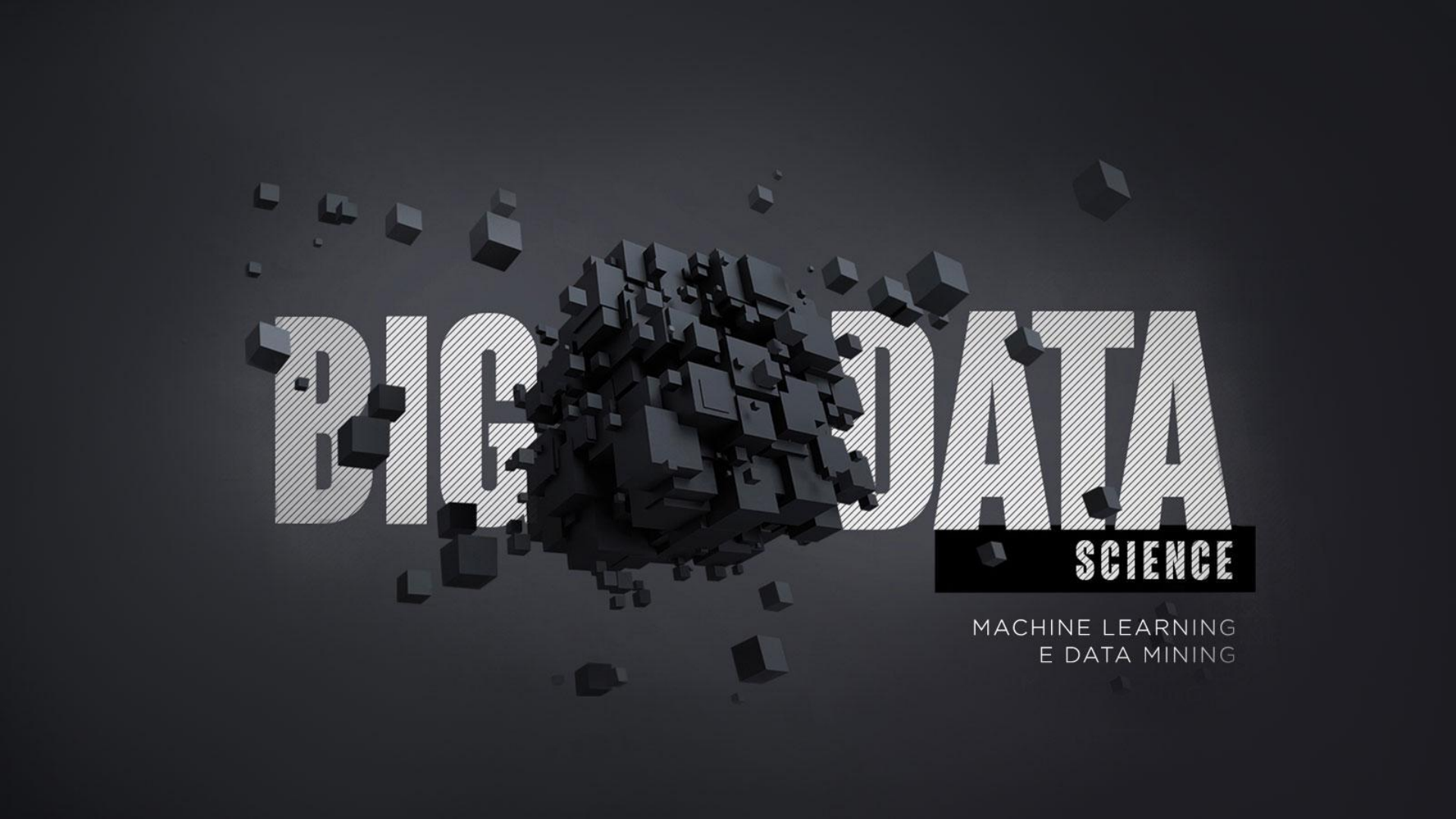

BIG DATA
SCIENCE
MACHINE LEARNING
E DATA MINING

BIG DATA SCIENCE

ESTATÍSTICA APLICADA

DIÓGENES JUSTO

SHORT BIO

DIÓGENES JUSTO

- Mestre em Economia Aplicada (UFRGS)
- MBE Economia (UFRGS)
- Especialista em Banking (FGV)
- Especialista em Data Science (John Hopkins University / Coursera).
- Bacharel em Matemática Aplicada e Computacional (UFRGS)

DIÓGENES JUSTO

- Cursos de especialização em Big Data, Machine Learning e Data Mining no MIT, Washington University, University of Illinois e Stanford
- Head Data & Analytics na Via Varejo - Profissional certificado PMP.
- 20 anos de experiência na área de TI, tendo atuado em desenvolvimento, infraestrutura, banco de dados e B.I., além de projetos.

