XPe

Aprenda com quem faz

Análise Estatística de Dados

Máiron César Simões Chaves

2023

SUMÁRIO

> Capítulo 1

Capítulo 1. Introdução à Disciplina

O objetivo deste curso é apresentar ao aluno métodos estatísticos de análise de dados que trazem abordagens científicas de como extrair conhecimento a partir de dados. Por exemplo, suponha que em uma empresa um treinamento foi aplicado à sua equipe vendas, e que a média de vendas após o treinamento é ligeiramente maior que a média de vendas antes do treinamento. Será que essa diferença foi devido ao acaso ou foi de fato um efeito do treinamento? O aluno aprenderá como utilizar métodos probabilísticos para analisar a incerteza sobre os fenômenos.

O curso compreende análises descritivas de dados, análises através de gráficos, compreensão das principais distribuições de probabilidades discretas e contínuas e testes de hipóteses. Além disso, demonstra a utilização da técnica de regressão linear tanto para explicar o impacto de uma ou mais variáveis X em uma variável Y quanto para prever os valores da variável Y baseado nos valores das variáveis X.

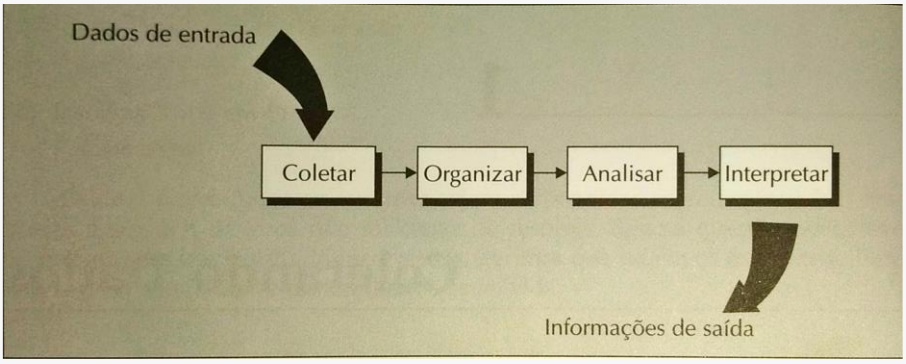
O curso foi cuidadosamente elaborado para sempre trazer exemplos claros e tangíveis, direcionados ao ambiente de negócios, para que o aluno não fique apenas na teoria abstrata. O rigor matemático será mantido apenas para formalizar conceitos e para mostrar ao aluno a origem de tais resultados, em momento algum a execução de cálculos matemáticos será exigida do aluno.

Para aplicar os métodos estatísticos aos dados, utilizaremos uma ferramenta computacional. Esta, por sua vez, será ensinada da forma mais didática possível. O objetivo do curso é ensinar a estatística ao aluno, portanto, seus esforços devem ser direcionados aos conceitos estatísticos, e não à ferramenta.

O que é a Estatística

A Estatística é um ramo das ciências exatas que visa obter conclusões a partir de dados, envolvendo técnicas para coletar, organizar, descrever, analisar e interpretar dados.

Figura 1 – O processo de análise de dados.



Fonte: Smailes & McGrane, 2012.

Em Pinheiro at al. (2009) temos a seguinte de definição de Estatística: “é identificar comportamentos médios, comportamentos discrepantes, comparar comportamentos, investigar a interdependência entre variáveis, revelar tendências, etc.”

Segundo Bussab e Morettin (1987), a Estatística pode ser dividida em dois grandes grupos: a Estatística Descritiva e a Estatística Inferencial.

A Estatística Descritiva tem como principal objetivo descrever as variáveis de um conjunto de dados, calculando medidas de centralidade e dispersão. Além do uso de gráficos e tabelas. Já a Estatística Inferencial, tem por objetivo a coleta, a redução, a análise e a modelagem dos dados, a partir do que, finalmente, faz-se a inferência para uma população da qual os dados (a amostra) foram obtidos. Um aspecto importante da modelagem dos dados é fazer previsões, a partir das quais se podem tomar decisões. Através de métodos probabilísticos, a Estatística nos permite compreender a incerteza que pode ocorrer em variáveis do mundo real.

O que é Ciência de Dados

Conforme definição em Provost e Fawcett (2016), a ciência de dados é um conjunto de métodos que cercam a extração do conhecimento a partir dos dados. O cientista de dados é o profissional que trabalha exclusivamente com a análise avançada de dados, é um profissional extremamente requisitado no mercado e que possui habilidades em diferentes áreas, sendo as principais: matemática, estatística, ciência da computação e negócios.

Ciência de Dados vs Estatística

A principal diferença entre Ciência de Dados e Estatística está na abordagem dos problemas. Na estatística "tradicional", nosso objetivo geralmente é definir uma população, coletar a amostra, organizar esses dados, aplicar métodos estatísticos para tirar conclusões e realizar um determinado estudo. Ou seja, tirar conclusões sobre uma população a partir de uma amostra.

Na Ciência de Dados, temos um grande volume de dados (nem sempre) que exigem maiores infraestruturas de TI para seu armazenamento, e a abordagem do Cientista de Dados é construir modelos preditivos, que uma vez validada sua capacidade preditiva, será colocado em produção para ser executado em tempo real. Por exemplo, para estimar as chances de uma nova transação de cartão de crédito ser ou não uma fraude, o modelo preditivo armazenado em nuvem pode receber os dados dessa transação via api, em formato json, processar esses dados e retornar a probabilidade de ser uma fraude. Caso a probabilidade seja alta, a transação pode ser bloqueada automaticamente.

O que são Medidas de Centralidade e Dispersão

Conforme definição em Smailes e McGrane (2012), as medidas de centralidade referem-se a valores típicos de uma variável, isto é, um valor em torno do qual uma grande proporção de outros valores está centralizada.

As medidas de centralidade que iremos aprender são: **Média Aritmética** e **Mediana**.

Também é importante saber como os dados se espalham ou o quão variadas são as observações em torno dessa medida central, e para isso, utilizamos as medidas de dispersão.

As medidas de dispersão (ou espalhamento) que iremos aprender são: **Variância**, **Desvio Padrão**, **Coeficiente de Variação**, **Amplitude** e **Quartis**.

Aplicar medidas de centralidade e de dispersão é uma forma eficiente de resumir dados e ajudar a revelar informações contidas neles, e assim utilizar o conhecimento para auxílio a tomada de decisão. Essas medidas nos ajudam a fazer uma sondagem dos dados, ou seja, tomar um primeiro contato com a informação disponível.

Para nosso estudo ficar mais palpável, tomemos um contexto. Imagine que você trabalha em um grande hipermercado varejista. É comum, nesse segmento, que preços dos produtos sejam remarcados diariamente. Então, você recebeu um conjunto de dados contendo o histórico de variação de preços para o produto café durante um mês. Portanto, cada linha do conjunto de dados é um dia, e cada valor é o preço praticado para o produto café.

Figura 2 – Variável “Preco_Cafe” contendo 30 observações.

Preco_Cafe
4,77
4,67
4,75
4,74
4,63
4,56
4,59
4,75
4,75
4,49
4,41
4,32
4,68
4,66
4,42
4,71
4,66
4,46
4,36
4,47
4,43
4,4
4,61
4,09
3,73
3,89
4,35
3,84
3,81
3,79

Para começar a explorar a variável, iremos tomar a **Média Aritmética**. A Média Aritmética é uma medida estatística que é calculada somando os valores da variável e dividindo pela quantidade de valores. Pode ser representada pela equação:

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

Onde x são os valores individuais de cada observação e n é a quantidade de observações.

Em estatística, as vezes trabalhamos com amostras, e as vezes trabalhamos com toda a população. Apesar do cálculo da Média Aritmética ser o mesmo para ambos os casos, é importante utilizar uma notação específica para quando a Média Aritmética é calculada sobre uma amostra e uma outra notação para quando for calculada sobre uma população. A notação $\bar{x}$ indica que a média se originou de uma amostra, e a notação μ

(letra grega mi) indica que a média se originou de uma população. População, nesse contexto, seria se estivéssemos trabalhando com todo histórico existente de preços do café. Não é o nosso caso, pois estramos trabalhando com uma amostra de trinta dias de variações de preço.

Aplicando a fórmula na nossa variável de estudo, que são as variações do preço do café, temos:

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

$$\bar{x} = \frac{132{,}79}{30}$$

$$\bar{x} = 4{,}42$$

A interpretação fica: O hipermercado geralmente pratica o preço de 4,42 reais para o café.

Agora que já sabemos o preço médio, precisamos de uma medida de dispersão para saber o quão os demais preços se destoam desse preço médio. Para esse objetivo, utilizaremos o **Desvio Padrão**, que pode ser calculado pela fórmula:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Onde: x são os valores individuais de cada observação, $\bar{x}$ é a média aritmética e n é a quantidade de observações.

Assim como na Média Aritmética, temos notações para desvio padrão amostral e populacional. A literatura nos sugere utilizar a notação S para Desvio Padrão amostral e a letra grega sigma σ para desvio padrão

populacional. Há uma pequena diferença entre o cálculo do desvio padrão da amostra S e no desvio padrão da população σ. No S utilizamos n-1 em seu denominador, já no σ utilizamos apenas o n.

Outra medida de dispersão muito utilizada na estatística é a variância, que nada mais é do que o desvio padrão elevado ao quadrado. Ou seja, o cálculo da variância e do desvio padrão é praticamente o mesmo, a diferença é que na variância não tomamos a raiz quadrada. No momento, preocuparemos apenas com o desvio padrão.

Calculando o desvio padrão dos preços, temos:

Preco_Cafe	Média	(Preço - Média)^2
4,77	4,4263	0,1181
4,67	4,4263	0,0594
4,75	4,4263	0,1048
4,74	4,4263	0,0984
4,63	4,4263	0,0415
4,56	4,4263	0,0179
4,59	4,4263	0,0268
4,75	4,4263	0,1048
4,75	4,4263	0,1048
4,49	4,4263	0,0041
4,41	4,4263	0,0003
4,32	4,4263	0,0113
4,68	4,4263	0,0643
4,66	4,4263	0,0546
4,42	4,4263	0,0000
4,71	4,4263	0,0805
4,66	4,4263	0,0546
4,46	4,4263	0,0011
4,36	4,4263	0,0044
4,47	4,4263	0,0019
4,43	4,4263	0,0000
4,4	4,4263	0,0007
4,61	4,4263	0,0337
4,09	4,4263	0,1131
3,73	4,4263	0,4849
3,89	4,4263	0,2877
4,35	4,4263	0,0058
3,84	4,4263	0,3438
3,81	4,4263	0,3799
3,79	4,4263	0,4049
		∑ = 3,00

Desvio Padrão:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

$$s = \sqrt{\frac{3{,}00}{30-1}}$$

$$s = \sqrt{0{,}10}$$

$$\underline{s = 0{,}31}$$

A interpretação fica: Os preços praticados para o café variam em média 0,31 centavos em torno do seu preço médio.

O Desvio Padrão nos dá a noção de variabilidade utilizando a própria unidade de medida da variável. No entanto, nem sempre o pesquisador compreende bem as unidades de medida, e pode ser difícil entender quão grande ou pequena é a dispersão da variável. Para isso, podemos adotar uma outra métrica, que nos dará a noção de dispersão em termos percentuais. Essa métrica é o Coeficiente de Variação. O Coeficiente de Variação é uma razão entre o Desvio Padrão e a Média Aritmética.

Uma vez que já calculamos a Média Aritmética e o Desvio Padrão, já conseguimos calcular o Coeficiente de Variação.

$$\text{Coeficiente de Variação: } \frac{s}{\bar{x}} * 100$$

$$\text{Coeficiente de Variação: } 7{,}01\ \%$$

$$\text{Coeficiente de Variação: } \frac{0{,}31}{4{,}42} * 100$$

A interpretação fica: Os preços praticados para o café variam em média 7,01% em torno do preço médio.

Vamos recapitular tudo que já aprendemos sobre a variável que estamos estudando.

O preço médio praticado por dia é de $\bar{x}$=4,77, como uma variação média de s=0,31 centavos para mais ou para menos, ou seja, o preço geralmente varia cv=7,01% em torno do preço médio.

Outro conjunto de medidas de dispersão bastante usado são os quartis. Eles são valores que dividem a variável em quatro partes iguais, e assim, cada parte representa 25% da variável.

Os quartis se dividem em Primeiro Quartil (Q1), Segundo Quartil (Q2) e Terceiro Quartil (Q3).

O Primeiro Quartil, ou Q1, é um valor que deixará 25% dos dados abaixo. O Segundo Quartil, ou Q2, é um valor que deixará 50% dos dados abaixo e 50% dos dados acima dele, ou seja, é um valor que corta os dados ao meio. Já o Terceiro Quartil, ou Q3, é um valor que deixará 75% dos dados abaixo dele.

Para calcular os quartis, os valores da variável estudada devem estar ordenados do menor para o maior. Vamos calcular os quartis para o histórico de variações diária do preço do café e, posteriormente, interpretá-los.

	Preco_Cafe
1	3,73
2	3,79
3	3,81
4	3,84
5	3,89
6	4,09
7	4,32
8	4,35
9	4,36
10	4,40
11	4,41
12	4,42
13	4,43
14	4,46
15	4,47
16	4,49
17	4,56
18	4,59
19	4,61
20	4,63
21	4,66
22	4,66
23	4,67
24	4,68
25	4,71
26	4,74
27	4,75
28	4,75
29	4,75
30	4,77

1º Quartil = (n*0,25) = (30*0,25) = 7,5

25% dos dados estão abaixo do primeiro quartil

2º Quartil Mediana = (n*0,5) = (30*0,5) = 15

A mediana corta os dados no meio, 50% estão acima e os outros 50% estão abaixo do valor mediano

3º Quartil = (n*0,75) = (30*0,75) = 22,5

75% dos dados estão abaixo do terceiro quartil

Repare que, ao calcular o primeiro quartil, aplicamos a fórmula n*0,25. Como temos 30 observações, substituímos e fica 30*0,25, que dá 7,5. Arredondando para o inteiro mais próximo, temos 8. Esse número nos diz que o valor que estiver na oitava posição é aquele que deixará 0,25 (ou 25%) dos dados abaixo dele. Nese caso é o 4,35.

A interpretação do Q1, Q2 e Q3 fica (respectivamente): 25% dos preços praticados são até R$4,35 (**Q1**), 50% dos preços praticados são até R$4,47 (**Q2**) e até 75% dos preços praticados são até R$ 4,67 **(Q3**)

Por fim, estudaremos mais uma medida de dispersão, que é a **Amplitude.** A Amplitude nada mais é do que o intervalo entre o maior valor e o menor valor. Como o menor preço foi de R$3,73 e o maior preço foi R$4,77, podemos dizer que na amostra estudada os preços variam entre R$3,73 e R$4,77.

Análise de Dados Através de Gráficos

Um gráfico é a maneira visual de exibir variáveis. Normalmente, é mais fácil para qualquer pessoa entender a mensagem de um gráfico do que aquela embutida em tabelas ou sumários numéricos.

Os gráficos são utilizados para diversos fins (Chambers et al., 1983):

a. Buscar padrões e relações;

b. Confirmar (ou não) expectativas que se tinha sobre os dados;

c. Descobrir novos fenômenos;

d. Confirmar (ou não) suposições feitas sobre os procedimentos estatísticos usados; ou simplesmente

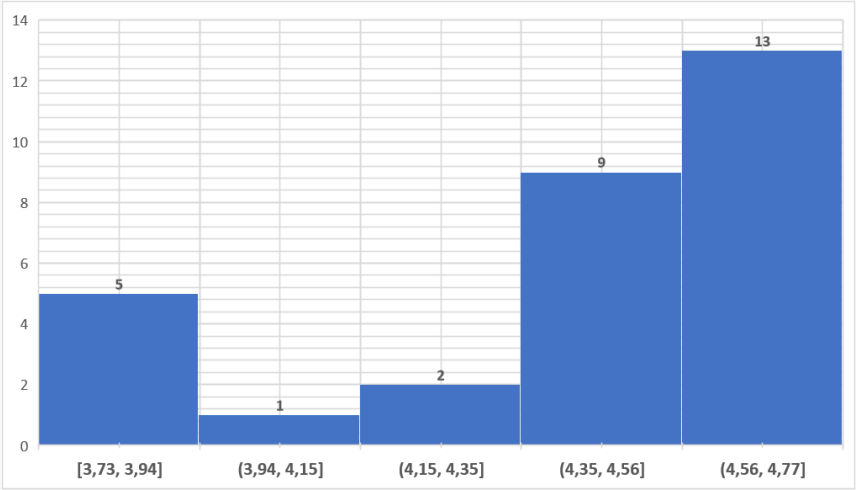
e. Apresentar resultados de modo mais fácil e rápido

f. Um gráfico bastante utilizado na estatística é o Histograma.

O **Histograma** é uma representação gráfica em barras de uma variável, dividida em classes. A altura de cada barra representa a frequência

com que o valor da classe ocorre. Vejamos um histograma para apresentar a variação do preço do café.

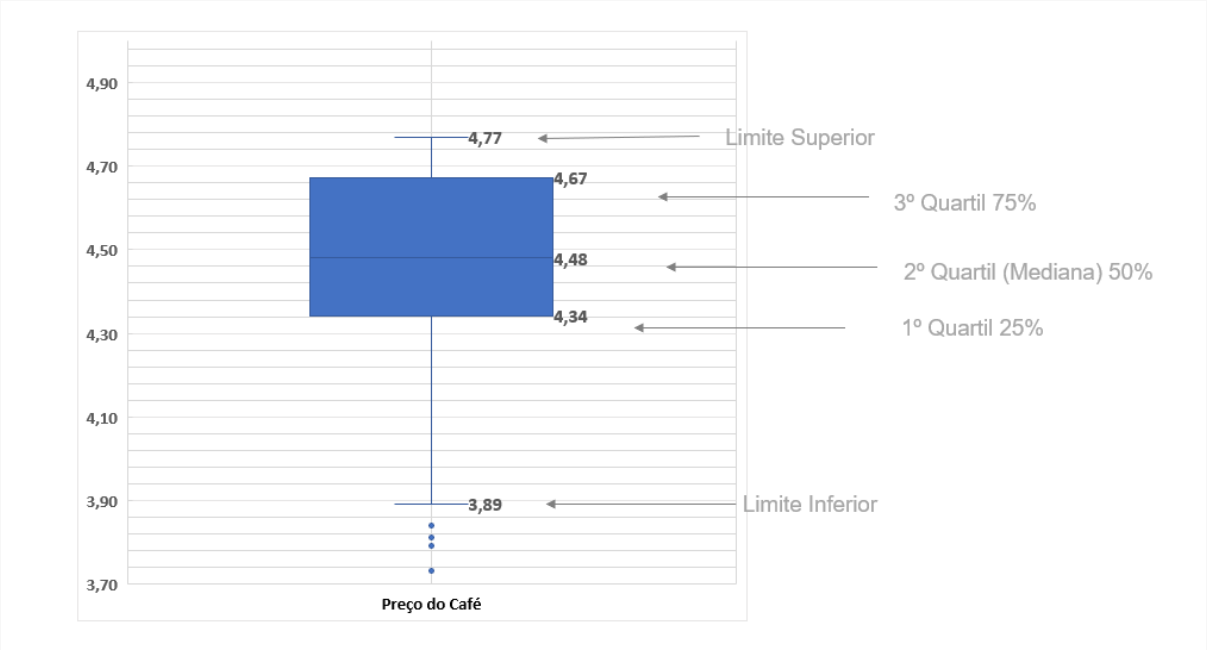
Figura 3 – Histograma dos Preços Praticados para o Café.



Repare que cada barra corresponde à frequência de um intervalo de preços. Interpretando as três primeiras barras da esquerda pra direita, temos: a primeira barra nos informa que tivemos cinco registros onde o preço praticado foi entre R$3,73 e R$3,94. A segunda barra nos informa que tivemos um registro onde o preço estava entre R$3,94 e R$4,15. E a terceira barra nos informa que tivemos dois registros onde o preço estava entre R$4,15 e R$4,35.

Outro gráfico bastante utilizado na Estatística é o **Boxplot**, apresentado por Tukey (Tukey, 1977), que é baseado nos quartis e são um modo rápido de visualizar a distribuição dos dados. Vejamos um boxplot para os preços do café.

Figura 4 – Boxplot dos Preços Praticados para o Café.



Além dos quartis, o boxplot também nos dá o limite inferior e o limite superior. No boxplot da figura 4 vemos que o limite superior R$4,77, ou seja, de acordo com a distribuição dos preços, um valor acima do de R$4,77 é um outlier. Já o limite inferior é R$3,89, ou seja, preços abaixo desse valor são considerados outliers.

Valores outliers, sejam superiores ou inferiores, devem ser investigados para compreender o que houve naquela observação, pode ter sido de fato um evento raro ou apenas um erro de digitação. No boxplot da figura 4 pode-se notar diversos pontos abaixo do limite inferior. Nesse caso, isso ocorreu em alguns dias em que o café estava promocionado, então é de esperar que o preço esteja abaixo do esperado.

O pesquisador pode desejar calcular o limite superior e inferior para identificar os outliers sem necessariamente querer utilizar um boxplot. Para isso, primeiro deve-se calcular o intervalo interquartil (IQR), que nada mais é que subtrair o terceiro quartil pelo primeiro quartil. Uma vez calculado o IQR, para chegar nos valores limites, a fórmula fica:

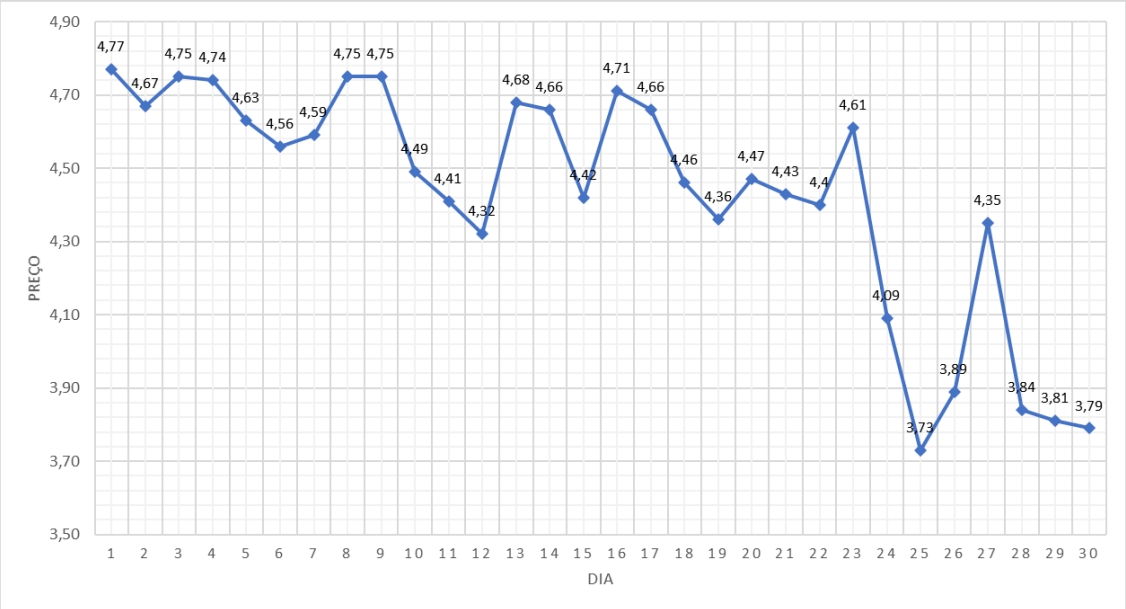
IQR = 3ºQuartil – 1ºQuartil

Limite Inferior = 1ºQuartil – (1.5 * IQR)

Limite Superior= 3ºQuartil + (1.5 * IQR)

Se desejarmos visualizar a evolução dos preços ao longo do tempo, é recomendado utilizar um **gráfico de linhas** (também chamado de gráfico de séries temporais). Ele é bastante simples. Basta plotar a variável no eixo vertical y e o tempo no eixo horizontal x. Cada ponto é representado por um marcador e ligado ao ponto seguinte por uma reta. Em nosso caso, que cada observação da base de dados é um dia de venda, e temos trinta dias observados, cada ponto do nosso gráfico será o preço praticado em um respectivo dia.

Figura 5 – Gráfico de linha da evolução dos preços do café durante os dias do mês.



Podemos observar que ao final do mês os preços vão abaixando. Pode ser devido alguma estratégia, pois em muitas empresas as vendas ao final do mês tendem a ser menores, devido ao fato de que os clientes já gastaram seu salário. Então, reduzir o preço, apesar de diminuir a margem de lucro, pode ser uma estratégia a ser considerada para manter o volume de vendas no fim do mês.

Um gráfico de linhas é muito útil para procurar padrões na variável longo do tempo, como tendências e padrões sazonais. Por exemplo, em uma loja de brinquedos, é comum esperar um pico todo dezembro. Para maiores detalhes sobre tendência e sazonalidade de uma série temporal, a literatura a seguir pode ser consultada: https://otexts.com/fpp2/tspatterns.html.

Supondo que você precise analisar se existe relação entre o preço do café com as vendas do café. Vamos adicionar mais uma variável, conforme a figura 6.

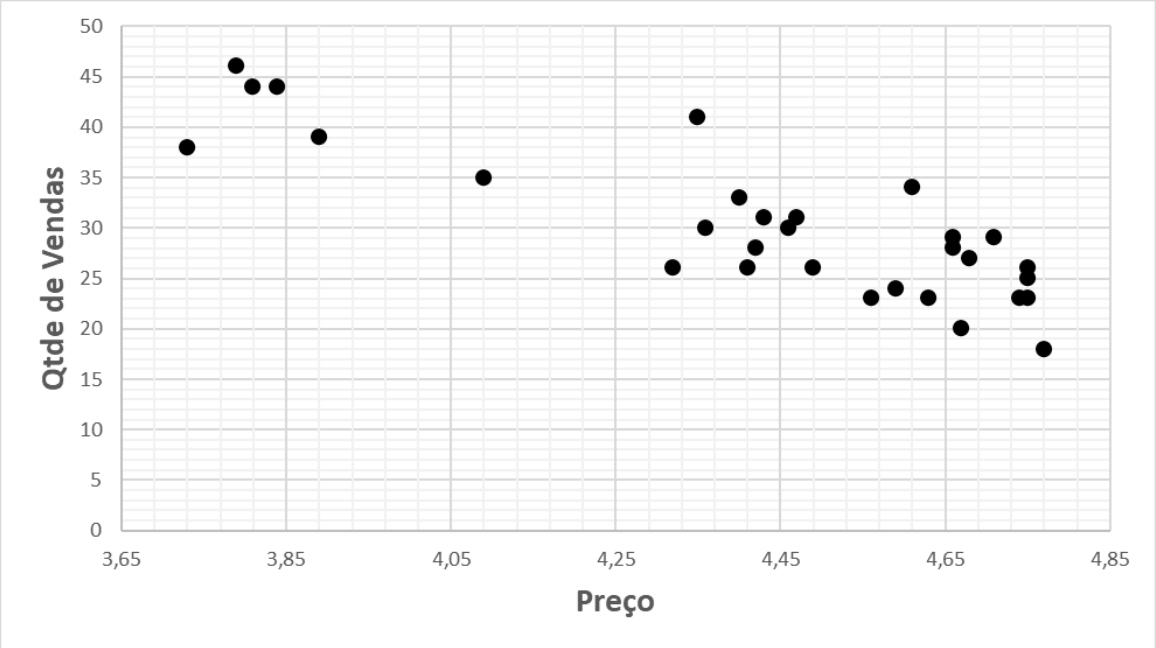
Figura 6 – Variável Preço do Café e a variável Vendas do Café.

Preco_Cafe	Vendas_Cafe
4,77	18
4,67	20
4,75	23
4,74	23
4,63	23
4,56	23
4,59	24
4,75	25
4,75	26
4,49	26
4,41	26
4,32	26
4,68	27
4,66	28
4,42	28
4,71	29
4,66	29
4,46	30
4,36	30
4,47	31
4,43	31
4,4	33
4,61	34
4,09	35
3,73	38
3,89	39
4,35	41
3,84	44
3,81	44
3,79	46

O comportamento esperado é de que quando o preço aumenta, as vendas diminuam. Uma forma de identificar a relação (ou a ausência de relação) entre um par de variáveis é através de um gráfico de dispersão, que exibe os valores de dados para um par de variáveis em suas coordenadas (x, y).

Geralmente, variável “resposta” é colocada no eixo y, e a variável “preditora” no eixo x. A variável resposta em nosso exemplo são as vendas, e a variável preditora (que também pode ser chamada de variável explicativa) é o preço. Então, colocaremos as vendas no eixo y e o preço no eixo x, pois queremos saber como as vendas se comportam na medida em que o preço varia.

Figura 7 – Relação entre o preço do café e as vendas do café.



No gráfico da figura 7, cada ponto é um dia de venda. No eixo x temos o preço, e no eixo y temos quantas foram as unidades vendidas por aquele preço. Como temos trinta observações em nosso conjunto de dados, temos trinta pontos em nosso gráfico de dispersão. Interpretando o gráfico, vemos que o comportamento é o esperado: se observamos os preços aumentando do início do eixo x até o seu final, podemos observar a quantidade vendida (eixo y) diminuindo.

Há inúmeras outras maneiras gráficas de exibir dados. Outras muito utilizadas são os gráficos de barras e o gráfico de setores (ou gráfico de

pizza). Entretanto, esses são mais intuitivos, e para direcionar nossos esforços focaremos nos que foram apresentados.

Quanto ao gráfico de setores, há bastante discussão acerca de quando usá-lo, ou até mesmo se realmente deve ser usado. Nesse link tem uma abordagem bastante interessante quanto a isso: https://bit.ly/38UcsKQ.

A ferramenta R

Visando atender as demandas atuais do mercado, para esta disciplina a ferramenta adotada será o R.

- R é uma linguagem e também um ambiente de desenvolvimento integrado para cálculos estatísticos e gráficos.
- Foi criada originalmente por Ross Ihaka e por Robert Gentleman no departamento de Estatística da Universidade de Auckland, Nova Zelândia.
- É utilizado por profissionais em diversas áreas, como estatística, ciências sociais, saúde, psicologia, computação, dentre outras.
- É grátis.

A interface padrão do R é um prompt de comandos sem muitos recursos interativos com o usuário, e para ganhar produtividade e facilitar o uso muitas das vezes o RStudio é utilizado.

Figura 8 – Tela do R.

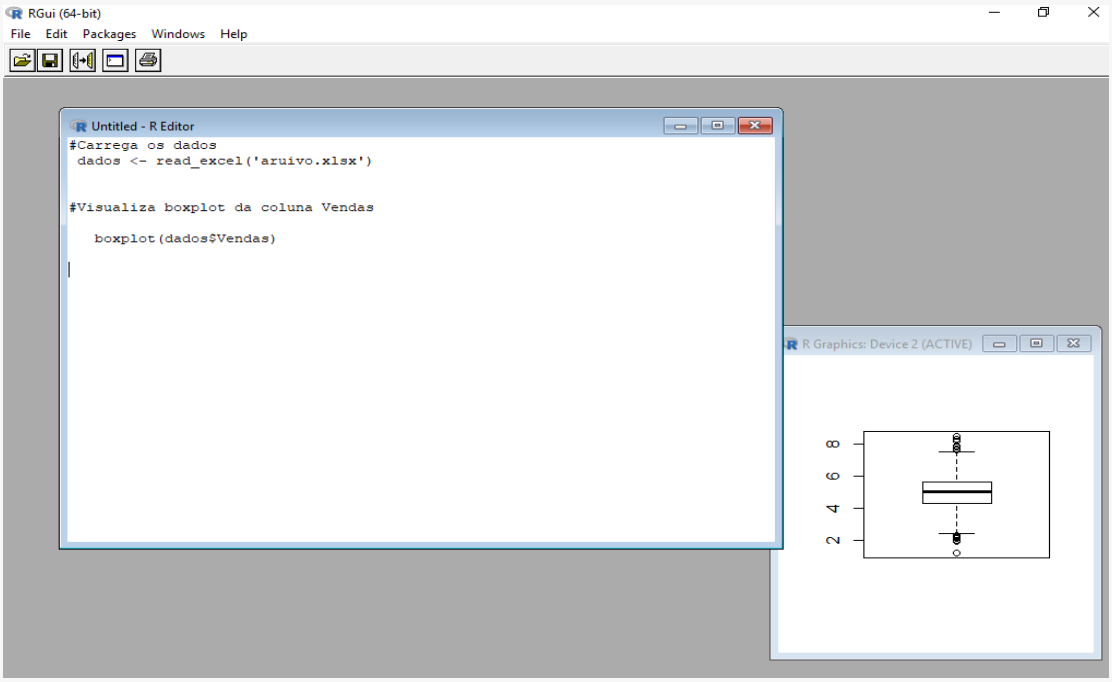
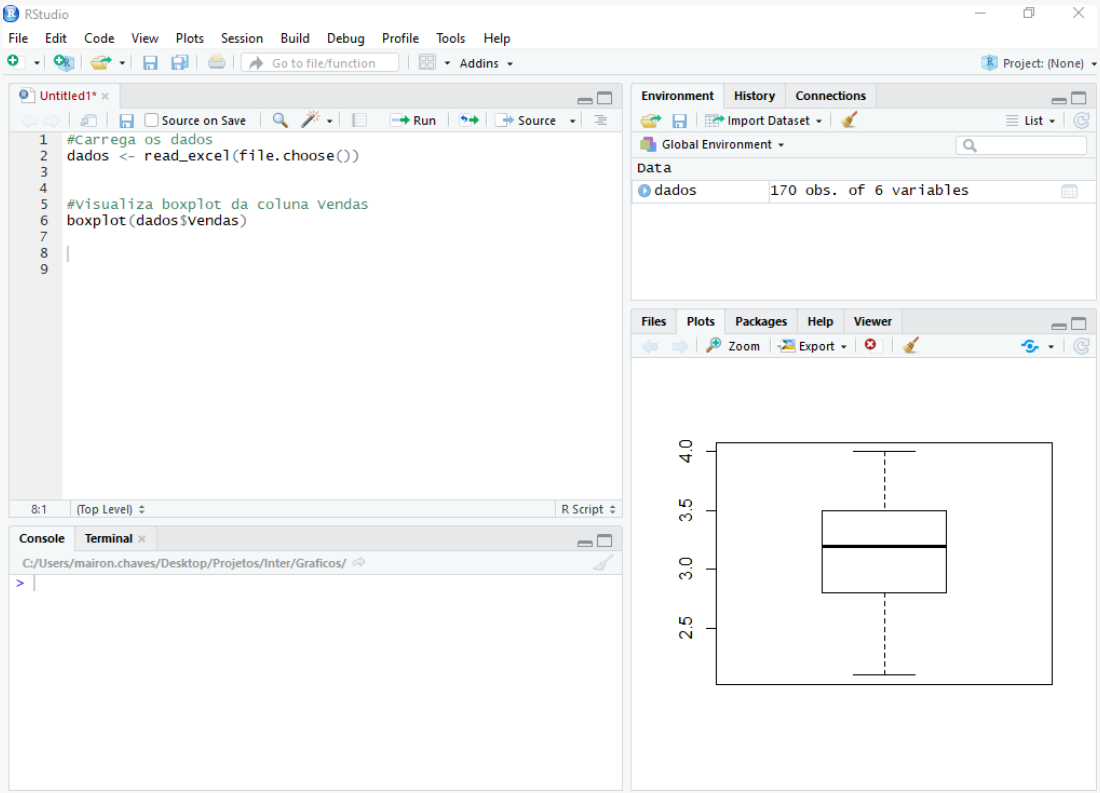


Figura 9 – Tela do RStudio.



Os objetos que mais utilizamos no R são:

Vetor - Uma coluna que representa alguma variável.

Preço
500,00
340,00
177,00
308,00

Matriz – Semelhante ao vetor, porém pode ter mais de uma variável, desde que sejam do mesmo tipo.

Preço	Lucro
500,00	123,12
340,00	201,00
177,00	78,00
308,00	234,20

Data Frame – Semelhante à matriz, porém aceita variáveis de todos os tipos.

Preço	Lucro	Categoria do Cliente	Data da Compra
500,00	123,12	A	01/05/2019
340,00	201,00	B	01/05/2019
177,00	78,00	A	02/05/2019
308,00	234,20	C	03/05/2019

Lista – Armazena outros objetos, que podem ser vetores, matrizes, data frames ou até mesmo outra lista.

Os principais tipos de dados são:

Numeric - Inteiro ou Decimal (int ou float).

Character – Texto (string).

Date Time – Data, hora.

Factor – Atribui codificação inteira ao dado.

Quando transformamos uma variável do tipo character para factor, internamente o R atribui uma codificação para ela. Isso é especialmente útil para algoritmos que não trabalham com valores categóricos e necessitam de dados de entrada numéricos.

Character

Estado
MG
SP
RJ
GO
MG
SP
MG
SP
GO
GO
MG
SP

Factor

Estado_MG	Estado_SP	Estado_RJ
1	0	0
0	1	0
0	0	1
0	0	0
1	0	0
0	1	0
1	0	0
0	1	0
0	0	0
0	0	0
1	0	0
0	1	0

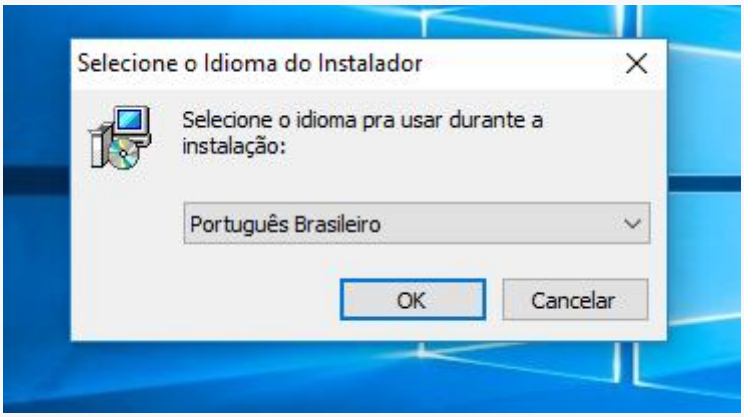
A instalação do R e do RStudio podem ser realizadas seguindo os passos a seguir:

Acesse a URL https://cran.r-project.org/bin/windows/base/ e clique em download.

Figura 10 – Instalando o R (parte 1 de 8).

