UNIVERSIDADE FEDERAL DO PIAUÍ

Reconhecimento de Padrões
Feature Selection

Profa: Deborah Magalhães

“

***Feature engineering** é o processo de*
transformar** dados **em** features **que melhor
representam o problema tratado,
*resultando na melhoria de **desempenho***
do algoritmo de aprendizado de máquina.

Feature Engineering Made Easy, 2018 (pg - 35)

“

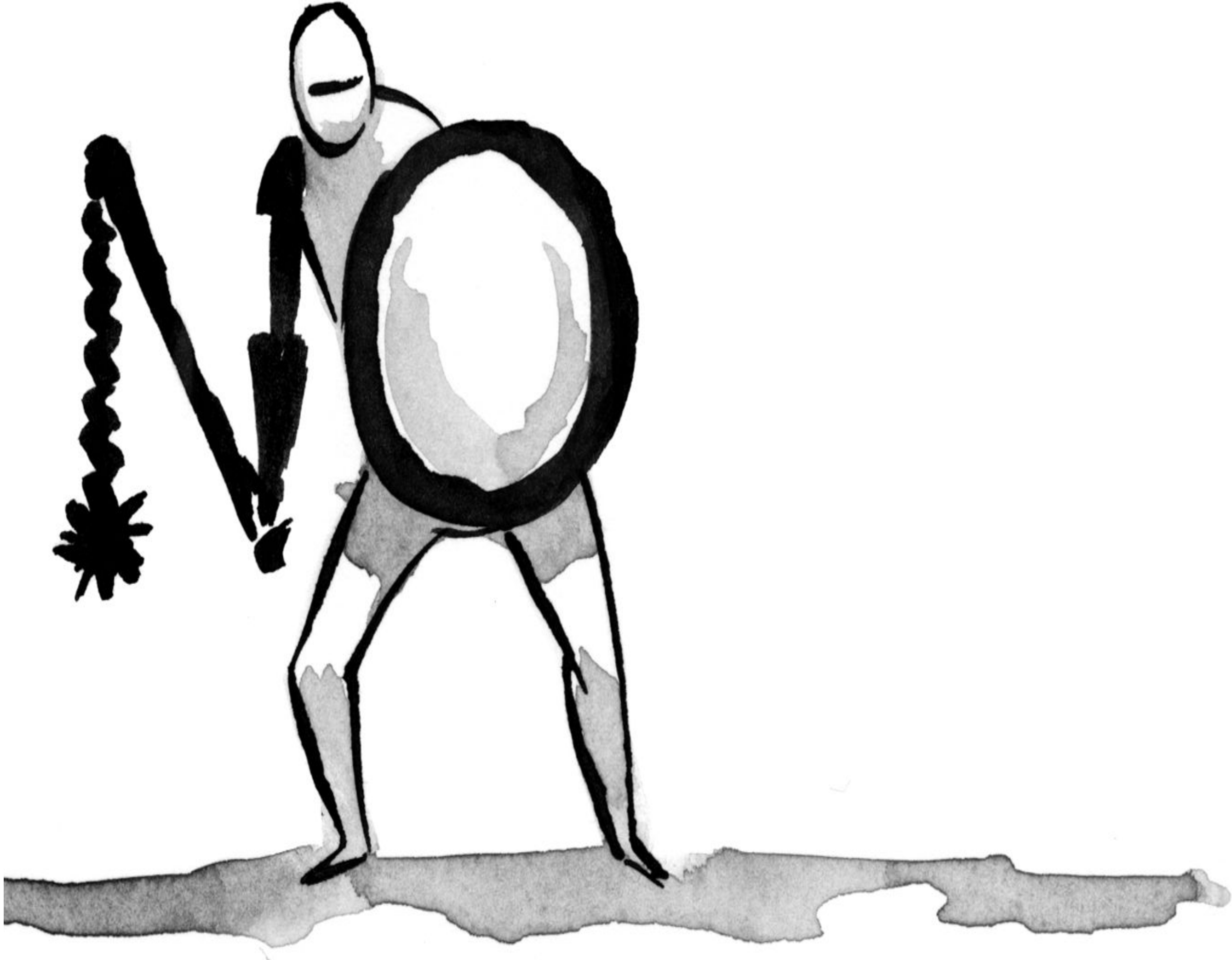
__Atributo__ é geralmente o termo dado a uma coluna de uma tabela de dados, enquanto __característica__ (feature) se refere apenas ao atributo que contribui para o sucesso do algoritmos de aprendizado de máquina.

