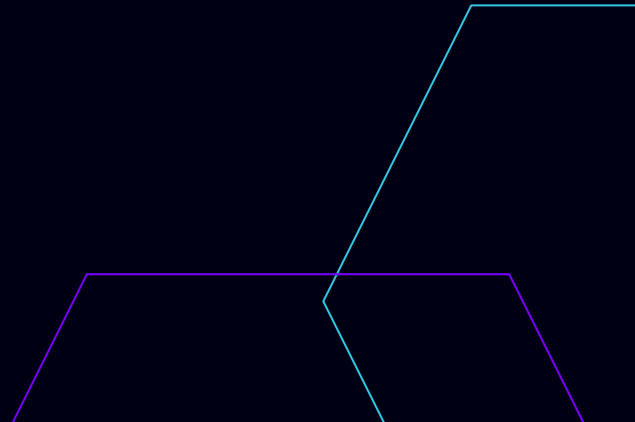


A series of overlapping, thin, light blue and purple lines forming a geometric pattern in the top-left corner of the slide.

AceLeraDev Data Science

Pensamento estatístico em Python

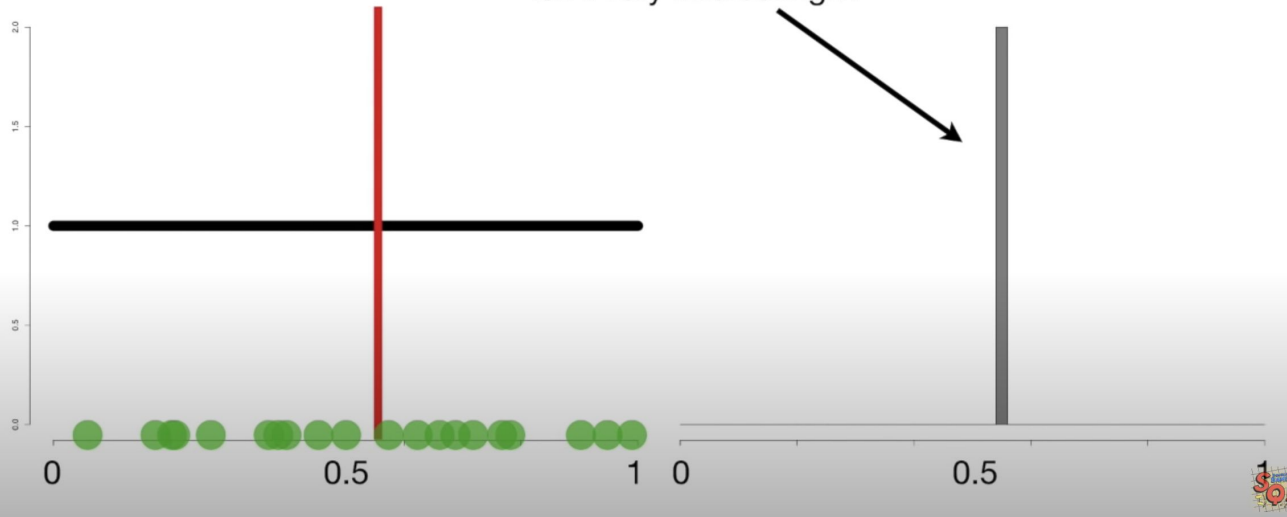
A series of overlapping, thin, light blue and purple lines forming a geometric pattern in the bottom-right corner of the slide.