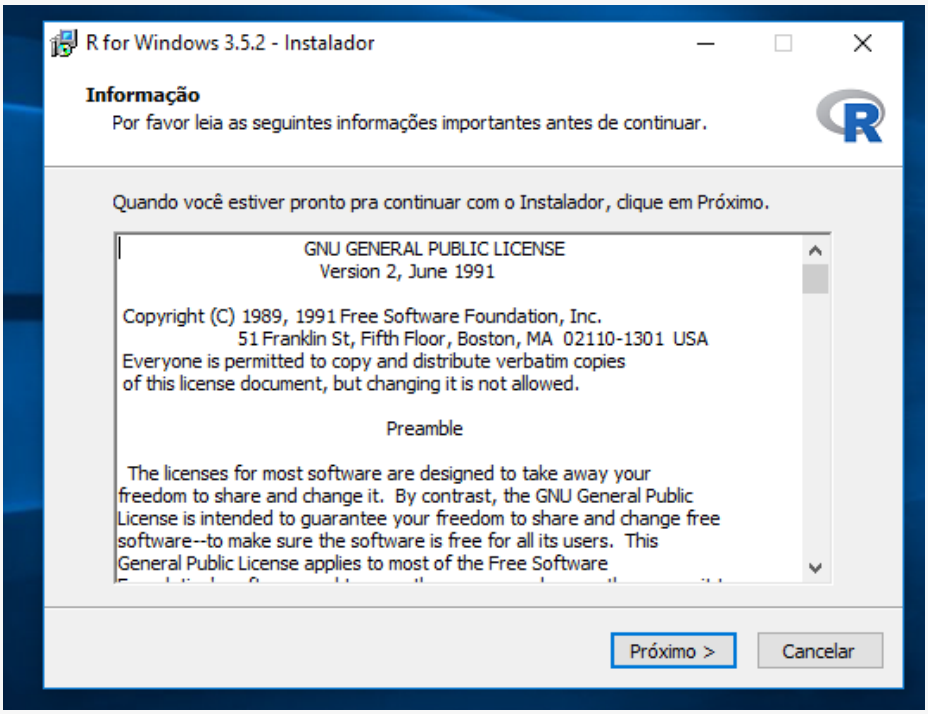
Escolha o idioma e clique em OK.

Figura 11 – Instalando o R (parte 2 de 8).



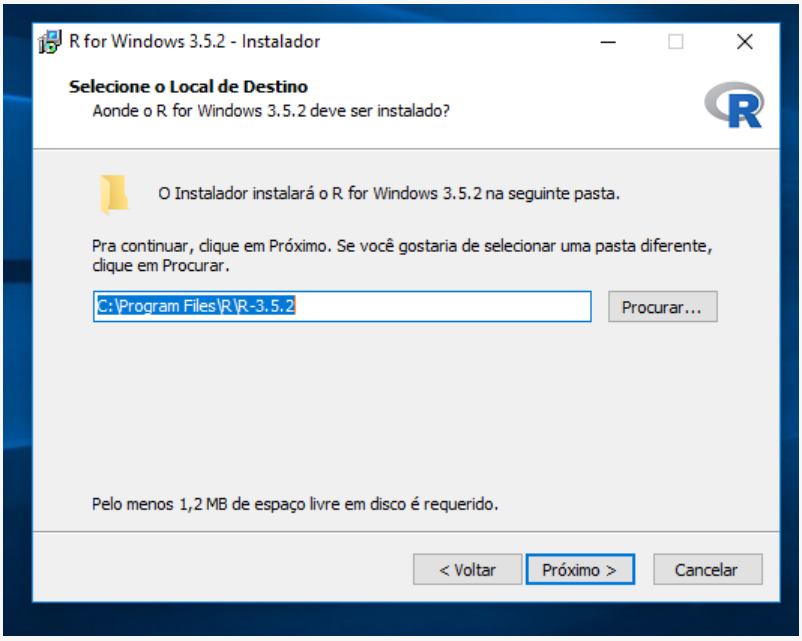
Leia a licença de uso caso tenha interesse e clique em próximo.

Figura 12 – Instalando o R (parte 3 de 8).



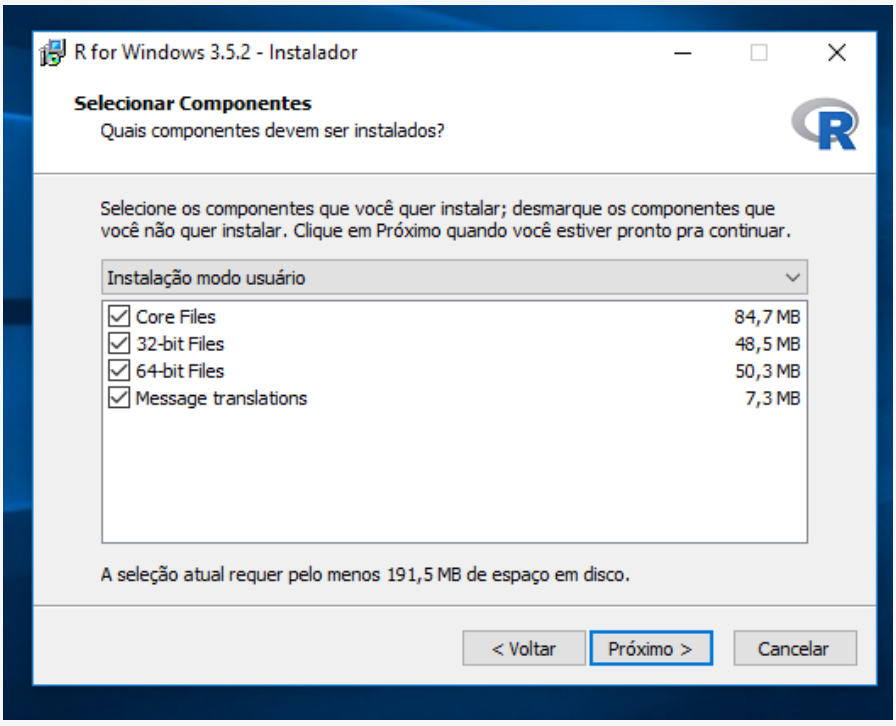
Selecione um diretório para instalação. Pode utilizar o diretório default.

Figura 13 – Instalando o R (parte 4 de 8).



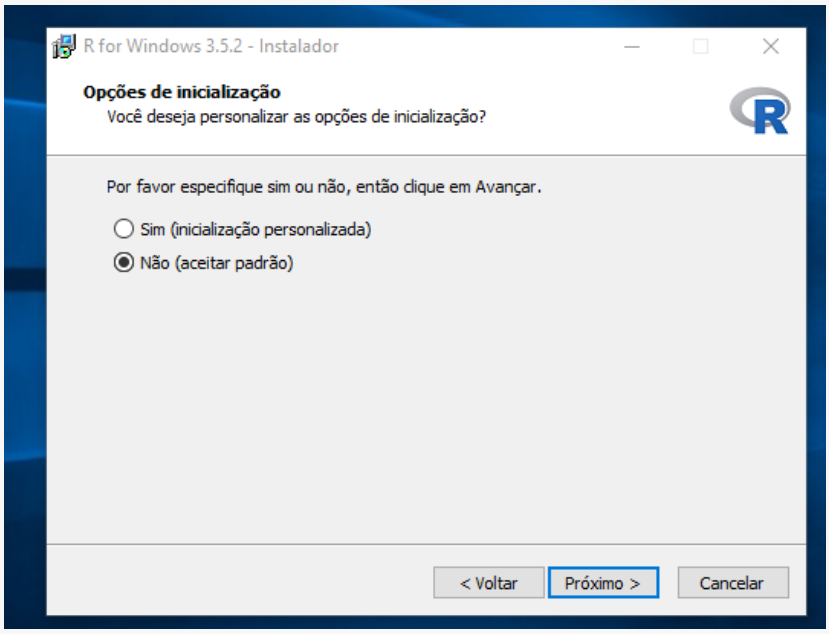
Selecione os componentes que serão instalados. Pode instalar todos.

Figura 14 – Instalando o R (parte 5 de 8).



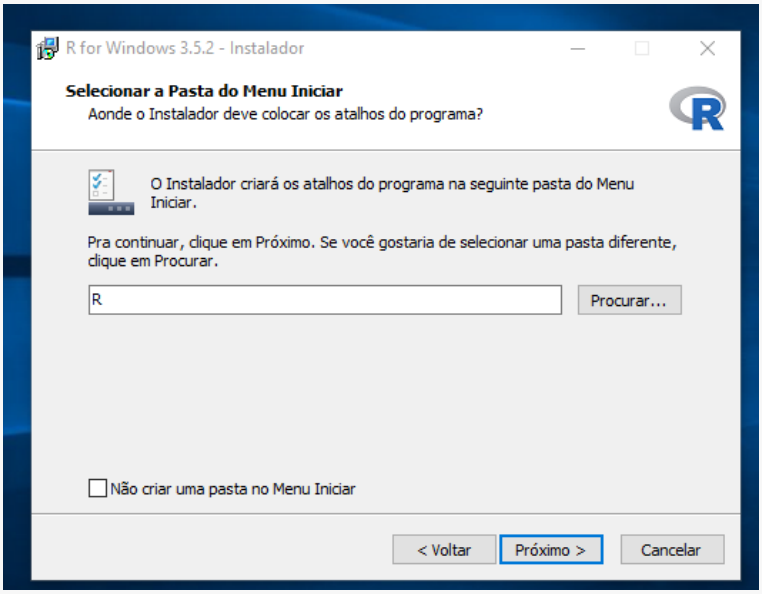
É possível customizar essa parte da instalação, mas neste momento é recomendado marcar a opção Aceitar Padrão.

Figura 15 – Instalando o R (parte 6 de 8).



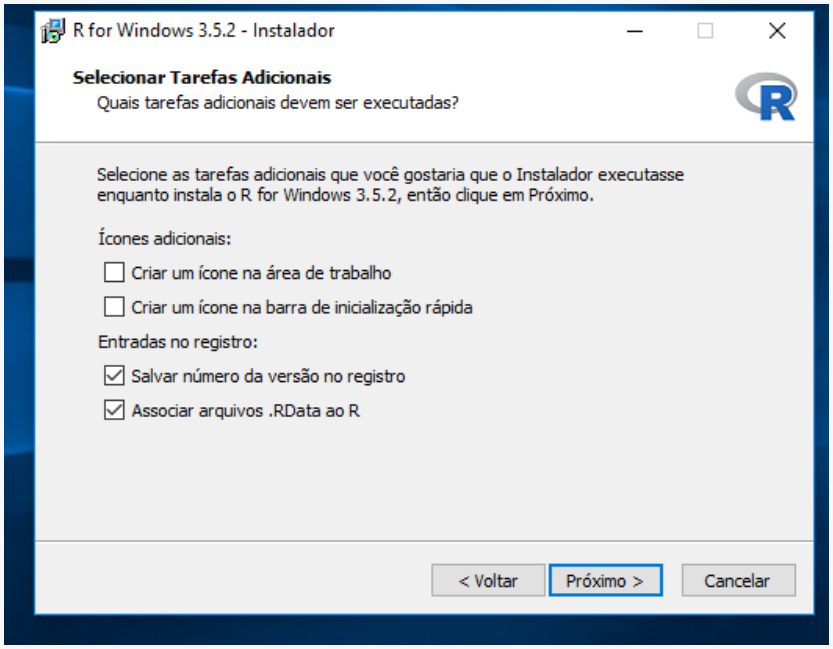
Pode customizar o nome que o atalho do R irá ter, mas é recomendado deixar as configurações default e seguir para a próxima tela.

Figura 16 – Instalando o R (parte 7 de 8).



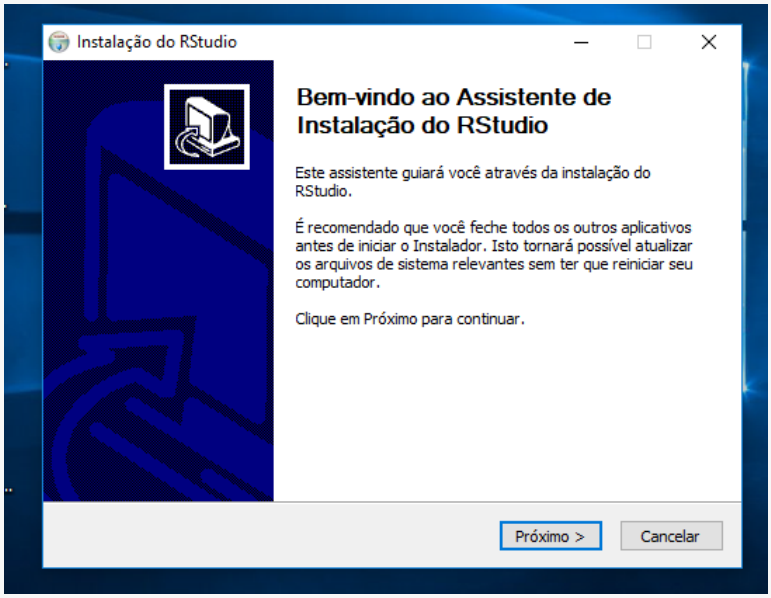
Em Ícones Adicionais, selecione se deseja criar atalhos e, em Entradas No Registro, marque as duas opções para que arquivos de extensão.RData sejam reconhecidos automaticamente pelo R.

Figura 17 – Instalando o R (parte 8 de 8).



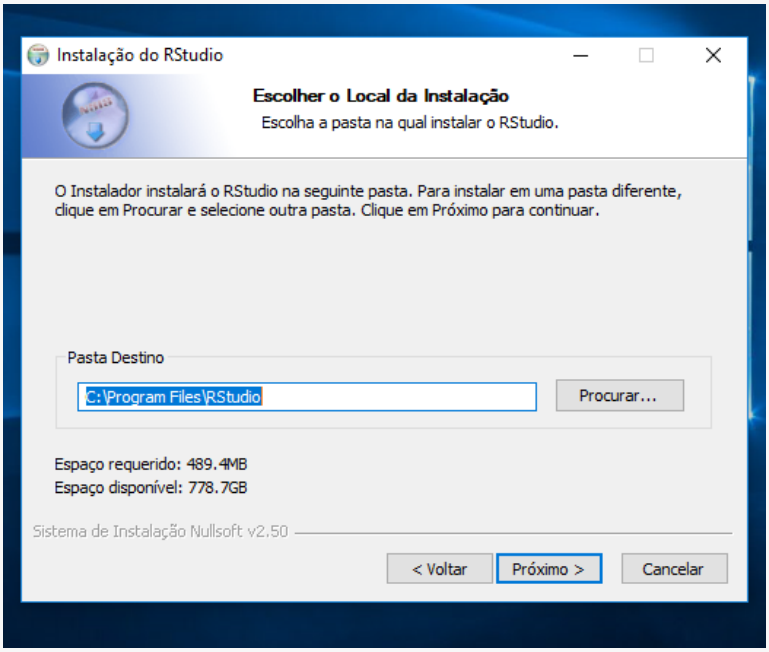
Uma vez instalado o R, iremos instalar o RStudio. Acesse a URL https://www.rstudio.com/products/rstudio/download/ e clique em download. Clique em baixar o primeiro item, é a versão gratuita do RStudio. Na tela em seguida, escolha o instalador mediante o sistema operacional de sua máquina, seja Windows, Linux ou Mac. Ao iniciar o instalador, os seguintes passos podem ser seguidos.

Figura 18 – Instalando o RStudio (parte 1 de 3).



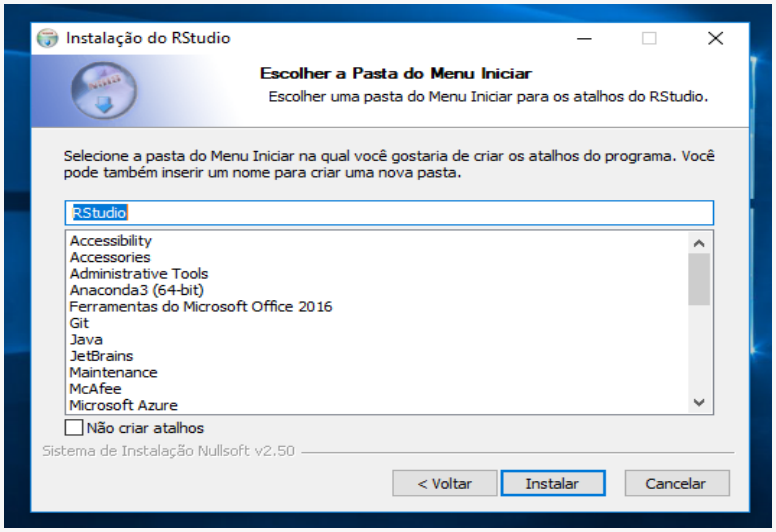
Selecione o diretório e clique em Próximo. É recomendado deixar o diretório default.

Figura 19 – Instalando o RStudio (parte 2 de 3).



Nessa etapa poderá ser criado atalho no menu iniciar. É recomendado utilizar o default e clicar em Instalar.

Figura 20 – Instalando o RStudio (parte 3 de 3).



Para iniciar o RStudio, basta clicar em seu logo na barra de tarefas ou buscar pelo software. O R irá rodar automaticamente dentro do RStudio.

Estatística Computacional – Análise Exploratória de Dados com o R

```
######## Análise exploratória de dados ########
##      AED - Capitulo 01  - Prof. Máiron Chaves     ###
#Copie este código, cole no seu R e execute para ver os resultados
rm(list=ls(all=TRUE)) #Remove objetos da memória do R
#Cria o data frame contendo o histórico de vendas do cafe
dados <- data.frame(Vendas_Cafe = c(18, 20, 23, 23, 23, 23, 24, 25, 26, 26, 26, 26, 27, 28, 28,
                                    29, 29, 30, 30, 31, 31, 33, 34, 35, 38, 39, 41, 44, 44, 46),
                    Preco_Cafe = c(4.77, 4.67, 4.75, 4.74, 4.63, 4.56, 4.59, 4.75, 4.75, 4.49,
                                   4.41, 4.32, 4.68, 4.66, 4.42, 4.71, 4.66, 4.46, 4.36, 4.47, 4.43,
                                   4.4, 4.61, 4.09, 3.73, 3.89, 4.35, 3.84, 3.81, 3.79),
                    Promocao = c("Nao", "Nao", "Nao", "Nao", "Nao", "Nao", "Nao", "Nao", "Sim",
                                 "Nao", "Sim", "Nao", "Nao", "Sim", "Sim", "Nao", "Sim", "Sim",
                                 "Sim", "Nao", "Nao", "Sim", "Sim", "Sim", "Nao", "Sim", "Sim",
```

```
                    "Sim", "Sim", "Sim"),
          Preco_Leite = c(4.74, 4.81, 4.36, 4.29, 4.17, 4.66, 4.73, 4.11, 4.21, 4.25,
                          4.62, 4.53, 4.44, 4.19, 4.37, 4.29, 4.57, 4.21, 4.77, 4, 4.31,
                          4.34, 4.05, 4.73, 4.07, 4.75, 4, 4.15, 4.34, 4.15) )
#visualiza a media (mean) e outras estatisticas descritivas das variaveis
summary(dados)
#Visualiza desvio padrao (standard deviation) das variaveis
sd(dados$Vendas_Cafe)
sd(dados$Preco_Cafe)
sd(dados$Preco_Leite)
#Visualiza atraves de um histograma a distribuicao da variavel Preco_Cafe
hist(dados$Preco_Cafe)
# Customizando o histograma
hist(dados$Preco_Cafe,
    col = 'blue',
    main = 'Distribuicao dos Preços Praticados para o Café')
#Visualiza o histograma das tres variaveis numericas na mesma pagina
par(mfrow=c(2,2)) #Configura layout para posicionar os graficos em duas linhas e duas
colunas
hist(dados$Vendas_Cafe,
    col = 'blue',
    main = 'Distribuicao das Vendas do Café')
hist(dados$Preco_Cafe,
    col = 'blue',
    main = 'Distribuicao dos Preços do Café')
hist(dados$Preco_Leite,
    col = 'blue',
    main = 'Distribuicao dos Preços do Leite')

dev.off() #limpa os graficos e volta o layout para configuracao normal
#Visualiza relacao entre as vendas do café o preço do café
plot(y = dados$Vendas_Cafe,
```

```
    x = dados$Preco_Cafe)
#Customiza o grafico
plot(y = dados$Vendas_Cafe,
    x = dados$Preco_Cafe,
    pch = 16,
    col = 'blue',
    xlab = 'Preço',
    ylab = 'Quantidade Vendidade',
    main = 'Relação entre o Preço e as Vendas do Café')
grid() #este comando adiciona linhas de grade ao grafico
#Colore os pontos em que havia promoção naquele dia
plot(y = dados$Vendas_Cafe,
    x = dados$Preco_Cafe,
    col = dados$Promocao,
    pch = 16,
    xlab = 'Preço',
    ylab = 'Quantidade Vendidade',
    main = 'Relação entre o Preço e as Vendas do Café')
#adiciona legenda
legend(x=4.4,y=45,
      c("Promoção","Sem_Promoção"),
      col=c("red","black"),
      pch=c(16,16))
grid()
#Cria uma nova variavel informando se naquele dia vendeu acima ou abaixo da media
historica
media <- mean(dados$Vendas_Cafe) #armazena a media em uma variavel
variavel <- ifelse(dados$Vendas_Cafe > media,
                   'Acima_da_media',
                   'Abaixo_da_media')
variavel <- factor(variavel) #converte nova variavel para factor
plot(variavel) #grafico com a qtde abaixo e acima da media
```

```
table(variavel) #visualiza a qtde abaixo e acima da media
#Gera boxplot das vendas
boxplot(dados$Vendas_Cafe)
#Gera boxplot do preco
boxplot(dados$Preco_Cafe)
#Gera boxplot comparativo das vendas quando houve promocao e de quando nao houve
boxplot(dados$Vendas_Cafe~dados$Promocao)
#Customizando o boxplot
boxplot(dados$Vendas_Cafe~dados$Promocao,
        col = 'gray',
        pch = 16,
        xlab = 'Promoção',
        ylab = 'Vendas',
        main = 'Vendas com promoção vs Vendas sem promoção')
```

> Capítulo 2

Capítulo 2. Distribuições de Probabilidade

As distribuições de probabilidades são funções matemáticas que nos ajudam a modelar a incerteza sobre variáveis do mundo real. Em ambiente de negócios, a incerteza está contida em diversos momentos, por exemplo:

- As vendas irão diminuir nos próximos seis meses?
- O percentual de turnover dos funcionários irá aumentar nos próximos meses?
- Será que nosso investimento em publicidade traz retorno nas vendas?
- Qual o melhor preço para colocarmos em um determinado produto?
- Qual o melhor dia da semana para promocionar um produto?
- Ao lançar um novo produto, em qual filial da nossa empresa teremos maior chance de sucesso?
- Um cliente tem mais chances de comprar um produto infantil quando a criança está junto?

Se aprendermos a modelar os fenômenos de forma probabilística, a incerteza continuará existindo, é claro, mas teremos melhores ferramentas para lidar com ela e subsidiar as tomadas de decisões.

O padrão matemático é que a probabilidade seja expressa como uma fração ou número decimal entre 0 e 1. Por exemplo, ao jogar uma moeda para cima e observar o resultado, teremos 0,5 (ou 50%) de chances de ter cara ou ter coroa. Iremos aprofundar nesses conceitos no decorrer desse capítulo e do curso.

Leis de Probabilidade e Diretrizes para sua Aplicação

Para encontrar a probabilidade de um determinado evento ocorrer, podemos utilizar a probabilidade frequentista. Sendo A um evento aleatório qualquer, podemos encontrar a probabilidade de A utilizando a probabilidade frequentista da seguinte forma:

$$P(A) = \frac{Número\ de\ Vezes\ que\ o\ evento\ A\ ocorreu}{Número\ total\ de\ observações}$$

De tal forma que:

$$0 \leq P(A) \leq 1$$

Reforçando o que foi dito no início deste capítulo, a probabilidade de um evento é sempre um valor entre 0 e 1 (ou em percentual 0% e 100%).

Utilizando um exemplo hipotético, vamos supor que desejamos saber a probabilidade de um cliente realizar uma compra ao entrar em nossa loja. Ao fazer um levantamento dos dados, observamos que 1000 clientes entraram em nossa loja, e desses, 500 compraram. Logo, a probabilidade de um cliente comprar fica:

$$P(Comprar) = \frac{Número\ total\ de\ clientes\ que\ entraram\ e\ compraram}{Número\ total\ de\ clientes\ que\ entraram}$$

$$P(Comprar) = \frac{500}{1000}$$

$$P(Comprar) = 0{,}5 \quad (50\%)$$

Ao trabalharmos com probabilidades, é fundamental definirmos o evento de interesse, (chamado de sucesso), que não necessariamente é algo bom, mas sim o que estamos interessados em estudar. Nesse caso, o evento de interesse é o cliente comprar.

A notação para o evento de interesse (sucesso) ocorrer é o número 1, e para o evento de não interesse (fracasso) é o 0. Em nosso exemplo, temos que:

$$P(1) = P(Comprar) = 0{,}5$$

Portanto, a probabilidade de o evento não ocorrer é o espaço complementar:

$$P(0) = P(Não\ Comprar) = 1 - P(Comprar) = 1 - 0{,}5 = 0{,}5$$

E se a probabilidade de comprar fosse 0,6? A probabilidade de não comprar seria:

$$P(Não\ Comprar) = 1 - 0{,}6 = 0{,}4$$

Como a probabilidade não passa de 1 (100%), a seguinte propriedade deve ser sempre respeitada:

$$P(Sucesso) + P(Fracasso) = P(1) + P(0) = 1$$

Ou seja, a probabilidade de o evento de interesse ocorrer mais a probabilidade de ele não ocorrer deve fechar em 1 (100%).

Duas regras fundamentais no estudo da teoria das probabilidades são as **Regras Aditivas** e as **Regras Multiplicativas**.

Para que não fique abstrato, vamos propor um contexto e, posteriormente, utilizá-lo como exemplo para entender o que são as Regras Aditivas e as Regras Multiplicativas.

Suponha que você tenha um restaurante em que uma promoção foi lançada. Como cortesia, para cada cliente será sorteada de forma aleatória uma sobremesa. Existem disponíveis dez sobremesas, nas quais:

- 1 sobremesa possui cobertura de menta.

- 2 sobremesas possuem cobertura de chocolate.
- 3 sobremesas possuem cobertura de morango.
- 1 sobremesa possui cobertura de chocolate e cobertura de morango.
- 3 sobremesas possuem cobertura de baunilha.

Qual a probabilidade do cliente receber uma sobremesa com cobertura de menta ou uma sobremesa com cobertura de chocolate?

O operador **ou** nos informa que pelo menos um dos eventos deve ocorrer. Podemos observar que o evento "receber sobremesa com cobertura de menta" e o evento "receber sobremesa com cobertura de chocolate" não podem ocorrer ao mesmo tempo, pois nenhuma sobremesa vai com as duas coberturas, portanto, eles são chamados de eventos **mutuamente exclusivos.**

Para aplicar a regra aditiva, devemos calcular a probabilidade de cada evento ocorrer.

Temos 10 tipos de sobremesas, nas quais:

1 possui cobertura de menta, portanto a probabilidade do cliente receber aleatoriamente uma sobremesa com cobertura de menta é $\frac{1}{10}$ (ou 10%).

3 sobremesas possuem cobertura de chocolate, portanto, a probabilidade de o cliente receber aleatoriamente uma sobremesa com cobertura de chocolate é de $\frac{3}{10}$ (ou 30%).

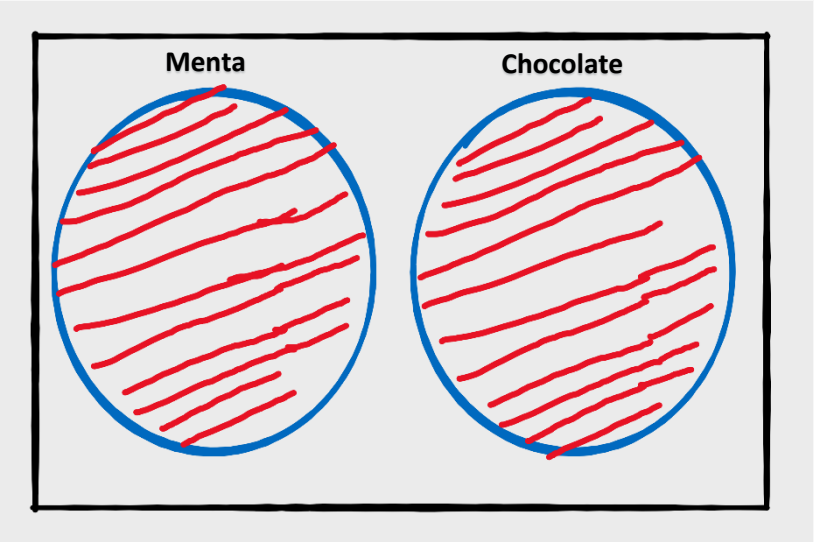
De acordo com a regra aditiva, para obter a probabilidade de um **ou** outro evento aleatório ocorrer, devemos somar suas respectivas probabilidades. Ou seja:

$$P(Menta)\ ou\ P(Chocolate) = P(Menta) + P(Chocolate)$$

Essa regra significa a união entre os dois eventos:

$$P(Menta)\ \cup\ P(Chocolate)$$

Figura 21 – Eventos mutuamente exclusivos.



Como os dois eventos não podem ocorrer ao mesmo tempo, o conjunto intersecção é igual ao conjunto vazio.

$$P(Menta \cap Chocolate) = \emptyset$$

Sendo assim, a probabilidade do cliente receber uma sobremesa com cobertura de menta ou uma sobremesa com cobertura de chocolate é:

$$P(Menta)\ ou\ P(Chocolate) = P(Menta) + P(Chocolate) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}\ (ou\ 40\%)$$

E se desejarmos saber a probabilidade de ocorrência de um **ou** outro evento que não sejam mutuamente exclusivos (ou seja, esses eventos podem ocorrer ao mesmo tempo)? Por exemplo:

Sendo A = rei, B = paus. Sabemos que em um baralho de 52 cartas existem 4 reis, logo p(A) = 4/52. Como também existem 13 cartas de paus,

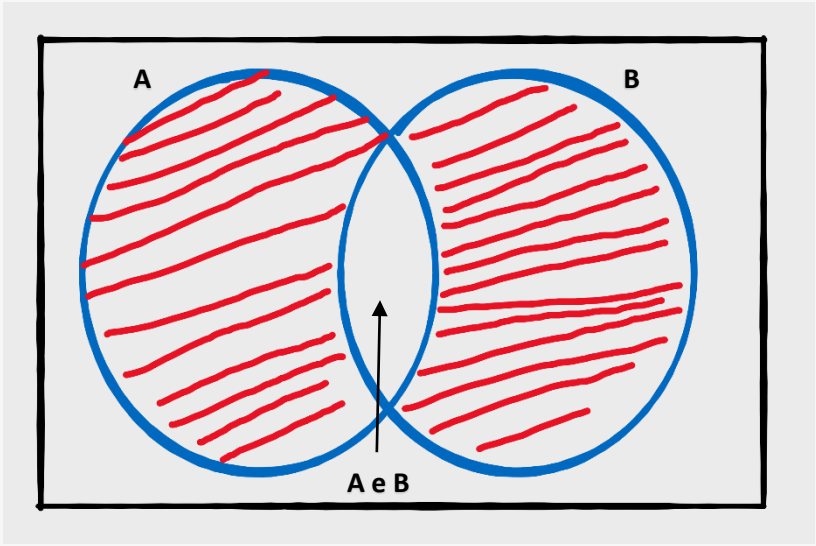
logo P(B)=13/52. Contudo, existe o rei de paus que é contabilizado tanto em A como em B.

Para o cálculo de P (A ou B) devemos desconsiderar a probabilidade de A e B ocorrerem juntos P(A e B), cujo valor é 1/52, p só há um rei de paus dentre as 52 cartas do baralho.

$$P(A\ ou\ B) = P(A) + P(B) - P(A\ e\ B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52}\ (ou\ 30{,}76\%)$$

Neste último exemplo, o conjunto intersecção foi removido, pois não estamos interessados na probabilidade de os dois eventos ocorrerem juntos.

Figura 22 – Removendo a intersecção da união dos dois eventos.



Uma vez que compreendemos a **Regra Aditiva** para obter a probabilidade de um ***ou*** outro evento aleatório ocorrer, vamos aprender como utilizar a **Regra Multiplicativa** para obter a probabilidade de um ***e*** outro evento aleatório ocorrer.

Supondo que além de sortear aleatoriamente uma das sobremesas para o cliente, também será sorteado um café.

Temos 2 tipos de cafés:

- Expresso
- Cappuccino

Qual a probabilidade do cliente receber uma sobremesa com cobertura de menta e em seguida um café expresso?

Veja que, agora, queremos obter a probabilidade de um evento ocorrer após um primeiro ter ocorrido, mas a probabilidade de um não influencia na probabilidade do outro. Portanto, o cálculo fica:

$$P(Menta)\ e\ P(Expresso) = P(Menta) * P(Expresso)$$

Como temos 2 tipos de cafés, a probabilidade do expresso ser sorteado é de $\frac{1}{2}$ (ou 50%). A probabilidade de uma sobremesa com cobertura de menta ser sorteada já sabemos que é de $\frac{1}{10}$.

Substituindo na fórmula fica:

$$P(Menta)\ e\ P(Expresso) = \frac{1}{2} * \frac{1}{10} = \frac{1}{20}\ (ou\ 5\%)$$

No entanto, e se os dois eventos foram dependentes, ou seja, dado que um evento ocorreu, a probabilidade do outro se modifica? Vamos supor que o cliente irá receber aleatoriamente duas sobremesas.

Dado que o cliente já recebeu uma sobremesa com cobertura de menta, qual a probabilidade de a próxima sobremesa sorteada ser de cobertura de baunilha?

Vamos por partes. Sabemos que 10 tipos de sobremesas podem ser selecionadas aleatoriamente, e sabemos que 1 tem cobertura de menta, então a probabilidade de uma sobremesa que tenha a cobertura de menta ser sorteada é de $\frac{1}{10}$.

Sabemos que existem 3 sobremesas que tem cobertura de baunilha, entretanto temos apenas 9 sobremesas restantes, pois uma já foi sorteada, que foi a que tem cobertura de menta.

Portanto, a probabilidade de sair aleatoriamente uma sobremesa com cobertura de baunilha dado que uma sobremesa com cobertura de menta já saiu, fica $\frac{3}{9}$ $(ou\ 33{,}33\%)$. Observe que o valor do denominador não é mais 10, e sim 9.

A notação para calcular a probabilidade de eventos dependentes fica:

$$P(Baunilha \mid Menta\ ) = P(Baunilha \mid Menta) * P(Menta)$$

Ou seja, a probabilidade de uma sobremesa com cobertura de baunilha ser sorteada dado que uma sobremesa com cobertura de menta já foi sorteada fica:

$$P(Baunilha \mid Menta\ ) = \frac{3}{9} * \frac{1}{10} = \frac{3}{90} (ou\ 3\%)$$

Variáveis Aleatórias Discretas e Contínuas

Antes de entrarmos nas distribuições de probabilidades, temos que entender os dois tipos mais importantes de variáveis aleatórias (ou v.a.).

Uma variável aleatória é classificada conforme a natureza do conjunto de valores que ela pode assumir. Os dois tipos de variáveis aleatórias mais importantes são as **variáveis aleatórias discretas** e as **variáveis aleatórias contínuas.**

- **Variável Aleatória Discreta**

Usualmente os valores de uma v.a. discreta são oriundos de um processo de contagem. Assumem valores inteiros.

Exemplos de v.a. discreta:

- Quantidade de ligações por dia que um call center recebe.
- Quantidade de clientes por hora que entram em uma loja.
- Quantidade de visitas “bem-sucedidas” (que geraram vendas) em cada n visitas realizadas por um vendedor.
- Quantidade média de veículos por hora que passam em um pedágio.

- **Variável Aleatória Contínua**

Os valores de uma v.a. contínua assumem um número infinito incontável de valores, ou seja, admite casas decimais. Geralmente são oriundos de algum processo de medição.

Exemplos de v.a. contínua:

- Tempo (em minutos) das ligações que um call center recebe.
- Valor (em reais) que os clientes compram em uma loja.
- Medições da altura (em cm) de uma determinada população.

Distribuições Discretas

Agora que já sabemos o que é uma variável aleatória discreta, vamos compreender algumas das principais distribuições de probabilidades discretas.

- **Experimento de Bernoulli**

É simplesmente a realização de uma tentativa de um experimento aleatório. Por exemplo, jogar a moeda para cima uma vez e observar se saiu cara ou coroa.

$$p(sucesso) = p(1) = p$$

$$p(fracasso) = p(0) = 1 - p$$

Onde p é a probabilidade de o sucesso ocorrer.

Vamos definir como sucesso para nosso exemplo sair coroa na face de cima da moeda. Sabemos que a moeda tem duas faces, uma cara e uma coroa. Então, as chances de uma coroa ocorrer é de $\frac{1}{2}$ (50%), e as chances de o sucesso não ocorrer, ou seja, do fracasso, é 1-p(sucesso).

São exemplos de experimentos de Bernoulli:

- Lançamento de uma moeda: sucesso = cara, fracasso = coroa
- Lançamento de um dado: sucesso = face 6, fracasso=faces 1, 2, 3, 4 ou 5
- Retirada (com reposição) de uma carta do baralho: sucesso=ás, fracasso=outra

Na prática, é o pesquisador que define qual evento será considerado o sucesso para determinado estudo.

- **Distribuição Binomial**

É o número de x sucessos em n tentativas. É a repetição de n experimentos de Bernoulli.

A função de probabilidade da distribuição Binomial é:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Onde **n** é o número de tentativas, **x** é o número de sucessos, **p** é a probabilidade de sucesso.

Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50%**(p).** Ao passar 10**(n)** clientes em nossa loja, qual a probabilidade de realizarmos 2 **(x)** vendas?

Para obter essa probabilidade, podemos substituir os valores na função de probabilidade da distribuição Binomial:

$$f(x) = \binom{10}{2} 0{,}5^2(1-0{,}5)^8$$

$$f(x) = 0{,}0439 \text{ (ou 4,39\%)}$$

Importante: Como foi citado no início desta apostila, as formulações matemáticas apresentadas têm o objetivo de manter o rigor na teoria estatística. Nosso foco não é nos cálculos, o R irá realizar todos para nós. Entretanto, ao aluno interessado na matemática, o professor da disciplina está à disposição.

- **Distribuição Geométrica**

É repetir um experimento de Beurnoulli x vezes até que o primeiro sucesso ocorra. Ou seja, é o número de fracassos até o primeiro sucesso.

A função de probabilidade da distribuição Geométrica é:

$$f(x) = (1-p)^{x-1}.p$$

Onde **x** é o número de tentativas, **p** é a probabilidade de sucesso.

Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50%**(p).** Qual a probabilidade da primeira venda ocorrer quando o quinto **(x)** cliente entrar na loja?

$$f(x) = (1-0{,}5)^{5-1}.0{,}5$$
$$f(x) = 0.03125 \text{ (ou 3,12\%)}$$

- **Distribuição Binomial Negativa**

É o número de x experimentos de Bernoulli até que uma quantidade r de sucessos ocorra. Pode ser vista como uma generalização da distribuição geométrica.

A função de probabilidade da distribuição Binomial Negativa é:

$$f(x) = \binom{x-1}{r-1}(1-p)^{x-r}.p^r$$

Onde **r** é a quantidade de sucessos, **x** é o número de tentativas, **p** é a probabilidade de sucesso.

Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50%**(p).** Qual a probabilidade de ter que entrar 8**(x)** clientes até que a segunda**(r)** venda ocorra?

$$f(x) = \binom{8-1}{2-1}(1-0{,}5)^{8-2}.0{,}5^2$$

$$f(x) = 0{,}02734 \text{ (ou 2,73\%)}$$

- **Distribuição de Poisson**

Expressa a probabilidade de um evento ou uma série de eventos ocorrerem em um determinado período de tempo ou espaço.

A função de probabilidade de distribuição de Poisson é:

$$f(x) = \frac{e^{-\boldsymbol{\lambda}}.\boldsymbol{\lambda}^x}{x!}$$

Onde **e**=2,71, **x!** é o fatorial de número de vezes que o evento ocorre, **λ** é o número de ocorrências de um evento aleatório em um determinado intervalo de tempo ou espaço.

Exemplo: Uma loja recebe em média, 6 **(λ)** clientes por minuto. Qual a probabilidade de que 5**(x)** clientes entrem em um minuto?

$$f(x) = \frac{e^{-6}.6^5}{5!}$$

$$f(x) = 0{,}1606 \text{ (ou 16,06\%)}$$

Distribuições Contínuas

Já sabemos o que é uma variável aleatória continua. Agora, vamos compreender algumas distribuições de probabilidades.

- **Distribuição Normal (Gaussiana)**

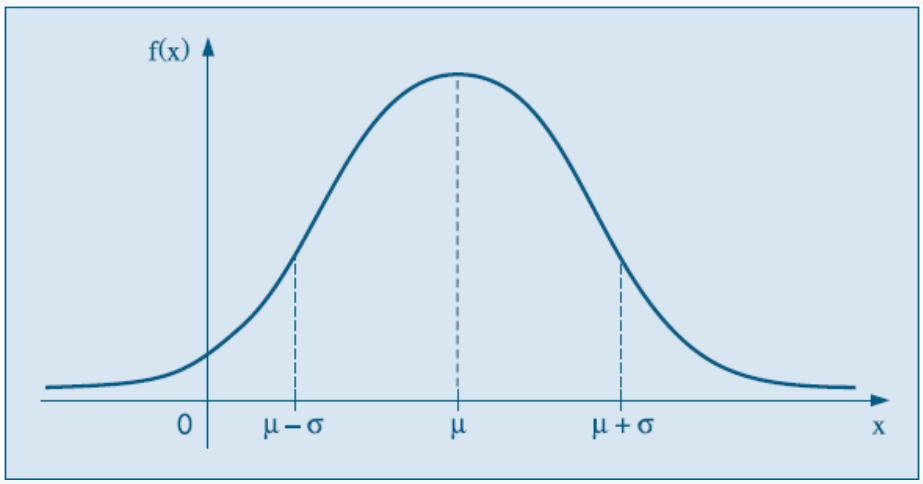
É uma das distribuições mais importantes. Conforme definição em Carvajal et al. (2009), a curva Normal ou Gaussiana descreve de forma muito adequada o comportamento de uma variável aleatória que se distribui de forma simétrica em relação a um valor central. Os dois parâmetros que a caracterizam são a média μ (que especifica o valor central) e a variância σ^2 (que define sua variabilidade em torno da média).

A função de densidade de uma v.a. que segue uma distribuição normal pode ser definida por:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Onde, **x** é o valor da variável aleatória, **μ** é a média, σ é o desvio padrão, π = 3,14, **e** = 2,71

Figura 23 – Distribuição Normal ou Gaussiana.



Fonte: Bussab & Morettin (1987).

Se uma v.a. chamada de X segue uma distribuição normal com média μ e desvio padrão σ, podemos representar pela notação: **X ~ N(μ, σ)**

Exemplo: Suponha que a distribuição dos salários dos funcionários de uma empresa siga uma distribuição normal com média μ=2.500 e desvio padrão σ= 170.

Ou seja: **Salário ~ N(μ=2500, σ=170)**.

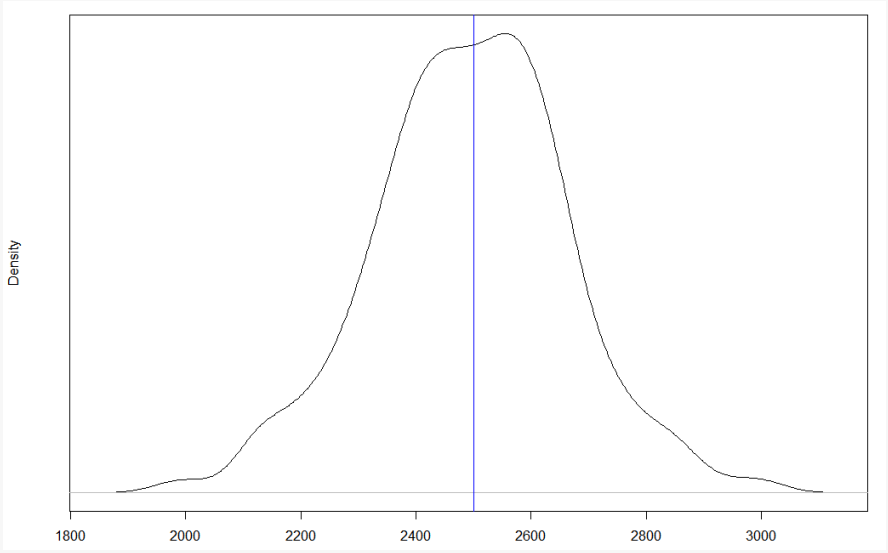
Ao selecionar aleatoriamente um indivíduo dessa população, qual a probabilidade de ter salário entre 2.400 e 2.600?

Antes de ir para a resposta, temos que entender alguns conceitos, pois nas distribuições contínuas, a interpretação é diferente de nas distribuições discretas.

Vamos por partes.

Primeiro, vamos plotar a distribuição dos salários. A posição da média está marcada com a barra azul.

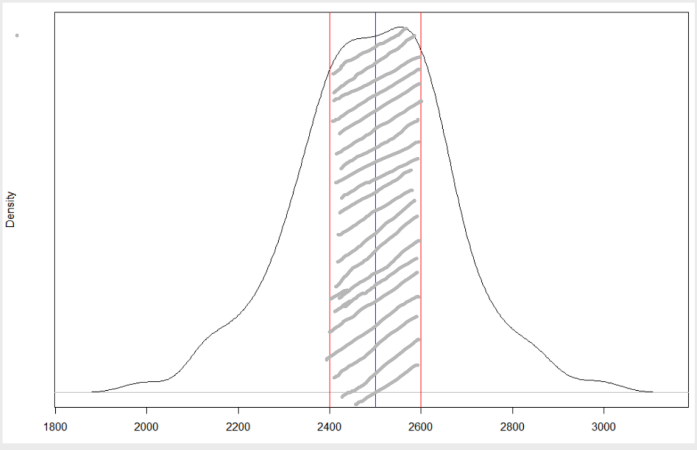
Figura 24 – Distribuição normal com μ=2500 e σ=170.



Quando trabalhamos com variáveis contínuas, sempre devemos utilizar um intervalo de valores para obter probabilidades, pois a probabilidade representa a área sob a curva de densidade, e a área sob a curva em um único ponto é zero. Portanto, não podemos utilizar a função de probabilidade para achar a probabilidade de o indivíduo ter por exemplo, o exato salário de 2.400. Mas podemos tomar um intervalo, como no enunciado acima, que nos pede para achar a probabilidade de um indivíduo ter o salário entre 2.400 e 2.600.

Vamos visualizar onde estão os valores 2.400 e 2.600 em relação à média.

Figura 25 – Distribuição normal com μ=2500 e σ=170, e de sombreado está o intervalo entre os valores 2.400 e 2.600.



Para calcular a área sob a curva para o intervalo sombreado, que corresponde ao intervalo entre 2.400 e 2.600, teríamos que aplicar uma integral definida.

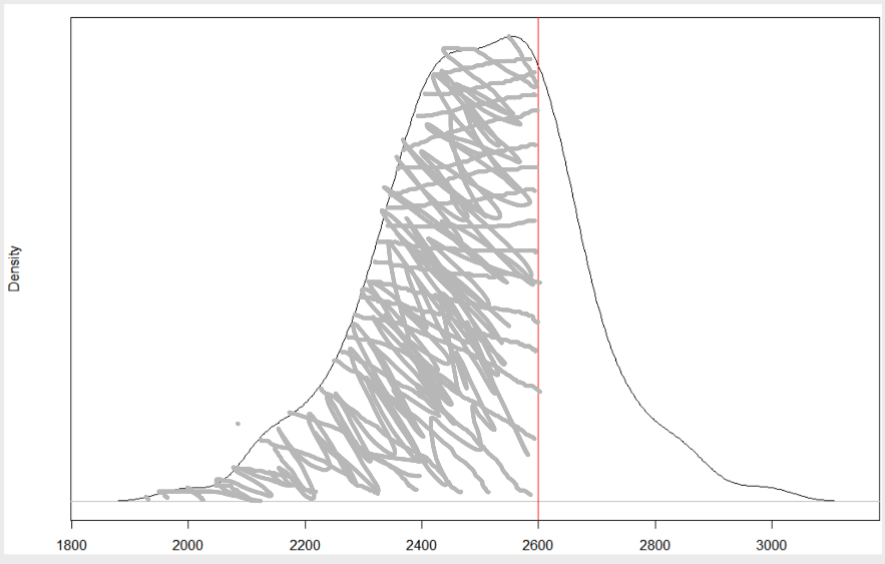
$$P(2.400 < Salario < 2.600) = \int_{2.400}^{2.600} f(x)\, dx,$$

Esse tipo de integral só é possui solução através de métodos numéricos, o que tornaria o trabalho bastante exaustivo na época que a distribuição normal foi descoberta. Para isso, na época, os pesquisadores criaram tabelas de probabilidades. Entretanto, para nossa felicidade, atualmente temos uma série de softwares que nos dão essa área sob a curva (probabilidade) facilmente.

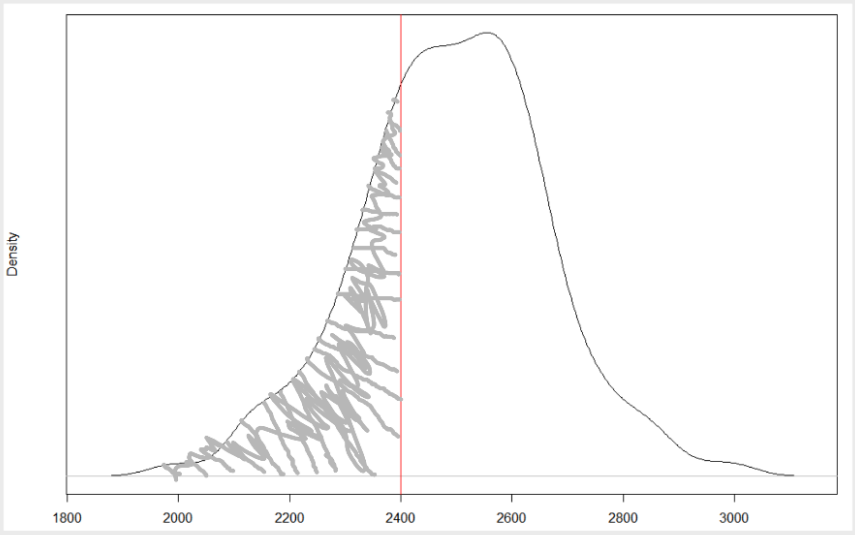
Enfim, para obtermos a probabilidade de um indivíduo dessa população com média μ=2.500 e σ=170 sorteado ao acaso ter o salário entre 2.400 e 2.600, devemos obter a probabilidade de um indivíduo ter o salário até 2.600 e subtrair pela probabilidade de o indivíduo ter o salário até 2.400.

Figura 26 – Sombreando as áreas de interesse.

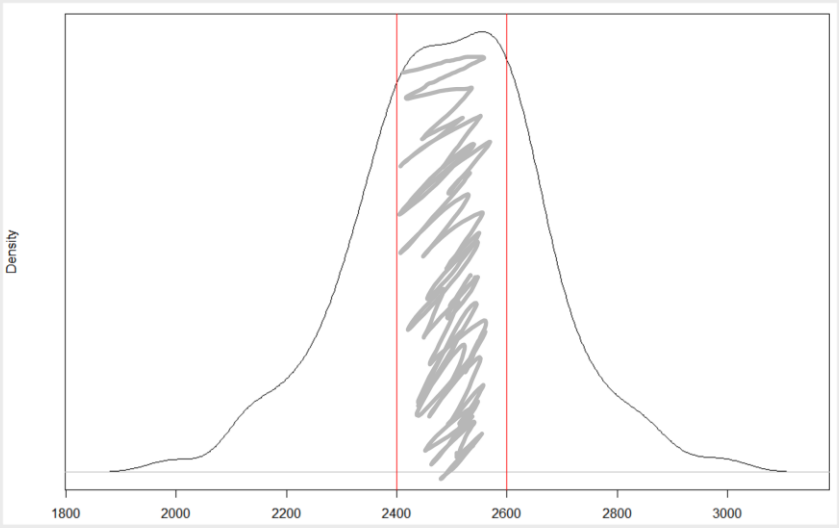
a) P(X < 2.600) , ou seja, Probabilidade do salário ser até 2.600



b) P(X< 2.400), ou seja, Probabilidade do salário ser até 2.400



c) P (2.400 ≤ X ≤ 2.600) = P(X < 2.600) – P(X < 2.400)



Utilizando o R para calcular as probabilidades, obteremos o resultado:

P (2.400 ≤ X ≤ 2.600) = P(X < 2.600) – P(X < 2.400)

P (2.400 ≤ X ≤ 2.600) = 0,7218 – 0,2781

P (2.400 ≤ X ≤ 2.600) = 0,4436

Conclusão: Ao retirarmos aleatoriamente um indivíduo dessa população, a probabilidade de ele ter um salário entre 2.400 e 2.600 é de 44,36%.

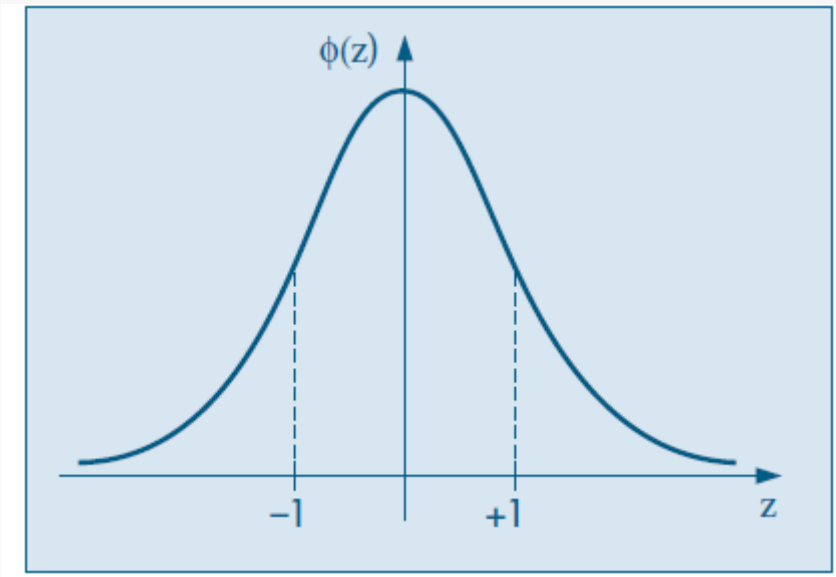
- **Distribuição Normal Padrão** (distribuição z)

É um caso especial da distribuição normal. Se uma variável aleatória segue uma distribuição normal, uma transformação (chamada de padronização) é aplicada de modo que essa variável tenha média zero e desvio padrão unitário. A equação a seguir demonstra como transformar uma variável aleatória normal em uma variável Z.

$$z = \frac{(x_i - \mu)}{\sigma}$$

Onde **x** é o i-ésimo valor da v.a., **μ** é a média da v.a. e **σ** é o desvio padrão da v.a..

Figura 27 – Distribuição Normal Padrão (distribuição Z).



Se uma v.a. chamada de X segue uma distribuição normal padrão com média zero e desvio padrão unitário, utilizamos a notação: **X ~ Z(μ=0, σ=1)**

Vamos aproveitar o contexto anterior para aplicar a distribuição Z e a sua tabela de probabilidades.

Exemplo: Suponha que a distribuição dos salários dos funcionários siga uma distribuição normal com média μ=2.500 e desvio padrão σ= 170.

Ou seja: **Salário ~ N(μ=2500, σ=170)**.

Ao selecionar aleatoriamente um indivíduo dessa população, qual a probabilidade de ter salário acima de 2.600?

O primeiro passo é padronizar o valor 2.600.

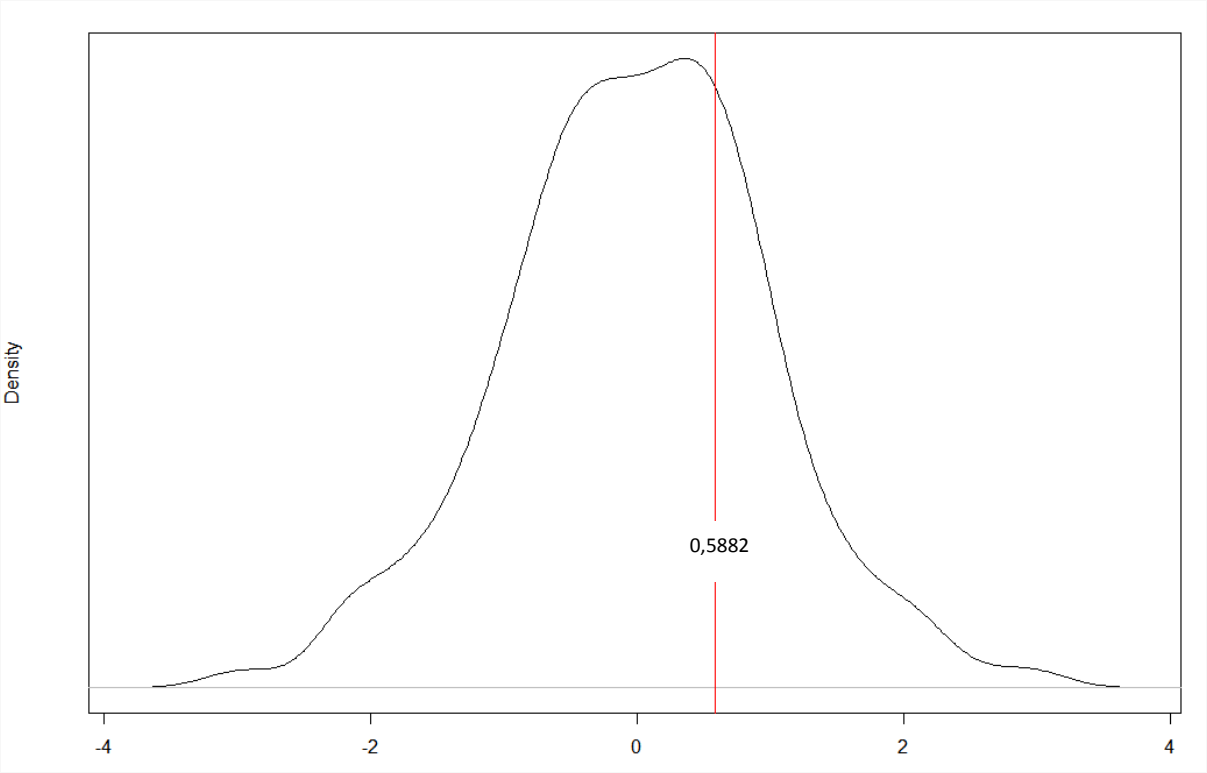
$$z = \frac{(x - \mu)}{\sigma}$$

$$z = \frac{(2600 - 2500)}{170}$$

$$z = 0{,}5882$$

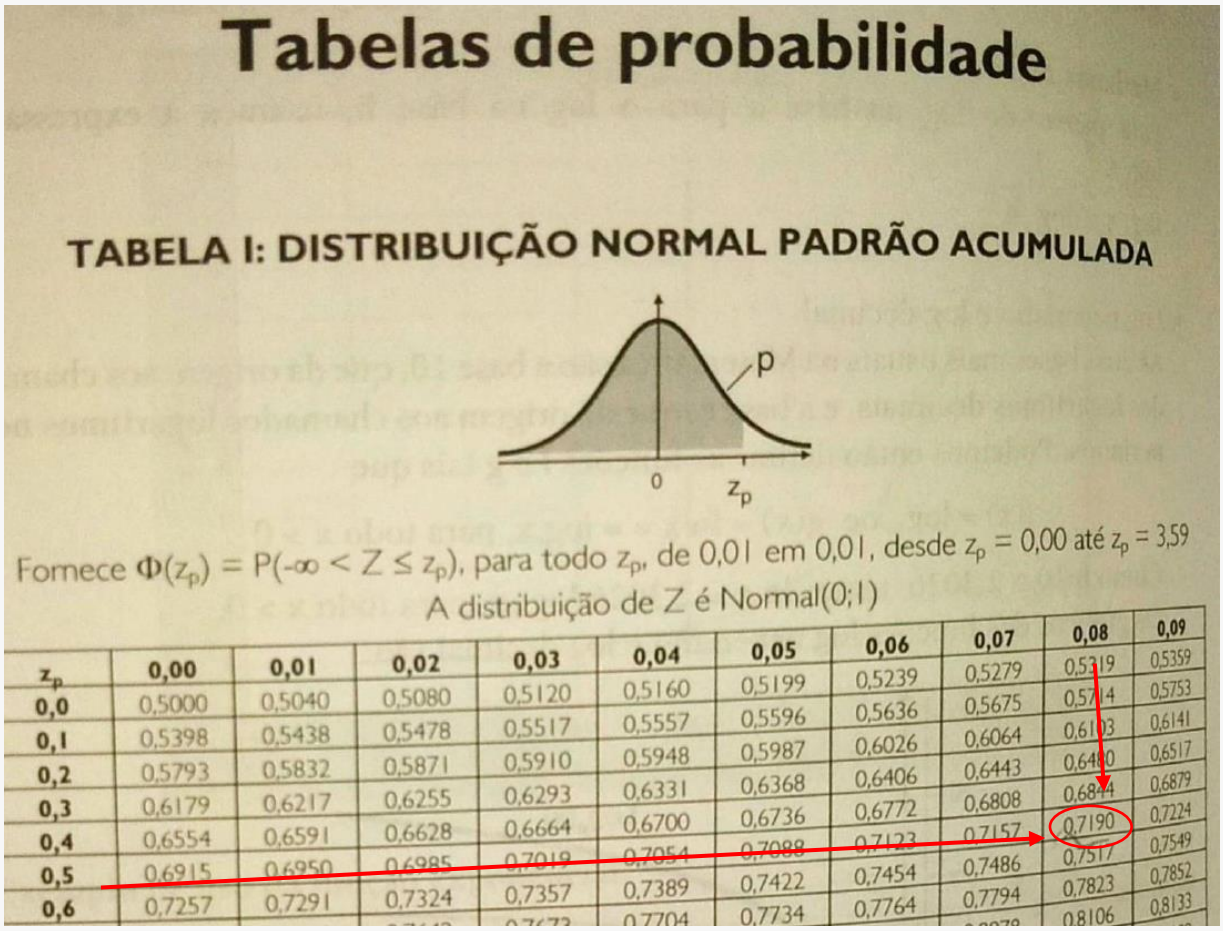
Já sabemos que a distribuição Z possui média μ=0 e σ=1. Vamos visualizar onde está nosso valor Z no gráfico:

Figura 28 – Visualizando Z na curva normal padrão.



O 0,5882 está na mesma posição em relação a média 0 do que o valor 2.600 em relação à média 2.500. Uma vez transformado, é possível utilizar a tabela a seguir para achar a área sob a curva (probabilidade). Devemos utilizar duas casas decimais para utilizar a tabela. Portanto, procuramos pelo valor 0,58.

Figura 29 – Tabela de Probabilidade da Distribuição Normal Padrão.



Tabelas de probabilidade

TABELA I: DISTRIBUIÇÃO NORMAL PADRÃO ACUMULADA

Fornece $\Phi(z_p) = P(-\infty < Z \leq z_p)$, para todo z_p, de 0,01 em 0,01, desde $z_p = 0,00$ até $z_p = 3,59$

A distribuição de Z é Normal(0;1)

z_p	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
								0,7794	0,7823	0,7852
					0,7704	0,7734	0,7764		0,8106	0,8133

Fonte: (Carvajal et al., 2009).

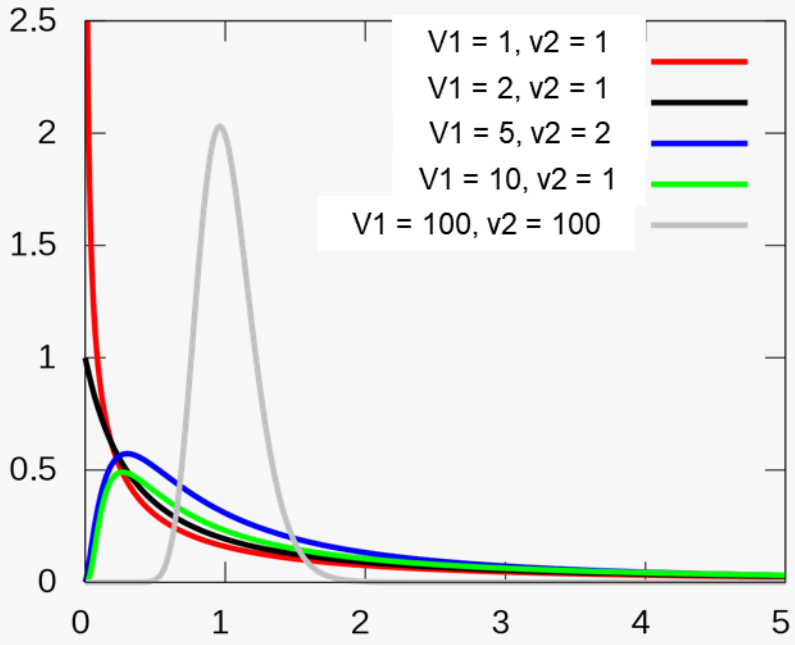
Na tabela acima, obtemos a probabilidade de 0,7190 para um valor Z=0,58. Portanto, a probabilidade de um indivíduo sorteado aleatoriamente ter um salário até 2.600 é de 71,9%.

Apesar do uso da tabela de probabilidades ainda ser extensivamente utilizado, principalmente em cursos de graduação de Estatística, a apresentação da tabela Z para obter probabilidades foi apresentada aqui apenas a título complementar. Em nosso curso, utilizaremos o R para calcular as probabilidades diretamente, sem o uso da tabela.

- **Distribuição F de Fisher-snedecor**

É uma distribuição positivamente assimétrica, não admite valores negativos. Geralmente é **utilizada para testar variâncias**, e depende de dois parâmetros chamados de graus de liberdade. Os graus de liberdade estão diretamente associados ao tamanho da amostra.

Figura 30 – Distribuição F com diferentes graus de liberdade.

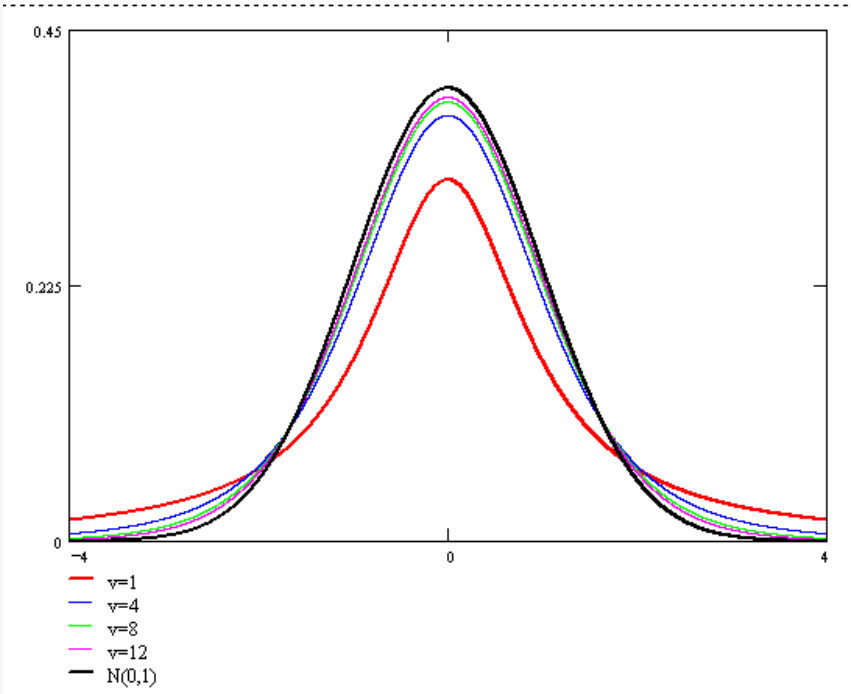


Se uma variável aleatória X segue uma distribuição F com v1 e v2 graus liberdades, então dizemos que: X ~ F(v1, v2).

- **Distribuição t de Student**

É simétrica e semelhante à curva normal padrão, e depende de um único parâmetro, que também é um grau de liberdade. É extensamente **utilizada para testar médias**.

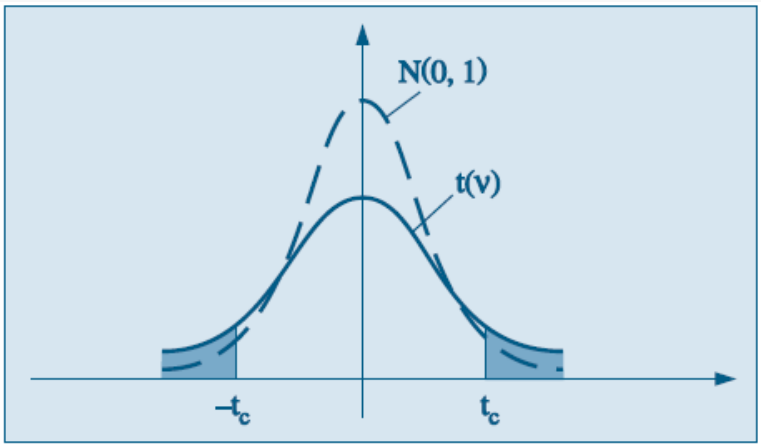
Figura 31– Distribuição t de Student com diferentes graus de liberdade e distribuição normal padrão em preto.



Observe que na medida que os graus de liberdade aumentam a distribuição Z se aproxima de uma distribuição normal padrão.

Na figura 32, é possível visualizar mais nitidamente o comportamento de uma distribuição t e de uma normal padrão. A distribuição t apresenta caudas mais longas de forma a comportar valores mais extremos.

Figura 32 – Distribuição t vs Distribuição normal padrão.



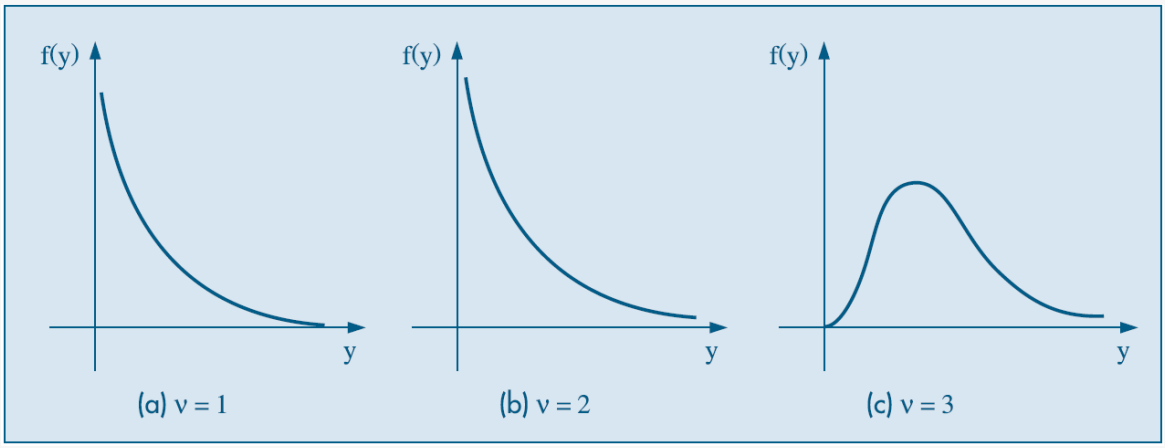
Fonte: (Bussab & Morettin, 1987).

Se uma variável aleatória X segue uma distribuição t com ν graus de liberdade, então dizemos que: X ~ t(ν).

- **Distribuição Qui-Quadrado**

O quadrado de uma v.a. com distribuição normal padrão é uma Qui-Quadrado. É frequentemente utilizado para **testar independência entre variáveis categóricas**.

Figura 33 – Distribuição Qui -Quadrado com 1, 2 e 3 graus de liberdade.



Fonte: (Bussab & Morettin, 1987).

Se uma v.a. contínua chamada de X, com valores positivos, tem uma distribuição qui-quadrado com ν graus de liberdade, dizemos que: X ~ χ2(ν).

Curiosidade: O quadrado de uma distribuição normal padrão é o mesmo que uma distribuição Qui-Quadrado com 1 grau de liberdade.

Estatística Computacional – Probabilidades com o R

```
##########################################################
##
#######      Distribuições de Probabilidades      #######
##      AED - Capitulo 02  - Prof. Máiron Chaves     ####
##########################################################
##
#Copie este código, cole no seu R e execute para ver os resultados
###################################
#### DISTRIBUIÇÃO BINOMIAL  ####
###################################
# Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50%.
# Ao passar 10 clientes em nossa loja, qual a probabilidade de realizarmos 2 vendas?
#Ou seja, queremos encontrar a probabilidade de dois sucessos, em dez tentativas. Cuja probabilidade de sucesso
# em cada tentativa é 50%


dbinom (x = 2-1, size = 10-1, prob = 0.5)
#Onde:
#  x é o número de sucessos,
#  size é o número de tentativas,
#  prob é a probabilidade de sucesso em cada tentativa


# A função a seguir gera quantidades aleatórias de sucesso oriundos de uma quantidade (size) de tentativas dada a probabilidade
#(prob) de sucesso.
# É útil para realizar experimentos. Podemos simular qual a frequencia esperada de vendas a cada dez clientes ?
```

```
#Ainda mantendo a probabilidade de sucesso (cliente comprar) de 50%

va_binomial <- rbinom(n = 30, size=10, prob=0.5)

#Onde:

# n é a quantidade de vezes que o experimento deve ser repetido

# size é o número de tentativas a cada experimento

# prob é o número de sucesso em cada uma das tentativas

hist(va_binomial) # A maior barra no histograma representa a quantidade esperada de vendas

#Ajuste o parametro n para 1000 e plote o histograma, observe como a distribuição binomial se aproxima da normal

# Podemos também querer a probabilidade de que até dois clientes comprem.

#Ao invés de saber a probabilidade de exatos dois comprarem.

#A probabilidade de até dois clientes comprarem é:

#(probabilidade de nenhum cliente comprar) + (probabilidade de um cliente comprar) + probabilidade de dois cliente comprarem)

#Formalizando: P(X<=2) = P(X=0) + P(X=1) + P(X=2)

pbinom(q = 2,size = 10, prob = 0.5)

#A probabilidade de que até dois clientes comprem ao entrarem dez clientes, é de 5,48%

####################################

#### DISTRIBUIÇÃO GEOMÉTRICA  ####

####################################

#Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50%.

#Qual a probabilidade da primeira venda ocorrer quando o quinto cliente entrar na loja?

dgeom(x = 5-1, prob = 0.5)

#Onde:

# x é o número de tentativas

# prob é a probabilidade de sucessos

# Podemos utilizar a mesma função para nos dar a probabilidade do sucesso ocorrer na primeira tentativa,

#Segunda tentativa, terceira tentativa ... até a décima tentativa.

va_geometrica <- dgeom(x = 1:10, prob = 0.5)

va_geometrica
```

```
plot(va_geometrica) #Veja como as probabilidades vão diminuindo. A probabilidade de sucesso de 50% é relativamente alta,

#então é muito provavel que o sucesso ocorra logo nas primeiras tentativas

# Podemos utilizar a distribuição geométrica acumulada para saber qual a probabilidade do primeiro sucesso

#ocorrer na primeira tentativa OU na segunda tentativa OU na terceira tentativa

#Formalizando, queremos: P(X<=3)

va_geometrica_acumulada <- pgeom(0:3, prob = 0.5)

plot(va_geometrica_acumulada)

##########################################

#### DISTRIBUIÇÃO BINOMIAL NEGATIVA  ####

##########################################

# Exemplo: Definindo como sucesso o cliente comprar, e supondo que a probabilidade de sucesso é 50%.

#Qual a probabilidade de ter que entrar 8 clientes até que a segunda venda ocorra?

dnbinom(x=2, size = 8, prob = 0.50)

#Onde:

# x é o número de sucessos

# size é a quantidade de tentativos

# prob é a probabilidade de sucesso

#Também pode ser calculado assim:

p <- 0.50

x <- 8

r <- 2

choose(x-1,r-1)*(p^r)*(1-p)^(x-r)

##########################################

####    DISTRIBUIÇÃO POISSON      ####

##########################################

# Exemplo: Uma loja recebe em média, 6 (λ) clientes por minuto. Qual a probabilidade de que 5(x) clientes

#entrem em um minuto?


dpois(x= 5,lambda = 6)
```

```
#Onde:

# x é a quantidade a ser testada

# lambda é a taxa média de ocorrêcia do evento em um determinado período de intervalo de tempo ou espaço

# Podemos utilizar a mesma funcao para obter a probabilidade de entrar um cliente, dois clientes... quinze clientes

va_poison <- dpois(x = 1:15, lambda = 6)

plot(va_poison)

# Observe que os valores se distribuiem simetricamente en tormo de seis, use acontece porque o paramentro

#lambda é a média (e também o desvio padrão) da distribuição de Poisson

# Também podemos obter a probabilidade acumulada de até 5 clientes entrarem na loja em um minuto

#Formalizando, queremos: P(X<=5)

va_poison <- ppois(1:5, lambda = 6)


plot(va_poison)

##########################################

####    DISTRIBUIÇÃO NORMAL      ####

##########################################

# Exemplo: Suponha que a distribuição dos salários dos funcionários de uma empresa sigam uma distribuição

#normal com média μ=2.500 e desvio padrão σ= 170.

# Ao selecionar aleatoriamente um indivíduo dessa população, qual a probabilidade de ter salário entre

#2.400 e 2.600 ?

# Precisamos achar a probabilidade do indivíduo ter um salário de até 2.600 e subtrair pela probabilidade do

#indivíduo ter o salário até 2.400



#P(X<=2600)

probabilidade_ate_2600 <- pnorm(q = 2600, mean = 2500, sd =170 )

#P(X<=2400)

probabilidade_ate_2400 <- pnorm(q = 2400, mean = 2500, sd =170 )
```

```
#P(X<=2600) - P(X<=2400)

probabilidade_ate_2600 - probabilidade_ate_2400

#Podemos gerar 100 números aleatórios para uma distribuição normal com média 2500 e desvio padrão 170

va_normal <- rnorm(n = 100, mean = 2500,sd = 170)

hist(va_normal)

##########################################

####  DISTRIBUIÇÃO NORMAL PADRÃO    ####

##########################################

# O comando scale() padroniza uma variável aleatória.

#Ao aplicar o comando na variável va_normal que acabmos de criar, ela ficará com média zero e desvio padrão unitário

va_normal_padrao <- scale(va_normal)

hist(va_normal_padrao)

# Exemplo: Suponha que a distribuição dos salários dos funcionários de uma empresa sigam uma distribuição

#normal com média μ=2.500 e desvio padrão σ= 170.

# Ao selecionar aleatoriamente um indivíduo dessa população, qual a probabilidade de ter

#salário acima de 2.600 ?

#Padronização

z <- (2600-2500)/170

pnorm(z, mean = 0, sd = 1)

#ou simplesmente

pnorm(z)

#Podemos também visualizar onde está o nosso valor Z em relação a média

plot(density(scale(va_normal))) #Plota curva de densidade

abline(v = 0,col = 'blue') #Gera uma linha sobre média, que é zero pois padronizamos a distribuição

abline(v = 0.58,col = 'red') #Gera uma linha sobre o valor z obtido

##########################################

####    DISTRIBUIÇÃO F        ####

##########################################


#Gerando uma amostra aleatória de 1000 número seguindo uma distribuição F
```

```
va_f <- rf( n= 1000, df1 =  5 , df2 =  33 )
# Onde:
# n é a quantidade de números a ser gerado
# df1 é o primeiro grau de liberidade
# df2 é o segundo grau de liberdade
hist(va_f)
#Vá aumentando os graus de liberdade e observe como a distribuição se aproxima da normal
#Informação Extra: Uma distribuição F é a razão entre duas chi-quadrado
##########################################
####     DISTRIBUIÇÃO T          ####
##########################################
#Gera  uma amostra aleatória de 1000 números seguindo uma distribuição T
va_t <- rt(1000, df = 2)

hist(va_t)
#Observe que a distribuição t, assim como a normal padrão, é centrada no zero
#Vá aumentando o grau de liberdade e observando o comportamento do histograma
##########################################
####    DISTRIBUIÇÃO QUI-QUADRADO    ####
##########################################
#Gera  uma amostra aleatória de 1000 números seguindo uma distribuição qui-quadrado
va_QuiQuadrado <- rchisq(1000,df = 3)
hist(va_QuiQuadrado)
```

> Capítulo 3

Capítulo 3. Intervalos de Confiança

Muitas vezes é interessante trabalhar com um intervalo de valores ao invés de com uma estimativa pontual. Por exemplo, imagine que você levantou um histórico de vendas mensais de vários anos e você precisa calcular a média aritmética de vendas em cada mês. Ao invés de dizer que o mês de julho vende em média 500 reais (estimativa pontual), você poderia dar uma estimativa intervalar. Por exemplo, no mês de julho, com 95% de confiança podemos dizer que a venda média varia de 460 a 540 reais (estimativa intervalar).

