

Probabilidade e Estatística

Estatística Descritiva!

O que é a estatística ?

Para muitos, a estatística não passa de conjuntos de tabelas de dados numéricos. Os estatísticos são pessoas que coletam esses dados.

- A estatística originou-se com a coleta e construção de tabelas de dados para os governos
- A situação evoluiu e esta coleta de dados representa somente um dos aspectos da estatística.

Definição de Estatística

A estatística é um conjunto de técnicas que permite, de forma sistemática, organizar, descrever, analisar e interpretar dados oriundos de estudos ou experimentos, realizados em qualquer área do conhecimento.

Áreas e aplicações da Estatística:

1. Estatística descritiva;
2. Probabilidade;
3. Inferência estatística;
4. Machine learning.

ESTATÍSTICA DESCRITIVA

A estatística descritiva é a etapa inicial da análise utilizada para descrever e resumir os dados. A disponibilidade de uma grande quantidade de dados e de métodos computacionais muito eficientes revigorou esta área da estatística.

PROBABILIDADE

A teoria de probabilidades nos permite descrever os fenômenos aleatórios, ou seja, aqueles em que está presente a incerteza.

INFERENCIA ESTATISTICA

E o estudo de técnicas que possibilitam a extrapolação, a um grande conjunto de dados, das informações e conclusões obtidas a partir da amostra.

Machine learning :

Método de análise de dados da área da Inteligência Artificial que automatiza a criação de modelos analíticos. Por meio de algoritmos que aprendem a partir de diversas bases de dados e de experiências acumuladas, o Machine Learning possibilita a predição e o aprendizado de certos padrões e comportamentos automaticamente, sem a intervenção humana.

Quando falamos em Machine Learning, estamos abordando também a estatística. Isso porque o Aprendizado de Máquina só pôde ser criado graças à ampla variedade de técnicas estatísticas desenvolvidas nos últimos tempos.

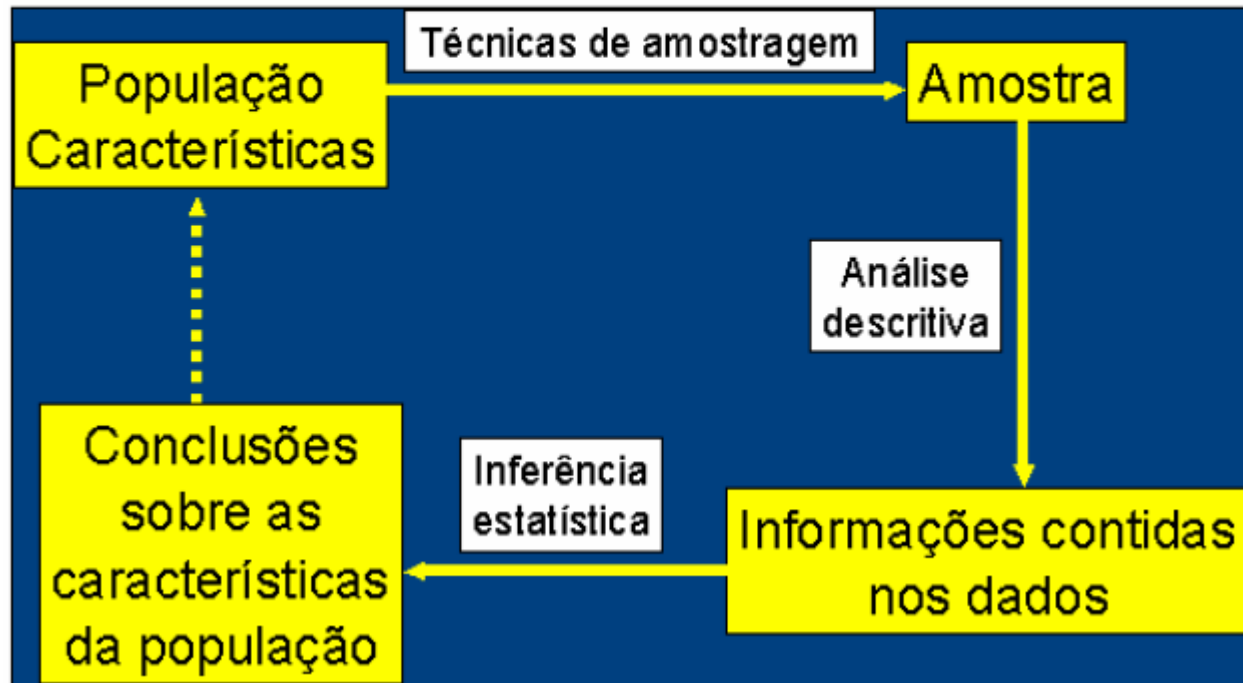
Machine learning :