Feature Engineering Made Easy, 2018 (pg - 47)

Say no to bad attributes

✓ Performance

✓ Time



Baseados em estatística

- Correlação de Pearson
- Testes de Hipótese

Baseados em modelos

- Recursive Feature Elimination (RFE)
- Information Gain (IG)

Pearson product-moment correlation coefficient (PPMCC) OU Coeficiente de Pearson

$$r_{xy} = \frac{\sum x_i y_i - n\ \overline{x}\ \overline{y}}{\sqrt{\sum x_i^2 - n\overline{x}^2}\sqrt{\sum y_i^2 - n\overline{y}^2}},$$

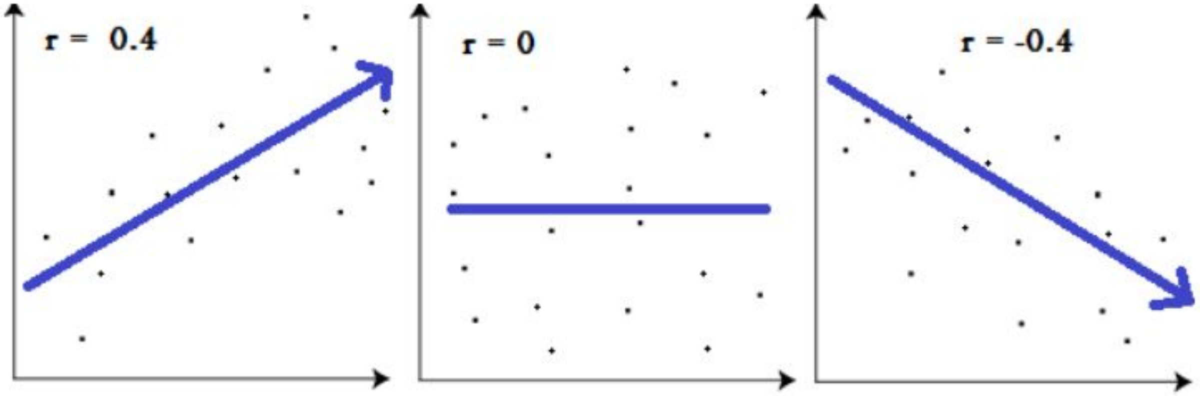
$$onde:$$

$$n : tamanho\ da\ amostra$$

$$x_i, y_i : valor\ da\ observacao\ indexado\ por\ i$$

$$\overline{x}, \overline{y} : media\ amostral$$

Coeficiente de Pearson



Força da Associação	Coeficiente (r)	
	Positiva	Negativa
Pequena	.1 a .3	-.1 a -.3
Média	.3 a .5	-.3 a -.5
Grande	.5 a 1.0	-.5 a -1.0

