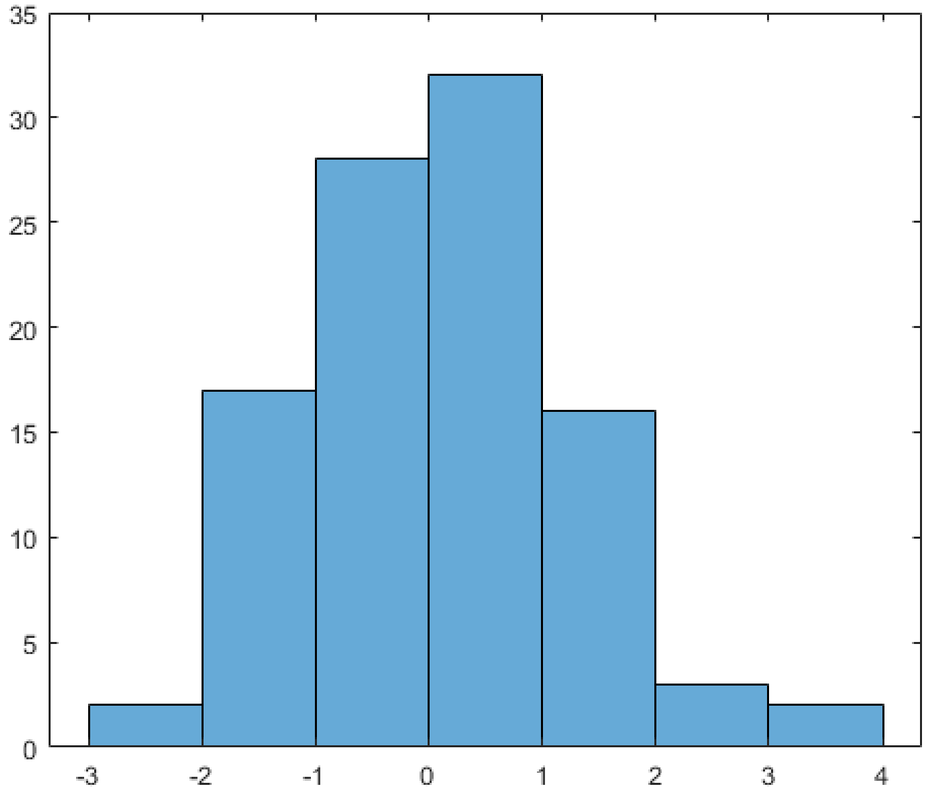
Análise Estatística para Data Science I com R e SAS

Análise Estatística para Data Science I com R e SAS

Visualizando e Descrevendo Dados Quantitativos

Data Science Academy

Visualizando e Descrevendo Dados Quantitativos



Na maior parte do tempo estaremos trabalhando com dados quantitativos e precisamos portanto saber como visualizá-los e descrevê-los.

Análise Estatística para Data Science I com R e SAS

Definindo Variáveis Quantitativas Discretas e Contínuas

Data Science Academy

Definindo Variáveis Quantitativas Discretas e Contínuas

Variável é a característica de interesse que é medida em cada elemento da amostra ou população. Como o nome diz, seus valores variam de elemento para elemento. As variáveis podem ter valores numéricos ou não numéricos.

Definindo Variáveis Quantitativas Discretas e Contínuas

Definindo Variáveis Quantitativas Discretas e Contínuas

Variáveis Quantitativas: são as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser discretas ou contínuas.

- **Discretas**: características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros. Geralmente são o resultado de contagens. Exemplos: número de filhos, número de carros, número de acessos.

Definindo Variáveis Quantitativas Discretas e Contínuas

Variáveis Quantitativas: são as características que podem ser medidas em uma escala quantitativa, ou seja, apresentam valores numéricos que fazem sentido. Podem ser discretas ou contínuas.

- **Contínuas**: características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores fracionais fazem sentido. Usualmente devem ser medidas através de algum instrumento. Exemplos: altura (régua), peso (balança), tempo (relógio), pressão arterial, salário.

Análise Estatística para Data Science I com R e SAS

Medidas de Tendência Central

Data Science Academy

Medidas de Tendência Central

A Estatística trabalha com diversas informações que são apresentadas por meio de gráficos e tabelas e com diversos números que representam e caracterizam um determinado conjunto de dados.

Medidas de Tendência Central

A Estatística trabalha com diversas informações que são apresentadas por meio de gráficos e tabelas e com diversos números que representam e caracterizam um determinado conjunto de dados.