Técnicas estatísticas que são geralmente aplicadas em cada tipo de aprendizado:

- Aprendizado Supervisionado: Árvores de Classificação, Suport Vector Machine (SVM), classificação (Regressão Logística, KNN-K vizinhos mais próximos), regressão (Regressão Linear, Splines, Árvores de Regressão, Redes Neurais);
- Aprendizado Não Supervisionado: redução de dimensionalidade (Análise de Componentes Principais, Escalonamento Multidimensional), análise de agrupamento (K-médias, Métodos Hierárquicos), regras de associação, sistemas de recomendação.

Graças ao aumento do poder de processamento dos computadores, as abordagens estatísticas e matemáticas evoluíram ao ponto de permitir o uso em Machine Learning.

Etapas da Análise Estatística

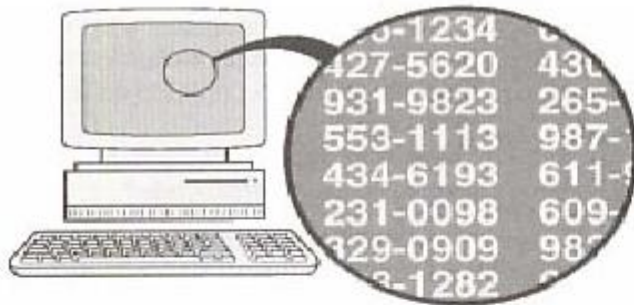


AMOSTRAGEM

Uma área importante em muitas aplicações Estatísticas é a da **Tecnologia de Amostragem**.

Exemplos de Aplicação:

- Pesquisa de mercado,
- Pesquisa de opinião,
- Avaliação do processo de produção,
- Praticamente em todo experimento.



Amostragem Aleatória

Cada elemento da população tem a mesma chance de ser escolhido.



Amostragem Estratificada

Classificar a população em, ao menos dois estratos e extrair uma amostra de cada um.



Amostragem Sistemática

Escolher cada elemento de ordem k .



Amostragem por Conglomerados

Dividir em seções a área populacional, selecionar aleatoriamente algumas dessas seções e tomar todos os elementos das mesmas.



Amostragem de Conveniência

Utilizar resultados de fácil acesso.



Exemplo 1

Numa pesquisa eleitoral, um instituto de pesquisa procura, com base nos resultados de um levantamento aplicado a uma amostra da população, prever o resultado da eleição.

