

# E.B.A - ESTATÍSTICA DO BÁSICO AO AVANÇADO

## COM RENATA BIAGGI



Renata Biaggi

# SUMÁRIO

1. Como usar esse e-book | 6
2. Excel e Google Sheets | 8
3. Python | 11
4. Introdução à Estatística | 21
5. Tipos de dados e algumas representações | 23
6. Medidas da estatística descritiva | 35
7. Introdução a probabilidade | 68
8. Distribuições discretas e contínuas | 84
9. Testes de hipótese - Conceitos fundamentais | 108
10. Intervalo de confiança para médias | 144  
Oseias Dias de Farias  
oseiasdiasdefarias@gmail.com  
021.399.242-66
11. Testes de hipótese para médias | 156
12. Usos e abusos do intervalo de confiança | 178
13. Intervalo de confiança para proporção | 187
14. Testes de hipótese para proporção | 189
15. Intervalo de confiança para variância | 196
16. Testes de hipótese para variância | 202
17. Testes de hipótese - categóricos ou proporção com mais de 2 categorias | 214
18. ANOVA - comparação com mais de 2 amostras | 223
19. Pensamento crítico: Rejeição da hipótese nula | 242
20. Pensamento crítico: Deep-dive no p-valor | 246
21. Testes não-paramétricos | 256
22. Teste AB - Desenhando um experimento | 274

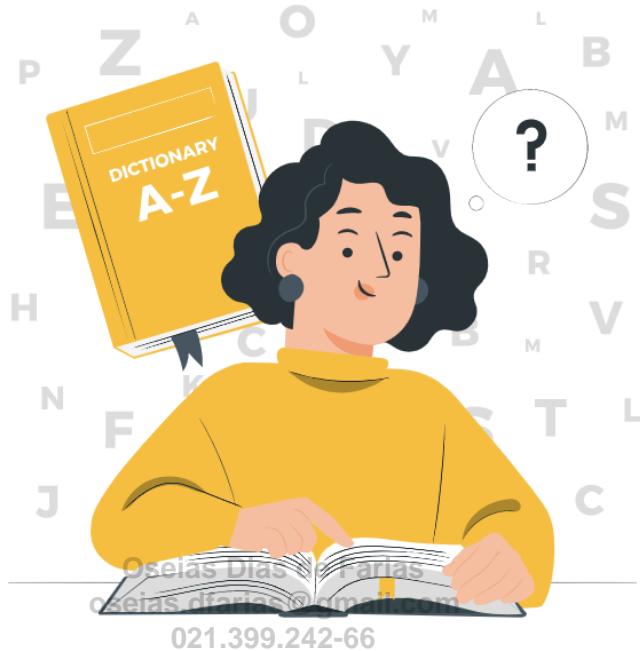


- 23. Correlação | 287**
- 24. Regressão Linear | 299**
- 25. Regressão Logística | 355**
- 26. The End | 368**
- 27. Referências bibliográficas | 370**

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66



# GLOSSÁRIO DE SÍMBOLOS



## ITEM DE UMA SEQUÊNCIA

Se temos um conjunto de dados, cada item pode ser representado como sendo  $x_1$ ,  $x_2$ ,  $x_3$ , etc. Por exemplo, temos o conjunto 10, 21, 32, 43, 58, então:

$$x_1 = 10, x_2 = 21, x_3 = 32, x_4 = 43, x_5 = 58$$

## SOMATÓRIO - $\Sigma$

Em matemática, somatório ou somatória é a adição de uma sequência de quaisquer tipos de números. O resultado é sua soma ou total.



$$\sum_{i=1}^n x_i = x_1 + x_2 + \cdots + x_n$$

O índice  $i$  significa a partir de qual elemento devemos começar a soma. O símbolo  $n$  indica que, nesse caso, vamos somar a sequencia toda - que tem um total de  $n$  elementos.

Vamos a alguns exemplos. Vamos supor que queremos somar os números 1, 2 e 3, ou seja, temos um total de 3 elementos para somar:

$$\sum_{i=1}^3 x_i = x_1 + x_2 + x_3 = 1 + 2 + 3 = 6$$

Muitas vezes vemos o somatório em expressões como:

$$\sum_{n=1}^3 2n - 1$$

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

Esse somatório indica que devemos somar a expressão  $2*n - 1$ , substituindo o  $n$  por cada um dos valores da série. Logo, para a série 1, 2 e 3 temos

$$\begin{aligned} & \sum_{n=1}^3 2\textcolor{brown}{n} - 1 \\ &= \underbrace{[2(\textcolor{brown}{1}) - 1]}_{n=1} + \underbrace{[2(\textcolor{brown}{2}) - 1]}_{n=2} + \underbrace{[2(\textcolor{brown}{3}) - 1]}_{n=3} \\ &= 1 + 3 + 5 \\ &= 9 \end{aligned}$$



## PRODUTÓRIO - $\Pi$

O produtório é a multiplicação de uma sequência de objetos matemáticos (números, funções, vetores, matrizes, etc.), chamados fatores, que tem como resultado o produto. É uma operação análoga ao somatório, embora seja menos utilizada quanto esse último. É representado pela letra grega pi maiúscula ( $\Pi$ ).

$$\prod_{i=1}^n x_i = x_1 \times x_2 \times \dots \times x_n$$

## OUTROS SÍMBOLOS IMPORTANTES

Símbolo	Parâmetro
$\mu$	Oseias Dias de Farias oseias.dfarias@gmail.com Média populacional .242-66
$\sigma$	Desvio-padrão populacional
$\sigma^2$	Variância populacional
$p$	Proporção populacional
$\bar{x}$	Média amostral
$s$	Desvio-padrão amostral
$s^2$	Variância amostral
$\hat{p}$	Proporção amostral
$\alpha$	Nível de significância



# 1. Como usar esse e-book



Bem-vindos(as) ao nosso e-book do E.B.A - Estatística do Básico ao Avançado! Sou a professora Renata Biaggi e estarei com vocês de perto nessa jornada em busca de mais conhecimento técnico e de tomadas de decisões bem fundamentadas.

O e-book é composta por todos os tópicos que vemos no curso E.B.A de uma forma muito detalhada e exemplificada. Para alunos do curso, minha sugestão é que vocês leiam o tópico referente a aula antes de assistirem as aulas referentes. Veja a tabelinha abaixo para se orientarem melhor

<b>Aula</b>	<b>Capítulo para ler</b>
Módulos de Python, Excel e Google Sheets	2, 3
Aula 1. Estatística Descritiva	4, 5, 6
Aula 2. Probabilidades e distribuições	7, 8,
Aula 3. Conceitos fundamentais de teste de hipótese	9
Aula 4. Intervalo de confiança e teste de hipótese para médias e proporções	10, 11, 12, 13, 14
Aula 5. Intervalo de confiança e teste de hipótese para variâncias, ANOVA e outras	15, 16, 17, 18, 19, 20
Aula 6. Testes não-paramétricos	21 Oseias Dias Farias <a href="mailto:oseias.dfarias@gmail.com">oseias.dfarias@gmail.com</a>
Aula 7. Experimentos e teste AB	22 021.399.212-66
Aula 8. Correlação e regressão linear	23, 24
Aula 9. Outras regressões e introdução a ML	25, 26

Ao longo do e-book vocês vão encontrar várias citações de autores, bem como indicações de livros e blogs para que vocês complementem os estudos - especialmente para os tópicos nos quais não vamos nos aprofundar durante o curso.

Apesar do intuito do nosso e-book não ser aprender como usar linguagens de programação ou ferramentas de uma forma geral, precisamos do auxílio de alguma interface para fazermos cálculos de uma forma mais rápida e para lidar com todos os nossos dados. Por isso, escrevi 2 tópicos no e-book abordando de forma bastante rápida as principais ferramentas que guiarão esse curso: Python e Excel. É importante ressaltar que para alunos do E.B.A,

criamos um módulo extra de python para quem nunca teve contato com a linguagem poder acompanhar o curso em Python caso queira. Ao final do módulo, indicamos materiais complementares caso você queira se aprofundar na linguagem. Depois desses dois tópicos, entraremos no conteúdo que aborda a nossa tão amada matemágica matemática.

Outro ponto importantíssimo de ressaltarmos é que neste e-book vamos detalhar os cálculos de todos os exemplos para demonstrarmos como usar cada fórmula. Na prática, dificilmente calcularemos qualquer coisa "na mão". Softwares como Excel ou linguagens como Python já tem esses cálculos intrínsecos, salvando bastante nosso tempo. Daí a importância dos capítulos introdutórios, uma vez que precisaremos manipular tais ferramentas para que elas façam todos os cálculos por nós.

Então Renata, pra quê eu preciso ver todas as fórmulas e entender os exemplos de cálculo? Porque aqui nós **não seremos ferramenteiros!** Nossa intuito é sair do curso entendendo de fato cada conceito que dá suporte às suas análises e modelos preditivos para que, quando vocês se deparam com situações complexas reais, vocês saibam agir sozinhos e possam pensar criticamente no que fazer.

Preparados? Bora lá!



## 2. Excel e Google Sheets



Na formação em dados, muitas ferramentas e linguagens de programação são essenciais para que o analista/cientista possa coletar, tratar, analisar e visualizar os dados. Sabe-se hoje em dia que uma das ferramentas mais populares é o *Excel* - que vem sendo cada vez mais substituído pelo *Google Sheets* - e que dispensa qualquer tipo de apresentação.

Na minha experiência como analista e, posteriormente, como cientista, posso afirmar que nunca deixei de usar o *Excel ou Google Sheets*. É um fato que quanto mais técnicos vamos nos tornando na nossa jornada profissional, mais usamos outras ferramentas que suportem uma gama melhor de estrutura - como veremos mais para frente que é o caso de Python. Porém, gosto de ressaltar a todos os meus alunos que considero o *Excel* uma das ferramentas mais democráticas que temos quando lidamos com dados - podemos coletar, apresentar, fazer dashboards, análises simples e, por vezes, análises mais complexas. Mais que isso, recomendo a todos os iniciantes na carreira de dados que comecem primeiro entendendo como essa ferramenta funciona para só depois se aprofundar em linguagens de programação.

Entretanto, reconheçamos também as graves limitações da ferramenta. Quem já trabalhou com essas o *Google Sheets* ou *Excel* sabe que, se estamos lidando com muitas linhas ou muitas colunas (ou seja, uma quantidade muito grande de dados), uma simples tarefa de cálculo de média pode se tornar um grande transtorno.

Outro ponto a se levar em conta é que a ferramenta também tem limitações de pacotes. Alguns testes estatísticos não têm o cálculo embutido na ferramenta e, portanto, a solução seria escrever a fórmula inteira (que pode ser bem complexa) ou partir para uma outra ferramenta mais amigável.

Além disso, temos uma grande desvantagem: *Excel* não é uma ferramenta gratuita. Apesar de termos o primo dele, *Google Sheets*, de forma gratuita, eu já adianto a vocês que os pacotes analíticos e estatísticos do *Google Sheets* nem se comparam ao do *Excel*.

**Testes estatísticos um pouco mais avançados serão possíveis de ser realizados de forma simples apenas no Excel.** Para eles, quem optar por seguir o curso usando essas ferramentas, vocês terão que deixar de lado o *Google Sheets* e partir para o *Excel*. Deixarei a indicação de pacote no *Excel* gratuito (chamado Real Statistics - suportado apenas pelo *Excel* e não pelo *Google Sheets*) que suporta diversos testes analíticos nos materiais complementares abaixo.

Dada todas as ponderações, chegamos a um impasse. Usamos ou não usamos o *Google Sheets/Excel*? Devemos nos aventurar em ferramentas mais completas e, portanto, mais complexas?

A resposta para essas perguntas não é simples e deve ser dita por cada um de vocês, considerando toda a bagagem e cenários profissionais que vocês têm em mãos. Porém, cabe a mim, como professora, dar todas as opções que julgo interessantes para que vocês tomem essa decisão.

**Devido a tantos "contras", o curso dará ênfase em Python (linguagem gratuita), porém também mostrarei como usar o Google Sheets para a maioria dos cálculos que eu fizer em Python. Para os cálculos que exijam**

o Excel especificamente, indicarei vídeos que demonstrem como usá-los de forma bastante detalhada.

## MATERIAL COMPLEMENTAR

SE SENTE UM POUCO INSEGURÓ(A) COM O GOOGLE SHEETS? DÊ UMA OLHADINHA NESSES TREINAMENTOS:

- INTRODUÇÃO AO GOOGLE SHEETS (GOOGLE PLANILHAS)[HASHTAG TREINAMENTOS]:  
[HTTPS://WWW.YOUTUBE.COM/WATCH?V=ZSQEWYIPVNS](https://www.youtube.com/watch?v=zsQEWYIPvNs)
- #04 - PLANILHAS GOOGLE DOCS - FÓRMULAS E FUNÇÕES /OPERAÇÕES MATEMÁTICAS [ALESSANDRO TROVATO]:  
[HTTPS://WWW.YOUTUBE.COM/WATCH?V=9XFAC607SAM&LIST=PL7IAT8C5WUMRW5IOSYLRTDVGEKNBTVR9Z&INDEX=4](https://www.youtube.com/watch?v=9xFAC607SAM&list=PL7IAT8C5WUMRW5IOSYLRTDVGEKNBTVR9Z&index=4)
- #05 - PLANILHAS GOOGLE DOCS - OPERAÇÕES MATEMÁTICAS (DIVISÃO, PORCENTAGEM E POTENCIACÃO)[ALESSANDRO TROVATO]  
[HTTPS://WWW.YOUTUBE.COM/WATCH?V=55KWXUTCGUQ&LIST=PL7IAT8C5WUMRW5IOSYLRTDVGEKNBTVR9Z&INDEX=5](https://www.youtube.com/watch?v=55KWXUTCGUQ&list=PL7IAT8C5WUMRW5IOSYLRTDVGEKNBTVR9Z&index=5)  
osetas.darias@gmail.com  
021.399.242-66
- #08 - PLANILHAS GOOGLE DOCS - FUNÇÕES ESTATÍSTICAS [ALESSANDRO TROVATO]  
[HTTPS://WWW.YOUTUBE.COM/WATCH?V=OCJLTZTM1h4&LIST=PL7IAT8C5WUMRW5IOSYLRTDVGEKNBTVR9Z&INDEX=8](https://www.youtube.com/watch?v=OCJLTZTM1h4&list=PL7IAT8C5WUMRW5IOSYLRTDVGEKNBTVR9Z&index=8)
- #15 - PLANILHAS GOOGLE DOCS - BUSCAR DADOS ENTRE PÁGINAS DA PLANILHA [ALESSANDRO TROVATO]  
[HTTPS://WWW.YOUTUBE.COM/WATCH?V=jFjI8XYXQ6w&LIST=PL7IAT8C5WUMRW5IOSYLRTDVGEKNBTVR9Z&INDEX=15](https://www.youtube.com/watch?v=jFjI8XYXQ6w&list=PL7IAT8C5WUMRW5IOSYLRTDVGEKNBTVR9Z&index=15)

QUEM OPTAR POR SEGUIR COM EXCEL (NÃO GOOGLE SHEETS), BAXE O PACOTE DE ANÁLISES ESTATÍSTICAS CHAMADO "REAL STATISTICS" - DISPONÍVEL APENAS PARA EXCEL

- BAXE O PACOTE:  
[HTTPS://WWW.REAL-STATISTICS.COM/FREE-DOWNLOAD/REAL-STATISTICS-RESOURCE-PACK/REAL-STATISTICS-RESOURCE-PACK-EXCEL-2007/](https://www.real-statistics.com/free-download/real-statistics-resource-pack/real-statistics-resource-pack-excel-2007/)
- VEJA COMO FAZER A INSTALAÇÃO E ATIVAÇÃO (FERNANDA PERES):  
[HTTPS://WWW.YOUTUBE.COM/WATCH?V=8G7RwDOM-MG](https://www.youtube.com/watch?v=8G7RwDOM-MG)



### 3. Python



Oseias Dias de Farias

Na formação em dados, muitas ferramentas e linguagens de programação são essenciais para que o analistacientista possa coletar, tratar, analisar e visualizar os dados. Sabe-se hoje em dia que uma das ferramentas mais populares é o Excel - que vem sendo cada vez mais substituído pelo Google Sheets. Entretanto, quem já trabalhou com essas ferramentas sabe que, se estamos lidando com muitas linhas ou muitas colunas, uma simples tarefa de cálculo de média pode se tornar um grande transtorno.

Dessa forma, cientistas e analistas que lidam com volumes enormes de dados muitas vezes partem para ferramentas que suportam isso muito bem - ainda que com certas limitações que vamos discutir mais pra frente. A mais famosa delas é a linguagem de programação Python.

Python é uma linguagem de programação de alto nível, ágil e interativa; considerada como uma das mais populares atualmente e é citada como a Linguagem de Programação do Ano de 2021. Além disso, é uma ferramenta acessível, gratuita e independente.

Uma grande vantagem do uso de Python são suas inúmeras **bibliotecas**. Bibliotecas são coleções de códigos pré-compilados que podem ser usados

posteriormente em um programa para algumas operações específicas bem definidas. Isso torna a programação Python mais simples e conveniente para o programador. O exemplo abaixo mostra a biblioteca Pandas (abreviada por pd) - uma das mais famosas em Python. Podemos usar a função `read_csv` (uma das inúmeras funções pertencente da biblioteca Pandas) para ler um arquivo formato `.csv` e transformá-lo em tabela dentro do Jupyter Notebook (veremos abaixo o que é um Jupyter Notebook).

### Células escritas em Python dentro do Jupyter Notebook

In [1]: `import pandas as pd`

In [2]: `students = pd.read_csv("students.csv")`  
`students.head()`

Out[2]:

	Name	RollNo	Date Of Admission	Emergency Contact
Oscias Dias de Farias				
0	Shubham	oseias.dias.de.farias@gmail.com 021.399.242-66	20-05-2012	9988776655
1	Gagan	2	20-05-2009	8364517829
2	Oshima	3	20-05-2003	5454223344
3	Vyom	4	20-05-2009	1223344556
4	Ankur	5	20-05-1999	9988776655

Para a utilização da linguagem precisamos de alguma **interface** que nos permita escrever nossos códigos e de alguma **memória de ambientes** que possa armazenar os dados/variáveis que geramos. Nesse contexto, existem duas interfaces bastante similares mas com diferentes fontes de armazenamento: O Google Colaboratory e o Jupyter Notebook local. Antes de analisarmos ambas as ferramentas, quero ressaltar que "local" significa que, por trás de toda a interface, a linguagem Python está executando os comandos e armazenando os dados no seu próprio computador, consumindo a memória da sua máquina.

Baixar cada uma dessas bibliotecas muitas vezes pode ser algo confuso e demandar tempo para preparar o ambiente de trabalho, por isso vamos abordar dois ambientes aqui (Anaconda e Colab) que já disponibilizam muitas dessas bibliotecas para nós. Nesses casos, precisamos apenas fazer o `import` das bibliotecas da forma:

```
import <nome da biblioteca> as <apelido que queremos dar aquela biblioteca>
```

## SOBRE A INTERFACE

A interface mais usada para rodarmos um código Python é um notebook. Um *notebook* integra o código e sua saída em um único documento que combina visualizações, texto narrativo, equações matemáticas e outras mídias avançadas. Em outras palavras: é um documento único onde você pode executar o código, exibir a saída e também adicionar explicações, fórmulas, gráficos e tornar seu trabalho mais transparente, comprehensível, repetível e compartilhável.

Oseias Dias da Faria  
oseias.dfarias@gmail.com  
021.399.242-66

O uso de *notebooks* é uma parte importante do fluxo de trabalho com dados em empresas de todo o mundo, e vamos utilizá-lo em nosso curso para desenvolver algumas atividades. Se seu objetivo é trabalhar com dados, o uso de um *notebook* acelerará seu fluxo de trabalho e facilitará a comunicação e o compartilhamento de seus resultados.

Nas sessões abaixo vamos mostrar um pouco mais sobre os notebooks.

## SOBRE O AMBIENTE

Existem vários ambientes para executar Python e, dentro desses ambientes, temos diversos prós e contras. Como comentado acima, vamos aqui abordar dois ambientes principais: o Anaconda e o Colab. A partir de agora vamos nos dedicar a entender um pouco como funcionam esses ambientes e memórias usadas para executar nossos comandos.

## ANACONDA

Os Jupyter Notebooks são *notebooks* totalmente gratuitos. Você pode baixar o software sozinho (basta ir até o site oficial <https://jupyter.org/> e escolher o instalador de acordo com seu sistema operacional), ou utilizá-lo como parte do kit de ferramentas **Anaconda**, através do Anaconda Navigator, que veremos abaixo que é a melhor forma de se usar o Jupyter Notebook.

Veja abaixo uma imagem de um Jupyter Notebook

### Interface Jupyter Notebook

The screenshot shows the Jupyter Notebook interface with the following details:

- Title Bar:** jupyter tutorial Last Checkpoint: 3 minutes ago (autosaved)
- Toolbar:** File, Edit, View, Insert, Cell, Kernel, Widgets, Help, Trusted, Python 3
- Code Cells:**
  - In [1]:

```
import matplotlib.pyplot as plt
import pandas as pd
pd.__version__
```
  - Out[1]: '0.24.1'
  - In [2]:

```
# ri stands for Rhode Island
ri = pd.read_csv('police.csv')
```
  - In [3]:

```
# what does each row represent?
ri.head()
```
  - Out[3]:

	stop_date	stop_time	county_name	driver_gender	driver_age_raw	driver_age	driver_race	violation_raw	violation	search_
0	2005-01-02	01:55	Nan	M	1985.0	20.0	White	Speeding	Speeding	

O objetivo do Anaconda é fornecer tudo o que você precisa (em termos de Python). A plataforma possui uma interface gráfica, o Anaconda Navigator, que facilita o acesso a diferentes aplicações de programação em Python. Ela também ajuda no gerenciamento e implantação de pacotes e bibliotecas, principalmente os relacionados à ciência de dados, com várias ferramentas e algoritmos de Machine Learning e IA prontos para serem usados.

Quando instalamos o Anaconda rodamos nossos códigos no Jupyter Notebook, todos os dados lidos ou criados são armazenados na memória no nosso próprio computador. E assim começa a grande desvantagem de usarmos o Jupyter **localmente** - ou seja, usando a memória da nossa máquina. Muitas vezes a memória RAM dos nossos computadores é bastante limitada, o que pode gerar uma incapacidade de processar tabelas muito grandes.

Essa limitação está relacionada à própria linguagem Python, não ao notebook nem ao Anaconda. A linguagem sempre armazena dados na máquina que está rodando - seja ela sua própria máquina física ou uma máquina virtual ("computadores" virtuais com capacidade de armazenamento muito maiores). Não vamos nos alongar muito nessas questões técnicas computacionais, mas é importante que vocês entendam que, caso o computador de vocês tenha uma memória pequena, pode não ser uma boa ideia usar o Anaconda e o notebook local.

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

**Nota importante:** Dissemos acima que a linguagem Python tem essa limitação de estar atrelada à memória do local que ela está sendo rodada (seu próprio computador ou uma máquina virtual). Porém é importante ressaltar que ainda que você esteja usando um notebook local (rodando no seu próprio computador) e seu computador tenha uma memória muito pequena, é muito provável que a capacidade ainda supere o Excel. Ou seja, Excel possivelmente ainda vai te dar mais trabalho caso você trabalhe com uma quantidade muito grande de dados.

## GOOGLE COLABORATORY (COLAB)

A instalação local do Python é uma etapa importante e que pode ser um pouco tediosa e complexa, embora a longo prazo seja importante. Por isso, surgiu o Anaconda, que já traz consigo bibliotecas e o próprio Python, sem nenhuma instalação extra. Entretanto, como comentamos, toda a memória consumida através do Anaconda será local - salvo casos em que possuímos máquinas virtuais externas atreladas a ele, e geralmente isso só acontece no mundo corporativo. Para superar esse problema e desatrelarmos nossos



programas da memória do nosso computador, o Google criou a ferramenta *Google Colaboratory (Colab)*.

O Colab é construído na nuvem (inclusive está atrelado ao nosso Google Drive, tal qual o Sheets, Docs, etc), ou seja, uma das suas grandes vantagens é possuir uma RAM própria, ou seja, não precisaremos usar a memória do nosso computador! Além disso, no Colab já temos várias bibliotecas pré instaladas e atualizadas, e tal qual o Anaconda, não precisaremos instalar muitas delas. Resumindo, a ferramenta permite que você escreva e execute Python em seu navegador, sem a necessidade de nenhuma instalação ou configuração específica. Para usá-lo, precisamos apenas ter uma conta no **gmail**.

A interface Colab é muito parecida com a do Jupyter Notebook, e não precisaremos baixar nenhum programa para acessá-lo.

## Interface Colab

The screenshot shows the Google Colaboratory interface. At the top, there's a navigation bar with 'File', 'Edit', 'View', 'Insert', 'Runtime', 'Tools', and 'Help'. Below the navigation bar, the user's name 'Oseias Dias de Farias' and email 'oseias.dfarias@gmail.com' are displayed, along with a phone number '021.399.242-66'. On the right side of the header, there are 'Share', 'Sign in', 'Connect', 'Editing', and a settings icon. The main content area has a title 'What is Colaboratory?' and a sub-section 'Getting started'. The 'Getting started' section contains a bulleted list: 'Zero configuration required', 'Free access to GPUs', and 'Easy sharing'. A note below states: 'Whether you're a student, a data scientist or an AI researcher, Colab can make your work easier. Watch [Introduction to Colab](#) to learn more, or just get started below!'.

**NOTA IMPORTANTE:** PARA ACOMPANHAR ESSE CURSO VOCÊ NÃO PRECISA TER NENHUM CONHECIMENTO PRÉVIO DE PYTHON! FAREMOS GRANDE PARTE DOS EXERCÍCIOS E RESOLUÇÕES TANTO EM PYTHON QUANTO NO GOOGLE SHEETS/EXCEL E VOCÊS PODERÃO ESCOLHER A MELHOR FERRAMENTA PARA VOCÊS. PARA PYTHON, A FERRAMENTA QUE USAREMOS AO LONGO DO CURSO SERÁ O GOOGLE COLAB, MAS CASO VOCÊ PREFIRA, FIQUE A VONTADE PARA USAR O JUPYTER NOTEBOOK. É IMPORTANTE RESSALTAR QUE O GOOGLE SHEETS É UMA FERRAMENTA COM ALGUMAS LIMITAÇÕES E PODE SER QUE NÃO EXISTAM ALGUNS CÁLCULOS PREVIAMENTE CALCULADOS EM SUAS FUNÇÕES NATIVAS. COM ISSO, APESAR DE SEMPRE QUE POSSÍVEL TENTARMOS PARTIR PARA AMBAS AS ABORDAGENS, TENHAM EM MENTE QUE TEREMOS ALGUMAS RESTRIÇÕES.



# INSTALAÇÕES

## ANACONDA

Para quem preferir partir para o Anaconda, basta acessar o site da empresa, fazer o download e seguir as instruções de instalação.

**SITE ANACONDA:** [HTTPS://WWW.ANACONDA.COM/PRODUCTS/DISTRIBUTION](https://www.anaconda.com/products/distribution)

**CASO HAJA DÚVIDAS, SUGERIMOS QUE ASSISTAM A ESSE VÍDEO:**

[HTTPS://WWW.YOUTUBE.COM/WATCH?V=\\_EKOZ5QBPKA](https://www.youtube.com/watch?v=_EkoZ5QBPKA)

## COLAB

A primeira coisa que precisamos fazer para usar o Google Colab é acessar o seguinte link:

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

**COLAB:** [HTTPS://COLAB.RESEARCH.GOOGLE.COM/NOTEBOOKS/INTRO.IPYNB](https://colab.research.google.com/notebooks/intro.ipynb)

Logo seremos direcionados para um notebook chamado *Welcome to Colaboratory*, que é um notebook explicando um pouco mais sobre o Colab.



## Interface notebook Welcome to Colaboratory

The screenshot shows the Colab interface with the title "Welcome To Colaboratory". The left sidebar has a "Table of contents" section with links to "Getting started", "Data science", "Machine learning", "More Resources", and "Featured examples". The main content area displays the "Welcome to Colab!" page, which includes a video thumbnail titled "3 Cool Google Colab Features" and a play button.

### What is Colab?

Colab, or "Colaboratory", allows you to write and execute Python in your browser, with

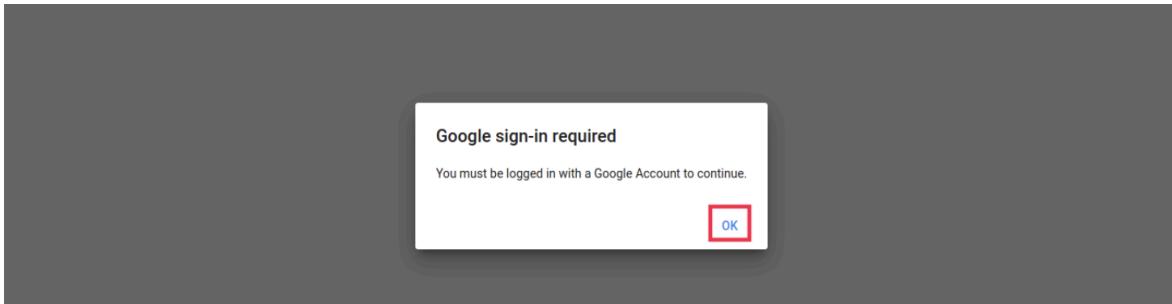
Para criarmos nosso próprio notebook no Colab, precisamos clicar em >File, na parte superior esquerda, logo em seguida em >New Notebook. Veja figura abaixo:

### Criando seu próprio notebook no Colab

The screenshot shows the Colab interface with the title "Oseias Dias de Farias" and email "oseias.dfarias@gmail.com". The left sidebar shows a list of actions: "New notebook" (highlighted with a red box), "Open notebook...", "Upload notebook...", "Rename...", "Move to trash", "Save a copy in Drive...", "Save a copy as a GitHub Gist...", "Save a copy in GitHub...", "Save" (with keyboard shortcut Ctrl+S), "Save and pin revision" (with keyboard shortcut Ctrl+M S), "Revision history", "Download.ipynb" (with keyboard shortcut Ctrl+P), "Download.py", "Update Drive preview", and "Print" (with keyboard shortcut Ctrl+P). The main content area displays the "What is Colaboratory?" page, which includes a list of features: "Zero configuration required", "Free access to GPUs", and "Easy sharing". It also encourages users to get started by watching an introduction video.

Feito esse processo, caso não esteja logado em uma conta google, o seguinte pop-up aparece:

### Pop-up inicial



Clique em OK para ser direcionado à tela de login do gmail. Agora, seu notebook está prontíssimo para usar.

### Colab notebook



## MATERIAL COMPLEMENTAR

QUER DEIXAR O EXCEL DE LADO E COMEÇAR A SE ARRISCAR MAIS EM PYTHON?

- COMECE ASSISTINDO NOSSO MÓDULO EXTRA DE PYTHON!

SE QUISER COMPLEMENTAR, LEIA ESSES ARTIGOS SUGERIDOS:

- 11 CONCEITOS DE PYTHON PARA INICIANTES:

[HTTPS://MEDIUM.COM/@URAPYTHON.COMMUNITY/11-CONCEITOS-PYTHON-PARA-INICIAINTES-EM-PYT  
HON-F87D8238FC8](https://medium.com/@urapython.community/11-conceitos-python-para-iniciantes-em-pyton-f87d8238fc8)

- BIBLIOTECA PANDAS:

[HTTPS://MEDIUM.COM/TECH-GRUPOZAP/INTRODU%C3%A7%C3%A3O-A-BIBLIOTECA-PANDAS-89FA8ED  
4FA38](https://medium.com/tech-grupozap/introdu%C3%A7%C3%A3o-a-biblioteca-pandas-89fa8ed4fa38)

- BIBLIOTECA NUMPY:

[HTTPS://MEDIUM.COM/@ALYSSONMACHADO388/BIBLIOTECA-NUMPY-DO-PYTHON-15BE85B02163](https://medium.com/@alyssonmachado388/biblioteca-numpy-do-python-15be85b02163)

- TUTORIAL COLAB: [HTTPS://WWW.ALURA.COM.BR/ARTIGOS/GOOGLE-COLAB-O-QUE-E-E-COMO-USAR](https://www.alura.com.br/artigos/google-colab-o-que-e-e-como-usar)

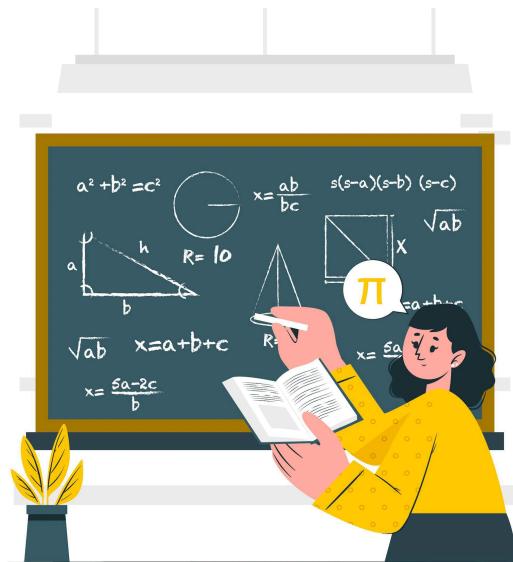
021.399.242-66

QUER SE APROFUNDAR MAIS AINDA E DOMINAR O ASSUNTO BEM O ASSUNTO? SUGERIMOS O CURSO GRATUITO DA DATA SCIENCE ACADEMY

- CURSO DSA: [HTTPS://WWW.DATASCIENCEACADEMY.COM.BR/COURSE/PYTHON-FUNDAMENTOS](https://www.datascienceacademy.com.br/course/python-fundamentos)



# 4. Introdução à Estatística



Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.24266

Em alguma fase de seu trabalho, você já deve ter se deparado com um problema que só poderia ser resolvido analisando e entendendo um conjunto de dados relevantes ao seus estudos. De forma geral, esses conjuntos de dados coletados precisam ser transformados em informações, para compará-los com outros resultados, ou ainda para julgar sua adequação a alguma teoria. Podemos dizer que a essência da estatística.

A estatística, de forma formal, é descrita como a ciência que se preocupa em desenvolver e estudar métodos para coletar, analisar, interpretar e apresentar dados empíricos. É um campo altamente interdisciplinar; a pesquisa em estatística encontra aplicabilidade em praticamente todos os campos de conhecimento humano - business, área da saúde, sociologia, entre outros.

Temos basicamente duas divisões quanto o assunto é estatística: a descrição e a inferência. A **descrição** é a forma que temos de resumir os nossos dados - e dedicamos alguns ao longo da e-book para realizar essa função.

Já a **inferência** está em tudo que tange à incerteza. Existem muitas situações que encontramos na ciência (ou mais geralmente na vida) em que o resultado é incerto. Em alguns casos, a incerteza é porque o resultado em

questão ainda não foi determinado (por exemplo, podemos não saber se choverá amanhã), enquanto em outros casos a incerteza é porque, embora o resultado já tenha sido determinado, não estamos cientes disso (por exemplo, podemos não saber se passamos em um exame específico).

Nesse contexto de incertezas, a probabilidade desempenha um papel fundamental e se dedica justamente a tentar medir de forma matemática quão improvável é um evento acontecer. Qualquer esforço de medição ou coleta de dados está sujeito a várias fontes de variação. Com isso quero dizer que, se a mesma medida/experimento fosse repetida com dados diferentes, a resposta provavelmente mudaria. Os estatísticos tentam entender e controlar (quando possível) as fontes de variação em qualquer situação - e a probabilidade entra com força nisso.

Dentro do cenário de incerteza, como fazemos para tomar uma decisão correta? Como manipular os dados de forma correta e tirar os insights adequados? A estatística e todos os seus campos vão nos ajudar nessa missão.

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

# 5. Tipos dados e algumas representações



## TIPOS DE DADOS

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

Dados são valores atribuídos a algo. Estes valores não precisam ser necessariamente números. Eles também podem ser, por exemplo, conceitos ou posições em um mapa. Dados podem ser medidos ou mensurados por meio de instrumentos, mas também podem ser atribuídos de forma arbitrária (ou seja, por opinião - por exemplo, pesquisas de satisfação).

Dividimos os dados em 2 categorias: numéricos (também chamados de quantitativos) e categóricos (também chamados de qualitativos). Os dados numéricos ainda podem ser divididos em discretos e contínuos, enquanto os categóricos podem ser divididos em nominal e ordinal.



**Variável qualitativa nominal** são valores que expressam atributos, sem nenhum tipo de ordem. Ex: cor dos olhos, sexo, estado civil, presença ou ausência.

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

**Variável qualitativa ordinal** são valores que expressam atributos, porém com algum tipo de ordem, ou grau. Ex: grau de escolaridade (1º grau, 2º grau, 3º grau, pós-graduação...); resposta de um paciente (nenhuma melhora, alguma melhora, muita melhora); classe social (alta, média, baixa).

**Variável quantitativa discreta** são valores observados somente em pontos isolados ao longo de uma escala de valores - ou seja, temos uma quantidade finita de dados. São valores positivos inteiros (incluindo o zero). Ex: Número de filhos(0, 1, 2, ...); Número de faltas; alunos com notas abaixo de 5,0.

**Variável quantitativa contínua** são valores em qualquer ponto fracionário ao longo de um intervalo especificado de valores - ou seja, temos uma quantidade quase infinita de dados. De forma geral, são números com casas depois da vírgula. Ex: temperatura do corpo; altura (em metros).



Para cada tipo de variável existem técnicas apropriadas para resumir as informações. Entretanto, verificaremos que técnicas usadas num caso podem ser adaptadas para outros.

Para finalizar, cabe uma observação sobre variáveis **qualitativas**. Em algumas situações podem-se atribuir valores numéricos às várias qualidades ou atributos (ou, ainda, classes) de uma variável qualitativa e depois proceder-se à análise como se esta fosse quantitativa, desde que o procedimento seja passível de interpretação.

Existe um tipo de variável qualitativa para a qual essa quantificação é muito útil: a chamada **variável booleana**. Uma variável booleana é aquela que pode assumir apenas dois valores. Esses valores geralmente são 0, como ausência, ou 1, como presença. Vamos supor que trabalhamos em um banco e precisamos coletar os dados de transações feitas no cartão de crédito de uma pessoa para analisarmos se houve compras fraudulentas - ou seja, compras feitas por um fraudador que estava tentando roubar o dinheiro do dono do cartão de crédito. Podemos classificar essas transações como 0 quando a transação não é uma fraude e 1 quando aquela transação é uma fraude. Essa é uma variável booleana.

Notas importantes: Para esse capítulo, vamos usar alguns exemplos da referência Bussab, W, Morettin, P. e da Knafllic, C.N.

## REPRESENTAÇÕES TABULAR DE FREQUÊNCIA DE CADA TIPO DE DADO

Quando se estuda uma variável, um dos grandes interesses é conhecer o comportamento dessa variável, analisando a ocorrência de seus possíveis valores. Vamos dar uma olhadinha em alguns tipos de **representação de frequência**, começando pelas variáveis categóricas.

Vamos começar com os **dados categóricos**. Suponhamos que fizemos uma pesquisa com 36 funcionários da seção de "orçamentos" de uma empresa para identificar o nível de escolaridade deles (ou seja, nosso dado de interesse é o nível de escolaridade - uma variável categórica). Nosso resultado foi:



Grau de instrução	Freqüência $n_i$	Proporção $f_i$	Porcentagem $100f_i$
Fundamental	12	0,3333	33,33
Médio	18	0,5000	50,00
Superior	6	0,1667	16,67
Total	36	1,0000	100,00

A coluna "freqüência" indica quantos funcionários tem cada um dos níveis. Por exemplo, 12 funcionários da nossa pesquisa estão no nível fundamental. A coluna "proporção" divide a freqüência encontrada pelo total que coletamos. Voltando ao exemplo do "fundamental", dos 36 funcionários que coletamos, 12 tinham ensino fundamental - ou seja, a freqüência de ocorrência do nível fundamental na nossa pesquisa é  $12/36 = 0,333$ . A coluna "porcentagem" multiplica a coluna "proporção" por 100.

As **proporções**, muitas vezes chamadas de **densidade**, são muito úteis quando se quer comparar resultados de duas pesquisas distintas. Por exemplo, suponhamos que se queira comparar a variável grau de instrução para empregados da seção de orçamento com todos os funcionários da empresa. Digamos que a empresa tenha 2.000 empregados.

Grau de instrução	Freqüência $n_i$	Porcentagem $100f_i$
Fundamental	650	32,50
Médio	1.020	51,00
Superior	330	16,50
Total	2.000	100,00



Não podemos comparar as colunas de frequência das duas tabelas que vimos, porque o número de empregados em cada uma é diferente. No entanto, podemos comparar as colunas de porcentagem, já que elas foram ajustadas para o mesmo total - 100%.

Seja qual for o tipo de categoria que estamos usando (nominal ou ordinal), o jeito de montar a tabela de frequências é o mesmo. Isso vale até para variáveis numéricas discretas. Como são números inteiros e limitados, a gente só conta quantas vezes cada um aparece. Por exemplo, se coletamos quantos filhos cada um dos 36 funcionários tem e descobrimos que 3 têm 4 filhos, a frequência para "4 filhos" é 3/36, ou seja, 8,33%.

Mas quando falamos de variáveis numéricas contínuas, como o salário, temos que ter um pouco mais de cuidado. Se usarmos o mesmo método, não vamos conseguir resumir bem as informações, já que os salários variam bastante. Aí entra a ideia de **dividir os salários em faixas** - chamado de **binarização** ou **discretização**. Por exemplo, uma faixa pode ser de 4 mil a 8 mil, outra de 8 mil a 12 mil, e assim por diante. Isso ajuda a transformar os dados contínuos em algo mais fácil de entender.

Celso Dantas Faria  
csejias.dfarias@gmail.com  
021.399.242-66

Classe de salários	Freqüência $n_i$	Porcentagem $100 f_i$
4,00 ← 8,00	10	27,78
8,00 ← 12,00	12	33,33
12,00 ← 16,00	8	22,22
16,00 ← 20,00	5	13,89
20,00 ← 24,00	1	2,78
<b>Total</b>	<b>36</b>	<b>100,00</b>

Agora que dividimos os salários em faixas, fica fácil contar quantos funcionários se encaixam em cada uma e calcular a porcentagem para cada faixa.

Mas, ao fazer isso com variáveis contínuas, como o salário, perdemos alguns detalhes. Por exemplo, não sabemos exatamente quais são os oito salários na faixa de 12 a 16 mil. Uma solução, sem perder muita precisão, é considerar que todos os salários dessa faixa são iguais ao ponto médio, ou seja, 14 mil. Vamos falar mais sobre isso quando discutirmos medidas de tendência central.

A escolha dos intervalos depende muito do que você conhece sobre seu negócio. Ferramentas como o Excel e Python podem ajudar muito, principalmente quando você quer visualizar os dados em gráficos, como o histograma.

Vamos agora mergulhar um pouco mais no mundo dos gráficos, focando em como eles representam a frequência e entender melhor o tal histograma para variáveis numéricas.

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

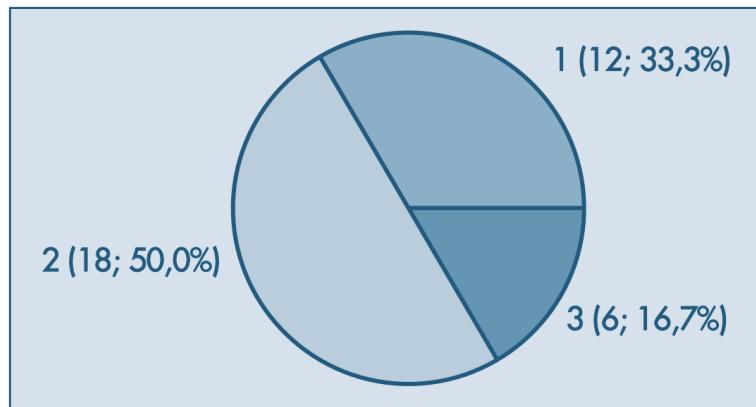
## REPRESENTAÇÃO GRÁFICA DE FREQUÊNCIA PARA VARIÁVEIS QUALITATIVAS

Existem vários tipos de gráficos para representar variáveis qualitativas. Vários são versões diferentes do mesmo princípio, logo nos limitaremos a apresentar dois deles: gráficos em barras e de “pizza”.

### GRÁFICO DE PIZZA

Vou apresentá-lo aqui por uma questão didática, mas nossa maior referência em visualização de dados condena veementemente o uso dos gráficos de pizza em uma apresentação pois “é difícil ler os gráficos de pizza. Quando os segmentos têm tamanhos parecidos, é difícil (senão impossível) dizer qual é o maior.” (Knafllic, C.N.). Quando a autora diz isso, não significa que os gráficos são complexos. Ela apenas quer dizer que esses gráficos podem acabar não passando a mensagem que você deseja. Recomendo fortemente a leitura da bibliografia Knafllic, C.N. caso você tenha interesse em melhorar suas apresentações.

Vamos ao gráfico de pizza. Esse tipo de gráfico destina-se a representar a porcentagem de cada categoria na base de dados. Consiste num círculo de raio arbitrário, representando o todo, dividido em setores, que correspondem às categorias. Para ilustrar, vamos usar como exemplo o grau de instrução dos empregados de "orçamento", exemplificada nas tabelas acima



1 = Fundamental, 2 = Médio e 3 = Superior

Oscar Dias de Faria

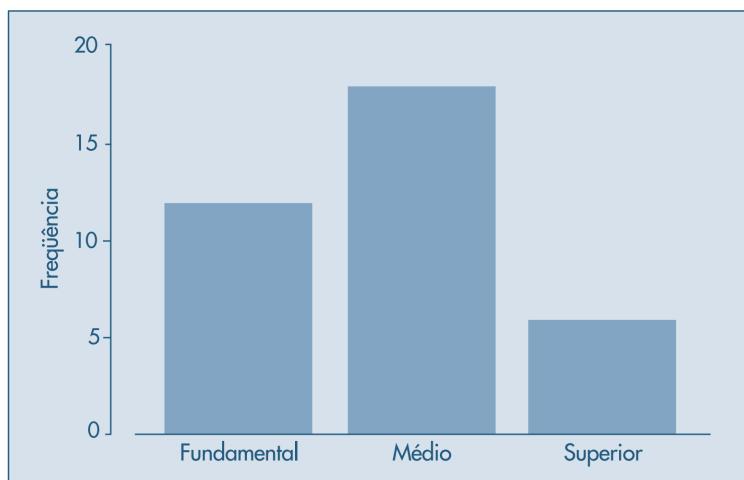
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242-66

## GRÁFICO DE BARRAS

Uma ótima alternativa ao gráfico de pizza é o gráfico de barras, que pode ser horizontal ou vertical. Vamos tomar a variável Y: grau de instrução dos empregados de "orçamento", exemplificada nas tabelas acima. O gráfico em barras consiste em construir retângulos ou barras, em que uma das dimensões é proporcional à magnitude a ser representada (frequência ou proporção), sendo a outra arbitrária, porém igual para todas as barras (mesma largura para todas as barras). Essas barras são dispostas paralelamente umas às outras, horizontal ou verticalmente.





Também podemos representá-lo de forma horizontal, e com percentuais. O exemplo abaixo foi retirado da referência Knaflic, C.N.

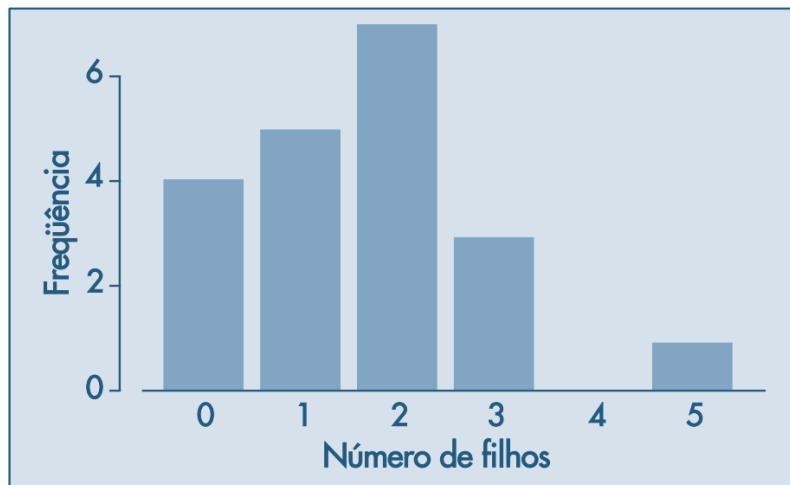


## REPRESENTAÇÃO GRÁFICA DE FREQUÊNCIA PARA VARIÁVEIS QUANTITATIVAS

A representação gráfica mais comum para uma variável quantitativa também é o gráfico de barras. Primeiro, vamos dar uma olhada nas variáveis **quantitativas discretas**.

Vamos supor que perguntamos a algumas pessoas na faixa de 20 a 35 anos quantos filhos elas têm. Esse dado é um valor discreto, pois a quantidade de filhos é finita e não tem-se casas decimais. As respostas dessas pesquisas foram 0 (nenhum filho), 1 (um filho), 2 (dois filhos) e assim por diante.

Da mesma forma que para as variáveis quantitativas, uma ótima forma de representação nesse caso são gráficos de barras. No eixo X temos a quantidade de filhos e no Y quantos funcionários tem cada uma dessas quantidades.

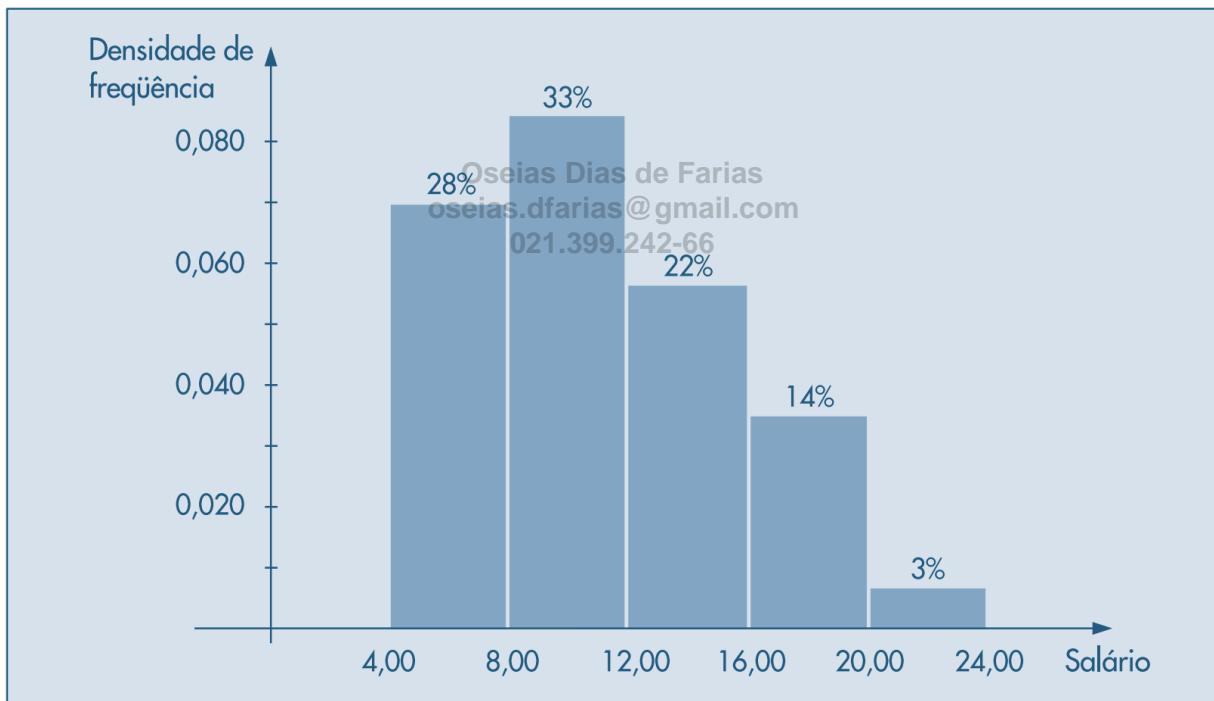


O artifício usado acima para representar uma **variável contínua** faz com que se perca muito das informações nela contidas. Uma alternativa a ser usada nestes casos é o gráfico conhecido como **histograma**.

O histograma é um gráfico de barras continua, com as bases proporcionais aos intervalos das classes e a área de cada retângulo proporcional à respectiva frequência. Pode-se usar tanto a frequência absoluta,  $f_i$  (contador), como a relativa,  $f_i$  (densidade de frequência).

Vamos ver abaixo um exemplo de distribuição de frequências da variável S, salário dos empregados da seção de orçamentos da Companhia MB.

Classes de salários	Ponto médio $s_i$	Freqüência $n_i$	Porcentagem $100f_i$
4,00 – 8,00	6,00	10	27,78
8,00 – 12,00	10,00	12	33,33
12,00 – 16,00	14,00	8	22,22
16,00 – 20,00	18,00	5	13,89
20,00 – 24,00	22,00	1	2,78
Total	–	36	100,00



Para facilitar o entendimento, foi colocada acima de cada retângulo a respectiva percentagem das observações (arredondada). Assim, por meio da figura, podemos dizer que 28% dos funcionários têm um salário entre 4 e 8, 33% dos funcionários entre 8-12 e assim por diante. Esse gráfico também é muito útil para entender os percentuais acumulados. Podemos somar os 2 primeiros retângulos e dizer que 61% (28%+33%) dos empregados têm salário

inferior a 12. Olhando os 2 últimos retângulos, podemos dizer que 17% (14% + 3%) possuem salário superior a 16.

Sabendo que a soma dos percentuais ( $f_i$ ) deve ser 100% (conforme mostra a tabela), dizemos que a soma de cada um desses retângulos deve ser 100%. Esse conhecimento será especialmente importante quando abordarmos **probabilidades**.

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66



# 6. Medidas da estatística descritiva



Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

As estatísticas descritivas resumem os dados de um grupo que você escolher. Ou seja, ela basicamente descreve como está sua distribuição.

Estatísticas descritivas descrevem uma amostra. Você simplesmente pega um grupo no qual está interessado, registra dados sobre cada dado do conjunto e, em seguida, usa estatísticas e gráficos resumidos para apresentar as propriedades do grupo. Com estatísticas descritivas, **não há incerteza** porque você está **descrevendo** apenas os dados que você realmente mede.

Por exemplo, se você medir a altura de dois grupos de pessoas, você conhece as médias precisas para ambos os grupos e pode afirmar sem incerteza qual deles tem a média mais alta. Você **não** está tentando inferir propriedades sobre uma população maior.

Além de frequência numérica como mostrado anteriormente, podemos representar e entender os números a partir de medidas únicas - medidas

centrais e medidas de dispersão. Vamos agora entender as principais medidas da estatística descritiva.

## OUTLIERS

Antes de prosseguirmos com as medidas da estatística descritiva, é importante que façamos a introdução de um conceito que será muito falado no curso: outliers.

Chamamos de outliers aqueles valores que são muito discrepantes de um conjunto de dados. Por exemplo, se coletarmos a altura de 1000 mulheres brasileiras aleatórias, teríamos algo como:

1.51 m, 1.52 m, 1.52 m, 1.55 m, ..., 1.79 m, 1.80 m, 1.90 m

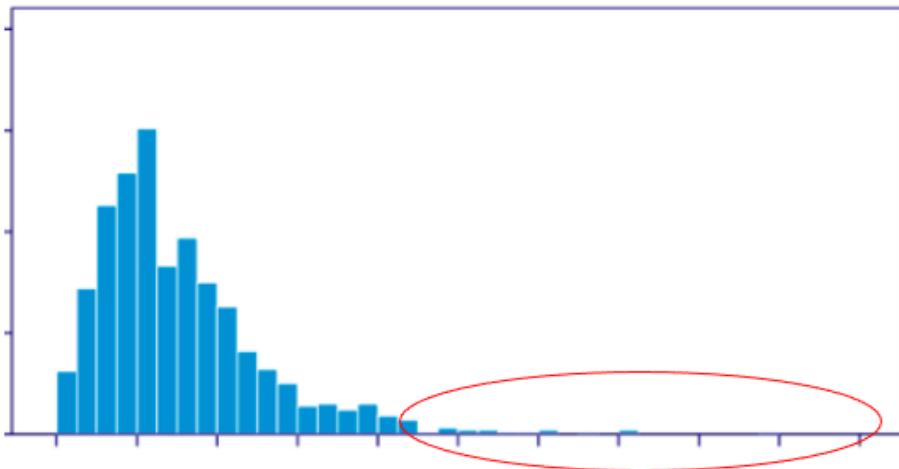
Em que a média é 1.60 m. Observando os dados acima, vemos que a mulher mais alta tem 1.90 m e destoa bastante da média. Dizemos então que esse dado é um outlier.

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66

Muitas vezes usamos a palavra outliers para valores destoantes que são possíveis de acontecer ou valores destoantes impossíveis de acontecer. No nosso caso, é perfeitamente possível encontrarmos uma mulher com 1.90 m de altura. Contudo, se por acaso nessa série encontrássemos uma mulher com 4.0 m, saberíamos que é um outlier impossível de acontecer - ou seja, provavelmente houve um erro quando computaram esse dado.

Em um histograma, os outliers podem ser facilmente vistos através da cauda longa do gráfico:





Falaremos mais sobre outliers quando abordarmos sobre Boxplots.

**NOTA IMPORTANTE: OUTLIERS PODEM SER VALORES MUITO PEQUENOS OU MUITO GRANDES. POR EXEMPLO, SE EXISTISSE NA SÉRIE UMA MULHER DE 1.30 M DE ALTURA, ELA TAMBÉM SERIA UM OUTLIER.**

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

## TENDÊNCIA CENTRAL

Uma medida de tendência central é um valor único que tenta descrever um conjunto de dados identificando sua posição central. São também chamadas de medidas de localização central.

Pense em como você descreve um número. Geralmente, a descrição é feita de acordo com seu valor. Por exemplo, para descrever o número 2, você pode mostrar dois dedos ou dizer  $1 + 1 = 2$ . Mas como você descreveria um grupo de dados? Neste caso, não adiantaria muito usar os dedos, e somar os números seria impossível. Usando medidas de tendência central, você pode descrever um grupo de dados em um único valor.

Existem três medidas de tendência central muito utilizadas no dia-a-dia: a média, a mediana e a moda.

**Média** — é o valor médio dos dados, calculado através da divisão da soma dos valores com o número total de valores.

**Mediana** — é o valor do meio na série de dados, quando os valores estão dispostos em ordem crescente ou decrescente.

**Moda** — é o valor mais comum (o que mais se repete) em uma série de dados.

Média, mediana e moda são todas medidas válidas de tendência central, mas sob diferentes condições. A escolha da medida de tendência central varia de acordo com o uso e necessidade. Nas próximas seções, veremos cada uma delas, aprenderemos como calculá-las e em quais condições utilizá-las.

## MÉDIA ARITMÉTICA

A média aritmética é a mais popular, e muitas vezes é simplesmente chamada de “média”.

Para calculá-la, some os valores de todos os termos e depois divida pelo número de termos. Exemplo:

Oseias Dias de Farias

Qual é a média de 2, 4, 6, 8 e 10?

021.399.242-66

Solução:

Primeiro, some todos os números.

$$2 + 4 + 6 + 8 + 10 = 30$$

Agora, divida por 5 (número total de observações).

$$\text{Média} = 30/5 = 6$$

## MÉDIA PONDERADA

A média ponderada é calculada quando determinados valores fornecidos em um conjunto de dados são mais “importantes” que os outros (possuem maior peso ou maior influência no resultado).

A fórmula é escrita como sendo:

$$W = \frac{\sum_{i=1}^n w_i X_i}{\sum_{i=1}^n w_i}$$



Em que  $W$  é a média ponderada e  $w_i$  é o peso dado àquele conjunto de dados.

**Exemplo 1.** Na escola, a média anual de cada matéria é calculada de acordo com os princípios da média ponderada. Considerando que o peso das notas esteja relacionado com o bimestre em questão, determine a média anual de sabendo que as notas em Matemática foram iguais a:

1º Bimestre: Nota 6- peso 1

2º Bimestre: Nota 7 - peso 2

3º Bimestre: Nota 8 - peso 3

4º Bimestre: Nota 9 - peso 4

$$W = \frac{6*1 + 7*2 + 8*3 + 9*4}{1+2+3+4} = \frac{80}{10} = 8$$

oséias.dfarias@gmail.com  
021.399.242-66

## MÉDIA GEOMÉTRICA/HARMÔNICA

A média geométrica é definida como a raiz enésima ( $n$ -ésima; de grau  $n$ ) do produto de cada valor, ou seja, de  $n$  números no conjunto de dados fornecido.

A média geométrica pode ser aplicada em qualquer conjunto de dados estatístico, mas normalmente ela é empregada na geometria. Há também aplicação em problemas da matemática financeira que envolvam taxa percentual acumulada, ou seja, porcentagem sob porcentagem. Além de ser a média mais conveniente para dados que se comportam como uma progressão geométrica. Sua fórmula é expressa como:



$$\left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \cdots x_n}$$

Essa fórmula pode ser bastante assustadora, mas vamos dar um exemplo para ficar mais claro.

**Exemplo 1.** Qual o valor da média geométrica entre os números 3, 8 e 9?

Como temos 3 valores, iremos calcular a raiz cúbica do produto.

$$M_G = \sqrt[3]{3 \cdot 8 \cdot 9} = \sqrt[3]{216} = 6$$

Simples, né? Agora vamos dar ~~um exemplo do mundo financeiro~~

~~oseias.dfarias@gmail.com~~

**Exemplo 2:** Um investimento rende no primeiro ano 5%, no segundo ano 7% e no terceiro ano 6%. Qual o rendimento médio desse investimento?

Para resolver esse problema devemos encontrar os fatores de crescimento.

- 1º ano: rendimento de 5% → fator de crescimento de 1,05 (100% + 5% = 105%)
- 2º ano: rendimento de 7% → fator de crescimento de 1,07 (100% + 7% = 107%)
- 3º ano: rendimento de 6% → fator de crescimento de 1,06 (100% + 6% = 106%)

$$M_G = \sqrt[3]{1,05 \cdot 1,06 \cdot 1,07} = \sqrt[3]{1,19091} = 1,05996$$

Para encontrar o rendimento médio devemos fazer:

$$1,05996 - 1 = 0,05996$$



Assim, o rendimento médio dessa aplicação, no período considerado, foi de aproximadamente 6%.

Notem que a média geométrica penaliza valores muito baixos. É o caso de quando temos um zero em uma das medidas. Multiplicando qualquer coisa por zero, a média geométrica será zero também (vocês já devem ter tido aquele querido professor que faz a média do semestre se baseando em médias geométricas, certo?).

Na prática, tirando o mundo financeiro, a média geométrica é pouco usada.

## MEDIANA

A mediana, em estatística, é o valor médio da lista de dados fornecida, quando organizados em uma ordem. A disposição dos dados ou observações pode ser feita em ordem crescente ou decrescente. Exemplo: A mediana de {2, 3, 4} é 3. Em matemática, a mediana também é um tipo de média, que é usada para encontrar o valor do centro (é encontrado ordenando-se todos os dados e escolhendo o que está no centro). Para um conjunto de dados, pode ser considerado como o valor "intermediário".

021.399.242-66

A mediana é menos afetada por discrepâncias e dados distorcidos. A característica básica da mediana na descrição de dados — em comparação com a média — é que ela não é distorcida por uma pequena proporção de valores extremamente grandes ou pequenos e, portanto, fornece uma melhor representação de um valor "típico".

A renda mediana, por exemplo, pode ser uma maneira melhor de sugerir o que é uma renda "típica", pois a distribuição de renda geralmente é muito distorcida, especialmente no Brasil (poucos ganham MUITO).

A mediana é de importância central em estatísticas robustas, pois é a estatística mais resistente, tendo um ponto de ruptura de 50%; a mediana não é um resultado arbitrariamente grande ou pequeno, desde que não mais da metade dos dados estejam contaminados.



Para determinar o valor mediano em uma sequência de números, esses números devem primeiro organizados, em ordem de valor do menor para o maior (mais comum) ou do maior para o menor.

Se houver uma quantidade ímpar de números, o valor mediano é o número que está no meio, com a mesma quantidade de números abaixo e acima. Se houver uma quantidade par de números na sequência, o par do meio deve ser determinado, somado e dividido por dois para que se encontre o valor mediano.

Sua fórmula ~~assustadora~~ é dada por:

$$\text{Med}(X) = \begin{cases} X\left[\frac{n}{2}\right] & \text{if } n \text{ is even} \\ \frac{(X\left[\frac{n-1}{2}\right] + X\left[\frac{n+1}{2}\right])}{2} & \text{if } n \text{ is odd} \end{cases}$$

Oscias Dias de Farias  
oseias.dfas@gmail.com  
021.399.242-66

Vamos a alguns exemplos.

**Exemplo 1.** Qual a mediana das alturas dos jogadores de um time de vôlei onde as alturas são: 1,97m; 1,87m; 1,99m; 2,01m; 1,83m?

Organizando os valores em ordem crescente:

1,83m; 1,87m; **1,97m**; 1,99m; 2,01m;

Verificamos que a quantidade de dados é ímpar. A mediana é, portanto, o valor do meio depois da ordenação. Logo, a mediana é **1,97m**

**Exemplo 2.** Calcule o valor da mediana da seguinte amostra de dados: (32, 27, 15, 44, 15, 32).

Primeiro precisamos colocar os dados em ordem, assim temos:

15, 15, 27, 32, 32, 44

Como essa amostra é formada por 6 elementos, que é um número par, a mediana será igual a média dos elementos centrais, ou seja:

$$M_d = \frac{27+32}{2} = \frac{59}{2} = 29,5$$

A mediana às vezes é usada em oposição à média, quando há valores discrepantes na sequência que podem distorcer o resultado. A mediana de uma sequência é menos afetada por **outliers** do que a média.

## MODA

A moda é o valor que aparece com mais frequência em um conjunto de valores de dados.

Assim como a média e a mediana, a moda é uma forma de expressar, em um número (geralmente) único, informações importantes sobre uma variável aleatória ou uma população.

Oseias Dias de Farias  
oseias.dfarias@gmail.com

021.399.242-66

Por exemplo, na sequência abaixo, 16 é a moda, pois aparece mais vezes no conjunto do que qualquer outro número:

3, 3, 6, 9, 16, 16, 16, 27, 27, 37, 48

Um conjunto de números pode ter mais de uma moda (isso é conhecido como bimodal — quando houver duas modas) se houver vários números que ocorram com igual frequência e mais vezes do que os outros no conjunto.

3, 3, 3, 9, 16, 16, 16, 27, 37, 48

No exemplo acima, tanto o número 3 quanto o número 16 são modas, pois cada um ocorre três vezes e nenhum outro número ocorre tão frequentemente.

Se nenhum número em um conjunto ocorrer mais de uma vez, esse conjunto não terá moda, como o exemplo abaixo:

3, 6, 9, 16, 27, 37, 48

Um conjunto de números com duas modas é chamado de bimodal; um conjunto de números com três modas é trimodal; e qualquer conjunto de números com mais de uma moda é multimodal.

A moda é a medida de tendência mais aplicada para séries de dados categóricos (“qualitativos”, que não possuem um ordenamento natural ou relações de grandeza entre si).

## MEDIDAS DE DISPERSÃO

O resumo de um conjunto de dados por uma única medida representativa de posição central esconde toda a informação sobre a dispersão do conjunto de observações. Por exemplo, suponhamos que cinco grupos de alunos submeteram-se a um teste, obtendo-se as seguintes notas:

**grupo A (variável X):** 3, 4, 5, 6, 7

**grupo B (variável Y):** 1, 3, 5, 7, 9

**grupo C (variável Z):** 5, 5, 5, 5, 5

**grupo D (variável W):** 3, 5, 5, 7

**grupo E (variável V):** 3, 5, 5, 6, 6

Vemos que a média em cada um dos grupos é 5,0. A identificação de cada uma destas séries por sua média (5, em todos os casos) nada informa sobre suas diferentes **variabilidades**. Notamos, então, a conveniência de serem criadas medidas que resumam quão dispersos são os dados de um conjunto de observações e que nos permita, por exemplo, comparar conjuntos diferentes de valores, como os dados acima, segundo algum critério estabelecido.

A **variabilidade/dispersão** é considerada significativa quando a variação ou falta de uniformidade no tamanho dos itens de uma série é grande. Se a variabilidade for menor, a dispersão é menos significativa. Se todos os dados forem idênticos, a dispersão é nula.

## ALCANCE

A medida mais simples de **dispersão absoluta** é o **alcance**. Ele é apenas o maior ponto de dados menos o menor. Podemos escrever isso como  $R = H - L$ , sendo  $H$  o valor máximo encontrado na amostra e  $L$  o valor mínimo. Curiosidade: a letra  $R$  foi usada para referenciar a palavra inglesa “range”, que pode ser usada para definir a amplitude dos dados.

Por exemplo, se as notas dos estudantes de matemática em uma prova é dada por:

Aluno A: 2 pontos  
Aluno B: 2 pontos  
Aluno C: 4 pontos  
Aluno D: 5 pontos  
Aluno E: 8 pontos  
Aluno F: 8 pontos  
Aluno G: 9 pontos

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66

O alcance nesse caso seria  $9 - 2 = 7$  pontos.

A medida de alcance pode ser interessante para entendermos, nesse caso, se o aluno com pior nota está em um "patamar" muito diferente do aluno com melhor nota. Nesse caso temos uma diferença de 7 pontos entre eles, o que é grande considerando que as notas variam de 0 a 10.

## DESVIOS E VARIÂNCIA

Um critério frequentemente usado para entender a dispersão de um conjunto é aquele que mede a dispersão dos dados **em torno de sua média** - ou seja, quanto cada um dos dados difere da média do seu grupo. Vamos voltar ao seguinte exemplo de notas um teste de grupos de pessoas:



**grupo A (variável X):** 3, 4, 5, 6, 7

**grupo B (variável Y):** 1, 3, 5, 7, 9

**grupo C (variável Z):** 5, 5, 5, 5, 5

**grupo D (variável W):** 3, 5, 5, 7

**grupo E (variável V):** 3, 5, 5, 6, 6

A média de cada grupo, como já calculamos antes, é 5. Olhando para o grupo C, vemos que os integrantes são muito homogêneos - ou seja, todos ali tiraram 5. Se observarmos por outro lado o grupo B, vemos que temos pessoas que tiraram uma nota super alta (nota 9) e alguém que não estudou e tirou 1 - ou seja, números que estão bem distantes da média do grupo como um todo.

Para analisarmos quão distantes cada um dos pontos está da média e compilamos isso em um único valor, duas medidas são as mais usadas:

**desvios e variância.**

*oséias.dfarias@gmail.com  
021.399.242-66*

Para o grupo A acima os desvios de cada ponto em relação a média ( $x_i - \text{média}$ ) são: -2, -1, 0, 1, 2. Se somarmos cada um desses elementos, teríamos que a soma dos desvios é igual a zero, que teoricamente indicaria que não há nenhum desvio na série. Mas isso não é verdade, certo? Vemos que os elementos não têm o mesmo valor! Isso acontece pois valores desviando para menos (valores menores que a média) se tornam negativos, enquanto valores desviando para mais (valores maiores que a média) tem sinal positivo. Somando-os, anulamos os efeitos e ficamos com uma ideia errada do desvio da série.

Para resolver esse problema temos duas opções: (a) considerar o total dos desvios em valor absoluto (módulo, ignorando o sinal); (b) considerar o total dos quadrados dos desvios, o que faria com que o sinal negativo se tornasse positivo. Para o grupo A teríamos, respectivamente,



$$\sum_{i=1}^5 |x_i - \bar{x}| = 2 + 1 + 0 + 1 + 2 = 6,$$

$$\sum_{i=1}^5 (x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10.$$

O uso desses totais pode causar dificuldades quando comparamos conjuntos de dados com números diferentes de observações, como os conjuntos A e D acima. Desse modo, é mais conveniente exprimir as medidas como médias (dividindo pela quantidade de dados), isto é, o **desvio médio** e a **variância** são definidos por

### Desvio médio

$$dm(X) = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

### Variância

$$var(X) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}$$

Em que  $x_i$  é a representação de cada dado,  $\bar{x}$  é a média de todos os dados juntos e  $n$  é o tamanho da amostra.

Para o grupo A temos

$$dm(X) = 6/5 = 1,2,$$

$$var(X) = 10/5 = 2,0$$

Para o grupo D temos



$$\text{dm}(W) = 4/4 = 1,0,$$

$$\text{var}(W) = 8/4 = 2,0.$$

Podemos dizer, então, que, de acordo com o desvio médio, o grupo D é mais homogêneo que A, enquanto ambos são igualmente homogêneos, segundo a variância.

A variância é uma medida de dimensão igual ao quadrado da dimensão dos dados (por exemplo, se os dados são expressos em  $cm$ , a variância será expressa em  $cm^2$ ), pode causar problemas de interpretação. Costuma-se usar, então, o **desvio padrão**, que é definido como a raiz quadrada positiva da variância.

$$dp(X) = \sqrt{\text{var}(X)}$$

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

Para o grupo A o desvio padrão é

$$dp(X) = \sqrt{\text{var}(X)} = \sqrt{2} = 1,41$$

Uma interpretação mais intuitiva para o desvio padrão é entendê-lo com uma medida de erro caso aproximarmos (ou estimarmos) todos os valores da série dados pela sua média. Quanto erraríamos se representássemos toda nossa série de dados pela média.

Por ser calculado com os quadrados dos desvios da média, tanto erros positivos quanto erros negativos contribuem para o valor final da medida.

Algumas observações sobre os desvios.

- O desvio padrão mede a variação dos dados com relação à média e tem a mesma unidade de medida que o conjunto de dados.
- Os desvios são sempre maior ou igual a 0. Quando  $s = 0$ , o conjunto de dados não apresenta variação (todos os elementos têm o mesmo valor).

- À medida que os valores se afastam da média (isto é, estão mais dispersos), o valor do desvio aumenta.

Veremos mais tarde que a variância de uma amostra será calculada usando-se o denominador  $n - 1$ , em vez de  $n$ . Para grandes amostras isso fará pouquíssima diferença.

**NOTA IMPORTANTE:** TANTO A VARIÂNCIA COMO O DESVIO MÉDIO SÃO MEDIDAS DE DISPERSÃO CALCULADAS EM RELAÇÃO À MÉDIA DAS OBSERVAÇÕES. ASSIM COMO A MÉDIA, A VARIÂNCIA (OU O DESVIO PADRÃO) SÃO AFETADOS EXAGERADAMENTE SE EXISTEM OUTLIERS EM NOSSA SÉRIE. OU SEJA, NOSSO DESVIO AUMENTA MUITO SE TEMOS APENAS 1 VALOR DISCREPANTE.

## COEFICIENTE DE VARIAÇÃO

O **coeficiente de variação (CV)**, a medida análoga de dispersão relativa, é apenas o desvio padrão dividido pela média aritmética. Para dar como uma porcentagem em vez de uma proporção, multiplique por 100%. A fórmula do CV pode ser vista abaixo:

Obs: Dias de Férias  
oseias.dfarias@gmail.com  
021.399.242-66

$$CV = \frac{\sigma}{\mu}$$

Em que  $\mu$  é a média de todos os dados e  $\sigma$  é o desvio-padrão.

O coeficiente de variação é muito interessante pois ele não está atrelado a unidade da medida e não precisamos ter ideia da grandeza da média para entender se o valor é grande ou pequeno.

Por exemplo, vamos supor que coletamos dados de salário mensal de um grupo de pessoas. Se eu te disser que o desvio-padrão é R\$ 500 você não tem ideia se esse desvio é grande ou pequeno, certo? Se a média for R\$ 1.000 reais, esse desvio parece enorme. Agora se a média for R\$ 100.000, esse desvio é bem menor. Logo, para entender o desvio-padrão, precisamos sempre olhar a média e fazer algumas continhas na nossa cabeça para entender quão grande é o desvio.



Com o coeficiente de variação, não precisamos olhar dois números separadamente e fazer continhos na nossa cabeça.

- Coeficiente de variação para média = R\$ 100.000 e desvio-padrão = R\$ 500

$$CV = R\$ 500/R\$ 100.000 = 0,005 = 0,5\%$$

- Coeficiente de variação para média = R\$ 1.000 e desvio-padrão = R\$ 500

$$CV = R\$ 500/R\$ 1.000 = 0,5 = 50\%$$

Percebem que só olhando para o CV já conseguimos entender que os dados são muito mais discrepantes no segundo caso?

O CV também é muito útil quando queremos entender qual distribuição tem maior variabilidade mas não estamos comparando coisas comparáveis. Por exemplo, se quisermos comparar uma distribuição de salário e uma distribuição de altura, como são medidas diferentes (uma em reais e outra em metros) não podemos comparar nem a média nem o desvio. Por outro lado, como o CV não possui unidade, podemos comparar os CVs dessas duas distribuições e entender qual delas é mais dispersa.

## VARIÂNCIA AGRUPADA

A variância agrupada (também conhecida como **variância combinada**, variância composta ou variância geral) é um método para estimar a variância de vários subconjuntos quando a média de cada conjunto pode ser diferente. Vamos a um exemplo para entender quando isso é aplicado.

Aproximadamente 1,7 milhões de estudantes fizeram a prova SAT em 2015. Cada estudante recebeu uma nota em análise crítica e em matemática. Aqui está o resumo das estatísticas para cada parte do teste em 2015:



Matéria	Média	Desvio-padrão
Análise crítica	$\mu_{CR} = 495$	$\sigma_{CR} = 116$
Matemática	$\mu_M = 511$	$\sigma_M = 120$
<b>Total</b>	$\mu_T = ?$	$\sigma_T = ?$

Considere que a nota final deve ser composta pela média de ambas as matérias - matemática e análise crítica. Considerando essas duas matérias, qual é a média e o desvio-padrão que representaria a nota geral dos alunos?

Aqui temos o caso de termos 2 conjuntos diferentes, com tamanhos diferentes. Como podemos calcular o desvio-padrão entre os dois conjuntos? Podemos usar a variância combinada!

Uma variância combinada é uma estimativa da variância da população obtida de duas variâncias da amostra quando se assume que os **dois conjuntos vêm da mesma população**. Nessa situação, nenhuma das variâncias dos conjuntos é uma estimativa melhor do que a outra, e as duas variâncias dos conjuntos fornecidas são "combinadas", em uma espécie de forma de média ponderada, para calcular a variância combinada. Sua fórmula pode ser representada por:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Em que  $s_p^2$  é a variância combinada,  $s_1^2$  é a variância do conjunto 1,  $s_2^2$  é a variância do conjunto 2,  $n_1$  é o tamanho do conjunto 1 e  $n_2$  é o tamanho do conjunto 2. Notem que se o tamanho das amostras for igual, a variância combinada é simplesmente a soma das variâncias de cada amostra.

Aplicando a raiz quadrada nos dados, temos o que chamamos de **desvio padrão agrupado** (ou seja, o desvio padrão de 2 amostras).



$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

No exemplo anterior sobre as provas SAT, temos um típico caso de uso da variância combinada. Para a média, a conta é bastante direta: A média de um estudante retirado da soma de duas variáveis aleatórias é igual a **média ponderada** das médias de cada uma das amostras. Ou seja, seja X e Y duas variáveis aleatórias

$$\bar{X} = \frac{w_1 \cdot X_1 + w_2 \cdot X_2 + \dots + w_n \cdot X_n}{w_1 + w_2 + \dots + w_n}$$

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

Nesse caso, como ambas as amostras têm o mesmo tamanho (mesma quantidade de estudantes realizou ambos os testes), temos que:

$$\text{Média} = (495 + 511)/2 = 503$$

Para encontrar o desvio-padrão, dado que é uma raiz quadrada (relembre a fórmula do desvio-padrão), não podemos apenas somá-los e dividi-los por 2. Para isso, precisamos primeiro passar os desvios para a variância e, então, calcular a raiz quadrada desse valor - de acordo com a fórmula do desvio padrão agrupado:

$$S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$



Novamente, nesse caso, como ambas as amostras têm o mesmo tamanho (mesma quantidade de estudantes realizou ambos os testes), temos que:

$$s_p = \sqrt{(s_1^2 + s_2^2)/2} = \sqrt{(116^2 + 120^2)/2} = 118,0169$$

Dessa forma, para os alunos da população considerando ambas as matérias, tiveram uma média de 503 e o desvio padrão de 118,0169.

## MEDIDAS SEPARATRIZES: QUARTIS E PERCENTIS

Tanto a média como o desvio padrão podem não ser medidas completas para representar um conjunto de dados, pois, além de serem afetados exageradamente por valores extremos, não conseguimos ter uma ideia da simetria da distribuição. Para contornar isso, usamos as medidas separatrizes.

Oseias Dias de Farias

[oseiasdfarias@gmail.com](mailto:oseiasdfarias@gmail.com)  
021.399.242-66

As medidas separatrizes são usadas em estatística para dividir o número total de observações de uma distribuição em certo número de partes iguais. As mais comumente usadas são: Quartis e Percentis. **É importante observar que os dados devem ser classificados em ordem crescente (uso mais comum) ou decrescente antes de calcular os valores da partição.**

Os Quartis dividem os dados em quatro partes iguais; os Decis os dividem em dez partes iguais; e os Percentis os dividem em cem partes iguais. Esses valores de separatrizes são usados para fragmentar uma distribuição em partes menores, tornando-as mais fáceis de medir, analisar e entender.

Os **quartis** dividem um conjunto de dados em **quatro** partes iguais. São três quartis que dividem todos os dados, com um quarto dos valores de dados em cada parte: Primeiro Quartil (Q1), Segundo Quartil (Q2) e Terceiro Quartil (Q3). O percentil refere-se ao percentual de dados totais acumulados em uma determinada porção. Por exemplo, o percentil 25 nos diz que 25% dos dados estão concentrados ali.



Os Q1, Q2 e Q3 também são chamados de quartil inferior, quartil médio (ou mediana) e quartil superior, respectivamente.

Vamos entender o conceito de cada quartil e percentil:

1. O **primeiro quartil (Q1)** separa a primeira parte de um quarto ( $1/4$ ) dos dados da parte superior de três quartos ( $3/4$ ), ou seja, 25% dos dados ficarão abaixo de Q1 e 75% ficarão acima dele. Esse conjunto também é chamado de **percentil 25** (25% dos dados concentrados nesse conjunto)
2. O segundo quartil (Q2) divide os dados em duas partes iguais. Ele separa a primeira metade dos dados da segunda metade, ou seja, 50% dos dados estão abaixo dele e os 50% restantes estão acima dele. O segundo quartil também é chamado de **mediana** dos dados ou de **percentil 50**.
3. O **terceiro quartil (Q3)** separa as três primeiras partes dos dados da última, ou seja, 75% dos dados ficarão abaixo dele e 25% acima dele. O Q3 também é chamado de **percentil 75**.