Teorema do limite central

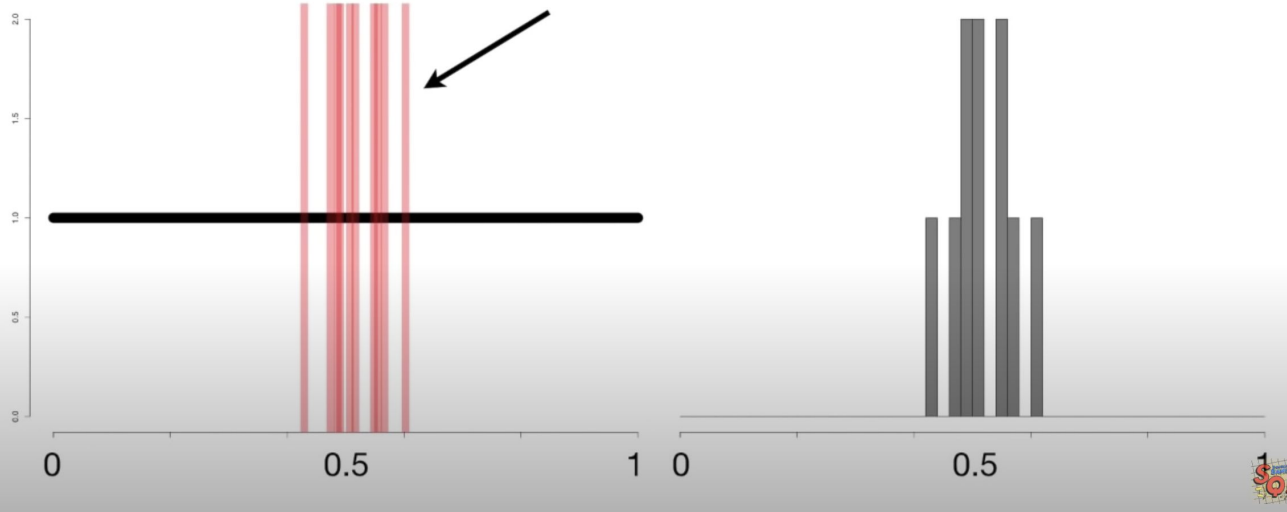
/

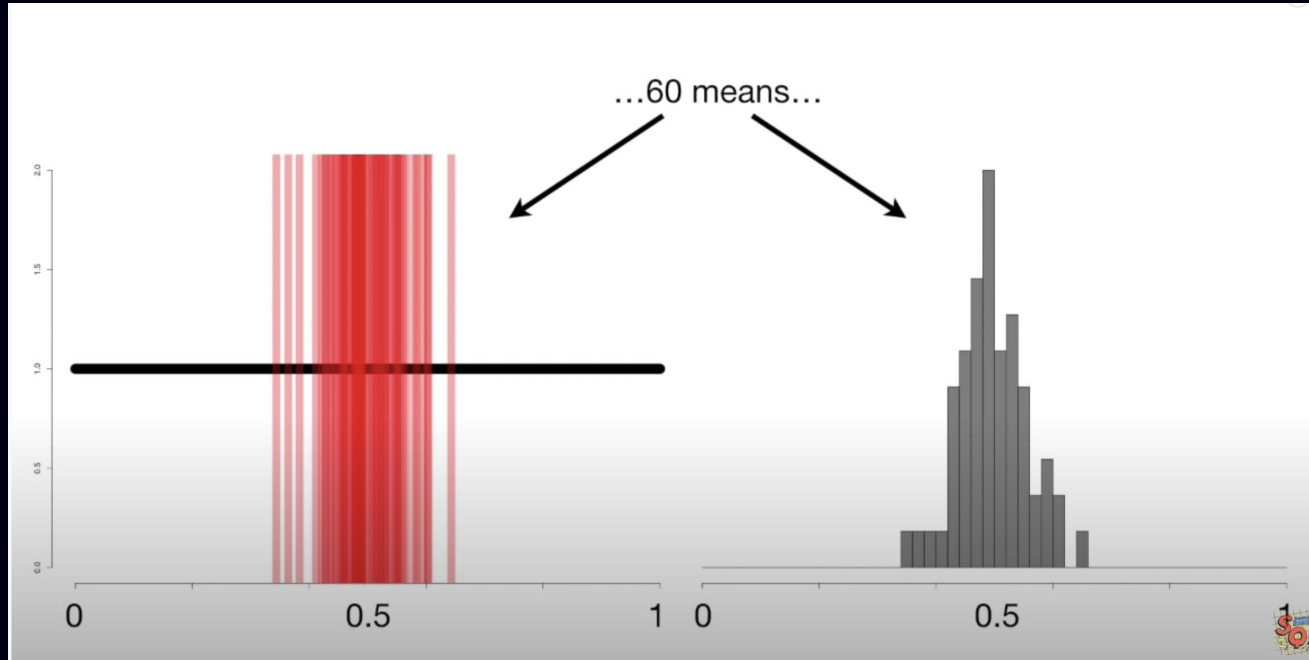
- Conceito básico para usado para vários teoremas estatísticos;
- **A distribuição das médias retiradas de uma amostra cuja população tenha qualquer distribuição terão uma distribuição normal.**

Since we only have one
mean value, the histogram
isn't very interesting...