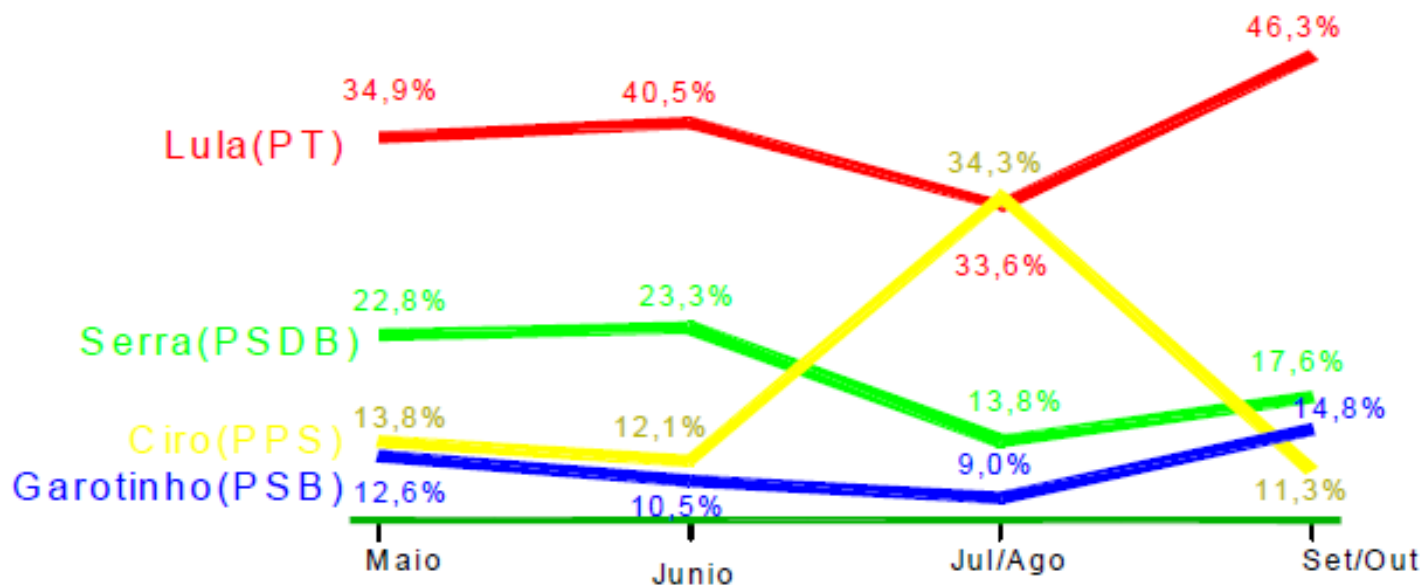
Na eleição Presidencial

Os Institutos de Pesquisa de opinião colhem periodicamente amostras de eleitores para obter as estimativas de intenção de voto da população. As estimativas são fornecidas com um valor e uma margem de erro.

O quadro do **Instituto Toledo & Associados**, a seguir refere-se à intenção de voto no 1º turno das eleições para o governo em 2002.

Intenção de voto para presidente do Brasil-2002

Voto estimulado, em % do total de votos. A última pesquisa ouviu 2.202 eleitores- Margem de erro de 2,09%



Confronto no segundo turno.

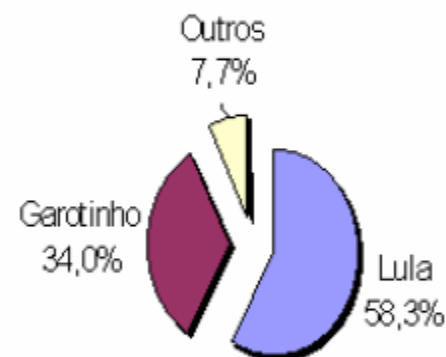
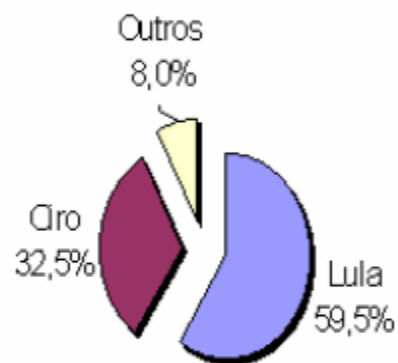
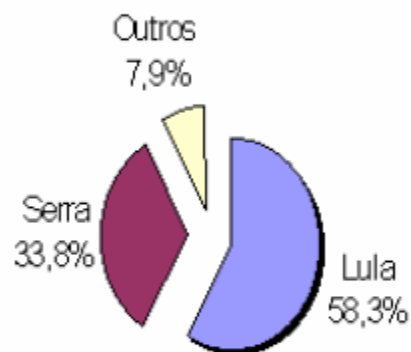


Gráfico de setores ou em forma de pizza

Tabela 1.1 Informação do estado civil, grau de instrução, número de filhos, idade e procedência de 36 funcionários sorteados ao acaso da empresa MB.(Bussab e Morettin)