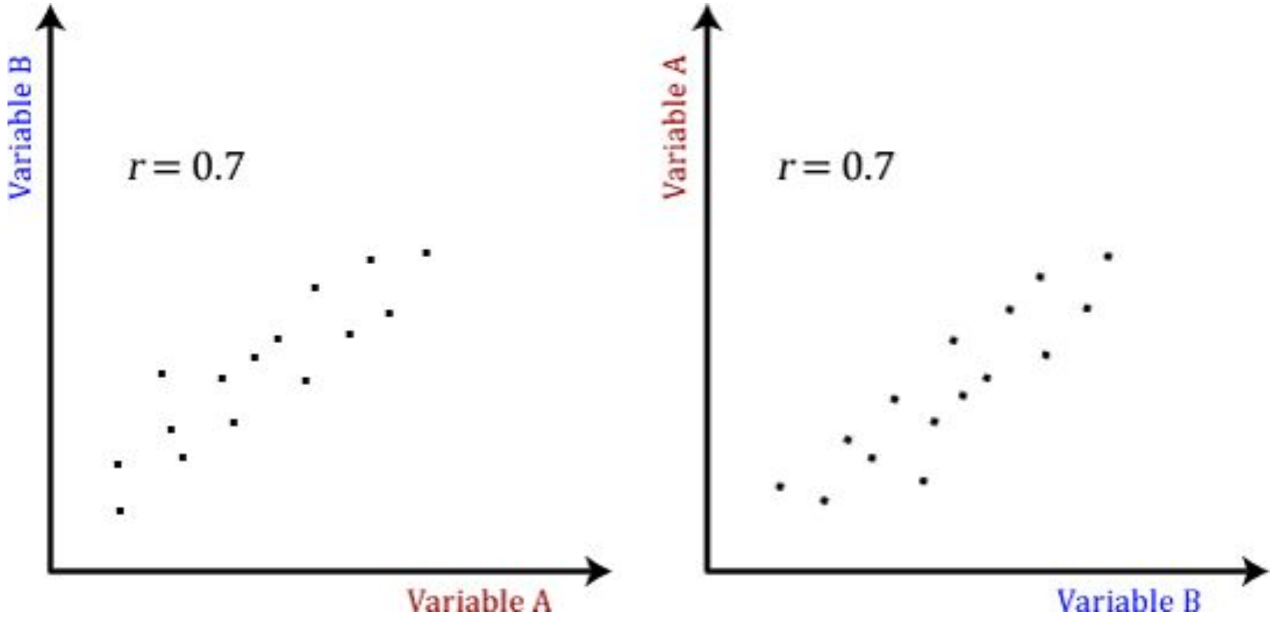
Desvantagens da Correlação de Pearson

#1 - Não se pode utilizar qualquer tipo de variável

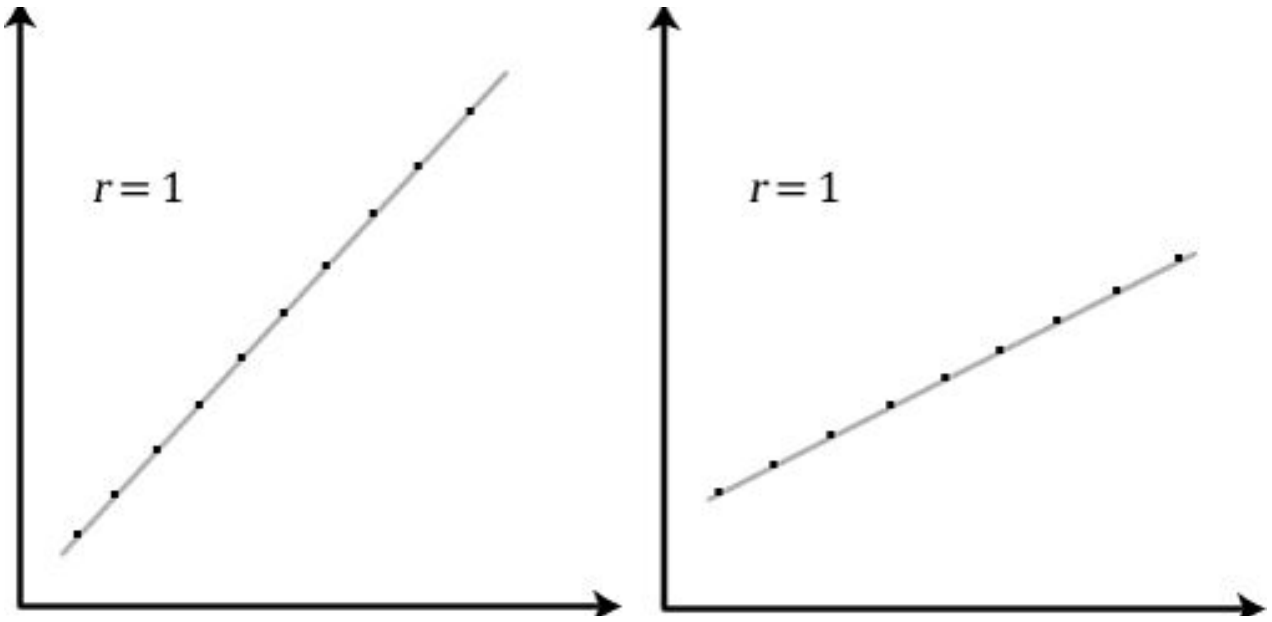
Quantitativa vs. Categória

Quantitativa	O que os dados representam?	Exemplos
Discreta	Número contável entre quaisquer dois valores	Número de reclamações de clientes, número de falhas de uma peça
Contínua	**Número infinito entre quaisquer dois valores**	**Comprimento, volume, saldo**

