tamanho da Amostra (n)

IN

# Nível de Confiança: z*

De

| Percentual de Confiança | Valor de Z* |
|---|---|
| 80 | 1,28 |
| 90 | 1,64 |
| 95 | 1,96 |
| 98 | 2,33 |
| 99 | 2,58 |

# Em números...

- Entre 63 e 67% dos entrevistados pretendem votar em Maria, com um nível de confiança de 95%

- Parâmetro: Intenção de Voto (proporção)
- Nível de Confiança: 95%
- Intervalo de Confiança: Entre 63 e 67%
- Erro padrão: 1,96
- Entrevistados (n): 1000
- Margem de Erro: +-2%

- Entre 63 e 67% dos entrevistados pretendem votar em Maria, com um nível de confiança de 95%

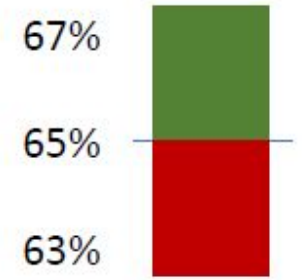
✓ Parâmetro: Intenção de Voto (proporção)

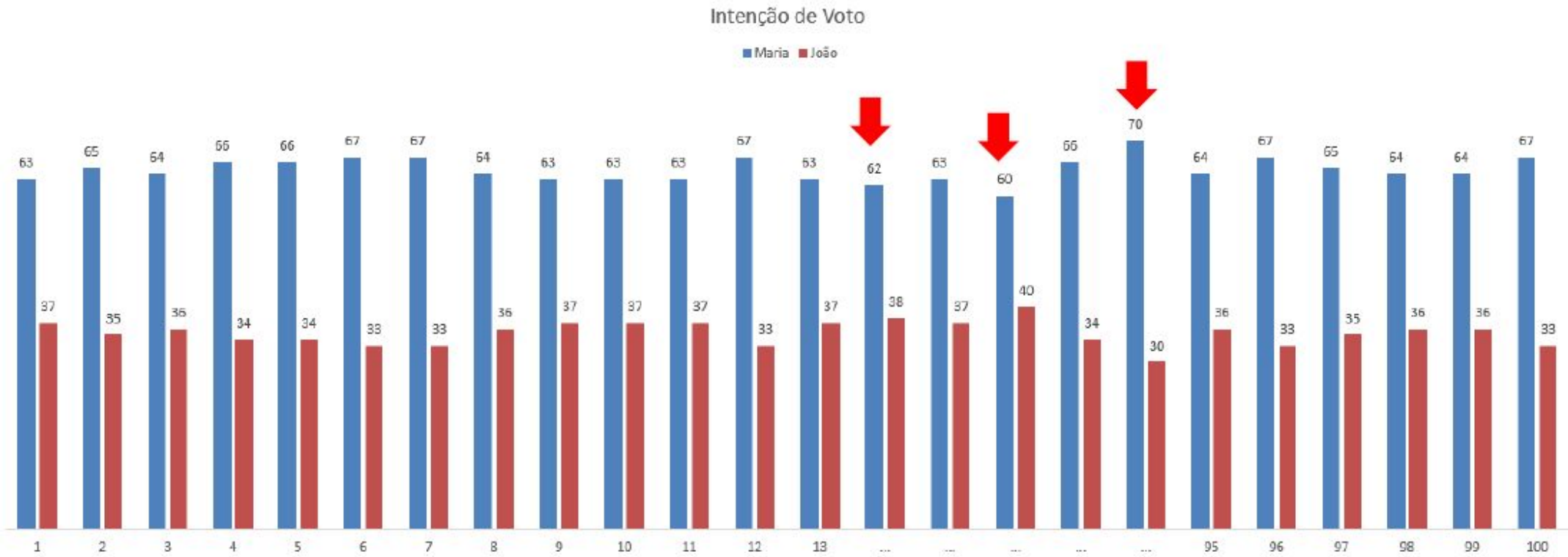
✓ Nível de Confiança: 95%

✓ Entrevistados (n): 1000

✓ Margem de Erro: +-2%

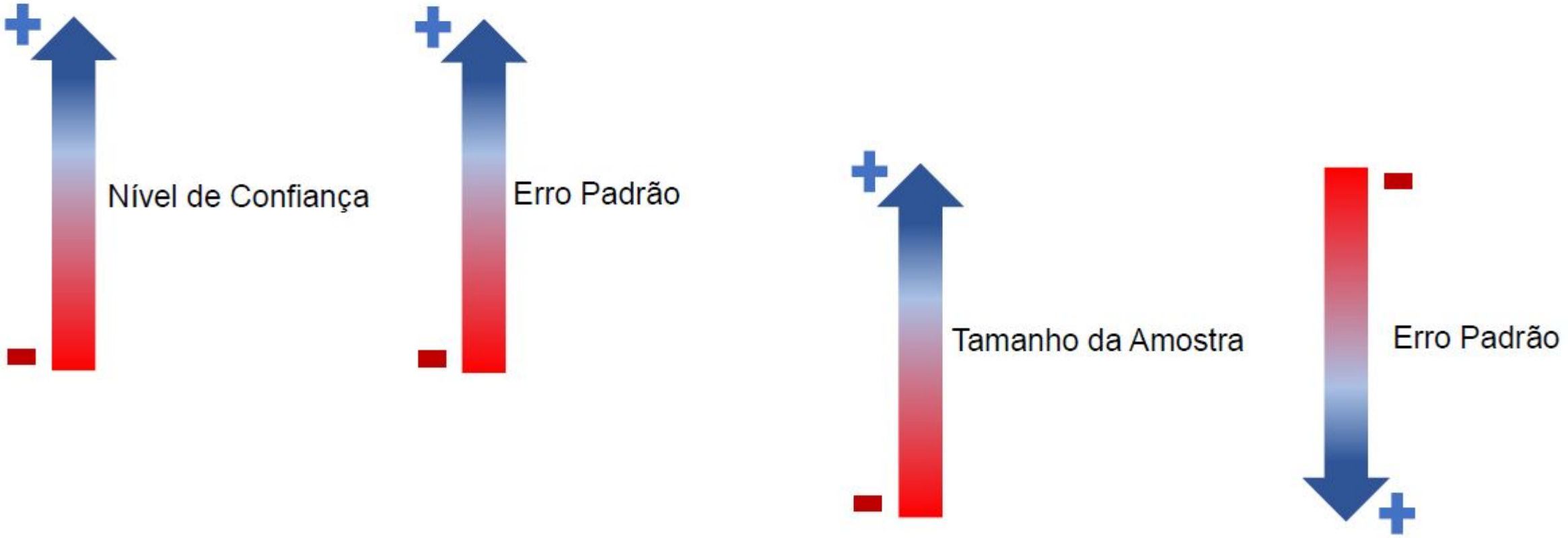
Intervalo de Confiança




PROF. FERNANDO AMARAL
www.datascientist.com.br

# COMPENSAÇÃO



# TIPOS DE INTERVALOS DE CONFIANÇA

- Intervalo de Confiança para a média
- Intervalo de Confiança para a proporção

# INTERVALOS DE CONFIANÇA PARA MÉDIA

- Queremos estimar o salário médio de um cientista de dados
- 100 pesquisados (n)
- Intervalo de confiança: 95%
- O desvio padrão é 1100,00
- A média é de R$ 5.800,00
- Valor de z* = 1,96
- **Margem de erro: $\pm$ 215,60**
- O salário médio de um cientista de dados é entre 5.584,40 e 6.015,60 com um nível de confiança de 95%

$$\bar{X} \pm Z * \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{X} \pm 1{,}96 \left(\frac{1100}{\sqrt{100}}\right)$$

$$\bar{X} \pm 1{,}96 * 110$$

$$\bar{X} \pm 215{,}60$$

# AUMENTANDO O INTERVALO DE CONFIANÇA

- 100 pesquisados (n)
- Intervalo de confiança: 99%
- O desvio padrão é 1100,00
- A média é de R$ 5.800,00

