Modelos Estatísticos

Disciplina: Modelos Estatísticos
Professora: Jéssica Assunção

Modelos Estatísticos e Tipos de Análise

Disciplina: Modelos Estatísticos

Professora: Jéssica Assunção

Modelos Estatísticos = Machine Learning

O que difere?

Modelagem Estatística
Técnicas baseadas em equações matemáticas para investigar a relação entre variáveis de um conjunto de dados
Fazer inferência de um parâmetro de uma população geral a partir de uma amostra utilizando métodos.
Análise de dados de um estudo de pesquisa clínica para testar a eficácia de uma droga e apresentar um parecer estatístico.
Utilizar um modelo para predizer a propensão de um cliente deixar de comprar ou sair da base de clientes.

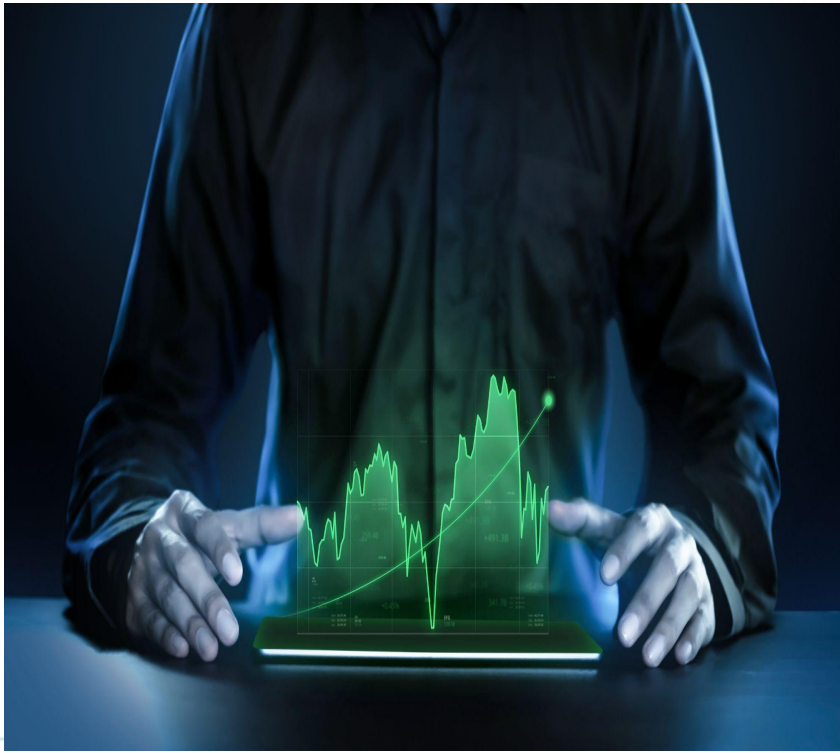
O que difere?

Machine Learning
Sistema composto de algoritmos e métodos matemáticos que podem aprender a partir dos dados, sem depender da programação baseada em regras pré-estabelecidas.
Algoritmos que possibilitam o desenvolvimento de um sistema adaptativo que utiliza dados para melhorar constantemente o seu desempenho, para fazer predição.
Sistema de recomendação de oferta na web que aprende a recomendar as melhores ofertas automaticamente, cada vez que o usuário clica no anúncio.
Sistema que aprende e aperfeiçoa, automaticamente, a predição do uso da memória e cpu de servidores.

Escopo da Modelagem Estatística

O que é um modelo e por que modelar?

Modelos estatísticos são a construção de hipóteses a partir da análise de dados, de sua relação e de outras variáveis para prever ou comprovar fatores.



Objetivos

- Estar equipado para escolher o melhor modelo de acordo com a sua necessidade
- Estar mais capacitado para realizar uma análise de dados
- Se tornar um melhor comunicador

Etapas de Modelagem Estatística

- Identifique o problema e escolha o modelo adequado
- Identifique as variáveis
- Formule o seu modelo
- Avalie seu modelo
- Comunique

Variáveis Quantitativas

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.

