# Estatística + R

Ana Paula Fernandes (DESCO/ICS/UFTM)

Atualizado em: 10/07/2025

# Contents

# Chapter 1

# Bem-vindos!

Esse livro *online* tem como propósito principal ser um guia para as aulas de estatística, referente as disciplinas de **Bioestatística** para os curso de Medicina e Educação Física e **Estatística Aplicada** para o curso de Psicologia da Universidade Federal do Triângulo Mineiro - UFTM. E como objetivo secundário, ser uma referência de consulta para todos os discentes que passaram por essas disciplinas, bem como, para todos que estão interessados em realizar análises de dados por meio da linguagem R e o ambiente de desenvolvimento RStudio.

Sugestões, correções ou qualquer outra forma de interação são sempre bem-vindas! Então, por favor, não hesite em me escrever (anapaula.fernandes@uftm.edu.br).

Para mais informações sobre minha trajetória acadêmica e profissional, acesse meu Currículo Lattes.

> Esta obra, registrada sob o **ISBN 978-65-01-56980-2**, traz conteúdos atualizados e essenciais para estudantes e profissionais interessados no tema.

# Chapter 2

# Introdução

Ao longo de algum tempo ministrando aulas de estatísticas conclui que estudar estatística com auxílio de recursos computacionais é bem mais eficaz, quero dizer, é mais fácil entender os conceitos téoricos, lidar com recusos visuais (gráficos) e, de fato, transformar o contéudo estudado na disciplina em uma ferramenta para pesquisas científicas, quando se trata de analisar dados.

Ministrando aulas para os cursos da área de saúde, esporte e psicologia sempre ouvi dos discentes que estatística é matemática, e sempre digo que estatística é estatística! É normal alguns discentes não assimilarem, em princípio, a importância da disciplina na grade do seu curso, e realmente, alguns acham até que é assunto que deveria ficar restrito aos curso das exatas. Assim, a primeira tarefa é sempre desconstruir essa ideia.

A estatística é **MULTIDICIPLINAR**, ela está em tudo na verdade… e para dizer uma coisa "bem chique" a estatística é a base da Inteligência Artificial. Advinha quem está por trás dos famosos algortimos das redes sociais? Ou das sugestões de filmes e músicas que aparecem no seu *streamming* favorito? Ou no ranque de busca realzada por meio do *Google*? Ou no *Chat* GTP?

E sendo um pouco mais "acadêmica", dentro do nosso propósito:

Qualquer competição ou treinamento esportivo está recheado de estatística, como medir o desempenho de um time ou atleta? Veja esse exemplo aqui:

> **Velocidade e resistência de velocidade de sprint em atletas de Futebol amador** http://www.rbff.com.br/index.php/rbff/article/view/866

Na medicina, estudos epidemiologicos, e claro, da medicina baseada em evidências, tem o suporte da estatística. Veja esse exemplo aqui:

> **Qualidade de Vida Relacionada à Saúde e Satisfação com o Tratamento Hospitalar de Adultos com Câncer: Estudo Observacional** https://rbc.inca.gov.br/index.php/revista/article/view/3554

Na psicologia a estatística é a ferramenta utilizada na psicometria. Veja esse exemplo:

> **Escala de Comportamentos Antissociais: construção e estudos psicométricos** https://periodicos.pucpr.br/psicologiaargumento/article/view/27071

Basta realizar uma busca com os termos estatística e um campo do seu curso que você se interessa, que você encontrará um artigo científico. E se você não encontrar, comece a escreve sobre o tema!

Quando olhamos os artigos acima, podemos ver que todos eles tem resultados **descritivos** e **inferenciais**. Discutiremos sobre estatística descritiva (de descrever - os dados amostrados para uma dada análise) e inferencial (de inferir - tirar conclusões a partir dos dados amostrados) no próximo tópico.

## 2.1 Atividade 1

**Busque um artigo do campo de seu interesse que utiliza a estatística.**

1. **Qual é o principal objetivo da pesquisa?**
    - Qual é a questão central que a pesquisa busca responder? Os autores apresentam claramente os objetivos do estudo e o que se espera alcançar com os resultados?
2. **Como a pesquisa foi conduzida?**
    - Quais métodos foram utilizados para coletar os dados? A pesquisa é qualitativa ou quantitativa? Os autores descrevem o procedimento de forma detalhada, incluindo a amostra, os instrumentos de coleta de dados e a análise estatística realizada?
3. **O que é apresentado por meio de tabelas ou gráficos?**
    - Quais informações estão sendo representadas visualmente? Os gráficos e tabelas são claros e bem organizados para facilitar a compreensão dos dados? Como os resultados são interpretados a partir dessas representações?
4. **Faça uma lista de termos relacionados à estatística encontrados no artigo.**
    - Identifique e liste os termos e conceitos estatísticos mencionados no artigo. Isso pode incluir testes de hipótese, modelos estatísticos, variáveis dependentes e independentes, intervalos de confiança, p-valor, etc.

Periódicos da área da Ciência dos Esportes

- RBFF - Revista Brasileira de Futsal e Futebol http://www.rbff.com.br
- RBME - Revista Brasileira de Medicina do Esporte https://www.scielo.br/j/rbme
- RBPE - Revista Brasileira de Psicologia do Esporte http://pepsic.bvsalud.org

Periódicos da área de Medicina

- RBC - Revista Brasileira de Cancerologia https://rbc.inca.gov.br/index.php/revista
- RBCMS - Revista Brasileira de Ciências Médicas e da Saúde http://www.rbcms.com.br
- Revista da Associação Brasileira de Saúde Coletiva https://cienciaesaudecoletiva.com.br

Periódicos da área de Piscologia

- Psicologia argumento https://periodicos.pucpr.br/psicologiaargumento
- Estudos de psicologia (Campinas) https://www.scielo.br/j/estpsi/
- Psicologia em foco https://revistas.fw.uri.br/index.php/psicologiaemfoco

Ou busque na ferramenta *Mendeley* https://www.mendeley.com

# Chapter 3

# Definições Iniciais

A **Estatística** é uma área fundamental para entender e analisar dados de diversas áreas do conhecimento.

## 3.1 Estatística Descritiva

### 3.1.1 Exemplo 1: Saúde Mental dos Estudantes

Suponha que você deseje realizar uma pesquisa na UFTM para avaliar a saúde mental dos estudantes. Os dados coletados provavelmente incluirão informações sobre os níveis de **estresse**, **ansiedade**, **depressão** e **bem-estar geral** dos alunos. Nesse contexto, a **Estatística Descritiva** seria uma ferramenta essencial para organizar esses dados e apresentá-los de maneira clara, objetiva e compreensível.

Utilizando a Estatística Descritiva, você poderia:

1. **Organizar e estruturar os dados**, criando tabelas que agrupem as respostas dos estudantes de forma sistemática, o que facilitaria a análise dos padrões e a visualização de tendências.
2. **Calcular medidas de tendência central**, como a **média** dos níveis de estresse ou ansiedade, permitindo que você obtenha uma visão geral dos aspectos mais comuns da saúde mental entre os estudantes.
3. **Elaborar gráficos**, como **gráficos de barras** ou **diagramas de caixa (boxplot)**, para representar visualmente os dados e facilitar a interpretação. Por exemplo, esses gráficos poderiam mostrar quantos alunos estão em diferentes faixas de intensidade de ansiedade (baixo, moderado ou alto).
4. **Calcular medidas de dispersão**, como o **desvio padrão**, para avaliar a variabilidade dos dados. Um desvio padrão alto, por exemplo, indicaria que há uma grande diferença entre os níveis de estresse ou ansiedade dos estudantes, o que poderia apontar para a necessidade de abordagens personalizadas em programas de apoio.
5. **Identificar padrões e correlações** nos dados, como a relação entre o nível de estresse e o número de horas de estudo, ou entre o bem-estar geral e a prática de atividades físicas, ajudando a identificar fatores que influenciam a saúde mental dos estudantes.

Essas análises permitiriam que você compreendesse melhor o panorama da saúde mental dos alunos da UFTM e fornecesse uma base sólida para tomar decisões informadas sobre possíveis ações e programas de apoio psicológico.

### 3.1.2 Exemplo 2: Prática de Atividades Físicas

Se o objetivo for entender a frequência com que os estudantes da UFTM praticam atividades físicas regularmente, podemos coletar dados específicos, como:

- **Percentual de alunos que praticam esportes regularmente**: A porcentagem de estudantes que praticam atividades físicas **mais de três vezes por semana**. Este dado oferece uma visão geral sobre a adesão à prática de exercícios e pode ser útil para identificar possíveis áreas de melhoria no incentivo à saúde física.

- **Distribuição por tipo de atividade física**: Podemos utilizar gráficos como **gráficos de pizza** ou **barras empilhadas** para mostrar a diversidade de atividades praticadas pelos estudantes. Por exemplo, os gráficos poderiam ilustrar a proporção de alunos que praticam **corrida**, **musculação**, **dança**, **futebol**, entre outras atividades, permitindo uma visão clara sobre as preferências esportivas da comunidade acadêmica.
- **Intensidade da prática de atividades físicas**: Coletar dados sobre a intensidade das atividades (leve, moderada ou intensa) também seria relevante, pois ajudaria a universidade a entender não só a frequência, mas também o impacto potencial dessas atividades na saúde física e mental dos estudantes.

Essas informações não apenas proporcionam um panorama claro sobre o comportamento dos estudantes em relação à saúde física, mas também podem embasar a criação de novos programas e campanhas para incentivar a prática de atividades físicas, promovendo o bem-estar geral da comunidade acadêmica.

## 3.2 Estatística Inferencial

### 3.2.1 Exemplo 1: Saúde Mental dos Estudantes

Suponha que você queira avaliar a saúde mental dos estudantes da UFTM e, para isso, tenha realizado uma pesquisa com uma amostra de 200 estudantes. Neste caso, a **Estatística Inferencial** será uma ferramenta essencial para generalizar os resultados dessa amostra para toda a população de estudantes da universidade.

- **Estimativa da média de ansiedade**: Se a média da amostra indicar que os estudantes têm um nível de ansiedade de 7 (em uma escala de 0 a 10), a Estatística Inferencial permitirá que você estime o nível médio de ansiedade de todos os estudantes da UFTM. A partir dessa amostra, podemos calcular um **intervalo de confiança**, que nos dirá com que precisão a média da amostra reflete a média da população inteira. Isso nos dá uma estimativa com uma margem de erro, ajudando a compreender a variabilidade e a precisão do resultado.
- **Testes de hipóteses**: Suponha que você queira comparar os níveis de ansiedade entre estudantes de diferentes cursos. Por exemplo, você deseja saber se os estudantes do curso de Medicina apresentam níveis de ansiedade significativamente mais altos do que os estudantes do curso de Engenharia. A Estatística Inferencial permite realizar um **teste de hipótese**, que nos ajuda a determinar se a diferença observada entre as médias dos dois grupos é **estatisticamente significativa** ou se pode ter ocorrido por acaso. Isso nos permite tomar decisões informadas sobre possíveis relações entre variáveis, com base em dados e não em suposições.

Esses exemplos demonstram como a Estatística Inferencial pode ser usada para tirar conclusões sobre toda a população de estudantes da UFTM, com base em uma amostra representativa. Isso é essencial para planejar intervenções e políticas de apoio à saúde mental de forma mais eficaz.

### 3.2.2 Exemplo 2: Prática de Atividades Físicas

Suponha que você esteja interessado em saber como os hábitos de atividade física variam entre os cursos da UFTM. A universidade realiza uma pesquisa com uma amostra de 300 estudantes, incluindo alunos de Medicina e de Educação Física. A **Estatística Inferencial** pode ser usada para tirar conclusões sobre toda a população de estudantes com base nessa amostra.

- **Estimar a proporção de estudantes ativos fisicamente**: Se 60% dos estudantes da amostra afirmam praticar atividades físicas regularmente, podemos usar a **probabilidade** para estimar a proporção de todos os estudantes da UFTM que têm esse hábito. A partir desses dados, também é possível calcular um **intervalo de confiança**, indicando com que precisão essa estimativa reflete a realidade da população.
- **Comparação entre cursos**: Podemos realizar um **teste de comparação de proporções** para verificar se a prática de atividades físicas é mais frequente entre estudantes de Educação Física do que entre estudantes de Medicina. Esse teste nos ajudaria a determinar se a diferença observada entre os dois cursos é estatisticamente significativa ou se pode ser explicada por variações aleatórias.

Essas análises ajudam a universidade a entender os padrões de comportamento relacionados à saúde física entre diferentes grupos acadêmicos e a planejar ações mais específicas para incentivar o bem-estar entre os estudantes.

## 3.3 Conexão com a Probabilidade

Esses dois tipos de análise estão profundamente conectados com a **Teoria de Probabilidade**, que nos permite estimar as chances de um evento acontecer e tomar decisões com base nesses dados. No caso da UFTM, podemos usar a **probabilidade** para prever, por exemplo, a chance de um estudante desenvolver sintomas de estresse durante o semestre, com base em comportamentos anteriores, como hábitos de estudo e participação em atividades de lazer.

## 3.4 População x Amostra

### 3.4.1 População

A **população** é o conjunto completo de indivíduos que queremos estudar. No contexto de uma pesquisa com estudantes da UFTM, a população pode ser definida como **todos os 6.900 estudantes** da instituição, considerando alunos da graduação, cursos técnicos e pós-graduação (dados do DRCA, em dezembro de 2024).

### 3.4.2 Amostra

Uma **amostra** é um grupo menor selecionado dessa população, que deve ser representativo o suficiente para permitir inferências sobre o todo. Por exemplo, em vez de aplicar um questionário para todos os 6.900 estudantes, podemos selecionar uma **amostra aleatória de 364 estudantes**.

Esse tamanho de amostra é adequado para garantir um **erro máximo de 5%**, com **95% de confiança** nos resultados. Isso significa que, com essa amostra, conseguimos estimar com boa precisão como é a situação da saúde dos estudantes da universidade como um todo.

### 3.4.3 Exemplo aplicado

Suponha que você queira avaliar o nível de bem-estar emocional dos estudantes da UFTM. Você aplica um questionário padronizado para uma amostra aleatória de **364 estudantes**. A partir das respostas, você calcula que a média de bem-estar emocional, numa escala de 0 a 10, é **6,8**.

Com esse resultado e a ajuda da **estatística inferencial**, você pode estimar com segurança que o nível médio de bem-estar emocional da **população inteira** de estudantes da UFTM está próximo de **6,8**, com uma **margem de erro de 5%**.

Isso significa que, com **95% de confiança**, o **intervalo de confiança** para a média do bem-estar emocional está entre **6,46 e 7,14**. Em outras palavras, se essa pesquisa fosse repetida várias vezes com diferentes amostras aleatórias de 364 estudantes, em 95% das vezes o valor da média real da população estaria dentro desse intervalo.

# Chapter 4

# Tamanho da Amostra

O cálculo do **tamanho da amostra** é um passo importante em qualquer pesquisa. Ele nos ajuda a determinar quantos participantes precisamos para que os resultados sejam representativos e estatisticamente confiáveis. Embora o cálculo envolva alguns conceitos de **estatística inferencial**, hoje em dia, esse processo pode ser muito mais simples, graças a ferramentas online e até mesmo com o auxílio de **Inteligência Artificial (IA)**.

## 4.1 Passos básicos para o cálculo do tamanho da amostra

Quando o **objetivo é estimar uma proporção**, o cálculo do tamanho amostral exige algumas informações essenciais:

1. **Tamanho da população**: Neste caso, os estudantes da UFTM, que somam cerca de **6.900**.
2. **Margem de erro**: A precisão com que queremos que nossa estimativa esteja próxima da realidade. Por exemplo, uma margem de erro de **5%** é comum em muitos estudos.
3. **Nível de confiança**: Normalmente, utiliza-se 95%, que indica a probabilidade de que o intervalo de confiança contenha o valor real.

A fórmula para calcular o tamanho da amostra é:

$$n = \frac{Z^2 \;\; p \;\; (1 \;\; p)}{E^2}$$

Onde:

- $n$ = tamanho da amostra
- $Z$ = valor da distribuição normal (geralmente 1,96 para 95% de confiança)
- $p$ = proporção estimada (por exemplo, 0,5 para o pior cenário)
- $E$ = margem de erro (por exemplo, 0,05 para 5%)

### 4.1.1 Exemplo de cálculo simples

Vamos calcular o tamanho da amostra para uma pesquisa na UFTM, onde sabemos que a população tem **6.900 estudantes**, a margem de erro é **5%**, o nível de confiança é **95%**, e vamos supor que queremos estimar uma proporção de **50%** (o pior cenário para garantir maior precisão).

Usando a fórmula, temos:

- $Z = 1,96$ (para 95% de confiança)
- $p = 0,5$ (proporção estimada)
- $E = 0,05$ (margem de erro de 5%)

Substituindo os valores na fórmula, obtemos:

$$n = \frac{1,96^2 \ 0,5 \ (1 \ 0,5)}{0,05^2} = 384$$

Logo, o tamanho da amostra necessário é **384** estudantes para garantir uma margem de erro de 5% com 95% de confiança.

**Observação:** A fórmula apresentada acima é utilizada para o cálculo do tamanho amostral quando se deseja estimar uma proporção, assumindo uma população infinita (ou muito grande). Para situações em que o objetivo é estimar a média populacional, a fórmula adequada é diferente e leva em conta o desvio padrão da variável em vez da proporção. Além disso, se a população for finita, é necessário aplicar o fator de correção amostral.

### 4.1.2 Ferramentas para cálculo de amostra

Nosso foco aqui não é ensinar a fazer os cálculos manualmente, mas compreender os conceitos envolvidos. Existem diversas ferramentas online gratuitas que fazem o cálculo automaticamente. Por exemplo:

## 4.2 Cálculo do Tamanho da Amostra

Para calcular o tamanho da amostra, podemos utilizar diferentes ferramentas, tanto no **R** quanto online. Aqui estão algumas opções:

- **No R**: Para calcular o tamanho da amostra, podemos utilizar o **pacote pwr**, que oferece funções específicas para realizar esses cálculos de forma simples e eficaz. O pacote é ideal para quem precisa calcular o tamanho da amostra para testes de hipóteses, como testes de médias, proporções ou análise de variância, ajustando facilmente os parâmetros conforme as necessidades da pesquisa.
- **Calculadora Amostral Online da USP Bauru**: Clique aqui para acessar - Uma ferramenta simples e eficiente para cálculos amostrais.
- **G*Power**: Um software gratuito que pode ajudar a calcular o tamanho da amostra dependendo do tipo de análise estatística que você irá realizar. Saiba mais sobre o G*Power.
- **OpenEpi – Sample Size**: Acesse aqui - Muito usado na área da saúde, permite calcular amostras para média, proporção, estudos de caso-controle, coorte, entre outros.

## 4.3 Conclusão

Em resumo, o cálculo do **tamanho da amostra** é um passo essencial para garantir que suas pesquisas sejam precisas e representativas. Embora os cálculos possam parecer complexos, hoje em dia, existem ferramentas poderosas e **Inteligência Artificial** para facilitar esse processo, economizando tempo e esforço na sua pesquisa.

## 4.4 Atividade 2

**Responda a partir do artigo buscado na atividade 1.**

1. **Os autores discutem o cálculo do tamanho da amostra no estudo?**
   - Você consegue identificar se os autores fornecem uma explicação clara sobre como determinaram o tamanho da amostra na pesquisa? Eles utilizaram alguma fórmula específica ou ferramenta estatística para esse cálculo?
2. **A população da pesquisa foi claramente definida pelos autores?**
   - Os autores descreveram com clareza quem compõe a população da pesquisa (por exemplo, características demográficas, contexto da amostra, entre outros)? A população foi bem delineada para garantir que os resultados sejam representativos?
3. **O tamanho da amostra utilizado no estudo foi adequado?**

- Com base no cálculo do tamanho de amostra, você acha que a amostra utilizada no estudo foi grande o suficiente para garantir a precisão dos resultados? Ela foi suficientemente representativa da população alvo da pesquisa?

4. **Realize o cálculo do tamanho de amostra para este estudo.**
   - Com as informações fornecidas pelos autores, você pode calcular o tamanho da amostra necessário para o estudo utilizando uma fórmula ou ferramenta de cálculo amostral. Considere aspectos como o nível de confiança, margem de erro e variabilidade dos dados.
5. **Os autores mencionaram se a pesquisa foi aprovada pelo Comitê de Ética?**
   - No artigo, há alguma referência sobre a aprovação ética da pesquisa? Os autores informaram se a pesquisa foi submetida e aprovada por um Comitê de Ética em Pesquisa (CEP)?
6. **Qual é a importância de submeter a pesquisa ao Comitê de Ética?**
   - Por que é essencial que a pesquisa seja submetida a um Comitê de Ética? Quais são as implicações éticas de não realizar essa submissão, especialmente em pesquisas que envolvem seres humanos?

**Importante**

Conheça o Comitê de Ética em Pesquisa (CEP) da UFTM

https://www.uftm.edu.br/comitesecomissoes/cep

# Chapter 5

# Técnicas de Amostragem

Em pesquisas, uma das etapas mais importantes é o **cálculo amostral**, que determina quantos participantes serão necessários para garantir que os resultados sejam confiáveis. Depois, é preciso escolher uma **técnica de amostragem**, ou seja, como vamos selecionar as pessoas ou unidades que farão parte da pesquisa. Uma parte essencial disso é garantir que o processo de amostragem evite **viés**, que ocorre quando a amostra não representa adequadamente a população.

Existem várias técnicas de amostragem, cada uma com suas características, vantagens e desvantagens.

---

## 5.1 Amostragem Aleatória Simples

Nesta técnica, cada elemento da população tem a mesma chance de ser selecionado.

> **Exemplo:** Se quisermos estudar a saúde mental dos estudantes da UFTM, podemos utilizar uma lista completa dos alunos e, de maneira aleatória (sorteio), selecionar os estudantes que farão parte da amostra, conforme o número determinado pelo cálculo amostral.

- **Vantagens:** É simples de aplicar e garante que todos têm a mesma chance de ser escolhidos, minimizando o viés.
- **Desvantagens:** Pode ser difícil de implementar se a população for muito grande ou se não tivermos uma lista completa de todos os membros da população.
- **Como evitar viés:** Garantir que a lista de onde serão sorteados os participantes seja completa e atualizada.

---

## 5.2 Amostragem Estratificada

Aqui, a população é dividida em grupos ou "estratos" (por exemplo, cursos ou faixas etárias) e, em seguida, amostras são selecionadas dentro de cada estrato.

> **Exemplo:** Suponha que queremos estudar a prática de atividade física entre os estudantes da UFTM. Podemos usar amostragem estratificada, dividindo os alunos por curso (estratos). Dentro de cada curso, selecionamos aleatoriamente um número proporcional de alunos, garantindo que todos os cursos estejam representados de forma adequada na amostra, refletindo a diversidade da universidade.

- **Vantagens:** Melhora a representatividade da amostra, especialmente quando diferentes subgrupos podem ter comportamentos ou características distintas. Ajuda a evitar viés ao garantir que todos os grupos da população estejam representados.
- **Desvantagens:** É necessário conhecer a população e seus estratos de antemão, o que pode ser difícil em alguns casos.

- **Como evitar viés:** A definição dos estratos deve ser precisa e relevante para a pesquisa.

---

## 5.3 Amostragem Sistemática

Na amostragem sistemática, escolhemos um ponto de partida aleatório e, a partir daí, selecionamos unidades a intervalos regulares de “n” unidades.

> **Exemplo:** Se quisermos estudar a prática de atividade esportiva entre os estudantes da UFTM e temos uma lista de todos os alunos, podemos selecionar a cada 10ž aluno da lista para participar da pesquisa.

- **Vantagens:** Fácil de implementar, especialmente em populações grandes, e garante uma distribuição uniforme dos selecionados ao longo da lista.
- **Desvantagens:** Pode ocorrer viés caso haja algum padrão na ordem da lista (por exemplo, se alunos de cursos específicos forem listados consecutivamente, a amostra pode não ser representativa).
- **Como evitar viés:** A lista de seleção deve ser aleatória e não seguir um padrão que favoreça um grupo específico.

---

## 5.4 Amostragem por Conglomerados

Nessa técnica, a população é dividida em grupos (chamados de conglomerados) e, em seguida, seleciona-se aleatoriamente alguns desses grupos para fazer parte da amostra.

> **Exemplo:** Em vez de selecionar alunos aleatoriamente, selecionar alguns cursos específicos da UFTM e estudar todos os alunos desses cursos. Essa técnica é útil quando a população é muito grande e difícil de acessar como um todo.

- **Vantagens:** Mais fácil de administrar em populações grandes, quando não se tem acesso a uma lista completa de todos os indivíduos.
- **Desvantagens:** Pode não ser tão representativa, pois estamos escolhendo grupos inteiros e não indivíduos aleatórios.
- **Como evitar viés:** Os conglomerados escolhidos devem representar adequadamente a diversidade da população.

