



Centro de Inovação em  
Inteligência Artificial  
para a Saúde

## Curso de Introdução à Análise de Dados em Saúde com Python

### Estatística Descritiva

**Prof. D.r Juliano Gaspar**

**E-mail:** julianogaspar@gmail.com

**Lattes:** <http://lattes.cnpq.br/3926707936198077>

**Orcid ID:** <https://orcid.org/0000-0003-0670-9021>



Centro de Inovação em  
Inteligência Artificial  
para a Saúde

Diretor do CI-IA Saúde

Prof. Virgílio Augusto Fernandes Almeida (UFMG)

Vice-diretor

Prof. Antônio Luiz Pinho Ribeiro (UFMG/Unimed-BH)

Coordenadora de educação e difusão do conhecimento

Profª Zilma Silveira Nogueira Reis (UFMG)

Coordenador de transferência de tecnologia

Prof. Gilberto Medeiros Ribeiro (UFMG)

Mais informações

<https://ciia-saude.medicina.ufmg.br>

Realização



Centro de Inovação em  
Inteligência Artificial  
para a Saúde

Apoio

U F M G

## Professores Conteudistas

**Prof. Juliano de Souza Gaspar**  
Faculdade de Medicina da UFMG

**Profa. Zilma Silveira Reis**  
Faculdade de Medicina da UFMG

**Profa. Ana Paula Couto da Silva**  
Faculdade de Ciências da Computação da UFMG

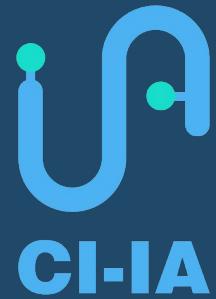
**Profa. Cristiane dos Santos Dias**  
Faculdade de Medicina da UFMG

## Alunos de Iniciação Científica

**André Soares da Silva**  
Graduando de Medicina da UFMG

**Yago Jean de Almeida Nogueira**  
Graduando de Medicina da UFMG

**Teresa Vitória Carvalho Rocha**  
Graduando de Engenharia de Produção da UFMG



# Estatística Descritiva

## Tabelas de Frequência (Variáveis categóricas)

N Válidos (%)

<b>Masculino</b>	853	51.69697
------------------	-----	----------

<b>Feminino</b>	797	48.30303
-----------------	-----	----------

N Válidos (%) Global (%)

<b>Masculino</b>	853	51.7	49.94
------------------	-----	------	-------

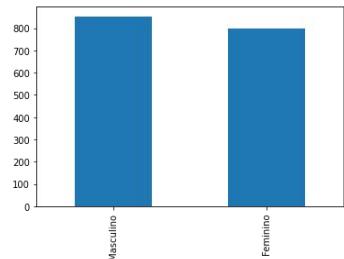
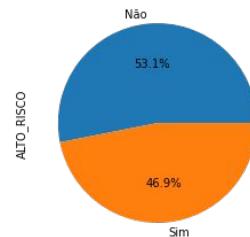
<b>Feminino</b>	797	48.3	46.66
-----------------	-----	------	-------

<b>NaN</b>	58	NaN	3.4
------------	----	-----	-----

<b>Total</b>	1708	100%	100%
--------------	------	------	------

Gráficos para variáveis categóricas

Pizza, colunas, barras



N Válidos (%) Global (%)

<b>Prematuro</b>	291	17.26	17.04
------------------	-----	-------	-------

<b>Termo</b>	925	54.86	54.16
--------------	-----	-------	-------

<b>Termo-precoce</b>	470	27.88	27.52
----------------------	-----	-------	-------

<b>NaN</b>	22	NaN	1.29
------------	----	-----	------

<b>Total</b>	1708	100%	100%
--------------	------	------	------

## Medidas de resumo (Variáveis numéricicas)

Conjunto de estatísticas descritivas que permitem uma avaliação concisa de **grandes quantidades** de valores.

1

Medidas de Tendência Central

2

Medidas de Dispersão

3

Medidas de Posição

PREMATURO_PED	IG_TERMO	HOUVE_CESAREA	HOUVE_LACERACAO	HOUVE_CM	PARIDADE	PESO_VIAVEIS	PESO_US	PESO_ALTA
Não	Termo-precoce	Não	Não	Sim	Primípara	3590.0	2590.0	3440.0
Sim	Prematuro	Não	Não	Nulípara	2660.0	1660.0	2510.0	
Não	Termo	Sim	Sim	Sim	Nulípara	3075.0	2075.0	2925.0
Sim	Prematuro	Não	Não	Sim	Nulípara	3505.0	2755.0	3355.0
Não	Termo-precoce	Não	Não	Sim	Multipara	3405.0	2405.0	3255.0
Não	Termo-precoce	Não	Não	Sim	Primípara	1625.0	1625.0	2475.0
Não	Termo	Não	Não	Sim	Primípara	2975.0	1965.0	2815.0
Não	Termo-precoce	Não	Não	Sim	Primípara	3570.0	2570.0	3420.0
Não	Termo	Não	Não	Sim	Multipara	2740.0	1740.0	2590.0
Não	Termo	Não	Não	Sim	Nulípara	2945.0	1945.0	2795.0
Não	Termo-precoce	Não	Não	Sim	Nulípara	3230.0	2230.0	3080.0
Não	Termo	Não	Não	Sim	Primípara	2980.0	1980.0	2830.0
Não	Termo	Não	Não	Sim	Primípara	3120.0	2120.0	2970.0
Não	Termo	Não	Não	Não	Nulípara	3055.0	2055.0	2905.0
Não	Termo-precoce	Não	Não	Não	Multipara	3550.0	2580.0	3430.0
Sim	Prematuro	Não	Não	Não	Nulípara	2150.0	850.5	2065.0
Não	Termo-precoce	Não	Não	Não	Nulípara	NaN	NaN	NaN
Não	Termo-precoce	Não	Não	Sim	Multipara	3130.0	2130.0	2980.0
Sim	Prematuro	Não	Não	Não	Nulípara	2325.0	1325.0	2175.0
Não	Termo	Não	Não	Não	Nulípara	3050.0	2050.0	2900.0
Não	Termo	Não	Não	Primípara	2895.0	1895.0	2745.0	
Não	NaN	Não	Não	Sim	Primípara	3190.0	2190.0	3040.0
Não	Prematuro	Não	Não	Não	Nulípara	2870.0	1870.0	2720.0
Não	NaN	Não	Não	Não	Nulípara	2720.0	1720.0	2570.0
Não	NaN	Não	Não	Primípara	3290.0	2290.0	3140.0	
Não	NaN	Não	Não	Sim	Nulípara	NaN	NaN	150.0
Não	NaN	Não	Não	Não	Nulípara	3160.0	2160.0	3010.0
Não	NaN	NaN	NaN	Primípara	3220.0	2220.0	3070.0	
Não	NaN	Não	Não	Nulípara	3060.0	2060.0	2910.0	



## Medidas de resumo

1

### Medidas de Tendência Central

- Média
- Mediana
- Moda

2

### Medidas de Dispersão

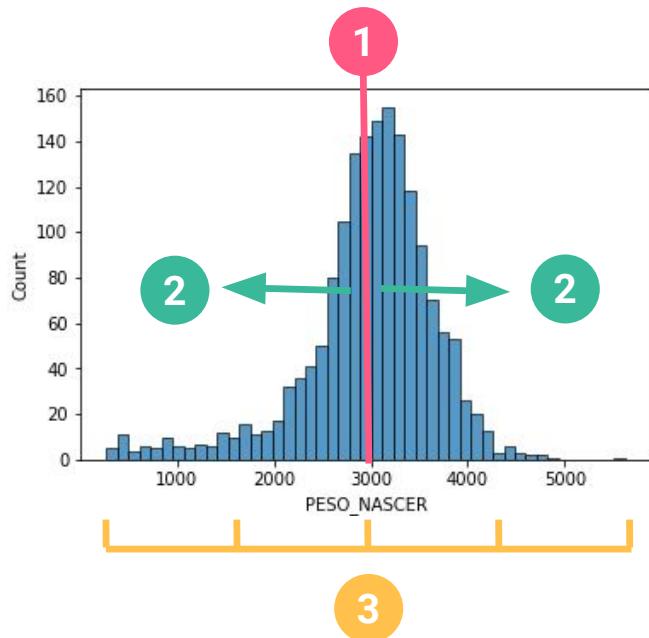
- Desvio Padrão
- Variância
- Mínimo, máximo (amplitude)

3

### Medidas de Posição

- Quartis      25%
- Decis          10%
- Percentis    1%

Gráficos para variáveis numéricas - distribuições  
Histograma, boxplot



## Medidas de tendência central: Média, Mediana e Moda

1

### Média

- Soma de todos os valores absolutos / número de observações
- Acompanha-se geralmente de medida de dispersão (Desvio padrão)
- Vantagem: algebricamente definida
- Desvantagens: distorcida por valores extremos

### Mediana

- Refere-se ao valor do meio, a partir dos dados ordenados em ordem crescente ( $p_{50}$ )
- Acompanha-se geralmente de medida de dispersão (amplitude)
- Se  $n=ímpar$  a mediana é o valor do meio. Se  $n=par$ , a mediana é a média dos valores centrais
- Vantagem: não é distorcida por valores extremos
- Desvantagens: leva em consideração a ordem e não os valores em si

### Moda

- É o valor mais frequente
- Pode ser usada para variáveis não-numéricas (categóricas ou nominais)
- O conjunto de dados pode ser amodal, unimodal, bimodal, multimodal

## Medidas de tendência central: Média, Mediana e Moda

2

### Desvio padrão

- O desvio padrão é raiz quadrada da variância
- Acompanha-se geralmente de medida de dispersão (Desvio padrão)
- Vantagem:
  - Usa todos os dados.
  - É definida algebricamente.
  - Mesma unidade que os dados e fácil de interpretar.
- Desvantagens: Sensível a valores extremos. Não apropriada em distribuições enviesadas.

### Variância

- É a soma dos quadrados dos desvios à média dividido pelo N° casos menos um.
- Vantagem: Usa todos os dados. Definida algebricamente.
- Desvantagens:
  - A unidade é o quadrado da unidade dos dados.
  - Sensível a valores extremos.
  - Não apropriada em distribuições enviesadas.

### Âmbito (amplitude) - Mínimo e Máximo

- É a diferença entre o valor maior e menor.
- Vantagem: Fácil de calcular.
- Desvantagem: Usa apenas dois valores. Distorcido por valores extremos.

## Medidas de tendência central: Média, Mediana e Moda

3

### Medidas de posição - Separatrizes

- Separatrizes são valores da distribuição que a dividem em partes quaisquer.
- Essas medidas, Quartis – Decis – Percentis, são juntamente com a Mediana.
- Conhecidas pelo nome genérico de separatrizes.

#### Mediana

- A mediana, apesar de ser uma medida de tendência central, também uma separatriz de ordem  $\frac{1}{2}$ .
- Divide a distribuição em duas partes iguais (50%).

#### Quartis

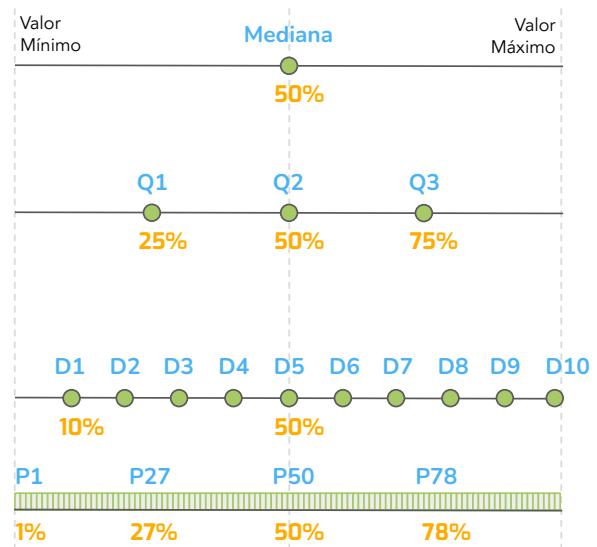
- Divide a distribuição em quatro partes iguais (25%).
  - Q1 (1º quartil ou quartil inferior)
  - Q2 (2º quartil ou quartil médio)
  - Q3 (3º quartil ou quartil superior)

#### Decis

- Divide a distribuição em 10 partes iguais (10%).

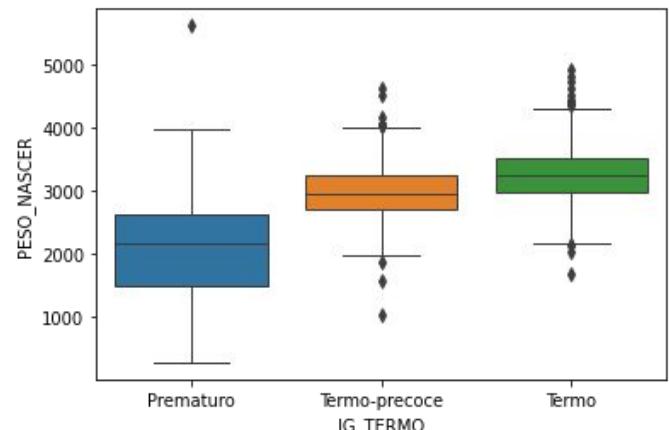
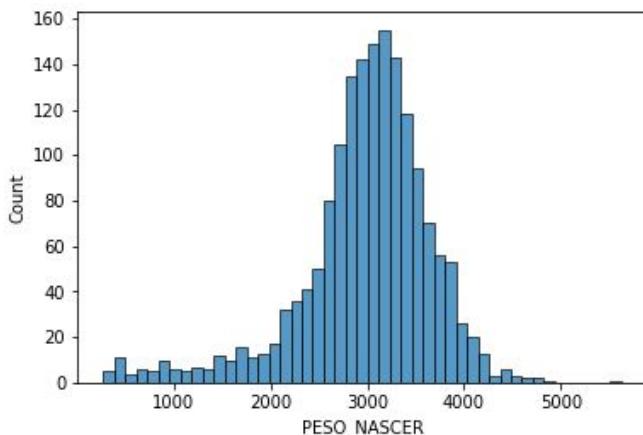
#### Percetis ou Centil

- Divide a distribuição em 100 partes iguais (1%).