Suponha que tenhamos os seguintes valores de uma variável X: 15, 5, 3, 8, 10, 2, 7, 11, 12. Aqui temos uma sequência de 9 valores

Ordenando os valores, obtemos as estatísticas de ordem  $x_1 = 2$ ,  $x_2 = 3$ , ...,  $x_9 = 15$ , ou seja, teremos  $2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15$ .

Usando a definição de mediana dada, teremos que mediana = Q2 = 8. Ou seja, 50% dos dados assumem valor até 8.

Suponha que queiramos calcular os dois outros quartis, Q1 e Q3. A ideia é dividir os dados em quatro partes: 2 3 5 7 8 10 11 12 15, em que o número 8 destacado é nossa mediana (número que divide os dados no meio).

Uma possibilidade razoável para obter o Q1 é considerar a mediana dos primeiros quatro valores para , ou seja

$$q_1 = \frac{3 + 5}{2} = 4,$$

Portanto, dizemos que 25% dos dados estão concentrados até o valor 4.

E a mediana dos últimos quatro valores para obter Q3 , ou seja,

$$q_3 = \frac{11 + 12}{2} = 11,5.$$

Portanto, dizemos que 75% dos dados assumem o valor de até 11.5.

Vamos observar agora uma outra sequência X: 15, 5, 3, 8, 10, 2, 7, 11, 12, 67

Ordenando os valores, obtemos as estatísticas de ordem  $x_1 = 2, x_2 = 3, \dots, x_{10} = 67$ , ou seja, teremos  $2 < 3 < 5 < 7 < 8 < 10 < 11 < 12 < 15 < 67$ .

021.399.242-66

Usando a definição de mediana dada, teremos que mediana =  $Q_2 = (8.+ 10)/2 = 9$

Um raciocínio similar ao anterior será aplicado. Porém, agora como temos 2 dados centrais (tivemos que fazer a média para encontrar a mediana), podemos abstrair os valores 8 e 10 para formar um único valor chamado de mediana. Nesse caso, teríamos:

$$2 < 3 < 5 < 7 < 9 < 11 < 12 < 15 < 67$$

Da mesma forma que o anterior Q1 será a mediana a partir do valor 2 até 9, ou seja, o valor central nesses dados.

$$2 < 3 < 5 < 7 < 9$$

Nesse conjunto, o número 5 é o valor central. Portanto,  $Q_1 = 5$ . Ou seja, 25% dos dados assumem até o valor 5.

Analogamente, Q3 será a mediana a partir de 9 até o valor 67.

$$9 < 11 < 12 < 15 < 67$$



Logo, no conjunto 15, 5, 3, 8, 10, 2, 7, 11, 12, 67, o Q3 é 12.

## RESISTÊNCIA DOS DADOS

Dizemos que uma medida de localização ou dispersão é **resistente** quando for pouco afetada por mudanças de uma pequena porção dos dados. A mediana é uma medida resistente, ao passo que a média não o é.

Para ilustrar este fato, considere as populações dos 30 municípios do Brasil. Se descartarmos Rio de Janeiro e São Paulo, a média das populações dos 28 municípios restantes é 100,6 e a mediana é 82,1. Para todos os dados, a média passa a ser 145,4, ao passo que a mediana será 84,3.

Note que a média aumentou bastante, influenciada que foi pelos dois valores maiores, que são muito discrepantes da maioria dos dados. Mas a mediana variou pouco. O desvio padrão também não é uma medida resistente.

Oseias Dias de Farias

Os valores de quartis/percentis também são bastante resistentes e com eles pode-se ter uma boa ideia da simetria da distribuição dos dados.

## SIMETRIA

Devido à resistência dos dados, em vários casos as medidas descritas acima podem não ser suficientes e, então, precisamos de medidas de simetria.

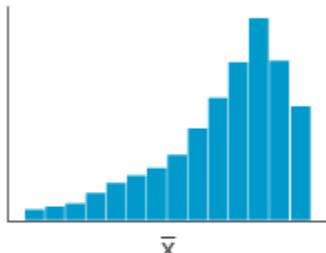
Quando uma distribuição é **simétrica** significa que as observações estão igualmente distribuídas em torno da média (metade acima e metade abaixo).

A assimetria de uma distribuição pode ocorrer de duas formas:

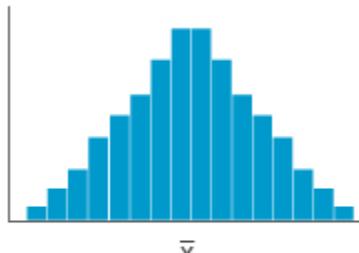
- quando os valores concentram-se à esquerda (assimetria com concentração à esquerda ou assimetria com cauda à direita);
- quando os valores concentram-se à direita (assimetria com concentração à direita ou com assimetria cauda à esquerda);

Ao definir a assimetria de uma distribuição, algumas pessoas preferem se referir ao lado onde está a concentração dos dados. Porém, outras pessoas

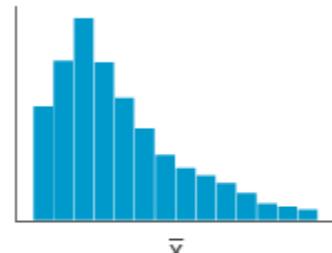
preferem se referir ao lado onde está faltando dados (cauda). As duas denominações são alternativas.



Assimétrica negativa



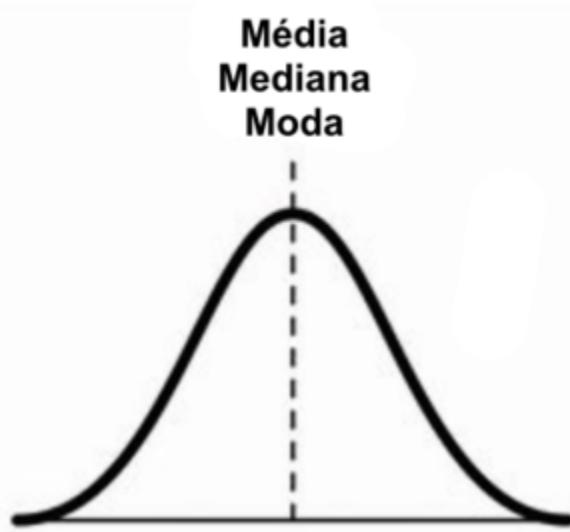
Simétrica



Assimétrica positiva

De maneira geral, encontramos uma perfeita simetria na distribuição normal (que vamos falar mais a frente), em que os dados estão mais concentrados em um ponto central e quanto mais distante da média, menor é a frequência dos dados. Ao traçarmos uma linha no meio da curva teremos dois lados espelhados. Se você tiver acesso às demais medidas descritivas, irá verificar que a média, a mediana e a moda também são iguais.

Oseias Dias de Farias  
oseias.dias@gmail.com  
021.399.242-66



**E por que é importante sabermos se os dados são simétricos ou não?**

Suponha que você faça parte do time de treinamento de uma empresa de telemarketing que está tentando entender a performance geral dos atendentes para dar treinamentos direcionados. Se você se basear apenas em média ou desvio-padrão, pode perder informações valiosas.

Por exemplo, se você plotar o gráfico de performance para os atendentes na forma de histograma e obtiver uma curva assimétrica negativa, poderá constatar que há poucos atendentes que possuem uma baixa performance quando comparado com a maioria e que provavelmente estão causando grandes desvios e baixas médias de performance geral. Uma possível solução de business para isso seria dar treinamentos para essas pessoas específicas ao invés de fazê-lo para o time inteiro.

Por outro lado, se você plotar o gráfico performance e constatar uma assimetria positiva, verá que a maioria dos seus atendentes tem uma performance baixa. Com isso, poderá pensar em medidas de melhoria de processo ou até entender se aqueles com performance muito elevada estão trapaceando o sistema de alguma forma (isso infelizmente acontece e como convededores de dados precisamos analisá-los de forma completa).

Mas será que precisamos sempre olhar os gráficos para entender isso? Nem sempre! Existem métricas específicas que traduzem assimetrias de uma forma numérica. Isso é muito útil especialmente se você tiver tantas métricas de performance e tantos times que olhar gráficos um a um seria quase inviável.

## SKEW (ASSIMETRIA)

Uma das formas adotadas para o cálculo da assimetria é através do **coeficiente de assimetria de Fisher**. O coeficiente vem a partir do terceiro momento de ordem superior em torno da média através de uma função geradora de momentos. Momentos são medidas resumo de uma distribuição, sendo 1º momento = média (valor esperado), 2º momento = variância, 3º momento = assimetria e 4º momento = curtose.

O Skew (assimetria) é medido por:

$$\text{Skew} = \frac{n}{(n-1)(n-2)} \sum \left( \frac{x_i - \bar{x}}{s} \right)^3$$

A interpretação, será:

- Skew = 0, a distribuição é simétrica;
- Skew > 0, a distribuição é assimétrica positiva (à direita);
- Skew < 0, a distribuição é assimétrica negativa (à esquerda).

Dizemos que os dados são aproximadamente normais (e, portanto, simétricos) se  $-1 < \text{Skew} < 1$ .

Vamos supor que para o caso dos atendentes de telemarketing o skew = 2,78. Podemos fazer a seguinte interpretação: o sinal positivo significa que a distribuição é assimétrica à direita e como 2,78 é maior que o intervalo de referência ( $-1 < \text{Skew} < 1$ ), os dados apresentam alto grau de assimetria.

**NOTA IMPORTANTE: PODEMOS USAR A FUNÇÃO SKEW PARA CALCULAR NO GOOGLE SHEETS**

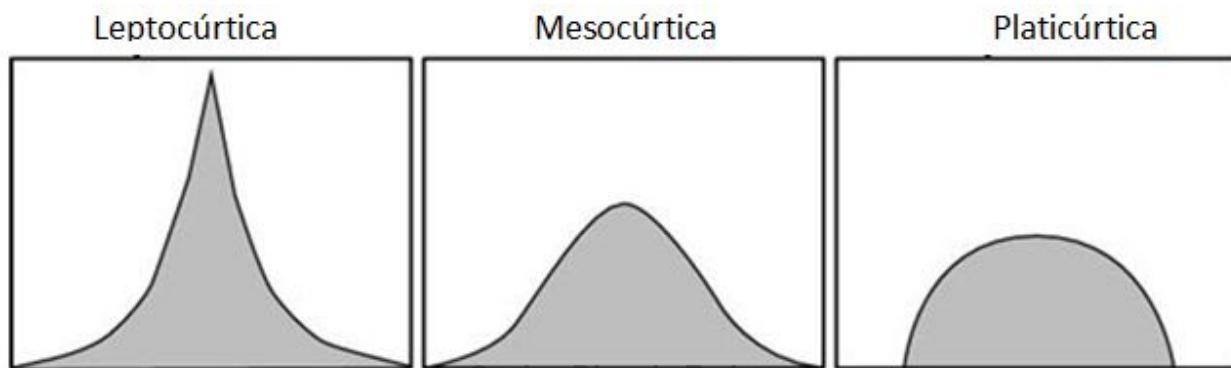
## CURTOSE

A curtose (kurtosis em inglês) representa o grau de achatamento da distribuição, isto é, quanto espalhados os dados estão em torno da média. Novamente, usamos a curva normal padrão como referência e podemos interpretar a curtose por meio de gráficos ou numericamente. Pode ser classificada em três tipos:

a) Mesocúrtica: que é própria curva normal padrão



- b) Platicúrtica: possui grau de achatamento maior que da curva normal padrão, o que nos indica que os dados estão mais espalhados (logo, o desvio padrão também é maior).
- c) Leptocúrtica: seu grau de achatamento é menor que o da curva normal padrão (curva mais pontiaguda), indica que os dados estão mais concentrados (desvio padrão menor)



Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

A curtose pode ser calculada pelo coeficiente de curtose de Fisher, que neste caso utiliza o quarto momento de ordem superior ao redor da média:

$$\text{Kurtosis} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left( \frac{x_i - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

Assim, se:

- Curtose = 0, a curva é normal padrão, isto é, mesocúrtica
- Curtose > 0, grau de achatamento baixo, a curva é leptocúrtica
- Curtose < 0, grau de achatamento alto, a curva é platicúrtica

Em alguns programas estatísticos, como o STATA, é comum encontrar a curtose da distribuição normal como  $K = 3$ . Neste caso, a interpretação é a mesma. Isto é:

- $K = 3$ , curva normal padrão
- $K > 3$ , curva leptocúrtica
- $K < 3$ , curva platicúrtica

Como interpretar na prática? Vamos supor que para o caso dos atendentes de telemarketing encontramos curtose = 10,82, o que nos indica que a curva é leptocúrtica, isto é, é menos achatada que a curva normal – o pico da distribuição é mais acentuado – então sabemos que os dados estão mais concentrados em um determinado ponto.

**NOTA IMPORTANTE:** PODEMOS USAR A FUNÇÃO KURT PARA CALCULAR NO GOOGLE SHEETS

## BOXPLOTS

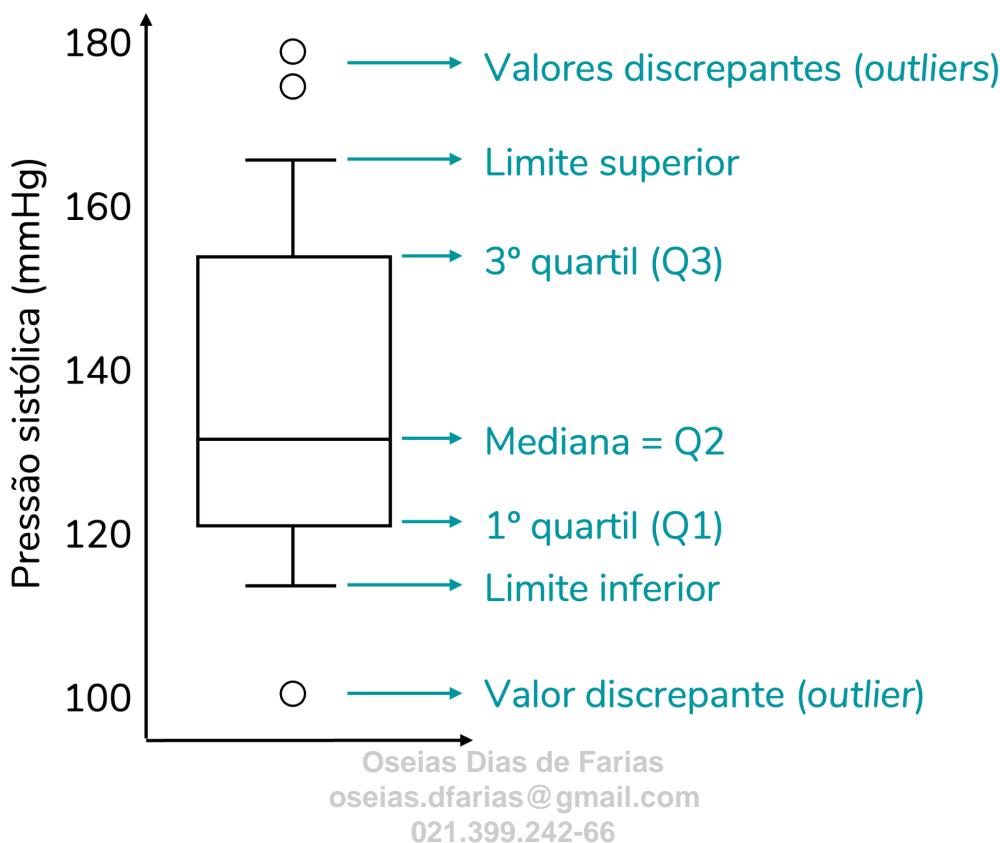
Outra forma muito interessante para entendermos a distribuição e simetria dos dados é através dos quartis e boxplots/. Vamos supor uma distribuição com vários números. Suponhamos que ordenamos de forma crescente esses números, da mesma forma que fizemos no exemplo dos quartis. O número de menor valor será chamado de  $x_1$ , o número de maior valor será chamado de  $x_n$  e os quartis serão chamados de  $q_1$ ,  $q_2$ ,  $q_3$ . Consideraremos que os **dados são simétricos** quando:

- a)  $q_2 - x_1 = x_n - q_2$ ;
- b)  $q_2 - q_1 = q_3 - q_2$ ;
- c)  $q_1 - x_1 = x_n - q_3$ ;

A diferença  $q_2 - x_1$  é chamada **dispersão inferior** e  $x_n - q_2$  é a **dispersão superior**. A condição (a) nos diz que estas duas dispersões devem ser aproximadamente iguais, para uma distribuição aproximadamente simétrica.

Podemos representar graficamente essas informações de quartis em um boxplot. No exemplo abaixo coletamos a pressão sistólica de 100 pessoas e plotamos os dados em um boxplot.





Notem que valores discrepantes (outliers) são representados por bolinhas e eles não entram na conta dos quartis. **O "limite inferior" para o boxplot não é o valor mínimo que temos nesse conjunto para os casos em que o valor mínimo do nosso conjunto seja um outlier.** Isso **não** significa que, caso queiramos escrever as estatísticas como mínimo, máximo, média, mediana e quartis, nós não consideramos esse outlier. A descrição das estatísticas do conjunto sempre vai considerar todos os valores, porém a representação no boxplot tem essa particularidade.

E como esse outlier é calculado? A partir dos cálculos de limite inferior e superior.

- Limite Inferior = Primeiro Quartil – 1,5 \* (Terceiro Quartil – Primeiro Quartil)
- Limite Superior = Terceiro Quartil + 1,5 \* (Terceiro Quartil – Primeiro Quartil)

Caso haja um dado menor do que o limite inferior ou maior que o limite superior, o boxplot considera isso um outlier.

Para exemplificar, vamos considerar que coletamos as notas de 12 alunos de um curso. Ordenamos essas idades em ordem crescente para facilitar a explicação.

Posição	Idade
1 <sup>a</sup>	18
2 <sup>a</sup>	19
3 <sup>a</sup>	21
4 <sup>a</sup>	21
5 <sup>a</sup>	21
6 <sup>a</sup>	22
7 <sup>a</sup>	22
8 <sup>a</sup>	22
9 <sup>a</sup>	23
10 <sup>a</sup>	23
11 <sup>a</sup>	24
12 <sup>a</sup>	27

Oseias Dias de Farias  
oseias.dfarias@gmail.com

021.399.242-66

O primeiro passo é calcular os quartis/percentis. Para encontrar o percentil 25, primeiramente precisamos encontrar em qual posição devemos buscar o valor. Podemos chegar a essa posição, multiplicando o percentil que queremos pelo tamanho da amostra e dividindo por 100.

Posição do Percentil 25 = Percentil \* Tamanho da Amostra / 100 =  $25 * 12 / 100 = 300/100 = 3$

Na posição 3, temos a idade de 21 anos. Sendo assim, o percentil 25 dessa amostra é 21 anos. Isso significa que pelo menos 25% dos indivíduos dessa amostra têm no máximo 21 anos.

E se o cálculo da posição de determinado percentil não resultar em um número inteiro? Podemos tirar a média dos dois valores intermediários, como mostramos no exemplo dos quartis. Lembre-se, dificilmente faremos esses cálculos "na mão" pois usaremos ferramentas como Excel e Python. Porém, para que haja um entendimento completo dos conceitos, é importante que vocês vejam como o cálculo é realizado.

Fazendo os cálculos dos outros quartis temos:

Variável	Mínimo	1º Quartil	2º Quartil	3º Quartil	Máximo
Idade	18	21	Oseias Dia oseias.dfarias@gmail.com 021.399.242-66	23	27

Agora calculando os limites para o boxplot:

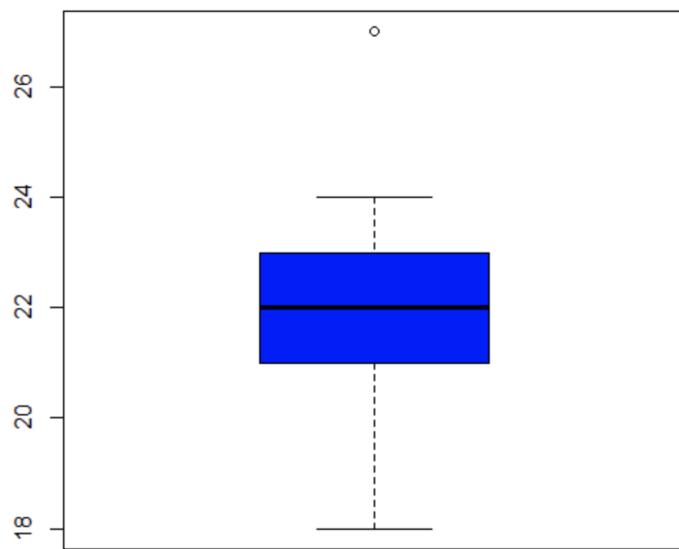
$$\text{Limite inferior} = \text{Primeiro Quartil} - 1,5 * (\text{Terceiro Quartil} - \text{Primeiro Quartil}) = 21 - 1,5 * (23 - 21) = 18$$

$$\text{Limite superior} = \text{Terceiro Quartil} + 1,5 * (\text{Terceiro Quartil} - \text{Primeiro Quartil}) = 23 + 1,5 * (23 - 21) = 26$$

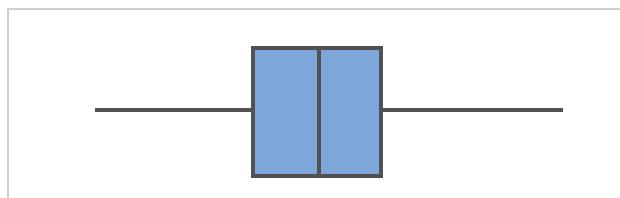
Vemos aqui que não temos outliers inferiores (nossa idade mínima é 18, que coincide com o limite inferior. Porém, por outro lado, o limite superior é 26, e a idade máxima é 27. Ou seja, teremos outliers superiores.

No gráfico Boxplot temos:

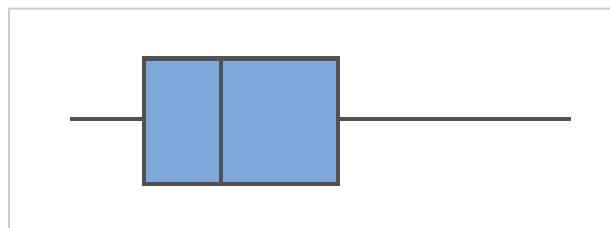




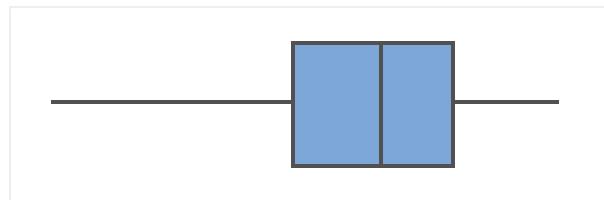
Também usamos o boxplot para entender a simetria dos nossos dados. Um conjunto de dados que tem uma distribuição **simétrica**, terá a linha da mediana no centro do retângulo, como o boxplot abaixo:



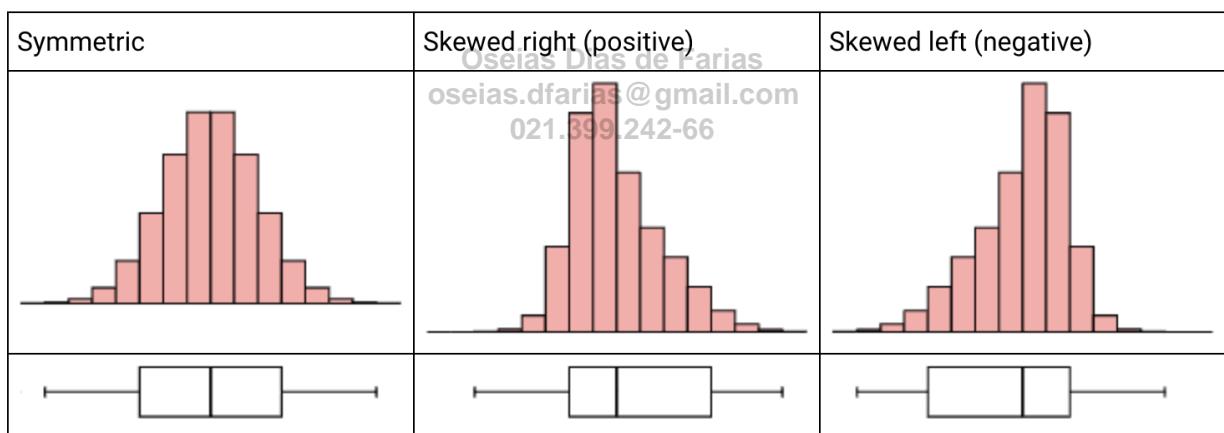
Quando a linha da mediana está próxima ao primeiro quartil, ou seja, quando existe uma cauda mais longa em números maiores (Q3 ou limite superior alongados), os dados são **assimétricos positivos**



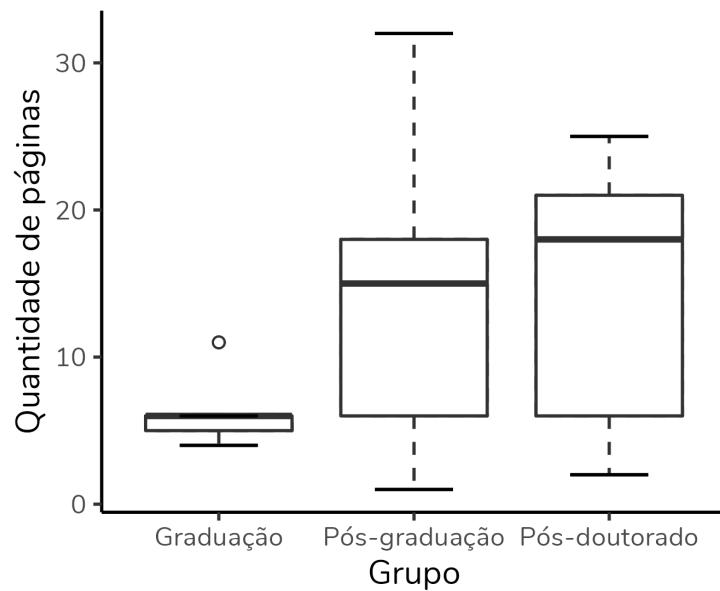
Quando a linha da mediana está próxima ao terceiro quartil, ou seja, quando existe uma cauda mais longa em números menores (Q1 ou limite inferior alongados), os dados são **assimétricos negativos**.



Também conseguimos fazer esse mesmo paralelo com histogramas



Os Boxplots são extremamente úteis para analisarmos nossos dados. Abaixo temos um 3 boxplots comparativos sobre a quantidade de páginas que alunos costumam ler quando estão na graduação, pós-graduação e pós-doutorado.



Baseado em todas as informações anteriores, qual informação você tira sobre esses dados?

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66



# 7. Introdução a probabilidade



A **teoria da probabilidade** é o campo da matemática que estuda experimentos ou fenômenos **aleatórios** e através dela é possível analisar as chances de um determinado evento ocorrer.

Muitos eventos não podem ser previstos com total certeza. Podemos prever apenas a chance de um evento ocorrer, ou seja, qual a probabilidade de acontecer, usando-o. A probabilidade pode variar de 0 a 1 (ou 0 a 100% dependendo da denotação), onde 0 significa que o evento é impossível e 1 indica que definitivamente aquilo vai acontecer.

Nas seções anteriores, falamos bastante sobre a distribuição de frequências e sua importância para avaliarmos a variabilidade das observações. A partir dessas frequentes podemos calcular medidas de posição e variabilidade, como média, mediana, desvio padrão etc. Essas frequências (relativas) são estimativas de probabilidades de ocorrências de certos eventos de interesse.

De maneira geral, a fórmula da probabilidade é:

$$P(A) = \frac{\text{Nº DE RESULTADOS FAVORÁVEIS}}{\text{Nº DE RESULTADOS POSSÍVEIS}}$$



Sendo A o evento que queremos prever. O numerador da equação acima representa a quantidade de vezes que acontece o evento A (resultados favoráveis) e o denominador representa a quantidade de espaços amostrais que temos (todos os resultados possíveis). Vamos a um exemplo

Há 6 travesseiros em uma cama, 3 são vermelhos, 2 são amarelos e 1 é azul. Qual é a probabilidade de escolher um travesseiro amarelo?

Resposta:

A probabilidade é igual ao número de travesseiros amarelos na cama dividido pelo número total de travesseiros, ou seja,  $2/6 = 1/3$ .

## EXPERIMENTO ALEATÓRIO

Um **experimento aleatório** é aquele que não é possível conhecer qual resultado será encontrado antes de realizá-lo. Os acontecimentos deste tipo quando repetidos nas mesmas condições, podem dar resultados diferentes e essa inconstância é atribuída ao acaso. Um exemplo de experimento aleatório é jogar um dado não viciado (dado que apresenta uma distribuição homogênea de massa) para o alto. Ao cair, não é possível prever com total certeza qual das 6 faces estará voltada para cima.

## ESPAÇO AMOSTRAL

Representado pela letra  $\Omega$  (ômega), o **espaço amostral** corresponde ao conjunto de todos os pontos amostrais, ou, resultados possíveis obtidos a partir de um experimento aleatório.

Por exemplo, ao retirar ao acaso uma carta de um baralho, o espaço amostral corresponde às 52 cartas que compõem este baralho.

No caso do dado, se o lançarmos para cima, qualquer uma das 6 faces podem ocorrer. Ou seja, nosso espaço amostral seria as 6 faces desse dado (face 1, face 2, ..., face 6). Nesse caso, a denotação do espaço amostral é:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$



Vamos a mais um exemplo para fixar bem esse conceito. Se lançarmos uma moeda para cima, podemos obter cara ou coroa. Nesse caso, nosso espaço amostral é:

$$\Omega = \{\text{cara, coroa}\}$$

Para facilitar, vou apelidar "cara" com a letra H e "coroa" com a letra T. Ou seja, teríamos:

$$\Omega = \{H, T\}$$

E se lançarmos 2 moedas simultaneamente para cima, qual seria nosso espaço amostral? Nesse caso, podemos obter cara em uma coroa em outra, cara em uma, cara em outra... e assim por diante. Logo, denotamos o espaço amostral como sendo:

$$\Omega = \{HT, HH, TH, TT\}$$

Agora vamos a um exemplo um pouco diferente. Considere o experimento que consiste em retirar uma lâmpada de um lote e medir seu "tempo de vida" antes de se queimar. Um espaço amostral conveniente é

$$\Omega = \{t \in \mathbb{R} : t \geq 0\}$$

Estranho né? Vou explicar. Primeiro, o símbolo  $\in$  significa "pertence" e o símbolo  $\mathbb{R}$  significa "reais" - ou seja, números reais. Quando falamos  $t \in \mathbb{R}$  dizemos que  $t$  (nossa variável de tempo) pertence aos reais, portanto,  $t$  pode ser qualquer valor desde que esteja no conjunto dos números reais. Também dizemos que  $t \geq 0$ , uma vez que aqui estamos interessados somente nos números reais não negativos (não faz sentido ter um tempo de vida negativo).

Se tivermos um evento A que indica "o tempo de vida da lâmpada é inferior a 20 horas", então

$$A = \{t : 0 \leq t \leq 20\}.$$

Ou seja, de todo o espaço amostral possível (qualquer valor de tempo possível), estaríamos interessados em encontrar quando o tempo é menor que 20 horas. Como o tempo nunca é negativo,  $t$  então tem que estar entre 0 a 20 horas. Esse é um exemplo de um espaço amostral contínuo, contrastado com os anteriores, que são discretos.



## TIPOS DE EVENTOS

Um evento, como já explicado, é qualquer subconjunto do espaço amostral de um experimento aleatório.

### EVENTO IMPOSSÍVEL

O conjunto do evento é vazio.

Exemplo: jogar um dado para o alto e tirar um número 8

### EVENTO COMPLEMENTAR

Os conjuntos de dois eventos formam todo o espaço amostral, sendo um evento complementar ao outro.

Exemplo: No experimento de lançar uma moeda, o espaço amostral é  $\Omega = \{\text{cara, coroa}\}$ .

Seja o evento A sair cara,  $A=\{\text{cara}\}$ , o evento B sair coroa é complementar ao evento A, pois,  $B=\{\text{coroa}\}$ . Juntos formam o próprio espaço amostral.

### EVENTO MUTUAMENTE EXCLUSIVO

Os conjuntos dos eventos não possuem elementos em comum. A intersecção entre os dois conjuntos é vazia.

Exemplo: Seja o experimento lançar um dado, os seguintes eventos são mutuamente exclusivos

A: ocorrer um número menor que 5,  $A=\{1, 2, 3, 4\}$

B: ocorrer um número maior que 5,  $A=\{6\}$

### EVENTOS DEPENDENTES E INDEPENDENTES

Em probabilidade, dizemos que dois **eventos são independentes** quando o fato de saber que um evento ocorreu não altera a probabilidade do outro evento.

**Eventos dependentes** são eventos que, caso um aconteça, a probabilidade do outro evento acontecer muda. Vamos pegar um exemplo bem simples.



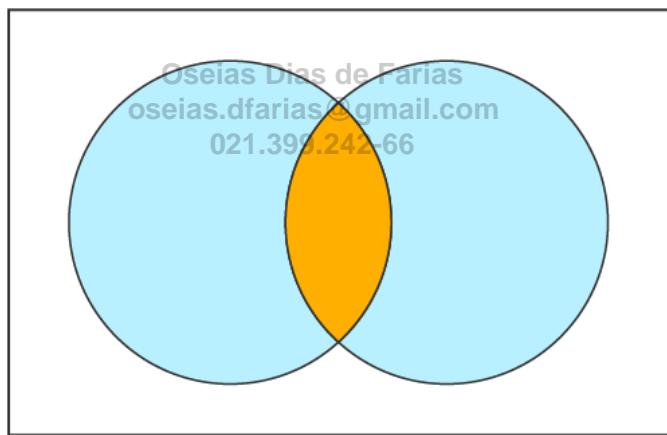
Quando você vai pegar o seu ônibus na parada, você precisa sair de casa horário pré determinado por você (evento A) para chegar no trabalho às 8h da manhã (evento B). Ou seja, caso você saia de casa em outro determinado horário, a probabilidade de você chegar no trabalho às 8h da manhã muda.

## PROBABILIDADE DA INTERSECÇÃO DE EVENTOS

Para dois eventos **independentes** a probabilidade da intersecção de dois eventos envolve a chance de o evento A **E** o evento B ocorrer. O cálculo é feito por:

$$P(A \cap B) = P(A) \times P(B)$$

$$P(A \cap B) = P(A) \times P(B)$$



Para **eventos dependentes**, a probabilidade passa a ser:

$$P(A \cap B) = P(A) \times P(B|A)$$

Em que  $P(B|A)$  é a probabilidade de um evento B acontecer dado que A já aconteceu - ou seja, a probabilidade de você chegar ao trabalho às 8h da manhã dado que você saiu de casa em um outro horário. Esse tipo de probabilidade é chamada de **probabilidade condicional** e vamos falar sobre falar sobre ela mais a frente

Vamos a um exemplo:

Qual é a probabilidade de selecionar uma carta vermelha **e** um 6 quando uma carta é selecionada aleatoriamente de um baralho de 52 cartas?

*Resposta:*

Sejam A e B as probabilidades individuais de obter uma carta vermelha e um 6, respectivamente.

Sabemos que o número de cartas vermelhos é 26 e que o número de cartas número 6 é 4

A probabilidade de obter uma carta vermelha de um baralho de 52 cartas,  $P(A) = 26/52$

A probabilidade de obter um 6 de um baralho de 52 cartas,  $P(B) = 4/52$

Usando a fórmula  $P(A \cap B)$ ,

$$P(A \cap B) = P(A) \times P(B)$$

Oseias Dias de Farias

$$P(A \cap B) = 26/52 \times 4/52 = 1/26 = 0,038$$

021.399.242-66

Ou seja, a probabilidade de retirarmos uma carta vermelha que seja 6 é de 3,8%..

## PROBABILIDADE DA UNIÃO DE DOIS EVENTOS

A probabilidade da união de dois eventos envolve a chance de o evento A **OU** de o evento B ocorrer. O cálculo é feito por:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$P(A \cap B)$  é a probabilidade da intersecção entre A e B.

Podemos dizer, então, que a probabilidade da união de dois eventos é calculada pela probabilidade do primeiro evento ocorrer mais a probabilidade do segundo evento ocorrer menos a probabilidade da intersecção de ambos, sendo que a probabilidade da intersecção de dois eventos é igual à probabilidade do primeiro e do segundo evento ocorrerem simultaneamente.



## Vamos a um exemplo

Uma urna contém 20 bolas numeradas de 1 a 20. Quando uma bola é retirada ao acaso, qual é a probabilidade do número ser múltiplo de 3 **ou** de 5?

Resposta:

- Evento A: Múltiplos de 3 no espaço amostral de 1 a 20: {3, 6, 9, 12, 15, 18}

Ou seja, de um total de 20 bolinhas, 6 são múltiplas de 3

$$P(A) = \text{probabilidade de ser múltiplo de } 3 = 6/20 = 0.3$$

- Evento B: Múltiplos de 5 no espaço amostral de 1 a 20: {5, 10, 15, 20}

Ou seja, de um total de 20 bolinhas, 4 são múltiplas de 5

$$P(B) = \text{probabilidade de ser múltiplo de } 5 = 4/20 = 0.2$$

- Evento  $A \cap B$ : Múltiplos de 3 e 5 simultaneamente no espaço amostral de 1 a 20: {15}

*Oscias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66*

$$P(A \cap B) = \text{probabilidade de ser múltiplo de 3 e 5 simultaneamente} = 1/20 = 0.05$$

Logo, a probabilidade de ser múltiplo de 3 ou de 5 é de:

$$P(A \cup B) = 0.3 + 0.2 - 0.05 = 0.45$$

Logo, há 45% de chance de tirarmos um múltiplo de 3 ou de 5.

## Vamos a um segundo exemplo

Considere o experimento: lançamento de um dado. Qual a probabilidade de sair um número maior que 5 **ou** um número ímpar?

Resposta:

- Evento A: Número maior que 5: {6}

Ou seja, de um total de 6 números (dado tem 6 faces), 1 número é maior que 5.



$P(A)$  = probabilidade de ser maior que 6 =  $1/6 = 0.166$

- Evento B: Número ímpar: {1, 3, 5}

Ou seja, de um total de 6 números, 3 são ímpares

$P(B)$  = probabilidade de ser ímpar =  $3/6 = 0.5$

- Evento  $A \cap B$ : Maior que 5 e ímpar: não existe

Nenhum valor é maior que 5 e é ímpar simultaneamente pois, o único valor maior que 5 é o número 6

$P(A \cap B) = 0$

Logo, a probabilidade de ser múltiplo de 3 ou de 5 é de:

$$P(A \cup B) = 0.166 + 0.5 - 0 = 0.666$$

Chamamos esse tipo de evento em que a probabilidade de intersecção dos eventos é 0 de **eventos mutuamente exclusivos**

021.399.242-66

## PROBABILIDADE CONDICIONAL

Se a probabilidade de ocorrência de um evento B interfere na probabilidade de ocorrência de um evento A, então dizemos que a probabilidade de A está **condicionada** à probabilidade de B e representamos por  $P(B/A)$ . Lê-se: probabilidade de A dado B. A probabilidade condicional pode ser calculada por:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Onde,

$P(A \cap B)$  é a probabilidade da interseção entre A e B.

$P(B)$  é a probabilidade do evento B.

Vamos a um exemplo



Dois dados são lançados ao acaso. Qual a probabilidade da soma ser igual a 6, dado que o primeiro dado saiu número menor que 3.

Resposta:

Aqui temos um exemplo clássico de probabilidade condicional, em que:

Evento A: “a soma ser igual a 6”

Evento B: “o primeiro dado é menor 3”

- $A = \{\text{soma igual a } 6\} = \{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\}$  -> 5 combinações

$P(A)$  - de um total de 36 combinações (6 faces de 2 dados), a soma pode ser 6 em apenas 5 casos.

Oseias Dias de Farias  
[oseias\\_dias\\_farias@gmail.com](mailto:oseias_dias_farias@gmail.com)  
021 399 242 66

Apesar de não ser importante para a probabilidade de A na fórmula de  $P(A|B)$ , saber o espaço amostral será útil para calcular  $P(A \cap B)$ . Veremos abaixo.

- $B = \{\text{primeiro dado} < 3\} = \{(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6), (2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6)\}$  -> 12 combinações

$P(B)$  - de um total de 36 combinações, ele é menor que 3 em apenas 12 casos. Ou seja,  $P(B) = 12/36 = 0.3$

- $P(A \cap B)$  = probabilidade da soma ser 6 e o primeiro dado ter um valor menor que 3.

Observando o espaço amostral de A, vemos que o primeiro dado é menor que 3 apenas para os casos (1,5) e (2,4). Ou seja, para todas as 36 combinações possíveis, somente em 2 casos essa condição será satisfeita. Ou seja,  $P(A \cap B) = 2/36 = 0.0555$

Logo,  $P(A|B) = 0.055/0.3 = 0.183$



Portanto, a probabilidade da soma ser 6 ao lançarmos dois dados uma vez que o primeiro dado tem valor menor que 3 é de 0.183.

### Vamos a um segundo exemplo

Duas cartas são selecionadas, sem reposição da primeira carta, de um baralho normal de 52 cartas. Encontre a probabilidade de selecionar um rei e depois uma rainha.

*Resposta:*

Queremos a intersecção de dois eventos (A e B acontecerem) e são eventos dependentes (seleção uma carta em seguida da outra sem reposição)

A: Selecionar um rei

B: Selecionar uma rainha

$$P(A \cap B) = P(A) \times P(B|A)$$

Oseias Dias de Farias

[oseias\\_dfarias@gmail.com](mailto:oseias_dfarias@gmail.com)  
021.399.242-66

Existem 4 reis e 4 rainhas no baralho. Na primeira retirada, temos 52 cartas. Logo, a probabilidade do evento A é:

$$P(A) = 4/52$$

Depois de retirar uma carta, restam 51 cartas. Dado que o primeiro evento aconteceu (retirar rei), continuam sobrando 4 rainhas.

$$P(B|A) = 4/51$$

$$\text{Logo, } P(A \cap B) = 4/52 \times 4/51 = 0,006$$

## TEOREMA DA PROBABILIDADE TOTAL

A definição formal do teorema da probabilidade total é a seguinte.

Seja um evento do espaço amostral  $\Omega$ , e  $\{A_i : i = 1, 2, 3, \dots, n\}$  um conjunto de eventos distintos cuja união é todo o espaço amostral, ou seja  $A_i$  é uma partição do espaço amostral  $\Omega$ . A probabilidade do evento pode ser calculada como segue:



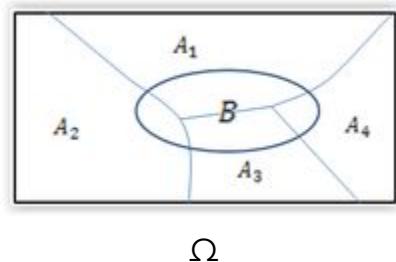
$$P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i)$$

Não entendeu nada? Vamos traduzir!

Primeira coisa que precisamos entender é que temos um espaço  $\Omega$  que é formado por vários eventos  $A_1, A_2, A_3, \dots$ , todos mutuamente exclusivos.



Agora, temos um evento  $B$  nesse espaço amostral que depende dos eventos anteriores



Podemos reescrever esse evento  $B$  de forma que:

$$B = (A_1 \cap B) + (A_2 \cap B) + (A_3 \cap B) + (A_4 \cap B)$$

Logo,

$$P(B) = P(A1 \cap B) + P(A2 \cap B) + P(A3 \cap B) + P(A4 \cap B)$$

Usando a probabilidade condicional  $P(Ai \cap B) = P(Ai) \times P(B|Ai)$ , reescrevemos a fórmula acima de forma que:

$$\sum_{i=1}^4 P(B|A_i) \cdot P(A_i)$$

Generalizando, temos que:

$$P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i)$$

Oseias Dias de Farias  
oseias.dfarias@gmail.com

Agora vamos a um exemplo de uso:

Considere um processo de fabricação de semicondutores. Nesses processos, quando um chip está sujeito a altos níveis de contaminação, a probabilidade de que ele cause defeito na produção é de 0,1. Se o chip não está sujeito a altos níveis de contaminação, a probabilidade dele causar defeito na produção é 0,005. Sabemos ainda que a probabilidade de um chip estar sob altos níveis de contaminação é 0,2. Estamos interessados no evento: o chip causa uma falha na produção (geral, estando ou não sob efeito de contaminação).

As condições a que esse evento está sujeito são: está sujeito a altos níveis de contaminação; não está sujeito a altos níveis de contaminação.

*Resposta:*



Nesse nosso caso,  $\Omega$  será composto por  $A_1 = \text{está sujeito a altos níveis de contaminação}$  e  $A_2 = \text{não está sujeito a altos níveis de contaminação}$ . Nossa evento de interesse é  $B = \text{o chip causa uma falha na produção}$ .

Usando a fórmula anterior, temos que:

$$P(B) = P(B|A_1)*P(A_1) + P(B|A_2)*P(A_2)$$

Sabemos que:

$P(B|A_1) = 0,1$  - Probabilidade do chip causar falha na produção dado que está sujeito a altos níveis de contaminação

$P(B|A_2) = 0,005$  - Probabilidade do chip causar falha na produção dado que não está sujeito a altos níveis de contaminação

$P(A_1) = 0,2$  - Probabilidade de um chip estar sob altos níveis de contaminação

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242-66

Agora vamos usar a lógica. O chip está ou não está sujeito a altos níveis de contaminação - não há um intermediário. Ou seja, a probabilidade de ele estar **OU** não estar sujeito a altos níveis de contaminação é 100%, uma vez que não existe outra possibilidade. Usamos a letra  $\Omega$  para indicar todo nosso espaço amostral possível. Logo

$$\Omega = A_1 \cup A_2$$

Usando a regra da união de dois eventos:

$$P(\Omega) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$$

$P(A_1 \cap A_2)$  é zero, uma vez que não existe uma opção de estar **E** não estar sob efeitos de alta contaminação.  $P(\Omega)$  é 100% (ou seja, 1), uma vez que é nosso espaço amostral inteiro. Como já dito,  $P(A_1) = 0,2$ . Então temos:

$$1 = 0,2 + P(A_2) \rightarrow P(A_2) = 1 - 0,2 = 0,8$$



Portanto, há 80% chance do chip não estar sujeito a altos níveis de contaminação. Agora, podemos voltar a nossa fórmula:

$$P(B) = P(B|A1)*P(A1) + P(B|A2)*P(A2) = 0,1*0,2 + 0,005*0,8 = 0,0235 = 2,35\%$$

Isso quer dizer que existe 2,35% de probabilidade do chip causar uma falha na produção no geral, considerando todos os cenários (contaminação ou não).

## TEOREMA DE BAYES

Uma das relações mais importantes envolvendo probabilidades condicionais é dada pelo Teorema de Bayes. A versão mais simples desse teorema é dada pela fórmula:

$$P(A | B) = \frac{P(B | A) \cdot P(A)}{P(B)}$$

Oseias Dias de Farias  
oseias.dfarias@gmail.com

Vimos que a informação muitas vezes é apresentada em forma de probabilidade condicional. Elas nos fornecem a probabilidade de um evento (no caso do chip, uma falha no processo) dada uma condição (estar contaminado). Pode ser que estejamos interessados em investigar: depois que o evento deu um resultado (falha no processo); qual a probabilidade de uma certa condição estar presente (alta contaminação)?

$P(A1|B)$  = Probabilidade do chip estar contaminado dado que aconteceu uma falha na produção.

**IMPORTANTE:** não confunda com  $P(B|A1) = 0,1$  - Probabilidade do chip causar falha na produção dado que está sujeito a altos níveis de contaminação

Usando o teorema de Bayes temos que:

$$P(A1|B) = P(B|A1)*P(A1)/P(B) = 0,1*0,2/0,0235 = 0,85 = 85\%.$$

Logo, existe 85% de probabilidade do chip estar contaminado dado que aconteceu uma falha na produção.

### Vamos a um outro exemplo:

Uma empresa oferece aos candidatos a uma vaga um curso de treinamento durante uma semana. No final do curso, eles são submetidos a uma prova e 25% são classificados como bons (B), 50% como médios (M) e os restantes 25% como fracos (F). A empresa pretende substituir o treinamento por um teste contendo questões referentes a conhecimentos gerais e específicos. Para isso, gostaria de conhecer qual a probabilidade de um indivíduo aprovado no teste ser considerado fraco, caso tivesse feito o curso. Assim, neste ano, fizeram um experimento. Antes do início do curso, os candidatos foram submetidos ao teste e receberam o conceito aprovado (A) ou reprovado (R). Posteriormente, fizeram o curso e, ao final, obtiveram-se as seguintes probabilidades condicionais:

$P(A|B) = 0,80$  - Probabilidade de ser aprovado dado que é "bom" (são as pessoas com resultado "bom" previamente que foram aprovadas)

$P(A|M) = 0,50$  - Probabilidade de ser aprovado dado que é "médio" (são as pessoas com resultado "médio" previamente que foram aprovadas)

$P(A|F) = 0,20$  - Probabilidade de ser aprovado dado que é "fraco" (são as pessoas com resultado "fraco" previamente que foram aprovadas)

Agora, a empresa quer entender qual seria a probabilidade de ser "fraco" dado que foi aprovado, ou seja  $P(F|A)$

Evento A: Probabilidade de ser aprovado

Evento F: Probabilidade de ser "fraco"

*Resposta:*

Usando teorema de Bayes temos que:

$$P(F|A) = P(A|F)*P(F)/P(A)$$

A probabilidade de ser "fraco" já foi dita no próprio enunciado  $\rightarrow P(F) = 0,25$ . Também temos  $P(A|F) = 0,2$ . Porém, qual seria  $P(A)$ ?

Note que a probabilidade de ser aprovado é uma probabilidade que ainda não temos. Porém, temos a probabilidade de ser aprovado dado um determinado cenário anterior: Aprovado dado que era bom, aprovado dado que era médio e aprovado dado que era fraco. Sabendo que só existem essas 3 possibilidades anteriores (ser bom, médio ou fraco), podemos usar aqui o teorema da probabilidade total.

$$P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i)$$

$$P(A) = P(A|B)*P(B) + P(A|M)*P(M) + P(A|F)*P(F) = 0,8*0,25 + 0,5*0,5 + 0,2*0,25 = 0,5.$$

Ou seja, a probabilidade de ser aprovado é de 50%.

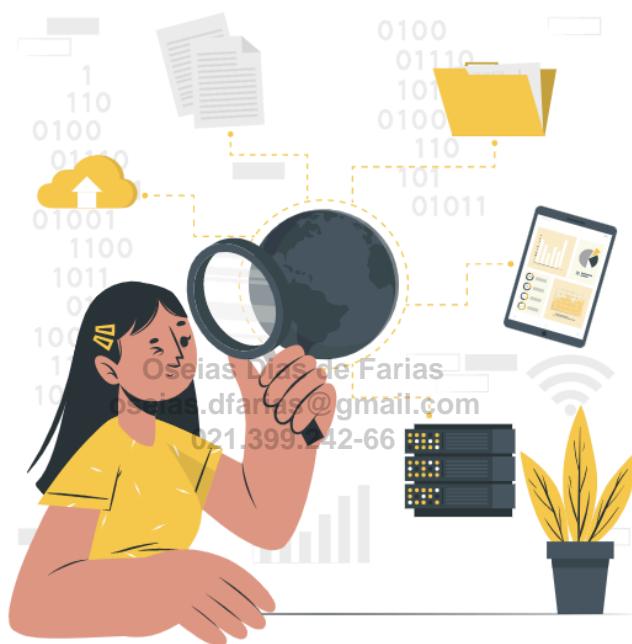
Voltando no teorema de Bayes:  
Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66

$$P(F|A) = P(A|F)*P(F)/P(A) = 0,2*0,25/0,5 = 0,1 = 10\%$$

Portanto, existe 10% de probabilidade do candidato ser "fraco" dado que foi aprovado (se eu escolhesse uma pessoa aleatória dentre os aprovados, haveria probabilidade de 10% de ela ser fraca)



# 8. Distribuições discretas e contínuas de probabilidade



No capítulo anterior vimos o que é um espaço amostral e como calcular a probabilidade de um evento. Neste capítulo vamos ampliar esses conceitos associando espaço amostral e eventos a valores numéricos.

Antes de iniciarmos, precisamos introduzir um conceito importante. Uma **variável aleatória** (abreviadamente, v.a.) é uma função que associa a cada elemento de um espaço amostral um número real.

Na prática, usualmente não existe a preocupação de se explicitar qual é o espaço amostral no qual está definida a variável aleatória. O que importa é definir o conjunto de valores reais que a variável pode admitir e explicitar como se calcula a probabilidade de que ela admita tais valores.

O conceito de variável aleatória é particularmente útil em situações nas quais se dispõe de um nível de conhecimento parcial ou incompleto do comportamento da grandeza que está sendo estudada. Essa incerteza pode ser então introduzida sob a forma de um modelo probabilístico.

Por exemplo, um engenheiro encarregado de realizar estudos em uma empresa mede o tempo que os operários gastam em executar certas tarefas. Naturalmente, para cada tarefa o tempo gasto depende da experiência e da destreza do operário. Suponha que, para uma particular tarefa, o tempo médio gasto é de 285 segundos. Aqui a variável aleatória é  $X$  = “tempo em segundos gasto na execução da tarefa”, e tudo indica que, para um operário novato, pouco treinado, é alta a probabilidade  $P(X > 285)$ .

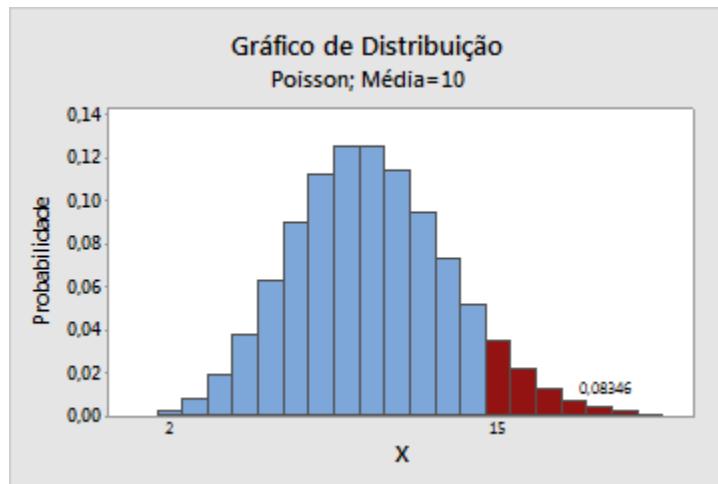
## FUNÇÃO DE PROBABILIDADE - VARIÁVEIS DISCRETAS

Dizemos que  $X$  é uma **v.a. discreta** se o número de valores que ela pode admitir é finito. Por exemplo, em uma linha de produção as peças produzidas são examinadas até que sejam encontradas 10 peças defeituosas, e então o número total de peças examinadas é anotado. Nesse caso, a v.a.  $X$  é o número total de peças examinadas. Nesse caso,  $X$  pode admitir os valores 10, 11, 12, 13, 14, ... Assim, claramente  $X$  é uma v.a. aleatória discreta.

Uma distribuição discreta descreve a probabilidade de ocorrência de cada valor de uma variável aleatória discreta. Com uma distribuição de probabilidade discreta, cada valor possível da variável aleatória discreta pode ser associado a uma probabilidade diferente de zero. Deste modo, uma distribuição de probabilidade discreta é, por vezes, apresentada em forma de tabela. Com uma distribuição discreta é possível calcular a probabilidade de que  $X$  é exatamente igual a algum valor.

Vamos a um exemplo:

O gráfico abaixo mostra a frequência de quantidade de reclamações por dia em um call center.



A probabilidade do call center receber 15 reclamações por dia é exatamente a frequência (ou seja, em torno de  $0,04 = 4\%$  de acordo com o gráfico).

As barras sombreadas neste exemplo representam o número de ocorrências quando as reclamações de clientes diárias são 15 ou mais. A altura das barras somam 0,08346; por conseguinte, a probabilidade de que o número de chamadas por dia seja de 15 ou mais é 8,35%.

021.399.242-66

Apresentamos a seguir alguns dos modelos probabilísticos discretos que costumam ser mais utilizados nas aplicações práticas da Estatística. Existem diversas modelagens para prever a probabilidade de variáveis discretas, como os modelos que envolvem ensaios de Bernoulli (Bernoulli, Binomial, Geométrico) e o modelo de Poisson. Aqui, vamos abordar rapidamente o modelo de Bernoulli e o modelo Binomial apenas, mas deixaremos dicas de onde vocês podem encontrar outras informações de outros tipos de modelos.

## MODELO DE BERNOUILLI

Num experimento aleatório é comum que estejamos interessados apenas na ocorrência de um resultado particular. Por exemplo:

1. Na seleção de um chip extraído de um lote, podemos querer saber somente se ele é perfeito ou não;
2. Na seleção de uma peça fabricada, queremos saber somente se ela satisfaz ou não às especificações exigidas pelo consumidor;



Em todos esses casos, o experimento realizado admite somente dois resultados possíveis.

Um experimento dessa natureza é chamado de “experimento de Bernoulli” ou, mais popularmente, “ensaio de Bernoulli”. Os dois resultados de um ensaio de Bernoulli são comumente chamados de “sucesso” e “fracasso”, sendo  $P$  a probabilidade de sucesso e  $(1-P)$  a probabilidade de fracasso.

Por exemplo, numa turma com 50 alunos, dos quais 30 são homens e 20 mulheres, escolhe-se um aluno ao acaso. Se levarmos em consideração apenas o sexo do aluno selecionado, isso trata-se de um ensaio de Bernoulli. Se considerarmos como sucesso a escolha de uma mulher, teremos  $p = 20/50 = 0,4$  e  $1 - p = 0,6$ .

## MODELO BINOMIAL

No modelo Binomial um mesmo experimento de Bernoulli é repetido  $n$  vezes, independentemente, e a v.a. de interesse representa o número de sucessos a serem obtidos nos  $n$  ensaios

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Sejam  $p$  e  $(1-p)$ , respectivamente, as probabilidades de sucesso e de fracasso em cada ensaio de Bernoulli. Se os resultados de cada ensaio são denotados por  $S$  (sucesso) e  $F$  (fracasso) teremos, para cada ensaio,  $P(S) = p$  e  $P(F) = 1 - p$ .

O espaço amostral do experimento resultante dos  $n$  ensaios de Bernoulli será composto por resultados que podem ser escritos como uma sequência de letras  $S$  e  $F$ . Em particular, um resultado com  $k$  sucessos e  $(n - k)$  fracassos pode ser descrito, sem perda de generalidade, como uma sequência de  $k$   $S$ 's, seguida de  $(n-k)$   $F$ 's, como a seguinte:

SSSSSS...SFFF...FF.

Como os  $n$  ensaios são independentes, a probabilidade de ocorrência desse resultado particular é:

$$p^k * (1 - p)^{n-k}$$

O evento “k sucessos e (n-k) fracassos” pode ocorrer de diversas outras maneiras. O cálculo do número de maneiras de se obter “k sucessos e (n – k) fracassos” é o número de combinações de n objetos tomados de k em k

$$P_x = \binom{n}{x} p^x q^{n-x}$$

O símbolo  $\binom{n}{x}$  indica combinação de n objetos tomados x a x. Por exemplo, vamos supor que temos um total de 50 elementos e o particionaremos em grupos de 5. Temos inúmeras possibilidades de formar de forma diferente esse grupo de 5. O cálculo de combinatória nos mostra quantas exatamente existem, de forma que:

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Oseias Dias de Farias  
 osseias.dafaria@gmail.com  
 021.399.242-66

Na fórmula acima representamos x como sendo r, que também é uma denotação possível que vocês devem encontrar em vários livros por aí. O símbolo “!” indica “ factorial. A fórmula de um factorial r!, por exemplo, é dada por:

$$r! = 1*2*3*4...*r$$

Essa fórmula acima pode ser simplificada por:

$$\binom{n}{r} = \frac{n(n-1)(n-2)....(n-r+1)}{1 \cdot 2 \cdot 3 ..... r}$$

Logo, voltando ao exemplo de termos 50 elementos e queremos formar 5 grupos, temos que podemos formá-lo:

$$\binom{50}{5} = \frac{50*(50-1)*(50-2)*...*(50-5+1)}{5!} = \frac{50*49*48*47*46}{1*2*3*4*5}$$



= 2.118.760 maneiras distintas

**NOTA IMPORTANTE:** É CONVENCIONADO QUE  $0! = 1$ . OU SEJA, NUNCA TEREMOS UMA DIVISÃO POR ZERO

Quando temos as combinações n tomado a n ou n tomado a 0, ficamos com:

$$\binom{n}{n} = \binom{n}{0} = \frac{n!}{0!n!} = 1$$

Vamos ver em um exemplo como o modelo binomial funciona:

Oseias Dias de Farias  
Geralmente, em cerca de 80% dos chamados que um certo técnico em computação recebe para resolver panes nos computadores de clientes ele constata que o problema decorreu da presença de algum vírus. Suponha que, em um determinado dia, esse técnico vai visitar seis desses clientes cujos computadores necessitam de conserto, e admita também que os seis clientes não se comunicam por meio de computador (o que garante a independência da existência de vírus em cada computador). Calcule a probabilidade de que:

- a) Pelo menos quatro entre os seis computadores estejam com vírus.
- b) No máximo dois dentre eles estejam com vírus.
- c) Todos os seis estejam com vírus.

Resposta:

Considere:

Sucesso = “o defeito no computador é devido a presença de vírus” ( $p = P(\text{sucesso}) = 0,80$ )

$X$  = número de computadores com vírus entre os 6 a serem consertados.

**a)**  $P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6) =$



$$= \binom{6}{4} 0,8^4 \times 0,2^2 + \binom{6}{5} 0,8^5 \times 0,2 + \binom{6}{6} 0,8^6 = 0,90112.$$

Isso significa que é bem alta a probabilidade de pelo menos quatro entre os seis computadores estarem com vírus (90,112%)

**b)**  $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$

$$= \binom{6}{0} 0,2^6 + \binom{6}{1} 0,8 \times 0,2^5 + \binom{6}{2} 0,8^2 \times 0,2^4 = 0,01696.$$

Este valor indica que é baixíssima a probabilidade de que no máximo dois deles estejam com vírus.

$$* 0,1^0 * 0,9^{(50-0)}$$

Oseias Dias de Farias

**NOTA IMPORTANTE:** NESSE CASO, A COMBINATÓRIA DE 6 ELEMENTOS TOMADOS DA O É IGUAL A 1  
[oseias.farias@gmail.com](mailto:oseias.farias@gmail.com)  
 021.399.242-66

**c)**  $P(X = 6) = 0,8^6 = 0,26214.$

Como dito anteriormente, existem diversos modelos probabilísticos para variáveis discretas. O modelo de **Poisson** prevê a distribuição do número de indivíduos ou outro dados de contagem por unidade de tempo ou espaço. Por exemplo

- Se os organismos distribuem-se independentemente no espaço, espera-se que a contagem de indivíduos por  $m^2$  se ajuste a uma distribuição de Poisson.
- Número de chamadas telefônicas que chegam a uma Central em um dado intervalo de tempo
- Número de navios que chegam ao cais de um porto em um dia
- Número de defeitos encontrados em uma geladeira recém-fabricada



O modelo **Geométrico** resulta de experimentos similares aos experimentos binomiais. Suponha novamente uma variável resposta pode assumir somente dois valores. Enquanto a distribuição binomial se preocupa com a probabilidade do número de sucessos em  $n$  tentativas, o modelo Geométrico conta o número de fracassos até a observação do primeiro sucesso. Alguns exemplos de quando podemos usá-lo:

- Suponha que um ambiente tenha  $N$  manchas de habitats que pode ser ou não ocupado por uma espécie. Queremos saber: quantas manchas de habitat devem ser avaliadas até que a espécie seja detectada?
- O engenheiro responsável pelo Controle da Qualidade de uma linha de produção examina, uma após a outra, as peças fabricadas. Se achar uma defeituosa, ele para a produção para detectar e corrigir as causas do defeito. Se após examinar 10 peças verificar que nenhuma é defeituosa, ele mantém a linha funcionando. Se a probabilidade de se achar uma peça defeituosa em cada exame é 0,05, qual é a probabilidade de: a) a produção ser parada antes que a quinta peça seja examinada? b) a produção não precisar ser parada?

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

Você pode encontrar mais detalhes sobre esses modelos na referência Pinheiro, J., Cunha, S., S. Santiago, Gomes, G. - Probabilidade e Estatística: Quantificando a incerteza.

## FUNÇÃO DE PROBABILIDADE - VARIÁVEIS CONTÍNUAS

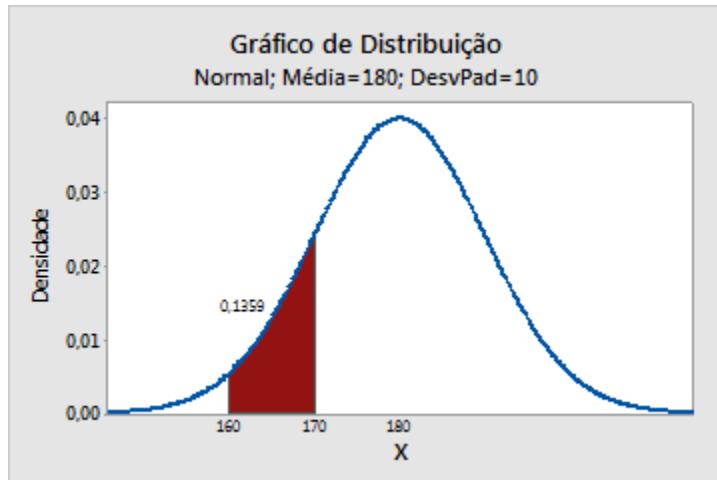
A distribuição contínua descreve as probabilidades dos possíveis valores de uma variável aleatória contínua. Uma variável aleatória contínua é uma variável aleatória com um conjunto de valores possíveis (conhecidos como intervalos) que é infinito e incontável.

As probabilidades de variáveis aleatórias contínuas ( $X$ ) são definidas como a **área sob a curva da sua distribuição**. Assim, apenas as faixas de valores podem ter uma probabilidade diferente de zero. A probabilidade de que uma variável aleatória contínua seja igual a algum valor é sempre zero.

Vamos a um exemplo:



A distribuição normal contínua pode descrever a distribuição de peso de indivíduos do sexo masculino adultos. Por exemplo, você pode calcular a probabilidade de que um homem pesa entre 160 e 170 libras



#### Oseias Dias de Farias

A região sombreada sob a curva, neste exemplo, representa o intervalo entre 160 e 170 libras (72 a 77 kg, aproximadamente). A área deste intervalo é 0,136; por conseguinte, a probabilidade de um homem ser selecionado aleatoriamente pesar entre 160 e 170 libras é de 13,6%. Toda a área sob a curva equivale a 1,0.

No entanto, a probabilidade de que  $X$  seja exatamente igual a algum valor é sempre zero porque a área sob a curva em um único ponto, que não tem nenhuma largura, é zero. Por exemplo, a probabilidade de um homem pesar exatamente 190 libras para a precisão infinita é zero. É possível calcular uma probabilidade não nula de que um homem pese mais do que 190 libras, ou menos do que 190 libras, ou entre 189,9 e 190,1 libras, mas a probabilidade de que ele pesa exatamente 190 libras é zero.

Como dito anteriormente, agora vamos definir uma probabilidade para um dado intervalo, ou seja,  $P(a \leq X \leq b)$  para dois números reais  $a$  e  $b$ . Isso é obtido ao substituir-se a função de probabilidade  $p$  por uma função  $f$ , chamada **função de densidade de  $X$** , ou simplesmente, **densidade de  $X$** .

Definindo formalmente esse conceito:

Dizemos que  $X$  é uma variável aleatória contínua se existe uma função  $f$ , chamada função de densidade de  $X$ , satisfazendo as seguintes condições:

1.  $f(x) \geq 0$  para todo  $x$  real
2.  $\int_{-\infty}^{\infty} f(x)dx = 1$
3. Para quaisquer  $a, b$  reais ( $a < b$ ),  $P(a \leq X \leq b) = \int_a^b (f(x)dx)$

Vamos traduzir esse trem! O primeiro ponto diz que essa função de densidade assume valor maior ou igual a zero para toda variável  $X$  existente. Ou seja, não há valores negativos para uma função de densidade.

O ponto dois mostra que a integral (símbolo  $\int$ ) dessa função deve ser igual a 1. Quando falamos de integral estamos automaticamente falando de "área embaixo da curva". Ou seja, se pegarmos a área inteirinha embaixo da curva, ela vai ser igual a 1. Não entraremos aqui no detalhe de cálculo de integrais, até porque muitas dessas fórmulas já são pré-calculadas na forma de tabela para nós (ufa!). Esse ponto 2 também é chamado de **Função de Distribuição Acumulada (FDA)**.

Sociedades de Fárias  
oseias.dfarias@gmail.com  
021.399.242-66

E os pontos 1 e 2 você já sabia! Afinal, uma probabilidade sempre está entre 0 e 1.

O ponto 3 diz que a probabilidade de uma faixa será exatamente a área embaixo da curva dentro dessa faixa. É exatamente o que falamos no exemplo acima sobre os pesos de homens.

Apresentaremos a seguir alguns dos modelos probabilísticos contínuos que costumam ser mais utilizados nas aplicações práticas da Estatística. Existem várias, como o modelo Uniforme, Exponencial, Normal, t-student, chi-quadrado, F, etc. Nesse capítulo vamos abordar apenas a Normal, mas nos capítulos de teste de hipótese ainda abordaremos a t-student, chi-quadrado e F.

## MODELO NORMAL

É, de longe, a distribuição mais importante da estatística. A curva Normal (também chamada de Gaussiana ou de paramétrica) descreve de forma

muito adequada o comportamento de uma variável que se distribui de forma simétrica em relação a um valor central. Os dois parâmetros que a caracterizam são  $\mu$ , que especifica a média, e  $\sigma$ , que define seu desvio-padrão. Tendo esses dois parâmetros definidos, a função é escrita como:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right\}, \text{ para todo } x \text{ real.}$$

Dessa forma, se tivermos a média e o desvio-padrão, conseguiremos traçar uma curva com o formato:



Lá no início dos anos 1800, supunha-se que todos os fenômenos da vida real devessem ajustar-se a uma curva em forma de sino; caso contrário, suspeitava-se de alguma anormalidade no processo de coleta de dados. Daí a designação de curva normal.

Ainda que hoje saibamos que isso não é verdade, a distribuição de probabilidade normal é importante na inferência estatística por três razões distintas:

- a) as medidas produzidas em diversos processos aleatórios seguem essa distribuição;
- b) as probabilidades normais podem ser usadas frequentemente como aproximações de outras distribuições de probabilidade, tais como a binomial e a de Poisson;

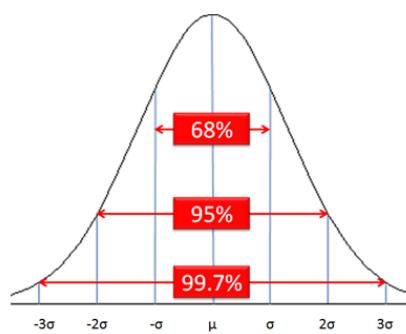
c) as distribuições de estatísticas da amostra, tais como a média e a proporção, frequentemente seguem a distribuição normal independentemente da distribuição da população. Veremos isso mais pra frente, quando falarmos sobre o Teorema do Limite Central.

As características da distribuição normal são:

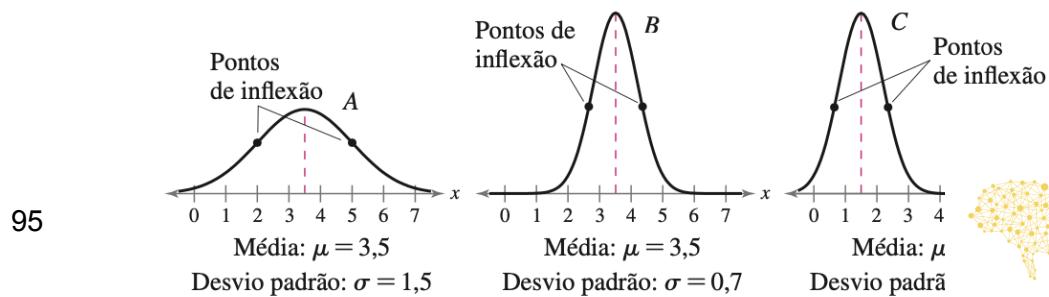
1. A média, a mediana e a moda são iguais - ou seja, não tem outliers!
2. Tem forma de sino e é simétrica em torno da média.
3. À medida que a curva normal se distancia da média, ela se aproxima do eixo  $x$ , mas sem tocá-lo.

Outra grande característica dela é poder afirmar que cerca de 68% dos dados encontram-se dentro do intervalo de  $\pm 1$  desvios padrão da média, 95% dos dados encontram-se dentro do intervalo de  $\pm 2$  desvios padrão da média e assim por diante, conforme mostra a figura abaixo.

Oseias Dias de Farias  
oseias.diasdefarias@gmail.com  
021.399.242-66



Uma **distribuição normal** pode ter qualquer média e qualquer desvio padrão positivo. Esses dois parâmetros, determinam o formato da curva normal. A média dá a localização da linha de simetria e o desvio padrão descreve o quanto os dados estão dispersos. Logo, observando a curva abaixo, sabendo que ela respeita todas as características de distribuição normal vistas acima, conseguimos ver que todas elas são distribuições normais!



Renata Biagi

Como para qualquer distribuição contínua de probabilidade, o valor da probabilidade pode somente ser determinado para um intervalo de valores da variável.

Dentro dessa classe de funções, temos também um tipo específico de curva normal. Se uma v.a. tem distribuição Normal com média igual a 0 (zero) e variância igual a 1 (um), diremos que ela tem distribuição **Normal Padrão** ou distribuição **Normal Reduzida**.

Qualquer curva normalmente distribuída pode se tornar uma Normal Padrão. Para isso, pegamos cada valor de X e o padronizamos pela seguinte fórmula:

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

Em que  $\mu$  é a média da sua distribuição atual  $\sigma$  é o desvio-padrão da distribuição atual. Nesse caso, Z será seu valor de X padronizado. E por que isso é legal? Pois temos uma tabela que indica a probabilidade para cada valor de Z, ou seja, não precisaremos fazer nenhum cálculo de probabilidade!

Um pedaço da tabela pode ser visto abaixo.

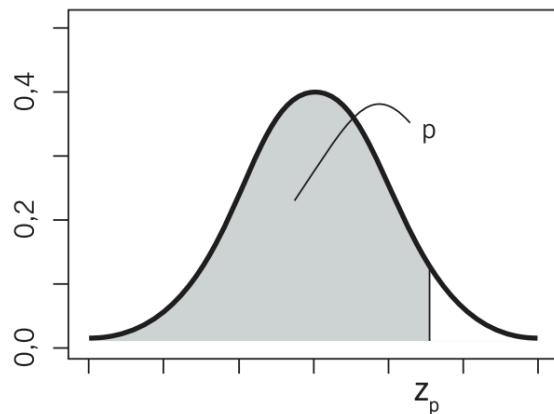
<b>z</b>	<b>0</b>	<b>0.01</b>	<b>0.02</b>	<b>0.03</b>	<b>0.04</b>	<b>0.05</b>	<b>0.06</b>	<b>0.07</b>	<b>0.08</b>	<b>0.09</b>
<b>+0</b>	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
<b>+0.1</b>	.53983	.54380	.54776	.55172	.55567	.55966	.56360	.56749	.57142	.57535
<b>+0.2</b>	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
<b>+0.3</b>	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
<b>+0.4</b>	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
<b>+0.5</b>	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
<b>+0.6</b>	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
<b>+0.7</b>	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
<b>+0.8</b>	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
<b>+0.9</b>	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
<b>+1</b>	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
<b>+1.1</b>	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298

Nessa tabela, caso obtenhamos um  $z = 0.56$ , podemos dizer que a probabilidade de termos um **z menor ou igual a** 0.56 é de 0.71226 - ou seja, a área embaixo da curva até o  $z = 0.56$  é de 0.71226.



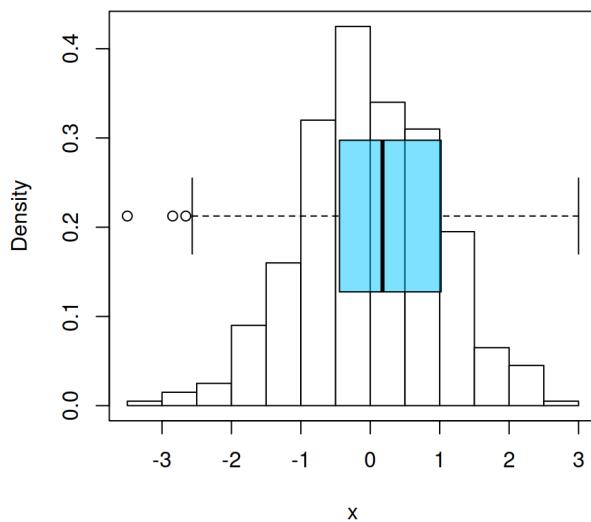
A notação que usamos para ler isso é:  $P(z < 0.56) = 0.71226$

Antes de partirmos para um exemplo de uso, note que acima falei "z menor ou igual a". É importante ressaltar aqui que essa tabela nos mostra somente a probabilidade **acumulada** até o valor de Z, como mostra a figura abaixo.



Isso é importante pois se quisermos calcular a probabilidade de uma **faixa** e não a acumulada, precisamos fazer algumas transformações. Para fazermos essas transformações, lembrem-se que a curva de probabilidade varia de 0 a 1, sendo que 0 é uma probabilidade nula de um valor acontecer e 1 uma probabilidade de 100% de que todos os valores estejam naquela faixa. Vocês entenderão mais sobre isso no exemplo 2.

Outra propriedade importante de lembrar é que em uma distribuição de probabilidade, a probabilidade acumulada  $P(z < z_c)$  para um determinado valor corresponde ao **percentil** em que aquele valor se encontra. Existe um paralelo muito grande entre o histograma e o boxplot que ajuda lembrar desse conceito.



Notem que se tivéssemos um boxplot traçado em cima do histograma, saberíamos exatamente qual é o percentil 25, 50, etc.

Mesmo sem ter o boxplot em cima da nossa distribuição, usando a tabela de Z conseguimos saber qual é o percentual correspondente a um Z calculado e, consequentemente, a um X que corresponde a esse Z. Por exemplo, considerando que temos uma distribuição normal e temos um  $X = 20$  que corresponde a um  $Z = 0.56$ . Sabendo que a probabilidade  $P(z < z_c)$  nesse  $z_c$  é de 0.71226 (de acordo com a tabela Z), podemos dizer que o  $x = 20$  está no percentil 71,226.

Por exemplo, considerando que temos uma distribuição normal e que nosso  $x = 20$  e, o z correspondente a esse x nessa distribuição é 0.56 (ou seja,  $z_c = 0.56$ ). Sabendo que a probabilidade  $P(z < z_c)$  nesse  $z_c$  é de 0.71226 (de acordo com a tabela z), podemos dizer que o  $x = 20$  está no percentil 71,226.

Agora vamos a outros exemplos de uso

### Exemplo 1.

Em uma população de homens com IMC médio = 29 e desvio-padrão = 6, qual é o IMC que representa o percentil 90?

*Resposta:*

Lembrando da fórmula:

$$z = \frac{x - \mu}{\sigma}$$

Queremos o X que representa o percentil 90. Sabemos que a média é 29 e desvio-padrão é 6. Logo, reescrevendo a fórmula acima e isolando o valor que queremos (x), temos:

$$X = Z^*6 + 29$$

Para encontrarmos o valor de Z, precisamos lembrar que queremos o Z que representaria o percentil 90, ou seja, a área embaixo da curva deverá ser 0,900. Como os valores de z são tabelados, voltamos a tabela e procuramos o valor mais próximo a 0,900.

$Z_a$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

Esse valor está entre 0.8997 e 0.9015. O mais próximo seria 0.8997, que corresponde a um z de 1.28. Logo,

$$x = 1,28 * 6 + 29 = 36,7$$

Portanto, o IMC que corresponde ao pctl 90 é de 36,7.

### Exemplo 2.

Suponha que o tempo X, em minutos, corresponde ao tempo que uma pessoa leva para executar determinada tarefa e varia conforme uma distribuição Normal com parâmetros  $\mu$  (média) e  $\sigma$  (desvio padrão). Suponha também que a probabilidade de que a tarefa seja executada em 70 minutos



no máximo é 0,75, e a probabilidade de que a tarefa seja executada em no máximo 50 minutos é 0,25.

- Determine os valores da média e desvio-padrão .
- Qual a porcentagem das pessoas que precisarão de mais de 85 minutos?

Resposta:

- Sabemos que  $X$  tem distribuição normal, então podemos padronizar sua distribuição para o  $Z$  normal padrão. A probabilidade acumulada quando  $X = 70$  minutos é 0,75. Então:

$P(Z \leq 70) = 0,75$ . Consultando a tabela, para essa probabilidade acumulada,  $Z = 0,67$ .

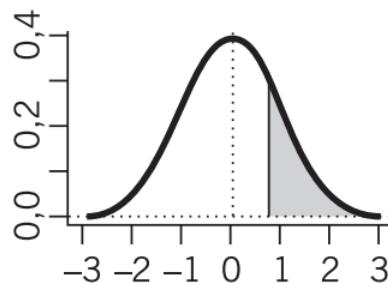
$$\text{Logo, } Z = \frac{70 - \mu}{\sigma} = 0,67$$

Da mesma forma, sabemos que a probabilidade acumulada quando  $X = 50$  minutos é 0,25. Da tabela, a probabilidade 0,25 corresponde a um  $Z = -0,67$ . Logo,

$$Z = \frac{50 - \mu}{\sigma} = -0,67$$

Temos agora 2 equações e 2 incógnitas. Resolvendo essas equações, temos  $\mu = 60$  e  $\sigma = 14,9$ , ambos em minutos.

- Aqui temos um problema: Não estamos em busca de uma probabilidade acumulada até certo valor - queremos saber a probabilidade a partir de um certo valor, ou seja, teríamos algo como:



Esse problema é fácil de resolver quando lembramos que a área total embaixo da curva é 1. Ou seja, ainda que a tabela Z apenas seja capaz de nos dar a probabilidade acumulada (área branca desse gráfico), sabemos que a área cinza deverá ser  $1 -$  área branca. Ou seja, conseguimos fazer o cálculo mesmo sem o valor exatamente tabelado.

Dessa forma,  $Z = (85-60)/14,9 = 1,677$

$$P(z \geq 1,677) = 1 - P(z \leq 1,677) = 0,9535$$

Você pode encontrar mais detalhes sobre esses e outros modelos probabilísticos na referência Pinheiro, J., Cunha, S., S. Santiago, Gomes, G. - Probabilidade e Estatística: Quantificando a incerteza.