EMENTA

- O ramo de estudo da estatística, sua relação com demais disciplinas (data mining, machine learning, data science e big data) e suas aplicações
- Estatística Descritiva - média, mediana, moda, variância e desvio padrão, outliers
- Aplicações
- Distribuição de dados, histograma, amostra, população e gráficos
- Análise de correlação
- Regressão Linear Simples
- Modelos regressivos: autoregressivo, multivariada
- Exercícios e Aplicações

AVALIAÇÕES

- Listas de EXERCÍCIOS
- MODELO REGRESSIVO elaborado em aula com uso do EXCEL
- Modelo regressivo elaborado em aula com uso do R

ESTATÍSTICA DESCRITIVA

É a ciência que estuda os fenômenos probabilísticos, de forma a descrever e compreender, com base nos dados e informações existentes, determinada situação.
Avaliando-se o grau de incerteza, será possível tentar estimar eventos futuros.

ESTATÍSTICA DESCRITIVA

FUNÇÕES BÁSICAS

MÉDIA (ARITMÉTICA)

Descreve um ponto médio da média.
É uma medida de tendência central.

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n} a_i = \frac{1}{n}\left(a_1 + a_2 + \cdots + a_n\right)$$

MÉDIA PONDERADA

Para cada dado, é considerado
um peso diferenciado.

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Onde w é o peso de cada valor x.

MODA

É o elemento do conjunto de dados que aparece o maior número de vezes. Um conjunto pode ser bimodal, multimodal ou amodal (no último caso, sem moda).
Moda(1,2,2,3) = 2

MEDIANA

Pressupõe a ordenação crescente dos dados. Para conjuntos com número ímpar de elementos, será o elemento central. Para conjuntos com número par, será a média dos dois elementos centrais.
Mediana(1,2,3,4,5) = 3

Mediana(1,2,3,4) = (2 + 3)/2 = 2,5

AMPLITUDE

Mede a distância entre maior valor e o menor valor.
Ou seja, a largura da faixa de valores nos quais todos os valores do conjunto estarão inseridos.
É uma medida de dispersão de dados.
Amplitude(X)=Máx(X)-Mín(X)

Ex: Amplitude(1,2,3) = 3 - 1 = 2

VARIÂNCIA

Também é uma medida de dispersão de dados. Mede, para cada ponto, a distância entre ele e a média (tendência central) do conjunto. Ao final obtém o valor médio destas distâncias.

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Onde μ é a média.
Obs: Utilizar 1/n para população, e 1/(n-1) para amostra.

DESVIO PADRÃO

Para a elaboração da fórmula da variância, foi incluído um quadrado propositalmente, o que submete o valor final a uma dimensão diferente (em geral maior) dos dados. O desvio padrão é a função similar a variância, porém com a vantagem de estar na mesma dimensão dos dados.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Onde μ é a média.

Obs: Utilizar 1/n para população, e 1/(n-1) para amostra.

POPULAÇÃO

Utilizamos o termo população para definir um conjunto completo de dados. Isto é, conseguimos obter todos os dados possíveis do universo em estudo.