Dentre todas as informações, podemos retirar valores que representem, de algum modo, todo o conjunto. Esses valores são denominados:

Medidas de Tendência Central

A Estatística trabalha com diversas informações que são apresentadas por meio de gráficos e tabelas e com diversos números que representam e caracterizam um determinado conjunto de dados.

Dentre todas as informações, podemos retirar valores que representem, de algum modo, todo o conjunto. Esses valores são denominados:

“Medidas de Tendência Central ou Medidas de Centralidade”.

Medidas de Tendência Central

As Medidas de Centralidade mais comuns são a Média Aritmética, a Mediana e a Moda.

Medidas de Tendência Central

Média

É uma das medidas de tendência central mais utilizadas no cotidiano.

É determinada pelo resultado da divisão do somatório dos números dados pela quantidade de números somados.

Por exemplo, vamos determinar a média dos números: 3, 12, 23, 15 e 2.

Para isso basta somarmos todos os números e dividirmos pela quantidade de números, ou seja:

$$\textbf{Média Aritmética} = \frac{3 + 12 + 23 + 15 + 2}{5} = 11$$

Medidas de Tendência Central

É a medida de tendência central que indica exatamente o valor central de um conjunto de dados quando organizados em ordem crescente ou decrescente.

Por exemplo, vamos considerar que um aluno tirou as seguintes notas em cinco provas de uma determinada disciplina: 5, 8, 7, 4 e 9.

Colocando as cinco notas em ordem crescente, por exemplo, obtemos: $4 < 5 < 7 < 8 < 9$.

A mediana é o valor que está no centro dessa sequência, ou seja, 7.

Medidas de Tendência Central

Mas se ao invés de cinco notas fossem seis?

Pois bem, nesse caso ao ordenarmos os números, teremos dois termos centrais ao invés de um. Por exemplo, digamos que as notas agora são: 5, 2 , 8, 7, 4 e 9.

Colocando em ordem crescente, temos: 2 < 4 < 5 < 7 < 8 < 9.

Aqui, os dois termos centrais seriam 5 e 7. Portanto, a Mediana desse conjunto de dados é a Média Aritmética dos dois termos centrais, ou seja:

$$\textbf{Mediana} = \frac{5+7}{2} = 6$$

Medidas de Tendência Central

É a medida de tendência central que consiste no valor observado com mais frequência em um conjunto de dados.

Por exemplo, digamos que o time do Barcelona em determinado torneio de futebol fez, em dez partidas, a seguinte quantidade de gols: 5, 4, 2, 1, 3, 7, 1, 1, 2 e 1.

Para essa sequência de gols marcados, a moda é de 1 gol, pois é o número que aparece mais vezes.

Outra situação comum seria se dentre 7 pessoas tomássemos suas idades, sendo: 15 anos, 20 anos, 32 anos, 13 anos, 5 anos, 43 anos e 90 anos.

Nesse caso, não há moda, pois nenhuma idade se repetiu mais vezes que a outra.

Medidas de Tendência Central

Quando um conjunto de dados não apresenta moda, dizemos que esse conjunto é **amodal**.

Caso exista uma moda, denominamos o conjunto de **Unimodal**.

Existindo duas modas, denominamos o conjunto de **bimodal** e assim sucessivamente.

Medidas de Tendência Central

Outras Medidas de Tendência Central

Medidas de Tendência Central

Medidas	Fórmula
Média aritmética	$\frac{x_1+\ldots x_n}{n}$
Média aritmética para dados agrupados	$\frac{f_1.x_1\ldots f_k.x_k}{f_1\ldots f_k}$
Média aritmética ponderada	$\frac{P_1.(X_1)+P_2.(X_2)}{P_{total}}$
Mediana	1) Se *n* é impar, o valor é central, 2) se *n* é par, o valor é a média dos dois valores centrais
Moda	Valor que ocorre com mais frequência.
Média geométrica	$G=\sqrt[n]{X_1.X}$
Média harmônica	$\frac{n}{\frac{1}{x_1}}+\frac{1}{x_1}$
Quartil	$Q=[p(sup)-0,25].x(inf)+[0,25-p(inf)]x\frac{(sup)}{p(sup)}-p(inf)$

Análise Estatística para Data Science I com R e SAS

Medidas de Dispersão

Data Science Academy

Medidas de Dispersão

Em Estatística, existem algumas medidas que servem para representar todo um conjunto de informações a partir dos dados, como média, mediana e moda.

