Manual de Análise de Dados

Prof. Steven Dutt-Ross

23/04/2020 updated:21 de julho de 2020

Contents

Chapter 1

Introdução

Este texto pretende ser um manual destinado ao auxílio às aulas de estatística para os(as) alunos(as) da UNIRIO com o objetivo de fornece os códigos para a linguagem R.

Esse guia foi montado em RMarkdown, usando o pacote **bookdown** (Xie, 2020). Esse material é 100% reprodutível.

AVISO

A ênfase deste texto será na prática com o uso da linguagem R. Você deve consultar os livros **Estatística Básica** (Bussab and Morettin, 2015) e **Métodos Estatísticos para as Ciências Sociais** (Agresti and Finlay, 2012) antes de usar as análises apresentadas aqui. Essas publicações são encontradas na biblioteca da UNIRIO.

Se você usar o código ou as informações deste guia em um trabalho publicado, solicito que cite-o como uma fonte nas referências bibliográficas.

DUTT-ROSS,S. Manual de Análise de Dados. Rio de Janeiro. 2020. mimeo. Disponível em: http://livro.metodosquantitativos.com/docs/bookdown.pdf

1.1 A estatística é importante

Esse texto foi planejado para introduzir os conceitos estatísticos com maior relevância para a vida cotidiana. Precisamos conjugar o rigor analítico com uma mentalidade prática.

> É fácil mentir com a estatística, mas é difícil dizer a verdade sem ela.

A estatística é uma maneira importante de se olhar o mundo. A vida é cheia de eventos ao acaso. Todos os dias a maioria de nós toma decisões que se baseia em uma avaliação de uma incerteza.

Além disso, a estatística é a base do método cientifico. Muitas vezes queremos ver como as explicações e as teorias se ajustam aos dados. Para tomar uma decisão (ou examinar argumento científico), precisamos de dispor de dados e métodos.

Assim, na última metade do século temos visto um aumento drástico do uso dos métodos estatísticos em todas as Ciências (sociais, humanas, medicina, biologia, e quase todas as outreas áreas). Existem várias razões para isso. 1. A pesquisa tem cada vez mais empregado uma abordagem quantitativa. 2. A ciência geralmente estuda as questões de interesse e análise evidências fornecidas pelos resultados empíricos. 3. o crescimento da internet resultou em um aumento da informação quantitativa disponível (Agresti and Finlay (2012)).

1.2 Por que analisar dados?

A análise estatística é sobre processo de descoberta. O desenvolvimento científico precisa gera um método para examinar objetos de interesse de uma maneira sistemática. Esse processo geralmente pecisa de evidência para apoiar um argumento. Um dos melhores métodos para estabelecer a evidência é por meio do exame dos números associados com o objeto de estudo. Esse exame é realizado através de análise estatística. A análise estatística é o ponto central da descoberta, e o domínio dela nos aproxima do método científico. A estatística apresenta muitas técnicas consagradas na academia para extrair informações significativas de dados brutos. por exemplo:

- Como fazer inferências sobre a natureza de uma população com base na observação de uma amostra estraída dela?
- Como prever taxas de ocorrências de eventos estocásticos (aleatórios)?
- E como entender e interpretar cálculos estatísticos efetuado por outras pessoas?

1.3 A grande pergunta

> Por que as graduações nas áreas de Ciências Sociais exigem um curso de estatística?

> Por que estudantes que desejam ingressar em uma profissão como Serviço Social ou Enfermagem tem que fazer a disciplina de estatística?

Estatística pode ser essencial na sua carreira. Os estudantes precisam estudar estatística por vários motivos:
1. Estudantes (e profissionais) devem estar aptos a ler e entender vários estudos estatísticos no seu campo de atuação. Tem que entender o vocabulário, símbolos, conceitos e procedimentos estatísticos utilizados nesses estudos.
2. Estudantes (e profissionais) podem ser chamados para conduzir uma pesquisa no seu campo. Dado que os procedimentos estatísticos são à base de todas as pesquisas quantitativas, para alcançar isso, eles devem estar aptos para desenhar experimento, coletar, organizar, analisar e resumir dados.
3. Eles também tem que estar apto a comunicar os resultados dos seus estudos em suas próprias palavras.
4. Estudantes (e profissionais) também podem usar o conhecimento adquirido dentro da estatística para ser melhores cidadãos e consumidores. Por exemplo, eles podem fazer entre decisões inteligentes sobre que produto comprar baseado no estudo do Consumidor, olhar os gastos do governo, etc…

> Entender Estatística é essencial aos cidadãos das sociedades atuais: para ser crítico em relação à informação disponível na sociedade, para entender e comunicar com base nessa informação, mas, também, para tomar decisões, uma vez que uma grande parte da organização dessas mesmas sociedades tem por base esses conhecimentos. (Shaughnessy, 1992, 1996).

Em outras palavras, quando você se formar, vai trabalhar em uma área que precisam usar métodos estatísticos ou pelo menos ler os relatórios que contém estatísticas. Em alguma fase da sua carreira, você vai ter que lidar com os métodos estatísticos.

Muitas pessoas encaram a estatística como assunto maçante/difícil. De fato algumas partes da estatística são difíceis, mas há outras que se apresentam de forma mais fácil e concisa. A reputação de dificuldade de estatística provém em grande parte da época em que não tinhamos acesso aos computadores.

Naquele tempo, os estatísticos eram forçados a efetuar os cálculos laboriosos manualmente. Hoje em computador faz essa parte do trabalho deixando os pesquisadores mais livres para estudar e entender o significado do que se passa. Nesse sentido, para fazer estatística, você deve aprender a programar em **R** ou em **Python**. Estas linguagens vão ajudá-lo a utilizar métodos para entender dados complexos para apoiar as suas conclusões.

> Em alguma fase do seu trabalho o pesquisador se vê as voltas com problemas de analisar e entender uma massa de dados relevantes

ao seu objeto de estudo. Se forem informações sobre uma amostra, ele necessitará resumir os dados para estes sejam informativos ou para compará-los com outros resultados, ou ainda para julgar sua adequação alguma teoria (Bussab & Morettin, 1987)

Chapter 2

Como os pesquisadores utilizam a Estatística?

Primeiro a gente usa a estatística para resumir dados. Mas não é só isso. Também usamos a estatística para mensurar a qualidade de uma inferência. Qualquer um pode extrapolar resultados. Todavia é bem difícil dizer o quanto essa inferência é boa. Por exemplo, se a gente tirasse uma amostra de 100 criminosos de Bangu 1 que foram para o programa de reabilitação e fizesse uma estimativa que 30% de todos os criminosos que participaram desse programa vão retornar a prisão, a gente gostaria de saber quanto essa estimativa se aproxima do verdadeiro percentual. O erro é de 5% e 10% ou 30%? De forma semelhante quando a gente tá pesquisando as características de uma população ou universo baseado em uma amostra, a gente devia saber qual a chance de conseguir uma conclusão incorreta.

AVISO
A estatística é como um bisturi: útil quando usada de forma correta e potencialmente desastrosa em mãos erradas.

O acesso fácil aos métodos estatísticos - usando softwares aonde você aperta um botão traz muitos riscos. Temos que ter muito cuidado com o excel, o SPSS e outras ferramentas automáticas.
Do político manipulador ao analista descuidado, do economista amador ao anunciante agressivo, temos infinitos exemplos de que pode dar errado quando o método estatístico é mal utilizado. Escolha seletiva supersimplificação, violação de pressupostos.
É comum a aplicação de métodos inapropriados e abordagens equivocadas. Um computador executar análises solicitadas sem ter condições se importar se os pressupostos requeridos para o uso adequado do método foram satisfeitos. Análises incorretas ocorrem quando os pesquisadores não tomam tempo suficiente para entender o método estatístico, as premissas para o seu uso ou se a abordagem é adequada.
Você irá precisar de um bom conhecimento de estatística para entender qual o método estatístico selecionar e como tirar conclusões válidas do resultado (Agresti and Finlay, 2012).

2.1 Antes de iniciar - Material do curso

Muitos livros de R foram disponibilizados pelos autores de forma gratuita na internet. Fiz repositório com todos os livros e apostilas de R em português que encontrei. Ele está disponível aqui https://github.com/DATAUNIRIO/R_Livros_e_Apostilas

A editora Springer disponibilizou mais de 500 livros gratuitamente. Muitos deles sobre a liguagem R. A lista completa pode ser acessada aqui

Os bancos de dados que vamos utilizar no curso estão disponíveis nesse repositório: https://github.com/DATAUNIRIO/Base_de_dados. Para fazer o download dos bancos de dados, copie o código abaixo e cole no script ou no console do R.

```
source('https://raw.githubusercontent.com/DATAUNIRIO/aulauniriov2/master/download_das_
```

2.1.0.1 O diretório de trabalho

No R, o diretório de trabalho é um conceito importante para entender.

É o local de onde o R estará procurando e salvando os arquivos. Quando você escreve código para o seu projeto, ele deve se referir aos arquivos em relação à raiz do seu diretório de trabalho e precisa colocar apenas de arquivos nesse local.

O RStudio facilita o acesso ao diretório de trabalho. Se você precisar verificar aonde está o diretório de trabalho, pode usar a função *getwd()*. Se, por algum motivo, o diretório de trabalho não for o que deveria ser, você poderá alterá-lo na interface do RStudio, navegando no menu chamado *Session* depois acesse o *Set working directory* e depois em *choose directory*. Como alternativa, você pode usar o comando *setwd("/caminho/para/seu/diretorio")* para redefinir seu diretório de trabalho por meio do uso do código.

2.1.0.2 Procurando ajuda no R

Use a interface de ajuda integrada do RStudio para procurar mais informações sobre as funções do R. Por exemplo, quero uma ajuda para a função média (*mean*)) no R. O Rstudio pode me ajudar como na figura abaixo.

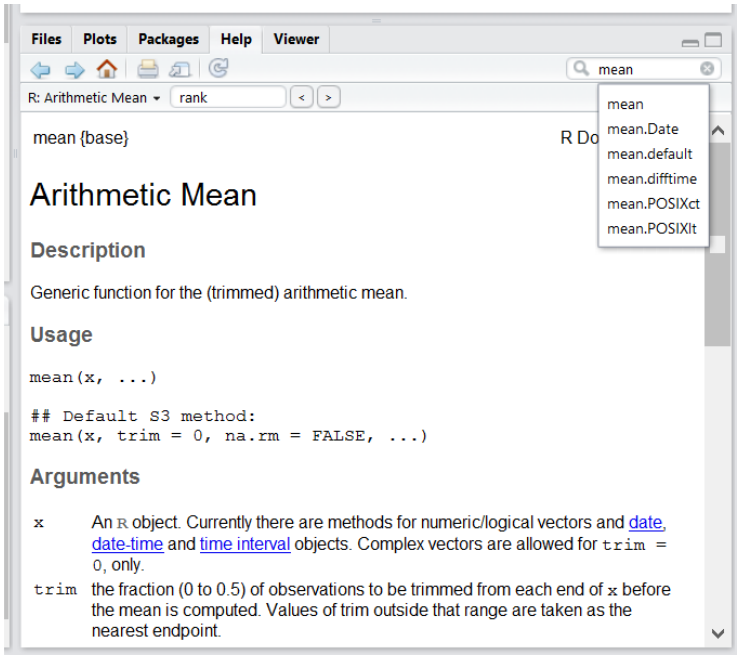


Figure 2.1: Help do R Studio

Como funciona a interface de ajuda do RStudio

Uma das maneiras mais rápidas de obter ajuda é usar a interface de ajuda do RStudio. Esse painel, por padrão, pode ser encontrado no painel inferior direito do RStudio. Como visto na captura de tela, digitando a palavra "Mean" (Média), o RStudio tenta também fornecer várias sugestões nas quais você pode estar interessado. A descrição é mostrada na janela direita inferior (Help).

"Eu sei o nome da função que quero usar, mas não sei como usá-la."

Se você precisar de ajuda com uma função específica, digamos *barplot()*, digite:

?barplot

Se você estiver procurando por uma função para executar uma tarefa específica, poderá usar a função nomeada *help.search()*, chamada pelo ponto de interrogação duplo **??**. O R vai procurar nos pacotes instalados páginas de ajuda que correspondem à sua solicitação de pesquisa. Por exemplo:

??kruskal

"Quero usar uma função que faça X, deve haver uma função para ela, mas não sei qual é o nome em inglês..."

Se você não encontrar o que está procurando na ajuda do R, use o google. Escreva algo do tipo *"barplot in R"* ou *kruskal in R* que provavelmente encontrará o que está procurando.

Uma pesquisa no Google o envia para a documentação de um pacote do R ou para um fórum em que uma outra pessoa já fez sua pergunta.

"Estou travado ... recebo uma mensagem de erro que não entendo

Comece com uma pesquisa da mensagem de erro no Google. No entanto, você pode ter uma mensagem de erro muito geral que pode não ser muito útil para diagnosticar um problema (por exemplo, "subscrito fora dos limites") . Se a mensagem for muito genérica, sugiro incluir o nome da função ou do pacote que está usando no R.

Se o erro persistir, uma outra estratégia muito utilizada é mudar a linguagem do R do português para o inglês, pesquisar a mensagem de erro em inglês e voltar para o português com as funções Sys.setenv(LANG = "en") e Sys.setenv(LANG = "pt")

Uma ótima referência (em inglês) de como começar no R pode ser encontrada aqui

Chapter 3

Tipo de dados

Os cinco tipos de dados mais comuns utilizados no R:
1. Lógico (logical)
2. Caractere (character)
3. Categórico (factor)
4. Numérico (numeric)
5. Inteiro (integer)

Os bancos de dados no R geralmente são uma combinação desses tipos diferentes. Exploramos, com mais detalhes, cada um dos tipos de dados.

3.0.0.0.0.1 Lógico Um dado lógico é um tipo de dado com apenas dois valores: Verdadeiro (TRUE) ou falso (FALSE).

```
#-------------------------------------------------------------

numero1 <- 117
numero2 <- 908

# o primero número é maior que o segundo número?
resultado <- numero1 > numero2
resultado
```

```
## [1] FALSE
```

```
class(resultado)
```

```
## [1] "logical"
```

```
# o primero número é menor ou igual ao segundo número?
resultado2 <- numero1 <= numero2
resultado2
```

```
## [1] TRUE
```

```
class(resultado2)
```

```
## [1] "logical"
```

```
x <- 0
as.logical(x)
```

```
## [1] FALSE
```

```
y <- 1
as.logical(y)
```

```
## [1] TRUE
```

3.0.0.0.0.2 Caractere O tipo de dado no formato "caractere" é utilizado para armazenar texto. A maneira mais simples de armazenar texto no R é usando aspas simples ou duplas.

```
#---------------------------------------------------------------

dados_texto <- "Um Lannister sempre paga suas dívidas"
dados_texto
```

```
## [1] "Um Lannister sempre paga suas dívidas"
```

```
class(dados_texto)
```

```
## [1] "character"
```

```
#-------------------------------------------------------------

vetor_de_texto <- c("Um Lannister sempre paga suas dívidas",
                    "O inverno está chegando")
vetor_de_texto
```

```
## [1] "Um Lannister sempre paga suas dívidas"
## [2] "O inverno está chegando"
```

```
class(dados_texto)
```

```
## [1] "character"
```

Se você deseja forçar qualquer tipo de dado a ser armazenado como caractere, você pode fazê-lo usando o comando *as.character*

```
#-------------------------------------------------------------
dados_numericos <- c(48, 59, 1, 53,3)
class(dados_numericos)
```

```
## [1] "numeric"
```

```
dados_caractere <-as.character(dados_numericos)
class(dados_caractere)
```

```
## [1] "character"
```

```
dados_caractere
```

```
## [1] "48" "59" "1"  "53" "3"
```

```
# Note que tudo dentro das aspas será considerado como caractere,
#não importa se ele é um número ou um texto.
```

3.0.0.0.0.3 Categórico (*factor*) Variáveis categóricas são um caso especial de variáveis de caracteres, no sentido em que também contém texto. No entanto, variáveis categóricas são usadas quando há um número limitado de cadeias exclusivas de caracteres para representar uma categoria.

Por exemplo, o gênero geralmente assume apenas dois valores, "feminino" ou "masculino" (e será considerado uma variável categórica). Para criar uma variável categóricas, use a função *as.factor*.

```
#-------------------------------------------------------------
#Ned Stark, Robert Baratheon, Arya Stark, Cersei Lannister,
#Daenerys Targaryen, Jaime Lannister


casas_got<-c("Stark", "Baratheon", "Stark", "Lannister",
             "Tyrell","Targaryen","Lannister")
class(casas_got)
```

```
## [1] "character"
```

```
casas_got_cat <-as.factor(casas_got)
class(casas_got_cat)
```

```
## [1] "factor"
```

```
# Para encontrar as categorias, use "levels"
sexo_got<-c("masculino","masculino","feminino",
            "feminino","feminino","masculino")
sexo_got<-as.factor(sexo_got)
levels(sexo_got)
```

```
## [1] "feminino"  "masculino"
```

```
levels(casas_got_cat)
```

```
## [1] "Baratheon" "Lannister" "Stark"     "Targaryen" "Tyrell"
```

3.0.0.0.0.4 Numérico No R, o tipo de dado mais comum é o numérico. Uma variável será armazenada como dado numérico, se os valores forem números ou se contiverem casas decimais. Por exemplo, as duas séries a seguir são armazenadas como numéricas por padrão:

```
#-------------------------------------------------------------
### vetor numérico sem valores decimais
dados_numericos <- c(48, 59, 1, 53,3)
dados_numericos
```

```
## [1] 48 59  1 53  3
```

```
class(dados_numericos)
```

```
## [1] "numeric"
```

```
### vetor numérico com valores decimais
dados_numericos_decimal <- c (48.4, 59.1, 1.2, 53.9,3.3)
dados_numericos_decimal
```

```
## [1] 48.4 59.1  1.2 53.9  3.3
```

```
class(dados_numericos_decimal)
```

```
## [1] "numeric"
```

```
### Para transformar outro tipode de dados em numerico
dados_caracteres <- c("48", "59", "1", "53","3")  # caracteres vem com aspas
class(dados_caracteres) # comando para identificar o tipo de variavel
```

```
## [1] "character"
```

```
dados_numericos_2 <- as.numeric(dados_caracteres)
class(dados_numericos_2)
```

```
## [1] "numeric"
```

3.0.0.0.0.5 Inteiro O tipo de dado inteiro é um caso especial de dados numéricos. Inteiros são os dados numéricos sem as casas decimais. Pode ser usado **se você tiver certeza de que os números armazenados nunca contêm decimais**.

Imagine que você esteja interessado em saber o número de crianças em uma família. Essa variável é discreta e nunca terá casas decimais. Não podemos ter 1,7 filhos (um virgula sete filhos). Portanto, ele pode ser armazenado como inteiro.

```
#---------------------------------------------------------------
criancas<- c("1", "4", "0", "3", "2")
class(criancas)
```

```
## [1] "character"
```

```
criancas_2 <- as.integer(criancas)
class(criancas_2)
```

```
## [1] "integer"
```

```
# o que acontece com o decimal se eu mudar para decimal?
dados_numericos_decimal <- c (48.4, 59.1, 1.2, 53.9,3.3)
dados_inteiros<-as.integer(dados_numericos_decimal)
# vamos perder a informação do decimal
dados_inteiros
```

```
## [1] 48 59  1 53  3
```

No R, existem outros tipos de dados que você pode usar. Vou falar de três áreas diferentes em que você pode trabalhar com dados diferentes no R, mas que é mais avançado que o escopo desse guia.

1. dados no formato JSON (computação)
2. dados no formato SHAPEFILE (geoprocessamento)
3. dados de Números Complexos (matématica)
4. dados textuais no formato de CORPUS (linguagem natural/análise de discursos)

Perguta de revisão

Qual o tipo de dado do vetor abaixo?

dados <- c(48, 59, “targaryen”, 1, 53,3)
class(dados)

Chapter 4

Banco de dados

O banco de dados (data.table) é a organização e armazenagem de informações sobre um deteminado escopo. De forma mais simples, é o agrupamento de dados que tratam do mesmo assunto.

Além disso, os bancos de dados (data.table) exibe as informações em forma de tabela, isto é, com linhas e colunas. Cada linha representa um registro/caso/observação e cada coluna representa um atributo. No nosso caso, chamamos esses atributos de variáveis.

4.1 Registros e variáveis

Os registros são objetos descritos por um conjunto de dados, podendo ser pessoas, animais, municípios, estados ou objetos.

Uma variável é qualquer característica do registro. Uma variável pode assumir valores diferentes para registros diferentes.

4.2 O que é um banco de dados

Iris Data: 50 flores de 03 espécies.

PhantomJS not found. You can install it with webshot::install_phantomjs(). If it is installed, please make sure the phantomjs executable can be found via the PATH variable.

Cada linha é um registro e cada coluna é um atributo (que chamamos de variável).