---

## 5.5 Amostragem por Conveniência

Na **amostragem por conveniência**, os participantes são selecionados por serem mais fáceis de acessar pelo pesquisador.

> **Exemplo:** O uso de formulários online, como o Google Forms, é uma forma comum de amostragem por conveniência. O pesquisador compartilha o formulário em redes sociais, grupos de WhatsApp ou por e-mail, e os primeiros que responderem entram para a amostra.

- **Vantagens:** Rápida, prática, econômica e permite coletar dados de muitas pessoas em pouco tempo.
- **Desvantagens:** Forte risco de viés, pois a amostra pode não representar a população como um todo, já que depende de quem tem acesso ao link, internet e interesse em responder.
- **Como evitar viés:** Embora não seja possível eliminar totalmente o viés, recomenda-se divulgar o formulário em diferentes canais e para públicos diversos, incentivando a participação de vários grupos.

**Atenção aos Vieses no Uso de Formulários Online:**

- **Viés de acesso:** Apenas pessoas com acesso à internet podem responder.
- **Viés de auto-seleção:** Aqueles mais motivados ou interessados no tema tendem a participar mais.

- **Viés de distribuição:** Se o link for enviado apenas a determinados grupos, outros podem ser sub-representados.

Portanto, ao usar formulários online, é importante considerar essas limitações e, se possível, adotar estratégias para ampliar o alcance e a diversidade dos respondentes.

---

## 5.6 Combinando Técnicas na Prática

Na prática, é comum que pesquisadores combinem diferentes técnicas de amostragem para aumentar a representatividade e reduzir o viés da amostra, adaptando o processo às características específicas da população e aos objetivos do estudo.

Por exemplo, pode-se aplicar a **amostragem estratificada** para garantir que todos os subgrupos relevantes (como cursos, faixas etárias ou regiões) estejam proporcionalmente representados na amostra. Em seguida, dentro de cada estrato, pode-se utilizar a **amostragem aleatória simples** para selecionar os participantes de forma justa e imparcial.

Além disso, em situações em que o acesso à população é limitado, pode-se recorrer à combinação de técnicas probabilísticas (como estratificada ou sistemática) com métodos não probabilísticos (como conveniência), sempre buscando estratégias para minimizar possíveis vieses e aumentar a diversidade dos participantes.

Ao combinar métodos, o pesquisador consegue contornar limitações práticas (como listas incompletas ou dificuldades de acesso) e, ao mesmo tempo, assegurar que a amostra reflita com maior fidelidade a diversidade da população, tornando os resultados mais robustos e confiáveis.

---

## 5.7 Evitar Viés na Amostragem

O **viés de amostragem** ocorre quando certos grupos da população têm mais chance de ser selecionados do que outros, o que pode distorcer os resultados da pesquisa. Para evitar viés, é essencial:

- Garantir que todos os grupos da população tenham uma chance igual ou proporcional de ser selecionados.
- Utilizar técnicas que levem em consideração as características específicas da população.
- Combinar diferentes métodos de amostragem para melhorar a representatividade.

---

## 5.8 Resumindo

- **Probabilística:** amostragem aleatória simples, sistemática, estratificada, por conglomerados.
- **Não probabilística:** amostragem por conveniência (incluindo Google Forms), intencional, por cotas.

  A escolha da técnica deve considerar os objetivos da pesquisa, os recursos disponíveis e a necessidade de representatividade da amostra.

# Chapter 6

# Ambiente Computacional

Existem diversos softwares dedicados à análise estatística, que vão desde planilhas eletrônicas, como o Excel, até programas mais robustos, como o SPSS. Abaixo, listamos algumas das principais ferramentas utilizadas:

## 6.1 Softwares pagos

- SPSS (IBM)
- Stata
- SAS
- JMP
- Prism
- Minitab
- Microsoft Excel

## 6.2 Softwares livres

- Jamovi
- OpenStat

## 6.3 Linguagens computacionais

- R
- Python

Neste curso, utilizaremos a linguagem **R**, desenvolvida especialmente para análise estatística. Quer entender por que essa escolha? Recomendamos a leitura: Por que usar R?

## 6.4 Preparando o Ambiente Computacional

Vamos preparar o ambiente computacional para realizar nossas análises com R.

> Para esclarecer: - **R** é uma linguagem de programação (mas não se preocupe, não vamos programar profundamente). - **RStudio** é o software onde os códigos R serão executados. É o que chamamos de IDE – *Integrated Development Environment* (Ambiente de Desenvolvimento Integrado).

## 6.5 Plano A: Instalação do R e RStudio

Nos laboratórios da UFTM, o R e o RStudio já estão instalados. No entanto, sugerimos que você também os instale em seu computador pessoal, pois nem sempre conseguiremos realizar todas as atividades em sala.

Siga estes passos:

1. Instale o R: https://cran.rstudio.com
2. Instale o RStudio Desktop: https://posit.co/download/rstudio-desktop

   Certifique-se de baixar versões compatíveis com o sistema operacional do seu computador.

Se tudo estiver correto, ao abrir o RStudio, você verá uma tela semelhante a esta:

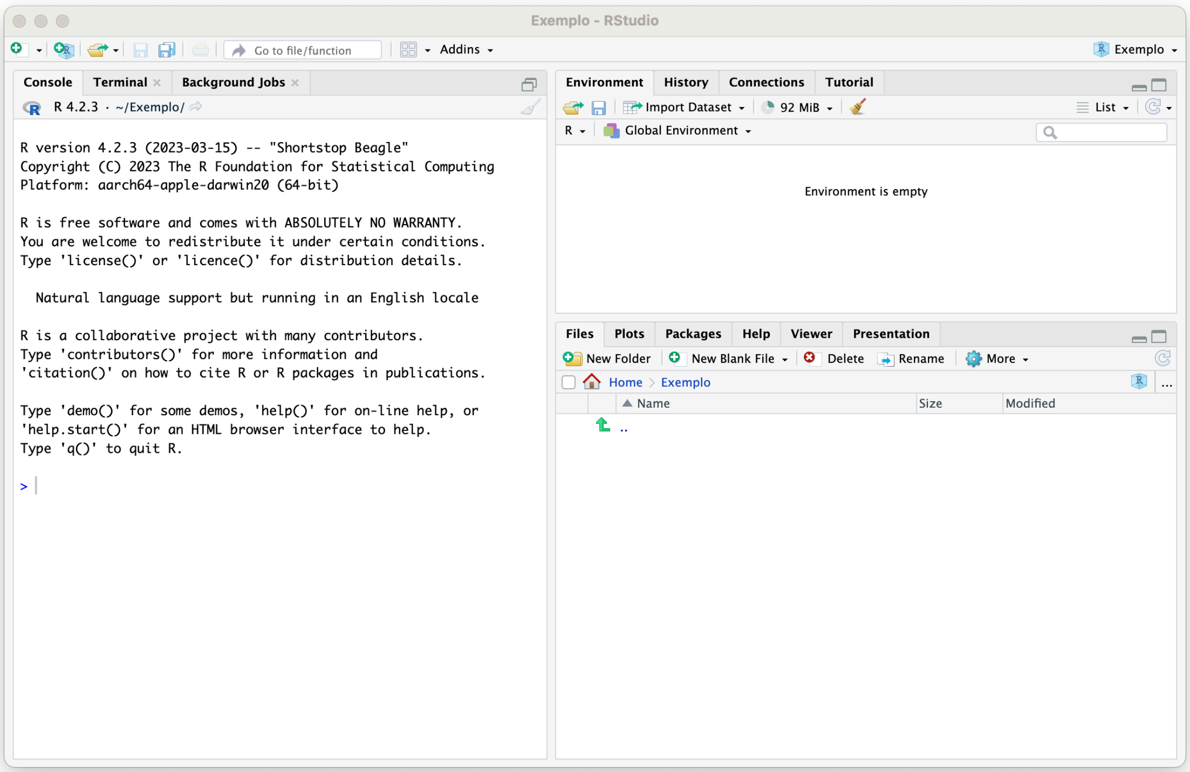



Figure 6.1: Figura: Tela inicial do RStudio

Caso enfrente dificuldades, siga para o Plano B.

## 6.6 Plano B: R e RStudio na Nuvem

O Plano B é utilizar o RStudio diretamente na nuvem, sem precisar instalar nada. Essa alternativa é excelente, mas requer uma boa conexão com a internet.

Siga os passos:

1. Acesse: https://posit.cloud
2. Faça login (você pode usar sua conta do Google, por exemplo)
3. Após o login, você verá esta tela:
4. Crie um novo projeto clicando em *New Project* e, depois, *New RStudio Project*:
5. Pronto! A interface do RStudio será carregada na nuvem:

   Vantagem: seus arquivos e análises ficam salvos na nuvem, em um local seguro e acessível de qualquer lugar.

---

Com o ambiente configurado, estaremos prontos para explorar o mundo da estatística com R!

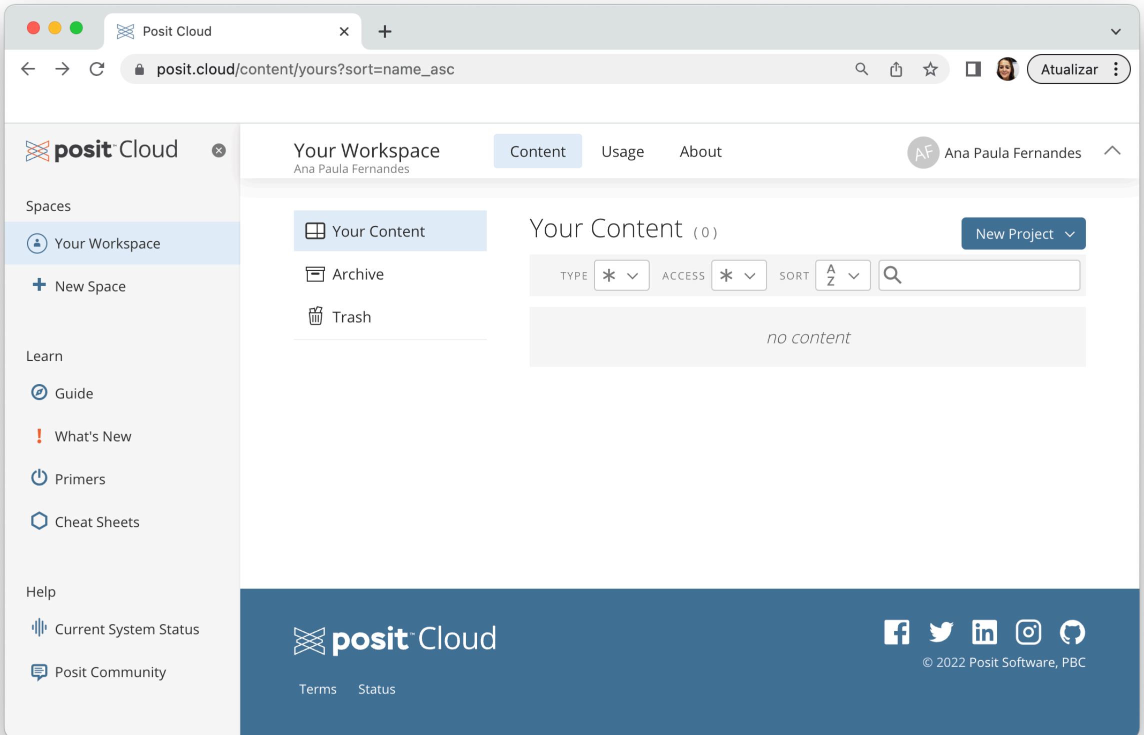



Figure 6.2: Figura: Tela inicial na nuvem da Posit

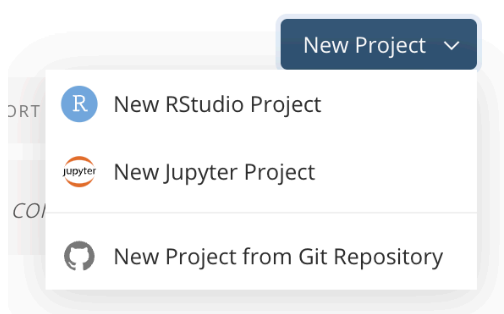



Figure 6.3: Figura: Botão de criação de um novo projeto

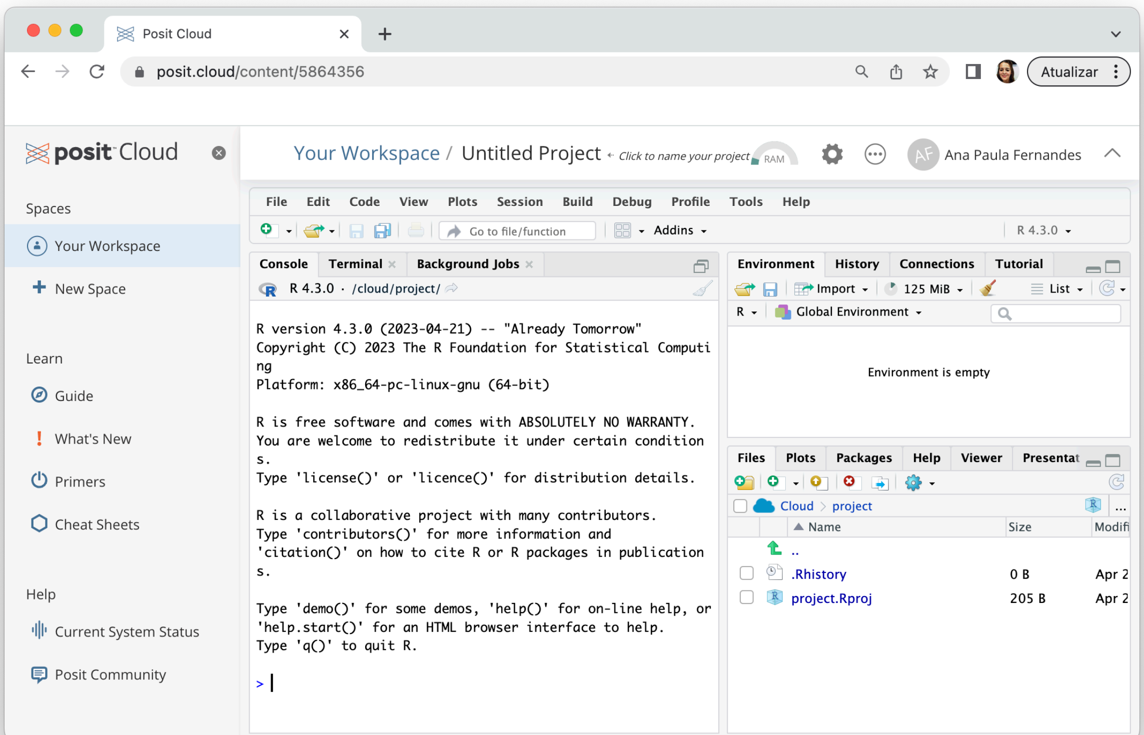



Figure 6.4: Figura: Tela do RStudio online na Posit Cloud

# Chapter 7

# Trabalhando no RStudio

Seja na versão instalada no seu computador (plano A) ou na nuvem (plano B), conheça melhor as áreas do RStudio:

1. **Console:** local onde serão apresentadas as respostas para códigos executados;
2. **Ambiente de memória (Environment):** é o cérebro do R, onde ficam registrados os objetos que ele reconhece.
3. A área de **Arquivos (Files), Gráficos (Plots), Pacotes (Packages), Ajuda (Help), Visualização (Viewer) e Apresentação (Presentation)**: mostram, respectivamente, os arquivos do diretório onde estão seus arquivos no computador, os gráficos, os pacotes, a ajuda, a janela de visualização e a apresentação.

A figura abaixo identifica cada uma dessas áreas:

Digitaremos os códigos da linguagem R, em um arquivo que chamamos de **script**. Para abrir um arquivo do tipo script R, faça:

1. Acesse a opção **File** no menu principal do RStudio;
2. Escolha a opção **New File**;
3. E depois a opção **R Script**.

Assim, na tela da IDE RStudio aparecerá uma nova área, que é a área do arquivo script, como mostra a figura.

Observe que o arquivo está sem um nome (**Untitled1**, sem título). Salve o arquivo atribuindo-lhe um nome adequado. Para isso, no menu principal, escolha *File*, depois *Save*.

> Dica: O ideal seria criar um **Projeto**. Veja a opção *File > New Project*.

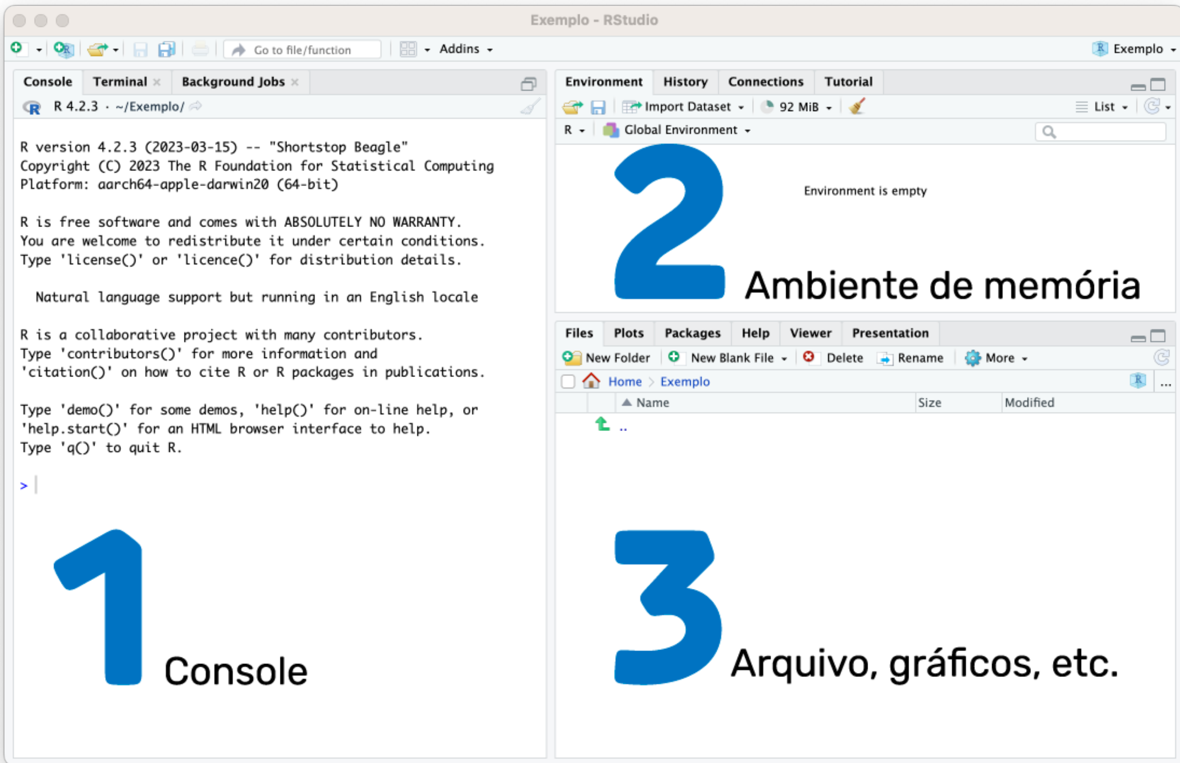



Figure 7.1: Figura: Identificação das áreas do RStudio

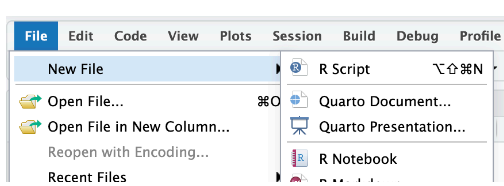



Figure 7.2: Figura: Como abrir um novo arquivo de script R

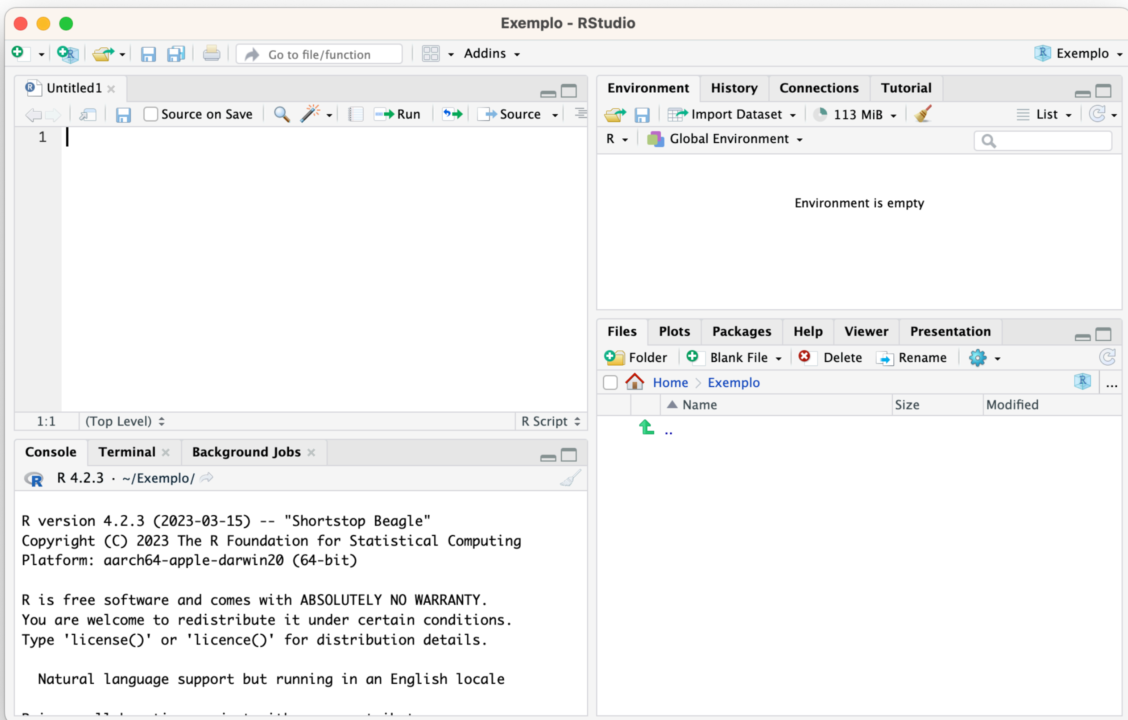



Figure 7.3: Figura: Como abrir um novo arquivo de script R

# Chapter 8

# Primeiros exercícios no R

Nos capítulos 6 e 7 vimos sobre o ambiente computacional (computador ou nuvem) e identificamos as 4 áreas da tela da interface do RStudio: **console**, **ambiente de memória**, **arquivos, gráficos, etc.** e **script**, assim estamos prontos para escrever alguns códigos e executá-los a partir da área de script.

> **Atenção:** TODOS os cógigos serão digitados no arquivo de script, seguindo uma sequência lógica de passos, ou seja, escreveremos um roteiro (*script*), como se fosse uma receita de bolo, isso é o que o pessoal da computação chama de algoritmo.

## 8.1 Exemplo 1

- Observe o código escrito na linha 1 do arquivo de script e o botão **Run** (primeira seta verde):

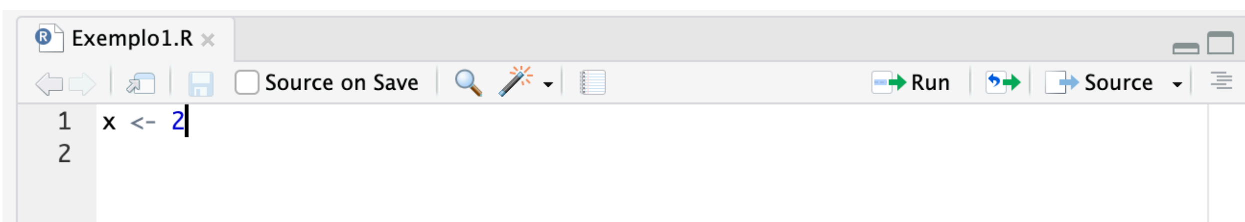



Figure 8.1: Figura: Primeiro exemplo de cógigo R

- O sequência de caracteres **<-** é o símbolo de atribuição no R.

  Pressionando as teclas ALT e - (menos) simultaneamente cria no script o sinal de atribuição.
- O código significa que criamos um objeto chamado **x** e atribuimos a esse objeto o valor 2.
- No entanto, o R ainda não sabe que o valor de x é igual a 2!
- Para registrar essa informação na memória do R, devemos executar essa linha.

  Para executar uma linha posicione o cursor na linha, e clique no botão **Run**

  Observe sempre o ambiente de memória (bem como o console) quando executar uma linha.

