Big Data

Wellington Franco
Universidade Federal do Ceará – UFC
Campus de Crateús
wellington@crateus.ufc.br

Distribuição dos Dados

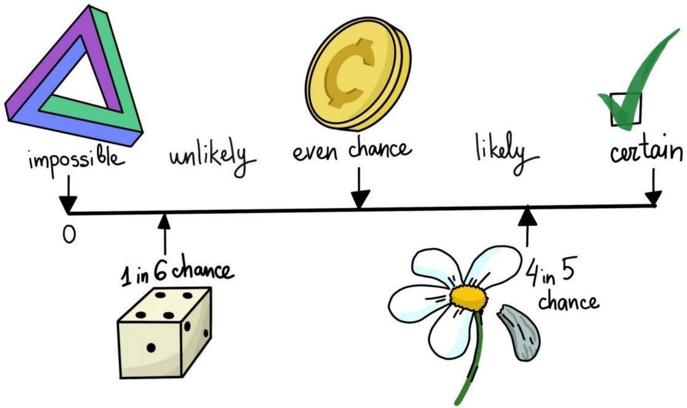
Ter um histórico estatístico do *dataset* pode ser muito benéfico na vida diária de um cientista de dados.

Sempre que começamos a explorar um novo *dataset*, precisamos primeiro fazer uma Análise Exploratória de Dados (EDA) para ter uma ideia das principais características do nosso conjunto.

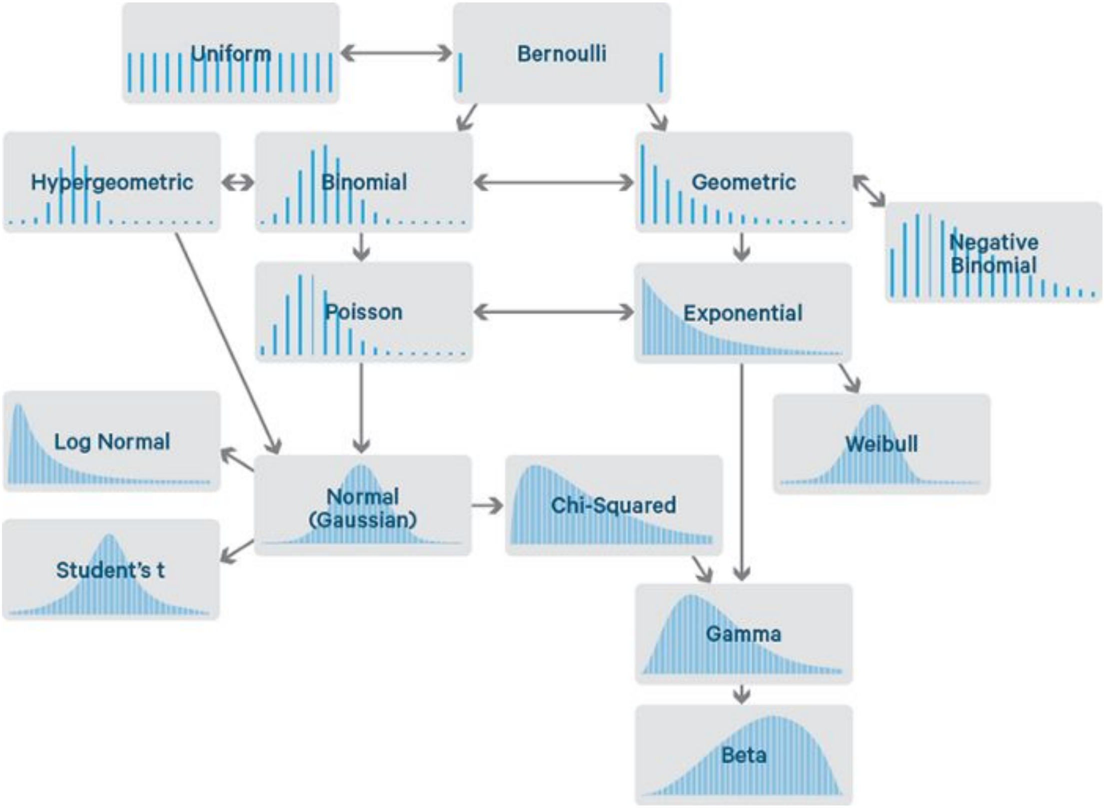
Se conseguirmos entender se existe algum padrão na distribuição de dados, podemos então personalizar nossos modelos de *Machine Learning* para melhor atender nosso estudo de caso.

Distribuição dos Dados

Dessa forma, conseguiremos um resultado melhor em menos tempo (reduzindo as etapas de otimização). De fato, alguns modelos de *Machine Learning* são projetados para funcionar melhor sob algumas suposições de distribuição. Portanto, saber com quais distribuições estamos trabalhando pode nos ajudar a identificar quais modelos são melhores para usar.

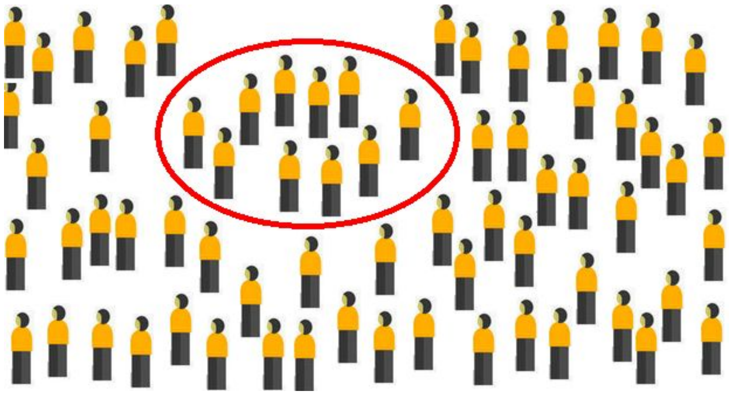


Distribuição dos Dados: visão geral



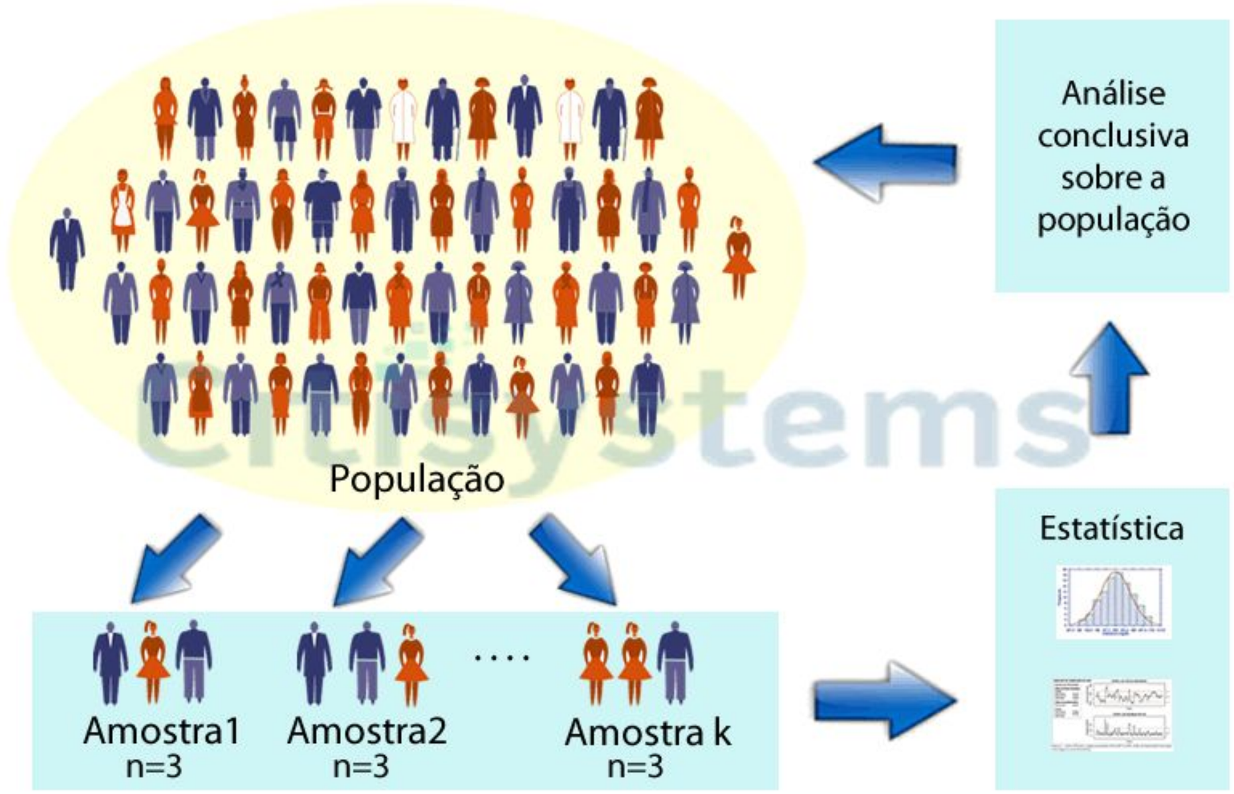
Distribuição dos Dados

Sempre que trabalhamos com um conjunto de dados, nosso conjunto de dados representa uma amostra de uma população.