Nº	Estado Civil	Grau de Instrução	No de filhos	Salário (X Sal. Min)	Idade anos meses	Região de procedência	
1	Solteiro	1º grau	-	4,00	26 03	Interior	
2	Casado	1º grau	1	4,56	32 10	Capital	
3	Casado	1º grau	2	5,25	36 05	Capital	
4	Solteiro	2º grau	-	5,73	20 10	Outro	
5	Solteiro	1º grau	-	6,26	40 07	Outro	
6	Casado	1º grau	0	6,66	28 00	Interior	
7	Solteiro	1º grau	-	6,86	41 00	Interior	
8	Solteiro	1º grau	-	7,39	43 04	Capital	
9	Casado	2º grau	1	7,59	34 10	Capital	
10	Solteiro	2º grau	-	7,44	23 06	Outro	
11	Casado	2º grau	2	8,12	33 06	Interior	
12	Solteiro	1º grau	-	8,46	27 11	Capital	
13	Solteiro	2º grau	-	8,74	37 05	Outro	
14	Casado	1º grau	3	8,95	44 02	Outro	
15	Casado	2º grau	0	9,13	30 05	Interior	
16	Solteiro	2º grau	-	9,35	38 08	Outro	
17	Casado	2º grau	1	9,77	31 07	Capital	
18	Casado	1º grau	2	9,80	39 07	Outro	
19	Solteiro	Superior	-	10,53	25 08	Interior	
20	Solteiro	2º grau	-	10,76	37 04	Interior	
21	Casado	2º grau	1	11,06	30 09	Outro	
22	Solteiro	2º grau	-	11,59	34 02	Capital	
23	Solteiro	1º grau	-	12,00	41 00	Outro	
24	Casado	Superior	0	12,79	26 01	Outro	
25	Casado	2º grau	2	13,23	32 05	Interior	
26	Casado	2º grau	2	13,60	35 00	Outro	
27	Solteiro	1º grau	-	13,85	46 07	Outro	
28	Casado	2º grau	0	14,69	29 08	Interior	
29	Casado	2º grau	5	14,71	40 06	Interior	
30	Casado	2º grau	2	15,99	35 10	Capital	
31	Solteiro	Superior	-	16,22	31 05	Outro	
32	Casado	2º grau	1	16,61	36 04	Interior	
33	Casado	Superior	3	17,26	43 07	Capital	
34	Solteiro	Superior	-	18,75	33 07	Capital	
35	Casado	2º grau	2	19,40	48 11	Capital	
36	Casado	Superior	3	23,30	42 02	Interior	

Estatística Descritiva

O que fazer com as observações que coletamos?



Primeira Etapa:

Resumo dos dados = Estatística descritiva

Variável

Qualquer característica associada a uma população

Classificação de variáveis

Qualitativa

Nominal

sexo, cor dos olhos

Ordinal

Classe social, grau de instrução

Quantitativa

Contínua

Peso, altura, salario

Discreta

Número de filhos, numero de carros

Medidas Resumo

Variáveis Quantitativas

MEDIDAS DE POSIÇÃO: Moda, Média, Mediana, Percentís, Quartis.


MEDIDAS DE DISPERSÃO: Amplitude, Intervalo-Interquartil, Variância, Desvio Padrão, Coeficiente de Variação.

Medidas de Posição

Moda(m_o): É o valor (ou atributo) que ocorre com maior frequência.

Ex: 4,5,4,6,5,8,4,4

M_o = 4



Variavel
qualitativa

Média aritmética simples

A média aritmética simples é obtida dividindo a soma de todos os valores que temos pela quantidade de valores. Geralmente expressamos a média pelo símbolo \overline{X} .

Suponhamos que existam uma quantidade n de dados $(x_1, x_2, x_3, \dots, x_n)$. A média entre esses dados será:

$$\overline{X} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Exemplo:

Um aluno obteve as seguintes notas durante um bimestre: 9.2, 8.5 e 8.4. Qual será a média de suas notas?

Temos 3 notas. Basta somá-las e dividir este resultado por 3:

$$\overline{X} = \frac{9,2 + 8,5 + 8,4}{3} = \frac{26,1}{3} = 8,7$$

A média será 8.7

Média aritmética ponderada

A média ponderada considera “pesos” para cada item, ou seja, em um conjunto de dados, cada item recebe uma importância. Vamos supor que tenhamos um conjunto com n dados $(x_1, x_2, x_3, \dots, x_n)$, onde cada dado receberá um peso, respectivamente $(p_1, p_2, p_3, \dots, p_n)$.

Cada item será multiplicado pelo seu peso. A média será dada pela divisão entre esta soma e a soma dos pesos considerados. A média entre esses dados será representada por \overline{P} e será dada por:

$$\overline{P} = \frac{x_1p_1 + x_2p_2 + x_3p_3 + \dots + x_np_n}{p_1 + p_2 + p_3 + \dots + p_n}$$

Exemplo: Uma aluna fez uma prova e obteve nota 9.1 e um trabalho, com nota 8,7. A média considera que a prova tenha peso 6 e o trabalho peso 4. Assim, a média dessa aluna será:

$$\overline{P} = \frac{9,1 \cdot 6 + 8,7 \cdot 4}{6 + 4} = \frac{54,6 + 34,8}{10} = \frac{89,4}{10} = 8,94$$

A média dessa aluna será 8,94.

Média geométrica

A média geométrica entre um conjunto de n dados é a raiz n -ésima da multiplicação desses dados.