Os intervalos de confiança sempre vêm com um nível de confiança associado. Geralmente 80%, 90%, 95% ou 99%. Quanto maior o nível de confiança, maior o intervalo. E quanto menor a amostra, maior o intervalo (ou seja, quanto menor a amostra, maior é a incerteza ao gerar o intervalo de confiança). Veremos isso com mais detalhes posteriormente.

Teorema Central do Limite

Conforme definido por Carvajal et al. (2009), o Teorema Central do Limite (TCL) afirma que independente de qual seja a distribuição original de uma variável aleatória, a distribuição de suas médias se aproxima de uma distribuição normal à medida que o tamanho n da amostra cresce.

Portanto, mesmo que a distribuição da variável aleatória que estamos estudando não siga uma distribuição normal ou tenha uma distribuição desconhecida, a distribuição de sua média terá distribuição normal à medida que n aumenta. Iremos aprender como gerar uma distribuição de médias a partir de uma amostra e prosseguir com os cálculos de intervalo de confiança e testes de hipóteses.

Intervalo de Confiança para Média

Conforme definições tragas em Smailes e McGrane (2012), os intervalos de confiança para média são usualmente utilizados para representar uma média populacional. Ou seja, o pesquisador define uma população, extrai uma amostra (pois obter os dados de toda população é bastante custoso em vários sentidos) e, a partir da média amostral, gera um intervalo de confiança para a população. Ou seja, a partir de uma amostra, o pesquisador consegue dizer que a média da população está entre um limite inferior e um limite superior, dado um nível de confiança.

No ambiente de negócios essa abordagem é válida, mas geralmente utilizamos os intervalos de confiança com outro propósito. O objetivo seria fornecer limites dentro do que podemos administrar nossa incerteza. Ao invés de tomar uma decisão em cima de um único valor para média aritmética, podemos tomar decisão sobre um intervalo.

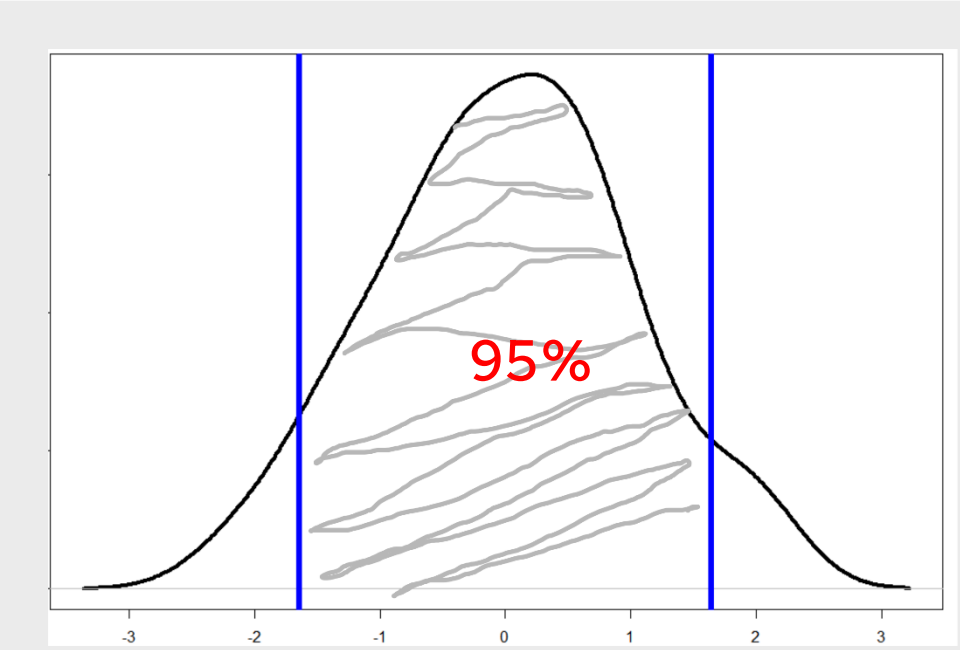
A fórmula para estimar um intervalo de confiança para média a partir de uma amostra é:

$$\bar{x} - 1{,}96 * \frac{s}{\sqrt{n}} < \mu > \bar{x} + 1{,}96 * \frac{s}{\sqrt{n}}$$

Onde $\bar{x}$ é a média amostral, s é desvio padrão amostral e n é o tamanho da amostra.

O valor 1,96 é oriundo da distribuição normal padrão, pois 95% dos dados na distribuição normal padrão estão entre -1,96 e 1,96. Portanto, se queremos que nosso intervalo tenha 95% de confiança, devemos utilizar 1,96. Se desejarmos que nosso intervalo tenha 90% de confiança, por exemplo, devemos substituir o 1,96 por 1,65, pois 90% dos dados na distribuição normal padrão estão entre -1,65 e 1,65. Esses valores são chamados de quantis

Figura 34 – Quantis para um intervalo de 95% de confiança na distribuição normal padrão.



Na figura 34 a primeira barra azul da esquerda para direita está em -1,96 e a segunda barra azul está em 1,96

Utilizando o histórico das vendas de café da figura 6 e assumindo que as vendas sigam uma distribuição normal, vamos calcular o intervalo de confiança para média de vendas.

A média de vendas é $\bar{x}$=30, o desvio padrão S=7,31 (supondo que também seja o desvio padrão populacional) e o tamanho da amostra é n = 30. Vamos adotar o nível de confiança de 95%. Consequentemente, utilizaremos o quantil de 1,96.

$$IC_{95\%} = \bar{x} - 1{,}96 * \frac{s}{\sqrt{n}} < \mu < \bar{x} + 1{,}96 * \frac{s}{\sqrt{n}}$$

$$IC_{95\%} = 30 - 1{,}96 * \frac{7{,}31}{\sqrt{30}} < \mu < 30 + 1{,}96 * \frac{7{,}31}{\sqrt{30}}$$

$$IC_{95\%} = 30 - 2{,}6158 < \mu < 30 + 2{,}6158$$

$$IC_{95\%} = 27{,}3842 < \mu < 32{,}6158$$

A resposta formal fica: Com 95% de confiança, a média de vendas está entre 27,38 e 32,61 unidades.

Intervalo de Confiança para Proporção

Podemos também calcular intervalo de confiança para uma proporção. Por exemplo, imagine que você está analisando as devoluções de um produto. Ao invés de colocar no seu relatório somente a proporção de clientes que devolveram, também pode colocar um intervalo de confiança. Vamos para um exemplo.

Suponha que n= 500 clientes foram escolhidos aleatoriamente, e que destes, 138 fizeram devolução do produto. Portanto, a proporção de clientes que realizaram devolução é 138/500 = 0,276 (ou 27,6%).

Podemos calcular um intervalo de 95% de confiança para essa proporção com a seguinte fórmula:

$$IC_{95\%} = \hat{p} - 1{,}96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1{,}96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Onde $\hat{p}$ é a proporção amostral, n é o tamanho da amostra, 1,96 é o quantil da distribuição Z para 95% de confiança.

Vamos substituir os valores do nosso exemplo na fórmula para obter o intervalo de confiança para a proporção de devoluções dos produtos.

$$IC_{95\%} = \hat{p} - 1{,}96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1{,}96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$IC_{95\%} = 0{,}276 - 1{,}96 * \sqrt{\frac{0{,}276(1-0{,}276)}{500}} < p < 0{,}276 + 1{,}96 * \sqrt{\frac{0{,}276(1-0{,}276)}{500}}$$

$$IC_{95\%} = 0{,}276 - 0{,}0391 < p < 0{,}276 + 0{,}0391$$

$$IC_{95\%} = 0{,}2369 < p < 0{,}3151$$

A resposta formal fica: Com 95% de confiança, a proporção de clientes que devolvem o produto é de 23,69% a 31,51%.

Como o intervalo de confiança para proporção usa a distribuição Z, também é necessário verificar se a variável estudada segue uma distribuição normal.

Intervalo de Confiança via Método Bootstrap

Grande parte dos livros clássicos de Estatística, escritos em eras não computadorizadas, utilizavam equações matemáticas para obter intervalos de confiança. Fórmulas como as que vimos acima, nas quais uma única amostra era extraída da população.

No entanto, e se a partir de uma amostra pudéssemos gerar 100, 1.000 ou 10.000 subamostras? Atualmente, com nossos computadores, podemos utilizar o método Bootstrap para realizar isso.

Segundo Bruce & Bruce (2019), o Bootstrap é uma ferramenta que pode ser usada para gerar intervalos de confiança para a maioria das estatísticas (média, mediana, desvio padrão, coeficientes de regressão etc.). O Bootstrap não exige nenhum pressuposto de normalidade para ser aplicado.

Dada uma (literalmente uma) amostra aleatória, o algoritmo Bootstrap para intervalo de confiança para média fica:

1- Extrair uma subamostra aleatória de tamanho n, com reposição.

2- Calcular a média da subamostra e registrar.

3- Repetir o passo 1 e 2 R vezes.

Ao final do processo, supondo que você repita o processo R=1.000 vezes, você terá uma variável com 1.000 valores (cada valor é a média calculada de uma subamostra). Essa nova variável gerada a partir das médias

da amostra única inicial é chamada de distribuição amostral. O Teorema Central do Limite nos garante que essa distribuição amostral seguirá uma distribuição normal, mesmo que a população não siga.

Para obter o intervalo de confiança de 95%, ordene as médias da menor para maior, para o limite inferior pegue o percentil 0,025 e para o limite superior pegue o percentil 0,975. Para 95% de confiança, utilizamos esses valores nos percentis, pois precisamos achar um intervalo de valores que 95% da nossa distribuição amostral gerada pelo Bootstrap esteja dentro dele.

Estatística Computacional – Intervalos de Confiança com o R

```
##########################################################
##
#######      Intervalo de Confiança          #######
##      AED - Capitulo 03  - Prof. Máiron Chaves     ####
##########################################################
##
#Copie este código, cole no seu R e execute para ver os resultados
########## Intervalo de confiança para média amostral pela distribuição Normal
Padrão ##############################
# Obter o intervalo de confiança para uma variável cuja média = 30, desvio padrão = 7,31 e n
= 30
#Temos que definir o nível de confiança do nosso intervalo.
#Podemos obter o valor do quantil para o nível de confiança desejado com a função qnorm()
#O quantil na distribuição normal padrão para 95% de confiança
ic <- 0.95
alfa <- 1-ic
1-(alfa/2)
qnorm(0.975)
#Vamos armazenar os valores em objetos
media <- 30
desvio_padrao_populacional <- 7.31
n <- 30
```

```
quantil_95 <- qnorm(0.975)

#Aplicando a fórmula vista na apostila fica:

Limite_Superior <- 30+quantil_95*(desvio_padrao_populacional/sqrt(n))

Limite_Inferior <- 30-quantil_95*(desvio_padrao_populacional/sqrt(n))

paste("Com 95% de confiança, podemos afirmar que a média varia entre",Limite_Inferior," e ",Limite_Superior)

########################  Intervalo de confiança para a média amostral pela distribuição t de Student ############

#A teoria nos diz para utilizar a distribuição t de Student quando não soubermos o desvio padrão populacional.

#Vamos assumir que o desvio padrão que temos é obtido a partir da amostra

#Vamos armazenar os valores em objetos

media <- 30

desvio_padrao_amostral <- 7.31

n <- 30

quantil_95_t <- qt(0.975,df = n-1)

#Aplicando a fórmula vista na apostila fica:

Limite_Superior_t <- 30+quantil_95_t*(desvio_padrao_amostral/sqrt(n))

Limite_Inferior_t <- 30-quantil_95_t*(desvio_padrao_amostral/sqrt(n))


paste("Com 95% de confiança, podemos afirmar que a média varia entre",Limite_Inferior_t," e ",Limite_Superior_t)

#Supondo que nossa variável já esteja em um data frame aqui no R, tem um comando para

#fornecer o intervalo de confiança de forma bem mais fácil

#Vamos gerar com o comando rnorm() uma variável aleatoria com média 30, desvio padrão 7,31 e n = 30

va <- rnorm(n = 30, mean = 30, sd = 7.31)

#Vamos visualizar a va gerada

hist(va)

#Calculando o intervalo de 95% de confiança com a distribuição t de Student

#com a funçao t.test()

IC <-t.test(va, conf.level = 0.95)

IC$conf.int

#Pronto, já temos o intervalo de confiança para média. Beeem mais fácil assim :)
```

```
########################  Intervalo de confiança para a proporção ############

#Utilizando o exemplo da apostila, onde calculamos o intervalo para proporção onde

# 138 de n = 500 clientes realizaram a devolução do produto

#Vamos armazenar os valores em objetos

devolucoes <- 138

n <- 500

quantil_95 <-qnorm(0.975)

proporcao_devolucoes <- devolucoes/n

#Aplicando a fórmula vista na apostila fica:

Limite_Superior_prop <-

 proporcao_devolucoes + quantil_95 *

 sqrt(proporcao_devolucoes*(1-proporcao_devolucoes)/n)

Limite_Inferior_prop <-

 proporcao_devolucoes - quantil_95 *

 sqrt(proporcao_devolucoes*(1-proporcao_devolucoes)/n)

paste("Com 95% de confiança, podemos afirmar que a proporção varia entre",Limite_Inferior_prop," e ",Limite_Superior_prop)

#Podemos obter o intervalo de confiança para proporção mais fácil pela função prop.test()

IC_proporcao <- prop.test(x = 138, n = 500, conf.level = 0.95)

IC_proporcao$conf.int

################  Intervalo de confiança para média via Bootstrap ############

#Vamos gerar uma va seguindo uma distribuição qui-quadrado

va <- rchisq(n = 60, df = 3)

#Observe o quão assimétrica é a va

hist(va)

#Inicializa variaveis

medias <- c() #Essa variável é um vetor para armazenar a média de cada subamostra bootstrap

R <- 1000 #Numero de subamostras extraídas para gerar a distribuição amostral de médias

#bootstrap

for (i in 1:R) {

 #Realiza uma subamostragem aleatória com reposição da va
```

```
  reamostra <- sample(va, size = 50, replace = T)
  #Armazena a média da subamostra
  medias[i] <- mean(reamostra)
}
#Distribuicao das médias das subamostras (distribuição amostral da média da va)
hist(medias)
#Observe que mesmo a variável original não seguindo uma distribuição normal, o Teorema Central do Limite
#nos garante que a distribuição das médias será normal se n é suficientemente grande
#A partir das médias geradas, precisamos achar dois valores, o que corta a cauda inferior
#e o que corta a cauda superior da distribuição. Lembrando que ela é simétrica
#Caso o intervalo desejado seja de 95% de confiança, temos que ordenar essa distribuição
#do menor valor para o maior e achar o valor que deixará 2,5% dos dados para trás e o
#valor que deixará 97,5% para trás
(1-0.95)/2
1-(1-0.95)/2
#Visualize o intervalo de confiança via bootstrap
quantile( medias, probs = c(0.025,0.975))
#Vamos realizar mais um experimento
#Geraremos uma va com média = 30 e desvio padrão amostral =7.31 e n = 30
va <- rnorm(n = 30, mean = 30, sd = 7.31)
#Iremos calcular o intervalo de confiança usando o Bootstrap e também com a distribuição
# t de Student. Compararemos os resultados.
#Inicializa variavel para armazenar as médias de cada subamostra
medias <- c()
R <- 10000 #Numero de subamostras extraídas para gerar a distribuição amostral de médias
#bootstrap
for (i in 1:R) {
  #Realiza uma subamostragem aleatória com reposição da va
   reamostra <- sample(va, size = 20, replace = T)
  #Armazena a média da subamostra
  medias[i] <- mean(reamostra)
```

```
}
#Distribuicao das médias das subamostras (distribuição amostral da média da va)
hist(medias)
#Limites inferior e superior do intervalo pelo bootstrap
quantile( medias, probs = c(0.025,0.975))
#Limites inferior e superior do intervalo via t de Student
IC<-t.test(va, conf.level = 0.95)
IC$conf.int
```

> Capítulo 4

Capítulo 4. Teste de Hipótese

É um procedimento estatístico que, por meio da teoria das probabilidades, auxilia na tomada de decisão no sentido de rejeitar ou não hipóteses em um experimento científico.

São exemplos de hipóteses:

- Quando colocamos o produto na posição A da gôndula, vende significativamente mais do que quando ele está na posição B?
- A equipe de vendas, após receber um treinamento, aumentou de forma significativa sua performance?
- Quando o cliente compra o produto A, ele também compra o produto B?
- A tendência de vendas ao longo dos meses parece estar aumentando, mas esse aumento é significativo ou são pequenas variações devido ao acaso?
- Clientes que entram em nossa loja acompanhados de criança compram mais do que aqueles que entram acompanhados de adultos?

Há algum tempo, testes de hipóteses eram utilizados praticamente só em laboratórios ou institutos de pesquisa, que coletavam os dados através de experimentos com pacientes ou aplicando questionários.

Hoje em dia, a ciência também está presente nas empresas e de questionários passamos para sistemas transacionais. O método estatístico já é amplamente utilizado no mundo dos negócios, e a tendência é que seu

uso aumente, já que estamos em uma era de tomada de decisão baseada em dados, e a Estatística é a ciência que analisa dados.

Passos para Execução de um Teste de Hipótese

Utilizaremos uma sequência de seis passos para nos auxiliar na elaboração de um teste de hipóteses. Para isso, vamos propor um contexto. Suponha que você e sua equipe estão analisando a melhor posição na gôndola para colocar um produto. A informação prévia é de que as vendas, quando o produto está na posição A da gôndola, possui média μ=R$140 com desvio padrão σ=R$27. Para comprovar essa hipótese, você conduziu um experimento. Colocou o produto na posição A da gôndola e observou n=15 dias de vendas. A média vendida nesses 15 dias foi de $\bar{x}$ =R$134.

E então, baseado na sua amostra, você rejeita ou não rejeita a hipótese afirmada no início do enunciado de que a venda média quando o produto está na posição A é R$140?

- **Passo 01 – Definir a hipótese nula (H_0) e a hipótese alternativa (H_1)**

Todo experimento científico gira sobre alguma hipótese a ser testada mediante fatos e evidências. Iremos definir como hipótese nula (H_0) a hipótese tida como verdade, e a hipótese alternativa (H_1) é a que será assumida como verdadeira caso existem evidências nos dados para rejeitar a hipótese nula. Vamos definir nossas hipóteses para o contexto proposto.

$$H_0: \mu = 140$$

$$H_1: \mu \neq 140$$

Nosso H_0 (tido como verdade até então) é de que a média (populacional) de vendas quando o produto está na posição A da gôndola é de R$140. Através da amostra que coletamos, iremos rejeitar ou não rejeitar essa hipótese probabilisticamente.

- **Passo 02 – Definir o nível de confiança e significância**

A ideia é a mesma do que vimos nos intervalos de confiança. Adotaremos uma distribuição de probabilidades e iremos definir o nível de confiança do nosso experimento. Aqui temos um novo conceito, que é o nível de significância. O nível de significância é a probabilidade de rejeitarmos H_0 mesmo ela sendo verdadeira. Isso é conhecido na Estatística como erro tipo I. Quando maior o nível de confiança, menor será a chance de cometer o erro tipo I. Por exemplo, se o nível de confiança adotado for 95%, automaticamente o nível de significância será 5%. Se o nível de confiança adotado for 99%, automaticamente o nível de significância será 1%. Utilizamos como notação a letra grega alfa α para representar o nível de significância.

- **Passo 03 – Calcular a estatística de teste**

Uma estatística de teste é um valor calculado a partir de uma amostra de dados. O seu valor é usado para decidir se podemos ou não rejeitar a hipótese nula.

A natureza do experimento irá direcionar qual distribuição de probabilidade usar. Por isso mesmo o pesquisador deve conhecer as distribuições para saber qual utilizar durante as situações cotidianas de pesquisa. Neste caso, estamos testando média e temos o desvio padrão da população. Portanto, adotaremos a distribuição normal padrão para nosso experimento. A fórmula para a estatística de teste usando a distribuição normal padrão é:

$$Z_{calculado} = \frac{(\bar{x} - \mu)}{\frac{\sigma}{\sqrt{n}}}$$

A fórmula para obter o Zcalculado é idêntica à da padronização, porém no denominador ponderamos o desvio padrão dividindo-o pela raiz quadrada de n.

Calculando o Z (nossa estatística de teste) para o exemplo proposto, fica:

$$Z_{calculado} = \frac{(134 - 140)}{\frac{27}{\sqrt{15}}}$$

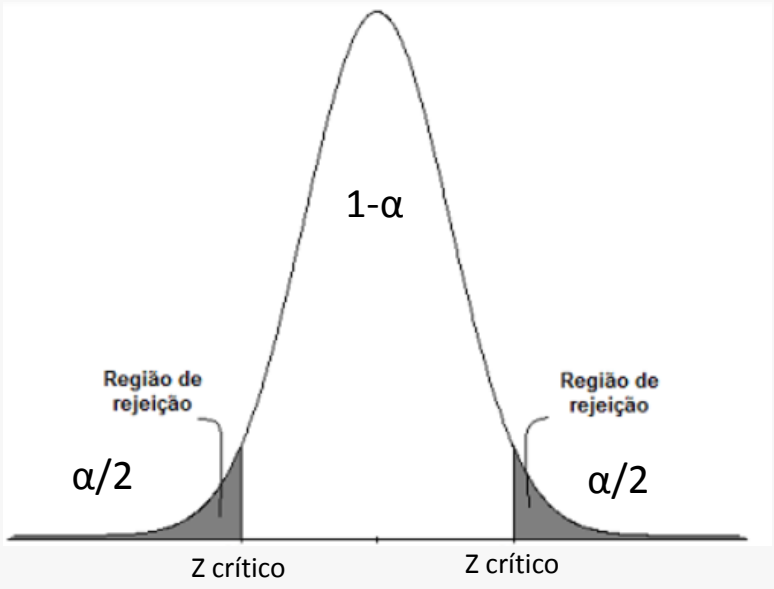
$$Z_{calculado} = -0{,}8606$$

Nos próximos passos iremos entender como utilizar o Zcalculado.

- **Passo 04 – Delimitar a região crítica**

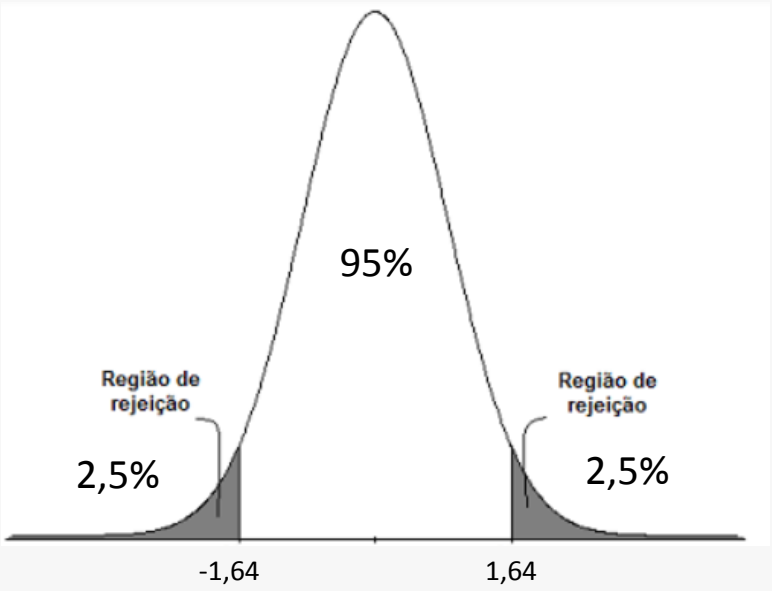
Baseado no nível de significância α que adotamos para o teste, iremos delimitar a região crítica. Iremos rejeitar H_0 se nossa estatística de teste se encontrar na região crítica. A região crítica também é conhecida como região de rejeição.

Figura 35 – Visualizando a região crítica bilateral na curva normal.



Adotando 95% de confiança para nosso teste (e consequentemente $\alpha = 5\%$), nossa região crítica ficaria:

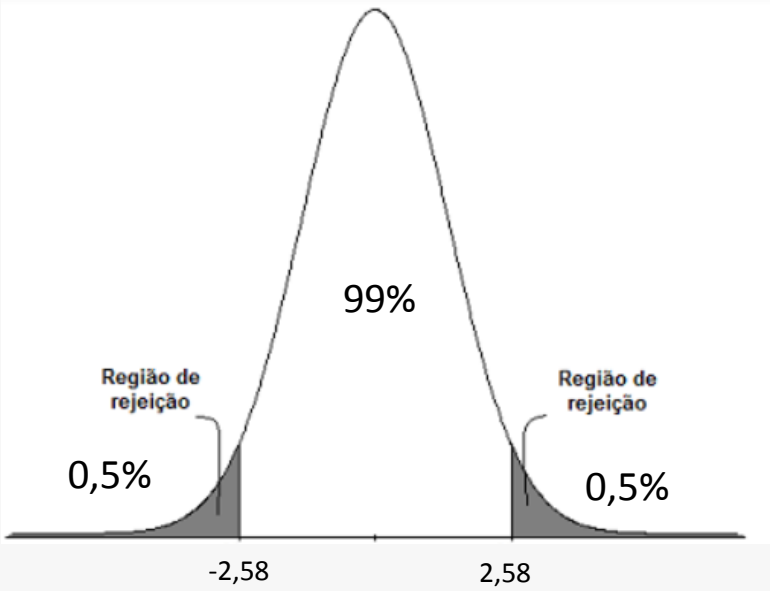
Figura 36 – Visualizando a região crítica bilateral na curva normal padrão para $\alpha = 5\%$.



Por que 1,96 é o valor que delimita a região crítica? Pois já vimos anteriormente que na curva normal padrão, 95% da amostra estará entre os quantis -1,96 e 1,96. Iremos rejeitar H_0 se o $Z_{calculado}$ estiver contido na região crítica, ou seja, se o $Z_{calculado}$ for menor que -1,96 ou maior que 1,96.

E se fôssemos mais exigentes e adotássemos 99% de confiança (e consequentemente $\alpha = 1\%$)? Podemos obter os valores críticos (quantis) utilizando o comando qnorm(0.995) no R. Obteríamos 2,58. Observe como a região crítica fica menor. Ou seja, nossas chances de rejeitar H_0 ficam menores. Isso nos diz que precisaremos de evidências mais fortes nos dados para poder obter uma estatística de teste que caia na região crítica.

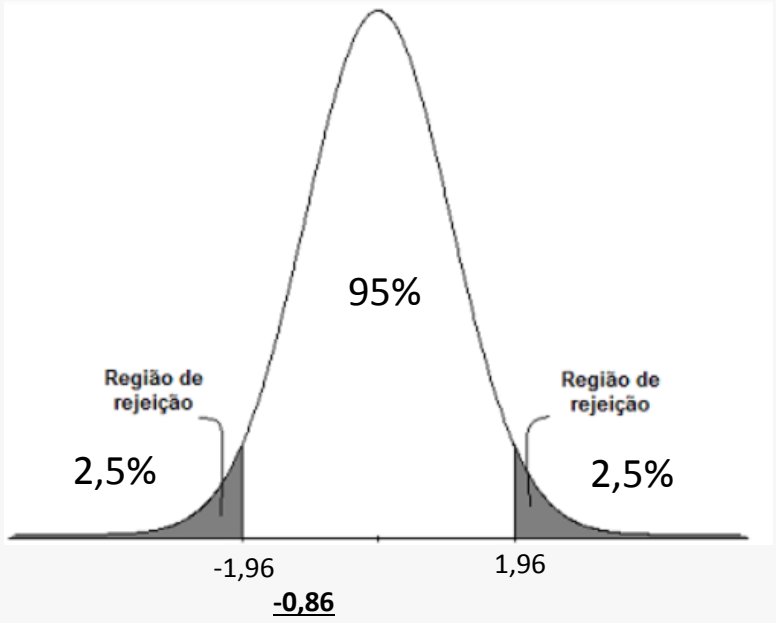
Figura 37 – Visualizando a região crítica bilateral na curva normal padrão para α = 1%.



- **Passo 05 - Obter o valor p**

O valor p (p-value) é a probabilidade observada nos dados de rejeitar a Hipótese Nula quando ela é verdadeira (erro tipo I). Ou ainda, a probabilidade da diferença ter ocorrido ao acaso. Seu valor é obtido a partir da estatística de teste calculada. Se o valor p for baixo o suficiente, devemos rejeitar H_0. Baixo o suficiente neste caso é quando obtemos um valor p abaixo do nível de significância α. Vamos visualizar onde está nosso $Z_{calculado}$ (que é -0,86) em relação aos quantis e a região crítica para 95% de confiança.

Figura 38 – $Z_{calculado}$ vs Região Crítica.



Observe que -0,86 está fora região crítica (ou região de rejeição). Isso nos diz que não temos evidências para rejeitar H_0. Em cursos ou disciplinas de Estatística é comum encerrar o teste por aqui, pois já é possível tirar uma conclusão. No entanto, em abordagens mais modernas com uso de computadores podemos obter o valor p, que será a probabilidade correspondente à estatística de teste na distribuição de probabilidades utilizada, que neste caso, é a normal padrão.

É comum em cursos tradicionais utilizar a tabela de probabilidades que vimos anteriormente, em que as integrais já estão calculadas. Em nosso caso, utilizaremos o R. Basta utilizar *pnorm(0.86)-pnorm(-0.86)* para identificar a probabilidade correspondente, que será 0,6102 (ou 61,02%). Observe que essa probabilidade é acima do nível de significância fixada que foi $\alpha = 5\%$. Nossa probabilidade de cometer o erro Tipo I é maior do que nós toleramos. Ou seja, como a probabilidade de estarmos errados ao rejeitar H_0 é alta, não devemos rejeitar.

- **Passo 06 – Rejeitar ou Não Rejeitar H_0**

O último dos 06 passos consiste em formalizar a conclusão para o teste. Nosso objetivo era testar se a média populacional das vendas é R$140. Para isso, coletamos uma amostra de dados: observamos o quanto vendemos em 15 dias e calculamos sua média, que foi R$134. Ao calcularmos a estatística de teste, vimos que ela caiu fora da região crítica que construímos utilizando a distribuição normal padrão a um nível de confiança de 95%. Ou seja, apesar da afirmação ser de que a média de vendas é R$140, ela tem um desvio padrão, e pelo fato de R$134 obter uma estatística de teste fora da região crítica, concluímos que a média amostrar obtida de R$134 é um valor que está dentro da variação natural da média populacional. Não temos evidências para discordar (rejeitar) da hipótese inicial de que a média das vendas quando o produto está na posição A é de R$140.

Já vimos que o Zcalculado está fora da região crítica e que, consequentemente, o valor p é maior do que o nível de significância fixado. Portanto, a resposta formal para concluir o teste fica:

Com 95% de confiança, não há evidências para rejeitar H_0. Ou seja, a média de vendas quando o produto está na posição A da gôndola é estatisticamente igual a R$140.

Testes unilaterais

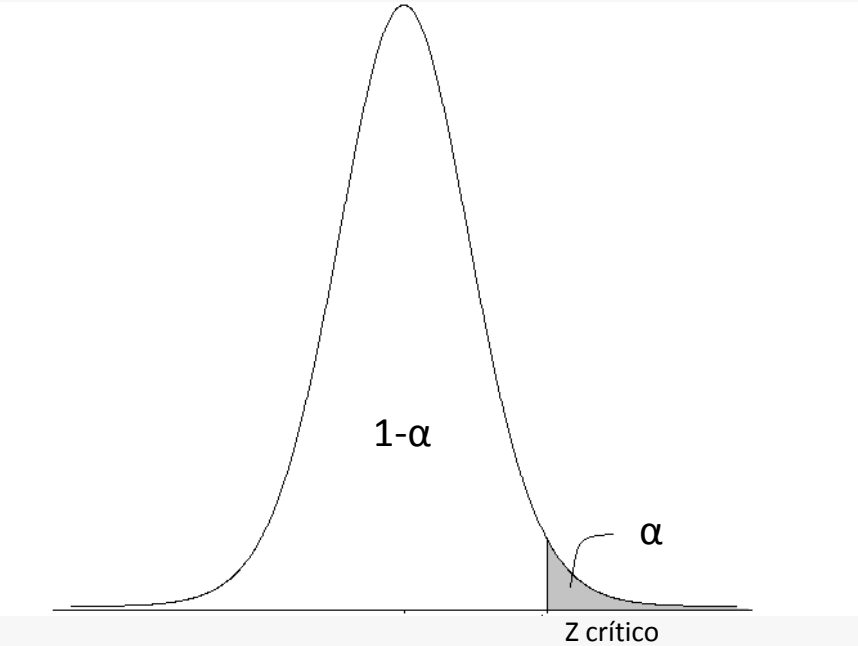
No exemplo anterior, utilizamos um teste de hipótese bilateral. Temos também a opção de utilizar teste de hipótese unilateral. Por exemplo, ao invés de testar se a média das vendas do produto na posição A da gôndola é igual ou diferente de R$140, podemos testar se a média de vendas é igual ou maior a R$140, ou ainda, se a média de vendas é igual ou menor a R$140. Vamos visualizar na curva normal como fica a região crítica para os testes unilaterais.

Um teste unilateral a direita fica:

H_0: $\mu = 140$

H_1: $\mu > 140$

Figura 39 – Região crítica para um teste unilateral a direita.



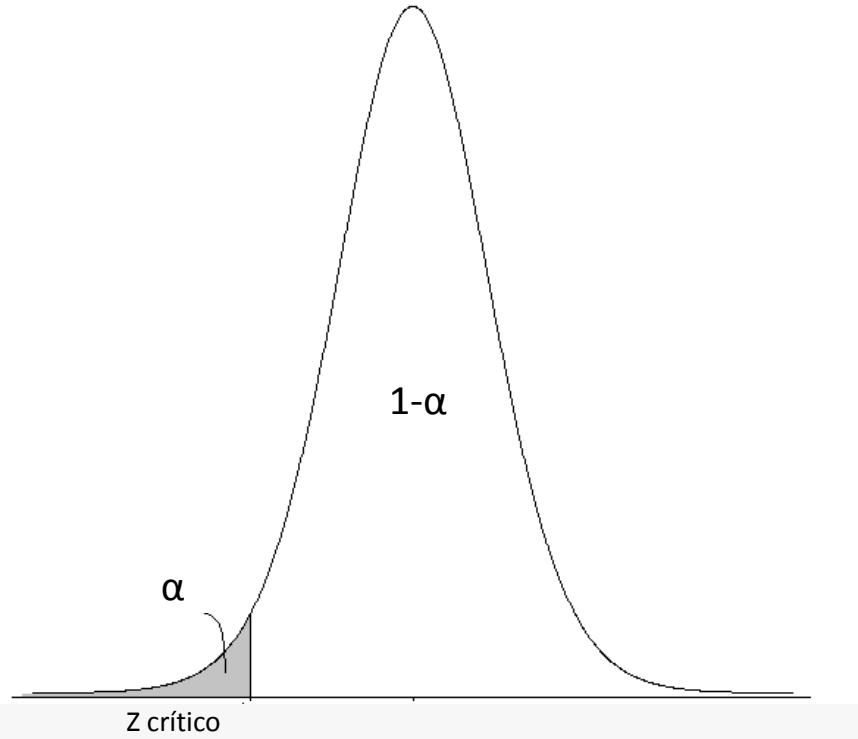
Dessa forma, rejeitaremos H_0 se a estatística de teste for maior que o Zcrítico.