## 8.2 Exemplo 2

Execute o seguinte cógigo no R.

```
idades <- c( 23, 18, 17, 25, 21, 19, 22, 24, 19, 19 )
```

- Esse código significa que foi criado um objeto chamado idade que armazena 10 valores: 23, 18, 17, 25, 21, 19, 22, 24, 19, 19, diferentemente do exemplo 1 em que x armazenava somente o valor 2.

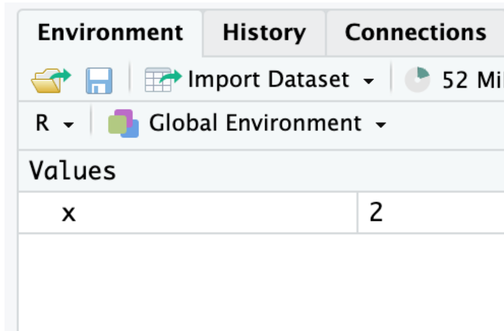



Figure 8.2: Figura: O objeto x é registrado na memória do R, armazenando o valor igual a 2

- Isso foi possível pois usamos a função **c( )**.
- observe que os valores foram colocado dentro dos parênteses da função **c( )**

  Com função **c( )** podemos **combinar** vários valores em um objeto, esse objeto recebe o nome de vetor ou lista.

## 8.3 Exemplo 3

Observe nesse código as funções:

- **max( )**
- **min( )**
- **range( )**
- **mean( )**
- **sd( )**

```
# criando o vetor idades
idades <- c( 23, 18, 17, 25, 21, 19, 22, 24, 19, 19 )

# maior valor
# função max( )
max(idades)
```

```
## [1] 25
```

```
# menor valor
# função min( )
min(idades)
```

```
## [1] 17
```

```
# faixa de valores
# função range( )
range(idades)
```

```
## [1] 17 25
```

```
# média (mean)
# função mean( )
mean(idades)
```

```
## [1] 20.7
```

```
# desvio padrão (standard deviation)
# função sd()
sd(idades)
```

```
## [1] 2.710064
```

Copie o código e cole no seu arquivo script, selecione todo conteúdo (CRTL+A) e execute todo o cógigo de uma única vez.

- Observe que as respostas apareceram no **console**, conforme mostrado na figura abaixo:

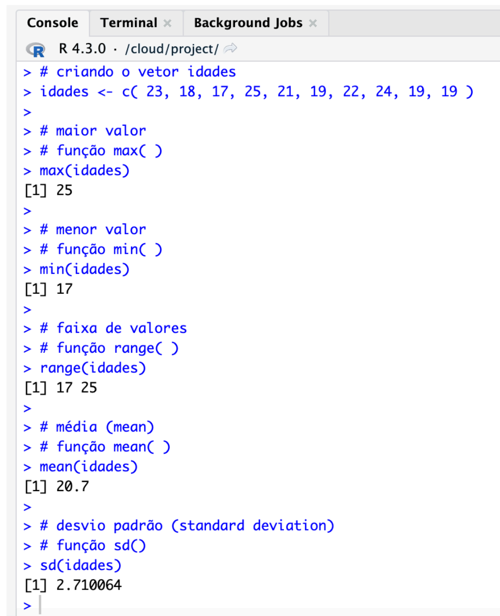



Figure 8.3: Figura: Como abrir um novo arquivo de script R

O símbolo # é o símbolo de comentário, isso significa que podemos escrever qualquer texto diferente do que o R sabe interpretar, e mesmo executando o código nenhum erro acontece!

**IMPORTANTE**: é uma boa prática comentar os trechos de códigos para deixar documentado qual é o objetivo do código.

# Chapter 9

# Tipos de Variáveis

As variáveis podem ser classificadas de acordo com sua natureza:

## 9.1 Variáveis Quantitativas

Expressam **quantidade** e podem ser divididas em:

- **Discreta**: Assume valores inteiros (contagem).
  - *Exemplo*: O número de filhos de uma pessoa. Você pode contar 0, 1, 2, 3 filhos, mas não pode ter 2,5 filhos.
- **Contínua**: Assume qualquer valor dentro de um intervalo específico (mensuração).
  - *Exemplo*: A altura de uma pessoa, que pode ser 1,70 m, 1,71 m, 1,711 m, e assim por diante, com infinitas possibilidades entre os valores.

## 9.2 Variáveis Qualitativas

Expressam **qualidade** e são representadas por **categorias** ou **rótulos**. São subdivididas em:

- **Nominal**: Categorias que não podem ser ordenadas.
  - *Exemplo*: O tipo de fruta que você prefere, como maçã, banana, laranja. Não faz sentido ordenar essas categorias de maneira que uma seja "maior" que a outra.
- **Ordinal**: Categorias que podem ser ordenadas, com uma graduação entre elas.
  - *Exemplo*: Níveis de satisfação de um serviço, como "satisfeito", "neutro" e "insatisfeito". Aqui, existe uma ordem em que "satisfeito" é maior que "neutro", e "neutro" é maior que "insatisfeito".

Uma aplicação comum das variáveis ordinais é a **Escala Likert**, que é amplamente utilizada em pesquisas de opinião para medir atitudes ou percepções. A escala Likert geralmente apresenta uma série de afirmativas, e o participante deve indicar seu nível de concordância com cada uma delas, com respostas em uma sequência ordenada.

*Exemplo*: Em uma pesquisa de satisfação de clientes, a pergunta poderia ser: "Concordo que a alimentação oferecida pelo hospital foi satisfatória." As opções de resposta poderiam ser:

1. **Discordo totalmente**
2. **Discordo parcialmente**
3. **Neutro**
4. **Concordo parcialmente**
5. **Concordo totalmente**

Essas respostas formam uma escala ordinal, pois a ordem das categorias reflete um aumento no nível de concordância, mas as diferenças entre elas não são necessariamente iguais.

**Dica**: Veja o vídeo do Prof. Heitor no **Canal Pesquise**: Assista aqui.

## 9.3 Aplicação das Variáveis em Estatística

A natureza da variável influencia o tipo de procedimento estatístico que será utilizado. Exemplos:

### 9.3.1 Estatística Descritiva:

- **Variáveis Qualitativas**: São representadas por sua **frequência absoluta** ou **percentual**.
  - *Exemplo*: Em uma pesquisa de preferência de frutas, pode-se dizer que 40% das pessoas escolheram maçã, 35% escolheram banana, e 25% escolheram laranja.
- **Variáveis Quantitativas**: São representadas por **medidas resumo**, como **média** e **desvio padrão**.
  - *Exemplo*: Se você calcular a média de altura de um grupo de pessoas, e o desvio padrão indicar que a altura das pessoas varia em torno de 5 cm da média.

### 9.3.2 Estatística Inferencial:

- **Teste Qui-Quadrado**: Usado para verificar se há associação entre as categorias de duas variáveis qualitativas.
  - *Exemplo*: Um teste Qui-Quadrado poderia ser utilizado para verificar se a escolha de fruta (maçã, banana, laranja) tem relação com o sexo (masculino, feminino) dos participantes.
- **Teste de Correlação de Pearson**: Mede a força da correlação linear entre duas variáveis quantitativas.
  - *Exemplo*: O teste de Pearson poderia ser utilizado para verificar se existe uma relação entre a altura e o peso das pessoas. Quanto maior a altura, maior o peso? Esse teste nos diria a força dessa relação.

## 9.4 Atividade 3

**Leia o artigo *Estado nutricional, tempo de internação e mortalidade em pacientes submetidos à cirurgia cardíaca em um hospital na cidade de Maceió***
Disponível em: RASBRAN, Revista da Associação Brasileira de Nutrição, 2023
Acesse também o portal: RASBRAN - Revista da Associação Brasileira de Nutrição.

### 9.4.1 Análise das Tabelas

- **Tabela 1 - Características clínicas dos pacientes submetidos à cirurgia cardíaca**: Esta tabela apresenta as características da amostra analisada na pesquisa.
  - **Tarefa**: Classifique as variáveis (características) em **qualitativas** e **quantitativas**.
  - **Observação**: Verifique como as variáveis foram resumidas. Elas foram apresentadas em **porcentagens**? Ou foram calculadas medidas de **média** e **desvio padrão**?
- **Tabela 2 - Associação entre estado nutricional, sexo, idade e tempo de internação hospitalar entre os pacientes submetidos à cirurgia cardíaca**

- **Tabela 3 - Associação entre evolução clínica, sexo, idade, tempo de internação hospitalar e estado nutricional entre os pacientes submetidos à cirurgia cardíaca**: Estas tabelas mostram os resultados de um **teste de hipótese** (parte da estatística inferencial).
  - **Tarefa**: Identifique qual **teste estatístico** foi aplicado.
  - **Objetivo**: Qual é o objetivo deste teste?

# Chapter 10

# Estatística Descritiva

A estatística descritiva permite **resumir, organizar e interpretar** dados de forma clara e objetiva. Para isso, utilizamos **medidas de tendência central**, **medidas de dispersão** e **medidas relativas de variabilidade**.

## 10.1 Medidas de Tendência Central (ou Posição)

### 10.1.1 Média

- **Definição:** Soma de todos os valores dividida pelo número de observações.
- **Interpretação:** Representa o valor médio ou típico do conjunto de dados.
- **Como reportar:**

  A média dos batimentos cardíacos foi de 58,6 bpm, indicando o valor médio da amostra analisada.

### 10.1.2 Mediana

- **Definição:** Valor central de um conjunto ordenado de dados.
- **Interpretação:** Divide o conjunto de dados ao meio, sendo útil quando há valores extremos (outliers).
- **Como reportar:**

  A mediana dos batimentos foi de 60,0 bpm, indicando que 50% dos indivíduos apresentaram valores abaixo ou iguais a esse valor.

### 10.1.3 Quartis

- **Definição:** Q1 (primeiro quartil) e Q3 (terceiro quartil) representam os valores que dividem os 25% e os 75% inferiores dos dados, respectivamente.
- **Interpretação:** Ajudam a entender a distribuição dos dados e identificar a dispersão em torno da mediana.
- **Como reportar:**

  O primeiro e o terceiro quartis foram 54,0 bpm e 64,0 bpm, respectivamente, revelando que 50% dos batimentos ficaram entre esses dois valores.

### 10.1.4 Moda

- **Definição:** Valor mais frequente do conjunto de dados.
- **Interpretação:** Indica o valor mais comum, embora possa não existir ou haver mais de uma moda.
- **Como reportar:**

  A moda foi 62 bpm, valor que ocorreu com maior frequência na amostra.

**Observação importante:**
- **Média e desvio padrão** são medidas que devem ser usadas juntas, especialmente para dados simétricos (distribuição simétrica) e sem valores extremos. - **Mediana e quartis** formam outro conjunto de medidas, mais apropriado quando há assimetria ou presença de outliers.

**Sugestão de vídeo:** Canal Pesquise - Tendência Central

## 10.2 Medidas de Dispersão (ou Variabilidade)

### 10.2.1 Amplitude

- **Definição:** Diferença entre o maior e o menor valor.
- **Interpretação:** Indica o intervalo total em que os dados variam.
- **Como reportar:**

  A amplitude foi de 36 bpm, com valores variando de 39 a 75 bpm.

### 10.2.2 Variância

- **Definição:** Média dos quadrados das diferenças entre os valores e a média.
- **Interpretação:** Mede a dispersão, mas sua unidade é o quadrado da unidade original.
- **Como reportar:**

  A variância foi de 98,8 bpmš, indicando a variabilidade dos batimentos em relação à média.

**Observação:**
A unidade da variância é expressa ao quadrado da unidade original dos dados (por exemplo, bpmš no caso de batimentos por minuto), o que pode dificultar sua interpretação direta.
Por isso, costuma-se utilizar o **desvio padrão**, que tem a **mesma unidade dos dados originais** e fornece uma noção mais intuitiva da dispersão dos valores em torno da média.

### 10.2.3 Desvio padrão (DP)

- **Definição:** Raiz quadrada da variância.
- **Interpretação:** Expressa, em média, o quanto os dados se afastam da média.
- **Como reportar:**

**Como reportar:**
O desvio padrão foi de 9,9 bpm, o que indica que, em média, os batimentos cardíacos dos indivíduos da amostra variam aproximadamente 9,9 unidades em relação à média.

### 10.2.4 Amplitude interquartil (IQR)

- **Definição:** Diferença entre o terceiro e o primeiro quartis (Q3 - Q1).

- **Interpretação:** Indica a dispersão dos 50% centrais dos dados.
- **Como reportar:**

  A amplitude interquartil foi de 10,0 bpm, mostrando a concentração dos valores médios.

**Sugestão de vídeo:** Canal Pesquise - Variabilidade

## 10.3 Medida Relativa de Variabilidade

### 10.3.1 Coeficiente de Variação (CV)

- **Definição:** Quociente entre o desvio padrão e a média, multiplicado por 100.
- **Interpretação:** Expressa a variabilidade dos dados em relação à média, permitindo comparar conjuntos com unidades diferentes.
- **Como reportar:**

  O coeficiente de variação foi de 16,9%, indicando que os dados são relativamente homogêneos.

**Observação:**
Um CV inferior a 25% geralmente indica homogeneidade; valores muito altos indicam alta variabilidade.

## 10.4 Apresentação dos Resultados

Uma maneira eficiente de apresentar estatísticas descritivas é organizar as variáveis em linhas, facilitando a visualização dos principais parâmetros de cada variável estudada. Veja abaixo uma sugestão de tabela para esse formato:

| Variável | n | Média ś DP | Mediana (Q1; Q3) | Mínimo | Máximo |
|---|---|---|---|---|---|
| Idade (anos) | 98 | 24,5 ś 4,2 | 24,0 (21,0; 28,0) | 18 | 35 |
| IMC (kg/mš) | 98 | 22,3 ś 3,1 | 21,9 (20,3; 23,7) | 17,0 | 31,5 |
| Pressão Sistólica | 98 | 118,5 ś 13,0 | 120 (110; 128) | 90 | 145 |
| Pressão Diastólica | 98 | 76,2 ś 9,1 | 76 (70; 82) | 60 | 98 |

Reportar uma medida de tendência central (como média ou mediana) junto com uma medida de dispersão (como desvio padrão, intervalo interquartil ou amplitude) é fundamental porque, isoladamente, a tendência central não fornece informações suficientes sobre o conjunto de dados.

Para variáveis qualitativas, a tabela pode ser organizada assim:

| Variável | Categoria | n | % |
|---|---|---|---|
| Sexo | Masculino | 52 | 53,1% |
| | Feminino | 46 | 46,9% |
| Tabagismo | Sim | 18 | 18,4% |
| | Não | 80 | 81,6% |

Essas tabelas permitem uma apresentação clara e objetiva das principais características da amostra analisada.

## 10.5 Funções no R

Com um vetor `x` contendo os dados, utilize:

| Medida | Código R |
|---|---|
| Média | `mean(x)` |
| Mediana | `median(x)` |
| Primeiro quartil (Q1) | `quantile(x, 0.25)` |
| Terceiro quartil (Q3) | `quantile(x, 0.75)` |
| Moda | `sort(table(x))` |
| Menor valor | `min(x)` |
| Maior valor | `max(x)` |
| Resumo geral | `summary(x)` |
| Amplitude | `range(x)` |
| Variância | `var(x)` |
| Desvio padrão | `sd(x)` |
| Amplitude interquartil | `IQR(x)` |
| Coeficiente de variação | `sd(x)/mean(x)*100` |

Calcular é importante, mas interpretar corretamente é essencial. Ao elaborar suas interpretações, descreva o que os números revelam sobre o fenômeno analisado.

## 10.6 Atividade 4

**Considere o objeto Batimentos, que é uma amostra de batimentos cardíacos de 20 homens.**

```
Batimentos <- c(62, 55, 56, 46, 75, 67, 62, 75, 60, 54, 69, 63, 39, 57, 40, 39, 64, 71, 61, 54)
```

- Obtenha as seguintes medidas:
  - Menor valor:
  - Maior valor:
  - Média:
  - Mediana:
  - Primeiro quartil:
  - Terceiro quartil:
  - Variância:
  - Desvio padrão:
  - Amplitude interquartil:
  - Coeficiente de varição:
- Escreva sobre o conjunto media e desvio padrão:

  A média dos dados foi de X (unidade), com um desvio padrão de Y (unidade), indicando que os valores estão, em geral, relativamente próximos/espalhados em torno da média. O desvio padrão reflete a quantidade de variabilidade ou dispersão dos dados em relação à média, e neste caso, a dispersão é baixa/média/alta, dependendo do valor de Y.

- Escreva sobre conjunto mediana e quartis:

  A mediana foi Z (unidade), e o intervalo interquartil (IQR), que representa a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), foi Q3 - Q1 (unidade). Isso indica que 50% dos dados estão concentrados nesse intervalo.

- Escreva sobre o coeficiente de variação:

  O coeficiente de variação (CV) foi calculado como X%, o que reflete a dispersão relativa dos dados em relação à média. Valores mais baixos de CV indicam que os dados estão mais concentrados em torno da média, enquanto valores mais altos indicam uma maior dispersão.

- Acrescente mais uma amostra com valor de batimento igual a 120, recalcule as medidas acima. Qual conjunto você consideraria mais adequado para resumir sua amostra, na presença desse valor discrepante (*outlier*)? A média (DP) ou mediana (1o.Q ; 3o.Q)? Explique.

# Chapter 11

# Importando banco de dados

Na prática, os dados que vamos analisar estarão armazenado em um **banco de dados**, um arquivo de banco de dados pode ser de diferentes tipos, por exemplo:

- Arquivo do tipo Excel (xls ou xlsx)
- Arquivo de texto separado por vírgulas (csv - *comma-separated values*)

  Existem várias fontes de dados abertas, onde podemos baixar um banco de dados para realizar analises estatísticas, aqui estão algumas delas:

- DataSus: https://datasus.saude.gov.br/transferencia-de-arquivos
- OMS: https://www.who.int/data/collections
- Kaggle: https://www.kaggle.com/datasets

  No link (google drive) existem alguns bancos que podemos usar para compreender como importar um banco de dados para o ambiente do RStudio: https://drive.google.com/drive/folders/1gyORbBEuKBstfSKULA58TLhawOXaY-st

## 11.1 Importando um banco csv

1. Faça *download* do banco de dados **mcdonald.csv** (fonte original: https://www.kaggle.com/datasets/mcdonalds/nutrition-facts)
2. Na área de ambinete de memória, localize **Import Dataset**, ao clicar nessa opção você terá o seguinte:

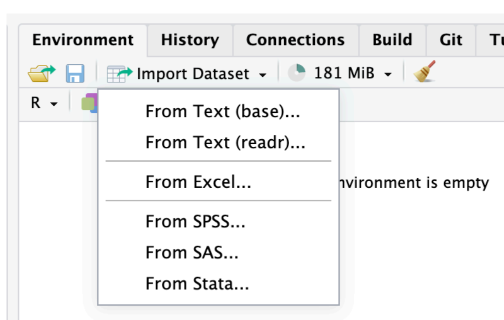



Figure 11.1: Figura: Importando banco de dados

- Como queremos importar um arquivo csv, a melhor opção é a segunda **From Text (readr)**
- ***readr*** é uma pacote do R que faz a leitura de arquivo csv (se o pacote ainda não estiver instalado no seu computador, o R fará a instalação, se você concordar!)

3. Clicando na opção **From Text (readr)**, no botão **browser** indidique onde (no seu computador) está localizado o arquivo a ser importado. A seguinte tela será apresentada:

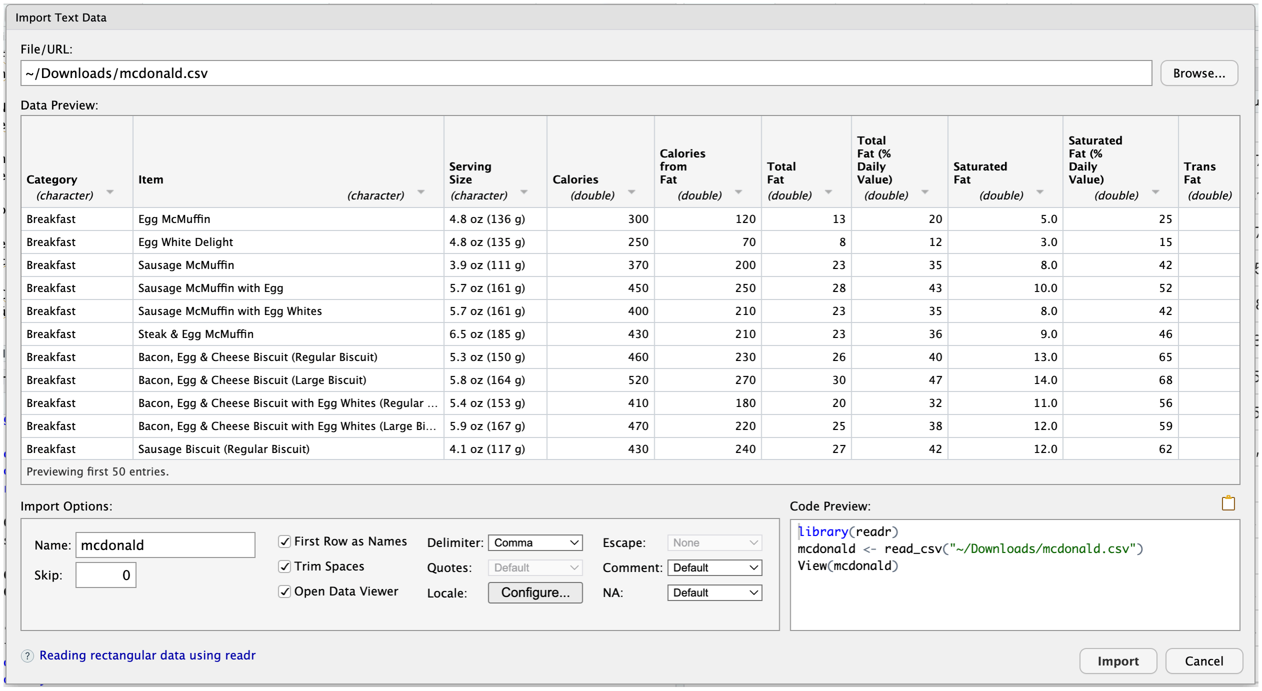



Figure 11.2: Figura: Prévia dos dados

- No quadro **Data Preview**, temos uma "prévia" com os nomes da variáveis, seus tipos computacionais e os primeiros valores que estão armazenados no banco de dados.
- No quadro **Import Options** temos as opções de importação, fique atento ao **Name** do seu banco de dados, geralmente usamos nomes sem espaços ou caracteres especiais (', ~ ou ç), é até permitido usar alguns desses caracteres especiais, mas evite.
- Ainda no quadro **Import Options**, observe que a opção **Open Data Viewer** está marcada, isso significa que ao importar o banco de dados, o arquivo de banco de dados será aberto pelo RStudio. Caso esteja trabalhando com bancos com muitos dados (como os bancos do dataSUS), talvez seja melhor desmarcar essa opção para não sobrecarregar o processamento do seu computador.
- O quadro **Code Preview** mostra como é a importação (leitura) do banco de dados via código. É interessante copiar esse trecho de código para o arquivo de script.

4. Clique no botão **Import** e observe que no ambiente de memória será criado o objeto do tipo **Data** com o nome do banco de dados que foi importado.

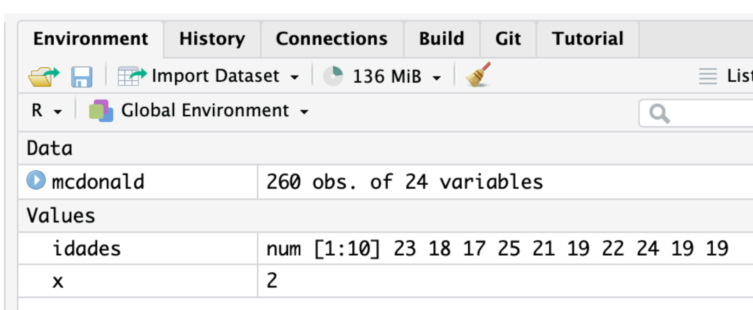



Figure 11.3: Figura: Import dataset

- Observe que esse objeto do tipo **Data** é diferente dos objetos do tipo **Values** que vimos nos exemplos iniciais.
- Ao clicar no ícone ao lado do nome do objeto, temos acesso ao nomes e tipos computacionais das variáveis, e ao clicar sobre nome do objeto, o banco será aberto!