## Medidas de resumo - Gráfico para variáveis numéricas

Gráficos para variáveis numéricas - distribuições  
Histograma, boxplot



## Medidas de resumo

### Distribuição



Distribuição Simétrica  
Ou tende p/ Curva Normal

### Tendência Central

Média  
**2977,9**

### Dispersão

Desvio padrão  
**691,1**

PESO_NASCER	
count	1680.0
mean	2977.9
std	691.1
min	270.0
25%	2710.0
50%	3065.0
75%	3390.0
max	5625.0



Distribuição Assimétrica

### Mediana

**2**

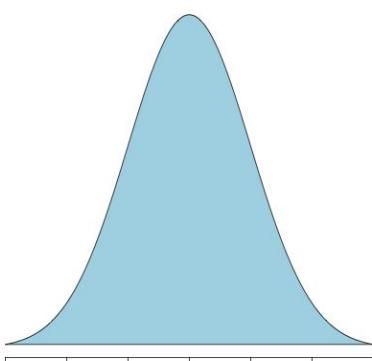
(Mínimo - Máximo)  
Amplitude / âmbito

**1 - 63**  
**62**

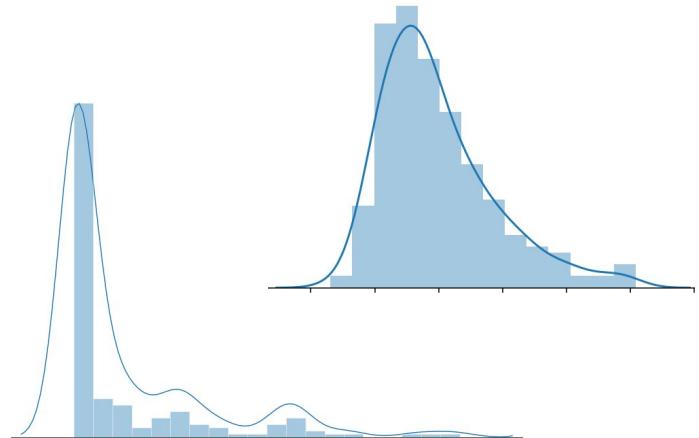
DURACAO_INT	
count	1708.0
mean	3.0
std	4.3
min	1.0
25%	2.0
50%	2.0
75%	3.0
max	63.0

## Distribuições

Histograma de frequência é um gráfico de barras em ordem crescente de valores.  
Na horizontal encontram-se as classes de valores e na vertical a sua frequência de ocorrência.



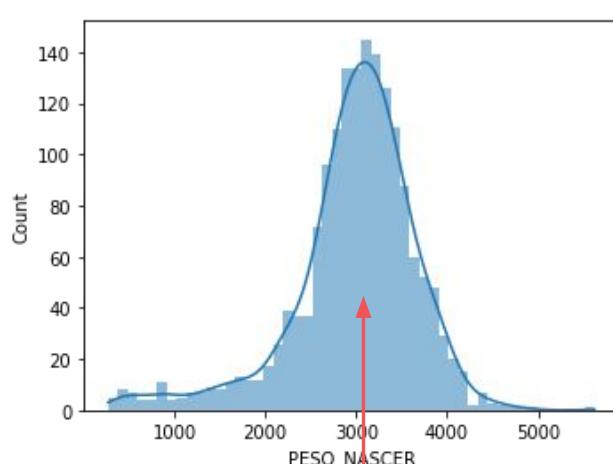
Distribuição Simétrica  
Distribuição Gaussiana



Distribuição Assimétrica  
Distribuição não-Gaussiana

## Distribuições

Distribuição simétrica



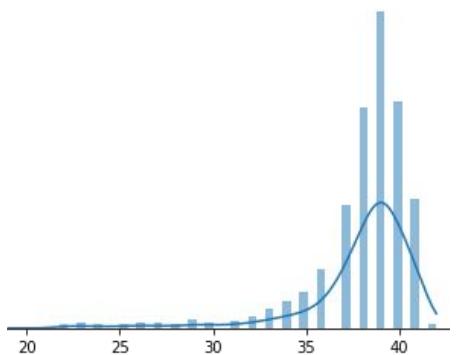
Média  
Mediana  
Moda

	PESO_NASCER
count	1680.0
mean	2977.9
std	691.1
min	270.0
25%	2710.0
50%	3065.0
75%	3390.0
max	5625.0

## Distribuições assimétricas

As distribuições consideradas **assimétricas** apresentam uma “cauda” em uma das extremidades, quando está à direita, é positivamente assimétrica, e se está à esquerda, é negativamente assimétrica.

Para verificar o tipo e o **grau da assimetria** da distribuição utiliza-se a medida estatística dimensional denominada de Coeficiente de Assimetria de Pearson.



$$|As| < 0,15$$

⇒ simétrica

$$0,15 < |As| < 1,0 \Rightarrow \text{assimetria moderada}$$

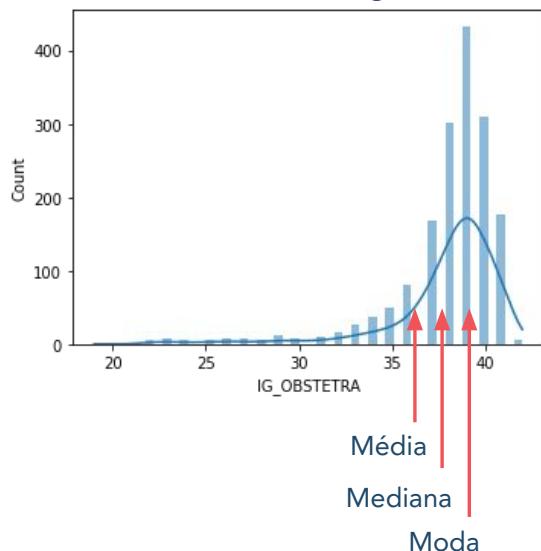
$$|As| > 1$$

⇒ assimetria forte

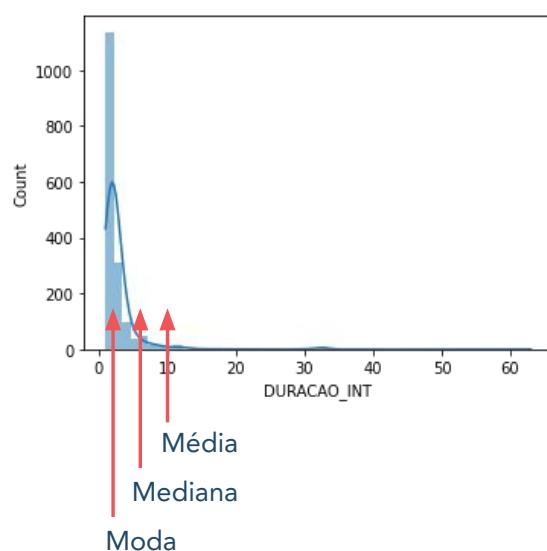
IG_OBSTETRA	
count	1686.0
mean	37.9
std	3.1
min	19.0
25%	37.0
50%	39.0
75%	40.0
max	42.0

## Distribuições assimétricas

Assimetria negativa

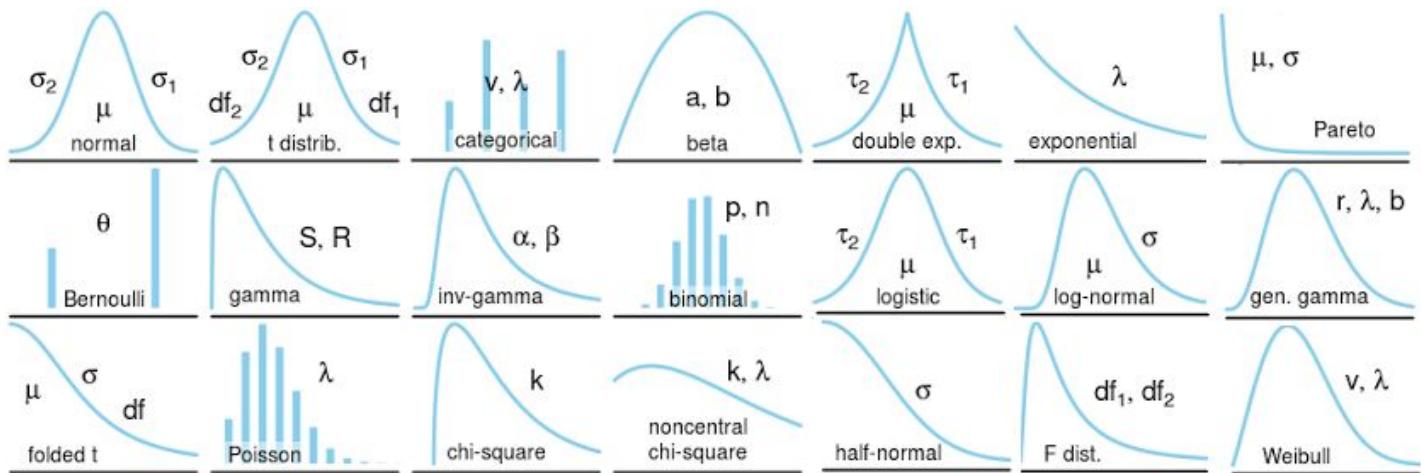


Assimetria positiva



## Distribuições

Existem diversos tipos de distribuições e ao longo do tempo foram sendo estudadas, denominadas e definidas matematicamente.

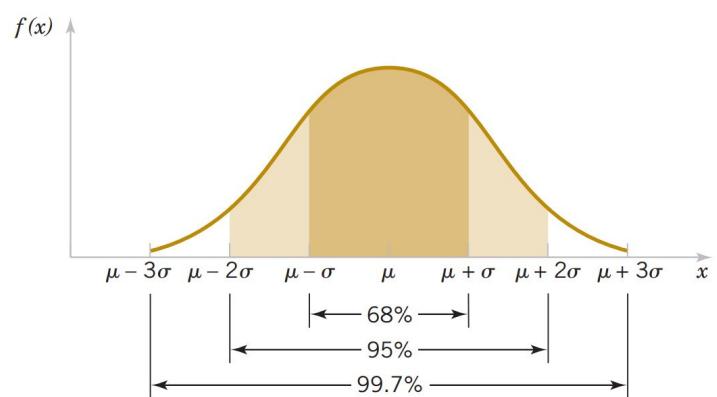
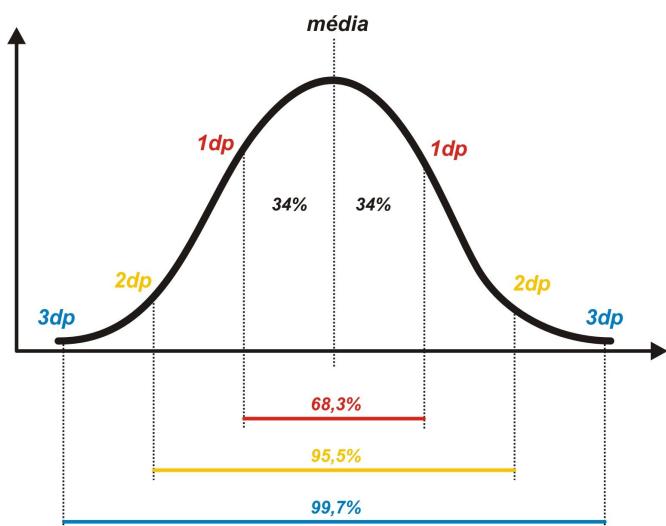


## Distribuição Normal

A distribuição normal possui dois parâmetros:

- a média ( $\mu$ )
- e a variância ( $\sigma^2 > 0$ ) que descreve o grau de dispersão.
- A dispersão é descrita em unidades padrão, ou seja desvio padrão ( $\sigma$ ).

Desvio padrão ( $\sigma$ )	
68,3%	<input type="checkbox"/> 1
95,5%	<input type="checkbox"/> 2
99,7%	<input type="checkbox"/> 3

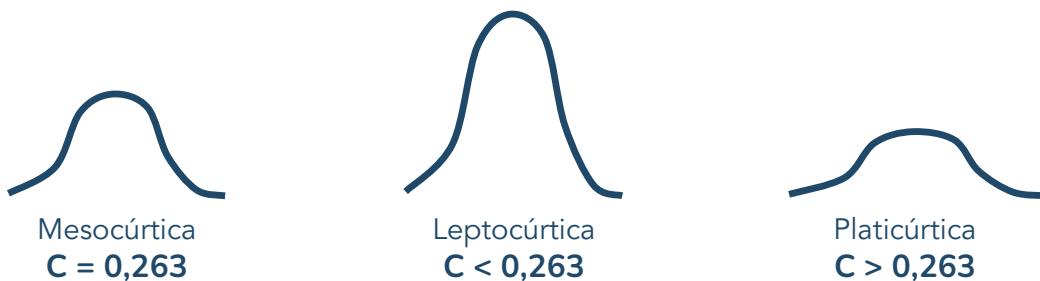


## Curtose - grau de achatamento

**Curtose** é o **grau de achatamento** de uma distribuição em relação a uma **distribuição padrão**, denominada de curva normal.