Um teste unilateral a esquerda fica:

H_0: $\mu = 140$

H_1: $\mu < 140$

Figura 40 – Região crítica para um teste unilateral a esquerda.



Dessa forma, rejeitaremos H_0 se a estatística de teste for menor que o Zcrítico.

Avaliando a normalidade de uma variável aleatória

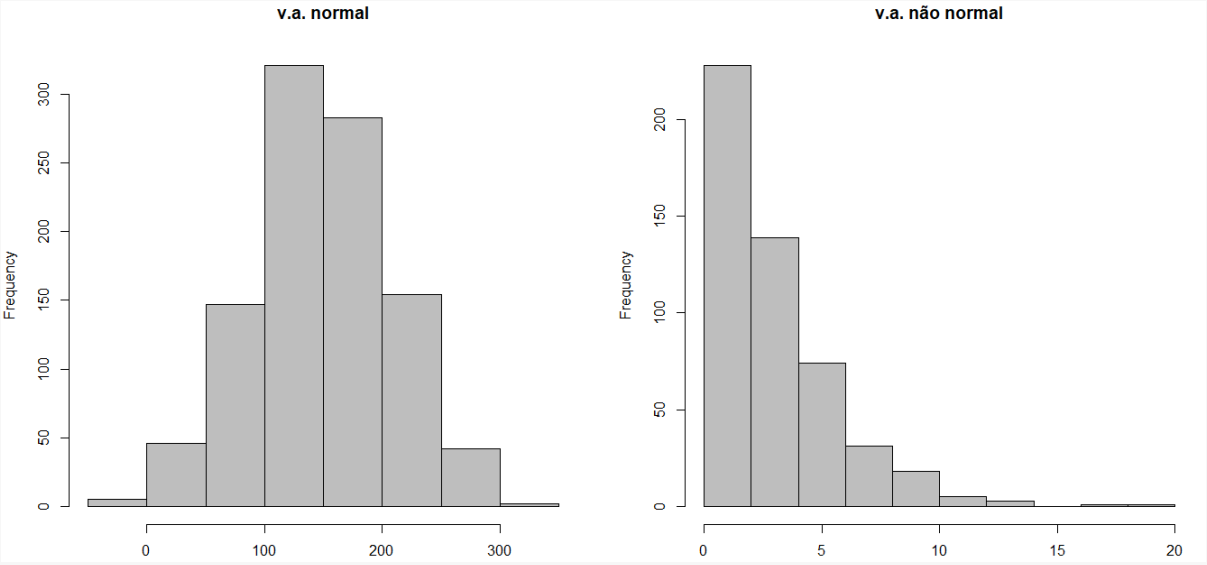
Vimos que muitas análises estatísticas necessitam que os dados sigam uma distribuição normal. Iremos aprender três ferramentas para diagnosticar se uma variável aleatória segue uma distribuição normal: o **Histograma**, o **QQ-Plot** e o **Teste Shapiro Wilk**.

- **Histograma para avaliar a normalidade de uma v.a.**

Observe os dois histogramas na figura 41. O primeiro segue uma distribuição normal com média 150. Note que os dados se distribuem simetricamente em torno do valor médio, esse padrão simétrico sugere

normalidade. Já o segundo histograma apresenta um padrão assimétrico, esse padrão sugere ausência de normalidade.

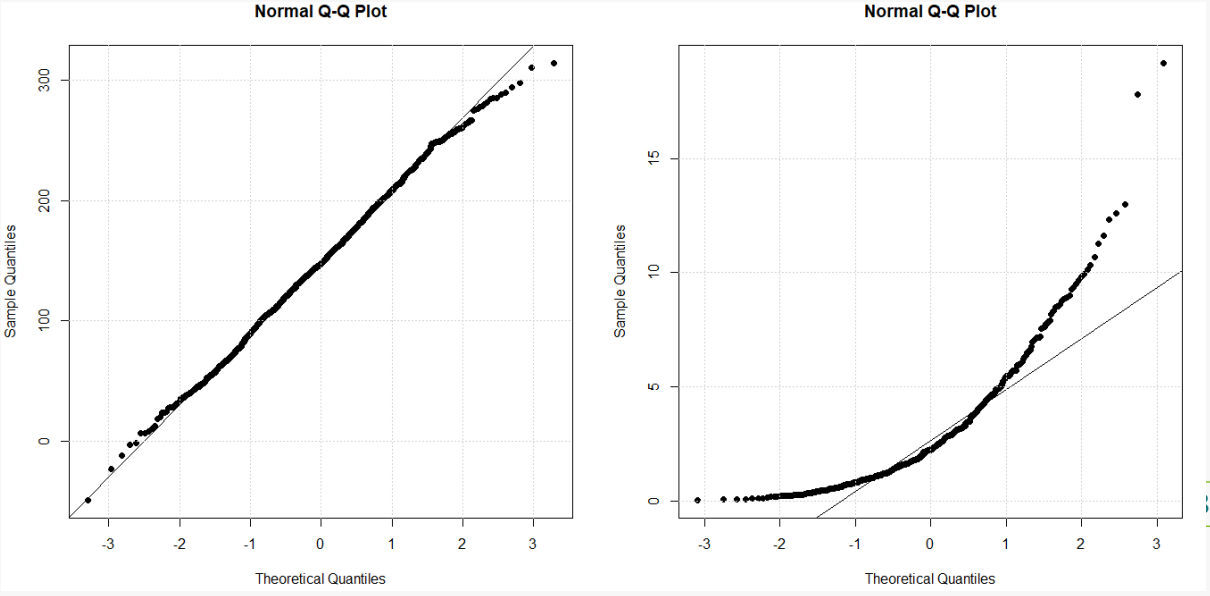
Figura 41 – Histograma de uma variável normal e de uma não normal.



- **QQ-Plot**

O Quantile-Quantile Plot (QQ-Plot) ordena os valores da v.a. do menor para o maior, plota os valores originais no eixo y, e plota os valores padronizados no eixo x. Se a v.a. for normal, ele vai apresentar um padrão linear. Veja na figura 42 um QQ-Plot para cada uma das v.a. do exemplo anterior na figura 41. Observe que a variável normal possui seus pontos seguindo a reta, já a não normal não segue a reta.

Figura 42 – QQ-Plot de uma v.a. normal e de uma não normal.



- **Teste Shapiro-WiIk**

É um teste de hipótese que testa:

H_0: A v.a. segue uma distribuição normal

H_1: A v.a. não segue uma distribuição normal

O procedimento é o mesmo que já conhecemos, uma estatística de teste será calculada e posteriormente o valor p. Entretanto, o cálculo para a estatística de teste do Shapiro-Wilk não é relevante para nosso estudo.

Vamos rodar o teste e interpretar o valor p. Vamos fixar α=5%: se o valor p for abaixo de 5% devemos rejeitar H_0, ou seja, rejeitar a hipótese de que a variável segue uma distribuição normal.

Figura 43 – Teste Shapiro-Wilk para v.a. normal e para a v.a. não normal.

```
        Shapiro-Wilk normality test

data:  va1
W = 0.9985, p-value = 0.5558
```

```
        Shapiro-Wilk normality test

data:  va2
W = 0.8414, p-value <0.0000000000000002
```

Na primeira v.a. (que já sabemos que segue uma normal) o valor p foi de 55,5%, que é acima do nosso nível de significância (5%), portanto com 95% de confiança podemos dizer que essa variável segue uma distribuição normal.

Já na segunda v.a. (que já sabemos que não segue uma normal) o valor p foi praticamente zero, abaixo do nosso nível de significância, portanto devemos rejeitar H_0. Podemos dizer com 95% de confiança que essa variável não segue uma distribuição normal. Observe que o valor p foi tão baixo que mesmo se nosso nível de confiança fosse 99% (consequentemente α=1%) ainda teríamos evidências seguras para rejeitar H_0.

Teste t para diferença de médias (duas amostras independentes)

Podemos utilizar a distribuição t para verificar se duas médias são estatisticamente diferentes ou se a diferença entre elas é devido ao acaso. Vamos propor um contexto.

Continuando nossos estudos da melhor posição da gôndola para colocar um produto, observamos n_1=25 dias de vendas do produto enquanto colocado na posição A. A média de vendas foi $\bar{x}_1$=R$150,1 e o desvio padrão foi s_1= 17. Também observamos n_2=30 dias de vendas do produto enquanto colocado na posição B: a média de vendas foi $\bar{x}_2$=R$182,1 e o desvio padrão foi s_2= 19,2.

Vamos assumir que as vendas sigam uma distribuição normal e que os desvios padrões populacionais são desconhecidos e diferentes. Desejamos saber, a partir das amostras coletadas, se a média das vendas do produto quando colocado na posição A é estatisticamente diferente da média de vendas do produto quando colocado na posição B. Ou seja, é um teste bilateral. Definindo H_0 e H_1, fica:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Ou traduzindo para o nosso contexto, fica:

$$H_0: \mu_{\text{Posição A}} = \mu_{\text{Posição B}}$$

$$H_1: \mu_{\text{Posição A}} \neq \mu_{\text{Posição B}}$$

Iremos assumir 95% para nosso teste. Consequentemente, nosso nível de significância será α =5%.

Iremos calcular a estatística de teste (t calculado) e os graus de liberdade. Para isso utilizaremos algumas fórmulas propostas por Bussab e Morettin (1987).

$$t_{calculado} = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$Graus\ de\ liberdade = \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

Todos os valores para os cálculos foram fornecidos no enunciado, então substituindo fica:

$$t_{calculado} = \frac{(150{,}1 - 182{,}1)}{\sqrt{\frac{17^2}{25} + \frac{19{,}1^2}{30}}}$$

$$t_{calculado} = -6{,}5527$$

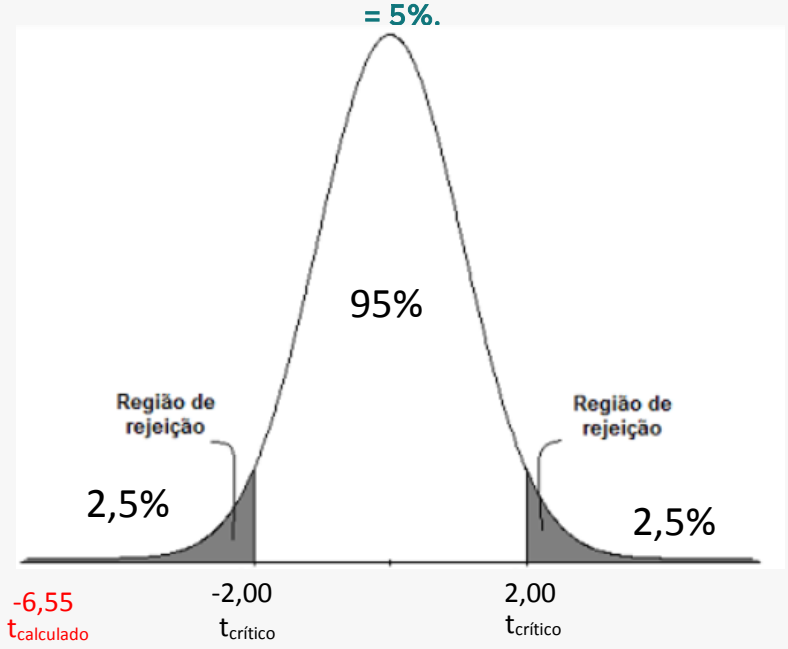
$$Graus\ de\ liberdade = \frac{(\frac{17^2}{25} + \frac{19{,}2^2}{30})^2}{\frac{(\frac{17^2}{25})^2}{25 - 1} + \frac{(\frac{19{,}2^2}{30})^2}{30 - 1}}$$

$$Graus\ de\ liberdade = 52{,}7831$$

Nossa estatística de teste $t_{calculado}$= -6,5527 segue uma distribuição t de Student com 52,7831 graus de liberdade.

Iremos, agora, construir nossa região crítica. Para isso, precisamos saber os valores do $t_{crítico.}$ Podemos utilizar o comando do R *qt(p=0.975, df=52.7831)*. O quantil obtido pelo R é de 2,0059 e como a distribuição t de Student é simétrica, o quantil inferior é negativo -2,0059. Vamos visualizar no gráfico.

Figura 44 – Região crítica para uma t de Student com 52,78 graus de liberdade. e α = 5%.



Podemos ver que o $t_{calculado}$ está contido na região crítica, então temos evidências para rejeitar H_0.

Podemos também obter o valor p para nossa estatística de teste com o comando 2*pt(q = -6.5527, df = 52.7831): obteremos que o valor p é aproximadamente zero. Ou seja, o valor p é menor que o nível de significância α = 5%, portanto H_0 deve ser rejeitada.

A resposta formal fica: Com 95% de confiança, há evidência para rejeitar a hipótese nula, ou seja, as vendas do produto na posição A são estatisticamente diferentes das vendas do produto na posição B.

No exemplo que acabamos de resolver, utilizamos um teste bilateral, ou seja, testamos a diferença. No entanto, podemos tranquilamente utilizar um teste unilateral para testar se a média de vendas do produto A é maior (unilateral a direita) ou menor (unilateral a esquerda) do que a média de vendas do produto na posição B.

Teste t para diferença de médias (duas amostras dependentes)

Podemos estar interessados em mensurar o efeito de um tratamento aplicado a uma amostra. Por exemplo, ao selecionar um grupo de vendedores e submetê-los a um treinamento de vendas (tratamento), podemos ter interesse em mensurar estatisticamente a eficácia do treinamento comparando a média de vendas antes com a média de vendas após o treinamento. Outro exemplo: um grupo de vinte pessoas obesas são candidatas de um programa de dieta (tratamento). Podemos ter interesse em mensurar a eficácia da dieta comparando estatisticamente a média de peso dos candidatos antes de serem submetidos a dieta com a média de peso após a dieta.

Vamos supor que esse programa de dieta é um produto que nossa empresa vem desenvolvendo, e desejamos avaliar sua efetividade antes de colocar o produto no mercado. Então, selecionamos vinte voluntários e observamos seus respectivos pesos. Aplicamos a dieta e observamos novamente seus pesos. Antes da dieta, os pesos possuíam média $\bar{x}_{antes}=123$ e desvio padrão $s_{antes}=18$. Após a dieta, os pesos apresentaram média $\bar{x}_{depois}=110$ e desvio padrão $s_{antes}=28$. O n=20 antes e depois, pois estamos avaliando a diferença nos mesmos indivíduos. No teste t pareado, o tamanho de n deve ser igual para antes e depois do tratamento.

Queremos saber se o peso médio dos indivíduos após a aplicação da dieta é estatisticamente menor do que o peso médio antes da dieta. Portanto, estamos trabalhando com um teste de hipótese unilateral a esquerda.

$$H_0: \mu_2 = \mu_1$$

$$H_1: \mu_2 < \mu_1$$

Ou traduzindo para o nosso contexto, fica:

$$H_0: \mu_{\text{Após a Dieta}} = \mu_{\text{Antes a Dieta}}$$

H_1: $\mu_{\text{Após a Dieta}} < \mu_{\text{Antes da Dieta}}$

Iremos adotar o nível de confiança de 90%. Consequentemente, nosso nível de significância será α=10%.

Importante: Para calcular a estatística de teste, precisamos calcular a diferença entre o peso de antes e de depois para cada um dos vinte indivíduos e, posteriormente, calcular a média e o desvio padrão das diferenças. Ou seja:

$$Média\ da\ Diferença = \bar{x}_{Diferença} = \sum_{i=1}^{n} \frac{(X_{i\ \ Antes} - X_{i\ \ Depois})}{n}$$

$$Desvio\ Padrão\ da\ Diferença = s_{Diferença} = \sqrt{\frac{\sum_{i=1}^{n}(Diferença_i - Média_{Diferença})^2}{n-1}}$$

Para simplificar, calcule uma nova coluna com as diferenças dos pesos e calcule a média e o desvio padrão para essa coluna. *No entanto, não se preocupe, veremos na aula com o R que é mais simples do que parece*. Vamos assumir que a diferença média ficou $\bar{x}_{\text{Diferença}} = -2{,}1728$ e o desvio padrão das diferenças foi $s_{\text{Diferença}} = 10{,}1803$. O n já sabemos que é vinte.

Detalhe: Precisamos que as diferenças sigam uma distribuição normal para aplicação do teste t pareado.

Nossa estatística de teste seguirá uma distribuição t de Student com n -1 graus de liberdade e pode ser obtida da seguinte forma:

$$t_{calculado} = \frac{\bar{x}_{diferença}}{\frac{s_{difença}}{\sqrt{n}}}$$

Substituindo fica:

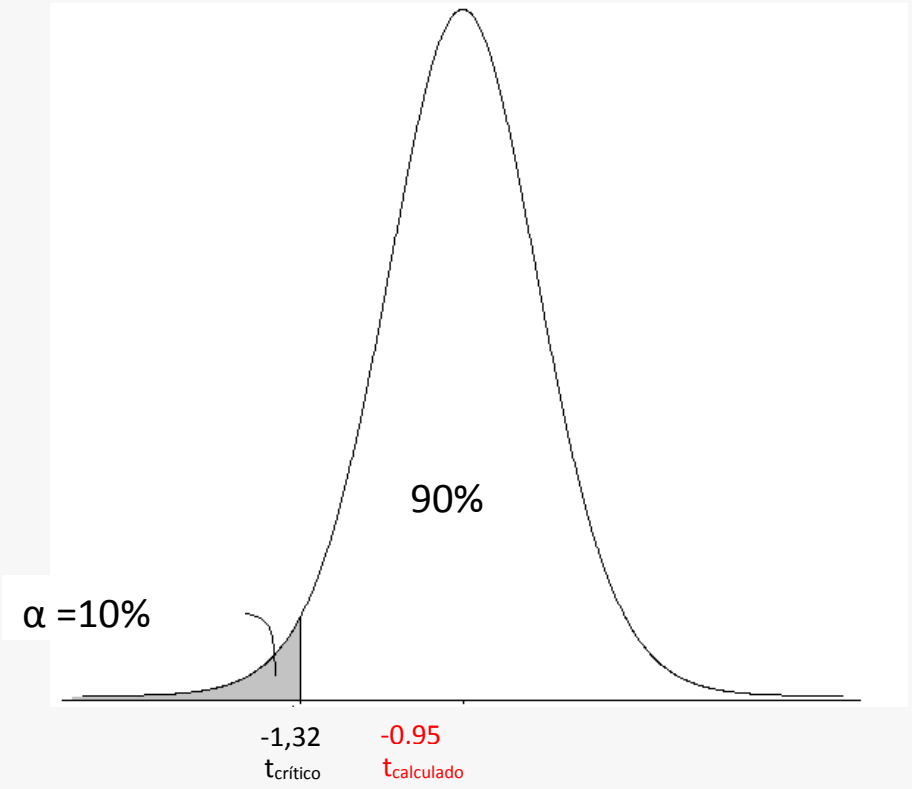
$$t_{calculado} = \frac{-2{,}1728}{\frac{10{,}1803}{\sqrt{20}}}$$

$$t_{calculado} = -0{,}9544$$

Nossa estatística de teste $t_{calculado}$= -0,9544 segue uma distribuição t de Student com 19 graus de liberdade.

Iremos agora construir nossa região crítica. Para isso, precisamos saber os valores do $t_{crítico.}$ Podemos utilizar o comando do R *qt(p=0.90, df=19)*. A probabilidade utilizada é 0,90, pois é o nível de confiança adotado. O quantil obtido pelo R é de 1,3277, entretanto, como nosso teste é unilateral a esquerda e a distribuição t e é simétrica, nosso quantil será negativo (-1,3277). Vamos visualizar no gráfico.

Figura 45 – Região crítica unilateral a esquerda para uma distribuição t de Student com 19 graus de liberdade ao nível de confiança de 90%.



Podemos ver que o tcalculado não está contido na região crítica, então não temos evidências para rejeitar H0. Podemos também obter o valor p para nossa estatística de teste com o comando pt(q = -0,9544, df = 19): obteremos que o valor p é de 0,1759 (17,59%). Ou seja, o valor p é maior que o nível de significância α = 10%, portanto H0 não deve ser rejeitada.

A resposta formal fica: Com 90% de confiança, não há evidências para rejeitar a hipótese nula, ou seja, o peso médio dos indivíduos após o tratamento com a dieta não é estatisticamente menor que a média do peso antes da dieta.

Alguns comentários: Nós vemos que há alguma diferença no peso médio após o tratamento, o peso médio sofreu uma ligeira redução. Porém, além desta ser 'pequena', o desvio padrão aumentou. Resumindo, mediante o valor p, temos evidências de que essas variações no peso foram devido ao acaso e não são significativas. A título de negócio, a dieta ainda não está pronta para ir ao mercado e deve ser revisada.

Teste Qui-Quadrado para independência entre variáveis categóricas

Em situações científicas, sejam elas em um laboratório ou em um departamento de marketing e vendas de uma empresa, é comum encontrarmos situações em que precisamos identificar se duas variáveis qualitativas são associadas ou independentes. Para essas situações, podemos utilizar o teste Qui-Quadrado. Vamos propor um exemplo.

A fim de conhecer melhor o comportamento dos nossos clientes, desejamos investigar se um produto vende mais quando o cliente adulto está acompanhado de uma criança. Para isso, observamos 50 clientes, com criança e sem criança, que compraram e não compraram o produto. Achamos que o cliente compra independentemente de estar ou não com criança. Os dados coletados estão dispostos em 50 linhas e 2 colunas, cada linha é um cliente observado e em cada coluna as características do cliente. Veja uma parte dos dados na figura 46.

Figura 46 – Características observadas dos clientes.

	Cliente	Comprou
1	Adulto_com_Crianca	Não_Comprou
2	Adulto_com_Crianca	Não_Comprou
3	Adulto_com_Crianca	Não_Comprou
4	Adulto	Não_Comprou
5	Adulto	Não_Comprou
6	Adulto	Não_Comprou
7	Adulto_com_Crianca	Comprou
8	Adulto_com_Crianca	Comprou
9	Adulto_com_Crianca	Comprou
10	Adulto_com_Crianca	Comprou
11	Adulto_com_Crianca	Comprou
12	Adulto_com_Crianca	Comprou
13	Adulto_com_Crianca	Comprou
14	Adulto_com_Crianca	Comprou
15	Adulto_com_Crianca	Comprou
16	Adulto_com_Crianca	Comprou
17	Adulto_com_Crianca	Comprou
18	Adulto_com_Crianca	Comprou
	...	...
49	Adulto	Comprou
50	Adulto	Comprou

A hipótese que o Teste qui-quadrado para independência de variáveis categóricas avalia é:

H_0: Não existe associação significativa entre as variáveis
H_1: Existe associação significativa entre as variáveis

Traduzindo pro contexto proposto fica:

H_0: O fato de o cliente estar ou não com criança não tem relação com o fato de comprar ou não comprar

H_1: O fato de o cliente estar ou não com criança tem relação com fato de comprar ou não comprar

Para compreendermos os cálculos para obter a estatística de teste, precisaremos dispor os dados em uma tabela 2(linhas)x2(colunas), também conhecida como tabela de contingência.

Figura 47– Tabela de contingência 2x2.

	Comprou	Não_Comprou	Total (Colunas)
Adulto	6	14	**20**
Adulto_com_Crianca	23	7	**30**
Total (linhas)	**29**	**21**	**50**

O próximo passo é calcular a frequência esperada para cada casela, pois se a distância entre a frequência observada e a esperada for "grande" o suficiente, teremos evidências para rejeitar H_0. Para calcular os valores esperados para cada casela podemos utilizar a fórmula:

$$E_{ij} = \frac{n_i * n_j}{n}$$

Onde i representa as linhas e j representa as colunas e n a quantidade de observações. Calcularemos juntos a frequência esperada para cada casela:

E_{11} = (29*20)/50 = 11,6

E_{12} = (21*20)/50 = 8,4

E_{21} = (29*30)/50 = 17,4

E_{22} = (21*30)/50 = 12,6

A tabela de valores esperados fica:

Figura 48 – Valores esperados para cada linha e coluna.

	Comprou	Não_Comprou	Total (Colunas)
Adulto	11,6	8,4	**20**
Adulto_com_Crianca	17,4	12,6	**30**
Total (linhas)	**29**	**21**	**50**

Interpretando os valores esperados:

O valor esperado para um adulto que compra (casela i=1,j=1), baseado no total de pessoas que compraram e no total de adultos, dado o total de observações na base de dados, é 11,6 pessoas.

O valor esperado para um adulto que não compra (casela i=1,j=2), baseado no total de pessoas que não compraram e no total de adultos, dado o total de observações disponível na base de dados, é 8,4 pessoas.

O valor esperado para um adulto com criança que compra (casela i=2,j=1), baseado no total de pessoas que compraram e no total de adultos com criança, dado o total de observações disponível na base de dados, é 17,4 pessoas.

O valor esperado para um adulto com criança que não compra (casela i=2,j=2), baseado no total de pessoas que não compraram e no total de adultos com criança, dado o total de observações disponível na base de dados, é 12,6 pessoas.

A estatística de teste qui-quadrado calculado fica:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Onde r e c são (respectivamente) as linhas e colunas da tabela de contingência.

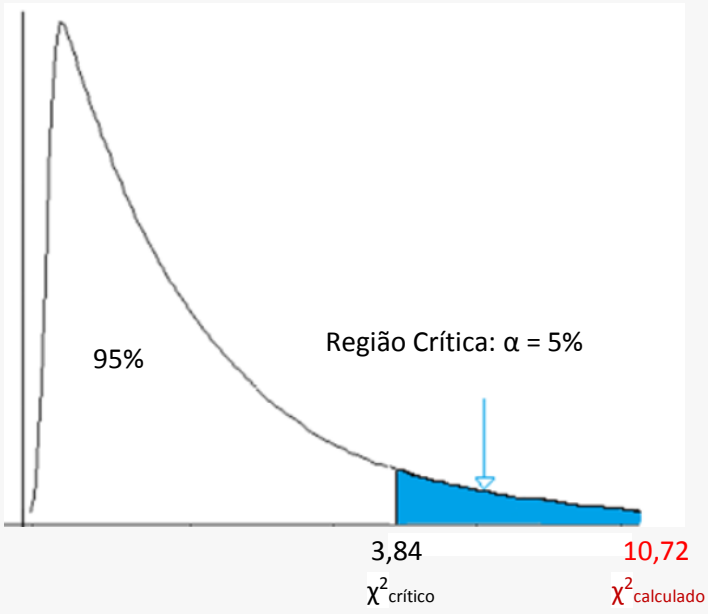
Ou seja, devemos subtrair o valor **observado** da linha=i e coluna=j pelo valor **esperado** da mesma casela de linha=i e coluna=j e, posteriormente, dividir pelo valor esperado dessa mesma casela de linha=i e coluna=j.

Nossa estatística de teste qui-quadrado com (linhas-1)*(colunas-1) graus de liberdade fica:

$$\chi^2 = 10{,}7279 \text{ com um 1 grau de liberdade}$$

A distribuição qui-quadrado conforme apresentado no capítulo 2, não é simétrica, então a região crítica será unilateral. Vamos construir a região crítica para 95% de confiança. Podemos obter o quantil referente ao valor crítico no R com o comando *qchisq(p=0.95,df = 1)*.

Figura 49 – Região crítica de 95% de confiança para uma distribuição qui-quadrado com 1 grau de liberdade.



Como nossa estatística de teste calculada está contida na região crítica, temos evidências para rejeitar a hipótese nula. Podemos também obter o valor p através do comando do R *1-pchisq(q=10,72,df = 1)*. O comando pchisq() nos dá a área sobre a curva qui-quadrado até o 10,72. Entretanto, queremos o complementar, por isso utilizamos 1-pchisq(). O valor p obtido é de 0,0010, que é menor do que nosso nível de significância.

A resposta formal fica: Com 95% de confiança, temos evidências para rejeitar a hipótese nula. Ou seja, o fato de o cliente estar ou não com criança tem relação com o fato de o cliente comprar ou não comprar.

Carvajal et al (2009) comenta que um dos pressupostos para o teste qui-quadrado para testar independência entre duas variáveis qualitativas é que a frequência esperada para cada casela seja ≥ 5, e recomenda que o número mínimo de dados observados seja n≥30.

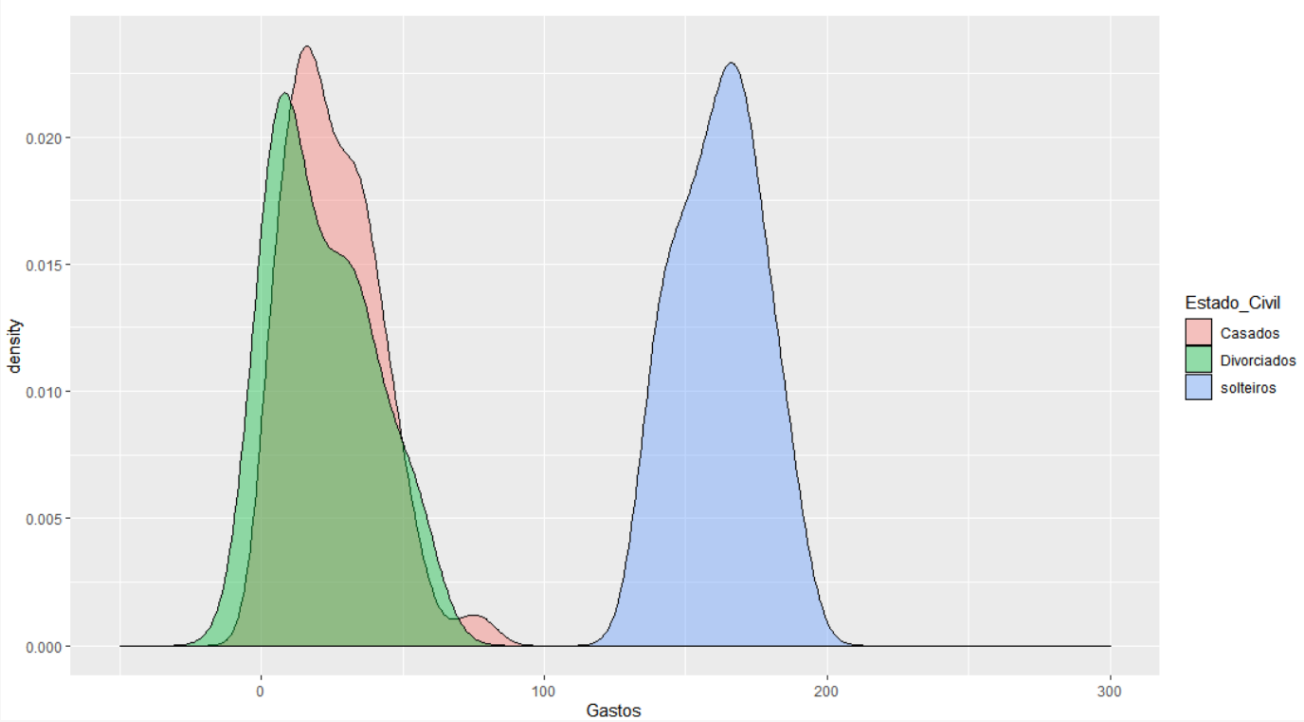
O teste qui-quadrado não se limita a tabelas de contingência 2x2, também pode ser aplicado a tabelas de ordem mais alta.

Teste F para análise de variância (ANOVA)

Há situações em que podemos desejar testar a diferença entre três ou mais médias e já vimos anteriormente que o teste t nos permite comparar apenas pares de médias. Podemos, então, utilizar uma análise de variância, também conhecida como ANOVA (Analysis of Variance), que utiliza um teste F para identificar se há variabilidade significativa ao realizar as comparações das médias das n populações. Vamos propor um exemplo para facilitar a compreensão.

Vamos supor que estamos pesquisando o gasto com uma determinada bebida para três populações (públicos) em um restaurante. Observamos os gastos com a bebida oriundos de n_1=17 solteiros, n_2=98 casados e n_3=15 divorciados. Vejamos em uma curva de densidade a distribuição dos gastos em cada uma das três populações.

Figura 50 – Distribuição com consumo para cada uma das três populações na amostra coletada.



Pelo gráfico da figura 50, vemos que a distribuição dos gastos entre as populações Casados e Divorciados possuem bastante intersecção e praticamente a mesma média de gastos (eixo x). Isso sugere que não possuem diferença se comparadas entre si. No entanto, se observarmos a distribuição dos gastos da população Solteira, vemos que tem muito pouca ou nenhuma intersecção com as demais populações (e possui maior média). Isso sugere que o consumo dos Solteiros difere do consumo dos Divorciados e Casados. Vamos formalizar nossas hipóteses e prosseguir com o teste fazendo uso dos passos que já aprendemos.

As hipóteses formais do teste F são:

H_0: As médias são iguais $(\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_n)$

H_1: Pelo menos uma das médias é diferente ($\mu_i \neq \mu_j$ para pelo menos um par de médias (i,j))

Vamos adotar 95% de confiança e α=5% para prosseguir com a execução do teste.

Para obter o F calculado, precisamos de alguns cálculos, pois ele é uma razão entre dois valores.

$$F_{calculado} = \frac{\frac{SS_{entre}}{m-1}}{\frac{SS_{dentro}}{n-m}}$$

Onde m é a quantidade de populações que estão sendo testadas, n é a quantidade de observações disponíveis, SS_{entre} é a soma dos quadrados das diferenças entre as médias de cada população em relação média global, SS_{dentro} é a soma dos quadrados das diferenças das observações dentro das populações em relação a média daquela população.

Para obter a $SS_{entre,}$ podemos utilizar a seguinte fórmula:

$$SS_{entre} = \sum_{i=1}^{m} n_i(\bar{Y}_i - \bar{Y})^2$$

Onde m é a quantidade de populações, n_i é a quantidade de observações da i-ésima população, $\bar{Y}_i$ é a média da i-ésima população e $\bar{Y}$ é a média global da variável estudada.

Para obter a $SS_{dentro,}$ podemos utilizar a seguinte fórmula:

$$SS_{dentro} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$$

Onde i é a i-ésima população, j é a j-ésima observação da i-ésima população, Y_{ij} é o valor de cada observação j da população i e $\bar{Y}_i$ é a média da i-ésima população.

O cálculo para a estatística de teste nesse exemplo fica:

$$F_{calculado} = \frac{\frac{SS_{entre}}{m-1}}{\frac{SS_{dentro}}{n-m}}$$

$$F_{calculado} = \frac{\frac{276.693}{3-1}}{\frac{32.665}{130-3}}$$

$$F_{calculado} = \frac{138.319,50}{257,2}$$

$$F_{calculado} = 538$$

Ou seja, nossa estatística F é de 537,7798 com 2 graus de liberdade no numerador e 127 graus de liberdade no denominador.

Usualmente os softwares estatísticos nos dão uma tabela com o resumo da ANOVA.

Figura 51 – Tabela ANOVA.

Fonte de Variação	Soma dos Quadrados (SS - Sum of Squares)	Graus de Liberdade	Quadrados médios (MS)	F
Entre populações	SS_{entre}	m-1	$MS_{entre} = \frac{SS_{entre}}{m-1}$	$\frac{MS_{entre}}{MS_{dentro}}$
Dentro das populações (erro)	SS_{dentro}	n-m	$MS_{dentro} = \frac{SS_{dentro}}{n-m}$	

Trazendo para nosso exemplo, a tabela da ANOVA fica:

Figura 52 – Tabela ANOVA para os gastos com bebidas nas populações avaliadas.

Fonte de Variação	Soma dos Quadrados (SS)	Graus de Liberdade	Quadrados médios (MS)	F
Entre populações	276.639	2	138.319	538
Dentro das populações (erro)	32.665	127	257	

O Fcrítico para uma distribuição F com α=5% com 2 graus de liberdade no numerador e 127 no denominador é 3,07. Podemos achar o F crítico utilizando o R *qf(p = 0.95,df1=2,df2=127).*

Podemos também obter valor p no R *1-pf(q=538,df1=2,df2=127)*. Temos que adicionar o 1- pois queremos a extremidade direita da curva F. O valor obtido é aproximadamente zero.

Para concluir o teste, tanto pelo $F_{calculado}$ quanto pelo valor P, temos evidências para rejeitar a hipótese nula.

A resposta formal fica: Com 95% de confiança, há evidências para rejeitar a hipótese nula. Ou seja, pelo menos uma das médias é estatisticamente diferente.