Usando esta amostra, podemos tentar entender seus principais padrões para que possamos usá-la para fazer previsões para toda a população (mesmo que nunca tenhamos tido a oportunidade de examinar toda a população).

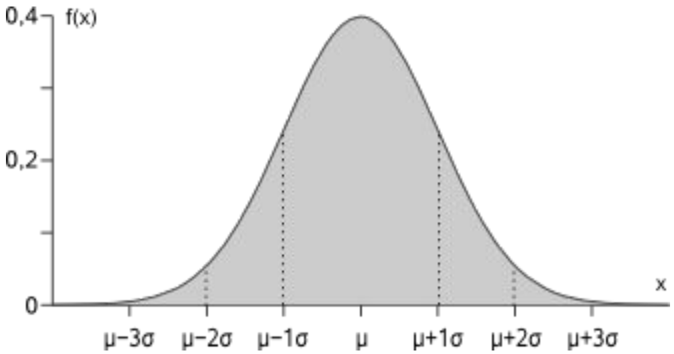
Distribuição dos Dados



Distribuição Normal

A distribuição normal é uma das distribuições mais falada na ciência de dados.

Muitos fenômenos comuns que ocorrem em nossa vida cotidiana seguem as Distribuições normais, como: a distribuição de renda na economia, os relatórios médios dos alunos, a altura média das populações, etc.



Distribuição Normal

Muitas técnicas em *machine learning* requerem que os dados estejam ajustados a uma distribuição normal para serem aplicadas, como por exemplo na hora em que vamos checar o grau de coeficiente de correlação entre variáveis usando a técnica de *Pearson*;

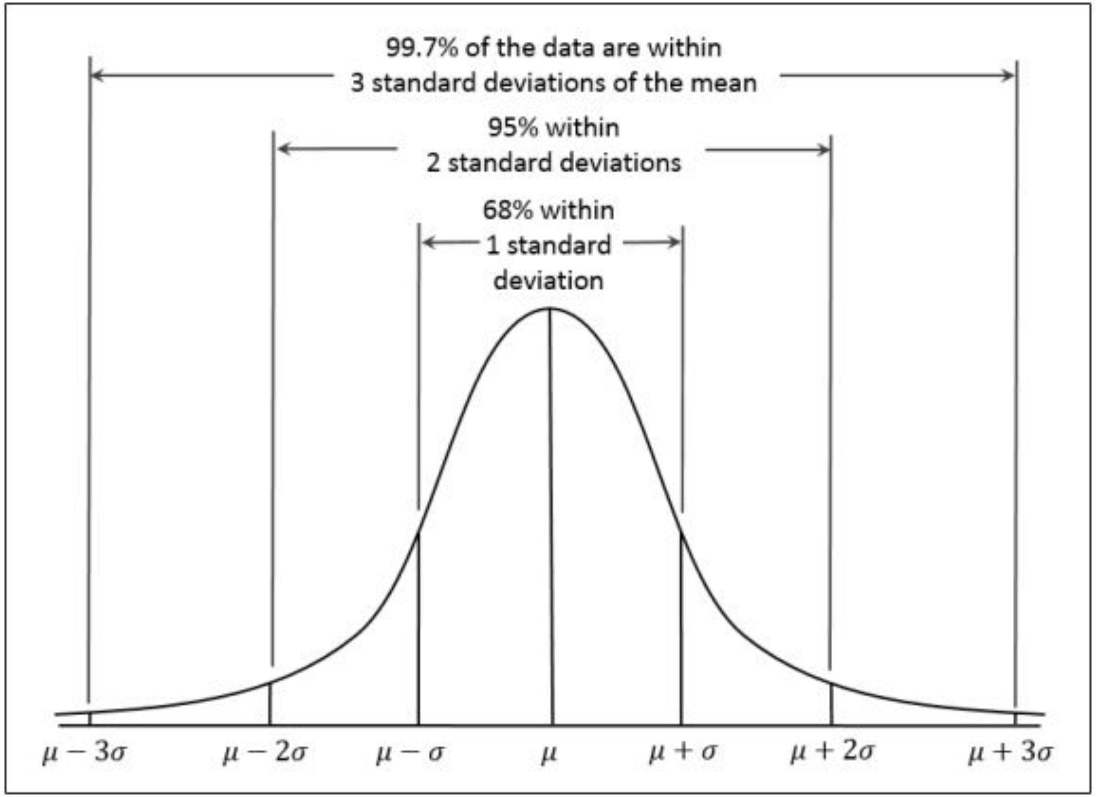
Quando uma técnica exige que a distribuição dos dados seja Normal, chamamos de **Paramétrica**; caso contrário, quando não há essa exigência, chamamos de **não-paramétrica**.

Distribuição Normal

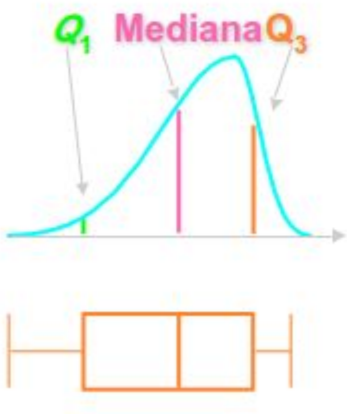
- **Erro:** a diferença entre o valor do dado e o valor predito (ou sua média);
- **Padronização (ou normalização):** subtrair do dado a média do conjunto e dividir pelo desvio padrão;
        - Sinônimo: *standardize*
- ***Z-score***: o resultado da padronização para cada dado do conjunto;
    - ***Standard normal***: uma distribuição normal com média = 0 e desvio padrão = 1;
- **QQ-Plot:** uma técnica de plot (dentre várias outras técnicas) para visualizar o quão próximo uma distribuição de amostra se aproxima da distribuição normal.

Distribuição Normal

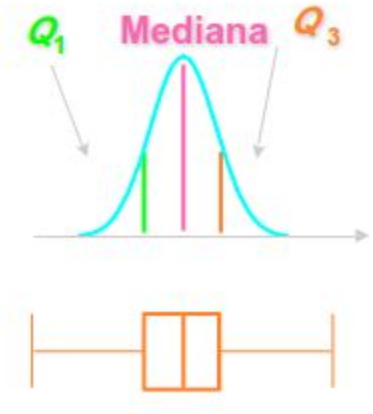
- Na distribuição normal, 68% dos dados estão concentrados a um desvio padrão da média;
- 95% dos dados estão a dois desvio padrão da média.



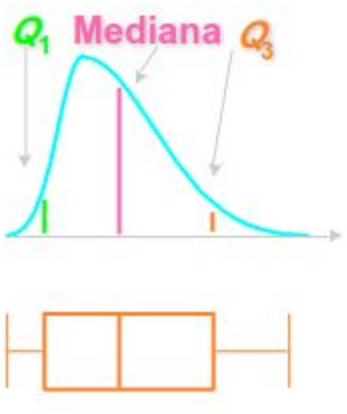
Distribuição Normal



Distribuição Não Normal



Distribuição Normal



Distribuição Não Normal

Observação

- Embora tenha o nome distribuição normal, na prática, o normal são os dados não seguirem essa distribuição;
- A utilidade da distribuição normal deriva do fato de muitas estatísticas serem normalmente distribuídas em sua distribuição amostral;
- Mesmo assim, as premissas de normalidade geralmente são o último recurso, usadas quando distribuições empíricas de probabilidade ou distribuições de *bootstrap* não estão disponíveis.

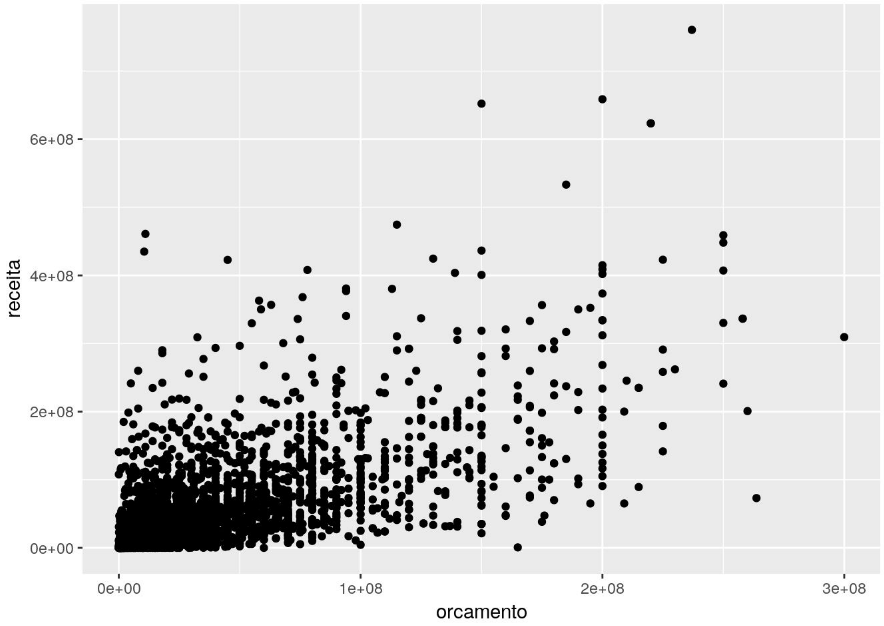
Distribuição Z e QQ-Plot

A distribuição Z é a normalização dos dados usando a técnica Z-*score* (subtrair cada elemento pela média do conjunto e dividir pelo desvio padrão).