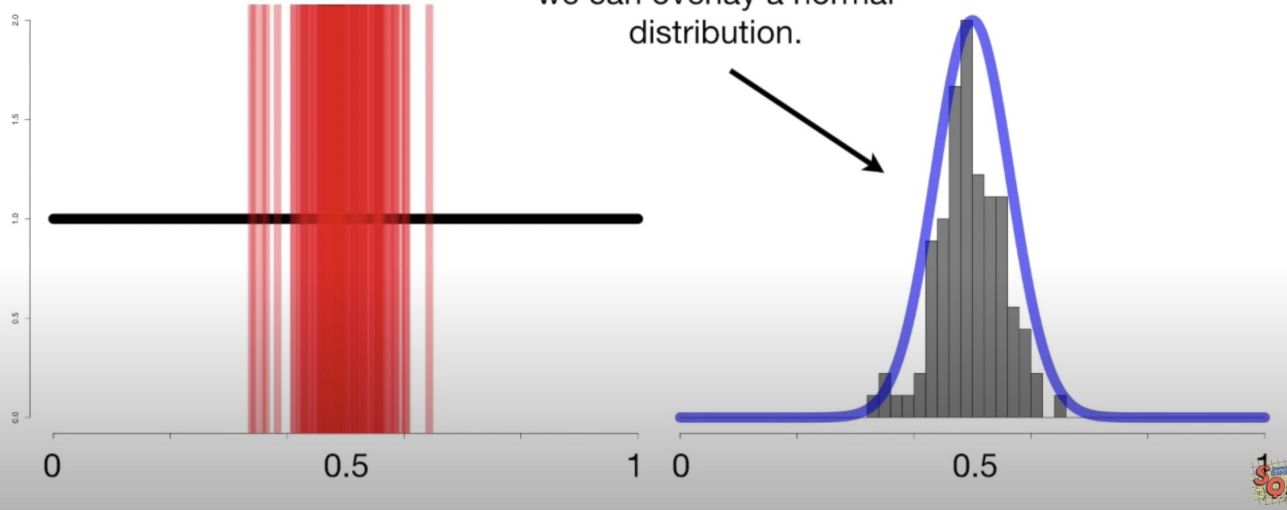


...but after we collect 10
more samples and
calculate 10 more means...





However, to make super easy to
see that the
means are normally distributed,
we can overlay a normal
distribution.