4.2.0.1 Exemplo de banco de dados: Províncias da Suiça (Swiss Data)

Cada linha é uma província e cada coluna é uma informação.

4.2.0.2 Exemplo de banco de dados: Midwest Data

Informações Demográficas dos municípios do Meio-Oeste. Cada linha é equivalente a um município e cada coluna é uma informação.

4.2.0.3 Exemplo de banco de dados: Mtcars Data

Banco de dados de Performance de carros (10 informações). Cada linha é um carro diferente e cada coluna é uma informação.

4.2.0.4 Geralmente acompanhado de um DICIONÁRIO DE DADOS

Por exemplo:
* mpg: Miles/(US) gallon
* cyl: Number of cylinders
* disp: Displacement (cu.in.)
* hp: Gross horsepower
* drat:Rear axle ratio
* wt: Weight (1000 lbs)
* qsec: 1/4 mile time
* vs: V/S
* am: Transmission (0 = automatic, 1 = manual)
* gear: Number of forward gears
* carb: Number of carburetors

Por exemplo, esse banco de dados que estamos trabalhando tem 32 carros e 11 variáveis. No R isso pode ser verificado pelo comando dim(mtcars), e names(mtcars).

```
dim(mtcars)
```

```
## [1] 32 11
```

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

Chapter 5

Testes de Hipóteses

Tradicionalmente, em um curso de Estatística há uma grande ênfase no teste de hipóteses e na tomada de decisões com base em p-valor. O teste de hipóteses é importante para determinar se há efeitos estatisticamente significativos.

5.0.0.1 Inferência estatística

A maior parte do que abordei até agora é sobre a produção de estatísticas descritivas: cálculo de médias, medianas, variância e desvio-padrão. Agora vamos abordar a inferência estatística: usando testes estatísticos para tirar alguma conclusão sobre os dados.

É natural para a maioria de nós utilizar astatísticas resumidas ou os gráficos, mas pular para a inferência estatística exige uma pequena mudança de perspectiva. A idéia de usar algum teste estatístico para responder a uma pergunta não é um conceito difícil, mas as próximas discussões ficarão um pouco teóricas.

Esse vídeo (em inglês) do Centro de Aprendizagem de Estatística explica bem a inferência estatística https://www.youtube.com/watch?v=tFRXsngz4UQ&list=PLp485I5glNfhYX6wTDB3wllRTtSJas_md

O teste de hipóteses é um procedimento estatístico que nos permite rejeitar ou não rejeitar uma hipótese através da evidência fornecida pela amostra. Em outras palavras, trata-se de uma metodologia estatística que nos auxilia a tomar decisões sobre uma ou mais populações, baseada na informação obtida da amostra.

5.0.0.2 Hipóteses Estatísticas

Uma suposição ou afirmação sobre parâmetros populacionais em forma numérica é chamada de hipótese estatística.

O Prof. Andre Zibetti afirma aqui que o tomar decisões, é comum a formulação de suposições ou de conjeturas sobre as populações de interesse, que, em geral, consistem em considerações sobre parâmetros (

$$\mu, \sigma 2$$

) das mesmas. Essas suposições, que podem ser ou não verdadeiras, são denominadas de Hipóteses. Em muitas situações práticas, o interesse do pesquisador é verificar a veracidade sobre um ou mais parâmetros populacionais (

$$\mu, \sigma 2$$

). Acredito que essa seja uma ótima definição.

5.0.0.3 Testando hipóteses

5.0.0.3.0.1 As hipóteses nulas e alternativas Os métodos estatísticos contam com o teste de uma hipótese nula, que possui uma formulação específica para cada teste. A hipótese nula sempre descreve o caso em que, por exemplo, dois grupos não são diferentes ou não há correlação entre duas variáveis, etc. A hipótese alternativa é contrária à hipótese nula e, portanto, descreve os casos em que há uma diferença entre grupos ou há uma correlação entre duas variáveis, etc.

Observe que as definições de hipótese nula e hipótese alternativa não têm nada a ver com o que você deseja encontrar ou não, ou o que é interessante ou não. Se você estivesse comparando a altura de homens e mulheres, a hipótese nula seria que a altura dos homens e a altura das mulheres não fossem diferentes.

Da mesma forma, se você estivesse estudando a renda de homens e mulheres, a hipótese nula seria que a renda de homens e mulheres não é diferente na população que você está estudando. Nesse caso, você pode estar esperando que a hipótese nula seja verdadeira, embora não fique surpreso se a hipótese alternativa for verdadeira. De qualquer forma, a hipótese nula assumirá a forma de que não há diferença entre grupos, não há correlação entre duas variáveis ou não há efeito dessa variável em nosso modelo.

5.0.0.4 Definição de p-valor

A maioria dos testes de hipóteses se baseia no uso do p-valor para avaliar se devemos rejeitar ou deixar de rejeitar a hipótese nula. Dado que a hipótese nula é verdadeira, o p-valor é definido como a probabilidade de obter um resultado igual ou mais extremo do que o que foi realmente observado nos dados.

5.0.0.5 Regra de decisão

O p-valor para os dados fornecidos será determinado pela realização do teste estatístico.

Este p-valor é então comparado com um valor de alpha pré-determinado. Geralmente, um valor alpha de 0,05 é usado, mas não há nada de mágico nesse valor. Dependendo do seu estudo, você pode usar 0,01 ou 0,1 como valor de alpha.

Se o p-valor para o teste for menor que alpha, rejeitamos a hipótese nula. Se o p-valor for maior ou igual a alpha, falhamos em rejeitar a hipótese nula.

5.0.0.6 Exemplo de lançamento de moeda

Para um exemplo do uso do p-valor para o teste de hipóteses, imagine que você tenha uma moeda que jogará 100 vezes.

A hipótese nula é que a moeda é justa - isto é, que é igualmente provável que a moeda caia sobre as faces cara/coroas. Temos 50% para cada face da moeda.

A hipótese alternativa é que a moeda não é justa. Digamos que, para esse experimento, você jogue a moeda 100 vezes e ela dê cara 94 vezes. O p-valor nesse caso seria a probabilidade de obter 94, 95, 96, 97, 98, 99 ou 100 caras, assumindo que a hipótese nula seja verdadeira (moeda honesta).

Usando a notação matemática:

$$H0 : p = 0,5 \ (hiptese \ nula)$$
$$H1 : p \neq 0,5 \ (hiptese \ alternativa)$$

Você pode imaginar que o p-valor para esses dados será bem pequeno. Se a hipótese nula for verdadeira e a moeda for justa, haveria uma baixa probabilidade de obter 94 ou mais caras (ou 94 ou mais coroas).

Usando um teste binomial, o valor-p é < 0,0000001. (Na verdade, R o informa como <2.2e-16, que é uma abreviação para o número na notação científica, 2,2 x 10-16, que é 0,00000000000000022, com 15 zeros após o ponto decimal)

Considerando um alpha de 0,05, uma vez que o p-valor é menor que alpha, rejeitamos a hipótese nula. Ou seja, concluímos que a moeda não é honesta.

```
binom.test(94, 100, 0.5)
```

```
##
##  Exact binomial test
##
```

```
## data:  94 and 100
## number of successes = 94, number of trials = 100, p-value <
## 0.00000000000000022
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.8739701 0.9776651
## sample estimates:
## probability of success
##                   0.94
```

Como o p-valor é menor que 0,05, vamos concluir que é improvável que uma moeda honesta gere 94 caras. Em outras palavras, vamos concluir que p != 0,5 (a moeda é desonesta).

5.0.0.7 Lançamento de outra moeda

Agora imagine que você queira testar se uma outra moeda é honesta. Assim, vamos jogá-la 1.000 vezes. O resultado foi 540 caras. Com essas informações, decida se a moeda é honesta ou não.

- A pergunta é: uma moeda honesta pode gerar 540 caras em 1.000 lançamentos? Vamos ver o gráfico.

```
library(ggplot2)
library(grid)

x1  <- 450:550
df <- data.frame(x = x1, y = dbinom(x1, 1000, 0.5))

ggplot(df, aes(x = x, y = y)) +
  geom_bar(stat = "identity", col = "#008080", fill = "#008080") +
  scale_y_continuous(expand = c(0.01, 0)) +
  xlab("Número de Caras") + ylab("Densidade") +
  labs(title = "Número de Caras em 1.000 lançamentos de uma moeda") +
  theme_bw(16, "serif") +
  theme(plot.title = element_text(size = rel(1.2), vjust = 1.5))
```

Em uma análise gráfica, parece improvável que uma moeda honesta gere 540 caras. Para ter certeza, vamos realizar o teste.

Primeiro passo: formular as hipóteses.

$$H0 : p = 0,5 \ (hiptese \ nula)$$

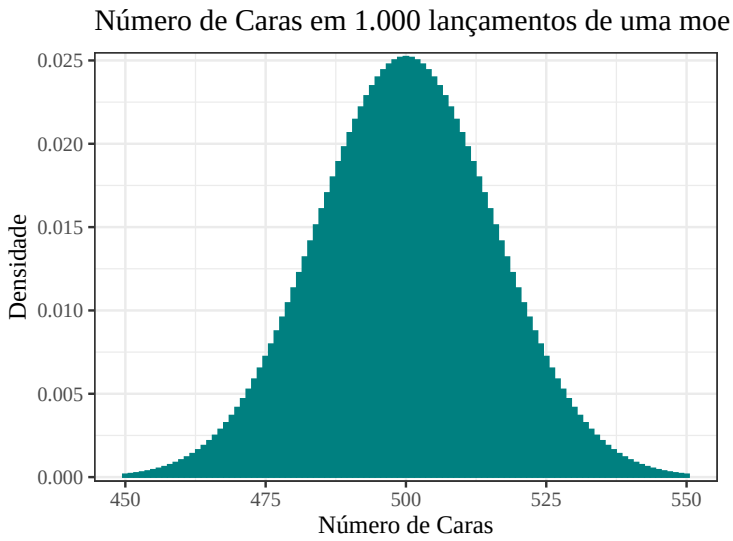


Figure 5.1: lançamento

$$H1 : p \neq 0,5 \ (hiptese \ alternativa)$$

Segundo passo: Criar uma regra de decisão.

Se p-valor $<$ 0,05 rejeito H0

Terceiro passo: Realizar o teste

```
binom.test(540, 1000, 0.5)
```

```
##
##  Exact binomial test
##
## data:  540 and 1000
## number of successes = 540, number of trials = 1000, p-value = 0.01244
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5085302 0.5712337
## sample estimates:
## probability of success
##                   0.54
```

Quarto passo: Concluir a partir do resultado

- p-valor = 0,01244
- alpha = 0,05

Como p-valor < 0,05,** Rejeito H0 **. Ou seja, "Essa moeda também não é honesta"

5.0.0.8 Espere, isso faz algum sentido?

Lembre-se de que a definição do p-valor é: Dado que a hipótese nula é verdadeira, o p-valor é definido como a probabilidade de obter um resultado igual ou mais extremo do que o que foi realmente observado nos dados.

Na prática, usamos os resultados dos testes estatísticos para chegar a conclusões sobre a hipótese nula. Tecnicamente, o p-valor não diz nada sobre a hipótese alternativa. Mas logicamente, se a hipótese nula é rejeitada, seu complemento lógico, a hipótese alternativa, é apoiado.

5.0.0.9 Estatística é como um júri

Observe o discurso utilizado ao testar a hipótese nula. Com base nos resultados de nossos testes estatísticos, rejeitamos a hipótese nula ou deixamos de rejeitar a hipótese nula.

Isso é um semelhante à abordagem de um júri em um julgamento. O júri encontra evidências suficientes para declarar alguém culpado ou falha em encontrar evidências suficientes para declarar alguém culpado. Ou ele é culpado ou não-culpado (*guilty or not guilty*).

Não-culpado é diferente de inocente. A sentença de absolvição do juri não é um certificado de inocência. Varias coisas podem acontecer: 1) o promotor não fez o trabalho direito, 2) a pessoa é inocente, 3) o policial contaminou a cena do crime, etc… Ou temos evidências sobre a culpa e ele é condenado ou alguma coisa aconteceu.

Não condenar alguém não é necessariamente o mesmo que declarar alguém inocente. Da mesma forma, se não conseguirmos rejeitar a hipótese nula, não devemos assumir que a hipótese nula é verdadeira. Pode ser que não tenhamos amostras suficientes para obter um resultado que nos permita rejeitar a hipótese nula, ou talvez haja alguns outros fatores que afetam os resultados pelos quais não consideramos. Isso é semelhante a uma postura de "inocente até que se prove o contrário".

IMPORTANTE

Por esse motivo, **NUNCA use o termo aceitar a hipótese nula**. Ou rejeitamos H0 ou não rejeitamos H0. Não rejeitar H0 é diferente de aceitar H0.

5.0.0.10 Erros do tipo 1 e tipo 2

Na maioria das vezes, os testes estatísticos que são utilizados são baseados em probabilidade, e os nossos dados sempre podem ser o resultado do acaso. Considerando o exemplo de lançamento de moeda acima, se jogássemos uma moeda 100 vezes e tivéssemos 94 caras, seríamos obrigados a concluir que a moeda não era justa. Mas 94 caras poderiam acontecer com uma moeda justa estritamente por acaso.

Podemos cometer dois tipos de erros ao testar a hipótese nula:

- Um erro do tipo I ocorre quando a hipótese nula é verdadeira, mas com base em nossa regra de decisão, rejeitamos a hipótese nula. Nesse caso, nosso resultado é um falso positivo; achamos que há um efeito (moeda injusta, associação entre variáveis, diferença entre grupos) quando realmente não existe. A probabilidade de cometer esse erro de tipo é alpha, o mesmo alpha que usamos em nossa regra de decisão.
- Um erro do tipo II ocorre quando a hipótese nula é falsa, mas com base em nossa regra de decisão, falhamos em rejeitar a hipótese nula. Nesse caso, nosso resultado é um falso negativo; falhamos em encontrar um efeito que realmente existe. A probabilidade de cometer esse tipo de erro é chamada beta.

A tabela a seguir resume esses erros. *inserir a tabela

5.0.0.10.1 Uma metáfora: Teste de hipóteses é como um julgamento

Imagine que você tenha dois professores (um carrasco e um mamata). Agora imagine que carrasco é tão rigoroso que acaba reprovando o melhor aluno da turma. Já o prof. mamata acabou de aprovar um dos piores da classe. O que é pior? Reprovar um estudante que sabe a materia ou aprovar um aluno que não sabe nada?

Procedimento parecido acontece nos testes de hipóteses. Se você aumentar muito o rigor do teste de hipóteses (alpha), somente evidências extraordinárias rejeitaria a hipótese nula. Se você aumentar a tolerância ao erro associado ao alpha e diminuir o erro relacionado ao beta, rejeitaria facilmente a hipótese nula mesmo quando há pouca evidência (na terminologia estatística, chamamos isso de tamanho do efeito).

Gostaria do seu *feedback*. Essa metáfora faz sentido?

5.0.0.11 Boas práticas para análises estatísticas

Ao analisar os dados, o analista não deve abordar a tarefa como faria um advogado da acusação. Ou seja, **o analista não deve procurar efeitos e testes significativos**, mas deve ser como um investigador independente, usando evidência para descobrir o que é mais provável de ser verdade.

5.0.0.12 O que é p-hacking?

Este seguimento é uma tradução de (Mangiafico, 2015)

É importante, ao abordar testes de hipóteses do seu banco de dados, evitar cometer p-hacking do p-valor.

Imagine o caso em que o pesquisador coleta muitas medidas diferentes em uma variedade de assuntos. O pesquisador pode ficar tentado a simplesmente fazer muitos testes e modelos diferentes para relacionar uma variável a outra, mas para todas as variáveis. Ele pode continuar fazendo isso até encontrar um teste com um p-valor significativo. Isso seria uma forma de fazer o p-hacking.

Como um valor alfa de 0,05 nos permite cometer um erro falso positivo cinco por cento do tempo, encontrar um p-valor abaixo de 0,05 após vários testes sucessivos pode ser simplesmente devido ao acaso. Algumas formas de hackers com p-valor são mais flagrantes. Por exemplo, se alguém coletar alguns dados, execute um teste e continue a coletar dados e execute os testes iterativamente até encontrar um p-valor significativo.

5.0.0.12.0.1 Consequência do p-hacking: Viés de publicação Este seguimento é uma tradução de (Mangiafico, 2015)

Uma questão relacionada à ciência é que existe um viés para publicar ou relatar apenas os resultados significativos. Isso também pode levar a uma inflação da taxa de falsos positivos.

Como um exemplo hipotético, imagine se atualmente há 20 estudos semelhantes sendo testados com um efeito semelhante - digamos o efeito dos suplementos de glucosamina na dor nas articulações.

Se 19 desses estudos não encontraram efeito e foram descartados, mas um estudo encontrou um efeito usando um alfa de 0,05 e foi publicado, isso é realmente algum suporte para que os suplementos de glucosamina diminuam a dor nas articulações?

5.0.0.13 Discussão opcional: métodos alternativos ao teste de hipóteses

A controvérsia do teste de hipóteses. Particularmente nos campos da psicologia e da educação, tem havido muitas críticas à abordagem do teste de hipóteses. Na minha leitura, as principais queixas contra o teste de hipóteses tendem a ser:

- Geralmente, usamos um alpha de 0,05 como um ponto de corte mágico. Não faz muito sentido chegar a uma conclusão se nosso p-valor 0,049 e a conclusão oposta se nosso p-valor for 0,051.
- O p-valor está desatualizado e não faz sentido com big data: tudo se torna estatisticamente significativo.
- Alunos e pesquisadores realmente não entendem o significado do p-valor.
- P-hacking

Qual abordagem podemos utilizar conjugado com o teste de hipóteses para mitigar esses problemas? Deixo essa pergunta para você. Pessoalmente temos dois métodos promissores. 1 - Método Bayesiano, 2 - Machine Learning

Essas críticas foram retiradas dos textos a seguir:

A Dirty Dozen: Twelve P-Value Misconceptions

Testing for Publication Bias in Political Science

The p-Value You Can't Buy

Misconceptions, Misuses, and Misinterpretations of P Values and Significance Testing

Show Me the Magnitude! The Consequences of Overemphasis on Null Hypothesis Significance Testing

The Enduring Evolution of the P Value Em breve vou desenvolver o restante do conteúdo sobre testes de hipóteses. Nesse momento sugiro ver os livros do (Bussab and Morettin, 2015) e (Agresti and Finlay, 2012).

Chapter 6

Qual Teste de Hipóteses deve ser utilizado?

Os testes de hipóteses podem ser divididos em dois grupos, conforme a fundamentação ou não dos seus cálculos nos pressupostos de que a **distribuição dos erros amostrais é normal, as variâncias são homogêneas e os erros independentes**.

O cumprimento ou não desses pré-requisito condiciona a escolha do teste pelo pesquisador, uma vez que, se forem preenchidos, ele poderá utilizar procedimetos estatísticos paramétricos, cujos testes são mais poderosos do que os da abordagem não-paramétrica. Caso não sejam atendidos, dois procedimentos são muito comuns:

1. Fazer uma transformação nos dados;
2. Utilizar um método não-paramétrico.

Todos os testes estatísticos têm pressuposições de base que precisam ser atendidas para que o teste de hipóteses forneça resultados válidos (sem erros inaceitáveis) em relação ao parâmetro que o teste está avaliando.

Sempre devemos verificar os pressupostos para a realização dos testes. Se eu ainda não te convenci a avaliar os pressupostos, deixa eu usar alguns exemplos:
Imagine que você está fazendo uma apresentação para o seu chefe e um grupo de pessoas. Você está tentando convence-los a investir mais no marketing (sua área). Para isso, você fez um questionário e está divulgando os resultados da sua pesquisa. Alguém levanta a mão e questiona os seus procedimentos. Ele fala que os seu resultados não são confiáveis porque você utilizou os procedimentos estatísticos errados. Algo do tipo "você não verificou a hipótese de normalidade!!!"
Imagine agora que você apresentou um preço menor que um intituto de pesquisa (exemplos: IBOPE/DATAFOLHA/GALLUP) e ganhou uma licitação do governo federal para realizar uma pesquisa quantitativa. Você desenvolveu o questionário, fez a coleta de dados, e fez a tabulação dos resultados. Chegou o dia de apresentar os resultados para o contratante. Nesse dia, o técnico do Instituto solicita o microfone e questiona os pressupostos da sua pesquisa.
Imagine agora que você está escrevendo um artigo, uma dissertação ou uma monografia. Você realizou um teste de hipóteses que você viu em um periódico científico, sem antes verificar os requisitos do teste. O parecerista ou o professor da banca afirma que os seus resultados não são válidos (podem ser válidos ou não.... não sabemos).
Situações que queremos evitar verificando pressupostos.
Todos esses exemplos são reais. Posso usar muitos outros exemplos, mas vou usar um que aconteceu comigo no ano passado. Submeti junto com amigos da ciência política um artigo para uma revista científica e recebi essa resposta do parecerista:

> since the authors use a panel, I recommend reporting autocorrelation and heteroskedasticity using post-estimation tests. These tests could help deciding which is the best model to run.