## TEOREMA DO LIMITE CENTRAL

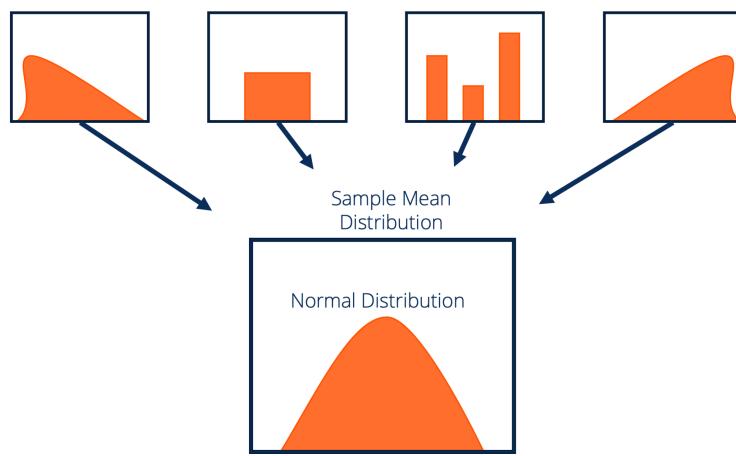
Vocês viram até agora que distribuições normais são ótimas para se trabalhar: temos suas propriedades, tabelas referências para encontrar probabilidades e um comportamento muito bem conhecido. Por conta disso, diversas estatísticas são descritas especialmente para distribuições normais.

Porém, em vários casos do dia a dia vamos nos deparar com **distribuições não normais**. Quando isso acontece, torna-se mais complexo e, em alguns casos, menos precisos, os métodos disponíveis para cálculo de probabilidade, teste de hipótese, entre outros.

Dessa forma, uma alternativa para garantir a normalidade dos dados é proveniente do **teorema do limite central**. O teorema nos diz que, para uma distribuição com **pelo menos 30 dados**, a **distribuição da média de subamostras** dessa distribuição seguirá uma normal.

Simplificando esse raciocínio: tendo os dados da distribuição, podemos selecionar várias subamostras dessa distribuição e calcular a média de cada uma dessas subamostras. Plotando em um gráfico a média de cada uma dessas subamostras, veremos uma distribuição normal, conforme exemplifica a imagem abaixo:





Para exemplificar o que isso quer dizer

Supondo que temos uma distribuição com esses dados:

**Distrib = [1, 1, 2, 3, 4, 5, 6, 7, 8, 8, 8, 8, 8, 8, 8, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9]**

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66

- 1) A amostra tem pelo menos 30 dados? Sim
- 2) Vamos selecionar várias subamostras aleatórias (podendo ter repetição) dessa distribuição:  
 $S_1 = [1, 1, 9]$   
 $S_2 = [3, 4, 9]$   
 $S_3 = [4, 6, 9]$   
.... até  $S_n$
- 3) Vamos calcular a média para cada uma dessas subamostras  
Média  $S_1 = 3,5$   
Média  $S_2 = 5,3$   
Média  $S_3 = 6,3$   
.... Calcular para todas as subamostras ( $S_n$ )
- 4) Plotando Média  $S_1$ , Média  $S_2$ , Média  $S_3$ ..., Média  $S_n$ , de acordo com o teorema do limite central, nós teremos uma normal

O aparecimento de uma distribuição normal proveniente da distribuição populacional que é distorcida tem algumas aplicações muito importantes na prática estatística. Muitas práticas em estatística, como aquelas que envolvem

testes de hipóteses ou intervalos de confiança, fazem algumas suposições sobre a população da qual os dados foram obtidos.

Outro fator importante do teorema do limite central é o cálculo da nova média e do novo desvio-padrão:

- 1) A média das médias amostrais será igual à média da distribuição original

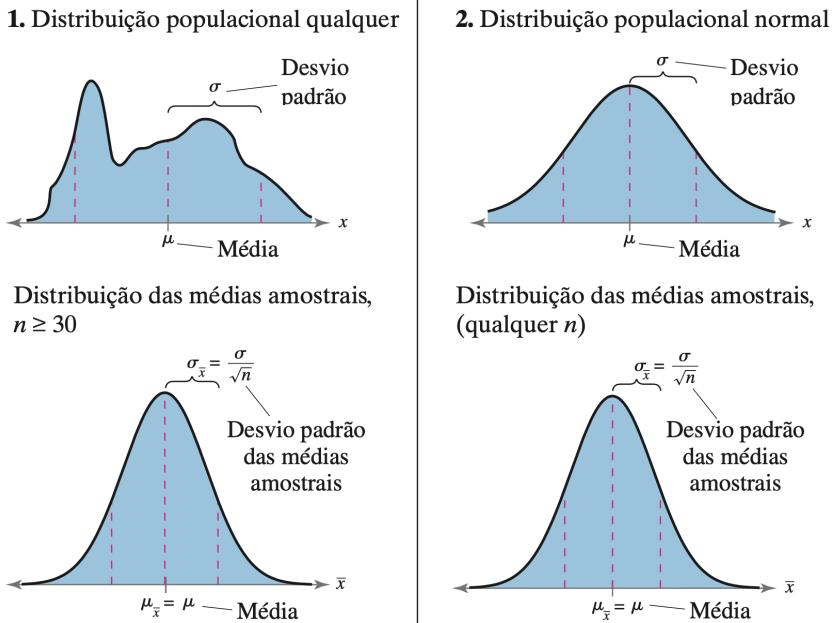
$$\mu_{\bar{x}} = \mu$$

- 2) A distribuição amostral das médias tem uma variância igual a  $1/n$  vezes a variância da população e um desvio padrão igual ao desvio padrão da população dividido pela raiz quadrada de  $n$

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

Oseias Dias Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Ou seja, a distribuição das médias amostrais tem a mesma média que a população, mas o seu desvio padrão é menor que o desvio padrão da população. Isso nos diz que a distribuição das médias amostrais tem o mesmo centro que a população, porém é mais concentrada. Além disso, a distribuição das médias amostrais torna-se cada vez menos dispersa (maior concentração em relação à média) conforme o tamanho  $n$  da amostra aumenta.



### Exemplo

O gasto médio com alojamento e refeição, por ano, em faculdades de quatro anos é de US\$ 9126 (supondo conjunto com mais de 1000 faculdades). Você seleciona aleatoriamente 40 dessas faculdades. Qual é a probabilidade de que a média de gastos com alojamento e refeição seja menor que US\$ 9400? Suponha que os gastos tem desvio padrão de US\$ 1500 na amostra original

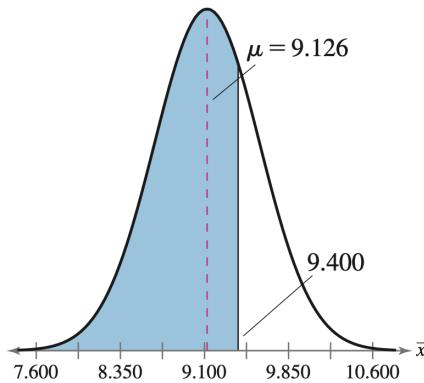
*Resposta:*

Selecionamos 40 amostras das 1000 originais. Como  $n > 30$  e como estamos querendo calcular a probabilidade da **média**, podemos usar o teorema do limite central.

De acordo com o teorema, a média das médias deve ser a mesma da população original. Logo, a média das médias amostrais deve ser US\$ 9126. O desvio-padrão das médias amostrais deve ser  $1500/\sqrt{40} \approx$  US\$ 237.17. Logo, agora temos uma distribuição normalizada com média US\$ 9126 e desvio-padrão US\$ 237.17. A partir dessa distribuição, podemos aplicar tudo que vimos sobre escore-z, pois essa é uma distribuição normalizada.

Nossa nova distribuição:





Queremos encontrar o valor de  $z$  que corresponde a 9400 e, posteriormente, na tabela de  $z$ , queremos encontrar a probabilidade acumulada até esse valor.

Logo:

$$Z = (9400 - 9126)/237.17 \approx 1.155.$$

De acordo com a tabela de  $z$ ,  $P(z < 1.155) \approx 0.876$ .

## TRANSFORMAÇÕES

Oseias Dias de Farias

Como já dissemos, vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma distribuição normal (em forma de sino) ou então mais ou menos simétrica. Mas, em muitas situações de interesse prático, a distribuição dos dados da amostra é assimétrica e pode conter valores atípicos, como vimos em exemplos anteriores.

Se estamos interessados em entender probabilidades de médias amostrais, conseguimos superar esse problema com o teorema do limite central. Entretanto, se estivermos interessados em qualquer outra medida, precisamos pensar em outras alternativas. O que se propõe é efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal. Aqui vamos ver algumas transformações mais usuais.

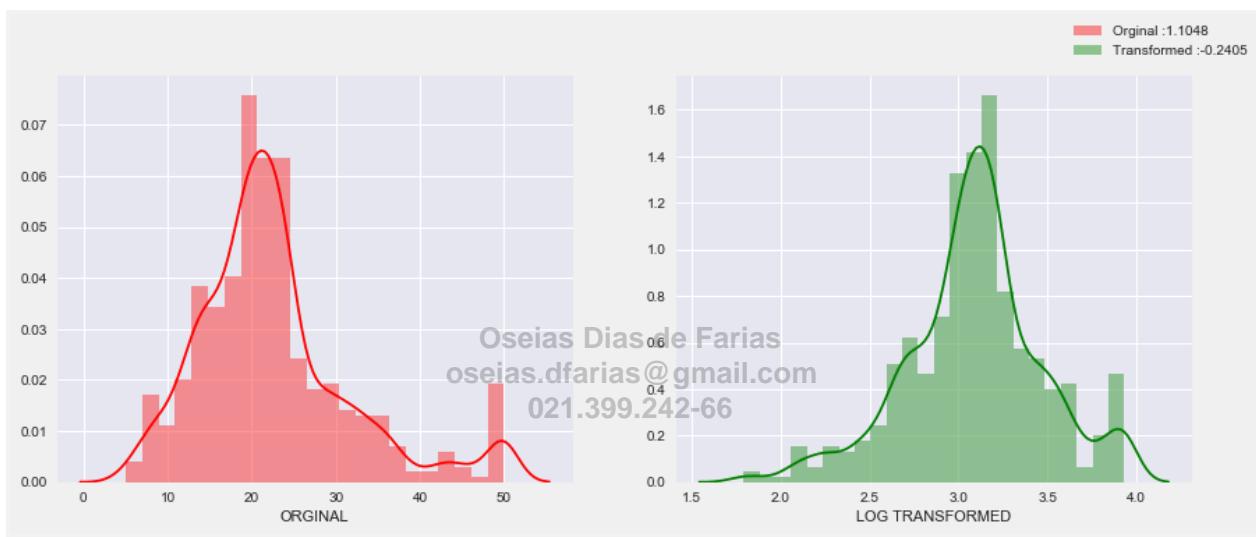
### 1. Transformação de Log

Variáveis numéricas podem ter distribuição altamente assimétrica e não normal (Distribuição Gaussiana) causada por outliers, distribuições altamente exponenciais, etc. Portanto, optamos pela transformação de dados.

Na transformação Log cada variável de x será substituída por  $\log(x)$  com base 10, base 2 ou log natural.

Em Python teríamos, caso quisessemos aplicar log em uma variável chamada "Target", teríamos

```
import numpy as np
log_target = np.log1p(df["Target"])
```



## 2. Box-cox

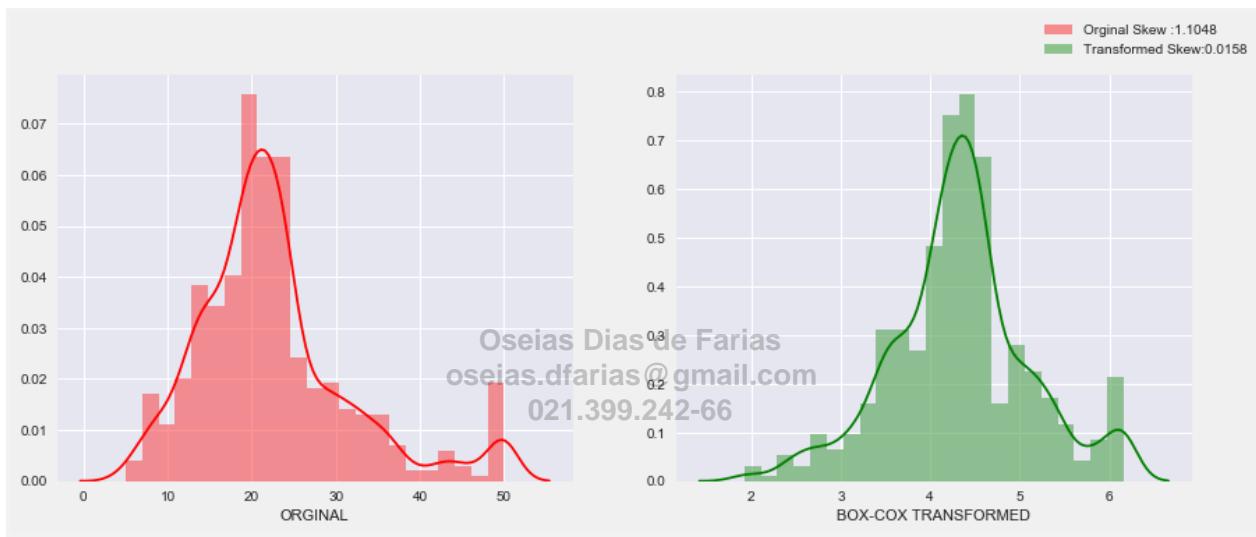
É uma das técnicas de transformação mais usadas. A transformação Box-cox funciona muito bem para muitas naturezas de dados. A imagem abaixo é a fórmula matemática para a transformação Box-cox.

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

Valores de lambda de -5 a 5 vão ser testados e o melhor valor para os dados é selecionado. O valor “Melhor” é aquele que resulta na menor assimetria da distribuição. A transformação de log ocorrerá quando tivermos lambda zero.

Em Python, faríamos:

```
from scipy.stats import boxcox  
bcx_target, lam = boxcox(df["Target"])  
#lam vai ser o lambda que der a menor assimetria
```



Aqui, notamos que a função Box-cox reduziu a assimetria e é quase igual a zero. Funcionou bem!

Vocês verão nos capítulos de teste de hipótese que existem diversos tipos de testes que podem ser aplicados **apenas** se a distribuição dos dados forem normais. Entretanto, se nossos dados não forem normalmente distribuídos, podemos usar essas transformações para que, após a transformação, o dado se torne normalmente distribuído.

# 9. Teste de hipótese - Conceitos fundamentais



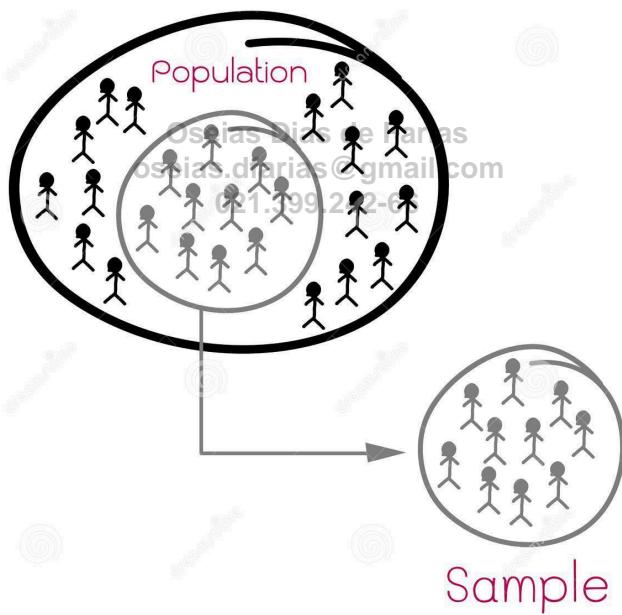
Notas importantes: A maioria dos exemplos dos capítulos a seguir foram retirados das referências Frost, J e Larson, R., Farber B.

Como falamos na seção de estatística descritiva, quando pensamos apenas em descrever um conjunto de dados, podemos usar quaisquer métricas da estatística (média, mediana, etc) para comparar dois conjuntos. Com estatísticas descritivas, **não há incerteza** porque você está **descrevendo** apenas os dados que você realmente mede.

No entanto, se você quiser fazer **inferências** sobre uma população, existem uma série de passos extras. Essa inferência é chamada de estatísticas inferenciais e é usada especialmente quando usamos amostras para descrever uma população.

Uma **população** é todo o grupo sobre o qual você deseja tirar conclusões. Uma **amostra** é o grupo específico do qual você coletará dados. O tamanho da amostra é sempre menor que o tamanho total da população. E por que coletamos amostras? Por que é, em muitos casos, impossível coletar dados da população inteira. Por exemplo, imaginem que queremos ter uma ideia da altura média de mulheres brasileiras de 18 a 60 anos. Concordam que é praticamente inviável coletarmos os dados de toda essa população? Ainda que pedíssemos para que cada mulher inputasse os dados em uma plataforma online, muitas teriam limitações com internet, de conhecimento, e até de vontade. Dessa forma, podemos coletar uma pequena amostra da população e garantir que essa amostra represente o todo.

### População (population) vs Amostra (sample)



Podemos medir médias, desvios e quaisquer outras métricas tanto na população (caso tenhamos) quanto na amostra. Para não nos confundirmos, vamos definir símbolos para quando estivermos falando sobre a amostra e para quando estivermos nos tratando sobre a população. Aqui está a simbologia mais adotada por autores da área:

Parâmetro	Símbolo - População	Símbolo - Amostra
Média	$\mu$	$\bar{x}$
Desvio-padrão	$\sigma$	$s$
Variância	$\sigma^2$	$s^2$
Proporção	$p$	$\hat{p}$
Quantidade de dados	$N$	$n$

Fazer inferências sobre uma população é particularmente importante em *business*, onde queremos aplicar os resultados a uma população maior, não apenas à amostra específica do estudo. Por exemplo, se estamos testando uma nova campanha de marketing, queremos saber se ele funciona apenas para um pequeno e seletivo grupo experimental. Queremos inferir que será eficaz para uma população maior. Queremos generalizar os resultados da amostra para pessoas fora da amostra. E é aí que entra o **teste de hipótese**.

O teste de hipóteses é um processo estatístico que permite usar uma amostra para tirar conclusões sobre uma população inteira. Mais especificamente, criamos **duas hipóteses** (hipótese nula e alternativa, que vamos falar mais sobre já) sobre a população e determinamos qual afirmação **os dados coletados das amostras suportam**. Esses procedimentos usam evidências (ou seja, os próprios dados) das amostras para fazer inferências sobre as características das populações.

## AMOSTRAGEM

Para escolher nossa amostra ideal, nós não podemos escolher um grupo conveniente. Em vez disso, existem diversas técnicas de **amostragem** que nos permitem ter confiança de que a amostra representa bem a população.

A amostragem correta é parte fundamental de quem trabalha com dados e, infelizmente, muitas vezes é negligenciada em cursos típicos. Compreender os diferentes métodos de amostragem e como eles estão sendo usados em nosso fluxo de trabalho pode, primeiro, nos ajudar a evitar possíveis vieses e, segundo, nos ajudar a escolher os métodos que melhoram a eficiência dos dados que amostramos.

Existem duas famílias de amostragem: **amostragem probabilística** e **amostragem as não probabilísticas**. As amostras selecionadas por critérios **não probabilísticos** por vezes não são representativas dos dados do mundo real e, portanto, se não forem bem justificadas, podem estar repletas de **vieses** de seleção (leia mais sobre vieses no artigo "Tipos de Vieses na Ciência de Dados" do material complementar). Um exemplo de amostra não probabilística é a **amostragem por conveniência**, em que você escolhe a amostra apenas por estarem disponíveis (por exemplo, você seleciona somente seus amigos). O problema disso é a falta de representatividade de uma população. Outro tipo de amostragem é a **amostragem por julgamento**, em que os especialistas decidem quais amostras incluir. A menos que o especialista seja praticamente uma máquina, capaz de entender exatamente qual é o perfil de sua população e selecionar uma amostra que a represente muito bem, esse tipo de amostragem pode conter muitos vieses.

Agora vamos abordar agora um pouco mais sobre alguns tipos mais comuns de **amostragem probabilísticas**.

## **AMOSTRAGEM ALEATÓRIA SIMPLES**

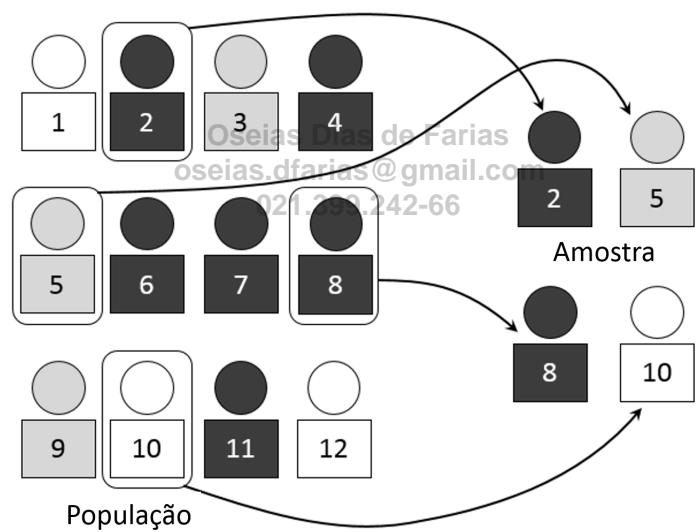
Na forma mais simples de amostragem aleatória, você dá a todas as amostras na população probabilidades iguais de serem selecionadas. Por exemplo, você seleciona aleatoriamente 10% de todas as amostras, dando a todas as amostras uma chance igual de 10% de serem selecionadas.

A vantagem deste método é que é fácil de implementar. A desvantagem é que categorias raras de dados podem não aparecer em sua seleção.

Considere o caso em que uma classe aparece apenas em 0,01% de sua população de dados. Se você selecionar aleatoriamente 1% de seus dados, é improvável que amostras dessa classe rara sejam selecionadas. Se seu intuito

é calcular médias ou medianas, por esses grupos serem muito pequenos pode ser que esse problema não seja um empecilho, uma vez que essas medidas de tendência central devem ser pouco impactadas. Contudo, se você quer representar desvios-padrão ou até fazer um modelo preditivo para identificar esses casos raros, essa pode não ser a melhor amostragem (a não ser que você garanta que o grupo sub representado continua na sua amostra!).

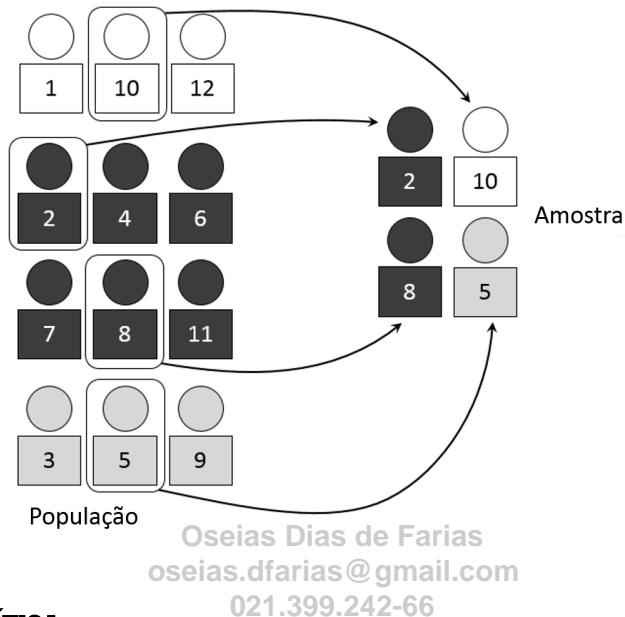
Ainda linkando a esse problema de "raridade", outro grande problema é que é difícil fazer esse tipo de amostragem em pesquisas reais. Imagine que você quer representar o Brasil todo, selecionou algumas pessoas, mas essa seleção acabou não pegando nenhuma pessoa, por exemplo, do Rio Grande do Sul. Nesse caso, sua amostra pode acabar sendo menos representativa do país pois a aleatoriedade acabou não selecionando um dos nossos estados.



## AMOSTRAGEM ESTRATIFICADA

Para evitar a desvantagem da amostragem aleatória simples, você pode primeiro dividir sua população nos grupos que lhe interessam e amostrar de cada grupo separadamente. Por exemplo, para amostrar 1% dos dados que têm duas classes, A e B, você pode amostrar 1% da classe A e 1% da classe B. Dessa forma, não importa quão rara seja a classe A ou B, você garantirá que as amostras dele serão incluídas na seleção. Cada grupo é chamado de estrato, e esse método é chamado de amostragem estratificada.

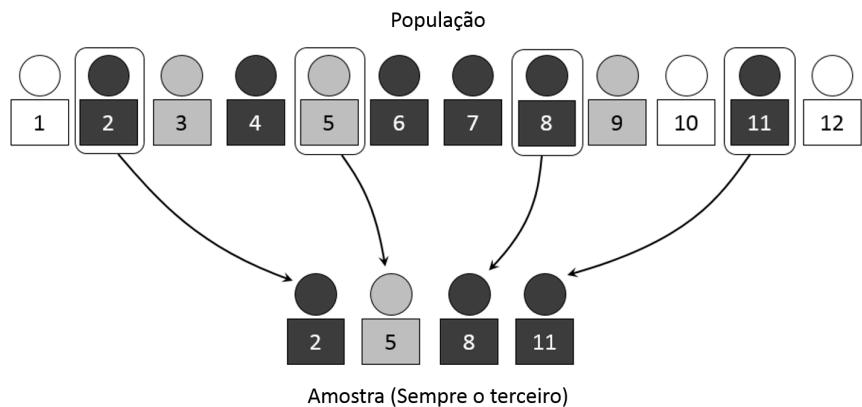
Uma desvantagem desse método de amostragem é que nem sempre é possível, como quando é impossível dividir todas as amostras em grupos. Isso é especialmente desafiador quando uma amostra pode pertencer a vários grupos, por exemplo, uma amostra pode ser tanto de classe A quanto de classe B.



## AMOSTRAGEM SISTÉMICA

É uma técnica dentro da categoria de amostragem probabilística – que requer certo controle do marco amostral entre os indivíduos selecionados junto com a probabilidade que sejam selecionados – consiste em escolher um indivíduo inicialmente de forma aleatória entre a população e, posteriormente, selecionar para amostra cada enésimo indivíduo disponível no marco amostral.

Por exemplo, imagine que você tem uma loja e quer selecionar uma amostra dos clientes que entram na loja. Você pode selecionar o segundo cliente que entrar na loja aquele dia e a segunda amostra será apenas o quinto cliente que entrar na loja aquele dia, e assim por diante.



Isso pode te atrapalhar um pouco caso você, por algum motivo, tenha problemas com essa amostra aleatória. Por exemplo, se você estiver pedindo feedback da sua loja, pode ser que algum cliente selecionado não queira responder sua pesquisa. Com isso, sua amostragem acaba sendo prejudicada.

Dito tudo isso, minha recomendação para vocês é que considerem bem o problema que têm em mãos, as possibilidades de manobras que os dados de vocês permitem e os prós e contras de cada método. Nada em estatística é considerado perfeito para todos os casos, por isso temos sempre que ponderar nossas escolhas.

## MATERIAL COMPLEMENTAR - AMOSTRAGEM

### TIPOS DE VIESES NA CIÊNCIA DE DADOS:

[https://www.linkedin.com/pulse/tipos-de-vieses-na-ci%C3%Aancia-dados-de-holanda-e-ayres-de-moura/?trk=public\\_profile\\_article\\_view](https://www.linkedin.com/pulse/tipos-de-vieses-na-ci%C3%Aancia-dados-de-holanda-e-ayres-de-moura/?trk=public_profile_article_view)

## O QUE É TESTE DE HIPÓTESE?

O teste de hipóteses é uma análise estatística que usa dados amostrais para comprovar uma afirmação sobre uma população. Vamos a alguns exemplos de usos:

1. Considere que um fabricante que anuncia que seu novo carro híbrido tem média de consumo de combustível de 12 km/L. Se você suspeitar que o consumo médio não é de 12 km/L, como você poderia mostrar

que o anúncio é falso? Obviamente você não pode testar todos os veículos, mas você ainda pode tomar uma decisão razoável sobre o consumo médio retirando uma amostra aleatória da população de veículos e medindo o consumo de cada um. Se a média da amostra diferir o suficiente da média do anúncio, você pode decidir que o anúncio está errado.

2. Você é analista de dados do squad de marketing de uma empresa que decidiu fazer uma campanha para lançar um novo produto. Esse produto pode atingir 2 grupos distintos: homens e mulheres. Para saber o público alvo da campanha, os marketeiros criaram uma pontuação (de 0 a 100) que mede a popularidade do produto e fizeram uma pesquisa. A popularidade entre as mulheres foi em média de 64 e entre os homens de 75. Será mesmo que essa média é maior para o grupo de homens? Podemos comprovar isso com testes de hipótese.

Para comprovarmos essas afirmações, precisamos seguir alguns passos:



Antes de explicarmos cada um dos passos, quero me adiantar (e muito!) para deixar um conceito bem claro desde já. O que faremos no Passo 4 é calcular uma **estatística de teste** para identificar se nossa afirmação é verdadeira ou não. Essa estatística de teste é um número calculado através de fórmulas matemáticas pré definidas baseadas em distribuições já bem conhecidas na estatística, como a distribuição normal, distribuição t, distribuição z, distribuição chi-quadrado, etc. Depois de calcularmos esse número que chamamos de estatística de teste, vamos compará-lo com um valor limite (chamado de nível de significância, que veremos nas próximas páginas) para verificar se nossa afirmação é verdadeira.

Como calcular cada estatística, como escolher o melhor teste, qual nível de significância e interpretação de resultados vamos ver nas cenas daqui a muitas páginas. Sei que isso pode gerar um pouco de ansiedade, mas

confiem em mim! Precisamos abordar alguns conceitos antes de entrarmos em toda a parte matemática. Vamos lá?

## AS HIPÓTESES

Toda vez que temos uma afirmação a comprovar, precisamos traduzir essa afirmação em duas hipóteses: a Hipótese Nula ( $H_0$ ) e a Hipótese Alternativa ( $H_a$ ), em que

1. Uma **hipótese nula ( $H_0$ )** é uma hipótese estatística que contém uma afirmação de igualdade, tal como  $\leq$ ,  $=$  ou  $\geq$ .
2. A **hipótese alternativa ( $H_a$ )** é o complemento da hipótese nula. É uma afirmação que é aceita como verdadeira se  $H_0$  for falsa e contém uma declaração de desigualdade estrita, tal como  $<$ ,  $\neq$  ou  $>$

Dessa forma, temos:

$$\begin{cases} H_0: \mu \leq k \\ H_a: \mu > k \end{cases} \quad \begin{cases} H_0: \mu \geq k \\ H_a: \mu < k \end{cases} \quad \begin{cases} H_0: \mu = k \\ H_a: \mu \neq k \end{cases}$$

Vamos dar alguns exemplos reais para esclarecer mais esses pontos:

- a) "Uma escola divulga que a proporção de seus estudantes que estão envolvidos em pelo menos uma atividade extracurricular é de 61%"

A afirmação "a proporção... é de 61%" pode ser escrita como  $p = 0,61$ . Seu complemento é  $p \neq 0,61$ . Como  $p = 0,61$  contém a afirmação de igualdade, ela se torna a hipótese nula. Logo, temos que

$$H_0: p = 0,61$$

$$H_a: p \neq 0,61.$$

- b) "Uma concessionária de automóveis anuncia que o tempo médio para uma troca de óleo é menor que 15 minutos"



A afirmação “a média... é menor que 15 minutos” pode ser escrita como  $\mu < 15$ . Seu complemento é média  $\geq 15$ , conforme. Como  $\mu \geq 15$  contém a igualdade, ela se torna a hipótese nula.

H0:  $\mu \geq 15$  minutos

Ha:  $\mu < 15$  minutos

- c) "Criamos um medicamento e testamos em um grupo de controle e um de teste. Queremos saber se há evidências suficientes para dizer que o grupo de teste (com média sendo  $\mu$  teste) teve uma melhora quando comparado ao grupo de controle (com média sendo  $\mu$  controle)"

Aqui temos que prestar bastante atenção, pois diferentemente das outras questões, estamos comparando 2 grupos. Como já dito, a hipótese nula precisa ter algum sinal de igualdade. Nesses casos, usualmente colocamos como hipótese nula como sendo "os efeitos são iguais", ou seja, assumimos que não há diferença estatística entre os dois grupos. Se essa é nossa hipótese nula, então nossa hipótese alternativa deve ser "os efeitos dos grupos são diferentes". Logo, dizemos que:

H0:  $\mu$  teste =  $\mu$  controle

Ha:  $\mu$  teste  $\neq \mu$  controle

Você pode pensar no H0 como a teoria padrão que requer evidência suficientemente forte em sua amostra para ser capaz de rejeitá-la. Por exemplo, quando você compara as médias de dois grupos, o valor nulo geralmente indica que a diferença entre as duas médias é igual a zero. Em outras palavras, os grupos não são diferentes.

O **efeito** é a diferença entre o valor da população e o valor da hipótese nula. O efeito também é conhecido como "efeito populacional" ou "diferença". Por exemplo, a diferença média entre o resultado de saúde de um grupo de tratamento e um grupo de controle é o efeito.

Normalmente, você não sabe o tamanho do efeito real. No entanto, você pode usar um teste de hipótese para determinar se um efeito existe e estimar seu tamanho. Por exemplo, se a média de um grupo for 10 e a média de outro grupo for 2, o efeito será 8. Com testes de hipótese vamos saber se esses



valores são **estatisticamente significativos** (de fato existe diferença) ou se esses valores são considerados estatisticamente iguais.

## EXEMPLO MASTER

A partir de agora, em quase todos os tópicos, vamos nos basear no exemplo retirado da referência Frost, J. para entender melhor cada tópico que formos abordar.

Um pesquisador está estudando os gastos com combustível para as famílias e quer determinar se o custo mensal mudou desde o ano passado, quando a média era de US\$ 260 por mês. A pesquisadora sorteia uma amostra aleatória de 25 famílias e analisa seus custos mensais para este ano.

Variável	Valor
$\bar{x}$	330.6
s	154.26
n	25

Obtivemos uma média amostral de 330,6. No entanto, é concebível que, devido ao erro amostral, a média da população seja apenas 260 (note que o desvio-padrão é muito alto!). Se o pesquisador sorteou outra amostra aleatória, a próxima média amostral pode estar mais próxima de 260. É impossível avaliar essa possibilidade olhando apenas a média amostral.

Logo, o pesquisador decide fazer um teste de hipótese. Para isso, ele traça as 2 hipóteses necessárias:

- Hipótese nula: A média da população é igual à média da hipótese nula (260).
- Hipótese alternativa: A média da população não é igual à média da hipótese nula (260).

É improvável que qualquer média amostral seja igual à média populacional devido ao **erro amostral**. No nosso caso, a média amostral de 330,6 é, numericamente, bastante diferente da média populacional para gastos com combustível do ano passado.

Imaginem o seguinte cenário agora. Temos disponíveis 1000 pesquisadores que poderão entrevistar, cada um, 25 pessoas diferentes. Portanto, cada um desses pesquisadores vai ter sua própria amostra de 25 pessoas. Se tirássemos a média de cada uma dessas amostras dos 1000 pesquisadores, provavelmente observaríamos um amplo espectro de médias amostrais. Poderíamos até representar graficamente a distribuição das médias amostrais desse processo.

Esse tipo de distribuição é chamado de **distribuição amostral**. Você obtém uma distribuição amostral extraindo muitas amostras aleatórias de uma população. E por que faríamos isso?

Porque essas distribuições amostrais permitem determinar a **probabilidade** de obter sua **estatística de amostra** e são cruciais para realizar testes de hipóteses.

Felizmente, não precisamos nos dar ao trabalho de coletar várias amostras aleatórias! Os procedimentos estatísticos estimam as distribuições amostrais usando as propriedades das amostras. Nos próximos tópicos vamos nos aprofundar bastante nisso. Por enquanto, quero que você se concentre na ideia de que a amostra coletada pelo estudo é apenas uma dentre um número infinito de amostras potenciais que poderia ter tirado. Esse é um conceito crucial em testes de hipóteses e estatísticas inferenciais.

## PREMISSE DA HIPÓTESE NULA

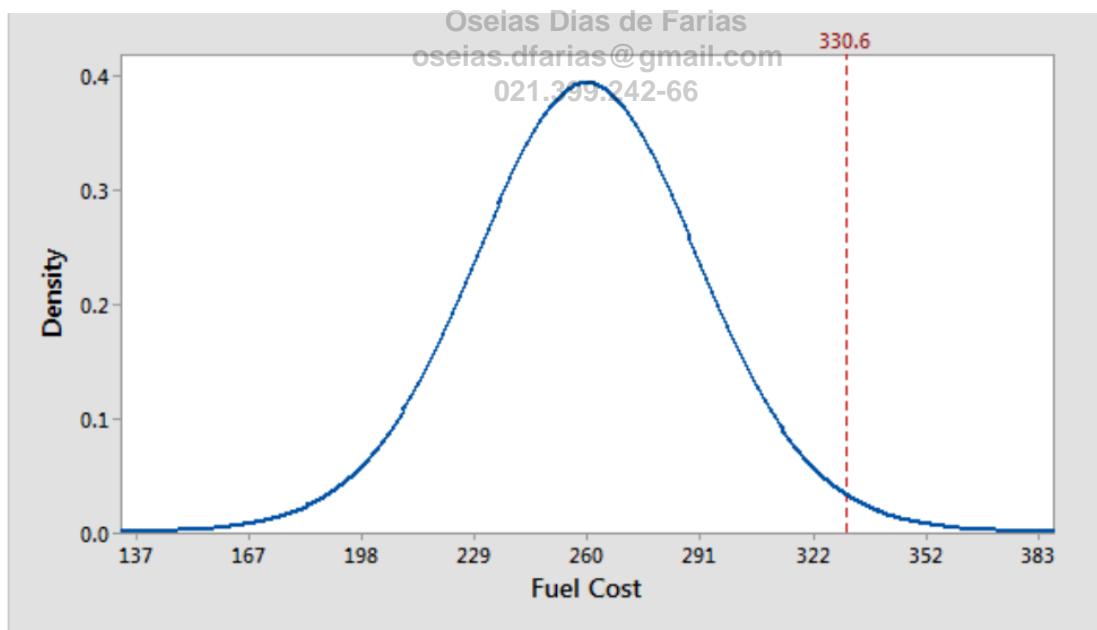
Como dissemos no exemplo anterior, queremos saber se o gasto médio com combustível neste ano (330,6) é diferente do ano passado (260).

Para responder a essa pergunta, faremos um **gráfico da distribuição de amostragem com base na suposição de que o custo médio de combustível para toda a população não mudou e ainda é 260**.

A grande premissa aqui é que os testes de hipótese sempre usam **distribuições de amostragem que assumem que a hipótese nula está correta**. Na prática, não sabemos se de fato  $H_0$  está correto ou não (inclusive estamos querendo comprovar isso!), mas o teste de hipótese exige um ponto de partida para que possa ser realizado, e esse ponto é assumir **inicialmente** que  $H_0$  está correto.

Assim, usamos o valor da hipótese nula como base de comparação para nosso valor amostral observado. O gráfico abaixo mostra um exemplo de distribuição de médias amostrais tendo em vista que a hipótese nula é verdadeira - ou seja, a média populacional é 260.

Podemos colocar nossa média amostral (330,6) nesta distribuição. Esse contexto mais amplo nos ajuda a ver quão improvável é nossa média amostral se a hipótese nula estiver correta ( $\mu = 260$ ).



A distribuição amostral indica que é relativamente improvável obter uma amostra de 330,6 se a média populacional for 260. Mas será que nossa média amostral é tão improvável assim a ponto de rejeitar que média populacional seja 260?

Em estatística, chamamos isso de **rejeitar a hipótese nula**. Se rejeitarmos o nulo para nosso exemplo, a diferença entre a média amostral (330,6) e 260 é **estatisticamente significativa**. Em outras palavras, os dados amostrais favorecem a hipótese de que a média populacional **não** é igual a 260.

No entanto, olhe novamente para o gráfico de distribuição amostral. Observe que não há um local específico na curva onde você possa tirar essa conclusão definitivamente. Há apenas uma diminuição consistente na probabilidade de observar médias amostrais que estão mais distantes do valor da hipótese nula. Como decidimos que uma média amostral está longe o suficiente?

Para responder a essa pergunta, precisaremos de mais ferramentas de teste de hipóteses! O procedimento de teste de hipótese quantifica a anormalidade de nossa amostra com uma probabilidade e então a compara com um padrão probatório. Esse processo permite que você tome uma decisão objetiva sobre a força da evidência.

Veremos em cada um dos tópicos abaixo como resolver esse problema a partir daqui.

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

## NÍVEL DE SIGNIFICÂNCIA (ALPHA OU $\alpha$ )

O nível de significância é um valor **arbitrário** (ou seja, escolhido por nós) que definimos **antes** do estudo. A grosso modo é a probabilidade de você dizer que há um efeito quando não há efeito. Níveis de significância mais baixos indicam que você precisa de evidências mais substanciais antes de rejeitar o nulo, ou seja, quanto menor o alpha, mais facilmente você **rejeita** a hipótese nula.

Por exemplo, um nível de significância de 0,05 significa um risco de 5% de decidir que **existe um efeito quando não existe**.

Mas vamos entrar em mais detalhes caso ainda tenha ficado confuso.

Seus dados fornecem evidências de um efeito (diferença de média entre dois grupos, por exemplo). O nível de significância é uma medida de quão forte a evidência da amostra deve ser antes de determinar se os resultados são **estatisticamente significativos**.

Ele define a linha entre a evidência ser forte o suficiente para concluir que o efeito existe na população versus é fraco o suficiente para que não possamos descartar a possibilidade de que o efeito amostral seja apenas um erro de amostragem aleatória.

Vamos fazer um paralelo com casos criminais. Casos criminais e casos civis uma quantidade mínima de provas para convencer um juiz ou júri a provar uma reclamação contra o réu. Para casos civis, a maioria dos estudiosos define que pelo menos 51% das provas apresentadas precisam apoiar o réu para que ele seja considerado culpado. Os casos criminais, que são mais graves, exigem provas mais substanciais, que devem ir além de uma dúvida razoável. A maioria dos estudiosos define esse padrão como sendo 90%, 95% ou mesmo 99% de certeza de que o réu é culpado.

O nível de significância pode ser associado a esse percentual de certeza que temos que ter para dizer se um réu é culpado ou não. Para aceitarmos o fato de que o réu é culpado, precisamos ter certa certeza (no caso criminal, 99% de certeza).

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

Assim como o "nível de certeza" varia de acordo com o tipo de processo judicial, você pode definir o nível de significância para um teste de hipótese dependendo das consequências de um falso positivo. Ao alterar o alpha, você aumenta ou diminui a quantidade de evidência necessária na amostra para concluir que o efeito existe na população.

O valor 0,05 é o alpha padrão que vocês vão encontrar em muitos estudos por aí. Aumentar o nível de significância de 0,05 para 0,10 diminui o "nível de certeza" para não rejeitar  $H_0$ . Por outro lado, diminuir de 0,05 para 0,01 aumenta a barra. Vamos dar uma olhadinha nas consequências disso.

## AUMENTANDO O NÍVEL DE SIGNIFICÂNCIA

Imagine que você está testando a força dos balões de festa. Você usará os resultados do teste para determinar qual marca de balões comprar. Um falso positivo aqui leva você a comprar balões que não são mais fortes. As desvantagens de um falso positivo são muito baixas. Consequentemente, você pode considerar diminuir a quantidade de evidência necessária alterando o nível de significância para 0,10. Como essa alteração diminui a



quantidade de evidência necessária, torna seu teste mais sensível à detecção de diferenças, mas também aumenta a chance de um falso positivo de 5% para 10%.

## DIMINUINDO O NÍVEL DE SIGNIFICÂNCIA

Por outro lado, imagine que você está testando a resistência do tecido para balões de ar quente. Um falso positivo aqui é muito arriscado porque vidas estão em jogo! Você quer ter muita certeza de que o material de um fabricante é mais forte que o outro. Nesse caso, você deve aumentar a quantidade de evidências exigidas alterando alfa para 0,01. Como essa alteração aumenta a quantidade de evidências necessárias, torna seu teste menos sensível à detecção de diferenças, mas diminui a chance de um falso positivo de 5% para 1%.

De novo, assim como várias questões de estatística, a escolha de um nível de significância adequado depende do seu problema. Quão problemático é um falso positivo? Você está disposto(a) a correr riscos?

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

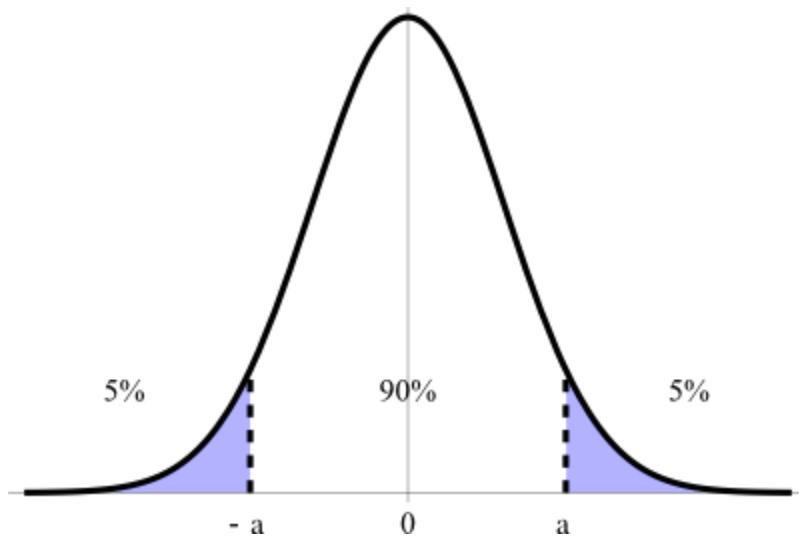
## VISUALIZANDO O NÍVEL DE SIGNIFICÂNCIA

Sei que toda essa história de nível de significância pode ser bem confusa, então vamos visualizá-lo de uma forma gráfica (e, prometo que nas próximas seções você vai entender isso muito melhor)

No gráfico de distribuição de probabilidade, o nível de significância define quão longe o valor da amostra deve estar antes que possamos rejeitar a hipótese nula. A área sob a curva que está sombreada é igual à probabilidade de que o valor da amostra caia nessas regiões se a hipótese nula estiver correta.

Para representar um nível de significância de 0,1, vou sombrear 10% da distribuição mais distante do valor nulo (5% de cada lado - vocês vão ver mais sobre isso no tópico "Testes Bicaudais e Unicaudais").





As duas regiões sombreadas no gráfico são equidistantes do valor central de  $H_0$ . Cada região tem uma probabilidade de 0,05, que soma ao total desejado de 0,1. Essas áreas sombreadas são ~~chamadas de~~ chamadas de **regiões críticas** para um teste de hipótese bicaudal (de novo, vocês vão entender o que é bicaudal no tópico "Testes Bicaudais e Unicaudais").

Se nossa estatística de teste calculada (vocês vão aprender a calcular mais pra frente) cair na região de rejeição, falamos que estamos muito "distantes" do que a hipótese nula considera correta - ou seja, rejeitamos  $H_0$ .

Para nosso "Exemplo Master" sobre os gastos com combustível, vamos definir que alpha será o padrão mais usado: 0,05. Ou seja, não rejeitamos  $H_0$  (média sendo 260) apenas se nossa estatística de teste não cair na região de rejeição.

Vamos agora aprender se nosso "Exemplo Master" é um teste uni ou bicaudal e onde estão essas tais regiões de rejeição para esse exemplo.

## TESTES BICAUDAIS E UNICAUDAIS

Esse passo é fundamental para definirmos bem alguns parâmetros do teste de hipótese mais adiante. Mas antes, vamos voltar algumas casas e relembrar o tópico "As Hipóteses" que vimos acima. Falamos que toda vez que temos

uma afirmação a comprovar, precisamos traduzir essa afirmação em duas hipóteses: a Hipótese Nula ( $H_0$ ) e a Hipótese Alternativa ( $H_a$ ), em que

1. Uma **hipótese nula ( $H_0$ )** é uma hipótese estatística que contém uma afirmação de igualdade, tal como " $\leq$ ", " $=$ " ou " $\geq$ ".
2. A **hipótese alternativa ( $H_a$ )** é o complemento da hipótese nula. É uma afirmação que é aceita como verdadeira se  $H_0$  for falsa e contém uma declaração de desigualdade estrita, tal como " $<$ ", " $\neq$ " ou " $>$ "

Agora reparem aqui comigo. Vocês notaram que nossa hipótese alternativa ( $H_a$ ) pode ter basicamente 3 sinais diferentes? Ela pode ter o sinal " $<$ ", " $\neq$ " ou " $>$ ". Essa é a chave de ouro para entender se nosso teste será bicaudal ou unicaudal.

## TESTE BICAUDAL

Se a hipótese alternativa  $H_a$  contém o símbolo “diferente de” ( $\neq$ ), então o teste de hipótese é um teste bicaudal.

Oceias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

Quando o  $H_a$  tem um sinal de " $\neq$ " não me comprometo em saber quem é maior e quem é menor, quero detectar apenas a diferença/igualdade (no caso,  $H_0$  terá o sinal de " $=$ "). O teste bicaudal é usado se os desvios do parâmetro estimado em qualquer direção de algum valor de referência são considerados teoricamente possíveis. Por exemplo, imagine que temos a seguinte situação:

*Tenho uma fábrica que produz bolachas em pacotes de 200 gramas cada. A agência que regulariza a venda de bolachas faz uma vistoria em minha fábrica para comprovar se não estou vendendo uma informação falsa aos consumidores. Logo, eles traçam a seguinte hipótese:*

$H_0$ : A média de peso em cada pacote é de 200 gramas  $\rightarrow \mu = 200g$

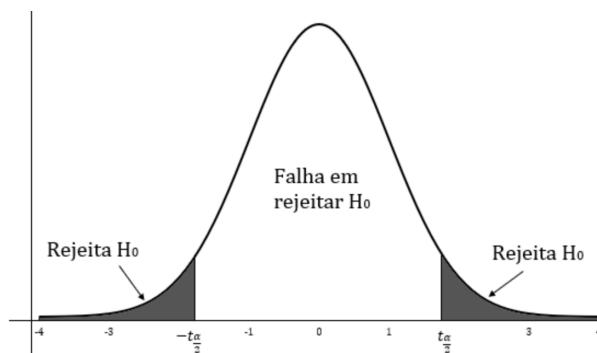
$H_a$ : A média de peso em cada pacote é de 200 gramas  $\rightarrow \mu \neq 200g$

Esse é um típico caso de um teste bicaudal. Sabemos disso pois  $H_a$  contém o símbolo " $\neq$ ", ou seja, tanto faz nesse caso se vendo mais ou menos, o que a agência está verificando é se estou passando a informação correta aos meus



consumidores (claro que como consumidores dificilmente nos importaríamos de vir mais produto, mas não é o que a agência está querendo verificar!)

E quais as consequências disso? Isso muda a forma que olhamos para o nível de significância. Se você estiver usando um nível de significância de 0.05, um teste bicaudal aloca **metade** de seu alpha para testar a significância estatística em uma direção e metade de seu alpha para testar a significância estatística na outra direção.



Oseias Dias de Farias

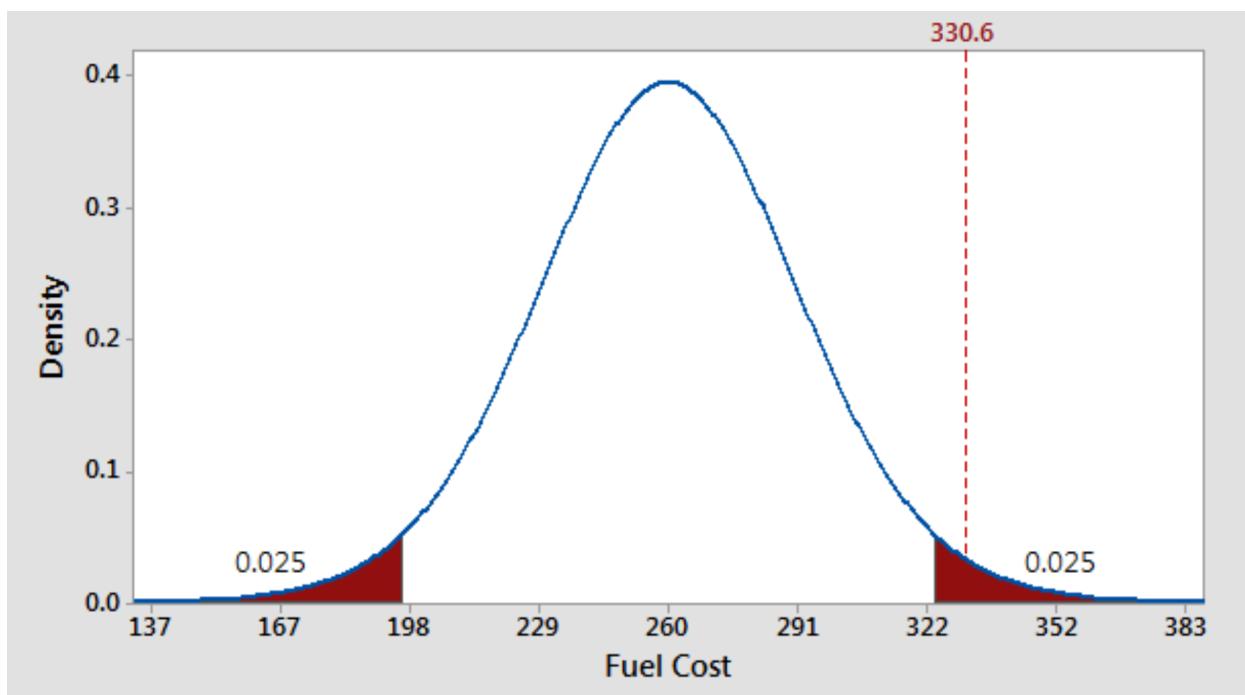
Isso significa que 0.025 (metade de 0.05) está em cada cauda da distribuição de sua estatística de teste. Ao usar um teste bicaudal, independentemente da direção do relacionamento que você supõe, você está testando a possibilidade do relacionamento em ambas as direções.

No nosso caso da fábrica de bolachas, a média é considerada significativamente diferente de 200 gramas se a **estatística de teste** estiver nos 2,5% superiores ou 2,5% inferiores de sua distribuição de probabilidade.

Vamos explicar melhor essa região sombreada com nosso caso do "Exemplo Master" em que temos que comprovar se o gasto médio com combustível foi de U\$260.

Como nossa hipótese alternativa tem um sinal "diferente de" então representamos a distribuição como sendo bicaudal. Nesse caso, vamos supor que queremos um alpha de 0.05 como dito anteriormente. Logo, nossa região crítica será:





As duas regiões sombreadas no gráfico são equidistantes do valor central da hipótese nula ( $H_0$  dizia que a média é 260). Cada região tem uma probabilidade de 0,025, que soma ao total desejado de 0,05. Essas áreas sombreadas são chamadas de **regiões críticas**.

A região crítica define valores de amostra que são **improváveis** o suficiente para justificar a rejeição da hipótese nula. Se a hipótese nula estiver correta e a média da população for 260, amostras aleatórias ( $n=25$ ) dessa população têm médias que caem nas regiões críticas 5% das vezes.

Se nossa **estatística de teste** (aquele número que vamos aprender a calcular com teste de hipótese) cair na região crítica, dizemos que nossa média amostral é **estatisticamente significativa** no nível de 0,05 porque cai na região crítica - ou seja, rejeitamos  $H_0$ . Em outras palavras, a média não será de 260.

Vamos aprender a calcular a estatística de teste ao longo do curso, não se preocupem!

## UNICAUDAL



Além do teste bicaudal, temos a categoria de testes unicaudais. Os testes são os testes que contém os sinais ">" ou "<" na hipótese alternativa.

Dizemos que um teste é **unicaudal à direita** quando a hipótese alternativa contém o sinal de o símbolo “maior que” (>). Por exemplo:

*Tenho uma fábrica de tênis e quero comprovar que meu tênis aumenta o desempenho dos atletas em 20%. Se com tênis comuns os atletas fazem um percurso determinado em 10 segundos, com meu tênis eles passarão a fazer em 8 segundos. Para isso, pedi para alguns atletas correrem esse percurso com tênis comuns e depois correrem com meus tênis. Logo, temos que:*

$H_0$ : A média de tempo é menor ou igual que 8 segundos  $\rightarrow \mu \leq 8$  s.

$H_a$ : A média de tempo é maior que 8 segundos  $\rightarrow \mu > 8$  s.

Esse é o típico caso de teste unicaudal à direita. Não estou aqui interessada em comprovar que a média de tempo está em uma determinada faixa de valor. Queremos comprovar que nosso tênis é melhor, ou seja, queremos comprovar que o percurso pode ser feito em menos tempo. Portanto,  $H_a$  tem o sinal de ">".

Observando a curva, podemos ver que se a média cair na região a **direita** (região sombreada de velocidades mais altas), rejeitamos  $H_0$  e dizemos que nosso tênis não melhora a performance do atleta.



Por outro lado, dizemos que um teste é **unicaudal à esquerda** quando a hipótese alternativa contém o sinal de o símbolo “menor que” (<). Por exemplo:

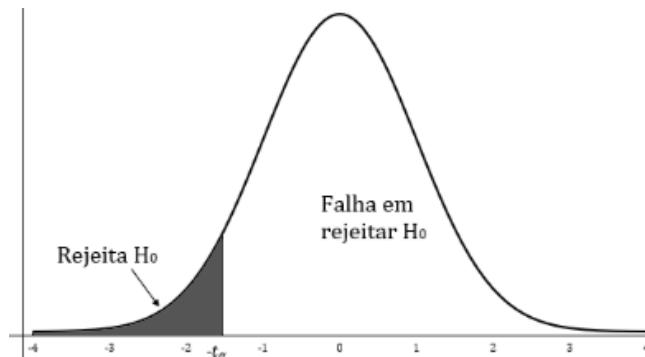
Tenho uma empresa de telemarketing e vejo que os atendentes têm performances muito baixas (performance = 1.2). Faço um treinamento em uma pequena amostra desses atendentes e vejo que agora tenho uma performance maior (performance = 3.5). Agora, preciso comprovar que meu teste foi de fato significativo:

$H_0$ : A média de performance dos atendentes é maior ou igual a 3.5  $\rightarrow \mu \geq 3.5$ .

$H_a$ : A média de performance dos atendentes é menor que 3.5  $\rightarrow \mu < 3.5$ .

Esse é o típico caso de teste unicaudal à esquerda. Não estou aqui interessada em comprovar que a média de performance está em uma determinada faixa de valor. Queremos comprovar que o treinamento dado foi efetivo, ou seja, queremos comprovar que a performance aumentou. Portanto,  $H_a$  tem o sinal de "<".

Observando a curva, podemos ver que se a média cair na região à esquerda (região sombreada de performances menores), rejeitamos  $H_0$  e dizemos que nosso treinamento não melhora a performance do time.



## P-VALOR

A definição formal de p-valor diz que "o p-valor representa a probabilidade de você obter o efeito observado em sua amostra, ou maior, se a hipótese nula estiver correta".

Em termos mais simples, o p-valor informa o quanto seus dados de amostra contradizem a hipótese nula. Se temos um p-valor baixo, temos evidência mais forte contra o  $H_0$ .

Na prática, usamos o p-valor para saber se devemos rejeitar ou não  $H_0$ , ou seja, se nossos dados nos mostram evidências suficientemente fortes para nos dizer se nossa estatística de teste cai na região de rejeição ou não.

Se o p-valor for **menor ou igual ao nível de significância**, você rejeita a hipótese nula. Quando o p-valor é **maior que o nível de significância**, **não** rejeitamos  $H_0$ . Vamos a alguns exemplos:

"Uma escola divulga que a proporção de seus estudantes que estão envolvidos em pelo menos uma atividade extracurricular é de 61%"

A afirmação "a proporção... é de 61%" pode ser escrita como  $prop = 0,61$ . Seu complemento é  $prop \neq 0,61$ . Como  $prop = 0,61$  contém a afirmação de igualdade, ela se torna a hipótese nula. Logo, temos que:

$H_0: prop = 0,61$

$H_a: prop \neq 0,61$ .

Oseias Dias de Farias

Aqui, queremos trabalhar com um nível de confiança de 95% - ou seja,  $\alpha = 0,05$ . Rodamos um teste de hipótese e constatamos que **o p-valor foi de 0,1**. Logo,  $p\text{-valor} > 0,05$  e, portanto, não rejeitamos  $H_0$ .

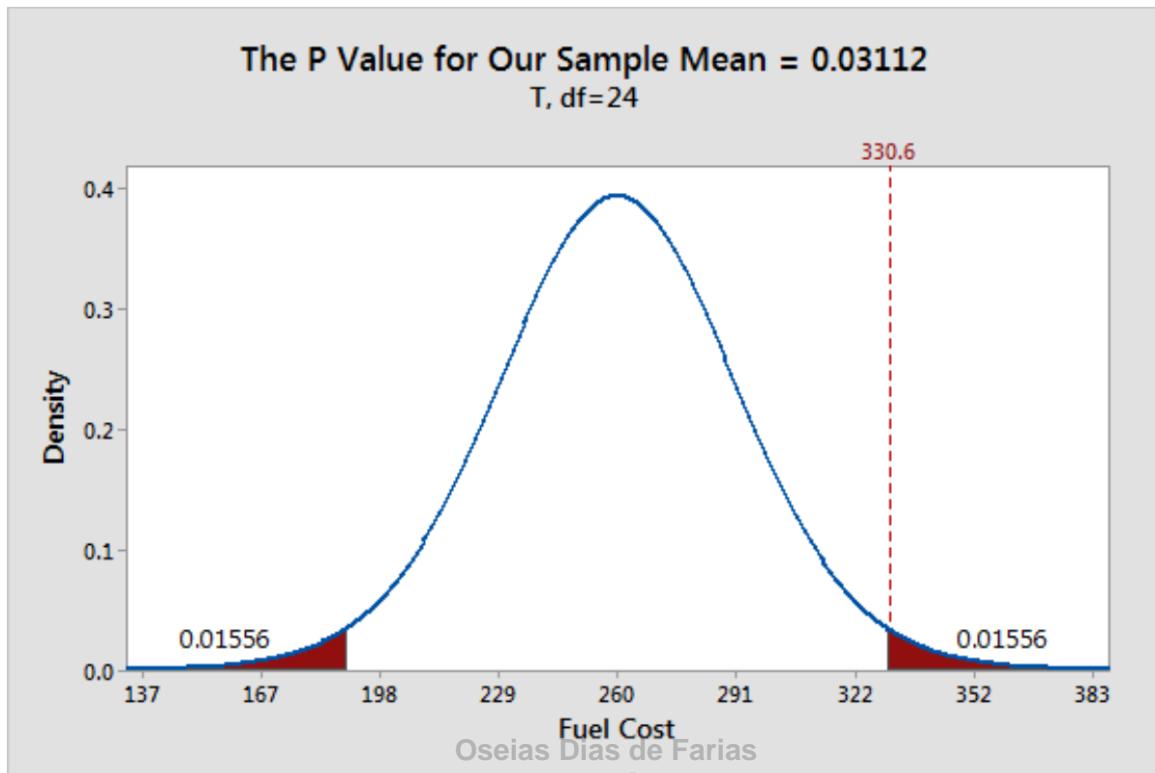
### Bora voltar no nosso exemplo de gasto de combustível?

Para entender o p-valor, precisamos primeiro calcular o efeito que está presente em nossa amostra considerando que  $H_0$  é verdade (média = 260).

O efeito é a distância entre o valor da amostra e o valor nulo:  $330,6 - 260 = 70,6$ .

Em seguida, vamos sombrear as regiões em ambos os lados da distribuição que estão pelo menos tão distantes quanto 70,6 do nulo ( $260 \pm 70,6$ ). Esse processo representa graficamente a probabilidade de observar uma média amostral extrema para os casos em que nosso  $H_0$  é verdade.





A probabilidade total das duas regiões sombreadas é 0,03112 ( $2 \times 0.01556$ ), que é a área embaixo da curva (região sombreada) que ultrapassa os limites  $260 \pm 70,6$  (lembrando aqui que estamos usando um teste bicaudal!)

Mas vocês bem sabem que é difícil medir uma área embaixo da curva, certo? Então como eu encontrei esse valor mágico de 0.01556? Com o teste de hipótese adequado para esse problema, que vocês só verão daqui a muitas páginas (*spoiler alert*: esse teste é o teste  $t - 1$  amostra). Por enquanto vamos nos concentrar em entender o que é o p-valor nesse contexto todo.

Voltando ao fato de que a soma das duas regiões é 0,03112. Lemos isso também como 3,112% de probabilidade de obter um valor nessas regiões (se não entendeu isso, volte algumas casas e vá para a sessão "Função de distribuição acumulada").

Lembram-se da nossa zona crítica? Pois bem, lá nos dissemos que caso H0 fosse verdadeiro, a média dos dados seria em torno de 260. Considerando essa distribuição fictícia de H0, se uma média amostral calculada ultrapassasse a zona crítica, diríamos que nossa amostra não concorda com H0. E é isso que acontece aqui!

Nossa média amostral está tão distante da média de H0 que ela caiu na região de rejeição (região onde temos valores super improváveis de acontecer **SE** H0 fosse verdade).

Nós aqui não estamos querendo provar se as amostras estão corretas - aqui estamos assumindo que fizemos de tudo para evitar os viéses amostrais e que essa amostra poderia representar bem a população. Pois bem, o que queremos agora com o teste de hipótese é comprovar se H0 está ou não correto. E vimos que, para um cenário de ele estar correto, nossa amostra não concordaria com ele. Portanto, sabendo que nossa amostra representa bem nossa população, nós **rejeitamos o H0**.

Oseias Dias de Farias  
Na prática, se o seu p-valor for menor ou igual ao seu nível alpha, rejeite a hipótese nula.  
oseias.diasdefarias@gmail.com  
021.399.242-66

Logo, com nível de significância de 0,05, rejeitamos o p-valor ( $p\text{-valor} \leq \alpha$ ). Nossos dados suportam a **hipótese alternativa**, que afirma que a média populacional **não** é igual a 260. **Podemos concluir que os gastos médios com combustível não são iguais a 260.** Como a média amostral (330.6) por si só é maior que 260, então concluímos que há **um aumento em gastos de combustível desde o ano passado.**

As estatísticas usam p-valores em todos os lugares. Você encontrará p-valores em testes t, testes de distribuição, ANOVA e análise de regressão. Eles se tornaram tão cruciais que ganharam vida própria. Eles podem determinar quais estudos são publicados, quais projetos recebem financiamento, se vamos investir dinheiro em determinados processos de melhoria ou campanhas de marketing e até quais membros do corpo docente universitário se tornam efetivos!

**Os p-valores NÃO são uma taxa de erro!**

Infelizmente, os p-valores são frequentemente mal interpretados. Um erro comum é que eles representam a probabilidade de rejeitar uma hipótese nula que é realmente verdadeira. A ideia de que os p-valores são a probabilidade de cometer um erro é **errada!**

Você não pode usar p-valores para calcular a taxa de erro diretamente por vários motivos.

Primeiro, os cálculos do p-valor assumem, inicialmente, que a hipótese nula está correta. Assim, do ponto de vista do p-valor, a hipótese nula é 100% verdadeira, então não há erro.

Em segundo lugar, os p-valores informam quão consistentes são seus dados de amostra com uma hipótese nula verdadeira. No entanto, quando seus dados são muito inconsistentes com a hipótese nula, os p-valores por si só não determinam qual das duas possibilidades a seguir é mais provável:

- a) A hipótese nula é verdadeira, mas sua amostra é incomum devido ao erro de amostragem aleatória.
- b) A hipótese nula é falsa.

A única coisa que p-valor vai nos dizer é que esses dados amostrais NÃO concordam com o H<sub>0</sub>. Assumindo que você fez uma amostragem correta, então iríamos direto para a letra b (a hipótese nula é falsa). Se você, por algum acaso desconfiar da sua amostragem, você deve aplicar o conhecimento especializado da área de estudo e avaliar resultados de estudos semelhantes. Depois de certo que sua amostragem está correta, aí sim você pode assumir a letra b como correta.

Vamos voltar ao nosso exemplo de combustível para solidificar o que é certo e o que é errado. Nosso p-valor era de 3,112%. **Se** H<sub>0</sub> fosse verdadeiro:

- **Correto:** Você obteria o efeito amostral (média) calculada ou maior, em 3,112% dos estudos por causa do erro amostral aleatório.
- **Incorreto:** há 3,112% de chance de cometer um erro ao rejeitar a hipótese nula.

## INTERVALO DE CONFIANÇA

Na estatística inferencial, um objetivo principal é estimar parâmetros populacionais. Esses parâmetros são os valores desconhecidos para toda a população, como a média populacional e o desvio padrão. Esses valores de parâmetros não são conhecidos e, normalmente, é impossível medir uma população inteira. O erro de amostragem produz incerteza, ou uma margem de erro, em torno de nossas estimativas.

Suponha que definamos nossa população como todos os jogadores de futebol do ensino médio. Em seguida, extraímos uma amostra aleatória dessa população e calculamos a altura média de 181 cm. Esta estimativa amostral de 181 cm é a melhor estimativa da altura média da população. Como a média é de uma amostra, é praticamente garantido que nossa estimativa do parâmetro populacional não esteja exatamente correta.

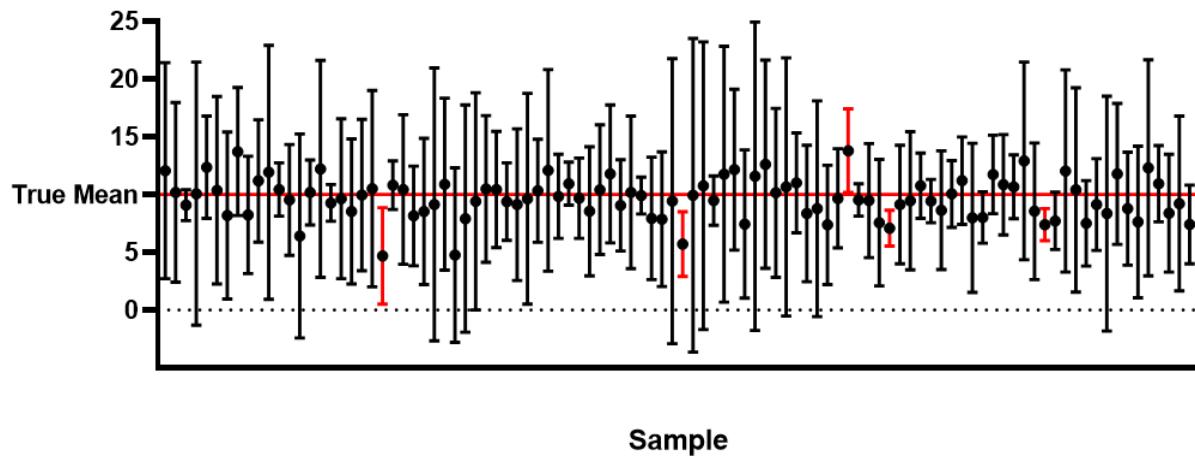
Os intervalos de confiança incorporam a incerteza e o erro da amostra para criar uma faixa de valores dentro da qual o valor real da população deve cair. Por exemplo, um intervalo de confiança de [176 186] indica que podemos ter certeza, a um determinado nível, de que a média populacional real está dentro desse intervalo.

Diferentes amostras aleatórias retiradas da mesma população podem produzir intervalos ligeiramente diferentes. Se você extrair muitas amostras aleatórias e calcular um intervalo de confiança para cada amostra, uma proporção específica dos intervalos conterá o parâmetro populacional. Essa porcentagem é o nível de confiança.

Por exemplo, um nível de confiança de 95% sugere que, se você extrair 20 amostras aleatórias da mesma população, espera-se que 19 dos intervalos de confiança incluam o valor da população, conforme mostrado abaixo.



### 95% confidence intervals for 100 samples with n=3 and mean=10



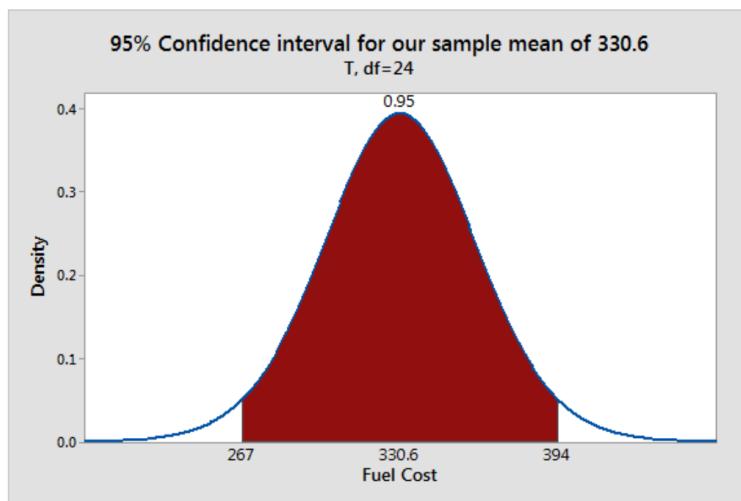
Os intervalos de confiança incluem a estimativa pontual para a amostra com uma margem de erro em torno da estimativa pontual. A estimativa pontual é o valor mais provável do parâmetro e é igual ao valor amostral. A margem de erro é responsável pela quantidade de dúvida envolvida na estimativa do parâmetro populacional. Quanto mais dispersão houver nos dados da amostra, menos precisa será a estimativa, o que faz com que a margem de erro se estenda além da estimativa pontual. Os intervalos de confiança ajudam você a navegar pela incerteza de quão bem uma amostra estima um valor para uma população inteira.

Com isso em mente, os intervalos de confiança podem ajudá-lo a comparar a precisão de diferentes estimativas. Suponha que dois estudos estimem a mesma média de 10. Parece que eles obtiveram os mesmos resultados. No entanto, usando intervalos de confiança de 95%, vemos que um intervalo é [5 15] enquanto o outro é [9 11]. O último intervalo de confiança é mais estreito, o que sugere que se trata de uma estimativa mais precisa.

Existem duas diferenças críticas entre os gráficos de distribuição amostral para **níveis de significância** e **intervalos de confiança**. A 95% de significância, gráfico de nível de significância centra-se no **valor nulo (premissa da hipótese nula)** e sombreamos os 5% externos da distribuição. Por outro lado, o gráfico do intervalo de confiança centra-se na **média da**

**amostra** e sombreados o centro 95% da distribuição.

Voltando ao nosso exemplo do preço do combustível, plotamos os 25 valores de gastos de combustíveis que tivemos na amostra:



Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242.88

A faixa sombreada das médias amostrais [267 392] cobre 95% desta distribuição amostral. Esse intervalo é o intervalo de confiança de 95% para nossos dados de amostra. Ou seja, em 95% das amostras a média de combustível ficou entre 267 e 392.

A verdade é que não sabemos se nossa média amostral está próxima da média populacional. No entanto, sabemos que a média amostral é uma **estimativa imparcial** da média populacional. Uma estimativa imparcial é aquela que não tende a ser muito alta ou muito baixa. Está correto na média. Os intervalos de confiança estão corretos em média porque usam estimativas de amostra que estão corretas em média. Dado o que sabemos, a média amostral é o valor mais provável para a média populacional.

Dada a distribuição amostral, não seria incomum que outras amostras aleatórias retiradas da mesma população tivessem médias dentro da área sombreada. Em outras palavras, dado que obtivemos a média amostral de 330,6, não seria surpreendente obter outras médias amostrais dentro da faixa sombreada - claro, se você coletou uma amostra sem viéses!

Se essas outras médias amostrais não forem incomuns, devemos concluir que esses outros valores também são prováveis candidatos para a média populacional. Há **incerteza** inerente quando você usa dados de amostra para fazer inferências sobre uma população inteira. Os intervalos de confiança ajudam a avaliar a quantidade de incerteza em suas estimativas de amostra.

**Intervalos de confiança e p-valor sempre concordam.** Se você quiser determinar se os resultados do seu teste são estatisticamente significativos, você pode usar p-valor com níveis de significância alpha ou intervalos de confiança. Essas duas abordagens sempre concordam.

A relação entre o nível de confiança e o nível de significância para um teste de hipótese é a seguinte:

$$\text{Nível de confiança} = 1 - \text{Nível de significância (alfa)}$$

Por exemplo, se seu nível de significância for 0,05, o nível de confiança equivalente será de 95%.

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Ambas as condições a seguir representam um teste de hipótese com resultados estatisticamente significativos:

- O p-valor é menor que o nível de significância.
- O intervalo de confiança exclui o valor da hipótese nula.

Além disso, é sempre verdade que quando o p-valor é menor que seu nível de significância, o intervalo exclui o valor da hipótese nula.

No exemplo do custo do combustível, nossos resultados de teste de hipótese são estatisticamente significativos porque o p-valor (0,03112) é menor que o nível de significância (0,05). Da mesma forma, o intervalo de confiança de 95% [267 394] exclui o valor da hipótese nula (260). **Usando qualquer um dos métodos, chegamos à mesma conclusão.**

Para entender a base desse acordo, precisamos lembrar como funcionam os níveis de confiança e os níveis de significância:

- Um nível de confiança determina a distância entre a média da amostra e os limites de confiança.
- Um nível de significância determina a distância entre a média amostral e as regiões críticas.

Ambos os conceitos especificam uma distância da média a um limite e essas são precisamente do mesmo comprimento.

## TIPOS DE ERROS

Antes de abordarmos mais profundamente as técnicas de teste de hipótese, quero que vocês tenham em mente os tipos de erros que podemos cometer em nossas análises.

Apesar de testes de hipótese serem uma ótima aproximação de como nossa população funciona de fato, existem desvantagens quando você usa amostras. As amostras que usamos são tipicamente uma porcentagem minúscula de toda a população. Consequentemente, elas ocasionalmente deturpam a população e podem gerar erros.<sup>6</sup>

Vamos levar em conta a situação do que chamamos de nosso "Exemplo Master"

- Hipótese nula ( $H_0$ ): A média da população é igual à média da hipótese nula (260) - Logo, não existe efeito (as médias são iguais)
- Hipótese alternativa ( $H_a$ ): A média da população **não** é igual à média da hipótese nula (260) - Logo, existe efeito (as médias são diferentes)

Idealmente, um teste de hipótese não falharia - ou seja, não rejeitaria a hipótese nula se de fato a média fosse 260 (igual a do ano anterior) e rejeitaria a hipótese nula se a média fosse diferente de 260.

Mas como dissemos anteriormente, infelizmente não vivemos num mundo ideal e testes de hipótese podem falhar.

Os estatísticos definem dois tipos de erros no teste de hipóteses. De forma não tão criativa, eles chamam esses erros de erros Tipo I e Tipo II. Ambos os tipos de erros estão relacionados a conclusões incorretas sobre a hipótese



nula. A tabela abaixo resume os quatro resultados possíveis para um teste de hipótese.

	Rejeita H0	Não rejeitar H0
H0 é verdadeiro (real)	Erro tipo I: Falso positivo (FP)	Acertamos \o/ Efeito não existe
H0 é falso (real)	Acertamos \o/ Efeito existe	Erro tipo II: Falso negativo (FN)

## ERRO TIPO I

Por que esses erros ocorrem? Tudo se resume a erro de amostra. Sua amostra aleatória superestimou o efeito por acaso. Esse tipo de erro não indica que os pesquisadores fizeram algo errado. O desenho experimental, a coleta de dados, a validação dos dados e a análise estatística podem estar corretos, mas esse erro ainda ocorre.

Oseias Dias de Farias  
oseias.dias@gmail.com  
021.399.242-66

Embora não saibamos quais estudos têm resultados falso-positivos, sabemos sua taxa de ocorrência. A taxa de ocorrência de erros do Tipo I é igual ao nível de significância do teste de hipótese, também conhecido como **alpha ( $\alpha$ )**.

O nível de significância é um padrão de evidência que você define para determinar se os dados de sua amostra são fortes o suficiente para rejeitar a hipótese nula. Os testes de hipóteses definem esse padrão usando a probabilidade de rejeitar uma hipótese nula verdadeira. Você define esse valor com base em sua **disposição de arriscar um falso positivo**.

Se você pensar nas distribuições de amostragem, faz sentido. As distribuições amostrais assumem que a hipótese nula está correta. O nível de significância define as regiões críticas. Portanto, quando a hipótese nula está correta, você espera que os resultados do teste caiam nas regiões críticas com uma probabilidade definida pelo nível de significância.

Quando o nível de significância é 0,05 e a hipótese nula é verdadeira, há 5% de chance de que o teste rejeite a hipótese nula incorretamente. Se você definir

alpha para 0,01, haverá 1% de um falso positivo. Se 5% é bom, então 1% parece ainda melhor, certo? Como você verá, há uma compensação entre os erros Tipo I e Tipo II. Se você mantiver todo o resto constante, ao reduzir a chance de um falso positivo, você aumenta a oportunidade de um falso negativo.

Os erros do tipo I são relativamente simples. Os estatísticos projetaram testes de hipóteses para incorporar tudo o que afeta essa taxa de erro, o que permite especificá-la para seus estudos. Se o seu projeto experimental é sólido, você coleta dados válidos e os dados satisfazem as suposições do teste de hipótese, a taxa de erro Tipo I é igual ao nível de significância que você especifica. No entanto, se houver um problema em uma dessas áreas, isso pode afetar a taxa de falsos positivos.

## ERRO TIPO II

A taxa de erro Tipo II é a probabilidade de um falso negativo. A probabilidade de cometer um erro do Tipo II é conhecida como **beta ( $\beta$ )**.

O que causa os erros do tipo II?  
Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Vamos revisitar brevemente os erros do Tipo I. Quando seu estudo faz tudo corretamente, o erro de amostragem é a única coisa que causa erros do Tipo I.

Em contrapartida, 3 motivos principais para acontecer o erro Tipo II - tamanhos de efeito pequenos, tamanhos de amostra pequenos e alta variabilidade de dados. Além disso, ao contrário dos erros do Tipo I, você não pode definir a taxa de erros do Tipo II para sua análise. Em vez disso, o melhor que você pode fazer é estimá-lo antes de iniciar seu estudo, aproximando as propriedades da hipótese alternativa que você está estudando. Quando você faz esse tipo de estimativa, é chamado de análise de poder (*Power Analysis*).

Ao estimar a taxa de erro Tipo II, seu software estatístico cria uma distribuição de probabilidade hipotética representando as propriedades de uma hipótese alternativa verdadeira. No entanto, quando você está realizando um teste de hipótese, você normalmente não sabe qual hipótese é verdadeira, muito menos as propriedades específicas da distribuição para a hipótese alternativa. Consequentemente, a taxa real de erro do Tipo II é geralmente desconhecida!