# AUMENTANDO O INTERVALO DE CONFIANÇA

- 100 pesquisados (n)
- Intervalo de confiança: 99%
- O desvio padrão é 1100,00
- A média é de R$ 5.800,00
- Valor de z* = 2,58 (era 1,96)

$$\bar{X} \pm Z * \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{X} \pm 2{,}58 \left(\frac{1100}{\sqrt{100}}\right)$$

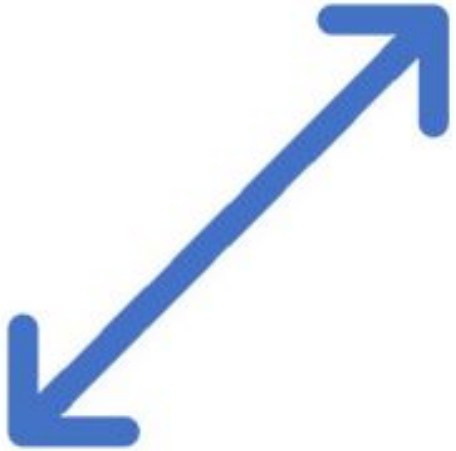
$$\bar{X} \pm 2{,}58 * 110$$

$$\bar{X} \pm 283{,}8$$

**Margem de erro:** $\pm$ 283,8

O salário médio de um cientista de dados é entre 5516,20 e 6083,80 com um nível de confiança de 99%

# INTERVALOS DE CONFIANÇA



Aumentando a margem de erro, é natural que as chances da minha amostra estarem dentro do intervalo, por isso eu tenho um intervalo de confiança maior

Da mesma forma, aumentando n, reduz a chance do efeito acaso, por isso minha margem de erro reduz

# INTERVALO DE CONFIANÇA - AUMENTANDO N

- Queremos estimar o salário médio dos cientistas de dados
- 1000 pesquisados (n)
- Intervalo de confiança: 95%
- O desvio padrão é 1100,00
- A média é de R$ 5.800,00
- Valor de $z^* = 1,96$

# INTERVALO DE CONFIANÇA PARA MÉDIA - AUMENTANDO N

- Queremos estimar o salário médio dos cientistas de dados
- 1000 pesquisados (n)
- Intervalo de confiança: 95%
- O desvio padrão é 1100,00
- A média é de R$ 5.800,00
- Valor de z* = 1,96
- **Margem de erro: +- 68,18**
- O salário médio de um cientista de dados é entre 5.731,82 e 5.868,18 com um nível de confiança de 95%

$$\bar{X} \pm Z * \left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{X} \pm 1{,}96 \left(\frac{1100}{\sqrt{1000}}\right)$$

$$\bar{X} \pm 1{,}96 * 31{,}62$$

$$\bar{X} \pm 68{,}18$$

# INTERVALO DE CONFIANÇA PARA PROPORÇÃO

Queremos estimar a proporção de eleitores que pretendem votar em Maria para prefeita

- 1000 pesquisados (n)
- Intervalo de confiança: 95%
- 650 Responde Maria.
- 330 Responde João.
- 20 Não sabe /Nenhum.
- Valor de $z^* = 1,96$

# INTERVALO DE CONFIANÇA PARA PROPORÇÃO

Queremos estimar a proporção de eleitores que pretendem votar em Maria para prefeita

- 1000 pesquisados (n)
- Intervalo de confiança: 95%
- 650 Responde Maria.
- 330 Responde João.
- 20 Não sabe /Nenhum.
- Valor de z* = 1,96

$$\hat{p} \pm z * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} \pm 1{,}96 * \sqrt{\frac{0{,}65\,(1-0{,}65)}{1000}}$$

$$\hat{p} \pm 0{,}029$$

$$\hat{p} \pm 0{,}03$$

Entre 62 e 68% dos entrevistados pretendem votar em Maria, com um nível de confiança de 95%

# TESTE DE HIPÓTESE

- Confirmar ou negar uma premissa usando uma amostra
  - Seu objetivo é averiguar se um determinado valor hipotético representa bem ou não uma determinada ocasião.