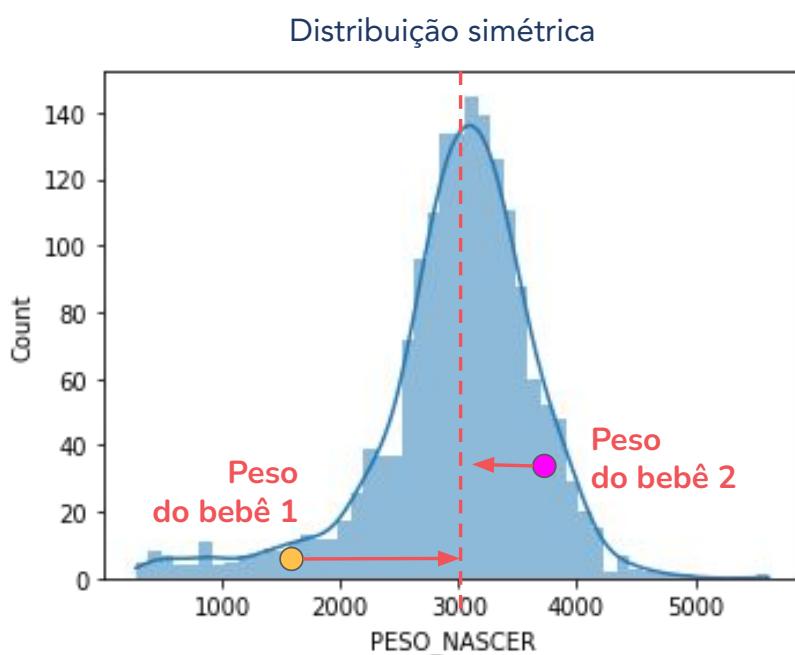
A classificação da distribuição da frequência, relativamente ao seu achatamento, pode ser feita através do cálculo do coeficiente percentílico de curtose:

$$C = \frac{Q_3 - Q_1}{2(P_{90} - P_{10})}$$



A **distribuição normal** possui uma curtose de 0,263.

## Score-Z - Medida de comparação em relação a dispersão



É possível saber o quanto uma medida de um paciente está distante média?

É possível comparar as distâncias de pacientes diferentes?

É possível saber qual paciente está mais próximo ou afastado da média?

## Outras medidas de tendência central e dispersão (Score Z)

O score-z é o quanto uma medida se afasta da média em termos de Desvios Padrão.

Informa o **quão comum ou extremo** um determinado valor é.

Quando o escore Z é positivo isto indica que o dado está acima da média e quando o mesmo é negativo significa que o dado está abaixo da média.

Ao converter os valores para scores-z padronizados em

- média = 0 e desvio padrão = 1

Permite aos pesquisadores **comparar as pontuações** nas escalas com unidades diferentes (o peso em quilogramas, altura vs em polegadas).

	PESO_NASCER	PESO_NASCER_zscore
<b>0</b>	3590.0	0.885759
<b>893</b>	2885.0	-0.134532
<b>1552</b>	2010.0	-1.400851
<b>948</b>	3260.0	0.408176
<b>487</b>	2450.0	-0.764073
<b>1547</b>	1710.0	-1.835017
<b>1038</b>	4020.0	1.508064
<b>393</b>	2240.0	-1.067990
<b>432</b>	2735.0	-0.351615
<b>1272</b>	3525.0	0.791689

Exemplo, qual bebê está mais distante dos valores médios?

- Bebê 1 com 400 gramas acima do peso médio
- Bebê 2 com 2 semanas acima da Idade Gestacional média

Só calculando o Score-Z para as duas medidas para chegar a essa resposta.

## Exemplo

Qual medidas de tendência central e de variabilidade foram utilizadas e porque?

**Table 1.** Baseline characteristics of the pregnancies and newborns.

Characteristics	Values, n	Statistics
<b>Maternal data</b>	702	N/A <sup>a</sup>
Maternal age (years), median (IQR)	702	27 (9)
First antenatal care assessment (weeks), median (IQR)	616	12 (4)
Absent recall of last menstrual period, n (%)	702	89 (12.7)
Reliable last menstrual period, n (%)	613	296 (42.2)
Diabetes, n (%)	701	103 (14.7)
Hypertensive disturbance during pregnancy, n (%)	702	1103 (14.7)
ACTFM <sup>b</sup> , n (%)	698	273 (35.1)
Multiple gestation, n (%)	702	74 (10.5)
<b>Neonatal data</b>	781	N/A
Reference gestational age at birth (weeks), median (IQR)	781	37.3 (6.3)
Gestational age at the first ultrasound assessment (weeks), median (IQR)	781	10.1 (3.6)
Gestational age at the second ultrasound assessment (weeks), median (IQR)	781	19.4 (4.3)
ACTFM exposure, n (%)	777	273 (35.1)
Major malformation, n (%)	781	8 (1.1)
1-min Apgar score, median (IQR)	775	8 (1)
5-min Apgar score, median (IQR)	777	9 (1)
Birth weight (g), median (IQR)	781	2740 (1498)
Sex (male), n (%)	781	390 (49.9)
Incubator accommodation at skin assessment, n (%)	781	239 (30.6)
NICU <sup>c</sup> at skin assessment, n (%)	781	280 (35.9)
Jaundice at skin assessment, n (%)	779	255 (32.7)

Fonte: Newborn Skin Maturity Medical Device Validation for Gestational Age Prediction: Clinical Trial  
J Med Internet Res 2022;24(9):e38727.  
URL: <https://www.jmir.org/2022/9/e38727>  
DOI: 10.2196/38727

Reis Z, Romanelli R, Guimarães R, Gaspar J, Neves G, do Vale M, Nader P, de Moura M, Vitral G, dos Reis M, Pereira M, Marques P, Nader S, Harff A, Beleza L, de Castro M, Souza R, Pappa G, de Aguiar R.

## Gráfico Boxplot - Medidas de resumo de variáveis numéricas

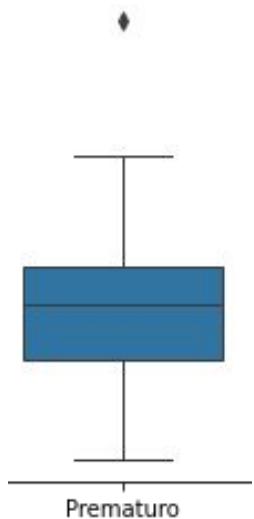
É um gráfico de dispersão unidimensional que exibe o resumo de um conjunto de dados numéricos no formato de caixas.

O boxplot introduzido pelo estatístico americano John Tukey em 1977 é a forma de representar graficamente os dados da distribuição de uma variável quantitativa em função de seus parâmetros.

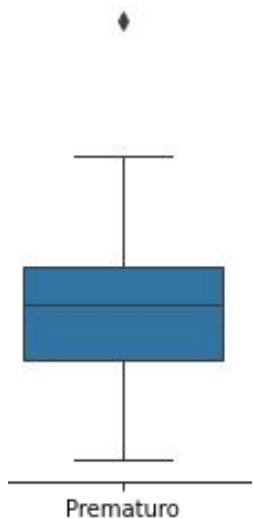
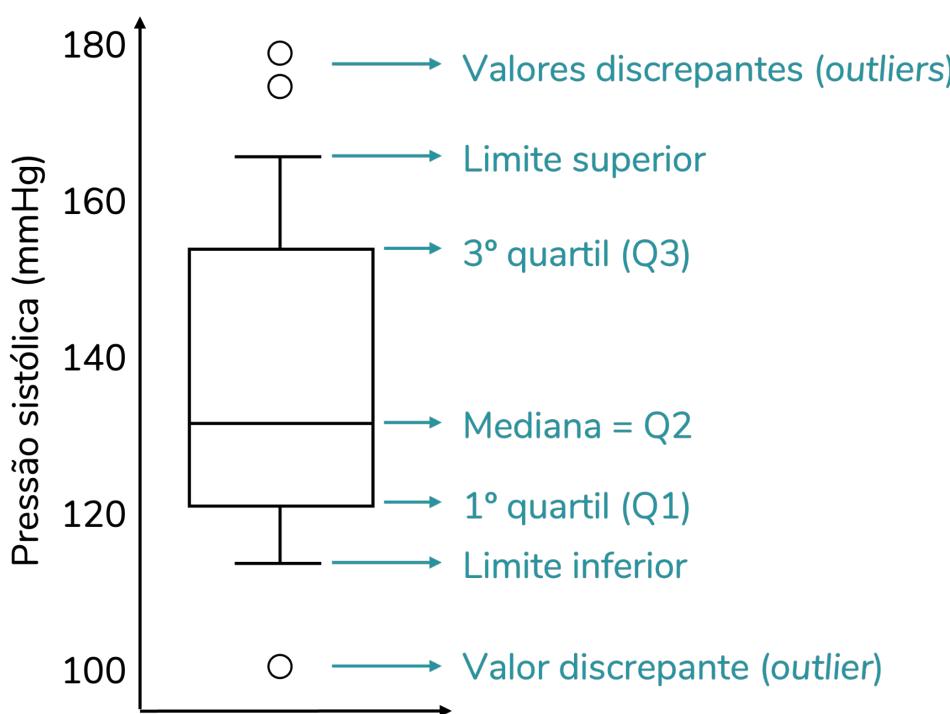
Os cinco itens ou valores: o menor valor ( $x_1$ ), os quartis ( $Q_1$ ,  $Q_2$  e  $Q_3$ ) e o maior valor ( $x_n$ ), são importantes para se ter uma ideia da posição, dispersão e assimetria da distribuição dos dados.

Na sua construção são considerados os quartis e os limites da distribuição, permitindo uma visualização do posicionamento da distribuição na escala da variável.

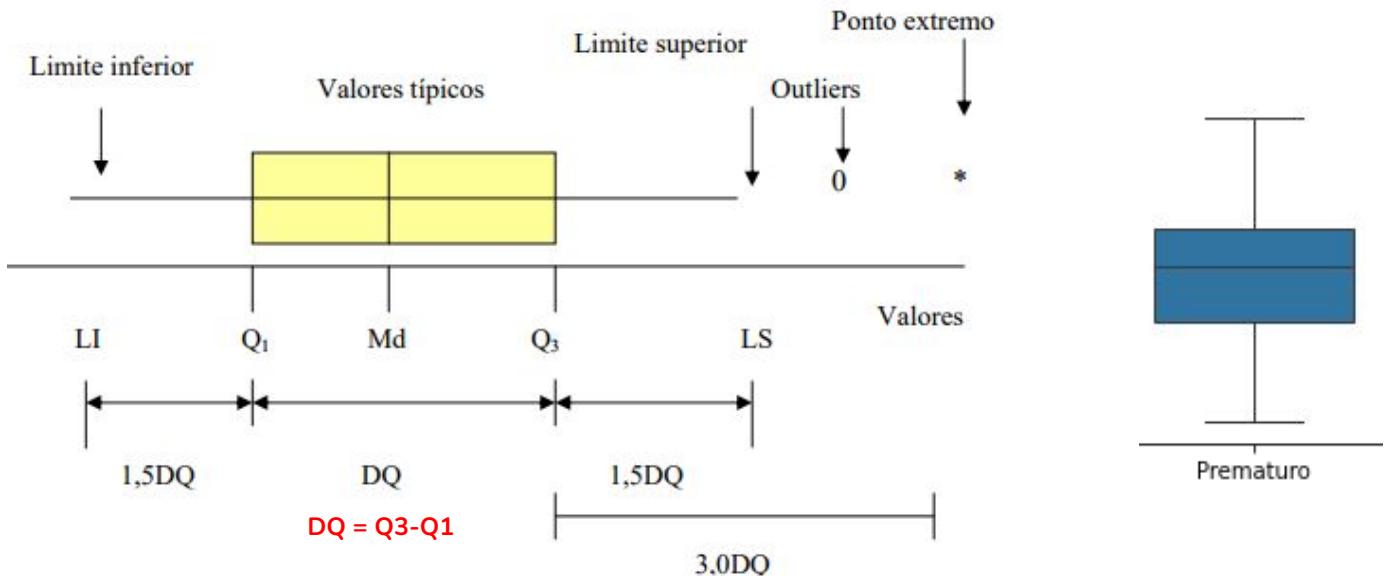
Exibe os quartis, a mediana, os valores extremos e atípicos.