Após a essa normalização, usa-se o QQ-Plot para determinar o quão próximo uma amostra está para uma distribuição normal.

Cuidado

A conversão de dados em *z-score* (isto é, padronizar ou normalizar os dados) **não torna os dados normalmente distribuídos**. Ele apenas coloca os dados na mesma escala da distribuição normal padrão, geralmente para fins de comparação.

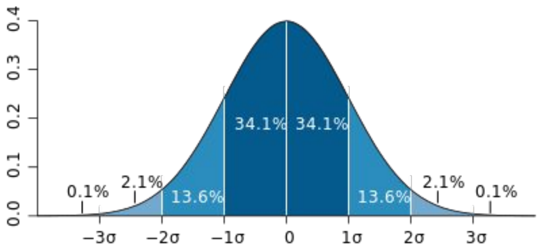


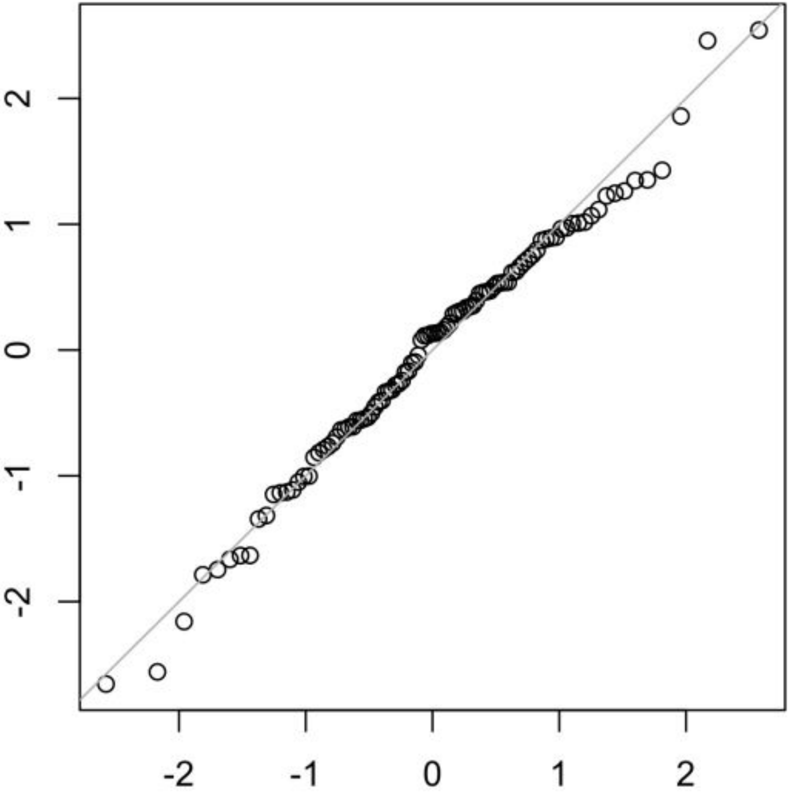
Distribuição Z e QQ-Plot

A figura ao lado mostra um QQ-Plot para uma amostra de 100 valores aleatoriamente gerada com uma distribuição normal.

O eixo X são as unidade de desvio padrão da média, e o eixo Y o valor do dado normalizado.

Perceba que a concentração de valores vai ficando maior entre -1 e 1 no eixo X.

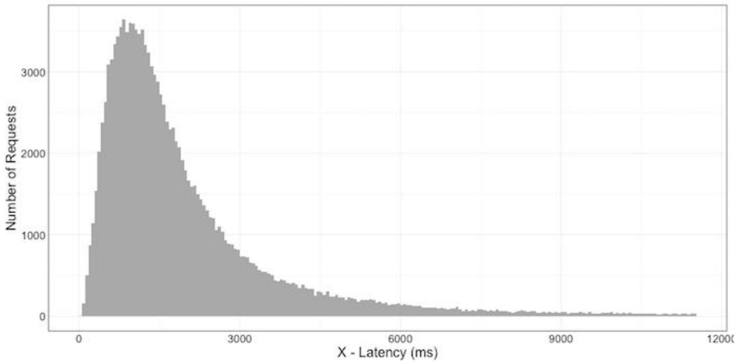




Long-Tailed (cauda longa)

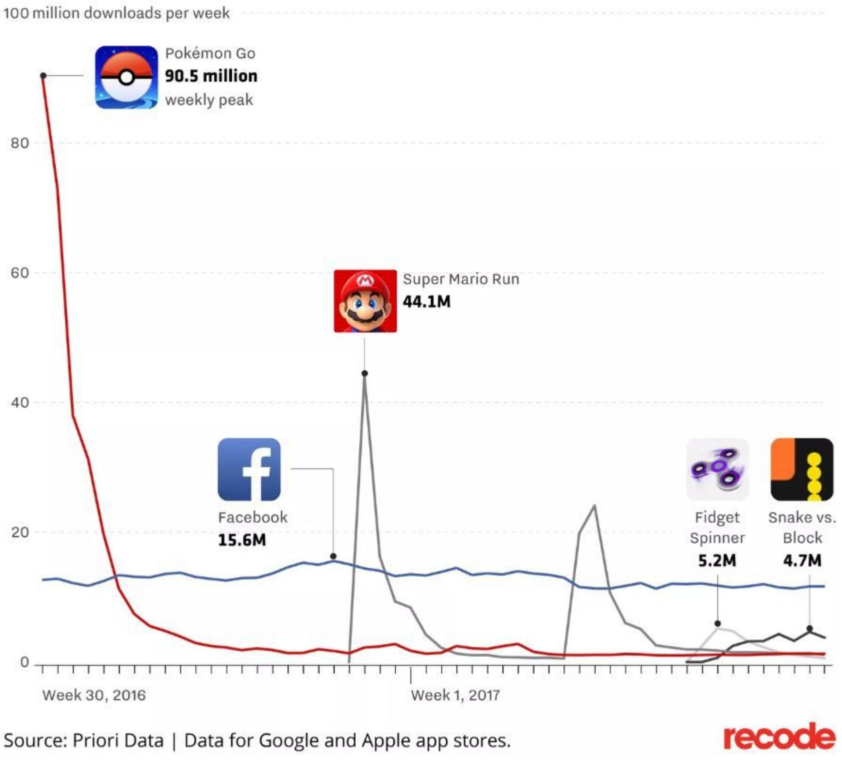
Apesar da importância histórica da distribuição Normal na estatística, na prática, os dados não seguem essa distribuição.

- *Tail* (cauda): a porção estreita e longa de uma distribuição de frequência, onde valores relativamente extremos ocorrem em baixa frequência;
- *skew*: onde a cauda de uma distribuição é maior do que a outra.
    - Sinônimo: assimétrica.



Long-Tailed (cauda longa)

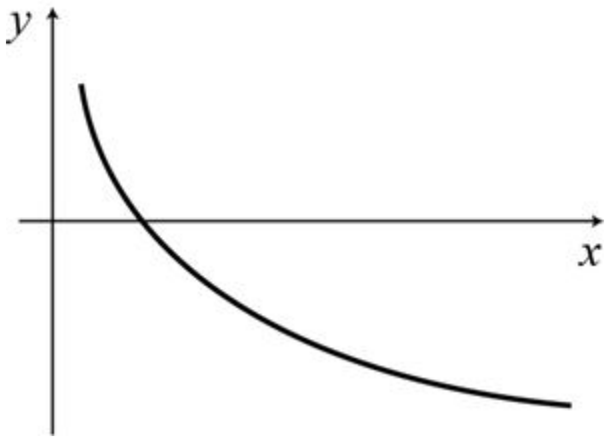
- Exemplo: Download do jogo Pokémon Go.
  - Fenômeno no seu lançamento, mas sua popularidade caiu drasticamente nos meses subsequentes.