AMOSTRA

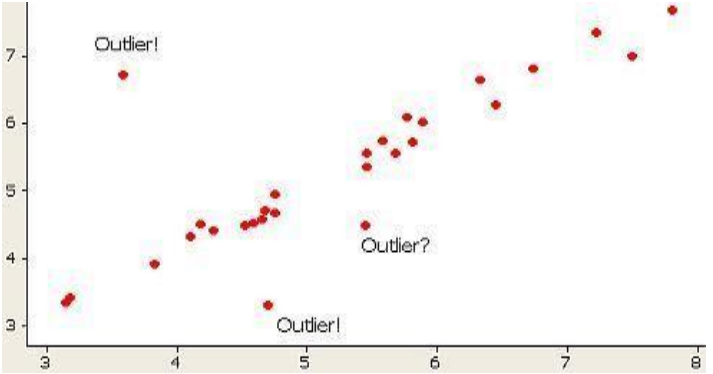
Nem sempre conseguimos obter todos os dados (por exemplo, pesquisa eleitoral), portanto utilizamos parte deles (a amostra) para tentar explicar o que pode ocorrer com a população. Chamamos, neste caso, a amostra de representativa.

Em algumas fórmulas estatísticas, como variância e desvio padrão, por exemplo, há diferença na fórmula.

OUTLIER

Quando um determinado elemento do conjunto de dados está muito longe ou fora de um determinado padrão, o chamamos de outlier.

Quando estamos realizando estimativas, a identificação dos outliers será útil para melhorarmos a precisão das previsões.



ESCORE Z

O teste Z, que produz o escore é uma forma de encontrar matematicamente outliers. Dizemos que um outlier, em geral, é aquele que está a mais de 3x o desvio padrão, mas este número (3) é calibrável de acordo com cada modelo.

$$z = \frac{x - \mu}{\sigma}$$

Onde μ é a média aritmética e σ é o desvio padrão. Um outlier será aquele na qual |z|>3.

EXERCÍCIO

Utilizando as estatísticas descritivas que estudamos até aqui, vamos calcular o risco e o retorno de uma carteira de investimentos.

Referências:

http://hcinvestimentos.com/2010/09/19/como-calcular-o-risco-volatilidade-de-um-investimento/

https://financascorporativas.files.wordpress.com/2010/08/retorno-e-risco-de-carteiras-de-investimentos6.pdf

EXERCÍCIO 1

Crie uma planilha no google drive e extraia as cotações com:

```
=GoogleFinance("ibov","price",DATE(2014, 1, 1),
DATE(2014, 5, 30))
=GoogleFinance("flry3.sa","close",DATE(2014,1,1),
DATE(2014,  5, 30))
=GoogleFinance("Alll3.sa","close",DATE(2014,1,1),
DATE(2014,  5, 30))
```

Obs: verifique em qual língua o seu drive está. Se for em inglês, como o meu, utilize vírgulas nas fórmulas acima. Caso contrário, substitua-as por ponto e vírgula.

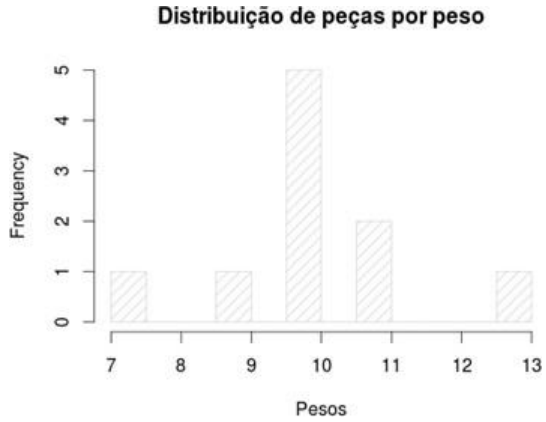
EXERCÍCIO 2

Crie um gráfico de linha que demonstre as variações ao longo do tempo, comparando as 3 séries de dados

Crie cabeçalhos na planilha para calcular individualmente o retorno de uma aplicação feita no início do período e sacada no final do período, bem como o retorno “ótimo”

Inclua no cabeçalho, o cálculo de “risco” (variação em relação a tendência central dada pelo desvio padrão)

HISTOGRAMA



Quando desejamos analisar as variações de um determinado evento, como por exemplo, defeitos na produção de uma peça, é útil utilizar um histograma.

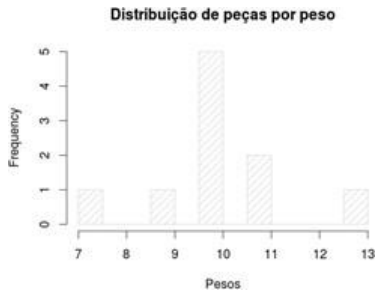
CRIAÇÃO

Pode-se observar, neste exemplo, que temos 1 peça com peso 7, 1 com peso 9, 5 peças com peso 10, 2 peças com peso 11 e 1 peça com peso 13. Suponhamos agora, que a tolerância de peso é 10%.