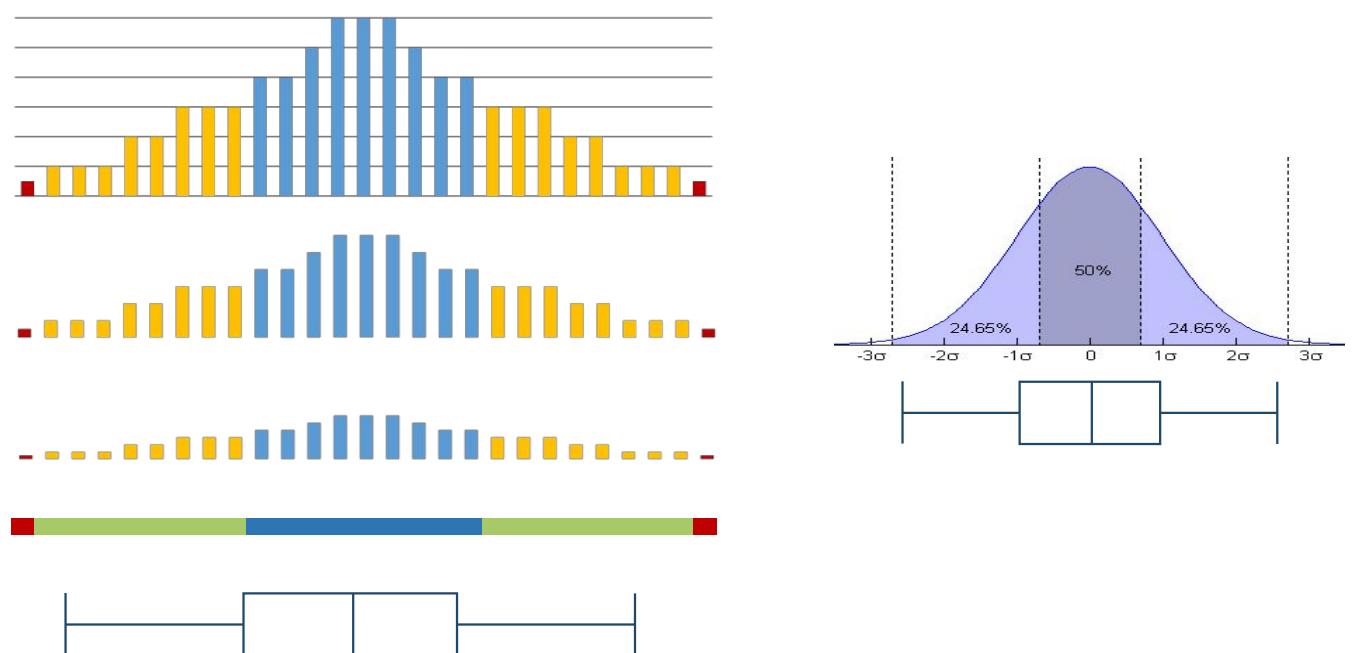
## Gráfico Boxplot - Medidas de resumo de variáveis numéricas



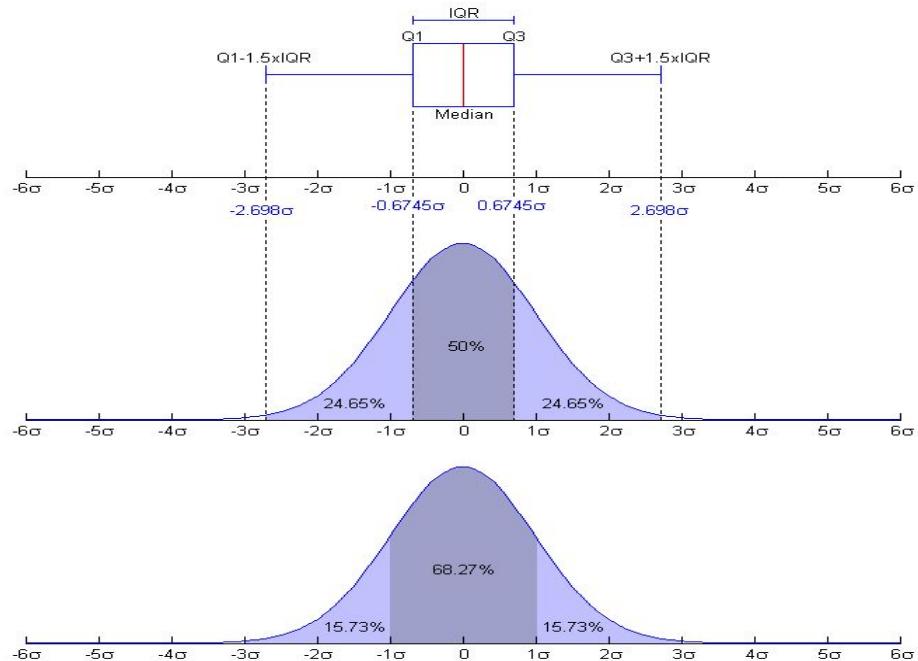
## Gráfico Boxplot - Medidas de resumo de variáveis numéricas



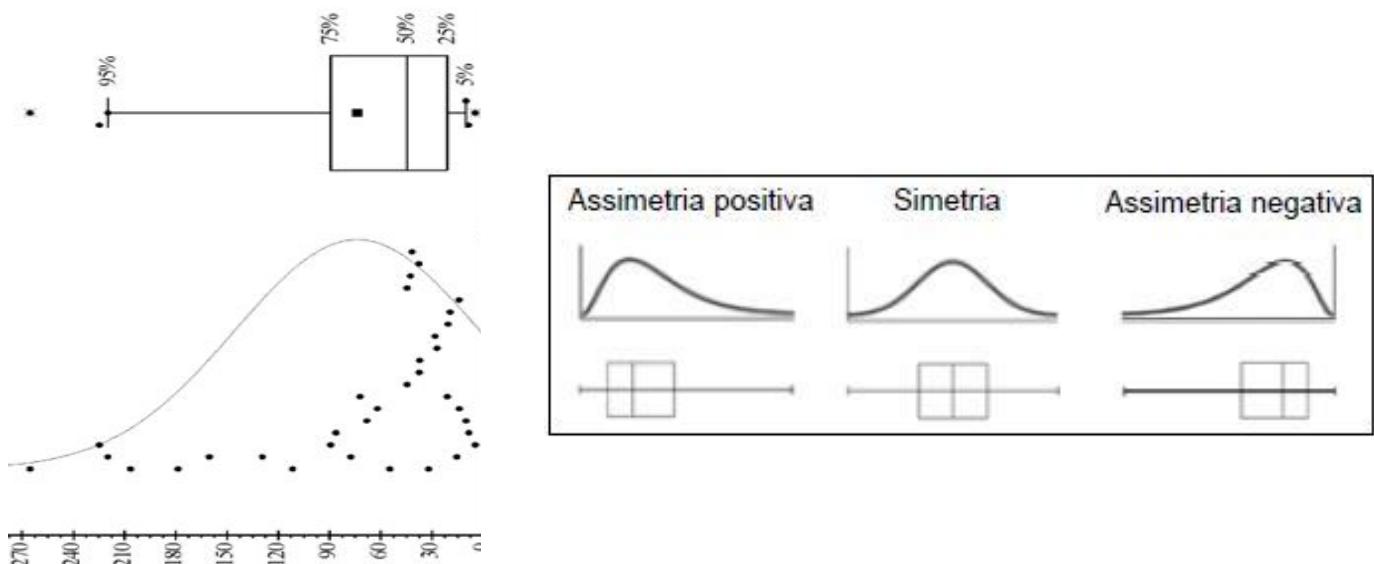
## Gráfico Boxplot - Medidas de resumo de variáveis numéricas



## Gráfico Boxplot - Medidas de resumo de variáveis numéricas



## Gráfico Boxplot - Medidas de resumo de variáveis numéricas



# Testes de hipóteses

Podemos inferir algo sobre a população a partir dos dados amostrais?

29



Introdução à Análise de Dados em Saúde com Python

## Medidas de resumo

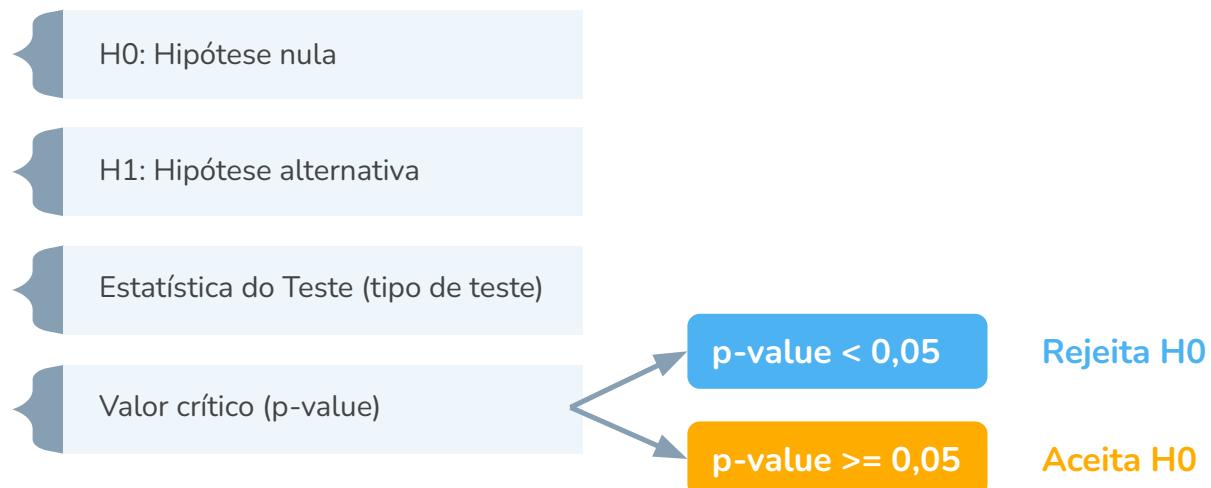
Inferências sobre a população a partir de dados amostrais.



PREMATURO_PED	IG_TERMO	HOUVE_CESAREA	HOUVE_LACERACAO	HOUVE_CM	PARIDADE	PESO_VIAVEIS	PESO_US	PESO_ALTA
Não	Termo-precoce	Não	Não	Sim	Primípara	3590.0	2590.0	3440.0
Sim	Prematuro	Não	Não	Não	Nulípara	2660.0	1660.0	2510.0
Não	Termo	Sim	Sim	Sim	Múltipara	3075.0	2075.0	2925.0
Não	Termo	Não	Sim	Sim	Primípara	3505.0	2755.0	3355.0
Sim	Termo	Não	Não	Sim	Nulípara	3405.0	2405.0	3255.0
Não	Termo	Não	Sim	Sim	Múltipara	3625.0	1625.0	2475.0
Sim	Termo	Não	Não	Sim	Primípara	2975.0	1965.0	2815.0
Não	Termo	Não	Sim	Sim	Primípara	3570.0	2570.0	3420.0
Não	Termo	Não	Não	Sim	Múltipara	2740.0	1740.0	2590.0
Não	Termo	Não	Sim	Não	Nulípara	2945.0	1945.0	2795.0
Não	Termo-precoce	Não	Sim	Sim	Nulípara	3230.0	2230.0	3080.0
Não	Termo	Não	Não	Sim	Primípara	2980.0	1980.0	2830.0
Não	Termo	Não	Sim	Sim	Primípara	3120.0	2120.0	2970.0
Não	Termo	Não	Não	Não	Nulípara	3055.0	2055.0	2905.0
Não	Termo-precoce	Não	Não	Não	Múltipara	3550.0	2580.0	3430.0
Sim	Prematuro	Não	Não	Não	Nulípara	2715.0	850.5	2065.0
Não	Termo-precoce	Não	Não	Não	Nulípara	NaN	NaN	NaN
Não	Termo-precoce	Não	Não	Sim	Múltipara	3130.0	2130.0	2980.0
Sim	Prematuro	Não	Não	Não	Múltipara	2325.0	1325.0	2175.0
Não	Termo	Não	Não	Não	Nulípara	3050.0	2050.0	2900.0
Não	Termo	Não	Não	Não	Primípara	2895.0	1895.0	2745.0
Não	NaN	Sim	Sim	Sim	Primípara	3190.0	2190.0	3040.0
Não	Prematuro	Não	Não	Não	Nulípara	2870.0	1870.0	2720.0
Não	NaN	Sim	Sim	Sim	Nulípara	2720.0	1720.0	2570.0
Não	NaN	Sim	Sim	Sim	Primípara	3290.0	2290.0	3140.0
Não	NaN	Sim	Sim	Sim	Nulípara	3160.0	2160.0	3010.0
Não	NaN	Sim	Sim	Sim	Primípara	3220.0	2220.0	3070.0
Não	NaN	Sim	Sim	Sim	Nulípara	3060.0	2060.0	2910.0

## Testes de Hipóteses

Para formular um teste de hipóteses é preciso identificar:



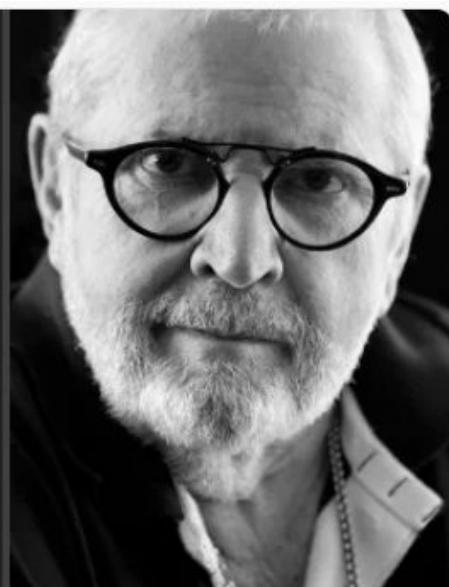
## Inferência Estatística - Testes de Hipóteses

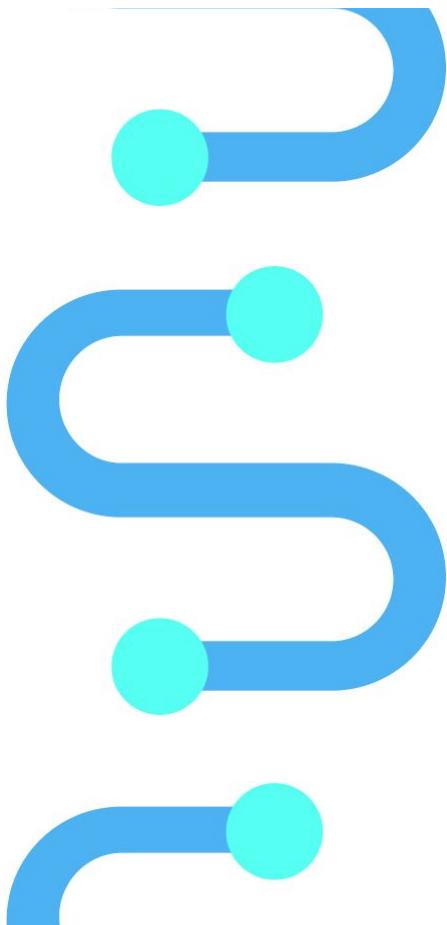
Decisões a partir de hipóteses previamente formuladas.