Variáveis Quantitativas: são as características que podem ser medidas em uma escala quantitativa:

1. **Variáveis discretas:** características mensuráveis que podem assumir apenas um número finito ou infinito contável de valores e, assim, somente fazem sentido valores inteiros.
2. **Variáveis contínuas**, características mensuráveis que assumem valores em uma escala contínua (na reta real), para as quais valores fracionais fazem sentido.

Variáveis Qualitativas

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.

Variáveis Qualitativas (ou categóricas): são as características que não possuem valores quantitativos, mas, ao contrário, são definidas por várias categorias, ou seja, representam uma classificação dos indivíduos. Podem ser nominais ou ordinais.

1. **Variáveis nominais**: não existe ordenação dentre as categorias. Exemplos: sexo, cor dos olhos, fumante/não fumante, doente/sadio.
2. **Variáveis ordinais**: existe uma ordenação entre as categorias. Exemplos: escolaridade (1o, 2o, 3o graus), estágio da doença (inicial, intermediário, terminal), mês de observação (janeiro, fevereiro,..., dezembro).

Tipos de Variáveis

Medidas de Tendência Central

- Moda: é o valor em que a **frequência dos seus dados é maior**. Então para encontrar a moda desse conjunto poderíamos fazer uma tabela de frequência e ver qual é o número mais frequente.

- Média: é o **resultado da soma de todos os valores dividido pela quantidade de entradas** do nosso conjunto de dados, ou seja, se somarmos todos os nossos valores e dividirmos pela quantidade teremos a média.

- Mediana: é o **valor que divide o nosso conjunto de dados em duas metades**. Para encontrar nossa mediana precisamos primeiro ordenar nossos dados.

Medidas de Tendência Central

Dados	Média	Moda	Mediana
[2,5,7,5,8,10,9]	6,25	10,0	7
[2,5,7,5,8,10,9,100]	18,25	100,00	7,5
[2,5,7,5,8,50,100,100]	34,62	100,00	7,5

Medidas de Dispersão

- Variância: é a média quadrática dos desvios tomados em relação à média aritmética.
- Desvio Padrão: é a raiz quadrática da variância.

Dados	Variância	Desvio Padrão
[2,5,7,5,8,10,9]	6,50	2,55
[2,5,7,5,8,10,9,100]	1097,59	33,13
[2,5,7,5,8,50,100,100]	1867,96	43,22

O que é uma distribuição?

Uma distribuição de probabilidade é um modelo matemático que relaciona um certo valor da variável em estudo com a sua probabilidade de ocorrência. Há dois tipos de distribuição de probabilidade:

1. **Distribuições Contínuas:** Quando a variável que está sendo medida é expressa em uma escala contínua, como no caso de uma característica dimensional.
2. **Distribuições Discretas:** Quando a variável que está sendo medida só pode assumir valores inteiros: 0, 1, 2, etc.