#2 - Não há diferença entre variáveis dependentes e independentes



#3 - Não representa a inclinação da linha de melhor ajuste



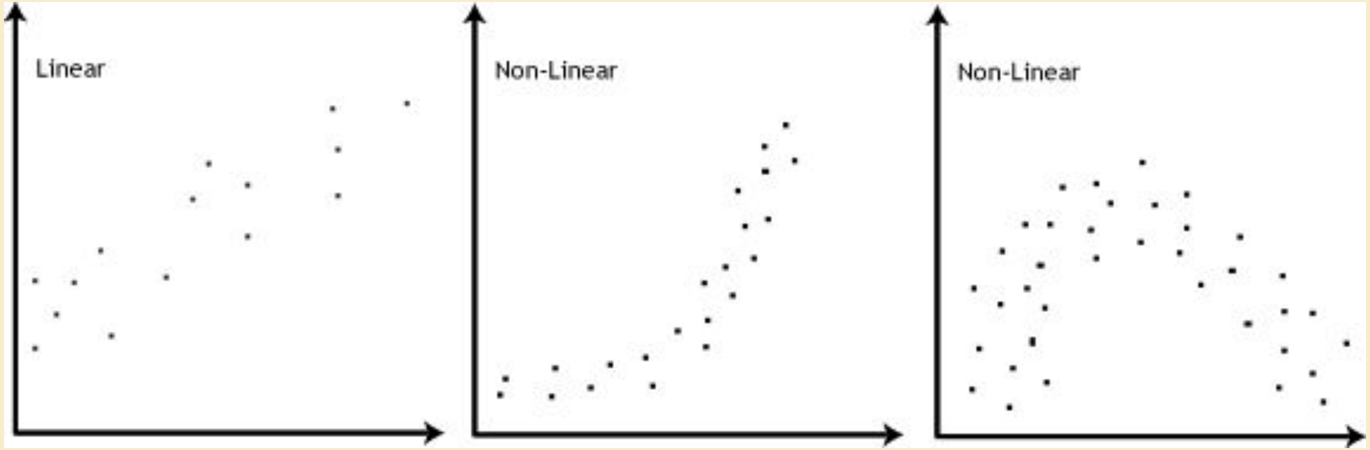
Premissas da correlação de Pearson

#1: as duas variáveis devem ser contínuas

#2: independência das observações

#3: as variáveis devem seguir a distribuição normal univariada

#4: as variáveis devem possuir uma relação linear



“

Hipótese:

Este atributo não tem relevância para o alvo.