- Esta premissa usa um parâmetro, por exemplo:
  - 56 % dos brasileiros não gostam de estatística

- Encontrar diferença não é tudo, é preciso saber se esta diferença é **estatisticamente significante**

# TESTE DE HIPÓTESE

- $H_0 = hipótese\ nula$ : Alegação que se quer testar
  - Presume-se que é verdadeira, a não ser que existam evidências para provar que não
  - Exemplo: $H_0 = \mu = 100$
- $H_a = hipótese\ alternativa$
  - Exemplos: $H_a \neq 100, \quad H_a > 100, \quad H_a < 100$

# TESTE DE HIPÓTESE - ERROS

- Erro do tipo 1: rejeitar $H_0$ quando não deveria
  - Chance de ocorrer igual a Alfa
- Erro do tipo 2: não rejeitar $H_0$ quando deveria ter rejeitado
  - Depende do tamanho da amostra
  - Ocorrem devido ao acaso

# TESTE DE HIPÓTESE - ERROS


# TESTE DE HIPÓTESE

- Score padrão: erros padrão que seus dados estão abaixo ou acima da média
- A versão padronizada de sua estatística é chamada de "estatística de teste"
- Olha na versão padronizada de Z. Se sua estatística de teste estiver próxima de zero ou num intervalo onde os resultados devem estar, então não se pode rejeitar $H_0$
- Se estive próximo a cauda, então podemos rejeitar $H_0$

# TESTE DE HIPÓTESE

- Alfa e valor - p
- Níveis de $\propto$ (alfa) :
  - 0,05 (normalmente usado)
  - ou 0,01
- Interpretar valor –p
  - -p ≥ alfa : não rejeita $H_0$
  - -p < alfa : rejeita $H_0$

# TESTE DE HIPÓTESE - ETAPAS

- 1. Definir o tamanho da sua amostra
- 2. Coletar dados
- 3. Calcular a média e o desvio padrão
- 4. Definir as duas hipóteses:
- 5. Definir seu $\propto$
- 6. Padronizar seus dados gerando a estatística de teste
- 7. Encontrar o valor -p na tabela Z
- 8. Comparar com seu $\propto$
- 9. Emitir seu veredito