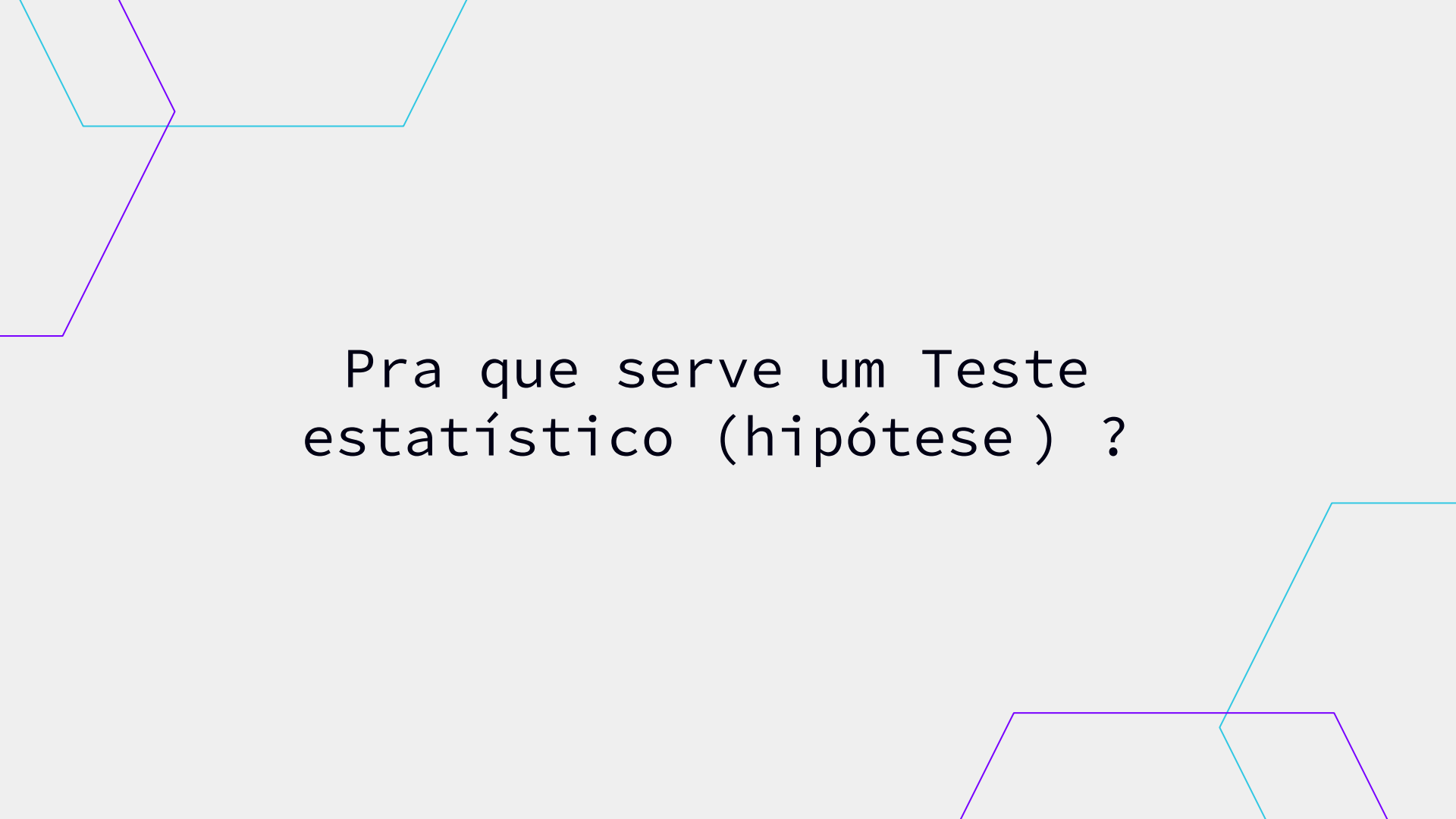


A series of overlapping, thin, light blue and purple lines forming a geometric pattern in the top-left corner of the slide.

AceLeraDev Data Science

Pensamento estatístico em Python

A series of overlapping, thin, light blue and purple lines forming a geometric pattern in the bottom-right corner of the slide.

The slide features decorative geometric lines in teal and purple. In the top-left corner, there are overlapping teal and purple lines forming a series of connected segments. In the bottom-right corner, there are similar teal and purple lines, also forming connected segments.

Pra que serve um Teste
estatístico (hipótese) ?

Teste estatístico (hipótese)

/

- Hipótese, estatístico, significância:
 - Permite tomada de decisão entre 2 ou mais hipóteses
 - Define-se uma Hipótese Nula (**H0**) e uma ou mais hipóteses alternativas (**H1,H2...HN**);
- **H0: é a assumida como verdadeira, aquilo de se quer testar;**
- **HN: considerada quando H0 é falsa (não possui relevância estatística)**

Teste estatístico (hipótese)

/

- **Testes de média:**

- A garrafa de cerveja padrão tem 600ml com uma amostra de 50 garrafas podemos continuar dizendo que uma garrafa possui 600ml?

- **Teste de proporções:**

- Uma fábrica declara que no máximo 5% da sua produção vem do defeito, em uma amostra de 100 unidades encontramos 7 defeituosas. Os números da fábrica estão corretos?

	Hipótese nula H_0 é verdadeira	Hipótese nula H_0 é falsa
Hipótese nula H_0 é rejeitada	Erro do tipo I	Não há erro
Hipótese nula H_0 não é rejeitada	Não há erro	Erro do tipo II

n = 165	Predicted: No	Predicted: Yes
Actual: No	50	10
	5	100
Actual: Yes		

ref

Type I error
(false positive)



Type II error
(false negative)



Figure 3.1 Type I and Type II errors

A series of overlapping geometric shapes, primarily triangles and quadrilaterals, in shades of teal and purple, located in the top-left corner of the slide.