Basicamente, esse texto diz que eu tenho que colocar no artigo a verificação dos pressupostos de autocorrelação/independência e de heterocedasticidade/homocedasticidade de variâncias. Como eu tinha já preparado o artigo avaliando essas questões, foi fácil colocar essas informações.
Esse é um ótimo exemplo para enfatizar que não será possível divulgar resultados estatísticos sem mostrar que você verificou os pressupostos do modelo. Inclusive nas ciências humanas e sociais.
:::: {.blackbox data-latex="“}
::: {.center data-latex=""}
O que quero deixar claro é que a ausencia de avaliação das premissas do teste de hipóteses abre espaço para dúvidas e questionamentos dos resultados da pesquisa.

Nas palavras de Patino e Ferreira (2018), é uma boa prática, como investigadores, saber se as premissas dos testes estatísticos usados para responder suas perguntas foram avaliadas e se foram ou não atendidas. Se as premissas dos testes tiverem sido atendidas, devemos relatar na seção de resultados do estudo. Isso garante à comunidade científica que os resultados do estudo atenderam critérios importantes relacionados a sua validade. **?**

A seguir vou apresentar um mapa para responder a pergunta título do capítulo: **Qual Teste de Hipóteses eu devo usar?**

A resposta depende da natureza das suas variáveis e dos pressupostos de cada teste.

6.1 Qualitativa vs Qualitativa

6.1.0.1 Variável Qualitativa versus Variável Qualitativa

Para escolher o teste, você precisa verificar se o valor esperado de cada célula é igual ou maior do que cinco.

Quando temos duas variáveis qualitativas, os principais procedimentos são:

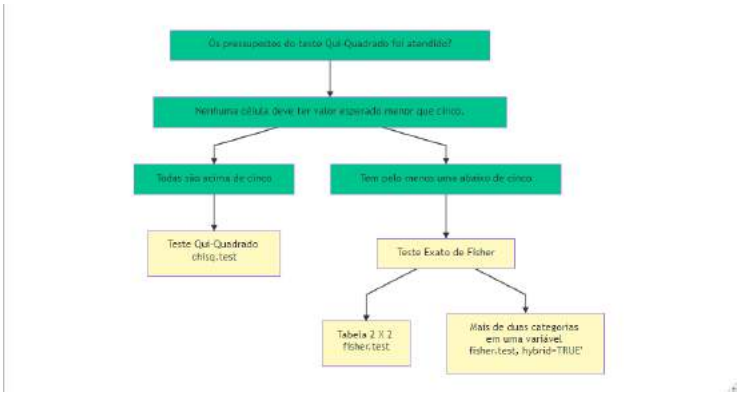


Figure 6.1: Escolha do teste para duas variáveis qualitativas

Uma versão aumentada dessa figura pode ser encontrada aqui

6.2 Quantitativa vs Qualitativa

6.2.0.1 Variável Quantitativa versus Variável Qualitativa

Nesse caso, você precisa verificar dois pressupostos:

1. Pressuposto de normalidade.
2. Pressuposto de homogeneidade de variâncias (igualdade de variâncias).

Quando temos uma variável quantitativa e uma uma variável qualitativa, os principais procedimentos são:

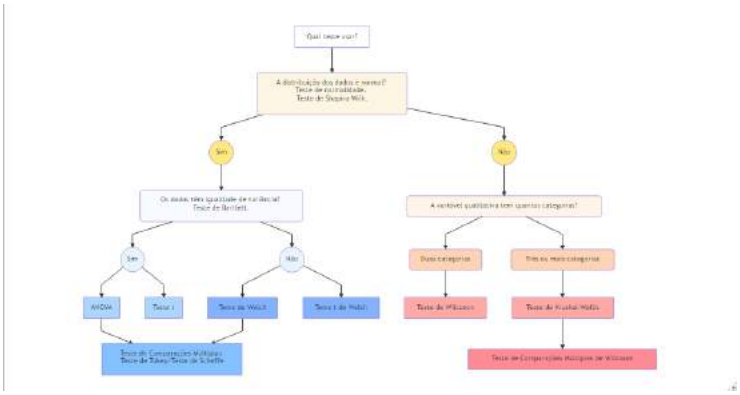


Figure 6.2: Escolha do teste para duas variáveis quantitativas

Uma versão aumentada dessa figura pode ser encontrada aqui

6.3 Quantitativa vs Quantitativa

Para escolher o teste, você precisa verificar o pressuposto de normalidade.

6.3.0.1 Variável Quantitativa versus Variável Quantitativa

Uma versão aumentada dessa figura pode ser encontrada aqui

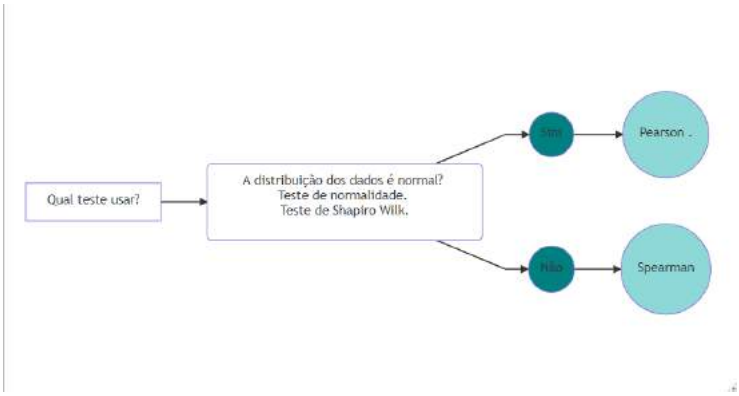


Figure 6.3: Escolha do teste para duas variáveis quantitativas

Chapter 7

Teste de Normalidade de Shapiro Wilk

7.0.0.1 Por que usar?

O Shapiro Wilk é um teste para aferir a normalidade dos dados (se a variável segue uma distribuição normal). A avaliação do pressuposto de normalidade é exigida pela maioria dos procedimentos estatísticos. A análise estatística paramétrica é um dos melhores exemplos para mostrar a importância de avaliar a suposição de normalidade.

A estatística paramétrica assume uma certa distribuição dos dados, geralmente a distribuição normal. Se a suposição de normalidade for violada, a interpretação, a inferência, e a conclusão a partir dos dados podem não ser confiáveis ou válidas.

Existem outros testes de normalidade (Jarque-Bera, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling). Recomendo o uso do Shapiro Wilk porque ele é geralmente poderoso, fácil de usar, e muitas pessoas estão familiarizadas com ele (diminuindo a necessidade de explicar em detalhes a abordagem se você o usar em um artigo). É realizado com a função *shapiro.test*.

7.0.0.2 Pressupostos

O teste assume que as observações são independentes.

7.0.0.3 Dados apropriados

- A variável dependente é quantitativa.

7.0.0.4 Hipóteses

Hipótese nula: os dados são normalmente distribuídos (os dados seguem uma distribuição normal).
Hipótese alternativa: os dados não são normalmente distribuídos (os dados não seguem uma distribuição normal).

7.0.0.5 Interpretação

Resultados significativos podem ser relatados como "os dados não seguem uma distribuição normal" ou como "a hipótese nula de que os dados são normalmente distribuídos foi rejeitada".

7.0.0.6 Exemplo de teste de Shapiro Wilk

Este exemplo apresenta os dados de uma amostra de 32 carros.
Esse teste responde à pergunta: "A variável preço do carro segue uma distribuição normal?"

O teste de Shapiro Wilk é realizado com a função *shapiro.test*, que produz um p-valor para a hipótese. Primeiro, os dados são resumidos. Em seguida, são examinados usando gráficos histograma, QQ-plot. Esses gráficos ajudam a decidir se a distribuição é normal.

```
#---------------------------------------------------------------
#### Crie o banco de dados
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")

### Verifique os dados
str(CARROS$Preco)
```

```
##  num [1:32] 160 160 108 258 360 ...
```

```
### Remova objetos desnecessários
remove(mtcars)

#### Resumo dos dados
summary(CARROS$Preco)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    71.1   120.8   196.3   230.7   326.0   472.0
```

```
#### Histograma
hist(CARROS$Preco, prob=TRUE, col="red",main = "Histograma do preço do carro")
lines(density(CARROS$Preco),lwd=3,col="blue")
```

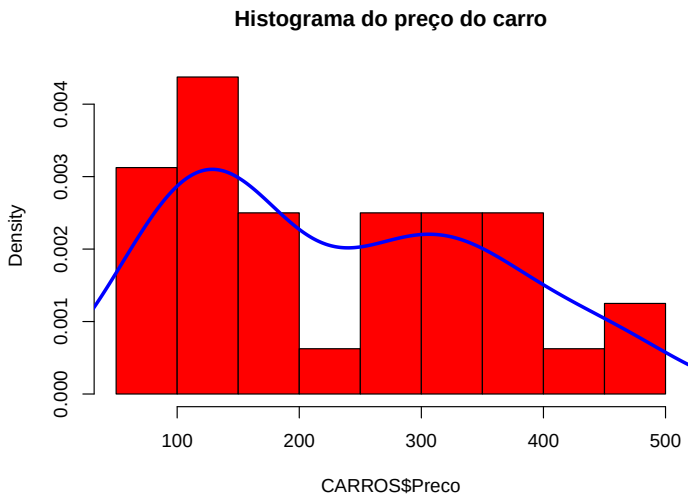


Figure 7.1: Teste de Shapiro Wilk

```
#### Grafico QQ-Plot
qqnorm(CARROS$Preco,xlab = "Quantis teóricos",
       ylab = "Quantis observados",main = "QQ-plot")
qqline(CARROS$Preco, col = 2)
```

```
#### Teste de Normalidade
shapiro.test(CARROS$Preco)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  CARROS$Preco
## W = 0.92001, p-value = 0.02081
```

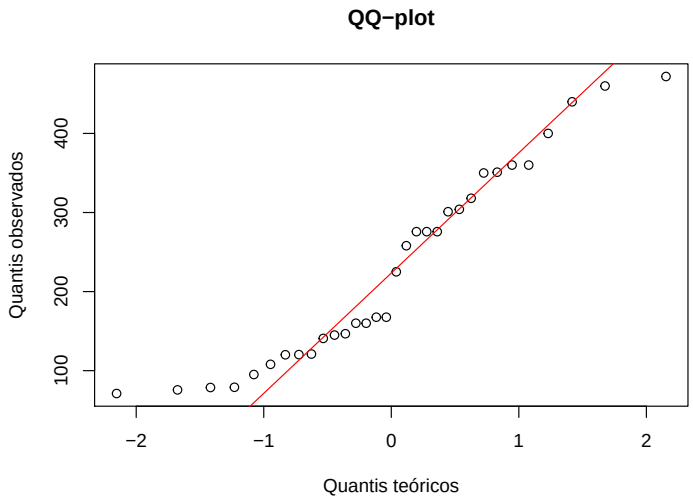


Figure 7.2: Teste de Shapiro Wilk

O preço do carro não segue uma distribuição normal. Possivelmente, os *outliers* pode estar influenciando o resultado. Seria interessante deletar os *outliers* e refazer o teste.

7.0.0.7 Limitações do teste

Cumpre registrar uma limitação do teste de shapiro wilk. Se a sua base de dados tiver mais de 5.000 observações, o teste de Shapiro Wilk pode indicar que o seu banco de dados não segue uma distribuição normal mesmo quando ela tem essa distribuição.

Tem uma ótima discussão no stackexchange (em inglês). Ela pode ser encontrada aqui. Reproduzi parte da discussão abaixo.

```
#-----------------------------------------------------------------
# semente para reproduzir o experimento
set.seed(981677672)

# procedimento para gerar 100 testes de
# Shapiro Wilk em cada distribuição normal.
# (com 10, 100,1.000,e 5.000 observações)
# a função rnorm gera um sorteio aleatório da distribuição normal
x <- replicate(100, {
  c(shapiro.test(rnorm(10)+c(1,0,2,0,1))$p.value,   #$
```

```
    shapiro.test(rnorm(100)+c(1,0,2,0,1))$p.value,
    shapiro.test(rnorm(1000)+c(1,0,2,0,1))$p.value,
    shapiro.test(rnorm(5000)+c(1,0,2,0,1))$p.value)
}
)
rownames(x) <- c("n10","n100","n1000","n5000")
rowMeans(x<0.05) # a proporção de desvios significativos da normalidade
```

```
##   n10  n100 n1000 n5000
##  0.03  0.02  0.19  0.83
```

A última linha verifica qual fração das simulações para cada o tamanho da amostra diverge significativamente da normalidade. Assim, em 83% dos casos, uma amostra de 5000 observações se desvia significativamente da normalidade de acordo com Shapiro-Wilk. No entanto, ao verificar os gráficos qq-plots, você nunca acreditaria no desvio da normalidade.

Chapter 8

Homogeneidade de variâncias

8.1 Teste de Bartlett

8.1.0.1 Por que usar?

O teste de Bartlett é usado para verificar se as amostras têm homogeneidade de variâncias (variâncias iguais). A avaliação do pressuposto de homogeneidade de variâncias é exigida pela maioria dos procedimentos estatísticos. Muitos testes estatísticos assumem que as variâncias são iguais entre grupos. O teste de Bartlett pode ser usado para verificar essa pressuposto. É realizado com a função *bartlett.test*.

8.1.0.2 Pressupostos

O teste assume que as observações são independentes.

8.1.0.3 Dados apropriados

- A variável dependente é quantitativa.
- A variável independente é qualitativa.

8.1.0.4 Hipóteses

Hipótese nula: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_k^2$ (as variâncias são iguais)
Hipótese alternativa: $\sigma_i^2 \neq \sigma_j^2$ (As variâncias são desiguais para pelo menos dois grupos)

8.1.0.5 Interpretação

Resultados significativos podem ser relatados como “os grupos têm variâncias desiguais”, “os dados são heterocedásticos”, ”a variável resposta viola o pressuposto de homogeneidade de variâncias.

8.1.0.6 Exemplo de teste de Bartlett

Este exemplo apresenta os dados de uma amostra de 32 carros.
Esse teste responde à pergunta: “A variância da variável Km/L (quantitativa) é igual para os dois grupos de carros (gasolina/álcool)?”

```
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")
CARROS$Tipodecombustivel<-as.factor(CARROS$Tipodecombustivel)
levels(CARROS$Tipodecombustivel) <- c('Gasolina','Álcool')

### Verifique os dados
str(CARROS$Kmporlitro)
```

```
##  num [1:32] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

```
str(CARROS$Tipodecombustivel)
```

```
##  Factor w/ 2 levels "Gasolina","Álcool": 1 1 2 2 1 2 1 2 2 2 ...
```

```
### Remova objetos desnecessários
remove(mtcars)

#### Resumo dos dados por grupo de combustivel
library(psych)
describeBy(CARROS$Kmporlitro,group = CARROS$Tipodecombustivel)
```

```
##
##  Descriptive statistics by group
## group: Gasolina
```

```
##    vars  n  mean   sd median trimmed  mad  min max range skew kurtosis   se
## X1    1 18 16.62 3.86  15.65   16.42 2.97 10.4  26  15.6 0.48    -0.05 0.91
## ------------------------------------------------------------
## group: Álcool
##    vars  n  mean   sd median trimmed mad  min  max range skew kurtosis   se
## X1    1 14 24.56 5.38   22.8   24.34   6 17.8 33.9  16.1 0.41     -1.4 1.44
```

```
#### Box-plot por grupo de combustivel
boxplot(CARROS$Kmporlitro~CARROS$Tipodecombustivel,
        horizontal = TRUE,col=c("skyblue","red"))
```

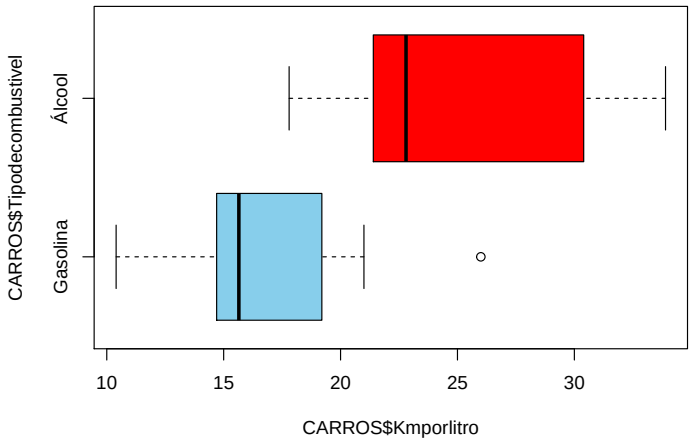


Figure 8.1: Teste

```
#### Teste de Bartlett
bartlett.test(CARROS$Kmporlitro~CARROS$Tipodecombustivel)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  CARROS$Kmporlitro by CARROS$Tipodecombustivel
## Bartlett's K-squared = 1.5849, df = 1, p-value = 0.2081
```

Falhamos em rejeitar a hipótese nula de homogeneidade de variâncias. Ou seja, os dois tipos de combustivel (Gasolina/Álcool) têm variâncias homogêneas para a variável Km/l.

8.1.0.7 Observação

Esse teste é equivalente ao Teste Breusch-Pagan de heterocedásticidade nos modelos lineares.

8.1.0.8 Limitação do teste

O teste de Bartlett é sensível a desvios da normalidade. Recomendo fazê-lo depois do teste de normalidade.

Se suas amostras são provenientes de uma distribuição que não é normal, o teste de Bartlett pode falhar. O teste de Levene é uma alternativa ao teste Bartlett, menos sensível a desvios da normalidade. Em outras palavras, se a distribuição não é normal, use o teste de levene.

8.1.0.9 Exemplo do Teste de Levene

```
library(car)
leveneTest(CARROS$Preco~CARROS$Tipodecombustivel)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  4.3041 0.0467 *
##       30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Chapter 9

Outliers

Os valores extremos (*outliers*) nos dados podem distorcer as previsões. Os *outliers* também podem ser chamados de valores discrepantes. Acredito que é essencial entender o impacto deles nos seus testes de hipóteses.

Cabe ao melhor julgamento do análista decidir como fazer o tratamento de *outliers*.

Os *outliers* podem influenciar drasticamente as estimativas da variância. Além disso, se você está construindo um modelo, eles podem modificar a inclinação da Reta.

9.0.0.1 Exemplo da influência do outlier

Para entender melhor as implicações dos *outliers*, compararei o ajuste de um modelo de regressão linear simples no conjunto de dados de carros com e sem *outliers*.

Para distinguir claramente o efeito, vamos introduzir manualmente os *outliers* no conjunto de dados. Depois disso, vamos fazer uma regressão linear nos dois bancos de dados.

```
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")
# Selecione somente o preço do carro e o HP (Horse Power - cavalos de potência)
CARROS<-CARROS[CARROS$HP<=334,c("Preco","HP")] # dados originais sem outliers
```

```
# Inserir outliers no banco de dados.
outliers_para_carros<-data.frame(Preco=c(319,590,500),
                                 HP=c(590, 686, 690))

# novo banco de dados com outliers
CARROS_OUTLIERS<-rbind(CARROS, outliers_para_carros)

### Remova objetos desnecessários
remove(mtcars,outliers_para_carros)

### Graficos com outliers.
par(mfrow=c(1, 2)) # grafico lado a lado

plot(CARROS_OUTLIERS$HP, CARROS_OUTLIERS$Preco, xlim=c(0, 700),
     ylim=c(0, 600), main="Com outliers", xlab="cavalos de potência",
     ylab="Preço do carro", pch=19, col="red")
abline(lm(Preco~HP, data=CARROS_OUTLIERS), col="blue", lwd=3, lty=2)

### Graficos sem outliers (dados originais)
plot(CARROS$HP, CARROS$Preco, xlim=c(0, 700), ylim=c(0, 700),
     main="Sem outliers", xlab="cavalos de potência",
     ylab="Preço do carro", pch=19, col="red") # um modelo melhor
abline(lm(Preco~HP, data=CARROS), col="blue", lwd=3, lty=2)
```

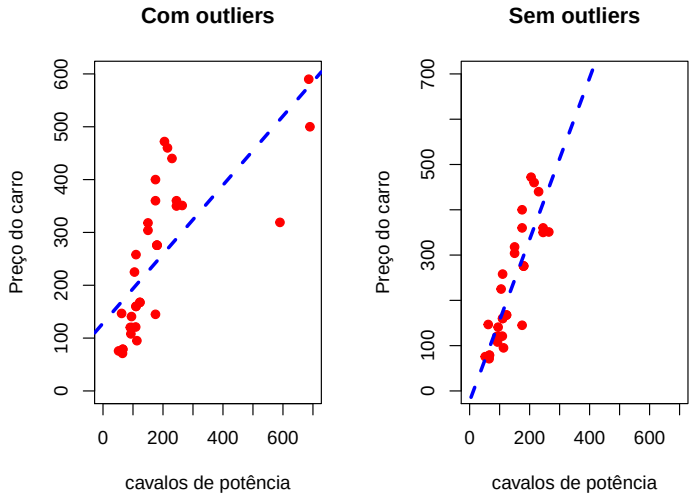


Figure 9.1: Impacto do outlier

9.0.0.2 Como encontrar os outliers?

Método básico

```
par(mfrow=c(1, 2))
valores_outliers <- boxplot.stats(CARROS_OUTLIERS$HP)$out
boxplot(CARROS_OUTLIERS$HP, main="Cavalos de potência",ylim=c(0, 700))
mtext(paste("Outliers: os valores", paste(valores_outliers, collapse=", "),
            "da variável HP"), cex=0.8)
boxplot(CARROS$HP, main="Cavalos de potência", ylim=c(0, 700))
```

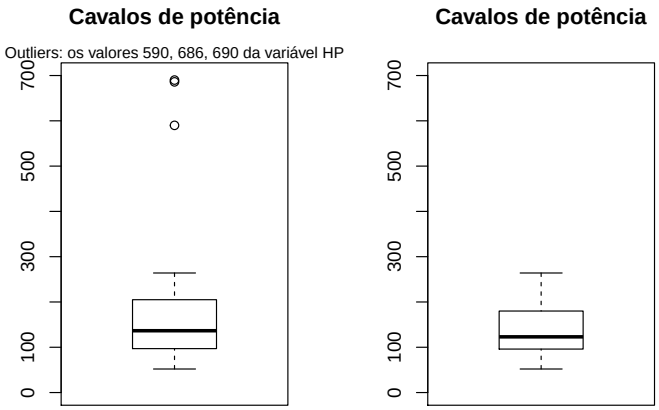


Figure 9.2: Outliers

```
par(mfrow=c(1, 1))

# Teste para Outliers e
# p-valor de Bonferonni para valores extremos
library(car)
modelo<-lm(CARROS$HP~CARROS$Preco)
outlierTest(modelo)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 29  2.62339           0.013932       0.4319
```

Ok. não temos *outliers*, mas podemos ter observações influêntes?