## 11.2 Importando um banco xls

Na área de ambiente de memória, localize **Import Dataset**, ao clicar sobre essa opção, escolha **From Excel...**

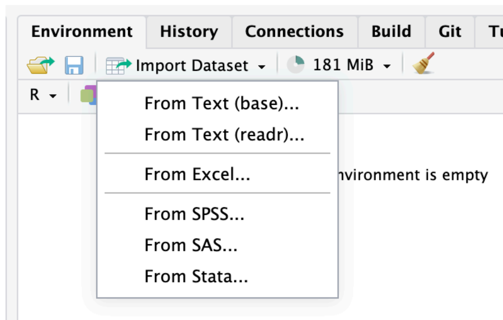



Figure 11.4: Figura: Importando banco de dados

- Se for a primeira vez que você estiver importando um arquivo Excel, pode ser necessária a instalação do pacote que fornece a biblioteca que tem a função de leitura de arquivo xls (**readxl**)! O RStudio mostrará um aviso parecido com este:

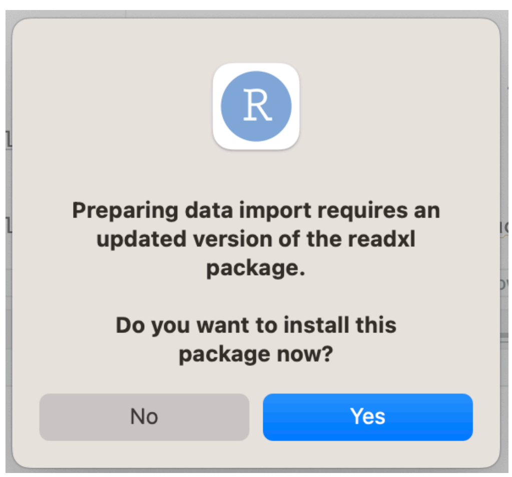



Figure 11.5: Figura: Aviso para instalação de pacote

## 11.3 Exemplo 1

Como obter a média da variável **Calories** que é uma coluna do objeto **mcdonald**, que por sua vez, é um objeto do tipo **Data**?

```
# Usamos o operador $
# Para calcular a média precisamos informar para função:
# mean( NOME DO BANCO $ NOME DA COLUNA ):
mean(mcdonald$Calories)
```

## 11.4 Exemplo 2

Como armazenar os valores de uma variável (coluna), em um objeto do tipo **Values** e depois calcular a média?

```
# Uso o operador <-
# Criamos o objeto
caloria <- mcdonald$Calories
# Agora podemos usar o objeto que criamos, por exemplo para calcular a média e o desvio padrão
mean(caloria)
sd(caloria)
```

## 11.5 Exemplo 3

O que acontece se usamos a função **summary()** para o objeto **mcdonald**, sem usar o operador, isto é sem indicar uma variável?

```
# No console será mostrado o resumo de todas as variáveis do banco!
summary(mcdonald)
```

> Essa forma de obter os resultados não é a melhor forma, vamos **instalar um pacote** para obter os resultados em uma tabela bem formatada que podemos copiar e colar diretamente para um editor de texto.

# Chapter 12

# Instalando pacotes

Quando instalamos nosso ambiente computacional R e RStudio, instalamos uma versão básica, onde apenas os recursos básicos do R estão diponíveis, o pacote básico (**base**) do R.

Os pacotes (**packages**) do R são compostos por uma biblioteca (**library**) que é um conjunto de funções. Por exemplo, do pacote **base** usamos as funções min(), max(), mean(), median(), table(), var(), sd(), summary(), etc.

Para ver a lista de funções que compõem a bilbioteca do pacote base, execute o código:

```
library(help = "base")
```

Os pacotes são análogos aos aplicativos que instalamos nos nossos celulares, são módulos que agregam funcionalidades específicas. Ao longo das nossas atividades usaremos alguns desses pacotes.

Como nesse momento estamos interessados em otimizar o trabalho para realizar uma análise descritiva dos dados, então vamos instalar um pacote chamado **gtsummary** (https://www.danieldsjoberg.com/gtsummary/).

> O pacote **gtsummary** nos fornecerá uma tabela resumo de todo banco de dados, otimizando bastante nosso trabalho de resumir o banco de dados.

- IMPORTANTE 1: instalamos um pacote apenas uma vez (como um aplicativo no celular… a gente só refaz a instalação se o app *bugar*!)
- IMPORTANTE 2: todas vez precisamos carregar o pacote com as funções que queremos usar por meio da função **library()**

Veja o código:

```
# comando para instalar o pacote gtsummary
install.packages("gtsummary")

# comando para carregar a biblioteca de funções do gtsummary
library(gtsummary)

# a função que vamos usar para gerar uma tabela que resume os dados é
# tbl_summary
tbl_summary(mcdonald)
```

- Ao executar **tbl_summary(mcdonald)** a tabela de resultados será mostrada na área de arquivos, gráficos, pacotes… na aba **Viewer**, no quadrante abaixo do ambiente de memória.
- Essa tabela pode ser copiada e colada para o editor de texto que você utiliza para escrever seus trabalhos, claro essa tabela pode ser melhorada!
- Observe no rodapé da tabela a seguinte legenda **n (%); Median (IQR)**, isso significa que para

- **variáveis qualitativas:** n é a contagem (frequência absoluta) e entre parenteses (%) é mostrado a porcentagem de cada categoria.
- **variveis quantitativas:** Median é a mediana e entre parenteses (IQR - de InterQuantile Range) estão o primeiro e terceiro quartil respectivamente.

## 12.1 Exemplo 1

Como mostrar o resultado com a média e desvio padrão?

```
# acrescente nos argumentos da função tbl_summary() a opção:
# statistic = list(all_continuous() ~ "{mean} ({sd})"
tbl_summary(
            mcdonald,
            statistic = list(all_continuous() ~ "{mean} ({sd})")
            )
```

## 12.2 Exemplo 2

Como selecionar somente algumas variáveis do banco de dados?

```
# Precisamos do pacote tidyverse, tire o símbolo de # se precisar instalar!
# install.packages("tidyverse")

# ative tidyverse
library(tidyverse)

# vamos usar a função select() do pacote tidyverse
dadosSelecionados <- select (mcdonald, Cholesterol, Sodium, Carbohydrates)

# faça uma tabela para o objeto dadosSelecionados
tbl_summary(dadosSelecionados)
```

## 12.3 Exemplo 3

Algumas vezes é mais fácil excluir algumas variáveis, por exemplo queremos todas, menos **Item** e **Serving Size**

```
# vamos usar a função select() do pacote tidyverse e colocar o sinal de menos (-)
# antes dos nomes das variáveis que queremos excluir
# IMPORTANTE: Serving Size é um nome de variável com espaço
# então devemos referênciá-la entre aspas: `Serving Size`
dadosSelecionados2 <- select (mcdonald, -Item, -`Serving Size`)

# faça uma tabela para o objeto dadosSelecionados2
tbl_summary(dadosSelecionados2)
```

## 12.4 Exemplo 4

Como selecinar um conjunto de variáveis que estão em sequência, por exemplo, de **Carbohydrates** a **Cholesterol (% Daily Value)**

```
# vamos usar a função select() do pacote tidyverse e colocar o sinal de dois pontos (:)
# entre a primeira variável e a última da sequência
# IMPORTANTE: Cholesterol (% Daily Value) é um nome de variável com espaço
# então devemos referênciá-la entre aspas: `Cholesterol (% Daily Value)`
```

```
dadosSelecionados3 <- select (mcdonald, Carbohydrates:`Cholesterol (% Daily Value)`)

# faça uma tabela para o objeto dadosSelecionados
tbl_summary(dadosSelecionados3)
```

Saiba mais sobre o Tidyverse https://www.tidyverse.org/packages/

## 12.5 Atividade 5

**Escolha outro banco de dados (você pode até criar um banco fictício!), faça uma tabela descritiva dos dados e escreva sobre os dados (um ou dois parágrafos), afinal, o nosso trabalho não é só obter a tabela, é dissertar sobre o que essa tabela revela sobre a amostra em estudo!**

# Chapter 13

# Gráficos

Nesse link https://r-graph-gallery.com/ está algumas possibilidades de gráficos que podemos fazer usando o R. Para fazer gráficos mais elaborados (aparentemente mais atrativos visualmente) usamos o pacote **GGPlot2** https://ggplot2.tidyverse.org/.

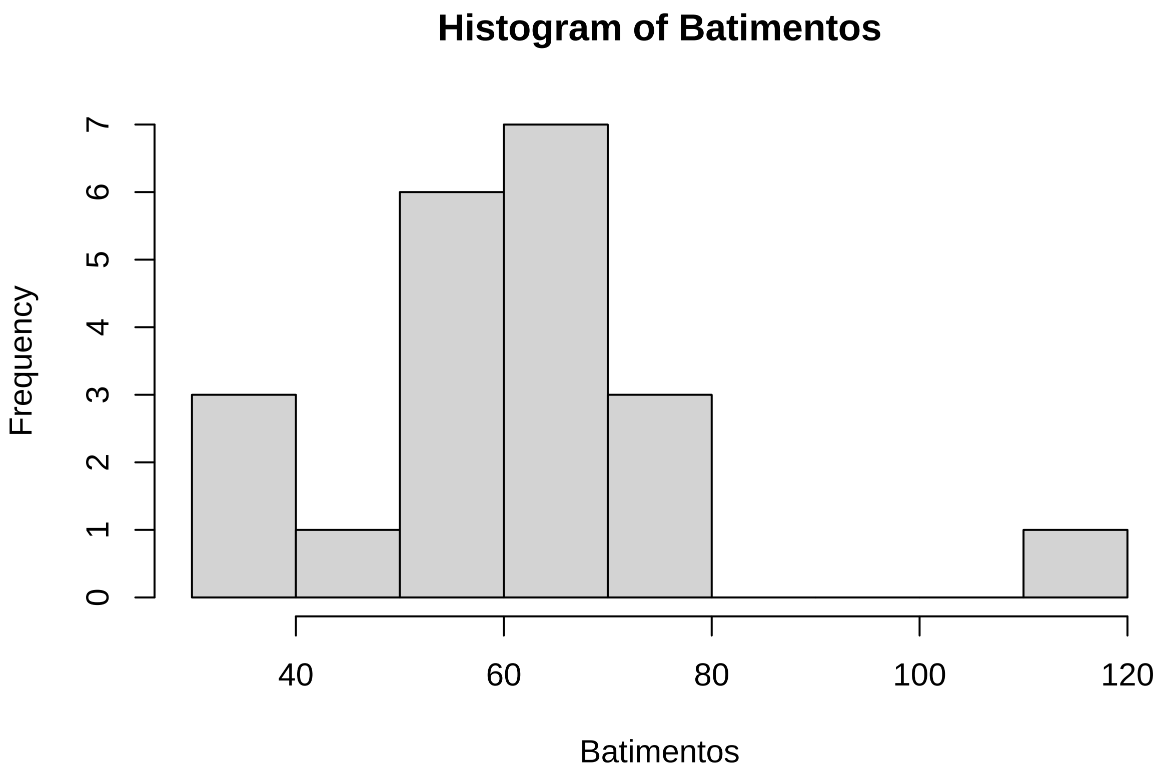
Focaremos nossa atenção em dois gráficos específicos para variáveis quantitativas: **Histograma** e **Boxplot**, em nem faremos nada atrativo, usaremos o pacote básico do R que nos fornece as funções **hist()** e **boxplot()**, pois o nosso obtivo para esse momento é simplesmente estudar a importância desses gráficos.

O que a gente levaria um tempinho... é simplesmente assim em código R:

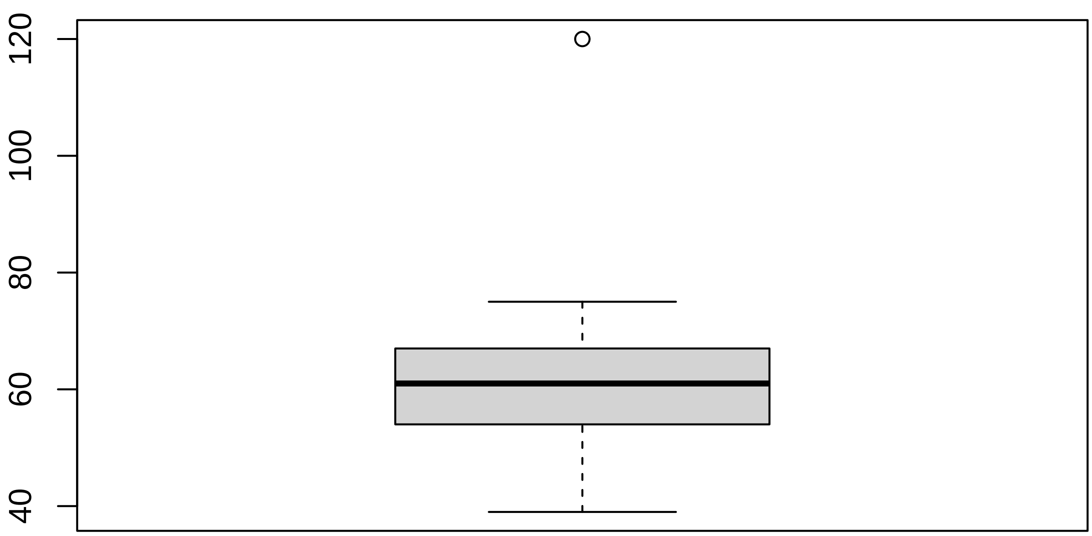
```
Batimentos <- c(62, 55, 56, 46, 75, 67, 62, 75, 60, 54, 69, 63, 39, 57, 40, 39, 64, 71, 61, 54, 120)

# Para fazer o Histograma de Batimentos
hist(Batimentos)

# Para fazer o Boxplot de Batimentos
boxplot(Batimentos)
```

Na área de gráficos (**Plots**), abaixo do ambiente de memória, serão mostrados os gráficos:

Histograma




Boxplot




Observações

- Os gráficos mostram a informação batimentos de duas formas diferentes, mas elas estão relacionadas!
- Observe que eixo horizontal do histograma corresponde ao eixo vertical do boxplot

## 13.1 Histograma

O histograma é um gráfico que usado para variáveis quantitativas contínua.

O histograma pode nos dar uma noção do tipo de **distribuição de probabibilidade** que os dados seguem.

A ideia desse gráfico é agrupar os dados em **classes** (cada barra do histograma é uma classe) e no eixo vertical tem-se a contagem (frequência) de quantos valores foram alocados em cada classe.

Para fazer a **leitura do histograma**:

- Identifique as classes no "eixo x"
- Identifique quantos elementos tem em cada classe no "eixo y"

Acredito que nesse exemplo, é fácil verificar:

- A segunda classe: 40 - 50 batimentos, que tem 1 elemento (verifique no objeto Batimentos)
- A terceira classe: 50 - 60 batimentos, que tem 6 elementos
- Então, a **aplitude das classes** é igual a 10. Logo, a primeira classe é de 30 - 40.
- As classes 80 - 90; 90 - 100 e 100 - 110 não tiveram ocorrências!
- A classe 110-120 possui 1 elemento, que é aquele valor discrepante em relação aos demais valores.

Se não for fácil identificar as classes (eixo x) você pode usar o comando abaixo:

```
# Para obter as "quebras" de cada classe
hist(Batimentos)$breaks
```

Se não for fácil identificar as frequencias (eixo y) você pode usar o comando abaixo:

```
# Para obter a frequência em cada classe
hist(Batimentos)$count
```

De fato, o que estamos lendo por meio do histograma é o que chamamos de **tabela de frequência**:

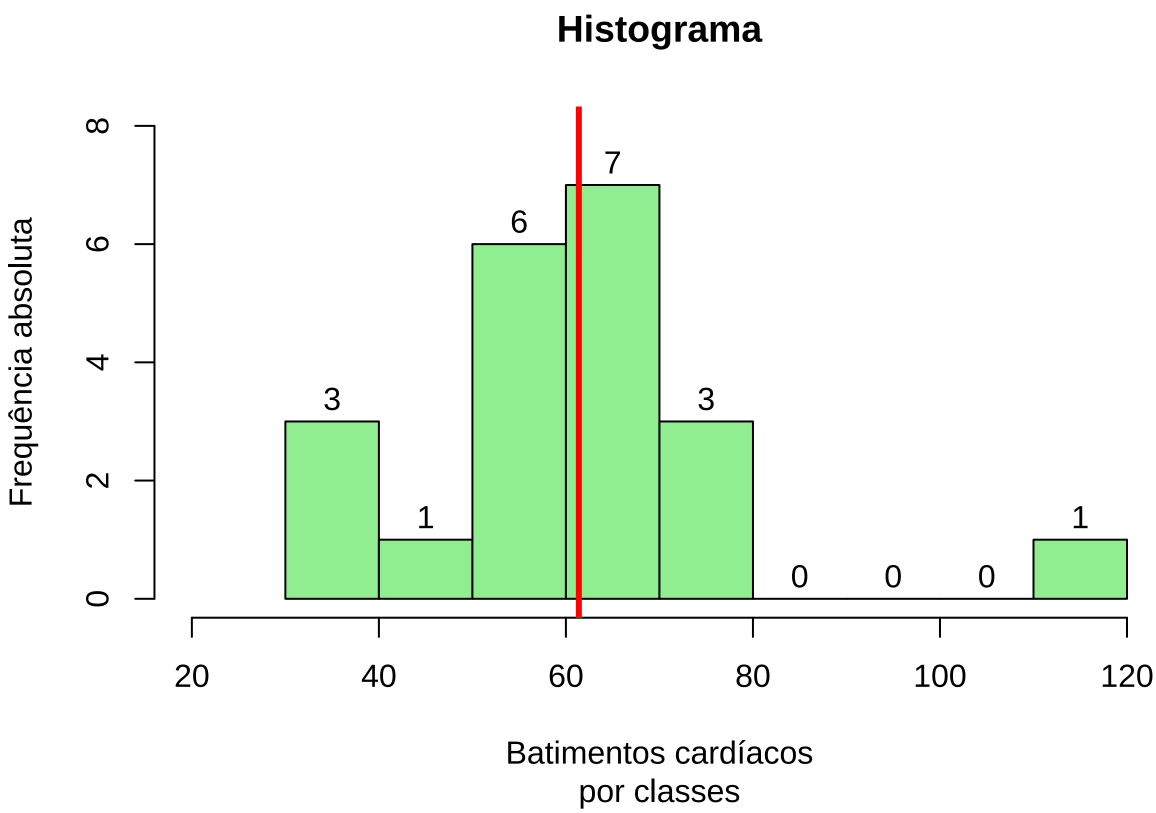
| Classe | Frequência |
|---|---|
| 30 - 40 | 3 |
| 40 - 50 | 1 |
| 50 - 60 | 6 |
| 60 - 70 | 7 |
| 70 - 80 | 3 |
| 80 - 90 | 0 |
| 90 - 100 | 0 |
| 100 - 110 | 0 |
| 110 - 120 | 1 |
| $n$ | 21 |

- Por meio do histograma ou da tabela podemos concluir que a classe modal (moda) é a classe de 60 - 70 batimentos;
- A frequência foi apresentada em termos absolutos mais pode ser transformada em frequência percentual.
- Quando estamos aprendendo a fazer um histograma manualmente, primeiro construímos essa tabela de frenquência, e para construí-la é necessário calcular o número ótimo de classes, umas das regras mais usada é a Regra Sturges (essa é opção padrão do R).

Podemos usar o pacote básico R para melhorar a aparência desse gráfico.

```
hisBat <- hist(Batimentos,
               main = "Histograma",
               xlab = "Batimentos cardíacos",
               sub = "por classes",
               ylab = "Frequência absoluta",
               xlim = c(20, 120),
               ylim = c(0, 8),
               col = "lightgreen")
text(hisBat$mids, hisBat$counts, labels=hisBat$counts, adj = c(0.5,-0.5))

# adicionar linha para indicar a média
abline(v = mean(Batimentos),
```

```
       col = "red",
       lwd = 3)
```




## 13.2 Boxplot

Boxplot ou diagrama de caixa, é um gráfico que mostra as medidas: menor valor, primeiro quartil, mediana, terceiro quartil e máximo valor.

- Valores discrepantes (*outliers*) são detectados pelo boxplot. Veja a figura:

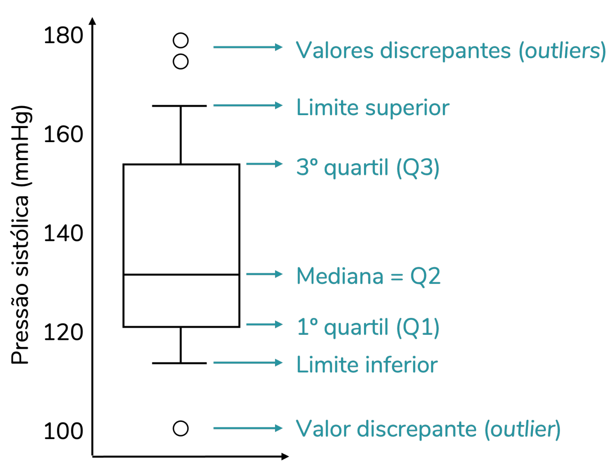



Figure 13.1: Figura: "Anatomia de um boxplot"

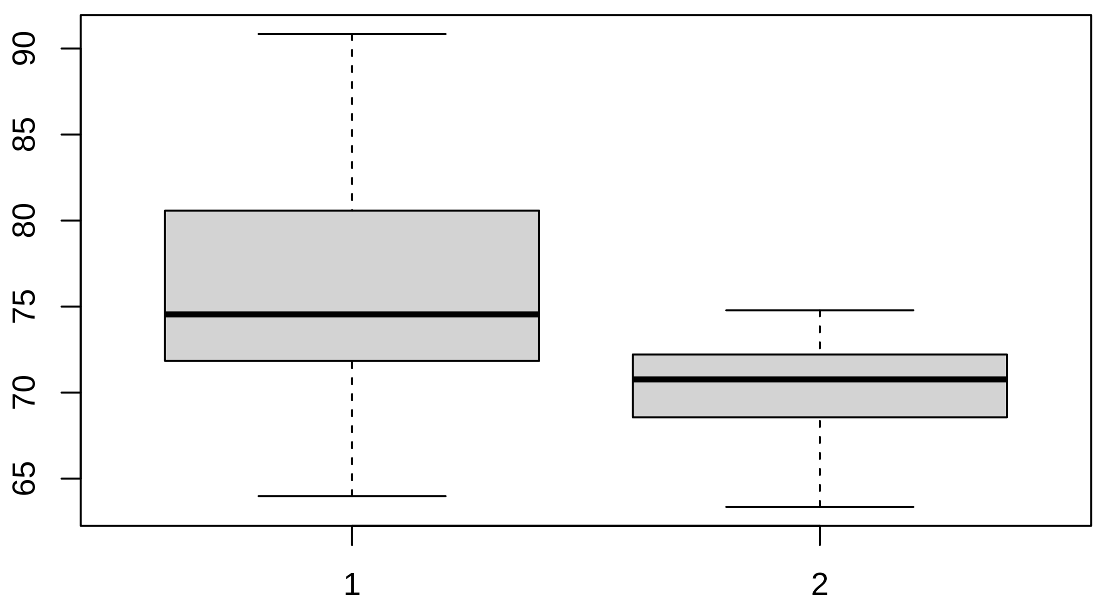
Essa figura foi retirada do site da Prof. Fernanda https://fernandafperes.com.br/blog/interpretacao-boxplot/ (uma exelente referência para estudar estatística!)

Geralmente eles são representados na vertical, mas também é comum a representação na horizontal.