Considere um conjunto de n dados $(x_1, x_2, x_3, \dots, x_n)$. A média geométrica entre estes dados será:

$$\overline{X} = \sqrt[n]{x_1 \cdot x_2 \cdot x_3 \cdot \dots \cdot x_n}$$

Exemplo. Qual a média geométrica entre 2, 8 e 32?

Temos três dados, então a média geométrica será a raiz cúbica de 2.8.32:

$$\overline{X} = \sqrt[3]{2 \cdot 8 \cdot 32} = \sqrt[3]{512} = 8$$

A média geométrica de 2, 8 e 32 será igual a 8.

Média harmônica

A média harmônica de um conjunto de n dados é obtida dividindo a quantidade de dados pela soma dos inversos dos dados.

Considerando um conjunto de n dados $(x_1, x_2, x_3, \dots, x_n)$, a média harmônica entre esses dados, indicada por H , será:

$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n}}$$

Por exemplo, dado um conjunto A (2, 3, 5, 6, 9), como ele possui cinco elementos, a média harmônica de A é calculada por:

$$M_h = \frac{5}{\frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{6} + \frac{1}{9}}$$

$$M_h = \frac{450}{118} = 3,81$$

Mediana

A mediana é o valor da variável que ocupa a posição central de um conjunto de n dados ordenados.

Posição da mediana: $(n+1)/2$

Ex: 2,5,3,7,8

Dados ordenados: 2,3,5,7,8 $\Rightarrow (5+1)/2=3$

$\Rightarrow Md = 5$

Ex: 3,5,2,1,8,6

Dados ordenados: 1,2,3,5,6,8 \Rightarrow

$(6+1)/2=3,5 \Rightarrow Md=(3+5)/2=4$

Variância

$$S^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Desvio padrão **S**

$$\text{Desvio Padrão : } S = \sqrt{\text{Variância}}$$

Cálculo da variância para o grupo 1:

G1:3, 4, 5, 6, 7: Vimos que: $\bar{x} = 5$

$$S^2 = \frac{(3-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2}{5-1} = \frac{10}{4} = 2,5$$

Desvio padrão $S = \sqrt{2,5} = 1,58$

$$G1 : S^2 = 2,5 \quad S = 1,58$$

$$G2 : S^2 = 10 \quad S = 3,16$$

$$G3 : S^2 = 0 \quad S = 0$$

Coeficiente de Variação (CV)

- É uma medida de dispersão relativa;
- Elimina o efeito da magnitude dos dados;
- Exprime a variabilidade em relação a média
- Útil Comparar duas ou mais variáveis

$$CV = \frac{S}{\bar{X}} \times 100 \%$$

Exemplo 4: Altura e peso de alunos

	Média	Desvio padrão	Coefficiente de variação
Altura	1,143m	0,063m	5,5%
Peso	50Kg	6kg	12%

Conclusão: Com relação as médias, os alunos são, aproximadamente, duas vezes mais dispersos quanto ao peso do que quanto a altura

ORGANIZAÇÃO E REPRESENTAÇÃO DOS DADOS

Uma das formas de organizar e resumir a informação contida em dados observados é por meio de tabela de frequências e gráficos.

Tabela de frequência relaciona categorias (ou classes) de valores, juntamente com contagem (ou frequências) do número de valores que se enquadram em cada categoria ou classe.

1. Variáveis qualitativas: Podemos construir tabela de frequência que os quantificam por categoria de classificação e sua representação gráfica é mediante gráfico de barras, gráfico setorial ou em forma de pizza.

Exemplo 1: Considere a variável grau de Instrução dos dados da tabela 1. (Variável qualitativa)

Tabela de frequência

Grau de instrução	Contagem	f_i	f_{r_i}	$f_{r_i} \%$
1o Grau		12	0,3333	33,3%
2o Grau		18	0,5000	50 %
Superior		6	0,1667	16.7%
total		n=36	1,0000	100%

f_i : **Frequência absoluta** da categoria i (número de indivíduos que pertencem à categoria i)

$f_{r_i} = \frac{f_i}{n}$: **Frequência relativa** da categoria i

$f_{r_i} \% = f_{r_i} * 100\%$: **Frequência relativa percentual** da categoria i

Representação gráfica de variáveis qualitativas

- Gráfico de Barras
- Diagrama circular, de sectores ou em forma de "pizza"