Transformações

A normalidade nas distribuições por muitas vezes é a mais preferível devido a algumas poderosas ferramentas da estatística paramétrica;

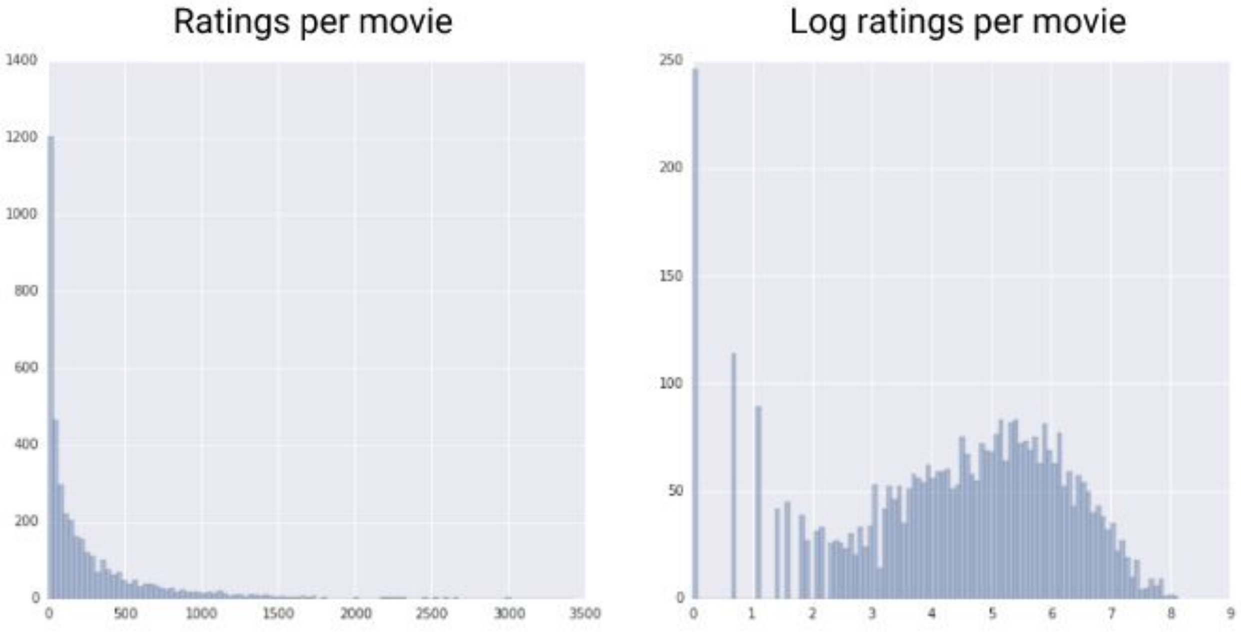
Uma das técnicas mais usadas na tentativa de normalizar os dados é a transformação.



Transformação logarítmica

A transformação logarítmica é dada por:

x’ = log(x)



Transformação logarítmica

Cuidado:

Não se deve esquecer portanto que, uma vez transformados os dados em logaritmos, a soma de dados logarítmicos não tem o mesmo valor que a soma de seus antilogaritmos, mas representa o produto destes, de modo que a média dos logaritmos não corresponde ao logaritmo da média de seus antilogaritmos. Portanto, a média deve ser computada pelos valores originais dos dados.

$$\log_a(b*c) = \log_a b + \log_a c$$

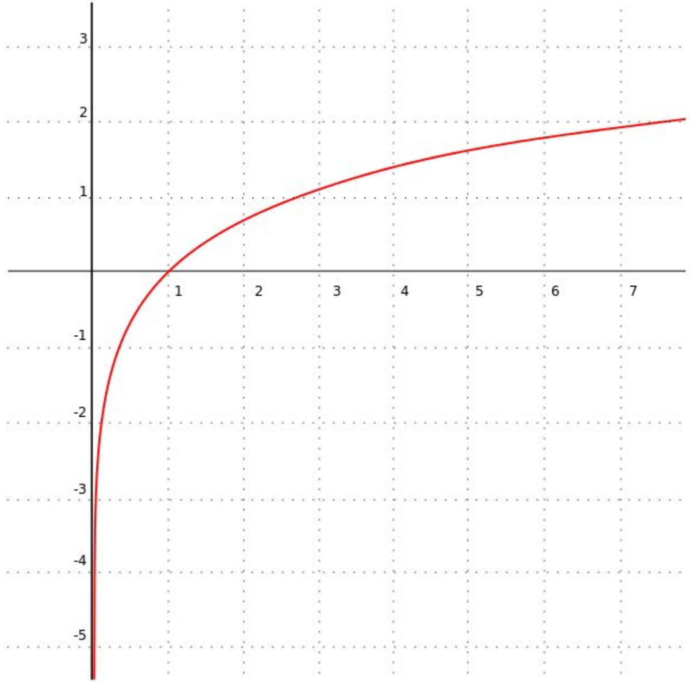
A única coisa que é mantida nesses casos é a hierarquia dos dados, pois quando um dado original é maior do que outro, os seus logaritmos mantêm essa mesma ordenação hierárquica, ainda que os próprios valores numéricos passem a ser diferentes

Transformação logarítmica

Cuidado:

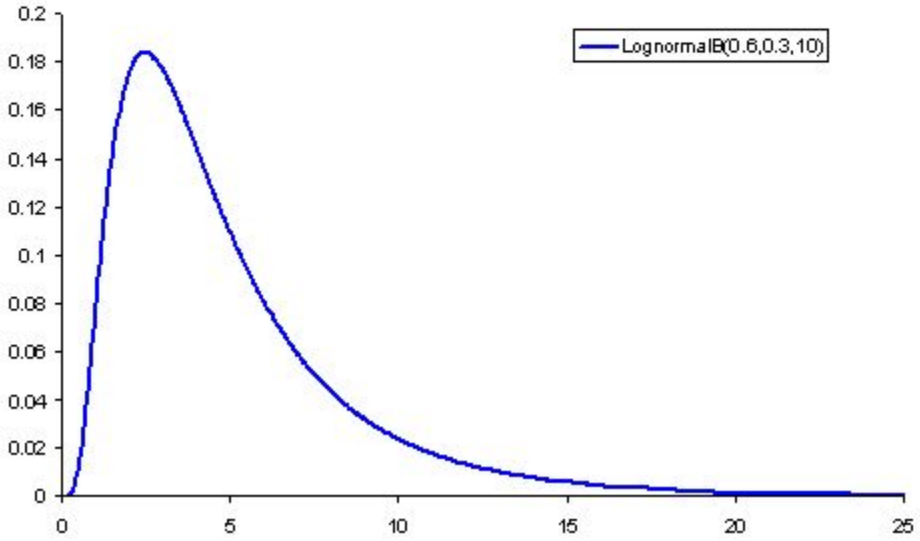
Se houver valor <= 0 no dataset, irá gerar um erro na hora da transformação, visto que não existe log para esse domínio. Portanto, o dataset deverá ser estritamente maior que zero.

Uma saída caso haja 0 no dataset é acrescentar +1 em todos os elementos: log(X**+1**)
Isso é válido pois log(1) = 0 e todos os outros elementos sofrerão *spread* de +1.



Algumas distribuições importantes da literatura

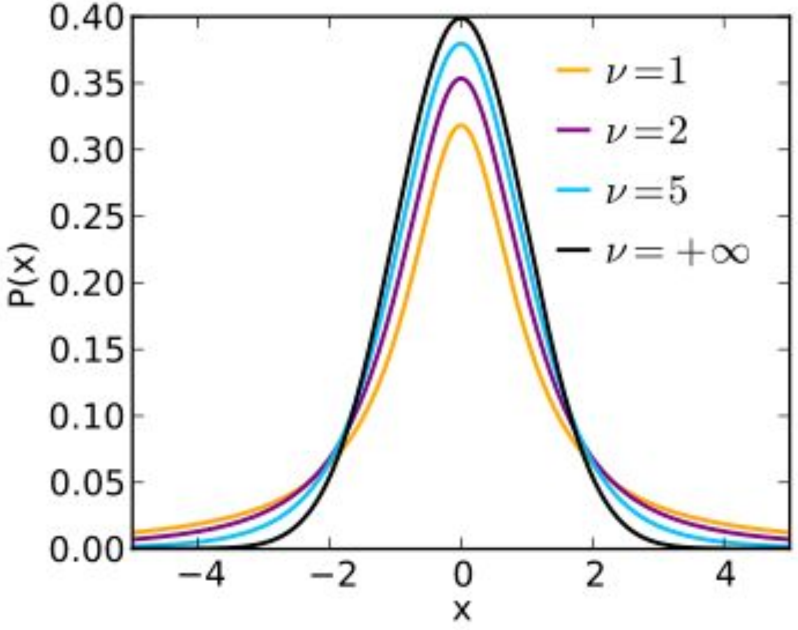
- **Lognormal**



Algumas distribuições importantes da literatura

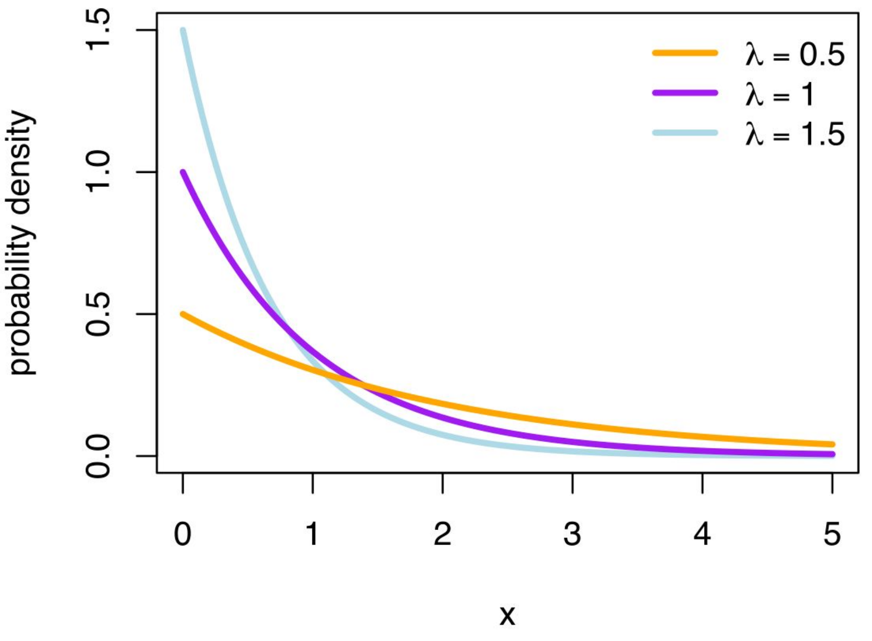
- **t de Student**

- Muito parecida com a distribuição Normal;
- Possui um parâmetro v, uma variável aleatória com distribuição Chi-quadrada com v graus de liberdade