Hipótese é uma coisa que não é, mas a gente faz de conta que é, para ver como seria se ela fosse.

Jô Soares

“ PENSADOR





## Testes de hipóteses

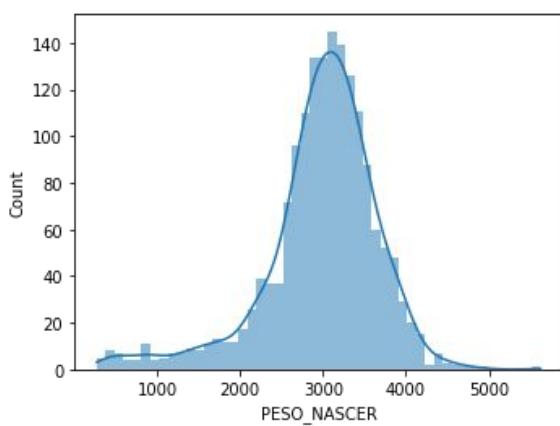
## Testes de normalidade



Introdução à Análise de Dados em Saúde com Python

### Distribuição Normal - Testes de normalidade

Como saber se um distribuição é normal?



#### Características da distribuição normal:

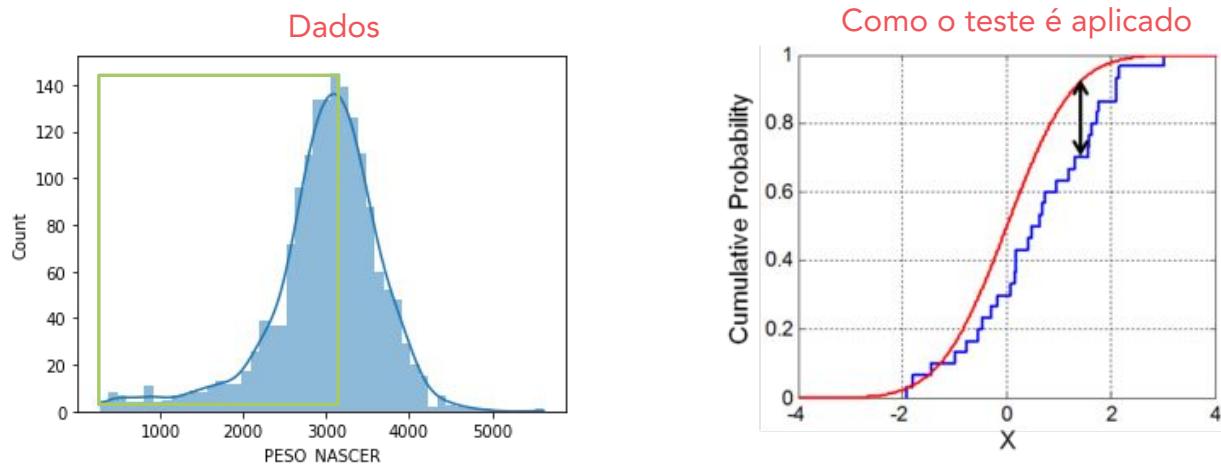
- A curva é simétrica em torno da média
- A média, mediana e a moda coincidem
- As extremidades podem se estender infinitamente
- Coeficiente de assimetria e curtose, padronizados pelo seu erro padrão estão entre -1,96 e + 1,96
- Aplicar testes de normalidade:
  - D'Agostino and Pearson's
  - Kolmogorov-Smirnov
  - Shapiro-Wilk

## Distribuição Normal - Testes de normalidade

Qual a lógica dos testes de hipótese que testam normalidade?

Considerando que:

- os dados tem sua distribuição específica (**linha azul**)
- a distribuição normal tem uma linha que pode ser calculada algebraicamente (**linha vermelha**)
- os testes de normalidade vão testar a probabilidade destas linhas (**azuis e vermelhas**) serem estatisticamente semelhantes.



## Testes de normalidade

Existem dezenas de testes de normalidade, apresentamos aqui alguns:

### Características dos testes de normalidade:

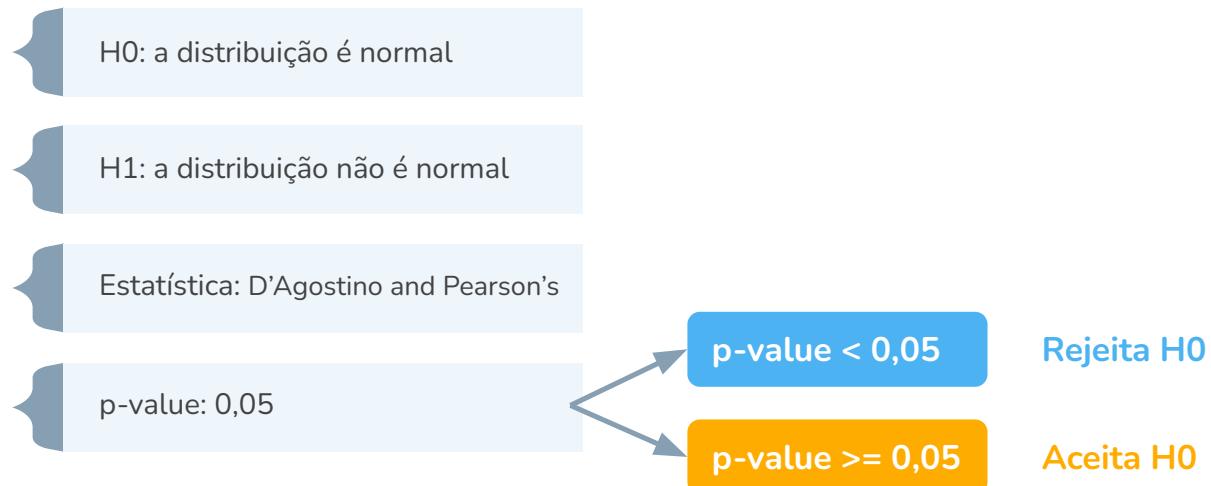
- D'Agostino and Pearson's (Normal Test)
  - Foi desenvolvido para lidar com amostras mais numerosas ( $n > 100$ )
  - Apresenta desempenho próximo ao do Shapiro-Wilk
- Shapiro-Wilk
  - Recomendado para amostras pequenas ( $< 30$  casos)
  - Não pode ser usado por amostras com valores nulos (zero)
- Kolmogorov-Smirnov
  - Para amostra com mais de 30 casos
  - Geralmente apresenta desempenho abaixo de outros testes
  - Recomenda-se o Kolmogorov-Smirnov com a correção de Lilliefors

Leitura recomendada:

Artigo de Hélio Miot, 2017. Avaliação da normalidade dos dados em estudos clínicos e experimentais.  
Fonte: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5915855/>

## Teste de Hipótese para avaliar se uma distribuição é normal

Teste de normalidade



## Estatística Inferencial Variáveis categóricas

## Relacionando variáveis - Tabela de Contingência ou Tabela Cruzada

Variáveis categóricas

```
✓ [48] 1 dados.SEXO.value_counts()
```

```
0s
Masculino    853
Feminino     797
Name: SEXO, dtype: int64
```

```
✓ [49] 1 dados.TIPO_PARTO.value_counts()
```

```
0s
Parto Normal    1056
Parto Cesáreo   652
Name: TIPO_PARTO, dtype: int64
```

Sexo do RN	Tipo de parto	
	Parto Cesáreo	Parto Normal
Feminino	274	523
Masculino	356	497

TIPO_PARTO	Parto Cesáreo	Parto Normal
SEXO		
Feminino	274	523
Masculino	356	497

## Relacionando variáveis - Tabelas de Contingência

As tabelas de contingência são utilizadas para registrar a ocorrência de duas variáveis aleatórias, simultaneamente.

Geralmente são usada para se estudar a relação entre duas variáveis categóricas descrevendo a frequências das categorias de uma das variáveis relativamente às categorias da outra.

Variável de desfecho

Variável Preditora	Presente	Ausente	Total
	Presente	B	A+B
Ausente	C	D	C+D
Total	A+C	B+D	A+B+C+D

## Tabelas de Contingência

Variável de desfecho  
**CM\_NEARMISS**

	Sim	Não	Total
Sim	A = 0	B = 87	A+B
Não	C = 17	D = 1604	C+D
Total	A+C	B+D	A+B+C+D

	CM_NEARMISS	Não	Sim
FORCEPS	1604	17	0
Não			
Sim			

## Medidas de Associação

### Hipótese da independência

Refere-se a um conjunto de testes estatísticos que busca avaliar a **existência de associação** entre dados aleatórios apresentados em categorias.

Os testes compararam as frequências **observadas** com as **esperadas**, em cada uma das caselas de contingência.

A probabilidade da associação em relação ao acaso é avaliada pela sua significância estatística.

A distribuição de probabilidade neste caso é a **qui-quadrado**.

**Qui-quadrado de Pearson**

**Teste exato de Fisher**

**Qui-quadrado de McNemar**

## Testes para Hipótese de Independência

### Qui-quadrado de Pearson

Observações independentes (variáveis diferentes)  
 Participante só entra uma vez  
 N fatores exposição x N fatores desfechos  
**Frequência esperada >5 (80%), qualquer freq. obs. >1**

**Sexo**  
 X  
**Tipo de parto**

### Teste exato de Fisher

Observações independentes (variáveis diferentes)  
 Participante só entra uma vez  
 1 fator exposição x 1 fator desfecho  
 Frequência esperada <5 , freq. obs. permite zero

**Gemelar**  
 X  
**Nearmiss**

### Qui-quadrado de McNemar

Observações dependentes (a mesma natureza da variável)  
 Participante só entra uma vez  
 Fator exposição (antes) x fator desfecho (depois)  
 Categorias devem ser identicas (mesmos valores)

**Baixo Apgar 1º min**  
 X  
**Baixo Apgar 5º min**

## Testes para Hipótese de Independência

Para formular um teste de hipóteses é preciso identificar:

{ H<sub>0</sub>: Não existe associação entre as variáveis

{ H<sub>1</sub>: Existe associação entre as variáveis

{ Teste: Qui-quadrado, Exato de Fisher, McNemar

{ Valor crítico (p-value)

p-value < 0,05

Rejeita H<sub>0</sub>

p-value >= 0,05

Aceita H<sub>0</sub>

## Testes para Hipótese de Independência

### Teste qui-quadrado de Pearson

Exemplo 1: Existe associação entre risco gestacional e tipo de parto?  
H0: o tipo de parto não é influenciado pelo risco gestacional

Teste: qui-quadrado  
p-value < 0,05, rejeita H0

Exemplo 2: Existe associação entre via de parto e infecção puerperal?  
H0: a infecção puerperal não é influenciada via de parto

Teste: qui-quadrado  
p-value < 0,05, rejeita H0

Exemplo 3: Existe associação entre baixo peso ao nascer e baixo apgar 5º minutos?  
H0: o apgar de 5º minuto não é influenciado pelo baixo peso ao nascer

Teste: qui-quadrado  
p-value < 0,05, rejeita H0

### Teste Exato de Fisher

Exemplo 1: Existe associação entre risco gestacional e tipo de parto?  
H0: o tipo de parto não é influenciado pelo risco gestacional

Teste: Exato de Fisher  
p-value < 0,05, rejeita H0

Exemplo 2: Existe associação entre via de parto e infecção puerperal?  
H0: a infecção puerperal não é influenciada via de parto

Teste: Exato de Fisher  
p-value < 0,05, rejeita H0

### Teste qui-quadrado de McNemar

Exemplo 1: Existe associação entre fazer uma cesárea (atual) com o histórico de cesáreas prévias da gestante?  
H0: não existe associação entre parto cesáreo prévio e a cesárea atual

Teste: McNemar  
p-value < 0,05, rejeita H0

## Fatores de risco para óbito em unidade de terapia intensiva neonatal

**Tabela 2 - Proporção de óbitos e altas para as variáveis peso ao nascer, Apgar 1º minuto e 5º minuto, doença da membrana hialina, sepse e malformação congênita e idade gestacional, para os recém-nascidos internados na unidade de terapia intensiva neonatal**

Variável (N)*	Alta	Óbito	Total	Valor de p
Peso (491)				<0,001
≥1500	300 (86,7)	46 (13,3)	346 (100,0)	
Muito baixo peso	62 (42,7)	83 (57,3)	145 (100,0)	
Apgar 1min (478)				<0,001
Boa vitalidade	247 (83,7)	48 (16,3)	295 (100,0)	
Má vitalidade	107 (58,5)	76 (41,5)	183 (100,0)	
Apgar 5min (480)				<0,001
Boa vitalidade	319 (78,4)	88 (21,6)	407 (100,0)	
Má vitalidade	36 (49,3)	37 (50,7)	73 (100,0)	
DMH (495)				<0,001
Não	320 (80,4)	78 (19,6)	398 (100,0)	
Sim	44 (46,8)	50 (53,2)	94 (100,0)	
Sepse (491)				0,065
Não	198 (77,7)	57 (22,3)	255 (100,0)	
Sim	166 (70,4)	70 (29,6)	236 (100,0)	
Malformação (488)				0,002
Não	318 (76,3)	99 (23,7)	417 (100,0)	
Sim	42 (59,2)	29 (40,8)	71 (100,0)	
Gestação (492)				<0,001
Termo	127 (87,6)	18 (12,4)	145 (100,0)	
Pré-termo	236 (68,0)	111 (32,0)	347 (100,0)	
Gestação (492)				<0,001
Pré-termo	295 (85,0)	57 (15,0)	348 (100,0)	
Muito pré-termo	67 (46,5)	77 (53,5)	144 (100,0)	

DMH - doença da membrana hialina. \*N=total de recém-nascidos em cada variável de estudo. Valores expressos em número (%).