Teremos, portanto, 8 peças em conformidade e duas consideradas como defeituosas, pois não atendem aos requisitos mínimos.

Peça	Peso
1	10
2	11
3	10
4	9
5	10
6	10
7	13
8	11
9	7
10	10

Peça	Peso
7	1
8	0
9	1
10	5
11	2
12	0
13	1



Vamos montar agora, uma tabela de frequência, elaborada a partir da contagem de ocorrências de cada tipo, como fizemos acima.

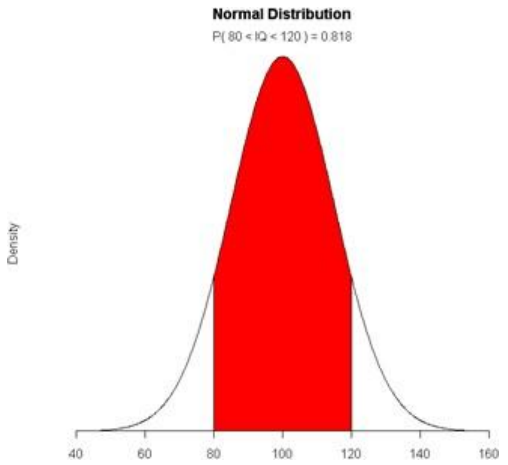
Um histograma será o gráfico de barras gerados a partir da tabela de frequências (quantidades) de determinadas ocorrências (peças produzidas por peso), que estão sendo analisadas para um processo (produção de peças). O eixo x (horizontal) representará os agrupamentos das ocorrências (neste caso, os agrupamentos por pesos). Já o eixo y (vertical) mostrará a Frequência (quantidades).

DISTRIBUIÇÃO

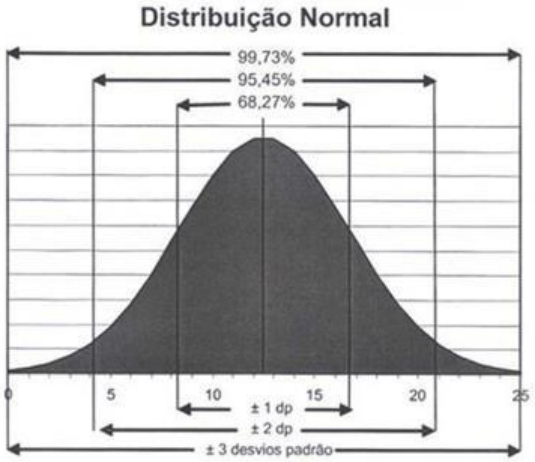
Chamamos de distribuição uma função que demonstra a probabilidade de uma variável apresentar determinado valor.

DISTRIBUIÇÃO NORMAL

Uma distribuição que se aproxime de uma curva perfeita em formato de sino simétrico, com elevação alta no ponto médio e curvas suaves a esquerda e direita, chamamos de distribuição normal, e assume o formato abaixo.

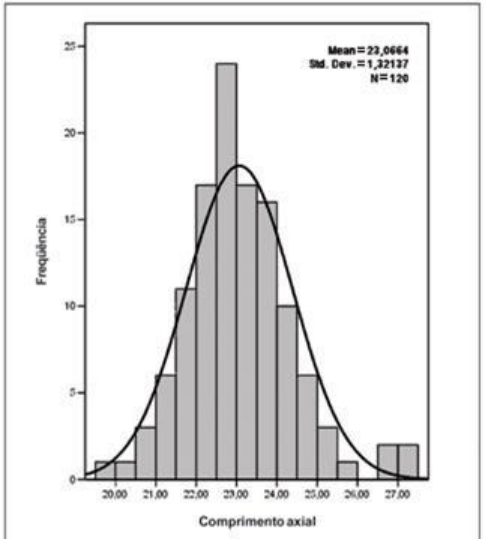


Uma distribuição normal apresenta algumas características como por exemplo o grau de confiança que que um novo valor esteja dentro de uma determinada faixa. Abaixo podemos constatar que entre -2 desvio padrão e +2, há um grau
de confiança de 95,45% (de que os valores de uma variável estarão presentes no intervalo).