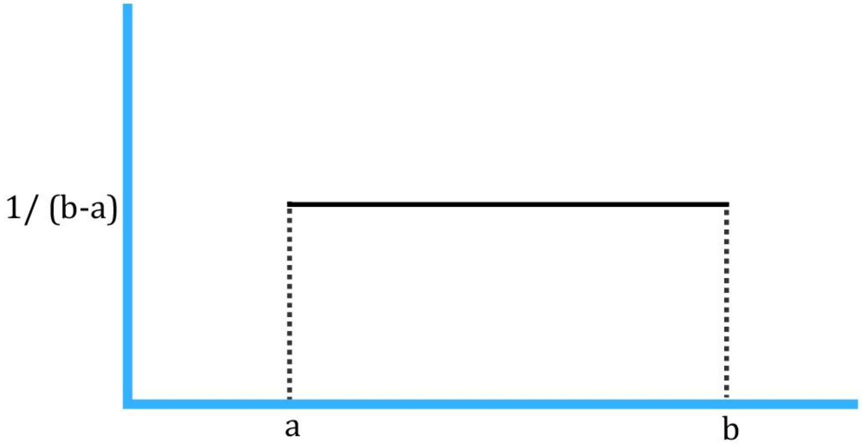
Algumas distribuições importantes da literatura

- **Exponencial**



Algumas distribuições importantes da literatura

- **Uniforme**



Transformação logarítmica

Dica de leitura:

https://towardsdatascience.com/transforming-skewed-data-73da4c2d0d16

https://medium.com/@kyawsawhtoon/log-transformation-purpose-and-interpretation-9444b4b049c9

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/

Bibliografia

1. **Practical Statistics for Data Scientists**. 50 essential concepts. Peter Bruce & Andrew Bruce. 1st ed. 2017.
2. **Bootstrap in Machine Learning.** https://www.analyticsvidhya.com/blog/2020/02/what-is-bootstrap-sampling-in-statistics-and-machine-learning/

3. Transformações logarítmicas:

    a. http://www.forp.usp.br/restauradora/gmc/gmc_livro/gmc_livro_cap13.html

    b. https://developers.google.com/machine-learning/data-prep/transform/normalization

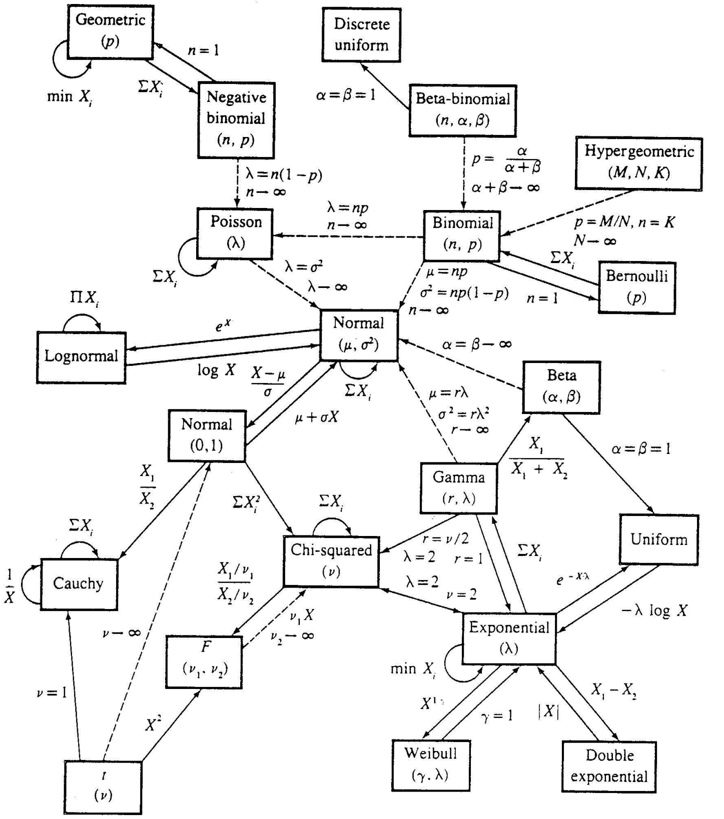
    c. https://towardsdatascience.com/log-transformation-base-for-data-linearization-does-not-matter-22eb3c1463d0

Testes de Hipótese

Introdução

Teste de hipótese é uma metodologia estatística que nos auxilia a tomar decisões sobre uma ou mais populações baseado na informação obtida da amostra.

630 *Table of Common Distributions*



Relationships among common distributions. Solid lines represent transformations and special cases, dashed lines represent limits. Adapted from Leemis (1986).

Teste de Hipótese

Nos permite verificar se os dados amostrais trazem evidência que apoiem ou não uma hipótese estatística formulada.

Ao tentarmos tomar decisões, é conveniente a formulação de suposições ou de conjeturas sobre as populações de interesse, que, em geral, consistem em **considerações sobre parâmetros** (μ, σ^2) das mesmas.

Essas suposições, que podem ser ou não verdadeiras, são denominadas de **Hipóteses Estatísticas**.

Teste de Hipótese

Portanto, **Teste de hipóteses** é um procedimento estatístico que permite tomar uma decisão entre duas ou mais hipóteses, utilizando os dados observados de um determinado experimento.

É importante destacar que:

Hipótese nula (H_0): É a hipótese assumida como verdadeira para a construção do teste. É a teoria, o efeito, ou a alternativa que se está interessado em testar.

Hipótese alternativa (H_1): É considerada quando a hipótese nula não tem evidência estatística e é formulada sendo contrária a hipótese nula.

Teste de Hipótese

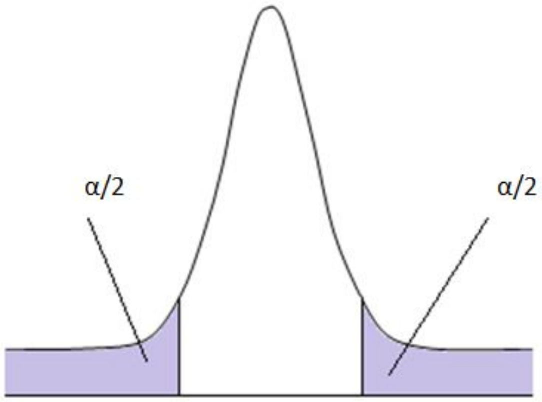
Nível de Significância (α):

É a probabilidade de rejeitar a hipótese nula quando ela é verdadeira.

Obs: Na linguagem coloquial, o termo "*significante*" quer dizer "algo importante" ao passo que, na linguagem estatística, esse termo tem o significado de "provavelmente verdadeiro" e, portanto, não resultante de uma situação aleatória.

Região Crítica: É o conjunto de valores assumidos pela variável aleatória ou estatística de teste para os quais a hipótese nula é rejeitada.

Nível de confiança: 1 - α



Teste de Hipótese

Para cientistas de dados, a distribuição mais desejada é a Normal (Gaussiana). Para isso, ao verificar a distribuição de um conjunto de dados, durante o teste de hipótese é comum usar como hipótese nula que tal distribuição é Normal.

A seguir, usaremos uma técnica para avaliar a distribuição do conjunto.