Como sabem, beta é a probabilidade de um falso negativo. Portanto,  $1 - \beta$  é a probabilidade de detectar corretamente um efeito. Os estatísticos referem-se a este conceito como **poder estatístico**. Os analistas normalmente estimam o poder em vez do beta diretamente.

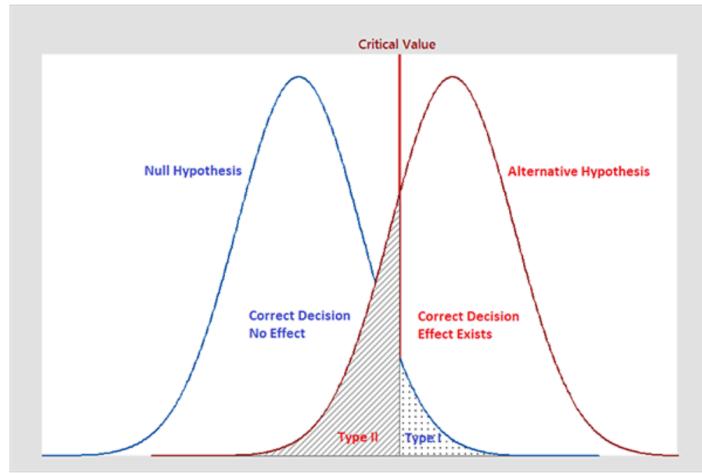
Ao projetar seu experimento, você pode inserir estimativas dos três fatores que afetam  $\beta$  (variabilidade, efeito e tamanho da amostra) no software estatístico (Python, Sheets ou algum software específico) e ele calcula o poder estimado para seu teste.

Suponha que você faça uma análise de poder para um próximo estudo e calcule um poder estimado de 90%. Para este estudo, a taxa de erro do Tipo II estimada é de 10% ( $1 - 0,9$ ). Tenha em mente que a variabilidade e o tamanho do efeito são estimativas e suposições - a única coisa que você sabe de fato é o tamanho da amostra. Consequentemente, a potência e a taxa de erro do Tipo II são apenas estimativas, e não algo que você define diretamente. Essas estimativas são tão boas quanto as entradas em sua análise de poder.

Baixa variabilidade e tamanhos de efeito maiores diminuem a taxa de erro do Tipo II, o que aumenta o poder estatístico. No entanto, os analistas pesquisadores geralmente têm menos controle sobre esses aspectos de um teste de hipótese.

Normalmente, temos mais controle sobre o tamanho da amostra, tornando-se a maneira crítica de gerenciar sua taxa de erro Tipo II. Mantendo tudo o mais constante, aumentar o tamanho da amostra reduz a taxa de erro do Tipo II e aumenta o poder.

O gráfico abaixo ilustra os dois tipos de erros usando duas distribuições de amostragem. A linha da região crítica representa o ponto em que você rejeita ou deixa de rejeitar a hipótese nula. É claro que, ao realizar o teste de hipóteses, você não sabe qual hipótese está correta. E as propriedades da distribuição para a hipótese alternativa são geralmente desconhecidas. No entanto, vamos usar este gráfico para entender a natureza geral desses erros e como eles estão relacionados.



A distribuição à esquerda representa a hipótese nula. Quando a hipótese nula está correta, você só precisa se preocupar com os erros do Tipo I, que é a parte sombreada da distribuição da hipótese nula. O restante da distribuição nula mostra a decisão correta de não rejeitar o nulo.

Por outro lado, quando a hipótese alternativa está correta, você precisa se preocupar com os erros do Tipo II. A região sombreada na distribuição de hipóteses alternativas representa a taxa de erro Tipo II. O restante da distribuição alternativa descreve a probabilidade de detectar corretamente um efeito – que é o poder estatístico.

Mover a linha de valor crítico é equivalente a **alterar o nível de significância**. Se você mover a linha para a esquerda, estará aumentando o nível de significância (por exemplo, α 0,05 a 0,10). Mantendo todo o resto constante, esse ajuste **aumenta a taxa de erro do Tipo I** enquanto **reduz a taxa de erro do Tipo II**. Mover a linha para a direita reduz o nível de significância (por exemplo, α 0,05 a 0,01), o que **diminui a taxa de erro tipo I**, mas **aumenta a taxa de erro tipo II**.

## UM ERRO É PIOR QUE O OUTRO?

Como você viu, a natureza dos dois tipos de erro, suas causas e a certeza de suas taxas de ocorrência são muito diferentes.

Uma dúvida comum é se um tipo de erro é pior que o outro? Os estatísticos projetaram testes de hipóteses para controlar os erros do Tipo I, enquanto os

erros do Tipo II são muito menos definidos. Consequentemente, muitos estatísticos afirmam que é melhor não detectar um efeito quando ele existe do que concluir que existe um efeito quando não existe. Em outras palavras, há uma tendência a assumir que os erros do Tipo I são piores.

No entanto, a realidade é mais complicada do que isso. Você deve considerar cuidadosamente as consequências de cada tipo de erro para seu teste específico.

Suponha que você esteja avaliando a detecção de um novo teste de câncer. A vida das pessoas depende dessa detecção. Um falso negativo neste cenário significa que não daremos chance da pessoa se tratar, uma vez que não detectamos a doença quando ela está presente.

Agora suponhamos que temos a opção de dar um pré-tratamento para alguém que ainda não desenvolveu nenhuma doença, e nosso teste tem como objetivo entender se essa pessoa desenvolveria ou não essa doença no futuro. Esse pré-tratamento é altamente invasivo e caro, além de poder colocar a vida da pessoa em risco caso ela não tenha predisposição a essa doença. Nesse caso, você deseja que o teste seja bastante assertivo com relação aos positivos, mesmo que isso signifique dar um falso negativo a alguém.

Dessa forma, a depender do seu caso você vai preferir penalizar um erro ou outro.



# 10. Intervalo de confiança para médias



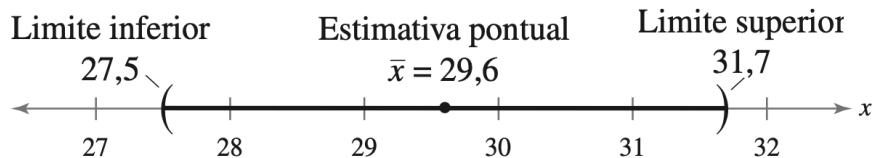
021.399.242-66

Como dissemos anteriormente, o intervalo de confiança sempre vai concordar com um teste de hipótese, então nada mais justo do que começar com ele.

Suponha que por exemplo você queira estimar a média da altura de mulheres brasileiras. Nesse caso, você coleta uma amostra sem vieses sistemáticos e encontra uma altura média de 1,63 m. Essa estimativa é chamada de **estimativa pontual** (ou por ponto). Obviamente, quanto maior o tamanho da sua amostra, mais próximo você de fato se aproxima da média real da altura de sua população. Entretanto, a verdade é que tentar representar toda uma população por apenas 1 número é bastante complicado e carrega muitos erros de estimação (afinal, qual é a chance de selecionar uma mulher aleatória do Brasil e ela ter exatamente 1,63m?).

Por conta disso, surge a ideia de construirmos um intervalo em torno da estimativa por ponto, de modo que a esse intervalo tenha uma probabilidade conhecida de conter o verdadeiro valor do parâmetro. Essa é a ideia da **estimativa intervalar**. Para formar uma estimativa intervalar, usamos a

**estimativa pontual** como o **centro** do intervalo e depois adicionamos e subtraímos uma margem de erro.



Chamamos de **intervalo de confiança** o intervalo que, com probabilidade conhecida, deverá conter o valor real do parâmetro. No caso da figura acima, o intervalo é [27,5, 31,7].

A **margem de erro** é o valor que somamos e subtraímos da estimativa pontual para obter o intervalo - no caso da figura acima, a margem de erro é 2,1 (somando e subtraindo 2,1 de 29,6 obtemos 31,7 e 27,5 respectivamente).

Antes de continuarmos falando como medir o intervalo de confiança, é importante frisar aqui que os intervalos que vamos calcular são **simétricos** em relação ao seu parâmetro e que, portanto, pressupõem uma **distribuição normal**. Por exemplo, voltando novamente ao exemplo acima, somando 2,1 a 29,6 obtemos 31,7 e subtraindo 2,1 de 29,6 obtemos 27,5. A margem de erro é a mesma "para mais" e "para menos", tornando o intervalo simétrico em relação à estimativa pontual.

Diferentes amostras aleatórias retiradas da mesma população podem produzir intervalos ligeiramente diferentes. Se você extrair muitas amostras aleatórias e calcular um intervalo de confiança para cada amostra, uma proporção específica dos intervalos conterá o parâmetro populacional. Essa porcentagem é o **nível de confiança (c)**.

Se considerarmos  $c = 100\%$ , podemos ter certeza que o verdadeiro valor do parâmetro estará nesse intervalo. Por exemplo, teríamos certeza que, se escolhessemos aleatoriamente uma mulher brasileira, com certeza a altura dela estaria nesse intervalo com  $c = 100\%$ . O problema é que esse intervalo seria infinito! É quase impossível garantirmos que temos em um intervalo uma população inteirinha, a menos que esse intervalo seja gigante - ou seja, que cubra todos os números de altura possíveis para um ser humano. Pois



bem, isso não é muito útil para nós, certo? Ter um intervalo muito grande (infinito) não é uma aproximação, é praticamente uma constatação - afinal, é óbvio que a altura de um ser humano está entre 0 a infinito e nem precisaríamos de uma amostra para dizer isso.

Por outro lado, se considerarmos um  $c$  muito pequeno (por exemplo,  $c = 5\%$ ), o seu intervalo vai ser bem menor, porém existe uma confiança muito pequena (5% nesse caso) do seu parâmetro não estar no seu intervalo. Suponhamos que o intervalo a 5% de confiança de altura das mulheres é 1,62 a 1,64. Essa altura não representa bem a maioria das mulheres no Brasil, certo? Apesar de termos um intervalo bem pequeno, ele não representa bem nossa população.

Grande parte dos estudos existentes considera  $c = 95\%$ , podendo variar de 90 a 99% a depender da precisão necessária para seu estudo.

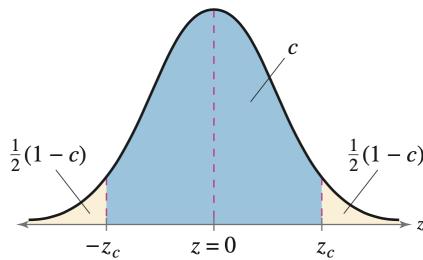
Como dissemos, para podermos aplicar o intervalo de confiança com as fórmulas a seguir, precisamos que a distribuição seja normal. Entretanto, você sabe do teorema do limite central que quando  $n \geq 30$ , a **média amostral** terá distribuição aproximadamente normal. Essa premissa é de suma importância quando queremos calcular um intervalo de confiança, pois tais intervalos são simétricos (mesma margem de erro para menos e para mais). Apenas podemos assumir intervalos de confiança simétricos se a própria distribuição é simétrica - ou seja, é uma normal.

**NOTA IMPORTANTE:** QUANDO FALAMOS DE TESTE DE HIPÓTESE E INTERVALO DE CONFIANÇA, ESTAMOS  
CONSIDERANDO QUE AS AMOSTRAS FORAM COLETADAS DE FORMA CORRETA E QUE ELAS REPRESENTAM  
BEM NOSSA POPULAÇÃO E NÃO HÁ VIESSES SISTEMÁTICOS! CASO VOCÊ DESCONFIE QUE HÁ, VOLTE ALGUMAS  
CASAS E FAÇA UMA REAMOSTRAGEM CONSIDERANDO TODOS OS CENÁRIOS POSSÍVEIS.



## INTERVALO DE CONFIANÇA PARA MÉDIAS QUANDO O DESVIO-PADRÃO POPULACIONAL É CONHECIDO

O **nível de confiança** corresponde à área sob a curva normal padrão entre os valores críticos. Aqui vamos chamar os **valores críticos** de  $-z_c$  e  $z_c$ .



Esses valores críticos, em geral, separam resultados prováveis (região central) de improváveis, ou incomuns (caudas).

Lembrando que a área total embaixo da curva é 1, e que **c é a área em azul**, então a área restante (**área bege**) é  $1 - c$ .

Sabendo que a curva é simétrica, a área em cada cauda (cada uma das áreas beges) é  $(1 - c)/2$ .

Por exemplo, se  $c = 95\%$ , então  $2.5\%$  da área está à esquerda e  $2.5\%$  está à direita. Dessa forma, **zc** deve ser a probabilidade para um determinado  $z$  em que a área seja azul +  $\frac{1}{2}(1-c)$ , ou seja,  $95\% + 2.5\% = 97.5\%$ . Logo, quando trabalhamos com  $95\%$  de confiança para encontrar um intervalo de confiança, a área acumulada até  $zc$  é  $97.5\%$ .

O  $z$  em questão aqui é o próprio  $z$  que corresponde a uma normal padrão, que vimos no capítulo de probabilidades e distribuição, o qual é tabelado. Quando o desvio-padrão da **população** é conhecido podemos usá-lo para representar os valores improváveis (a área bege do gráfico acima).

Para facilitar, abaixo segue o valor de  $z$  para os **intervalos de confiança** mais usuais:

Percentage Confidence	z*-Value
80	1.28
90	1.645
95	1.96
98	2.33
99	2.58

Quando trabalhamos com 95% de confiança em um teste bicaudal, verificamos que  $P(z < z_c) = 0.9750$  para um  $z = 1.96$ . Dessa forma, falamos que  $z_c = 1.96$ .

Você também pode encontrar os valores de  $z$  para cada confiança nesse site: <https://www.omnicalculator.com/statistics/critical-value>

No site acima, escolha:

---

What distribution?

Z (standard normal) ▾

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

What type of test?

Two-tailed ▾

O significance level (nível de significância) deve ser  $1-\alpha$ . Para 95%, deverá ser 0.05.

*Disclaimer: Quando escolhemos o parâmetro "two-tailed" ele já distribui os 5% em ambas as caudas e, por isso, precisamos apenas passar 5% ao invés de 2,5%.*

Dado um nível de confiança  $\alpha$ , a **margem de erro E** (às vezes chamada também de erro máximo da estimativa ou tolerância de erro) é a maior distância possível entre a estimativa pontual e o valor do parâmetro que ela está estimando. Sua fórmula é dada por:



$$E = z_c \frac{\sigma}{\sqrt{n}}$$

Usando uma estimativa pontual e uma margem de erro, você pode construir uma estimativa intervalar de um parâmetro populacional. Essa estimativa intervalar é chamada de **intervalo de confiança** e é representada por

$$\text{Confidence Interval} = \bar{X} \pm Z \times \frac{\sigma}{\sqrt{n}}$$

Em que  $CI$  é o intervalo de confiança,  $\bar{x}$  é a média amostral,  $\sigma$  é o desvio-padrão populacional e  $n$  é o tamanho da amostra.

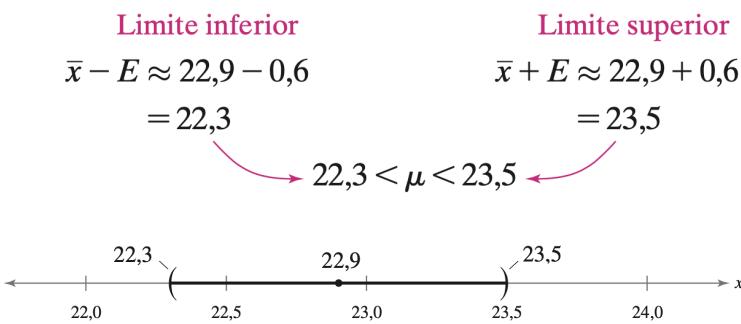
**Exemplo:** O diretor de admissões de uma faculdade deseja estimar a idade média de todos os estudantes atualmente matriculados. Em uma amostra aleatória de 20 estudantes, a idade média encontrada é de 22,9 anos e a distribuição é normal. De estudos anteriores, o desvio padrão populacional conhecido é de 1,5 ano, e a população é normalmente distribuída. Construa um intervalo de confiança de 90% da idade média da população.

**Resposta:** Primeiro, nossa distribuição é normal. Segundo, o desvio-padrão **populacional** é conhecido. Terceiro, a amostra é aleatória. Logo, podemos usar o intervalo de confiança com  $z$ .

Usando  $n = 20$ ,  $\bar{x} = 22,9$ ,  $s = 1,5$  e  $zc = 1,645$  (90% de confiança), a margem de erro no intervalo de confiança de 90% é:

$$E = z_c \frac{\sigma}{\sqrt{n}} = 1,645 \cdot \frac{1,5}{\sqrt{20}} \approx 0,6.$$





Ou seja, se coletarmos amostras de 100 estudantes, 90 intervalos de confiança conterão a média populacional real. Como nosso caso temos apenas 1 amostra, podemos dizer que estamos 90% confiantes que nosso intervalo contém a média real. Em outras palavras, com 90% de confiança, você pode dizer que a idade média de todos os estudantes está entre 22,3 e 23,5 anos.

## DISTRIBUIÇÃO T-STUDENT

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242-66

Em muitas situações da vida real, o **desvio padrão da população é desconhecido**. Então, como podemos construir um intervalo de confiança para uma média populacional? Para uma variável aleatória que é normalmente distribuída (ou aproximadamente normalmente distribuída), a variável média amostral comporta-se tal qual outro modelo, a **distribuição t**, também chamado de **t-student**.

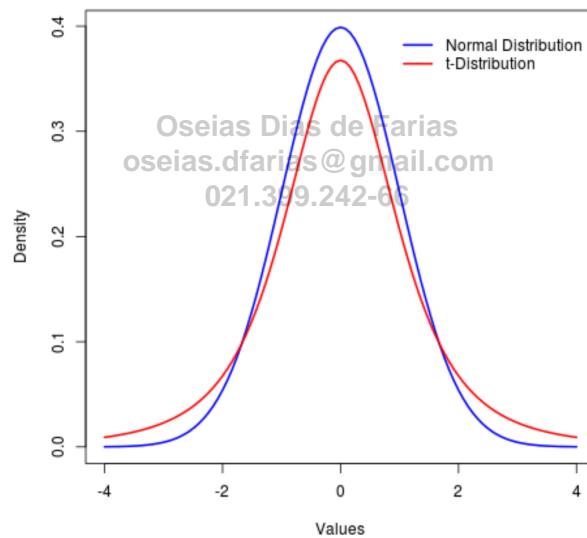
A distribuição t é uma **família de curvas**, cada uma determinada por um parâmetro chamado de **graus de liberdade**. Os graus de liberdade são o número de escolhas livres deixadas depois que uma estatística amostral tal como  $x$  é calculada. Quando usamos a distribuição t para estimar uma média populacional, os graus de liberdade são iguais ao tamanho da amostra menos um ( $n-1$ ).

A distribuição t distribui-se também simetricamente com média 0, porém não é uma normal. O valor de t é dado por:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Em que  $s$  é o desvio-padrão **amostral** (note que não é o populacional!). Quando nossa amostra é muito grande,  $s$  tende a se aproximar do desvio-padrão populacional e, portanto,  $t$  passa a assumir exatamente a mesma fórmula de  $z$  - nesse caso, a distribuição passa a ser normal.

A distribuição  $t$  ao primeiro olhar realmente se parece com a distribuição normal, mas não é (a não ser para uma quantidade de amostras grandes, quando  $s$  se aproxima de  $\sigma$ ). Em geral,  $t$  tem causas mais grossas do que a normal.



Da mesma forma que o  $z$ , os valores de  $t$  também são tabelados. A partir do grau de liberdade e de um nível de confiança desejável, conseguimos verificar o  $t$ -tabelado a partir da tabela- $t$ .

**Exemplo:** Encontre o valor crítico  $t_c$  para um nível de confiança de 95% quando o tamanho da amostra é 15.

*Resposta:*

Como  $n = 15$ , os graus de liberdade são:

$$\text{Grau liberdade} = n - 1 = 15 - 1 = 14.$$

Usando GL. = 14 e c = 0,95, você pode encontrar o valor crítico  $t_c$ , como mostrado pelas áreas destacadas na tabela

	Nível de confiança, c	0,80	0,90	0,95	0,98	0,99
g.l.	Unilateral, $\alpha$	0,10	0,05	0,025	0,01	0,005
	Bilateral, $\alpha$	0,20	0,10	0,05	0,02	0,01
1		3,078	6,314	12,706	31,821	63,657
2		1,886	2,920	4,303	6,965	9,925
3		1,638	2,353	3,182	4,541	5,841
12		1,356	1,782	2,179	2,681	3,055
13		1,350	1,771	2,160	2,650	3,012
14		1,345	1,761	2,145	2,624	2,977
15		1,341	1,753	2,131	2,602	2,947
16		1,337	1,746	2,120	2,583	2,921

Da tabela, você pode ver que  $t_c = 2,145$ .

Porém, ao longo do curso não usaremos a tabela! Vamos usar Sheets e Python para coletar esses valores quando necessário.

**Oseias Dias de Farias**  
**INTERVALO DE CONFIANÇA PARA MÉDIAS QUANDO O DESVIO-PADRÃO POPULACIONAL  
NÃO É CONHECIDO**

Construir um intervalo de confiança quando o desvio-padrão populacional **não** é conhecido usando a distribuição t é similar a construir um intervalo de confiança quando o desvio-padrão é conhecido usando a distribuição normal — mudando apenas o parâmetro z para t. A fórmula abaixo é usada para encontrar a margem de erro

$$E = t_c \frac{s}{\sqrt{n}}$$

**Exemplo:** Você seleciona aleatoriamente 16 cafeterias e mede a temperatura do café vendido em cada uma delas. A temperatura média da amostra é 162,0 oF (fahrenheit) com desvio padrão de 10,0 oF. Construa um intervalo de confiança de 95% para a temperatura média da população de cafés vendidos.



Suponha que as temperaturas tenham distribuição aproximadamente normal.

Resposta: Como o desvio-padrão populacional é desconhecido, a amostra é aleatória e as temperaturas têm distribuição aproximadamente normal, use a distribuição t. Sendo  $n = 16$ ,  $x = 162,0$ ,  $s = 10,0$ ,  $c = 0,95$  e g.l. = 15, você pode os valores de t nesse site: <https://www.omnicalculator.com/statistics/critical-value>. Escolhemos os seguintes parâmetros:

What distribution?	<u>t-Student</u> ▾
What type of test?	<u>Two-tailed</u> ▾

Em "degrees of freedom" devemos colocar nossos graus de liberdade (em nosso exercício, será 15) e em significance level devemos colocar 1-c, nesse caso será 0.05 (uma vez que c é 95%).

Oseias Dias de Farias  
Logo, t será aproximadamente 2,131  
[oseias.diasdefarias@gmail.com](mailto:oseias.diasdefarias@gmail.com)  
021.399.242-66

What distribution?	<u>t-Student</u> ▾
What type of test?	<u>Two-tailed</u> ▾
Degrees of freedom (d)	15
Significance level	0.05

The test statistic follows the t-distribution with 15 degrees of freedom.

Critical value:  $\pm 2.1314$

Critical region:

$(-\infty, -2.1314] \cup [2.1314, \infty)$

Voltando a fórmula, temos:



$$E = t_c \frac{s}{\sqrt{n}} = 2,131 \cdot \frac{10,0}{\sqrt{16}} \approx 5,3.$$

Com isso, conseguimos calcular nosso intervalo de confiança:

<b>Limite inferior</b> $\bar{x} - E \approx 162 - 5,3 = 156,7$	<b>Limite superior</b> $\bar{x} + E \approx 162 + 5,3 = 167,3$
$156,7 < \mu < 167,3$	

Com 95% de confiança, você pode dizer que a temperatura média da população de cafés vendidos está entre 156,7 oF e 167,3 oF.

## O INTERVALO DE CONFIANÇA CONCORDA MESMO COM O TESTE DE HIPÓTESE?

oseias Das De Farias  
oseias.dfarias@gmail.com  
021 399 242-66

Vocês lembram que comentamos que o intervalo de confiança sempre concorda com o teste de hipótese? Pois bem, vamos mostrar isso com um exemplo!

Voltando ao nosso exemplo de média de gastos de combustível, temos:

Variável	Valor
$\bar{x}$	330.6
s	154.2
n	25

Usando o 95% de confiança, para g.l. = 25 - 1 = 24 o valor de t é 2,0639



What distribution?	<u>t-Student</u> ▾
What type of test?	<u>Two-tailed</u> ▾
Degrees of freedom (d)	24
Significance level	0.05
The test statistic follows the t-distribution with 24 degrees of freedom.	
Critical value: ±2.0639	

Com esse valor de t, temos:

$$E = t * \frac{s}{\sqrt{n}} = 2,0639 * \frac{154,2}{\sqrt{25}} = 63,6$$

Logo, o intervalo de confiança é:

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242-66

Limite inferior = 330,6 - 63,6 = 267

Limite superior = 330,6 + 63,6 = 394,2

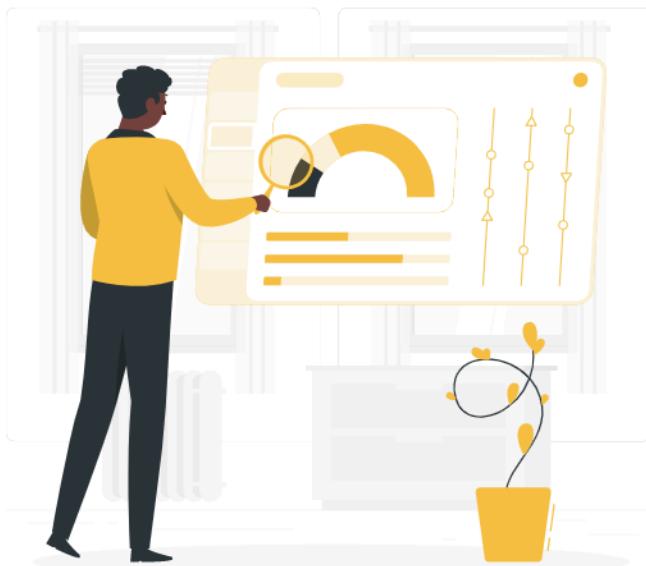
Intervalo = [267 394,2]

Nossa ideia inicial era tentar ver, a 95%, se a hipótese nula estava correta, ou seja, se a média de gastos era de 260. Nos capítulos anteriores fizemos um teste de hipótese (que não mostramos os cálculos, mas obtivemos um p-valor) que rejeitava a hipótese nula a 95% de certeza.

Aqui podemos ver que o limite inferior ainda é superior a 260 a 95% de confiança. Como 260 não está dentro desses limites, dizemos que de fato a média dessa amostra não é 260. Porém, já deixo um spoiler, muito cuidado ao interpretar o intervalo de confiança! Vocês verão mais sobre isso no capítulo "Usos e abusos do intervalo de confiança"



# 11. Teste de hipótese para médias

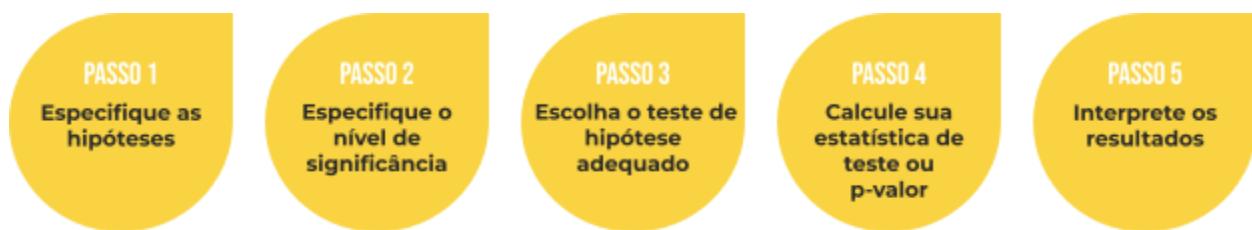


Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

031 3000 0200

E finalmente chegamos nele, o tão aguardado teste de hipótese! Se vocês bem se lembram, usamos o teste de hipótese quando temos apenas amostras e nosso intuito é fazer afirmações sobre toda uma população. Devido aos erros aleatórios intrínsecos que acontecem quando usamos uma amostra, não podemos fazer afirmações diretamente em uma amostra sem um teste de hipótese adequado (ou o intervalo de confiança!).

Para realizar um teste de hipótese seguimos os seguintes passos:



Até agora falamos sobre os passos 1 e 2. Depois de construir as hipóteses nula e alternativa e especificar o nível de significância, o próximo passo em um teste de hipótese é obter uma **amostra aleatória** da população e calcular as estatísticas amostrais de interesse para aquele teste (tais como *média*,

proporção, desvio-padrão, etc), correspondentes aos parâmetros na hipótese nula. Vamos a alguns casos de comparações que podemos usar um teste de hipótese:

### 1) Médias

- a) Queremos comprovar se nossa média de gastos de combustível é US\$ 260. Não conseguimos coletar a população toda, por isso coletamos uma amostra e a média de gastos foi US\$ 330,6. Será que podemos afirmar que a média populacional não é US\$ 260? Precisamos de um teste de hipótese (ou intervalo de confiança) para comparar e ver se devemos rejeitar a hipótese nula ou não.
- b) Queremos testar 2 métodos de ensino e, para isso, selecionamos aleatoriamente 2 grupos de pessoas. No grupo 1 aplicamos o método A e no grupo 2 aplicamos o método B. Depois, esses 2 grupos fazem uma prova e calculamos a média da nota de cada um dos grupos. Nesse cálculo, vimos que a média do grupo 1 é maior do que a do grupo 2. Será que podemos afirmar que o método A é melhor que o B? Para esses grupos, com certeza podemos! Mas será que isso é aplicável a toda uma população de interesse? Precisamos de um teste de hipótese ou comparar os intervalos de confiança para tomarmos essa decisão.

### 2) Proporções

- a) Temos 2 grupos de pessoas escolhidas aleatoriamente e queremos testar uma vacina. O grupo A toma a vacina e o grupo B não toma. Medimos então a proporção de infectados no grupo A e no grupo B e vimos que a proporção é menor no grupo que tomou a vacina (grupo A). Será que podemos estender esse resultado à população? Precisamos de um teste de hipótese ou comparar os intervalos de confiança para tomarmos essa decisão.
- b) Fizemos uma melhoria em um site de e-commerce e agora queremos comprovar que a proporção de pessoas que efetivam a compra agora é de 20%. Podemos coletar uma amostra, medir a proporção de efetivação de compra e rodar um teste de hipótese ou comparar os intervalos de confiança para vermos se de fato essa efetivação é de 20% para a população com base



apenas na amostra.

### 3) Desvio-padrão

- a) Temos um processo industrial com alta variabilidade, que faz com que os produtos não sigam um padrão. Fizemos uma melhoria para reduzir a variabilidade, testamos uma amostra e agora precisamos comprovar se essa redução se estende a população.

Notem que em todos os exemplos podemos comparar uma amostra a um valor fixo ou 2 amostras. Mais para frente vocês também verão que também poderemos comparar mais que duas amostras.

Independentemente se queremos comparar médias, proporções ou desvios, temos uma denominação chamada **teste de hipótese 1 amostra** (hypothesis testing 1 sample) quando temos uma 1 amostra e queremos compará-la a um valor fixo. Também podemos comparar 2 amostras e, nesse caso, temos o que chamamos de **teste de hipótese 2 amostra** (hypothesis testing 2 samples/)

A estatística amostral de interesse é chamada de **estatística de teste**. Sob a suposição de que a hipótese nula é verdadeira, o valor específico da variável de teste (a média, por exemplo) é então transformada em uma estatística de teste, tal como z, t, f ou  $\chi^2$ . A estatística de teste padronizada é usada na tomada de decisão sobre a rejeição ou não da hipótese nula.

Nessa seção vamos falar sobre a **média** especificamente (1 sample e 2 samples), mas nas próximas sessões falaremos sobre comparações de proporção e desvio-padrão. Para a média, nossas estatísticas de teste são a **z** e a **t**, a depender do caso.

**NOTA IMPORTANTE:** QUANDO FALAMOS DE TESTE DE HIPÓTESE E INTERVALO DE CONFIANÇA, ESTAMOS CONSIDERANDO QUE AS AMOSTRAS FORAM COLETADAS DE FORMA CORRETA E QUE ELAS REPRESENTAM BEM NOSSA POPULAÇÃO E NÃO HÁ VIESSES SISTEMÁTICOS! CASO VOCÊ DESCONFIE QUE HÁ, VOLTE ALGUMAS CASAS E FAÇA UMA REAMOSTRAGEM CONSIDERANDO TODOS OS CENÁRIOS POSSÍVEIS.



# 1 SAMPLE

## DESVIO-PADRÃO POPULACIONAL CONHECIDO

Quando temos um desvio-padrão populacional conhecido, padronizamos nossa média com o **teste z**.

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Podemos calcular a área embaixo da curva para valores mais extremos de z (lembrem-se que os valores de z são tabelados). Essa área para valores mais extremos corresponde ao nosso p-valor.

Vamos direto a um exemplo para deixar isso mais claro.

Oseias Dias de Farias

[oseias.diasfarias@gmail.com](mailto:oseias.diasfarias@gmail.com)

021.399.242-66

**Exemplo:** Em corrida de carros, o pit stop é o local em que um veículo vai para trocar pneus, abastecer, efetuar reparos e outros ajustes mecânicos. A eficiência de uma equipe que realiza esses ajustes pode afetar o resultado de uma corrida. Uma equipe afirma que seu tempo médio no pit stop (para 4 trocas de pneus e abastecimento) é **menor que 13 segundos**. Uma amostra aleatória de **32 tempos** de pit stop tem uma **média de 12,9 segundos**. Suponha que o **desvio padrão populacional é de 0,19 segundos**. Há evidência suficiente para concordar com a afirmação para 99% de confiança?

Resposta:

Para 99% de confiança,  $\alpha = 1 - 0.99 = 0.01$

Como o desvio-padrão **populacional** é conhecido (desvio = 0,19), a amostra é aleatória e  $n = 32 \geq 30$  (garantindo a normalidade pelo teorema do limite central), você pode usar o **teste z**. A afirmação é “o tempo médio no pit stop é menor que 13 segundos”. Então, as hipóteses nula e alternativa são:

$H_0$ : média  $\geq 13$  segundos



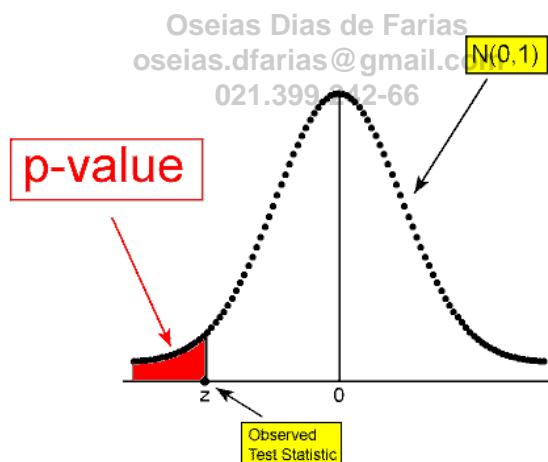
$H_a$ : média < 13 segundos

Notem que aqui estamos falando de um **teste unicaudal à esquerda!**

Logo, nosso **z calculado** é:

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \\ &= \frac{12,9 - 13}{0,19 / \sqrt{32}} \\ &\approx -2,98. \end{aligned}$$

Vamos relembrar rapidamente o que é o p-valor. Para isso, vamos usar a figura abaixo para exemplificar.



A curva acima representa o cenário em que  $H_0$  seria verdadeiro (lembrem-se, esse cenário é fictício pois ainda veremos se vamos rejeitar ou não  $H_0$ ). O **z calculado é o "observed test statistic"** e o p-valor seria a área vermelha. Dessa forma, dizemos que o p-valor é a probabilidade (área embaixo da curva) de obtermos o valor amostral que tivemos (no caso, 13 segundos) ou mais extremo (no caso, < 13 segundos) **SE** a hipótese nula estiver correta. Quando essa probabilidade é muito pequena, nós rejeitamos  $H_0$ .

E o que é "muito pequeno"? É aí que entra o nosso  $\alpha$ , que podemos entender como uma probabilidade limite para não rejeitamos o  $H_0$ . Se o p-valor é menor do que  $\alpha$ , a probabilidade de ocorrer um valor igual ou mais extremo quando  $H_0$  é verdadeiro é TÃO pequena que não podemos não rejeitar  $H_0$  e, portanto, dizemos que  $H_0$  é mentira (em palavras mais bonitas, rejeitamos  $H_0$ ).

Dito isso, vocês devem estar pensando "ok, mas eu só tenho o z (observed test statistic) e não o p-valor (área embaixo da curva). E agora?". Agora a estatística te dá 2 opções:

- 1) Você pode encontrar o z que corresponde a  $\alpha$  (esse z é chamado de **z crítico**) e compará-lo com z calculado. Se z calculado for menor que o z crítico, dizemos que a probabilidade mínima para não rejeitamos  $H_0$  ( $\alpha$ ) não foi atendida, e, portanto, rejeitamos o  $H_0$ .
- 2) Você pode encontrar a probabilidade (área embaixo da curva) que corresponde o z calculado. Essa área é justamente o seu **p-valor!** Com o p-valor, você pode compará-lo com  $\alpha$  e seguir o mesmo raciocínio do item anterior.

Oscar Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Vamos fazer ambas as formas para demonstrar.

- 1) Nossa z crítico é o z que corresponde a  $\alpha = 1\%$ , unicaudal à esquerda (left-tail). Usando o site que já conhecemos <https://www.omnicalculator.com/statistics/critical-value> temos:

What distribution? Z (standard normal) ▾

What type of test? Left-tailed ▾

Significance level 0.01

The test statistic follows the standard normal distribution  $N(0,1)$ .

Critical value: -2.3263

Critical region:  $(-\infty, -2.3263]$

*To increase the precision with which the critical values are calculated, click the advanced mode.*

$Z$  crítico = -2,3263. Portanto,  $z$  crítico é maior do que  $z$  calculado (-2,98) e, portanto, o  $z$  calculado cai na zona de rejeição. Ou seja, há evidência suficiente ao **nível de significância de 1%** para **concordar** com a afirmação de que o tempo médio no pit stop é menor que 13 segundos (Ha).

2) Aqui vamos usar um novo site:

<https://www.calculator.net/z-score-calculator.html>

Apontando  $z$  = -2,98, podemos usar o item "Z-score and Probability Converter" para calcular a área embaixo da curva. Nesse caso, basta substituir  $z$  por -2,98 e clicar em "calculate"

Z-score, <b>Z</b>	-2.98
Probability, <b>P(x&lt;Z)</b>	
Probability, <b>P(x&gt;Z)</b>	
Probability, <b>P(0 to Z or Z to 0)</b>	
Probability, <b>P(-Z&lt;x&lt;Z)</b>	Oseias Dias de Farias oseias.dfarias@gmail.com
Probability, <b>P(x&lt;-Z or x&gt;Z)</b>	021.399.242-66
<input style="background-color: #2e6b2e; color: white; padding: 5px 10px; border-radius: 5px; border: none; font-weight: bold; margin-right: 10px;" type="button" value="Calculate"/> <input style="background-color: #ccc; border: none; border-radius: 5px; padding: 5px 10px;" type="button" value="Clear"/>	

O resultado será

Given  $Z$  = -2.98,

$$P(x < Z) = 0.0014412$$



$$P(x > Z) = 0.99856$$



$$P(Z < x < 0) = 0.49856$$



$$P(-Z < x < Z) = 0.99712$$



$$P(x < -Z \text{ or } x > Z) = 0.0028825$$

A primeira figura é exatamente o que estamos procurando: valores mais extremos que  $z$  e apenas menores do que  $z$ .

Nesse caso, temos que a área embaixo da curva é  $0,0014412 = 0,14412\%$ . Esse é o nosso p-valor.

Uma vez que o p-valor é **menor** que 1%, você **rejeita** a hipótese nula. Ou seja, há evidência suficiente ao **nível de significância de 1%** para **concordar** com a afirmação de que o tempo médio no pit stop é menor que 13 segundos (Ha).

## DESVIO-PADRÃO POPULACIONAL DESCONHECIDO

Da mesma forma que acontece para o intervalo de confiança, é muito mais comum os casos de não termos o desvio-padrão populacional conhecido e termos apenas o desvio-padrão amostral. Semelhante ao intervalo de confiança, nesses casos também usamos a distribuição t-student para nossos cálculos. Nesse caso, temos que calcular o t por:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Em que s representa o desvio-padrão amostral.

Vamos a um exemplo

**Exemplo:** Uma indústria afirma que o nível médio do pH da água em um rio próximo é de 6,8. Você seleciona aleatoriamente 39 amostras de água e mede o pH de cada uma. A média amostral e o desvio padrão são de 6,7 e 0,35, respectivamente. Há evidência suficiente para rejeitar a afirmação da indústria considerando nível de significância  $\alpha = 0,05$ ?

Resposta:

Notem que aqui só temos o desvio padrão amostral! Também sabemos que  $n > 30$  e estamos querendo fazer os cálculos de médias amostrais, então podemos usar o teorema do limite central que garante normalidade na distribuição de médias amostrais. Portanto, podemos usar o teste t.

$H_0$  : média = 6,8 (Afirmação)

$H_a$  : média  $\neq$  6,8

O teste é bilateral (two-tail), o nível de significância é 0,05 e os graus de liberdade são g.l. =  $39 - 1 = 38$ .

Da mesma forma que antes, temos 2 opções de cálculo

- 1) Você pode encontrar o t que corresponde a  $\alpha$  (esse t é chamado de **t crítico**) e compará-lo com t calculado. Se t calculado for menor que o t crítico, dizemos que a probabilidade mínima para não rejeitamos H<sub>0</sub> ( $\alpha$ ) não foi atendida, e, portanto, rejeitamos o H<sub>0</sub>.
- 2) Você pode encontrar a probabilidade (área embaixo da curva) que corresponde o t calculado. Essa área é justamente o seu **p-valor!** Com o p-valor, você pode compará-lo com  $\alpha$  e seguir o mesmo raciocínio do item anterior.

Agora, vamos usar apenas a abordagem do p-valor.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$
$$= \frac{6,7 - 6,8}{0,35 / \sqrt{39}}$$

$$\approx -1,784.$$

Como  $\sigma$  é desconhecido e  $n \geq 30$ , use o teste t.

Useias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Suponha que  $\mu = 6,8$ .

Arredonde para três casas decimais.

Agora vamos usar o seguinte site para ver qual seria o p-valor correspondente a esse t: <https://www.socscistatistics.com/pvalues/tdistribution.aspx>



## P Value from T Score Calculator

This should be self-explanatory, but just in case it's not: your *t*-score, degrees of freedom in the *DF* box (*N* - 1 for single sample and dependent samples), select your significance level and whether you're testing a one-tailed or two-tailed hypothesis (if you're not sure, go with the defaults), then press the button.

If you need to derive a T Score from raw data, [then you can find t tests here](#).

[Report a T-Test Result \(APA\)](#)

T Score:

DF:

Significance Level:

.01

.05

.10

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

One-tailed or two-tailed hypothesis?:

One-tailed

Two-tailed

The *p*-value is .082206.

The result is *not* significant at *p* < .05.

[Calculate](#)

Não esqueçam que o teste é bicaudal (two-tailed).

Dessa forma, como *p*-valor = 0,0822 dizemos que não devemos rejeitar a hipótese nula. Ou seja, não podemos afirmar que a média da população não é 6,8.

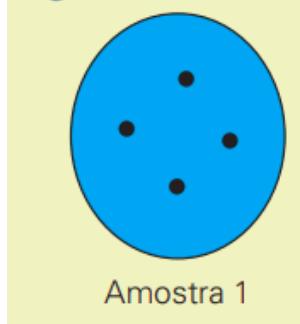


## 2 SAMPLE

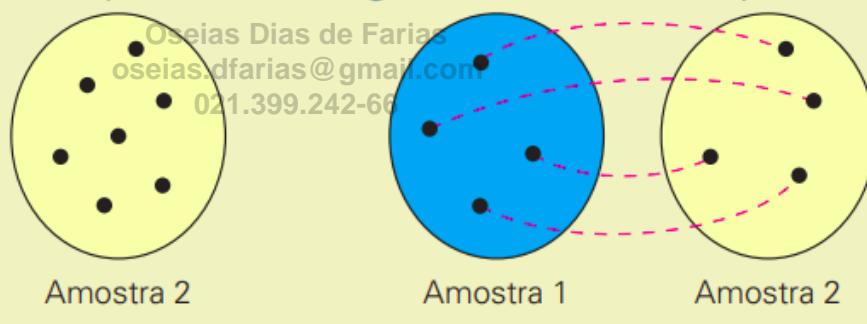
Na seção anterior você estudou métodos para testar uma afirmação sobre o valor de um parâmetro populacional. Agora você aprenderá como testar uma afirmação comparando 2 médias de duas populações. Antes de aprender como testar a diferença entre dois parâmetros, você precisa entender a diferença entre amostras independentes e **amostras dependentes**.

Duas **amostras são independentes** quando a amostra selecionada de uma população não é relacionada à amostra selecionada da segunda população. Duas **amostras são dependentes** quando cada elemento de uma amostra corresponde a um elemento da outra amostra. Amostras dependentes também são chamadas de amostras pareadas ou amostras emparelhadas.

**Figura 8.1** Amostras independentes.



**Figura 8.2** Amostras dependentes.



Por exemplo

- 1) Amostra 1: pesos de 65 calouros universitários antes do início das aulas.  
Amostra 2: pesos dos mesmos 65 calouros após o primeiro ano.

As amostras são dependentes. Como os pesos dos mesmos estudantes são medidos, as amostras são relacionadas. As amostras podem ser pareadas em relação a cada estudante.

- 2) Amostra 1: pontuações de 38 homens adultos em um teste psicológico para transtorno do déficit de atenção com hiperatividade.  
Amostra 2: pontuações de 50 mulheres adultas em um teste psicológico para transtorno do déficit de atenção com hiperatividade.



As amostras são independentes. Não é possível formar pares entre os elementos das amostras, pois os tamanhos das amostras são diferentes e os dados representam pontuações para indivíduos diferentes

## UMA VISÃO GERAL DO TESTE DE HIPÓTESE USANDO DUAS AMOSTRAS

Nesta seção, você aprenderá como testar uma afirmação comparando as médias de duas populações usando amostras independentes.

Por exemplo, um provedor de serviço de internet está desenvolvendo um plano de marketing para determinar se há diferença nos tempos que estudantes universitários do sexo masculino e feminino passam conectados à internet por dia. A única maneira de se concluir com certeza que há diferença é fazendo um censo de todos os universitários, calculando os tempos médios diários que os estudantes do sexo masculino e do sexo feminino ficam conectados e encontrando a diferença. É claro que não é prático fazer esse censo. No entanto, é possível determinar com algum grau de certeza se tal diferença existe.

Oseias Dias de Farias  
oseias.dfarias@gmail.com

Para determinar se existe uma diferença, o provedor de serviço de internet começa assumindo que não há diferença no tempo médio das duas populações. Isto é:

$$\mu_1 - \mu_2 = 0.$$

Então, retirando uma amostra aleatória de cada população, um teste de hipótese baseado nas duas amostras é realizado usando a estatística de teste:

$$x_1 - x_2$$

Lembrando que:

- A hipótese nula  $H_0$  é uma hipótese estatística que geralmente diz que não há diferença entre os parâmetros de duas populações. A hipótese nula sempre contém o símbolo de igualdade
- A hipótese alternativa  $H_a$  é uma hipótese estatística que é verdadeira quando  $H_0$  é falsa.

Para cada cenário, temos uma fórmula para calcular o teste de hipótese de 2 médias.

## **DESVIO-PADRÃO POPULACIONAL CONHECIDO E AMOSTRAS INDEPENDENTES**

Podemos usar um teste z para a diferença entre duas médias populacionais  $\mu_1$  e  $\mu_2$ , quando as amostras são independentes. As condições a seguir são necessárias para realizar tal teste.

1. Os desvios padrão populacionais são conhecidos.
2. As amostras são selecionadas aleatoriamente.
3. As amostras são independentes.
4. As populações são normalmente distribuídas ou cada tamanho de amostra é de pelo menos 30.

Quando esses requisitos são satisfeitos, a distribuição amostral para  $x_1 - x_2$ , ou seja, a diferença das médias das amostras, é uma distribuição normal com média e erro padrão conforme mostrado abaixo:

Em palavras	Em símbolos
<p>A média da diferença das médias amostrais é a diferença presumida entre as duas médias populacionais. Quando nenhuma diferença é presumida, a média é 0.</p> <p>A variância da distribuição amostral é a soma das variâncias das distribuições amostrais individuais para <math>\bar{x}_1</math> e <math>\bar{x}_2</math>. O erro padrão é a raiz quadrada dessa soma.</p>	<p>Média = <math>\mu_{\bar{x}_1 - \bar{x}_2}</math>  <math>= \mu_{\bar{x}_1} - \mu_{\bar{x}_2}</math>  <math>= \mu_1 - \mu_2</math></p> <p>Erro padrão = <math>\sigma_{\bar{x}_1 - \bar{x}_2}</math>  <math>= \sqrt{\sigma_{\bar{x}_1}^2 + \sigma_{\bar{x}_2}^2}</math>  <math>= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}</math></p>

O z a ser calculado nesse caso é:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$



Vamos explicar melhor a fórmula e o teste através de um exemplo.

Um grupo de analistas afirmou que existe diferença entre as médias dos gastos em cartões de crédito em São Paulo e no Rio de Janeiro. Eles se basearam em amostras (independente) de 250 pessoas de cada um desses estados e tiverem esse resultado:

*Média de gastos no cartão para 250 pessoas no Rio de Janeiro:*

$$x_1 = \text{R\$ } 4.777$$

*Média de gastos no cartão para 250 pessoas em São Paulo:*

$$x_2 = \text{R\$ } 4.866$$

Esses analistas, como não estudaram estatística a fundo, viram apenas a média e disseram que o pessoal de São Paulo gasta mais que o pessoal do Rio.

Você, como está estudando comigo, desconfia desse resultado. Então você descobre, de estudos anteriores, que você pode aproximar o desvio-padrão populacional para o Rio para R\$ 1.045 e para São Paulo para R\$ 1.350. Ve

As condições a seguir são necessárias para realizar testes de hipótese de amostras independentes:

1. Os desvios padrão populacionais são conhecidos
2. As amostras são selecionadas aleatoriamente
3. As amostras são independentes
4. As populações são normalmente distribuídas ou cada tamanho de amostra é de pelo menos 30

Seguindo todos os requisitos necessários, você pode usar o teste z para testar a diferença entre duas médias:



$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Sabendo que o problema do gasto médio no cartão cobre todos os requisitos acima e que agora você tem os desvios-padrão de cada uma das amostras, você decide aplicar um teste a 95% de confiança com a seguinte hipótese (lembrem-se que a hipótese de "igualdade" deve estar sempre na hipótese nula):

$$H_0: \mu_1 = \mu_2 \quad \text{e} \quad H_a: \mu_1 \neq \mu_2.$$

Na fórmula acima,  $\mu_1 - \mu_2$  é nossa “premissa da hipótese nula”. Dessa forma, usamos a hipótese nula como valor  $\mu_1 - \mu_2$ , ou seja, se  $\mu_1 - \mu_2 = 0$  pois, em nossa hipótese nula, estamos afirmando que  $\mu_1 = \mu_2$  (e vamos provar se isso é verdade já já). Notem que esse teste é bicaudal!

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021 299 312 66

O nível de significância é 0,05 (ou seja, 1-0,95)

Usando a fórmula, temos:

$$\begin{aligned} z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ &= \frac{(4.777 - 4.866) - 0}{\sqrt{\frac{1.045^2}{250} + \frac{1.350^2}{250}}} \\ &\approx -0,82. \end{aligned}$$

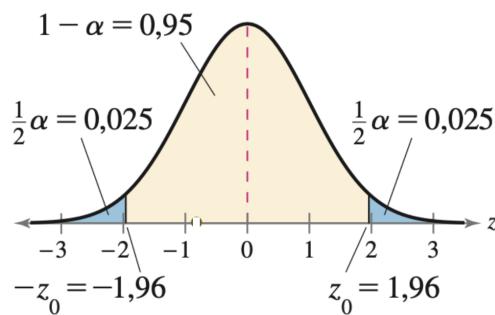
Calcule no site caso prefira:

<https://www.statology.org/two-sample-z-test-calculator/>



A área que comprehende  $z = -0.82$  é 0,4108 (bicaudal), que corresponde ao nosso p-valor. O p-valor calculado é **maior** do que nosso nível de significância (0.05), ou seja **não** devemos rejeitar a hipótese nula. Portanto, a 95% de confiança, dizemos que a média é igual em ambos os grupos. Portanto, São Paulo não tem maiores gastos do que o Rio de Janeiro.

Outra forma de ver o resultado é pela **zona de rejeição**. Como o teste é bilateral (hipótese nula tem um " $\neq$ "), temos:



- Área azul deve ter 0.025 pois a confiança é de 95% e o teste é bilateral
- Pela tabela de z, devemos calcular a área acumulada. Ou seja, a área azul à esquerda + área bege. Dessa forma, a área acumulada é 0.975.
- O valor de z que corresponde a essa área é 1,96 (procurado na tabela)
- Dada a simetria da distribuição normal, sabemos que  $-z$  é -1,96.

As áreas de rejeição da hipótese são as áreas em azul. Sabemos que a área bege vai de -1,96 até 1,96 e que o nosso z calculado é de -0,82. Como o z calculado não está dentro das áreas de rejeição **não** devemos rejeitar a hipótese nula. Portanto, a 95% de confiança, dizemos que a média é igual em ambos os grupos. Portanto, São Paulo não tem maiores gastos do que o Rio de Janeiro.

## DESVIO-PADRÃO POPULACIONAL DESCONHECIDO E AMOSTRAS INDEPENDENTES

Em muitas situações da vida real os desvios padrão populacionais **não** são conhecidos. Logo, usamos o **teste t** para testar a diferença entre duas médias populacionais com desvio-padrão populacional desconhecido. Para usar um teste *t*, as condições a seguir são necessárias:

- Os desvios padrão populacionais são desconhecidos

2. As amostras são selecionadas aleatoriamente
3. As amostras são independentes
4. As amostras são normalmente distribuídas ou cada tamanho de amostra é de pelo menos 30

O teste t para 2 médias é escrito como:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}}$$

Em que  $s_{\bar{x}_1 - \bar{x}_2}$  é definido de 2 formas:

- Se as **variâncias populacionais são consideradas iguais**, então as variâncias das duas amostras são combinadas para se calcular uma estimativa conjunta do desvio padrão. Nesse caso,  $s_{\bar{x}_1 - \bar{x}_2}$  é:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \quad \text{Oseias.Dias de Farias} \quad 1/1$$

$$\text{e g.l} = n_1 + n_2 - 2.$$

- Se as **variâncias populacionais não são iguais ou se não sabemos**, então  $s_{\bar{x}_1 - \bar{x}_2}$  é:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

O grau de liberdade nesse caso é dado por:

$$v = \frac{(A + B)^2}{A^2/(n - 1) + B^2/(m - 1)}$$



Em que  $A = s_1^2/n_1$  e  $B = s_2^2/n_2$  e  $n = n_1 + n_2$

..

Como esse valor é geralmente fracionário, arredonde para o inteiro mais próximo para obter o número de graus de liberdade

### **Exemplo**

Os resultados de um teste estadual de matemática para amostras aleatórias de estudantes ensinados por dois professores diferentes na mesma escola. Podemos concluir que há diferença nas pontuações médias dos testes de matemática para todos os estudantes dos dois professores? Use confiança de 90%. Suponha que as populações são normalmente distribuídas e que as variâncias populacionais não são iguais.

<b>Professor 1</b>	<b>Professor 2</b>
$\bar{x}_1 = 473$	$\bar{x}_2 = 459$
$s_1 = 39,7$	$s_2 = 24,5$
$n_1 = 8$	$n_2 = 18$

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

*Resposta:*

Nossa hipótese deve ser:

$$H_0: \mu_1 = \mu_2 \quad \text{e} \quad H_a: \mu_1 \neq \mu_2.$$

Para variâncias não iguais:

$$\begin{aligned}
 t &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{Use o teste } t \text{ (as variâncias } \text{não } \text{são iguais).} \\
 &= \frac{(473 - 459) - 0}{\sqrt{\frac{(39,7)^2}{8} + \frac{(24,5)^2}{18}}} \quad \text{Suponha que } \mu_1 = \mu_2, \text{ então } \mu_1 - \mu_2 = 0. \\
 &\approx 0,922. \quad \text{Arredonde para três casas decimais.}
 \end{aligned}$$



O grau de liberdade deve ser calculado por:

$$v = \frac{(A + B)^2}{A^2/(n - 1) + B^2/(m - 1)}$$

$$A = 39,7^2 / 8 = 4,96$$

$$B = 24,5^2 / 18 = 33,34$$

$$n = 8$$

$$m = 18$$

$$\text{Logo, } gl = (4,96 + 33,34)^2 / [4,96^2/(8-1) + 33,34^2/(18-1)] = 1466,89/[3,51+65,38] = 21,29 = \text{aprox. } 21$$

Podemos procurar na nossa tabela t o valor de t = 0.922 com gl = 21 para encontrar a área acumulada ou então usar nossa calculadora: <https://select-statistics.co.uk/calculators/two-sample-t-test-calculator/>

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.333124200

What is the sample mean of population 1?	473	i
What is the sample mean of population 2?	459	i
What is the sample standard deviation of population 1?	39,7	i
What is the sample standard deviation of population 2?	24,5	i
How big is the sample from population 1?	8	i
How big is the sample from population 2?	18	i
The p-value is	0.379	i

P-valor é maior que alpha (0.1). Ou seja, não rejeitamos a hipótese nula ao nível de significância de 10%, portanto dizemos que a esse nível de confiança, as médias são estatisticamente iguais.

## DESVIO-PADRÃO POPULACIONAL DESCONHECIDO E AMOSTRAS DEPENDENTES

Para realizar um teste de hipótese usando duas amostras dependentes, você usará uma técnica diferente. Você calculará primeiro a diferença  $d$  entre os elementos de cada par de dados:

$d = (\text{valor do dado na primeira amostra}) - (\text{correspondente valor do dado na segunda amostra}).$

Para esse teste, alguns critérios são importantes:

1. As amostras são selecionadas aleatoriamente.
2. As amostras são dependentes (emparelhadas)
3. As populações são normalmente distribuídas ou o número  $n$  de pares de dados é pelo menos 30.

Quando essas condições são satisfeitas, a distribuição amostral para  $d$  é aproximada por uma distribuição  $t$  com  $n - 1$  graus de liberdade, em que  $n$  é o número de pares de dados.

Oseias Dias de Farias  
oseias.dfarias@gmail.com

021.399.242-66

Para esses casos, precisamos seguir os seguintes passos:

1. Calcular  $\bar{d}$  por:

$$\bar{d} = \frac{\Sigma d}{n}$$

2. Calcular  $s_d$

$$s_d = \sqrt{\frac{\Sigma d^2 - \left[ \frac{(\Sigma d)^2}{n} \right]}{n - 1}}$$

3. Calcular  $t$  por



$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

### Exemplo

Um fabricante de tênis afirma que os atletas podem aumentar a altura de seus saltos verticais quando usam seu tênis por alguns meses. Os saltos de 8 atletas aleatoriamente selecionados são medidos. Após usarem os calçados por 8 meses, os saltos são novamente medidos. Com  $\alpha = 0,10$ , há evidência suficiente para aceitar a afirmação do fabricante? Suponha que os saltos são normalmente distribuídos

*Resposta:*

Atleta	1	2	3	4	5	6	7	8
Altura do salto vertical (antes de usar o calçado)	24	22	25	28	35	32	30	27
Altura do salto vertical (após usar o calçado)	26	25	25	29	33	34	35	30

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Como as amostras são aleatórias e dependentes, pois são as mesmas pessoas sendo avaliadas em um antes x depois, e as populações, normalmente distribuídas, você pode usar o teste  $t$ .

A afirmação é que “os atletas podem aumentar a altura de seus saltos verticais”. Em outras palavras, o fabricante afirma que a altura do salto vertical de um atleta antes de usar o calçado será menor que a altura após usar o calçado. Cada diferença é dada por:

**$d = (\text{altura do salto antes do calçado}) - (\text{altura do salto após o calçado})$ . As hipóteses nula e alternativa são:**

$$H_0: \mu_d \geq 0 \quad \text{e} \quad H_a: \mu_d < 0. \quad (\text{Afirmação.})$$

Logo, aqui estamos falando de um teste unilateral à esquerda (Ha contém " $<$ ").

Como o teste é unilateral à esquerda,  $\alpha = 0,10$  e g.l. =  $8 - 1 = 7$ , o valor crítico é  $t_0 = -1,415$  (vide site - teste unilateral). A região de rejeição é  $t < -1,415$ .

Usando as fórmulas anteriores:

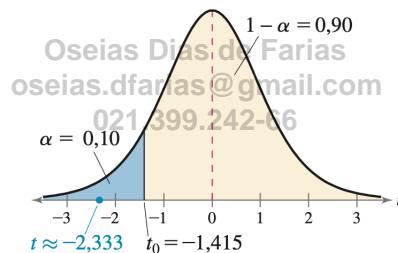
$$\bar{d} = \frac{\Sigma d}{n} = \frac{-14}{8} = -1,75.$$

$$s_d = \sqrt{\frac{\Sigma d^2 - \left[ \frac{(\Sigma d)^2}{n} \right]}{n-1}} = \sqrt{\frac{56 - \frac{(-14)^2}{8}}{8-1}} \approx 2,1213.$$

$$t = \frac{\bar{d} - \mu_d}{s_d / \sqrt{n}}$$

$$\approx \frac{-1,75 - 0}{2,1213 / \sqrt{8}}$$

$$\approx -2,333.$$



Logo, como  $t$  calculado (-2,33) está na região de rejeição, há evidência suficiente, ao nível de significância de 10%, para concordar com a afirmação do fabricante de calçados de que os atletas podem aumentar a altura de seus saltos verticais usando o calçado de treinamento do fabricante.

Você pode chegar no mesmo resultado encontrando o p-valor para o  $t$  calculado e comparando com  $\alpha = 0.1$ .



# 12. Usos e abusos do intervalo de confiança



As seções anteriores mostraram diferentes maneiras de comparar médias usando testes t e z.

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66

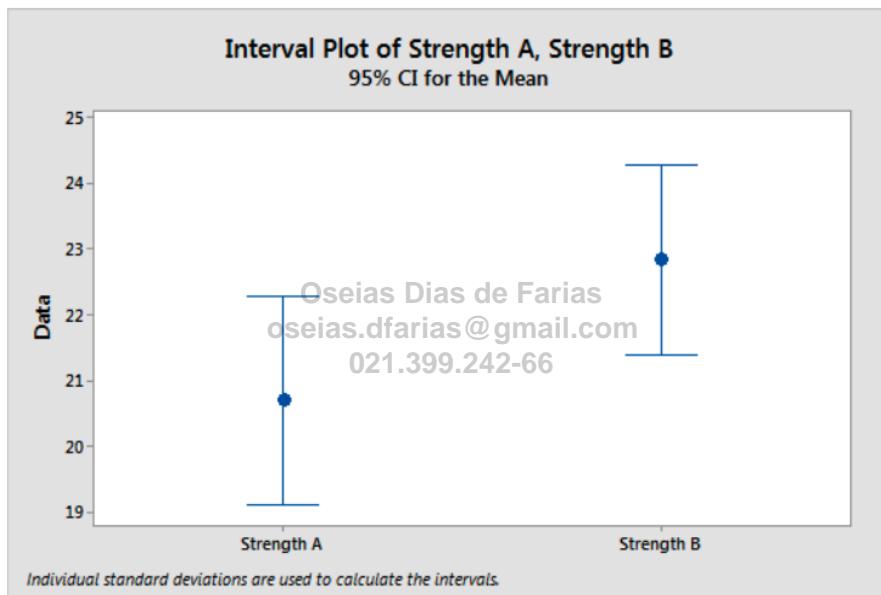
Agora, quero mostrar uma maneira de NÃO comparar dois meios que já vi pessoas usarem com muita frequência.

Muitos frequentemente comparam os intervalos de confiança para **duas** amostras (2 samples) para determinar se a diferença entre duas médias ou proporções (vocês verão isso a seguir) é estatisticamente significativa. Se esses intervalos se sobrepõem, eles concluem que a diferença entre os grupos não é estatisticamente significativa. Se não houver sobreposição, a diferença é significativa. A ideia por trás disso é: se um intervalo de confiança de uma amostra contém valores do intervalo de confiança de outra amostra, está provado que não há diferença estatística entre esses dois grupos.

Embora esse método visual de avaliar a sobreposição seja fácil de executar, ele reduz sua capacidade de detectar diferenças. Felizmente, existe uma solução simples para esse problema que permite realizar uma avaliação visual simples e ainda assim não diminuir o poder de sua análise.

Vou começar mostrando o problema em ação e explicar por que isso acontece. Em seguida, prosseguiremos para um método alternativo fácil que evita esse problema.

Determinar se os intervalos de confiança se sobrepõem é uma abordagem **excessivamente conservadora** para identificar diferenças significativas entre os grupos. É verdade que **quando os intervalos de confiança não se sobrepõem, a diferença entre os grupos é estatisticamente significativa**. **No entanto, quando há alguma sobreposição, a diferença ainda pode ser significativa**.



Ao ver como esses intervalos se sobrepõem, você conclui que a diferença entre as médias dos grupos não é estatisticamente significativa. Afinal, se eles estão sobrepostos, eles não são diferentes, certo?

Essa conclusão parece lógica, mas não é necessariamente verdadeira. Os resultados do teste t de 2 amostras são estatisticamente significativos com um valor p de 0,044. Apesar dos intervalos de confiança sobrepostos, a diferença entre essas duas médias é estatisticamente significativa.

Este exemplo mostra como o método de sobreposição de IC falha em rejeitar a hipótese nula com mais frequência do que o teste de hipótese

correspondente. O uso desse método diminui sua capacidade de detectar diferenças, fazendo com que você perca descobertas essenciais.

Essa aparente discrepância entre os intervalos de confiança e os resultados dos testes de hipóteses pode te deixar surpreso, já que dissemos anteriormente que o intervalo de confiança e o teste de hipótese sempre concordam.

O problema ocorre porque **não estamos comparando os intervalos de confiança corretos com o resultado do teste de hipóteses**. Os resultados do teste se aplicam à **diferença entre as médias**, enquanto os ICs se **aplicam à estimativa da média de cada grupo**, não à diferença entre as médias. Estamos comparando maçãs com bananas, então não é surpresa que os resultados sejam diferentes.

Para obter resultados consistentes, devemos usar intervalos de confiança para **diferenças entre as médias dos grupos**.

Esse tipo de IC sempre concordará com o teste de 2 amostras – apenas certifique-se de usar a combinação equivalente de nível de confiança e nível de significância (por exemplo, 95% e 5%). Agora estamos comparando maçãs com maçãs!

Usando o mesmo conjunto de dados, o intervalo de confiança abaixo apresenta uma faixa de valores que provavelmente contém a diferença média para toda a população. A interpretação continua sendo uma simples avaliação visual. Zero representa nenhuma diferença entre as médias. O intervalo contém zero? Se não incluir zero, a diferença é estatisticamente significativa porque o intervalo não exclui nenhuma diferença. De relance, podemos dizer que a diferença é estatisticamente significativa.

Vamos a um exemplo:

### **Amostra A**

Média 22.84

Desvio-padrão amostral: 3.08

n: 20

### **Amostra B**



Média 20.69

Desvio-padrão amostral: 3.41

n: 20

Vamos supor que as populações têm distribuições normais, as amostras são independentes, o desvio-padrão populacional é desconhecido e que:

H0: Médias são iguais

Ha: Média são diferentes

### Teste de hipótese

Nesse caso, usamos o teste t para variâncias populacionais não iguais

Usando o site para cálculo do p-valor (2 sample t-test)

<https://select-statistics.co.uk/calculators/two-sample-t-test-calculator/>

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

What is the sample mean of population 1?

What is the sample mean of population 2?

What is the sample standard deviation of population 1?

What is the sample standard deviation of population 2?

How big is the sample from population 1?

How big is the sample from population 2?

The p-value is

Ou seja, p-valor menor que alpha (0.05) e, portanto, rejeitamos a hipótese nula (médias não são iguais).

## INTERVALO DE CONFIANÇA PARA DIFERENÇA ENTRE DUAS MÉDIAS

Dependendo dos tipos de amostra e se o desvio padrão da população é conhecido ou não, usaremos um teste z ou um teste t. Porém, a grande chave aqui é que precisamos construir o intervalo de confiança para a diferença.

Um intervalo de confiança (C.I.) para uma diferença entre médias é um intervalo de valores que provavelmente contém a verdadeira diferença entre duas médias populacionais com um certo nível de confiança.

Para estimar essa diferença, coletamos uma amostra aleatória de cada população e calculamos a média para cada amostra. No entanto, não sabemos com certeza se a diferença nas médias da amostra corresponde à verdadeira diferença nas médias da população e é por isso que eles podemos criar um intervalo de confiança para a diferença entre as duas médias. Isso fornece um intervalo de valores que provavelmente conterá a verdadeira diferença entre as médias da população.

Oseias Dias de Farias

Supondo um teste t, teríamos que:

021.399.242-66

$$\text{Lower Limit} = M_1 - M_2 - (t_{\text{CL}}) (S_{M_1 - M_2})$$

$$\text{Upper Limit} = M_1 - M_2 + (t_{\text{CL}}) (S_{M_1 - M_2})$$

Onde  $M_1 - M_2$  é a diferença entre as médias amostrais,  $t_{\text{CL}}$  é o t para o nível de confiança desejado e  $S_{m_1 - m_2}$  é o erro padrão estimado da diferença entre as médias amostrais. O erro padrão é o desvio-padrão dividido por raiz de n.

Isso também pode ser escrito como  $(M_1 - M_2) \pm t_{\text{CL}} * S_{m_1 - m_2}$

O primeiro passo é calcular a estimativa do erro padrão da diferença entre médias  $S_{m_1 - m_2}$ .

Para variâncias populacionais iguais:

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$



Para variâncias populacionais não iguais:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Em ambos os casos, o grau de liberdade é  $n_1+n_2-2$ .

Caso suas amostras sejam dependentes, seu intervalo de confiança será :

$$\bar{x}_d \pm t^* \left( \frac{s_d}{\sqrt{n}} \right)$$

Em que:

$$s_d = \sqrt{\frac{\sum d^2 - \left[ \frac{(\sum d)^2}{n} \right]}{n-1}}$$

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Onde  $d$  é a diferença de cada amostra (antes e depois).

E  $t^*$  é o tcrítico para g.l =  $n - 1$  com o nível de significância que escolher.

Voltando ao exemplo entre a força de dois materiais, temos:

### Amostra A

Média 22.84

Desvio-padrão amostral: 3.08

n: 20

### Amostra B

Média 20.69

Desvio-padrão amostral: 3.41

n: 20



Lembrando que são 2 amostras independentes.

$$g.l = 20+20-2 = 38$$

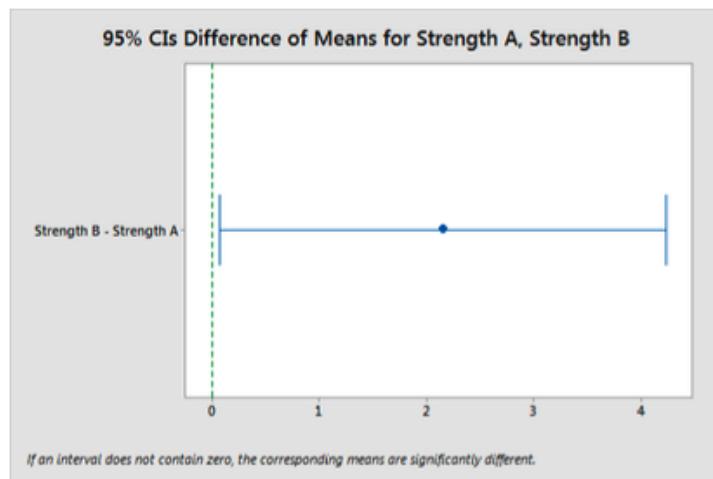
Usando nossa calculadora, temos que tcrítico é 2.0244

What distribution?	t-Student ▾
What type of test?	Two-tailed ▾
Degrees of freedom (d)	38
Significance level	0.05

The test statistic follows the t-distribution with 38 degrees of freedom.  
Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66  
Critical value:  $\pm 2.0244$

Logo, o intervalo de confiança é:

$$(M_1 - M_2) \pm tc^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = (22.84 - 20.69) \pm 2.0244 * \sqrt{\frac{3.08^2}{20} + \frac{3.41^2}{20}} = 2,15 \pm 2,079 \\ = [0.071 \text{ a } 4.229]$$



A interpretação continua sendo uma simples avaliação visual. Zero representa nenhuma diferença entre as médias. O intervalo contém zero? **Se não incluir zero, a diferença é estatisticamente significativa** porque o intervalo não exclui nenhuma diferença.

Além de fornecer uma avaliação visual simples, o intervalo de confiança da diferença apresenta informações cruciais que nem os intervalos de confiança do grupo individuais nem o p-valor fornecem. Ele responde à pergunta, com base em nossa amostra, **quão grande é a diferença entre as duas populações?** Como qualquer estimativa, há uma margem de erro em torno da estimativa pontual da diferença. É importante levar em consideração essa margem de erro antes de agir com base nas descobertas.

Para o nosso exemplo, a estimativa pontual da diferença média é de 2,15 e podemos ter 95% de confiança de que a diferença da população está dentro do intervalo de 0,071 a 4,229.

Como em todos os intervalos, a largura do intervalo para a diferença média revela a precisão da estimativa. Intervalos mais estreitos sugerem uma estimativa mais precisa. E você pode avaliar se toda a gama de valores é praticamente significativa.

Quando o intervalo é muito amplo (impreciso) para ser útil e/ou o intervalo inclui diferenças que não são significativas na prática, você tem motivos para hesitar antes de tomar decisões com base nos resultados. Esses tipos de resultados de IC indicam que você pode não obter benefícios significativos, mesmo que a diferença seja estatisticamente significativa.

Não existe um método estatístico para responder a perguntas sobre quão precisa uma estimativa deve ser ou quão grande um efeito deve ser para ser útil na prática. Você precisará aplicar seu conhecimento da área de assunto ao intervalo de confiança da diferença para responder a essas perguntas.

Para o exemplo nesta seção, é importante observar que a extremidade inferior do IC está muito próxima de zero. Não será surpreendente se a diferença populacional real cair perto de zero, o que pode não ser significativo na prática, apesar do resultado estatisticamente significativo. Se você está pensando em mudar para o Grupo B para um produto mais forte, a melhoria média pode ser muito pequena para ser significativa.

Ao comparar grupos, avalie os intervalos de confiança dessas diferenças em vez de comparar os intervalos de confiança de cada grupo. Esse método é simples e ainda fornece informações valiosas adicionais.

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66



# 13. Intervalo de confiança para proporção



A **estimativa pontual para  $p$** , a proporção populacional de sucessos, é dada pela proporção de sucessos em uma amostra e é denotada por:

$$\hat{p} = \frac{x}{n}$$

Em que  $x$  é o número de “sucessos” em uma amostra e  $n$  é o tamanho da amostra. A estimativa pontual para a proporção populacional de não sucessos é  $q = 1 - p$ .

A **margem de erro** para esse caso será de:

$$E = z_c \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

**Exemplo:** Em uma pesquisa com 1.000 adolescentes americanos, 372 disseram que possuem smartphones. 1) Encontre uma **estimativa pontual** para a proporção populacional de adolescentes americanos que **não** possuem smartphones; 2) Construa um intervalo de confiança a 95% para a proporção populacional de adolescentes americanos que possuem smartphones.

*Resposta:*

$$1) p = 372/1000 = 0,372 \text{ ou } 37,2\%$$

Logo, a proporção de adolescente que **não** possui smartphones é de:

$$q = 1 - 0,372 = 0,628 = 62,8\%$$

$$2) E = zc * \sqrt{(p * q / n)} = 1,96 * \sqrt{(0,628 * 0,372) / 1000} = 0,030$$

Logo,

**Limite inferior** =  $p - E \approx 0,372 - 0,030 = 0,342$

Oseias Dias da Farias  
oseias.dfarias@gmail.com

021.399.242-66

**Limite superior** =  $p + E \approx 0,372 + 0,030 = 0,402$

Portanto, com 95% de confiança, você pode dizer que a proporção populacional de adolescentes americanos que possuem smartphones está entre 34,2% e 40,2%.



# 14. Teste de hipótese para proporção



Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

## 1 SAMPLE

Nas seções anteriores você aprendeu como realizar um teste de hipótese para uma média populacional. Nesta seção, aprenderá como testar uma proporção populacional  $p$ .

Testes de hipótese para proporções podem ser usados, por exemplo, quando políticos querem saber a proporção de seus eleitores que são a favor de certo projeto de lei, quando engenheiros de qualidade testam a proporção de peças defeituosas, quando queremos saber a proporção de pessoas a mais que compram em um novo site, e assim por diante.

O teste z para uma proporção  $p$  é um teste estatístico para uma proporção populacional. O teste z pode ser usado quando uma **distribuição binomial** é dada tal que  $n*p \geq 5$  e  $n*(1-p) \geq 5$  (condição para assumir normalidade). A estatística de teste é a proporção amostral  $\hat{p}$  e a estatística de teste padronizada é :

$$z = \frac{\hat{p} - p}{\sqrt{pq/n}}$$

Em que  $\hat{p}$  é a proporção de uma amostra,  $p$  seria a proporção de uma população,  $q$  é  $(1-p)$  e  $n$  é o tamanho da amostra. Vamos a um exemplo:

Um pesquisador afirma que menos de 40% dos proprietários de celular nos Estados Unidos usam seus aparelhos para a maioria de suas navegações online. Em uma amostra aleatória de 100 adultos, 31% dizem que usam seus aparelhos para a maioria de suas navegações online. Considerando o nível de significância  $\alpha = 0,01$ , há evidência suficiente para concordar com a afirmação do pesquisador?

*Resposta:*

Oseias Dias de Farias

Os produtos  $np = 100(0,40) = 40$  e  $nq = 100(0,60) = 60$  são ambos maiores que 5. Então, você pode usar um teste z. A afirmação é: "menos de 40% usam seus aparelhos para a maioria de suas navegações online". Então, as hipóteses nula e alternativa são:

$$H_0: p \geq 0,4 \quad \text{e} \quad H_a: p < 0,4 \quad (\text{Afirmação.})$$

Esse é um teste unilateral à esquerda (sinal de  $H_a$  é " $<$ ").

Calculando o z:

$$\begin{aligned} z &= \frac{\hat{p} - p}{\sqrt{pq/n}} \\ &= \frac{0,31 - 0,4}{\sqrt{(0,4)(0,6)/100}} \\ &\approx -1,84. \end{aligned}$$

Como  $np \geq 5$  e  $nq \geq 5$ , você pode usar o teste z.

Suponha que  $p = 0,4$ .

Arredonde para duas casas decimais.



Vamos usar o site para entender qual seria o p-valor equivalente a esse z.  
<https://www.calculator.net/z-score-calculator.html>

### Z-score and Probability Converter

Please provide any one value to convert between z-score and probability. This is the equivalent of referencing a z-table.

#### Result

Given Z = -1.84,

$$P(x < Z) = 0.032884$$



$$P(x > Z) = 0.96712$$



$$P(Z < x < 0) = 0.46712$$



$$P(-Z < x < Z) = 0.93423$$



$$P(x < -Z \text{ or } x > Z) = 0.065768$$



Z-score, Z	<input type="text" value="-1.84"/>
Probability, P(x < Z)	<input type="text"/>
Probability, P(x > Z)	<input type="text"/>
Probability, P(0 to Z or Z to 0)	<input type="text"/>
Probability, P(-Z < x < Z)	<input type="text"/>
Probability, P(x < -Z or x > Z)	<input type="text"/>

**Calculate**  **Clear**

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Como o teste é unilateral a esqueda, estamos interessados no primeiro  $P(x < Z)$  dado, ou seja, p-valor = 0,032884. Lembrando que alpha = 0,01

Como p-valor > alpha, não rejeitamos a hipótese nula. Ou seja, não há evidência suficiente, ao nível de significância de 1%, para concordar com a afirmação de que menos de 40% dos proprietários de telefone celular nos Estados Unidos usam seus aparelhos para a maioria de suas navegações online.

## 2 SAMPLES

Se uma afirmação é feita sobre dois parâmetros populacionais  $p_1$  e  $p_2$ , então os possíveis pares de hipóteses nula e alternativa são:

$$\begin{cases} H_0: p_1 = p_2 \\ H_a: p_1 \neq p_2 \end{cases}, \quad \begin{cases} H_0: p_1 \leq p_2 \\ H_a: p_1 > p_2 \end{cases}, \quad \text{e} \quad \begin{cases} H_0: p_1 \geq p_2 \\ H_a: p_1 < p_2 \end{cases}.$$

Independentemente de quais hipóteses você use, para esse teste **sempre assuma em  $H_0$  que não há diferença entre as proporções populacionais ( $p_1 = p_2$ )**. Isto é, o teste será realizado assumindo que  $H_0$  é verdade, ou seja, sempre vamos assumir  $\mathbf{p1 = p2}$ .

As condições a seguir são necessárias para usar um teste z para testar tal diferença.

1. As amostras são selecionadas aleatoriamente
2. As amostras são independentes
3. As amostras são grandes o suficiente para usar uma distribuição amostral normal.

Z pode então ser calculado por:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p} \bar{q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}},$$

Em que  $\hat{p}_1$  e  $\hat{p}_2$  são as proporções amostrais,  $p_1 - p_2$  é a diferença a ser comprovada. Se a hipótese nula declara  $p_1 = p_2$ , então  $p_1 - p_2$  é igual a 0. Lembrem-se que para esse teste,  $H_0$  sempre declarará igualdade, então em qualquer que seja o caso,  $p_1 - p_2$  será 0.

Além disso,  $\bar{p}$  é dado por

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

Em que  $x_1$  e  $x_2$  é número de sucessos em cada amostra (número de casos da proporção) e  $n_1$  e  $n_2$  é o tamanho de cada amostra

E  $\bar{q}$  é dado por:

$$\bar{q} = 1 - \bar{p}$$

### **Exemplo:**

Um estudo com 150 proprietários de carros de passageiros e 200 proprietários de caminhonetes, selecionados aleatoriamente, mostra que 86% dos ocupantes de carros de passageiros e 74% dos ocupantes de caminhonetes usam cinto de segurança. Com  $\alpha = 0,10$ , você pode rejeitar a afirmação de que a proporção de pessoas que usam cinto de segurança é a mesma para os carros de passageiros e as caminhonetes?