Distribuição Binomial

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0,1,\ldots,n$$

$\mu = np$

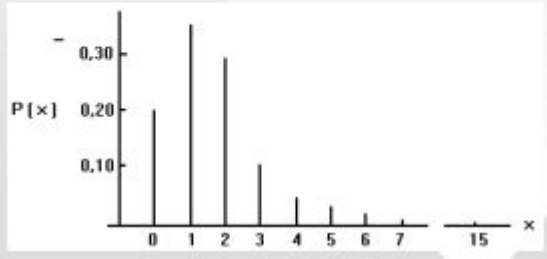
$\sigma^2 = np(1 - p)$

n = número de repetições

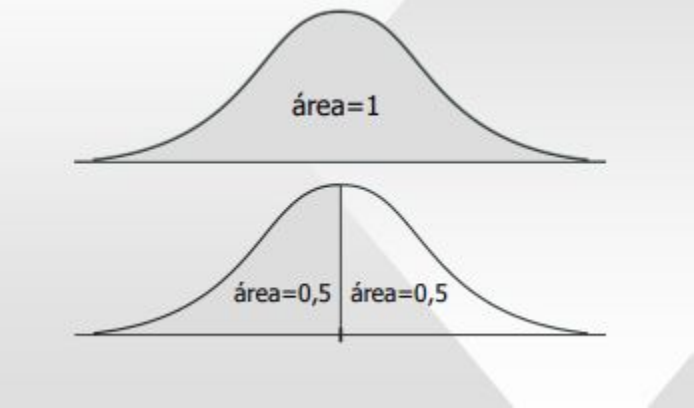
p = probabilidade de sucesso

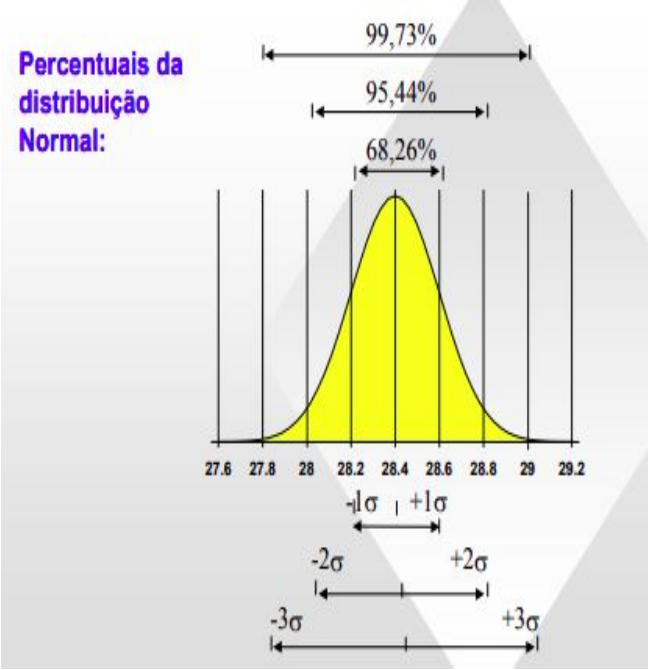
$$\binom{15}{1} = \frac{15!}{1!(15-1)!} = 15$$

$$\hat{P}(1) = \binom{15}{1} x\, 0{,}10^1\, x\, (1-0{,}10)^{15-1} = 15 \times 0{,}10\, x\, 0{,}23 = 0{,}34$$