Medidas de Dispersão

Em Estatística, existem algumas medidas que servem para representar todo um conjunto de informações a partir dos dados, como média, mediana e moda.

Existem ainda outras medidas responsáveis por ilustrar o grau de variação entre as informações do conjunto. São elas: amplitude, desvio, variância e desvio padrão.

Medidas de Dispersão

Em Estatística, existem algumas medidas que servem para representar todo um conjunto de informações a partir dos dados, como média, mediana e moda.

Existem ainda outras medidas responsáveis por ilustrar o grau de variação entre as informações do conjunto. São elas: amplitude, desvio, variância e desvio padrão.

Essas últimas são chamadas medidas de dispersão.

Medidas de Dispersão

A média é uma medida de tendência central usada para representar os valores de um conjunto utilizando apenas um número.

No entanto, imagine que o diretor de uma escola precisa descobrir qual turma teve melhor desempenho durante o ano letivo, cada qual com professores diferentes. Perceba que a média das duas turmas é a mesma 7,2, o que dificulta a análise.

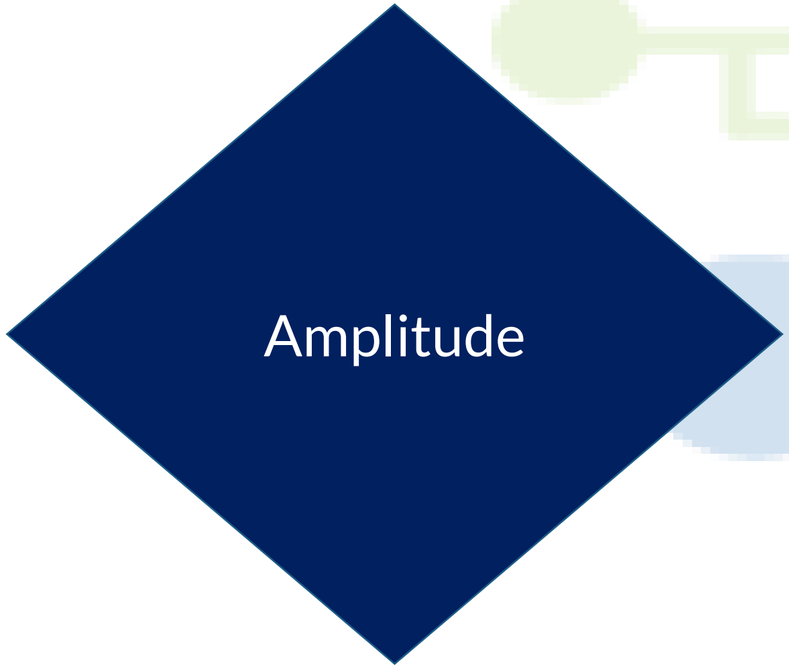
Podemos usar as medidas de dispersão, que indicam o quão distante está cada uma das notas desses alunos da média obtida.

Sala 1			Sala 2	
Aluno	Nota		Aluno	Nota
1	7		1	7
2	5		2	8
3	8		3	7,5
4	9		4	9
5	6		5	7
6	7		6	6
7	9		7	8
8	10		8	6
9	9,5		9	6,5
10	5		10	6
11	5,5		11	7
12	6		12	8,5
13	4,5		13	7
14	8		14	7,5
15	8,5		15	7
Média	7,2			7,2

Medidas de Dispersão

Amplitude, Desvio, Variância e Desvio Padrão

Medidas de Dispersão



Em um conjunto de informações numéricas, a primeira medida de dispersão é chamada amplitude e é obtida a partir da diferença entre a maior informação da lista e a menor.

Por exemplo, considere as notas de dois alunos na disciplina de Estatística:

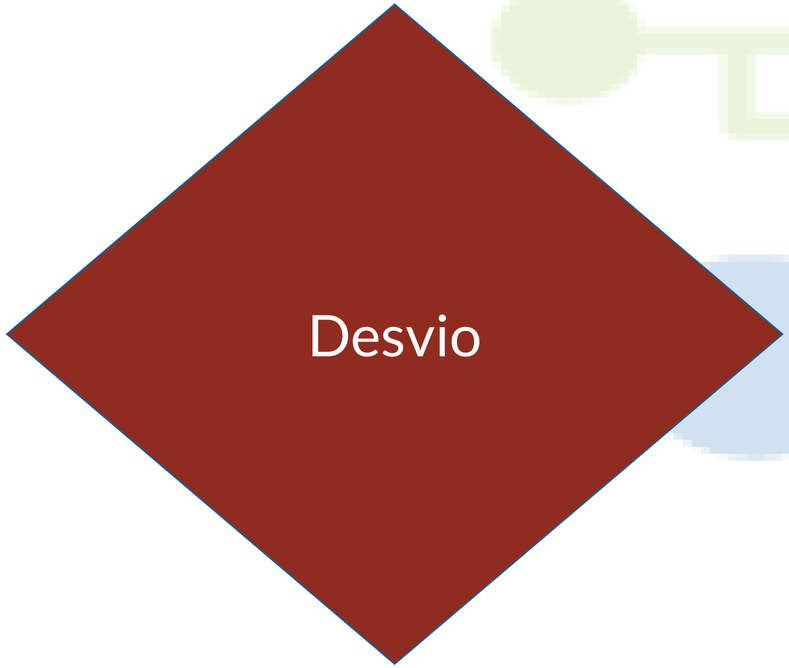
João: 6,5; 6,5; 6,0 e 5,0
Maria: 1,0; 4,0; 9,0 e 10,0

A média dos dois alunos é 6,0, mas observe a amplitude das notas deles:

João: Média 6,0; amplitude = 6,5 – 5,5 = 1,0

Maria: Média 6,0; amplitude = 10,0 – 1,0 = 9,0

Medidas de Dispersão



Em um conjunto de informações numéricas, o desvio é a "distância" de cada uma dessas informações até a média aritmética delas. Matematicamente, o desvio é obtido subtraindo cada um dos valores de um conjunto de informações da média aritmética desse conjunto.

Assim, os desvios devem ser calculados para cada elemento desse conjunto. No exemplo anterior, seriam quatro desvios para o primeiro aluno e outros quatro desvios para o segundo aluno. Por exemplo:

Notas do João: 6,5 6,5; 6,0 e 5,0. Média: 6,0. Desvios:

d1 = 6,5 – 6,0 = 0,5
d2 = 6,5 – 6,0 = 0,5
d3 = 6,0 – 6,0 = 0
d4 = 5,0 – 6,0 = – 1,0

Observe que o sinal nos desvios é importante. É ele que determina, por meio do desvio, se a nota tirada é maior ou menor que a média.

Medidas de Dispersão

Variância

Define-se a Variância, como a medida que se obtém somando os quadrados dos desvios das observações da amostra, relativamente à sua média, e dividindo pelo número de observações da amostra menos um.

A Variância é uma média das distâncias calculada a partir dos quadrados dos desvios em relação à média aritmética simples. Ou seja, calculamos a diferença entre cada indivíduo da amostra e a média aritmética simples e elevamos ao quadrado. Em seguida, somamos todos os valores obtidos e dividimos pelo tamanho da amostra menos uma unidade. A fórmula matemática da variância é:

$$s^2 = \frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}$$

A rigor, o denominador desta expressão deveria ser n. Entretanto, por razões relacionadas à inferência estatística, pode-se mostrar que é conveniente dividir a soma dos quadrados das diferenças por n – 1.

Medidas de Dispersão

Desvio Padrão

É a raiz quadrada da Variância. O Desvio Padrão possui a mesma unidade de medida dos dados e é a medida que, efetivamente, é utilizada como síntese da dispersão ou variabilidade. É essa medida que mede a concentração dos valores dos indivíduos da amostra em relação à média aritmética simples.

Em outros termos, pode-se dizer que quanto menor for o desvio padrão, mais representativa é a média aritmética simples; pois, neste caso, a baixa dispersão indica que a maioria das medidas dos indivíduos da amostra estão razoavelmente próximas da média e, portanto, esta representa bem o conjunto de dados. A fórmula do desvio padrão é:

$$s = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}}$$

Análise Estatística para Data Science I com R e SAS

Medidas de Posição Relativa – Quartis e Percentis

Data Science Academy

Medidas de Posição Relativa – Quartis e Percentis

Separatrizes separam a distribuição em partes percentualmente iguais.

As mais utilizadas são: Quartis e Percentis

Medidas de Posição Relativa – Quartis e Percentis

Quartis

São valores que dividem um conjunto de elementos ordenados em quatro partes iguais, ou seja, cada parte contém 25% desses elementos.

Há, portanto, três quartis: Q1, Q2 e Q3.