AceLeraDev Data Science

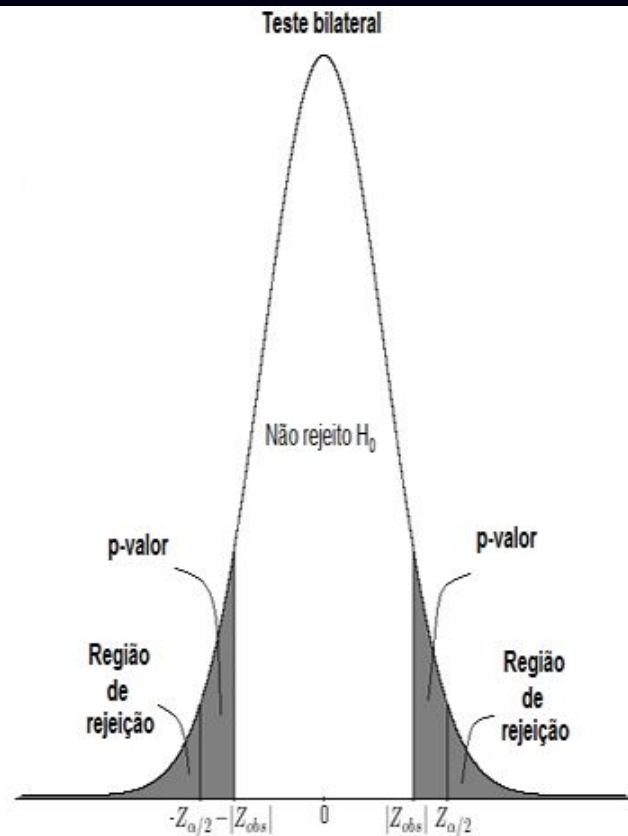
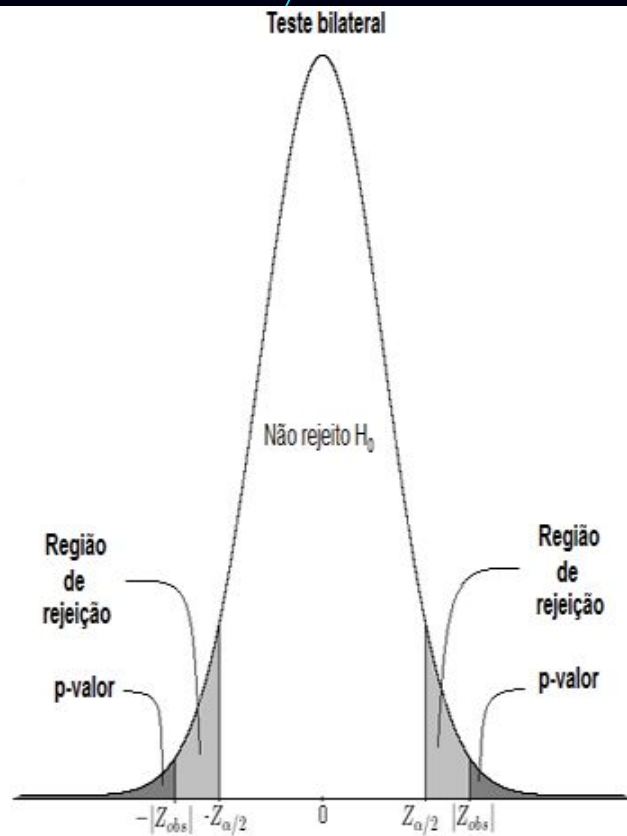
Pensamento estatístico em Python

A series of overlapping geometric shapes, primarily triangles and quadrilaterals, in shades of teal and purple, located in the bottom-right corner of the slide.

P-Valor (p-value)

/

- Probabilidade de obter uma estatística de teste **igual ou maior** que a observada em uma amostra **para H0 (hipótese nula)**;
- Quanto **menor o valor de p** maior a chance de se rejeitar a hipótese nula;
- Define-se alpha antes do experimento!!! (p-value hacking)
- [ref](#)
- [ref 2](#)



[ref](#)

P-VALUE

INTERPRETATION

0.001	}	HIGHLY SIGNIFICANT
0.01		
0.02		
0.03		
0.04	}	SIGNIFICANT
0.049		
0.050	}	OH CRAP. REDO CALCULATIONS.
0.051	}	ON THE EDGE OF SIGNIFICANCE
0.06		
0.07	}	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08		
0.09		
0.099	}	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1		



AceLeraDev Data Science

Pensamento estatístico em Python (Aula 1)



T-test (Student T-test)

/

- Baseia-se na **distribuição t de Student**
 - Distribuição simétrica;
 - Semelhante à curva normal;
 - Grau de liberdade;
- Comparar 2 grupos
 - Utilizando a média entre os valores;
 - Desvio padrão e variância
- [ref](#)