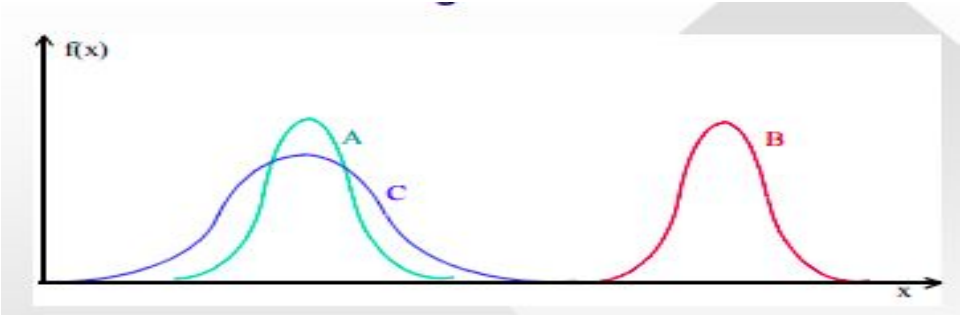
Distribuições Normal

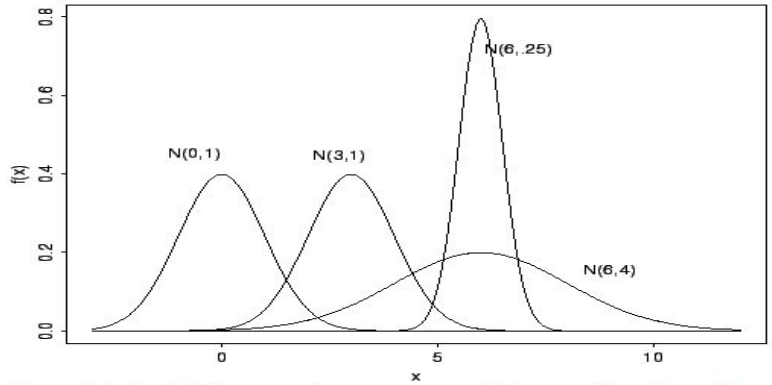




Distribuição Normal

Amostras	Dados	Localização ($\bar{x}$)	Variabilidade (R)
A	10 12 14 16 18	$\bar{x} = 14$	$\bar{R} = 8$
B	22 24 26 28 30	$\bar{x} = 26$	$\bar{R} = 8$
C	6 10 14 18 22	$\bar{x} = 14$	$\bar{R} = 16$

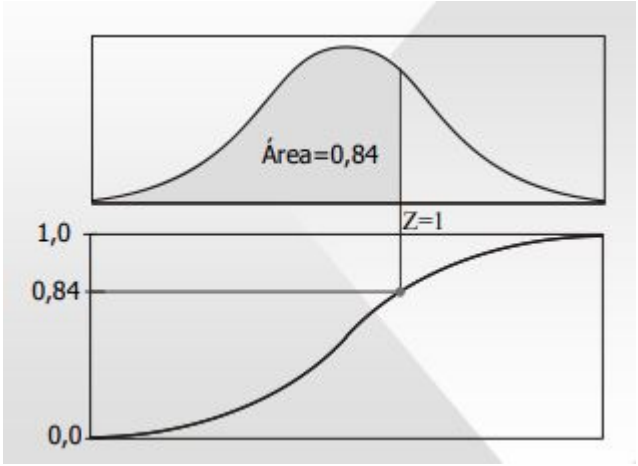




Distribuições Normal

$$f(x) = \frac{1}{\sigma\sqrt{2\Pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$P\{X \le x\} = P\left\{Z \le \frac{x-\mu}{\sigma}\right\} = F(Z) \Rightarrow \text{Tabelado}$$

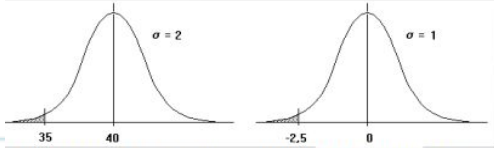


$P\{X \ge 35\} = 1 - P\{X \le 35\}$

$P\{X \le 35\} = P\left\{Z \le \frac{35-40}{2}\right\} = P\{Z \le -2{,}5\}$

Tabela: F(-2,5) = 0,0062

Assim a resposta é 1 - 0,0062 = 99,38%



Salário versus Anos de Experiência

	YearsExperience	Salary
0	1.1	39343.0
1	1.3	46205.0
2	1.5	37731.0
3	2.0	43525.0
4	2.2	39891.0

Covariância

- A covariância entre duas variáveis (X, Y) é uma medida de variabilidade conjunta dessas duas variáveis aleatórias.
- Quando a covariância entre essas variáveis é positiva os dados apresentam tendência positiva na dispersão.
- Quando o valor da covariância é negativo, o comportamento é análogo, no entanto, os dados apresentam tendências negativas.