# TESTE DE HIPÓTESE

## Fórmula para Estatística de Teste

*Média:*

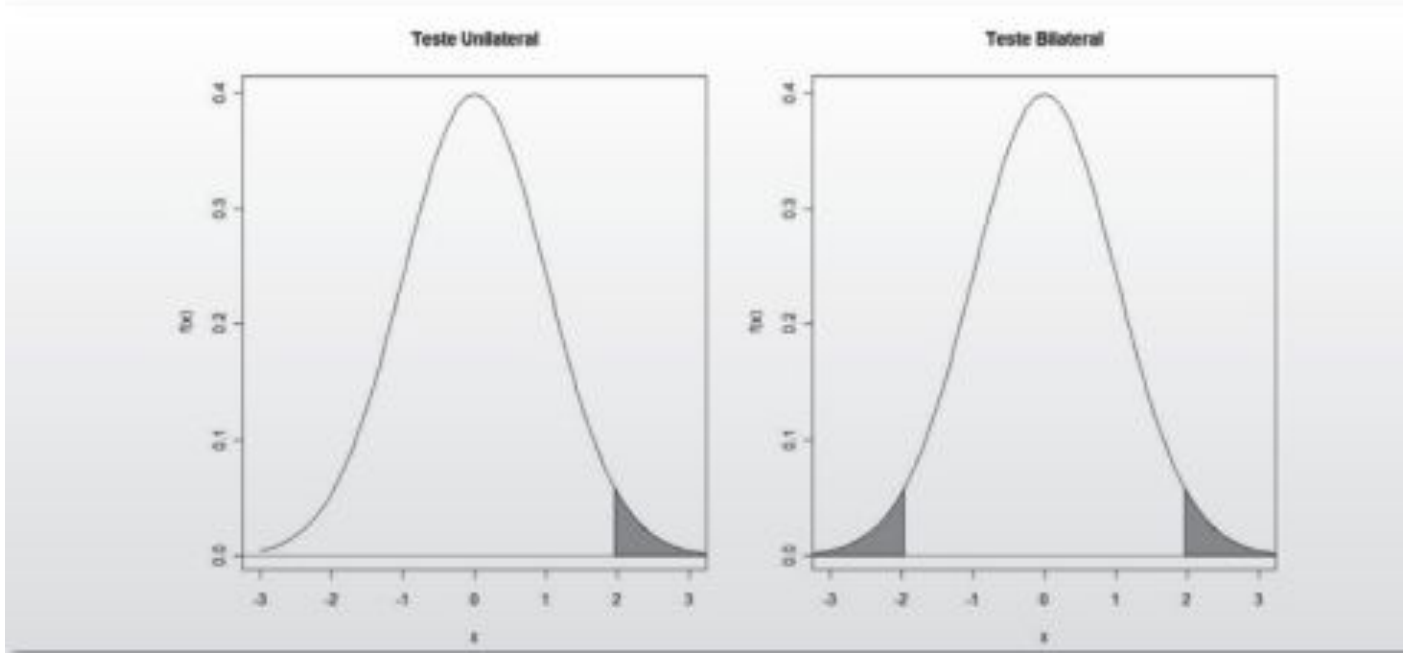
$$Z = \frac{\bar{X} - \mu_o}{\frac{\sigma}{\sqrt{n}}}$$

*Proporção:*

$$\mathrm{P} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_{0(1-p_0)}}{n}}}$$

# TESTE DE HIPÓTESE

- O teste de hipóteses pode ser
  - Unilateral
  - Bilateral




  - Vamos ver como usar cada caso nos próximos exemplos

# TESTE DE HIPÓTESE UNILATERAL

- Exemplo: Uma companhia de cigarros anuncia que o índice médio de nicotina dos cigarros que fabrica apresenta-se abaixo de 23 mg por cigarro. Um laboratório realiza seis análises desse índice, obtendo uma media de 24,17 e um desvio padrão de 8,87 . Pode-se aceitar a afirmação do fabricante?

# TESTE DE HIPÓTESE BILATERAL

- Exemplo: um fabricante afirma que suas lâmpadas duram em média 1000 horas. Você deseja testar se a média de duração real é diferente de 1000 horas.

Você coleta uma amostra de 30 lâmpadas e encontra que a média amostral é 980 horas com um desvio padrão de 50 horas.

# EXEMPLO PROPORÇÃO

Um instituto aponta que 75% dos entrevistados pretendem votar em Maria.
Sabendo que p=77 e n=100

Hipóteses
Ho = p = 0,75
Ha = p < 0,75

$$P = \frac{\hat{p} - p_0}{\sqrt{\frac{p_{0(1-p_0)}}{n}}}$$

$$P = \frac{0,77 - 0,75}{\sqrt{\frac{0,75 * 0,25}{100}}}$$

$$P = \frac{0,02}{0,0433}$$

$$P = 0,461$$

# EXEMPLO PROPORÇÃO

Um instituto aponta que 75% dos entrevistados pretendem votar em Maria.
Sabendo que p=77 e n=100