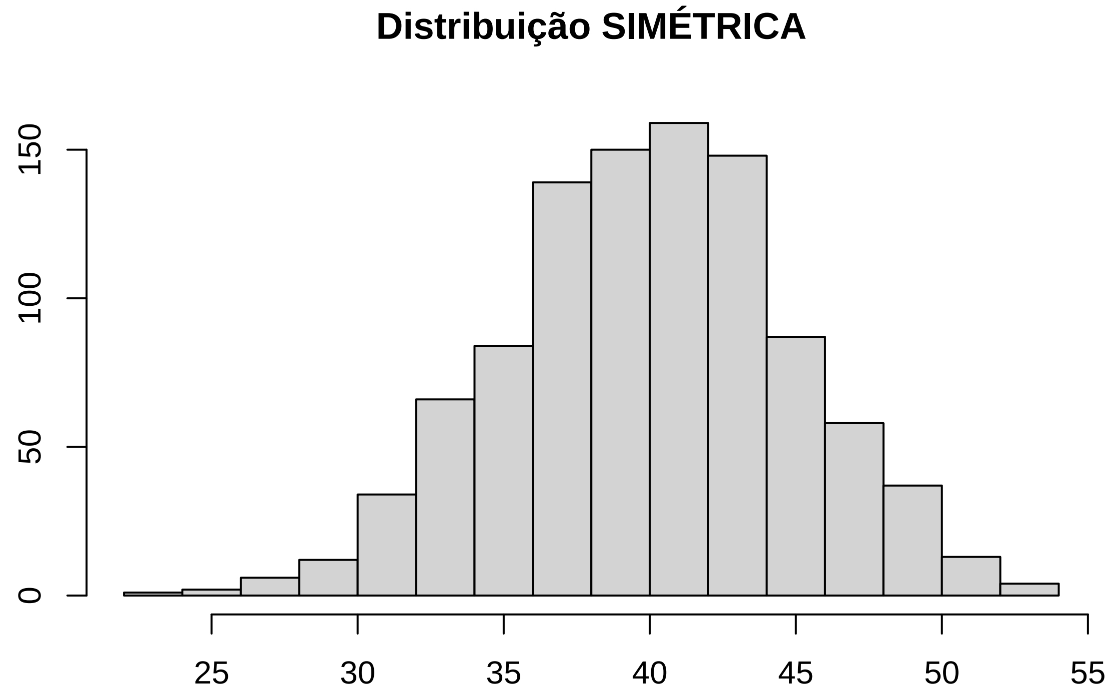
```
# Para fazer o Boxplot de Batimentos na horizontal
boxplot(Batimentos, horizontal = TRUE)
```

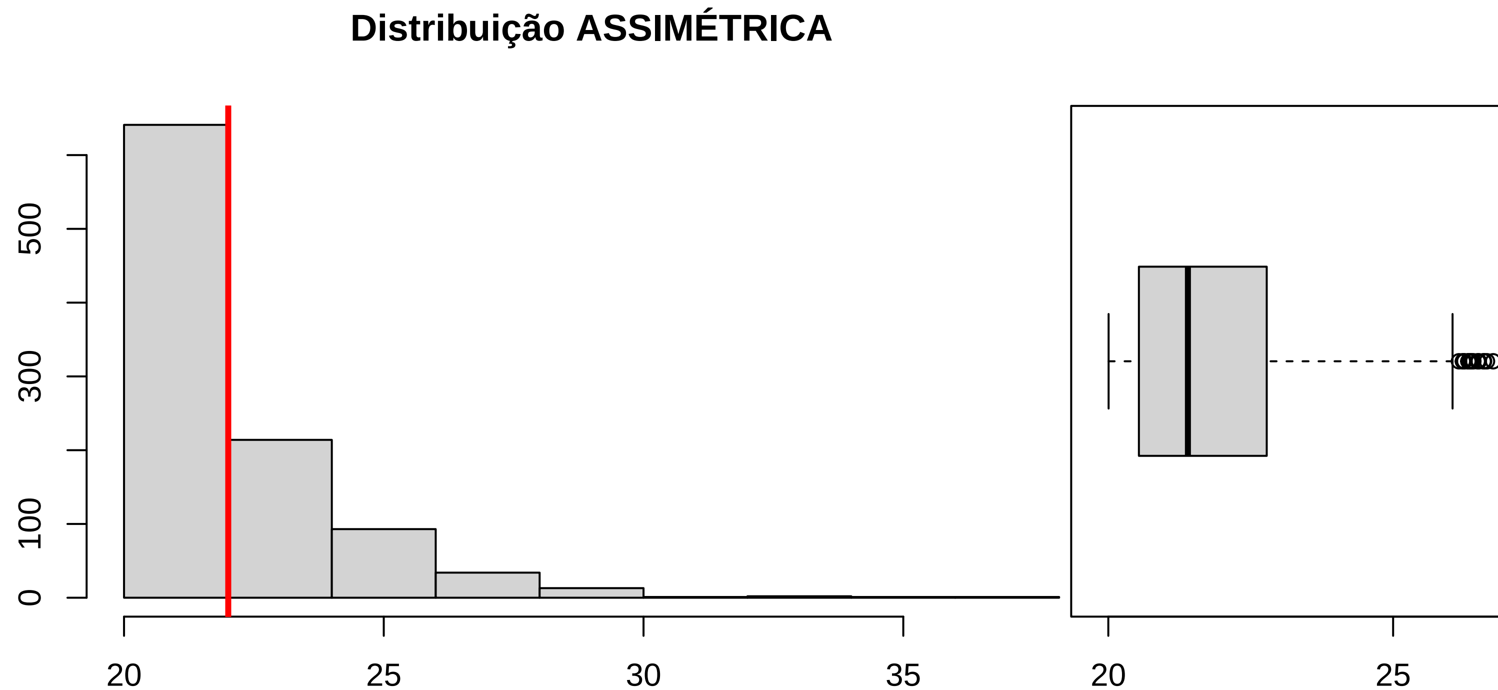
É uma forma de comparar dois grupos em relação a uma medida, por exemplo os batimentos cardiacos de grupo de homens e de mulheres

```
# Geração de amostras simuladas
set.seed(1)
BatimentosMulheres <- rnorm(30, 70, 3)
BatimentosHomens <- rnorm(30, 75, 8)
# Boxplot para os dois grupos Homens e Mulheres
boxplot(BatimentosHomens, BatimentosMulheres)
```




O boxplot também pode nos informar se uma distribuição de probabilidade é simétrica ou não. Analise os gráficos abaixo, veja a conexão entre histograma e boxplot.




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   20.54   21.40   22.01   22.78   36.42
```

## 13.3 Atividade 6

**Para o banco de dados escolhido na atividade 5, faça gráficos como o histograma e boxplot, além disso, pesquise outras formas de fazer gráficos no R.**

# Chapter 14

# Pacote `ggplot`

O `ggplot2` é um dos pacotes mais populares do R para a criação de gráficos estatísticos sofisticados e visualmente atraentes. Baseado no conceito da "Gramática dos Gráficos", o `ggplot2` permite construir gráficos de maneira flexível e modular, combinando camadas (layers) de dados, geometrias, escalas e temas.

## 14.1 Principais Vantagens

- Produz gráficos de alta qualidade e personalizáveis.
- Permite adicionar camadas de informação facilmente.
- Suporta uma variedade de tipos de gráficos.

## 14.2 Exemplos de Gráficos Simples e Bonitos com ggplot2

Antes de começar, certifique-se de instalar e carregar o pacote:

```
install.packages("ggplot2")
library(ggplot2)
```

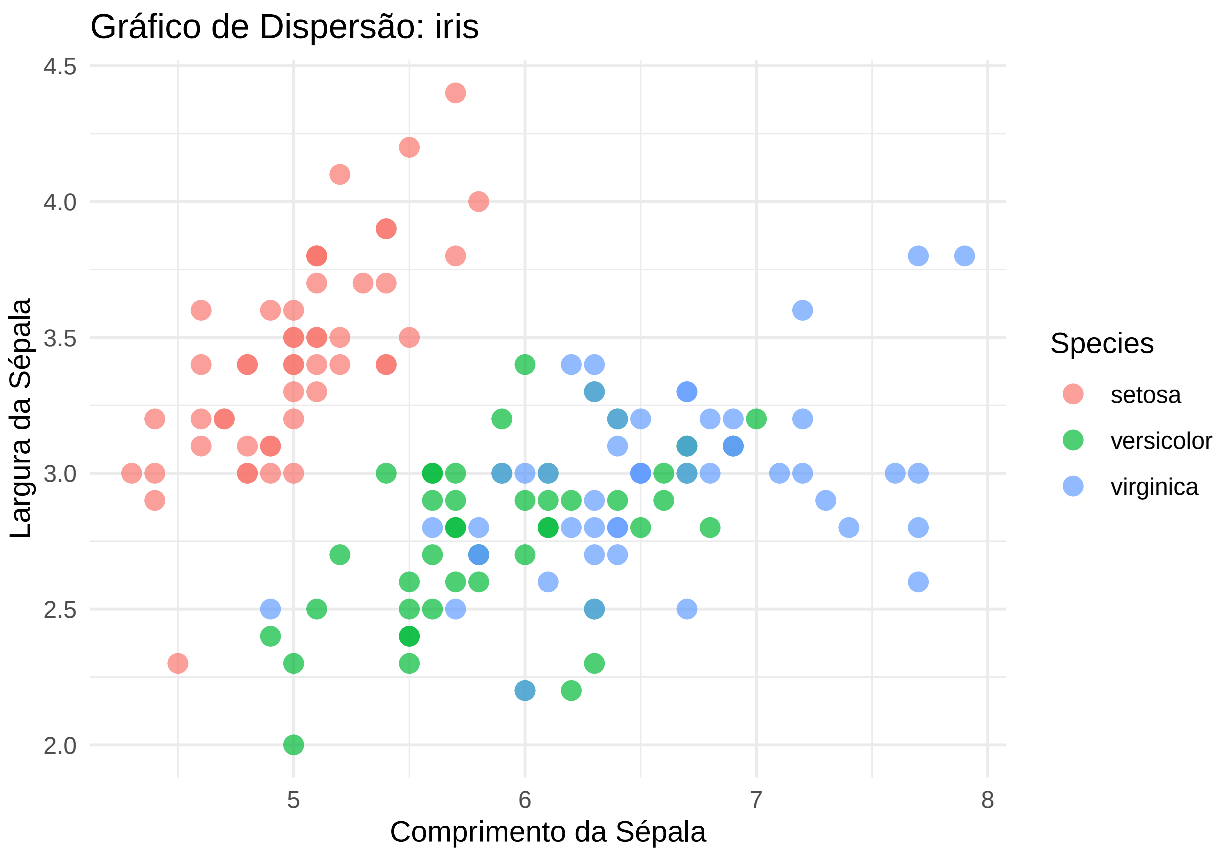
Para ilustrar alguns gráficos utilizando o `ggplot2`, empregaremos o conjunto de dados *iris*, que já está disponível nativamente no R.

O *iris* é um dos bancos de dados mais clássicos e utilizados em estatística e aprendizado de máquina, trazendo informações sobre 150 flores de três espécies diferentes, com quatro variáveis quantitativas (comprimento e largura das sépalas e pétalas). Ele é amplamente utilizado para exemplos de análise exploratória de dados, classificação e visualização de padrões.

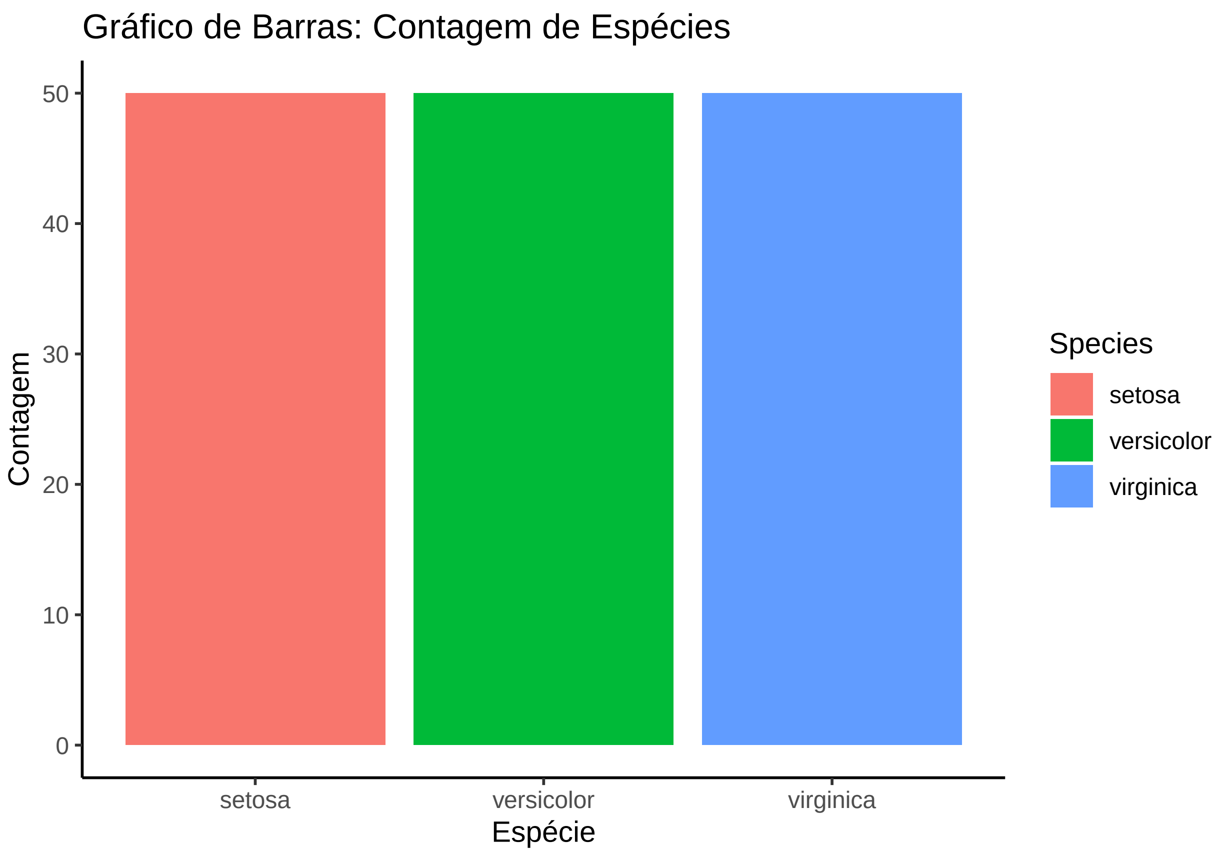
Além do iris, o R oferece outros conjuntos de dados nativos bastante úteis para exercícios e demonstrações em diversas áreas, como *mtcars* (carros), *ToothGrowth* (crescimento de dentes em cobaias), *sleep* (efeito de medicamentos no sono), *airquality* (qualidade do ar em Nova Iorque) e *CO2* (absorção de CO2 em plantas).

### 14.2.1 Gráfico de Dispersão (Scatterplot)

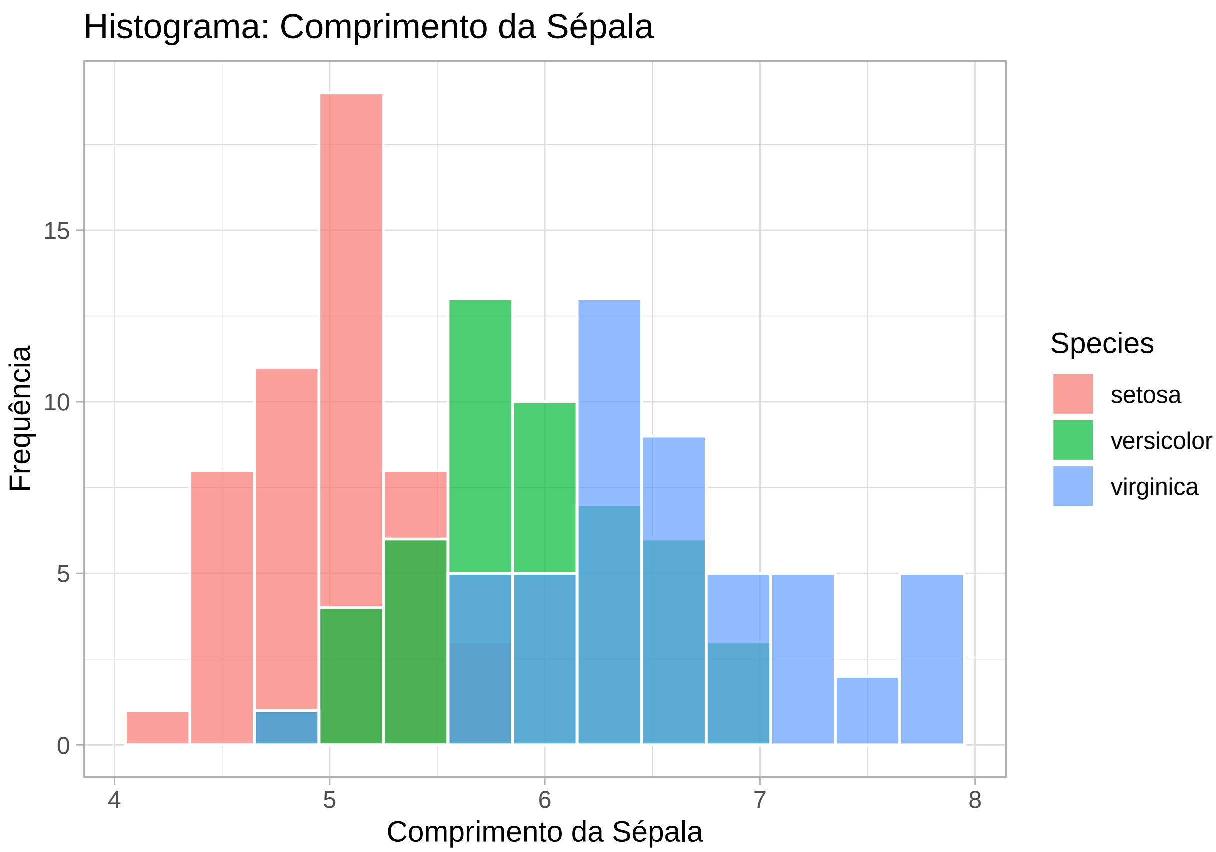
```
library(ggplot2)
# Exemplo com o dataset iris
ggplot(iris, aes(x = Sepal.Length, y = Sepal.Width, color = Species)) +
  geom_point(size = 3, alpha = 0.7) +
  theme_minimal() +
  labs(title = "Gráfico de Dispersão: iris",
       x = "Comprimento da Sépala",
       y = "Largura da Sépala")
```



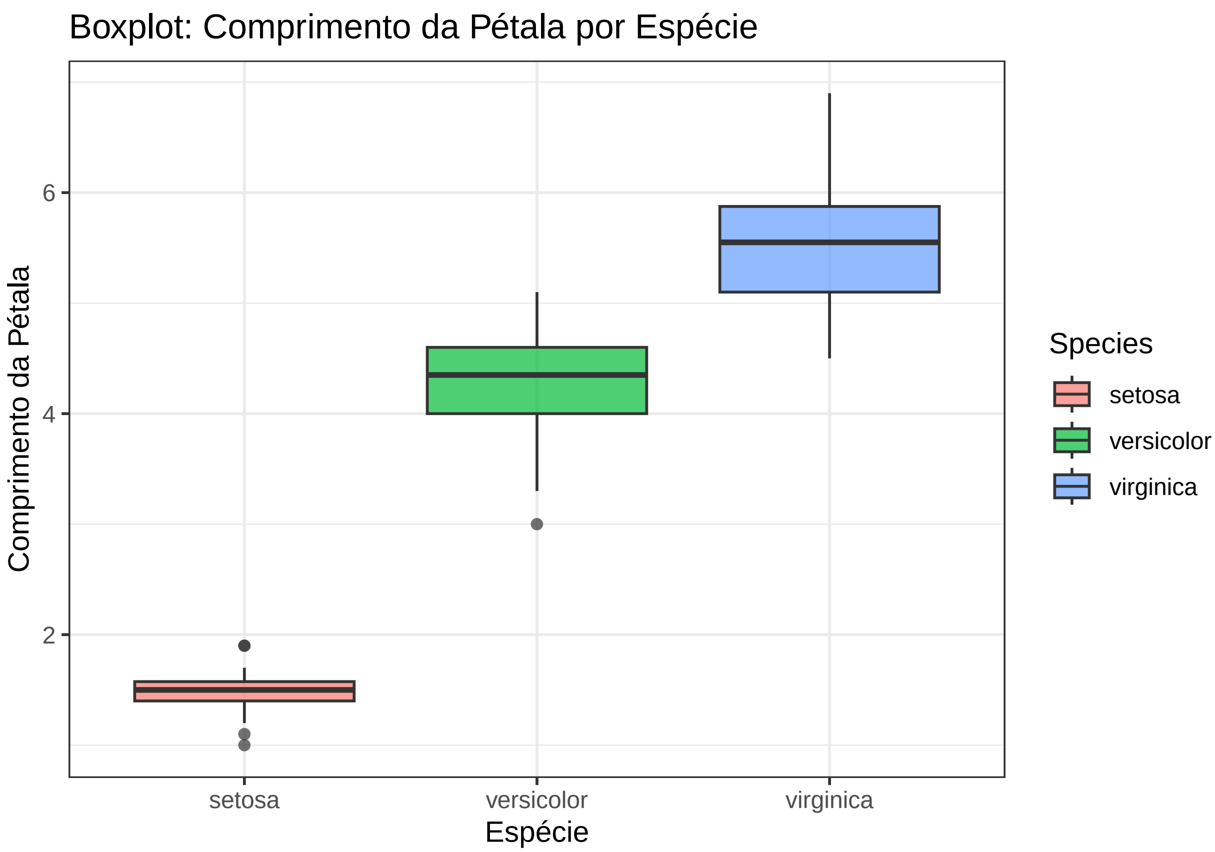

```
# Contagem das espécies no dataset iris
ggplot(iris, aes(x = Species, fill = Species)) +
  geom_bar() +
  theme_classic() +
  labs(title = "Gráfico de Barras: Contagem de Espécies",
       x = "Espécie",
       y = "Contagem")
```




```r
# Distribuição do comprimento da sépala
ggplot(iris, aes(x = Sepal.Length, fill = Species)) +
  geom_histogram(binwidth = 0.3, color = "white", alpha = 0.7, position = "identity") +
  theme_light() +
  labs(title = "Histograma: Comprimento da Sépala",
       x = "Comprimento da Sépala",
       y = "Frequência")
```