Carros de passageiros	Caminhonetes
$n_1 = 150$	$n_2 = 200$
$\hat{p}_1 = 0,86$	$\hat{p}_2 = 0,74$
$x_1 = 129$	$x_2 = 148$

*Resposta:*

Da tabela acima,  $x_1$  e  $x_2$  é o total de passageiros que dizem usar cinto de segurança (note que **sempre**  $n_1 * \hat{p}_1 = x_1$  e  $n_2 * \hat{p}_2 = x_2$ ).

De acordo com as fórmulas acima:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{129 + 148}{150 + 200} = \frac{277}{350} \approx 0,7914$$



$$\bar{q} = 1 - \bar{p} \approx 1 - 0,7914 = 0,2086$$

De acordo com o enunciado:

$$H_0: p_1 = p_2 \quad (\text{afirmação}) \quad \text{e} \quad H_a: p_1 \neq p_2$$

Como queremos provar que as diferenças de proporções é 0 na população

$$p_1 - p_2 = 0$$

Logo,

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p} \bar{q} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \approx \frac{(0,86 - 0,74) - 0}{\sqrt{(0,7914)(0,2086) \left( \frac{1}{150} + \frac{1}{200} \right)}} \approx 2,73.$$

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Usando o site <https://www.calculator.net/z-score-calculator.html> temos que:



### Z-score and Probability Converter

Please provide any one value to convert between z-score and probability. This is the equivalent of referencing a z-table.

#### Result

Given Z = 2.73,

$$P(x < Z) = 0.99683$$



$$P(x > Z) = 0.0031667$$



$$P(0 < x < Z) = 0.49683$$



$$P(-Z < x < Z) = 0.99367$$



$$P(x < -Z \text{ or } x > Z) = 0.0063334$$



Z-score, Z	<input type="text" value="2.73"/>
Probability, $P(x < Z)$	<input type="text"/>
Probability, $P(x > Z)$	<input type="text"/>
Probability, $P(0 < x < Z)$	<input type="text"/>
Probability, $P(-Z < x < Z)$	<input type="text"/>
Probability, $P(x < -Z \text{ or } x > Z)$	<input type="text"/>

**Calculate** **Clear**

Oseias Dias de Farias

Como o teste é bicaudal, p-valor é a soma das duas áreas vermelhas da imagem, ou seja, 0.0063334.  
021.399.242-66

Como p-valor é menor que alpha, rejeitamos a hipótese nula. Ou seja, há evidência suficiente, ao nível de significância de 10%, para rejeitar a afirmação de que a proporção de pessoas que usam cinto de segurança é a mesma para carros de passeio e caminhonetes.

## INTERVALO DE CONFIANÇA PARA A DIFERENÇA DE DUAS PROPORÇÕES

Da mesma forma que para a média, também podemos calcular o intervalo de confiança para a diferença de proporções. Nesse caso temos:

$$(\text{sample difference}) \pm (\text{critical value}) \left( \begin{array}{c} \text{standard error} \\ \text{of difference} \end{array} \right)$$

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}}$$



# 15. Intervalo de confiança para a variância



Na vida real, é necessário controlar o quanto um processo ou uma métrica varia. Por exemplo, o fabricante de uma peça de automóvel deve produzir milhares de peças para ~~os serem fausadas no processo de fabricação~~ ~~031.399.242-66~~. É importante que as peças variem muito pouco dentro do intervalo especificado. Como você pode medir, e consequentemente controlar, a quantidade de variação nas peças? Você pode usar uma **distribuição qui-quadrado** (representado por  $\chi^2$ ) para construir um intervalo de confiança para a variância e o desvio padrão.

## DISTRIBUIÇÃO CHI-QUADRADO ( $\chi^2$ )

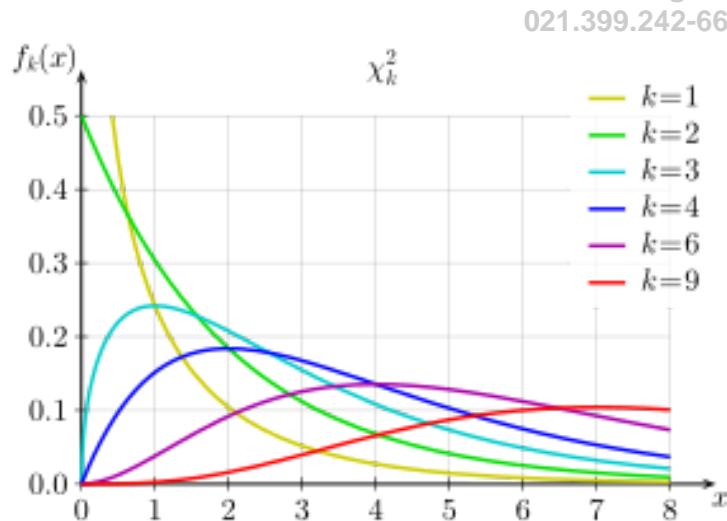
A distribuição chi-quadrado **não é simétrica**. Se a variável aleatória  $x$  tem uma distribuição normal com desvio padrão  $s$ , então:

$$\chi^2 = \frac{(n - 1) s^2}{\sigma^2}$$



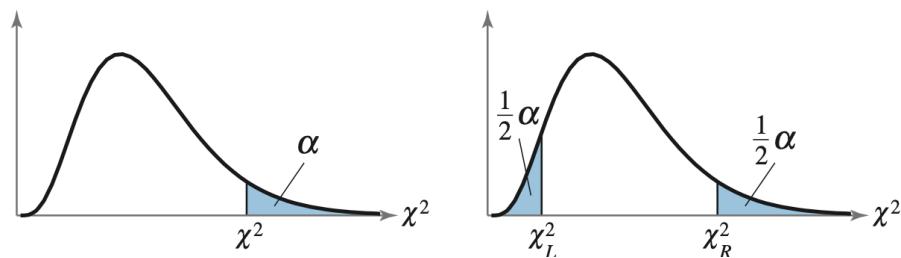
resulta uma distribuição qui-quadrado com  $n - 1$  graus de liberdade, para amostras de qualquer tamanho  $n > 1$ . A seguir algumas propriedades da distribuição qui-quadrado.

1. Todos valores de  $\chi^2$  são maiores ou iguais a 0.
  2. A distribuição qui-quadrado é uma família de curvas, cada uma determinada pelos graus de liberdade. Para construir um intervalo de confiança para  $\sigma$ , use a distribuição qui-quadrado com graus de liberdade iguais ao tamanho da amostra menos um.
- $g.l. = n - 1$
3. A área total abaixo de cada curva da distribuição qui-quadrado é igual a 1.
  4. A distribuição qui-quadrado é assimétrica positiva.
  5. A distribuição qui-quadrado é diferente para cada número de graus de liberdade. Conforme os graus de liberdade aumentam, a distribuição qui-quadrado se aproxima de uma distribuição normal.



Diferente das distribuições z e t que vimos até agora, a qui-quadrado tem dois valores críticos para cada nível de confiança. O valor  $\chi^2_R$  representa valor crítico da cauda direita e  $\chi^2_L$  representa valor crítico da cauda esquerda. A tabela do qui-quadrado lista valores críticos de  $\chi^2$  para vários graus de liberdade e áreas. **Cada área listada na linha do topo da tabela representa a**

**região sob a curva qui-quadrado à direita do valor crítico, diferente das distribuições z e t.**



Graus de liberdade	$\alpha$									
	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025	0,01	0,005
1	—	—	0,001	0,004	0,016	2,706	3,841	5,024	6,635	7,879
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378	9,210	10,597
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348	11,345	12,838
4	0,207	0,297	0,484	0,711	1,064	7,779	9,488	11,143	13,277	14,860
5	0,412	0,554	0,831	1,145	1,610	9,236	11,071	12,833	15,086	16,750

Oseias Dias de Farias

**Exemplo:** Encontre os valores críticos  $\chi^2_R$  e  $\chi^2_L$  para um intervalo de confiança de 95% quando o tamanho da amostra é 18.

Resposta:

$$g.l. = n - 1 = 18 - 1 = 17$$

Nosso nível de confiança é 95%, ou seja,  $c = 95\%$ .

As áreas a direita de  $\chi^2_R$  e  $\chi^2_L$  são:

$$\chi^2_L = \frac{1+c}{2} = \frac{1+0,95}{2} = 0,975$$

$$\chi^2_R = \frac{1-c}{2} = \frac{1-0,95}{2} = 0,025$$

Usando  $g.l. = 17$  e as áreas 0,975 e 0,025, você pode encontrar os valores críticos, conforme as áreas destacadas na tabela.



Graus de liberdade	$\alpha$							
	0,995	0,99	0,975	0,95	0,90	0,10	0,05	0,025
1	—	—	0,001	0,004	0,016	2,706	3,841	5,024
2	0,010	0,020	0,051	0,103	0,211	4,605	5,991	7,378
3	0,072	0,115	0,216	0,352	0,584	6,251	7,815	9,348

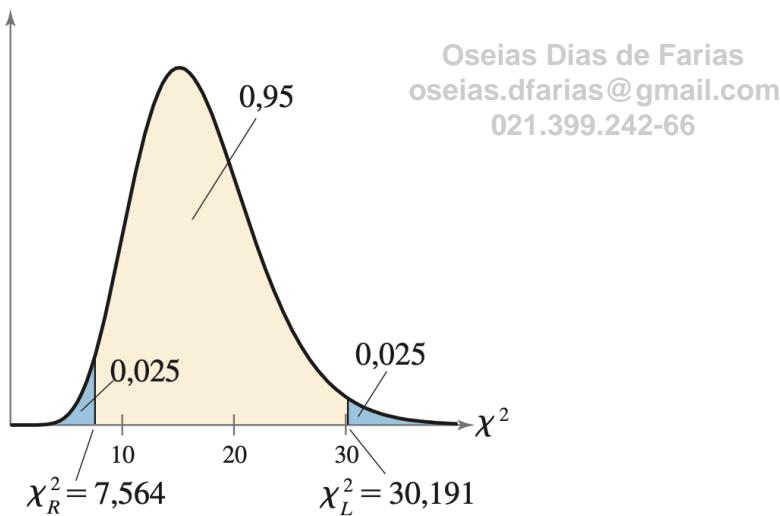
  

15	4,601	5,229	6,262	7,261	8,547	22,307	24,996	27,488
16	5,142	5,812	6,908	7,962	9,312	23,542	26,296	28,845
17	5,697	6,408	7,564	8,672	10,085	24,769	27,587	30,191
18	6,265	7,015	8,231	9,390	10,865	25,989	28,869	31,526
19	6,844	7,633	8,907	10,117	11,651	27,204	30,144	32,852
20	7,434	8,260	9,591	10,851	12,443	28,412	31,410	34,170

$\chi^2_L$        $\chi^2_R$

Da tabela, podemos ver que  $\chi^2_R = 30,191$  e  $\chi^2_L = 7,564$

Então, para uma curva da distribuição qui-quadrado com 17 graus de liberdade (g.l.), 95% da área sob a curva está situada entre 7,564 e 30,191, conforme mostrado na figura abaixo.



## INTERVALO DE CONFIANÇA USANDO O CHI-QUADRADO ( $\chi^2$ )

Você pode usar os valores críticos  $\chi^2_R$  e  $\chi^2_L$  para construir intervalos de confiança para a variância e desvio padrão de uma população. A melhor estimativa pontual para a variância é  $\sigma^2$  e a melhor estimativa pontual para o desvio padrão é  $\sigma$ . **Como a distribuição qui-quadrado não é simétrica, o intervalo de confiança para  $s^2$  não pode ser escrito como  $s^2 \pm E$ .** Você deve



separar os cálculos para os limites do intervalo de confiança, conforme apresentado na próxima definição.

Para **variância**:

$$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$

Para o **desvio padrão**:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_R}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_L}}$$

**Exemplo:** Você seleciona aleatoriamente e pesa as 30 unidades de uma amostra de um antialérgico. O desvio padrão da amostra é de 1,20 miligramas. Supondo que os pesos são normalmente distribuídos, construa intervalos de confiança de 99% para a variância e o desvio padrão da população.

As áreas à direita de  $\chi^2_R$  e  $\chi^2_L$  são:

Área à direita de  $\chi^2_R = (1 - 0,99)/2 = 0,005$   
Oseias Dias de Farias  
oseias.dias@gmail.com  
021.399.242-66

Área à direita de  $\chi^2_L = 0,99 + 0,005 = 0,995$

Usando os valores  $n = 30$ , g.l. = 29 e os valores das áreas, temos de acordo com a tabela de qui-quadrado:

$$\chi^2_R = 52,336 \text{ e } \chi^2_L = 13,121$$

Esses valores podem ser encontrados nesse site também:  
<https://www.danielsoper.com/statcalc/calculator.aspx?id=12>

Degrees of freedom:  ?  
Probability level:  ?  
**Calculate!**

Chi-square ( $X^2$ ) value: **13.12114889**

Com esses valores críticos e  $s = 1,20$ , o intervalo de confiança para a variancia é:

$$\begin{array}{ll}
 \text{Limite inferior} & \text{Limite superior} \\
 \frac{(n-1)s^2}{\chi_R^2} = \frac{(30-1)(1,20)^2}{52,336} & \frac{(n-1)s^2}{\chi_R^2} = \frac{(30-1)(1,20)^2}{13,121} \\
 \approx 0,80 & \approx 3,18 \\
 \end{array}$$

$0,80 < \sigma^2 < 3,18$

E para o desvio padrão:

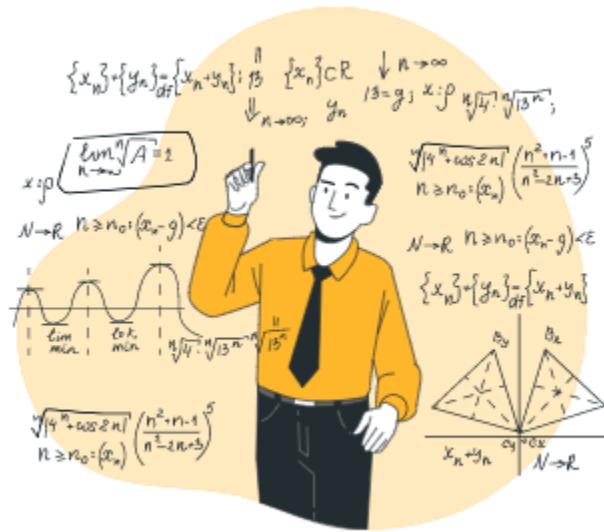
$$\begin{array}{cc}
 \text{Limite inferior} & \text{Limite superior} \\
 \sqrt{\frac{(30-1)(1,20)^2}{52,336}} < \sigma < \sqrt{\frac{(30-1)(1,20)^2}{13,121}} \\
 0,89 < \sigma < 1,78
 \end{array}$$

Logo, com 99% de confiança, podemos dizer que a variância populacional está entre 0,80 e 3,18, e o desvio padrão populacional entre 0,89 e 1,78 miligramas.

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66



# 16. Teste de hipótese para a variância



Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

Na vida real, é importante produzir resultados previsíveis, consistentes. Por exemplo, considere uma empresa que fabrica bolas de golfe. O fabricante deve produzir milhões de bolas de golfe, cada uma tendo o mesmo tamanho e o mesmo peso. Há uma tolerância de variação muito pequena. Para uma população normalmente distribuída, você pode testar a variância e o desvio padrão do processo usando a distribuição qui-quadrado com  $n - 1$  graus de liberdade.

## 1 SAMPLE

Para testar uma variância ou um desvio padrão de uma população que é normalmente distribuída, você pode usar o teste qui-quadrado. **O teste qui-quadrado para uma variância ou desvio padrão não é tão robusto quanto os testes para a média da população ou a proporção da população.** Sempre assumimos nos testes de média e proporção que a população seja normalmente distribuída, ou então aproximadamente distribuída. Aqui com o qui-quadrado, **é essencial, ao realizar um teste qui-quadrado para uma variância ou desvio padrão, que a população seja normalmente**

**distribuída.** Os resultados podem ser equivocados caso a população não seja normal. O teste é dado por:

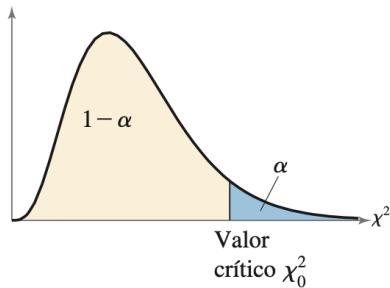
$$\chi^2 = \frac{(n - 1) s^2}{\sigma^2}$$

Antes de aprender como realizar o teste, vamos aprender como encontrar os valores críticos, conforme mostrado nas instruções.

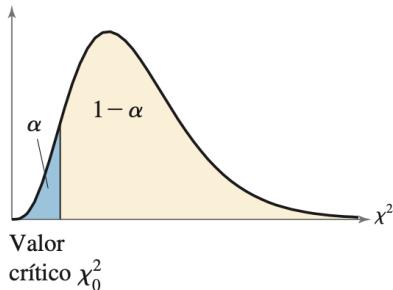
1. Especifique o nível de significância alpha.
  2. Determine os graus de liberdade g.l. = n – 1.
  3. Os valores críticos para a distribuição qui-quadrado são encontrados na em uma tabela. Porém, alternativamente, podemos usar o site:  
<https://www.danielsoper.com/statcalc/calculator.aspx?id=12>
- Para encontrar o valor "Probability level":  
021.399.242-66
- a. teste unilateral à direita, use o valor que corresponde a g.l. e alpha.
  - b. teste unilateral à esquerda, use o valor que corresponde a g.l. e 1 – alpha.
  - c. teste bilateral, use os valores que correspondem a g.l. e  $\frac{1}{2}*\alpha$ , e g.l. e  $1 - \frac{1}{2}*\alpha$



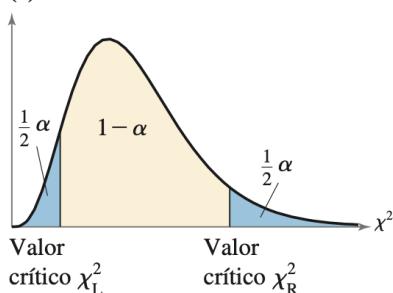
**(a) Teste unilateral à direita**



**(b) Teste unilateral à esquerda**



**(c) Teste bilateral**



Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

Da mesma forma que os testes z e t, você pode:

- 1) Encontrar o  **$\chi^2$  crítico** que corresponde a  $\alpha$  e compará-lo com chi-calculado calculado.
  - Se  $\chi^2$  calculado **menor**  $\chi^2_L$  (esquerda) ou **maior** que  $\chi^2_R$  (direita), dizemos que a probabilidade mínima para não rejeitar  $H_0$  ( $\alpha$ ) não foi atendida, e, portanto, rejeitamos o  $H_0$ .
- 2) Encontrar a probabilidade (área embaixo da curva) que corresponde ao  $\chi^2$  calculado.
  - Essa área é justamente o seu **p-valor!** Com o p-valor, você pode compará-lo com  $\alpha$



**Exemplo:** Uma empresa de processamento de laticínios afirma que a variância da quantidade de gordura no leite integral processado por ela é não mais que 0,25. Você suspeita que essa afirmação esteja errada e descobre que uma amostra aleatória de 41 recipientes de leite tem um variância de 0,27. Para um nível de significância  $\alpha = 0,05$ , há evidência suficiente para rejeitar a afirmação da empresa? Suponha que a população é normalmente distribuída.

Como a amostra é aleatória e a população é normalmente distribuída, você pode usar o teste qui-quadrado. A afirmação é “a variância é não mais que 0,25”. Então, as hipóteses nula e alternativa são:

$$H_0: \sigma^2 \leq 0,25 \quad (\text{Afirmação}) \quad \text{e} \quad H_a: \sigma^2 > 0,25.$$

O teste é unilateral à direita, o nível de significância é  $\alpha = 0,05$  e os graus de liberdade são  $g.l. = 41 - 1 = 40$ . Calculando o valor de qui-quadrado:

. Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66  
Use o teste qui-quadrado.

$$\chi^2 = \frac{(n - 1) s^2}{\sigma^2}$$

$$= \frac{(41 - 1)(0,27)}{0,25} \quad \text{Suponha que } \sigma^2 = 0,25.$$

$$= 43,2$$

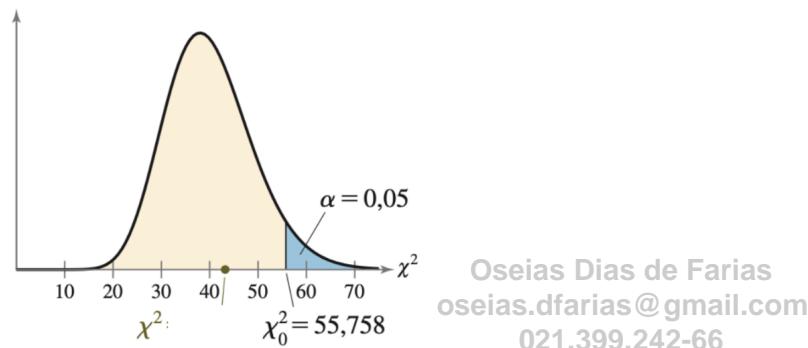
### 1) Abordagem usando o **$\chi^2$ crítico**

O teste é unilateral à direita, o nível de significância é  $\alpha = 0,05$  e os graus de liberdade são  $g.l. = 41 - 1 = 40$ . Logo, o valor crítico é 55,578. Usei a calculadora <https://www.danielsoper.com/statcalc/calculator.aspx?id=12>

Degrees of freedom:	40	?
Probability level:	0.05	?
<b>Calculate!</b>		

**Chi-square ( $\chi^2$ ) value: 55.75847928**

A figura abaixo mostra a localização da região de rejeição e a estatística de teste padronizada  $\chi^2$ . Como  $\chi^2$  não está na região de rejeição, você não rejeita a hipótese nula.



*Interpretação* Não há evidência suficiente, ao nível de significância de 5%, para rejeitar a afirmação da empresa de que a variância da quantidade de gordura no leite integral é não mais que 0,25.

## 2) Abordagem p-valor

Da mesma forma, podemos usar a abordagem do p-valor para entender se rejeitamos ou não a hipótese nula. Nesse caso, vamos usar a seguinte calculadora:

<https://www.vrcbuzz.com/chi-square-test-calculator-for-variance-with-examples/>

Passamos os parâmetros da variância populacional (o que estamos querendo comparar, os 0.25) e passamos todos os outros parâmetros comuns. Nesse caso, ele inclusive já calculou o chi-quadrado para nós.

Aqui precisamos calcular o desvio-padrão. 0.5196 é o desvio-padrão para variância 0.27 ( $\sqrt{0.27}$ ) e 0.5 é o desvio-padrão para variância 0.25 ( $\sqrt{0.25}$ )

Chi Square test Calculator for variance	
Population Standard Deviation ( $\sigma$ )	0.5
Sample Size ( $n$ )	41
Sample Standard Deviation ( $s$ )	0.51961524227
Level of Significance ( $\alpha$ )	0.05
Tail :	<input type="radio"/> Left tailed <input checked="" type="radio"/> Right tailed <input type="radio"/> Two tailed
<b>Calculate</b>	
<b>Results</b>	
Test Statistics $\chi^2$ :	43.2
Degrees of Freedom:	40
$\chi^2$ -critical value(s):	55.7585
p-value:	0.3362 Oseias Dias de Farias dfarias@gmail.com 021.399.242-66

P-valor é maior do que alpha (0.05). Portanto, não rejeitamos a hipótese nula.

*Interpretação* Não há evidência suficiente, ao nível de significância de 5%, para rejeitar a afirmação da empresa de que a variância da quantidade de gordura no leite integral é não mais que 0,25.

**Exemplo 2.** Um fabricante de artigos esportivos afirma que a variância da força de uma certa linha de pesca é de 15,9. Uma amostra aleatória de 15 rolos de linha tem uma variância de 21,8. Para o nível de significância  $\alpha = 0,05$ , há evidência suficiente para rejeitar a afirmação do fabricante? Suponha que a população é normalmente distribuída.

Como a amostra é aleatória e a população é normalmente distribuída, você pode usar o teste qui-quadrado. A afirmação é “a variância é de 15,9”. Então, as hipóteses nula e alternativa são:

$$H_0: \sigma^2 = 15,9 \quad (\text{Afirmação}) \quad \text{e} \quad H_a: \sigma^2 \neq 15,9.$$

O teste é bilateral, o nível de significância é  $\alpha = 0,05$  e os graus de liberdade são

$$g.l. = 15 - 1 = 14.$$

Aqui, vou usar a abordagem do p-valor apenas, mas na figura vou mostrar como seria a região de rejeição. Com a calculadora <https://www.vrcbuzz.com/chi-square-test-calculator-for-variance-with-examples/>

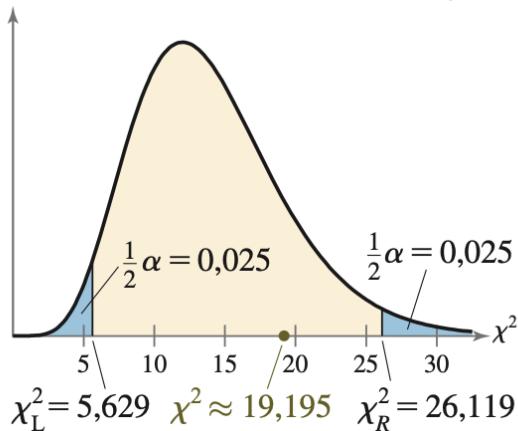
Desvio-padrão populacional a ser testado: raiz(15.9) = 3.98

Desvio-padrão amostral medido: raiz(21.8) = 4.66

Chi Square test Calculator for variance	
Population Standard Deviation ( $\sigma$ )	3.98
Sample Size ( $n$ )	15
Sample Standard Deviation ( $s$ )	4.66 oseias.dfarias@gmail.com 021.399.242-66
Level of Significance ( $\alpha$ )	0.05
Tail :	<input type="radio"/> Left tailed <input type="radio"/> Right tailed <input checked="" type="radio"/> Two tailed
<b>Calculate</b>	
<b>Results</b>	
Test Statistics $\chi^2$ :	19.1926
Degrees of Freedom:	14
$\chi^2$ -critical value(s):	5.6287 26.1189
p-value:	0.1577

O p-valor é **maior** do que alpha (0.05). Logo, não rejeitamos a hipótese nula. Não há evidência suficiente, ao nível de significância de 5%, para rejeitar a afirmação de que a variância da força da linha de pesca é de 15,9.

A figura da região de rejeição ficaria:



## 2 SAMPLES

Para determinar se as variâncias populacionais são iguais, você pode fazer um teste  $F$  com duas amostras. O teste  $F$  é definido por:

Oseias Dias de Farias

oseias.dfarias@gmail.com

021.399.242-66

$$F = \frac{s_1^2}{s_2^2}$$

Em que  $s_1$  é o desvio-padrão de uma amostra e  $s_2$  é o desvio-padrão de outra amostra. Necessariamente,  $s_1$  é maior que  $s_2$  - é assim que você escolherá quem é a primeira e quem é a segunda amostra.

## DISTRIBUIÇÃO F

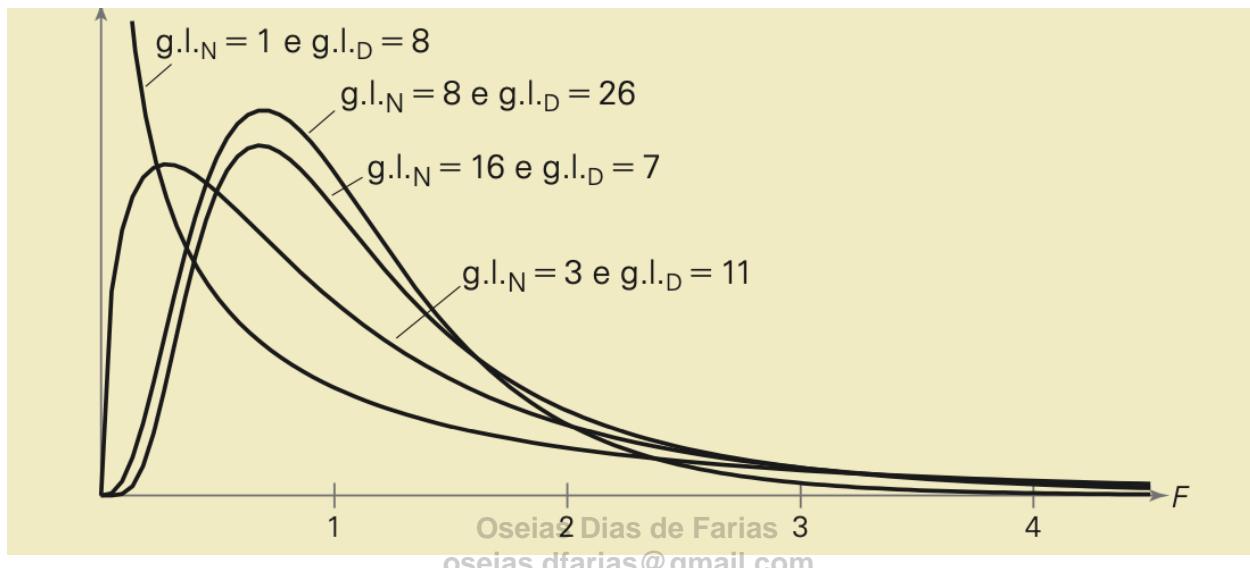
Antes de entrarmos no teste, vamos ver como é a distribuição  $F$ .

1. A distribuição  $F$  é uma família de curvas, cada uma determinada por dois tipos de graus de liberdade: os graus de liberdade correspondentes à variância no numerador, denotado por g.l.N, e os graus de liberdade correspondentes à variância no denominador, denotado por g.l.D.
2. A distribuição  $F$  é positivamente assimétrica
3. A área total sob cada curva de uma distribuição  $F$  é igual a 1.



4. Todos os valores de F são maiores ou iguais a 0.
5. Para todas as distribuições F, o valor médio de F é aproximadamente igual a 1

Abaixo podemos ver uma distribuição F para diferentes graus de liberdade



Como dissemos, o numerador sempre vai ser a variância maior das 2 amostras. Portanto, F sempre será maior que 1.

## COMPARANDO DUAS VARIÂNCIAS COM O TESTE F

Para realizar esse teste, as seguintes **condições** devem

1. As amostras devem ser aleatórias.
2. As amostras devem ser independentes.
3. Cada população deve ter uma distribuição normal.

F calculado será:

$$F = \frac{s_1^2}{s_2^2}$$



Em que o **numerador é a variância da amostra com maior variância** é a variância da amostra com menor variância.

O numerador tem g.l.N = n1 – 1 graus de liberdade e o denominador tem g.l.D = n2 – 1 graus de liberdade, em que n1 é o tamanho da amostra 1 e n2 é o tamanho da amostra 2.

**Exemplo:** Um gerente de restaurante está criando um sistema que se destina a diminuir a variância do tempo que os clientes esperam antes de suas refeições serem servidas. Com o antigo sistema, uma amostra aleatória de 10 clientes teve uma variância de 400. Com o novo sistema, uma amostra aleatória de 21 clientes teve uma variância de 256. Para  $\alpha = 0,10$ , há evidência suficiente para convencer o gerente a mudar para o novo sistema? Suponha que ambas as populações são normalmente distribuídas.

Resposta:

Como  $400 > 256$ ,  $s_1^2 = 400$  (sistema antigo) e  $s_2^2 = 256$  (novo sistema). Portanto,  $s_1^2$  e  $s_2^2$  representam as variâncias da amostra e da população do sistema antigo, respectivamente.

*Oseias Dins de Farias  
oseias.dfarias@gmail.com  
021.399.242-66*

Com a afirmação “A variância dos tempos de espera no novo sistema é menor que a variância dos tempos de espera no sistema antigo”, as hipóteses nula e alternativa são:

$$H_0: \sigma_1^2 \leq \sigma_2^2 \quad \text{e} \quad H_a: \sigma_1^2 > \sigma_2^2.$$

Note que o teste é **unilateral à direita com  $\alpha = 0,10$** , e os graus de liberdade são g.l.N = n1 – 1 = 10 – 1 = 9, e g.l.D = n2 – 1 = 21 – 1 = 20. A estatística de teste é:

$$F = \frac{s_1^2}{s_2^2} = \frac{400}{256} \approx 1,56.$$

Calculando o p-valor com a calculadora <https://www.socscistatistics.com/pvalues/fdistribution.aspx> temos:

F-ratio value:   
 DF- numerator:   
 DF- denominator:

Significance Level:

- .01
- .05
- .10

The  $p$ -value is .194701. The result is *not* significant at  $p < .10$ .

### Exemplo 2:

Você quer comprar ações em uma empresa e está decidindo entre duas ações diferentes. Como o risco de uma ação pode estar associado ao desvio padrão dos preços de fechamento diários, você seleciona aleatoriamente amostras dos preços de fechamento diários para cada ação e obtém os resultados mostrados na tabela abaixo. Com  $\alpha = 0,05$ , você pode concluir que uma das duas ações é um investimento mais arriscado? Suponha que os preços de fechamento das ações são normalmente distribuídos.

Oscar Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Ação A	Ação B
$n_2 = 30$	$n_1 = 31$
$s_2 = 3,5$	$s_1 = 5,7$

Como  $5,7 > 3,5$ , então nossa amostra 1 será a com  $s=5,7$  (Ação B) e amostra 2 será a com  $s = 3,5$  (Ação A).

Com a afirmação “uma das duas ações é um investimento mais arriscado”, as hipóteses nula e a alternativa são:

$$H_0: s_1^2 = s_2^2$$

$$H_a: s_1^2 \neq s_2^2$$

Note que o teste é bilateral com  $1/2\alpha = (0,05)/2 = 0,025$ , e os graus de liberdade são  $g.l.N = n_1 - 1 = 31 - 1 = 30$ , e  $g.l.D = n_2 - 1 = 30 - 1 = 29$ .



Com nossa calculadora, temos p-valor:

#### P-Value from F-Ratio Calculator (ANOVA)

This should be self-explanatory, but just in case it's not: your *F*-ratio value goes in the top box, stick your degrees of freedom for the numerator (between-treatments) in the *DF-* numerator box, degrees of freedom for the denominator (within-treatments) in the *DF-* denominator box, enter your significance level, then press the "Calculate" button.

If you need to derive an *F*-ratio value from raw data, [you can find an ANOVA calculator here](#).

<i>F</i> -ratio value:	2.65
<i>DF-</i> numerator:	30
<i>DF-</i> denominator:	29

Significance Level:

- .01  
 .05  
 .10

The *p*-value is .00514. The result is significant at  $p < .05$ .

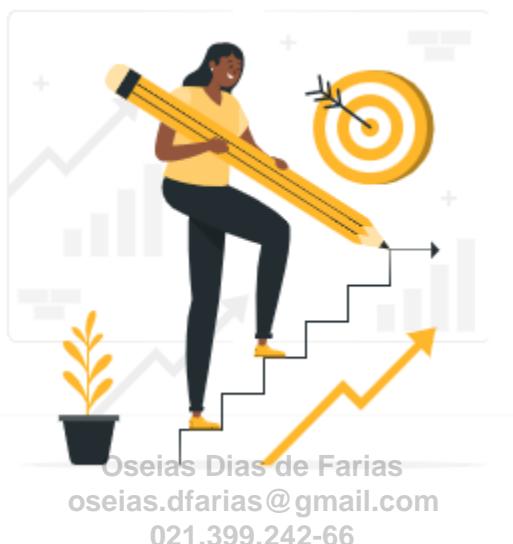
Como p-valor é **menor** do que alpha (0.05), então rejeitamos a hipótese nula. Há evidência suficiente, ao nível de significância de 5%, para confirmar a afirmação de que uma das duas ações é um investimento mais arriscado.

Oseias Diaz de Farias  
oseias.dfarias@gmail.com

021.399.242-66



# 17. Teste de hipótese - categóricos ou proporção com mais de 2 categorias



Nas seções anteriores você aprendeu como testar uma hipótese que compara a proporção em duas categorias (binomial). Agora vamos ver como comparar proporções com mais de 2 categorias.

O **teste qui-quadrado** é usado para testar se uma distribuição de frequência observada (valores amostrais) se ajusta a uma distribuição esperada (valores fixos que queremos comparar).

Geralmente, a hipótese nula estabelece que a distribuição de frequência se ajusta à distribuição esperada e a hipótese alternativa estabelece que a distribuição de frequência não se ajusta.

Para realizar o teste qui-quadrado para a qualidade do ajuste, as seguintes **condições** devem ser satisfeitas:

1. As frequências observadas devem ser obtidas de uma amostra aleatória.
2. Cada frequência esperada deve ser maior ou igual a 5.

## 1 SAMPLE

Se as condições acima são satisfeitas, então a distribuição amostral para o teste é aproximada por uma distribuição qui-quadrado com  $k - 1$  graus de liberdade, sendo  $k$  o número de categorias. A estatística de teste é:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Em que  $O$  representa a frequência observada de cada categoria e  $E$  representa a frequência esperada de cada categoria.

Quando as frequências observadas estão muito próximas das frequências esperadas, as diferenças entre  $O$  e  $E$  serão pequenas e a estatística de teste qui-quadrado será próxima de 0. Portanto, não se rejeita a hipótese nula.

Porém, quando há grandes discrepâncias entre as frequências observadas e as frequências esperadas, as diferenças entre  $O$  e  $E$  serão grandes, resultando em uma estatística de um qui-quadrado grande. Uma estatística de teste qui-quadrado grande é uma evidência para rejeitar a hipótese nula.

### Exemplo

Uma associação de comércio varejista afirma que os meios de preparação de imposto são distribuídos conforme a tabela abaixo. Uma consultoria de impostos seleciona aleatoriamente 300 adultos e pergunta como eles preparam seus impostos. Os resultados encontram-se na tabela abaixo. Para  $\alpha = 0,01$ , teste a afirmação da associação.

Distribuição esperada para os meios		Resultados da pesquisa ( $n = 300$ )	
Contador	24%	Contador	61
À mão	20%	À mão	42
Programa de computador	35%	Programa de computador	112
Amigo/familiar	6%	Amigo/familiar	29
Consultoria de impostos	15%	Consultoria de impostos	56

*Resposta:*

$H_0$ : a distribuição esperada dos métodos de preparação de impostos é: 24% por contador (72 do total de 300), 20% (60 do total de 300) à mão, 35% (105 do total de 300) com programa de computador, 6% (18 do total de 300) por amigo ou familiar e 15% (45 do total de 300) com consultoria de impostos.

$H_a$ : a distribuição dos métodos de preparação de impostos difere da distribuição esperada.

Como há 5 categorias, a distribuição qui-quadrado tem g.l. =  $k - 1 = 5 - 1 = 4$  graus de liberdade.

A estatística de teste qui-quadrado é:

$$\begin{aligned}\chi^2 &= \sum \frac{(O - E)^2}{E} && \text{Oseias Dias de Farias} \\ &= \frac{(61 - 72)^2}{72} + \frac{(42 - 60)^2}{60} + \frac{(112 - 105)^2}{105} && \text{oseias.dfarias@gmail.com} \\ &\quad + \frac{(29 - 18)^2}{18} + \frac{(56 - 45)^2}{45} && 021.399.242-66 \\ &\approx 16,958.\end{aligned}$$

Usando nossa calculadora :

<https://www.socscistatistics.com/pvalues/chidistribution.aspx>

Temos que:



Chi-square score:

16.95

DF:

4

Significance Level:

- 0.01
- 0.05
- 0.10

The P-Value is .001977. The result is significant.

Como p-valor é **menor** do que alpha (0.01), rejetamos a hipótese nula. Há evidência suficiente, ao nível de significância de 1%, para rejeitar a afirmação de que a distribuição dos meios de preparação de imposto observada e a distribuição esperada da associação são as mesmas.

## 2 SAMPLES

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.333.242-00

Suponha que um médico queira determinar se há uma relação entre o consumo de cafeína e o risco de ataque cardíaco. Essas variáveis são independentes ou dependentes? Nesta seção, você aprenderá como usar o teste qui-quadrado para entender diferentes proporções em 2 amostras e fazer associações a partir disso.

Para realizar um teste qui-quadrado para independência, você vai utilizar dados amostrais que estão organizados em uma tabela de contingência.

Essa tabela deve mostrar as frequências observadas (amostrais) para duas amostras. As frequências observadas são organizadas em r linhas e c colunas. A interseção de uma linha e uma coluna é chamada de célula.

### Exemplo

A Tabela abaixo exemplifica uma tabela de contingência 2 x 5.

Ela tem duas linhas e cinco colunas, e mostra os resultados de uma amostra aleatória de 2.200 adultos classificados por duas variáveis: forma favorita de tomar sorvete e gênero



Forma favorita de tomar sorvete					
Gênero	Copo	Casquinha	Sundae	Sanduíche	Outro
Masculino	592	300	204	24	80
Feminino	410	335	180	20	55

A partir dessa coleta de dados, você deve encontrar a frequência esperada para cada conjunto, que é:

$$\text{Frequência esperada } E_{r,c} = \frac{(\text{soma da linha } r) \cdot (\text{soma da coluna } c)}{\text{tamanho da amostra}}$$

Para isso, primeiro somamos os dados de cada linha e de cada coluna:

Oscias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Forma favorita de tomar sorvete						
Gênero	Copo	Casquinha	Sundae	Sanduíche	Outro	Total
Masculino	592	300	204	24	80	1.200
Feminino	410	335	180	20	55	1.000
Total	1.002	635	384	44	135	2.200

Agora você precisa calcular  $E_{r,c}$  para cada linha e cada coluna

$$E_{1,1} = \frac{1.200 \cdot 1.002}{2.200} \approx 546,55$$

$$E_{1,2} = \frac{1.200 \cdot 635}{2.200} \approx 346,36$$

$$E_{1,3} = \frac{1.200 \cdot 384}{2.200} \approx 209,45$$

$$E_{1,4} = \frac{1.200 \cdot 44}{2.200} = 24$$

$$E_{1,5} = \frac{1.200 \cdot 135}{2.200} \approx 73,64$$

$$E_{2,1} = \frac{1.000 \cdot 1.002}{2.200} \approx 455,45$$

$$E_{2,2} = \frac{1.000 \cdot 635}{2.200} \approx 288,64$$

$$E_{2,3} = \frac{1.000 \cdot 384}{2.200} \approx 174,55$$

$$E_{2,4} = \frac{1.000 \cdot 44}{2.200} = 20$$

$$E_{2,5} = \frac{1.200 \cdot 135}{2.200} \approx 61,36$$

Dessa forma, você tem uma tabela Er,c para cada um desses dados. As frequências esperadas (calculadas a partir das amostras) estão entre parênteses e é o dado calculado acima.

O dado que não está em parêntesis é o que chamamos de frequência observada, ou seja, é a frequência que observamos na amostra que colhemos (é a nossa tabela original)

Maneira favorita de tomar sorvete						
Gênero	Copo	Casquinha	Sundae	Sanduíche	Outro	Total
Masculino	592 (546,55)	300 (346,36)	204 (209,45)	24 (24)	80 (73,64)	1.200
	410 (455,45)	335 (288,64)	180 (174,55)	20 (20)	55 (61,36)	1.000
Total	1.002	635	384	44	135	2.200

Agora vamos entender se essas amostras são independentes. Primeiro, como qualquer teste de hipótese, precisamos ter nosso  $H_0$  e nosso  $H_a$



$H_0$ : as variáveis maneira favorita de tomar sorvete e gênero são independentes - ou seja, as proporções são estatisticamente **iguais** em ambas as populações.

$H_a$ : as variáveis maneira favorita de tomar sorvete e gênero são dependentes - ou seja, ou seja, as proporções são estatisticamente **diferentes** em ambas as populações.

Queremos ver se não rejeitamos  $H_0$  a 99% de confiança, por exemplo

**Para usar o teste chi-quadrado várias categorias em 2 amostras (independência):**

1. Determinar o grau de liberdade:  $g.l. = (r - 1)(c - 1)$
2. Determine o valor crítico usando a tabela chi-quadrado
3. Determine a região de rejeição.
4. Determine a estatística de teste e resuma a distribuição amostral.

A tabela de contingência "maneira favorita de tomar sorvete por gênero" tem duas linhas e cinco colunas, então a distribuição qui-quadrado tem  $(r - 1)(c - 1) = (2 - 1)(5 - 1) = 4$  graus de liberdade.

Agora, precisamos determinar o chi-quadrado calculado

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Para isso, para cada  $O$  e  $E$  da tabela acima, temos que calcular  $O-E$ , eleva-lo ao quadrado, dividir por  $E$  e somar todos os valores. Segue o compilado para cada um dos valores:

<b>O</b>	<b>E</b>	<b>O - E</b>	<b>(O - E)<sup>2</sup></b>	<b><math>\frac{(O - E)^2}{E}</math></b>
592	546,55	45,45	2065,7025	3,7795
300	346,36	-46,36	2149,2496	6,2052
204	209,45	-5,45	29,7025	0,1418
24	24	0	0	0
80	73,64	6,36	40,4496	0,5493
410	455,45	-45,45	2065,7025	4,5355
335	288,64	46,36	2149,2496	7,4461
180	174,55	5,45	29,7025	0,1702
20	20	0	0	0
55	61,36	-6,36	40,4496	0,6592
		Oseias Dias de Farias oseias.dfarias@gmail.com 021.399.242-66	$\chi^2 = \sum \frac{(O - E)^2}{E} \approx 23,487$	

Chi-quadrado  $\approx 23,487$ . Usando nossa calculadora de chi-quadrado:

<https://www.socscistatistics.com/pvalues/chidistribution.aspx>

[Report a Chi-Square Result \(APA\)](#)

Chi-square score:   
 DF:

Significance Level:

- 0.01
- 0.05
- 0.10

The P-Value is .000101. The result is significant.

[Calculate](#)



Como p-valor é menor do que alpha (0.01), você rejeita a hipótese nula. Logo, há evidência suficiente, ao nível de significância de 1% para concluir que as variáveis maneira favorita de tomar sorvete e gênero são dependentes - ou seja, as proporções das categorias em cada um dos grupos é diferente.

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66



# 18. ANOVA - comparação com mais de 2 amostras



Neste capítulo, continuaremos com dados contínuos e comparando as médias dos grupos. Cobrimos os testes t em profundidade ao longo da primeira metade deste livro, então não vou revisitá-los aqui. No entanto, se você se lembrar, você pode usar testes t para comparar as médias de dois grupos no máximo. O que você faz se tiver três grupos ou mais? Use análise de variância (ANOVA)!

## ANOVA COM UM FATOR

Análise de variância com um fator é uma técnica de teste de hipótese usada para comparar as médias de três ou mais populações. A análise de variância geralmente é abreviada como ANOVA. A ANOVA com um fator requer **um fator categórico** para a variável independente **e uma variável contínua para a variável dependente**. Os valores do fator categórico dividem os dados contínuos em grupos. O teste determina se as diferenças médias entre esses grupos são estatisticamente significativas. Por exemplo, se o tipo de

fertilizante for sua variável categórica, você pode avaliar se as diferenças entre as médias de crescimento das plantas para pelo menos três fertilizantes são estatisticamente significativas. Para esse teste, usamos a **distribuição F**.

Tecnicamente, você pode usar ANOVA um fator para comparar apenas dois grupos. No entanto, se você tiver dois grupos, normalmente usará um teste t de duas amostras.

As hipóteses padrão para ANOVA 1 fator são as seguintes:

- H0: Todas as médias do grupo são iguais.
- Ha: Nem todas as médias do grupo são iguais.

Se o valor-p for menor que seu nível de significância (geralmente 0,05), rejeite a hipótese nula. Seus dados de amostra suportam a hipótese de que a média de pelo menos uma população é diferente das outras médias populacionais.

## Suposições

Para resultados confiáveis de ANOVA unidirecional, seus dados devem atender às seguintes suposições:

- Amostras aleatórias
- Grupos independentes
- A variável dependente é contínua

Oseias Dias de Farias  
osseias.dfarias@gmail.com  
021.399.242-66

A variável dependente é o resultado que você está medindo. O procedimento compara as médias do grupo desta variável. Por exemplo, o salário é uma variável contínua e você pode comparar os salários médios por grupos.

## A variável independente é categórica

Os níveis da variável categórica definem os grupos que você está comparando. Por exemplo, o curso superior é uma variável categórica. As variáveis categóricas na ANOVA também são conhecidas como fatores.

## Seus dados de amostra devem seguir uma distribuição normal ou cada grupo tem mais de 15 ou 20 observações

Os procedimentos de ANOVA pressupõem que seus dados seguem a distribuição normal. No entanto, como você viu para os testes t, você pode dispensar essa suposição se o tamanho da amostra for grande o suficiente (teorema do limite central).

Para ANOVA 1 fator, quando você tem de 2 a 9 grupos e cada grupo é menor que 15, seus dados podem ser distorcidos e os resultados do teste ainda serão confiáveis. Quando você tem 10-12 grupos, você deve ter pelo menos 20 por grupo para dispensar a suposição de normalidade.

Se seus dados não forem normais e seus tamanhos de amostra forem menores do que essas diretrizes, os resultados do teste podem não ser confiáveis.

### **Os grupos devem ter variações aproximadamente iguais ou usar a ANOVA de Welch**

A forma padrão do teste F de ANOVA 1 fator assume que a variância dentro de cada uma das populações é igual. A diretriz padrão é que você pode assumir que as variâncias da população são iguais se nenhum grupo em sua amostra tiver o dobro da variância de outro grupo.

No entanto, se você não tiver certeza de que as variâncias são iguais, use a ANOVA de Welch, que não assume variâncias iguais. Veremos isso em uma outra sessão.

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

A estatística de teste para um teste ANOVA com um fator é a razão de duas variâncias: a variância entre amostras e a variância dentro das amostras.

$$\text{Estatística de teste} = \frac{\text{variância entre amostras}}{\text{variância dentro das amostras}}$$

A **variância entre amostras** mede as diferenças relacionadas ao tratamento dado a cada amostra. Essa variância, às vezes chamada de quadrado médio entre, é denotada por MSb, MQe ou SE2.

A **variância dentro das amostras** mede as diferenças relacionadas aos valores dentro da mesma amostra e é geralmente devido a erro amostral. Essa variância, às vezes chamada de quadrado médio dentro, é denotada por MSw, MQd, SD2 ou MSe.

As fórmulas necessárias para o cálculo da ANOVA são:

Variação	Soma dos quadrados	Graus de liberdade	Quadrados médios	$F$
<b>Entre</b>	$SS_B$	$g.l._N = k - 1$	$MS_B = \frac{SS_B}{g.l._N}$	$\frac{MS_B}{MS_W}$
<b>Dentro</b>	$SS_W$	$g.l._D = N - k$	$MS_W = \frac{SS_W}{g.l._D}$	

Da mesma forma, a notação  $SSW$  representa a soma dos quadrados dentro das amostras.

$$SS_W = (n_1 - 1) s_i^2 + (n_2 - 1) s_2^2 + \dots + (n_k - 1) s_k^2$$

$$= \sum (n_i - 1) s_i^2$$

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

$$SS_B = n_1 (\bar{x}_1 - \bar{\bar{x}})^2 + n_2 (\bar{x}_2 - \bar{\bar{x}})^2 + \dots + n_k (\bar{x}_k - \bar{\bar{x}})^2$$

$$= \sum n_i (\bar{x}_i - \bar{\bar{x}})^2$$

$g.l.N = k - 1$  (Graus de liberdade do numerador)

$g.l.D = N - k$  (Graus de liberdade do denominador)

Em que  $k$  é o número de amostras e  $N$  é a soma dos tamanhos das amostras

Por fim, calculamos  $F$ . Se  $F$  estiver dentro da zona de rejeição ou se p-valor menor que alpha, rejeitamos  $H_0$

## Exemplo

Um pesquisador médico quer determinar se há diferença nas durações médias de tempo que três tipos de analgésicos levam para aliviar a dor de cabeça. Várias pessoas que sofrem com dores de cabeça são selecionadas aleatoriamente e tomam um dos três medicamentos. Cada pessoa registra o tempo (em minutos) que o medicamento levou para começar a fazer efeito. Os resultados estão na tabela abaixo. Para o nível de significância  $\alpha = 0,01$ , você pode concluir que pelo menos um tempo médio é diferente dos demais? Suponha que cada população de tempos é normalmente distribuída e que as variâncias populacionais são iguais.

Medicamento 1	Medicamento 2	Medicamento 3
12	16	14
15	14	17
17	Oseias Dias de Farias oseias.dfarias@gmail.com 021.315242-66	20
12	19	15
$n_1 = 4$	$n_2 = 5$	$n_3 = 4$
$\bar{x}_1 = \frac{56}{4} = 14$	$\bar{x}_2 = \frac{85}{5} = 17$	$\bar{x}_3 = \frac{66}{4} = 16,5$
$s^2_1 = 6$	$s^2_2 = 8,5$	$s^2_3 = 7$

Resposta:

As hipóteses nula e alternativa são:

$$H_0: m_1 = m_2 = m_3.$$

$H_a$ : Pelo menos uma média é diferente das demais. (Afirmação.)

Como há  $k = 3$  amostras,  $g.l.N = k - 1 = 3 - 1 = 2$ . A soma dos tamanhos das amostras é  $N = n_1 + n_2 + n_3 = 4 + 5 + 4 = 13$ . Então,  $g.l.D = N - k = 13 - 3 = 10$ .

Agora veremos qual é o  $F$  para nosso teste (amostras)

$$\bar{\bar{x}} = \frac{\Sigma x}{N} = \frac{56 + 85 + 66}{13} \approx 15,92$$

$$MS_B = \frac{SS_B}{\text{g.l.N}} = \frac{\sum n_i (\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

$$\approx \frac{(14 - 15,92)^2 + 5 (17 - 15,92)^2 + 4 (16,5 - 15,92)^2}{3 - 1}$$

$$= \frac{21,9232}{2} = 10,9616$$

$$MS_W = \frac{SS_W}{\text{g.l.D}} = \frac{\sum (n_i - 1)s_i^2}{N - k}$$

$$= \frac{(4-1)(6) + (5-1)(8,5) + (4-1)(7)}{13 - 3}$$

$$= \frac{73}{10} = 7,3$$

Usando  $MS_B \approx 10,9616$  e  $MS_W = 7,3$ , a estatística de teste é:

$$F = \frac{MS_B}{MS_W} \approx \frac{10,9616}{7,3} \approx 1,50.$$

Usando nossa calculadora:

<https://www.socscistatistics.com/pvalues/fdistribution.aspx>

Temos:



## P-Value from F-Ratio Calculator (ANOVA)

This should be self-explanatory, but just in case it's not, stick your degrees of freedom for the numerator (between-treatments) and degrees of freedom for the denominator (within-treatments), enter your significance level, then press the "Calculate" button.

If you need to derive an *F*-ratio value from raw data, [you can do so here](#).

*F*-ratio value:

DF - numerator:

DF - denominator:

Significance Level:

- .01
- .05
- .10

The *p*-value is .269329. The result is *not* significant at

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242.66

Como o *p*-valor é maior do que alpha (.01) você não rejeita a hipótese nula. Portanto, dizemos que as médias populacionais são estatisticamente iguais.

## POST HOC ANOVA

Os testes post hoc são parte integrante da ANOVA. Quando você usa ANOVA para testar a igualdade de pelo menos três médias de grupos, resultados estatisticamente significativos indicam que nem todas as médias de grupos são iguais. No entanto, os resultados da ANOVA não identificam quais diferenças particulares entre pares de médias são significativas. Use testes post hoc para explorar as diferenças entre as médias de vários grupos enquanto controla a taxa de erro do experimento.

Nesta seção, mostrarei o que são análises post hoc, os benefícios críticos que elas fornecem e ajudarei você a escolher a correta para seu estudo.

Começaremos com um exemplo de ANOVA 1 fator e depois o usaremos para ilustrar três testes post hoc.



Imagine que estamos testando quatro materiais que estamos considerando para fazer uma peça de produto. Queremos determinar se as diferenças médias entre as forças desses quatro materiais são estatisticamente significativas. Obtemos os seguintes resultados de ANOVA de uma via.

A	B	C	D
40	26.2	36	38
36.9	24.9	39.4	40.8
33.4	30.3	36.3	45.9
42.3	37.9	29.5	40.4
39.1	32.6	34.9	39.9
34.7	37.5	39.8	41.4

Performando a ANOVA com 1 fator, obtivemos  $p\text{-valor} = 0.004$ , ou seja, rejeitamos a hipótese nula que afirma que todas as médias são iguais. Porém, quais são os pares de amostras que têm diferenças na média?

Oseias Dias de Farias

[oseias\\_dfarias@gmail.com](mailto:oseias_dfarias@gmail.com)

021.399.242-66

- i) Eles informam quais médias de grupo são significativamente diferentes de outras médias de grupo
- ii) Eles também controlam a taxa de erro do experimento.

Qual é essa taxa de erro experimental? Para cada teste de hipótese que você executa, há uma taxa de erro do tipo I, que seu nível de significância (alfa) define. Em outras palavras, há uma chance de você rejeitar uma hipótese nula que é realmente verdadeira – um falso positivo. Quando você realiza apenas um teste, a taxa de erro do tipo I é igual ao seu nível de significância, que geralmente é de 5%. No entanto, à medida que você realiza mais e mais testes, sua chance de um falso positivo aumenta. Se você realizar testes suficientes, você aumenta muito sua chance de obter um falso positivo! A taxa de erro para uma família de testes é sempre maior do que para um teste individual.

No contexto ANOVA, você deseja comparar as médias do grupo. Quanto mais grupos você tiver, mais testes de comparação você precisará realizar. Para



nosso exemplo de ANOVA com quatro grupos (A B C D), precisaremos fazer as seis comparações a seguir.

- A-B
- A-C
- A-D
- B-C
- B-D
- C-D

Nosso experimento inclui essa família de seis comparações. Infelizmente, a taxa de erro do experimento aumenta com base no número de grupos em seu experimento.

A tabela abaixo mostra como o aumento do número de grupos em seu estudo faz com que o número de comparações aumente, o que, por sua vez, aumenta a taxa de erro experimental. Abaixo temos um resumo de taxa de erro por quantidade de grupo que teríamos quando pensamos em 95% para cada teste:

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

All Pairwise Comparisons Alpha = 0.05		
Groups	Comparisons	Experimentwise Error Rate
2	1	0.05
3	3	0.142625
4	6	0.264908109
5	10	0.401263061
6	15	0.53670877
7	21	0.659438374
8	28	0.762173115
9	36	0.842220785
10	45	0.900559743
11	55	0.940461445
12	66	0.966134464
13	78	0.981700416
14	91	0.990606054
15	105	0.995418807

A tabela ilustra sucintamente o problema que os testes post hoc resolvem. Normalmente, ao realizar uma análise estatística, você espera uma taxa de

falso positivo de 5%, ou qualquer valor que você definir para o nível de significância. Como mostra a tabela, quando você aumenta o número de grupos de 2 para 3, a taxa de erro quase triplica de 0,05 para 0,143. E, rapidamente piora a partir daí!

Essas taxas de erro são muito altas! Ao ver uma diferença significativa entre os grupos, você teria sérias dúvidas sobre se era um falso positivo em vez de uma diferença real.

Se você usar testes t de 2 amostras para comparar sistematicamente todas as médias dos grupos em seu estudo, encontrará esse problema. Você definiria o nível de significância para cada teste (por exemplo, 0,05) e, em seguida, o número de comparações determinará a taxa de erro do experimento, conforme mostrado na tabela.

Felizmente, os testes post hoc usam uma abordagem diferente. Para esses testes, você define a taxa de erro do experimento que deseja para todo o conjunto de comparações. Em seguida, o teste post hoc calcula o nível de significância para todas as comparações individuais que produzem a taxa de erro experimental especificada. [oseias.dias.de.farias@gmail.com](mailto:oseias.dias.de.farias@gmail.com) 021.399.242-66

Vamos voltar ao nosso exemplo de 4 grupos:

e especificar que a família de seis comparações deve produzir coletivamente uma taxa de erro familiar de 0,05. O teste post hoc que usarei é o método de Tukey. Há uma variedade de testes post hoc que você pode escolher, mas o método de Tukey é o mais comum para comparar todos os pares de grupos possíveis.

Existem duas maneiras de apresentar resultados de testes post hoc—valores de p ajustados e intervalos de confiança simultâneos. Vou mostrar os dois abaixo.

#### Valores P ajustados

A tabela abaixo mostra as seis diferentes comparações em nosso estudo, a diferença entre as médias dos grupos e o valor de p ajustado para cada comparação.

## TESTE DE TUKEY

O teste post hoc que vamos usar é o **método de Tukey**. Há uma variedade de testes post hoc que você pode escolher, mas o método de Tukey é o mais comum para comparar todos os pares de grupos possíveis.

Para usar o teste de Tukey, precisamos:

- As amostras são extraídas independentemente umas das outras.
- As amostras são aleatórias
- Os dados em cada grupo são de uma população normalmente distribuída.
- As populações das quais os dados de cada grupo foram extraídos têm variâncias iguais.
- Os tamanhos de amostra de todos os grupos são iguais. (Se os grupos tiverem tamanhos de amostra diferentes, é realizado um teste de Tukey-Kramer).

O teste é dado por  $qs$ , também chamado de *qscore*:  
Oseias Dias de Farias  
021.399.242-66

$$qs = \frac{Y_A - Y_B}{SE}$$

Em que  $Ya$  é a média do grupo A,  $Yb$  é a média do grupo B e  $SE$  é o erro padrão das duas médias.  $SE$  é dado por:

$$SE_{ANOVA} = \sqrt{\frac{MS_E}{n_j}}$$

$MS_E$  é a variância dentro das amostras e  $n_j$  é a quantidade de amostras em cada conjunto.

Depois, podemos comparar o valor de qscore com o valor de qcritico para entendermos se rejeitamos ou não H0.

Q critico depende da confiança que queremos (em geral, 95%), do g.l, que é dado por N-k, em que N é o total de amostras somadas em cada grupo e k é a quantidade de grupos que temos. Se qcrtico for **menor** que qscore, rejeitamos H0 e falamos que para aquele grupo as diferenças são estatísticas.

Vamos voltar ao nosso exemplo:

A	B	C	D
40	26.2	36	38
36.9	24.9	39.4	40.8
33.4	30.3	36.3	45.9
42.3	37.9	29.5	40.4
39.1	32.6	34.9	39.9
34.7	37.5	39.8	41.4

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021 322 242 66

Vimos que os grupos tem médias diferentes, ou seja, nem todas são iguais. Agora precisamos entender qual delas não é igual. Vamos ao nosso passo a passo:

1. Calcule a média de cada grupo:

A	B	C	D
40	26.2	36	38
36.9	24.9	39.4	40.8
33.4	30.3	36.3	45.9
42.3	37.9	29.5	40.4
39.1	32.6	34.9	39.9
34.7	37.5	39.8	41.4
37.73	31.57	35.98	41.07

Em que a última linha é a média dos valores de cada grupo

## 2. Separe os pares

Pares
A vs B
A vs C
A vs D
B vs C
B vs D
C vs D

## 3. Obtenha os valores absolutos das diferenças das médias:

Pares	Abs(dif)
A vs B	6.17
A vs C	1.75
A vs D	3.33
B vs C	4.42
B vs D	9.50
C vs D	5.08

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Por exemplo A vs B tem médias:

37.73	31.57
-------	-------

$$\text{Dif} = 37.73 - 31.57 = 6.17$$

Nesse caso a diferença é positiva, mas caso não seja, precisamos modificar o sinal para deixar todas as diferenças positivas.

## 4. Calcule o erro padrão: $\text{sqrt}(\text{mse}/\text{ni})$ .

$$\text{MSE} = 15.60$$

$$\text{Erro padrão} = \text{sqrt}(15.60/6)$$

Pares	Abs(dif)	erro padrao
A vs B	6.17	1.612628117
A vs C	1.75	1.612628117
A vs D	3.33	1.612628117
B vs C	4.42	1.612628117
B vs D	9.50	1.612628117
C vs D	5.08	1.612628117

5. calculamos o qscore = abs(dif)/erro padrao

Pares	Abs(dif)	erro padrao	qscore
A vs B	6.17	1.612628117	3.82398558
A vs C	1.75	1.612628117	1.085185097
A vs D	3.33	1.612628117	2.067019233
B vs C	4.42	1.612628117	2.738800483
B vs D	9.50	1.612628117	5.891004813
C vs D	5.08	1.612628117	3.15220433

6. Compare com qcrítico, em que qcrítico é o q para 95% confiança, 4 grupos (k=4) e gl = 6\*4 -3 (N-k) e pode ser obtido:

<https://www.socscistatistics.com/pvalues/qcalculator.aspx>

Logo, qcrítico é 3.96:



### Tukey Q Calculator

This tool will calculate critical values ( $Q_{.05}$  and  $Q_{.01}$ ) for the Studentized normally used in the calculation of Tukey's HSD.

The calculator is easy to use. Just input the number of groups in your自由度 (normally the total number of subjects minus the number of groups) both these values need to be integers. That's all there is to it - just press ready.

#### The Calculator

Number of means/groups (k)	<input type="text" value="4"/>
Degrees of freedom (N - k):	<input type="text" value="20"/>

#### Result

$Q_{.05} = 3.96$

$Q_{.01} = 5.02$

7. qscore pode ser comparado ao qcrítico para acharmos o p-valor. Basta subtrair qcrítico e qscore

Pares	Abs(dif)	erro padrao	qscore	qcrítico	qcrítico-qscore
A vs B	6.17	1.612628117	3.82398558	3.96	0.1360144197
A vs C	1.75	1.612628117	1.085185097	3.96	2.874814903
A vs D	3.33	1.612628117	2.067019233	3.96	1.892980767
B vs C	4.42	1.612628117	2.738800483	3.96	1.221199517
B vs D	9.50	1.612628117	5.891004813	3.96	-1.931004813
C vs D	5.08	1.612628117	3.15220433	3.96	0.8077956703

8. Se qcrítico menor qscore: rejeita  $H_0$  e, portanto, há diferença estatística de médias

Pares	Abs(dif)	erro padrao	qscore	qcrítico	qcrítico-qscore
A vs B	6.17	1.612628117	3.82398558	3.96	0.1360144197
A vs C	1.75	1.612628117	1.085185097	3.96	2.874814903
A vs D	3.33	1.612628117	2.067019233	3.96	1.892980767
B vs C	4.42	1.612628117	2.738800483	3.96	1.221199517
B vs D	9.50	1.612628117	5.891004813	3.96	-1.931004813
C vs D	5.08	1.612628117	3.15220433	3.96	0.8077956703



Logo, o único grupo que tem médias estatisticamente diferentes é o B e o D, a 95% de confiança.

Outra forma de verificar isso seria através do p-valor. Caso você o tivesse, se p-valor do grupo for menor do que alpha, rejeitamos a hipótese nula. Ou seja, aquele grupo tem médias estatisticamente diferentes.

Para consultar como aplicar o teste Tukey-Kramer, acesse o site:  
<https://www.automateexcel.com/stats/tukey-kramer-test/>

Existem diversos testes post hoc.

### Dunnet

Se o seu estudo tiver um grupo de controle e vários grupos de tratamento, talvez seja necessário comparar os grupos de tratamento apenas com o grupo de controle. Use o método de Dunnett quando o seguinte for verdadeiro:

- Antes do estudo, você sabe qual grupo (controle) deseja comparar com todos os outros grupos (tratamentos).
- Você não precisa comparar os grupos de tratamento entre si.

Leia mais sobre esse teste nos seguintes materiais:

- <https://www.statology.org/dunnettts-test/>

### Hsu's MCB

Se o objetivo do seu estudo for identificar o melhor grupo, talvez não seja necessário comparar todos os grupos possíveis. As comparações múltiplas de Hsu com os melhores (MCB) identificam os grupos que são os melhores, insignificanteamente diferentes dos melhores e significativamente diferentes dos melhores.

Use o MCB de Hsu quando você:

- Não sabe de antemão qual grupo você deseja comparar com todos os outros grupos.

- Não precisa comparar grupos que não são os melhores com outros grupos que não são os melhores.
- Pode definir “o melhor” como o grupo com a média mais alta ou com a média mais baixa.

O MCB de Hsu compara cada grupo ao grupo com a melhor média (mais alta ou mais baixa). Usando este procedimento, você pode acabar com vários grupos que não são significativamente diferentes do melhor grupo. Tenha em mente que o grupo que é realmente melhor em toda a população pode não ter a melhor média amostral devido ao erro de amostragem. Os grupos que não são significativamente diferentes do melhor grupo podem ser tão bons ou até melhores do que o grupo com a melhor média amostral.

Por exemplo, o grupo D é o melhor grupo geral porque tem a maior média (41,07). O procedimento compara D a todos os outros grupos. Nesse caso, os melhores grupos seriam D e outro grupo que supere muito o D estatisticamente falando (cuidado, pois não necessariamente é o que tem a maior média além de D devido aos desvios-padrões).

021.399.242-66

Leia mais sobre esse teste nos seguintes materiais:

- <https://support.minitab.com/en-us/minitab/21/help-and-how-to/statistics/al-modeling/anova/supporting-topics/multiple-comparisons/what-is-hsu-s-mcb/>
- <https://www.real-statistics.com/one-way-analysis-of-variance-anova/unplanned-comparisons/hsus-mcb/>

## Games Howell

O teste post-hoc de Games-Howell é outra abordagem **não paramétrica** para comparar combinações de grupos ou tratamentos. Embora bastante semelhante ao teste de Tukey em sua formulação, o teste de Games-Howell **não assume variâncias iguais, tamanhos amostrais iguais e nem amostras normais**. O teste foi projetado com base na correção dos graus de liberdade de Welch:



$$df' = \frac{\left( \frac{s_i^2}{n_i} + \frac{s_j^2}{n_j} \right)^2}{\frac{(s_i^2)^2}{n_i} + \frac{(s_j^2)^2}{n_j}} \cdot \frac{n_i - 1}{n_i} + \frac{n_j - 1}{n_j}$$

Leia mais sobre o teste nesse site:

<https://aaronschlegel.me/games-howell-post-hoc-multiple-comparisons-test-python.html#:~:text=The%20Games%2DHowell%20test%20is,variances%20or%20equal%20sample%20sizes.>

### **Resumo de todos os testes para ANOVA 1 fator:**

<https://www.ibm.com/docs/en/spss-statistics/saas?topic=anova-one-way-post-hoc-tests>

## **INTRODUÇÃO ANOVA COM DOIS FATORES**

Dias de Farias  
oseias.dfarias@gmail.com

021.399.242-66

Use ANOVA 2 fatores para avaliar as diferenças entre as médias dos grupos que são definidas por dois fatores categóricos. Como todos os testes de hipóteses, a ANOVA 2 fatores usa dados amostrais para inferir as propriedades de toda a população.

Para realizar essa análise, você precisará de duas variáveis categóricas, também chamadas de fatores. Esses fatores são suas variáveis independentes. Cada fator tem um número finito de valores possíveis, que são conhecidos como níveis. Por exemplo, gênero é um fator categórico que tem os dois níveis de masculino e feminino.

Você também precisa de uma variável de resultado contínua, que é a variável dependente. A ANOVA de 2 fatores determina se as diferenças médias entre esses grupos são estatisticamente significativas.

Por exemplo, avaliaremos se os dois fatores categóricos de gênero e curso superior correspondem a diferenças de renda, uma variável contínua.



A ANOVA 2 fatores começa a abordar aspectos dos modelos lineares de mínimos quadrados (que vocês vão ver em regressão). Por isso, não a abordaremos agora, porém voltaremos nesse assunto após a matéria de regressão.

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66



# 19. Pensamento crítico: Rejeição da hipótese nula



"Deixar de rejeitar a hipótese nula quando é uma maneira de afirmar que os resultados de seu teste de hipótese não são estatisticamente significativos". O que isso significa exatamente?

Embora a não rejeição da hipótese nula pareça mais direta, ela não é estatisticamente precisa!

Antes de prosseguir, vamos recapitular algumas informações necessárias. Quando a hipótese nula tem a **igualdade** de médias ou proporções, temos:

- A hipótese nula afirma que não há efeito ou relação entre as variáveis.
- A hipótese alternativa afirma que o efeito ou relação existe.

Assumimos que a hipótese nula está correta até que tenhamos evidências suficientes para sugerir o contrário.

Depois de realizar um teste de hipótese, há apenas dois resultados possíveis.

- Quando seu valor-p é menor ou igual ao seu nível de significância, você rejeita a hipótese nula. Os dados favorecem a hipótese alternativa. Seus resultados são estatisticamente significativos.
- Quando seu valor-p é maior que seu nível de significância, você deixa de rejeitar a hipótese nula.

Para entender por que não aceitamos o nulo, considere que você não pode provar uma negativa. **A falta de evidência não é prova de que algo não existe. Você só não provou que existe.** Pode existir, mas seu estudo perdeu. Essa é uma diferença enorme, e é a razão para a redação complicada. Vejamos várias analogias.

Em um julgamento, começamos com a suposição de que o réu é inocente até que se prove o contrário. O promotor deve trabalhar duro para exceder um padrão probatório para obter um veredito de culpado. Se o promotor não cumprir esse ônus, isso não prova que o réu é inocente. Em vez disso, não havia provas suficientes para concluir que ele é culpado.

Talvez o promotor tenha conduzido uma investigação de má qualidade e perdido as pistas? Ou, o réu cobriu com sucesso seus rastros? Consequentemente, o veredito nesses casos é “inocente”. Esse julgamento não diz que o réu é inocente, apenas que não havia provas suficientes para afastar o júri da suposição padrão de inocência.

Quando você está realizando testes de hipóteses em estudos estatísticos, em muitos casos você pode desejar encontrar um efeito ou relação entre as variáveis. A posição padrão em um teste de hipótese é que a hipótese nula está correta. Como em um processo judicial, a evidência da amostra deve exceder o padrão probatório, que é o nível de significância, para concluir que existe um efeito.

O teste de hipótese avalia a evidência em sua amostra. Se o seu teste não detectar um efeito, isso não é prova de que ele não existe. Significa apenas

que sua amostra continha uma quantidade insuficiente de evidências para concluir que ela existe. Como as espécies “extintas” ou o promotor que perdeu as pistas, o efeito pode existir na população geral, mas não em sua amostra específica. Consequentemente, os resultados do teste não rejeitam a hipótese nula, que é análoga a um veredito de “inocente” em um julgamento. Simplesmente não havia evidência suficiente para mover o teste de hipótese da posição padrão de que o nulo é verdadeiro.

O ponto crítico dessas analogias é que a falta de evidências não prova que algo não existe – apenas que você não o encontrou em sua investigação específica. **Portanto, você nunca aceita a hipótese nula por completo.**

No caso de dizer que as amostras são estatisticamente iguais em H<sub>0</sub> (não há efeito), aceitar a hipótese nula indicaria que você provou que um efeito não existe. Como você viu, não é bem assim. Você não pode provar um negativo. Em vez disso, a força de sua evidência fica aquém de ser capaz de rejeitar o nulo. Consequentemente, deixamos de rejeitá-lo.

Deixar de rejeitar o nulo indica que nossa amostra não forneceu evidências suficientes para concluir que o efeito existe. No entanto, ao mesmo tempo, **essa falta de evidência não prova que o efeito não existe.** Quais são as possíveis interpretações de não rejeitar a hipótese nula? Vamos trabalhar com eles.

Primeiro, é possível que o efeito de fato não exista na população, então seu teste de hipótese não o detectou na amostra. Faz sentido, certo? Embora essa seja uma possibilidade, não termina aí.

Outra possibilidade é que o efeito existe na população, mas o teste não o detectou por vários motivos. Esses motivos incluem o seguinte:

- O tamanho da amostra era muito pequeno para detectar o efeito.
- A variabilidade nos dados era muito alta. O efeito existe, mas o ruído em seus dados inundou o sinal (efeito).
- Por acaso, você coletou uma amostra aleatória não representativa da sua população. Ao lidar com amostras aleatórias, o acaso sempre desempenha

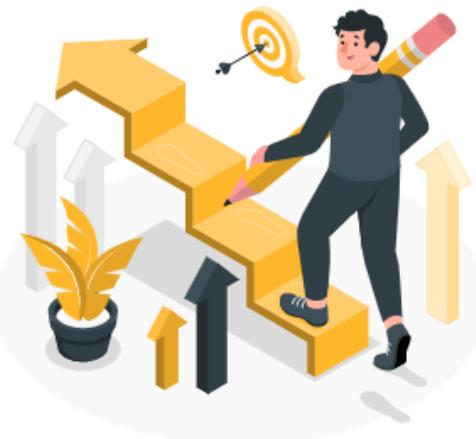
um papel nos resultados. A sorte do sorteio pode ter feito com que sua amostra não refletisse um efeito que existe na população.

Observe como os estudos que coletam uma pequena quantidade de dados ou dados de baixa qualidade tendem a perder um efeito que existe? Esses estudos tiveram uma capacidade inadequada para detectar o efeito. Certamente não queremos tomar resultados de estudos de baixa qualidade como prova de que algo não existe!

No entanto, não detectar um efeito (rejeitar  $H_0$ ) não significa necessariamente que um estudo seja de baixa qualidade. O acaso no processo de amostragem pode funcionar contra até mesmo os melhores projetos de pesquisa!

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66

# 20. Pensamento crítico: Deep dive no p-valor



Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.333.242-68

Antes de avançarmos e falarmos sobre os outros testes de hipótese e intervalos de confiança para além da média, agora que vocês viram um teste de hipótese na prática precisamos levantar uma discussão.

Muitos dos testes que vocês farão na vida de vocês será comparando 2 ou mais amostras. Vimos que quando é esse o caso, nossa hipótese nula geralmente traz “média 1 = média 2”. Quando fazemos alguma melhoria e testamos essas duas amostras, frequentemente torcemos para que H<sub>0</sub> seja rejeitado, uma vez que queremos de fato que essa melhoria seja impactante.

## MAS AFINAL, O QUE SIGNIFICA UM P-VALOR ALTO?

P-valores altos (não rejeitar o H<sub>0</sub>) não provam que não há efeito. P-valores altos indicam que sua evidência não é forte o suficiente para sugerir que existe um efeito na população. Um efeito pode existir, mas é possível que o tamanho do efeito seja muito pequeno, o tamanho da amostra seja muito pequeno ou haja muita variabilidade para o teste de hipótese detectá-lo.

Embora você possa não gostar de obter resultados que não sejam **estatisticamente significativos**, esses resultados podem impedir que você

tire conclusões precipitadas e que tome decisões com base em ruído aleatório em seus dados! Valores altos de  $p$  ajudam a evitar erros dispendiosos. Afinal, se você basear suas decisões em erros aleatórios, você não obterá os benefícios esperados. Esta proteção aplica-se a estudos sobre métodos de ensino, eficácia de medicamentos, força do produto e assim por diante.

## SIGNIFICÂNCIA PRÁTICA VERSUS ESTATÍSTICA

Anteriormente vimos como um efeito relativamente grande na sua amostra pode realmente ser um erro aleatório (lembrem-se dos boxplots com diversos efeitos mostrados na aula 3). Vimos como os valores- $p$  altos podem protegê-lo de tirar conclusões precipitadas com base no erro.

Agora imagine que você acabou de realizar um teste de hipótese e seus resultados são estatisticamente significativos. Oba! Esses resultados são importantes, certo? Não tão rápido. A significância estatística não significa necessariamente que os resultados são significativos no mundo real. Você pode ter resultados significativos para um efeito pequeno.

021.399.242-66

Vamos falar agora sobre as diferenças entre significância prática e significância estatística e como determinar se seus resultados são significativos no mundo real.

### Significância estatística

O procedimento de teste de hipótese determina se os resultados da amostra que você obtém são prováveis se você assumir que a hipótese nula é correta para a população. Se os resultados forem improváveis (caem na zona de rejeição), você pode rejeitar a hipótese nula e concluir que existe um efeito estatisticamente significativo.

Consequentemente, pode parecer lógico que os  $p$ -valor e a significância estatística estejam relacionados à importância. No entanto, isso é falso porque as condições além de tamanhos de efeito grandes, podem produzir valores  $p$  minúsculos, o que nos levaria a rejeitar  $H_0$  e assumir que o efeito é estatisticamente significativo.



Isso acontece quando o **tamanho da amostra é muito grande** e/ou os dados têm **baixa variabilidade**. Vamos ver o porquê.

À medida que o tamanho da amostra aumenta, o teste de hipóteses ganha maior **poder estatístico** para detectar pequenos efeitos.

Com um tamanho de amostra grande o suficiente, o teste de hipótese pode detectar um efeito tão minúsculo que não tem sentido prático.

Quando seus dados de amostra têm baixa variabilidade, os testes de hipóteses podem produzir estimativas mais precisas do efeito da população. Essa precisão permite que o teste detecte pequenos efeitos.

A significância estatística indica apenas que você tem evidências suficientes para concluir que existe um efeito. É uma definição matemática que não sabe nada sobre a área temática e o que constitui um efeito importante.

### **Significância prática**

Tamanho importa!

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Enquanto a significância estatística se refere à existência de um efeito (rejeição de H<sub>0</sub>), a significância prática refere-se à sua magnitude. No entanto, nenhum teste estatístico pode dizer se o efeito é grande o suficiente para ser importante em sua área de estudo. Em vez disso, você precisa aplicar seu conhecimento da área de assunto e experiência para determinar se o efeito é grande o suficiente para ser significativo no mundo real.

Como você faz isso? Acho que é útil identificar os menores tamanho do efeito que ainda tem algum significado prático. Mais uma vez, este processo requer que você use seu conhecimento do assunto para fazer essa determinação. Se o tamanho do efeito do seu estudo for maior que este menor efeito significativo, seus resultados são praticamente significativos.

Por exemplo, suponha que você esteja avaliando um programa de treinamento comparando os resultados dos testes dos participantes do programa com aqueles que estudam por conta própria. Além disso, decidimos que a diferença entre esses dois grupos devem ter pelo menos cinco pontos para representar um tamanho de efeito praticamente

significativo. Um efeito de 4 pontos ou menos é muito pequeno para ser relevante.

Após a realização do estudo, a análise encontra uma diferença entre os dois grupos. Os participantes do programa de estudos marcaram uma média de 3 pontos a mais em um teste de 100 pontos.

Enquanto estes resultados podem ser estatisticamente significativos (rodamos o teste de hipótese para comprovar), a diferença de 3 pontos é menor do que nosso limite de 5 pontos. Consequentemente, nosso estudo fornece evidências de que esse efeito existe, mas é muito pequeno para ser significativo na realidade mundial. O tempo e o dinheiro que os participantes gastam no treinamento programa não valem uma melhoria média de apenas 3 pontos.

### **Nem todas as diferenças estatisticamente significativas são interessantes!**

Isso é bem direto ao ponto. Infelizmente, há uma pequena complicação. O tamanho do seu efeito é apenas uma estimativa porque vem de uma amostra. Graças ao erro de amostragem, há uma margem de erro em torno dele.

Precisamos de um método para determinar se o efeito estimado ainda tem significância prática quando você considera essa margem de erro.