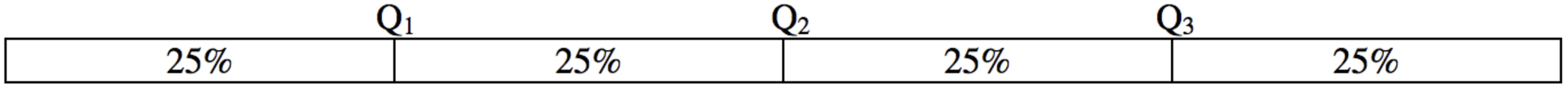
- Q1 – é chamado de primeiro quartil, ou seja, valor que deixa 25% dos elementos à sua esquerda e 75% dos elementos à sua direita. Q1 significa um quarto.

- Q2 – é chamado de segundo quartil e coincide com a mediana (Q2 = Md), ou seja, 50% dos elementos estão à sua esquerda e 50% à sua direita.

- Q3 – é chamado de terceiro quartil, ou seja, valor que deixa 75% dos elementos à sua esquerda e 25% à sua direita. Q3 significa três quartos.

Quando os dados são agrupados para determinar os quartis, usamos a mesma técnica do cálculo da mediana.

Medidas de Posição Relativa – Quartis e Percentis



Medidas de Posição Relativa – Quartis e Percentis

Percentis

Denominamos percentis os noventa e nove valores que separam uma série em 100 partes iguais. A notação que usaremos para os percentis será Pi, onde o índice i indica a ordem do percentil considerado. Podemos também conceituar como sendo a medida que divide a amostra em 100 partes iguais.

Exemplo:

- P10 indica que 10% dos dados estão ordenados à sua esquerda e 90% à direita de P10.
- P20 indica que 20% dos dados estão ordenados à sua esquerda e 80% à sua direita.

Medidas de Posição Relativa – Quartis e Percentis

Percentis

Importante: Não confundir percentis com percentagens. Um percentil é relacionado somente com a posição relativa de uma observação quando comparada com os outros valores. Desse modo se um estudante que acerta 75% de um teste, mas cuja nota é o percentil 40, significa que somente 40%, da turma tiveram nota pior que aquele estudante e 60% saíram-se melhor.

Percentis são válidos apenas para dados ordinais, intervalores e proporcionais, se usado com dados ordinais, deve-se ter cuidado na hora de interpolar entre dois valores, porque estes não estão bem definidos.

Análise Estatística para Data Science I com R e SAS

Gráficos Para Variáveis Quantitativas

Data Science Academy

Gráficos Para Variáveis Quantitativas

Para variáveis qualitativas usamos normalmente esses dois tipos de gráficos:

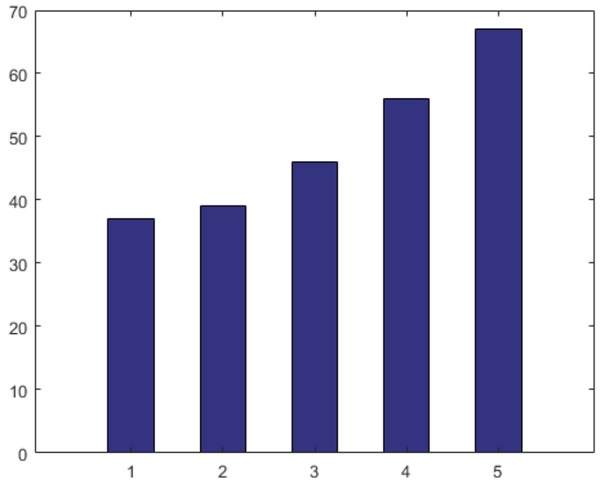


Gráfico de Barras

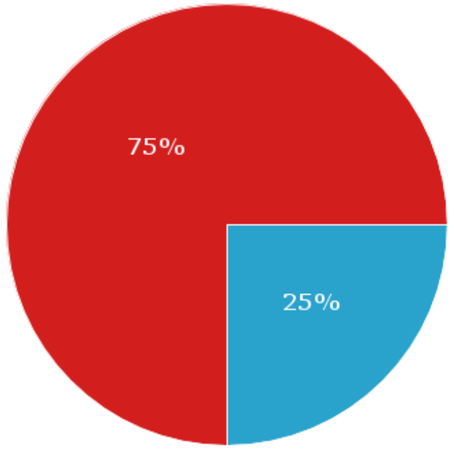
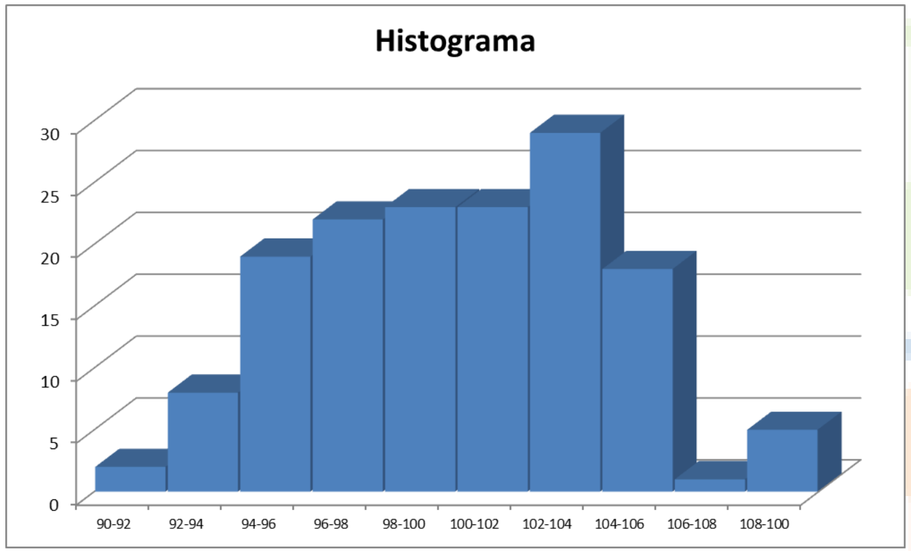


Gráfico de Pizza

Gráficos Para Variáveis Quantitativas

E quais tipos de gráficos usamos para variáveis quantitativas?

Gráficos Para Variáveis Quantitativas



Histograma

Um histograma é formado por um conjunto de barras justapostas, cujas bases se localizam sobre o eixo horizontal, de tal modo que seus pontos médios coincidem com os pontos médios dos intervalos de classe. O número de barras é igual ao número de classes da distribuição.

Gráficos Para Variáveis Quantitativas

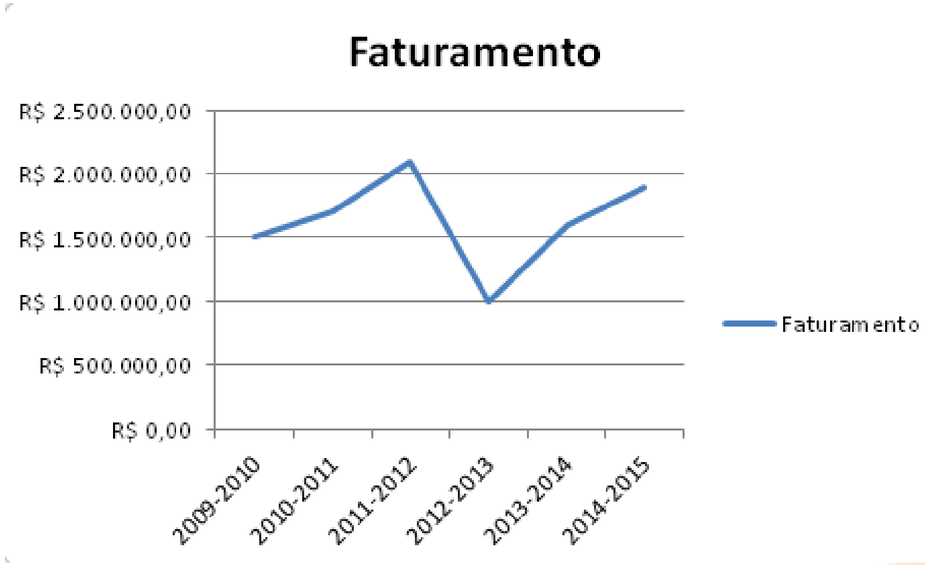
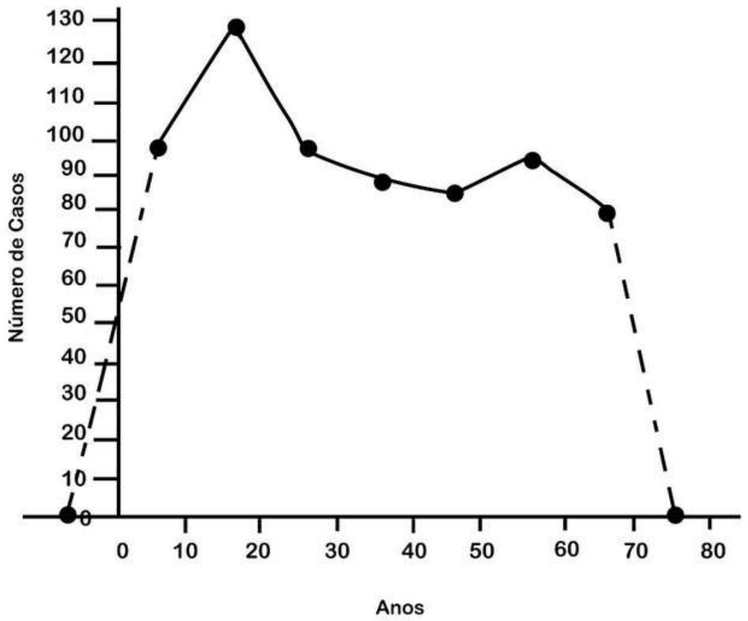