Estatística Computacional – Teste de Hipótese com o R

```
#############################################################
##
#######      Teste de Hipótese              #######
##      AED - Capitulo 04  - Prof. Máiron Chaves     ####
#############################################################
##
#Copie este código, cole no seu R e execute para ver os resultados
rm(list = ls()) #Limpa memória do R
########## Avaliando a normalidade de uma variável aleatória
##############################
set.seed(10)
#Gera v.a. que segue distribuição normal com n = 70, média = 40 e desvio padrão = 8
va_normal <- rnorm(n = 70, mean = 25, sd = 8)
#Gera v.a. que segue uma distribuição F (não normal) com n = 15, 2 graus de liberdade no
numerados e 10 graus de liberdade no denominador
va_nao_normal <- rf(n =15, df1 =2, df2 = 10)
#Visualize o histograma das variáveis geradas
#Observe como os dados se distribuem em torno do valor médio na va normal
hist(va_normal)
#Observe como os dados não se distribuem em torno de um valor médio exibindo padrão
assimétrico
hist(va_nao_normal)
```

```
# Visualize o QQ-Plot
# Observe como os pontos de dados seguem a linha reta qq norm da va normal
qqnorm(va_normal)
qqline(va_normal) #Este comando é para adicionar a linha
# Observe como os pontos de dados não seguem a linha reta na va não normal
qqnorm(va_nao_normal)
qqline(va_nao_normal) #Este comando é para adicionar a linha
# Vamos aplicar o teste de hipóteses Shapiro Wilk. O teste funciona sob as hipóteses
# H0: A variável segue uma distribuição normal
# H1: A variável não segue uma distribuição normal

# Fixe um nível de significância alfa e analise o p valor (p-value) do Shapiro Wilk
#Se o p-value for menor que alfa a hipótese nula deve ser rejeitada
shapiro.test(va_normal)



shapiro.test(va_nao_normal)
################ Teste t para diferença de médias (duas amostras independentes) ################
#Iremos simular o exemplo da apostila
#Iremos testar se:
# H0: As vendas na posição A são iguais as Vendas na Posição B
# H1: As vendas na posição A são diferentes das vendas na posição B
rm(list = ls()) #Limpa objetos da memória do R
mu1 <- 150.1 #Armazena as média de vendas na posição A
mu2 <- 182.1 #Armazena as média de vendas na posição B

s1 <- 17 #Armazena o desvio padrão das vendas na posição A
s2 <- 19.2 #Armazena o desvio padrão das vendas na posição B

n1 <- 25 #Armazena a quantidade observações registradas para de vendas na posição A
n2 <- 30  #Armazena a quantidade observações registradas para de vendas na posição B
```

```
#Calcula nossa estatística de teste. Que é o t calculado

t <- (mu1 - mu2)  / sqrt( s1^2/n1 + s2^2/n2)

t #Visualize o valor de t calculado

#Calcula os graus de liberdade da estatística de teste

gl <-  (s1^2/n1 + s2^2/n2)^2  /( (s1^2/n1)^2 / (n1-1) + (s2^2/n2)^2 / (n2-1) )

gl #Visualize a quantidade de graus de liberdade

#Obtem o quantil (t crítico) para uma distribuição t com gl graus de liberdade. A um alfa de 5%

quantil <- qt(0.975,df = gl)

quantil #Visualize o t crítico

#Esse é o aspecto de uma distribuição t com n=53 observações e com n - 1 graus de liberdade

plot(density(rt(n = 53,df = gl)),xlim = c(-7,7))

#Observe onde estão os valores críticos que acabamos de encontrar

abline(v = quantil,col = 'blue',lwd = 2)

abline(v = -quantil,col = 'blue',lwd = 2)

abline(v = t, col = 'red')# Observe como o tcalculado é muito menor que o tcrítico. Está na região de rejeição

#Obtendo o valor p

#P(Tcalculado > Tcritico)

2*pt(q = t, df = gl)

#Agora vamos realizar o mesmo teste de hipótese  utilizando a função nativa do R t.test()


vendas_A <- rnorm(n= 25, mean = 150.1, sd = 17)

vendas_B <- rnorm(n = 30, mean = 182.1, sd = 19.2)

#Observe no output desta função, que ela já nos da tudo pronto, t calculado e valor p

t.test(vendas_A,vendas_B, alternative = 'two.sided')

#Esse é o aspecto de uma distribuição t com n observações e com n - 1 graus de liberdade

n <- 5

plot(density(rt(n = n,df = n-1)))

#Altere o valor de n de 5 em 5 observe que a medida que os graus de liberdade aumenta a distribuição se aproxima da normal. Como os valores são gerados aleatoriamente
```

poderemos ter curvas diferentes para um mesmo valor de n, mas a medida que n cresce o comportamento simétrico tende a estabilizar.

```
################ Teste t para diferença de médias (duas amostras dependentes) ################

#Iremos simular o exemplo da apostila

# H0: O peso médio após a dieta é igual ao peso médio antes da dieta

# H1: O peso médio após a dieta é menor do que o peso médio antes da dieta

rm(list = ls()) #Limpa memória do R

#Iremos utilizar uma biblioteca adicional para gerar valores aleatórios que sigam uma distribuição normal entre um intervalo de valor para simular os pesos

#A biblioteca chama 'truncnorm'. Basta instalar com o comando abaixo install.packages().

#Uma vez instalada não há mais necessidade de instalar novamente. Basta carregar com o comando library()

install.packages('truncnorm')

library(truncnorm)

set.seed(100)

#Gera uma amostra aleatória, seguindo uma distribuição normal cujo valor mínimo é 100 e o valor máximo é 140.

#O valor de n=20, média = 123 e desvio padrão 18

#Com essa v.a. iremos simular os pesos dos indivíduos antes da dieta

antes_da_dieta <- rtruncnorm(n=20, a=100, b=140, mean=123, sd=18)

#Gera uma amostra aleatória, seguindo uma distribuição normal cujo valor mínimo é 110 e o valor máximo é 130.

#O valor de n=20, média = 110 e desvio padrão 28

#Com essa v.a. iremos simular os pesos dos indivíduos após a dieta

depois_da_dieta <- rtruncnorm(n=20, a=110, b=130, mean=110, sd=28)

#Calcula a diferença depois da dieta e antes da dieta, para cada indivíduo

diferenca <- depois_da_dieta-antes_da_dieta

#Visualiza a distribuicao da diferença de pesos

hist(diferenca)

shapiro.test(diferenca) #Avalie a normalidade da distribuição da diferença

#Aplica test t com os seguintes argumentos

t.test(depois_da_dieta,antes_da_dieta,

    paired = TRUE, #Pareado
```

```
    alternative = "less", #Unilateral a esquerda
    conf.level = 0.9 #90 porcento de confiança
   )
#O comando t.test() acima nos da tudo que precisamos para executar e concluir o teste. Mas a título de conhecimento, podemos realizar o teste passo a passo
#Calcula a média das diferenças
media <- mean(diferenca)
#Desvio padrão das diferenças
desvio_padrao <- sd(diferenca)
#Quantidade de indivíduos
n <- 20
#Obtem o t calculado
t_calculado <- media / (desvio_padrao/sqrt(n))
#Obtem o valor p para o t calculado com n - 1 graus de liberdade.
pt(q = t_calculado, df = n-1)
#Podemos também obter o t crítico para uma distribuição t com 19 (n-1=20-1) graus de liberdade ao nível de confiança de 90%

tcrítico_teste_t_pareado <- -qt(p = 0.9, df = 19) #Devido ao teste ser unilateral a esquerda a distribuição t ser simétrica, nossa estatística de teste será negativa

#Observe que o t calculado é maior que o t critico. Como estamos em um teste unilateral a esquerda o t calculado
#estará fora da região de rejeição caso seja maior que o t crítico
t_calculado < tcrítico_teste_t_pareado
?t.test #Maiores informações sobre o comando t.test()
################## Teste Qui-Quadrado para associação entre variáveis categóricas ##################
#Iremos simular o exemplo da apostila
# H0: O fato do cliente estar ou não com criança não tem relação com o fato de comprar ou não comprar
# H1: O fato do cliente estar ou não com criança tem relação com fato de comprar ou não comprar
rm(list = ls())
#Vamos gerar um data frame contendo os dados da pesquisa
```

```
dados <- data.frame(

 Cliente = c("Adulto_com_Crianca", "Adulto_com_Crianca", "Adulto_com_Crianca",

             "Adulto", "Adulto", "Adulto", "Adulto_com_Crianca", "Adulto_com_Crianca",

             "Adulto_com_Crianca", "Adulto_com_Crianca", "Adulto_com_Crianca",

             "Adulto_com_Crianca", "Adulto_com_Crianca", "Adulto_com_Crianca",

             "Adulto_com_Crianca", "Adulto_com_Crianca", "Adulto_com_Crianca",

             "Adulto_com_Crianca", "Adulto_com_Crianca", "Adulto_com_Crianca",

             "Adulto_com_Crianca", "Adulto", "Adulto", "Adulto", "Adulto",

             "Adulto_com_Crianca", "Adulto_com_Crianca", "Adulto_com_Crianca",

             "Adulto_com_Crianca", "Adulto", "Adulto_com_Crianca", "Adulto",

             "Adulto", "Adulto_com_Crianca", "Adulto_com_Crianca", "Adulto_com_Crianca",

             "Adulto", "Adulto_com_Crianca", "Adulto", "Adulto", "Adulto",

             "Adulto","Adulto","Adulto","Adulto","Adulto","Adulto"),

 Comprou = c("Não_Comprou", "Não_Comprou", "Não_Comprou", "Não_Comprou",

             "Não_Comprou", "Não_Comprou", "Comprou", "Comprou", "Comprou",
```

```
        "Comprou", "Comprou", "Comprou", "Comprou", "Comprou", "Comprou",

        "Comprou", "Comprou", "Comprou", "Comprou", "Comprou", "Comprou",

        "Comprou", "Comprou", "Comprou", "Não_Comprou", "Não_Comprou",

        "Não_Comprou", "Não_Comprou", "Comprou", "Não_Comprou", "Comprou",

        "Comprou", "Não_Comprou", "Não_Comprou", "Não_Comprou", "Não_Comprou",

        "Não_Comprou", "Comprou", "Comprou", "Não_Comprou", "Não_Comprou",

        "Não_Comprou", "Não_Comprou", "Não_Comprou",
"Comprou","Comprou","Comprou","Comprou","Comprou","Comprou")


)


#Visualiza o conjunto de dados

View(dados)

#Gera tabela de contigência 2x2

tabela <- table(dados$Cliente,dados$Comprou)

tabela

barplot(tabela)

#O valor críticico para uma distribuição qui-quadrado com (linhas-1)*(colunas-1)=1 grau de liberdade ao nível de confiança de 95%

qchisq(p=0.95,df = 1)

#O valor p unilateral fica

1-pchisq(q=10.728,df=1) #Mesmo que o nível de confiança fosse 99%, ainda teríamos evidências para rejeitar H0

#Assim como fizemos no test t, podemos usar um comando direto no R para realizar o teste qui-quadrado chisq.test()

teste<-chisq.test(tabela,correct = F)

teste

#Visualiza valores observados. Que nada mais é do que a tabela original
```

```
teste$observed
#Visualiza valores esperados
teste$expected

?chisq.test #Maiores informações sobre o comando chisq.test()

################ ANOVA ############################
# Vamos utilizar o exemplo da apostila
#H0: Não há diferença no valor médio gasto com bebidas em nenhuma das populações
#H1: Há diferença no valor médio gasto com bebidas em pelo menos uma das populações
rm(list = ls())
#Gera um data frame contendo os dados da pesquisa
dados_anova <- data.frame(Gastos = c(174.770021661909, 161.329206619394,
                                     153.679900850863, 163.790338797433, 141.363480335882,
175.351592994046,
                                     185.793398289321, 184.720273514352, 163.400459287948,
170.202462740626,
                                     150.8549565713, 167.583106239899, 140.190492201897,
157.440088617225,
                                     171.596654773339, 138.885665257324, 147.942698809323,
9.87474262516482,
                                     50.5645554670016, 14.2586307887884, 8.5061846804934,
25.0875496696788,
                                     17.0661987504312, 41.3867417301938, 20.8113941426179,
60.1224674502026,
                                     35.5154028285664, 23.7622285692359, 34.6086119259266,
30.4321086925016,
                                     27.8188980544904, 37.4729772794009, 30.7229538650678,
48.0452539322412,
                                     78.9197865324734, 42.4926762466659, 8.81227865272712,
39.5751781629677,
                                     37.1329656327517, 15.8016718071775, 5.74735216885902,
38.684069121093,
                                     30.9398891106907, 34.7370783113952, 13.2630510987537,
19.6212096123791,
                                     16.716945267481, 24.4037922212213, 4.63398786180773,
32.9436217626275,
```

21.511905851158, 31.4997283634204, 26.6610570873775, 34.6304034101472,

16.2704826042681, 11.2323425300881, 18.023244405391, 15.4790632095655,

8.25633422881043, 27.9053307974433, 72.3298402892867, 4.7263338963663,

14.4153129255327, 41.2234268777169, 50.5684226296565, 19.8344282661234,

8.81306901471397, 19.5112436004646, 55.6251926080436, 16.7592556127806,

20.3176176298076, 31.2073058210955, 17.0613250010048, 47.8590627884627,

2.59778754862417, 35.9470130480825, 2.39404093355522, 9.38425601777391,

25.2455048267186, 16.1960287769175, 43.530118783298, 32.7250288712979,

5.43268078364765, 44.5365791890593, 32.9831443965413, 28.2104605365607,

3.18609515001209, 14.3698142789208, 39.9617218607622, 50.564581262513,

10.4634451365926, 36.4842442182048, 13.1330189654278, 8.93702642184252,

12.1501174131844, 22.2552757873296, 15.1407470062459, 11.7525513477354,

16.2990775324815, 24.4627568806115, 2.87916580644454, 44.5453919973285,

38.0393535792355, 32.1985589022666, 0.357075783631849, 22.0703974352325,

50.7486034030794, 18.604230207709, 5.83122133978906, 19.9252025339318,

6.8366108202567, 27.5834177510951, 41.9303025963975, 3.077799353254,

28.0507001837521, 33.0042729903, 50.7366690908169, 30.1697285113061,

6.53184416916073, 7.53469171526227, 5.49225229796712, 9.53198727121377,

6.59266645551752, 19.8423174628847, 0.781567028951091, 22.1605754480815,

5.90830712162365, 54.3457453874529, 33.3341495203441, 37.2034845899045

```
), Estado_Civil = c("solteiros", "solteiros", "solteiros", "solteiros",
            "solteiros", "solteiros", "solteiros", "solteiros", "solteiros",
            "solteiros", "solteiros", "solteiros", "solteiros", "solteiros",
            "solteiros", "solteiros", "solteiros", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Casados", "Casados", "Casados", "Casados", "Casados", "Casados",
            "Divorciados", "Divorciados", "Divorciados", "Divorciados", "Divorciados",
            "Divorciados", "Divorciados", "Divorciados", "Divorciados", "Divorciados",
            "Divorciados", "Divorciados", "Divorciados", "Divorciados", "Divorciados"))
#Visualiza o conjunto de dados
View(dados_anova)
#Podemos utilizar os recursos de visualização da biblioteca ggplot2 para visualizar a
distribuição dos gastos nas populações
install.packages("ggplot2")
library(ggplot2)

ggplot(data = dados_anova, aes(x = Gastos, fill = Estado_Civil)) +
  geom_density(alpha=0.4)+
```

```
 xlim(-50,300)
#É bastante comum também analisarmos a variabilidade nas distintas populações com uso de boxplot
boxplot(dados_anova$Gastos ~ dados_anova$Estado_Civil)



#Com o comando aov(), o R gera a tabela da ANOVA completa
anova <- aov(Gastos~ #Variável resposta
        Estado_Civil, #Fator que queremos testar se exerce influencia na variável resposta
        data = dados_anova)
#Visualize a tabela da ANOVA. Observe o F calculado e o valor p ( Pr > F)
summary(anova)
# O valor p é praticamente zero. Mesmo que nosso nível de confiança fosse 99,9% ainda teríamos evidências para rejeitar H0
```

> Capítulo 5

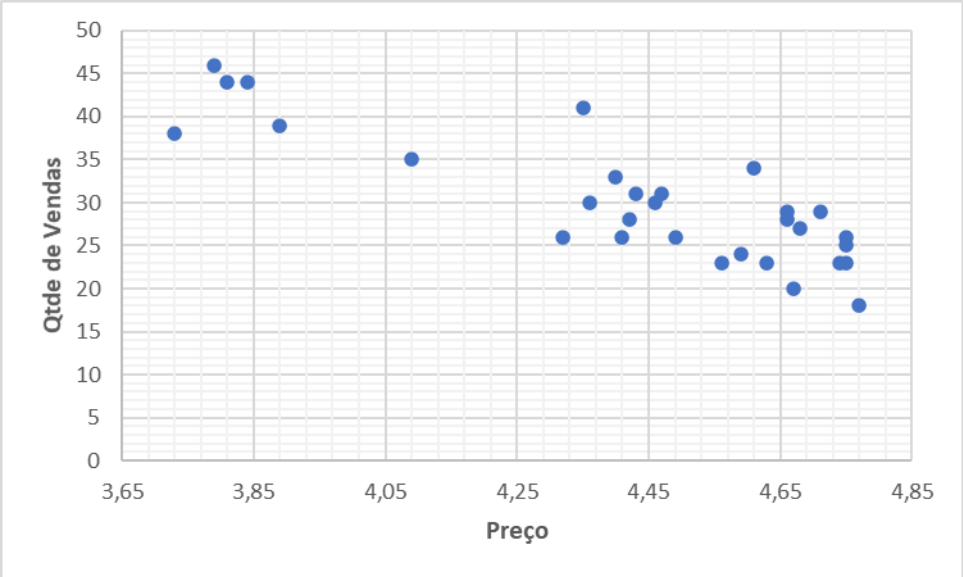
Capítulo 5. Regressão Linear

A Regressão Linear permite gerar um modelo matemático através de uma reta que explique a relação linear entre variáveis. No caso mais simples, teremos a relação entre uma variável explicativa X e uma variável resposta Y. O modelo estatístico de Regressão Linear com duas variáveis pode ser representado pela seguinte equação (MONTGOMERY; RUNGER. 2003):

$$\hat{Y} = \beta_0 + (\beta_1 * X_1) + \varepsilon$$

Onde β_0 é o termo de intercepto (em outras palavras, é o valor de Y quando X = 0). β_1 é a inclinação da reta, representa a mudança média prevista em y resultante do aumento de uma unidade em X. ε é um termo erro aleatório com média μ zero e variância σ^2 constante.

Retomando o gráfico da figura 7, onde visualizamos a relação entre o preço do café e suas vendas.



Existe relação entre o preço do café e a quantidade vendida? Como o preço explica as vendas do café? Posso utilizar o preço para prever as vendas?

Ao longo deste capítulo iremos aprender as respostas para essas questões.

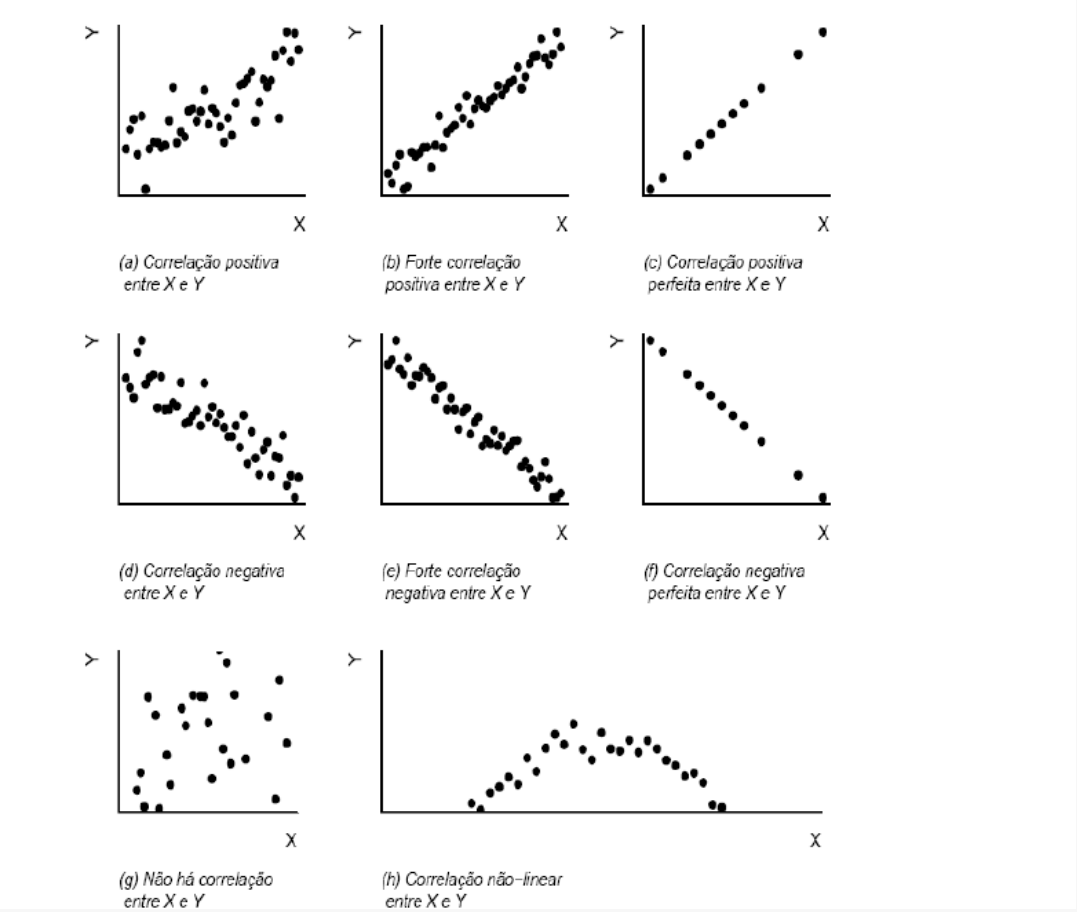
Correlação Linear

Para medir a força da correlação entre duas variáveis quantitativas, pode ser utilizado o coeficiente de correlação de Pearson.

O coeficiente de correlação de Pearson é um valor numérico que vai de -1 a 1. Quanto mais próximo de -1 for a correlação, mais forte será a correlação de forma negativa (quando uma variável aumenta a outra diminui). Quanto mais próxima de 1, mais forte será a correlação entre as duas variáveis de forma positiva (quando uma variável aumenta a outra também aumenta).

Antes de calcular a correlação entre o preço do café e suas vendas, vamos visualizar alguns exemplos hipotéticos de gráficos de dispersão para termos uma melhor noção do diagnóstico da correlação entre um par de variáveis através da análise gráfica.

Figura 53 – Analisando a correlação pelo gráfico de dispersão.



Vemos que a relação entre o preço do café e suas vendas têm um padrão parecido com o gráfico (d) da figura 53. Portanto, é uma correlação negativa. Esse é o comportamento esperado, já que quando os preços aumentam, as vendas tendem a diminuir.

Para mensurar a força da correlação entre um par de variáveis, podemos calcular o coeficiente de correlação linear de Pearson com a seguinte equação:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}}$$

Substituindo, fica:

$$\rho = \frac{-58{,}01}{\sqrt{[3{,}0078]\,[1550]}}$$

$$\rho = \frac{-58{,}01}{68{,}2806}$$

$$\rho = -0{,}84$$

O coeficiente de correlação (denotado pela letra grega rho) entre o preço do café e suas vendas é ρ =-0,84. Vamos utilizar a tabela abaixo para nos orientar na interpretação desse valor.

Figura 54 – Interpretando valores do coeficiente de correlação.

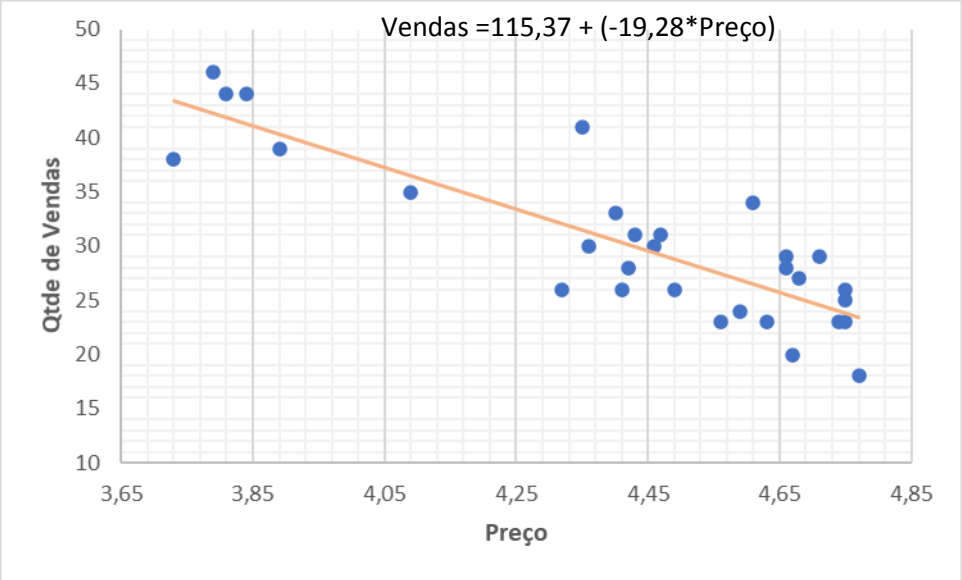
Valor do Coeficiente	Interpretação
Entre 0,10 e 0,29	Correlação positiva fraca
Entre 0,30 a 49	Correlação positiva moderada
Entre 0,5 e 1	Correlação positiva alta
Entre -0,10 e -0,29	Correlação negativa fraca
Entre -0,30 a -49	Correlação negativa moderada
Entre -0,5 e -1	Correlação negativa alta

Fonte: (COHEN, Multiple Commitment in the workplace: an integrative approach, 2003).

Regressão Linear Simples e Regressão Linear Múltipla

A regressão linear simples é quando temos apenas uma variável preditora e uma variável resposta. É o exemplo do preço do café e suas vendas. Nosso modelo considera apenas o preço para explicar e prever a variação nas vendas. Vamos ajustar uma reta de regressão na figura 55 para modelarmos estatisticamente a relação entre o preço do café e a quantidade de vendas.

Figura 55 – Ajustando uma regressão linear entre o preço do café e suas vendas.



Conforme já analisamos anteriormente, a correlação entre as duas variáveis é negativa, ou seja, quando uma aumenta (preço) a outra diminui (vendas). Então, nossa reta é decrescente. Temos também a equação de primeiro grau que descreve a reta, ou seja, descreve a relação entre o preço e as vendas.

$$Vendas\ do\ Café = 115{,}37 + (-19{,}28 * Preço\ do\ Café)$$

Onde 115,37 é o intercepto B0 e -19,28 é o coeficiente angular B1. Ou seja, a cada real que aumenta no preço, as vendas caem, em média, em 19,28 unidades.

Devemos sempre estar atentos à unidade de medida das variáveis, pois a interpretação do coeficiente angular é, sempre, dado um aumento unitário na unidade de medida da variável preditora, a variável resposta cresce (ou decresce se o coeficiente angular for negativo) tantas unidades de medida.

Se quisermos prever o volume de uma árvore (medido em pés cúbicos) baseado na sua circunferência (medida em polegadas) e ajustássemos uma reta de regressão para isso, supondo que a equação obtida fosse:

$$Volume = -36{,}94 + (5{,}06 * Circunferência)$$

O coeficiente angular 5,06 é positivo, portanto, a interpretação fica: à cada polegada aumentada na circunferência, o volume da árvore aumenta, em média, 5,06 pés cúbicos.

Voltando ao nosso exemplo principal, que são as vendas do café em função de seu próprio preço. Podemos querer incluir mais uma variável em nosso estudo. Por exemplo, suspeita-se que na nossa empresa, o leite seja um produto complementar ao café, ou seja, o cliente que compra café tende a comprar leite. Se essa hipótese for verdadeira, ao realizar mudanças no preço do leite as vendas do café também poderão ser impactadas. Vamos incluir mais uma variável no nosso modelo, que será o preço do leite.

Repare que dessa forma, já estamos trabalhando com três variáveis, pois temos duas variáveis preditoras (preço do café, preço do leite) e uma variável resposta (quantidade vendida do café). Quando temos duas ou mais variáveis preditoras, estamos trabalhando com um modelo de regressão linear múltipla.

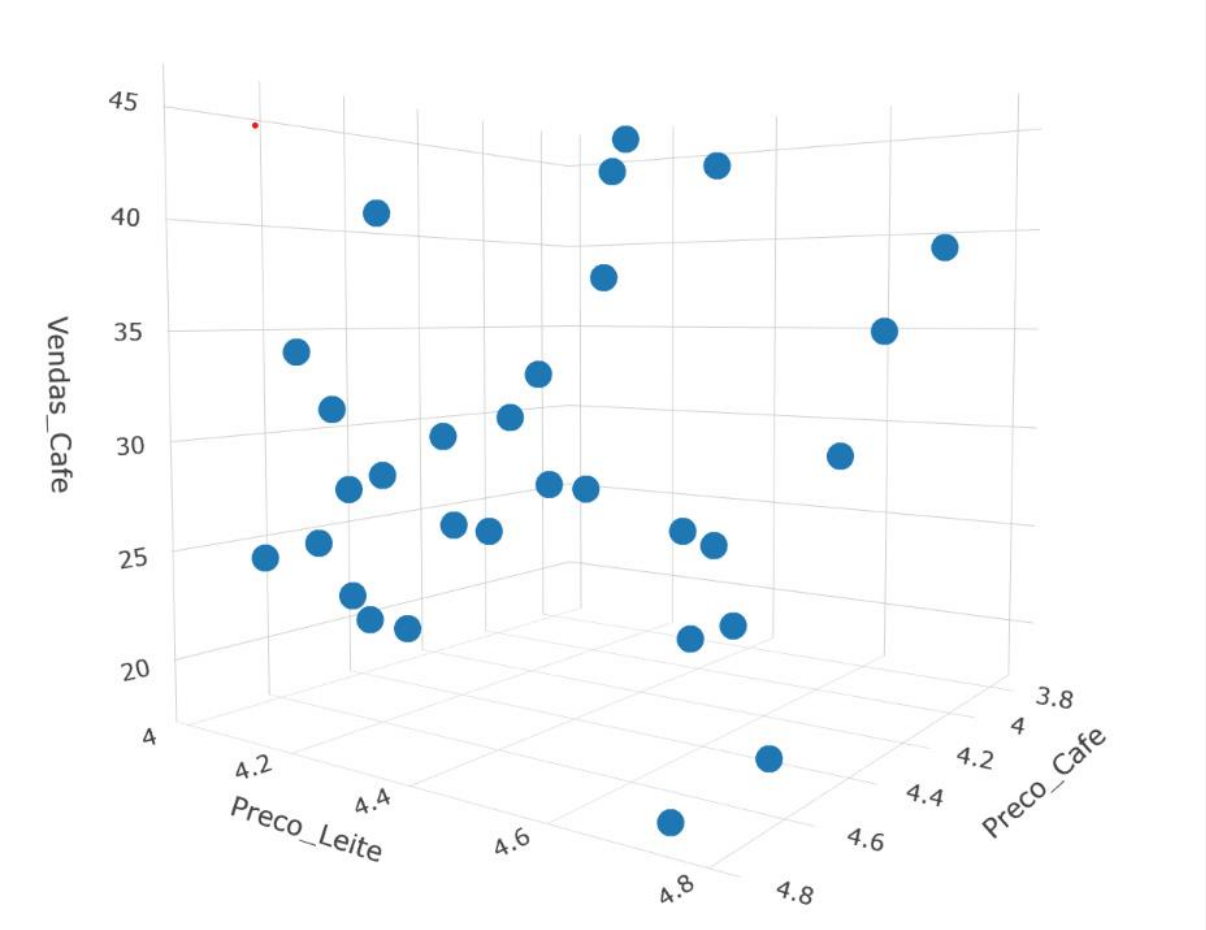
A regressão linear múltipla pode comportar p variáveis preditoras ao invés de somente uma, como na regressão linear simples.

$$\hat{Y} = \beta_0 + (\beta_1 * X_1) + (\beta_2 * X_2) + \cdots + (\beta_p * X_p) + \varepsilon$$

Algumas literaturas sugerem ter pelo menos dez observações para cada preditor adicionado.

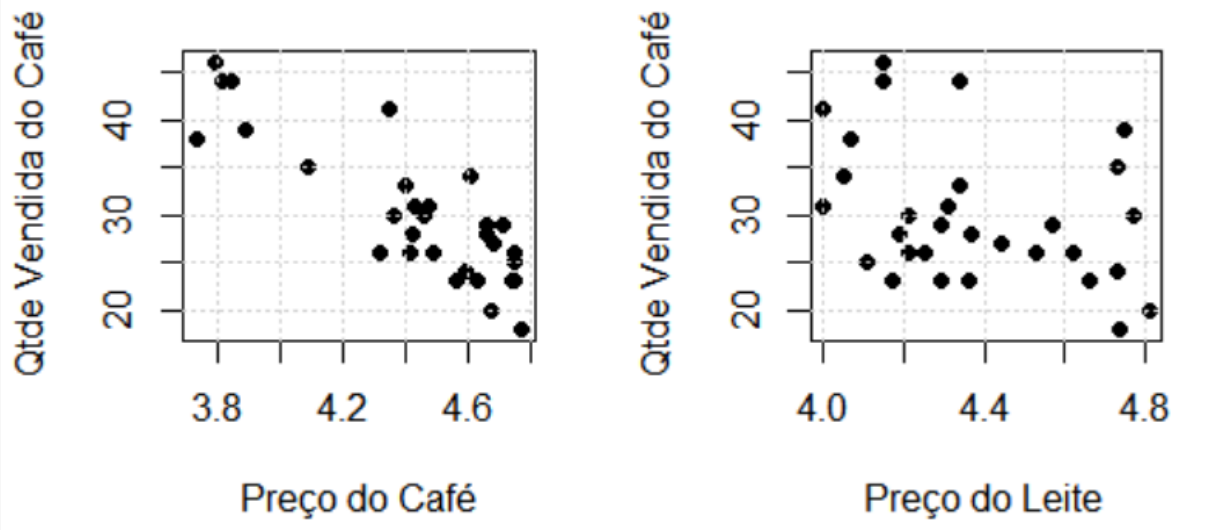
Quando trabalhamos com três dimensões, ainda é possível visualizar graficamente. Acima disso não é mais possível, pois nós seres humanos não enxergamos mais de três dimensões.

Figura 56 – Relação entre Vendas do Café, Preço do Café e Preço do Leite, em um gráfico tridimensional.



É possível ver que quando o preço do leite aumenta as vendas do café caem, assim como quando o preço do café aumenta as vendas do café também caem. No entanto, muitas vezes o gráfico em 3D pode não ser muito intuitivo de interpretar. Uma forma simples é apresentar os gráficos aos pares de variáveis. Por exemplo:

Figura 57 – Relação entre Vendas do Café, Preço do Café e Preço do Leite, em dois gráficos bidimensionais.



A equação de regressão múltipla ficou:

$$Vendas\ do\ Café = 151{,}31 + (-18{,}72 * Preço\ do\ Café) + (-8{,}78 * Preço\ do\ Leite)$$

Observe que o coeficiente do Preço do Café no modelo múltiplo ficou ligeiramente diferente do coeficiente no modelo simples. Isso é comum pois adicionamos mais uma variável, e os coeficientes são estimados em conjunto. O método utilizado para estimação dos parâmetros (coeficientes) é o método dos Mínimos Quadrados Ordinários. Ele é um método de otimização numérica que pode ser resolvido por algoritmos numéricos como o método do gradiente, e também possui solução analítica, ou seja, é possível estimar os coeficientes utilizando fórmula fechada. Em notação matricial, o algoritmo de mínimos quadrados fica:

$$\beta = (X^T X)^{-1}(X^T Y)$$

Onde X é a matriz $n_x p$ (n é a quantidade de observações da amostra e p a quantidade de preditores) e Y é uma matriz coluna $n_x 1$ contendo os valores da variável resposta na amostra.

Aos alunos interessados na álgebra linear por trás do algoritmo, podem aprofundar seus conhecimentos neste link.

Utilizando Variável Categórica em um Modelo de Regressão Linear

É bastante comum o uso de variáveis categóricas como preditoras em um modelo de regressão linear. Ainda no nosso contexto das vendas do café, vamos supor que desejamos adicionar uma nova variável indicando em quais dias o café estava em promoção. Veja na figura a seguir:

Figura 58 – Adicionando a variável categórica Promoção.

Promocao	Preco_Cafe	Vendas_Cafe
Nao	4,77	18
Nao	4,67	20
Nao	4,75	23
Nao	4,74	23
Nao	4,63	23
Nao	4,56	23
Nao	4,59	24
Nao	4,75	25
Sim	4,75	26
Nao	4,49	26
Sim	4,41	26
Nao	4,32	26
Nao	4,68	27
Sim	4,66	28
Sim	4,42	28
Nao	4,71	29
Sim	4,66	29
Sim	4,46	30
Sim	4,36	30
Nao	4,47	31
Nao	4,43	31
Sim	4,4	33
Sim	4,61	34
Sim	4,09	35
Nao	3,73	38
Sim	3,89	39
Sim	4,35	41
Sim	3,84	44
Sim	3,81	44
Sim	3,79	46

Como vimos anteriormente, para estimar os coeficientes betas da equação de regressão, são necessários cálculos entre matrizes, e não é possível realizar cálculos sobre a palavra ‘Sim’ e ‘Não’. Portanto, para uma

variável categórica que pode assumir dois níveis, o mais comum é transformar em binário. A variável assume o valor 1 quando há promoção e 0 quando não há promoção.

Figura 59 – Convertendo a variável promoção para binário.

Promoção_Binária	Promocao	Preco_Cafe	Vendas_Cafe
0	Nao	4,77	18
0	Nao	4,67	20
0	Nao	4,75	23
0	Nao	4,74	23
0	Nao	4,63	23
0	Nao	4,56	23
0	Nao	4,59	24
0	Nao	4,75	25
1	Sim	4,75	26
0	Nao	4,49	26
1	Sim	4,41	26
0	Nao	4,32	26
0	Nao	4,68	27
1	Sim	4,66	28
1	Sim	4,42	28
0	Nao	4,71	29
1	Sim	4,66	29
1	Sim	4,46	30
1	Sim	4,36	30
0	Nao	4,47	31
0	Nao	4,43	31
1	Sim	4,4	33
1	Sim	4,61	34
1	Sim	4,09	35
0	Nao	3,73	38
1	Sim	3,89	39
1	Sim	4,35	41
1	Sim	3,84	44
1	Sim	3,81	44
1	Sim	3,79	46

Dessa forma, podemos remover da matriz de dados a Promoção em categorias e manter somente a Promoção ajustada para binário para rodar o algoritmo de regressão linear.

Outra situação comum no uso de variáveis categóricas é quando ela pode assumir mais de dois níveis. Por exemplo, se desejarmos incluir a variável Dia da Semana pra capturar o efeito de cada dia da semana nas vendas:

Figura 60 – Adicionando a variável Dia da Semana.

Promoção_Binária	Dia_da_Semana	Preco_Cafe	Vendas_Cafe
0	Segunda	4,77	18
0	Terca	4,67	20
0	Quarta	4,75	23
0	Quinta	4,74	23
0	Sexta	4,63	23
0	Sabado	4,56	23
0	Domingo	4,59	24
0	Segunda	4,75	25
1	Terca	4,75	26
0	Quarta	4,49	26
1	Quinta	4,41	26
0	Sexta	4,32	26
0	Sabado	4,68	27
1	Domingo	4,66	28
1	Segunda	4,42	28
0	Terca	4,71	29
1	Quarta	4,66	29
1	Quinta	4,46	30
1	Sexta	4,36	30
0	Sabado	4,47	31
0	Domingo	4,43	31
1	Segunda	4,4	33
1	Terca	4,61	34
1	Quarta	4,09	35
0	Quinta	3,73	38
1	Sexta	3,89	39
1	Sabado	4,35	41
1	Domingo	3,84	44
1	Segunda	3,81	44
1	Terca	3,79	46

A variável Dia da Semana é uma variável categórica que pode assumir 7 níveis: Segunda, Terça, Quarta, Quinta, Sexta, Sábado e Domingo. Portanto, precisaremos 6 variáveis binárias para representar o dia da semana. Sempre que uma variável categórica possuir mais de 2 níveis categóricos, precisaremos m-1 colunas binárias, sendo m a quantidade de níveis categóricos que a variável pode assumir.