Um intervalo de confiança é um intervalo de valores que provavelmente contém o valor da população. Eu escrevi sobre intervalos de confiança anteriormente, então não vou me alongar aqui. A ideia central é que os intervalos de confiança incorporam a margem de erro criando um intervalo em torno do efeito estimado.

É provável que o valor da população caia dentro desse intervalo. Sua tarefa é determinar se todos, alguns ou nenhum desses intervalos representam efeitos com significância prática.

Vamos a um exemplo:

Suponha que conduzimos dois estudos sobre 2 programas de treinamento descritos acima. Ambos os estudos são estatisticamente significativos e produzem uma média de 9 pontos acima do que os estudantes tiravam

antes. Esse aumento parece bom porque são maiores do que nosso menor tamanho de efeito significativo de 5.

No entanto, essas estimativas não incorporam a margem de erro. Os intervalos de confiança (ICs) para esse aumento de média em ambos os estudos abaixo fornecem essa informação crucial.

Método A: [3 15]

Método B: [7 11]

O IC do Método A se estende de valores que são muito pequenos para serem significativos (<5) para aqueles que são grandes o suficiente para serem significativos. Mesmo que o método seja estatisticamente significativo e o efeito estimado é 9, o IC cria dúvidas sobre se o efeito populacional real é grande suficiente para ser importante.

Por outro lado, o IC para o Método B contém apenas efeito significativo tamanhos. Podemos estar mais confiantes de que o tamanho do efeito da população é grande o suficiente para nos importarmos!

021.399.242-66

Intervalos de confiança são ótimos porque você pode usá-los para determinar a significância estatística e importância prática. Os intervalos de confiança focam no tamanho do efeito e da incerteza em torno da estimativa, em vez de apenas se o efeito existe.

## CUIDADO PARA NÃO TOMAR CONCLUSÕES PRECIPITADAS

Precisamos lembrar agora dos nossos tipos de erros.

	<b>Rejeita H<sub>0</sub></b>	<b>Não rejeita H<sub>0</sub></b>
<b>H<sub>0</sub> é verdadeiro (real)</b>	<b>Erro tipo I: Falso positivo (FP)</b>	<b>Acertamos \o/ Efeito não existe</b>
<b>H<sub>0</sub> é falso (real)</b>	<b>Acertamos \o/ Efeito existe</b>	<b>Erro tipo II: Falso negativo (FN)</b>



Sempre que rejeitamos H<sub>0</sub> (resultados estatisticamente significativos) corremos o risco de estar cometendo o erro do tipo I, ou seja, os falsos positivos

Neste contexto, um falso positivo ocorre quando você obtém um p-valor baixo e, sem saber, rejeita uma hipótese nula que é realmente verdadeira. Você conclui que existe um efeito (médias não são iguais) na população quando não existe.

Do ponto de vista científico, as altas taxas de falsos positivos são problemáticas por causa dos resultados enganosos. Do ponto de vista prático, se você estiver usando um teste de hipótese para melhorar um produto ou processo, você não obterá os benefícios esperados se os resultados do teste forem falsos positivos. Isso pode te custar muito dinheiro!

Essas dicas ajudarão você a desenvolver uma compreensão dos resultados do seu teste. Vou usar um estudo real de vacina contra a AIDS realizado na Tailândia para trabalhar com essas considerações. O estudo obteve um valor de p de 0,039, o que parece ótimo (mostra aqui que as médias de quem tomou e quem não tomou não são iguais). No entanto, depois de ler o que escrevemos, você pode pensar de forma diferente.

### Dica 1: P-valores menores são interessantes

Os analistas geralmente veem os resultados estatísticos como significativos ou não. O foco está em saber se o p-valor é menor que o nível de significância porque os resultados estatisticamente significativos são altamente valorizados. Infelizmente, esse processo de decisão binária é uma simplificação excessiva porque nenhum nível de significância específico determina corretamente quais estudos têm efeitos populacionais reais 100% das vezes. Em vez disso, precisamos nos concentrar em entender a relação entre as taxas de falso-positivos e p-valores.

Existem vários estudos de simulação que mostram que taxas mais baixas de falsos positivos estão associadas com p-valores menores. Por exemplo, um valor de p próximo a 0,05 geralmente tem uma taxa de falso positivo de 25-50%. No entanto, um valor de p de 0,0027 geralmente tem um taxa de falso positivo de cerca de 4,5%. Essa taxa de falso positivo está próxima da taxa que é muitas vezes erroneamente atribuído a um valor p de 0,05.

*P-valores mais baixos indicam evidências mais fortes contra a hipótese nula e uma menor probabilidade de um falso positivo.*

É importante falar que não há relação diretamente calculável entre os p-valores e a taxa de falsos positivos. No entanto, a simulação estudos e abordagens Bayesiana podem produzir estimativas aproximadas de a taxa de falsos positivos.

Para ajudar a evitar resultados enganosos, você deve considerar o valor exato do p-valor. Usando a abordagem binária de uma determinação sim ou não de significância estatística é muito simplista.

O estudo da vacina contra a AIDS tem um valor de p de 0,039. Com base nas informações acima, devemos ser cautelosos com esse resultado.

### **Dica 2: A replicação é crucial**

Na dica anterior, referi-me aos resultados de um único estudo. Realisticamente, você precisa ~~de replicá-las~~ resultados estatisticamente significativos várias vezes antes de poder ter confiança nas conclusões.  
~~social@faria@gmail.com  
021.399.242-66~~

No ambiente de alta pressão no meio corporativo para obter p-valores, um único p-valor é frequentemente considerado conclusivo. No entanto, Ronald Fisher, um grande matemático do século passado, desenvolveu p-valores com a noção de que eles são apenas uma parte do processo científico que inclui experimentação, análise e replicação.

"Um fato científico deve ser considerado como estabelecido experimentalmente somente se um experimento adequadamente projetado raramente deixa de fornecer esse nível de significância." –Ronald Fisher

O ideal é que façamos experimentação repetida com resultados consistentemente significativos para ter certeza de que a hipótese alternativa está correta.

Para o estudo da vacina contra a AIDS, o experimento tailandês é o primeiro estudo de vacinação contra a AIDS a produzir resultados estatisticamente significativos. Outros pesquisadores não conseguiram o replicar, então precisamos ser cautelosos com resultados. Esta vacina não construiu um histórico de resultados significativos.

### **Dica 3: o tamanho do efeito é importante**

A alta pressão para obter p-valores estatisticamente significativos desvia a atenção tanto do tamanho do efeito quanto da precisão da estimativa.

Você pode ter resultados de teste estatisticamente significativos mesmo quando os tamanhos dos efeitos são muito pequenos para serem significativos na prática. Além disso, um p-valor significativo não indica necessariamente que a análise possa estimar o tamanho do efeito com alta precisão.

Para dar mais ênfase ao tamanho e à precisão do efeito, use intervalos de confiança.

Considere se o tamanho do efeito é grande suficiente para ser importante na prática.

Infelizmente, o intervalo de confiança para a eficácia do estudo da vacina contra AIDS se estende de 1% a 52%. A vacina pode funcionar quase nenhuma vez até a metade das vezes. O intervalo de confiança revela que o tamanho do efeito estimado é pequeno e impreciso.

### **Dica 4: A plausibilidade da hipótese alternativa é importante**

À medida que avaliamos os p-valores em testes de hipóteses, há uma tendência achar que p-valores semelhantes em todos os estudos dão suporte comparável para a hipótese alternativa. Por exemplo, um p-valor de 0,04 em um estudo parece fornecer a mesma evidência que um valor de p de 0,04 em outro estudo.

No entanto, estudos de simulação mostram que a plausibilidade da hipótese alternativa do estudo afeta consideravelmente o falso positivo

Por exemplo, com um p-valor de 0,05, uma hipótese alternativa altamente plausível está associada a uma taxa de falsos positivos de pelo menos 12%. Em comparação, uma alternativa implausível tem uma taxa de pelo menos 76%!

Se você está estudando uma hipótese alternativa improvável e você obtém um p-valor significativo, há uma probabilidade maior de que a hipótese alternativa não esteja correta.

Um p-valor significativo não nos absolve de usar nosso sentido ao interpretar os resultados. Se você ouvir falar de um estudo surpreendente que produz resultados maravilhosos, pode ser interessante esperar até que os outros estudos repliquem antes de confiar nele!

Nenhum estudo de outras vacinas contra a AIDS forneceu evidências suficientes para rejeitar a hipótese nula. Este padrão demonstra que é improvável que a hipótese alternativa esteja correta para o estudo tailandês. Nesse cenário, podemos esperar taxas de falso-positivos em torno de 75%!

### **Dica 5: Use sua experiência**

Você deve aplicar seu conhecimento da área de assunto a todas as faces do teste de hipóteses para evitar resultados enganosos. Os pesquisadores e analistas devem usar sua expertise para avaliar a validade do desenho experimental, mecanismos propostos por trás do efeito, significado prático do efeito, a plausibilidade da hipótese alternativa e assim por diante.

### **Avaliando os resultados do teste de hipóteses para a vacina contra a AIDS**

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242-66

Vimos o seguinte:

- O valor-p de 0,039 não é uma evidência convincente por si só.
- A vacina não tem um histórico comprovado de resultados.
- O intervalo de confiança indica que o efeito estimado é pequeno e impreciso.
- Estudos de outras vacinas contra a Aids não tiveram resultados significativos, sugerindo que a hipótese alternativa na Tailândia é improvável.

Tomando todos esses pontos juntos, as considerações adicionais devem nos tornar cautelosos sobre resultados potencialmente enganosos. Em outras palavras, não devemos abrir uma garrafa de champanhe e começar a produzir a vacina em massa ainda. Precisamos esperar e ver se outros estudos replicaram esses resultados. Também precisamos ficar de olho no efeito tamanho em estudos futuros para determinar se a eficácia da vacina é significativa na prática.



Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

## 21. Testes não-paramétricos



Até agora examinamos testes de hipóteses paramétricos. Esse tipo de teste usa distribuições de amostragem para calcular probabilidades e determinar significância estatística, como as distribuições t, F, binomial e qui-quadrado. A obtenção de resultados válidos para esses testes pode depender se seus dados seguem uma distribuição específica, como a distribuição normal. Embora, vimos como você pode renunciar a essa suposição em alguns casos, especialmente considerando o teorema do limite central.

Existe outro tipo de teste de hipótese que não assume que seus dados seguem uma distribuição específica: os não paramétricos. Embora esses testes não exijam que seus dados sigam uma distribuição específica, existem outras suposições.

Os testes não paramétricos são um universo paralelo aos testes paramétricos. Na tabela abaixo, temos alguns de/para de testes paramétricos - não paramétricos.

Oseias Dias de Farias Testes Estatísticos oseias.dfarias@gmail.com			
Paramétricos		021.399.242-66	
Independentes	Vinculados	Independentes	Vinculados
2 amostras	2 amostras	2 amostras	2 amostras
Teste <i>t</i> (Student)	Teste <i>t</i> (Student)	Mann-Whitney T. da Mediana $\chi^2$ (2 x 2) Proporções Exato (Fisher)	Wilcoxon T. dos sinais Mac Nemar Binomial
Análise de variância	Análise de variância	Mais de duas Kruskal-Wallis Mediana (m x n) $\chi^2$ (m x n) Nemenyi	Mais de duas Cochran Friedman



\*Vinculados = Dependentes

Muito se discute sobre as vantagens e desvantagens dos testes paramétricos em contraste com os não paramétricos. Vamos ver algumas delas.

## VANTAGENS DOS TESTES PARAMÉTRICOS

- 1) Testes paramétricos podem fornecer resultados confiáveis com distribuições distorcidas e não normais

Muitas pessoas não estão cientes desse fato, mas análises paramétricas podem produzir resultados confiáveis mesmo quando seus dados contínuos são distribuídos de forma não normal. Você só precisa ter certeza de que o tamanho da sua amostra atende aos requisitos para cada análise.

- 2) Testes paramétricos têm maior poder estatístico

Na maioria dos casos, os testes paramétricos têm mais poder. Se um efeito realmente existe, uma análise paramétrica tem maior probabilidade de detectá-lo.

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242-66

## VANTAGENS DOS TESTES NÃO PARAMÉTRICOS

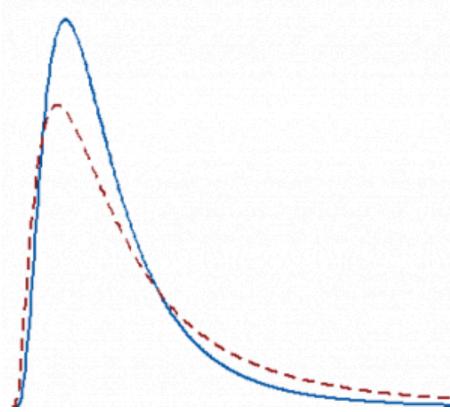
- 1) Testes não paramétricos avaliam a mediana que pode ser melhor para algumas áreas de estudo

Agora estamos chegando ao motivo mais interessante para quando usar um teste não paramétrico.

Para alguns conjuntos de dados, as análises não paramétricas fornecem uma vantagem porque **avaliam a mediana em vez da média**. A média nem sempre é a melhor medida de tendência central para uma amostra (lembrem-se da seção de estatística descritiva). Mesmo que você pode realizar uma análise paramétrica válida em dados distorcidos, isso não significa necessariamente ser o melhor método. Vamos dar um exemplo usando a distribuição de salários.

Os salários tendem a ser uma distribuição assimétrica à direita. A maioria dos salários se agrupa em torno da mediana, que é o ponto em que metade está

acima e metade está abaixo. No entanto, há uma longa cauda que se estende até as faixas salariais mais altas. Essa cauda longa afasta a média do valor mediano central:



Essas duas distribuições têm medianas aproximadamente iguais, mas médias diferentes.

#### Oseias Dias de Farias

Nessas distribuições, se vários indivíduos de renda muito alta se juntarem à amostra, a média aumenta significativamente, embora a renda da maioria das pessoas não mude. Eles ainda se aglomeram ao redor da mediana.

Nesta situação, os resultados dos testes paramétricos e não paramétricos podem fornecer resultados diferentes, e ambos podem estar corretos! Para as duas distribuições, se você extrair uma grande amostra aleatória de cada população, a diferença entre as médias é estatisticamente significativa. Apesar disso, a diferença entre as medianas não é estatisticamente significativa.

Para distribuições assimétricas, mudanças na cauda afetam substancialmente a média. Testes paramétricos podem detectar essa mudança média. Por outro lado, a mediana é relativamente inalterada e uma análise não paramétrica pode indicar legitimamente que a mediana não mudou significativamente.

Você precisa decidir se a média ou mediana é a melhor para o seu estudo e qual tipo de diferença é mais importante detectar.

- 2) Testes não paramétricos são válidos quando nosso tamanho de amostra é pequeno e seus dados são potencialmente não normais

Use um teste não paramétrico quando o tamanho da amostra não for grande o suficiente para atender aos requisitos da tabela acima e você não tiver certeza de que seus dados seguem a distribuição normal. Com tamanhos de amostra pequenos, esteja ciente de que os testes de normalidade podem ter poder insuficiente para produzir resultados úteis.

Esta situação é difícil. Análises não paramétricas tendem a ter um poder menor no início, e um tamanho de amostra pequeno só agrava esse problema.

- 3) Testes não paramétricos podem analisar dados ordinais e valores discrepantes

Os testes paramétricos podem analisar apenas dados contínuos e as descobertas podem ser excessivamente afetadas por discrepâncias. Por outro lado, os testes não paramétricos também podem analisar dados ordinais e não serem enganados por outliers. Às vezes, você pode remover de forma legítima os valores discrepantes do seu conjunto de dados se eles representarem condições incomuns (erros). No entanto, às vezes os valores discrepantes são uma parte genuína da distribuição de uma área de estudo e você não deve removê-los.

Você deve verificar as suposições para análises não paramétricas porque os vários testes podem analisar diferentes tipos de dados e têm diferentes habilidades para lidar com valores discrepantes.

É importante ressaltar que nada do que vimos anteriormente vai mudar. Ainda escrevemos  $H_0$  e  $H_a$  da mesma forma que víhamos escrevendo, e rejeitamos  $H_0$  se p-valor for melhor do que o nível de significância.

Por enquanto vamos abordar os testes Wilcoxon, Mann-Whitney, Kruskal-Wallis e Kolmogorov-Smirnov.

## KRUSKAL-WALLIS



O teste de Kruskal-Wallis (teste H) é um teste de hipótese **para mais de 2 amostras independentes, que é usado quando as suposições para uma análise de variância de um fator não são atendidas (ANOVA 1 fator)**. Como o teste de Kruskal-Wallis é um teste não paramétrico, os dados usados não precisam ser distribuídos normalmente, ao contrário da análise de variância. O único requisito é que os dados estejam em escala ordinal ou sejam contínuos.

O teste de Kruskal-Wallis não trabalha com as hipóteses de comparação dos parâmetros, não testa a hipótese de igualdade de médias e nem testa a igualdade de medianas, como muitos acreditam. O teste de Kruskal-Wallis é indicado para testar a hipótese de que três ou mais populações têm distribuição igual ou não.

A rigor, o teste de Kruskal-Wallis avalia diferenças de médias de ordens (postos), as quais não são necessariamente iguais às medianas dos grupos, uma vez que é atribuído a cada observação seu posto (ordem). Deve-se ordenar as observações em ordem crescente, independente dos grupos, e em seguida fazer a distribuição dos postos, onde o menor valor recebe o posto ou ordem 1, o segundo 2 e assim sucessivamente, até que todas as observações tenham sido consideradas.

Sua fórmula é:

$$H = \left[ \frac{12}{(N.(N+1))} \right] \cdot \left[ \frac{\sum R_1^2}{n_1} + \frac{\sum R_2^2}{n_2} + \frac{\sum R_3^2}{n_3} \right] - 3 \cdot (N + 1)$$

Onde: N é o número dados em todos os grupos  
n é o número de sujeitos em cada grupo  
 $\Sigma R$  é a somatória dos postos em cada grupo

Quando há observações repetidas, é indicado que seja atribuído o valor médio dos postos entre estas observações. Por isso, pode ocorrer do resultado

do teste apontar para a diferença significativa entre os grupos, mas as medianas serem iguais ou relativamente próximas. Quando isso ocorre, o teste de Kruskal-Wallis testa, em simultâneo, a mediana e as formas de distribuição.

Quando os grupos apresentam a mesma forma de distribuição de probabilidade, o resultado do teste de Kruskal-Wallis pode ser interpretado com base na mediana.

Vamos a um exemplo para clarear as ideias. A tabela abaixo mostra dados de 3 grupos de sujeitos relativos ao número de vezes que os mesmos realizam algum tipo de compra no shopping durante um mês. • Os grupos apresentam a mesma distribuição?

G1	Oseias D. de Farias oseias.dfarias3@gmail.com 021.399.242-66	G3
20	12	8
4	21	22
7	9	10
2	0	5
17	14	6
3	1	20

Resolvendo:

- 1) Fazer a ordenação dos postos (OP) de 1 a 18 de todos os dados (temos 18 dados acima)
- 2) Quando existir empates divide o posto pelo número de empates para continuar a ordenação.
- 3) Calcule a soma dos postos para cada grupo.

G1	OP	G2	OP	G3	OP
20	15,5º	12	12º	8	9º
4	5º	21	17º	22	18º
7	8º	9	10º	10	11º
2	3º	0	1º	5	6º
17	14º	14	13º	6	7º
3	4º	1	2º	20	15,5º

$$\sum 49,5 \quad \sum 55,0 \quad \sum 66,5$$

Oseias D'Farias

[oseias.dfaras@gmail.com](mailto:oseias.dfaras@gmail.com)

021.399.242-66

- Aplicando na fórmula:

$$H = \left[ \frac{12}{(18 \cdot (18 + 1))} \right] \cdot \left[ \frac{(49,5)^2}{6} + \frac{(55,0)^2}{6} + \frac{(66,5)^2}{6} \right] - 3 \cdot (18 + 1)$$

$$\bullet H = \left( \frac{12}{342} \right) \cdot \left( \frac{9897,5}{6} \right) - 3 \cdot (19)$$

$$H = 0,0350 \cdot 1649,5 - 57 = 57,81 - 57 = 0,73$$

Da mesma forma que todos os outros valores (t, z, etc) são tabelados, o valor de H também é. Nesse caso, p-valor é 0.64 e, portanto, não rejeitamos a hipótese nula a 95% de confiança. Logo, não existem diferenças estatisticamente significativas entre os grupos.



Não se preocupem, calcularemos o p-valor e a estatística de teste usando Python e Excel.

## WILCOXON

O teste de **postos sinalizados de Wilcoxon** é um teste estatístico não paramétrico que compara **dois grupos pareados (dependentes)**. Da mesma forma que o anterior, ele também calcula a diferença de postos entre conjuntos de pares e analisa essas diferenças para estabelecer se elas são estatisticamente significativas ou não.

Hipótese nula: Não há diferença entre as duas populações

Hipótese alternativa: Existe uma diferença nas populações.

Vocês podem calcular o teste usando tanto Python quanto Excel. Caso haja interesse, as fórmulas por trás desse teste podem ser vistas aqui: <https://datatab.net/tutorial/wilcoxon-test>

## MANN-WHITNEY

O teste U de Mann-Whitney é a contrapartida não paramétrica do **teste t para amostras independentes**; e, por ser não paramétrico, está sujeito a suposições menos rigorosas do que o teste t. Portanto, o teste U de Mann-Whitney é sempre utilizado quando o requisito de distribuição normal para o teste t não é atendido.

Para poder calcular um teste U de Mann-Whitney, apenas duas amostras aleatórias independentes com características pelo menos em escala ordinal devem estar disponíveis. As variáveis não precisam satisfazer nenhuma curva de distribuição. E da mesma forma que os anteriores, ele também é baseado na ideia de posto.

Hipótese nula: Não há diferença entre as duas populações



Hipótese alternativa: Existe uma diferença nas populações.

Vocês podem calcular o teste usando tanto Python quanto Excel. Caso haja interesse, as fórmulas por trás desse teste podem ser vistas aqui: <https://datatab.net/tutorial/mann-whitney-u-test>

## KOLMOGOROV-SMIRNOV (KS)

Diferente dos demais, o KS não tem o objetivo de comparar valores amostrais. A ideia aqui é aplicamos o teste de aderência de Kolmogorov-Smirnov para verificar se determinada amostra vem de população com distribuição específica. Essa “distribuição específica” é, na maioria das vezes, a distribuição normal. Nesses casos, podemos dizer que estamos usando o teste de normalidade de Kolmogorov-Smirnov.

Geralmente realizamos um teste de normalidade porque alguns testes de hipóteses pressupõem que seus dados seguem uma distribuição normal. Embora, como você aprendeu, você pode dispensar a suposição de normalidade quando o tamanho da amostra for grande o suficiente graças ao teorema do limite central.

A hipótese nula do KS é que a amostra segue a mesma distribuição que é normal. A hipótese alternativa diz que as duas distribuições são diferentes. Portanto, se queremos confirmar a normalidade de uma variável, precisamos que o valor de p seja maior que 0,05 (ou seja qual for o valor de significância estabelecido).

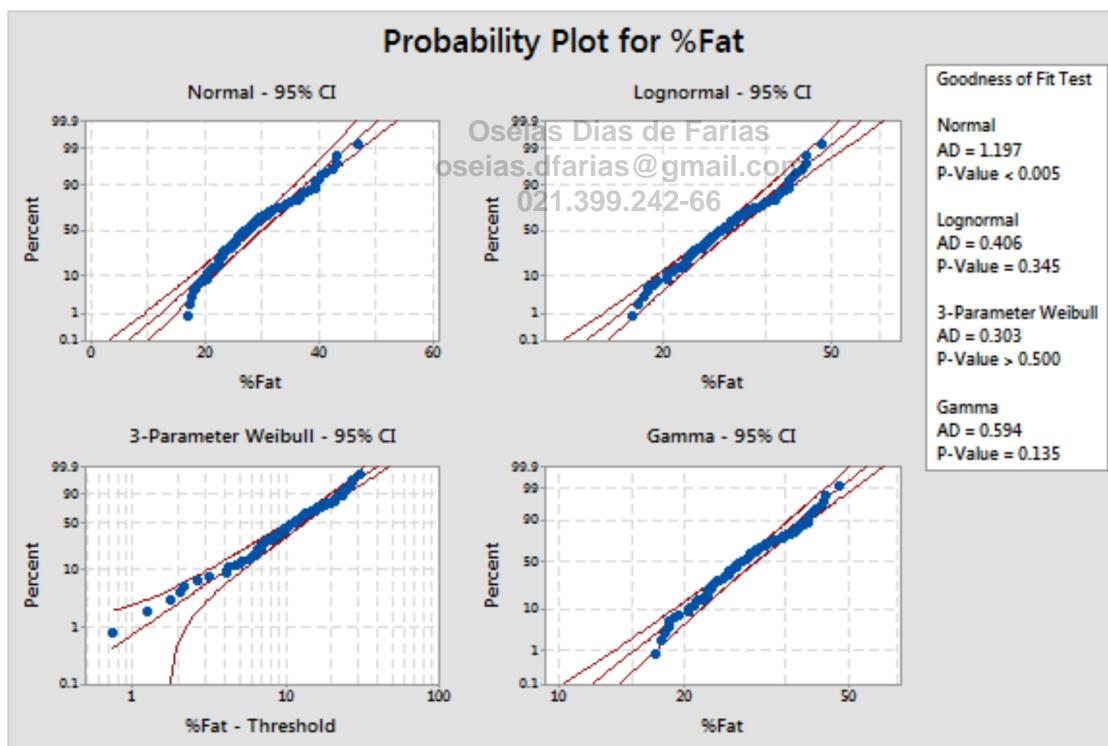
Os gráficos de probabilidade podem ser a melhor maneira de determinar se seus dados seguem uma distribuição específica. Se seus dados seguirem a linha reta no gráfico, a distribuição se ajustará aos seus dados. Este processo é simples de fazer visualmente. Informalmente, esse processo é chamado de teste do “fat pencil”. Se todos os pontos de dados se alinharem dentro da área de um lápis colocado sobre a linha reta central, você pode concluir que seus dados seguem a distribuição.

Esses gráficos são especialmente úteis nos casos em que os testes de distribuição são muito poderosos. Os testes de distribuição são como outros

testes de hipóteses. À medida que o tamanho da amostra aumenta, o poder estatístico do teste também aumenta. Com tamanhos de amostra muito grandes, o teste pode ter tanto poder que desvios triviais da distribuição produzem resultados estatisticamente significativos. Nesses casos, seu valor-p será menor que o nível de significância, mesmo quando seus dados seguirem a distribuição.

A solução é avaliar os gráficos de probabilidade para identificar a distribuição de seus dados. Se os pontos de dados caírem ao longo da linha reta, você pode concluir que os dados seguem essa distribuição mesmo que o valor-p seja estatisticamente significativo.

No gráfico abaixo, vocês podem ver os pontos em azul e um teste de ajuste em 4 tipos de curva: Normal, Lognormal, Wibull e Gama.

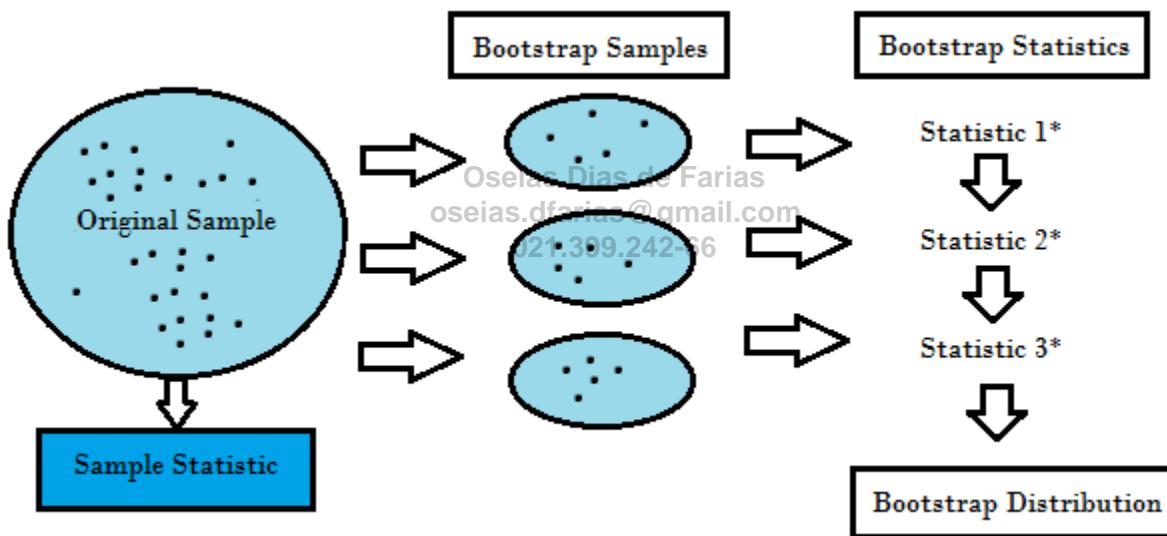


Os pontos de dados para a distribuição normal não seguem a linha central. No entanto, os pontos de dados seguem a linha muito de perto para as distribuições lognormal e Weibull de três parâmetros. A distribuição gama

não segue a linha central tão bem quanto as outras duas, e seu valor p é menor.

## BOOTSTRAP

Bootstrapping é um procedimento estatístico que reamostra um único conjunto de dados para criar muitas amostras "simuladas". Esse processo permite calcular erros padrão, construir intervalos de confiança e realizar testes de hipóteses para vários tipos de estatísticas de amostra. Métodos de bootstrap são abordagens alternativas para testes de hipóteses paramétricas e são conhecidos por serem mais fáceis de entender e válidos para mais condições.



Ambos os métodos paramétricos e não-paramétricos usam amostras para fazer inferências sobre populações. Para atingir esse objetivo, esses procedimentos tratam a amostra única que um estudo obtém como apenas uma das muitas amostras aleatórias que o estudo poderia ter coletado.

A partir de uma única amostra, você pode calcular uma variedade de estatísticas de amostra, como média, mediana e desvio padrão, mas vamos nos concentrar na média aqui.

Suponha que a gente repita seu estudo muitas vezes. Nessa situação, a média varia de amostra para amostra e forma uma distribuição das médias

amostrais. Os estatísticos referem-se a este tipo de distribuição como distribuição amostral. Como você viu ao longo deste e-book, as distribuições de amostragem são cruciais porque colocam o valor de sua estatística de amostra no contexto mais amplo de muitos outros valores possíveis.

Embora realizar um estudo muitas vezes seja inviável, tanto testes paramétricos/não paramétricos quanto métodos de bootstrapping podem estimar distribuições amostrais. Usando o contexto maior que as distribuições de amostragem fornecem, esses procedimentos podem construir intervalos de confiança e realizar testes de hipóteses.

## DIFERENÇAS ENTRE BOOTSTRAPPING E TESTE DE HIPÓTESE PARAMÉTRICA/NÃO PARAMÉTRICO

A principal diferença entre bootstrapping e estatísticas tradicionais é como elas estimam as distribuições de amostragem.

Os procedimentos de teste de hipóteses paramétricas requerem equações para distribuições de probabilidade que estimam distribuições de amostragem usando as propriedades dos dados da amostra, o desenho experimental e uma estatística de teste. Para obter resultados válidos, você precisará usar a estatística de teste adequada e satisfazer as suposições.

O método bootstrap usa uma abordagem totalmente diferente para estimar as distribuições de amostragem. Esse método pega os dados de amostra que um estudo obtém e os **reamostra** repetidamente para criar muitas amostras simuladas - ou seja, pega essa amostra maior e quebra em várias amostras aleatórias menores. Cada uma dessas amostras simuladas tem suas próprias propriedades, como a média. Ao representar graficamente a distribuição dessas médias em um histograma, você pode observar a distribuição amostral da média. Você não precisa se preocupar com estatísticas de teste, fórmulas e suposições.

O procedimento bootstrap usa essas distribuições de amostragem como base para intervalos de confiança e testes de hipóteses. Vejamos como funciona esse processo de reamostragem.

## CÓMO O BOOTSTRAPPING REAMOSTRA SEUS DADOS PARA CRIAR CONJUNTOS DE DADOS SIMULADOS

Bootstrapping reamostra o conjunto de dados original com substituição muitas de vezes para criar conjuntos de dados simulados. Esse processo envolve o desenho de amostras aleatórias do conjunto de dados original. Veja como funciona:

- 1) O método bootstrap tem uma probabilidade igual de desenhar aleatoriamente cada ponto de dados original para inclusão nos conjuntos de dados reamostrados.
- 2) O procedimento pode selecionar um ponto de dados mais de uma vez para um conjunto de dados reamostrado. Esta propriedade é o aspecto “com reposição” do processo.
- 3) O procedimento cria conjuntos de dados reamostrados com o mesmo tamanho do conjunto de dados original.

O processo termina com seus conjuntos de dados simulados com muitas combinações diferentes dos valores que existem no conjunto de dados original. Cada conjunto de dados simulado tem seu próprio conjunto de estatísticas de amostra, como a média, mediana e desvio padrão. Os procedimentos de bootstrapping usam a distribuição das estatísticas de amostra entre as amostras simuladas como distribuição de amostragem.

## EXEMPLO DE AMOSTRAS DE BOOTSTRAP

Vamos trabalhar com um caso fácil. Suponha que um estudo colete cinco pontos de dados e crie quatro amostras bootstrap, conforme mostrado abaixo.

Original	Bootstrap1	Bootstrap2	Bootstrap3	Bootstrap4
1	1	2	1	1
2	1	3	2	1
3	3	3	3	1
4	3	3	5	4
5	5	4	5	5

Este exemplo simples ilustra as propriedades de amostras de bootstrap. Os conjuntos de dados reamostrados têm o **mesmo tamanho do conjunto de dados original** e **contêm apenas valores que existem no conjunto original**. Além disso, esses **valores podem aparecer com mais ou menos frequência**.

**nos conjuntos de dados reamostrados do que no conjunto de dados original.** Por fim, o processo de reamostragem é **aleatório** e poderia ter criado um conjunto diferente de conjuntos de dados simulados.

Claro, em um estudo real, você esperaria ter um tamanho de amostra maior e criaria milhares de conjuntos de dados reamostrados. Dado o enorme número de conjuntos de dados reamostrados, você sempre usará um computador para realizar essas análises.

## QUÃO BEM O BOOTSTRAPPING FUNCIONA?

A reamostragem envolve a reutilização de seu conjunto de dados várias vezes. Quase parece bom demais para ser verdade! No entanto, usar o poder dos computadores para reamostrar aleatoriamente seu único conjunto de dados para criar milhares de conjuntos de dados simulados produz resultados significativos.

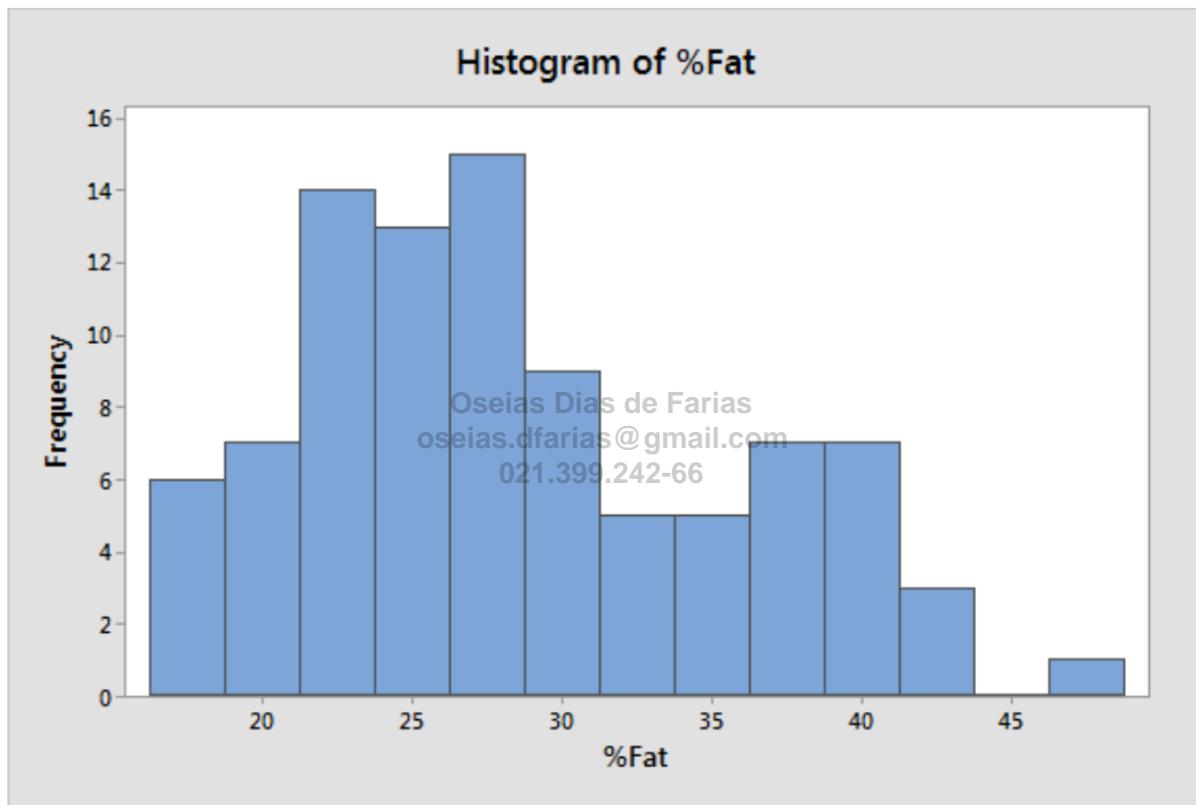
O método bootstrap existe desde 1979 e seu uso aumentou. Vários estudos ao longo das décadas seguintes determinaram que as distribuições de amostragem bootstrap se aproximam das distribuições de amostragem corretas.

Para entender como funciona, lembre-se de que o bootstrap não cria novos dados. Em vez disso, trata a amostra original como uma *proxy* para a população real e, em seguida, extrai amostras aleatórias dela. Consequentemente, a suposição central para bootstrapping é que a amostra original representa com precisão a população real.

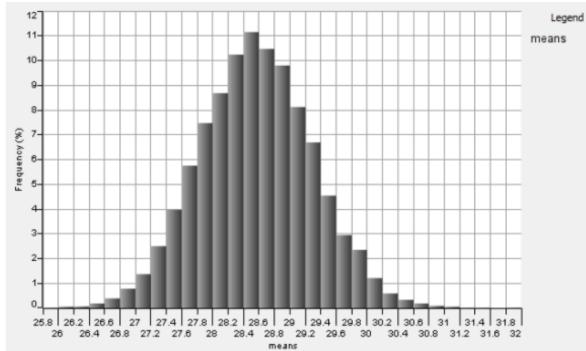
O processo de reamostragem cria muitas amostras possíveis que um estudo poderia ter desenhado. As várias combinações de valores nas amostras simuladas fornecem coletivamente uma estimativa da variabilidade entre amostras aleatórias retiradas da mesma população. O alcance dessas amostras potenciais permite que o procedimento construa intervalos de confiança e realize testes de hipóteses. É importante ressaltar que, à medida que o tamanho da amostra aumenta, o bootstrap converge para a distribuição amostral correta na maioria das condições.

## EXEMPLO DE USO DE BOOTSTRAPPING PARA CRIAR INTERVALOS DE CONFIANÇA

Para este exemplo, usarei bootstrap para construir um intervalo de confiança para um conjunto de dados que contém as porcentagens de gordura corporal de 92 meninas adolescentes. Esse dados não seguem a distribuição normal. Por não atender à suposição de normalidade das estatísticas tradicionais, é um bom candidato para bootstrapping. No entanto, o grande tamanho da amostra nos permite contornar essa suposição. O histograma abaixo mostra a distribuição dos dados da amostra original.



Usando o Excel ou Python, podemos pegar o conjunto de dados original e o reamostrar com substituição quantas vezes forem necessárias. Suponhamos que fizemos 1000 reamostragens. Este processo produz 1000 amostras bootstrap com 92 observações em cada. O programa calcula a média de cada amostra e traça a distribuição dessas 1000 médias no histograma abaixo. Os estatísticos referem-se a esse tipo de distribuição como distribuição amostral das médias. Os métodos de bootstrapping criam essas distribuições usando reamostragem, enquanto os métodos tradicionais usam equações para distribuições de probabilidade.



Para criar o intervalo de confiança bootstrap, simplesmente usamos percentis. Para um intervalo de confiança de 95%, precisamos identificar os 95% centrais da distribuição. Para fazer isso, use o percentil 97,5 e o percentil 2,5 ( $97,5 - 2,5 = 95$ ). Em outras palavras, se ordenarmos todas as médias da amostra de baixo para alto e, em seguida, cortar os 2,5% mais baixos e os 2,5% mais altos das médias, os 95% médios permanecem.

No nosso caso, o intervalo de confiança a 95% da média é [27,16 30,01]. Podemos ter 95% de confiança de que a média populacional está dentro dessa faixa.  
 Oseias Dias de Farias  
 Oseias.diasdefarias@gmail.com  
 021.399.242-66

Esse intervalo tem a mesma largura que o intervalo de confiança tradicional para esses dados e difere apenas em alguns pontos percentuais. Os dois métodos são muito próximos.

Observe como a distribuição de amostragem no histograma se aproxima de uma distribuição normal, mesmo que a distribuição de dados subjacente seja distorcida. Esta aproximação ocorre graças ao teorema do limite central. À medida que o tamanho da amostra aumenta, a distribuição amostral converge para uma distribuição normal, independentemente da distribuição de dados subjacente (com algumas exceções).

## BENEFÍCIOS DO BOOTSTRAPPING SOBRE AS ESTATÍSTICAS TRADICIONAIS

Esse processo é muito mais fácil de compreender do que as equações necessárias para distribuições de probabilidade que os métodos paramétricos usam. No entanto, bootstrapping oferece mais benefícios do que apenas ser fácil de entender!

Bootstrapping não faz suposições sobre a distribuição de seus dados. Você simplesmente reamostra seus dados e usa qualquer distribuição de amostragem que surja. Então, você trabalha com essa distribuição, seja ela qual for, como fizemos no exemplo.

Por outro lado, os métodos tradicionais geralmente assumem que os dados seguem a distribuição normal ou alguma outra distribuição. Para a distribuição normal, o teorema do limite central pode permitir que você ignore essa suposição quando tiver um tamanho de amostra grande o suficiente. Consequentemente, você pode usar bootstrap para uma variedade maior de distribuições, distribuições desconhecidas e tamanhos de amostra menores. Tamanhos de amostra tão pequenos quanto 10 podem ser usados.

Nesse sentido, todos os métodos tradicionais utilizam equações que estimam a distribuição amostral para uma estatística amostral específica quando os dados seguem uma distribuição particular. Infelizmente, não existem fórmulas para todas as combinações de estatísticas amostrais e distribuições de dados! Por exemplo, não há distribuição amostral conhecida de medianas para algumas distribuições, o que torna o bootstrap a análise perfeita para isso. Outras análises têm pressupostos como igualdade de variâncias. No entanto, nenhuma dessas questões são problemas para bootstrapping.

## **PARA QUAIS ESTATÍSTICAS DE AMOSTRA POSSO USAR O BOOTSTRAPPING?**

Embora esta visão geral se concentre na média da amostra, o método bootstrap pode analisar uma ampla gama de estatísticas e propriedades da amostra. Essas estatísticas incluem a média, mediana, moda, desvio padrão, análise de variância, correlações, coeficientes de regressão, proporções, variância em dados binários e estatísticas multivariadas entre outros.

Nos links abaixo você poderá ver como fazer um bootstrap em Excel e Python.

Python:

<https://www.statology.org/bootstrapping-in-python/#:~:text=Bootstrapping%20is%20a%20method%20that,replacement%20from%20a%20given%20dataset.>

Excel: <https://www.statology.org/bootstrapping-in-excel/>

# 22. Teste AB | Desenhando um experimento



Experimentos são usados para estudar relações causais. Você manipula uma ou mais variáveis independentes e mede seu efeito em uma ou mais variáveis dependentes. Design experimental significa criar um conjunto de procedimentos para testar uma hipótese.<sup>21,22,23,24,25,26</sup> Um bom projeto experimental requer uma forte compreensão do sistema que você está estudando. Por exemplo, podemos pensar nos seguintes objetos de estudo:

## Política

- Qual candidato tem maior chance de vencer?

## Marketing

- Aplicando diferentes campanhas promocionais, será que aumentaremos nossa venda?

## Fraude

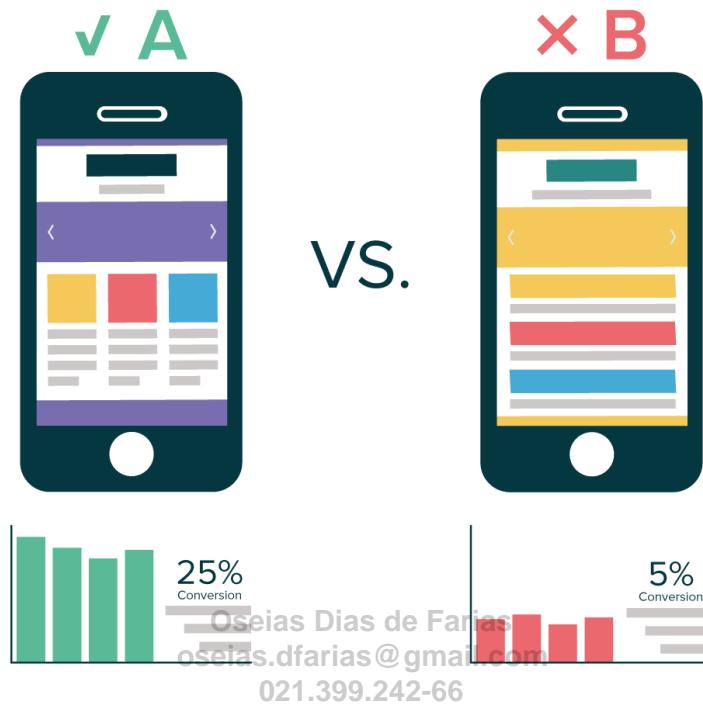
- Aplicando um modelo para detecção, será que reduziremos nossas fraudes em sistema?

## UX design

- Um site com layout diferente pode tornar a experiência do usuário melhor e aumentar nossa métrica de engajamento?

Até agora vocês viram como comprovar uma hipótese dado que já temos as amostras e resultados de cada população. Mas e se nosso intuito é

exatamente desenhar um teste para pegar o resultado de cada amostra? Como devemos fazer? Por onde começar?



Para chegarmos nesse ponto, precisamos passar por algumas etapas:

1	2	3	4
<p><b>Escolha sua métrica e meça seu processo atual</b></p> <p>Qual é sua métrica de sucesso? Até onde você quer chegar? Como está o estado atual dessa métrica? Essa etapa envolve um grande conhecimento de negócios e análises exploratórias</p>	<p><b>Pense em quais serão suas amostras e qual será seu objetivo no estudo</b></p> <p>Como suas amostras precisam ser? Qual tipo de estudo vai te responder o que você precisa? Cuidado com os vieses!</p>	<p><b>Calcule o tamanho mínimo da amostra</b></p> <p>Sabendo como está a métrica atual e onde quer chegar, podemos usar alguns cálculos para encontrar o tamanho mínimo de amostra necessária para medir a diferença que se objetiva</p>	<p><b>Realize seu teste e compare os resultados</b></p> <p>Com os resultados em mãos, chegou a hora de usar todo o ferramental estatístico visto até agora para comparar os resultados e ver se há mudanças estatisticamente significativas</p>

## PASSO 1: ESCOLHA SUA MÉTRICA E MEÇA SEU PROCESSO ATUAL

KPI vem da sigla em inglês para Key Performance Indicator, ou seja, Indicador-chave de Performance. É uma forma de medir se uma ação ou um

conjunto de iniciativas está efetivamente atendendo aos objetivos propostos pela organização.

Existem milhares de indicadores que podem ser medidos. Estamos em uma época em que o fluxo de informação é imenso e constante! O ponto central é saber escolher quais são os indicadores a serem medidos.

A depender do contexto e do que você almeja atingir, seu KPI vai mudar. Por exemplo, um indicador chave para testes de vacina é a eficácia da vacina (% de infectados dado todos do grupo). Para uma melhoria de um site de e-commerce visando maior conversão, seu indicador é justamente a conversão de clientes (% pessoas que compraram no e-commerce de todas que entraram no site).

Deixarei aqui 2 sugestões de livros para quem quer se aprofundar no assunto

- Key Performance Indicators (Kpi): The 75 Measures Every Manager Needs to Know - Bernard Marr
- Indicadores de desempenho - Andresa S. N. Francischini

Outro passo importante aqui é sua meta - onde você quer chegar. Esse passo é importante para coletar o tamanho mínimo da amostra, então discutiremos isso quando chegarmos no passo 3.

## PASSO 2: PENSE EM QUAIS SERÃO SUAS AMOSTRAS E QUAL O OBJETIVO DO SEU ESTUDO

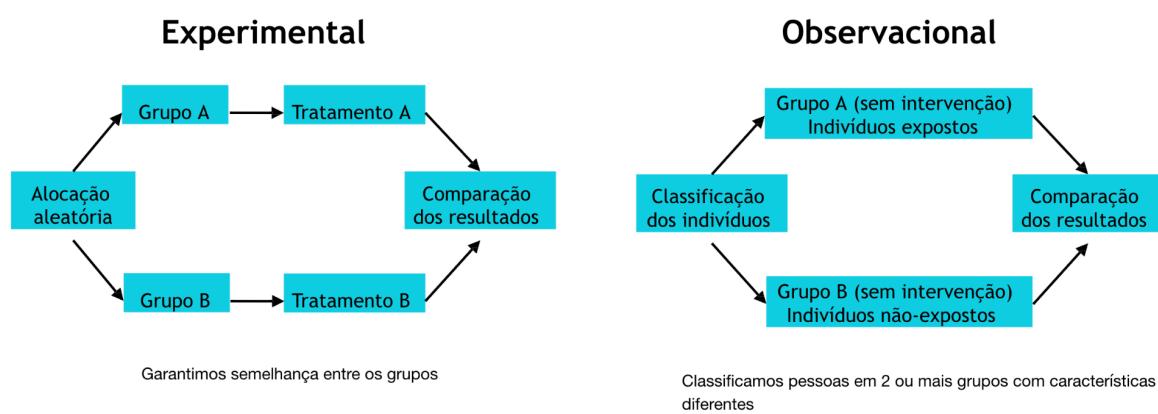
A escolha da amostra é um ponto crucial para que vieses sejam evitados. Falamos sobre isso na seção de "Conceitos Fundamentais", quando introduzimos os tipos de amostragem.

Além da amostragem, precisamos falar também sobre outro conceito: Estudo experimental x Estudo observacional. Como dissemos, fazemos estudos para reunir informações e tirar conclusões. O tipo de conclusão a que chegamos depende do método de estudo utilizado.

Em um **estudo observacional** (empírico), nós medimos ou entrevistamos membros de uma amostra sem tentar afetá-los. Esse tipo de estudo tem o

objetivo de analisar as associações entre variáveis. Não há intervenção do pesquisador sobre as amostras coletadas, pois os grupos já são pré-existentes. Por exemplo, um estudo considerou uma amostra aleatória de adultos e perguntou a eles sobre seus hábitos antes de dormir. Os dados demonstraram que as pessoas que bebem uma xícara de chá antes de dormir eram mais propensas a dormir mais cedo que aquelas que não bebiam chá. Este estudo foi uma pesquisa para ver se as pessoas bebiam chá ou não, e quando elas iam para a cama. As pessoas não foram aleatoriamente atribuídas aos grupos.

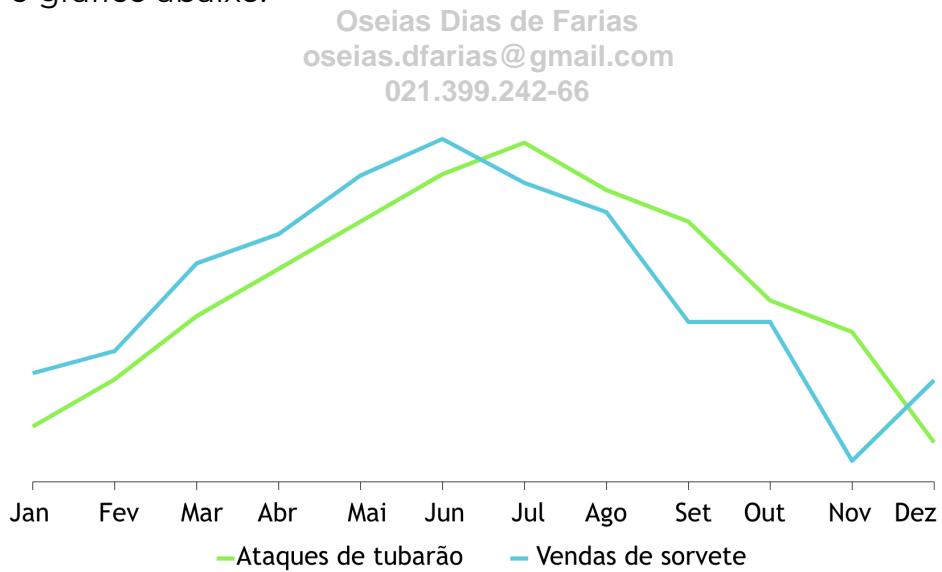
Em um **estudo experimental** (experimento controlado), atribuímos pessoas ou coisas a grupos e administrarmos algum tratamento a um dos grupos, enquanto o outro grupo não o recebe. Esse tipo de estudo tem o objetivo de analisar o efeito de uma variável sobre outra. Há intervenção intencional do pesquisador sobre as amostras coletadas. Por exemplo, outro estudo considerou um grupo de adultos e os dividiu, aleatoriamente, em dois grupos: um deles devia beber chá todas as noites por uma semana, enquanto o outro não devia beber chá naquela semana. Em seguida, os pesquisadores compararam os horários em que os membros de cada grupo adormeceram. Este estudo atribuiu as pessoas aleatoriamente aos grupos. Um grupo recebeu um tratamento e o outro grupo não - ou seja, houve intervenção do pesquisador.



Estudos experimentais são ótimos, pois conseguimos formar amostras aleatórias e fazer todos os testes necessários. Porém, muitas vezes o que temos em mão são apenas os estudos observacionais. Por exemplo,

imaginem que queremos estudar a influência do hábito de fumar no câncer de pulmão. Nós não podemos separar grupos perfeitamente comparáveis e dizer para o grupo A passar a fumar e para o grupo B não fumar. Aqui trabalhamos com questões de ética. Nesse caso, temos que usar as amostras que já estão à disposição - pessoas que fumam e pessoas que não fumam. Porém, quem nos garante que esse grupo é comparável? Quem nos garante que não existe um gene que torna a pessoa mais propensa a fumar e, que esse gene também deixa a pessoa mais propensa a ter câncer de pulmão? Quando trabalhamos com estudos observacionais, não conseguimos garantir necessariamente a semelhança entre nossas amostras.

Quando lidamos com estudos observacionais, precisamos ter um cuidado especial com vieses, em especial o que chamamos de viés de confundimento. Um confundidor é uma variável adicional, além da variável independente, que tem efeito sobre a variável dependente, fazendo com que seja inferida erroneamente uma associação entre as mesmas. Para falar sobre isso, vamos observar o gráfico abaixo.



Vemos acima uma clara **correlação** entre ataque de tubarão e venda de sorvete, porém temos que tomar muito cuidado com o que vamos concluir. Obviamente, o aumento de ataque de tubarão não está causando um aumento na venda de sorvete e vice-versa. A questão aqui é a variável escondida, chamada de confundidora. Nesse caso, essa variável é o verão nos

EUA. Nos meses mais quentes, as pessoas tomam mais sorvete, aumentando a venda. Nesses meses, as pessoas costumam ir mais a praias, se expondo mais aos ataques de tubarões. Não se esqueçam também que correlação não implica em causalidade.

Voltando ao nosso estudo observacional, é comum termos variáveis não mapeadas que podem causar o efeito que estamos vendo. E como controlar esse viés nesse tipo de estudo?

1. Controlar (incluir na análise) variáveis confundidoras conhecidas

- Temos que garantir que coletamos todas as variáveis confundidoras
- Iremos supor que o efeito de confundimento de variáveis não coletadas é mínimo

2. Matching: selecionar o grupo controle de acordo com características do grupo de exposição (ex.: idade, gênero, localização, etc.)

Desvantagem: pode levar a viés de seleção e precisamos saber todas as características que podem ser importantes

Oseias Dias de Farias  
oseias.dias@outlook.com  
021.399.242-66

Feito isso, vamos partir para um dos passos mais importantes: Calcular o tamanho mínimo da amostra

### PASSO 3: CALCULE O TAMANHO MÍNIMO DA AMOSTRA

Para esse passo, vamos relembrar rapidamente os tipos de erros que temos.

Os estatísticos definem dois tipos de erros no teste de hipóteses. De forma não tão criativa, eles chamam esses erros de erros Tipo I e Tipo II. Ambos os tipos de erros estão relacionados a conclusões incorretas sobre a hipótese nula. A tabela abaixo resume os quatro resultados possíveis para um teste de hipótese.



	<b>Rejeita H0</b>	<b>Não rejeita H0</b>
<b>H0 é verdadeiro (real)</b>	<b>Erro tipo I: Falso positivo (FP)</b>	<b>Acertamos \o/ Efeito não existe</b>
<b>H0 é falso (real)</b>	<b>Acertamos \o/ Efeito existe</b>	<b>Erro tipo II: Falso negativo (FN)</b>

Quando seu estudo faz tudo corretamente, o erro de amostragem é a única coisa que causa erros do Tipo I.

Em contrapartida, existem 3 motivos principais para acontecer o erro Tipo II - tamanhos de efeito pequenos, tamanhos de amostra pequenos e alta variabilidade de dados. Além disso, ao contrário dos erros do Tipo I, você não pode definir a taxa de erros do Tipo II para sua análise. Em vez disso, o melhor que você pode fazer é estimá-lo antes de iniciar seu estudo, aproximando as propriedades da hipótese alternativa que você está estudando. Quando você faz esse tipo de estimativa, é chamado de análise de poder (*Power Analysis*).

oséias.darias@gmail.com

Ao estimar a taxa de erro Tipo II, seu software estatístico cria uma distribuição de probabilidade hipotética representando as propriedades de uma hipótese alternativa verdadeira. No entanto, quando você está realizando um teste de hipótese, você normalmente não sabe qual hipótese é verdadeira, muito menos as propriedades específicas da distribuição para a hipótese alternativa. Consequentemente, a taxa real de erro do Tipo II é geralmente desconhecida!

Como sabem, beta é a probabilidade de um falso negativo. Portanto,  $1 - \beta$  é a probabilidade de detectar corretamente um efeito. Os estatísticos referem-se a este conceito como **poder estatístico**. Os analistas normalmente estimam o poder em vez do beta diretamente.

O **poder estatístico** é o oposto dos erros do Tipo II, tanto matematicamente ( $1 - \beta$ ) quanto conceitualmente. Poder é a capacidade do teste de detectar um efeito que existe na população. Em outras palavras, o teste rejeita corretamente uma falsa hipótese nula.

Por exemplo, se seu estudo tem 80% de poder, ele tem 80% de chance de detectar um efeito que existe. Deixe este ponto ser um lembrete de que

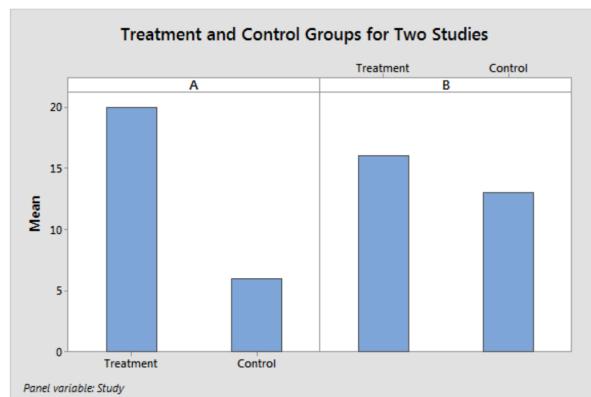
quando você trabalha com amostras, nada é garantido! Quando existe um efeito na população, seu estudo pode não detectá-lo porque você está trabalhando com uma amostra. As amostras contêm erros amostrais, que podem ocasionalmente fazer com que uma amostra aleatória represente incorretamente a população. 80% de potência é uma referência padrão para estudos. No entanto, você precisará considerar os padrões para seu campo ou indústria.

Como você aprendeu nas seções anteriores, embora vários fatores afetem o poder, os pesquisadores têm o maior controle sobre o **tamanho da amostra**. Determinar um bom tamanho de amostra para um estudo é sempre uma questão importante. Afinal, usar o tamanho de amostra errado pode condenar seu estudo desde o início. Felizmente, a análise de poder pode encontrar a resposta para você. A análise de poder combina análise estatística, conhecimento da área de assunto e seus requisitos para ajudá-lo a obter o tamanho ideal da amostra.

Como você verá nesta seção, tanto os estudos com potência insuficiente quanto com potência excessiva são problemáticos. Vamos aprender como encontrar o tamanho de amostra certo para o seu estudo!

## FATORES ENVOLVIDOS NA SIGNIFICÂNCIA ESTATÍSTICA

Observe o gráfico abaixo e identifique qual estudo encontrou um efeito de tratamento real e qual não. Dentro de cada estudo, a diferença entre o grupo de tratamento e o grupo de controle é a estimativa amostral do tamanho do efeito.



Algum estudo obteve resultados significativos? Os efeitos estimados em ambos os estudos podem representar um efeito real ou um erro amostral aleatório. Você não tem informações suficientes para fazer essa determinação. Os testes de hipóteses incorporam essas considerações para determinar se os resultados são estatisticamente significativos.

- Tamanho do efeito: quanto maior o tamanho do efeito, menor a probabilidade de ser um erro aleatório. É claro que o Estudo A exibe um efeito mais substancial na amostra – mas isso é insuficiente por si só.
- Tamanho da amostra: tamanhos de amostra maiores permitem que testes de hipóteses detectem efeitos menores. Se o tamanho da amostra do Estudo B for grande o suficiente, seu efeito mais modesto pode ser estatisticamente significativo.
- Variabilidade: Quando os dados de sua amostra têm mais variabilidade, é mais provável que o erro de amostragem aleatória produza diferenças consideráveis entre os grupos experimentais, mesmo quando não há efeito real. Se os dados da amostra no Estudo A tiverem variabilidade suficiente, o erro aleatório pode ser responsável pela grande diferença.

021.399.242-66

O teste de hipóteses pega todas essas informações e as usa para calcular o p-valor — que você usa para determinar a significância estatística. A principal conclusão é que a significância estatística de qualquer efeito depende coletivamente do tamanho do efeito, do tamanho da amostra e da variabilidade presente nos dados da amostra. Consequentemente, você não pode determinar o tamanho correto da amostra sem essas informações porque os três fatores estão interligados.

## OBJETIVOS DE UMA ANÁLISE DE PODER

A análise de poder envolve tomar essas três considerações, adicionar conhecimento da área de assunto para estabelecer um tamanho de amostra. Durante esse processo, você deve confiar muito em sua experiência para fornecer estimativas razoáveis dos valores de entrada.

À medida que você aumenta o tamanho da amostra, o teste de hipóteses ganha maior capacidade de detectar pequenos efeitos. Esta situação parece fantástica. No entanto, tamanhos de amostra maiores custam mais dinheiro.

Em muitos casos, a coleta de amostras pode custar caro. Por exemplo, se você está fazendo uma pesquisa com pessoas, vai precisar de mais tempo e mais investimento se quiser uma amostra muito grande. Quando você está desenhando um experimento (antes mesmo de ter as amostras) você não quer coletar uma amostra grande e cara apenas para detectar um efeito pequeno demais para ser útil! Nem você quer um estudo de baixo poder que tenha uma baixa probabilidade de detectar um efeito. Seu objetivo é coletar uma amostra grande o suficiente para ter poder suficiente para detectar um efeito significativo, mas não grande demais caso isso represente um desperdício.

Você vai precisar especificar três fatores para calcular o tamanho da amostra. Por exemplo, se você especificar o menor tamanho de efeito que é praticamente significativo, variabilidade e poder, você calculará o tamanho de amostra necessário.

É agora que precisaremos falar sobre a meta - ou objetivo - para especificar o menor tamanho de efeito que é praticamente significativo. Nós não sabemos até onde podemos chegar em uma experimentação, mas precisamos ter uma estimativa de onde queremos chegar. Por exemplo, suponhamos que vamos fazer um novo site para um e-commerce para aumentar a conversão de clientes (% de pessoas que compram no site considerando todas que entram no site). Queremos fazer isso pois acreditamos que o site atual é muito complicado e não tem um visual atrativo, e os clientes entram no site mas não efetuam a compra.

Nesse exemplo acima, nós não sabemos quanto vamos aumentar de conversão com um novo site - e nem se vamos. Porém, precisamos desenhar um teste que garanta uma amostra mínima para comprovar um certo efeito (um aumento de conversão) caso haja realmente esse efeito. Em outras palavras, não queremos que o tamanho da amostra influencie nosso resultado. Queremos que, caso o site realmente melhore a conversão, tenhamos garantido uma amostra grande o suficiente para nos mostrar esse efeito.



Mas como calcular a amostra mínima se não temos esse efeito? Essa etapa nem sempre é fácil, pois envolverá seu conhecimento do seu business. quando trabalhamos em empresas, geralmente temos metas a cumprir. Por exemplo, a meta para aquele semestre pode ser aumentar a conversão em 20%. Ou seja, nossa amostra teria que ter um tamanho bom o suficiente para comprovar efeito caso ele realmente exista. Você também pode usar sua experiência para identificar a menor diferença que ainda seja significativa para sua aplicação. Em outras palavras, você considera diferenças menores como irrelevantes. Não valeria a pena gastar recursos para detectá-los.

O valor de poder é onde especificamos a probabilidade de que o teste de hipótese detecte a diferença na amostra se essa diferença existir na população (evitar o erro tipo II). Se você mantiver os outros valores de entrada constantes e aumentar o poder do teste, o tamanho da amostra necessária também aumentará. O valor adequado para entrar neste campo depende das normas em sua área de estudo ou indústria. Os valores de potência comuns são 80% e 90%.

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.353.242-06

Para fazer o cálculo de tamanho mínimo, a melhor opção que conheço é o software G\*Power. É um software alemão gratuito e vocês podem baixá-lo nesse site:

<https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower>

Vá até parte de Download e selecione o seu sistema operacional para completar o download.

## Download

By downloading G\*Power you agree to these terms of use:

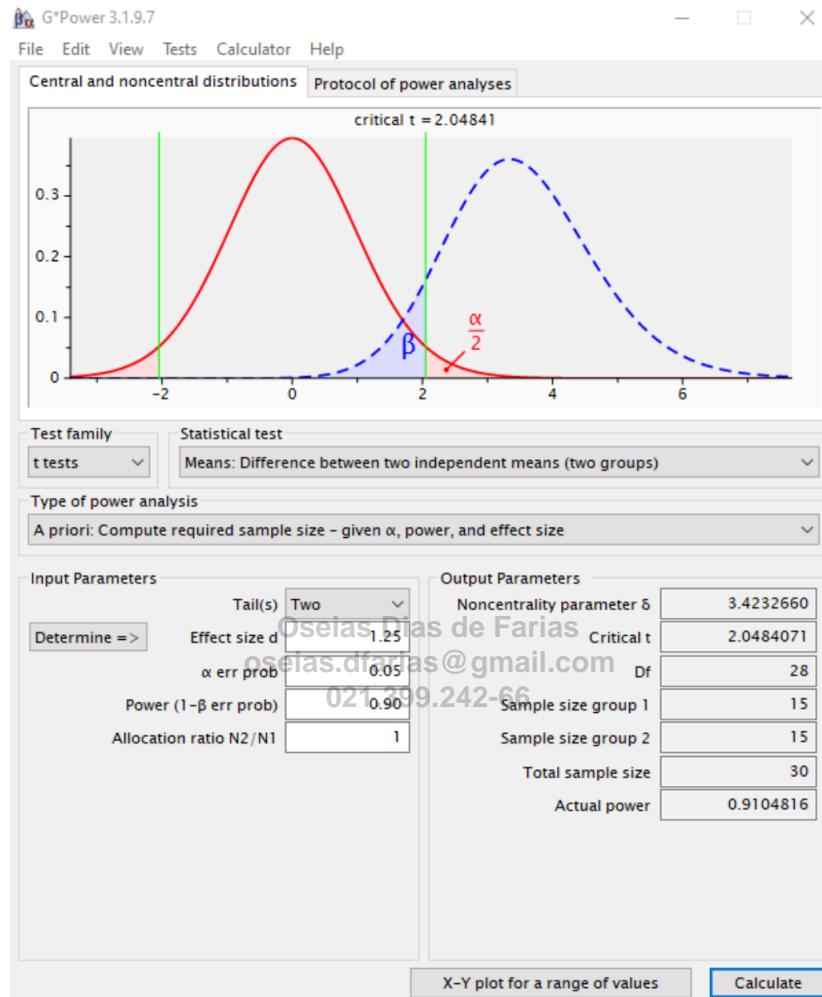
1. G\*Power is free for everyone. Commercial distribution is strictly prohibited.
2. G\*Power is distributed from this website. If you wish to distribute G\*Power in some other way, then you need to seek permission from the authors. Please [✉](#) send us an e-mail in which you specify how and for what purpose you intend to distribute G\*Power.
3. You may use screenshots of G\*Power without asking for permission.
4. Considerable effort has been put into program development and evaluation, but there is no warranty whatsoever.

[🔗](#) Download G\*Power 3.1.9.7 for Windows XP, Vista, 7, 8, 10 and 11 (about 20 MB). Please make sure to choose "unpack with folders" in your unzip tool.

[🔗](#) Download G\*Power 3.1.9.6 for Mac OS X 10.7 to 12 (about 2 MB).



O G\*Power tem uma infinidade de testes para realizar o cálculo. Abaixo, selecionei o t-test para comparar 2 médias de 2 grupos diferentes. Selecionamos o campo "A priori" para calcular o tamanho da amostra.



O  $d$  é o efeito padronizado para médias. É simplesmente a diferença média dividida pelo desvio padrão. No exemplo, inseri uma diferença de 5 e um desvio padrão de 4. Isso equivale a um  $d$  de  $5/4 = 1,25$ . Também optei por um poder de 90% e um alpha de 5% (nível de significância). No termo "allocation ratio" você especifica se quer que uma amostra deverá ser maior do que a outra (controle maior do que teste, por exemplo). Coloque 1 se você quiser que as amostras tenham o mesmo tamanho.

No meu caso, a ferramenta indica que preciso ter uma amostra mínima com 30 dados - 15 em cada um dos grupos. Ou seja, caso eu queira realizar um

teste que tenha um grupo de tratamento (ou teste) e um de controle, e queira ver uma diferença de pelo menos 5 pontos entre eles, sabendo que atualmente eles tem um desvio-padrão = 4, eu preciso desse tamanho mínimo de amostra.

Obtemos do lado direito os tamanhos de grupo e resultados de poder. O G\*Power inclui até um belo gráfico na parte superior que ilustra sua análise de poder. É semelhante ao gráfico de taxa de erro que apresentei anteriormente neste capítulo.

Outros tipos de testes estatísticos têm medidas diferentes para efeitos padronizados. Você só precisa saber como converter do efeito bruto (diferença) para o efeito padronizado, como fizemos para o d. O manual do G\*Power poderá te orientar quanto a isso. Na ferramenta, procure o campo "Help" e clique em "Download the G\*Power manual (PDF)".

## PASSO 4: REALIZE O TESTE E COMPARE OS RESULTADOS

Oseias Dias de Farias  
021.399.242-66

Agora chegou a hora realizar o teste com esse tamanho de amostra (ou maior, se preferir) e usar tudo que aprendemos anteriormente com testes de hipótese. Continuando com o teste de ter um novo site para o e-commerce, precisaríamos separar 15 pessoas (evitando vieses) que veriam o site antigo e 15 pessoas que veriam o site novo. Mediremos a conversão de cada um desses grupos, e, para compará-las, usamos o teste de hipótese adequado - conforme visto em seções anteriores.

## MATERIAL COMPLEMENTAR

Caso tenha interesse em aprofundar seus conhecimentos em testes AB, recomendo fortemente o curso gratuito feito pelo Google em parceria com a Udacity: <https://www.udacity.com/course/ab-testing--ud257>

Disponível apenas em inglês.

# 23. Correlação



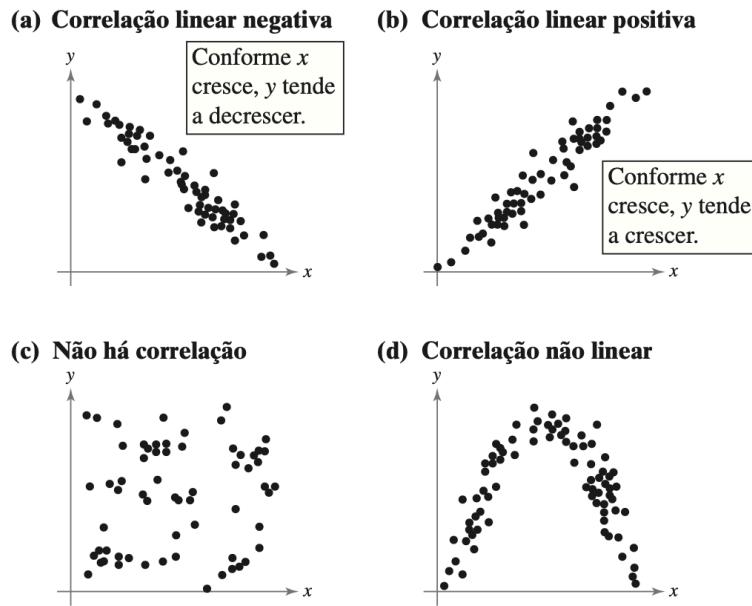
## CORRELAÇÃO

Oseias Dias de Farias

[oseias\\_dfarias@gmail.com](mailto:oseias_dfarias@gmail.com)  
021.399.242-66

Suponha que um inspetor de segurança queira determinar se existe relação entre o número de horas de treinamento para um funcionário e o número de acidentes envolvendo este funcionário. Ou suponha que uma psicóloga queira saber se existe relação entre o número de horas que uma pessoa dorme a cada noite e o tempo de reação dessa pessoa. Como ele ou ela determinaria se existe alguma relação?

Uma **correlação** é uma relação entre duas variáveis. Os dados podem ser representados por pares ordenados  $(x, y)$ , sendo  $x$  a **variável independente** (ou **explanatória**) e  $y$  a **variável dependente** (ou **resposta**). Dizemos que:



## COEFICIENTE DE PEARSON

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

Interpretar a correlação usando um gráfico dispersão pode ser subjetivo. Uma maneira adequada de obter a direção e medir a força de uma correlação **linear** entre duas variáveis é calcular o **coeficiente de correlação de Pearson**, que é uma medida da força e da direção de uma relação linear entre duas variáveis. Sua fórmula é dada por:

$$r = \frac{n\sum xy - \sum x\sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \sqrt{n\sum y^2 - (\sum y)^2}}$$

O coeficiente de correlação de Pearson é representado pela letra grega rho ( $\rho$ ) para o parâmetro populacional e  $r$  para uma estatística amostral. Este

O coeficiente de relação é um único número que mede tanto a força quanto a direção da relação linear entre duas variáveis contínuas. Os valores podem variar de -1 a +1.

- Força: Quanto maior o valor absoluto do coeficiente de correlação, mais forte é a relação. Os valores extremos de -1 e 1 indicam uma relação

perfeitamente linear onde uma mudança em uma variável é acompanhada por uma mudança perfeitamente consistente no outro. Para esses relacionamentos, todos os pontos de dados caem em uma linha. Na prática, você não verá nenhum tipo de relacionamento perfeito. Quanto mais próximo de zero, pior é a correlação.

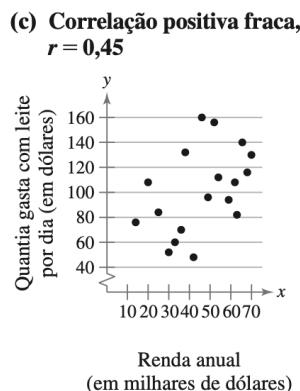
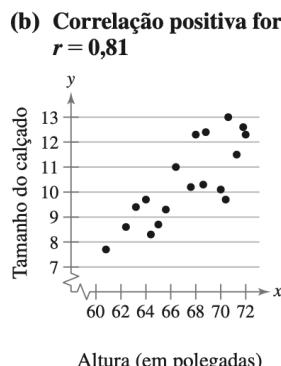
- Direção: O sinal do coeficiente de correlação representa a direção do relacionamento. Coeficientes positivos indicam que quando o valor de uma variável aumenta, o valor da outra variável também tende a aumentar. Relacionamentos positivos produzem uma inclinação ascendente em um gráfico de dispersão. Coeficientes negativos representam indicam que quando o valor de uma variável aumenta, o valor da outra variável tende a diminuir. As relações negativas produzem uma inclinação descendente.

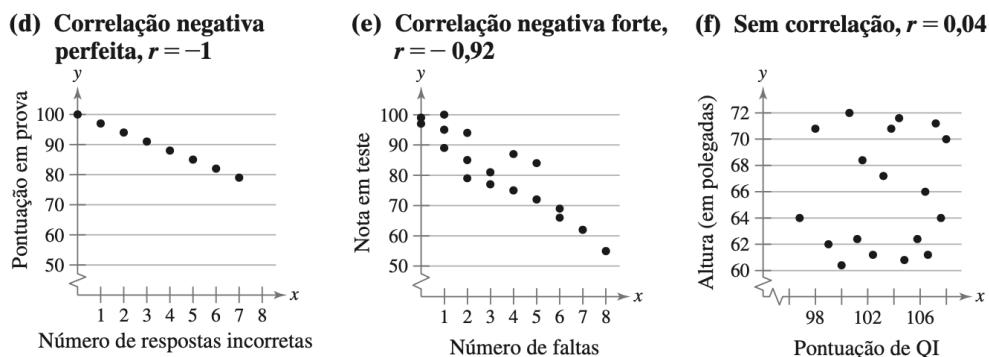
Uma interpretação errônea comum é assumir que os coeficientes de correlação negativos indicam que não há relação. Afinal, uma correlação negativa soa supostamente como nenhum relacionamento. No entanto, os gráficos de dispersão para as ~~correlações negativas~~ mostram relações reais.

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021 323 242 66

Para coeficientes de correlação negativos, valores altos de uma variável estão associados a valores baixos de outra variável. Por exemplo, há uma correlação negativa entre faltas escolares e notas.





Lembrando que os coeficientes de correlação de Pearson medem apenas relações lineares. Consequentemente, se seus dados contiverem uma relação curvilínea, o coeficiente de correlação não a detectará. Por exemplo, a correlação para os dados no gráfico de dispersão abaixo são zero. No entanto, há uma relação entre as duas variáveis - não é apenas linear.



## COEFICIENTE DE SPEARMAN

A correlação de Pearson avalia a relação linear entre duas variáveis contínuas. Uma relação é linear quando a mudança em uma variável é associada a uma mudança proporcional na outra variável.

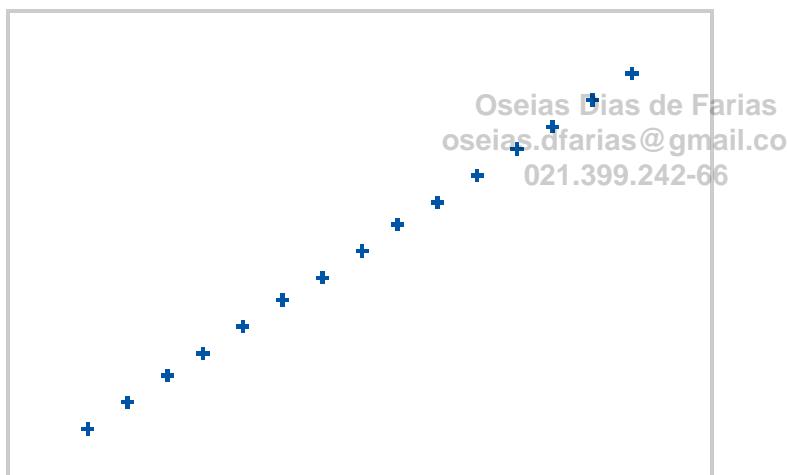
Por outro lado, a correlação de Spearman avalia a relação monotônica entre duas variáveis contínuas ou ordinais. Em uma relação monotônica, as variáveis tendem a mudar juntas, mas não necessariamente a uma taxa constante. O coeficiente de correlação de Spearman baseia-se nos valores classificados de cada variável, em vez de os dados brutos.

Sua fórmula é dada por:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

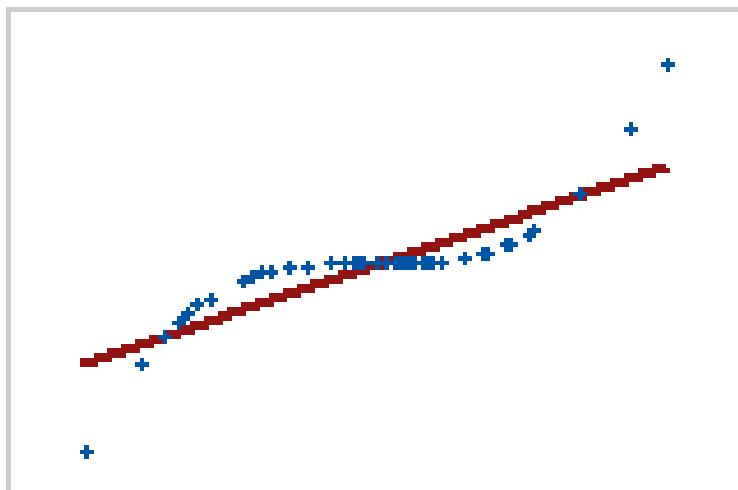
Em que  $d$  é a diferença de dois rankings (semelhante a ideia de Posto, já discutida anteriormente) e  $n$  é o número de observações.

Os coeficientes de correlação de Pearson e Spearman podem variar em valor de -1 a +1. Para o coeficiente de correlação de Pearson ser +1, quando uma variável aumenta, as outras variáveis aumentam por uma quantidade consistente. Este relacionamento forma uma linha perfeita. O coeficiente de correlação de Spearman também é +1 neste caso.