ANOVA (ANalysis Of VAriance) entre duas amostras distintas

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_1 : \sigma_1^2 > \sigma_2^2, \; unicaudal \; a \; dir.$$

$$Teste \; estatistico : F = \frac{\sigma_1^2}{\sigma_2^2}, \; sendo \; v = n - 1$$

$$C : \left\{ f > f_{\alpha, v_1, v_2} \quad unicaudal \; a \; dir. \right.$$

Exemplo: com os dados das amostras abaixo, teste H_0: $\sigma_1^2 \leq \sigma_2^2$ e H_1: $\sigma_1^2 > \sigma_2^2$. O que podemos concluir sobre a hipótese nula? Considere α=0.05.

Amostra 1			
19	20	29	23
29	26	22	19
26	13	34	16
19	30	27	25

Amostra 2			
41	47	41	44
43	44	50	47
44	46	41	47
48	44	46	44

$H_0: \sigma_1^2 \leq \sigma_2^2$ e $H_1: \sigma_1^2 > \sigma_2^2$, $\alpha = 0.05$

$$Teste\ estatistico: F = \frac{\sigma_1^2}{\sigma_2^2}$$

$$= \frac{30.1210}{6.5273}$$

$$= 4.6146$$

$$f_{0.05,15,15} = 2.403$$

Limites unilaterais da distribuição F de Fisher-Snedecor ao nível de 5% de probabilidade.

GL	V1															
V2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	244.7	245.4	[illegible]	248.0
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396	19.405	19.412	19.419	19.424	19.429	19.446
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.763	8.745	8.729	8.715	8.703	8.660
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912	5.891	5.873	5.858	5.803
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.704	4.678	4.655	4.636	4.619	4.558
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.027	4.000	3.976	3.956	3.938	3.874
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.603	3.575	3.550	3.529	3.511	3.445
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.313	3.284	3.259	3.237	3.218	3.150
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.102	3.073	3.048	3.025	3.006	2.936
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.943	2.913	2.887	2.865	2.845	2.774
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.818	2.788	2.761	2.739	2.719	2.646
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.717	2.687	2.660	2.637	2.617	2.544
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.635	2.604	2.577	2.554	2.533	2.459
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.565	2.534	2.507	2.484	[illegible]	2.388
15	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	2.403	2.328

https://www.statology.org/f-distribution-calculator/

$H_0: \sigma_1^2 \leq \sigma_2^2$ e $H_1: \sigma_1^2 > \sigma_2^2$, $\alpha = 0.05$

$$Teste\ estatistico: F = \frac{\sigma_1^2}{\sigma_2^2}$$

$$= \frac{30.1210}{6.5273}$$

$$= 4.6146$$

$$f_{0.05,15,15} = 2.403$$

Conclusão: 4.61 > 2.403, portanto F está dentro da região crítica, logo H_0 **é rejeitada**

Limites unilaterais da distribuição F de Fisher-Snedecor ao nível de 5% de probabilidade.

GL	V1															
V2	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	20
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.0	243.9	244.7	245.4	245.9	248.0
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396	19.405	19.412	19.419	19.424	19.429	19.446
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.763	8.745	8.729	8.715	8.703	8.660
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912	5.891	5.873	5.858	5.803
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.704	4.678	4.655	4.636	4.619	4.558
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.027	4.000	3.976	3.956	3.938	3.874
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.603	3.575	3.550	3.529	3.511	3.445
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.313	3.284	3.259	3.237	3.218	3.150
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.102	3.073	3.048	3.025	3.006	2.936
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.943	2.913	2.887	2.865	2.845	2.774
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.818	2.788	2.761	2.739	2.719	2.646
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.717	2.687	2.660	2.637	2.617	2.544
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.635	2.604	2.577	2.554	2.533	2.459
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.565	2.534	2.507	2.484	2.463	2.388
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.507	2.475	2.448	[illegible]	2.403	2.328