A distância de Cook

A distância de Cook mede a influência da observação sobre todos os valores ajustados.

```
plot(modelo, which=4)
```

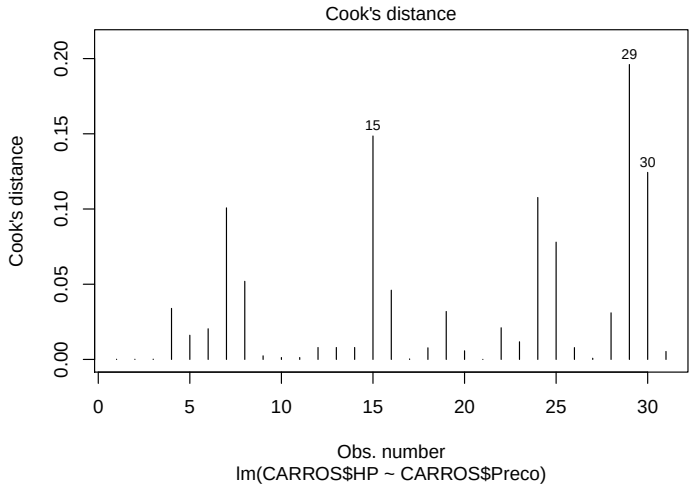


Figure 9.3: Outliers

```
# Influence Plot
influencePlot(modelo, id.method="identify", sub="o tamanho do círculo é
              proporcional à distância de Cook")
```

```
##       StudRes        Hat     CookD
## 15 -1.2707146 0.15816753 0.1485416
## 16 -0.7275545 0.14606554 0.0460183
## 29  2.6233896 0.06413611 0.1960559
## 30  2.4263435 0.04704242 0.1243523
```

```
# vale a pena inverstigar as observações 15,16,29,e 30
```

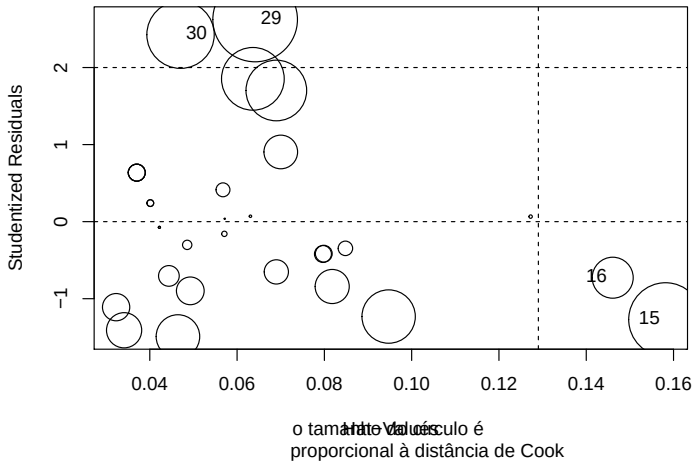


Figure 9.4: Outliers

Além da abordagem básica, os pacotes do R chamados de *outliers* e *OutlierDetection* podem ajuda-lo a decidir se uma observação é um outlier ou não

```
#-----------------------------------------------------
# Teste
#-----------------------------------------------------
library(outliers)
# teste se é um outlier
chisq.out.test(CARROS_OUTLIERS$Preco)
```

```
##
##  chi-squared test for outlier
##
## data:  CARROS_OUTLIERS$Preco
## X-squared = 5.7167, p-value = 0.0168
## alternative hypothesis: highest value 590 is an outlier
```

```
chisq.out.test(CARROS$Preco)
```

```
##
##  chi-squared test for outlier
##
```

```
## data:  CARROS$Preco
## X-squared = 3.7773, p-value = 0.05195
## alternative hypothesis: highest value 472 is an outlier
```

```
# remove o outlier do banco de dados
banco_sem_outliers<-rm.outlier(CARROS_OUTLIERS)
head(banco_sem_outliers)
```

```
##   Preco  HP
## 1   160 110
## 2   160 110
## 3   108  93
## 4   258 110
## 5   360 175
## 6   225 105
```

```
#-------------------------------------------------
# Grafico para mostrar os outliers
#-------------------------------------------------
library(OutlierDetection)
# Univariado - Outlier Detection
UnivariateOutlierDetection(CARROS_OUTLIERS$Preco)
```

```
## $`Outlier Observations`
## [1] 590 500
##
## $`Location of Outlier`
## [1] 33 34
##
## $`Scatter plot`
```

```
UnivariateOutlierDetection(CARROS_OUTLIERS$HP)
```

```
## $`Outlier Observations`
## [1] 686 690
##
## $`Location of Outlier`
## [1] 33 34
##
## $`Scatter plot`
```

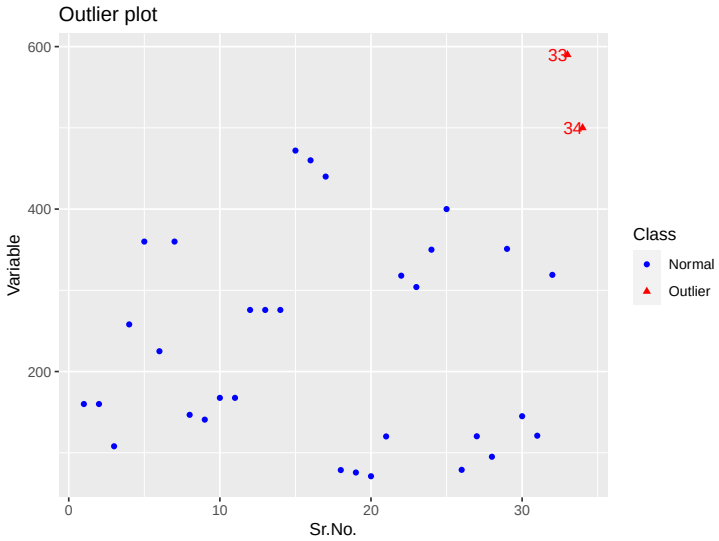


Figure 9.5: Outliers

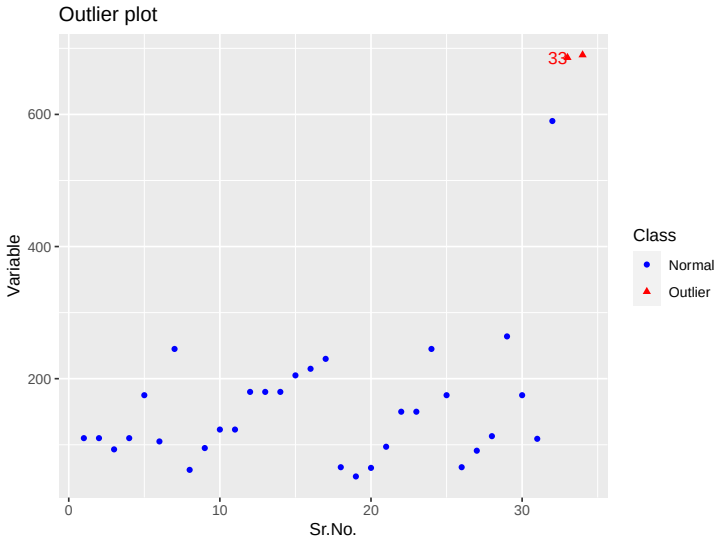


Figure 9.6: Outliers

```
# Bivariado - Outlier Detection
OutlierDetection(CARROS_OUTLIERS)
```

```
## $`Outlier Observations`
##   Preco  HP
## 1   319 590
## 2   590 686
## 3   500 690
##
## $`Location of Outlier`
##  1  2  3
## 32 33 34
##
## $`Scatter plot`
```

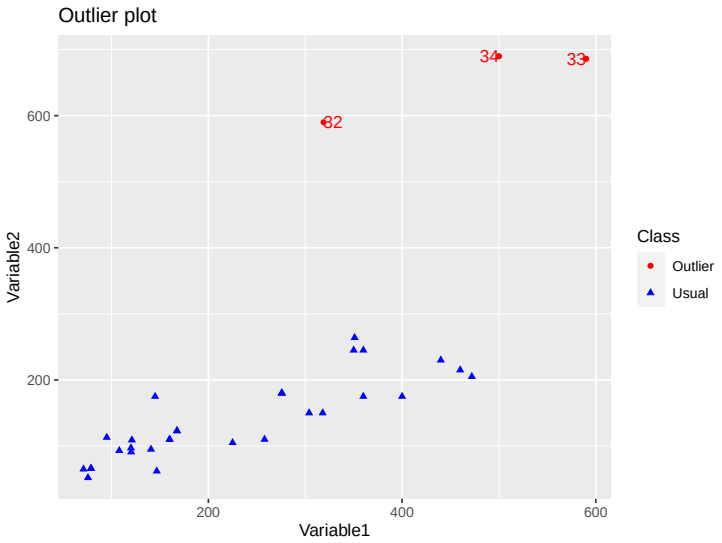


Figure 9.7: Outliers

Parece que o teste detectou que as observações 32,33, e 34 são *outliers*.

Chapter 10

Avaliando a Independência

Pressuposto de independência: as observações são independentes uma da outra.

Regra geral: para verificar a independência, faça um gráfico da variável resposta (ou dos resíduos) com a ordem da observação). Um padrão não aleatório sugere falta de independência.

Dependência de tempo ou variáveis espaciais são fontes comuns de falta de independência entre as observações.

10.0.0.1 Gráfico de resíduos vs ordem da observação

Como usar um "gráfico de resíduos versus ordem" como uma maneira de detectar uma forma específica de não independência dos termos de erro, ou seja, correlação serial. Se os dados forem obtidos em uma sequência de tempo (ou espaço), um gráfico de resíduos versus ordem ajuda a verificar se há alguma correlação entre os termos de erro que estão próximos um do outro na sequência.

Então, o que é esse gráfico de resíduos versus ordem? Como o próprio nome sugere, é um gráfico de dispersão com resíduos no eixo y e a ordem em que os dados foram coletados no eixo x. Aqui temos dois gráficos de resíduos versus ordem. O primeiro independente e o segundo com autocorrelação.

Caso você esteja em um modelo de regressão, pode realizar um teste de Durbin-Watson para autocorrelação dos erros.

```
##### Teste de erros autocorrelacionados
durbinWatsonTest(fit)
```

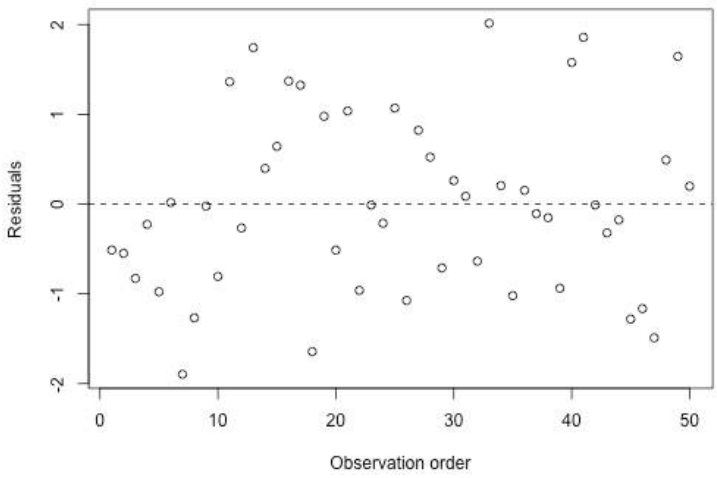


Figure 10.1: resíduos vs ordem da observação

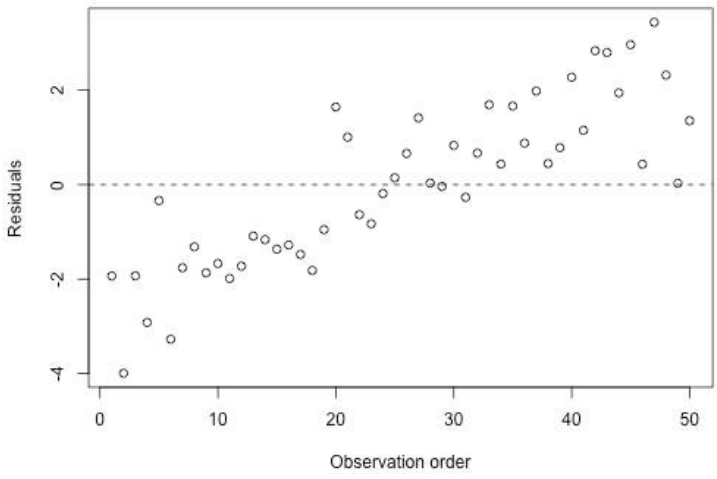


Figure 10.2: resíduos vs ordem da observação

Chapter 11

Associação entre duas variáveis qualitativas

11.1 Teste Qui-Quadrado

O teste do Qui-Quadrado para independência compara duas variáveis qualitativas em uma tabela de contingência para verificar se elas estão relacionadas. O Teste Qui-Quadrado mede como os valores esperados se comparam aos dados reais observados.

11.1.0.1 Dados apropriados

- Variáveis categóricas
- Pelo menos cinco observações em cada celula da tabela de contigência (contagem esperada em cada celula)

11.1.0.2 Hipóteses

Hipótese nula: Não existe associação entre as variáveis
Hipótese alternativa: Existe associação entre as variáveis

11.1.0.3 Interpretação

Resultados significativos podem ser relatados como “há associação entre as duas variáveis”.

```
### Tabela para o teste
M <- as.table(rbind(c(762, 327, 468), c(484, 239, 477)))
### Rótulos para tabela
dimnames(M) <- list(sexo = c("Feminino", "Masculino"),
                    partido = c("PT","Outro partido", "PSDB"))

#### Gráfico
library("gplots")
balloonplot(t(M), main ="Exemplo do Agresti", xlab ="", ylab="",
            label = FALSE, show.margins = FALSE)
```

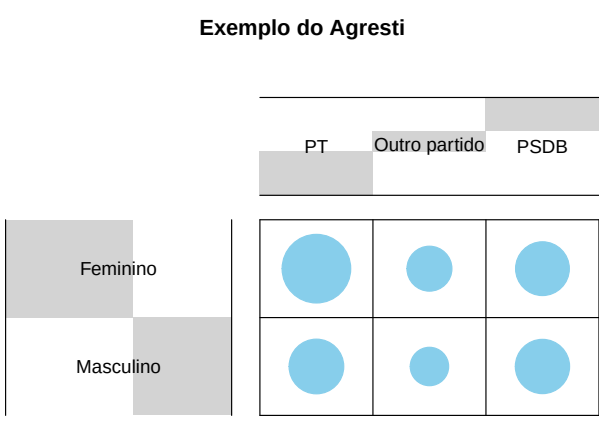


Figure 11.1: Teste Qui-quadrado

```
#### Resumo dos dados
M
```

```
##           partido
## sexo         PT Outro partido PSDB
##   Feminino  762           327  468
##   Masculino 484           239  477
```

```
#### Teste Qui-Quadrado
Xsq <- chisq.test(M)
Xsq
```

```
##
##  Pearson's Chi-squared test
##
## data:  M
## X-squared = 30.07, df = 2, p-value = 0.0000002954
```

```
#### Valores observados (mesmo valor que a tabela M)
Xsq$observed
```

```
##            partido
## sexo         PT Outro partido PSDB
##   Feminino  762           327  468
##   Masculino 484           239  477
```

```
#### Valores esperados sob a hipótese nula
Xsq$expected
```

```
##            partido
## sexo              PT Outro partido     PSDB
##   Feminino  703.6714      319.6453 533.6834
##   Masculino 542.3286      246.3547 411.3166
```

```
# Devem ser maiores que cinco
# (se tiver uma célula menor que cinco, você deve usar o teste exato de fisher)
```

Como todos os valores esperados são maiores que cinco, o teste Qui-Quadrado é adequado. O P-valor é menor que 0,05, logo, concluimos que existe associação entre as duas variáveis qualitativas. Em outras palavras, há prefência dos sexos pelos partidos políticos.

11.1.0.4 Limitação do teste qui-quadrado

O teste qui-quadrado pode ser usado apenas com números. Eles não podem ser usados para porcentagens, proporções, médias ou outras estatísticas. Assim, se

você tiver 10% de 200 pessoas, precisará convertê-lo em um número (20) antes de poder executar o teste.

Referências do teste qui-quadrado Assumptions of the Chi-square The chi-square test of independence

11.2 Teste Exato de Fisher

11.2.0.1 Dados apropriados

- Variáveis categóricas com dois níveis (exemplo: feminino/masculino)
- Qualquer contagem em cada celula

11.2.0.2 Hipóteses

Hipótese nula: Não existe associação entre as variáveis
Hipótese alternativa: Existe associação entre as variáveis

11.2.0.3 Interpretação

Mesma interpretação do teste qui-quadrado. O teste exato de Fisher é interpretado da mesma forma que o teste qui-quadrado.

11.2.0.4 Exemplo do teste exato de Fisher

Um exemplo interessante de como podemos devenvolver testes para tudo é apresentado em (Agresti, 2002). Uma ótima referência para esse experimento é o livro de Salsburg (2009).

O gosto do chá muda de acordo com a ordem em que as ervas e o leite são colocados? Uma britânica diz ser uma especialista em chá. Ela afirmou ser capaz de distinguir se leite ou chá foi adicionado à xícara primeiro. Leite sobre o chá ou chá sobre o leite.

Vamos construir um experimento para verificar isso? Para testar, ela recebeu 8 xícaras de chá, das quais quatro o chá foi adicionado antes do leite.

A hipótese nula é a de que não há associação entre a verdadeira ordem dos ingredientes e a opinião da mulher, a hipótese alternativa de que existe uma associação positiva (que a razão de chances é maior que 1).

```
### Tabela para o teste
resultado_xicaras <-matrix(c(3, 1, 1, 3),
       nrow = 2, dimnames =
         list(opiniao = c("Leite", "Chá"),
              verdadeiro_result = c("Leite", "Chá")))

resultado_xicaras
```

```
##        verdadeiro_result
## opiniao Leite Chá
##   Leite     3   1
##   Chá       1   3
```

```
#### Teste Exato de Fisher
Teste_fisher <- fisher.test(resultado_xicaras, alternative = "greater")
Teste_fisher
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  resultado_xicaras
## p-value = 0.2429
## alternative hypothesis: true odds ratio is greater than 1
## 95 percent confidence interval:
##  0.3135693       Inf
## sample estimates:
## odds ratio
##   6.408309
```

11.2.0.5 Teste exato de Fisher com mais de duas categorias

Se você tiver tabelas maiores que 2 por 2 (mais de duas categorias em uma das variáveis), socê pode usar uma aproximação híbrida.

```
### Tabela para o teste
Tabela <- as.table(rbind(c(2,3,10,6,1), c(1,6,7,14,12)))

### Rótulos para tabela
dimnames(Tabela) <- list(sexo = c("Feminino", "Masculino"),
                    likert = c("Concordo Totalmente","Concordo",
                               "Nem concordo nem discordo",
```

```
                                "Discordo", "Discordo Totalmente"))
Tabela
```

```
##            likert
## sexo        Concordo Totalmente Concordo Nem concordo nem discordo Discordo
##   Feminino                    2        3                        10        6
##   Masculino                   1        6                         7       14
##            likert
## sexo        Discordo Totalmente
##   Feminino                    1
##   Masculino                  12
```

```
Teste_fisher_hybrid <- fisher.test(Tabela,  hybrid=TRUE)
Teste_fisher_hybrid
```

```
##
##  Fisher's Exact Test for Count Data hybrid using asym.chisq. iff
##  (exp=5, perc=80, Emin=1)
##
## data:  Tabela
## p-value = 0.03019
## alternative hypothesis: two.sided
```

Chapter 12

Variáveis qualitativa e quantitativa

12.1 Abordagem paramétrica

12.1.1 Teste t de Student para duas amostras

O Teste t de Student compara duas médias e mostra se as diferenças entre essas médias são significativas. Este teste permite que você avalie se a diferença entre duas médias ocorreram por acaso ou não. Este teste deve ser utilizado somente após a avaliação dos pressupostos de Normalidade, de homogeneidade de variâncias (variâncias iguais), e independência. Ele e realizado com a função *t.test*.