Fonte: RISSO, Susana de Paula; NASCIMENTO, Luiz Fernando C.. Fatores de risco para óbito em unidade de terapia intensiva neonatal, utilizando a técnica de análise de sobrevida. *Rev. bras. ter. intensiva*, São Paulo , v. 22, n. 1, p. 19-26, Mar. 2010 .

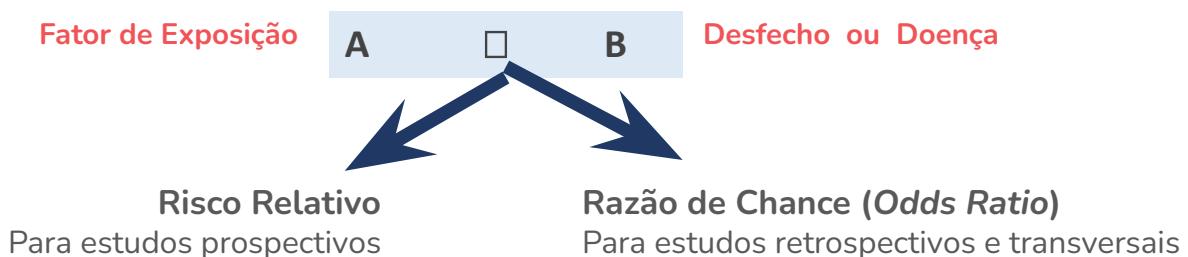
## Força de Associação

Magnitude ou força da associação

No entanto, para determinar a magnitude da associação entre a exposição (ou intervenção) e os desfechos (resultados) de interesse utilizamos medidas relativas e absolutas que quantificam esta associação.

O **Risco Relativo** e a **Razão de Chances** são as mais frequentes.

E quando quisermos saber o quanto a variável A influência na variável B ?



## Fatores de exposição e Doença

Evento A

### Fatores de exposição

- Idade
- Sexo
- Hábitos
- Vícios
- Classe social
- Ocupação
- Estado nutricional
- Uso de uma droga
- Realização de uma atividade
- ...



Evento B

### Doenças / desfechos

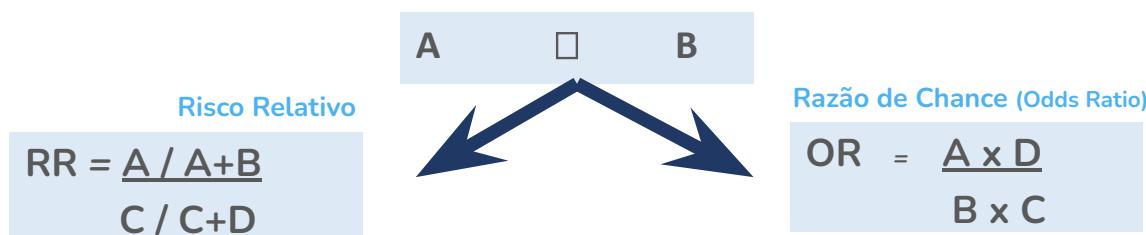
- Diabetes
- AVC
- Cancer
- Gastrite
- Pré-eclâmpsia
- Alívio da dor
- Redução do tempo de internamento
- Morte
- Acidentes
- ...



## Tabela de contigência 2 x 2

A partir da tabela de contigência conseguimos de forma mais fácil aplicar as fórmulas de Risco Relativo e Razão de Chance.

		Variável de desfecho		
		Presente	Ausente	Total
Variável Preditora	Presente	A	B	A+B
	Ausente	C	D	C+D
	Total	A+C	B+D	A+B+C+D



## Interpretação do RR e do OR

- O RR ou OR, normalmente é maior que 1, pois supostamente ele calcula o valor que a exposição de um indivíduo a um fator de risco pode aumentar a frequência da exposição.
- Quando o valor do risco relativo for menor que 1, nós chamamos de fator de prevenção / proteção.
  - Exemplo quando calculamos o RR para saber do impacto da toma de cálcio diário por mulheres com mais de 40 anos de idade.
- Quando o RR for igual (próximo) a 1, não podemos dizer que o fator põe em risco ou previne o indivíduo de um determinado evento. Ou seja, não se comprova os efeitos.

<b>RR ou OR &gt; 1</b>	<b>Fator de risco</b>
<b>RR ou OR &lt; 1</b>	<b>Fator de proteção</b>
<b>RR ou OR = 1</b>	<b>Sem impacto de risco ou proteção</b>

## Interpretação do Risco Relativo e do Odds Ratio



**RR < 1**

- FATOR DE PROTEÇÃO



**RR = 1**

- SEM EFEITO



**RR > 1**

- FATOR DE RISCO

**RR = 0,73**

Então:  $1 - 0,73 = 0,27$

Proteção de 27% menos risco

**RR = 1,32**

Então =  $1,32 - 1 = 0,32$

Risco de 32% a mais

Grupo a tem 1,32 vezes o risco do grupo B

Pode inferir relação de causa-efeito



**OR < 1**

- FATOR DE PROTEÇÃO



**OR = 1**

- SEM EFEITO



**OR > 1**

- FATOR DE RISCO

**OR = 0,73**

Então:  $1 - 0,73 = 0,27$

Proteção de 27% menos chance

**OR = 1,32**

Então =  $1,32 - 1 = 0,32$

Chance de 32% a mais

Grupo a tem 1,32 vezes a chance do grupo B

Não avalia relação de causa-efeito  
É uma aproximação do risco relativo

## Exemplos

Tabela II - Variáveis associadas independentemente com IAM\*

Variável independente	OR	IC 95%	p
Tabagismo ( $\geq 5$ cigarros/dia vs. nunca)	5,86	3,20-10,57	< 0,00001
Relação cintura-quadril (tercil 1 vs. tercil 3)	4,27	2,28-8,00	< 0,00001
Antecedentes de hipertensão arterial (presente vs. ausente)	3,26	1,95-5,46	< 0,00001
Relação cintura-quadril (tercil 1 vs. tercil 2)	3,07	1,66-5,66	0,0003
LDL-colesterol ( $< 100$ mg/dl vs. $\geq 100$ mg/dl)	2,75	1,45-5,19	0,0018
Antecedentes de diabetes (presente vs. ausente)	2,51	1,13-5,56	0,0230
História familiar de insuficiência coronariana (presente vs. ausente)	2,33	1,44-3,75	0,0005
HDL-colesterol ( $\geq 40$ mg/dl vs. $< 40$ mg/dl)	0,53	0,32-0,87	0,011

OR - "Odds ratio" (razão de chances); IC - Intervalo de confiança de 95%; \* Análise multivariada por meio de regressão logística não-condicional.

Arquivos Brasileiros de Cardiologia - Volume 84, Nº 3, Março 2005

- 1) Qual a natureza das variáveis independentes (preditoras) neste contexto?
- 2) Qual a variável de desfecho?
- 3) Qual a medida de efeito (força de associação) utilizada?
- 4) Que teste de hipótese foi utilizado?
- 5) Quais são as variáveis que representam fatores que elevam a chance de infarto e quais são as que reduzem (proteção)?

## Intervalo de Confiança do RR / OR

Variável independente	OR	IC 95%	p
Tabagismo ( $\geq 5$ cigarros/dia vs. nunca)	5,86	3,20-10,57	< 0,00001
Relação cintura-quadril (tercil 1 vs. tercil 3)	4,27	2,28-8,00	< 0,00001
Antecedentes de hipertensão arterial (presente vs. ausente)	3,26	1,95-5,46	< 0,00001
Relação cintura-quadril (tercil 1 vs. tercil 2)	3,07	1,66-5,66	0,0003
LDL-colesterol ( $< 100$ mg/dl vs. $\geq 100$ mg/dl)	2,75	1,45-5,19	0,0018
Antecedentes de diabetes (presente vs. ausente)	2,51	1,13-5,56	0,0230
História familiar de insuficiência coronariana (presente vs. ausente)	2,33	1,44-3,75	0,0005
HDL-colesterol ( $\geq 40$ mg/dl vs. $< 40$ mg/dl)	0,53	0,32-0,87	0,011

OR - "Odds ratio" (razão de chances); IC - Intervalo de confiança de 95%; \* Análise multivariada por meio de regressão logística não-condicional.

Arquivos Brasileiros de Cardiologia - Volume 84, Nº 3, Março 2005

IC: É uma faixa de valores usada para se estimar o verdadeiro valor de um parâmetro populacional (nível de confiança 95%).

É proporcional ao tamanho da amostra e frequência do parâmetro na população

$$\text{IC 95\% do } \ln(\text{RR}) = \ln(\text{RR}) \pm 1,96 \sqrt{\frac{b/a}{a+b} + \frac{d/c}{c+d}}$$

## Fatores para ventilação mecânica em lactentes com doença respiratória aguda baixa

Tabela I – Associação entre variáveis epidemiológicas e clínicas e evolução para ventilação mecânica em 152 lactentes com doença respiratória aguda baixa

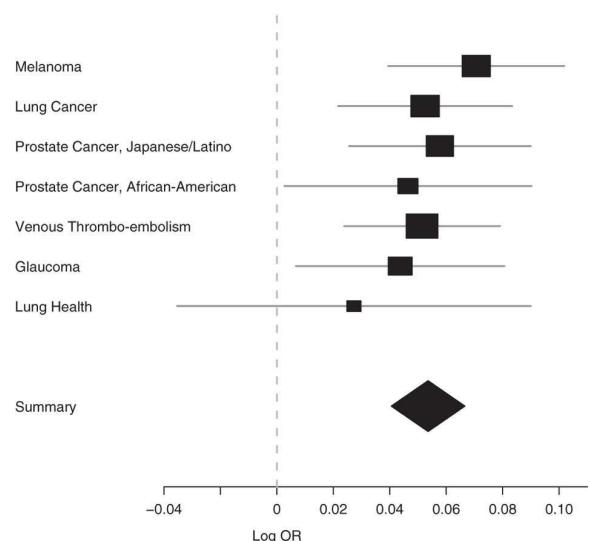
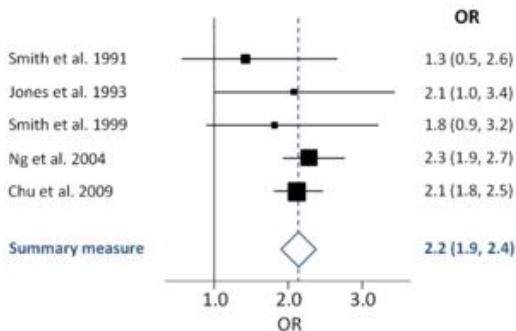
Variável	Grupo ventilação mecânica (presença/ausência)	Grupo sem ventilação mecânica (presença/ausência)	Risco relativo (IC 95%)*
Idade < 3 meses	9/12	43/88	2,35 (1,06-5,22)
Aleitamento materno < 1 mês	14/7	45/86	3,15 (1,35-7,35)
Cianose	1/20	0/131	7,55 (5,01-11,36)
Internação > 10 dias	17/4	19/112	13,69 (4,92-38,09)
Oxigenoterapia > 10 dias	16/5	13/118	13,57 (5,41-34,03)
Uso de antimicrobianos	13/8	40/91	3,03 (1,34-6,86)
Reinternação	2/19	1/130	5,23 (2,12-12,91)
Sexo masculino	14/7	79/52	1,26 (0,54-2,96)
Etnia caucasóide	16/5	81/50	1,81 (0,70-4,68)
Peso de nascimento < 2500 g	4/16	17/114	1,55 (0,57-4,18)
Idade gestacional < 35 semanas ao nascer	3/17	9/122	2,04 (0,69-5,99)
Doenças prévias	7/14	33/98	1,40 (0,61-3,22)
Síbilo anterior	3/18	40/91	0,42 (0,13-1,36)
Tabagismo familiar	4/17	54/77	0,38 (0,12-1,08)
Mais de 5 moradores no domicílio	13/8	67/64	1,46 (0,64-3,32)
Idade materna < 20 anos	3/17	38/93	0,47 (0,15-1,32)
Escolaridade materna < 5 anos	3/17	32/97	0,57 (0,18-1,85)
Taquicardia	12/9	46/85	2,16 (0,97-4,81)
Saturação de O <sub>2</sub> < 90%	8/13	25/106	2,22 (1,00-4,90)
Corticoterapia	5/16	17/114	1,85 (0,75-4,53)

\* Associações significativas destacadas em negrito

RICCETTO, Adriana Gut Lopes et al . Fatores prognósticos para ventilação mecânica em lactentes com doença respiratória aguda baixa. Rev. Assoc. Med. Bras., São Paulo , v. 52, n. 5, p. 342-346, Oct. 2006 .