```
# Distribuição do comprimento da pétala por espécie
ggplot(iris, aes(x = Species, y = Petal.Length, fill = Species)) +
  geom_boxplot(alpha = 0.7) +
  theme_bw() +
  labs(title = "Boxplot: Comprimento da Pétala por Espécie",
       x = "Espécie",
       y = "Comprimento da Pétala")
```




O `ggplot2` oferece diversas opções de personalização, como cores, temas e anotações, permitindo criar gráficos bonitos e informativos para diferentes finalidades.

## 14.3 Exemplos de Bancos de Dados do R

| Banco de Dados | Pacote | Área/Descrição |
|---|---|---|
| iris | datasets | Botânica, estatística, aprendizado de máquina (flores) |
| mtcars | datasets | Automóveis, regressão, análise multivariada |
| airquality | datasets | Qualidade do ar, saúde ambiental |
| ToothGrowth | datasets | Farmacologia, saúde, crescimento de dentes |
| sleep | datasets | Psicologia, farmacologia, estudo do sono |
| ChickWeight | datasets | Nutrição, crescimento animal |
| USArrests | datasets | Sociologia, estatísticas criminais dos EUA |
| CO2 | datasets | Biologia, fisiologia vegetal |
| BOD | datasets | Biologia, demanda bioquímica de oxigênio |
| Boston | MASS | Imobiliário, regressão, análise multivariada |
| lung | survival | Medicina, análise de sobrevivência (dados de câncer de pulmão) |
| bfi | psych | Psicologia, personalidade (Big Five Inventory) |
| NHANES | NHANES | Saúde pública, epidemiologia (pesquisa nacional dos EUA) |
| titanic | titanic | Sobrevivência, estatística, aprendizado de máquina |
| worldcup | faraway | Esportes (Copa do Mundo de Futebol) |

**Nota:**
Os bancos do pacote `datasets` já vêm instalados por padrão no R. Outros, como `MASS`, `psych`, `NHANES`, `survival`, `titanic` e `faraway`, podem ser instalados via `install.packages("nome_do_pacote")`.
Verifique sempre a documentação do pacote para acessar o nome correto dos bancos de dados e exemplos de uso.

# Chapter 15

# Distribuição de Probabilidade

Uma distribuição de probabilidade é um modelo matemático que descreve a relação entre os valores possíveis de uma variável e as probabilidades de ocorrência desses valores. Essencialmente, ela nos permite prever a probabilidade de eventos com base em um conjunto de dados.

Entre as distribuições mais comuns, destacam-se:

- Distribuição Normal (ou Gaussiana)
- Distribuição Binomial
- Distribuição Poisson
- Distribuição Exponencial
- Distribuição Uniforme
- Distribuição Qui-quadrado
- Distribuição t-Student
- Distribuição Gama, entre outras.

Cada tipo de distribuição possui características próprias e é aplicada em diferentes contextos. A distribuição normal é uma das mais amplamente utilizadas, especialmente para modelar fenômenos naturais e sociais, como a altura de indivíduos ou o tempo de vida de produtos.

**Propriedades Gerais das Distribuições de Probabilidade**

- **Área total sob a curva é igual a 1:** Isso significa que a soma de todas as probabilidades possíveis de ocorrência dos eventos é igual a 100%.
- **A área sob a curva representa a probabilidade de um evento.** Por exemplo, a probabilidade de um evento ocorrer entre dois valores quaisquer pode ser calculada pela área sob a curva entre esses dois pontos.

## 15.1 Distribuição Normal

A distribuição normal é uma das mais conhecidas na estatística. Ela modela fenômenos que seguem um comportamento simétrico em torno de uma média. A distribuição normal tem várias aplicações, desde a análise de medidas biológicas (como a pressão arterial) até a previsão de fenômenos econômicos (como o preço das ações).

**A distribuição normal é definida por dois parâmetros principais:**

- **Média ():** Representa o valor central da distribuição.
- **Desvio Padrão ():** Mede a dispersão dos dados em relação à média.

**Características da Distribuição Normal**

- **Forma de sino:** A distribuição normal tem a forma de um sino simétrico, com a maior concentração de dados perto da média.
- **Simetria:** A distribuição é simétrica em relação à média.

- **Média, Mediana e Moda:** Para uma distribuição normal, a média, mediana e moda são "iguais" e localizam-se no centro da distribuição.

  **Teoricamente o comportamento da Distribuição Normal é dado por:**

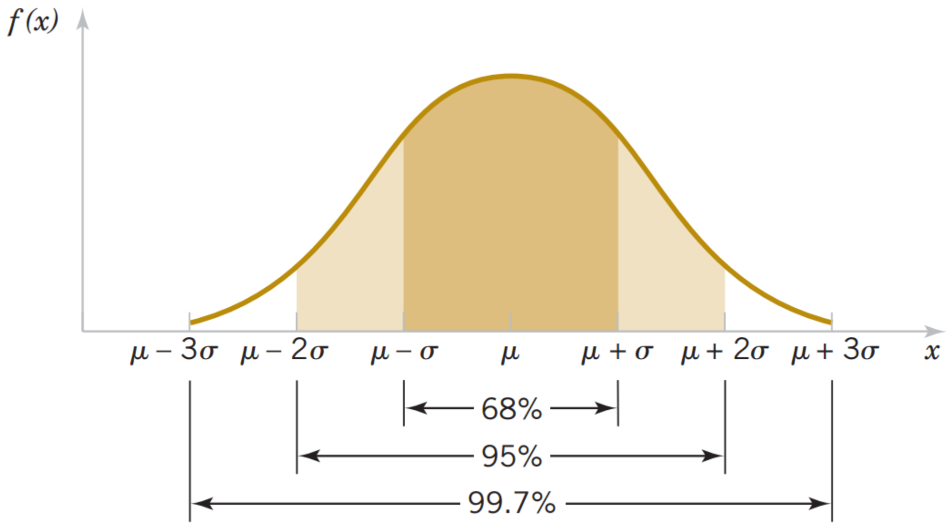



Figure 15.1: Figura: Distribuição Normal teórica (Fonte: https://www.inf.ufsc.br/~andre.zibetti/probabilidade/figures/normal.PNG)

Onde:

- A média está representada por
- O devio padrão está representado por

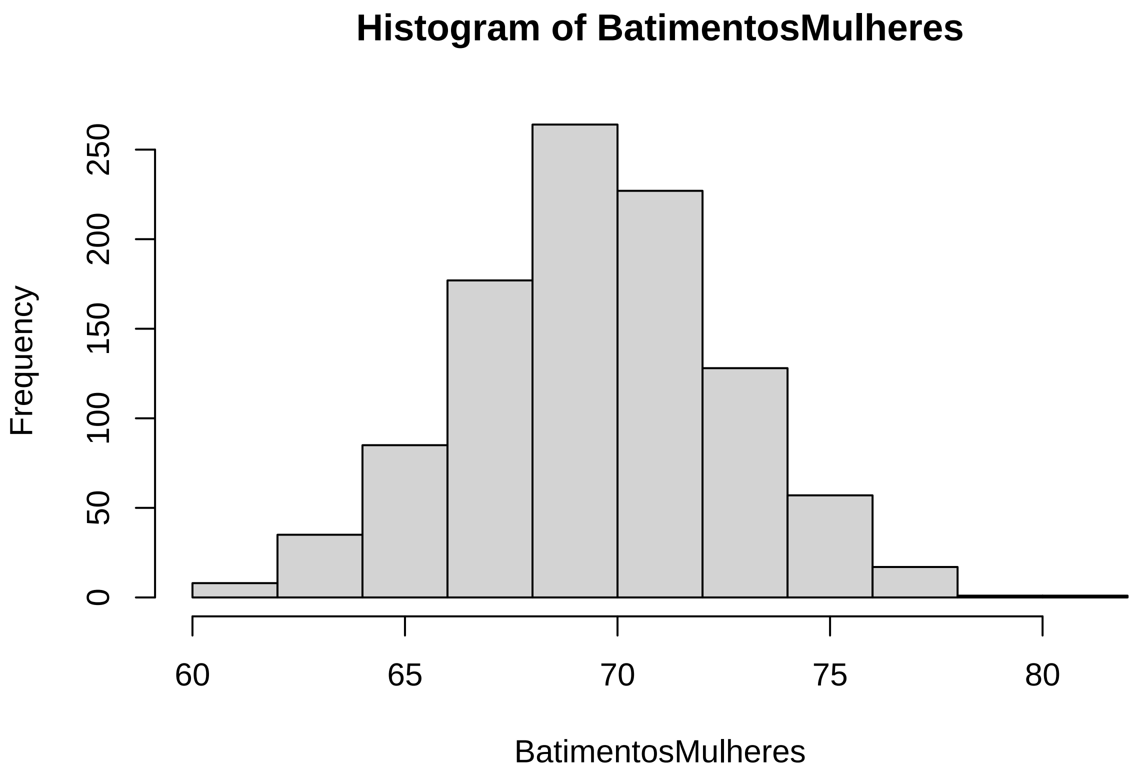
  **Regra empirica 68% - 95% - 99,7%** Se os dados seguem uma distribuição normal, é possível fazer afirmações sobre a concentração dos dados em torno da média, conforme a regra empírica:

- *68%* dos dados estão no intervalo de uma vez o desvio padrão ( ś 1).
- *95%* dos dados estão no intervalo de duas vezes o desvio padrão ( ś 2).
- *99,7%* dos dados estão no intervalo de três vezes o desvio padrão ( ś 3).

**Importante:** Dados fora do intervalo de ś 3 são considerados raros.

Exemplo de distribuição Normal, com dados simulados usado a função rnorm().

```
# semente de geração de números aleatórios
set.seed(1)

# Será simulada uma amostra com a seguinte característica:
# 1000 valores
# média ~ 70
# desvio padrão ~ 3
# A função rnorm() gera números randômicos com comportamento de uma distribuição Normal
BatimentosMulheres <- rnorm(1000, 70, 3)

# Arrendodamento com nenhuma casa depois da vírugula
BatimentosMulheres <- round(BatimentosMulheres,0)

# histograma
hist(BatimentosMulheres)
```



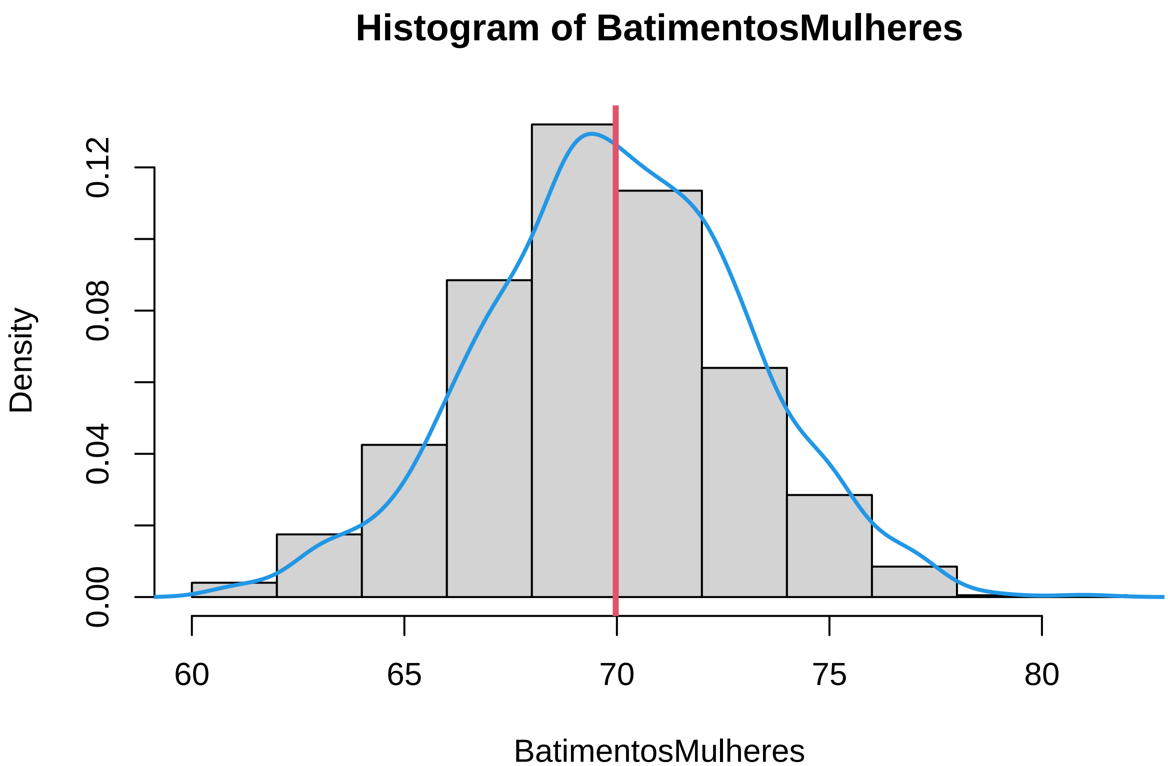

```
# Classes e frequencias do histograma
hist(BatimentosMulheres)$breaks
```

```
##  [1] 60 62 64 66 68 70 72 74 76 78 80 82
```

```
hist(BatimentosMulheres)$count
```

```
##  [1]   8  35  85 177 264 227 128  57  17   1   1
```

```
# histograma e curva de densidade (da Dist. Normal)
hist(BatimentosMulheres, prob = TRUE)
lines(density(BatimentosMulheres), col = 4, lwd = 2)

# idicação da média
abline(v = mean(BatimentosMulheres), col = 2, lwd = 3)
```




```
# medidas resumo
summary(BatimentosMulheres)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   61.00   68.00   70.00   69.97   72.00   81.00
```

```
sort(table(BatimentosMulheres), decreasing = T)
```

```
## BatimentosMulheres
##  69  70  71  72  68  67  73  66  74  75  65  64  63  76  77  61  62  78  79  81
## 138 126 116 111  95  82  82  56  46  41  29  18  17  16  15   4   4   2   1   1
```

```
sd(BatimentosMulheres)
```

```
## [1] 3.10594
```

```
# CV em %
sd(BatimentosMulheres)/mean(BatimentosMulheres)*100
```

```
## [1] 4.438833
```

As funções **pnorm()** e **dnorm()** são usadas para calcular a probabilidade de um evento que segue uma distribuição, a qual conhecemos a média e o desvio padrão.

**Exemplo:** Sabendo que os batimentos cardíacos de mulheres de 18 a 65 anos tem média de 70bmp e desvio padrão igual a 3bmp.

Calcule as probabilidades:

- de uma mulher ter batimentos inferior a 70bmp, ou seja, $P(x < 70)$:

```
# pnorm(): Calcula a probabilidade acumulada até um valor específico. Ou seja, retorna a probabilidade de

# observação: a resposta é 0.5 pois a média 70.
pnorm(70, 70, 3)
```

```
## [1] 0.5
```

- de uma mulher ter batimentos superior a 70bmp, ou seja, $P(x > 70)$:

```
# observação: 1 é o valor da área total
# a pnorm() fornece a área á esquerda
1 - pnorm(70, 70, 3)
```

```
## [1] 0.5
```

- de uma mulher ter batimentos igual a 70bmp, ou seja, $P(x = 70)$:

```
# dnorm(): Calcula a densidade de probabilidade de um valor específico em uma distribuição normal. En
dnorm(70, 70, 3)
```

```
## [1] 0.1329808
```

- de uma mulher ter batimentos entre 67 e 73bmp $P(67 < x < 73)$:

```
# Observe que estamos testando a regra empírica (68%)
pnorm(73, 70, 3) - pnorm(67, 70, 3)
```

```
## [1] 0.6826895
```

- de uma mulher ter batimentos entre 67 e 73bmp $P(64 < x < 76)$:

```
# Observe que estamos testando a regra empírica (95%)
pnorm(76, 70, 3) - pnorm(64, 70, 3)
```

```
## [1] 0.9544997
```

- de uma mulher ter batimentos entre 61 e 79bmp $P(61 < x < 79)$:

```
# Observe que estamos testando a regra empírica (99,7%)
pnorm(79, 70, 3) - pnorm(61, 70, 3)
```

```
## [1] 0.9973002
```

- de uma mulher ter batimentos maior que 90bmp $P(x > 90)$:

```
# Um evento raro
1 - pnorm(90, 70, 3)
```

```
## [1] 1.308398e-11
```

- de uma mulher ter batimentos menor que 65bmp $P(x < 65)$:

```
pnorm(65, 70, 3)
```

```
## [1] 0.04779035
```

## 15.2 Gráfico QQ

Uma maneira comum de verificar a normalidade de uma distribuição é utilizando o gráfico QQ. Ele compara os quantis de uma amostra com os quantis de uma distribuição normal padrão, e é muito útil para identificar se os dados seguem uma distribuição normal.

**Distribuição Normal Padrão**

A distribuição normal padrão tem duas características importantes:

- Média () igual a 0
- Desvio padrão () igual a 1

**Cálculo do Escore Z**

Qualquer distribuição normal pode ser transformada na distribuição normal padrão utilizando o escore Z. O **escore Z** é calculado pela fórmula:

$$z = \frac{(x \quad )}{}$$

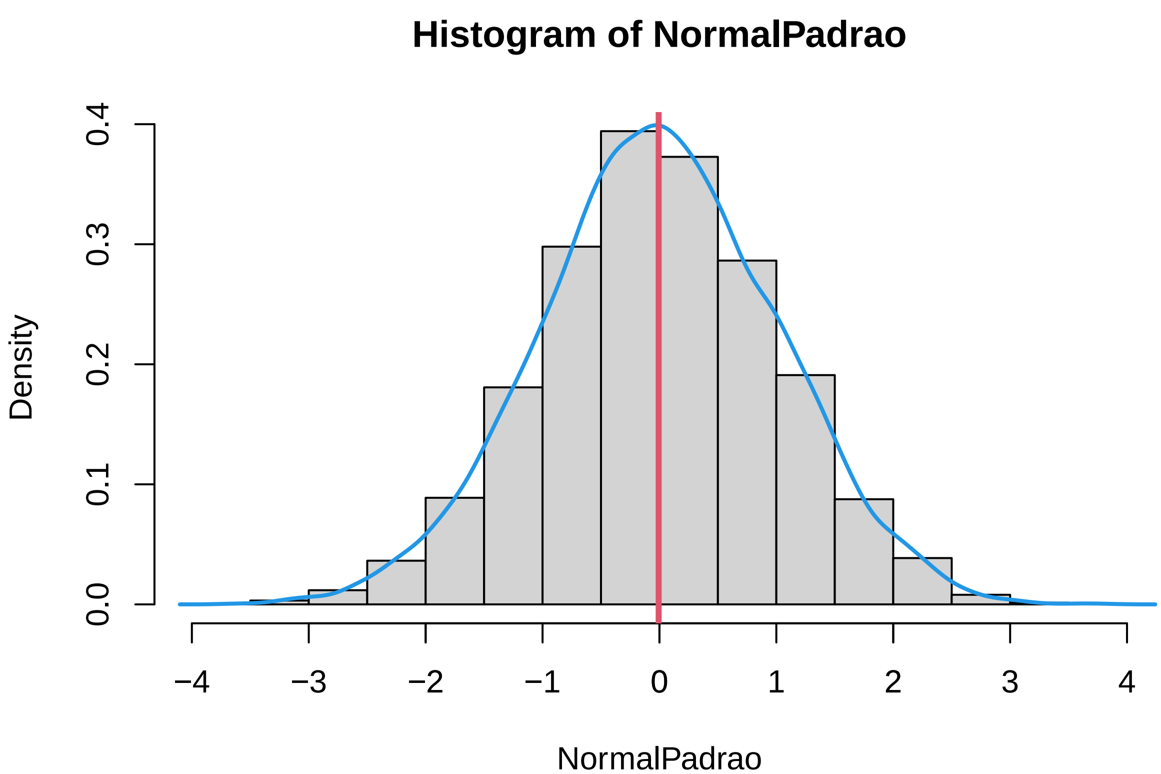
Onde:

- $x$ é o valor da observação.
- é a média da distribuição.
- é o desvio padrão da distribuição.

O gráfico QQ realiza exatamente o que a fórmula do escore Z descreve: ele transforma os dados da amostra para a distribuição normal padrão (média 0 e desvio padrão 1) e então compara esses valores transformados com os quantis da distribuição normal teórica. Se os pontos formarem uma linha reta, isso sugere que os dados seguem uma distribuição normal.

Veja um exemplo no R para uma Distribuição Normal Padrão:

```r
set.seed(1)
# rnorm(10000, 0, 1): normal padrão média 0, dp=1
# ou simplesmente rnorm(10000)
NormalPadrao <- rnorm(10000)
hist(NormalPadrao, probability = T)
lines(density(NormalPadrao), col = 4, lwd = 2)
axis(side = 1, at = seq(-3, 3, by = 1), labels = seq(-3, 3, by = 1))
abline(v = mean(NormalPadrao), col = 2, lwd = 3)
```

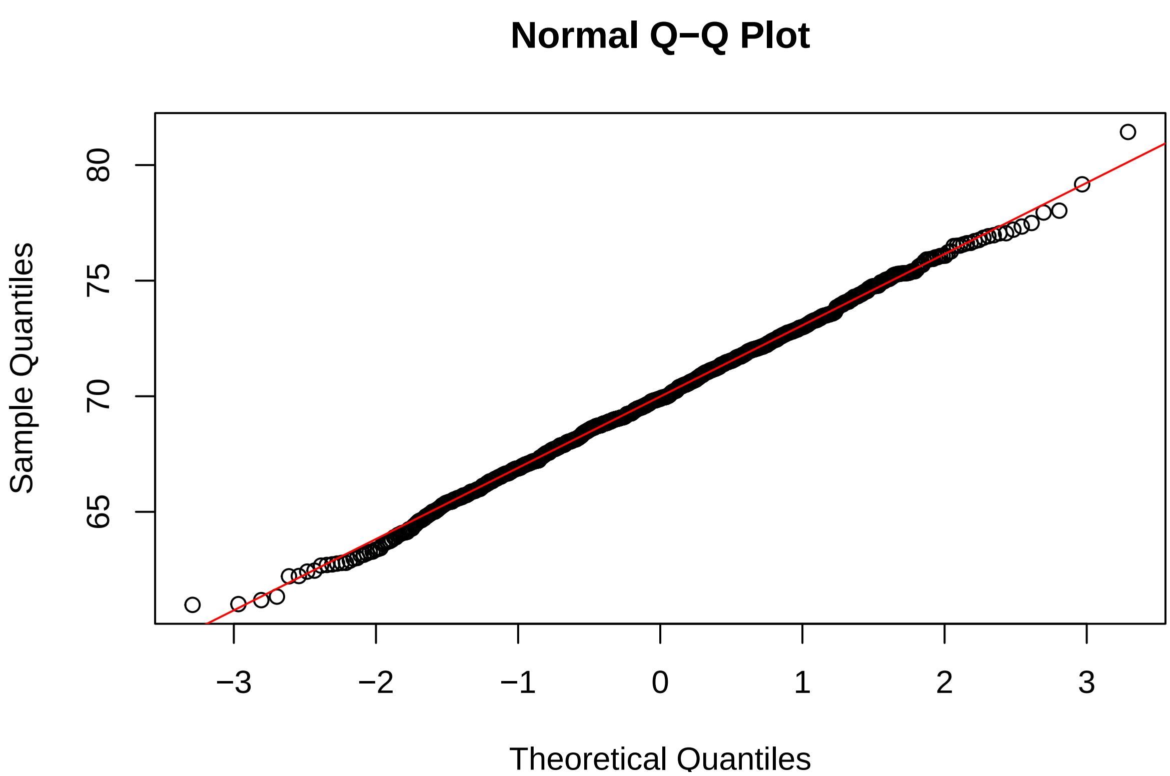



No gráfico QQ, os quantis da amostra são comparados com os quantis da distribuição normal padrão. Se a amostra segue uma distribuição normal, os pontos no gráfico QQ devem se alinhar aproximadamente a uma linha reta.

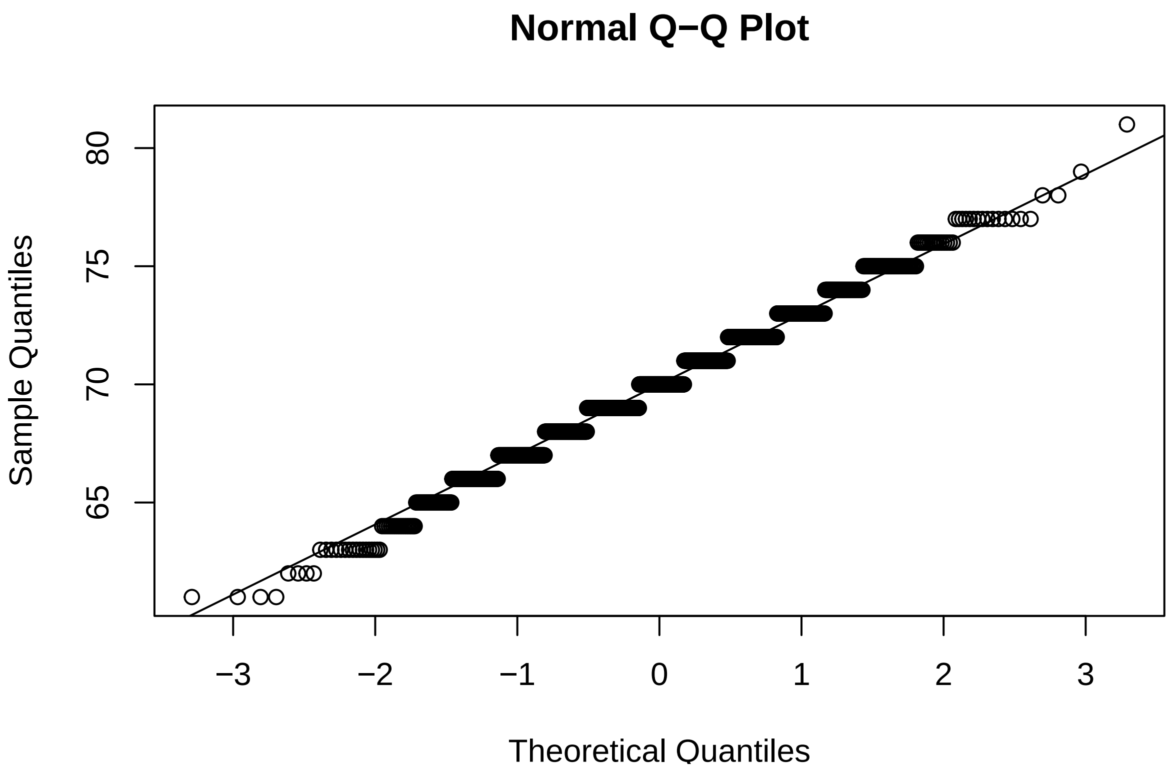
Se os pontos se afastam dessa linha reta de forma sistemática, isso indica que os dados não seguem uma distribuição normal. O tipo de desvio pode dar pistas sobre a natureza dessa não-normalidade (por exemplo, se os pontos se curvam para cima ou para baixo, pode indicar assimetria ou caudas pesadas).

Distrbuição normal é usada para dados contínuos!