Gráfico de Linha

O gráfico de linha consiste em uma linha poligonal para representar uma série estatística.

As séries são comumente conhecidas como séries temporais, cronológicas ou históricas. Essas séries são um conjunto de observações de uma mesma variável quantitativa (discreta ou contínua) obtidas ao longo de um período de tempo. Assim esse gráfico representa a aplicação de funções em um sistema de coordenadas cartesianas.

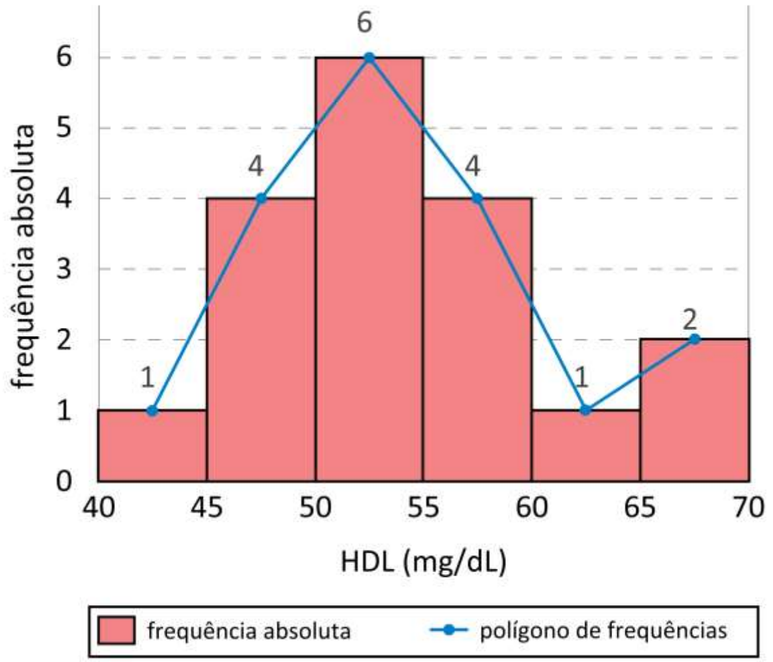
Gráficos Para Variáveis Quantitativas



Polígono de Frequência

O polígono de frequência é um gráfico em linha e as frequências são marcadas sobre perpendiculares ao eixo horizontal, levantadas pelos pontos médios dos intervalos de classe.

Gráficos Para Variáveis Quantitativas

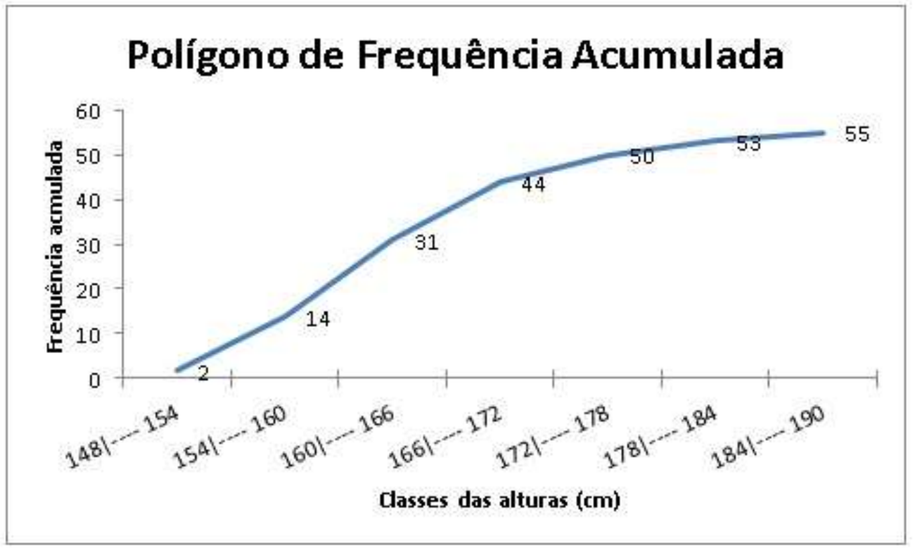


Polígono de Frequência

O polígono de frequência é um gráfico em linha e as frequências são marcadas sobre perpendiculares ao eixo horizontal, levantadas pelos pontos médios dos intervalos de classe.