## Intervalo de Confiança do RR / OR

### Gráfico de folhas (Forest Plot)



Cochrane  
Brazil

<http://www.cochrane.org/> <http://brazil.cochrane.org/>

Estatística Inferencial  
Variáveis numéricas x categóricas  
Testes de Médias e de Medianas

## Medidas de resumo

### Distribuição

PESO\_NASCER

```
count    1680.0
mean     2977.9
std      691.1
min      270.0
25%     2710.0
50%     3065.0
75%     3390.0
max     5625.0
```



Distribuição Simétrica  
Ou tende p/ Curva Normal

### Tendência Central

Média

**2977,9**

### Dispersão

Desvio padrão

**691,1**

DURACAO\_INT

```
count    1708.0
mean     3.0
std      4.3
min      1.0
25%     2.0
50%     2.0
75%     3.0
max     63.0
```



Distribuição Assimétrica

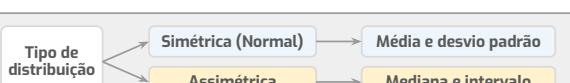
Mediana

**2**

(Mínimo - Máximo)  
Amplitude / âmbito

**1 - 63  
62**

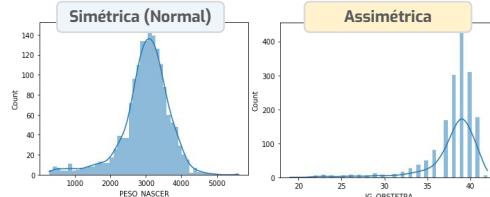
## Testes de médias e medianas



Interpretar o teste

p-value < 0,05

Rejeita H0 - hipótese nula



1 variável categórica como fator de exposição

Testes de médias e de medianas  
Variável numérica X categórica

2 categorias

>2 categorias

Variáveis independentes

Variáveis dependentes (antes e depois)

H0: Não existe relação, as médias/medianas são iguais

H0: Não existe relação, as médias/medianas são iguais

Em qual grupo está a diferença?

Distribuição normal

Outras distribuições

Levene  
H0: as variâncias são iguais

Teste t (var. = ou ≠)

Mann-Whitney

Teste t pareado

Wilcoxon

ANOVA unifatorial

Kruskal-Wallis

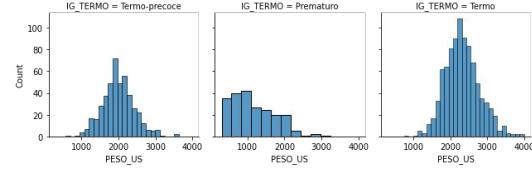
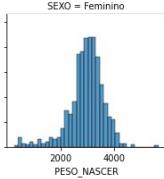
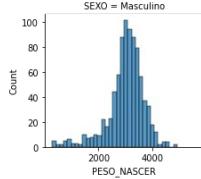
ANOVA medidas repetidas

Friedman

Tukey-HSD

Dunn

Dunn



# Estatística Inferencial

## Variáveis numéricas x numéricas

### Testes de Correlações



Introdução à Análise de Dados em Saúde com Python

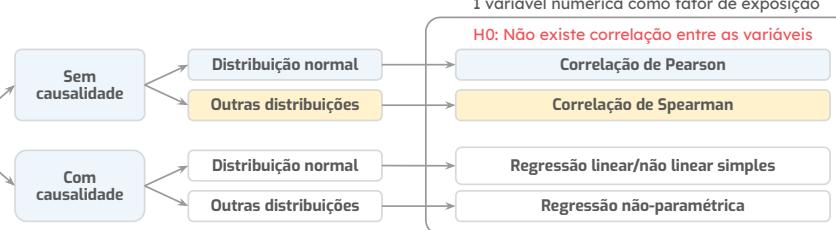
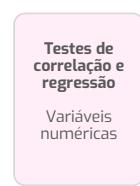
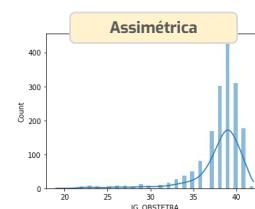
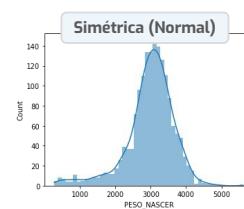
## Correlações - Variáveis numéricas



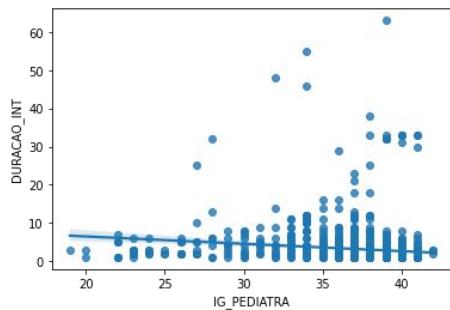
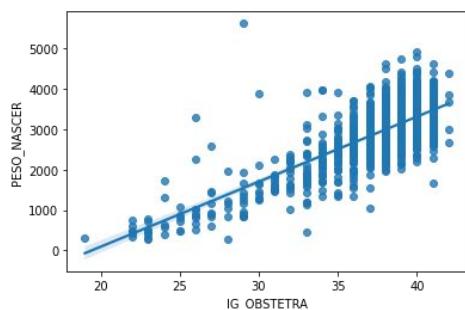
Interpretar o teste

p-value < 0,05

Rejeita H0 - hipótese nula

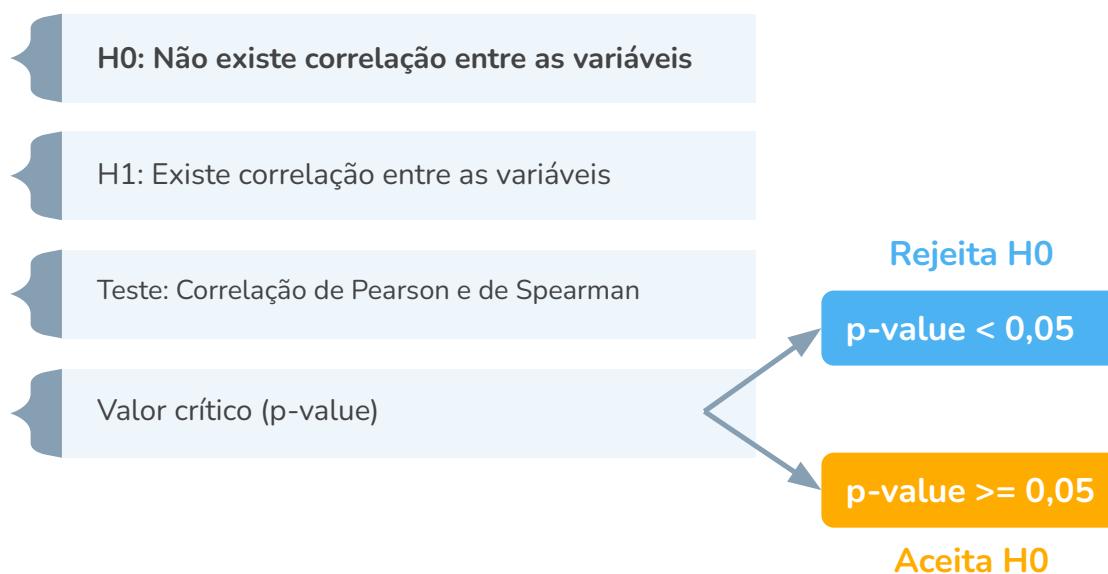


## Correlações - Variáveis numéricas



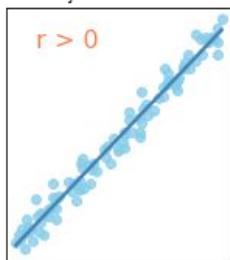
## Testes para Hipótese de Independência

Para formular um teste de hipóteses é preciso identificar:



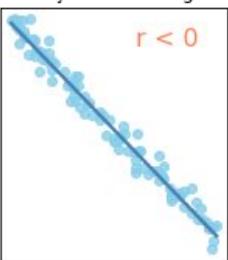
## Correlações - Variáveis numéricas

Correlação Linear Positiva



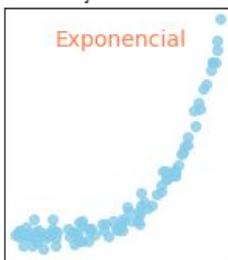
Correlação Linear Positiva

Correlação Linear Negativa



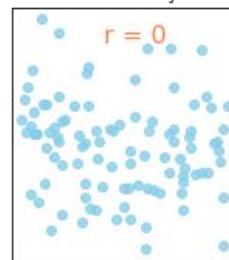
Correlação Linear Negativa

Correlação Não-Linear



Correlação Não-Linear

Sem Correlação



Sem Correlação

**H0:** A concordância foi ao acaso, ou seja,  $r = 0$ Valores possíveis para  $r$ : -1 até +1

<b>0</b>	<b>ausência de correlação</b>
<b>&gt; 0,70</b>	<b>forte correlação</b>
<b>0,30 a 0,7</b>	<b>Correlação moderada</b>
<b>0 a 0,30</b>	<b>Fraca correlação.</b>

## Correlações - Variáveis numéricas

### Teste Correlação de Pearson

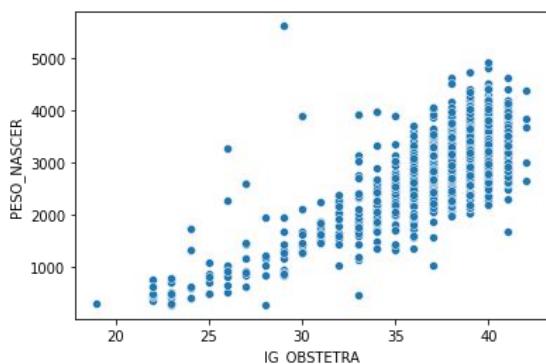
Existe correlação entre a idade gestacional e o peso do recém nascido?

**H0:** Não existe correlação entre IG e o Peso

Significância: 0,05

p-value &lt; 0,05, rejeita a Hipótese Nula (H0)

Se existe correlação, de quanto é essa correlação?



Teste: Correlação de Pearson  
p-value < 0,05, rejeita H0

p-value < 0,001  
 $r = 0,745$  (correlação forte)

## Correlações - Variáveis numéricas

### Teste Correlação de Pearson

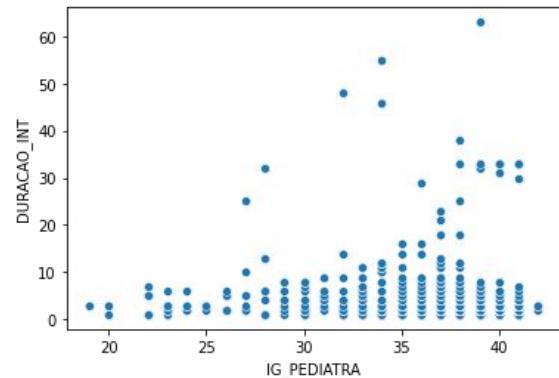
Existe correlação entre a idade gestacional e a duração na internação?

H0: Não existe correlação entre IG e a duração na internação

Significância: 0,05

p-value < 0,05, rejeita a Hipótese Nula (H0)

Se existe correlação, de quanto é essa correlação?



Teste: Correlação de Spearman  
p-value < 0,05, rejeita H0

p-value < 0,001  
r = -0,159 (correlação fraca)

**Estatística Inferencial**  
**Teste de concordâncias**  
**Kappa de Cohen**

## Testes Kappa

Existe concordância entre os avaliadores? Ou foi ao acaso?

H0: A concordância foi ao acaso.

Significância: 0,05

p-value < 0,05, rejeita a Hipótese Nula (H0)

Se existe concordância, de quanto é essa essa concordância?

	Observação 1	
	Teste +	Teste -
Observação 2	Teste +	a      b
	Teste -	c      d
Observações concordantes		



## Exemplo

		Juiz 1 - Avaliador 1		
		Leve	Moderada	Grave
Juiz 2 Avaliador 2	Leve	14	0	0
	Moderada	0	12	0
	Grave	1	0	13

Concordância simples:  
 $(14+12+13) / 40 = 97,5\%$

Teste: Kappa de Cohen  
p-value < 0,05, rejeita H0

## Testes Kappa

Avaliar a concordância entre o diagnóstico de prematuridade feito pelo obstetra, com o do pediatra.



H0: A concordância foi ao acaso.

Significância: 0,05

p-value < 0,05, rejeita a Hipótese Nula (H0)

Se existe concordância, de quanto é essa essa concordância?