Utiliza-se este teste quando:
* Os dados seguem a Distribuição normal;
* Temos dois grupos;
* Concluimos no teste de Bartlett que os dois grupos têm variâncias iguais.

12.1.1.1 Dados apropriados

- A variável dependente é quantitativa.
- A variável independente é qualitativa com dois níveis.

12.1.1.2 Hipóteses

Hipótese nula:

$$\mu_1 = \mu_2$$

Hipótese alternativa:

$$\mu_1 \neq \mu_2$$

12.1.1.3 Interpretação

Resultados significativos podem ser relatados como "Houve uma diferença significativa entre as médias dos dois grupos".

12.1.1.4 Exemplo do teste t de Student

Este exemplo apresenta os dados de uma amostra de 32 carros.
Esse teste responde à pergunta: "A média da variável Km/L (quantitativa) é igual para os dois grupos de carros (gasolina/álcool)?"

```
#-----------------------------------------------------------------
### Banco de dados
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")
CARROS$Tipodecombustivel<-as.factor(CARROS$Tipodecombustivel)
levels(CARROS$Tipodecombustivel) <- c('Gasolina','Álcool')

### Verifique os dados
str(CARROS$Kmporlitro)
```

```
##  num [1:32] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

```
str(CARROS$Tipodecombustivel)
```

```
##  Factor w/ 2 levels "Gasolina","Álcool": 1 1 2 2 1 2 1 2 2 2 ...
```

```
### Remova objetos desnecessários
remove(mtcars)

#### Resumo dos dados
library(psych)
describeBy(CARROS$Kmporlitro,group = CARROS$Tipodecombustivel)
```

```
##
##  Descriptive statistics by group
## group: Gasolina
##    vars  n  mean   sd median trimmed  mad  min max range skew kurtosis   se
## X1    1 18 16.62 3.86  15.65   16.42 2.97 10.4  26  15.6 0.48    -0.05 0.91
## ------------------------------------------------------------
## group: Álcool
##    vars  n  mean   sd median trimmed mad  min  max range skew kurtosis   se
## X1    1 14 24.56 5.38   22.8   24.34   6 17.8 33.9  16.1 0.41     -1.4 1.44
```

```
#### Box-plot
boxplot(CARROS$Kmporlitro~CARROS$Tipodecombustivel,
        horizontal = TRUE,col=c("skyblue","red"))
```

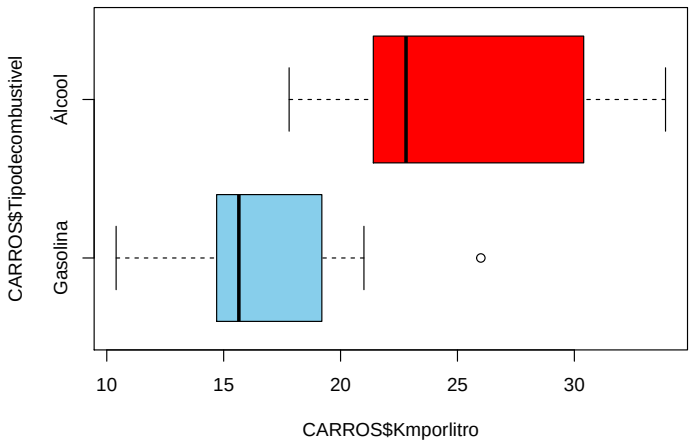


Figure 12.1: Teste de Shapiro Wilk

```
#### Teste t de Student
t.test(CARROS$Kmporlitro~CARROS$Tipodecombustivel, var.equal=TRUE)
```

```
##
##  Two Sample t-test
##
## data:  CARROS$Kmporlitro by CARROS$Tipodecombustivel
## t = -4.8644, df = 30, p-value = 0.00003416
```

```
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.274221  -4.606732
## sample estimates:
## mean in group Gasolina   mean in group Álcool
##               16.61667               24.55714
```

12.1.1.5 Teste t pareado

O teste t pareado é utilizado com amostras dependentes(pareadas). Em amostras pareadas para cada observação realizamos duas mensurações. As medidas são tomadas em um único indivíduo em dois pontos distintos no tempo. Em geral, observações pareadas correspondem a medidas tomadas antes e depois de uma intervenção.

Veja o exemplo a seguir, no qual existem um grupo de pacientes que teve alguma medida realizada antes da intervenção e outra depois da intervenção. Por exemplo, uma métrica de popularidade de políticos (de 0 a 100) antes e depois de uma campanha de marketing uma medida em uma comunidade antes e depois de uma política pública. O importante aqui é entendermos que as duas medidas são realizadas num mesmo grupo de pessoas, antes e depois de uma intervenção.

O teste t para amostras indepententes não exige que os grupos tenham o mesmo tamanho, mas observe que o teste t pareado exige que existam exatamente o mesmo número de medidas antes e depois. Para usar o teste t pareado, iremos simplesmente alterar o argumento paired:

```r
#-----------------------------------------------------------------
### Banco de dados
antes  <- c(21, 28, 24, 23, 23, 19, 28, 20, 22, 20, 26, 26)
depois <- c(26, 27, 23, 25, 25, 29, 30, 31, 36, 23, 32, 22)
dados<-data.frame(antes,depois)

## Grafico simples da diferenca
Diferenca <- (dados$depois - dados$antes)
plot(Diferenca,pch = 16,ylab="Diferença (Depois/antes)")
abline(0,0, col="blue", lwd=2)
```

```r
#### Teste t pareado
t.test(dados$antes, dados$depois, paired = TRUE)
```

```
##
##  Paired t-test
```

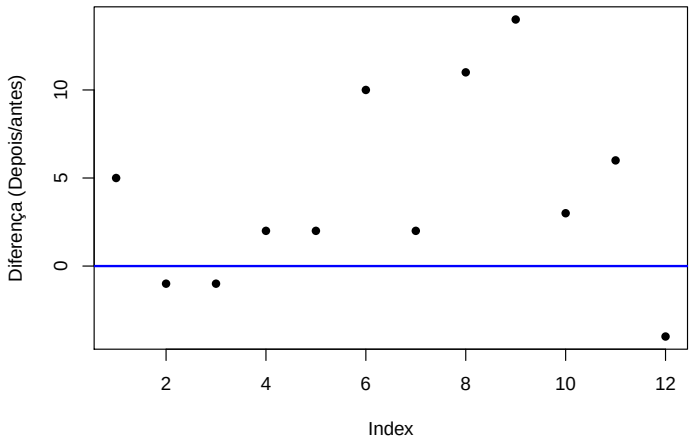


Figure 12.2: Teste

```
##
## data:  dados$antes and dados$depois
## t = -2.6353, df = 11, p-value = 0.02319
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.493713 -0.672954
## sample estimates:
## mean of the differences
##               -4.083333
```

12.1.1.6 Interpretação do resultado do teste-t pareado

O teste é interpretado da mesma forma que o teste t para duas amostras independentes.

12.1.2 Teste-t de Welch de duas amostras (Teste t de Student com variâncias desiguais)

O teste-t de Welch compara duas médias e mostra se as diferenças entre essas médias são significativas. O teste-t de Welch é uma adaptação do teste t de Student, que é mais confiável quando as duas amostras têm variâncias desiguais.

Em outras palavras, este teste permite que você avalie se a diferença entre duas médias ocorreram por um mero por acaso ou não. Este teste deve ser utilizado

somente após a avaliação do pressuposto de homogeneidade de variâncias (variâncias iguais).

Utiliza-se este teste quando:
* Os dados seguem a Distribuição normal;
* Temos dois grupos;
* Concluimos no teste de Bartlett que os dois grupos têm variâncias diferentes.

É realizado com a função *t.test*.

12.1.2.1 Dados apropriados

- A variável dependente é quantitativa.
- A variável independente é qualitativa com dois níveis.

12.1.2.2 Hipóteses

Hipótese nula:

$$\mu_1 = \mu_2$$

Hipótese alternativa:

$$\mu_1 \neq \mu_2$$

12.1.2.3 Interpretação

Resultados significativos podem ser relatados como "Houve uma diferença significativa entre as médias dos dois grupos".

```
#-----------------------------------------------------------------
### Banco de dados
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")
CARROS$Tipodecombustivel<-as.factor(CARROS$Tipodecombustivel)
levels(CARROS$Tipodecombustivel) <- c('Gasolina','Álcool')

### Verifique os dados
str(CARROS$Kmporlitro)
```

```
##  num [1:32] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

```
str(CARROS$Tipodecombustivel)
```

```
##  Factor w/ 2 levels "Gasolina","Álcool": 1 1 2 2 1 2 1 2 2 2 ...
```

```
### Remova objetos desnecessários
remove(mtcars)

#### Resumo dos dados
library(psych)
describeBy(CARROS$Kmporlitro,group = CARROS$Tipodecombustivel)
```

```
##
##  Descriptive statistics by group
## group: Gasolina
##    vars  n  mean   sd median trimmed  mad  min max range skew kurtosis   se
## X1    1 18 16.62 3.86  15.65   16.42 2.97 10.4  26  15.6 0.48    -0.05 0.91
## ------------------------------------------------------------
## group: Álcool
##    vars  n  mean   sd median trimmed mad  min  max range skew kurtosis   se
## X1    1 14 24.56 5.38   22.8   24.34   6 17.8 33.9  16.1 0.41     -1.4 1.44
```

```
#### Box-plot
boxplot(CARROS$Kmporlitro~CARROS$Tipodecombustivel,
        horizontal = TRUE,col=c("skyblue","red"))
```

```
#### Teste t de Student com variâncias desiguais (Teste-t de Welch)
t.test(CARROS$Kmporlitro~CARROS$Tipodecombustivel, var.equal=FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  CARROS$Kmporlitro by CARROS$Tipodecombustivel
## t = -4.6671, df = 22.716, p-value = 0.0001098
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.462508  -4.418445
## sample estimates:
## mean in group Gasolina   mean in group Álcool
##               16.61667               24.55714
```

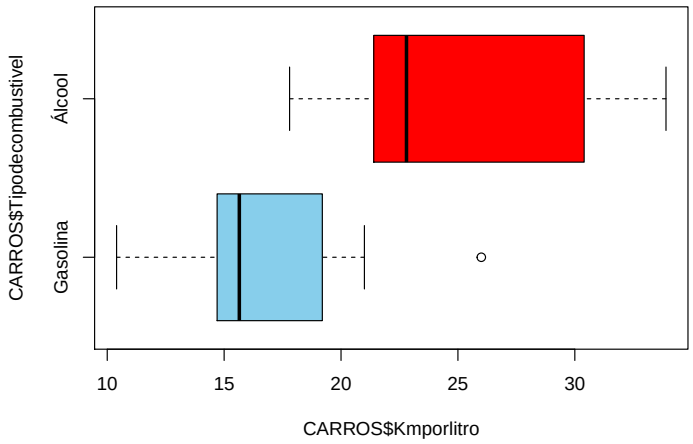


Figure 12.3: Teste

12.1.3 Análise de Variância - ANOVA

A ANOVA é parecida com o teste t de Student, mas pode comparar mais de duas médias e mostra se as diferenças entre essas médias são significativas. Este teste permite que você avalie se a diferença entre as médias ocorreram por acaso ou não.

Este teste deve ser utilizado somente após a avaliação dos pressupostos de Normalidade, de homogeneidade de variâncias (variâncias iguais), e independência. Ele e realizado com a função *aov*.

Utiliza-se este teste quando:
* Os dados seguem a Distribuição normal;
* Temos dois grupos ou mais;
* Concluimos no teste de Bartlett que os grupos têm variâncias iguais.

12.1.3.1 Dados apropriados

- A variável dependente é quantitativa.
- A variável independente é qualitativa.

12.1.3.2 Hipóteses

Hipótese nula:

$$\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

(as médias são iguais)
Hipótese alternativa:

$$\mu_i \neq \mu_j$$

(pelo menos uma média diferente para qualquer i e j).

12.1.3.3 Interpretação

Resultados significativos podem ser relatados como "Houve uma diferença significativa entre pelo menos duas médias. Devemos realizar os testes post hoc".

12.1.3.4 Exemplo da ANOVA

```
#------------------------------------------------------------
### Banco de dados
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")
CARROS$Cilindros<-as.factor(CARROS$Cilindros)

### Verifique os dados
str(CARROS$Kmporlitro)
```

```
##  num [1:32] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

```
str(CARROS$Cilindros)
```

```
##  Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
```

```
levels(CARROS$Cilindros)
```

```
## [1] "4" "6" "8"
```

```
### Remova objetos desnecessários
remove(mtcars)

#### Resumo dos dados
library(psych)
describeBy(CARROS$Kmporlitro,group = CARROS$Cilindros)
```

```
##
##  Descriptive statistics by group
## group: 4
##    vars  n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 11 26.66 4.51     26   26.44 6.52 21.4 33.9  12.5 0.26    -1.65 1.36
## ------------------------------------------------------------
## group: 6
##    vars n  mean   sd median trimmed  mad  min  max range  skew kurtosis   se
## X1    1 7 19.74 1.45   19.7   19.74 1.93 17.8 21.4   3.6 -0.16    -1.91 0.55
## ------------------------------------------------------------
## group: 8
##    vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
## X1    1 14 15.1 2.56   15.2   15.15 1.56 10.4 19.2   8.8 -0.36    -0.57 0.68
```

```
#### Box-plot
boxplot(CARROS$Kmporlitro~CARROS$Cilindros,
        horizontal = TRUE,col=c("skyblue","red","green"))
```



```
#### Verificando os pressupostos de normalidade (resíduos com
# distribuição normal) e homogeneidade de variâncias (variância
# constante)
modelo <- aov(Kmporlitro~Cilindros,data = CARROS)
residuos<-residuals(modelo)
shapiro.test(residuos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.97065, p-value = 0.5177
```

```
bartlett.test(residuos~CARROS$Cilindros)
```

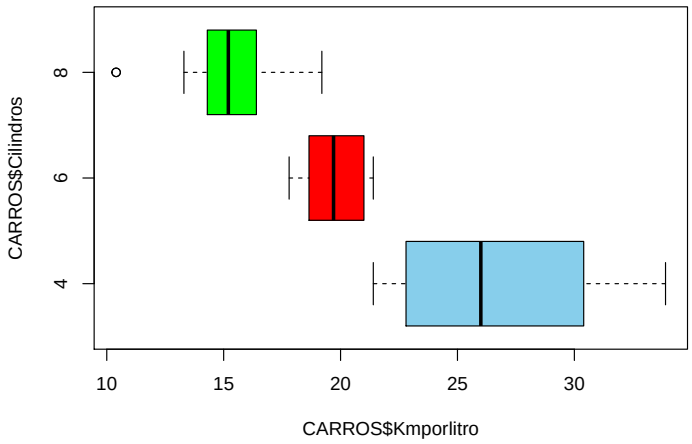


Figure 12.4: Teste

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residuos by CARROS$Cilindros
## Bartlett's K-squared = 8.3934, df = 2, p-value = 0.01505
```

```
#### Os residuos tem distribuição normal, mas a variancia
# não é constante.
# vamos fazer uma transformação LOG
modelo2 <- aov(log(Kmporlitro)~Cilindros,data = CARROS)
residuos<-residuals(modelo2)
shapiro.test(residuos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.96444, p-value = 0.3616
```

```
bartlett.test(residuos~CARROS$Cilindros)
```

```
##
```

```
##  Bartlett test of homogeneity of variances
##
## data:  residuos by CARROS$Cilindros
## Bartlett's K-squared = 4.7326, df = 2, p-value = 0.09383
```

```
#### Ótimo! os residuos tem distribuição normal e
# a variancia é constante.
summary(modelo2)
```

```
##             Df Sum Sq Mean Sq F value        Pr(>F)
## Cilindros    2 2.0081  1.0040   39.31 0.00000000552 ***
## Residuals   29 0.7407  0.0255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A partir deste resultado, podemos ver que a variável cilindros é significativa, mas não sabemos quais das três médias são diferentes entre si.

No entanto, isso exigirá três testes (quatro vs seis, seis vs oito e quatro vs oito), portanto, desejamos mostrar quais dessas diferenças são significativas. Isso pode ser realizado com os testes *post-hoc*. Os testes *post-hoc* muitas vezer são chamados de testes a posteriori porque eles devem ser realizados depois da ANOVA.

Podemos ver as diferencas entre as médias e o p-valor usando o comando *TukeyHSD* (*Tukey Honest Significant Differences*). Se quiser saber mais, esta é uma ótima referencia

```
#-----------------------------------------------------------------
TukeyHSD(modelo2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(Kmporlitro) ~ Cilindros, data = CARROS)
##
## $Cilindros
##           diff        lwr         upr     p adj
## 6-4 -0.2900143 -0.4808396 -0.09918904 0.0021808
## 8-4 -0.5702825 -0.7293036 -0.41126146 0.0000000
## 8-6 -0.2802682 -0.4629695 -0.09756690 0.0019875
```

Podemos ver que as médias de Km/l com 6 e 4, 8 e 4, e 8 e 6 cilindros são significativas.

12.1.4 ANOVA de Welch

Um procedimento considerado equivalente a transformação dos dados é a Anova de Welch para variâncias desiguais.

12.1.4.1 Dados apropriados

- A variável dependente é quantitativa.
- A variável independente é qualitativa.

12.1.4.2 Hipóteses

Hipótese nula:

$$\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$$

(as médias são iguais)
Hipótese alternativa:

$$\mu_i \neq \mu_j$$

(pelo menos uma média diferente para qualquer i e j).

12.1.4.3 Interpretação

Resultados significativos podem ser relatados como “Houve uma diferença significativa entre pelo menos duas médias.”.

12.1.4.4 Exemplo da ANOVA de Welch

```
#-------------------------------------------------------------
### Banco de dados
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")
CARROS$Cilindros<-as.factor(CARROS$Cilindros)

### Verifique os dados
str(CARROS$Kmporlitro)
```

```
##  num [1:32] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

```
str(CARROS$Cilindros)
```

```
##  Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
```

```
levels(CARROS$Cilindros)
```

```
## [1] "4" "6" "8"
```

```
### Remova objetos desnecessários
remove(mtcars)

#### Resumo dos dados
library(psych)
describeBy(CARROS$Kmporlitro,group = CARROS$Cilindros)
```

```
##
##  Descriptive statistics by group
## group: 4
##    vars  n  mean   sd median trimmed  mad  min  max range skew kurtosis   se
## X1    1 11 26.66 4.51     26   26.44 6.52 21.4 33.9  12.5 0.26    -1.65 1.36
## ------------------------------------------------------------
## group: 6
##    vars n  mean   sd median trimmed  mad  min  max range  skew kurtosis   se
## X1    1 7 19.74 1.45   19.7   19.74 1.93 17.8 21.4   3.6 -0.16    -1.91 0.55
## ------------------------------------------------------------
## group: 8
##    vars  n mean   sd median trimmed  mad  min  max range  skew kurtosis   se
## X1    1 14 15.1 2.56   15.2   15.15 1.56 10.4 19.2   8.8 -0.36    -0.57 0.68
```

```
#### Box-plot
boxplot(CARROS$Kmporlitro~CARROS$Cilindros,
        horizontal = TRUE,col=c("skyblue","red","green"))
```

```
#### Verificando os pressupostos de normalidade (resíduos com
# distribuição normal) e homogeneidade de variâncias (variância
# constante)
```

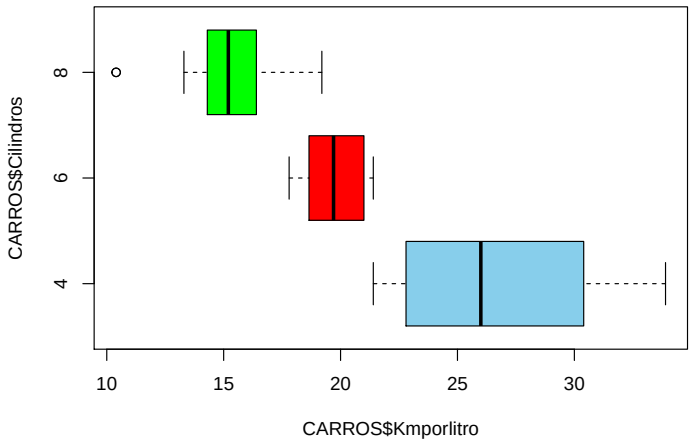


Figure 12.5: Teste

```
modelo <- aov(Kmporlitro~Cilindros,data = CARROS)
residuos<-residuals(modelo)
shapiro.test(residuos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.97065, p-value = 0.5177
```

```
bartlett.test(residuos~CARROS$Cilindros)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residuos by CARROS$Cilindros
## Bartlett's K-squared = 8.3934, df = 2, p-value = 0.01505
```

```
#### Os residuos tem distribuição normal, mas a variancia
# não é constante.
```

```r
# vamos utilizar a ANOVA de Welch

modelo2 <- oneway.test(Kmporlitro~Cilindros,data = CARROS)
modelo2
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Kmporlitro and Cilindros
## F = 31.624, num df = 2.000, denom df = 18.032, p-value = 0.000001271
```

o comando *TukeyHSD* não funciona com a ANOVA de Welch. Ainda não encontrei um pacote do R para facilitar a aplicação dos teste post-hoc. Desso modo, vou deixar a discussão sobre como implementar esses testes aqui

12.1.5 ANOVA em Blocos

A ANOVA em Blocos determina se há diferenças entre os grupos em um *design* de blocos. Nesse tipo de projeto, temos duas variáveis independentes: uma variável que serve como tratamento ou grupo e a outra serve como variável de bloco.