https://www.statology.org/f-distribution-calculator/

ANOVA (ANalysis Of VAriance) entre duas amostras distintas

$$H_0 : \sigma_1^2 \geq \sigma_2^2$$
$$H_1 : \sigma_1^2 < \sigma_2^2, \ \ unicaudal \ a \ esq.$$

$$Teste \ estatistico : F = \frac{\sigma_1^2}{\sigma_2^2}, \ \ sendo \ v = n - 1$$

$$C : \left\{ f < f_{1-\alpha, v_1, v_2}, unicaudal \ a \ esq. \right.$$

Exemplo: com os dados das amostras abaixo, teste H_0: $\sigma_1^2 \geq \sigma_2^2$ e H_1: $\sigma_1^2 < \sigma_2^2$. O que podemos concluir sobre a hipótese nula? Considere α=0.05.

	Amostra 1	Amostra 2
n	25	50
v	24	49
μ	2.5	5.0
σ^2	0.02	0.16
σ	0.14	0.40

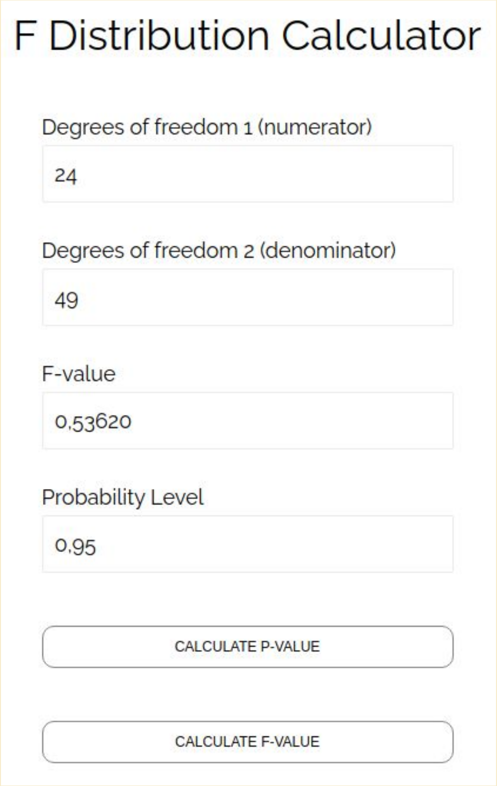
$H_0: \sigma_1^2 \geq \sigma_2^2$ e $H_1: \sigma_1^2 < \sigma_2^2$, $\alpha = 0.05$

$$Teste\ estatistico : F = \frac{\sigma_1^2}{\sigma_2^2}$$

$$= \frac{0.02}{0.16}$$

$$= 0.125$$

$$f_{0.95,24,49} = 0.5362$$



https://www.statology.org/f-distribution-calculator/

$H_0: \sigma_1^2 \geq \sigma_2^2$ e $H_1: \sigma_1^2 < \sigma_2^2$, $\alpha = 0.05$

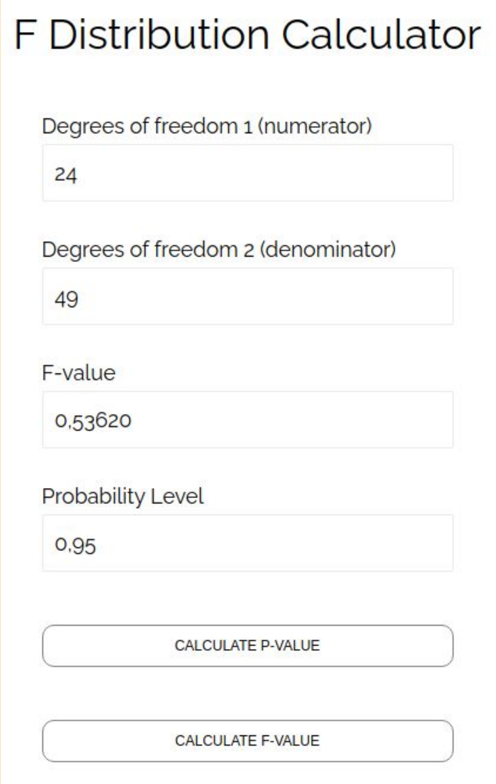
$$Teste\ estatistico : F = \frac{\sigma_1^2}{\sigma_2^2}$$