Hipóteses
Ho = p = 0,75
Ha = p < 0,75

$Valor - P$

*Buscar 0,46 na tabela Z*

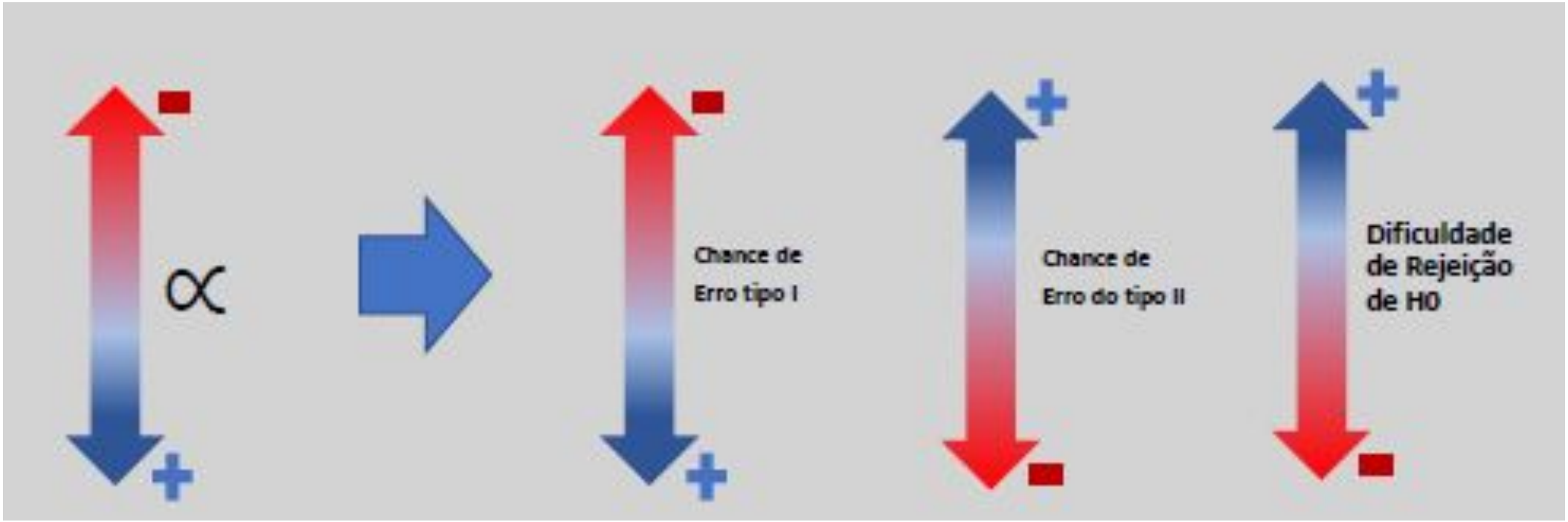
$Z = 0{,}6772$

$Valor\text{-}P = 0{,}6772$

Valor-P ~ Alfa
0,6772 > 0,05

Ho não é rejeitado
De acordo com o estudo, não foi possível rejeitar a hipótese de 75% dos eleitores tem intenção de votar em Maria.

# CONSEQUENCIA DAS ESCOLHAS DE ALFA




Ideal é: amostra grande e alfa pequeno

# EXEMPLO

- Em estudo afirma que, em média crianças de 6 anos pesam 22 Kg . Uma escola pesou seus alunos de 6 anos e obteve uma média de 23Kg, com um desvio padrão de 4 . A escola possui 100 alunos .A escola quer descobrir se a o estudo é correto.

# MÉTRICAS DE ERROS

- Previsão de valores numéricos (reais, inteiros)
- Métricas diferentes da previsão de categorias

- Uso:
  - Regressão linear
  - Regressão ML
  - Series Temporais
  - Etc.

# MÉTRICAS DE ERROS

Quando uma previsão é feita...