É nas diferenças entre tratamentos ou grupos que estamos interessados. Não estamos interessados nas diferenças entre os blocos, mas queremos que nossas estatísticas levem em conta as diferenças nos blocos (controle o efeito do bloco).

```r
#-------------------------------------------------------------------
# ANOVA com Blocos
tratamento   <- factor(rep(1:4, each = 4))
bloco <- factor(rep(1:4, times = 4))
y <- c(9.3, 9.4, 9.6, 10, 9.4, 9.3, 9.8, 9.9,
       9.2, 9.4, 9.5, 9.7,9.7, 9.6, 10, 10.2)
banco <- data.frame(y, tratamento, bloco)
## Analyze data ####
modelo <- aov(y ~ tratamento + bloco, data = banco)
summary(modelo)
```

```
##             Df Sum Sq Mean Sq F value    Pr(>F)
## tratamento   3  0.385 0.12833   14.44  0.000871 ***
## bloco        3  0.825 0.27500   30.94 0.0000452 ***
## Residuals    9  0.080 0.00889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12.1.6 ANOVA com dois fatores: o desenho fatorial

Esse seguimento é uma tentativa de tradução do livro Mangiafico (2015). Se você quiser ler o documento original, clique aqui

A ANOVA com dois fatores geralmente vêm de experimentos com um planejamento fatorial. Um planejamento fatorial possui pelo menos duas variáveis independentes.

O exemplo de ganho de peso abaixo mostra dados fatoriais. Neste exemplo, existem três observações para cada combinação de dieta e país.

Com esse tipo de dado, geralmente estamos interessados em testar o efeito de cada variável independente (efeitos principais) e depois o efeito de sua combinação (efeito de interação).

Um modo de ver esse efeito é por meio do gráfico de interação. Ele mostra o valor médio ou mediano da variável de resposta para cada combinação das variáveis independentes. Esse tipo de gráfico, especialmente se incluir barras de erro para indicar a variabilidade dos dados dentro de cada grupo, nos dá uma compreensão do efeito dos principais fatores e de sua interação.

Quando a interação é significativa, não podemos realizar testes post-hoc para os efeitos principais. Os resultados podem ser mascarados pelo efeito da interação. Com um desenho fatorial, existem diretrizes para determinar quando fazer testes post-hoc. De uma forma simplificada, as diretrizes são apresentadas a seguir:

• Quando nem os efeitos principais nem o efeito de interação são estatisticamente significativos, nenhum teste post-hoc de separação média deve ser realizado.
• Quando um ou mais dos efeitos principais são estatisticamente significativos e o efeito da interação tem o p-valor acima de 0,05, os testes pós-hoc devem ser realizado apenas com efeitos principais significativos. • Quando o efeito da interação é estatisticamente significativo, os testes pós-hoc devem ser realizado apenas no efeito da interação. Este é o caso mesmo quando os principais efeitos também são estatisticamente significativos. Isso porque os efeitos principais podem ser mascarados pelo efeito da interação.

12.1.6.1 Exemplo ANOVA com dois fatores sem efeito de interação

Imagine para este exemplo um experimento em que as pessoas foram submetidas a uma das três dietas para incentivar a perda de peso. O peso ganho será a variável dependente quantitativa. Temos duas variáveis independentes: 1. a variável categórica chamada dieta com três níveis diferentes e 2. o país em que os indivíduos vivem (variável categórica com dois níveis).

Esse tipo de análise faz certas suposições sobre a distribuição dos dados (linearidade, independência, normalidade e homocedasticidade), mas, por simplicidade,

este exemplo ignorará a necessidade de determinar se os dados atendem a essas suposições.

```
#---------------------------------------------------------------
Entrada =("
dieta    pais  mudanca_peso
 A       BR       0.120
 A       BR       0.125
 A       BR       0.112
 A       UK       0.052
 A       UK       0.055
 A       UK       0.044
 B       BR       0.096
 B       BR       0.100
 B       BR       0.089
 B       UK       0.025
 B       UK       0.029
 B       UK       0.019
 C       BR       0.149
 C       BR       0.150
 C       BR       0.142
 C       UK       0.077
 C       UK       0.080
 C       UK       0.066
")

Dados <- read.table(textConnection(Entrada),header=TRUE)

### Transfomar character em categórica
Dados$pais <- as.factor(Dados$pais)
Dados$dieta <- as.factor(Dados$dieta)

### Verifique os dados
str(Dados)
```

```
## 'data.frame':    18 obs. of  3 variables:
##  $ dieta       : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 2 2 2 2 ...
##  $ pais        : Factor w/ 2 levels "BR","UK": 1 1 1 2 2 2 1 1 1 2 ...
##  $ mudanca_peso: num  0.12 0.125 0.112 0.052 0.055 0.044 0.096 0.1 0.089 0.025 ...
```

```
### Remova objetos desnecessários
remove(Entrada)
```

12.1.6.2 Gráfico de interação

A função *action.plot* cria um gráfico de interação simples para dados com duas independentes categóricas. Para o significado das opções, consulte *?action.plot*.

Esse estilo de gráfico de interação não mostra a variabilidade da média de cada grupo, portanto, é difícil usar esse estilo de gráfico para determinar se há diferenças significativas entre os grupos.

O gráfico mostra que o ganho de peso médio para cada dieta foi menor no Reino Unido em comparação com o Brasil. E que essa diferença era relativamente constante para cada dieta, como é evidenciado pelas linhas paralelas na trama. Isso sugere que não há efeito de interação grande ou significativo. Ou seja, a diferença entre as dietas é constante entre os países.

```
#-----------------------------------------------------------------
interaction.plot(x.factor     = Dados$pais,
                 trace.factor = Dados$dieta,
                 response     = Dados$mudanca_peso,
                 fun = mean,type="b",
                 col=c("black","red","green"),
                 pch=c(19, 17, 15),
                 fixed=TRUE,leg.bty = "o")
```

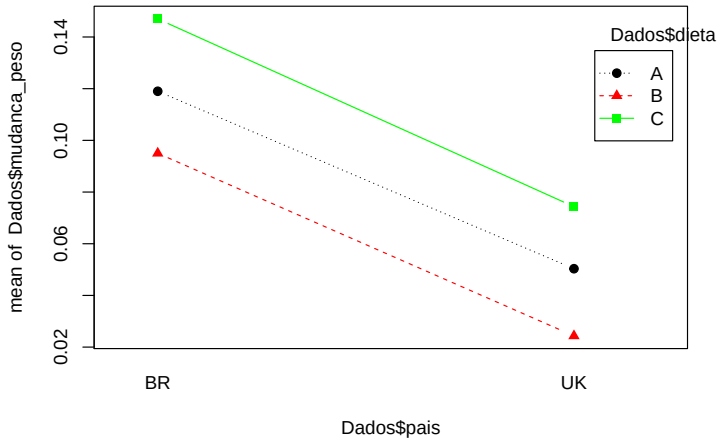


Figure 12.6: interação

12.1.6.3 Como especificar o modelo linear e realizar uma ANOVA

O modelo linear é especificado com: * a variável dependente (y) é a *mudanca_peso*. * as variáveis independentes são *país* e *dieta*. e * o termo de interação para *país* e *dieta* é adicionada ao modelo (*país:dieta*).

A tabela ANOVA indica que os efeitos principais (efeitos das variáveis) são significativos, mas que o efeito de interação não é.

```
#-------------------------------------------------------------
# Two Way Factorial Design
modelo <-aov(mudanca_peso ~ dieta + pais + dieta:pais,data = Dados)
#forma alternativa
#modelo <-aov(mudanca_peso ~ dieta*pais,data = Dados)
summary(modelo)
```

```
##             Df   Sum Sq  Mean Sq F value           Pr(>F)
## dieta        2 0.007804 0.003902 114.205 0.00000001546589 ***
## pais         1 0.022472 0.022472 657.717 0.00000000000752 ***
## dieta:pais   2 0.000012 0.000006   0.176            0.841
## Residuals   12 0.000410 0.000034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12.1.6.4 Teste post-hoc com lsmeans

Como os efeitos principais foram significativos, queremos realizar testes de comparação par-a-par (*post-hoc*) para cada variável independente (efeito principal).

Para isso, usaremos o pacote *lsmeans*. A fórmula na função *lsmeans* indica que comparações aos pares devem ser realizadas para a variável *País* e para a variável *Dieta*.

```
#-------------------------------------------------------------

library(lsmeans)

lsmeans(modelo,
        pairwise ~ pais,
        adjust="tukey")
```

```
## $lsmeans
##  pais lsmean      SE df lower.CL upper.CL
```

```
##  BR   0.1203 0.00195 12   0.1154   0.1253
##  UK   0.0497 0.00195 12   0.0447   0.0546
##
## Results are averaged over the levels of: dieta
## Confidence level used: 0.95
## Conf-level adjustment: sidak method for 2 estimates
##
## $contrasts
##  contrast estimate      SE df t.ratio p.value
##  BR - UK    0.0707 0.00276 12 25.646  <.0001
##
## Results are averaged over the levels of: dieta
```

```
lsmeans(modelo,
        pairwise ~ dieta,
        adjust="tukey")
```

```
## $lsmeans
##  dieta lsmean      SE df lower.CL upper.CL
##  A     0.0847 0.00239 12   0.0781   0.0913
##  B     0.0597 0.00239 12   0.0531   0.0663
##  C     0.1107 0.00239 12   0.1041   0.1173
##
## Results are averaged over the levels of: pais
## Confidence level used: 0.95
## Conf-level adjustment: sidak method for 3 estimates
##
## $contrasts
##  contrast estimate      SE df t.ratio p.value
##  A - B       0.025 0.00337 12   7.408 <.0001
##  A - C      -0.026 0.00337 12  -7.704 <.0001
##  B - C      -0.051 0.00337 12 -15.112 <.0001
##
## Results are averaged over the levels of: pais
## P value adjustment: tukey method for comparing a family of 3 estimates
```

12.1.6.5 Exemplo ANOVA com dois fatores com efeito de interação

Vamos usar o mesmo exemplo, mas vamos colocar um país a mais: a Argentina

```
#------------------------------------------------------------
Entrada =("
```

```
dieta    pais  mudanca_peso
 A       BR      0.120
 A       BR      0.125
 A       BR      0.112
 A       UK      0.052
 A       UK      0.055
 A       UK      0.044
 A       AR      0.080
 A       AR      0.090
 A       AR      0.075
 B       BR      0.096
 B       BR      0.100
 B       BR      0.089
 B       UK      0.025
 B       UK      0.029
 B       UK      0.019
 B       AR      0.055
 B       AR      0.065
 B       AR      0.050
 C       BR      0.149
 C       BR      0.150
 C       BR      0.142
 C       UK      0.077
 C       UK      0.080
 C       UK      0.066
 C       AR      0.055
 C       AR      0.065
 C       AR      0.050
 C       AR      0.054
")


Dados <- read.table(textConnection(Entrada),header=TRUE)

### Transfomar character em categórica
Dados$pais <- as.factor(Dados$pais)
Dados$dieta <- as.factor(Dados$dieta)

### Verifique os dados
str(Dados)
```

```
## 'data.frame':    28 obs. of  3 variables:
##  $ dieta       : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 2 ...
##  $ pais        : Factor w/ 3 levels "AR","BR","UK": 2 2 2 3 3 3 1 1 1 2 ...
```

```
##  $ mudanca_peso: num  0.12 0.125 0.112 0.052 0.055 0.044 0.08 0.09 0.075 0.096 ...
```

```
### Remova objetos desnecessários
remove(Entrada)
```

12.1.6.6 Gráfico de interação

O gráfico sugere que o efeito da dieta não é consistente nos três países. Embora a Dieta C tenha apresentado o maior ganho médio de peso para o Brasil e o Reino Unido, para a Argentina ela tem uma média menor que a Dieta A. Isso sugere que pode haver um efeito de interação significativo, mas precisaremos fazer um teste estatístico para confirmar esta hipótese .

```
#-----------------------------------------------------------------
interaction.plot(x.factor     = Dados$pais,
                 trace.factor = Dados$dieta,
                 response     = Dados$mudanca_peso,
                 fun = mean,type="b",
                 col=c("black","red","green"),
                 pch=c(19, 17, 15),
                 fixed=TRUE,leg.bty = "o")
```

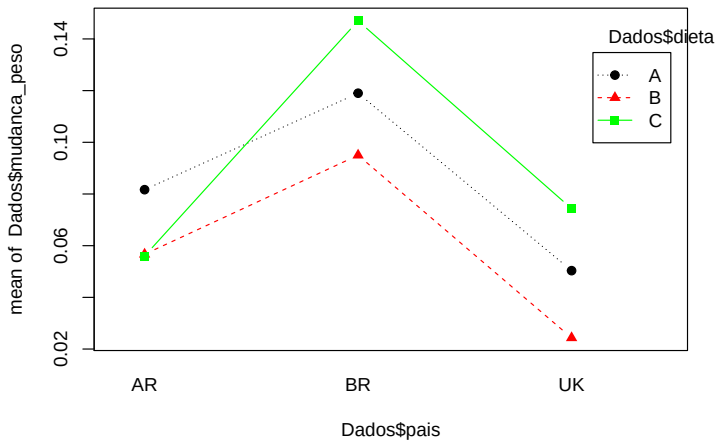


Figure 12.7: interação

A tabela ANOVA indica que os efeitos principais e o efeito de interação são significativos.

```
#-----------------------------------------------------------------
# Two Way Factorial Design
modelo <-aov(mudanca_peso ~ dieta + pais + dieta:pais,data = Dados)
#forma alternativa
#modelo <-aov(mudanca_peso ~ dieta*pais,data = Dados)
summary(modelo)
```

```
##             Df   Sum Sq  Mean Sq F value              Pr(>F)
## dieta        2 0.004811 0.002406   59.72 0.00000000639732710 ***
## pais         2 0.025676 0.012838  318.71 0.00000000000000243 ***
## dieta:pais   4 0.004016 0.001004   24.93 0.00000024772486736 ***
## Residuals   19 0.000765 0.000040
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12.1.6.7 Teste post-hoc com lsmeans

Como o efeito da interação foi significativo, gostaríamos de comparar todas as médias de grupos da interação. Embora os efeitos principais tenham sido significativos, o conselho típico é não realizar comparações aos pares dos efeitos principais quando a interação deles for significativa. Isso ocorre porque o foco nos grupos na interação descreve melhor os resultados da análise.

Usaremos o mesmo pacote *lsmeans* e vamos fazer os contrastes para a interação País e Dieta, indicadas com *país:dieta*.

```
#-----------------------------------------------------------------

library(lsmeans)

marginal <- lsmeans(modelo,
                    pairwise ~ pais:dieta,
                    adjust="tukey")
marginal$contrasts
```

```
##  contrast     estimate      SE df t.ratio p.value
##  AR,A - BR,A -0.037333 0.00518 19  -7.204 <.0001
##  AR,A - UK,A  0.031333 0.00518 19   6.046 0.0002
##  AR,A - AR,B  0.025000 0.00518 19   4.824 0.0030
##  AR,A - BR,B -0.013333 0.00518 19  -2.573 0.2600
```

```
##  AR,A - UK,B  0.057333 0.00518 19  11.064 <.0001
##  AR,A - AR,C  0.025667 0.00485 19   5.295 0.0011
##  AR,A - BR,C -0.065333 0.00518 19 -12.608 <.0001
##  AR,A - UK,C  0.007333 0.00518 19   1.415 0.8786
##  BR,A - UK,A  0.068667 0.00518 19  13.251 <.0001
##  BR,A - AR,B  0.062333 0.00518 19  12.029 <.0001
##  BR,A - BR,B  0.024000 0.00518 19   4.631 0.0045
##  BR,A - UK,B  0.094667 0.00518 19  18.268 <.0001
##  BR,A - AR,C  0.063000 0.00485 19  12.997 <.0001
##  BR,A - BR,C -0.028000 0.00518 19  -5.403 0.0009
##  BR,A - UK,C  0.044667 0.00518 19   8.619 <.0001
##  UK,A - AR,B -0.006333 0.00518 19  -1.222 0.9414
##  UK,A - BR,B -0.044667 0.00518 19  -8.619 <.0001
##  UK,A - UK,B  0.026000 0.00518 19   5.017 0.0020
##  UK,A - AR,C -0.005667 0.00485 19  -1.169 0.9539
##  UK,A - BR,C -0.096667 0.00518 19 -18.654 <.0001
##  UK,A - UK,C -0.024000 0.00518 19  -4.631 0.0045
##  AR,B - BR,B -0.038333 0.00518 19  -7.397 <.0001
##  AR,B - UK,B  0.032333 0.00518 19   6.239 0.0002
##  AR,B - AR,C  0.000667 0.00485 19   0.138 1.0000
##  AR,B - BR,C -0.090333 0.00518 19 -17.432 <.0001
##  AR,B - UK,C -0.017667 0.00518 19  -3.409 0.0577
##  BR,B - UK,B  0.070667 0.00518 19  13.637 <.0001
##  BR,B - AR,C  0.039000 0.00485 19   8.046 <.0001
##  BR,B - BR,C -0.052000 0.00518 19 -10.035 <.0001
##  BR,B - UK,C  0.020667 0.00518 19   3.988 0.0177
##  UK,B - AR,C -0.031667 0.00485 19  -6.533 0.0001
##  UK,B - BR,C -0.122667 0.00518 19 -23.671 <.0001
##  UK,B - UK,C -0.050000 0.00518 19  -9.649 <.0001
##  AR,C - BR,C -0.091000 0.00485 19 -18.773 <.0001
##  AR,C - UK,C -0.018333 0.00485 19  -3.782 0.0272
##  BR,C - UK,C  0.072667 0.00518 19  14.023 <.0001
##
## P value adjustment: tukey method for comparing a family of 9 estimates
```

```r
# Gráfico de interação com o pacote phia
#O pacote phia pode ser usado para criar gráficos de interação.
library(phia)

# Divide o dispositivo gráfico em 04 gráficos por janela
par(mfrow=c(2,2))
IM <- interactionMeans(modelo)

### Gráfico de interação
plot(IM)
```

```
### Retorne o dispositivo gráfico ao seu estado original
#de um gráfico por janela
par(mfrow=c(1,1))
```

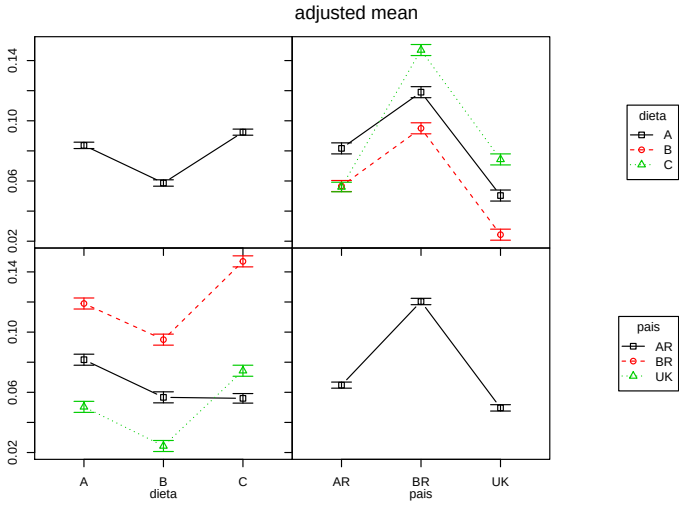


Figure 12.8: gráfico de interação

12.1.6.8 Indo além: outros tipos de ANOVA

Existem muitos outros tipos de ANOVA. Como por exemplo:

1. Experimentos em parcelas subdivididas (split-plot)
2. ANCOVA - Análise de Covariância
3. MANOVA - Análise multivariada da variância
4. ANOVA com medidas repetidas

Todos esses modelos são utilizados em estatística experimental. Um bom livro em R sobre os outros modelos ANOVAs pode ser encontrado aqui.

Um livro de referência para essa área é o livro do (Montgomery, 2019) chamado *Design and Analysis of Experiments*. Vou apenas apresentar os comandos para a realização desses modelos.