### Prematuro Pediatra

	Não	Sim
Não	1387	4
Sim	23	268

Teste: Kappa de Cohen  
p-value < 0,05, rejeita H0

p-value < 0,001  
kappa = 0,942  
(concordância quase perfeita)

## Interpretação do coeficiente Kappa ou Correlações

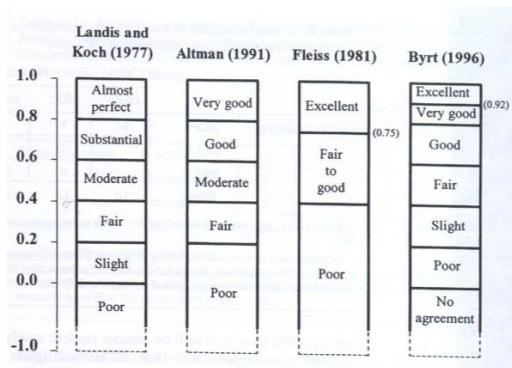
### Interpretation of Kappa

	Poor	Slight	Fair	Moderate	Substantial	Almost perfect
Kappa	0.0	.20	.40	.60	.80	1.0

Kappa	Agreement
< 0	Less than chance agreement
0.01–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–0.99	Almost perfect agreement

Viera, Anthony J., and Joanne M. Garrett. "Understanding interobserver agreement: the kappa statistic." *Fam Med* 37, no. 5 (2005): 360-363.  
<https://www.ncbi.nlm.nih.gov/pubmed/15883903>:

## Interpretação do coeficiente Kappa ou Correlações



**Quadro 2** Categorização da concordância intra-observador e interobservador pelo índice kappa e coeficiente de correlação intraclasses.

Índice kappa ( $\kappa$ ) ou coeficiente de correlação intraclasses ( $r$ )	Grau de concordância
1,00	Perfeito
0,81 – 0,99	Quase perfeito
0,61 – 0,80	Substancial
0,41 – 0,60	Moderado
0,21 – 0,40	Mediano
0,00 – 0,20	Insignificante

# Dados populacionais e amostrais



Introdução à Análise de Dados em Saúde com Python

## Tipos de dados quanto a sua origem: população e amostra



**Dados populacionais**

Coleção completa de todos os elementos (sujeitos) a serem estudados.

Inferências sobre a população a partir de dados amostrais.



**Dados amostrais**

Subconjunto de membros selecionados de uma população.

## Tipos de dados quanto a sua origem: população e amostra

Quanto **maior o tamanho da amostra**, maior será o **poder de generalização** das conclusões que obtivermos ao analisá-la.

Isso acontece, porque um número maior de indivíduos permite se aproximar mais fielmente do tamanho real da população.

O **desvio padrão** é um estimador para o **verdadeiro desvio padrão da população**, que é um tipo de parâmetro estatístico.

### Dicas:

Ao fazer um estudo estatístico, tente:

1º **identificar a população** a qual se desejar inferir algo.

2º definir como selecionar uma amostra (ou mais de uma) que permita fazer inferências adequadas a respeito da população em questão (**maior ALEATORIEDADE possível**).

3º tenha em mente que o que você quer é tirar conclusões que **se apliquem à uma população** a partir de uma amostra.

## Intervalo de confiança

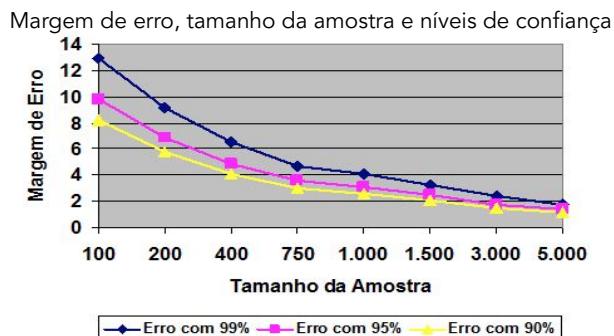
É um conjunto de valores no qual observamos um certo grau de confiança em que se presume estar o parâmetro de interesse, ou seja, o verdadeiro valor populacional.

Em estudos sobre saúde, geralmente opta-se por um **intervalo de confiança de 95%**.

Este conceito é necessário no contexto na pesquisa científica em saúde, porque geralmente analisamos amostras para fazer inferência a respeito de toda população.

O **Intervalo de Confiança** é influenciado por

- Tamanho da amostra
- Margem de erro
- Nível de confiança



## Fórmulas para cálculos amostrais

**Quadro 1.** Fórmulas para cálculo do tamanho de amostras para descrição de variáveis quantitativas e qualitativas em uma população.

	Variável quantitativa	Variável qualitativa
População infinita	$n = \left( \frac{Z\alpha/2 \cdot \delta}{E} \right)^2$	$n = \left( \frac{Z\alpha/2 \cdot \sqrt{p \cdot q}}{E} \right)^2$
População finita (<10000)	$n = \frac{N \cdot \delta^2 \cdot (Z\alpha/2)^2}{(N-1) \cdot (E)^2 + \delta^2 \cdot (Z\alpha/2)^2}$	$n = \frac{N \cdot p \cdot q \cdot (Z\alpha/2)^2}{(N-1) \cdot (E)^2 + p \cdot q \cdot (Z\alpha/2)^2}$

n – tamanho da amostra;  $Z_{\alpha/2}$  – valor crítico para o grau de confiança desejado, usualmente 1,96 (95%);  $\delta$  – desvio padrão populacional da variável; E – erro padrão, usualmente  $\pm 5\%$  da proporção dos casos (precisão absoluta), ou  $\pm 5\%$  da média ( $1,05 \times \text{média}$ ); N – tamanho da população (finita); p – proporção de resultados favoráveis da variável na população; q – proporção de resultados desfavoráveis na população ( $q=1-p$ ).

**Fonte:** Hélio Amante Miot. 2011. Tamanho da amostra em estudos clínicos e experimentais.

<https://doi.org/10.1590/S1677-54492011000400001>

## Intervalo de confiança

População	Margem de erro desejada			
	1%	3%	5%	10%
< 1.000			222	83
1.000			286	91
3.000		1.364	353	97
4.000		1.538	364	98
5.000		1.667	370	98
7.000		1.842	378	99
10.000	5.000	2.000	383	99
20.000	6.667	2.222	392	100
50.000	8.333	2.381	397	100
100.000	9.091	2.439	398	100
>100.000	10.000	2.500	400	100

Fonte: Arkin, H., & Colton, R. R. (1971). *Tables for statisticians*. Barnes and Noble.

**01. População**

É o conjunto total de indivíduos ou parâmetros que devem ser investigados. Por exemplo, todos os funcionários de uma empresa; toda a população de uma cidade.

**02. Margem de erro**

É o índice de variação dos resultados de uma pesquisa. Por exemplo, um erro amostral de 5% indica que o resultado poderá variar cinco pontos percentuais para mais ou para menos em sua pesquisa.

**04. Distribuição da população**

É o grau de homogeneidade da população, considerando aspectos relevantes tais como nível sociocultural, gênero, idade, entre outros. Por exemplo, uma pesquisa realizada numa cidade inteira requer um tratamento mais heterogêneo que uma pesquisa realizada dentro de uma empresa, onde a população pode estar distribuída de forma mais homogênea. Na prática, quanto menos variada é a população, menor é a amostra necessária.

**Fonte e Calculadora Amostral:** <https://comentto.com/calculadora-amostral/>

**Calculadora Amostral**

População	<input type="text"/>
Erro amostral (%)	<input type="text"/>
Nível de confiança	<input type="text" value="90%"/>
Distribuição da população	<input type="text" value="Mais homogênea (80/20)"/>
<b>CALCULAR</b>	

**Exemplo****Parâmetros de Pesquisa**

Universo: 1000 funcionários  
Margem de erro: 5%  
Nível de confiança: 95%  
Distribuição: Homogênea  
Amostra: 198

**Margem de Erro**

Indica que o resultado da pesquisa pode variar em 5%, ou seja, se a satisfação final obtida for de 80%, podemos dizer que a margem de erro para a satisfação pode variar de 75% a 85%.

**Nível de Confiança**

Significa dizer que, aplicando-se a outro grupo de funcionários da mesma empresa, ela deverá permanecer com pelo menos 95% dos resultados iguais ao da primeira pesquisa, respeitando-se a margem de erro anterior.

**Amostra**

Estatisticamente, considerando os parâmetros de pesquisa estipulados acima, constatamos que representando toda a empresa, se faz necessário entrevistarmos o número de 198 funcionários.

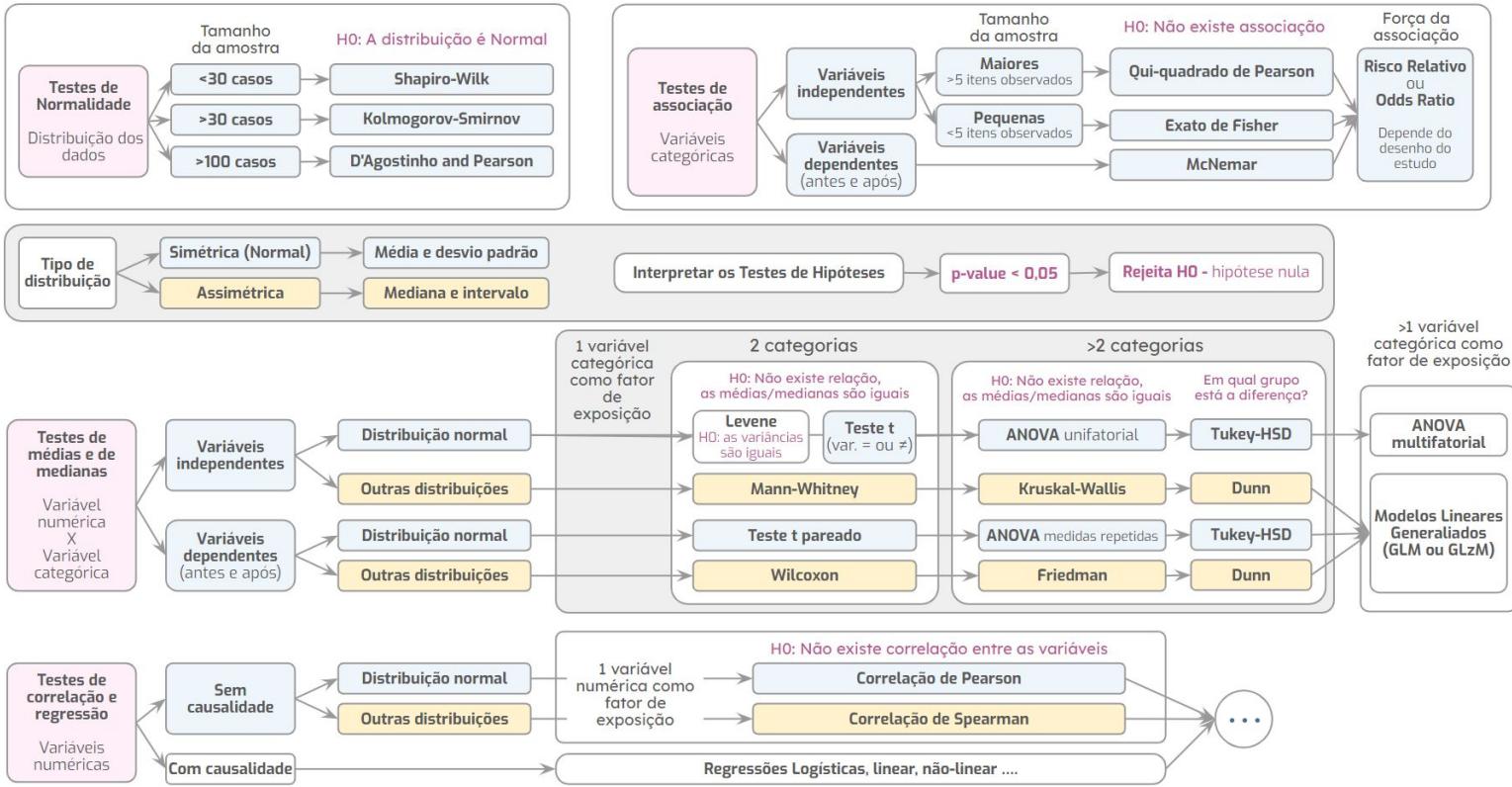
# Que teste devo usar?

## Como descrever meus dados e quais testes estatísticos devo usar?

Choosing the Right Statistical Test? Análise introdutória de dados em Saúde

### Como citar:

Gaspar, Juliano de Souza, 2023. Como descrever meus dados e quais testes estatísticos devo usar? Adaptado de Marco Mello e Jutta Schmid. DOI: <https://doi.org/10.5281/zenodo.7855276>



# Obrigado

Prof. D.r Juliano Gaspar

E-mail: [julianogaspar@gmail.com](mailto:julianogaspar@gmail.com)

Lattes: <http://lattes.cnpq.br/3926707936198077>

Orcid ID: <https://orcid.org/0000-0003-0670-9021>

## **Referências e materiais utilizados**

Com base nos incisos VI e VIII do artigo 46 da Lei 9.610, os autores do curso, valeram-se dos artigos relativos às exceções existentes na Lei de Direito Autoral que permitem a utilização de obras e/ou trecho de obras para fins didáticos.

As ilustrações utilizadas no curso foram de produção própria, desenvolvidas com a expertise acadêmica dos autores, repositórios de imagens livres ou obtidas através contratação de serviços de design e parceiras acadêmicas. As imagens fotográficas usadas foram as do acervo do Centro de Informática em Saúde da UFMG ou obtidas em repositórios livres ou adquiridas com recursos do projeto.

Mais informações

<https://ciia-saude.medicina.ufmg.br>

Realização



Apoio