- Existe uma diferença entre a previsão e o que ocorreu...
- Temos que medir esta diferença
  - Para saber a qualidade do nosso modelo
  - Para podermos melhora-lo
  - Para podermos fazer benchmarks

# MÉTRICAS DE ERROS

DIFERENÇA:

| Previsto | Realizado |
|---|---|
| 3,34 | 3,00 |
| 4,18 | 4,00 |
| 3,00 | 3,00 |
| 2,99 | 3,00 |
| 4,51 | 4,50 |
| 5,18 | 4,00 |
| 8,18 | 4,50 |

# MÉTRICAS DE ERROS - Mean Erro (ME)

Dependente de Escala

É a média da diferença entre realizado e previsto

| Previsto | Realizado | Dif. |
|---|---|---|
| 3,34 | 3,00 | -0,34 |
| 4,18 | 4,00 | -0,18 |
| 3,00 | 3,00 | 0 |
| 2,99 | 3,00 | 0,01 |
| 4,51 | 4,50 | -0,01 |
| 5,18 | 4,00 | -1,18 |
| 8,18 | 4,50 | -3,68 |
| | | -5,38 |

$$\mathrm{ME} = \frac{\sum_{i=1}^{n} y_i - x_i}{n},$$

$$\mathrm{ME} = \frac{-5{,}38}{7} = -0{,}76$$

# MÉTRICAS DE ERROS - Mean Absolute Error (MAE)

É a média da diferença absoluta entre o realizado e o previsto

| Previsto | Realizado | Dif. Absoluta |
|---|---|---|
| 3,34 | 3,00 | \|0,34\| |
| 4,18 | 4,00 | \|0,18\| |
| 3,00 | 3,00 | \|0\| |
| 2,99 | 3,00 | \|0,01\| |
| 4,51 | 4,50 | \|0,01\| |
| 5,18 | 4,00 | \|1,18\| |
| 8,18 | 4,50 | \|3,68\| |
| | | 5,4 |

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

$$\mathrm{MAE} = \frac{5,4}{7} = 0,77$$

# MÉTRICAS DE ERROS - Root Mean Squared Error (RMSE)

Independente de Escala

É o desvio padrão da amostra da diferença entre o previsto e o teste

| Previsto | Realizado | Dif. ao Quad. |
|---|---|---|
| 3,34 | 3,00 | 0,1156 |
| 4,18 | 4,00 | 0,0324 |
| 3,00 | 3,00 | 0 |
| 2,99 | 3,00 | 1E-04 |
| 4,51 | 4,50 | 1E-04 |
| 5,18 | 4,00 | 1,3924 |
| 8,18 | 4,50 | 13,5424 |
| | | 15,083 |

$$RMSE = \sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$$

$$RMSE = \sqrt{\frac{15,083}{7}}$$

$$RMSE = 1,46$$

# MÉTRICAS DE ERROS - Mean Percentage Error (MPE)

Independente de Escala (%)

É a diferença percentual de erro

| Previsto | Realizado | Erro % |
|---|---|---|
| 3,34 | 3,00 | -11,3333 |
| 4,18 | 4,00 | -4,5 |
| 3,00 | 3,00 | 0 |
| 2,99 | 3,00 | 0,333333 |
| 4,51 | 4,50 | -0,22222 |
| 5,18 | 4,00 | -29,5 |
| 8,18 | 4,50 | -81,7778 |
| | | -127 |

$$\mathrm{MPE} = \frac{100\%}{n} \sum_{t=1}^{n} \frac{a_t - f_t}{a_t}$$

$$\mathrm{MPE} = \frac{-127}{7}$$

$$\mathrm{MPE} = -18{,}14$$

# MÉTRICAS DE ERROS - Mean Absolute Percentage Error (MAPE)

Independente de Escala (%)