Figura 61 – Criando m-1 variáveis binárias para o Dia da Semana.

Promoção_Binária	Dia_da_Semana	Segunda	Terca	Quarta	Quinta	Sexta	Sabado	Preco_Cafe	Vendas_Cafe
0	Segunda	1	0	0	0	0	0	4,77	18
0	Terca	0	1	0	0	0	0	4,67	20
0	Quarta	0	0	1	0	0	0	4,75	23
0	Quinta	0	0	0	1	0	0	4,74	23
0	Sexta	0	0	0	0	1	0	4,63	23
0	Sabado	0	0	0	0	0	1	4,56	23
0	Domingo	0	0	0	0	0	0	4,59	24
0	Segunda	1	0	0	0	0	0	4,75	25
1	Terca	0	1	0	0	0	0	4,75	26
0	Quarta	0	0	1	0	0	0	4,49	26
1	Quinta	0	0	0	1	0	0	4,41	26
0	Sexta	0	0	0	0	1	0	4,32	26
0	Sabado	0	0	0	0	0	1	4,68	27
1	Domingo	0	0	0	0	0	0	4,66	28
1	Segunda	1	0	0	0	0	0	4,42	28
0	Terca	0	1	0	0	0	0	4,71	29
1	Quarta	0	0	1	0	0	0	4,66	29
1	Quinta	0	0	0	1	0	0	4,46	30
1	Sexta	0	0	0	0	1	0	4,36	30
0	Sabado	0	0	0	0	0	1	4,47	31
0	Domingo	0	0	0	0	0	0	4,43	31
1	Segunda	1	0	0	0	0	0	4,4	33
1	Terca	0	1	0	0	0	0	4,61	34
1	Quarta	0	0	1	0	0	0	4,09	35
0	Quinta	0	0	0	1	0	0	3,73	38
1	Sexta	0	0	0	0	1	0	3,89	39
1	Sabado	0	0	0	0	0	1	4,35	41
1	Domingo	0	0	0	0	0	0	3,84	44
1	Segunda	1	0	0	0	0	0	3,81	44
1	Terca	0	1	0	0	0	0	3,79	46

Observe que não foi criada uma coluna binária pro indicador quando é o domingo. Não há necessidade, pois a forma de sinalizar o domingo é quando todas as outras binárias do Dia da Semana estiverem zeradas. Nesse caso, o intercepto da equação irá capturar o efeito do Domingo sobre as vendas. Após incluir as m-1 variáveis binárias, a variável original categórica do Dia da Semana pode ser removida da matriz de dados para prosseguir com o ajuste do modelo de regressão.

Explicação vs Predição

Outro fator bastante positivo da regressão linear, e que faz dela um algoritmo muito utilizado tanto na estatística “tradicional” quanto em machine learning, é que o algoritmo nos traz informações inferenciais, ou seja, podemos extrapolar conclusões para a população a partir da amostra e também podemos utilizá-lo para modelagem preditiva. Vai depender do objetivo do pesquisador.

Vamos ajustar um modelo de regressão múltipla no qual a variável resposta é a quantidade vendida do café e nossas variáveis preditoras (também chamadas de variáveis explicativas) são o preço do café, o preço

do leite (produto complementar) e uma variável binária informando se o café estava ou não em promoção no dia. A equação obtida por mínimos quadrados foi:

$$Vendas\ do\ Café = 137{,}3 + (-16{,}11 * Preço\ do\ Café) + (4{,}14 * Promoção) + (-8{,}71 * Preço\ do\ Leite)$$

Podemos interpretar a equação da seguinte forma:

β_0 = Intercepto = É o valor que a variável resposta assume quando os preditores estão zerados.

β_1 = Preço do Café = Mantendo as demais variáveis constantes, para cada aumento unitário nessa variável, ou seja, a cada real aumentado no preço do café, as vendas caem em média 16,11 unidades.

β_2 = Promoção = Mantendo as demais variáveis constantes, quando o Café está em promoção, são vendidas em média 4,14 unidades a mais em relação a quando não está em promoção.

β_3 = Preço do Leite = Mantendo as demais variáveis constantes, para cada aumento unitário nessa variável, ou seja, a cada real aumentado no preço do Leite, as vendas do café caem em média 8,71 unidades.

Uma vez que já sabemos como as variáveis preditoras impactam nas vendas do café, podemos utilizar das previsões para simular cenários.

Qual será a venda estimada caso coloquemos o café em promoção, ao valor de R$2,25, e o leite acima de seu preço médio, custando R$5,00?

Basta substituir na equação: (oriente-se pelas cores)

$$Vendas = 137{,}37 + (-16{,}11 * Preço\ do\ Café) + (4{,}14 * Promoção) + (-8{,}71 * Preço\ do\ Leite)$$

$$Vendas = 137{,}37 + (-16{,}11 * 2{,}25) + (4{,}14 * 1) + (-8{,}71 * 5{,}00)$$

$$Vendas = 62$$

Vamos simular mais um cenário para fixar.

Qual será a venda estimada caso coloquemos o café sem promoção, ao valor de R$4,37, e o leite em seu preço médio R$4,42?

$$Vendas = 137{,}37 + (-16{,}11 * Preço\ do\ Café) + (4{,}14 * Promoção) + (-8{,}71 * Preço\ do\ Leite)$$

$$Vendas = 137{,}37 + (-16{,}11 * 4{,}37) + (4{,}14 * 0) + (-8{,}71 * 4{,}42)$$

$$Vendas = 29$$

Na prática, decisões de contexto e estratégias de negócio devem ser mescladas às decisões matemáticas na tomada de decisão. Estatisticamente falando, vemos que o primeiro cenário nos oferecerá maior quantidade de vendas.

Diagnóstico do Ajuste do Modelo de Regressão Linear

Ao ajustar um modelo de regressão linear sobre um conjunto de dados, alguns pressupostos devem ser atendidos, principalmente quando o pesquisador tem o objetivo inferencial, ou seja, de verificar se o impacto das variáveis preditoras na variável resposta daquela amostra pode ser extrapolado para a população. Por isso, ao rodarmos uma regressão linear utilizando o R, ele nos dá um resumo do modelo, incluindo um teste t de Student para cada coeficiente beta (os betas estão na coluna Estimate).

Figura 62 – Output da regressão linear fornecido pelo R.

```
Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)   137.37      10.83   12.68 0.0000000000012 ***
Preco_Cafe    -16.12       1.65   -9.79 0.0000000003258 ***
PromocaoSim     4.15       1.04    3.99              0.00048 ***
Preco_Leite    -8.71       1.90   -4.58              0.00010 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.61 on 26 degrees of freedom
Multiple R-squared:  0.886,     Adjusted R-squared:  0.872
F-statistic: 67.1 on 3 and 26 DF,  p-value: 0.00000000000225
```

Marcado de azul, temos o coeficiente de cada variável preditora. Observe que são os mesmos da equação que trabalhamos anteriormente para realizar previsões e simular cenários.

Para cada coeficiente beta, ou seja, para cada variável preditora, é feito um teste t de Student para testar:

Figura 63 – Teste t de Student para cada coeficiente.

Test t:
H0: $\beta=0$
H1: $\beta\neq0$

Ou seja

H0: A variável preditora não tem relação significativa com a variável resposta
H1: A variável preditora tem relação significativa com a variável resposta

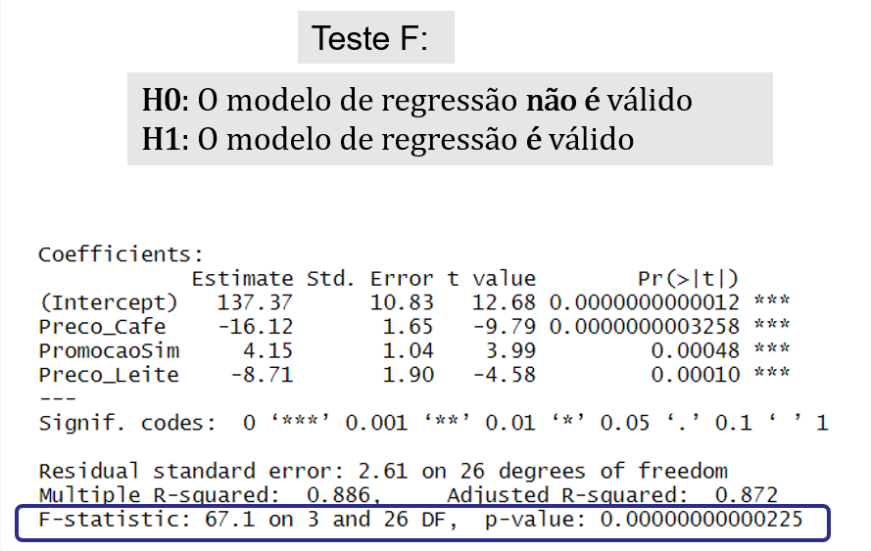
```
Coefficients:
             Estimate Std. Error t value        Pr(>|t|)
(Intercept)   137.37      10.83   12.68 0.0000000000012 ***
Preco_Cafe    -16.12       1.65   -9.79 0.0000000003258 ***
PromocaoSim     4.15       1.04    3.99         0.00048 ***
Preco_Leite    -8.71       1.90   -4.58         0.00010 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.61 on 26 degrees of freedom
Multiple R-squared:  0.886,     Adjusted R-squared:  0.872
F-statistic: 67.1 on 3 and 26 DF,  p-value: 0.00000000000225
```

Na coluna t value, temos o t calculado (Estimate dividido pelo Std. Error) e, em seguida, temos o valor p (Pr(>|t|). Se fixamos um nível de significância, por exemplo α = 5%, rejeitaremos H_0 se o valor p for abaixo de 5% (0,05). Se rejeitarmos, significa que o coeficiente daquela variável não é estatisticamente diferente de zero. Ou seja, aquela variável preditora não tem impacto significativo na variável resposta. Geralmente, nesses casos a preditora é removida do modelo pelo pesquisador e um novo modelo sem ela é ajustado.

Temos também um teste F, idêntico a uma ANOVA.

Figura 64 – Teste F para ajuste geral do modelo de regressão linear.



Veja que temos um F calculado de 67,1 com 3 e 26 graus de liberdade, o que nos retorna um valor p de aproximadamente zero. Ao rejeitarmos H_0 após fixar um valor de alfa, temos evidências de que o modelo de regressão linear ajustado é válido. Caso contrário, precisaríamos rever as variáveis, coletar novas, buscar mais amostra etc.

Temos também duas métricas importantes, que são o R^2 (R quadrado) e o R^2 ajustado. O R^2 é uma medida percentual, ou seja, varia de zero a um, e nos diz o quanto da variação da variável resposta o modelo ajustado explica. O R^2 também é chamado de coeficiente de determinação.

Figura 65 – R^2 e R^2 ajustado.

```
Coefficients:
             Estimate Std. Error t value        Pr(>|t|)
(Intercept)   137.37      10.83   12.68 0.0000000000012 ***
Preco_Cafe    -16.12       1.65   -9.79 0.0000000003258 ***
PromocaoSim     4.15       1.04    3.99          0.00048 ***
Preco_Leite    -8.71       1.90   -4.58          0.00010 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.61 on 26 degrees of freedom
Multiple R-squared:  0.886,     Adjusted R-squared:  0.872
F-statistic: 67.1 on 3 and 26 DF,  p-value: 0.00000000000225
```

O R^2 obtido foi de 0,886 (88,6%). Podemos interpretar que o modelo ajustado com esses preditores, consegue explicar 88,6% das vendas do café. A parte não explicada é devido a fatores aleatórios e a variáveis não inclusas no modelo. Resumindo, quanto maior o R^2, melhor.

O R^2 ajustado é uma métrica interessante para comparar modelos, pois ela penaliza a adição de novas variáveis preditoras. Sempre que adicionarmos uma nova variável preditora no modelo, o R^2 irá aumentar automaticamente. Entretanto, se essa nova variável não contribuir de forma significativa para o modelo, o R^2 ajustado irá diminuir.

O R^2 pode ser obtido pela equação:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Onde $\hat{Y}_i$ é a predição para i-ésima observação e $\bar{Y}$ é a média da variável reposta. Ou seja, o R^2 é uma razão entre a variação dos valores preditos em torno da média com os valores originais em torno na média.

O R^2 ajustado pode ser obtido pela equação:

$$R^2_{ajustado} = 1 - \left(\frac{n-1}{n-k}\right)(1 - R^2)$$

Onde n é a quantidade de observações e k é a quantidade de preditores.

- **Diagnóstico dos resíduos**

Outro ponto extremamente importante para avaliar a qualidade do ajuste de um modelo de regressão linear é o diagnóstico de resíduos. O resíduo nada mais é do que o erro de predição. É valor original de Y subtraído pelo valor de Y estimado pela equação, denotado por Y chapéu $\hat{Y}$.

O resíduo geralmente é denotado pela letra grega épsilon.

$$\varepsilon = y - \hat{y}$$

Onde y é o valor original de Y e $\hat{Y}$ é o valor estimado pela equação de regressão.

Um dos pressupostos para o ajuste de um modelo de regressão linear é que os resíduos sigam uma distribuição normal, com média zero e variância constante.

$$\varepsilon \sim N(\mu=0, \sigma=1)$$

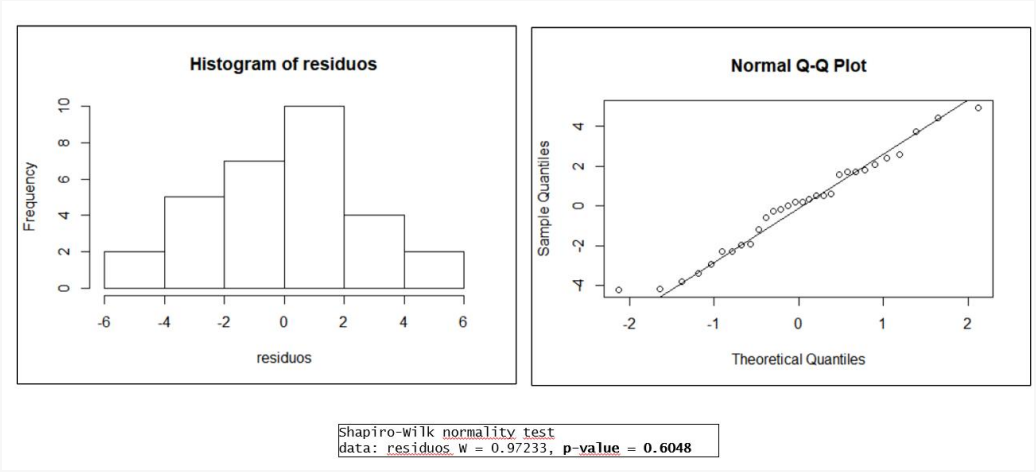
Vamos avaliar a normalidade dos resíduos do modelo de regressão linear que ajustamos para explicar/prever as vendas do café.

Figura 66 – Obtendo os resíduos (Vendas_Cafe – Predicao).

Vendas_Cafe	Predicao	Residuos
18	19.19589	-1.195890576
20	20.19798	-0.197978235
23	22.82836	0.171643614
23	23.59930	-0.599295399
23	26.41762	-3.417615962
23	23.27762	-0.277622271
24	22.18431	1.815685228
25	25.00605	-0.006051994
26	28.28236	-2.282362378
26	27.97733	-1.977332768
26	30.19121	-4.191205838
26	28.27845	-2.278445815
27	23.25978	3.740216509
28	29.90724	-1.907236212
28	32.20772	-4.207717203
29	24.08285	4.917151872
29	26.59714	2.402861112
30	32.95671	-2.956705420
30	29.69051	0.309490313

Vamos utilizar as ferramentas que já conhecemos para avaliar a normalidade da coluna Resíduos.

Figura 67 – Diagnóstico de normalidade dos resíduos.




```
Shapiro-Wilk normality test
data: residuos W = 0.97233, p-value = 0.6048
```

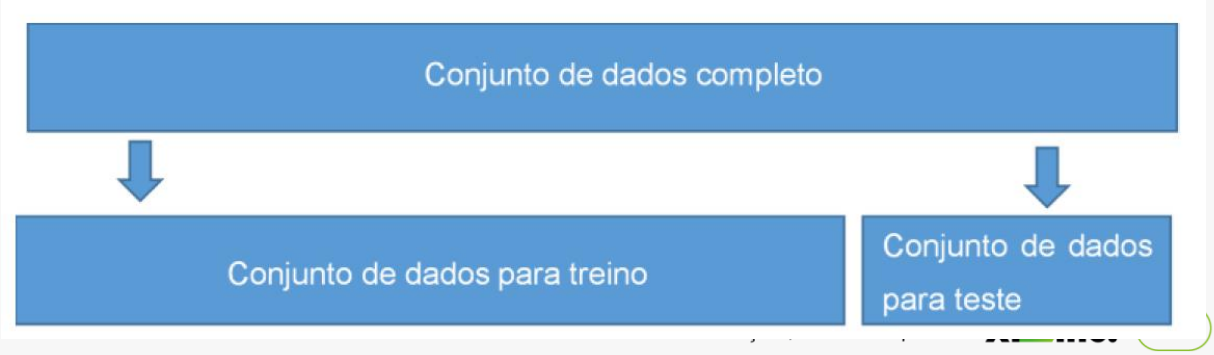
Temos evidências de que os resíduos seguem uma distribuição normal. Esse pressuposto de normalidade é principalmente em função dos testes t e F que o modelo realiza para avaliar a significância estatística dos coeficientes e a qualidade de ajuste do modelo de regressão de forma geral, pois já vimos no capítulo de teste de hipóteses que precisamos de normalidade nos dados para prosseguir com esses testes. No caso da regressão linear, a normalidade é exigida nos resíduos e não necessariamente nas variáveis.

É preferível que a variável resposta Y possua distribuição normal, mas não é um pressuposto, o importante é que os resíduos sigam uma distribuição normal.

O pressuposto de normalidade dos resíduos não é tão obrigatório caso o objetivo do pesquisador seja somente realizar previsões. No entanto, para realizar inferências para a população, a normalidade é um pressuposto que não deve ser violado. Principalmente para amostras de n≤30.

Caso o objetivo do pesquisador seja realizar predições, é extremamente aconselhado que ele separe um percentual do conjunto de dados para ajustar a regressão, aplique o modelo nos dados que ficaram de fora e calcule o R^2 nesses dados. Esse procedimento é chamado de hold-out, e é importante para evitar o overfitting, que é quando um modelo estatístico se super ajusta aos dados de treinamento apresentando um R^2 altíssimo, porém quando chegam dados que ele ainda não conhece o modelo tem dificuldade em acertar nas previsões.

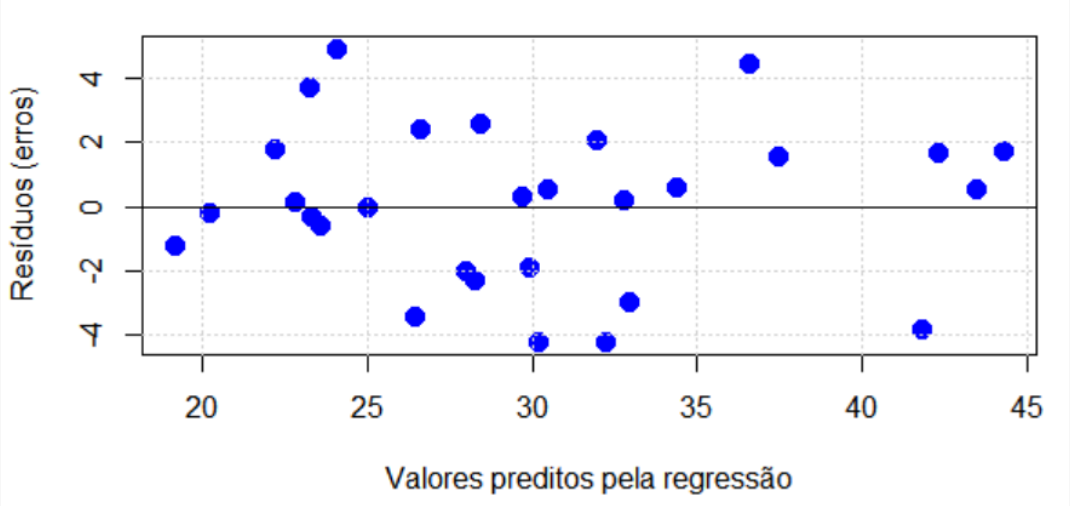
Figura 68 – Metodologia hold-out.



Outras metodologias e métricas pra validar a capacidade preditiva de um algoritmo podem ser acessadas neste link.

Outro pressuposto que deve ser avaliado é o de homocedasticidade dos resíduos. Ou seja, a variância dos resíduos deve ser constante na medida em que os valores preditos aumentam. Esse diagnóstico pode ser realizado de forma gráfica, plotando os resíduos no eixo Y e os valores preditos pela equação de regressão no eixo X. Vamos visualizar como fica o gráfico para o modelo que ajustamos para as vendas do café.

Figura 69 – Diagnóstico de homocedasticidade dos resíduos.

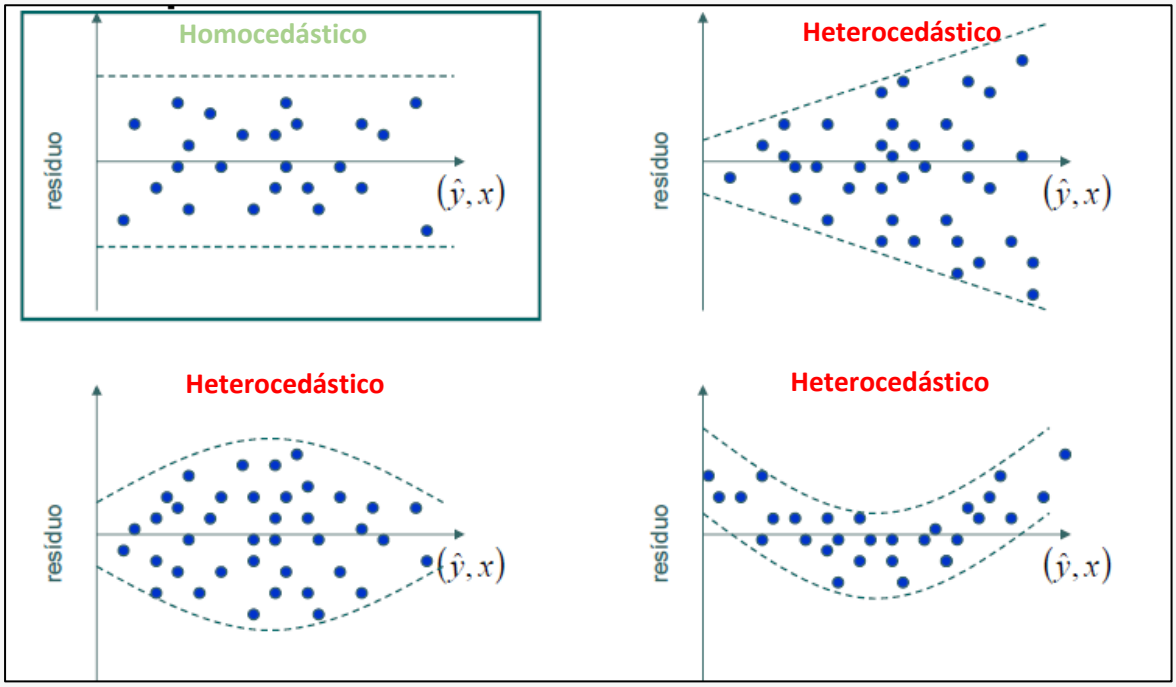


A ideia é que os erros de predição (resíduos) se mantenham constantes a medida que os valores preditos vão aumentando, para garantir que o modelo tenha boa capacidade preditiva para valores baixos e não se perca quando os valores forem altos, e vice versa. Uma forma de se orientar pelo gráfico é se os pontos estão distribuídos no mesmo padrão aleatório do início ao fim do gráfico em torno do zero.

Vamos ver alguns padrões hipotéticos de quando os resíduos (erros) não são homocedásticos, ou seja, não são constantes. O primeiro gráfico é um padrão de resíduos esperado, se distribuindo aleatoriamente em torno

da reta no zero e mantendo a variação em uma amplitude constantes. Os outros três gráficos são exemplos de resíduos não homocedásticos.

Figura 70 – Exemplos hipotéticos de resíduos não homocedásticos (heterocedástico).



Uma possível solução quando ocorre um padrão heterocedástico é aplicar alguma transformação matemática na variável resposta (ou nas preditoras) para tentar estabilizar a variância dos resíduos. Por exemplo, ao invés de trabalhar com as vendas, podemos trabalhar com o logaritmo natural das vendas ou com a raiz quadrada das vendas.

Seleção automática de variáveis preditoras

É comum nos depararmos com algum contexto em que tenhamos um grande conjunto de variáveis preditoras, e pode ser inviável ficar analisando o valor p de cada uma e reajustando modelos. Então, pode ser interessante fazer uso de alguma técnica de seleção automática de variáveis. Veremos sobre o algoritmo chamado Stepwise.

Ele possui três abordagens:

Método Forward – O método inicia com um modelo vazio, contendo apenas o intercepto. Escolhe a variável candidata para entrar sendo aquela que trará a melhor qualidade de ajuste. Assim que é acrescentada, uma variável nunca é removida.

Método Backward – O método inicia com um modelo cheio contendo todos preditores. Escolhe qual será o próximo a sair baseado em algum critério de qualidade de ajuste. Uma vez removida, a variável não retorna ao modelo.

Método Both – Uma combinação dos dois métodos anteriores. Começa com um modelo vazio. Entretanto, uma variável que for adicionada em uma interação pode ser removida do modelo nas próximas interações. Ou seja, as variáveis podem sair e voltar para o modelo durante as interações do algoritmo. É a alternativa mais cara computacionalmente, porém mais eficiente.

Métricas de qualidade de ajuste podem ser variadas, por exemplo, o AIC (*Akaike information criterion), o BIC (Bayesian information criterion) e o R2 ajustado.*

Por default, o R utiliza o AIC, ou seja, o algoritmo Stepwise vai retornar o modelo de regressão linear em que o conjunto de variáveis preditoras presente no modelo dará o menor valor de AIC. É uma métrica que não tem interpretação direta, apenas quanto menor o valor de AIC, melhor é o ajuste do modelo.

O cálculo do AIC pode ser dado pela equação:

$$AIC = -2\ln\left(\frac{\sum_i^n (y_i - \hat{y}_i)^2}{n}\right) + 2k$$

Onde n é a quantidade de observações na amostra e k é a quantidade de preditores no modelo de regressão ajustado.

Em resumo, independente de qual das três abordagens Stepwise o pesquisador utilize, em cada etapa do algoritmo ele irá ajustar um novo modelo de regressão considerando variáveis diferentes, e para cada modelo o AIC será calculado. O modelo final escolhido pelo algoritmo Stepwise é aquele cujo conjunto de variáveis preditoras utilizadas resultou no menor valor de AIC.

Estatística Computacional – Regressão Linear com o R

```
#############################################################
##
#######          Regressao Linear            #######
##      AED - Capitulo 05  - Prof. Máiron Chaves     ####
#############################################################
##
#Copie este código, cole no seu R e execute para ver os resultados
rm(list = ls()) #Limpa memória do R
library(ggplot2) #Biblioteca pra gerar visualizacoes mais sofisticadas
library(plotly)  #Biblioteca pra gerar visualizacoes mais sofisticadas
#Cria o data frame
dados <- data.frame(Vendas_Cafe = c(18, 20, 23, 23, 23, 23, 24, 25, 26, 26, 26, 26, 27, 28, 28,
                                    29, 29, 30, 30, 31, 31, 33, 34, 35, 38, 39, 41, 44, 44, 46),
                    Preco_Cafe = c(4.77, 4.67, 4.75, 4.74, 4.63, 4.56, 4.59, 4.75, 4.75, 4.49,
                                   4.41, 4.32, 4.68, 4.66, 4.42, 4.71, 4.66, 4.46, 4.36, 4.47, 4.43,
                                   4.4, 4.61, 4.09, 3.73, 3.89, 4.35, 3.84, 3.81, 3.79),
                    Promocao = c("Nao", "Nao", "Nao", "Nao", "Nao", "Nao", "Nao", "Nao", "Sim",
                                 "Nao", "Sim", "Nao", "Nao", "Sim", "Sim", "Nao", "Sim", "Sim",
                                 "Sim", "Nao", "Nao", "Sim", "Sim", "Sim", "Nao", "Sim", "Sim",
                                 "Sim", "Sim", "Sim"),
                    Preco_Leite = c(4.74, 4.81, 4.36, 4.29, 4.17, 4.66, 4.73, 4.11, 4.21, 4.25,
                                    4.62, 4.53, 4.44, 4.19, 4.37, 4.29, 4.57, 4.21, 4.77, 4, 4.31,
```

```
                4.34, 4.05, 4.73, 4.07, 4.75, 4, 4.15, 4.34, 4.15) )
View(dados)
#Explorando os dados
#Relacao entre preco do cafe e suas vendas
plot(y = dados$Vendas_Cafe,
     x = dados$Preco_Cafe,
     main = 'Relação entre Vendas do Café VS Preço do Café',
     xlab = 'Preço do Café',
     ylab = 'Qtde Vendida do Café',
     pch = 16)
grid()
#E possivel gerar grafico mais sofisticado utilizando a biblioteca ggplot
g1 <- ggplot(data = dados, aes(y = Vendas_Cafe, x = Preco_Cafe)) + geom_point()
#Podemos adicionar uma reta de regressao com o argumento geom_smooth
g1 + geom_smooth(method = 'lm')
ggplotly(g1) #este comando vem da biblioteca plotly. Passe o cursor do mouse no pontos
do gráfico
#Visualiza a correlacao de Pearson entre as vendas e o preco do cafe
cor(dados$Vendas_Cafe, dados$Preco_Cafe) #Observe que é o mesmo valor que calculamos
na apostila
#Relacao entre preco do leite e as vendas do café
plot(y = dados$Vendas_Cafe,
     x = dados$Preco_Leite,
     main = 'Relação entre Vendas do Café VS Preço do Leite',
     xlab = 'Preço do Leite',
     ylab = 'Qtde Vendida do Café',
     pch = 16)
grid()
#Coeficiente de correlacao preco de leite e as vendas do cafe
cor(dados$Preco_Leite, dados$Vendas_Cafe)
#Grafico 3D entre as vendas do cafe, preco do cafe e preco do leite
#O gráfico e interativo, arraste-o com o mouse
plot_ly(dados, z = ~Vendas_Cafe,
```

```
      x = ~Preco_Cafe,
      y = ~Preco_Leite) %>% add_markers()
#Relacao entre vendas com promocao e sem promocao
boxplot(dados$Vendas_Cafe ~ dados$Promocao)
#Tambem podemos utilizar ggplot e o plotly
g2 <- ggplot(data = dados, aes(y = Vendas_Cafe, x = Promocao, col = Promocao)) +
geom_boxplot()
ggplotly(g2)


#Podemos configurar a tela para exibir multiplos graficos
par(mfrow = c(2,2))
 plot(y = dados$Vendas_Cafe,
      x = dados$Preco_Cafe,
      pch = 16,
      main = 'Vendas Cafe vs Preco Cafe')
  plot(y = dados$Vendas_Cafe,
      x = dados$Preco_Leite,
      pch = 16,
      main = 'Vendas Cafe vs Preco Leite')
 boxplot(dados$Vendas_Cafe ~ dados$Promocao,
         main = 'Vendas Cafe vs Promocao')
 hist(dados$Vendas_Cafe,
      main = 'Distribuicao das vendas do cafe')
dev.off()
#Ajusta um modelo de regressao linear multipla
modelo <- lm(Vendas_Cafe ~  Preco_Cafe + Preco_Leite + Promocao, data = dados)
#Visualiza resumo do ajuste do modelo
summary(modelo)

#Diagnostico de residuos
par(mfrow = c(2,2))
```

```
    plot(modelo,pch = 16)
dev.off()
#E se chegasse novos dados para realizarmos predicoes
#Iremos criar um data frame sem a variavel resposta vendas do cafe, pois ela sera estimada pela equacao de regressao que ajustamos
dados_para_predicao <- data.frame(Preco_Cafe = c(4.77, 4.67, 4.75),
                                  Promocao = c("Nao", "Nao", "Sim"),
                                  Preco_Leite = c(4.74, 4.81, 4.36) )
#Observe que nao ha variavel resposta 'Vendas do Cafe'
View(dados_para_predicao)
#Estima a variavel resposta pra cada observacao do novo data frame
predicoes <- predict(modelo, newdata = dados_para_predicao)
View (data.frame(dados_para_predicao, predicoes))
## Metodo Stepwise para selecao automatica de variaveis
nova_variavel = rpois(n = 30, lambda = 2)
fit2 <- lm(Vendas_Cafe ~ Preco_Cafe +  Promocao +  Preco_Leite + nova_variavel, data = dados)
summary(fit2) #Observe o p valor da nova variavel, nao e significativo
fit2 <- step(fit2, direction = 'both')
summary(fit2) #Observe que o stepwise removeu a nova variavel
```

> Capítulo 6

Capítulo 6. Regressão Logística

É muito comum, no mundo real, nos deparamos com situações em que precisamos prever uma categoria de saída ao invés de um número. Por exemplo, supondo que trabalhamos em uma financeira e chega um novo cliente para nos solicitar um empréstimo, de acordo com as características do cliente, qual a probabilidade dele ser Adimplente (1) e qual a probabilidade dele ser Inadimplente (0). Se conseguirmos prever que este cliente tem maior probabilidade de ser Adimplente ficaremos mais seguros em conceder o empréstimo.

Outro exemplo no setor de RH, ao contratar um novo funcionário para equipe de vendas, de acordo as características desse candidato e suas respostas nos testes, qual a probabilidade de ele pertencer a categoria de Bom Vendedor (1) e qual a probabilidade dele pertencer a categoria de Mal Vendedor (0). Se conseguirmos prever quais candidatos tem maior probabilidade de ser um Bom Vendedor podemos priorizá-los nas contratações.

No entanto, vimos que a Regressão Linear deve ser utilizada quando a variável resposta é numérica (de preferência contínua), para prever uma classe, podemos utilizar a Regressão Logística, que faz parte dos modelos lineares generalizados e é uma variação da regressão de mínimos quadrados ordinários. Portanto, apesar de conter a palavra Regressão em seu nome, a Regressão Logística é utilizada para tarefa de classificação, pois é baseada na distribuição binomial.

A regressão logística modela a probabilidade de uma observação pertencer a uma determinada categoria de saída, que deve ser binária. As categorias são 1 = Ocorrerá o evento de interesse e 0 = Não ocorrerá o evento de interesse.

O modelo de Regressão Logística ajustado tem a forma:

$$\hat{y} = \Pr(y = 1 \mid x)$$

Onde X é matriz de preditores, β são os coeficientes da equação e 'e' é a função exponencial.

A função custo que a regressão logística tem que minimizar (também chamada de Binary Cross Entropy) é dada por:

$$\min_{\beta,\beta_0} = \frac{1}{N}\sum_{i=1}^{N} y_i * \log(\hat{y}_i) + (1 - y_i) * \log(1 - \hat{y}_i)$$

Onde N é a quantidade de observações no conjunto de treinamento, yi é o i-ésimo valor da variável resposta (que assume 0 ou 1) e $\hat{y}_i$ é a probabilidade estimada para a i-ésima observação (que assume algum valor entre 0 e 1)

A estimação dos coeficientes β, ao contrário da regressão linear que tem fórmula fechada para encontrar os coeficientes da equação **através do método dos mínimos quadrados ordinários**, a regressão logística não tem, então é necessário fazer uso de algum **método** interativo. **Na regressão logística, é utilizado o método dos mínimos quadrados reponderados interativos, também conhecido como método da máxima verossimilhança. Algum algoritmo de otimização numérica deve ser utilizado**, como o método de Newton-Raphson.

Interpretando o modelo ajustado

Assim como na Regressão Linear, a Regressão Logística serve tanto para explicação como para predição. Vamos propor um contexto para que a teoria fique mais tangível.

Suponha que você trabalha em uma grande empresa do segmento de RH e deseja agilizar suas contratações para uma determinada empresa que é um grande cliente. Este cliente demanda sempre uma grande quantidade de contratações simultâneas, pois sempre que fecha um novo contrato uma nova operação é inicializada e há uma grande demanda por novos funcionários.

Processos seletivos onde n candidatos são colocados em uma sala e submetidos a diversos testes, podem ter um alto custo além de exigir que o RH corrija questão por questão de cada teste dos n candidatos, elevando o tempo gasto.

Portanto, a empresa de RH reuniu em um conjunto de dados as notas obtidas nas avaliações em processos seletivos anteriores e vinculou a performance dos candidatos, contendo n = 699 observações e rotulando como 'Boa' os funcionários que tiveram bom desempenho e 'Ruim' os funcionários que não obtiveram bom desempenho.

Utilizaremos a Regressão Logística tanto para mensurar o impacto de cada preditor na variável resposta como também a utilizaremos para prever se um novo candidato será da classe 'Boa' ou da classe 'Ruim'.

Figura 71 – Conjunto de dados reunido pelo RH.

	Prova_Logica	Redacao	Psicotecnico	Dinamica_Grupo	Fit_Cultural	Ingles	Avaliacao_RH	Auto_Avaliacao	Classe
1	2	1	1	1	2	1	2	1	Ruim
2	2	1	1	1	2	1	3	1	Ruim
3	5	1	1	1	2	1	2	1	Ruim
4	5	4	6	8	4	1	8	1	Boa
5	5	3	3	1	2	1	2	1	Ruim
6	2	3	1	1	3	1	1	1	Ruim
7	3	5	7	8	8	9	7	1	Boa
8	2	5	6	1	6	1	7	7	Boa
9	1	9	8	7	6	4	7	1	Boa
...									
697	6	1	1	1	8	1	7	1	Boa
698	5	7	1	1	5	1	1	1	Boa
699	1	1	1	1	2	1	2	1	Ruim

Vamos examinar o output que o R nos fornece para regressão logística. Para simplificar, por enquanto, vamos utilizar as pontuações do candidato em *Prova_Logica*, *Redacao* e *Auto_Avaliacao* para explicar/prever a Classe.