Covariância

$$Cov_{xy} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{(n-1)} = \frac{\Sigma xy - n\bar{x}\bar{y}}{(n-1)}$$

Covariância

Ano	Consumo (X)	Taxa de Juros (Y)	XY
1	800	10	8000
2	700	11	7700
3	600	13	7800
4	500	14	7000
Média	**650**	**12**	**7625**
Covariância			

Cov(X,Y) = $7625 - 650 \times 12 = -175$

Correlação

- Correlação é uma versão em escala de covariância que assume valores em [−1,1]

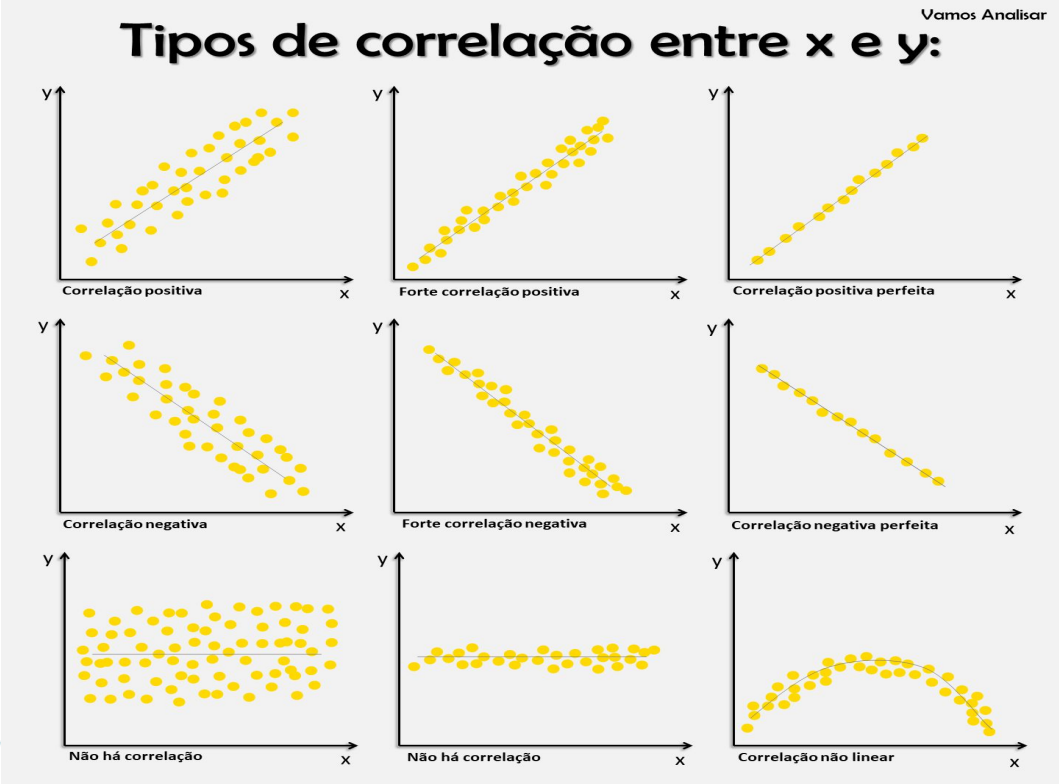
- Com uma correlação de ± 1 indicando associação linear perfeita e 0 indicando nenhuma relação linear.

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Correlação

- Para p = 1, tem-se uma correlação perfeita entre as duas variáveis.
- Se p = - 1, há uma correlação perfeita entre as variáveis, no entanto, essa correlação é negativa.
- Caso p = 0, as duas variáveis não dependem linearmente uma da outra.

Correlação


Vamos Analisar
Tipos de correlação entre x e y:
y
x
Correlação positiva
Forte correlação positiva
Correlação positiva perfeita
Correlação negativa
Forte correlação negativa
Correlação negativa perfeita
Não há correlação
Não há correlação
Correlação não linear

Referências Bibliográficas

- Introdução à Estatística - Mário F. Triola
- Noções de Probabilidade e Estatística - Marco Nascimento Magalhães
- Modelos de Regressão em R - Écio Souza Diniz
- Análise de Séries Temporais - Pedro A. Moretin / Clélia M. C. Toloi
- Análise e Previsões de Séries Temporais: Os modelos ARIMA - Reinaldo Castro Souza / Maria Emília Camargo