É a diferença absoluta percentual de erro

| Previsto | Realizado | Erro abs. | Erro % abs. |
|---|---|---|---|
| 3,34 | 3,00 | 0,1156 | 0,1133333 |
| 4,18 | 4,00 | 0,0324 | 0,045 |
| 3,00 | 3,00 | 0 | 0 |
| 2,99 | 3,00 | 1E-04 | 0,0033333 |
| 4,51 | 4,50 | 1E-04 | 0,0022222 |
| 5,18 | 4,00 | 1,3924 | 0,295 |
| 8,18 | 4,50 | 13,5424 | 0,8177778 |
| | | | 1,2766667 |

$$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right|$$

$$\text{MAPE} = \frac{1{,}2766667}{7}$$

$$\text{MAPE} = 0{,}18$$

# MÉTRICAS DE ERROS

| Previsto | Realizado | Diferença | Dif. Abs. | Dif. Quad. | Erro % | Erro % abs |
|---|---|---|---|---|---|---|
| 3,34 | 3 | -0,34 | 0,34 | 0,1156 | -11,3333 | 11,33333 |
| 4,18 | 4 | -0,18 | 0,18 | 0,0324 | -4,5 | 4,5 |
| 3 | 3 | 0 | 0 | 0 | 0 | 0 |
| 2,99 | 3 | 0,01 | 0,01 | 1E-04 | 0,33333 | 0,333333 |
| 4,51 | 4,5 | -0,01 | 0,01 | 1E-04 | -0,22222 | 0,222222 |
| 5,18 | 4 | -1,18 | 1,18 | 1,3924 | -29,5 | 29,5 |
| 8,18 | 4,5 | -3,68 | 3,68 | 13,5424 | -81,7778 | 81,77778 |

| | |
|---|---|
| ME | -0,76857 |
| MAE | 0,77143 |
| RMSE | 1,46789 |
| MPE | -18,1429 |
| MAPE | 18,2381 |

# MÉTRICAS DE ERROS - EXEMPLO

Calcule o Root Mean Squared Error (RMSE) da seguinte análise:

| Previsto | Realizado |
|---|---|
| 3,12 | 3,00 |
| 6,18 | 7,20 |
| 12,30 | 11,00 |
| | |

# REFERÊNCIAS


- FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de Análise de Dados:** Estatística e Modelagem Multivariada com Excel, SPSS e Stata. Rio de Janeiro: Ltc, 2020.
- CIFERRI, Cristina Dutra de Aguiar; CIFERRI, Ricardo Rodrigues. **Modelagem Multidimensional.** São Paulo: Usp, 2020. 14 slides, color. Disponível em: <http://wiki.icmc.usp.br/images/6/6a/SCC5911-02-ModelagemMultidimensional.pdf>. Acesso em: 10 jan. 2020.
- RESENDE, Tânia. **Modelagem multidimensional conceitos básicos.** São Paulo: Slideshare, 2016. 28 slides, color. Disponível em: <https://pt.slideshare.net/TANIARESENDE/modelagem-multidimensional-conceitos-bsicos>. Acesso em: 18 fev. 2020.
- JARDIM, Edgar Silveira; OLIVEIRA, Marcus Vinícius Abreu de; MORAVIA, Rodrigo Vitorino. Diferença Entre Banco de Dados Relacional e Banco de Dados Dimensional. **Revista Pensar Tecnologia**, Belo Horizonte, v. 2, n. 4, p.1-17, julho 2015. Mensal. Disponível em: <http://revistapensar.com.br/tecnologia/pasta_upload/artigos/a122.pdf>. Acesso em: 18 fev. 2020.
- SERGENTI, Alexsandro. **Modelagem Relacional e Multidimensional:** uma análise envolvendo Sistemas de Apoio a decisão. 2015. Disponível em: <https://www.linkedin.com/pulse/modelagem-relacional-e-multidimensional-uma-an%C3%A1lise-de-sergenti/>. Acesso em: 18 fev. 2020.