```r
#-------------------------------------------------------------
### Banco de dados
data(iris)

### Verifique os dados
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```r
# ANCOVA - Análise de Covariância
# duas variáveis independentes: uma quantitativa e uma qualitativa
fit <- aov(Sepal.Length~Petal.Width+Species, data=iris)
summary(fit)
```

```
##              Df Sum Sq Mean Sq F value               Pr(>F)
## Petal.Width   1  68.35   68.35 295.427 <0.0000000000000002 ***
## Species       2   0.03    0.02   0.075                0.928
## Residuals   146  33.78    0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# MANOVA - Análise multivariada da variância
# 3 Variáveis dependentes e uma independente
Y <- cbind(iris$Sepal.Length,iris$Sepal.Width,iris$Petal.Length)
fit <- manova(Y ~ Species, data=iris)
summary(fit)
```

```
##            Df Pillai approx F num Df den Df                Pr(>F)
## Species     2 1.1325   63.532      6    292 < 0.00000000000000022 ***
## Residuals 147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12.2 Abordagem não-paramétrica

Quando os pressupostos (normalidade ou homogêneidade de variâncias) são violados, é comum procurar um método não-paramétrico. A tabela a seguir mostra os equivalentes não-paramétricos para os procedimentos paramétricos. Em seguida, vamos descrever cada um dos métodos

Paramétrico	Não.Paramétrico
Teste t, Teste-t de Welch	Mann-Whitney (Wilcoxon)
Teste t pareado	Wilcoxon pareado
ANOVA, ANOVA de Welch	Teste Kruskal–Wallis
ANOVA em Blocos	Teste de Friedman
ANOVA com dois fatores: desenho fatorial	Aligned Ranks Transformation ANOVA

Figure 12.9: Comparação de abordagens

12.2.1 Teste U de Mann-Whitney (Wilcoxon) de duas amostras

Esse seguimento é uma tentativa de tradução do livro Mangiafico (2015). Se você quiser ler o documento original, clique aqui.

12.2.1.1 Quando usar este teste

O teste U de Mann-Whitney (Wilcoxon) de duas amostras é tipicamente um teste de igualdade estocástica entre duas distribuições de dados. Esse teste é baseado em classificação que compara valores para dois grupos. Um resultado significativo sugere que os valores para os dois grupos são diferentes.

Sem outras suposições sobre a distribuição dos dados, o teste de Mann-Whitney não aborda hipóteses sobre as medianas dos grupos. Em vez disso, o teste aborda se é provável que uma observação em um grupo seja maior que uma observação no outro. Às vezes, isso é afirmado como teste se uma amostra possui dominância estocástica em comparação com a outra.

O teste assume que as observações são independentes. Ou seja, não é apropriado para séries temporais, dados emparelhados ou dados de medidas repetidas.

12.2.1.2 Dados apropriados

Dados de duas amostras. Ou seja, dados com apenas dois grupos.

- A variável dependente é quantitativa
- A variável independente é qualitativa com dois níveis. Ou seja, dois grupos.
- As observações entre os grupos são independentes. Evite séries temporais e dados espaciais (geralmente possuem autocorrelação temporal e espacial).

12.2.1.3 Hipóteses

• Hipótese nula: os dois grupos são amostrados de populações com distribuições idênticas.
• Hipótese alternativa: os dois grupos são amostrados de populações com distribuições diferentes.

12.2.1.4 Interpretação

Resultados significativos podem ser relatados como: "Os valores da variável resposta do grupo A foram significativamente diferentes dos do grupo B."

12.2.1.5 Outras notas e testes alternativos

O teste U de Mann-Whitney (Wilcoxon) pode ser considerado equivalente ao teste de Kruskal-Wallis com apenas dois grupos.

O teste de mediana de Mood compara as medianas de dois grupos. Todavia, o teste de de Mann-Whitney (Wilcoxon) é considerado mais poderoso que o teste de mediana de Mood.

12.2.1.6 Exemplo de teste de Mann-Whitney (Wilcoxon) para duas amostras

Este exemplo apresenta os dados das palestras motivacionais das famílias Stark e Targaryen do *Game of Thrones.*Após a palestra foi distribuido um questionário para avaliar o palestrante com uma nota de 1 até 5.

Responde à pergunta: "As pontuações dos Stark são significativamente diferentes da família Targaryen?"

O teste de Mann-Whitney (Wilcoxon) é realizado com a função *wilcox.test*, que produz um p-valor para a hipótese.

Primeiro, os dados são resumidos e examinados usando gráficos de barras para cada grupo.

```
#---------------------------------------------------------------
### Banco de dados
Entrada <- ("
 Palestrante Nota
 Stark 3
 Stark 5
 Stark 4
 Stark 4
 Stark 4
 Stark 4
 Stark 4
 Stark 4
 Stark 5
 Stark 5
 Targaryen 2
 Targaryen 4
 Targaryen 2
 Targaryen 2
 Targaryen 1
 Targaryen 2
 Targaryen 3
 Targaryen 2
 Targaryen 2
 Targaryen 3
")

Dados <- read.table(textConnection(Entrada), head = TRUE)

#### Crie uma nova variável que seja a nota da avaliação
# da palestra como um fator ordenado
Dados$Avalicao = as.factor(Dados$Nota)

### Verifique os dados
str(Dados)
```

```
## 'data.frame':    20 obs. of  3 variables:
##  $ Palestrante: Factor w/ 2 levels "Stark","Targaryen": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Nota       : int  3 5 4 4 4 4 4 4 5 5 ...
##  $ Avalicao   : Factor w/ 5 levels "1","2","3","4",..: 3 5 4 4 4 4 4 4 5 5 ...
```

```
summary(Dados)
```

```
##     Palestrante      Nota      Avalicao
```

```
##  Stark    :10   Min.   :1.00   1:1
##  Targaryen:10   1st Qu.:2.00   2:6
##                 Median :3.50   3:3
##                 Mean   :3.25   4:7
##                 3rd Qu.:4.00   5:3
##                 Max.   :5.00
```

```
### Remova objetos desnecessários
rm(Entrada)
```

Observe que a variável que queremos observar é Avaliação. As contagens para Avaliacao são tabuladas cruzadamente sobre os valores de Palestrante. A função prop.table converte uma tabela em proporções. A opção margem = 1 indica que as proporções são calculadas para cada linha.

```
#---------------------------------------------------------------
#### Resumo dos dados tratando as pontuações como categórico

# Numero Absoluto
tabela <- table(Dados$Palestrante,Dados$Avalicao)
tabela
```

```
##
##             1 2 3 4 5
##   Stark     0 0 1 6 3
##   Targaryen 1 6 2 1 0
```

```
#proporcao
prop.table(tabela,margin = 1)*100
```

```
##
##              1  2  3  4  5
##   Stark      0  0 10 60 30
##   Targaryen 10 60 20 10  0
```

```
# As duas medianas
library(psych)
describeBy(Dados$Nota, group = Dados$Palestrante)
```

```
##
```

```
##  Descriptive statistics by group
## group: Stark
##    vars  n mean   sd median trimmed mad min max range  skew kurtosis  se
## X1    1 10  4.2 0.63      4    4.25   0   3   5     2 -0.09    -0.93 0.2
## ------------------------------------------------------------
## group: Targaryen
##    vars  n mean   sd median trimmed mad min max range skew kurtosis   se
## X1    1 10  2.3 0.82      2    2.25   0   1   4     3 0.58    -0.45 0.26
```

12.2.1.7 Exemplo de teste de Mann-Whitney (Wilcoxon) de duas amostras

Este exemplo usa a notação de fórmula indicando que avaliação é a variável dependente e Palestrante é a variável independente. A opção data = indica o banco de dados que contém as variáveis. Para o significado de outras opções, consulte *?Wilcox.test*.

```
#------------------------------------------------------------

wilcox.test(Nota ~ Palestrante, data = Dados)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Nota by Palestrante
## W = 95, p-value = 0.0004713
## alternative hypothesis: true location shift is not equal to 0
```

12.2.2 Teste de Kruskal-Wallis

O teste de Kruskal-Wallis é um teste semelhante ao teste de Mann-Whitney (Wilcoxon), mas pode ser aplicado a variáveis categóricas com mais de dois grupos.

Sem outras suposições sobre a distribuição dos dados, o teste de Kruskal-Wallis não aborda hipóteses sobre as medianas dos grupos. Em vez disso, o teste aborda se é provável que uma observação em um grupo seja maior que uma observação no outro.

O teste assume que as observações são independentes. Ou seja, não é apropriado para observações emparelhadas ou dados de medidas repetidas. É realizado com a função *kruskal.test*.

12.2.2.1 Testes post-hoc

O resultado do teste de Kruskal-Wallis informa se existem diferenças entre os grupos, mas não informa quais grupos são diferentes de outros grupos. Para determinar quais grupos são diferentes dos outros, é possível realizar testes post-hoc.

12.2.2.2 Dados apropriados

- A variável dependente é ordinal, discreta ou continua.
- Variável independente é um fator com três ou mais níveis. Ou seja, três ou mais grupos.
- As observações entre os grupos são independentes.

12.2.2.3 Hipóteses

• Hipótese nula: os grupos são amostrados de populações com distribuições idênticas. • Hipótese alternativa: os grupos são amostrados de populações com diferentes distribuições.

12.2.2.4 Interpretação

Resultados significativos podem ser relatados como "Houve uma diferença significativa nos valores entre os grupos". Os grupos tem distribuições diferentes.

A análise post-hoc permite dizer "Houve uma diferença significativa nos valores entre os grupos A e B." e assim por diante.

12.2.2.5 Exemplo de teste de Kruskal–Wallis

Este exemplo apresenta os dados das palestras motivacionais das famílias Stark, Targaryen e Lannister do *Game of Thrones*. Após a palestra foi distribuido um questionário para avaliar o palestrante com uma nota de 1 até 5. Esse teste responde à pergunta: "As pontuações são significativamente diferentes entre os três palestrantes?"

O teste de Kruskal-Wallis é realizado com a função *kruskal.test*, que produz um p-valor para a hipótese. Primeiro, os dados são resumidos e examinados usando gráficos de barras para cada grupo.

```
#------------------------------------------------------

Entrada =("
```

```
 Palestrante Nota
 Stark      3
 Stark      5
 Stark      4
 Stark      4
 Stark      4
 Stark      4
 Stark      4
 Stark      4
 Stark      5
 Stark      5
 Targaryen    2
 Targaryen    4
 Targaryen    2
 Targaryen    2
 Targaryen    1
 Targaryen    2
 Targaryen    3
 Targaryen    2
 Targaryen    2
 Targaryen    3
 Lannister    4
 Lannister    4
 Lannister    4
 Lannister    4
 Lannister    5
 Lannister    3
 Lannister    5
 Lannister    4
 Lannister    4
 Lannister    3
")

Dados <- read.table(textConnection(Entrada), head = TRUE)

#### Crie uma nova variável que seja a categórica
Dados$Avaliacao <- as.factor(Dados$Nota)

### Verifique os dados
str(Dados)
```

```
## 'data.frame':    30 obs. of  3 variables:
##  $ Palestrante: Factor w/ 3 levels "Lannister","Stark",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ Nota       : int  3 5 4 4 4 4 4 4 5 5 ...
```

```
##  $ Avaliacao  : Factor w/ 5 levels "1","2","3","4",..: 3 5 4 4 4 4 4 4 5 5 ...
```

```
summary(Dados)
```

```
##     Palestrante      Nota     Avaliacao
##  Lannister:10   Min.   :1.0   1: 1
##  Stark    :10   1st Qu.:3.0   2: 6
##  Targaryen:10   Median :4.0   3: 5
##                 Mean   :3.5   4:13
##                 3rd Qu.:4.0   5: 5
##                 Max.   :5.0
```

```
### Remova objetos desnecessários
rm(Entrada)

#### Resumo dos dados tratando as pontuações como categóricas
# Numero Absoluto
tabela <- table(Dados$Palestrante,Dados$Avaliacao)
tabela
```

```
##
##             1 2 3 4 5
##   Lannister 0 0 2 6 2
##   Stark     0 0 1 6 3
##   Targaryen 1 6 2 1 0
```

```
#proporcao
prop.table(tabela,margin = 1)*100
```

```
##
##              1  2  3  4  5
##   Lannister  0  0 20 60 20
##   Stark      0  0 10 60 30
##   Targaryen 10 60 20 10  0
```

```
# As três medianas
library(psych)
describeBy(Dados$Nota, group = Dados$Palestrante)
```

```
##
##  Descriptive statistics by group
## group: Lannister
##    vars  n mean   sd median trimmed mad min max range skew kurtosis   se
## X1    1 10    4 0.67      4       4   0   3   5     2    0    -0.97 0.21
## ------------------------------------------------------------
## group: Stark
##    vars  n mean   sd median trimmed mad min max range  skew kurtosis  se
## X1    1 10  4.2 0.63      4    4.25   0   3   5     2 -0.09    -0.93 0.2
## ------------------------------------------------------------
## group: Targaryen
##    vars  n mean   sd median trimmed mad min max range skew kurtosis   se
## X1    1 10  2.3 0.82      2    2.25   0   1   4     3 0.58    -0.45 0.26
```

Este exemplo usa a notação de fórmula indicando que Nota é a variável dependente e Palestrante é a variável independente. A opção data = indica o quadro de dados que contém as variáveis. Para o significado de outras opções, consulte *?kruskal.test*.

```
#------------------------------------------------------------
kruskal.test(Nota ~ Palestrante, data = Dados)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Nota by Palestrante
## Kruskal-Wallis chi-squared = 16.842, df = 2, p-value = 0.0002202
```

12.2.2.6 Teste post-hoc: testes de Mann-Whitney para comparações múltiplas

A função *pairwise.wilcox.test* produz uma tabela de p-valor comparando cada par de grupos. Para evitar a inflação das taxas de erro do tipo I, é possível fazer ajustes no p-valor usando a opção *p.adjust.method*. O método *fdr* é usado. Consulte *?p.adjust* para obter detalhes sobre os métodos de ajuste de p-valor disponíveis. O código cria uma matriz de p-valor.

12.2.2.7 Pairwise Mann–Whitney

```
#------------------------------------------------------------
```

```
PMW <- pairwise.wilcox.test(Dados$Nota,
                            Dados$Palestrante,
                            p.adjust.method="fdr")
# Adjusts p-values for multiple comparisons;
# See ?p.adjust for options

PMW
```

```
##
##  Pairwise comparisons using Wilcoxon rank sum test
##
## data:  Dados$Nota and Dados$Palestrante
##
##           Lannister Stark
## Stark     0.5174    -
## Targaryen 0.0012    0.0012
##
## P value adjustment method: fdr
```

O p-valor da diferença entre Stark e Lannister é 0,5174. Isso indica que não existe diferença entre as distribuições de notas dessas duas casas.

12.2.3 Teste de Friedman

O teste de Friedman determina se há diferenças entre os grupos em um *design* de blocos. Nesse projeto, uma variável serve como variável de tratamento ou grupo e outra variável serve como variável de bloco. É nas diferenças entre tratamentos ou grupos que estamos interessados. Não estamos interessados nas diferenças entre os blocos, mas queremos que nossas estatísticas levem em conta as diferenças nos blocos. O teste de Friedman é muitas vezes utilizado como uma análise de variância não-paramétrica com uma variável de bloco. Ou seja, é uma versão não-paramétrica de uma ANOVA com Blocos Aleatorizados.

Isso significa que, embora a ANOVA tem o pressuposto de uma distribuição normal e variâncias iguais (dos resíduos), o teste de Friedman está livre dessas restrições. O preço dessa liberdade paramétrica é a perda de potência (do teste de Friedman em comparação com as versões paramétricas da ANOVA).

Para um exemplo dessa estrutura, veja os dados da família Belcher abaixo.

O avaliador é considerado a variável de bloqueio, e cada avaliador tem uma observação para cada instrutor. O teste determinará se há diferenças entre os valores para o Instrutor, levando em consideração qualquer efeito consistente de um Avaliador.

Em outros casos, a variável de bloco pode ser a turma em que as classificações foram feitas ou a escola em que as classificações foram realizadas. Se você estava testando diferenças entre currículos ou outros tratamentos de ensino com diferentes instrutores, diferentes instrutores podem ser usados como blocos.

O teste de Friedman determina se há uma diferença sistemática nos valores da variável dependente entre os grupos.

12.2.3.1 Testes post-hoc

O resultado do teste de Friedman informa se existem diferenças entre os grupos, mas não informa quais grupos são diferentes de outros grupos. Para determinar quais grupos são diferentes dos outros, é possível realizar testes post-hoc.

12.2.3.2 Dados apropriados

- Dados organizados em um design de bloco completo não replicado
- A variável dependente é ordinal, discreta ou contínua
- A variável independente de tratamento ou grupo é categórica com dois ou mais níveis. Ou seja, dois ou mais grupos.
- Variável de bloqueio é categórica com dois ou mais níveis.
- Os blocos são independentes um do outro e não têm interação com tratamentos.

12.2.3.3 Hipóteses

Hipótese nula: as distribuições (sejam elas quais forem) são as mesmas para cada grupo nos blocos. **Hipótese alternativa**: As distribuições para cada grupo entre os blocos são diferentes.

12.2.3.4 Interpretação

Resultados significativos podem ser relatados como “Houve uma diferença significativa nos valores da variável resposta entre os grupos”.

12.2.3.5 Outras anotações e testes alternativos

O teste Quade é usado para os mesmos tipos de dados e hipóteses, mas pode ser mais poderoso em alguns casos. Foi sugerido que o teste de Friedman pode ser preferível quando há um número maior de grupos (cinco ou mais), enquanto o Quade é preferível para menos grupos.

12.2.3.6 Exemplo do teste de Friedman

```
#---------------------------------------------------------------

Entrada =("
 Instrutor        Avaliador  Nota
 'Bob Belcher'        a       4
 'Bob Belcher'        b       5
 'Bob Belcher'        c       4
 'Bob Belcher'        d       6
 'Bob Belcher'        e       6
 'Bob Belcher'        f       6
 'Bob Belcher'        g      10
 'Bob Belcher'        h       6
 'Linda Belcher'      a       8
 'Linda Belcher'      b       6
 'Linda Belcher'      c       8
 'Linda Belcher'      d       8
 'Linda Belcher'      e       8
 'Linda Belcher'      f       7
 'Linda Belcher'      g      10
 'Linda Belcher'      h       9
 'Tina Belcher'       a       7
 'Tina Belcher'       b       5
 'Tina Belcher'       c       7
 'Tina Belcher'       d       8
 'Tina Belcher'       e       8
 'Tina Belcher'       f       9
 'Tina Belcher'       g      10
 'Tina Belcher'       h       9
 'Gene Belcher'       a       6
 'Gene Belcher'       b       4
 'Gene Belcher'       c       5
 'Gene Belcher'       d       5
 'Gene Belcher'       e       6
 'Gene Belcher'       f       6
 'Gene Belcher'       g       5
 'Gene Belcher'       h       5
 'Louise Belcher'     a       8
 'Louise Belcher'     b       7
 'Louise Belcher'     c       8
 'Louise Belcher'     d       8
 'Louise Belcher'     e       9
 'Louise Belcher'     f       9
```

```
 'Louise Belcher'     g      8
 'Louise Belcher'     h     10
")

Dados = read.table(textConnection(Entrada),header=TRUE)

Dados$Nota_cat = as.factor(Dados$Nota)

### Verifique os dados
str(Dados)
```

```
## 'data.frame':    40 obs. of  4 variables:
##  $ Instrutor: Factor w/ 5 levels "Bob Belcher",..: 1 1 1 1 1 1 1 1 3 3 ...
##  $ Avaliador: Factor w/ 8 levels "a","b","c","d",..: 1 2 3 4 5 6 7 8 1 2 ...
##  $ Nota     : int  4 5 4 6 6 6 10 6 8 6 ...
##  $ Nota_cat : Factor w/ 7 levels "4","5","6","7",..: 1 2 1 3 3 3 7 3 5 3 ...
```

```
summary(Dados)
```

```
##           Instrutor   Avaliador       Nota        Nota_cat
##  Bob Belcher   :8   a      : 5   Min.   : 4.000   4 : 3
##  Gene Belcher  :8   b      : 5   1st Qu.: 6.000   5 : 6
##  Linda Belcher :8   c      : 5   Median : 7.000   6 : 8
##  Louise Belcher:8   d      : 5   Mean   : 7.075   7 : 4
##  Tina Belcher  :8   e      : 5   3rd Qu.: 8.000   8 :10
##                     f      : 5   Max.   :10.000   9 : 5
##                     (Other):10                    10: 4
```

```
### Remova objetos desnecessários
rm(Entrada)

### Teste de Friedman
friedman.test(Nota ~ Instrutor | Avaliador,
              data = Dados)
```

```
##
##  Friedman rank sum test
##
## data:  Nota and Instrutor and Avaliador
## Friedman chi-squared = 23.139, df = 4, p-value = 0.0001188
```

```
# Comparações pareadas usando o teste de Conover para um projeto de bloco completo balanceado
library(PMCMR)
PFCT <- posthoc.friedman.conover.test(y    = Dados$Nota,
                                      groups = Dados$Instrutor,
                                      blocks = Dados$Avaliador,
                                      p.adjust.method="fdr")
# P-valor para comparações pareadas
PFCT
```

```
##
##  Pairwise comparisons using Conover's test for a two-way
##                      balanced complete block design
##
## data:  Dados$Nota , Dados$Instrutor and Dados$Avaliador
##
##                Bob Belcher Gene Belcher Linda Belcher Louise Belcher
## Gene Belcher   0.17328     -            -             -
## Linda Belcher  0.00002796  0.00000117   -             -
## Louise Belcher 0.00000190  0.00000011   0.27303       -
## Tina Belcher   0.00037     0.00000877   0.31821       0.05154
##
## P value adjustment method: fdr
```

12.2.4 ANOVA de transformação de postos alinhados

Esse seguimento é uma tentativa de tradução do livro Mangiafico (2015). Se você quiser ler o documento original, clique aqui.