```
# Gráfico QQ
set.seed(1)
BatimentosMulheres <- rnorm(1000, 70, 3)
qqnorm(BatimentosMulheres)
qqline(BatimentosMulheres, col="red")
```




```
# Faça o arredondamento
BatimentosMulheresR <- round(BatimentosMulheres,0)
qqnorm(BatimentosMulheresR)
qqline(BatimentosMulheresR)
```




A distribuição Normal é uma distribuição para modelar variáveis **CONTÍNUAS**!

## 15.3 Atividade 7

**Utilizando o banco de dados escolhido para as atividades 5 e 6, gere gráficos QQ para as variáveis quantitativas. Em seguida, avalie visualmente se essas variáveis seguem uma distribuição normal.**

# Chapter 16

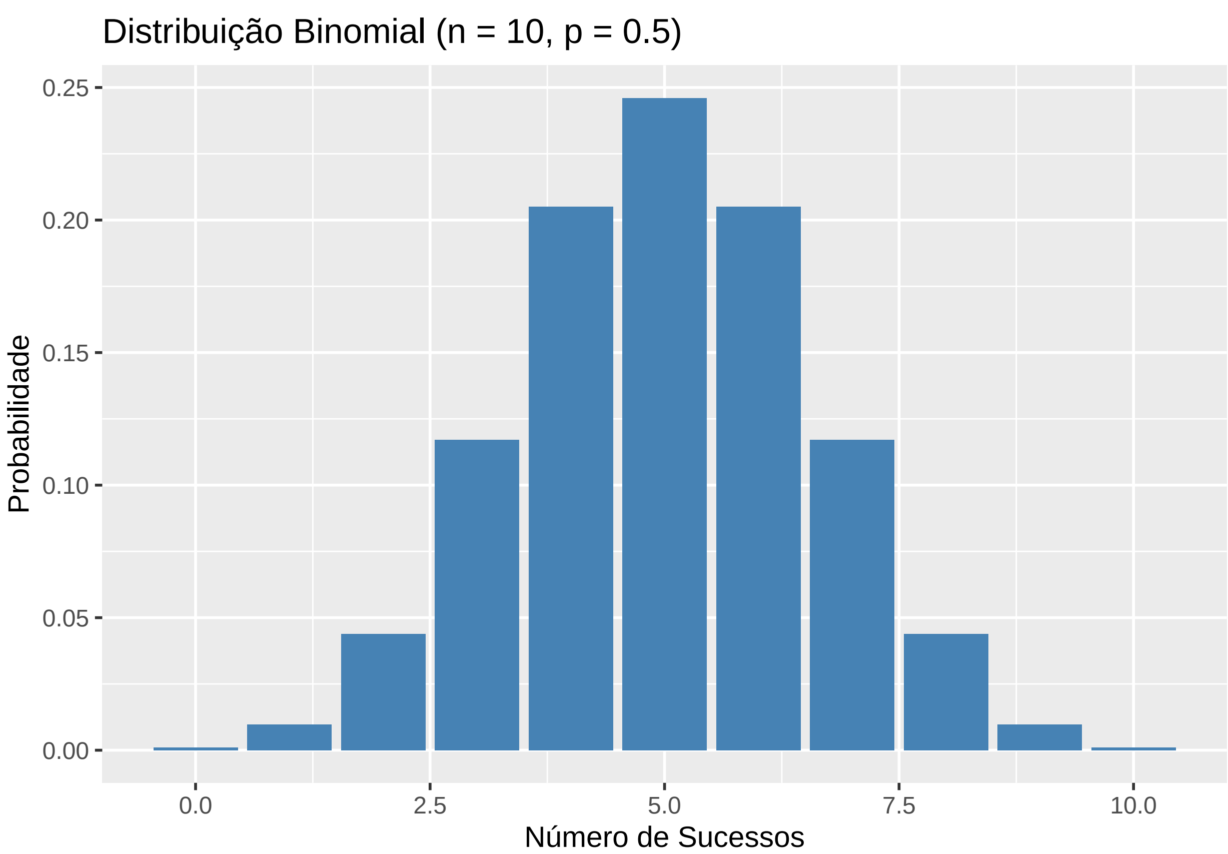
# Exemplos de distribuições

## 16.1 Distribuições Discretas

### 16.1.1 Binomial

Modela o número de sucessos em n tentativas com probabilidade p.

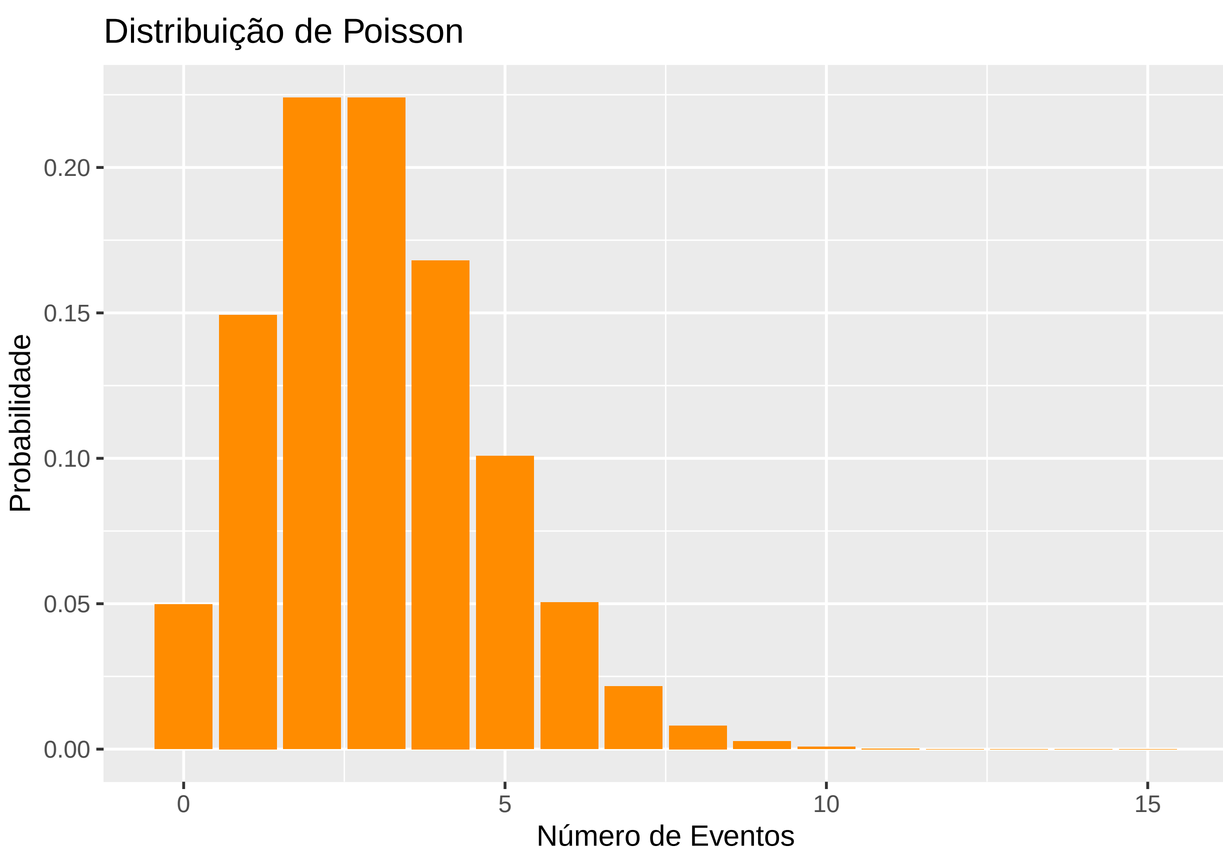
```r
n <- 10
p <- 0.5
x <- 0:n
y <- dbinom(x, size = n, prob = p)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  labs(title = "Distribuição Binomial (n = 10, p = 0.5)", x = "Número de Sucessos", y = "Probabilidad
```



### 16.1.2 Poisson

Modela o número de eventos raros num intervalo fixo.

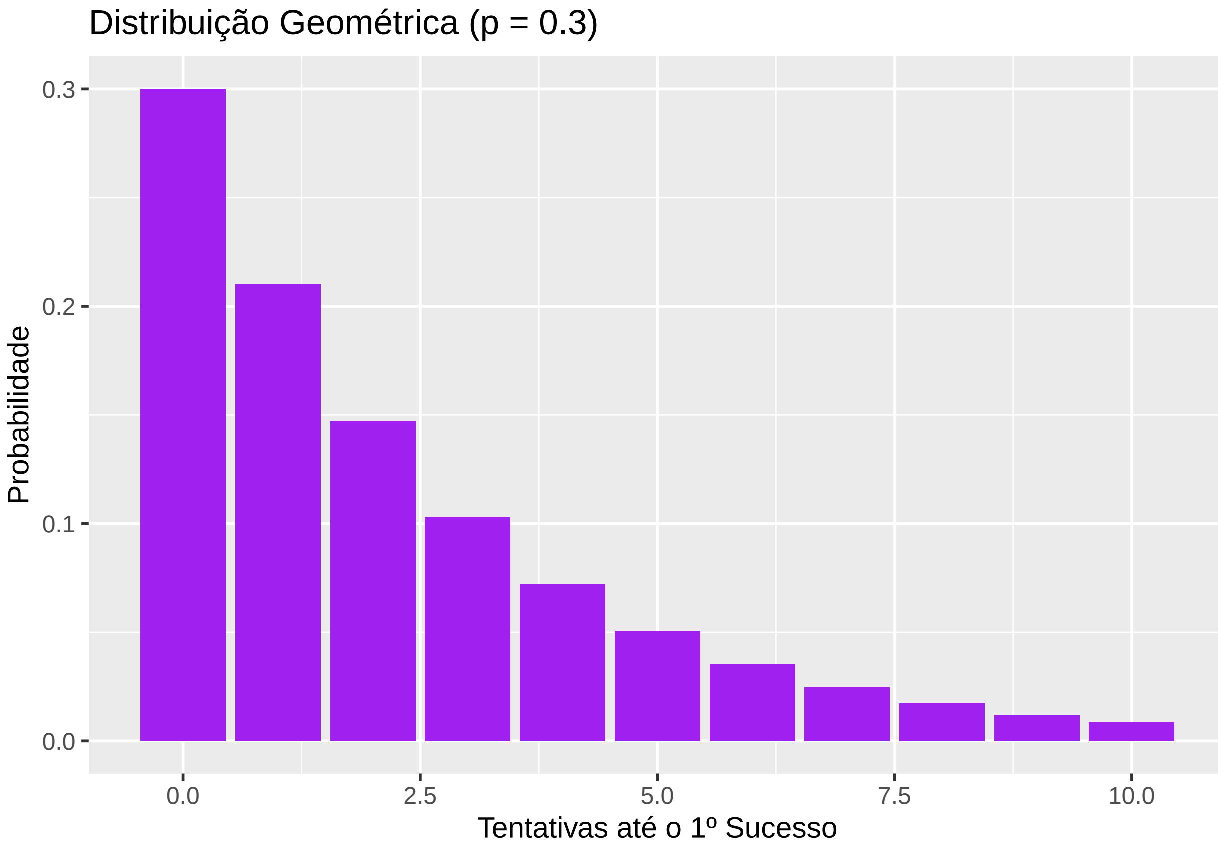
```
lambda <- 3
x <- 0:15
y <- dpois(x, lambda)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_bar(stat = "identity", fill = "darkorange") +
  labs(title = "Distribuição de Poisson", x = "Número de Eventos", y = "Probabilidade")
```




### 16.1.3 Geométrica

Modela o número de falhas antes do primeiro sucesso.

```
p <- 0.3
x <- 0:10
y <- dgeom(x, prob = p)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_bar(stat = "identity", fill = "purple") +
  labs(title = "Distribuição Geométrica (p = 0.3)", x = "Tentativas até o 1ž Sucesso", y = "Probabilidade"
```

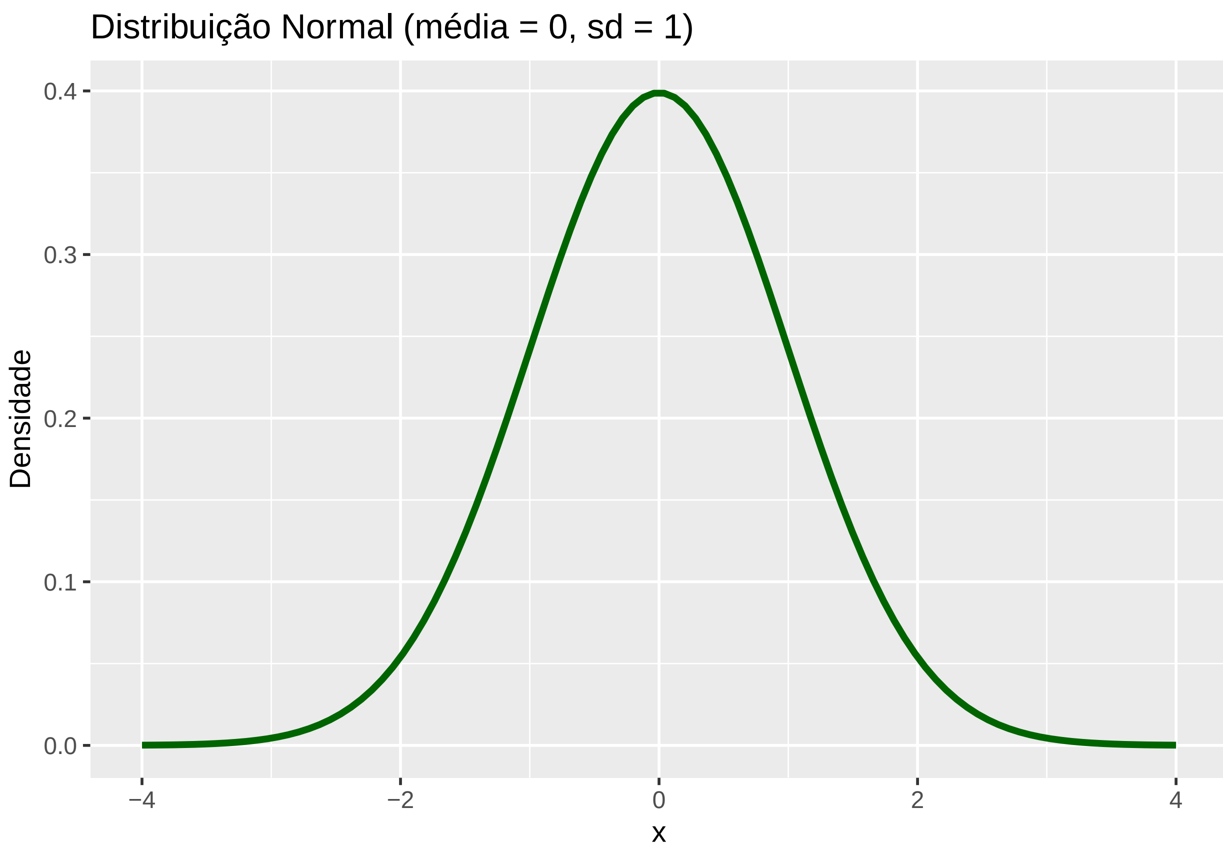



## 16.2 Distribuições Contínuas

### 16.2.1 Normal

Modela fenômenos naturais e erros de medida.

```
x <- seq(-4, 4, length.out = 100)
y <- dnorm(x)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "darkgreen", linewidth = 1.2) +
  labs(title = "Distribuição Normal (média = 0, sd = 1)", x = "x", y = "Densidade")
```

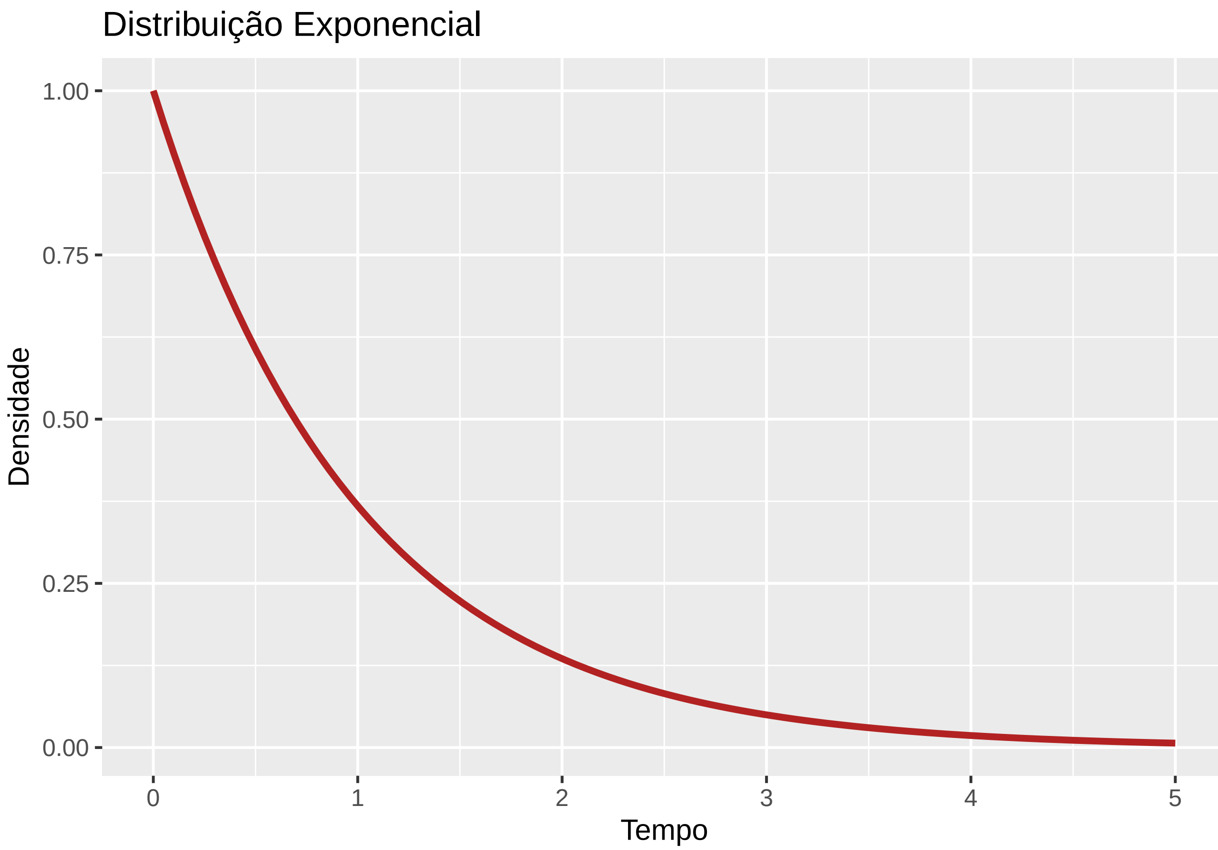



### 16.2.2 Exponencial

Tempo até um evento ocorrer.

```
lambda <- 1
x <- seq(0, 5, length.out = 100)
y <- dexp(x, rate = lambda)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "firebrick", size = 1.2) +
  labs(title = "Distribuição Exponencial", x = "Tempo", y = "Densidade")
```

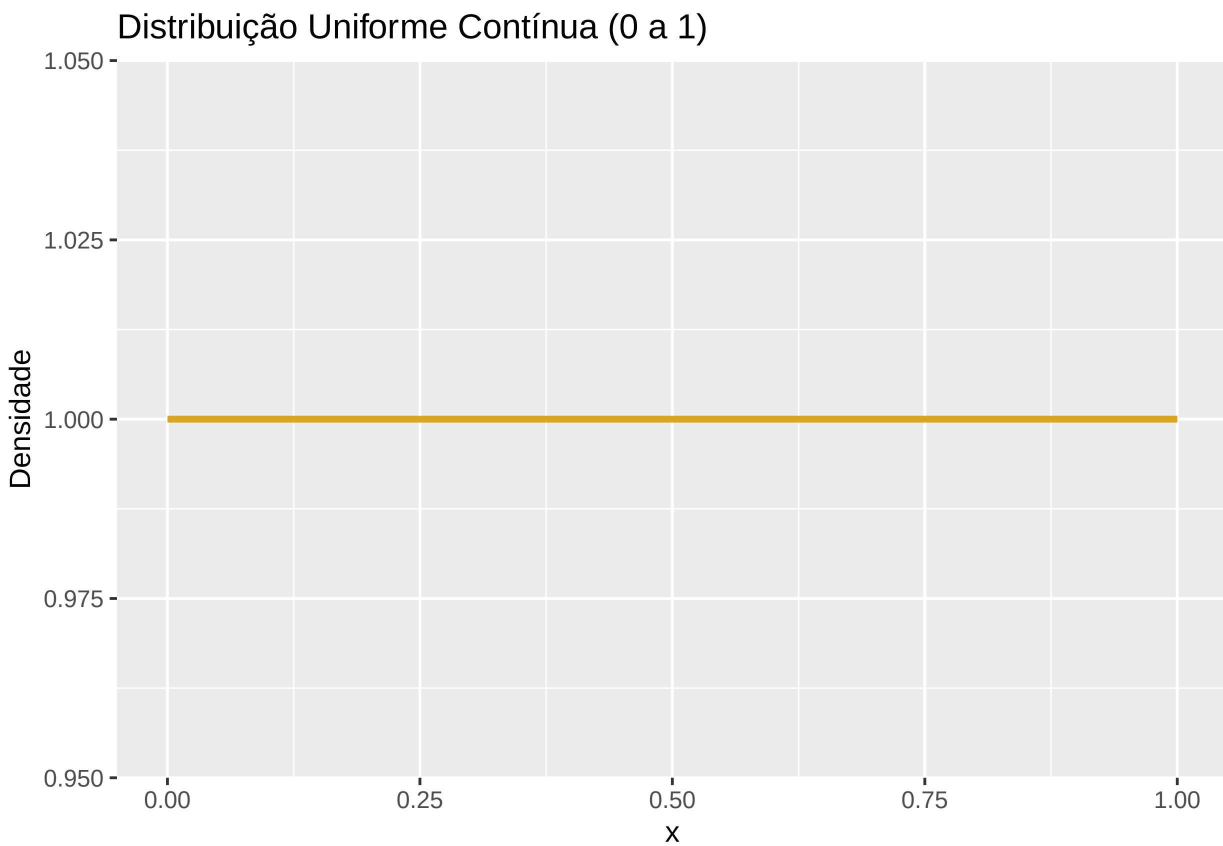
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```




### 16.2.3 Uniforme Contínua

Todos os valores têm a mesma chance.

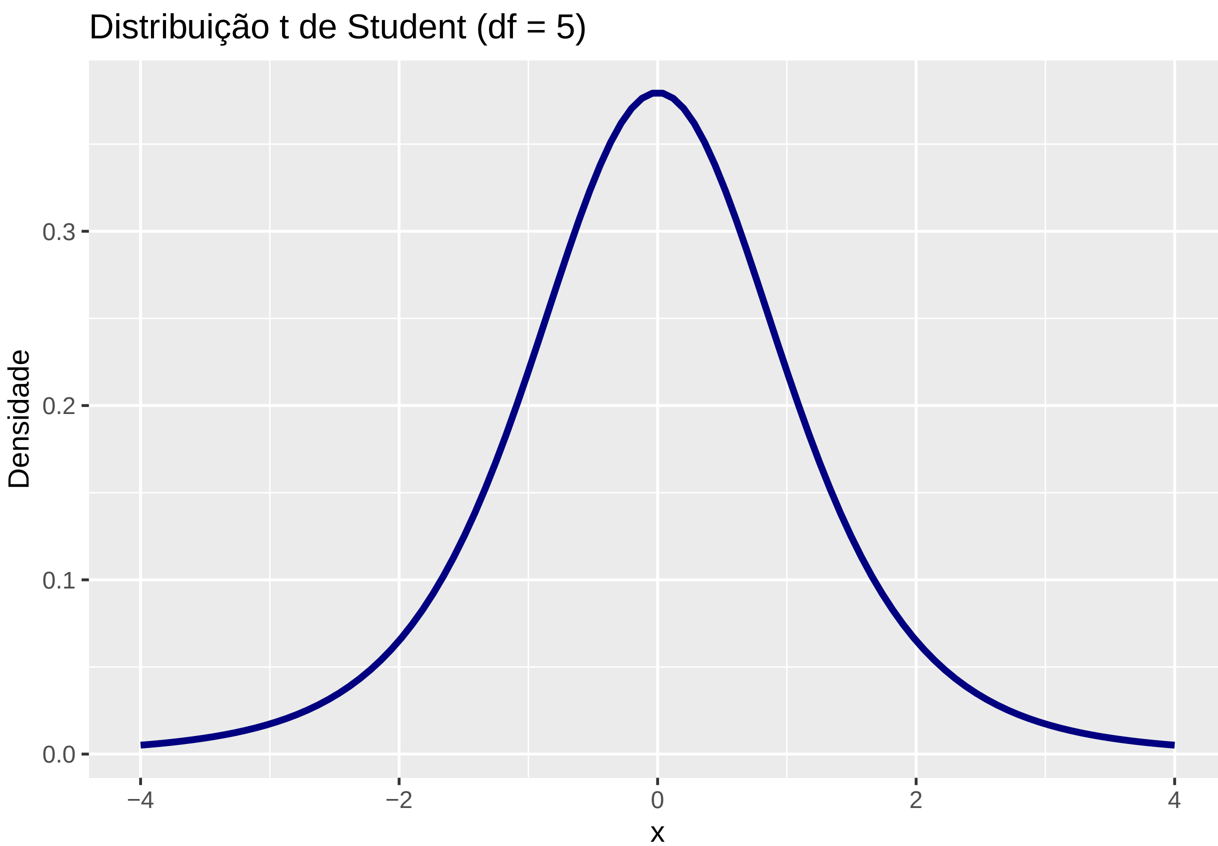
```
x <- seq(0, 1, length.out = 100)
y <- dunif(x, min = 0, max = 1)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "goldenrod", size = 1.2) +
  labs(title = "Distribuição Uniforme Contínua (0 a 1)", x = "x", y = "Densidade")
```




### 16.2.4 t de Student

Usada em testes com amostras pequenas.

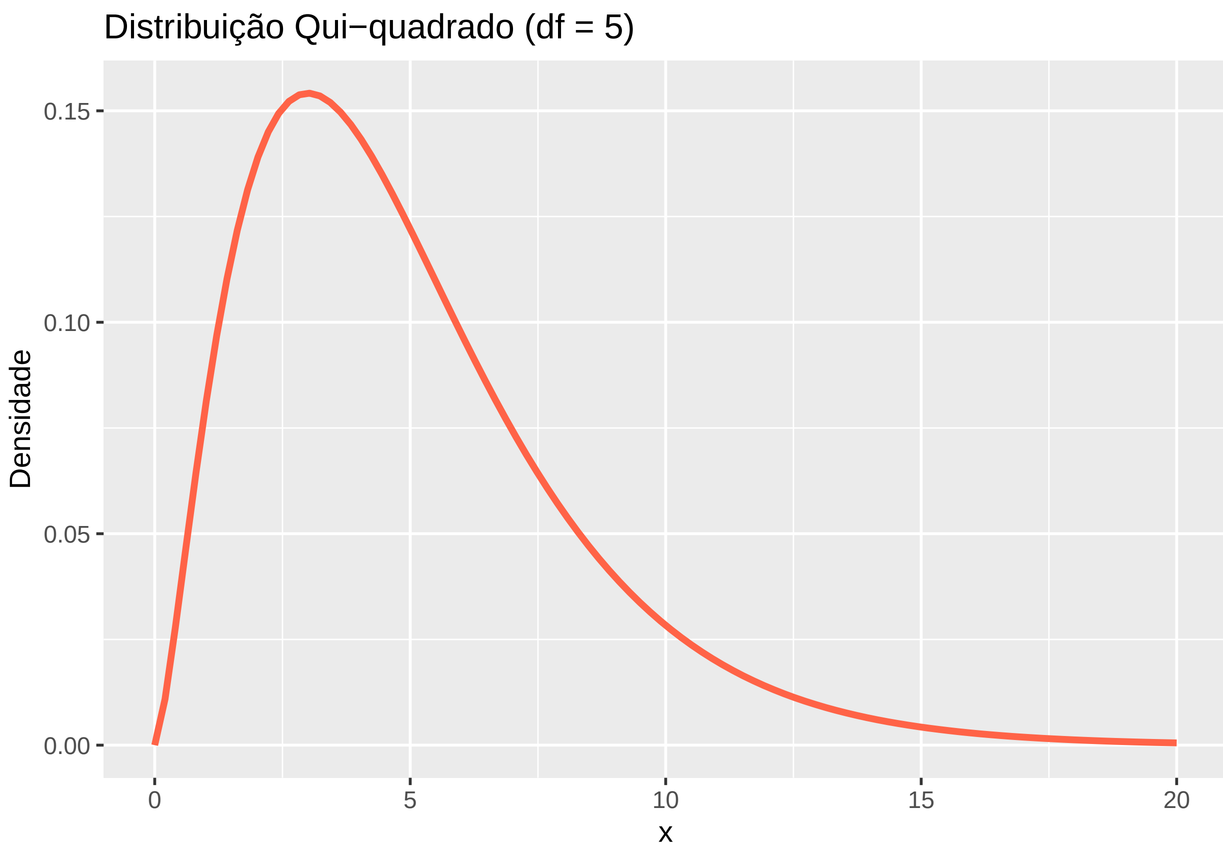
```
x <- seq(-4, 4, length.out = 100)
y <- dt(x, df = 5)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "navy", size = 1.2) +
  labs(title = "Distribuição t de Student (df = 5)", x = "x", y = "Densidade")
```




### 16.2.5 Qui-Quadrado (š)

Usada para testes de aderência e independência.

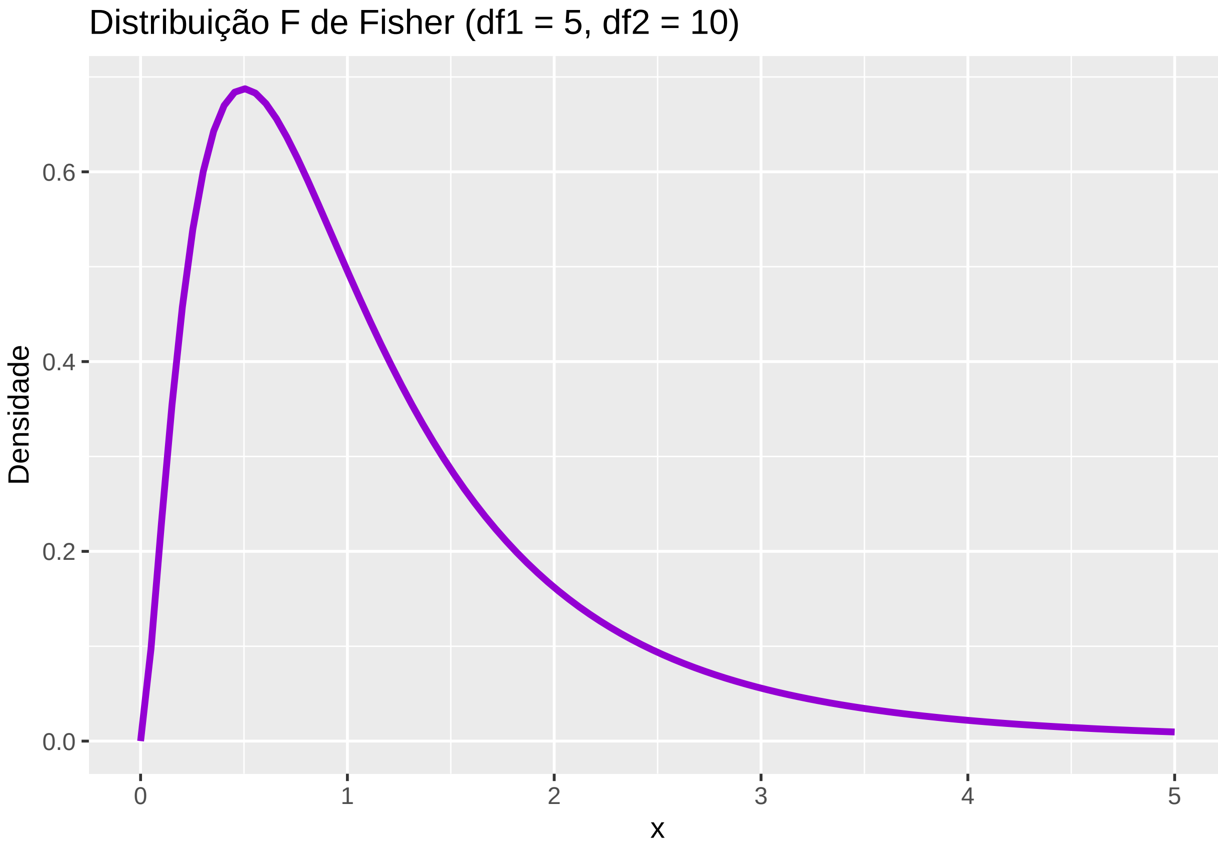
```
x <- seq(0, 20, length.out = 100)
y <- dchisq(x, df = 5)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "tomato", size = 1.2) +
  labs(title = "Distribuição Qui-quadrado (df = 5)", x = "x", y = "Densidade")
```




### 16.2.6 F de Fisher

Usada em ANOVA para comparar variâncias.