Teste t de Student

UNIVESP

Shapiro-Wilk

/

- Teste de Normalidade
 - Utiliza uma amostra de uma população para validar se a mesma está **distribuída normalmente**
 - Calcular a probabilidade usando essa distribuição normal
- Variância, média e desvio padrão
- H_0 : Amostra **provém** de uma normal;
- H_1 : Amostra **não provém** de uma normal;
- **Não funciona muito bem com mais de 5000 amostras.**

Jarque-Bera

/

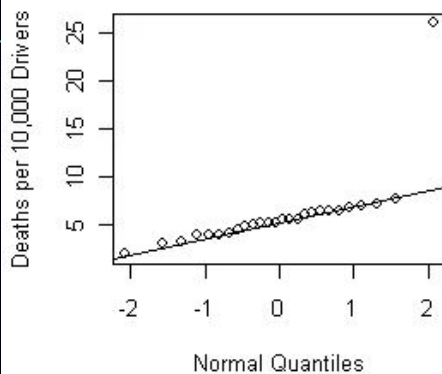
- Teste de normalidade
 - Validar se existe desvio padrão;
- Curtose e Assimetria
- H_0 : Amostra **provém** de uma normal;
- H_1 : Amostra **não provém** de uma normal;
- [ref](#)

Gráfico Q-Q (QQ-Plot)

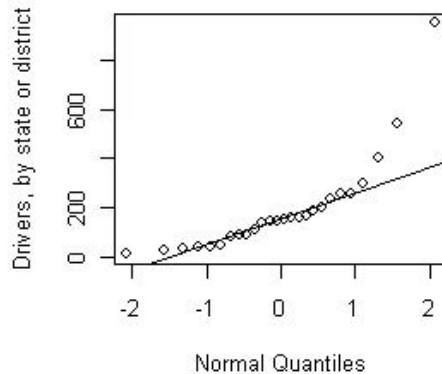
/

- Gráfico de Quantil-Quantil (Quartil-Quartil);
- Compara a distribuição de duas probabilidades;
 - Entre duas variáveis;
 - Entre uma variável e “quartis teóricos”
- Ajuda a validar se um distribuição é normal.

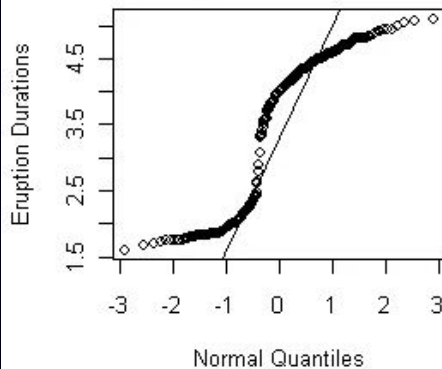
**Deaths per 10,000 Drivers,
by State or District**



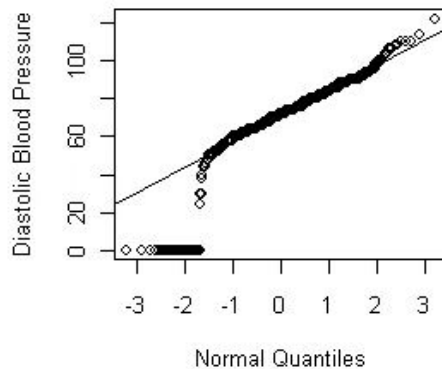
**Drivers by State or District,
Tens of Thousands**