$$= \frac{0.02}{0.16}$$

$$= 0.125$$

$$f_{0.95,24,49} = 0.5362$$

Conclusão: 0.125 < 0.5362, portanto F está dentro da região crítica, logo **H_0 é rejeitada!**



https://www.statology.org/f-distribution-calculator/

ANOVA (ANalysis Of VAriance) entre duas amostras distintas

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2, \;\; bicaudal$$

$$Teste\ estatistico : F = \frac{\sigma_1^2}{\sigma_2^2}, \;\; sendo\ v = n - 1$$

$$C : \begin{cases} f > f_{\alpha/2, v_1, v_2} \\ f < f_{1-\alpha/2, v_1, v_2} \end{cases}$$

Exemplo: com os dados das amostras abaixo, teste H_0: $\sigma_1^2 = \sigma_2^2$ e H_1: $\sigma_1^2 \neq \sigma_2^2$. O que podemos concluir sobre a hipótese nula? Considere $\alpha = 0.05$.

Amostra 1				
19	23	19	29	23
23	25	22	26	24
20	21	26	22	25
20	23	24	24	27
16	30	20	18	20

Amostra 2			
21	22	23	24
30	21	20	23
21	21	21	28
23	25	24	21

$H_0: \sigma_1^2 = \sigma_2^2$ e $H_1: \sigma_1^2 \neq \sigma_2^2$, $\alpha = 0.05$

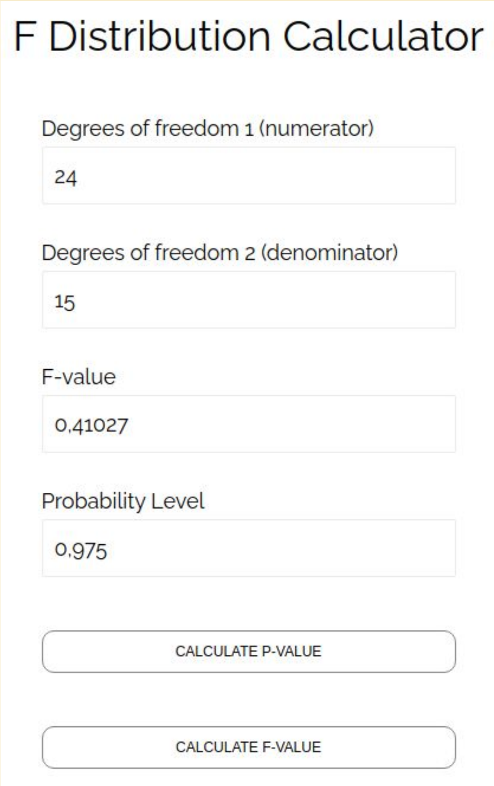
$$Teste\ estatistico: F = \frac{\sigma_1^2}{\sigma_2^2}$$