O gráfico abaixo mostra um histograma (barras) sobreposto com uma curva normal, demonstrando sua semelhança.

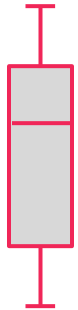
http://shabal.in/visuals/histogram2density.gif



BOX PLOT

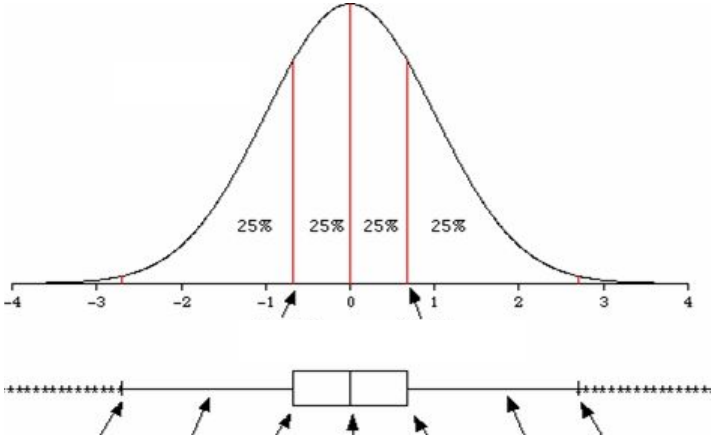
- Também chamado de DIAGRAMA DE CAIXA OU EXTREMOS E QUARTIS, é um gráfico que apresenta a distribuição de valores em um eixos, segmentado em geral por:
  - Menor valor, 1.o Quartil, Mediana, 3.o Quartil e Maior valor

- Desta forma, teremos 4 "partes", cada qual com 25% DOS DADOS.



BOX PLOT E DISTRIBUIÇÃO NORMAL

Um box plot pode representar uma distribuição. Se ele tiver um formato "comportado" (simétrico, contraído no meio - entre 1.o e 3.o quartil), tal qual uma curva normal, podemos dizer que ele representa uma distribuição normal. Caso contrário, o box ficará esticado ou transladado, demonstrando a distribuição.



Min 25% 1.o quartil Mediana 3.o quartil 25% Max

Q1 E Q3

Q1 = (n+1)/4 valor na ordem de classificação

Q3 = 3(n+1)/4 valor na ordem de classificação

DISTRIBUIÇÃO NORMAL

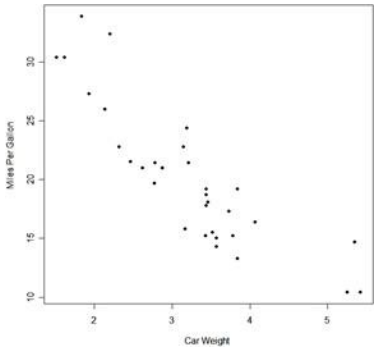
Obs 1: a função que gera a curva normal é dada por:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

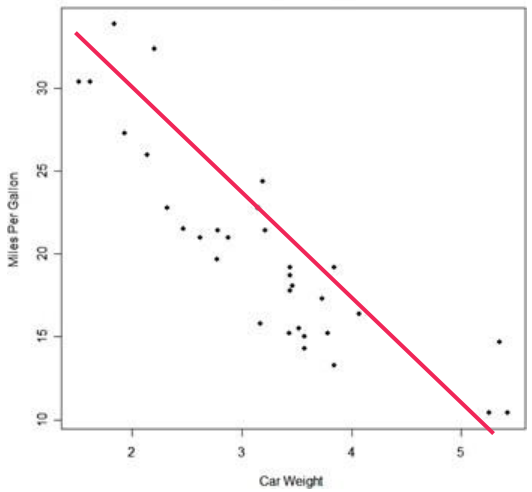
Obs 2: utilizamos a letra grega sigma para representar o desvio padrão. Daí deriva o nome do programa 6Sigma, utilizando como motivador para o nome 6*desvio padrão, ou seja, um grau de confiança de 99,99966% (3,4 defeitos por milhão)