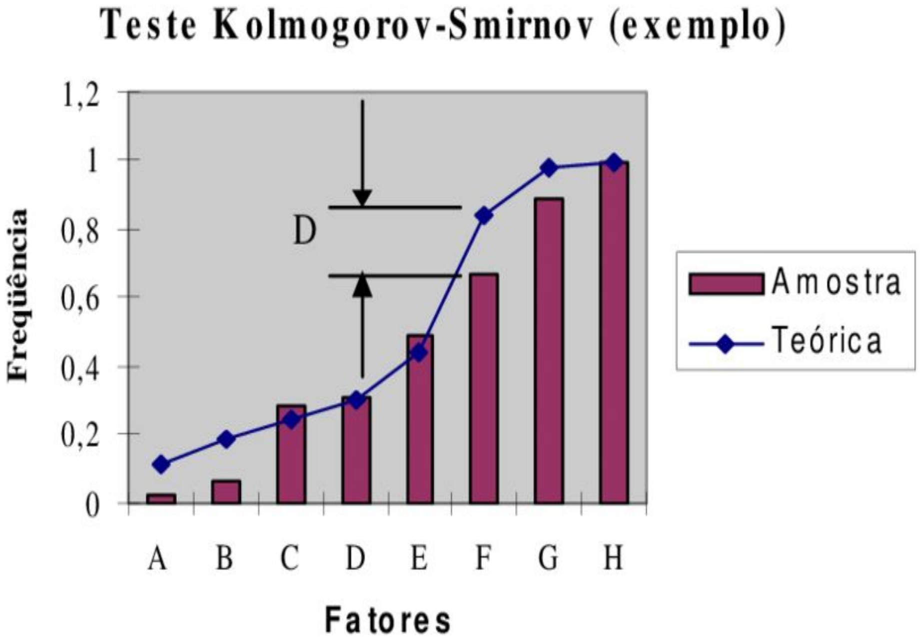
Teste de Kolmogorov-Smirnov

O teste de Kolmogorov-Smirnov pode ser utilizado para avaliar as hipóteses abaixo.

Hipóteses:

- **H_0**: Os dados seguem uma distribuição normal
- **H_1**: Os dados não seguem uma distribuição normal

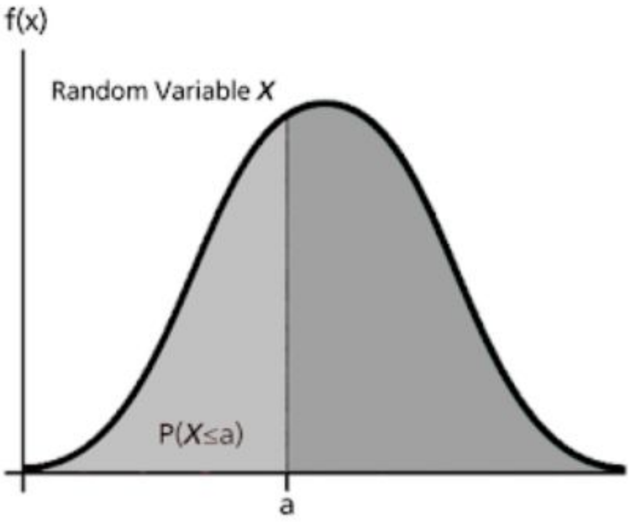
Este teste observa a diferença absoluta entre a função de distribuição acumulada assumida para os dados, no caso a distribuição Normal, e a função de distribuição empírica dos dados. Como critério, comparamos esta diferença com um valor crítico, para um dado nível de significância.

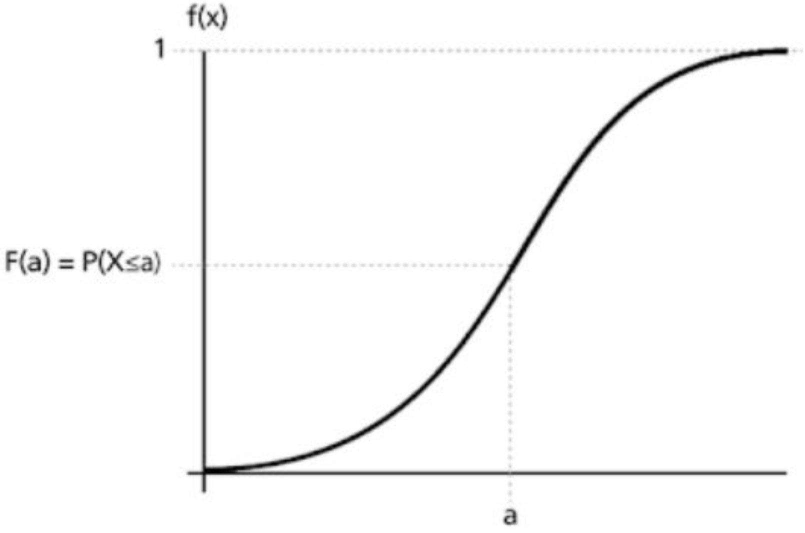


Teste de Kolmogorov-Smirnov

Atenção:

PDF (Função Densidade de Probabilidade) x **CDF** (Função de Distribuição Acumulada)





Teste de Kolmogorov-Smirnov

Considere uma amostra aleatória simples X_1, X_2, ..., X_n de uma população com função de distribuição acumulada Fx desconhecida. A estatística para o teste é:

$$D_n = \sup_x |F_n(x) - F(x)|$$

Esta função corresponde a **distância máxima vertical** entre os gráficos **F(x)** e **Fn(x)** sobre amplitude dos possíveis valores de x. Em **Dn** temos que:

- **F(x):** Representa a função de distribuição acumulada assumida para dados. No caso, a distribuição normal.
- **F_n(x):** Representa a função de distribuição acumulada empírica dos dados.

Teste de Kolmogorov-Smirnov

Queremos testar a hipótese $\mathbf{H_0: F_n(x) = F(x)}$ contra a $\mathbf{H_1: F_n(x) \neq F(x)}$

Para isto, tomamos $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ as observações aleatórias ordenadas de forma crescente da população com função de distribuição contínua F(x). No caso de análise da normalidade dos dados, assumimos F(x) a função de distribuição normal.

A função de distribuição acumulada assumida para os dados é definida por $F(x_{(i)}) = \mathcal{P}(X \leq x_{(i)})$

Teste de Kolmogorov-Smirnov

Além disso, haja visto que a função de distribuição empírica F_n é descontínua e a função da nossa distribuição hipotética é contínua, vamos considerar as seguintes estatísticas:

$$D^+ = \sup_{x_{(i)}} |F(x_{(i)}) - F_n(x_{(i)})|$$

$$D^- = \sup_{x_{(i)}} |F(x_{(i)}) - F_n(x_{(i-1)})|$$

Essas estatísticas medem as distâncias (verticais) entre os gráficos das duas funções, teórica e empírica, nos pontos $x_{(i-1)}$ e $x_{(i)}$. Com isso, podemos utilizar como estatística de teste:

$$D_n = \max(D^+, D^-)$$

Se D_n é maior que o valor crítico, rejeitamos a hipótese de normalidade; caso contrário, não rejeitamos.

Teste de Kolmogorov-Smirnov

O valor crítico pode ser encontrado em valores tabelados, bastando apenas saber o tamanho do conjunto e o nível de significância (que na literatura escolhe-se 0,05 por *default*)

	Nível de Significância α			
n	0,2	0,1	0,05	0,01
5	0,45	0,51	0,56	0,67
10	0,32	0,37	0,41	0,49
15	0,27	0,30	0,34	0,40
20	0,23	0,26	0,29	0,36
25	0,21	0,24	0,27	0,32
30	0,19	0,22	0,24	0,29
35	0,18	0,20	0,23	0,27
40	0,17	0,19	0,21	0,25
45	0,16	0,18	0,20	0,24
50	0,15	0,17	0,19	0,23
Valores maiores	$\frac{1,07}{\sqrt{n}}$	$\frac{1,22}{\sqrt{n}}$	$\frac{1,36}{\sqrt{n}}$	$\frac{1,63}{\sqrt{n}}$

Teste de Kolmogorov-Smirnov

Resumo das estatísticas de teste:

x(ordenado)	$F_n(x)$	$F(x) = \mathbb{P}\left(z_{(i)} \leq \frac{x_{(i)} - \overline{x}}{s}\right)$	$\mid F(x_{i)}) - F_n(x_{(i)}) \mid$	$\mid F(x_{(i)}) - F_n(x_{(i-1)}) \mid$
$x_{(1)}$	$\frac{1}{n}$	$F(x) = \mathbb{P}\left(z_{(1)} \leq \frac{x_{(1)} - \overline{x}}{s}\right)$	$\mid F(x_{(1)}) - F_n(x_{(1)}) \mid$	$\mid F(x_{(1)}) - 0) \mid$
$x_{(2)}$	$\frac{2}{n}$	$F(x) = \mathbb{P}\left(z_{(2)} \leq \frac{x_{(2)} - \overline{x}}{s}\right)$	$\mid F(x_{(2)}) - F_n(x_{(2)}) \mid$	$\mid F(x_{(2)}) - F_n(x_{(1)}) \mid$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$				
$x_{(n-1)}$	$\frac{n-1}{n}$	$F(x) = \mathbb{P}\left(z_{(n)} \leq \frac{x_{(n-1)} - \overline{x}}{s}\right)$	$\mid F(x_{(n-1)}) - F_n x_{(n-1)}) \mid$	$\mid F(x_{(n-1)}) - F_n(x_{(n-2)}) \mid$
$x_{(n)}$	1	$F(x) = \mathbb{P}\left(z_{(n-1)} \leq \frac{x_{(n)} - \overline{x}}{s}\right)$	$\mid F(x_{(n)}) - F_n(x_{(n)}) \mid$	$\mid F(x_{(n)}) - F_n(x_{(n-1)}) \mid$

Tabela 6.2.1: Estatísticas de teste.