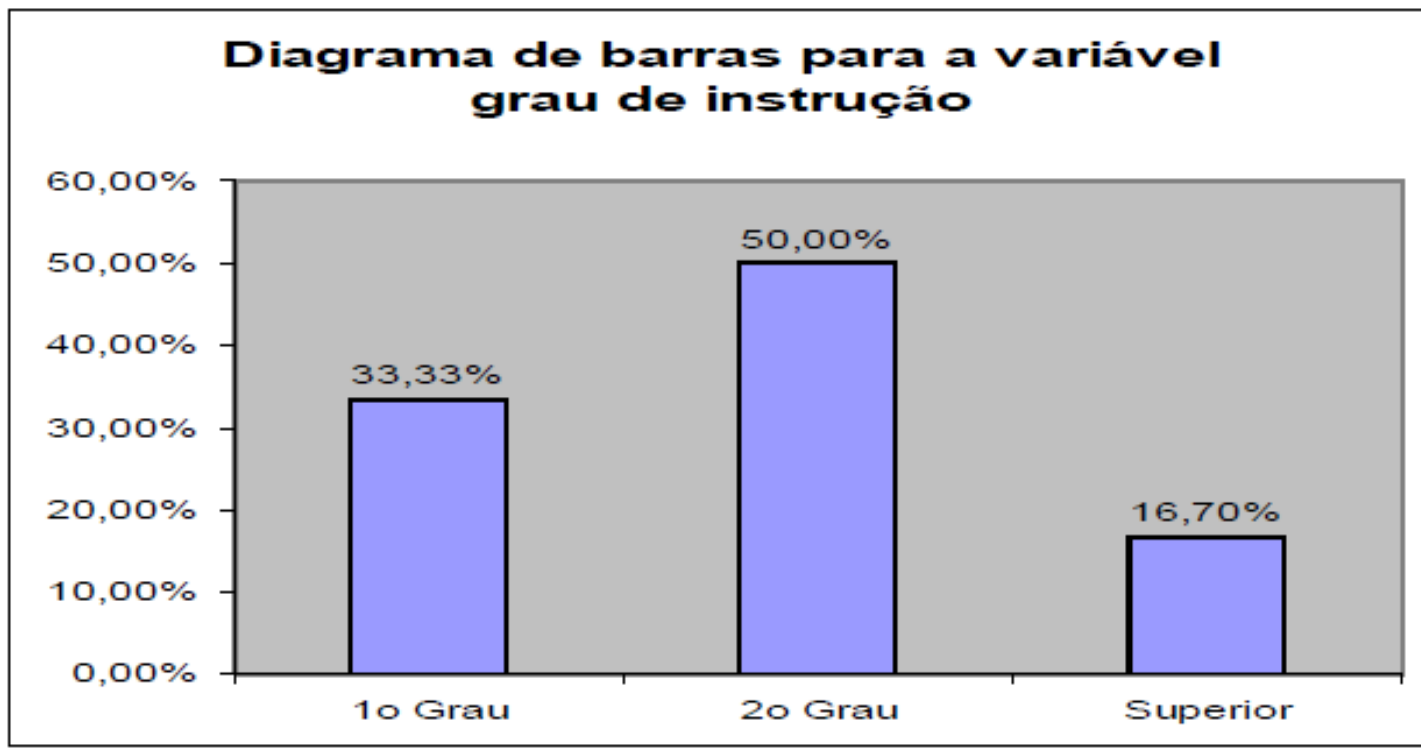


Diagrama circular para a variável grau de instrução

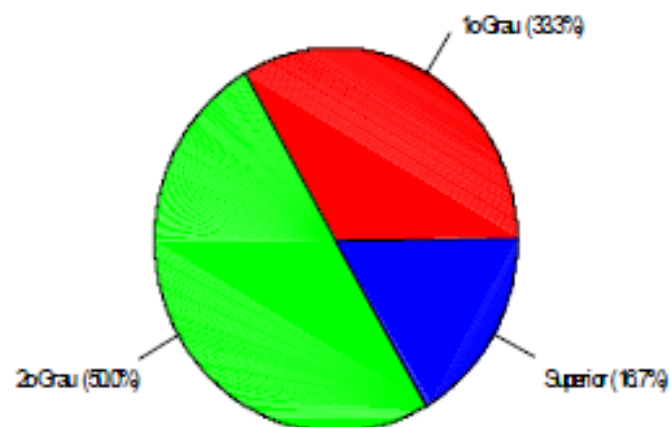


Diagrama circular para a variável grau de instrução