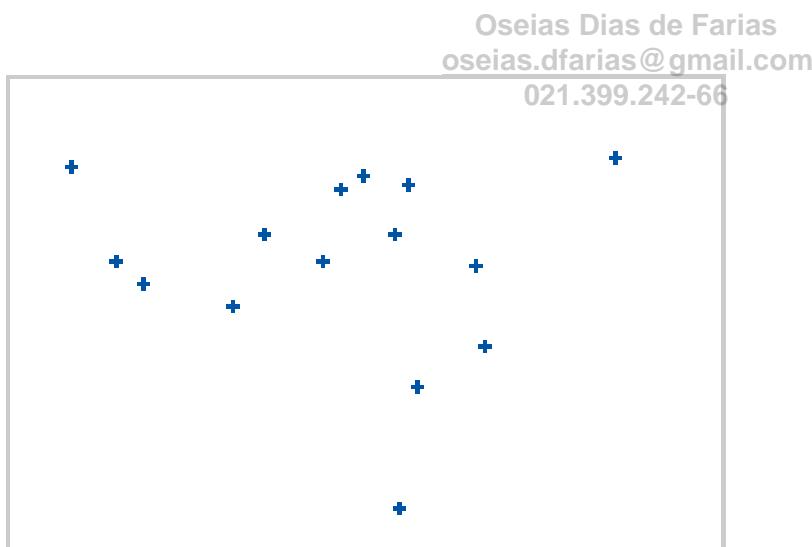
Pearson = +1, Spearman = +1

Se a relação é que uma variável aumenta quando a outra aumenta mas a quantidade não é consistente, o coeficiente de correlação de Pearson é positivo mas menor que +1. O coeficiente de Spearman ainda será +1 neste caso.



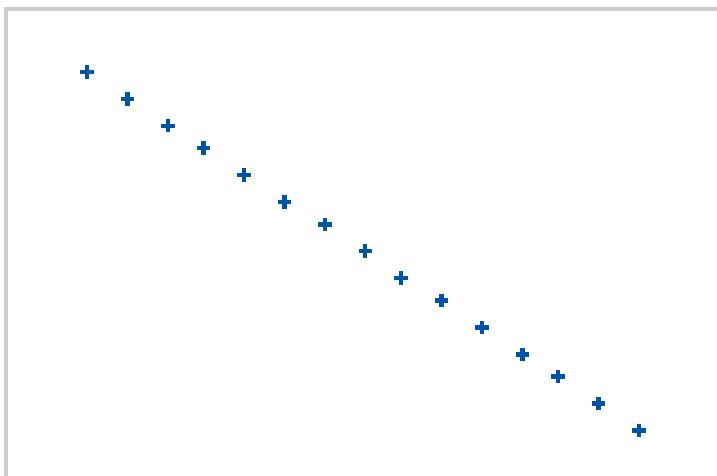
Pearson = +0,851, Spearman = +1

Quando uma relação é aleatória ou inexistente, os dois coeficientes de correlação se aproximam de zero.



Pearson = -0,093, Spearman = -0,093

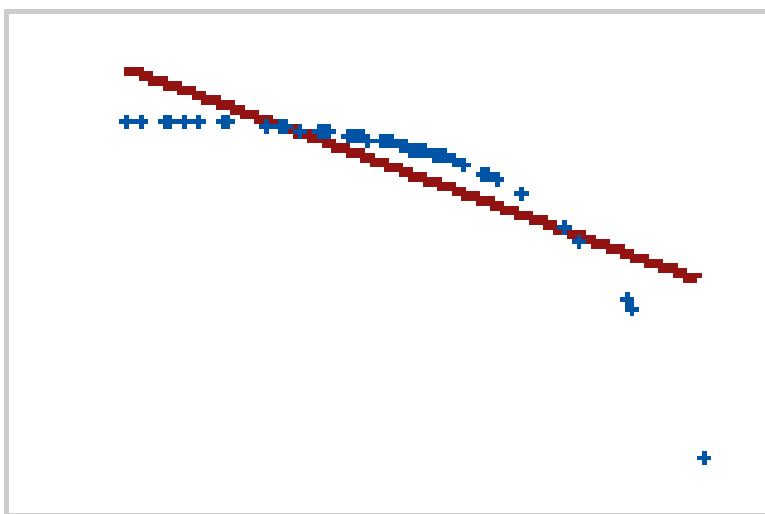
Se o relacionamento é uma linha perfeita para uma relação decrescente, ambos os coeficientes de correlação são -1.



Pearson = -1, Spearman = -1

Se o relacionamento é aquela variável que diminui quando as outras aumentam, mas a quantidade não é consistente, então o coeficiente de correlação de Pearson é negativo, mas maior que -1. O coeficiente de Spearman ainda é igual a -1, neste caso,

Oseias Dias de Farias  
oseias.dias@gmail.com  
021.399.242-66

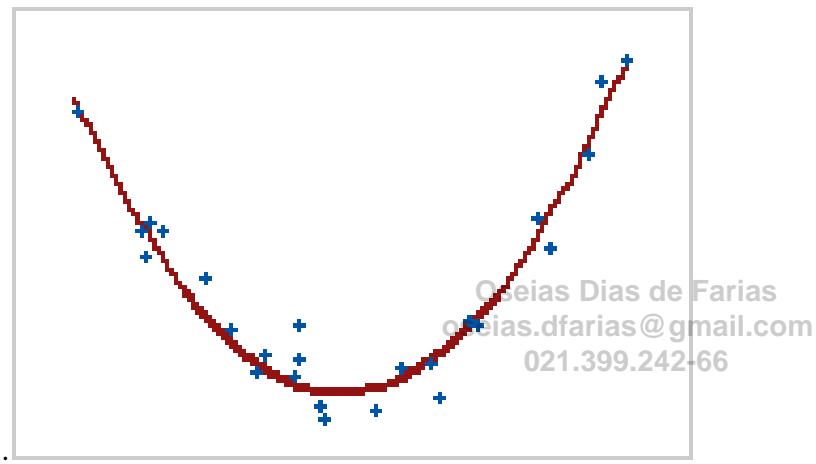


Pearson = -0,799, Spearman = -1



## OUTROS RELACIONAMENTOS NÃO LINEARES

Os coeficientes de correlação de Pearson medem somente relações lineares. Os coeficientes de correlação de Spearman medem somente relações monotônicas. Por isso, é possível que exista uma relação significativa mesmo que os coeficientes de correlação sejam 0. Examine um gráfico de dispersão para determinar a forma da relação

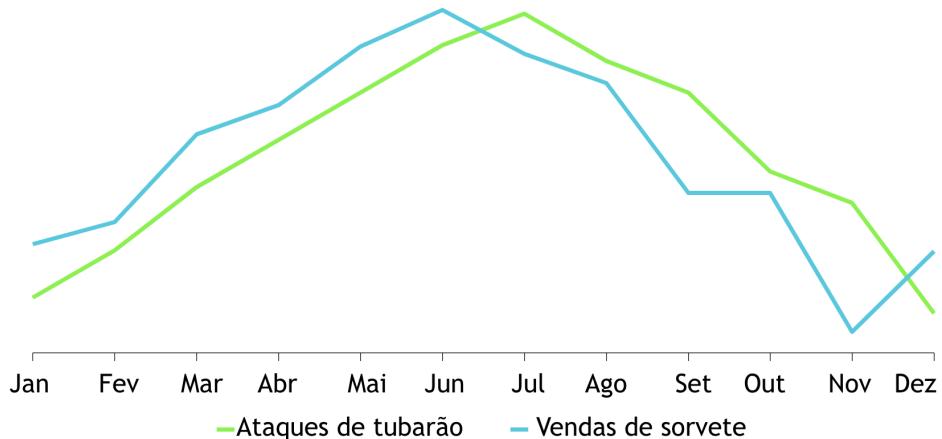


Coeficiente 0

Este gráfico mostra uma relação muito forte. O coeficiente de Pearson e o coeficiente de Spearman são aproximadamente 0.

## CORRELAÇÃO VERSUS CAUSA

Outro ponto importante de lembrar - que já foi discutido no capítulo "Desenhando um experimento" é que **correlação não implica em causa**. Vimos que ataques de tubarão x venda de sorvete são variáveis correlacionadas, mas uma não causa a outra. Nesse caso, tínhamos uma variável confundidora: verão nos EUA.



## TESTE DE HIPÓTESE PARA COEFICIENTES

Você pode usar um teste de hipóteses para determinar se o coeficiente de correlação amostral  $r$  fornece evidência suficiente para concluir que o coeficiente de correlação populacional  $\rho$  é significativo. Um teste de hipótese para  $r$  pode ser unilateral ou bilateral. As hipóteses nula e alternativa para os testes estão a seguir.

$$\begin{cases} H_0: \rho \geq 0 \text{ (não há correlação negativa significativa)} \\ H_a: \rho < 0 \text{ (correlação negativa significativa)} \end{cases}$$

Teste unilateral à esquerda

$$\begin{cases} H_0: \rho \leq 0 \text{ (não há correlação positiva significativa)} \\ H_a: \rho > 0 \text{ (correlação positiva significativa)} \end{cases}$$

Teste unilateral à direita

$$\begin{cases} H_0: \rho = 0 \text{ (não há correlação significativa)} \\ H_a: \rho \neq 0 \text{ (correlação significativa)} \end{cases}$$

Teste bilateral

Considerando um teste bilateral, na hipótese nula, um coeficiente de correlação de zero indica que não existe correlação. Em outras palavras, saber o valor de uma variável não fornece informações sobre o valor da outra.

variável. À medida que uma variável aumenta, a outra variável não tende a aumentar ou diminuir.

Para a hipótese alternativa, um coeficiente de correlação diferente de zero indica que o valor de uma variável fornece informações sobre o valor provável da outra variável – existe uma correlação. À medida que o valor de uma variável aumenta, o valor da outra variável tende a aumentar ou diminuir a uma taxa previsível. Observe que este é um teste bicaudal para que possa detectar correlações positivas e negativas. Portanto, a redação “não é igual” na hipótese alternativa.

**Se o p-valor for menor que seu nível de significância (por exemplo, 0,05), você pode rejeitar a hipótese nula, portanto dizemos que a correlação é estatisticamente significativa.** Sua amostra fornece evidências fortes o suficiente para concluir que o coeficiente de correlação populacional não é igual a zero.

## PEARSON

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

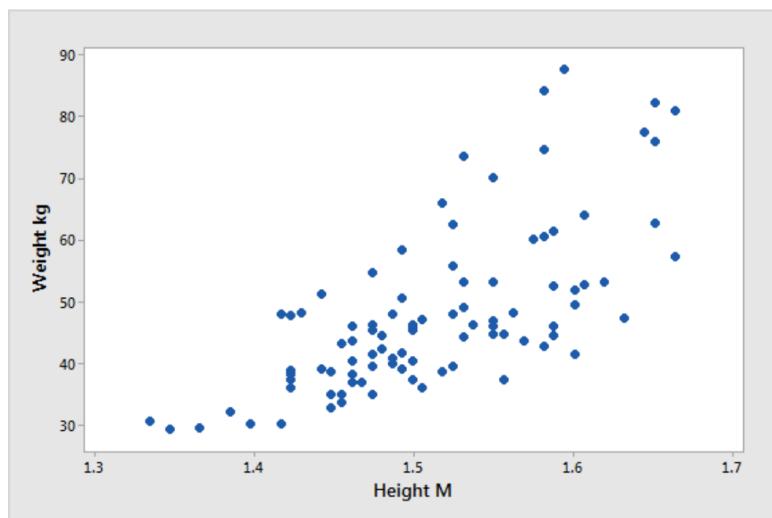
Para resultados confiáveis do teste de correlação da Pearson, seus dados devem atender às seguintes suposições:

- Duas variáveis contínuas
- Relação linear
- Amostras aleatórias e independentes
- Os dados seguem uma distribuição normal bivariada ou você tem pelo menos 25 observações

Se você tiver pelo menos 25 observações, o p-valor é válido para dados que se afastam da distribuição normal. Com menos observações, o p-valor pode não ser preciso para distribuições não normais.

Muitíssimo cuidado aqui! Isso não significa que os dados precisam ser normais ou ter pelo menos 25 dados para que você possa avaliar a **correlação**. A correlação por si só pode ser feita sempre que seus dados forem numéricos, mas o **teste de hipótese** em questão só poderá ser feito nessas condições.

Vamos ver um exemplo de correlação de altura e peso. O gráfico e os resultados estatísticos estão abaixo.



**Correlation: Height M, Weight kg**

Pearson correlation of Height M and Weight kg = 0.694  
P-Value = 0.000  
Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

A correlação para esses dados é de 0,694. É um valor positivo, que indica que à medida que a altura aumenta, o peso tende a aumentar. Você pode ver essa relação no gráfico. A força da correlação é moderada. Não é tão forte que os pontos de dados estejam abraçando firmemente uma linha. No entanto, não é tão fraco que parece uma bolha amorfa.

O p-valor de 0,000 é menor que nosso nível de significância de 0,05. A evidência amostral é forte o suficiente para rejeitar a hipótese nula e concluir que a correlação existe na população.

A fórmula para o teste pode ser vista abaixo, em que  $r$  é o coeficiente de correlação.

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

## SPEARMAN

O teste de Spearman é não paramétrico, e com isso tem menos suposições do que o teste de Pearson. As únicas coisas que o teste exigem são: amostras aleatórias e independentes. Sua fórmula é dada por:

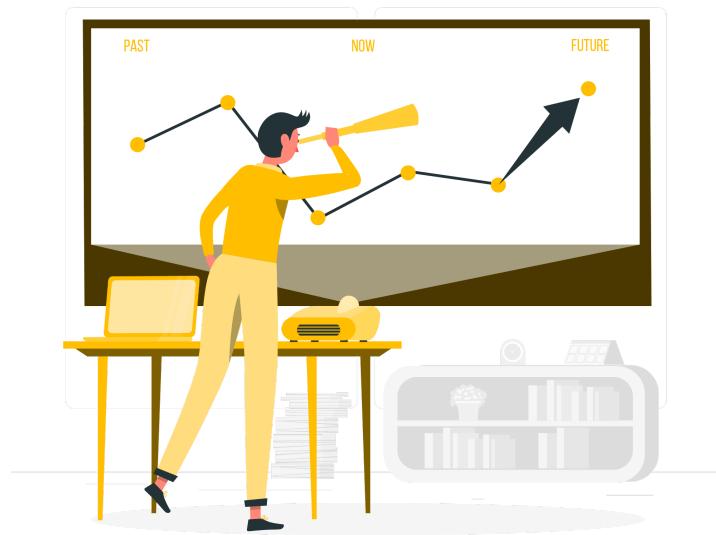
$$t = \frac{r_s}{\sqrt{(1 - r_s^2)/(n - 2)}}$$

Em que  $r_s$  é o coeficiente de correlação de spearman.

A ideia é a mesma que Pearson. Se p-valor for menor do que alpha, rejeitamos a hipótese nula e, portanto, o valor é estatisticamente significativo.



# 24. Regressão Linear



Oseias Dias de Farias  
Notas importantes: A maioria dos exemplos do capítulo foram retirados das  
[oseias.diasdefarias@gmail.com](mailto:oseias.diasdefarias@gmail.com)  
referências Frost, J [2].  
021.399.242-66

Em estatística, regressão é uma técnica que permite quantificar e inferir a relação de uma variável dependente (variável de resposta) com variáveis independentes (variáveis explicativas). Por exemplo, suponha que você seja um gerente de vendas tentando prever os números do próximo mês. Você sabe que dezenas, talvez até mesmo centenas de fatores, como o tempo para a promoção, um competidor, um boato de um modelo novo e melhorado, podem afetar o número. Talvez as pessoas da sua organização tenham até uma teoria sobre o que terá o maior efeito nas vendas, mas é a partir de uma análise de regressão que conseguiremos:

- 1) Fazer um forecast (previsão) de vendas
- 2) Entender quais variáveis realmente impactam no preço

Na análise de regressão, esses fatores são chamados de variáveis explicativas ou independentes. O principal fator que você está tentando entender ou prever é sua variável dependente. No exemplo acima, a variável dependente é a venda mensal. E, então você tem suas variáveis independentes – os fatores

que você suspeita ter um impacto em sua variável dependente, como o tempo para a promoção, um competidor, um boato de um modelo novo e melhorado, etc.

É importante ressaltar que a partir de agora começamos a colocar nossos pés em *machine learning* - área de domínio dos cientistas de dados. O escopo do nosso curso não vai cobrir todos os algoritmos que existem para *machine learning* (são muitos e precisaríamos de anos de estudo), porém vamos cobrir 2 dos principais deles: A regressão linear e logística.

Pensando no exemplo de peso x altura dito anteriormente, podemos usar essa equação para entender o quanto o peso aumenta com cada unidade adicional de altura e fazer previsões para alturas específicas.

A análise de regressão nos permite expandir a correlação de outras maneiras.

Se tivermos mais variáveis que expliquem as mudanças no peso, podemos incluí-las no modelo e potencialmente melhorar nossas previsões.

Oseias Dias de Farias

E, se a relação for curva, ainda podemos ajustar um modelo de regressão para os dados.

[oseias\\_farias@gmail.com](mailto:oseias_farias@gmail.com)  
021.399.242-66

Além disso, uma forma do coeficiente de correlação de Pearson aparece na análise de regressão. O  $R^2$  (R-quadrado) é uma medida primária de quão bem um modelo de regressão ajusta os dados. Essa estatística representa a porcentagem de variação em uma variável que outras variáveis explicam. Para um par de variáveis, R-quadrado é simplesmente o quadrado da correlação de Pearson. Por exemplo, o quadrado do coeficiente de correlação altura-peso de 0,705 produz um R-quadrado de 0,497, ou 49,7%. Em outras palavras, a altura explica cerca de metade da variabilidade do peso.

Mas estamos nos antecipando. Eu vou cobrir R-quadrado em muito mais detalhes mais para frente.

## CONCEITOS FUNDAMENTAIS

### VARIÁVEIS DEPENDENTES



A variável dependente é uma variável que você deseja explicar ou prever usando o modelo. Os valores desta variável dependem de outras variáveis. Também é conhecida como variável de resposta, variável de resultado e é comumente denotada usando um Y. Tradicionalmente, representamos graficamente variáveis dependentes e o eixo vertical, ou Y.

## VARIÁVEIS INDEPENDENTES

Variáveis independentes são as variáveis que você inclui no modelo para explicar ou prever mudanças na variável dependente. Em experimentos, variáveis independentes são sistematicamente definidas e alteradas pelos pesquisadores. No entanto, em estudos observacionais, os valores das variáveis independentes não são estabelecidas por pesquisadores, mas sim observadas. Essas variáveis também são conhecidas como variáveis preditoras, variáveis de entrada, e são comumente denotadas usando Xs. Nos gráficos, a colocamos comumente no eixo horizontal ou X.

## REGRESSÃO SIMPLES VERSUS MÚLTIPLA

seias Dias de Farias  
oseias.dfarias@gmail.com  
021 399 242-66

Ao incluir uma variável independente no modelo, você está realizando uma regressão simples. Para mais de uma variável independente, é regressão múltipla. Apesar dos nomes diferentes, é essencialmente a mesma análise com as mesmas interpretações e suposições.

## OBJETIVOS DA ANÁLISE DE REGRESSÃO

A análise de regressão descreve matematicamente as relações entre variáveis independentes e uma variável dependente. Use a regressão para dois objetivos principais:

1. Compreender as relações entre essas variáveis.

Como as mudanças nas variáveis independentes se relacionam com mudanças na variável dependente?

2. Para prever a variável dependente inserindo valores para as variáveis independentes na equação de regressão.

## EQUAÇÃO DA REGRESSÃO

Imagine que você quer prever algo, como o preço de uma casa baseado em coisas como seu tamanho, idade e localização. A equação de regressão é como uma receita que te ajuda a fazer essa previsão. Ela combina essas informações (tamanho, idade, localização) de maneiras específicas para te dar uma estimativa do preço da casa.

Uma regressão pode ser multivariada ou univariada

- **Univariada ou simples:** Aqui, estamos focados em apenas uma variável para prever outra. Imagine que para prever o preço de uma casa você usa somente a quantidade de quartos que ela tem. Isso é uma regressão univariada
- **Multivariada ou múltipla:** Aqui usamos mais de apenas 1 variável para prever a outra. No exemplo da casa, poderíamos usar quantidade de quartos, localização, idade, etc

Essa relação entre a variável dependente com as independentes podem ser expressos em uma equação da seguinte forma

Simple  
Linear  
Regression

$$y = b_0 + b_1 * x_1$$

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Multiple  
Linear  
Regression

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

A equação tem uma parte que é o ponto de partida (chamado de intercepto), que seria o preço base da casa se todas as outras informações fossem zero - esse é o  $b_0$  indicado acima. Depois, para cada variável independente da casa (como tamanho ou idade), você adiciona um pouco ao preço base, dependendo de quanto essa característica afeta o preço. Essa "influência" é representada por números na equação chamados coeficientes -  $b_1$ ,  $b_2$ ,  $b_3$ , etc

Os valores  $x_1$ ,  $x_2$ , etc são os valores das suas variáveis independentes.

Por exemplo, supondo que rodamos um modelo de predição de casa com uma base de dados própria, teríamos algo assim:

$$Y = b_0 + b_1 * (\text{idade da casa}) + b_2 * (\text{quantidade de quartos}) + \dots$$

$Y$  é o valor da casa, tudo que está em parênteses são nossas variáveis que vão nos ajudar a falar o preço da casa. Tendo os coeficientes, se tivéssemos

O último termo chamamos de Random Component, ou simplesmente erro. O erro do modelo é a diferença entre os valores observados dos dados e os valores que o modelo de regressão prevê. Quando você usa um modelo de regressão para prever um valor de  $y$  com base em  $x$ , o modelo calcula uma estimativa com base na relação linear que ele encontrou nos dados. No entanto, essas previsões raramente são perfeitamente precisas, e o erro do modelo captura essas imprecisões.

## **EXEMPLO DE UMA ANÁLISE DE REGRESSÃO**

as Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Antes de entrarmos nas questões mais técnicas, quero mostrar aqui um exemplo de uma análise usando regressão.

Suponha que um pesquisador estude a relação entre potência e a saída de uma lâmpada. Neste estudo, a saída de luz é a variável dependente porque depende da potência. A potência é a variável independente.

Após realizar a análise de regressão, o pesquisador compreenderá a natureza da relação entre essas duas variáveis. Essa relação é estatisticamente significativa? Que efeito a potência tem na saída de luz? Para uma determinada potência, quanta saída de luz o modelo prevê?

Especificamente, a equação de regressão descreve a mudança média em saída de luz para cada aumento de um watt. O p-valor indica se a relação é estatisticamente significativa. E o pesquisador pode inserir valores de potência na equação para prever a saída de luz.

A análise de regressão pode lidar com muitas coisas. Por exemplo, você pode usar análise de regressão para fazer o seguinte:

1. Modelar múltiplas variáveis independentes
2. Incluir variáveis contínuas e categóricas
3. Modelar relações lineares e curvilíneas

Alguns exemplos do que podemos fazer com regressão:

1. O status socioeconômico tem relação com desempenho educacional?
2. A educação e o QI tem relação com o salário de alguém?
3. Os hábitos de exercício e dieta tem relação com o peso?
4. Beber café e fumar cigarros estão relacionados ao risco de mortalidade?

Vamos a um exemplo. Vamos supor que sua variável dependente é renda e suas variáveis independentes incluem QI (em pontos) e educação (em anos), você pode ver uma saída como esta:

Coefficients					
	Osejas Dias de Farias osejas.dfarias@gmail.com 021.33324266				
Term	Coef	SE Coef	T	P	
Constant	483.670	39.5671	12.2241	0.000	
IQ	4.796	0.9511	5.0429	0.000	
Education	24.215	1.9405	12.4785	0.000	

Os baixos valores de p indicam que tanto a educação quanto o QI são estatisticamente significativos. O coeficiente de QI (4,796) indica que cada ponto de QI adicional aumenta sua renda em uma média de aproximadamente \$ 4,80 se as outras variáveis forem constantes. Além disso, o coeficiente de educação (24.215) indica que um ano adicional de educação aumenta os ganhos médios em \$ 24,22, mantendo o outras variáveis constantes. O uso da análise de regressão oferece a capacidade de separar os efeitos de variáveis e avaliar o papel que cada uma desempenha. Porém, eu já aviso aqui, **você só vai conseguir fazer essas inferências se, e somente se, algumas premissas forem satisfeitas.** Vamos falar disso mais para frente.

Com isso, você também é capaz de prever - sim, prever! - qual seria a renda de uma pessoa dado que você sabe o QI e o nível de educação da pessoa. É aqui que entra a segunda grande potência de uma regressão linear: a **predição**.

A grande beleza é que para a predição você não precisa que essas premissas sejam satisfeitas. Porém, para as inferências, como disse acima, vai precisar, ok? Mas relaxa que já já você vai entender isso melhor

## ENTENDENDO REGRESSÃO MAIS A FUNDO

Agora quero explicar uma série de conceitos para vocês. Para isso, vou partir de um exemplo.

Vamos supor que queremos saber quantos cartões de crédito uma família vai usar dentro de um período de tempo (1 ano). Coletamos dados do primeiro ano de todas as famílias que já são nossos clientes,

Family ID	Number of credit cards used Y	Family size $X_1$	Family income (\$000) $X_2$	Number of Automobiles owned $X_3$
Oseias Dias de Farias oseias.dfas@gmail.com 021.399.242-66				
1	4	2	14	1
2	6	2	16	2
3	6	4	14	2
4	7	4	17	1
5	8	5	18	3
6	7	5	21	2
7	8	6	17	1
8	10	6	25	2

Variável dependente: Number of credit cards used (Y) - Número de cartão de créditos usados dentro do período de 1 ano

Variáveis independentes: Family size (tamanho da família), family income (salário total da família), number of automobiles (quantidade de automóveis da família).

Primeiro, precisamos setar um **modelo baseline**, ou seja, se eu não tivesse meu modelo para fazer a predição, qual seria a quantidade de cartões iria prever? Isso é feito se baseando somente na média - ou seja,

$y = 7$  (média de cartões considerando esse grupo de famílias)

Percebiam que para algumas famílias esse número seria um ótimo preditor (ex: família de ID 4), mas para outras esse número seria um péssimo preditor, pois eles estão muito longe da média (ex: família ID 1). Isso é totalmente esperado, já que simplesmente estamos coletando a média sem entender as outras variáveis da família.

Agora nossa tentativa vai ser encontrar um modelo que melhore esse cenário para cada família.

Antes de mais nada, vamos descobrir quanto estamos errando pra cada família?

**Regression variate:  $\hat{Y} = Y$**   
 Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
**Prediction equation:  $\hat{Y} = 7$**   
 091.399.242-66

Family ID	Number of credit cards used Y	Baseline prediction (56/8=7)	Prediction error	Prediction error squared
1	4	7	-3	9
2	6	7	-1	1
3	6	7	-1	1
4	7	7	0	0
5	8	7	1	1
6	7	7	0	0
7	8	7	1	1
8	10	7	3	9
<b>Total</b>	<b>56</b>		<b>0</b>	<b>22</b>



Aqui simplesmente subtrai cada linha de cartões por 7. Na coluna “prediction error” temos quanto longe estamos do nosso baseline para cada cliente. A coluna “prediction error squared” é simplesmente a coluna “prediction error” elevada ao quadrado.

Olhem como elevar o valor ao quadrado é relevante! Se não o fizéssemos, quando somássemos a coluna “prediction error” veríamos que, overall, não estamos errando nada - um modelo aparentemente perfeito. Mas aí que mora o perigo. Os valores negativos e positivos se anulam na soma, por isso precisamos elevar ao quadrado para fazer com que esse efeito desapareça.

### **E se usarmos uma variável apenas para prever?**

Vamos supor agora que ao invés de simplesmente usarmos a média, usaríamos isso daqui:

$$Y = b_0 + b_1 * X_1$$

Em que  $y$  é a quantidade de cartões que queremos prever e  $x_1$  poderia ser uma das variáveis, como tamanho da família.  
seuendfrio@gmail.com  
021.399.242-66

Usando um método chamado OLS que vamos explicar a seguir, somos capazes de encontrar o  $b_0$  e o  $b_1$  da equação. Vamos supor que fiz a regressão linear e temos agora:

$$Y = 2.87 + 0.97 * X_1$$

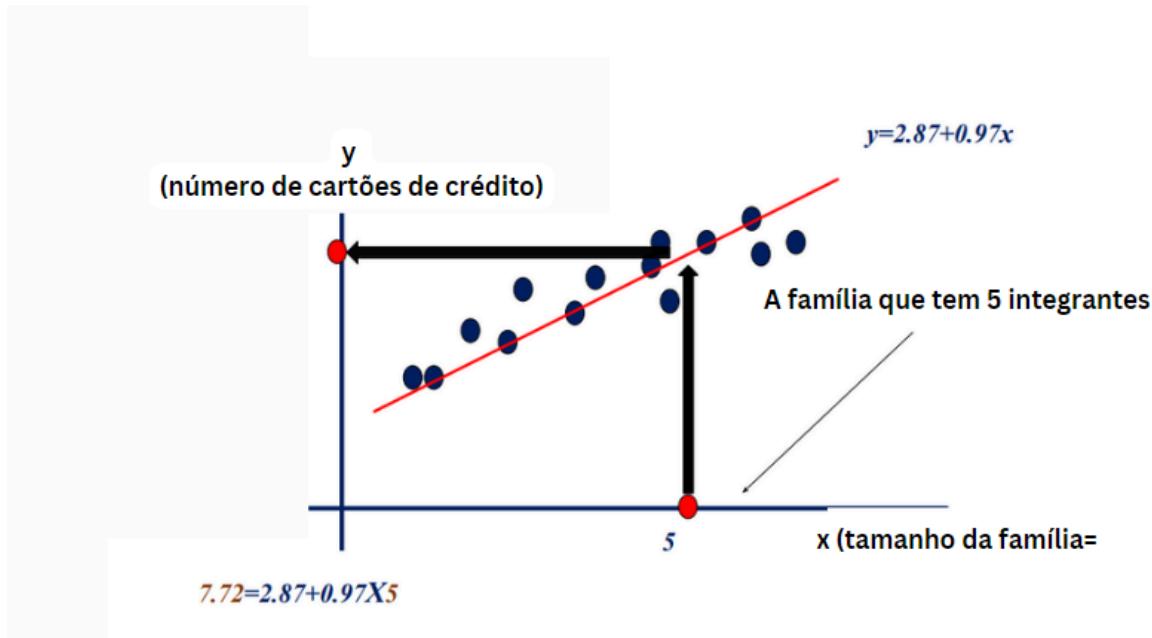
Agora, se você tiver o tamanho da família, vai ser capaz de dizer quantos cartões serão usados!

Por exemplo, se tivermos uma família com 5 integrantes:

$$Y = 2.87 + 0.97 * 5 = 7.72$$

Ou seja, estimamos que essa família vai consumir um total de 7.72 cartões em média.

Veja abaixo uma figura representativa



Sempre haverá uma diferença entre a previsão (7,72) e o valor verdadeiro (neste caso 8). Isso é chamado de **erro de previsão**. A diferença entre os pontos (real e previsto) é o que chamamos de **resíduo**

Antes de avançarmos, vamos entender como conseguimos achar os coeficientes da função - OLS

## MÍNIMOS QUADRADOS ORDINÁRIOS (OLS)

Antes de mais nada, para ajudar a garantir que seus resultados sejam válidos para regressão linear, considere os seguintes princípios ao coletar dados, realizar análise e interpretação dos resultados.

1. As variáveis independentes podem ser contínuas ou categóricas.
2. A variável dependente deve ser contínua. Se não for contínua, você provavelmente precisará usar um tipo diferente de análise de regressão porque é improvável que seu modelo satisfaça as suposições do OLS e possa produzir resultados nos quais você não pode confiar.

Use as melhores práticas ao coletar seus dados. Seguem alguns pontos a considerar:



1. Confirme se os dados representam sua população de interesse.
2. Colete uma quantidade suficiente de dados que lhe permita ajustar um modelo que é apropriadamente complexo para a área de assunto e fornece a precisão necessária para os coeficientes e previsões (veremos isso mais a seguir)
3. Meça todas as variáveis com a mais alta exatidão e precisão possível.

Vamos agora a algumas **definições essenciais para entender o OLS:**

### **1. Valores Observados e Ajustados**

Os **valores observados** da variável dependente são os valores da variável dependente que você registra durante seu estudo ou experimento juntamente com os valores das variáveis independentes. Esses valores são denotado usando Y.

Os **valores ajustados** são os valores que o modelo **prevê** para o variável usando as variáveis independentes. Se você inserir valores para o variáveis independentes na equação de regressão, você obtém o valor ajustado. Valores previstos e valores ajustados são sinônimos. É comum denotarmos valores ajustados como  $\hat{Y}$ .

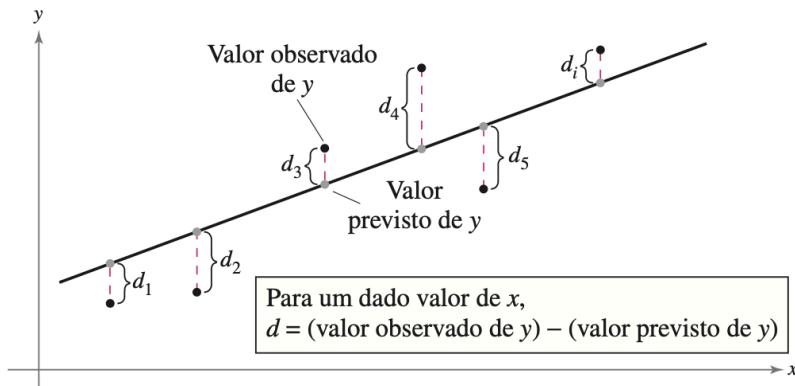
### **2. Resíduos**

Para entender o quanto bem o seu modelo se ajusta aos dados, você precisa avaliar as diferenças entre os valores observados e os valores ajustados. Essas diferenças representam o erro no modelo. Nenhum modelo é perfeito. Os valores observados e ajustados nunca corresponderão exatamente. No entanto, os modelos podem ser bons o suficiente

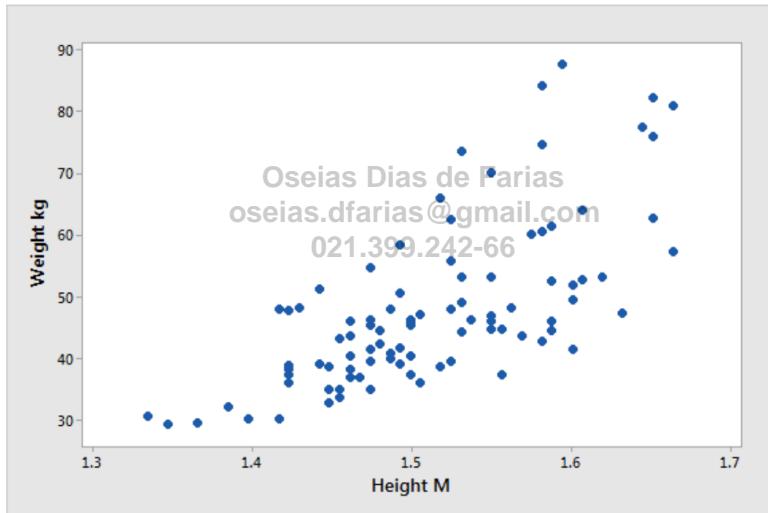
Essa diferença é conhecida como residual. Um resíduo é a distância entre um valor observado e o valor ajustado correspondente. Para calcular a diferença matematicamente, é simples subtração:

Resíduos = Valor observado – Valor ajustado.

Graficamente, os resíduos são as distâncias verticais entre os valores e os valores ajustados. No gráfico, a linha representa os valores ajustados do modelo de regressão.



Agora, vamos voltar ao conjunto de dados de altura e peso para o qual calculamos a correlação.



**Correlation: Height M, Weight kg**

Pearson correlation of Height M and Weight kg = 0.694  
P-Value = 0.000

O objetivo da análise de regressão é traçar uma linha entre esses dados pontos que minimizem a distância total dos pontos da linha. Em outras palavras, você quer ter um  $y$  predito mais próximo possível do  $y$  real - ou seja, você quer minimizar o resíduo.

Porém, você não pode simplesmente somar os resíduos porque o positivo e o negativo valores se anularão mesmo quando tendem a ser relativamente

amplos. Em vez disso, a regressão OLS eleva ao quadrado esses resíduos para que sejam sempre positivos. Desta forma, o processo pode somá-los sem cancelar um ao outro.

Primeiro, obtemos o resíduos entre os valores observados e ajustados usando subtração simples e, em seguida, apenas elevamos ao quadrado. Simples! Um ponto de dados com um resíduo de 3 terá um erro quadrado de 9. Um resíduo de -4 produz um erro quadrado de 16.

Então, o procedimento de mínimos quadrados ordinários soma esses erros quadrados, como mostrado na equação abaixo:

$$\sum (y - \hat{y})^2$$

O OLS desenha a linha que minimiza a soma dos erros ao quadrado (SSE).

SSE é uma medida de variabilidade. À medida que os pontos se afastam da linha ajustada, SSE aumenta. Como os cálculos usam diferenças quadradas, a variância está em unidades quadradas em vez das unidades originais dos dados. Para um dado definido, valores SSE menores sinalizam que as observações caem mais perto do valores ajustados.

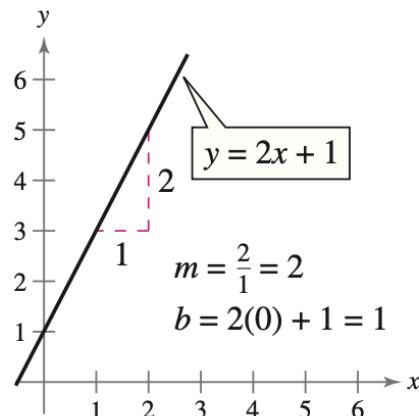
Agora vamos partir para o nosso matematiquês. Porém, não se preocupem pois esse matematiquês será apenas para te explicar como funciona a regressão - nunca calcularemos isso "na mão".

Já mostramos anteriormente a equação da regressão linear, mas aqui está uma nova versão daquela mesma equação

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

onde  $\alpha$  é o intercepto,  $\beta$  é a inclinação da linha e  $y$  é a saída prevista (ajustado). Os valores ótimos de  $\alpha$  e  $\beta$  precisam ser encontrados, para então, termos a equação da reta  $y$ . Tendo a equação da reta, para qualquer valor de  $x$  somos capazes de calcular o  $y$  ajustado.

A inclinação de uma reta é a razão da variação de  $y$  sobre a variação de  $x$ . O intercepto no eixo  $y$  é o valor de  $y$  no ponto onde a reta cruza esse eixo. Isto é, o valor de  $y$  quando  $x = 0$ . Por exemplo, o gráfico de  $y = 2x + 1$  é mostrado abaixo. A inclinação da reta é 2 e o intercepto em  $y$  é 1.



Em álgebra, usamos dois pontos para determinar a equação de uma reta. Em estatística, você vai usar todos os pontos do conjunto de dados para determinar a equação da reta de regressão.

A fórmula do OLS para encontrar os coeficientes envolve cálculos de derivadas e não abordaremos aqui. Porém, vamos mostrar a solução.

$$\hat{\beta} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \sum_{i=1}^n (y_i - \hat{\beta} * x_i) \Rightarrow \hat{\alpha} = \bar{y} - \hat{\beta} * \bar{x}$$

Esta é a solução de Mínimos Quadrados Ordinários (OLS) - que é a solução analítica.

Apenas para exemplificar, vamos voltar nesse exemplo:

Coefficients					
Term	Coef	SE Coef	T	P	
Constant	483.670	39.5671	12.2241	0.000	
IQ	4.796	0.9511	5.0429	0.000	
Education	24.215	1.9405	12.4785	0.000	

Nesse caso, o processo do OLS foi feito. O que temos é uma reta do tipo:

$$\text{Renda} = 483.67 + 4.796 * \text{IQ} + 24.215 * \text{Education}$$

Lembra que eu falei que vocês podem fazer predição com a regressão linear? É aqui que entra! Se você souber o QI de uma pessoa e o Education dela, basta substituir na fórmula acima que você terá uma aproximação da renda.

É claro que nenhum modelo é perfeito! Você viram ali que não dá para ajustar uma reta **PERFEITA** nos pontos, pois eles mesmos não formam uma reta. Mas com isso você terá um valor aproximado da renda.

### **E quando podemos usar a regressão linear?**

Você **sempre** pode usar a regressão linear para fazer previsões, **mas nem sempre pode usar para fazer inferência!** Você só poderá usar para fazer as inferências se algumas premissas forem cumpridas.

Vai parecer um pouco estranho mas para entender se suas premissas forem cumpridas, você terá que fazer uma regressão linear. É o que chamamos de processo iterativo. Primeiro vamos traçar um modelo, depois vamos dar uma olhada na saída estatística, e verificar o resíduo. A partir dessa análise do resíduo saberemos se podemos usar essa regressão ou não. Isso acontece pois várias premissas para podermos usar o OLS dependem do erro (resíduo), o qual só obtemos se traçarmos o modelo.

Antes de falarmos sobre as premissas, quero falar sobre os resíduos. Os gráficos residuais exibem os valores residuais no eixo y e os valores ajustados e ordem de tempo ou outra variável no eixo x. Se os resíduos mostrarem padrões em vez de aleatoriedade, você não pode confiar nos coeficientes de regressão e outros resultados numéricos. Vamos entender isso



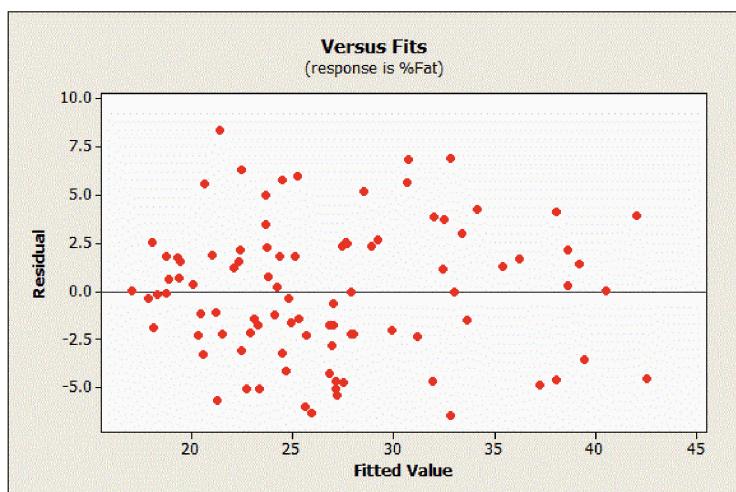
Ao olhar para gráficos de resíduos, você simplesmente quer determinar se os resíduos são consistentes com o erro aleatório. Vou usar uma analogia de rolar um dado. Você não deve ser capaz de usar uma rolagem para prever o resultado da próxima rolagem porque é essa próxima rolagem é aleatório. Portanto, se você anotar várias rolagens, verá apenas resultados aleatórios. Se você começar a ver padrões, saberá que algo está errado com seu modelo de como o dado funciona. Você acha que é aleatório, mas não é. Se você fosse um jogador, usaria essas informações para ajustar como joga para corresponder melhor aos resultados reais do dado.

Você também pode aplicar essa ideia a modelos de regressão. Se você observar uma série de erros, essa série deveria ser aleatória. Se houver padrões nos erros, você poderá usar um erro para prever outro. Tal como acontece com a analogia do dado, se existem padrões nos resíduos, você precisa ajustar seu modelo. Mas, não se preocupe, isso significa apenas que você pode melhorar o ajuste do modelo movendo essa previsibilidade para o lado determinístico das coisas (ou seja, suas variáveis independentes).

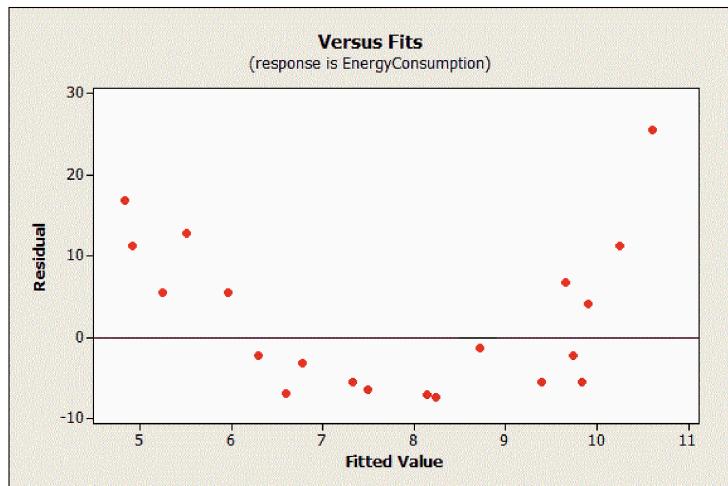
#### Oseias Dias de Farias

Como você determina se os resíduos são aleatórios na análise de regressão? É bem simples, basta verificar se eles estão espalhados aleatoriamente em torno de zero para todo o intervalo de valores ajustados. Os testes de hipótese em regressão assumem que os resíduos seguem uma distribuição normal e que o grau de espalhamento é o mesmo para todos os valores ajustados.

O resíduo deveria se parecer com isso:



Agora observem o gráfico abaixo:



O gráfico acima claramente tem um padrão! Para o modelo acima, se você souber o valor ajustado, poderá usá-lo para prever o resíduo. Por exemplo, valores ajustados próximos de 5 e 10 tendem a ter resíduos positivos. Valores ajustados próximos a 7 tendem a ter valores negativos. Se eles fossem realmente aleatórios, você não seria capaz de fazer essas previsões.

Vamos ver agora as 7 premissas para podemos usar a inferência quando temos uma regressão linear. Novamente, essas premissas são importantes apenas se você quiser analisar as inferências de uma regressão (p-valor, etc). Se seu intuito com a regressão é predição, então essas premissas não serão relevantes e você não precisa satisfazê-las - exceto uma análise de resíduo, mas falaremos sobre isso quando falarmos sobre R-quadrado

## 1 - LINEARIDADE

Em estatística, um modelo de regressão é linear quando as variáveis dependentes e a independente tem uma relação linear. Isso nos deixa com uma equação tipo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon$$

Na equação, os betas ( $\beta$ s) são os parâmetros que o OLS estima. Epsilon ( $\epsilon$ ) é o erro aleatório

Veja que os valores de X não tem exponenciais, ou seja, se alterarmos X o Y vai ser alterado linearmente.

Se você ajustar um modelo linear a um dado não linearmente relacionado, o modelo será incorreto e, portanto, não confiável - p-valores não confiáveis. Ao usar o modelo para extração, é provável que você obtenha resultados errôneos.

E aqui eu já te adianto que se você não tiver uma relação linear, muito provavelmente sua predição também não será muito boa. Mas fique tranquilo que você já já vai descobrir como analisar se sua predição é boa ou não

## 2 - O TERMO DE ERRO TEM UMA MÉDIA POPULACIONAL DE ZERO

Esse é exatamente o ponto que falamos acima sobre a aleatoriedade do resíduo. O termo de erro é o número que considera qualquer variação no "Y", ou variável dependente, que as variáveis independentes não conseguem mostrar. Em circunstâncias ideais, o acaso determina o valor do termo de erro. Para atender a essa expectativa, uma boa suposição é que a média populacional do termo de erro seja igual a zero.

## 3 - AS VARIÁVEIS INDEPENDENTES NÃO ESTÃO CORRELACIONADAS COM O TERMO DE ERRO

Se uma variável independente está correlacionada com o termo de erro, podemos usar a variável independente para prever o termo de erro, o que viola a noção de que o termo de erro representa um erro aleatório imprevisível.

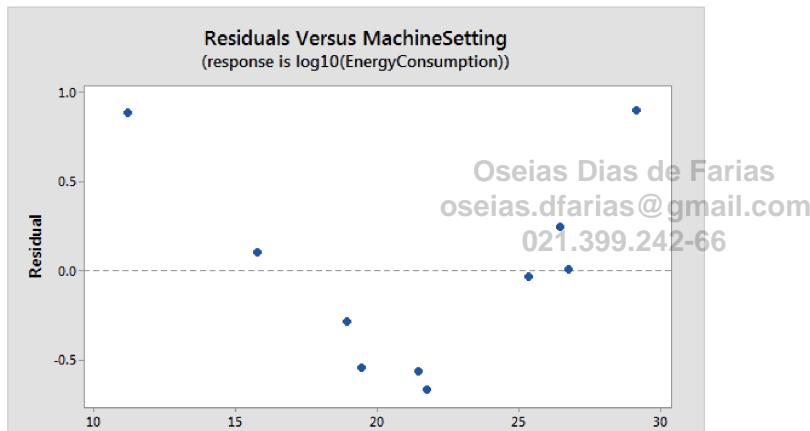
Essa suposição também é chamada de **exogeneidade (exogeneity)**. Por outro lado, quando existe esse tipo de correlação, que viola a suposição, há **endogeneidade (endogeneity)**.

Violações dessa suposição acontecem quando há viés de variável confundidora, curvatura modelada incorretamente ou erro de medição nas variáveis independentes.

A violação dessa suposição distorce a estimativa do coeficiente. Para entender por que esse viés ocorre, lembre-se de que o termo de erro sempre explica parte da variabilidade na variável dependente. No entanto, quando uma variável independente é correlacionada com o termo de erro, OLS atribui incorretamente parte da variância que o termo de erro realmente explica à variável independente.

Para verificar essa suposição, faça um gráfico dos resíduos por cada variável. O gráfico deve exibir aquela boa aleatoriedade que comentei anteriormente. Se houver um padrão, seu modelo tem um problema.

O gráfico de resíduos abaixo mostra um exemplo em que o Consumo de Energia (a variável independente) se correlaciona com os resíduo.

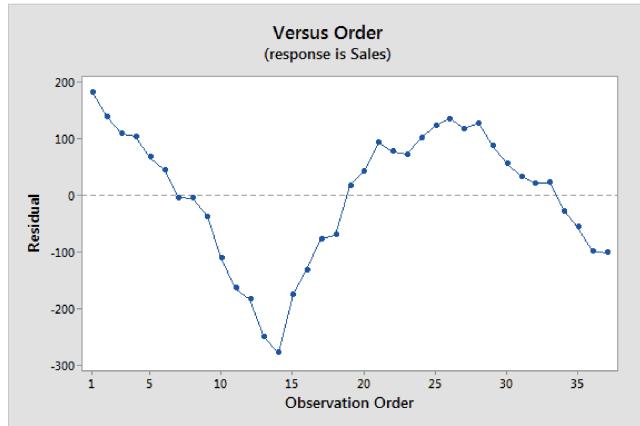


#### 4 - AS OBSERVAÇÕES DO TERMO DE ERRO NÃO SÃO CORRELACIONADAS ENTRE SI

Uma observação do termo de erro não deve prever a próxima observação. Por exemplo, se o erro de uma observação for positivo e isso aumentar sistematicamente a probabilidade de que o erro seguinte seja positivo, essa é uma correlação positiva. Se for mais provável que o erro subsequente tenha o sinal oposto, essa é uma correlação negativa.

Por exemplo, se as vendas forem inesperadamente altas em um dia, provavelmente serão mais altas que a média no dia seguinte. Essa não é uma expectativa irracional para algumas áreas temáticas, como taxas de inflação, PIB, desemprego e assim por diante.

Avalie essa suposição fazendo um gráfico dos resíduos na ordem em que os dados foram coletados. No gráfico para um modelo de vendas abaixo, parece haver um padrão cíclico com um correlação positiva



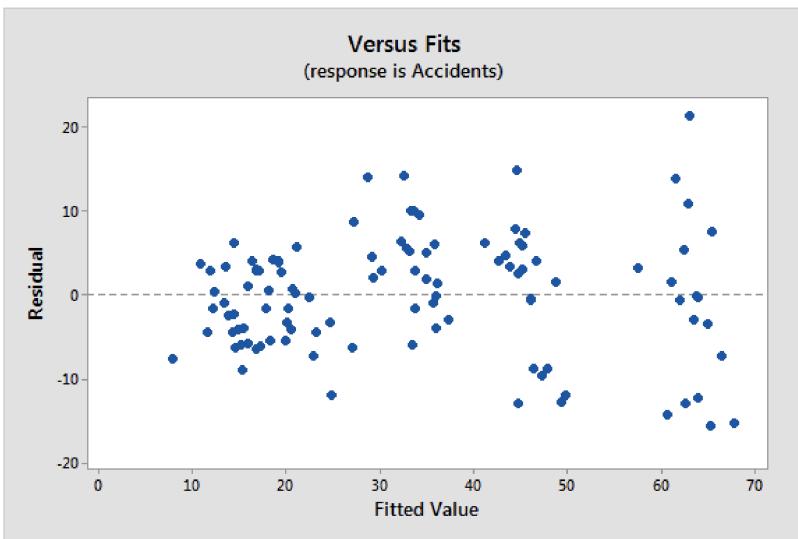
Para o modelo de vendas acima, precisamos adicionar variáveis que explicam o padrão cíclico.

Oseias Dias de Farias  
**5 - O TERMO DE ERRO TEM UMA VARIÂNCIA CONSTANTE ( HOMOCEDASTICIDADE)**  
021.399.242-66

A variância dos erros deve ser consistente para todas as observações. Em outras palavras, a variância não muda para cada observação ou para uma série de observações. Esta condição é conhecida como **homocedasticidade** (mesma dispersão). Se a variação mudar, nos referimos a isso como **heterocedasticidade** (dispersão diferente).

A maneira mais fácil de verificar essa suposição é criar uma relação de resíduos versus gráfico de valores ajustados. Nesse tipo de gráfico, a heterocedasticidade aparece como uma forma de cone onde a dispersão dos resíduos aumenta em uma direção.

No gráfico abaixo, a dispersão spread dos resíduos aumenta à medida que os valores ajustados aumentam.



Especificamente, a heterocedasticidade é uma mudança sistemática na dispersão dos resíduos ao longo do intervalo de valores medidos. A heterocedasticidade é um problema porque a regressão dos mínimos quadrados ordinários (OLS) assume que todos os resíduos são extraídos de uma população que tem uma variância constante (homocedasticidade).

[cessias.maria@gmail.com](mailto:cessias.maria@gmail.com)

021.399.242-66

A heterocedasticidade ocorre com mais frequência em conjuntos de dados que têm um grande intervalo entre o maior e o menor valores observados. Embora existam inúmeras razões pelas quais a heterocedasticidade pode existir, uma explicação comum é que a variância do erro muda proporcionalmente com um fator. Este fator pode ser uma variável no modelo. Em alguns casos, a variância aumenta proporcionalmente com esse fator, mas permanece constante como uma porcentagem.

Por exemplo, uma mudança de 10% em um número como 100 é muito menor do que uma mudança de 10% em um número grande como 100.000. Nesse cenário, você espera ver resíduos maiores associados a valores mais altos. É por isso que você precisa ter cuidado ao trabalhar com amplas faixas de valores!

Vamos dar uma olhada em um exemplo clássico de heterocedasticidade. Se você modelar o consumo das famílias com base na renda, descobrirá que a variabilidade no consumo aumenta à medida que a renda aumenta. As famílias de baixa renda são menos variáveis em termos absolutos porque



precisam se concentrar nas necessidades e há menos espaço para diferentes hábitos de consumo. As famílias de maior renda podem comprar uma grande variedade de itens de luxo, ou não, o que resulta em uma maior disseminação dos hábitos de consumo.

Sempre que você violar uma suposição do OLS, há uma chance de que você não possa confiar nos resultados estatísticos.

Por que corrigir esse problema? Existem duas grandes razões pelas quais você deseja a homocedasticidade:

- Embora a heterocedasticidade não cause viés nas estimativas dos coeficientes, ela as torna menos precisas. A menor precisão aumenta a probabilidade de que as estimativas dos coeficientes sejam mais longe do valor populacional correto.
- A heterocedasticidade tende a produzir p-valores menores do que deveriam ser. Esse efeito ocorre porque a heterocedasticidade aumenta a variância das estimativas dos coeficientes, mas o procedimento OLS não detecta esse aumento. Este problema pode levar você a concluir que um termo de modelo é estatisticamente significativo quando na verdade não é significativo.

## **6 - NENHUMA VARIÁVEL INDEPENDENTE É UMA FUNÇÃO LINEAR PERFEITA DE OUTRAS VARIÁVEIS EXPLICATIVAS (SEM MULTICOLINEARIDADE)**

Essa é uma das premissas que pode ser avaliada antes de termos um modelo, pois ela não depende do erro.

A correlação perfeita ocorre quando duas variáveis têm um coeficiente de correlação de Pearson de +1 ou -1. Quando uma das variáveis muda, outra variável também muda em uma proporção completamente fixa.

A correlação perfeita sugere que duas variáveis são formas diferentes de a mesma variável. Por exemplo, jogos ganhos e jogos perdidos têm uma correlação negativa perfeita (-1). A temperatura em Fahrenheit e Celsius tem uma correlação positiva perfeita (+1).

Os mínimos quadrados não conseguem distinguir uma variável da outra quando estão perfeitamente correlacionadas. Se você especificar um modelo

que contenha variáveis independentes com correlação perfeita, seu software estatístico não poderá ajustar o modelo e exibirá uma mensagem de erro.

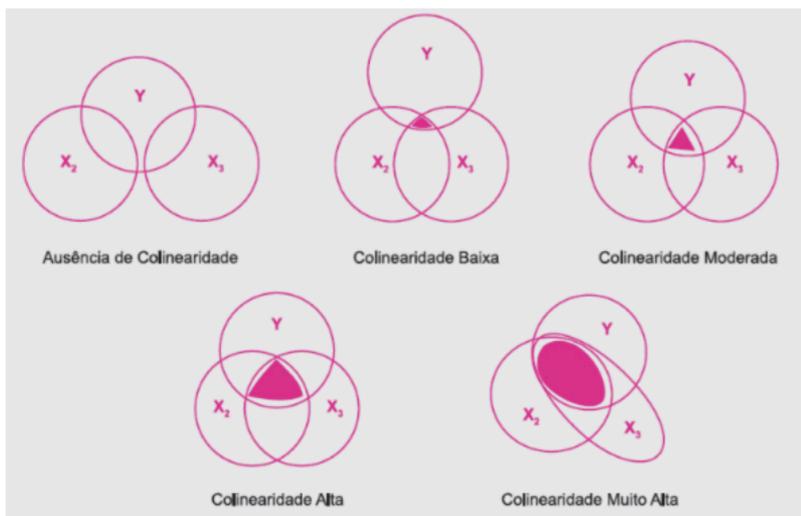
No entanto, seu software estatístico pode ajustar modelos de regressão OLS com relacionamentos imperfeitos, mas fortes, entre as variáveis independentes. Se essas correlações forem altas o suficiente, elas podem causar problemas. Os estatísticos referem-se a essa condição como **multicolinearidade**, e ela reduz a precisão das estimativas na regressão linear.

**E por que a multicolinearidade é um problema?** Um dos principais objetivos da análise de regressão é isolar a relação entre cada variável independente e a variável dependente. A interpretação de um coeficiente de regressão é que ele representa a mudança média na variável dependente para cada mudança de 1 unidade em uma variável quando você mantém todas as outras variáveis independentes constantes.

Essa última parte é crucial para nossa discussão sobre multicolinearidade. A ideia é que você possa alterar o valor de uma variável independente e não os outros. No entanto, quando as variáveis independentes são correlacionadas, isso indica que as mudanças em uma variável estão associadas ao deslocamentos em outra variável. Quanto mais forte a correlação, mais difícil é mudar uma variável sem mudar outra. Torna-se difícil para o modelo estimar a relação entre cada variável independente e a variável dependente independentemente porque as variáveis independentes tendem a mudar em conjunto.

Veja no gráfico abaixo uma visualização prática do conceito de multicolinearidade:





À medida que a colinearidade vai aumentando entre as variáveis preditoras, perceba que o efeito compartilhado entre elas começa a ser refletido em Y de maneira exagerada, tornando difícil detectar o efeito individual de cada uma delas na variável resposta.

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021 329 242 66

**Que problemas a multicolinearidade causa?** A multicolinearidade causa os dois tipos básicos de problemas a seguir:

- As estimativas de coeficiente podem oscilar muito com base em qual outras variáveis independentes estão no modelo. Os coeficientes tornam-se muito sensíveis a pequenas mudanças no modelo.
- A multicolinearidade reduz a precisão dos coeficientes de estimativa, o que enfraquece o poder estatístico de sua regressão. Você pode não ser capaz de confiar nos p-valores para identificar variáveis independentes que são estatisticamente significativas.

#### Teste de multicolinearidade com fatores de inflação de variância (VIFs)

O fator de inflação de variância (VIF) identifica a correlação entre as variáveis independentes e a força dessa correlação.

As ferramentas que estamos usando no curso (Excel e Python) calculam um VIF para cada variável independente. O cálculo é:

$$VIF_i = \frac{1}{1 - R_i^2}$$

O termo R ao quadrado é o coeficiente de determinação, que representa a proporção da variação na variável dependente ou resposta que é explicada pela(s) variável(is) independente(s). Veremos o que ele significa mais para frente.

O VIF somente pode ser usado para variáveis numéricas e equações lineares.

Os VIFs começam em 1 e não têm limite superior. **Um valor de 1 indica que não há correlação entre esta variável independente e quaisquer outras.** **VIFs entre 1 e 5 sugerem que há uma correlação moderada**, mas não é grave o suficiente para justificar medidas corretivas. **VIFs maiores que 5 representam níveis críticos de multicolinearidade** onde os coeficientes são mal estimados e os p-valores são questionáveis.

A avaliação de VIFs é particularmente importante para estudos observacionais porque esses estudos são mais propensos a ter multicolinearidade.

Vamos a um exemplo. Vou usar aqui uma análise de regressão para modelar a relação entre as variáveis independentes (atividade física, percentual de gordura corporal, peso, e uma variável criada que é %gordura\*peso) e a variável dependente (densidade mineral óssea do colo do fêmur). Fizemos a análise e obtivemos:



## Regression Analysis: Femoral Neck versus %Fat, Weight kg, Activity

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.555785	0.138946	27.95	0.000
%Fat	1	0.009240	0.009240	1.86	0.176
Weight kg	1	0.127942	0.127942	25.73	0.000
Activity	1	0.047027	0.047027	9.46	0.003
%Fat*Weight kg	1	0.041745	0.041745	8.40	0.005
Error	87	0.432557	0.004972		
Total	91	0.988342			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.155	0.132	1.18	0.243	
%Fat	0.00557	0.00409	1.36	0.176	14.93
Weight kg	0.01447	0.00285	5.07	0.000	33.95
Activity	0.000022	0.000007	3.08	0.003	1.05
%Fat*Weight kg	-0.000214	0.000074	-2.90	0.005	75.06

Oseias Dias de Farias  
oseias.dfarias@gmail.com

021.399.242-66

**P-valor:** Esses resultados mostram que Peso (Weight), Atividade (Activity) e %gordura\*peso (%fat\*weight) são estatisticamente significativos, pois p-valor abaixo de 0,05.

No entanto, os **VIFs** indicam que nosso modelo tem multicolinearidade severa para algumas das variáveis independentes. Observe que Activity tem um VIF próximo de 1, o que mostra que a multicolinearidade não a afeta e podemos confiar nesse coeficiente e p-valor sem mais nenhuma ação. No entanto, os coeficientes e p-valor para os outros termos são suspeitos!

Além disso, pelo menos parte da multicolinearidade em nosso modelo é a tipo estrutura pois incluímos o termo de interação%gordura\*peso (%fat\*weight). Claramente, há uma correlação entre o termo de interação e ambos os termos do efeito principal.

## Como lidar com a multicolinearidade?



1. Remova alguma das variáveis independentes altamente correlacionadas. Você pode fazer uma correlação de Pearson para todas as variáveis, duas a duas, para ver quais se correlacionam.
2. Combine linearmente as variáveis independentes, então você pode adicioná-las em conjunto.
3. A regressão LASSO e Ridge são formas avançadas de análise de regressão que estão além do escopo deste curso, mas podem lidar com multicolinearidade. Se você souber como executar a regressão linear de mínimos quadrados, poderá lidar com essas análises com apenas um pouco de estudo adicional.

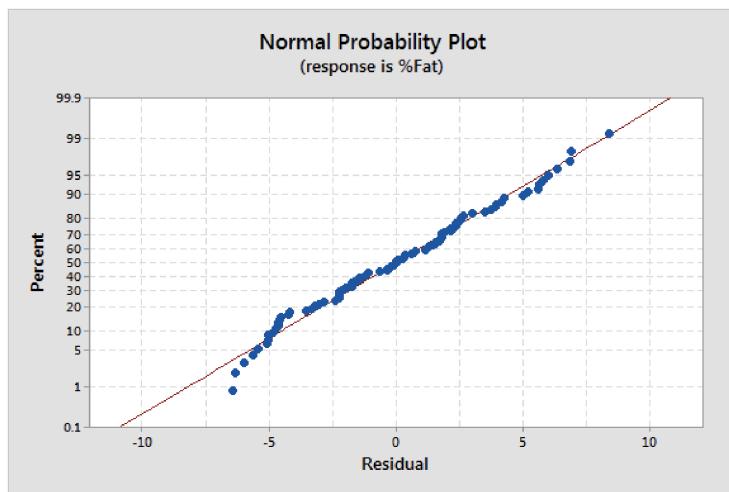
Ao considerar uma solução, lembre-se de que todas elas têm desvantagens. Se você pode aceitar coeficientes menos precisos ou um modelo de regressão com um alto R-quadrado (veremos mais para frente o que é), mas quase nenhuma variável estatisticamente significativa, então não fazer nada sobre a multicolinearidade pode ser a melhor solução.

**LEIA MAIS SOBRE LASSO E RIDGE AQUI:** Oseias Dias de Farias  
[HTTPS://MEDIUM.COM/TURING-TALKS/TURING-TALKS-20-REGRESS%C3%A3O-DE-RIDGE-E-LASSO-A0FC467B5629](https://medium.com/turing-talks/turing-talks-20-regress%C3%A3o-de-ridge-e-lasso-a0fc467b5629)  
021.399.242-66

## 7 - O TERMO DE ERRO É NORMALMENTE DISTRIBUÍDO (OPCIONAL)

OLS não exige que o termo de erro siga uma distribuição normal. No entanto, satisfazer essa suposição permite que você realize testes de hipótese (analisar o p-valor) e gerar intervalos de confiança e intervalos de previsão confiáveis.

A maneira mais fácil de determinar se os resíduos seguem um padrão normal é avaliar um gráfico de probabilidade normal. Se os resíduos seguem a linha reta neste tipo de gráfico, eles são normalmente distribuídos.



## ENTENDENDO SE NOSSA REGRESSÃO LINEAR FOI BEM SUCEDIDA

Existe uma métrica muito valiosa quando falamos sobre regressão linear, mas antes de falarmos sobre ela, precisamos definir alguns conceitos.

Oseias Dias de Farias

[oseias\\_dfarias@gmail.com](mailto:oseias_dfarias@gmail.com)  
021.399.242-66

**Total Sum of Squares (SST)** | Soma Total de Quadrados (SST): É literalmente a variabilidade da variável dependente (volte lá na estatística descritiva, se necessário). É a soma do quadrado distâncias entre os valores observados e a média da variável dependente - quanto distante cada valor está da média-

$$\sum_{i=1}^n (y_i - \bar{y})^2$$

**Regression Sum of Squares (SSR)** | Soma de Regressão ao quadrado: E se ao invés do dado real tivéssemos o dado que nosso modelo previu? Essa é a variabilidade que seu modelo explica sobre seu dado. Se o SSR fosse zero, isso significaria que não há variabilidade explicada pelo modelo de regressão. Em outras palavras, o modelo não está capturando ou explicando nenhuma parte da variação nos dados. Isso pode indicar que o modelo não está ajustando corretamente aos dados ou que não está conseguindo capturar a relação entre as variáveis independentes e dependentes.



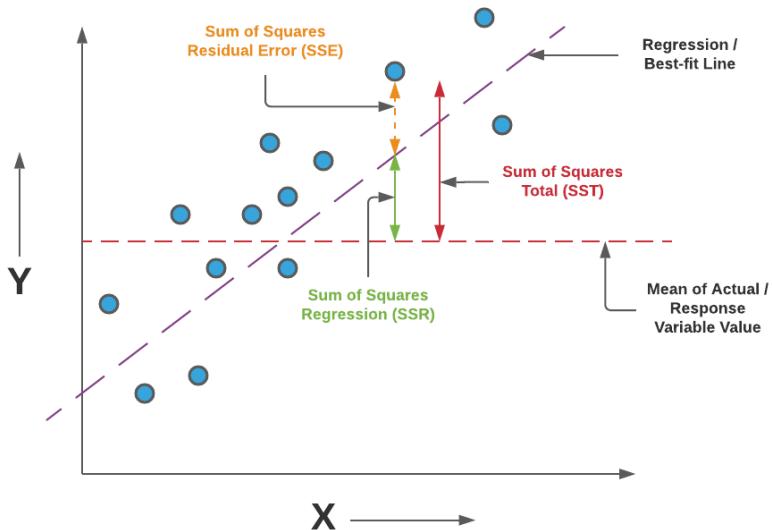
$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

**Sum of Squared Errors (SSE)** | Esta é a variabilidade dos dados que **não** é explicada pelo modelo de regressão. É a soma dos quadrados das diferenças entre os valores observados e os valores previstos pelo modelo. Se o SSE fosse zero, isso significaria que o modelo se ajusta perfeitamente aos dados, explicando toda a variabilidade presente. No entanto, na prática, sempre haverá algum erro residual devido à aleatoriedade ou à complexidade dos dados.

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Oseias Dias de Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

Veja a imagem abaixo com todas essas somas



Note que todas essas somas tem uma relação entre si. Se somarmos o SSR com o SSE, teremos no final das contas o SST

$$SST = SSE + SSR$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- SSE representa a variabilidade que seu modelo não explica. Quanto menor, melhor
- RSS representa a variabilidade que seu modelo explica. Quanto maior, melhor
- SST representa a variabilidade inerente a sua variável dependente.

## R-QUADRADO ( $R^2$ )

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

O  **$R^2$  (R-quadrado)**, também chamado de coeficiente de determinação, é uma das formas de entender se o seu modelo está bem ajustado ou não. A fórmula do coeficiente de determinação é uma medida estatística que indica a quanto da variabilidade total é explicada pelo seu modelo

$$R^2 = \frac{\text{Variance explained by the model}}{\text{Total variance}}$$

Em outras palavras, usando as definições de erro que vimos acima:

$$R^2 = \frac{SSR}{SST} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$

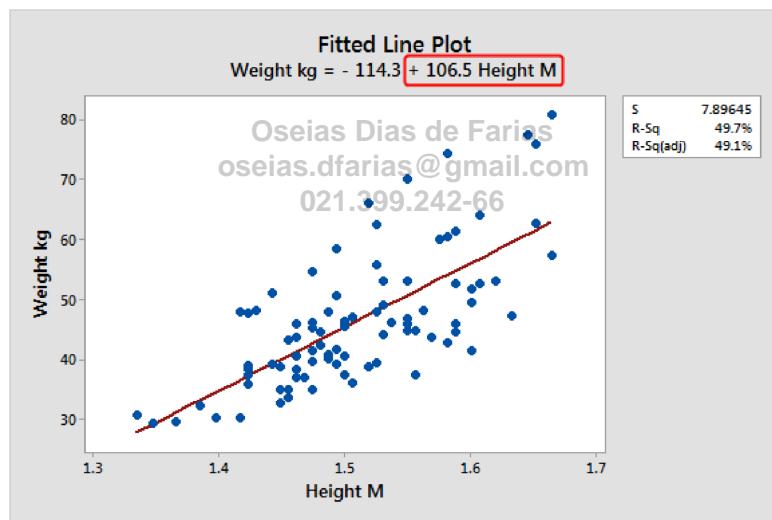
Lembrando que SSR é a variabilidade do dado que seu modelo explica. Se ela fosse zero, seu modelo não estaria explicando a variabilidade do dado e seu



R-quadrado seria 0. Se o SSR/SST fosse 1, ou seja, a variabilidade do dado que seu modelo explica é igual a variabilidade total, seu R-quadrado seria 1. Isso geralmente é considerado bom, pois o modelo está ajustando bem aos dados e capturando a relação entre as variáveis independentes e dependentes.

O  $R^2$  varia entre 0 e 1, por vezes sendo expresso em termos percentuais. Ele expressa a quantidade da variância dos dados que é explicada pelo modelo linear. Assim, quanto maior o  $R^2$ , mais explicativo é o modelo linear, ou seja, melhor ele se ajusta à amostra. Por exemplo, um  $R^2 = 0,8234$  significa que o modelo linear explica 82,34% da variância da variável dependente a partir dos regressores (variáveis independentes) incluídas naquele modelo linear.

Voltando ao nosso problema de peso x altura, teríamos a seguinte reta de ajuste:



Esta linha produz um SSR maior do que qualquer outra linha que você possa desenhar através dessas observações.

Visualmente, vemos que a linha ajustada tem uma inclinação positiva que corresponde à correlação positiva que obtivemos anteriormente. A linha segue os pontos de dados, o que indica que o modelo se ajusta aos dados.

A inclinação da linha é igual ao coeficiente circulado em vermelho. Este coeficiente indica quanto peso médio tende a aumentar à medida que aumentamos a altura.

Também poderíamos inserir um valor de altura na equação e obter uma previsão para o peso médio.

Cada ponto na linha ajustada representa o peso médio para uma determinada altura. No entanto, como qualquer média, há variabilidade em torno da média. Observe como há uma dispersão de pontos de dados ao redor da linha. Você pode avaliar essa variabilidade escolhendo um ponto na linha e observando o intervalo de pontos de dados acima e abaixo desse ponto. Por fim, a distância entre cada ponto de dados e a linha é o resíduo para essa observação.

### **Valores baixos de R-quadrado são sempre um problema?**

Não! Modelos de regressão com baixos valores de R-quadrado podem ser perfeitamente bons modelos por várias razões.

Alguns campos de estudo têm uma quantidade inherentemente maior de variação inexplicável. Nessas áreas, seus valores de R<sup>2</sup> serão menores.

Oseias Dias de Farias

[oseias\\_dias\\_farias@gmail.com](mailto:oseias_dias_farias@gmail.com)

021 399 242-66

Por exemplo, estudos que tentam explicar o comportamento humano geralmente têm valores de R<sup>2</sup> inferiores a 50%. As pessoas são apenas mais difíceis de prever do que coisas como processos físicos.

Felizmente, se você tiver um valor baixo de R-quadrado, mas o valor de variáveis independentes são estatisticamente significativas, você ainda pode tirar conclusões importantes sobre as relações entre as variáveis.

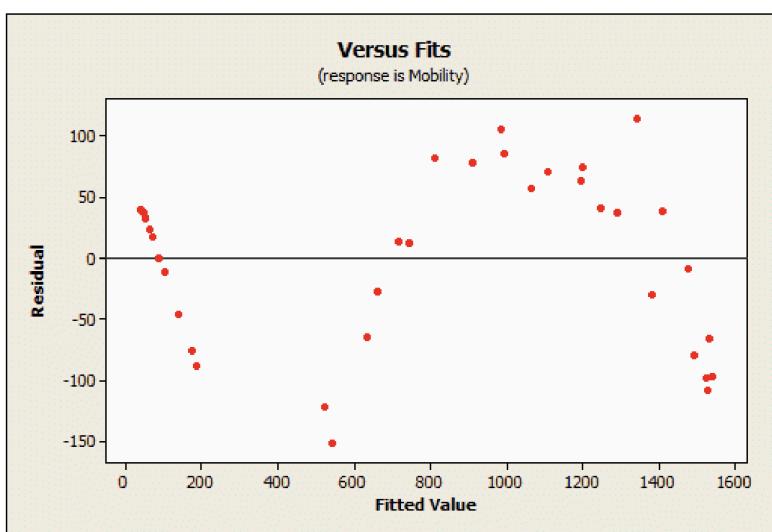
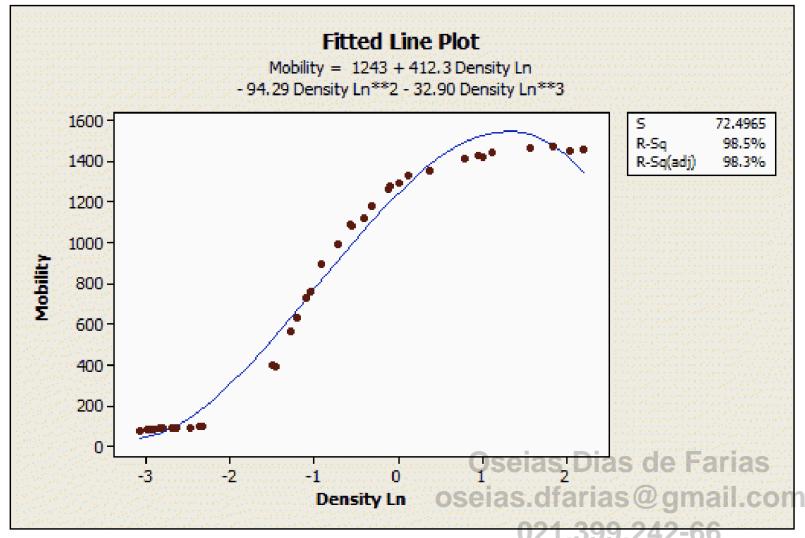
Estatisticamente coeficientes significativos continuam a representar a mudança na variável dependente dado um deslocamento de uma unidade na variável independente. Ser capaz de tirar conclusões como esta é vital. Em poucas palavras, se seu objetivo principal é entender a natureza dos relacionamentos em seus dados, um R-quadrado baixo provavelmente não é um problema!

Existe um cenário em que pequenos valores de R-quadrado podem causar problemas. Se você precisar gerar previsões relativamente precisas (estreitos intervalos de previsão), um R<sup>2</sup> baixo pode ser um obstáculo.

### **Valores altos de R-quadrado são sempre ótimos?**



Não! Um modelo de regressão com um alto valor de R-quadrado pode ter vários problemas. Você provavelmente espera que um R<sup>2</sup> alto indique um bom modelo, mas examine os gráficos a seguir. O gráfico de linha ajustada modela a associação entre mobilidade eletrônica e densidade.



Os dados no gráfico com a linha ajustada seguem uma relação de ruído muito baixo, e o R-quadrado é de 98,5%, o que parece fantástico. No entanto, a linha de regressão consistentemente, em faixas determinadas de valores,

superestima e subestima os dados ao longo do curva, que é enviesada. O correto seria algo mais aleatório em torno da primeira curva.

O gráfico Resíduos versus Ajustes enfatiza isso com o padrão indesejado. Um modelo não enviesado tem resíduos que são aleatoriamente espalhados em torno de zero. Padrões residuais não aleatórios indicam um ajuste ruim apesar de um R<sup>2</sup> alto. Verifique sempre seus resíduos!

Esse tipo de viés ocorre quando seu modelo linear é subespecificado. Em outras palavras, faltam variáveis independentes significativas, ou termos polinomiais. Para produzir resíduos aleatórios, tente adicionar termos ao modelo ou ajustar a um modelo não linear (polinomial, por exemplo).

Lembra que falamos que previsões não precisam que premissas sejam satisfeitas? De fato não precisam, porém é muito bom você olhar para o resíduo justamente para identificar esse tipo de viés.

### R-quadrado nem sempre é simples

Oseias Dias de Farias

[oseias\\_dfarias@gmail.com](mailto:oseias_dfarias@gmail.com)  
021 399 242-66

À primeira vista, o R-quadrado parece uma estatística fácil de entender que indica quão bem um modelo de regressão se ajusta a um conjunto de dados, no entanto não nos conta toda a história. Para obter uma ideia completa, você deve considerar os valores de R<sup>2</sup> em combinação com gráficos de resíduos, outras estatísticas e um profundo conhecimento do assunto.

### Problemas do R-quadrado

Toda vez que você adiciona uma variável para o modelo o R-quadrado aumenta, mesmo que essa variável não seja boa para seu modelo.

Um modelo de regressão que contém mais variáveis independentes do que outro modelo pode parecer que fornece um ajuste melhor simplesmente porque contém mais variáveis. Mas isso nem sempre é verdade, pois aquela variável pode apenas ser um ruído. O R-quadrado não detecta isso

Quando um modelo contém um número excessivo de variáveis independentes e termos polinomiais, ele se torna excessivamente customizado para e tende a se ajustar às peculiaridades dos seus dados, inclusive aos ruídos aleatórios em sua amostra, ao invés de refletir toda a população. O que no fundo ele está fazendo é decorando seus dados da sua

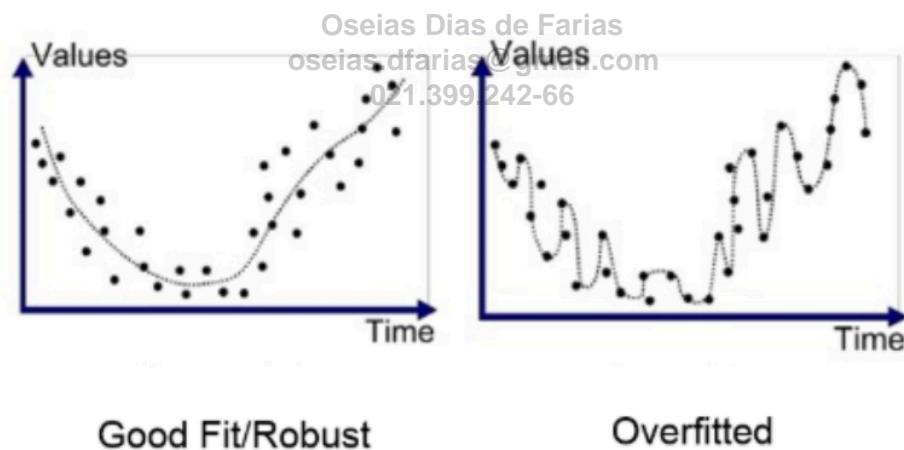


amostra, e não aprendendo o padrão para generalizamos para uma população. Chamamos isso de **overfitting**. Esse overfitting produz valores de R-quadrado enganosamente altos e uma capacidade reduzida para previsões precisas futuras.

A principal característica a ser considerada em qualquer modelo é a sua capacidade de **generalização**. Ou seja, ao receber novos conjuntos de dados referentes ao problema para o qual ele foi construído, ele deve ser capaz de fazer previsões coerentes acerca desse conjunto, previsões que se aproximam bem com a realidade.

O overfitting ocorre principalmente quando o modelo, ao invés de generalizar (“aprender”), “memoriza” os dados que recebeu durante o treinamento. Ou seja, ele não consegue fazer previsões de dados que nunca viu antes, pois ele não aprendeu um padrão geral de reconhecimento.

Veja como isso se reflete graficamente:



## R-QUADRADO AJUSTADO ( $R^2$ )

Devido aos problemas do R-quadrado, usamos o R-quadrado ajustado para comparar a qualidade do ajuste para regressão que contêm números diferentes de variáveis independentes.

Digamos que você esteja comparando um modelo com 5 variáveis independentes com um modelo com 1 variável e o modelo de 5 variáveis tenha uma maior R-quadrado. O modelo com cinco variáveis é realmente um

modelo melhor, ou só tem mais variáveis? Para determinar isso, basta comparar os valores de R-quadrado ajustado!