OBS: O valor de $\mathbb{P}\left(Z_{(i)} \leq \frac{x_{(i)} - \bar{x}}{s}\right)$ é encontrado na tabela da distribuição normal padrão.

Teste de Kolmogorov-Smirnov

A tabela da Distribuição Normal Padrão pode ser acessada aqui:

http://wiki.icmc.usp.br/images/f/f9/Tabela_Normal.pdf

Tabela da Distribuição Normal Padrão

P(Z<z)

Precisão das casas decimais

z	0,0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
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
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441

Valor para 1,57:

Exemplo

Exemplo disponível no Notebook *Curso Big Data - Modulo 1 - Aula 3 - Parte 4.ipynb*

In [2]:

```
# Gerando um dataframe inicial para os seguintes dados abaixo

Data_df = pd.DataFrame(data=[52, 50, 44, 50, 42, 30, 36, 34, 48, 40, 55, 40,
          30, 36, 40, 42, 55, 44, 38, 42, 40, 38, 52, 44,
         52, 34, 38, 44, 48, 36, 36, 55, 50, 34, 44, 42])
Data_df
```

Out[2]:

	0
0	52
1	50
2	44
3	50
4	42
5	30
6	36
7	34
8	48
9	40
10	55

Exemplo retirado do youtube, porém, feito no Excel:
https://www.youtube.com/watch?v=-DrHe2IOD34

Exemplo

Calculando a Frequência Absoluta:

In [3]:
```
# Computando a Frequência Absoluta (Fabs) e colocando o resultado em um novo dataframe

table_df = Data_df.groupby([0]).size().reset_index(name='Fabs')
table_df
```

Out[3]:

	0	Fabs
0	30	2
1	34	3
2	36	4
3	38	3
4	40	4
5	42	4
6	44	5
7	48	2
8	50	3
9	52	3
10	55	3

Frequência absoluta é saber quantas ocorrências de cada valor possui no conjunto de dados.

Exemplo

Renomeando a coluna "0" gerada pelo DataFrame para X_i por questão de praticidade na interpretação :)

```
In [4]: # Renomeando a coluna de 0 para X_i (lê-se: x índice i)

        newcols = {
            0: 'Xi'
        }
        table_df.rename(columns=newcols, inplace=True)
        table_df
```

Out[4]:

	Xi	Fabs
0	30	2
1	34	3
2	36	4
3	38	3
4	40	4
5	42	4
6	44	5
7	48	2
8	50	3
9	52	3
10	55	3

Exemplo

Calculando a Frequência Acumulada:
É a soma da Frequência Absoluta atual ($Fabs_i$) com a anterior ($Fabs_{(i-1)}$)

```
In [5]: # Calculando a Frequência Acumulada (Fac)

        table_df['Fac'] = table_df['Fabs'].cumsum()
        table_df
```

Out[5]:

	Xi	Fabs	Fac
0	30	2	2
1	34	3	5
2	36	4	9
3	38	3	12
4	40	4	16
5	42	4	20
6	44	5	25
7	48	2	27
8	50	3	30
9	52	3	33
10	55	3	36

Exemplo

```
In [6]: # Calculando a coluna Fracionária: Total acumulado dividido pelo valor máximo total

        table_df['Frac'] = table_df['Fac']/table_df['Fac'].max()
        table_df
```

Out[6]:

	Xi	Fabs	Fac	Frac
0	30	2	2	$\boxed{0.055556}$ 2/36
1	34	3	5	0.138889
2	36	4	9	0.250000
3	38	3	12	0.333333
4	40	4	16	0.444444
5	42	4	20	0.555556
6	44	5	25	0.694444
7	48	2	27	0.750000
8	50	3	30	0.833333
9	52	3	33	0.916667
10	55	3	36	1.000000

Calcular a coluna Fracionária:
Frequência acumulada atual dividida pelo valor da maior frequência acumulada.

Exemplo

Calculando Z-score: $\mathbb{P}\left(Z_{(i)} \leq \frac{x_{(i)}-\bar{x}}{s}\right)$

Onde:

x(i) é o i-ésimo elemento da tabela

x-barra é a média amostral

s é o desvio padrão amostral

Z(i) é o resultado da normalização

$\mathcal{P}$() é a probabilidade da Distribuição Normal Padrão

```
In [7]: # Calculando a média

        mean = Data_df.mean()
        mean[0]

Out[7]: 42.638888888888886
```

```
In [8]: # Calculando o desvio padrão

        std = Data_df.std()
        std[0]

Out[8]: 7.099910573995292
```

Exemplo

```
table_df['Zi'] = table_df['Xi'].apply(lambda x: (x - mean)/std)
table_df
```

Out[9]:

	Xi	Fabs	Fac	Frac	Zi
0	30	2	2	0.055556	-1.780148
1	34	3	5	0.138889	-1.216760
2	36	4	9	0.250000	-0.935067
3	38	3	12	0.333333	-0.653373
4	40	4	16	0.444444	-0.371679
5	42	4	20	0.555556	-0.089985
6	44	5	25	0.694444	0.191708
7	48	2	27	0.750000	0.755096
8	50	3	30	0.833333	1.036789
9	52	3	33	0.916667	1.318483
10	55	3	36	1.000000	1.741023

$$\mathbb{P}\left(\boxed{Z_{(i)}} \leq \frac{x_{(i)}-\bar{x}}{s}\right)$$

Exemplo

A função de probabilidade da Distribuição Normal Padrão pode ser implementada por:

```
In [10]: import scipy.special as scsp
         def zScoreToPvalue(z):
             # Retornar p-value a partir do z-score
             return 0.5 * (1 + scsp.erf(z / np.sqrt(2)))
```

$$\boxed{\mathbb{P}}\left(Z_{(i)} \leq \frac{x_{(i)} - \bar{x}}{s}\right)$$

Exemplo

In [11]:
```
table_df['FracEsp'] = table_df['Zi'].apply(lambda x: zScoreToPvalue(x))
table_df
```

Out[11]:

	Xi	Fabs	Fac	Frac	Zi	FracEsp
0	30	2	2	0.055556	-1.780148	0.037526
1	34	3	5	0.138889	-1.216760	0.111848
2	36	4	9	0.250000	-0.935067	0.174877
3	38	3	12	0.333333	-0.653373	0.256758
4	40	4	16	0.444444	-0.371679	0.355066
5	42	4	20	0.555556	-0.089985	0.464149
6	44	5	25	0.694444	0.191708	0.576015
7	48	2	27	0.750000	0.755096	0.774904
8	50	3	30	0.833333	1.036789	0.850083
9	52	3	33	0.916667	1.318483	0.906329
10	55	3	36	1.000000	1.741023	0.959160

Aplicando a função em cada Z_i

z	0,0	0,01	0,02	0,03
0,0	0,5000	0,5040	0,5080	0,5120
0,1	0,5398	0,5438	0,5478	0,5517
0,2	0,5793	0,5832	0,5871	0,5910
0,3	0,6179	0,6217	0,6255	0,6293
0,4	0,6554	0,6591	0,6628	0,6664
0,5	0,6915	0,6950	0,6985	0,7019
0,6	0,7257	0,7291	0,7324	0,7357
0,7	0,7580	0,7611	0,7642	0,7673
0,8	0,7881	0,7910	0,7939	0,7967
0,9	0,8159	0,8186	0,8212	0,8238
1,0	0,8413	0,8438	0,8461	0,8485

Exemplo

Calculando **D⁻** :

In [12]:
```
# Result1 = FracEsp - Frac
table_df['D_negativo'] = abs(table_df['FracEsp']-table_df['Frac'])
table_df
```

Out[12]:

	Xi	Fabs	Fac	Frac	Zi	FracEsp	D_negativo
0	30	2	2	0.055556	-1.780148	0.037526	0.018030
1	34	3	5	0.138889	-1.216760	0.111848	0.027041
2	36	4	9	0.250000	-0.935067	0.174877	0.075123
3	38	3	12	0.333333	-0.653373	0.256758	0.076575
4	40	4	16	0.444444	-0.371679	0.355066	0.089379
5	42	4	20	0.555556	-0.089985	0.464149	0.091406
6	44	5	25	0.694444	0.191708	0.576015	0.118430
7	48	2	27	0.750000	0.755096	0.774904	0.024904
8	50	3	30	0.833333	1.036789	0.850083	0.016750
9	52	3	33	0.916667	1.318483	0.906329	0.010338
10	55	3	36	1.000000	1.741023	0.959160	0.040840

$\mid F(x_{i)}) - F_n(x_{(i)}) \mid$
$\mid F(x_{(1)}) - F_n(x_{(1)}) \mid$ $\mid F(x_{(2)}) - F_n(x_{(2)}) \mid$ $\vdots$
$\mid F(x_{(n-1)}) - F_n x_{(n-1)}) \mid$ $\mid F(x_{(n)}) - F_n(x_{(n)}) \mid$