A partir de um histograma, a construção de um polígono torna-se ainda mais fácil, pois é preciso somente encontrar o ponto médio do topo de cada retângulo e realizar uma pequena marcação ou um ponto. Em seguida, desenha-se uma reta reunindo os pontos identificados.

Gráficos Para Variáveis Quantitativas



Polígono de Frequência Acumulada

O polígono de frequência acumulada, também conhecido como Ogiva de Galton, é traçado marcando-se as frequências acumuladas sobre perpendiculares ao eixo horizontal, levantadas nos pontos correspondentes aos limites superiores dos intervalos de classe.

Para construir esse tipo de gráfico, portanto, é preciso ter em mãos as frequências acumuladas das classes. Em cada classe, fazemos uma marcação (ponto) do valor da frequência e, depois, a cada dois pontos, unimos com uma reta.

Gráficos Para Variáveis Quantitativas

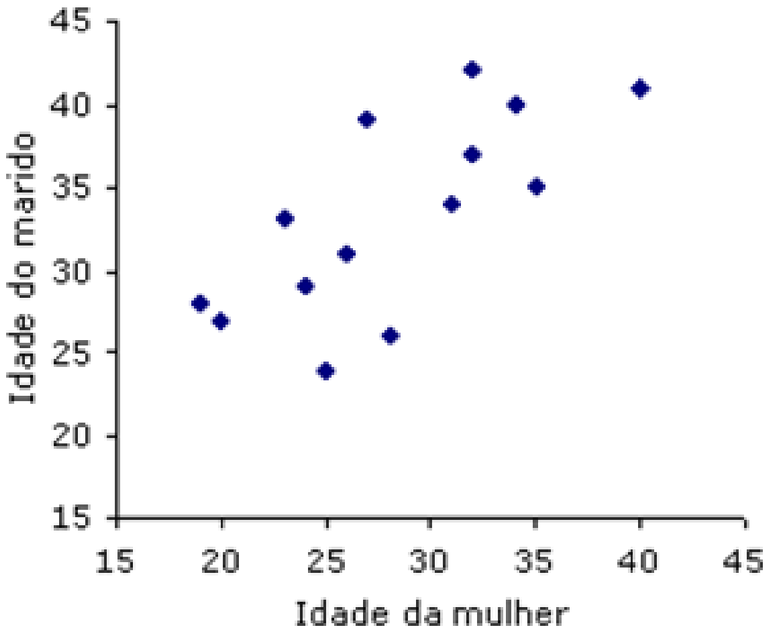


Gráfico de Dispersão

O gráfico de dispersão, frequentemente utilizado para fins científicos, é uma representação da variação de uma variável quantitativa em função de outra variável.

Assim, a partir de um gráfico de dispersão, é possível analisar a relação entre duas variáveis quantitativas por meio da plotagem dos valores dessas variáveis em um sistema de eixos. O eixo x é destinado para uma variável e o y para a outra variável.

Gráficos Para Variáveis Quantitativas

Dicas Para Escolha do Gráfico

Variáveis Quantitativas Discretas	Variáveis Quantitativas Contínuas
Se a variável for discreta é costume usar diagrama de barras (como para variáveis qualitativas). Tipicamente: • Barras são horizontais para variáveis qualitativas. • Barras são verticais para variáveis quantitativas. Tanto se pode fazer o diagrama de barras usando frequências absolutas, como frequências relativas, como percentagens. Para as variáveis quantitativas discretas (ou qualitativas com escala ordinal) faz sentido calcular as frequências acumuladas e podem representar-se usando diagrama de dispersão.	Dados quantitativos (variável contínua) • Histograma • Polígonos de Frequência • Polígonos de Frequência Acumulada Para séries, usamos: • Gráfico de Linha • Gráfico de Dispersão

É um prazer ter você aqui!

Muito Obrigado!

Pela Confiança em Nosso Trabalho.

Continue Trilhando Uma Excelente Jornada de Aprendizagem!

Data Science Academy