O R-quadrado ajustado se ajusta ao número de termos no modelo. É importante ressaltar que seu valor aumenta apenas quando o novo termo melhora o ajuste do seu modelo mais do que o esperado. O valor R-quadrado ajustado na verdade diminui quando o termo não melhora o modelo.

O exemplo abaixo mostra como o R-quadrado ajustado aumenta até um ponto e depois diminui. Por outro lado, R-quadrado continua aumentando com cada variável independente adicional.

Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7

Oseias Dias de Farias

Neste exemplo, deveríamos apenas incluir 3 variáveis independentes no modelo de regressão.

[oseias\\_dfarias@gmail.com](mailto:oseias_dfarias@gmail.com)  
021.399.242-66

## INTERVALO DE CONFIANÇA PARA REGRESSÃO

No contexto de regressão, lembre-se de que também estamos usando nossa amostra para calcular os coeficientes de regressão (os Ws, da equação de reta), que são as estimativas pontuais dos parâmetros populacionais. A estimativa da nossa amostra do coeficiente de altura é 106,5. No entanto, se coletarmos várias amostras aleatórias da mesma população, cada amostra produzirá sua estimativa própria para o coeficiente de altura. E, nós não sabemos o valor verdadeiro. Por isso, calculamos um intervalo de confiança para cada coeficiente!

Por exemplo, para o caso de peso x altura, temos:



#### Coefficients

Term	Coef	SE Coef	95% CI	T-Value	P-Value	VIF
Constant	-114.3	17.4	(-149.0, -79.7)	-6.55	0.000	
Height M	106.5	11.6	( 83.5, 129.5)	9.22	0.000	1.00

Podemos estar 95% confiantes de que o valor real da população para o coeficiente de altura está entre 83,5 e 129,5.

Quando um IC **exclui** zero, os resultados são estatisticamente significativos. Um intervalo de confiança de 95% sempre concordará com um teste de hipótese que usa um nível de significância de 0,05, como já vimos em teste de hipótese.

Na saída acima, o IC exclui zero, que corresponde ao p-valor (0,000) que é menor que o nível de significância (0,05). Ou seja, o coeficiente de altura é estatisticamente significativo.

A largura de um intervalo de confiança revela a precisão da estimativa. Faixas mais estreitas sugerem uma estimativa mais precisa.

021.399.242-66

## TESTE F E TESTE T NO CONTEXTO DE REGRESSÃO

O teste F de significância geral (tabela "Analysis of Variance" acima) indica se seu modelo de regressão linear fornece um ajuste melhor aos dados do que um modelo que não contém variáveis independentes.

O R-quadrado lhe diz quão bem seu modelo se ajusta aos dados - muito útil na previsão - e o teste F está relacionado a ele.

Um teste F é um tipo de teste estatístico muito flexível. Você pode usá-lo em uma ampla variedade de configurações.

Os testes F podem avaliar muitas variáveis simultaneamente, o que lhes permite comparar os ajustes de diferentes modelos lineares. Em contraste, os testes t (tabela "Coefficients" acima) podem avaliar apenas um termo de cada vez.

O teste F geral compara o modelo que você especifica com o modelo com sem variáveis independentes, que também é conhecido como interceptação.



Como já dissemos, o teste F tem as duas hipóteses a seguir:

- A hipótese nula afirma que o modelo sem variáveis independentes se ajusta aos dados tão bem quanto ao seu modelo.
- A hipótese alternativa diz que seu modelo se ajusta aos dados melhor do que o modelo somente de interceptação.

#### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	12833.9	4278.0	57.87	0.000
East	1	226.3	226.3	3.06	0.092
South	1	2255.1	2255.1	30.51	0.000
North	1	12330.6	12330.6	166.80	0.000
Error	25	1848.1	73.9		
Total	28	14681.9			

Compare o p-valor do teste F com o seu nível de significância. Se o p-valor for menor que o nível de significância, seus dados de amostra fornecem evidências suficientes para concluir que seu modelo de regressão se ajusta aos dados melhor do que o modelo sem variáveis independentes.

021.399.242-66

Isso significa que as variáveis independentes em seu modelo melhoraram o ajuste!

De um modo geral, se nenhuma de suas variáveis independentes for estatisticamente significativa, o teste F também não será estatisticamente significativo.

O teste t é o teste feito em cada um dos coeficientes (East, South, North, no exemplo acima). Lembrando que a variável é significativa quando p-valor for menor do que seu alpha.

Ocasionalmente, os testes t para coeficientes e o teste F geral podem produzir resultados conflitantes. Essa discordância pode ocorrer porque o F-test avalia todos os coeficientes em conjunto enquanto o teste t examina os coeficientes individualmente. Por exemplo, o teste F geral pode descobrir que os coeficientes são significativos em conjunto, enquanto os testes-t podem falhar em encontrar significância individualmente.



Esses resultados de teste conflitantes podem ser difíceis de entender, mas pense sobre isso desta forma. O teste F soma o poder preditivo de todas as variáveis independentes e determina que é improvável que todos os coeficientes sejam iguais a zero. No entanto, é possível que cada variável não seja preditiva o suficiente por si só para ser estatisticamente significativa. Em outras palavras, sua amostra fornece evidências suficientes para concluir que seu modelo é significativo, mas não o suficiente para concluir que qualquer variável individual é significativa. Porém, quando em conjunto elas tem uma força maior, fazendo com que o teste F indique que o modelo é significativo.

## INTERPRETANDO RESULTADOS DE UMA REGRESSÃO

Como já falamos antes, uma variável dependente pode ser descrita por uma ou mais variáveis independentes. Vamos aqui usar um exemplo. Suponha que os custos com ar condicionado possam ser descritos pela temperatura que você coloca no ar condicionado e pelo isolamento do ambiente. Vamos supor que fizemos um modelo e a equação retornada foi:

Oseiás Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

$$\text{Custo com ar condicionado} = 2 * \text{Temperatura} - 1.5 * \text{Isolamento}$$

O custo será em reais, a temperatura em graus celsius e o isolamento em espessura (centímetros)

O sinal do coeficiente para Temperatura é positivo, o que indica uma relação positiva entre Temperatura e Custos. À medida que a temperatura aumenta, o mesmo acontece com os custos de ar condicionado. Mais especificamente, o valor do coeficiente de 2 indica que para cada aumento de 1 grau, a média valor dos custos de ar condicionado aumenta em 2 reais.

Por outro lado, o coeficiente negativo para o isolamento representa uma relação negativa entre os custos de isolamento e ar condicionado. À medida que a espessura do isolamento aumenta, os custos de ar condicionado diminuem. Para cada aumento de 1 centímetro, os custos de ar condicionado caem em 1,50 reais.

Além disso, ambos são efeitos principais, o que indica que se você alterar o valor, digamos, do isolamento, a relação entre os custos da temperatura e da condição do ar permanece a mesma. E são efeitos lineares. Para cada

aumento de 1°C na temperatura, os custos do ar condicionado serão sempre aumentados em R\$ 2. Não importa se a temperatura aumenta de 20 a 21°C ou de 30 a 31°C. Esse grau extra custa R\$ 2!

No entanto, você não pode estender essa interpretação para fora do alcance de seus dados. Se para fazer essa equação você mediu apenas até 30 °C, você não pode presumir que a mesma relação é verdadeira em 35 °C.

Vamos supor agora que fizemos um modelo de regressão qualquer, usando agora 3 variáveis: East, South e North. Obtivemos o seguinte resultado:

Coefficients					
Term	Coef	SE Coef	T	P	
Constant	389.166	66.0937	5.8881	0.000	
East	2.125	1.2145	1.7495	0.092	
South	5.318	0.9629	5.5232	0.000	
North	-24.132	1.8685	-12.9153	0.000	

#### Oseias Dias de Farias

Como vimos em correlação, se o p-valor de uma variável for menor que seu nível de significância, os dados da amostra fornecem evidências suficientes para rejeitar a hipótese nula para toda a população. Logo, seus dados favorecem a hipótese de que **há uma correlação diferente de zero**. Esta variável é estatisticamente significativa e provavelmente uma adição valiosa para seu modelo de regressão.

Por outro lado, um p-valor maior que o nível de significância indica que não há evidências suficientes em sua amostra para concluir que o coeficiente não é igual a zero.

O exemplo de saída de regressão mostra que o South e o North são estatisticamente significativas porque seus p-valores são iguais a 0,000. Por outro lado, East não é estatisticamente significativa porque seu p-valor (0,092) é maior que o nível de significância (0,05). Nesse caso, poderíamos remover a variável East e avaliar como o modelo se comporta (avaliar R ajustado e teste F da regressão)

## PREPARAÇÃO DE DADOS PARA ENTRADA NO MODELO



## SCALING VARIÁVEIS NUMÉRICAS

Os exemplos acima usam os valores brutos das variáveis independentes para se adequar ao modelo. Por exemplo, o modelo de altura e peso usa os valores reais de altura e peso para cada pessoa.

Usar os valores brutos é muitas vezes apropriado, e permite a interpretação mais natural dos resultados. No entanto, muitas vezes precisamos "recodificar" os dados para obter informações valiosas.

A recodificação envolve pegar os valores originais e convertê-los matematicamente em outros valores. Embora esses métodos de recodificação façam com que você interprete alguns dos resultados de forma diferente, os p-valores e todas as medidas de ajuste permanecem as mesmas.

### Padronizando as Variáveis Contínuas

Padronizar seus dados contínuos pode ser útil em algumas circunstâncias. Para padronizar uma variável, você toma cada valor observado para uma variável, subtrai a média da variável e depois divide pelo desvio padrão da variável.

$$x_{scaled} = \frac{x - mean}{sd}$$

Quando você padroniza uma variável, o valor codificado denota onde a observação cai na distribuição total, indicando o número de desvios padrão acima ou abaixo da média da variável. O sinal indica se a observação está acima ou abaixo da média, e o número indica o número de desvios padrão.

Suponha que temos uma medida de comprimento e o comprimento médio é 10 e o desvio padrão é 3.

Vamos padronizar o valor de três observações de comprimento para mostrar como isso funciona: (valor bruto - média variável) / desvio padrão variável

$$16: (16 - 10) / 3 = 2$$

$$10: (10 - 10) / 3 = 0$$

$$7: (7 - 10) / 3 = -1$$

A primeira observação tem um valor de comprimento bruto de 16, que é recodificado para um valor padronizado de 2. Este valor indica que a observação foi um comprimento que é 2 desvios padrão acima do comprimento médio.

A segunda observação tem um valor não codificado de 10 e um valor padronizado de 0. Valores padronizados de zero indicam que o valor original é exatamente igual à média.

O terceiro valor bruto é 7, que é recodificado para um valor padronizado de -1. Esta observação é um desvio padrão abaixo da média.

021.399.242-66

### Interpretando Coeficientes Padronizados

Quando você ajusta o modelo usando as variáveis independentes padronizadas, os coeficientes são agora coeficientes padronizados.

Vamos a um exemplo já dado antes:

$$\text{Custos de Ar Condicionado} = 3 * \text{Temperatura} - 4 * \text{Isolamento}$$

O coeficiente padronizado para Temperatura (3) indica que para cada aumento de desvio padrão na temperatura, a média dos custos com ar condicionado aumentam em \$3. E, para isolamento, cada aumento de desvio padrão na espessura reduz os custos em US\$ 4.

A padronização coloca todas as variáveis na mesma escala para que você possa comparar a magnitude dos resultados. No exemplo acima, a temperatura e a espessura do isolamento são tipos de variáveis completamente diferentes. Qual deles tem um efeito maior? Você não pode usar os coeficientes brutos para fazer essa determinação porque eles estão



usando unidades diferentes (Celsius vs. centímetros). No entanto, a padronização coloca todos em uma escala consistente, o que permite comparar os coeficientes padronizados.

Para o exemplo de ar condicionado, os valores absolutos dos coeficientes padronizados indicam que para um aumento de um desvio padrão, a espessura do isolamento (-4) afeta os custos mais do que a temperatura (3).

Padronizar os valores de suas variáveis contínuas também pode fazer mais fácil de entender em alguns casos. Temperatura em Celsius e espessura em centímetros são ambos concretos, fáceis de entender. No entanto, algumas variáveis podem ter significados difíceis de entender.

Imagine que você está trabalhando com uma escala psicológica de ansiedade que vai de 12 a 48. O que representa um aumento de uma unidade? O que é considerado uma mudança substancial usando essas unidades sem sentido? Você não pode responder a nenhuma dessas perguntas sem entender a distribuição de pontuações.

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242-06

O uso de valores padronizados e coeficientes padronizados remove as unidades sem sentido e permite comparar pontuações para toda a distribuição.

## CENTRALIZAÇÃO DE VARIÁVEIS NUMÉRICAS

Centralizar as variáveis também é conhecido como padronizar as variáveis subtraindo a média. Esse processo envolve calcular a média para cada variável independente contínua e subtrair a média de todos os valores observados dessa variável. Em seguida, use essas variáveis centralizadas em seu modelo.

Existem outros métodos de padronização, como já falamos na seção de "scaling" mas a vantagem de apenas subtrair a média é que a interpretação dos coeficientes permanece a mesma. Os coeficientes continuam a representar a mudança média na variável dependente dada uma mudança de 1 unidade na variável independente.



Por exemplo, vamos pegar a coluna %Fat. A média dessa coluna é 28,57. Pegando a coluna e subtraindo a média dela mesma, temos uma nova coluna (%Fat - média %Fat).

%Fat	média %Fat	%Fat - média %Fat
25.3	28.57	-3.27
29.3		0.73
37.7		9.13
32.8		4.23
24.6		-3.97
26.5		-2.07
21.2	Oseias Dias de Farias oseias.dfarias@gmail.com 021.399.242-66	-7.37

Essa nova coluna entrará no seu modelo, ao invés de %Fat. Vamos ajustar o mesmo modelo, mas usando as variáveis independentes centradas.

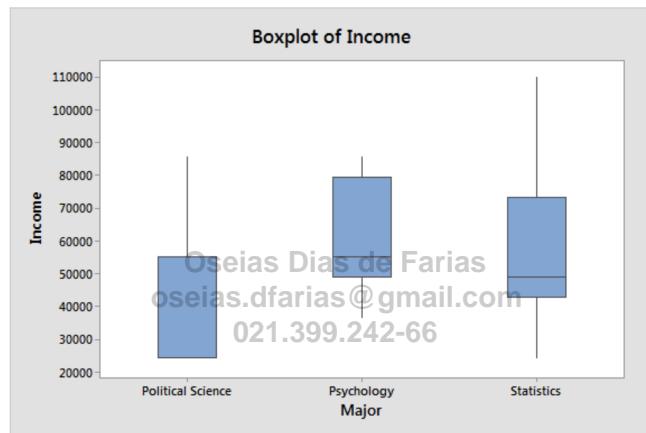
## TRABALHANDO COM VARIÁVEIS CATEGÓRICAS

Para variáveis categóricas, você tem o nome da variável e os níveis dessa variável. A tabela a seguir mostra exemplos de diversas variáveis categóricas e seus níveis.

Graduação	Gênero de filme	Gênero biológico
Engenharia	Ficção	Feminino
Psicologia	Drama	Masculino
Estatística	Comédia	

Com variáveis contínuas, você pode plotá-las em um gráfico de dispersão e ver como uma variável muda à medida que você aumenta o valor da outra variável. No entanto, com variáveis categóricas, você está lidando com grupos que você não pode aumentar incrementalmente. Consequentemente, você interpreta variáveis categóricas de forma diferente na análise de regressão.

Os níveis de variáveis categóricas representam grupos (ex: engenharia, psicologia, estatística) e você pode plotá-los usando um boxplot, como mostrado abaixo. A análise de regressão estima as diferenças médias entre esses grupos e determina se elas são estatisticamente significativas.



Incluir variáveis categóricas em um modelo de regressão permite determinar se as diferenças neste tipo de gráfico são estatisticamente significativas.

### Codificando Variáveis Categóricas (Variáveis dummy)

Nenhum software ou linguagem de programação pode pegar uma variável categórica e analisá-la diretamente. Em vez disso, ele converte variáveis categóricas em variáveis indicadoras usando um esquema de codificação (0, 1).

As variáveis indicadoras, também conhecidas como variáveis dummy, são colunas de 1s e 0s que indicam a presença ou ausência de uma característica. O 1 indica a presença enquanto um 0 representa sua ausência. Para mostrar como isso funciona, vou começar com o gênero biológico.

Imagine que temos uma tabela com o nome de clientes o gênero biológico:

Cliente	Gênero Biológico
Alan Turing	Masculino
Marie Curie	Feminino
Ada Lovelace	Feminino
Albert Einstein	Masculino

A codificação em variável dummy dessa tabela acima ficaria:

Cliente	Masculino	Feminino
Alan Turing	1	0
Marie Curie	0	1
Ada Lovelace	0 Oseias Dias de Farias oseias.dfarias@gmail.com	1
Albert Einstein	1 021.399.242-66	0

As colunas Masculino e Feminino são as variáveis indicadoras baseadas na coluna Gênero Biológico. A coluna Masculino contém 1s para observações que correspondem a homens e 0s para mulheres. O padrão oposto se aplica à coluna Feminino.

Observe como essas duas colunas fornecem informações completamente redundantes? Uma coluna prevê a outra coluna perfeitamente. Estatísticos referem-se a isso como multicolinearidade perfeita, o que cria um erro se você incluir ambas em um modelo de regressão. Para uma variável categórica, você deve omitir uma das variáveis indicadoras subjacentes do modelo, que se tornará a **referência**. Nesse caso, podemos ter apenas:

Cliente	Masculino
Alan Turing	1
Marie Curie	0

Ada Lovelace	0
Albert Einstein	1

Ou seja, Feminino é nossa coluna de referência.

Agora vamos olhar uma tabela com "College Major" sendo o curso de graduação e as colunas subsequentes como sendo cada curso (psicologia, ciências políticas e estatística)

College Major	Psychology	Political Science	Statistics
Statistics	0	0	1
Psychology	1	0	0
Statistics	0	0	1
Political Science	0	1	0
Psychology	1	0	0

Nesta tabela, College Major é a variável categórica. Cada célula contém 1s somente quando essa propriedade está presente e 0 caso contrário. Para cada linha, deve haver um único valor de 1, e todos os outros valores são 0s. Em outras palavras, os grupos são **mutuamente exclusivos**.

Assim como no gênero, se você incluir todas as variáveis do indicador, estará fornecendo informações redundantes e não poderá realizar a análise. Se você olhar para quaisquer três colunas, sempre poderá descobrir o valor da quarta coluna. Suponha que excluamos a coluna Psicologia. Na primeira linha, vemos o 1 nas estatísticas, então sabemos que a psicologia deve ser zero. Na segunda linha, Ciência Política, Engenharia e Estatística têm 0s. Assim, a Psicologia deve ter 1. Novamente, teremos que remover uma variável indicadora para realizar a análise.

Para todas as variáveis categóricas, você deve sempre remover um nível da análise. A removida será seu nível de referência.

## Interpretando os resultados para variáveis categóricas

Como uma variável categórica geralmente representa várias variáveis indicadoras, teremos um teste F nesse grupo de variáveis indicadoras. Ao contrário dos testes t, os testes F podem avaliar vários termos de modelos simultaneamente, o que lhes permite comparar os ajustes de diferentes modelos lineares. Nesta situação, um teste F compara o ajuste do modelo com o conjunto de variáveis indicadoras que corresponde a uma variável categórica a um modelo sem esse conjunto de variáveis indicadoras.

Se o seu p-valor for menor que o nível de significância, você pode rejeitar a hipótese nula e concluir que a variável categórica melhora o ajuste do modelo.

Em seguida, a análise compara cada nível com o nível de referência usando testes t. Enquanto o teste F informa sobre a variável categórica como um todo, os testes t permitem explorar as diferenças entre as médias do grupo e o nível de referência. Os coeficientes representam a diferença entre a média de cada nível e a média do nível de referência. Consequentemente, se seu p-valor for menor que seu nível de significância, você pode rejeitar a hipótese nula e concluir que a média do nível é significativamente diferente da média do nível de referência.

Vamos a um exemplo. Temos uma tabela com salário (Income), curso da graduação (Major) e anos de experiência (Experience). Abaixo mostro o começo dessa tabela.

Income	Major	Experience
36669	Political Science	4
24446	Political Science	1
61115	Political Science	5
24446	Political Science	1
24446	Political Science	2
24446	Political Science	2

Se olharmos a ANOVA, teríamos:

#### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	4768836128	1589612043	3.50	0.030
Experience	1	2252342774	2252342774	4.96	0.035
Major	2	3762711560	1881355780	4.14	0.027
Error	26	11809775734	454222144		
Lack-of-Fit	13	7078720982	544516999	1.50	0.239
Pure Error	13	4731054752	363927289		
Total	29	16578611861			

A variável anos de experiência é estatisticamente significativa. No entanto, vamos nos concentrar na variável categórica de Major, que circulei. Você pode ver que esta variável usa 2 graus de liberdade ao contrário do Experience, que usa apenas 1. Lembre-se, o Major tem três níveis e excluímos o Statistics do modelo para usá-lo como nível de referência. Consequentemente, o modelo inclui duas variáveis indicadoras para representar toda a variável categórica de Major, o que explica por que ele usa dois graus de liberdade. Se sua variável categórica tem muitos níveis, ele usará muitos graus de liberdade, o que pode ser problemático quando o tamanho da amostra é pequeno.

Observando o resultado do teste F circulado na saída anterior, vemos que Major é estatisticamente significativo em geral. Melhora o modelo. Agora, vamos olhar a tabela de coeficientes

#### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	49064	7470	6.57	0.000	
Experience	5085	2284	2.23	0.035	1.15
Major					
Political Science	-27195	9813	-2.77	0.010	1.34
Psychology	-5368	9916	-0.54	0.593	1.27



Eu circulei a saída relacionada ao Major. Como Estatística ("Statistics") é o nível de referência que excluímos da análise, a tabela não o exibirá. Os coeficientes de Ciência Política (Political Science) e Psicologia (Psychohology) indicam como os salários médios desses cursos se comparam com os rendimentos salários do curso de Estatística. Os coeficientes negativos indicam que estes cursos têm rendimentos médios mais baixos do que Estatística. A partir dos coeficientes, aprendemos o seguinte:

- A renda média dos graduados em Ciência Política é de US\$ 27.195 a menos do que a renda média dos graduados em Estatística.
- A renda média para graduandos em Psicologia é \$ 5.368 menor do que a renda média dos graduados em Estatística.

Em seguida, observe os p-valores para os testes t. Esses p-valores determinam se as diferenças médias são estatisticamente significativas.

O coeficiente de Ciência Política é estatisticamente significativo. Consequentemente, podemos rejeitar a hipótese nula de que a diferença média entre Ciência Política e Estatística é zero (lembre-se que Estatística é nosso nível de referência)

Oseias Dias de Farias  
oseias.diasdefarias@gmail.com  
021.399.242-66

Por outro lado, a diferença entre os salários médios da Psicologia e Estatística não é estatisticamente significativa. Temos evidências insuficientes para concluir que esses meios são diferentes. Em outras palavras, a diferença observada de -\$5368 pode representar um erro aleatório. Se coletássemos outra amostra aleatória e realizássemos a análise novamente, essa diferença poderia desaparecer.

Se ajustarmos o modelo usando um nível de referência diferente, a significância geral de Major na tabela ANOVA permanecerá a mesma, como R-quadrado. Por outro lado, as comparações entre níveis específicos mudarão porque estaríamos comparando os principais a um nível de referência diferente. Por exemplo, se Ciência Política for o nível de referência, tanto a Psicologia como a Estatística teriam rendimentos médios que são significativamente maiores do que ele. No entanto, o quadro geral permanece o mesmo.

Use o nível de referência que torna o sentido mais intuitivo para sua pergunta de pesquisa.

Como você aprendeu na seção de variáveis contínuas, veja como interpretar seu coeficiente positivo. Para cada aumento de um ano na experiência, a renda média aumenta em uma média de \$ 5.085, mantendo Major constante.

Agora, veremos uma maneira diferente de representar os resultados na equação de regressão quando você tem variáveis categóricas. Como visto acima, a participação em diferentes cursos está relacionada a diferentes rendas médias. Por exemplo, os graduados em Ciências Políticas têm uma renda média mais baixa do que os graduados em Estatística. Essa diferença média é - \$ 27.195. Como esse valor não muda, podemos subtraí-lo da constante na equação de regressão (que pela tabela é 49064) e criar uma equação para a ciência política especificamente.

De modo mais geral, as variáveis indicadoras deslocarão a linha de regressão para cima e para baixo no eixo y para grupos específicos pelo valor do coeficiente para a variável indicadora correspondente. Consequentemente, você pode obter equações separadas para cada nível categórico com diferentes constantes, como mostrado abaixo.

Regression Equation

Oscar Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Major

Political Science    Income = 21869 + 5085 Experience

Psychology

Income = 43696 + 5085 Experience

Statistics

Income = 49064 + 5085 Experience

As diferenças entre as constantes correspondem aos coeficientes para Psicologia e Ciência Política.

## ANOVA 2 FATORES (OU MAIS)

Usamos a ANOVA 2 fatores para avaliar as diferenças entre as médias dos grupos que são definidas por dois fatores. Como todos os testes de hipóteses, a ANOVA 2 fatores usa dados amostrais para inferir as propriedades de toda a população.



Os fatores são suas variáveis independentes. O número de fatores em sua análise determina o nome da análise ANOVA. A ANOVA 1 fator usa um fator. A ANOVA 2 fatores tem dois. E assim por diante.

Você também precisa de uma variável de resultado contínua, que é a variável dependente.

Para resultados confiáveis de ANOVA, seus dados devem atender às seguintes suposições:

- O resultado ou variável dependente é contínua.
- Resíduos aleatórios com variância constante

Imagine que estamos avaliando os salários anuais (Income), que é nossa variável dependente contínua. Nossos dois fatores são gênero (Gender) e curso superior (Major). Para esta análise, usaremos os três cursos: estatística, psicologia e ciência política. A combinação desses dois fatores (2 gêneros X 3 majors) produz os seguintes seis grupos. Cada grupo contém 20 observações.

Oseias Dias de Farias  
oseias.dfarias@gmail.com

Male / Statistics: \$77,743	Female / Statistics: \$74,074
Male / Psychology: \$69,766	Female / Psychology: \$65,320
Male / Political Science \$62,015	Female / Political Science \$55,195

O valor em dólares indica a renda média de cada grupo. A ANOVA de duas vias determina se as diferenças observadas entre as médias fornecem evidências fortes o suficiente para concluir que as médias populacionais são diferentes. Vamos fazer a análise!

#### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Gender	1	593002242	593002242	25.80	0.000
Major	2	6009140933	3004570466	130.70	0.000
Gender*Major	2	88232238	44116119	1.92	0.151
Error	114	2620748173	22989019		
Total	119	9311123587			

Olhe na coluna p-valor na tabela Analysis of Variance. Como os **p-valores tanto para Gênero quanto para Major são menores do que nosso nível de significância, esses fatores são estatisticamente significativos**. Esses são os principais efeitos no modelo.

Por outro lado, o efeito de interação (Gender\*Major) não é significativo porque seu p-valor (0,151) é maior que nosso nível de significância. Como o efeito de interação não é significativo, podemos nos concentrar apenas nos efeitos principais.

Oseias Dias de Farias

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66

Os efeitos principais são a parcela da relação entre uma variável independente e a variável dependente que não muda com base nos valores das demais variáveis do modelo. Por exemplo, o efeito de Gender sobre a renda média não muda de acordo com o ensino superior (Major) que a pessoa cursa. É um efeito consistente em todos os cursos (majors). Nesse caso, os homens têm uma renda média mais alta e esse efeito é consistente entre os cursos.

## OK, MAS ENTÃO QUAL É A DIFERENÇA ENTRE UMA ANOVA COM MÚLTIPLOS FATORES E UMA REGRESSÃO LINEAR?

ANOVA e regressão linear são baseados no mesmo modelo (linear), mas eles focam em aspectos diferentes na análise do modelo.

Os modelos ANOVA são usados quando as variáveis preditoras são categóricas. Exemplos de variáveis categóricas incluem nível de escolaridade, cor dos olhos, estado civil, etc. Modelos de regressão são usados quando as variáveis preditoras são contínuas ou categóricas.



Na avaliação de impacto de variáveis, a ANOVA e regressão OLS são iguais **nos casos em que seus preditores são categóricos** (em termos das inferências que você está extraíndo da estatística de teste - ou seja, entender os fatores mais relevantes). Não há nada que uma ANOVA possa dizer que a regressão não possa derivar. O contrário, porém, não é verdadeiro. ANOVA não pode ser usada para análise com variáveis contínuas. Como tal, a ANOVA pode ser classificada como a técnica mais limitada.

Entretanto, a ANOVA tem uma vantagem sobre a regressão: o termo de interação entre 2 variáveis categóricas é automaticamente colocado (Gender\*Major, no caso acima). Dessa forma, conseguimos avaliar as interações entre variáveis categóricas também!

Vamos a um outro exemplo que isso pode ser útil. Imagine que estamos realizando um teste de sabor para determinar qual alimento e de condimento produz o maior satisfação. Faremos uma análise onde nossa variável dependente é Satisfação. Nossas duas variáveis independentes são ambas variáveis categóricas: Alimentos e Condimento.

Por trás dos panos, a ANOVA multiplica as duas variáveis para calcular o valor do termo de interação. Para manter as coisas simples, vamos incluir apenas dois alimentos (sorvete e cachorro quente) e dois condimentos (molho de chocolate e mostarda) em nossa análise.

Dadas as especificidades do exemplo, um efeito de interação não seria surpreendente. Se alguém lhe perguntar: "Você prefere ketchup ou chocolate na sua comida?" Sem dúvida, você responderá: "Depende do tipo de comida!" Essa é a natureza "depende" de um efeito de interação. Você não pode responder à pergunta sem saber mais informações sobre a outra variável no termo de interação – que é o tipo de comida no nosso exemplo!

Os p-valores na saída abaixo nos dizem que o efeito da interação (Alimento\*Condimento) é estatisticamente significativo. Consequentemente, sabemos que a Satisfação que você obtém com o condimento depende do tipo de alimento e vice-versa.



#### Factor Information

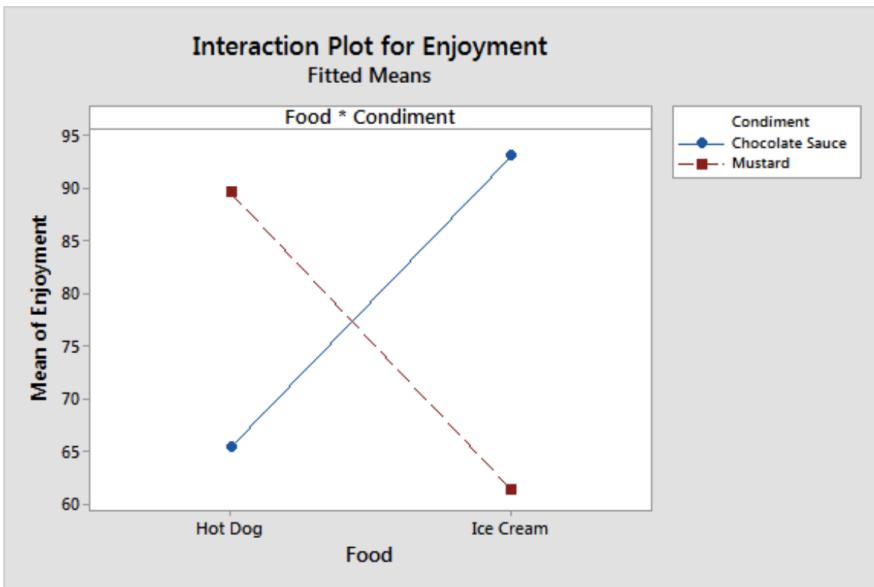
Factor	Type	Levels	Values
Food	Fixed	2	Hot Dog, Ice Cream
Condiment	Fixed	2	Chocolate Sauce, Mustard

#### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Food	1	1.6	1.6	0.06	0.801
Condiment	1	277.5	277.5	11.07	0.001
Food*Condiment	1	15695.8	15695.8	626.15	0.000
Error	76	1905.1	25.1		
Total	79	17880.0			

Mas, como interpretamos o efeito de interação e realmente entendemos o que os dados estão dizendo? A melhor maneira de entender esses efeitos é com um tipo especial de gráfico **gráfico de interação**. Este tipo de gráfico exibe os valores ajustados da variável dependente no eixo y enquanto o eixo x mostra os valores da primeira variável independente. As várias linhas representam valores da segunda variável independente.



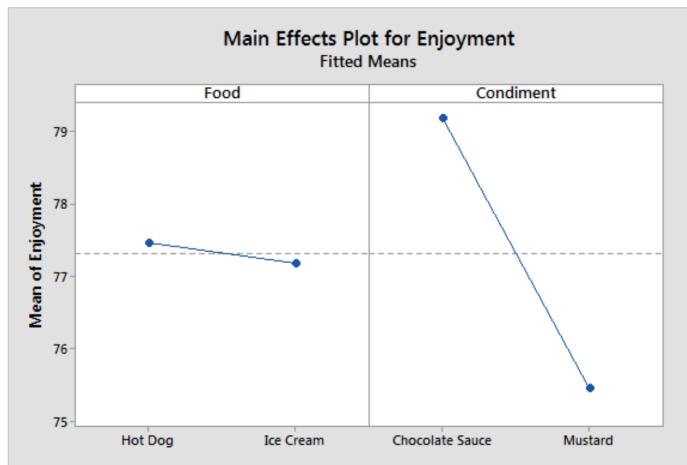


Em um gráfico de interação, linhas paralelas ao eixo X indicam que não há efeito de interação, enquanto inclinações diferentes sugerem que um pode estar presente.

As linhas cruzadas no gráfico sugerem que existe um efeito de interação, o que o p-valor é significativo para o termo Alimentos\*Condimentos. O gráfico mostra que os níveis de Satisfação (eixo y) são maiores para a calda de chocolate quando a comida for sorvete (Ice Cream). Por outro lado, os níveis de satisfação são maiores para mostarda quando a comida é um cachorro-quente (hot dog). Se você colocar mostarda no sorvete ou calda de chocolate em cachorros-quentes, você não ficará feliz! Qual condimento é melhor? Depende do tipo de alimento, e temos estatísticas para demonstrar esse efeito.

Quando você tem efeitos de interação estatisticamente significativos, não se pode interpretar os efeitos principais sem considerar as interações. No exemplo anterior, você não pode responder à pergunta sobre qual condimento é melhor sem saber o tipo de alimento. Mais uma vez, “depende”.

Suponha que queremos maximizar a satisfação escolhendo o melhor alimento e o melhor condimento. No entanto, imagine que esquecemos de incluir o efeito de interação e avaliamos apenas os efeitos principais. Nós faremos nossa decisão com base nos principais gráficos de efeitos abaixo.



Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Com base nesses dados, escolheríamos cachorros-quentes com calda de chocolate porque cada um deles produz maior Satisfação. Isso não é uma boa escolha, apesar do que os efeitos principais mostram! Quando você tem interações estatisticamente significativas, não pode interpretar o efeito principal sem considerar os efeitos de interação.

Um equívoco comum é que um efeito de interação indica que as próprias variáveis independentes estão correlacionadas. Isso não é necessariamente verdade. Um efeito de interação refere-se à relação entre cada variável independente e a variável dependente. Especificamente, um efeito de interação indica que a relação entre uma variável independente e a variável dependente muda com base no valor de pelo menos uma outra variável independente. Essas variáveis independentes não precisam estar correlacionadas para que esse efeito ocorra.

Variáveis independentes correlacionadas é outro fenômeno, que é chamada de multicolinearidade, que já foi abordado aqui.

Uma preocupação comum ocorre ao interpretar interação significativa de efeitos quando os efeitos principais não são significativos.

Para entender a resposta, vamos refrescar a memória sobre cada tipo de efeito.

- Efeito principal: a porção do efeito de uma variável independente na variável dependente que não depende dos valores das outras variáveis do modelo.
- Efeito de interação. A porção do efeito de uma variável independente que depende do valor de pelo menos uma outra variável independente no modelo.

Além disso: O efeito total de uma variável independente = efeito principal + interação efeito

Oseias Dias de Farias

[oseias\\_dias\\_farias@gmail.com](mailto:oseias_dias_farias@gmail.com)  
021.399.242-66

Quando qualquer tipo de efeito não é estatisticamente significativo, você não tem evidências suficientes para concluir que o efeito é diferente de zero. Quando um efeito não é significativo, você pode zerá-la (remover da equação de regressão).

## MULTICOLINEARIDADE EM TERMOS DE INTERAÇÃO

Agora imagine que você viu que o efeito de interação é significativo e que você queira levar esse termo para sua regressão. O que fazer com a multicolinearidade?

Usarei uma análise de regressão para modelar a relação entre as variáveis independentes (atividade física - activity , percentual de gordura corporal - %fat, peso - weight e a interação entre peso e gordura corporal - %fat\*weight) e a variável dependente (densidade mineral óssea do colo do fêmur).

Aqui estão os resultados da regressão:

## Regression Analysis: Femoral Neck versus %Fat, Weight kg, Activity

### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.555785	0.138946	27.95	0.000
%Fat	1	0.009240	0.009240	1.86	0.176
Weight kg	1	0.127942	0.127942	25.73	0.000
Activity	1	0.047027	0.047027	9.46	0.003
%Fat*Weight kg	1	0.041745	0.041745	8.40	0.005
Error	87	0.432557	0.004972		
Total	91	0.988342			

### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

### Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	0.155	0.132	1.18	0.243	
%Fat	0.00557	0.00409	1.36	0.176	14.93
Weight kg	0.01447	0.00285	5.07	0.000	33.95
Activity	0.000022	0.000007	3.08	0.003	1.05
%Fat*Weight kg	-0.000214	0.000074	-2.90	0.005	75.06

Esses resultados mostram que Peso, Atividade e a interação entre eles são estatisticamente significativos. O percentual de gordura corporal não é estatisticamente significativo. No entanto, os VIFs indicam que nosso modelo possui multicolinearidade severa para algumas das variáveis independentes.

Observe que a atividade (activity) tem um VIF próximo a 1, o que mostra que a multicolinearidade não afeta e podemos confiar nesse coeficiente e p-valor sem nenhuma ação adicional. No entanto, os coeficientes e p-valores para os outros termos são suspeitos!

Incluímos também o termo de interação gordura corporal\*peso (%fat\*weight). Claramente, há uma correlação entre o termo de interação e ambos os termos de efeito principal. Os VIFs refletem essas relações.



Existe um método para remover esse tipo de multicolinearidade estrutural de forma rápida e fácil!

### **Centralize as variáveis independentes para reduzir a multicolinearidade.**

Centralizar as variáveis também é conhecido como padronizar as variáveis subtraindo a média. Esse processo envolve calcular a média para cada variável independente contínua e subtrair a média de todos os valores observados dessa variável. Em seguida, use essas variáveis centralizadas em seu modelo.

Existem outros métodos de padronização, como já falamos na seção de "scaling" mas a vantagem de apenas subtrair a média é que a interpretação dos coeficientes permanece a mesma. Os coeficientes continuam a representar a mudança média na variável dependente dada uma mudança de 1 unidade na variável independente.

Por exemplo, vamos pegar a coluna %Fat. A média dessa coluna é 28,57. Pegando a coluna e subtraindo a média dela mesma, temos uma nova coluna (%Fat - média %Fat).

%Fat	média %Fat	%Fat - média %Fat
25.3	28.57	-3.27
29.3		0.73
37.7		9.13
32.8		4.23
24.6		-3.97
26.5		-2.07
21.2		-7.37



Essa nova coluna entrará no seu modelo, ao invés de %Fat. Vamos ajustar o mesmo modelo, mas usando as variáveis independentes centradas.

### Regression Analysis: Femoral Neck versus %Fat S, Weight S, Activity S

#### Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.55578	0.138946	27.95	0.000
%Fat S	1	0.04786	0.047863	9.63	0.003
Weight S	1	0.30473	0.304728	61.29	0.000
Activity S	1	0.04703	0.047027	9.46	0.003
%Fat S*Weight S	1	0.04175	0.041745	8.40	0.005
Error	87	0.43256	0.004972		
Total	91	0.98834			

#### Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.0705118	56.23%	54.22%	50.48%

Oseias Dias de Farias oseias.dfarias@gmail.com 021 399 242-66						
Coefficients	Term	Coef	SE Coef	T-Value	P-Value	VIF
	Constant	0.82161	0.00973	84.40	0.000	
	%Fat S	-0.00598	0.00193	-3.10	0.003	3.32
	Weight S	0.00835	0.00107	7.83	0.000	4.75
	Activity S	0.000022	0.000007	3.08	0.003	1.05
	%Fat S*Weight S	-0.000214	0.000074	-2.90	0.005	1.99

A diferença mais aparente é que os VIFs estão todos abaixo de valores satisfatórios; eles são todos menores que 5. Podemos ver que há alguma multicolinearidade em nossos dados, mas não é severa o suficiente para garantir outras medidas corretivas.

Podemos comparar duas versões do mesmo modelo, uma com alta multicolinearidade e outra sem. Esta comparação destaca seus efeitos.

A primeira variável independente que veremos é Atividade. Essa variável foi a única a quase não apresentar multicolinearidade no primeiro modelo. Compare os coeficientes de atividade e os valores-p entre os dois modelos e você verá que eles são iguais (coeficiente = 0,000022, valor-p = 0,003). Isso



ilustra como apenas as variáveis altamente correlacionadas são afetadas por seus problemas.

Além disso, %Fat é significativo no segundo modelo, embora não fosse no primeiro modelo. Não apenas isso, mas o sinal do coeficiente para %Fat mudou de positivo para negativo!

A menor precisão do modelo com multicolinearidade, os sinais trocados e a falta de significância estatística são problemas típicos associados à multicolinearidade.

Agora, dê uma olhada nas tabelas Summary of Model para ambos os modelos. Você notará que o erro padrão da regressão (S) , R-quadrado , R-quadrado ajustado e R-quadrado previsto são todos idênticos. A multicolinearidade não afeta as previsões ou a qualidade do ajuste. **Se você quiser apenas fazer previsões, o modelo com multicolinearidade severa também é bom! Porém, se quiser entender como cada variável impacta no seu modelo, um modelo com alta colinearidade pode não ser muito bom.**



## REGRESSÃO PARA O MODELO RENDA X GÊNERO E CURSO SUPERIOR

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	67543	438	154.32	0.000	
Gender					
Female	-2223	438	-5.08	0.000	1.00
Major					
Political Science	-8940	619	-14.44	0.000	1.33
Psychology	574	619	0.93	0.356	1.33
Gender*Major					
Female Political Science	-1189	619	-1.92	0.057	1.33
Female Psychology	800	619	1.29	0.199	1.33

Na tabela "Coefficients", cada coeficiente representa a diferença entre um valor e a média geral. O p-valor correspondente indica se essa diferença é estatisticamente significativa. Por exemplo, o coeficiente de "Female" é -2223, ou seja, as mulheres ganham \$ 2.223 a menos do que o salário médio geral. Os graduados em ciência política ganham US\$ 8.940 a menos do que a média global. Os p-valores indicam que essas diferenças são estatisticamente significativas. Por outro lado, os graduandos em psicologia ganham US\$ 574 a mais do que a média geral, mas essa diferença não é significativa (0,356).

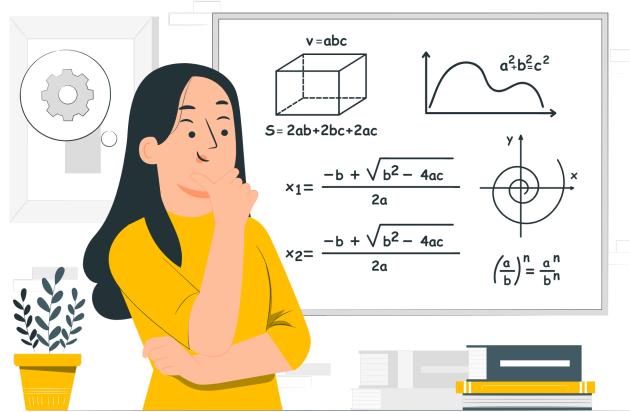
Mais para frente vamos falar sobre codificação de variáveis categóricas e vocês entenderão o porque o gênero "Male" e a graduação "Statistics" não estão sendo mostradas. Por hora, vamos nos concentrar em entender as tabelas.

O efeito de interação (Gênero\*Maior) não é significativo porque seu p-valor (0,151) é maior que nosso nível de significância. Como o efeito de interação não é significativo, podemos nos concentrar apenas nos efeitos principais.

Graças aos p-valores na tabela ANOVA, sabemos que ambos os padrões de efeito principal neste gráfico são estatisticamente significativos. Sem os resultados de teste significativos, os padrões podem ser atribuídos a erros aleatórios.



# 25. Regressão Logística



Antes de falarmos sobre regressão logística, precisamos falar sobre "modelos de classificação".

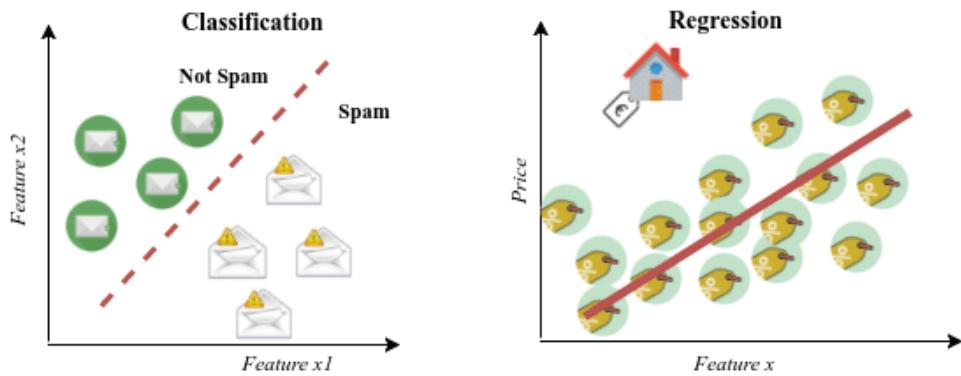
## CLASSIFICAÇÃO

Oseias Dias de Farias  
[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)  
021.399.242-66

Um modelo de classificação também usa variáveis independentes ( $X$ s) para modelar variável dependente ( $y$ ). Porém, a grande chave aqui é o que é o nosso  $y$ . Quando falamos de modelos de regressão linear e polinomial, nosso  $y$  necessariamente era uma variável contínua. Quando falamos em classificação, nosso  $y$  passa a ser uma classe e nosso objetivo passa a ser prever a probabilidade de determinado ponto pertencer a determinada classe.

**Classificação binária:** Classifica os dados em duas classes como Sim/Não, bom/ruim, alto/baixo, sofre de uma determinada doença ou não, etc. A figura abaixo representa o modelo de classificação representando as linhas que separam duas classes diferentes.





Por exemplo, um modelo pode ler um e-mail e classificá-lo como spam ou não — classificação binária. Alternativamente, um modelo pode ler uma imagem médica, digamos uma mamografia, e classificá-la como benigna ou maligna.

**Classificação multiclasse:** Classifica os dados em três ou mais classes; Classificação de documentos, categorização de produtos (produto de beleza, eletrônico, eletrodomésticos, etc)

Os algoritmos de classificação, como a regressão logística, geram uma pontuação de probabilidade que atribui alguma probabilidade à entrada pertencente a uma categoria. Essa probabilidade é então mapeada para um mapeamento binário, assumindo que a classificação é binária (maligno ou benigno, spam ou não spam).

No exemplo de spam, um modelo pode ler um e-mail e gerar uma pontuação de probabilidade de 92% de spam, sugerindo que há uma chance muito alta de que esse e-mail seja realmente spam. Uma pontuação próxima de 0 indica que o e-mail provavelmente não é spam, enquanto uma pontuação mais próxima de 100 indica que o e-mail é muito provavelmente spam.

## PROBLEMAS COM EXEMPLOS DE CLASSIFICAÇÃO DO MUNDO REAL

1. Previsão do comportamento do cliente: os clientes podem ser divididos em grupos com base em seus hábitos de compra, hábitos de navegação na loja online e outros fatores. Modelos de classificação, por exemplo, podem ser usados para avaliar se um consumidor provavelmente comprará coisas adicionais. Se o modelo de

categorização sugerir que eles farão mais compras, você poderá oferecer ofertas e descontos especiais. Alternativamente, se for descoberto que eles provavelmente abandonarão seus hábitos de compras em um futuro próximo, você pode salvá-los para mais tarde, disponibilizando suas informações prontamente.

2. Classificação de imagens: Para categorizar as fotos em categorias distintas, um modelo de classificação multiclasse pode ser desenvolvido. Por exemplo, podemos fazer um modelo para ajudar a automatizar a classificação de fotos de cães e gatos (binário) ou cães, gatos, pássaros, jacarés, etc (multiclasse).
3. Classificação de texto da web: usa categorias pré-determinadas aprendidas com dados anteriores para classificar o texto da web ou atribuir tags ao texto da web. Os modelos de classificação, por exemplo, podem ser usados para classificar o material da web em uma das três categorias: esportes, entretenimento ou tecnologia.
4. Previsão de desligamento de cliente (churn): um modelo de classificação binária pode ser usado para prever se um cliente provavelmente irá ou não desistir da assinatura de um produto em um futuro próximo.
5. Detecção de fraude de cartão de crédito: Para detecção de fraude de cartão de crédito, pode ser empregado um modelo de classificação binária, no qual os dados históricos de transações de um cliente são avaliados

## REGRESSÃO LOGÍSTICA

Regressão logística entra no contexto de modelos de classificação. É um dos algoritmos usados para podermos prever essas classes dadas todas as variáveis independentes. A fórmula geral da regressão logística é:

$$\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

Em que



- $p$  é a probabilidade de ocorrência do evento de interesse
- $\log(p/(1-p))$  é o logit da probabilidade, a transformação logarítmica da razão de chances.

Se arranjarmos essa equação usando o logaritmo natural (na base e) temos que:

$$P(Y = 1|X) = \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1}$$

Vocês também verão a equação acima representada por:

$$P = \frac{1}{1+e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n)}}$$

Oseias Dias da Farias  
 oseias.dfarias@gmail.com  
 021.399.242-66

Agora quero explicar um pouco os detalhes para vocês

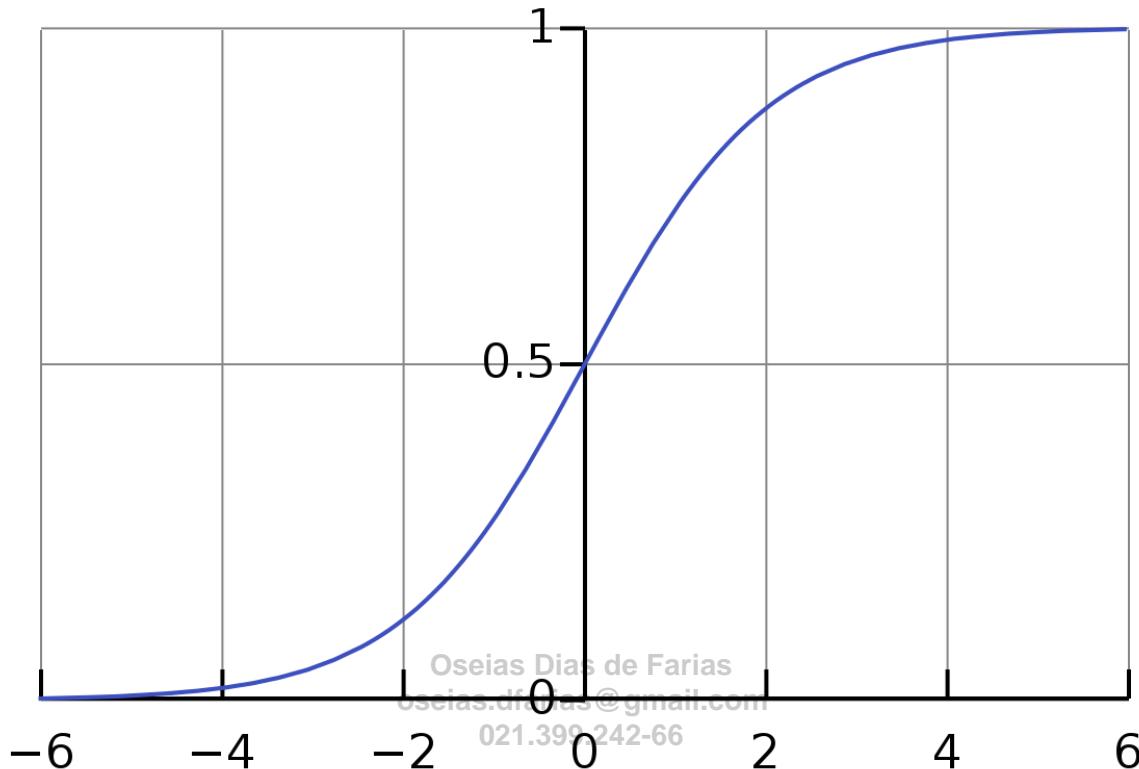
Primeiro, se você não conhece essa letra “e” na parte de baixo da fração, ele representa o número de Euler e tem um valor constante aproximado de 2,718. Ele é uma constante matemática tipo o pi - aquele que vale, aproximadamente, 3,14.

Para quem ainda não conhece, essa equação é a que chamamos de **equação sigmoid**:

$$y = \frac{1}{1 + e^{-x}}$$

Em que  $y$  é nosso  $P$  e o  $x$  é a função igualzinha a **regressão linear**.

Graficamente, essa é a carinha da sigmoid



Embora utilizamos o termo “regressão” para problemas onde queremos prever valores contínuos, como na regressão linear, ele também é utilizado na regressão logística porque fazemos uma **modelação do valor contínuo** previsto para se ajustar a problemas de classificação. E isso é feito justamente através da sigmoid! Essa função é usada para converter qualquer valor real para um valor entre 0 e 1, o que a torna perfeita para modelar probabilidades, ou seja, o y da função abaixo será o x da função sigmoide. Veja abaixo

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + \dots + b_n * x_n$$

Vamos dar um exemplo mais palpável:

Imagine que os pesos  $b_1$ ,  $b_2$ ,  $b_3$  e  $b_0$  forem, respectivamente, 0.3, 0.9, 0.6 e -5,5. Esses pesos foram encontrados, por exemplo, com o OLS. Agora, vamos

olhar para um exemplo (uma instância) específica do nosso modelo. Abaixo, você consegue ver a instância:

$$X_1 = 7$$

$$X_2 = 8$$

$$X_3 = -3$$

Note que multiplicaremos  $x_1$  por  $b_1$ ,  $x_2$  por  $b_2$ ,  $x_3$  por  $b_3$  e somaremos com o  $b_0$ . Ficando com o resultado abaixo:

$$y = 7 * 0.3 + 8 * 0.9 + -3 * 0.6 + (-5.5)$$

Realizando os cálculos, chegamos no resultado abaixo:

$$y = 2.1 + 7.2 - 1.8 - 5.5 = 2$$

Ou seja, para uma determinada instância, o resultado da nossa regressão foi igual a 2. Entretanto, esse valor não será nosso valor de saída, isto é, o output da nossa função. Na verdade, como disse, ele será o input da função sigmoid (o  $x$ ). Observe a função abaixo:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

Agora, iremos substituir o  $x$  por 2, que é o resultado da expressão anterior. O resultado da função sigmoid quando substituímos  $x$  por 2 é 0.881.

$$\sigma(2) = 0.881$$

Observe que ele gerou um número maior do que 0 e menor do que 1. Mas será que isso sempre ocorre?

Para validar nossa hipótese, vamos fazer a seguinte observação. Note que a fórmula possui o número 1 no numerador (parte de cima da fração) e o número  $1 + e^{-x}$  no denominador (parte de baixo da fração).

Além disso, repare que  $e^{-x}$  será sempre positivo, visto que o número de Euler é uma constante positiva. Com isso, o denominador será sempre maior que o numerador, ou seja,  $1 + e^{-x} > 1$ . Dessa forma, ele sempre será maior que 0 e menor que 1: nos limites de uma probabilidade.

Podemos concluir, portanto, que a função sigmoid vai retornar um número entre 0 e 1, ou seja, enquadra-se perfeitamente nos limites de uma probabilidade!

Quanto maior o  $x$ , mais a função fica próxima de 1 (alta probabilidade). Quanto menor o  $x$ , mais a função fica próxima de 0 (baixa probabilidade). E quando o  $x$  é igual a 0, temos uma probabilidade igual a  $0.5 = 50\%$ .

Mas o que significa essa probabilidade?

A probabilidade gerada pela função sigmoid representa a probabilidade de pertencer à classe 1. Ou seja, neste exemplo, a instância que fizemos o cálculo possui 88.1% de chance de pertencer à classe 1.

Oseias Dias de Farias  
oseias.dias@farias@gmail.com  
021.399.242-66

Mas até que valor de probabilidade será aceito como classe 1? Se a probabilidade gerada for  $0.2 = 20\%$  ele ainda será da classe 1?

Como vocês podem imaginar, não é bem assim. Jogamos esse resultado da função sigmoid em uma outra função. Se o valor da probabilidade for maior ou igual a 0.5 (50%), atribuímos à classe 1. Caso contrário, ou seja, caso o valor da probabilidade seja menor que 0.5 (50%), atribuímos à classe 0. Observe a função abaixo:

$$f(x) = \begin{cases} 0, & \text{se } x < 0.5 \\ 1, & \text{se } x \geq 0.5 \end{cases}$$

Como disse, assim como um modelo de regressão linear, o modelo da regressão logística irá gerar como saída o somatório do produto das features

com seus respectivos pesos mais o bias (viés). Entretanto, esse não será o output final do nosso modelo. O resultado irá sofrer uma “modulação”, na verdade, o termo correto seria uma transformação não linear - feita pela função sigmoid que mostramos acima.

Agora que já vimos como o algoritmo funciona, vamos falar sobre como ele é treinado, ou seja, **como achamos os coeficientes da equação** ( $b_0, b_1, b_2\dots$ ).

Já sabemos que ele associa à classe 1 outputs com altas probabilidades e à classe 0, outputs com baixa probabilidade. Então, o objetivo do treinamento é encontrar todos os pesos ( $b_1, b_2, b_3$  etc.) tais que a probabilidade seja maior ou igual a 0.5 para as classes iguais a 1 e menor que 0.5 para as classes iguais a 0.

Para uma probabilidade ser maior ou igual a 0.5, basta que o resultado daquela expressão parecida com a regressão linear seja maior ou igual a 0. Analogamente, para uma probabilidade gerada pela função sigmoid ser menor que 0.5, basta que o resultado da expressão seja menor que 0.

Oseias Dias de Farias  
oseias.dfarias@gmail.com

021.399.242-66

Abaixo, você vê a função de perda utilizada frequentemente, a log-loss:

$$Custo(p) = \begin{cases} -\log(p), & \text{se classe} = 1 \\ -\log(1-p), & \text{se classe} = 0 \end{cases}$$

Imagine que a probabilidade gerada para uma instância com classe real igual a 1 seja 0.2. Como pode imaginar, nosso modelo errou. O custo desse erro é  $-\log(0.2) = 1.609$ . Se a probabilidade fosse 0.005, ou seja, 0.5%, o resultado do custo seria  $-\log(0.005) = 5.298$ . Note que quanto maior o erro, mais o custo irá aumentar.

Para recapitular, “função de perda” é o nome dado à função de custo para apenas uma instância. Quando queremos mensurar o erro de todas as instâncias do nosso dataset, utilizamos a função de custo.



Neste caso, a função de custo será a média da função de perda para todas os exemplos do nosso dataset. Além disso, podemos simplificar o cálculo da **função de perda** com a seguinte fórmula:

$$Logloss = -\frac{1}{n} \sum_{i=0}^n [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

A função de custo é formulada como a **negativa da log-verossimilhança**. Ou seja, minimizar a função de custo acima significa **maximizar log-verossimilhança**. Maximizar a função de verossimilhança (**maximum likelihood - MLE**) diretamente pode ser complexo devido à sua natureza não-linear e ao produto de muitos termos probabilísticos. Portanto, é comum maximizar o logaritmo da função de verossimilhança (log-verossimilhança), que transforma o produto de probabilidades em uma soma, simplificando a matemática. A maximização é feita em relação aos coeficientes  $\beta$ .

Oseias Dias de Farias

[oseiasdfarias@gmail.com](mailto:oseiasdfarias@gmail.com)  
021 399 242-66

A intuição de máxima verossimilhança para a regressão logística é que um procedimento de busca valores para os coeficientes (valores Beta) que minimizam o erro nas probabilidades previstas pelo modelo para aqueles nos dados (por exemplo, probabilidade de 1 se os dados são da classe 1).

O processo de encontrar os valores para minimizar ou maximizar uma função é chamado de **Otimização**. Para qualquer problema de otimização, você pode tratá-lo em uma abordagem analítica ou em uma abordagem de aproximação numérica.

Uma **abordagem analítica** é determinística, o que significa que existe uma solução de forma fechada para resolver o problema de otimização. Você pode encontrar a expressão matemática da solução exata .

No entanto, uma solução de forma fechada raramente é obtida na modelagem estatística. **A regressão linear é um dos casos raros. A regressão logística não é um desses casos.**

Uma **abordagem de aproximação numérica** muito conhecida é chamada de **Gradient Descent**. Não é escopo do nosso curso falarmos sobre o gradiente

descendente - isso é escopo de machine learning! Porém, aqui vou abordar brevemente apenas para matar a curiosidade de quem tem interesse.

Em matemática, gradiente descendente é um algoritmo de otimização iterativa de primeira ordem para encontrar um mínimo local de uma função diferenciável. Uma descida de gradiente típica segue as etapas listadas abaixo.

- 1.** Inicialmente, podemos considerar  $\beta_1$  e  $\beta_0$  quaisquer valores (por exemplo,  $\beta_1 = 0$  e  $\beta_0 = 0$ ). Seja  $L$  a taxa de aprendizado.  $L$  determina a magnitude das alterações que aplicaríamos para atualizar os parâmetros. Quanto maior a taxa de aprendizado, mais rápido a função de custo se aproximaria do ponto mínimo. Se  $L$  for muito grande, seu otimizador estará disparando sobre a “curva” e perderá o ponto mínimo. Por outro lado, se  $L$  for muito pequeno, levaria muito tempo para atingir os mínimos. Portanto, precisamos definir a taxa de aprendizado estrategicamente.
- 2.** Calcule a derivada parcial da função de custo em relação a cada parâmetro (que discuti acima). Insira os valores de  $\beta_1$ ,  $\beta_0$ ,  $X_i$ ,  $Y_i$  em cada derivada parcial e calcule seus valores. Valores derivados determinariam a direção das mudanças nesses parâmetros.
- 3.** Atualize os parâmetros

Tanto a regressão linear quanto a logística podem usar esse método para encontrar os valores de coeficiente. Somente a regressão linear pode usar o OLS. Quando temos muitas variáveis independentes, o método do gradiente descendente é preferível para regressão linear por ser computacionalmente mais simples.

Se quiserem ler mais sobre gradiente descendente, [fizemos um texto aqui](#) para os alunos do PED, no módulo de Machine Learning.

Também indico os seguintes artigos:

<https://medium.com/@bruno.dorneles/regress%C3%A3o-linear-com-gradiente-descendente-d3420b0b0ff#:~:text=O%20Gradiente%20Descendente%20%C3%A9%20um,melhor%20se%20ajusta%20aos%20dados.>

[https://towardsdatascience.com/linear-regression-vs-logistic-regression-ols-ma](https://towardsdatascience.com/linear-regression-vs-logistic-regression-ols-maximum-likelihood-estimation-gradient-descent-bcfac2c7b8e4)  
[ximum-likelihood-estimation-gradient-descent-bcfac2c7b8e4](https://towardsdatascience.com/linear-regression-vs-logistic-regression-ols-maximum-likelihood-estimation-gradient-descent-bcfac2c7b8e4)

## PREMISSAS PARA UTILIZAÇÃO DE REGRESSÃO LOGÍSTICA

Para podermos usar a regressão logística, precisamos:

- Independência das observações: As observações devem ser independentes umas das outras.
- Ausência de multicolinearidade: As variáveis independentes não devem estar altamente correlacionadas entre si.
- Amostra de tamanho grande: O tamanho da amostra deve ser grande o suficiente para estimar corretamente os parâmetros do modelo. Uma diretriz geral é ter pelo menos 10-15 observações por preditor no modelo.
- Ausência de viés de variáveis omitidas: O modelo deve incluir todas as variáveis relevantes. Omitir variáveis importantes pode levar a estimativas de parâmetros viciadas (viés de confundimento).

## AVALIAÇÃO DE UM MODELO DE CLASSIFICAÇÃO

Alguns modelos podem nos fornecer um compilado de resultados muito semelhante ao que encontramos em regressão linear. Por exemplo, suponha que queremos prever se uma pessoa passará em uma prova (0 se não passou e 1 se passou) a partir de horas estudadas e do método de estudo (método A ou B). Uma parte da tabela pode ser vista abaixo:



result	hours	method
0	1	A
1	2	A
0	2	A
0	2	B
0	3	B

Em seguida, usamos um modelo de regressão linear. O resultado foi:

Logit Regression Results						
Dep. Variable:	result	No. Observations:	20			
Model:	Logit	Df Residuals:	17			
Method:	MLE	Df Model:	2			
Date:	Mon, 22 Aug 2022	Pseudo R-squ.:	0.1894			
Time:	09:53:35	Log-Likelihood:	-11.156			
converged:	True	LL-Null:	-13.763			
Covariance Type:	nonrobust	LLR p-value:	0.07375			
	coef	std err	z	P> z	[0.025	0.975]
Intercept	-2.1569	1.416	-1.523	0.128	-4.932	0.618
method[T.B]	0.0875	1.051	0.083	0.934	-1.973	2.148
hours	0.4909	0.245	2.002	0.045	0.010	0.972

Os valores na coluna coef da saída nos informam a alteração média nas probabilidades logísticas de aprovação no exame.

- Os coeficientes estão associados ao aumento de probabilidade

$$P = \frac{e^{b_0 + b_1*x_1 + b_2*x_2}}{1 + e^{b_0 + b_1*x_1 + b_2*x_2}}$$



Em que  $b_0$  é o intercepto (-2.1569),  $b_1$  é o coeficiente da feature 1 (por exemplo, método B: 0.0875) e  $b_2$  é o coeficiente da feature 2 (por exemplo, método B: 0.4909) e

Os valores em  $P>|z|$  coluna representam os p-valores para cada coeficiente.

- O método de estudo tem um p-valor de 0,934. Como esse valor não é menor que 0,05, significa que não há uma relação estatisticamente significativa entre as horas estudadas e a aprovação ou não do aluno no exame.
- As horas estudadas têm um p-valor de 0,045. Como esse valor é menor que 0,05, significa que há uma relação estatisticamente significativa entre as horas estudadas e a aprovação ou não do aluno no exame.

Para avaliar a qualidade do modelo de regressão logística, podemos observar duas métricas na saída:

Oseias Dias de Farias

#### 1. Pseudo R-quadrado

[oseias.dfarias@gmail.com](mailto:oseias.dfarias@gmail.com)

021.399.242-66

Esse valor pode ser considerado como o substituto do valor R-quadrado para um modelo de regressão linear. É calculado como a razão da função logarítmica maximizada do modelo nulo para o modelo completo.

Esse valor pode variar de 0 a 1, com valores mais altos indicando um melhor ajuste do modelo.

Neste exemplo, o pseudo valor de R-quadrado é 0,1894, é ligeiramente baixo. Isso nos diz que as variáveis preditoras no modelo não fazem um bom trabalho em prever o valor da variável de resposta.

#### 2. P-valor de LLR

Esse valor pode ser considerado como o substituto do p-valor para o p-valor F geral de um modelo de regressão linear.

Se esse valor estiver abaixo de um certo limite (por exemplo,  $\alpha = 0,05$ ), podemos concluir que o modelo geral é "útil" e é melhor para prever os



valores da variável de resposta em comparação com um modelo sem variáveis de previsão.

Neste exemplo, o valor-p LLR é 0,07375. Dependendo do nível de significância que escolhermos (por exemplo, 0,01, 0,05, 0,1), podemos ou não concluir que o modelo como um todo é útil.

Outra forma muito interessante de fazer a avaliação - inclusive mais comum - é a avaliação a partir da comparação entre as classes preditas pelo modelo e as classes verdadeiras de cada exemplo. Existem muitas métricas para medir quanto bom seu modelo é, e todas as métricas de classificação têm como objetivo comum medir quanto distante o modelo está da classificação perfeita, porém fazem isto de formas diferentes. Aqui vamos abordar as 2 principais: acurácia e acurácia balanceada.

## Matriz de confusão

Matriz de confusão é um tabela que mostra as frequências de classificação para cada classe do modelo. Ela vai nos mostrar as frequências:

- Verdadeiro positivo (true positive — TP): ocorre quando no conjunto real, a classe 1 foi prevista corretamente. Por exemplo, quando a mulher está grávida e o modelo previu corretamente que ela está grávida.
- Falso positivo (false positive — FP): ocorre quando no conjunto real, a classe 1 foi prevista incorretamente. Exemplo: a mulher não está grávida, mas o modelo disse que ela está.
- Falso verdadeiro (true negative — TN): ocorre quando no conjunto real, a classe 0 foi prevista corretamente. Exemplo: a mulher não estava grávida, e o modelo previu corretamente que ela não está.
- Falso negativo (false negative — FN): ocorre quando no conjunto real, a classe 0 foi prevista incorretamente. Por exemplo, quando a mulher está grávida e o modelo previu incorretamente que ela não está grávida.

Ao final teremos para o conjunto acima



		Valores preditos	
		Grávida	Não Grávida
Valores reais	Grávida	3	1
	Não grávida	2	4

Assim, nosso modelo:

- Previu grávida 3 vezes corretamente
- Previu não grávidas 4 vezes corretamente
- Previu grávida 1 vez incorretamente
- Previu não grávida 2 vezes incorretamente

É a partir da matriz de confusão que quase todas as nossas métricas são calculadas.

## Acurácia

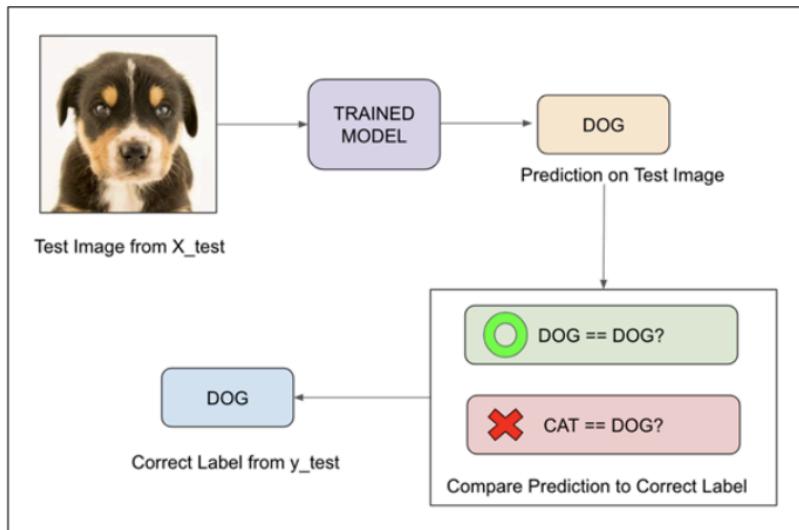
A acurácia simplesmente mede a frequência com que o classificador prevê corretamente. Podemos definir acurácia como a razão entre o número de previsões corretas e o número total de previsões.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Quando qualquer modelo fornece uma taxa de precisão de 99%, você pode pensar que o desempenho do modelo é muito bom, mas isso nem sempre é verdade e pode ser enganoso em algumas situações. Vou explicar isso com a ajuda de um exemplo.

Considere um problema de classificação binária, onde um modelo pode alcançar apenas dois resultados, qualquer um dos modelos fornece uma previsão correta ou incorreta. Agora imagine que temos uma tarefa de classificação para prever se uma imagem é um cachorro ou gato, conforme mostrado na imagem abaixo. Em um algoritmo de aprendizado supervisionado, primeiro ajustamos/treinamos um modelo nos dados de

treinamento e, em seguida, testamos o modelo nos dados de teste. Assim que tivermos as previsões do modelo a partir dos dados do  $X_{test}$ , as comparamos com os valores  $y_{train}$  (os rótulos corretos).



Oseias Dias de Farias  
Alimentamos a imagem do cão no modelo de treinamento. Suponha que o modelo preveja que este é um cachorro e, em seguida, comparamos a previsão com o rótulo correto. Se o modelo prevê que esta imagem é um gato e, em seguida, a comparamos novamente com o rótulo correto, ela estaria incorreta.

Repetimos esse processo para todas as imagens nos dados de teste. Eventualmente, teremos uma contagem de correspondências corretas e incorretas. Mas, na realidade, é muito raro que todas as correspondências incorretas ou corretas tenham o mesmo valor. Portanto, uma métrica não contará toda a história.

A acurácia é útil quando a classe alvo está *bem balanceada*, mas não é uma boa escolha para as **classes não balanceadas**. Imagine o cenário em que tínhamos 99 imagens do cachorro e apenas 1 imagem de um gato presente em nossos dados de treinamento. Então, nosso modelo sempre predizia o cachorro e, portanto, obtivemos 99% de acurácia, mas esse modelo é ruim pois simplesmente está prevendo tudo como a classe de maior quantidade.. Na vida real, os dados estão quase sempre desbalanceados. Por exemplo, e-mail de spam, fraude de cartão de crédito e diagnóstico médico. Portanto,

se quisermos fazer uma melhor avaliação do modelo a acurácia não é indicada para a maioria dos casos.

Uma alternativa à acurácia é utilizar a **acurácia balanceada** que não é influenciada pelo desbalanceamento das classes, porque os cálculos ocorrem em cima da taxa de verdadeiros positivos e verdadeiros negativos, como demonstrado na equação abaixo.. Logo, conseguindo chegar a um valor mais correto em relação aos acertos do modelo em relação às classes.

$$\text{Acurácia Balanceada} = \frac{1}{2} \left( \frac{VP}{VP + FN} + \frac{VN}{VN + FP} \right)$$

Para exemplificar, serão calculados os valores da acurácia e acurácia balanceada utilizando os valores da tabela abaixo para os cálculos.

	Não	Sim
Não	101668	3
Sim	36	95

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

O resultado encontrado para a acurácia é de 0,9996, ou seja, podemos logo imaginar que praticamente acertou todas as classes e o modelo está ótimo. Contudo, a maior parte dos acertos vieram da classe majoritária, enviesando o resultado.

Quando utilizamos a acurácia balanceada, no qual é levada em conta os acertos de cada classe de forma igualitária, o valor encontrado é de 0,8626. Isto mostra um valor mais próximo do quanto o modelo consegue acertar cada classe.

Apesar disso, mesmo usando ainda a acurácia balanceada, ainda temos uma visão global de acerto de todas as classes, então não conseguimos verificar o quanto um modelo acertou ou errou em relação a uma determinada classe do nosso interesse. Portanto, tenha o cuidado de avaliar seu modelo como um todo - taxa de verdadeiros positivos, falsos positivos, verdadeiros negativos

e falsos negativos.

Caso tenha interesse em saber mais sobre classificação, leia os seguintes artigos:

<https://www.google.com/url?q=https://towardsdatascience.com/introduction-to-logistic-regression-66248243c148&sa=D&source=docs&ust=1673476365364827&usg=AOvVaw1PRO1ZnINDNomx7d7UHJZr>

<https://medium.com/kunumi/m%C3%A9tricas-de-avalia%C3%A7%C3%A3o-em-machine-learning-classifica%C3%A7%C3%A3o-49340dcdb198>

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66



# 26. The End



É com muita alegria que anuncio que você cumpriu com sucesso seu e-book de estatística. Agora domina completamente pontos como estatística descritiva, gráficos, probabilidade, testes de hipótese, testes AB, regressões lineares e logística. Você tem em mãos a arma mais poderosa que eu poderia te dar no mundo de dados, ~~para se tornar um~~ ~~profissional~~ para quaisquer aprendizado posterior que você venha a ter. **021.399.242-66**

Agora é com você! Quer aprender técnicas mais robustas? Técnicas que fogem do analítico e vão para aproximação numérica? Minha recomendação é que você inicie seus estudos em machine learning! Para isso, deixo aqui uma sugestão de roteiro e cursos gratuitos para você começar a mergulhar no mundo da ciência de dados.

Veja aqui as indicações de trilha para analistas e cientistas de dados:



**Analistas  
de Dados**

Indicações de cursos no link da bio!

- 1 Excel  
Youtube
- 2 Estatística  
E.B.A - Estatística do Básico ao Avançado
- 3 SQL  
Udemy
- 4 Ferramenta para  
data viz  
DSA
- 5 Raciocínio  
Analítico  
E.B.A - Estatística do Básico  
ao Avançado
- 6 Python  
E.B.A e DSA

**DICA DE MILHÕES** *Oseias Dias de Faria*  
**oseias.dfarias@gmail.com**  
 Recomendo aprender Python em  
 conjunto com estatística e raciocínio  
 analítico!

**Cientistas  
de Dados**

Indicações de cursos no link da bio

- 1 Excel Básico  
Youtube
- 2 Estatística  
E.B.A - Estatística do Básico ao Avançado
- 3 Python  
E.B.A e DSA
- 4 Raciocínio Analítico  
E.B.A - Estatística do Básico ao Avançado
- 5 SQL  
Udemy + Livro
- 6 Algebra linear  
e Cálculo  
Youtube
- 7 Machine  
learning  
Coursera + Livro
- 8 Noções de MLOps  
Livro

Agora veja as indicações de cursos para cada uma dessas trilhas [CLIQUE AQUI](#)

## 👉 Acabou, Jéssica?

Não, meus caros, não acabou!

Com tudo que você vai aprender, você vai ter o conhecimento necessário para desempenhar a função e aplicar para qualquer vaga - garanto que vai se sair bem na parte técnica se você aprender o que mencionei na profundidade necessária.

Mas acredite, isso não te garante uma vaga! No mundo de hoje, nós precisamos de fato estar preparados para uma vaga - saber o que responder, como responder, ter um conhecimento profundo, etc

Estamos vivendo tempos complicados para todos - nós de dados não estamos a salvo dessa. Alta competição, menos vagas (especialmente para júnior). Por isso que, além desse aprendizado todo você precisa investir MUITO em:

- **Saber responder as perguntas técnicas e resolver os cases**

Parece meio óbvio, mas quando estamos estudando as vezes caímos na armadilha de não nos aprofundarmos tanto quanto deveríamos para responder o que vier pela frente durante o processo seletivo. Afinal, quanto é necessário? Será que já sei tudo? Será que estou fazendo da forma que é esperado? Calma, vou te trazer a solução já já!

- **Formar um portfólio**

Oseias Dias de Farias

[oseias\\_dias\\_farias@gmail.com](mailto:oseias_dias_farias@gmail.com)

021 399 242-66

O portfólio é sua vitrine. Você vai colocar ali os projetos mais relevantes que fez - e seja ousado! Não me vem colocar o projetinho famoso do Titanic que isso não pega mais! Tente ser criativo. É difícil ter uma ideia boa, eu sei. Mas te juro que vai fazer diferença na sua procura.

Não conseguiu ter uma ideia bacana? Use e abuse dos sites de projetos que coloquei acima, veja alguns vídeos de ideias interessantes no Youtube. Não precisa inventar a roda, mas sim fazer algo relevante e que não seja batido!

Fora isso, faça projetos end-to-end. Você que quer ser analista de dados, pense projetinhos que você faça todo a coleta de dados (com web scraping, por exemplo), transformações dos dados com SQL, análise estatísticas (descritiva, inferência, etc), ganho em métricas de business ou financeiras do seu projeto e/ou recomendações futuras - as quais devem fazer sentido de acordo com seus dados (por exemplo, recomendamos focar a campanha de marketing no grupo X pois há maior conversão da métrica Y, conforme visto nos dados).

E por fim, lembre-se, não é sobre quantidade. É sobre qualidade

- **Personal branding**



Já o personal branding é o famoso "vender o seu peixe". Mas não pense que isso significa postar certificados no Linkedin.

A lógica do personal branding se baseia na ideia de que todos deixamos uma marca e entregamos valor. Quando um recrutador faz uma busca por candidatos, uma das coisas que brilha aos olhos é justamente o que você entrega à comunidade de dados. Não tem a ver com ser vaidoso, tem a ver com ajudar, ser participativo e ativo na comunidade e entregar valor. É assim que você vai se destacar!

[Veja esse post](#) para pegar algumas dicas de como você pode fazer o seu personal branding

⚠ Já deixo um GRANDE aviso aqui: De nada adianta se você não tiver conhecimento e vontade de aprender. O personal branding tem que ser trabalhado junto com sua rotina de estudos!

É por isso que, para te ajudar nessa jornada eu criei o **P.E.D - Preparatório para Entrevistas em Dados**. Veja mais detalhes abaixo

Oseias.D.Farias@gmail.com  
021.399.242-66

## 👉 Se preparando para seu processo seletivo | P.E.D

O P.E.D é a mentoria em grupo que vai ajudar você a se preparar para conseguir a tão sonhada vaga em dados

- ✓ Trilhas com **conteúdo profundo**, misturando teoria e prática
- ✓ Perguntas para **entrevistas reais**
- ✓ **Cases** para seu **portfólio**
- ✓ **Workshops** e **palestras** exclusivas
- ✓ **Convidados ilustres** da área de dados para dar palestras e aulas exclusivas
- ✓ **Lives** para mostrar possíveis resoluções de cases e tirar dúvidas
- ✓ **Explicação de todo o conteúdo** que envolve o tema da pergunta
- ✓ **Comunidade exclusiva** para alunos no Discord com acesso direto à prof Renata Biaggi

- ✓ Oportunidade de fazer **networking** e construir **grupos de estudos**
- ✓ **Suporte a dúvidas** quando você precisar

Quer se juntar a nós? Acesse o link abaixo 👉

[www.renatabiaggi.com/ped](http://www.renatabiaggi.com/ped)

Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66



# 27. Referências Bibliográficas



Oseias Dias de Farias  
oseias.dfarias@gmail.com  
021.399.242-66

Bussan, W., Morettin, P. - Estatística Básica - Editora Saraiva - 2010 - 6th ed

Frost, J. - Hypothesis testing - An intuitive guide for making data driven decisions - Jim Frost - 2020 - 1st ed

Frost, J. [2] - Regression Analysis: An Intuitive Guide for Using and Interpreting Linear Models - Jim Frost - 2020 - 1st ed

Huyen, C. - Designing machine learning systems - Editora O'Reilly - 2022 - 1st ed

Knafllic, C.N - Storytelling com dados: Um guia sobre visualização - Editora Alta Books - 2019

Larson, R., Farber B. - Estatística Aplicada - Editora Pearson - 2016 - 6th ed

Pinheiro, J., Cunha, S., S. Santiago, Gomes, G. - Probabilidade e Estatística: Quantificando a incerteza - Elsevier Editora Ltda - 2012