DIAGRAMA DE DISPERSÃO

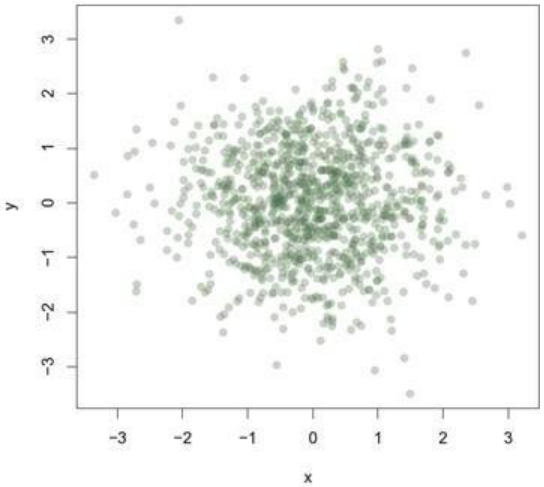
Veículo	Veículo	Veículo
Carro 1	17	4
Carro 2	26	2
...	...	...



Uma forma de avaliar dados ou informações que tem uma certa relação entre si (chamamos de correlação) é utilizar um gráfico de dispersão. Para gerar este gráfico (em inglês Scatterplot) utilizaremos os valores de duas variáveis com intuito de compará-las.



Repare que podemos traçar uma linha reta no sentido diagonal decrescente. Quanto mais os pontos da dispersão se aproximarem de uma reta, maior será a correlação entre as variáveis. Chamamos a primeira variável independente (aquela que deve ser a causa) e a segunda, variável de dependente (aquela que estamos tentando explicar, neste caso, milhas por galão, ou seja, o efeito da causa).



Quando as variáveis não tem correlação nenhuma entre si, o gráfico não apresenta uma tendência ou aproxima-se com uma reta, como no exemplo ao lado

COVARIÂNCIA

A estatística (métrica) covariância indica o grau de interdependência entre duas variáveis aleatórias.

$$Cov(X,Y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})}{n}$$

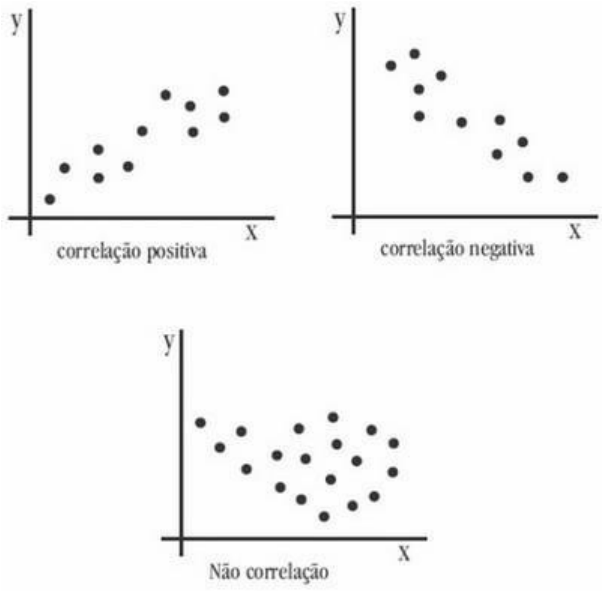
Obs: para cov de uma população utilize n, para amostra, utilize n-1

CORRELAÇÃO LINEAR

A estatística (métrica) correlação linear, coeficiente de correlação ou simplesmente correlação, indica a força e relacionamento linear de duas variáveis aleatórias.
É representada por r ou ϱ (rô, letra grega).

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \cdot \sigma_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X) \cdot Var(Y)}}$$

Obs: Os valores possíveis de r são entre -1 e 1. Quanto |r| mais próximo de 1, maior o relacionamento entre as variáveis.



O sinal positivo ou negativo de R indica a orientacão da reta de regressão.

INSIGHT

Se encontrarmos valores com um alto coeficiente de correlação, poderemos supor que um influencia em outro (relação causa-efeito).
Assim sendo, a partir de estimativas futuras da variável independente, calcularemos a previsão da variável dependente.