**Old Faithful Geyser
Eruption Durations**

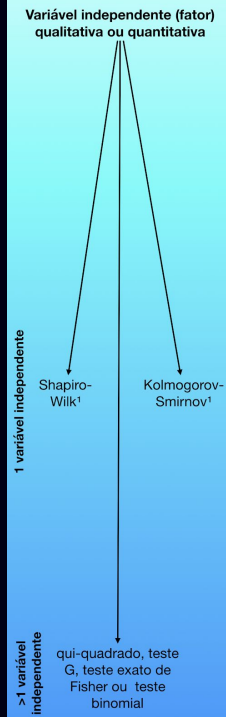


**Pima Indians Data
Diastolic Blood Pressure**



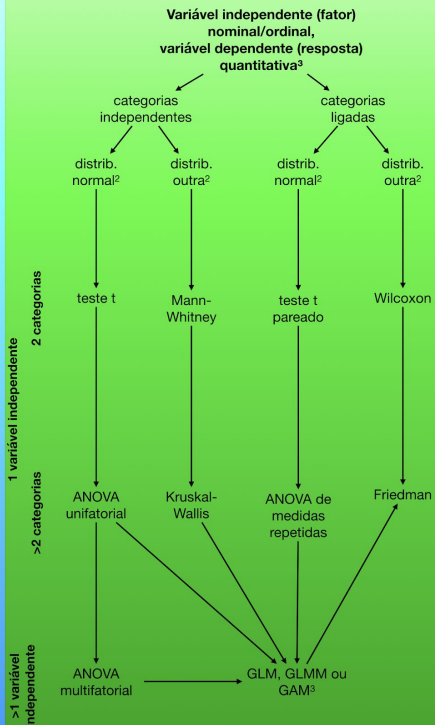
Qual teste estatístico devo usar?

Qual é a distribuição dos meus dados?



1. Costumam ser usados para testar a normalidade dos dados.

Há uma relação entre as minhas variáveis?



2 categorias

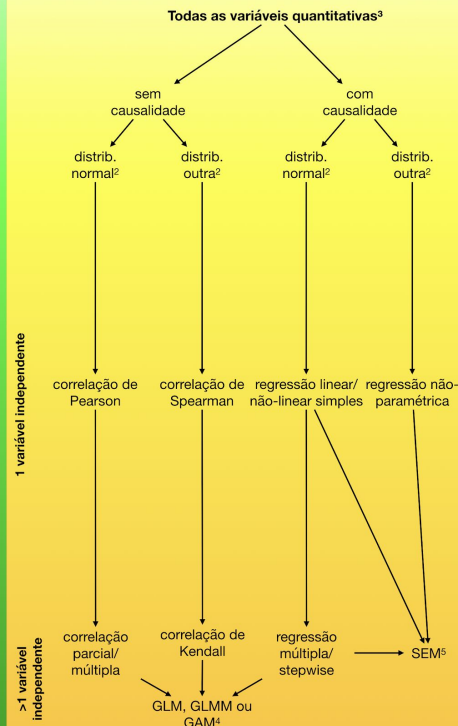
>2 categorias

>1 variável independente

2. Cheque sempre os **pressupostos** de cada teste. Em alguns casos, a distribuição da variável independente deve ser normal. Em outros, quem deve ser normal é a distribuição das diferenças ou dos erros. Fique atento!

3. Quando a distribuição (variáveis, diferenças, erros etc.) não é normal, pode-se usar um modelo linearizado (GLM), baseado em **outra distribuição** de probabilidade teórica. Quando há **fatores qualitativos e quantitativos** no mesmo teste, pode-se usar um modelo misto (GLMM). Modelos aditivos (GAM) podem ser usados, quando se nota que a relação não parece ser linear.

Por Marco A. R. Mello (adaptado de Jutta Schmid)
Versão 6, 24/05/2019
<https://marcoarmello.wordpress.com>



1 variável independente

>1 variável independente

4. Se a a variável independente for quantitativa, mas a variável dependente for nominal e **binária** (e.g., sim ou não), você pode usar uma **regressão logística** ou GLM com função logit.

5. Quando o modelo inclui relações diretas e indiretas entre as variáveis, recomenda-se usar **modelos de equações estruturais** (SEM). Essa família de testes inclui análise de caminhos (PA) e análise de variáveis latentes (LatVaAn), por exemplo.

Links

/

- <https://machinelearningmastery.com/a-gentle-introduction-to-normality-tests-in-python/>
- https://pt.wikipedia.org/wiki/Testes_de_hip%C3%B3teses
- <https://marcoarmello.wordpress.com/2012/05/17/qualteste/>
- https://en.wikipedia.org/wiki/Jarque%E2%80%93Bera_test
- <https://analysereal.com/tag/jarque-bera/>
- <https://www.azziengenharia.com/single-post/2017/02/08/0-Teste-de-Jarque-Bera-para-a-Verifica%C3%A7%C3%A3o-da-Hip%C3%B3tese-de-Normalidade-dos-Res%C3%ADduos-em-Regress%C3%B5es-para-Avalia%C3%A7%C3%A3o-de-Im%C3%B3veis>
- <http://www.portalection.com.br/inferencia/64-teste-de-shapiro-wilk>