A ANOVA de transformação de postos alinhados - ART-ANOVA (Aligned Ranks Transformation ANOVA) é uma abordagem não paramétrica que permite múltiplas variáveis independentes, interações e medidas repetidas.

Neste modelo a variável dependente precisa ter uma natureza quantitativa. Ou seja,no processo, os dados ordinais precisariam ser tratados como numéricos. O pacote ARTool torna o uso dessa abordagem no R relativamente fácil.

Algumas notas sobre o uso do ARTool:
• Todas as variáveis independentes devem ser nominais.
• Todas as interações de variáveis independentes fixas precisam ser incluídas no modelo
• Comparações post-hoc podem ser realizadas para os efeitos principais.
• Para efeitos de interações, comparações post-hoc podem ser realizadas para modelos com duas variáveis independentes. • Para modelos de efeitos fixos, o eta-quadrado pode ser calculado como um tamanho de efeito.

```
#-----------------------------------------------------------------

### Banco de dados
Location = c(rep("Olympia" , 6), rep("Ventura", 6),
             rep("Northampton", 6), rep("Burlington", 6))
Tribe  = c(rep(c("Jedi", "Sith"), 12))
Midichlorians = c(10,  4, 12,  5, 15,  4, 15,  9, 15, 11, 18, 12,
                  8, 13,  8, 15, 10, 17, 22, 22, 20, 22, 20, 25)
Dados <- data.frame(Tribe, Location, Midichlorians)
str(Dados)
```

```
## 'data.frame':    24 obs. of  3 variables:
##  $ Tribe        : Factor w/ 2 levels "Jedi","Sith": 1 2 1 2 1 2 1 2 1 2 ...
##  $ Location     : Factor w/ 4 levels "Burlington","Northampton",..: 3 3 3 3 3 3 4 4
##  $ Midichlorians: num  10 4 12 5 15 4 15 9 15 11 ...
```

```
### Remova objetos desnecessários
remove(Location,Tribe,Midichlorians)

### Aligned ranks anova
library(ARTool)
modelo <- art(Midichlorians ~ Tribe + Location + Tribe:Location,
              data = Dados)

### Art ANOVA
anova(modelo)
```

```
## Analysis of Variance of Aligned Rank Transformed Data
##
## Table Type: Anova Table (Type III tests)
## Model: No Repeated Measures (lm)
## Response: art(Midichlorians)
##
##                  Df Df.res F value        Pr(>F)
## 1 Tribe           1     16  3.0606      0.099364   .
## 2 Location        3     16 34.6201 0.00000031598 ***
## 3 Tribe:Location  3     16 29.9354 0.00000084929 ***
## ---
## Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

12.2.4.1 Comparações post-hoc para efeitos principais

As comparações post-hoc para efeitos principais podem ser tratadas pelo pacote *emmeans* da maneira usual, exceto que a função artlm deve ser usada para ajustar primeiro um modelo que pode ser passado para emmeans. Os valores estimados na saída emmeans devem ser ignorados.

```
#------------------------------------------------------------

library(emmeans)
modelo.lm = artlm(modelo, "Location")
marginal = emmeans(modelo.lm,~ Location)
pairs(marginal,adjust = "tukey")
```

```
##  contrast                 estimate   SE df t.ratio p.value
##  Burlington - Northampton    10.83 1.78 16  6.075  0.0001
##  Burlington - Olympia        17.83 1.78 16 10.000  <.0001
##  Burlington - Ventura         7.33 1.78 16  4.112  0.0041
##  Northampton - Olympia        7.00 1.78 16  3.925  0.0060
##  Northampton - Ventura       -3.50 1.78 16 -1.963  0.2426
##  Olympia - Ventura          -10.50 1.78 16 -5.888  0.0001
##
## Results are averaged over the levels of: Tribe
## P value adjustment: tukey method for comparing a family of 4 estimates
```

```
#Comparações post-hoc para interações
#Os valores estimados na saída emmeans devem ser ignorados.

modelo.int = artlm(modelo, "Tribe:Location")
marginal = emmeans(modelo.int, ~ Tribe:Location)
contrast(marginal, method="pairwise", adjust="none")
```

```
##  contrast                           estimate   SE df t.ratio p.value
##  Jedi,Burlington - Sith,Burlington    -6.500 2.68 16 -2.428  0.0273
##  Jedi,Burlington - Jedi,Northampton    5.667 2.68 16  2.117  0.0503
##  Jedi,Burlington - Sith,Northampton  -11.167 2.68 16 -4.171  0.0007
##  Jedi,Burlington - Jedi,Olympia      -10.333 2.68 16 -3.860  0.0014
##  Jedi,Burlington - Sith,Olympia        4.667 2.68 16  1.743  0.1005
##  Jedi,Burlington - Jedi,Ventura       -8.667 2.68 16 -3.237  0.0052
##  Jedi,Burlington - Sith,Ventura        1.000 2.68 16  0.374  0.7136
##  Sith,Burlington - Jedi,Northampton   12.167 2.68 16  4.545  0.0003
##  Sith,Burlington - Sith,Northampton   -4.667 2.68 16 -1.743  0.1005
##  Sith,Burlington - Jedi,Olympia       -3.833 2.68 16 -1.432  0.1714
```

```
##  Sith,Burlington - Sith,Olympia        11.167 2.68 16  4.171  0.0007
##  Sith,Burlington - Jedi,Ventura        -2.167 2.68 16 -0.809  0.4302
##  Sith,Burlington - Sith,Ventura         7.500 2.68 16  2.802  0.0128
##  Jedi,Northampton - Sith,Northampton  -16.833 2.68 16 -6.288  <.0001
##  Jedi,Northampton - Jedi,Olympia      -16.000 2.68 16 -5.977  <.0001
##  Jedi,Northampton - Sith,Olympia       -1.000 2.68 16 -0.374  0.7136
##  Jedi,Northampton - Jedi,Ventura      -14.333 2.68 16 -5.354  0.0001
##  Jedi,Northampton - Sith,Ventura       -4.667 2.68 16 -1.743  0.1005
##  Sith,Northampton - Jedi,Olympia        0.833 2.68 16  0.311  0.7596
##  Sith,Northampton - Sith,Olympia       15.833 2.68 16  5.914  <.0001
##  Sith,Northampton - Jedi,Ventura        2.500 2.68 16  0.934  0.3643
##  Sith,Northampton - Sith,Ventura       12.167 2.68 16  4.545  0.0003
##  Jedi,Olympia - Sith,Olympia           15.000 2.68 16  5.603  <.0001
##  Jedi,Olympia - Jedi,Ventura            1.667 2.68 16  0.623  0.5423
##  Jedi,Olympia - Sith,Ventura           11.333 2.68 16  4.233  0.0006
##  Sith,Olympia - Jedi,Ventura          -13.333 2.68 16 -4.981  0.0001
##  Sith,Olympia - Sith,Ventura           -3.667 2.68 16 -1.370  0.1897
##  Jedi,Ventura - Sith,Ventura            9.667 2.68 16  3.611  0.0023
```

```
#Comparações post-hoc para outras interações
#Os valores estimados na saída emmeans devem ser ignorados.

modelo.diff = artlm(modelo, "Tribe:Location")
marginal = emmeans(modelo.diff, ~Tribe:Location)
contrast(marginal, method="pairwise", interaction=TRUE)
```

```
##  Tribe_pairwise Location_pairwise        estimate   SE df t.ratio p.value
##  Jedi - Sith    Burlington - Northampton    10.33 3.79 16  2.729  0.0148
##  Jedi - Sith    Burlington - Olympia      -21.50 3.79 16 -5.679  <.0001
##  Jedi - Sith    Burlington - Ventura      -16.17 3.79 16 -4.270  0.0006
##  Jedi - Sith    Northampton - Olympia     -31.83 3.79 16 -8.408  <.0001
##  Jedi - Sith    Northampton - Ventura     -26.50 3.79 16 -7.000  <.0001
##  Jedi - Sith    Olympia - Ventura           5.33 3.79 16  1.409  0.1781
```

12.2.4.1.1 Eta quadrado parcial O eta-quadrado parcial pode ser calculado como uma estatística de tamanho de efeito para a Art ANOVA.

12.2.4.1.2 Interpretação de eta-quadrado A interpretação dos tamanhos dos efeitos varia necessariamente de acordo com a disciplina e as expectativas do experimento, mas, para estudos comportamentais, são seguidas as diretrizes propostas por Cohen (1988). Elas não devem ser consideradas universais.

```
Small         Medium    Large
```

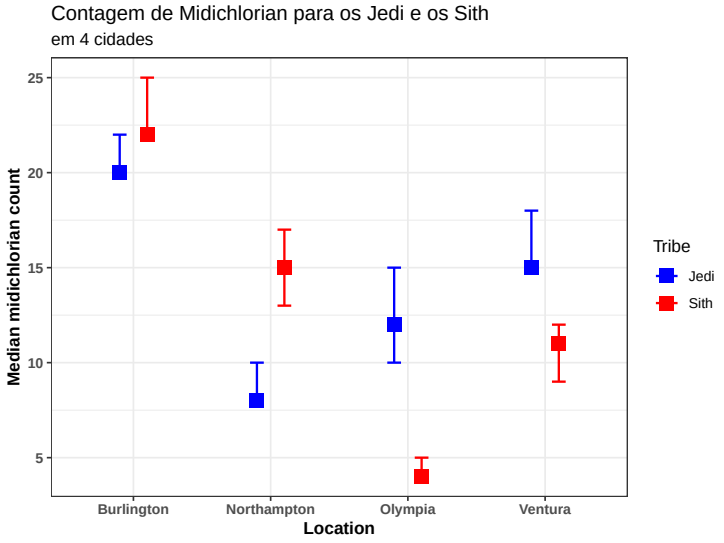
eta-squared 0.01–< 0.06 0.06–< 0.14 0.14 Source: Cohen (1988).

```
#-------------------------------------------------------------
Resultado <- anova(modelo)

Resultado$eta.quadrado.parcial <- with(Resultado, `Sum Sq`/(`Sum Sq` + `Sum Sq.res`))

Resultado
```

```
## Analysis of Variance of Aligned Rank Transformed Data
##
## Table Type: Anova Table (Type III tests)
## Model: No Repeated Measures (lm)
## Response: art(Midichlorians)
##
##                  Df Df.res F value        Pr(>F) eta.quadrado.parcial
## 1 Tribe           1     16  3.0606      0.099364              0.16057   .
## 2 Location        3     16 34.6201 0.00000031598              0.86651 ***
## 3 Tribe:Location  3     16 29.9354 0.00000084929              0.84878 ***
## ---
## Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Referências - parte 1

- Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences, 2nd Edition. Routledge.

- Kay, M. 2019. Contrast tests with ART. cran.r-project.org/web/packages/ARTool/vignettes/art-contrasts.html.

- Kay, M. 2019. Effect Sizes with ART. https://cran.r-project.org/web/packages/ARTool/vignettes/art-effect-size.html.

- Kay, M. 2019. Package 'ARTool'. cran.r-project.org/web/packages/ARTool/ARTool.pdf.

- Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In Conference on Human Factors in Computing Systems (pp. 143–146). faculty.washington.edu/wobbrock/pubs/chi-11.06.pdf.

- Wobbrock, J. O., Findlater, L., Gergle, D., Higgins, J. J., & Kay, M. 2018. ARTool: Align-and-rank data for a nonparametric ANOVA. University of Washington. depts.washington.edu/madlab/proj/art/index.html.

Chapter 13

Associação entre duas variáveis quantitativas

13.1 Teste de correlação de pearson

O Teste de correlação de pearson faz a mensuração da associação linear entre duas variáveis quantitativas. Além disso, ele testa se essa correlaçao linear é significativa.

13.1.0.1 Dados apropriados

Utiliza-se este teste quando:
* Os dados seguem uma distribuição normal bi-variada.

13.1.0.2 Hipóteses

Hipótese nula: $\rho = 0$
Hipótese alternativa: $\rho \neq 0$

13.1.0.3 Interpretação

Resultados significativos podem ser relatados como “A correlação é diferente de zero. As variáveis são correlacionadas”.

13.1.0.4 Exemplo

13.1.0.5 Exemplo do teste de correlação

Este exemplo apresenta os dados de uma amostra de 32 carros.
Esse teste responde à pergunta: "A variável Km/L (quantitativa) é correlacionada com o log do preço (quantitativa)"

```
#------------------------------------------------------------
### Banco de dados
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")
CARROS$log_preco<-log(CARROS$Preco)
### Verifique os dados
str(CARROS$Kmporlitro)
```

```
##  num [1:32] 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
```

```
str(CARROS$log_preco)
```

```
##  num [1:32] 5.08 5.08 4.68 5.55 5.89 ...
```

```
### Remova objetos desnecessários
remove(mtcars)

#### Resumo dos dados
library(ggplot2)
ggplot(CARROS) +
  aes(x = Kmporlitro, y = log_preco) +
  geom_point(size = 3L, colour = "#cb181d") +
  theme_minimal()
```

```
#### Teste de correlação
cor.test(CARROS$log_preco,CARROS$Kmporlitro)
```

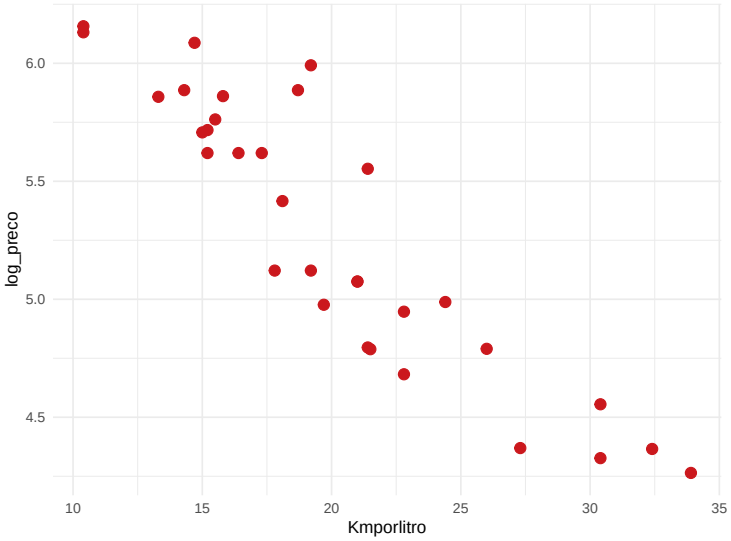


Figure 13.1: Diagrama de dispersão

```
##
##  Pearson's product-moment correlation
##
## data:  CARROS$log_preco and CARROS$Kmporlitro
## t = -11.805, df = 30, p-value = 0.00000000000084
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9540391 -0.8167641
## sample estimates:
##       cor
## -0.9071119
```

13.2 Teste de correlação de spearman

O Teste de correlação de spearman também faz a mensuração da associação linear entre duas variáveis quantitativas. Todavia, neste teste, as variaveis devem ser no mínimo do tipo ordinal.

13.2.0.1 Dados apropriados

Utiliza-se este teste quando:
* Os dados **não** seguem uma distribuição normal bi-variada.

13.2.0.2 Hipóteses

Hipótese nula: Não há associação [monotônica] entre as duas variáveis.
Hipótese alternativa: Há associação [monotônica] entre as duas variáveis.

13.2.0.3 Interpretação

Resultados significativos podem ser relatados como "Há associação entre as duas variáveis".

13.2.0.4 Exemplo do teste de correlação

Este exemplo apresenta os dados de uma amostra de 32 carros.
Esse teste responde à pergunta: "A variável HP (quantitativa) é correlacionada com o preço (quantitativa)"

```
#-------------------------------------------------------------
### Banco de dados
data(mtcars)
CARROS<-mtcars
colnames(CARROS) <- c("Kmporlitro","Cilindros","Preco","HP",
                      "Amperagem_circ_eletrico","Peso","RPM",
                      "Tipodecombustivel","TipodeMarcha",
                      "NumdeMarchas","NumdeValvulas")

### Verifique os dados
str(CARROS$HP)
```

```
##  num [1:32] 110 110 93 110 175 105 245 62 95 123 ...
```

```
str(CARROS$Preco)
```

```
##  num [1:32] 160 160 108 258 360 ...
```

```
### Remova objetos desnecessários
remove(mtcars)

#### Resumo dos dados
library(ggplot2)
ggplot(CARROS) +
```

```
  aes(x = HP, y = Preco) +
  geom_point(size = 3L, colour = "darkblue") +
  theme_minimal()
```

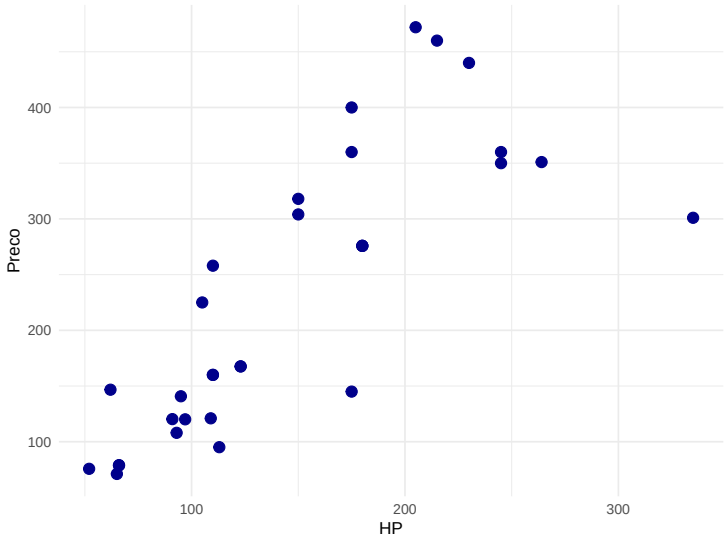


Figure 13.2: Diagrama de dispersão

```
#### Teste de correlação
cor.test(CARROS$HP,CARROS$Preco,method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  CARROS$HP and CARROS$Preco
## S = 812.71, p-value = 0.0000000006791
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.8510426
```

Em breve vamos ter um módulo 2 contendo: 1. Regressão simples e multipla
2. Diagnostico em modelos de regressão
3. Modelos lineares Generalizados

Por enquanto, sugiro olhar as apresentações:

https://dataunirio.github.io/Modelos_Lineares/

https://dataunirio.github.io/Regressao_Diagnostico/

Bibliography

Agresti, A. (2002). *Categorical data analysis, second edition.* Editora Wiley, New York.

Agresti, A. and Finlay, B. (2012). *Métodos Estatísticos para as Ciências Sociais.* Editora Penso.

Bussab, W. and Morettin, P. (2015). *Estatística Básica.* Editora Saraiva.

Mangiafico, S. (2015). *An R Companion for the Handbook of Biological Statistics, version 1.3.2.*

Montgomery, D. (2019). *Design and Analysis of Experiments.* Editora John Wiley and Sons.

Salsburg, D. (2009). *Uma senhora toma chá. Como a estatística revolucionou a ciência no século XX.* Editora Zahar.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown.* R package version 0.18.