Parece ter sido algo parecido que
Galton (primo de Charles Darwin) utilizou
para propor a "Regressão a Média", baseado
em estudos aplicados a área da saúde.

O método estatístico da regressão
(e suas ariantes) é um dos mais utilizados
para análise de relacionamento de eventos
em diversos campos de aplicação: saúde, política,
economia[1]…

1 - em economia este campo de estudos se chama econometria.)

IDEIA BÁSICA

Dadas duas variáveis aleatórias (x e y), supomos (hipótese) que uma pode ser a explicação da outra (p.ex.: PIB é explicado pelo tráfego de veículos). Quanto maior o r (coef. corr. linear, melhor a previsibilidade).

Procuraremos uma reta y=ax+b, que passe mais próximo possível do conjunto de dados (visualmente observável a partir de um diagrama de dispersão).

REGRESSÃO LINEAR

REGRESSÃO LINEAR

Chamaremos de modelo de regressão linear uma aplicação, com equação no seguinte formato:

$$y = \alpha.x + \beta$$

A reta será uma aproximação, portanto, para cada ponto há uma diferença entre o ponto real e o ponto expressado por esta equação, Incluiremos, portanto, este fator de erro.

$$y_i = \alpha.x_i + \beta + \varepsilon_i$$

PARA O MODELO:

$$y_i = \alpha.x_i + \beta + \varepsilon_i$$

Obs 1: alfa será o mesmo para todos valores. Utilizamos a notação indicial, pois posteriormente incluiremos outros fatores alfa.

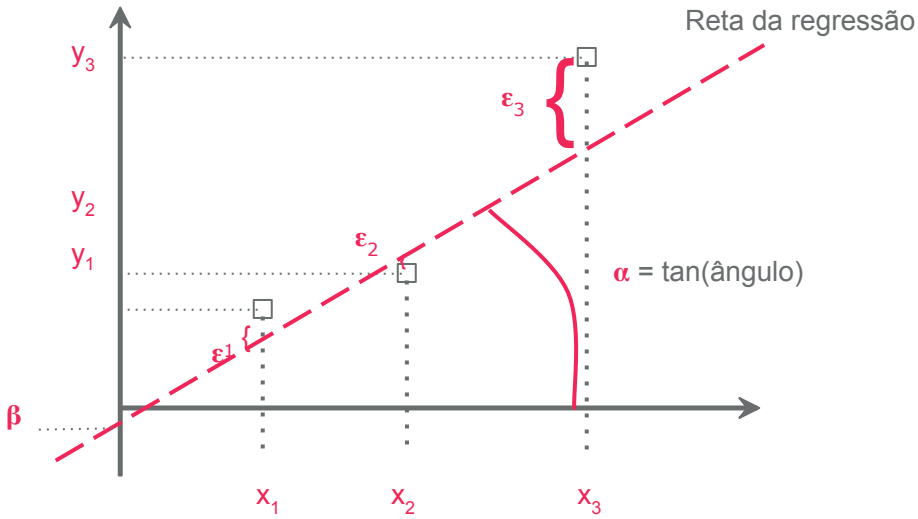
Obs 2: épslon tem índice i pois cada valor é independente, sendo correto incluir também o índice em y e x, portanto.

Obs 3: alfa, em geometria, é chamado de coeficiente angular e beta, coeficiente linear ou intercepto.

Obs 4: a variável x é dita independente e y dependente (depende de x). Ou ainda x exógena (exo=fora do modelo) e y endógena (endo=explicada dentro do modelo).

PARA O MODELO:

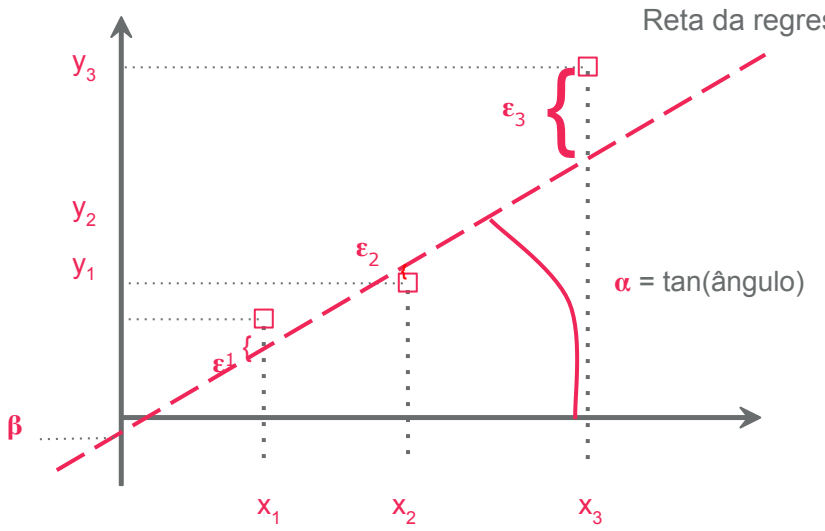
$$y_i = \alpha.x_i + \beta + \varepsilon_i$$