Figura 72 – Output da Regressão Logística fornecido pelo R.

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.84584    0.27818 -13.825  < 2e-16 ***
Prova_Logica     0.21431    0.05186   4.132 3.59e-05 ***
Redacao          0.69144    0.07437   9.298  < 2e-16 ***
Auto_Avaliacao   0.42384    0.06730   6.298 3.02e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Marcados de azul, temos os coeficientes para montar a equação, ou seja, são os coeficientes β. Quando o coeficiente é positivo, nos diz que na medida que o preditor aumenta, a probabilidade do evento de interesse (classe positiva) ocorrer também aumenta. Quando o coeficiente é negativo nos diz que na medida que o preditor aumenta, a probabilidade do evento de interesse (classe negativa) ocorrer diminui. Lembrando que neste caso o nosso evento de interesse é prever a classe ‘Boa’.

Ao contrário da Regressão Linear, a interpretação não é direta, para saber o quanto o preditor impacta nas chances do evento de interesse ocorrer, devemos aplicar a função exponencial ao coeficiente β.

Vamos interpretar os coeficientes (com exceção do β_0, que não tem uma interpretação muito prática na Regressão Logística):

β_1= Prova_Logica = exp(0,2143) = $_{2,7182}$^0,2143 = **1,23** – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na prova de lógica, o candidato aumenta em média 1,23 vezes as chances de pertencer a classe ‘Boa’.

β_2= Redacao = exp(0,6914) = $_{2,7182}$^0,6914= **1,99**– Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na prova de redação, o candidato aumenta em média, 1,99 vezes as chances de pertencer a classe 'Boa'.

β_3= Auto_Avaliacao = exp(0,4238) = $_{2,7182}$^0,4238= **1,5278** – Ou seja, mantendo as demais variáveis constantes, para cada ponto a mais na auto avaliação, o candidato aumenta em 1,52 vezes as chances de pertencer a classe 'Boa'.

Observe na figura 73 que também temos um teste hipótese para cada coeficiente β.

Figura 73 – Teste de hipótese para os coeficientes β.

Test t:
H0: $\beta=0$
H1: $\beta\neq0$

Ou seja

H0: A variável preditora não tem relação significativa com a variável resposta
H1: A variável preditora tem relação significativa com a variável resposta

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.84584    0.27818 -13.825  < 2e-16 ***
Prova_Logica  0.21431    0.05186   4.132 3.59e-05 ***
Redacao       0.69144    0.07437   9.298  < 2e-16 ***
Auto_Avaliacao 0.42384   0.06730   6.298 3.02e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Na coluna z value temos o z calculado (Estimate dividido pelo Std. Error) e, em seguida, temos o valor p (Pr(>|t|) . Se fixamos um nível de significância, por exemplo, α = 5%, rejeitaremos H_0 se o valor p for abaixo de 5% (0,05). Se rejeitarmos significa que o coeficiente daquela variável não é estatisticamente diferente de zero. Ou seja, aquela variável preditora não tem impacto significativo na variável resposta. Geralmente, nesses casos a variável preditora é removida do modelo pelo pesquisador e um novo modelo sem ela é ajustado.

A equação linear utilizando os coeficientes fornecidos pelo R ficaria:

$$\hat{y} = -3{,}8458 + (0{,}2143 * Prova\_Logica) + (0{,}6914 * Redacao) + (0{,}4238 * Auto\_Avaliacao)$$

Entretanto, para que o resultado da predição fique sempre entre 0 e 1, devemos incluir a equação na função logística.

$$\hat{y} = \frac{e^{-3{,}8458+(0{,}2143*Prova\_Logica)+(0{,}6914*Redacao)+(0{,}4238*Auto\_Avaliacao)}}{1 + e^{-3{,}8458+(0{,}2143*Prova\_Logica)+(0{,}6914*Redacao)+(0{,}4238*Auto\_Avaliacao)}}$$

Supondo que o candidato tire 3 em Prova_Logica, 5 em Redacao e 1 em Auto_Avaliacao. A probabilidade dele(a) pertencer a classe de interesse 'Boa' fica:

$$\hat{y} = \frac{e^{-3{,}8458+(0{,}2143*3)+(0{,}6914*5)+(0{,}4238*1)}}{1 + e^{-3{,}8458+(0{,}2143*3)+(0{,}6914*5)+(0{,}4238*1)}}$$

$$\hat{y} = \frac{1{,}9697}{2{,}9697}$$

$$\hat{y} = 0{,}6632\ (ou\ 66{,}32\%)$$

Portanto, um candidato com essa pontuação nas três variáveis teria 66,32% de probabilidade de pertencer à classe 'Boa'.

Avaliando a Performance Preditiva do modelo

O output da Regressão Logística é uma probabilidade, então um ponto de corte (threshold) deve ser definido de forma que, se a probabilidade estimada for acima do ponto de corte, a predição será considerada como pertencente ao evento positivo 1 (neste contexto, é a classe 'Boa'), caso contrário, a predição será considerada como pertencente ao evento negativo 0 (neste contexto, é a classe 'Ruim'). Se adotarmos o critério de que se a probabilidade for acima de 50%, consideraremos a classe positiva e caso seja abaixo, consideraremos a classe negativa. Teríamos a seguinte classificação para cada observação:

Figura 74 – Classe predita considerando 0,5 como ponto de corte.

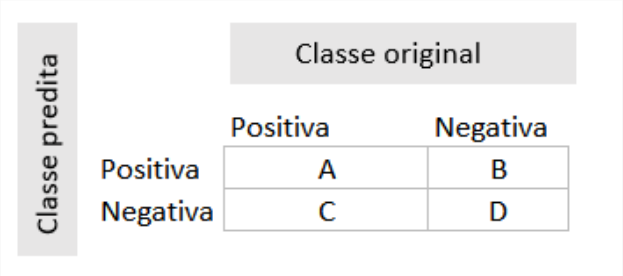
	Prova_Logica	Redacao	Auto_Avaliacao	Classe	Probabilidade	Classe_Predita
1	2	1	1	Ruim	0.09096250	Ruim
2	2	1	1	Ruim	0.09096250	Ruim
3	5	1	1	Ruim	0.15989589	Ruim
4	5	4	1	Boa	0.60235935	Boa
5	5	3	1	Ruim	0.43140265	Ruim
6	2	3	1	Ruim	0.28514808	Ruim
7	3	5	1	Boa	0.66331842	Boa
8	2	5	7	Boa	0.95288247	Boa
9	1	9	1	Boa	0.95325780	Boa
10	4	1	1	Ruim	0.13315862	Ruim
11	5	1	1	Ruim	0.15989589	Ruim
12	8	1	9	Boa	0.91487406	Boa
...						
697	6	1	1	Boa	0.19081999	Ruim
698	5	7	1	Boa	0.92341050	Boa
699	1	1	1	Ruim	0.07472674	Ruim

Para conferir os erros e acertos da Regressão Logística, podemos comparar a coluna Classe com a coluna Classe_Predita. Entretanto, fazer essa análise para muitas observações não é viável, então, para este fim, iremos resumir essas duas colunas em uma tabela chamada de Matriz de Confusão.

A Matriz de confusão é uma tabela que nos informa os erros e acertos de predição do algoritmo. Normalmente, em suas linhas tem a classe que foi predita e nas colunas as classes originais que cada observação pertence. A matriz de confusão é calculada utilizando apenas duas colunas, a variável resposta que já estava no conjunto de dados e uma nova coluna informando a classe que o algoritmo predisse para cada observação. Independente do contexto, sempre que trabalhamos com uma tarefa de classificação, deve ser definida qual será a classe positiva e a negativa. A classe positiva não necessariamente é algo bom, mas é o que se busca predizer com o algoritmo, neste contexto, a classe positiva é o candidato pertencer à classe 'Boa', pois o objetivo do algoritmo é predizer se o

candidato terá uma boa performance, e não predizer que o candidato não terá uma boa performance.

Figura 75 – Estrutura geral de uma Matriz de Confusão para um classificador binário.



A célula A trará a quantidade de observações preditas como pertencentes da classe positiva e que realmente eram da classe positiva. Ou seja, o quanto o algoritmo acertou para a classe positiva (verdadeiro positivo).

A célula B trará a quantidade de observações preditas como pertencentes da classe positiva, mas que originalmente pertencem a classe negativa. Ou seja, o quanto o algoritmo errou para a classe positiva (falso positivo).

A célula C trará a quantidade de observações preditas como pertencentes da classe negativa, mas que originalmente pertencem a classe positiva. Ou seja, o quanto o algoritmo errou para a classe negativa (falso negativo).

A célula D trará a quantidade de observações preditas como pertencentes da classe negativa e que realmente eram da classe negativa. Ou seja, o quanto o algoritmo acertou para a classe negativa (verdadeiro negativo).

Observe que os elementos da diagonal principal da matriz (células A e D) correspondem a quantidade de acertos, e os elementos fora da diagonal principal (células B e C) correspondem aos erros.

A Matriz de Confusão para as predições da figura 74, fica:

Figura 76 – Matriz de Confusão considerando 0,5 como ponto de corte.

	Classe_Original	
Classe_Predita	Boa	Ruim
Boa	169	23
Ruim	72	435

A partir da Matriz de Confusão, iremos avaliar a taxa de acerto geral (Acurácia), a taxa de acerto para classe positiva (Sensitividade) e a taxa de acertos para classe negativa (Especificidade).

Acurácia – Soma dos elementos da diagonal principal/Soma de todos os elementos.

$$Acurácia = \frac{169 + 435}{169 + 23 + 72 + 435} = 86{,}40\%$$

Sensitividade - A Sensitividade responde a seguinte pergunta: De todas as observações da classe 'Boa', quantas o algoritmo classificou como 'Boa'?

$$Sensitividade = \frac{Verdadeiros\ Positivos}{Verdadeiros\ Positivos + Falsos\ Negativos}$$

$$Sensitividade = \frac{169}{169 + 72} = 70{,}12\%$$

Especificidade – A Especificidade responde a seguinte pergunta: De todas as observações da classe 'Ruim', quantas o algoritmo classificou como 'Ruim'?

$$Especificidade = \frac{Verdadeiros\ Negativos}{Verdadeiros\ Negativos + Falsos\ Positivos}$$

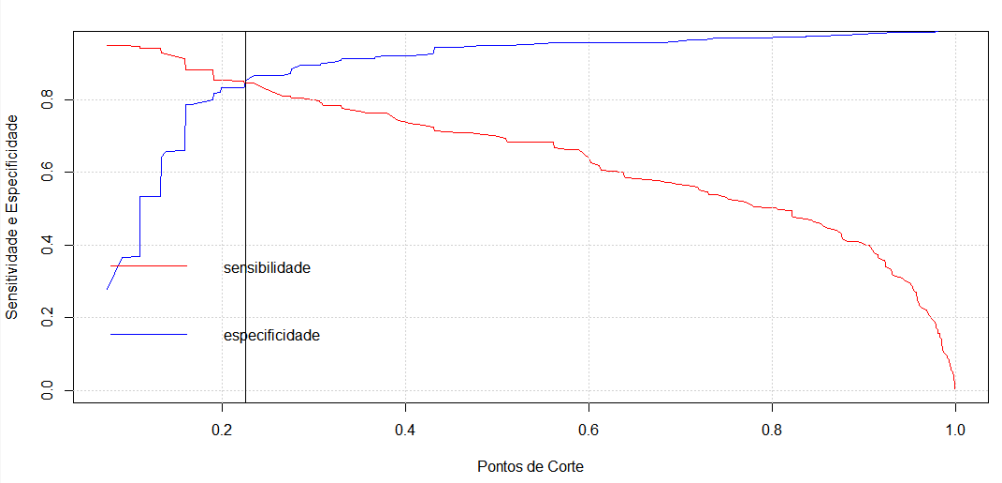
$$Especificidade = \frac{435}{435 + 23} = 94{,}97\%$$

Como podemos perceber, o algoritmo teve uma performance preditiva melhor para a classe negativa. Para tentar resolver este problema, podemos tentar variar o ponto de corte. Veremos como achar o ponto de corte ideal para equilibrar a Sensitividade a Especificidade no próximo tópico.

Análise de Sensibilidade e Especificidade

A ideia da Análise de Sensibilidade e Especificidade é simular várias matrizes de confusão, através de vários pontos de corte diferentes e identificar aquela matriz de confusão que nos dará tanto a maior Sensibilidade quanto a maior Especificidade. Observe na figura 77.

Figura 77 – Sensibilidade e Especificidade para diversos pontos de corte.



Na intersecção das duas curvas é o ponto de corte que nos dará a maior Sensitividade em equilíbrio com a maior Especificidade. Neste caso o ponto de corte é 0,225, ou seja, caso a probabilidade seja acima de 22,5% classificaremos o indivíduo como pertencente da classe positiva ('Boa'), caso contrário, classe negativa ('Ruim').

A matriz de confusão, considerando o ponto de corte de 22,5%, fica:

Figura 78 – Matriz de Confusão considerando 22,5% como ponto de corte.

	Classe_Original	
Classe_Predita	Boa	Ruim
Boa	204	75
Ruim	37	383

A Sensibilidade e a Especificidade ficaram:

$$Sensitividade = \frac{204}{204 + 37} = 84{,}64\%$$

$$Especificidade = \frac{383}{383 + 75} = 83{,}62\%$$

Observe que, comparando com o ponto de corte de 50%, tivemos a Sensitividade de 70,12% e a Especificidade de 94,97%, o ponto de corte de 22,5% nos forneceu uma Sensitividade de 86,64% e a Especificidade de 83,62%. Foi necessário reduzir um pouco a Especificidade para ganhar na Sensitividade.

A análise de Sensitividade e Especificidade nos fornece o ponto de corte de forma matemática para equilibrar a Sensitividade e a Especificidade, mas na prática não devemos desprezar o contexto ao definir o ponto de corte, é sempre interessante analisar, o que é mais custoso, um falso positivo ou um falso negativo? E, dessa forma, o pesquisador pode ir variando o ponto de corte, observando que sempre que a Sensitividade aumentar a Especificidade cairá, e vice-versa.

Estatística Computacional – Regressão Logística no R

```
#############################################################
##

#######          Regressao Logística          #######

##      AED - Capitulo 06  - Prof. Máiron Chaves     ####

#############################################################
##

#Copie este código, cole no seu R e execute para ver os resultados
```

```
rm(list = ls()) #Limpa memória do R
#Instala e carrega biblioteca para gerar a curva ROC
Install.packages('pROC') #Instala
library(pROC) #Carrega
#Monte o dataset
dados <- data.frame(Prova_Logica = c(2, 2, 5, 5, 5, 2, 3, 2, 1, 4,
                                     5, 8, 1, 1, 3, 4, 3, 2, 1, 1, 8, 8, 1, 2, 1, 5, 3, 3, 5, 4, 4,
                                     1, 8, 3, 2, 3, 3, 2, 1, 1, 5, 4, 1, 5, 3, 1, 4, 6, 1, 1, 8, 1,
                                     1, 5, 1, 5, 3, 1, 1, 8, 1, 1, 1, 1, 1, 2, 1, 5, 5, 4, 2, 1, 8,
                                     4, 5, 1, 3, 3, 3, 5, 3, 1, 7, 1, 1, 2, 9, 5, 3, 1, 5, 1, 4, 2,
                                     1, 4, 3, 3, 8, 1, 1, 8, 5, 1, 1, 1, 5, 8, 5, 1, 4, 2, 5, 4, 5,
                                     3, 3, 5, 5, 5, 5, 5, 8, 5, 4, 9, 8, 1, 3, 4, 2, 5, 1, 4, 3, 5,
                                     5, 5, 6, 4, 3, 5, 7, 1, 8, 5, 7, 3, 2, 3, 2, 5, 5, 5, 5, 4, 4,
                                     8, 1, 1, 2, 5, 3, 2, 7, 4, 1, 1, 1, 4, 5, 1, 1, 8, 3, 6, 8, 3,
                                     1, 3, 3, 2, 8, 4, 1, 1, 1, 1, 1, 2, 3, 4, 6, 2, 3, 3, 4, 2, 1,
                                     5, 2, 4, 3, 3, 1, 3, 3, 3, 1, 3, 5, 6, 1, 5, 1, 5, 4, 3, 1, 6,
                                     1, 4, 9, 3, 3, 2, 1, 1, 4, 3, 1, 3, 1, 1, 3, 7, 8, 1, 3, 5, 6,
                                     3, 6, 5, 8, 5, 1, 1, 4, 2, 1, 8, 7, 5, 1, 1, 1, 6, 5, 7, 3, 3,
                                     5, 1, 3, 5, 1, 8, 8, 1, 2, 3, 3, 3, 3, 7, 1, 9, 8, 4, 1, 7, 1,
                                     1, 1, 5, 1, 1, 5, 3, 5, 1, 3, 6, 2, 1, 3, 4, 5, 6, 1, 5, 1, 5,
                                     1, 1, 5, 1, 1, 1, 5, 1, 3, 7, 1, 4, 3, 7, 1, 1, 5, 4, 1, 1, 3,
                                     5, 4, 2, 1, 5, 1, 1, 1, 8, 8, 5, 1, 2, 1, 6, 8, 3, 1, 5, 1, 5,
                                     1, 4, 4, 8, 1, 1, 1, 5, 1, 1, 5, 4, 6, 8, 1, 3, 1, 6, 1, 1, 1,
                                     1, 1, 8, 1, 5, 3, 1, 4, 4, 7, 2, 3, 3, 5, 8, 3, 1, 4, 1, 5, 1,
                                     7, 2, 6, 4, 1, 3, 1, 8, 5, 5, 5, 3, 1, 4, 5, 3, 6, 1, 3, 3, 5,
                                     4, 5, 3, 1, 4, 5, 1, 3, 6, 1, 1, 1, 3, 1, 1, 1, 1, 3, 1, 5, 4,
                                     8, 1, 5, 6, 6, 4, 2, 5, 5, 6, 1, 2, 5, 6, 4, 2, 1, 2, 7, 2, 8,
                                     1, 3, 1, 1, 1, 6, 1, 4, 3, 1, 5, 2, 1, 1, 5, 4, 3, 1, 3, 1, 1,
                                     2, 4, 5, 4, 5, 4, 3, 7, 4, 1, 7, 1, 8, 5, 3, 3, 5, 1, 2, 1, 5,
                                     4, 4, 4, 3, 6, 8, 8, 1, 1, 1, 8, 8, 1, 5, 1, 4, 5, 1, 4, 1, 4,
                                     1, 5, 1, 4, 4, 1, 1, 2, 5, 1, 4, 4, 5, 4, 1, 1, 5, 3, 5, 4, 1,
                                     1, 5, 3, 3, 1, 5, 5, 4, 5, 7, 2, 5, 9, 4, 6, 1, 1, 1, 1, 8, 7,
```

2, 3, 2, 9, 3, 1, 5, 1, 3, 5, 1, 4, 6, 3, 1, 5, 1, 3, 9, 1, 4,

8, 8, 1, 3, 2, 3, 3, 1, 5, 1, 1, 3, 1, 5, 3, 4, 3, 4, 1, 4, 3,

5, 1, 5, 2, 5, 8, 9, 4, 7, 4, 8, 5, 7, 5, 3, 5, 6, 3, 1, 5, 6,

4, 3, 5, 1, 2, 4, 1, 1, 2, 9, 5, 1, 5, 1, 2, 1, 5, 1, 2, 3, 5,

1, 1, 5, 3, 9, 5, 6, 9, 6, 1, 1, 9, 1, 3, 5, 4, 8, 4, 2, 7, 1,

6, 3, 7, 8, 1, 5, 5, 6, 1, 4, 4, 3, 1, 5, 5, 8, 3, 3, 1, 9, 5,

5, 5, 5, 1, 9, 6, 3, 1, 1, 2, 4, 6, 5, 7, 6, 5, 1), Redacao = c(1,

1, 1, 4, 3, 3, 5, 5, 9, 1, 1, 1, 1, 1, 4, 3, 3, 1, 1, 1, 1, 7,

1, 1, 8, 1, 1, 1, 3, 8, 3, 1, 5, 3, 3, 1, 2, 7, 1, 1, 1, 1, 1,

1, 7, 1, 1, 8, 7, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 1, 8, 5,

1, 1, 1, 1, 1, 2, 1, 1, 6, 1, 1, 1, 1, 1, 1, 1, 4, 1, 8, 5, 1,

5, 8, 1, 1, 1, 5, 1, 1, 1, 2, 3, 3, 1, 3, 8, 1, 4, 6, 1, 1, 1,

1, 3, 1, 1, 2, 3, 2, 1, 1, 1, 1, 4, 1, 1, 1, 1, 4, 1, 8, 1, 5,

1, 1, 1, 1, 1, 3, 6, 1, 2, 5, 6, 1, 2, 2, 1, 8, 1, 4, 6, 9, 3,

1, 1, 1, 1, 1, 1, 1, 1, 1, 7, 1, 1, 1, 1, 1, 1, 6, 1, 1, 1, 1,

1, 3, 5, 1, 1, 1, 1, 3, 1, 4, 1, 1, 1, 2, 1, 3, 1, 1, 1, 4, 5,

1, 1, 6, 1, 3, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3,

6, 1, 8, 1, 1, 5, 1, 8, 2, 6, 1, 5, 1, 6, 1, 1, 1, 1, 1, 1, 1,

1, 3, 1, 4, 8, 1, 1, 1, 8, 1, 3, 1, 6, 3, 1, 1, 1, 1, 1, 1, 1,

1, 1, 1, 1, 1, 1, 5, 1, 1, 3, 1, 1, 7, 4, 1, 1, 1, 1, 6, 1, 3,

1, 4, 1, 1, 7, 2, 6, 4, 1, 1, 1, 1, 1, 4, 7, 1, 3, 1, 1, 9, 1,

1, 1, 1, 1, 1, 1, 3, 1, 3, 1, 3, 1, 1, 3, 2, 1, 1, 1, 5, 3, 1,

1, 2, 1, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 4, 8, 1, 7, 1,

1, 3, 8, 1, 1, 1, 1, 1, 4, 1, 1, 1, 2, 2, 7, 1, 3, 1, 1, 1, 4,

2, 4, 2, 2, 5, 3, 1, 1, 1, 5, 1, 9, 1, 1, 3, 2, 1, 1, 5, 1, 2,

1, 3, 8, 1, 5, 1, 4, 3, 1, 8, 1, 6, 5, 1, 1, 1, 1, 1, 4, 5, 1,

7, 8, 1, 4, 1, 1, 1, 1, 4, 1, 1, 2, 1, 8, 2, 6, 2, 1, 4, 1, 1,

1, 1, 1, 4, 1, 1, 1, 1, 1, 8, 1, 1, 1, 3, 1, 1, 1, 8, 1, 1, 1,

3, 1, 1, 1, 1, 1, 6, 1, 1, 1, 1, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1,

3, 1, 2, 7, 2, 1, 1, 1, 1, 1, 2, 2, 1, 3, 1, 1, 3, 1, 1, 5, 1,

7, 1, 1, 1, 3, 6, 1, 1, 1, 1, 1, 1, 1, 1, 4, 1, 6, 8, 8, 7, 2,

1, 1, 1, 1, 1, 1, 1, 5, 1, 1, 5, 1, 1, 1, 1, 9, 1, 8, 1, 1, 2,

4, 1, 1, 6, 2, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 6, 1, 2,

1, 1, 5, 4, 1, 8, 4, 6, 6, 1, 1, 1, 9, 1, 1, 1, 1, 1, 8, 1, 1,

1, 1, 1, 3, 1, 1, 4, 1, 1, 3, 4, 1, 1, 3, 2, 3, 1, 2, 1, 1, 1,

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 3, 1, 4, 1, 4, 2, 1, 6, 1,

4, 2, 2, 1, 1, 1, 4, 1, 1, 1, 1, 1, 6, 1, 1, 1, 3, 2, 8, 1, 1,

1, 1, 1, 2, 3, 1, 1, 1, 1, 1, 1, 6, 1, 6, 7, 1, 1, 5, 1, 2, 5,

1, 1, 1, 1, 1, 2, 1, 3, 1, 1, 1, 8, 7, 1, 1, 1, 1, 4, 1, 6, 1,

2, 8, 4, 7, 1, 1, 1, 5, 1, 1, 2, 1, 1, 7, 1, 1, 1, 4, 1, 1, 3,

1, 5, 1, 7, 1), Auto_Avaliacao = c(1, 1, 1, 1, 1, 1, 1, 7, 1,

1, 1, 9, 1, 1, 1, 3, 6, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,

1, 1, 3, 1, 1, 1, 1, 9, 1, 1, 1, 1, 1, 1, 8, 1, 1, 9, 7, 1, 2,

1, 1, 1, 1, 1, 1, 1, 1, 7, 1, 3, 8, 1, 1, 1, 1, 1, 1, 6, 1, 1,

1, 2, 1, 1, 1, 1, 1, 1, 3, 1, 8, 3, 1, 6, 1, 6, 1, 1, 3, 1, 1,

1, 1, 8, 5, 3, 3, 1, 1, 3, 1, 1, 1, 1, 1, 6, 1, 2, 1, 1, 1, 1,

1, 4, 1, 6, 1, 1, 1, 2, 3, 1, 4, 7, 6, 1, 1, 1, 1, 1, 1, 9, 1,

2, 1, 1, 1, 1, 2, 1, 8, 1, 8, 1, 3, 2, 1, 1, 1, 1, 2, 6, 1, 1,

1, 2, 1, 3, 1, 1, 8, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 9,

1, 5, 1, 1, 1, 1, 1, 5, 1, 1, 1, 3, 5, 1, 1, 1, 1, 1, 1, 3, 1,

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 6, 1, 8, 1, 2, 5, 2, 1,

1, 7, 1, 1, 1, 8, 1, 1, 5, 1, 1, 1, 3, 1, 1, 1, 9, 1, 1, 1, 5,

9, 1, 5, 3, 4, 1, 1, 1, 1, 1, 1, 7, 1, 1, 1, 1, 1, 3, 3, 1, 3, 1,

1, 1, 1, 1, 1, 3, 8, 1, 4, 1, 4, 1, 1, 1, 6, 1, 3, 1, 1, 2, 3,

1, 1, 1, 1, 1, 1, 1, 1, 9, 1, 1, 2, 2, 8, 1, 1, 1, 1, 1, 4, 1,

1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 4, 3, 1, 1, 4, 1, 6, 2, 1, 5, 3,

1, 1, 2, 1, 1, 1, 1, 1, 1, 8, 3, 4, 8, 1, 3, 7, 7, 1, 1, 2, 1,

1, 1, 1, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 9, 1, 1, 1,

1, 8, 1, 7, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 9, 6, 3, 3,

1, 2, 1, 8, 7, 1, 1, 1, 1, 1, 6, 3, 1, 4, 5, 1, 6, 1, 1, 1, 1,

3, 1, 1, 2, 1, 6, 3, 7, 1, 8, 3, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,

1, 8, 1, 1, 1, 9, 1, 1, 1, 7, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,

1, 1, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 1, 6, 1, 1, 1, 1, 1, 1, 2,

1, 1, 1, 2, 1, 7, 1, 1, 4, 2, 1, 5, 1, 5, 1, 1, 1, 4, 6, 1, 1,

```
  4, 1, 2, 1, 1, 1, 9, 8, 1, 9, 1, 1, 1, 1, 1, 1, 1, 3, 1, 1, 1,
  1, 1, 4, 1, 1, 1, 1, 7, 1, 1, 1, 1, 1, 3, 1, 1, 8, 1, 1, 6, 1,
  1, 1, 2, 1, 1, 1, 1, 1, 1, 2, 5, 1, 6, 3, 1, 4, 3, 1, 7, 5, 1,
  1, 1, 1, 1, 1, 2, 1, 5, 9, 1, 1, 1, 2, 1, 1, 1, 1, 3, 1, 8, 1,
  1, 2, 5, 1, 1, 5, 1, 4, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 1, 1, 3, 1, 1, 3, 1, 5, 1, 8, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 6, 1, 1, 1, 2, 1, 1, 2, 1, 1, 1, 1, 2, 1, 1, 1, 1,
  1, 1, 1, 6, 1, 9, 5, 3, 1, 8, 1, 1, 3, 1, 1, 1, 1, 1, 1, 1, 2,
  1, 1, 1, 8, 1, 1, 1, 1, 1, 1, 1, 7, 1, 1, 1, 1, 1, 1, 1, 1, 1,
  1, 1, 1, 1, 2, 8, 1, 1, 1, 6, 1, 1, 1, 1, 9, 1, 1, 1),
Classe = c("Ruim",
    "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Boa", "Boa", "Ruim",
    "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
    "Ruim", "Ruim", "Boa", "Boa", "Ruim", "Ruim", "Boa", "Ruim",
    "Ruim", "Ruim", "Boa", "Boa", "Ruim", "Ruim", "Boa", "Boa", "Ruim",
    "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Ruim", "Ruim",
    "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Boa", "Ruim", "Boa", "Ruim",
    "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa",
    "Boa", "Boa", "Boa", "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim",
    "Boa", "Ruim", "Boa", "Boa", "Ruim", "Ruim", "Ruim", "Ruim",
    "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa", "Boa", "Ruim",
    "Boa", "Boa", "Boa", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
    "Ruim", "Ruim", "Ruim", "Boa", "Boa", "Boa", "Ruim", "Ruim",
    "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Boa", "Boa", "Ruim",
    "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
    "Ruim", "Boa", "Ruim", "Boa", "Ruim", "Boa", "Boa", "Boa", "Ruim",
    "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
    "Boa", "Boa", "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim",
    "Boa", "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim",
    "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
    "Ruim", "Ruim", "Ruim", "Boa", "Boa", "Ruim", "Ruim", "Ruim",
    "Ruim", "Ruim", "Boa", "Boa", "Boa", "Ruim", "Boa", "Boa", "Ruim",
```

"Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Ruim", "Ruim", "Boa", "Boa", "Ruim", "Ruim", "Boa", "Ruim",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim",
"Boa", "Boa", "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Ruim", "Boa",
"Ruim", "Boa", "Ruim", "Boa", "Ruim", "Boa", "Ruim", "Ruim",
"Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Boa", "Boa", "Ruim", "Ruim", "Boa", "Boa", "Ruim", "Ruim", "Ruim",
"Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa",
"Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Boa", "Boa", "Boa",
"Boa", "Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Boa", "Boa", "Ruim", "Boa", "Boa", "Ruim", "Ruim", "Ruim", "Ruim",
"Ruim", "Boa", "Boa", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
"Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Boa", "Boa", "Ruim",
"Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim",
"Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Boa", "Ruim", "Boa", "Boa", "Boa", "Ruim", "Ruim", "Ruim", "Ruim",
"Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Boa", "Boa", "Ruim",
"Boa", "Ruim", "Boa", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
"Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa",
"Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Boa",
"Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Ruim", "Boa",
"Ruim", "Boa", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
"Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa",
"Ruim", "Boa", "Boa", "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Ruim",
"Ruim", "Boa", "Boa", "Ruim", "Boa", "Boa", "Ruim", "Boa", "Boa",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa", "Ruim", "Boa",
"Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
"Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",

"Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa",
"Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim",
"Boa", "Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Boa", "Ruim",
"Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa",
"Boa", "Boa", "Boa", "Boa", "Boa", "Boa", "Ruim", "Ruim", "Ruim",
"Ruim", "Ruim", "Boa", "Ruim", "Boa", "Ruim", "Ruim", "Boa",
"Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa", "Ruim",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim",
"Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa",
"Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa", "Boa", "Ruim",
"Boa", "Boa", "Ruim", "Boa", "Boa", "Boa", "Boa", "Ruim", "Ruim",
"Ruim", "Boa", "Ruim", "Ruim", "Boa", "Boa", "Ruim", "Boa", "Ruim",
"Ruim", "Boa", "Ruim", "Boa", "Boa", "Boa", "Ruim", "Boa", "Ruim",
"Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Boa", "Ruim", "Boa",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim",
"Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Ruim",
"Ruim", "Ruim", "Boa", "Boa", "Boa", "Ruim", "Boa", "Ruim", "Boa",
"Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa",
"Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Boa", "Ruim", "Ruim",
"Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Boa",
"Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Ruim", "Boa", "Ruim", "Boa", "Boa", "Boa", "Boa", "Boa", "Boa",
"Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Ruim", "Boa", "Ruim",
"Ruim", "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Boa", "Ruim",
"Ruim", "Ruim", "Boa", "Boa", "Ruim", "Boa", "Ruim", "Ruim",
"Boa", "Boa", "Boa", "Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Ruim",
"Ruim", "Ruim", "Ruim", "Boa", "Ruim", "Boa", "Ruim", "Boa",
"Ruim", "Ruim", "Boa", "Ruim", "Boa", "Boa", "Boa", "Ruim"))

```
#Converte variavel resposta para factor
dados$Classe <- factor(dados$Classe, levels = c('Ruim','Boa'))

#Pequena analisa exploratoria
dados %>% group_by(Classe) %>% summarise_all("mean")

#Ajusta regressao logistica
fit <- glm(Classe ~ Prova_Logica + Redacao + Auto_Avaliacao ,
           data = dados,
           family = binomial)

#Visualiza resumo do modelo ajustado
summary(fit)

#Aplica exponenciacao nos coeficientes para interpretar
exp(fit$coefficients)
#Curva ROC
prob = predict(fit, newdata = dados, type = "response")
roc = roc(dados$Classe ~ prob, plot = TRUE, print.auc = TRUE)
#Obtem a predicao/probabilidade para cada observacao
Probabilidade <- predict(fit, newdata= dados,type = 'response')

#Se a probabilidade for maior que 50% classifica como 'Boa'
Classe_Predita <- ifelse(Probabilidade > 0.5,"Boa","Ruim")

#Visualiza data frame com as predicoes
View(data.frame(dados,Probabilidade,Classe_Predita))


#Gera matriz de confusao
```

```
confusao <- table(Classe_Predita = Classe_Predita, Classe_Original =
relevel(dados$Classe,ref = 'Boa'))

#Armazena valores da matriz de confusao
vp <- confusao[1,1];vp
fn <- confusao[2,1];fn

vn <- confusao[2,2];vn
fp <- confusao[1,2];fp

#Calcula acuracia
acuracia <- sum(diag(confusao))/ sum(confusao);acuracia

#Calcula  Sensitividade
sensitividade <- vp /(vp+fn)


#Cacula Especificidade
especificidade <- vn / (vn + fp)



#Analise de Sensitividade e Especificidade

limiares <- sort(Probabilidade)

acuracia <- c()
sensitividade <- c()
especificidade <- c()


for ( i in 1:length(limiares)) {
```

```
  limiar_atual <- limiares[i]

  Classe_Predita <- ifelse(Probabilidade > limiar_atual,'Boa' , 'Ruim')

  #Gera matriz de confusao
  confusao <- table(Classe_Predita = Classe_Predita, Classe_Original =
relevel(dados$Classe,ref = 'Boa'))

  #Armazena valores da matriz de confusao
  vp <- confusao[1,1];vp
  fn <- confusao[2,1];fn

  vn <- confusao[2,2];vn
  fp <- confusao[1,2];fp

  #Calcula acuracia
  acuracia[i] <- sum(diag(confusao))/ sum(confusao);acuracia

  #Calcula  Sensitividade
  sensitividade[i] <- vp /(vp+fn)

  #Calcula Especificidade
  especificidade[i] <- vn / (vn + fp)

}

plot(y = sensitividade[1:698] , x = limiares[1:698],  type="l", col="red", ylab = 'Sensitividade e
Especificidade', xlab= 'Pontos de Corte')
```

```
grid()
lines(y = especificidade[1:698], x = limiares[1:698], type = 'l',col="blue" )

legend("bottomleft", c("sensibilidade","especificidade"),
       col=c("red","blue"), lty=c(1,1),bty="n", cex=1, lwd=1)

abline(v=0.225)
```

```
#Obtem novamente as probabilidades para classificar baseado no ponto de corte 22,5%
Probabilidade <- predict(fit, newdata= dados,type = 'response')
Classe_Predita <- ifelse(Probabilidade > 0.225,"Boa","Ruim")

View(data.frame(dados,Probabilidade,Classe_Predita))

#Visualiza matriz de confusao final
confusao <- table(Classe_Predita = Classe_Predita, Classe_Original =
relevel(dados$Classe,ref = 'Boa'))

#Armazena valores da matriz de confusao
vp <- confusao[1,1];vp
fn <- confusao[2,1];fn

vn <- confusao[2,2];vn
fp <- confusao[1,2];fp
```

```
#Calcula acuracia
acuracia <- sum(diag(confusao))/ sum(confusao);acuracia


#Calcula  Sensitividade
sensitividade <- vp /(vp+fn)


#Cacula Especificidade
especificidade <- vn / (vn + fp)
```

Referências

BERRY; LINOFF.*Data Mining Techniques for Marketing.* New Jersey: Wiley, 1997.

BOLDRINI, C.; FIGUEIREDO, W. *Álgebra Linear.* São Paulo: HARBRA, 1986.

BRUCE, P.; BRUCE, A. *Estatística Prática para Cientistas de Dados.* Rio de Janeiro: Alta Books, 2019.

BUSSAB, W., & MORETTIN, P. *Estatística Básica.* São Paulo: Atual, 1987.

CHAMBERS, J. M. *Graphical Methods for Data Analysis.* New York: Chapman and Hall/CRC, 1983.

COHEN, J. *Statistical Power Analysis for the Behavioral Sciences.* New Jersey: Lawrence Erlbaum Associates, 1988.

COHEN, J. *Multiple Commitment in the workplace: an integrative approach.* New Jersey: Lawrence Erlbaum Associates, Inc., 2003.

FLEISS, J.; COHEN, J. The Equivalence of Wheighted Kappa and the Intraclass Correlation Coeficient As Measures of Reability. *Education and Psychological Measurement*, 1973.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning.* New York: Springer, 2009.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning.* New York: Springer, 2013.

MINGOTI, S. A. *Análise de Dados Através de Métodos de Estatística Multivariada.* Belo Horizonte: Editora UFMG.

PINHEIRO, J.; CUNHA, S.; CARVAJAL, S.; GOMES, G. *Estatística Básica: A Arte de Trabalhar com Dados.* Rio de Janeiro: Elsevier, 2009.

PORTER, D. C.; GUARAJATI, D. N. *Econometria Básica.* Porto Alegre: AMGH Editora Ltda, 2011.

PROVOST, F.; FAWCETT, T. *Data Science para Negócios.* Rio de Janeiro: Alta Books, 2016.

SMAILES, J.; McGRANE, A. *Estatística Aplicada a Administração com Excel.* São Paulo: Atlas S.A, 2012.

TURKEY, J. *Exploratory Data Analysis.* Boston: Addison-Wesley Pub. Co., 1977.