Exemplo

Calculando **D⁺** :

In [14]:
```python
# Criando uma coluna de zeros
table_df['D_positivo'] = 0
table_df
```

Out[14]:

	Xi	Fabs	Fac	Frac	Zi	FracEsp	D_negativo	D_positivo
0	30	2	2	0.055556	-1.780148	0.037526	0.018030	0
1	34	3	5	0.138889	-1.216760	0.111848	0.027041	0
2	36	4	9	0.250000	-0.935067	0.174877	0.075123	0
3	38	3	12	0.333333	-0.653373	0.256758	0.076575	0
4	40	4	16	0.444444	-0.371679	0.355066	0.089379	0
5	42	4	20	0.555556	-0.089985	0.464149	0.091406	0
6	44	5	25	0.694444	0.191708	0.576015	0.118430	0
7	48	2	27	0.750000	0.755096	0.774904	0.024904	0
8	50	3	30	0.833333	1.036789	0.850083	0.016750	0
9	52	3	33	0.916667	1.318483	0.906329	0.010338	0
10	55	3	36	1.000000	1.741023	0.959160	0.040840	0

In [15]:
```python
for i in range(table_df['Frac'].shape[0]):
    if i > 0:
        table_df['D_positivo'].iloc[i] = table_df['FracEsp'].iloc[i] - table_df['Frac'].iloc[i-1]
    else:
        table_df['D_positivo'].iloc[i] = table_df['FracEsp'].iloc[i]
```

$\mid F(x_{(i)}) - F_n(x_{(i-1)}) \mid$
$\mid F(x_{(1)}) - 0) \mid$ $\mid F(x_{(2)}) - F_n(x_{(1)}) \mid$ $\vdots$
$\mid F(x_{(n-1)}) - F_n(x_{(n-2)}) \mid$ $\mid F(x_{(n)}) - F_n(x_{(n-1)}) \mid$

Exemplo

Calculando $\mathbf{D^+}$:

```
In [16]: table_df
Out[16]:
```

	Xi	Fabs	Fac	Frac	Zi	FracEsp	D_negativo	D_positivo
0	30	2	2	0.055556	-1.780148	0.037526	0.018030	0.037526
1	34	3	5	0.138889	-1.216760	0.111848	0.027041	0.056292
2	36	4	9	0.250000	-0.935067	0.174877	0.075123	0.035988
3	38	3	12	0.333333	-0.653373	0.256758	0.076575	0.006758
4	40	4	16	0.444444	-0.371679	0.355066	0.089379	0.021733
5	42	4	20	0.555556	-0.089985	0.464149	0.091406	0.019705
6	44	5	25	0.694444	0.191708	0.576015	0.118430	0.020459
7	48	2	27	0.750000	0.755096	0.774904	0.024904	0.080460
8	50	3	30	0.833333	1.036789	0.850083	0.016750	0.100083
9	52	3	33	0.916667	1.318483	0.906329	0.010338	0.072996
10	55	3	36	1.000000	1.741023	0.959160	0.040840	0.042494

$\mid F(x_{(i)}) - F_n(x_{(i-1)}) \mid$
$\mid F(x_{(1)}) - 0) \mid$ $\mid F(x_{(2)}) - F_n(x_{(1)}) \mid$ $\vdots$
$\mid F(x_{(n-1)}) - F_n(x_{(n-2)}) \mid$ $\mid F(x_{(n)}) - F_n(x_{(n-1)}) \mid$

Exemplo

Calculados D^- e D^+ , vamos obter o maior valor de D e calcular o valor de *p-value* (de acordo com o tamanho do conjunto e o nível de significância definido em 0.05)

```
In [17]: # Calcular o máximo valor da coluna Result1 e depois o máximo da coluna Result 2
         # E, por fim, retornar o maior dos dois
         D = ( table_df[['D_negativo','D_positivo']].max() ).max()
         D

Out[17]: 0.1184298337535814
```

```
In [18]: from scipy.stats import ksone

         def ks_critical_value(n_trials, alpha):
             return ksone.ppf(1-alpha/2, n_trials)
```

```
In [19]: # n-trials: quantidade de dados
         # alpha: Um bom valor para o nível de significância do teste é com um alfa = 0,05 para assegurar 95% de confiança
         p_value = ks_critical_value(Data_df.shape[0], 0.05)
         p_value

Out[19]: 0.22119105379255108
```

Exemplo

Por fim, fazemos a checagem para tirar a conclusão:

```
In [20]: if D < p_value:
             print('Os dados seguem uma distribuição normal')
         else:
             print('Os dados não seguem uma distribuição normal')

         Os dados seguem uma distribuição normal
```

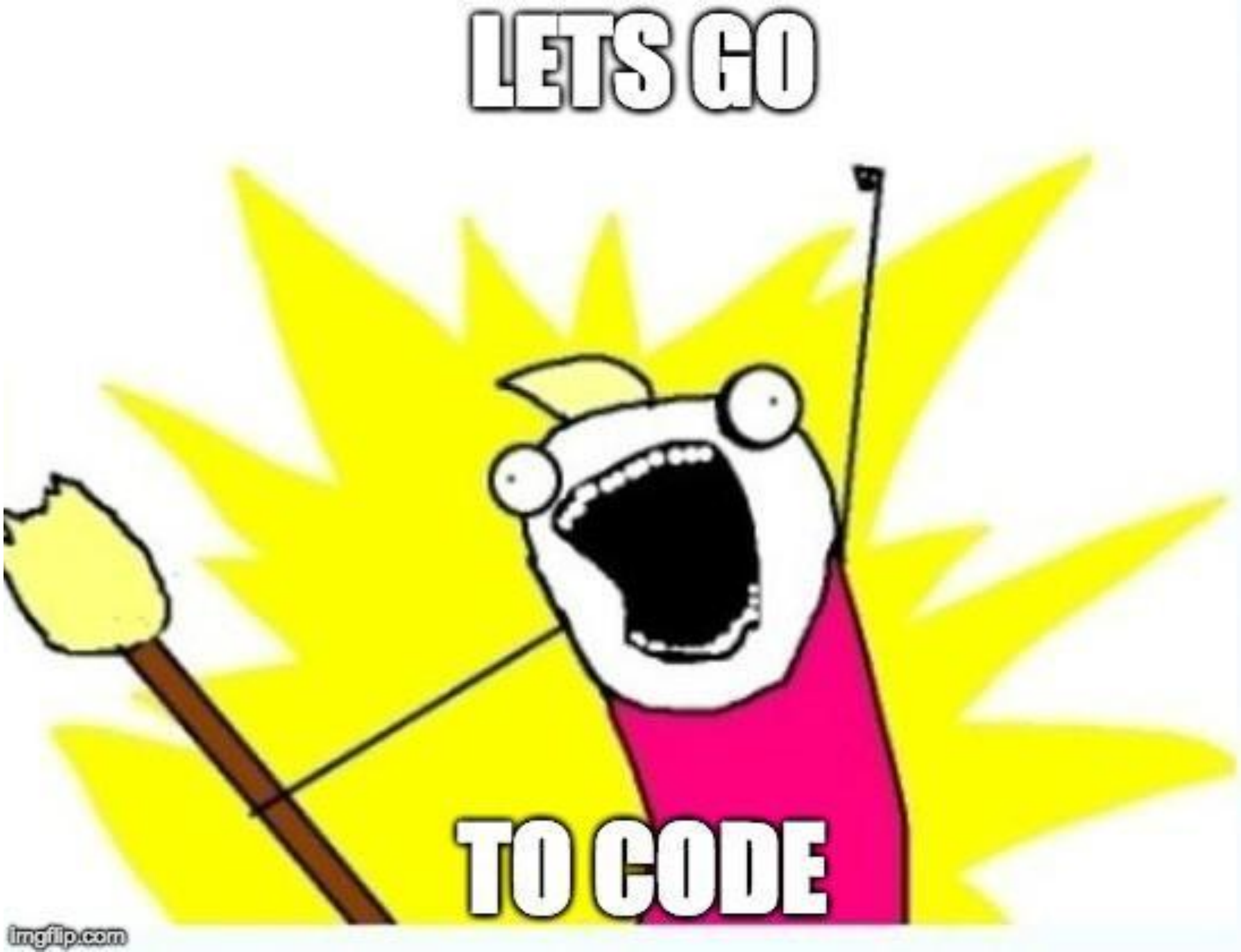
Bibliografia

Testes de Hipótese: https://www.inf.ufsc.br/~andre.zibetti/probabilidade/teste-de-hipoteses.html

https://blog.minitab.com/pt/como-compreender-os-testes-de-hipoteses-niveis-de-significancia-alfa-e-valores-p-na-estatistica

Kolmogorov-Smirnov (K-S): http://www.portalaction.com.br/inferencia/62-teste-de-kolmogorov-smirnov

Exemplo de K-S (aplicado no Excel): https://www.youtube.com/watch?v=-DrHe2IOD34


LETS GO
TO CODE
imgflip.com