PERG.: MAS COMO DETERMINAR OS COEFICIENTES ALFA E BETA, A PARTIR DOS DADOS?

Res.: Através de um método matemático chamado mínimos quadrados, onde procura-se minimizar os erros (epslon),

daí o nome do método.



MÍNIMOS QUADRADOS

- Dado seu rigor matemático acadêmico, a explicação do método foge do escopo da disciplina em nível de pós-graduação. Para maiores referências, sugere-se consultar: http://livrommq.blogspot.com.br/

- Dentro do escopo da disciplina, utilizaremos o EXCEL como ferramenta de cálculo (que se utiliza do método em suas funções):

  - Suplemento ANÁLISE DE DADOS, Opção Regressão Linear

  - Funções:
    para beta, INTERCEPT (INTERCEPÇÃO) e;
    para alfa, SLOPE (INCLINAÇÃO).

R

INTRODUÇÃO

O que é R: Uma engine para cálculos estatísticos interpretados.

Um pouco diferente de uma linguagem de programação tradicional (Java, .Net) pois ele trabalhar por padrão em modo debug (interativo).

- Baseado no S
- Ferramenta de estudo estatístico, cálculos
- Similar ao MatLab

R e R STUDIO

- Ambos são FREE-SOFWARE
- R É A ENGINE (com um front-end simplificado)
- R STUDIO É A IDE, isto é, a interface de desenvolvimento

PACOTES COM MAIS FUNCIONALIDADES

- A instalação padrão vem com o PACOTE BÁSICO (leve)
- Pacotes de EXTENSÃO do R
- Há packages para GRÁFICOS, MINNING, ETC

PACOTES COM MAIS FUNCIONALIDADES

- Para instalar um package: install.packages("<packagename>")
- Para usar (instanciar) um package: library(<packagename>)
- Manter PACKAGES ATUALIZADOS
- TOOLS / Check for Package Update

BÁSICAS CARACTERÍSTICAS

- Trabalha em memória
  - Se carregar 4Gb de dados, pode usar toda memória da máquina
- Há versão servidor
- Roda em 64bits (endereça mais que 4Gb RAM)
- Permite programação (criação de rotinas)

ANALISANDO DADOS

- Utilizando mtcars
- O que são os dados: help(mtcars)
- Olhando os dados: mtcars
- Resumo dos dados: summary(mtcars)
- Olhando uma coluna: mtcars$mpg
- Gráfico: plot(mtcars$mpg)

ENTRADA E SAÍDA DE DADOS

- write.csv(mtcars, file = "mtcars.csv")
- myData <- read.csv("mtcars.csv")

TRABALHANDO COM VARIÁVEIS

- Atribuição: x<- 1
- Mostrar valor: x
- Operações:
  - x+2, x*2, x-2, x/2
  - x^2, sqrt(x)

TRABALHANDO COM VETORES

- v <- mtcars$mpg
- v * 2, v^2, v/2
- Criando um vetor: v2 <- c(1,2,3,4,5)
- Sum(v2)
- Combinando
  - v3 <- c(2,3,4,5,6)
  - c(v2,v3)
  - rbind(v2,v3)
  - cbind(v2,v3)

FUNÇÕES ESTATÍSTICAS

- Média: mean(v2)
- Mediana: median(v2)
- Histograma: hist(swiss$Examination)
- Desvio-Padrão: sd(v2)

Obrigado

profDiogenes.Justo@fiap.com.br

SHIFT
FIAP