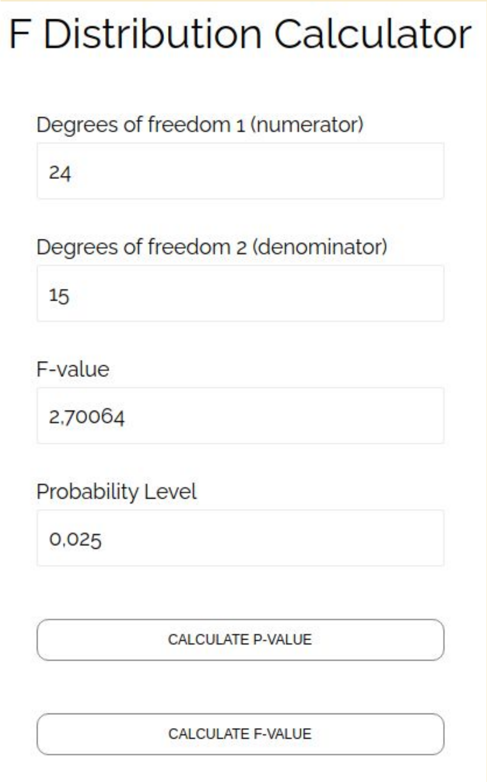
$$= \frac{11.0624}{7.125}$$

$$= 1.5526$$

$$f_{0.025,24,15} = 2.7006$$

$$f_{0.975,24,15} = 0.4102$$





$H_0: \sigma_1^2 = \sigma_2^2$ e $H_1: \sigma_1^2 \neq \sigma_2^2$, $\alpha = 0.05$

$$Teste\ estatistico: F = \frac{\sigma_1^2}{\sigma_2^2}$$

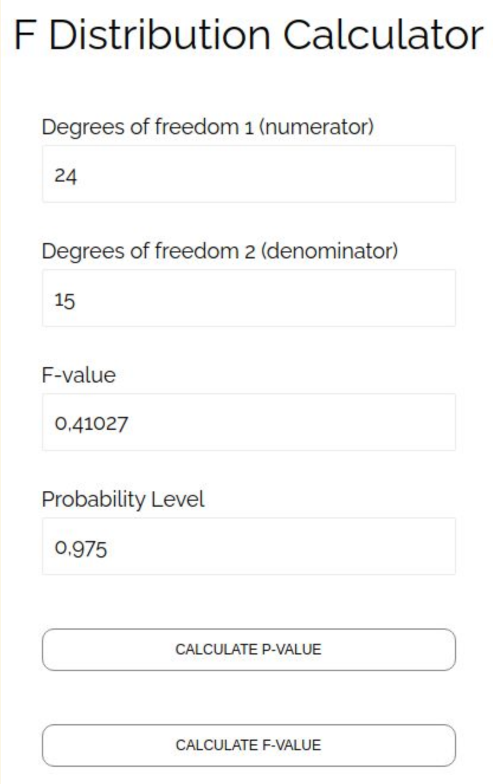
$$= \frac{11.0624}{7.125}$$

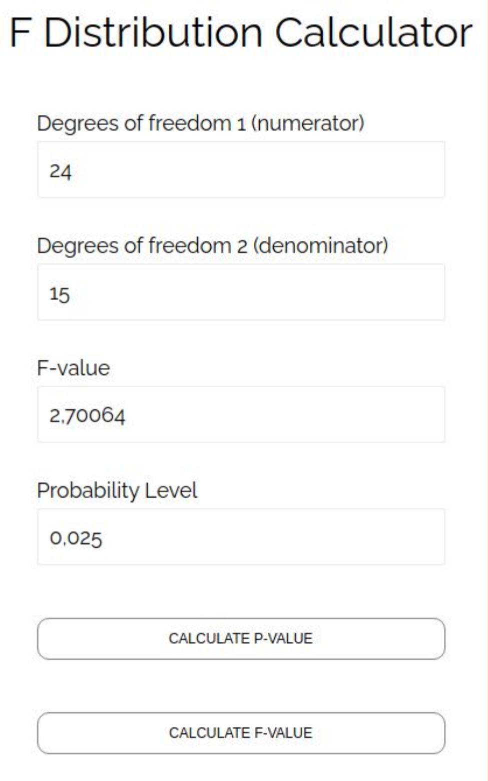
$$= 1.5526$$

$$f_{0.025,24,15} = 2.7006$$

$$f_{0.975,24,15} = 0.4102$$

Conclusão: F está fora da região crítica, logo **H_0 não pode ser rejeitada!**





Muito Obrigada!

Se você tiver qualquer dúvida ou sugestão:

- deborah.vm@ufpi.edu.br

UNIVERSIDADE FEDERAL DO PIAUÍ