```
x <- seq(0, 5, length.out = 100)
y <- df(x, df1 = 5, df2 = 10)

ggplot(data.frame(x, y), aes(x, y)) +
  geom_line(color = "darkviolet", size = 1.2) +
  labs(title = "Distribuição F de Fisher (df1 = 5, df2 = 10)", x = "x", y = "Densidade")
```




Cada distribuição tem uma aplicação específica conforme o tipo de variável (discreta ou contínua), o contexto do problema e as suposições do modelo. Entender suas formas e usos ajuda na escolha correta da análise estatística.

# Chapter 17

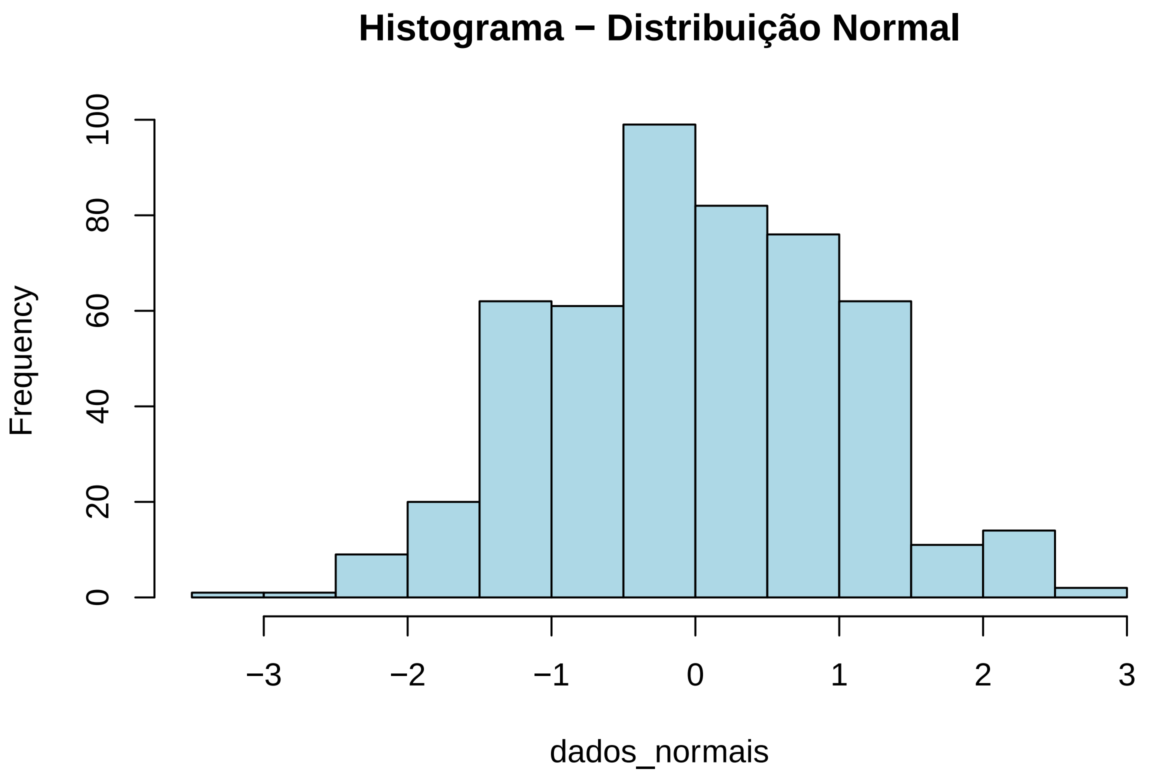
# Normal ou Não?

Antes de avançarmos para a aplicação dos testes estatísticos, é importante avaliarmos se os dados seguem ou não uma distribuição normal. Essa verificação inicial é muito importante porque muitos testes paramétricos, como o teste t e a ANOVA, partem da suposição de normalidade dos dados. Compreender esse aspecto permite escolher os métodos estatísticos mais adequados, aumentando a confiabilidade dos resultados e evitando interpretações equivocadas que poderiam comprometer a análise.

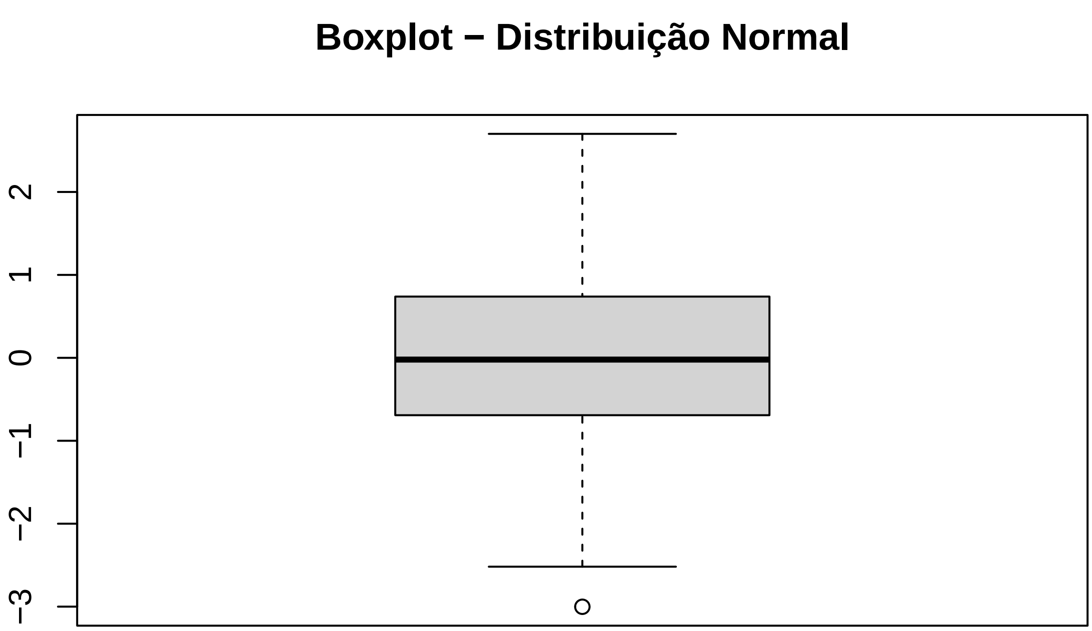
Esta seção apresenta uma análise visual da normalidade dos dados por meio de gráficos QQ (Quantil-Quantil), com a comparação de três cenários distintos: dados que seguem uma distribuição normal, dados que levantam dúvidas quanto à normalidade e dados que claramente não seguem essa distribuição.

- Dados normalmente distribuídos
- Dados que geram dúvida quanto à normalidade
- Dados que claramente não seguem distribuição normal

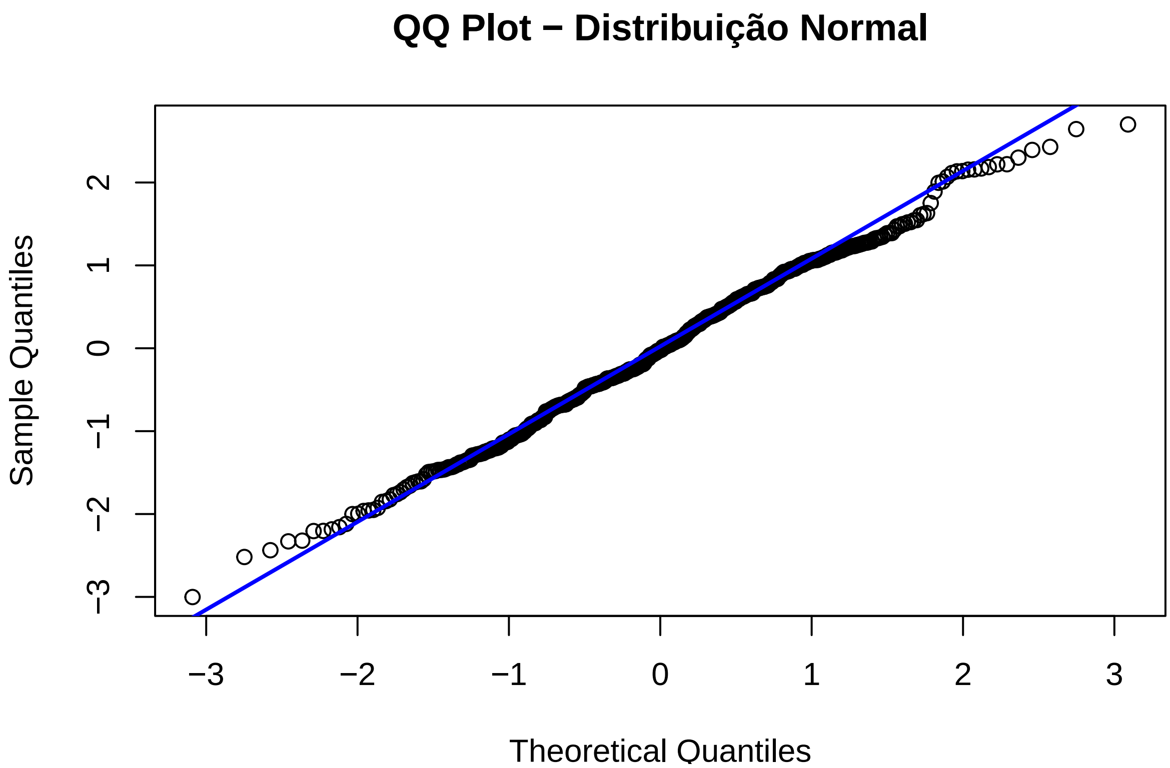
```r
set.seed(10)
dados_normais <- rnorm(500, mean = 0, sd = 1)

hist(dados_normais, main = "Histograma - Distribuição Normal", col = "lightblue")
```




```r
boxplot(dados_normais, main = "Boxplot - Distribuição Normal")
```

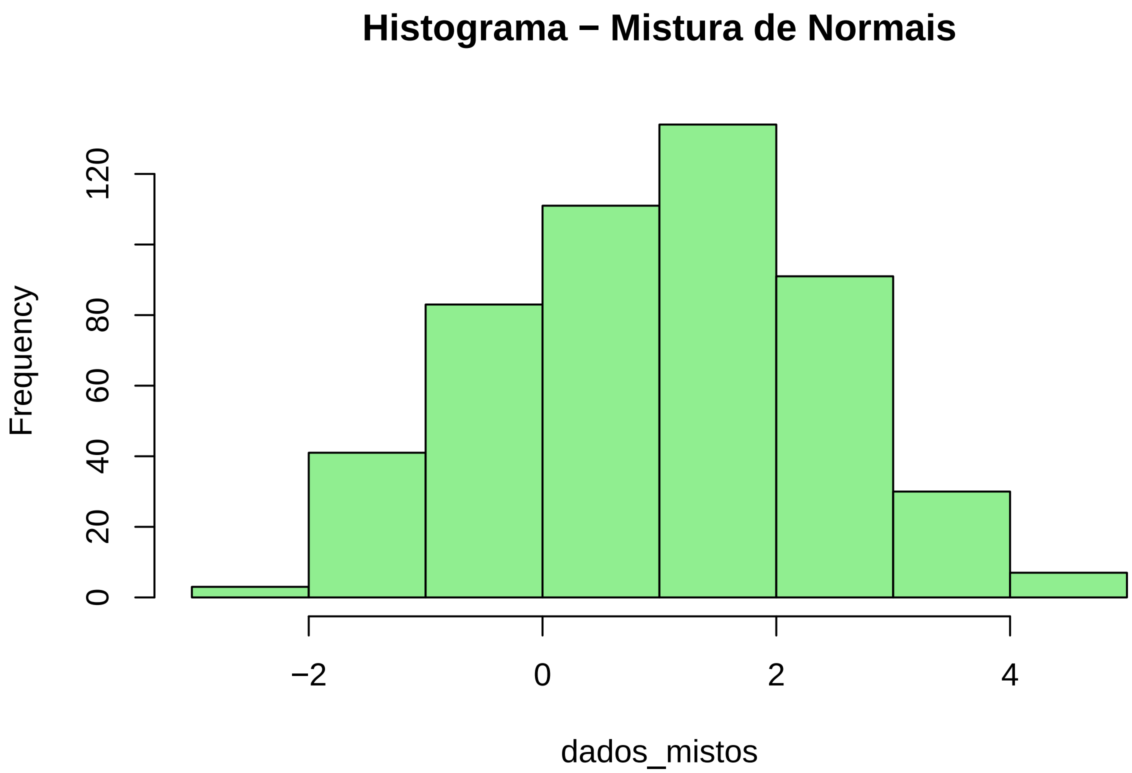



```r
qqnorm(dados_normais, main = "QQ Plot - Distribuição Normal")
qqline(dados_normais, col = "blue", lwd = 2)
```

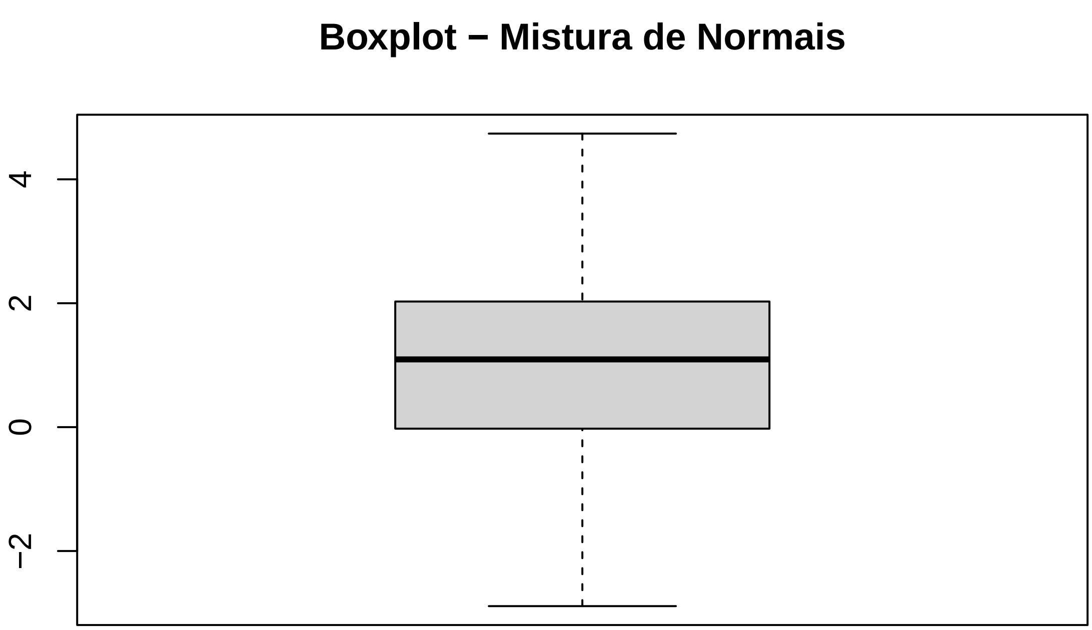



*Observação:* Os pontos seguem de perto a linha, indicando que os dados são normalmente distribuídos.

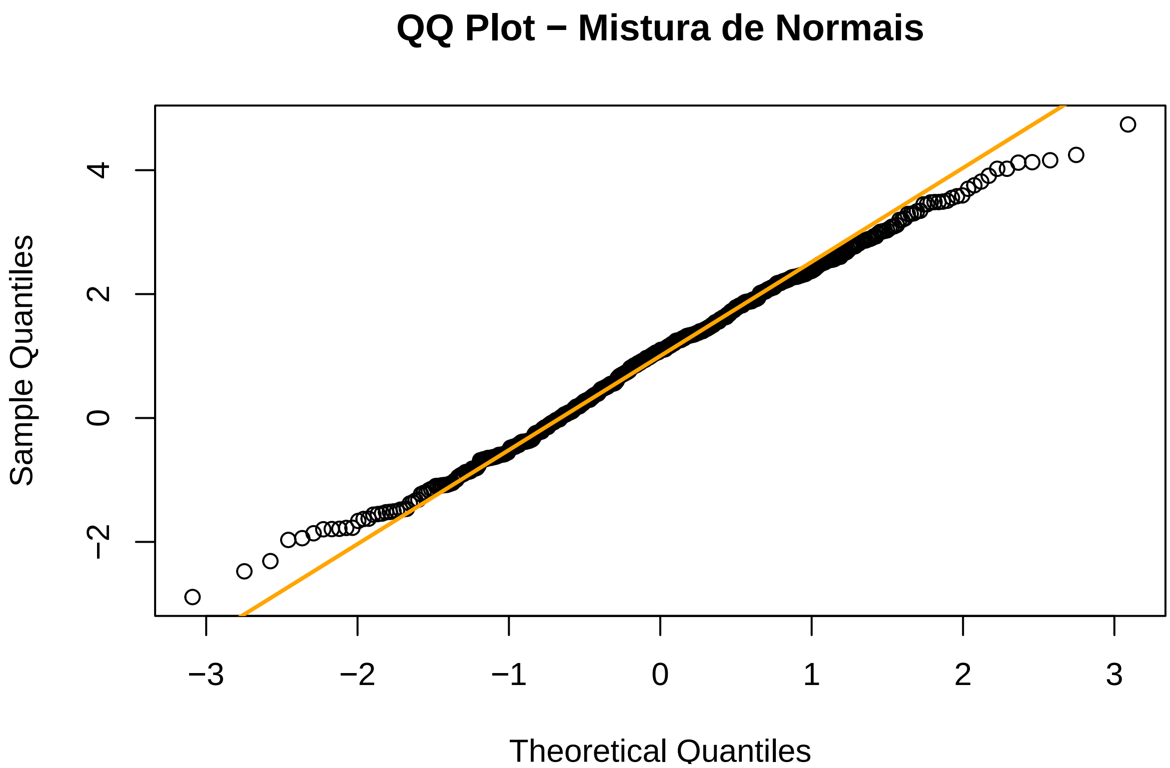
```r
set.seed(20)
dados_mistos <- c(rnorm(250, 0, 1), rnorm(250, 2, 1))

hist(dados_mistos, main = "Histograma - Mistura de Normais", col = "lightgreen")
```




```r
boxplot(dados_mistos, main = "Boxplot - Mistura de Normais")
```

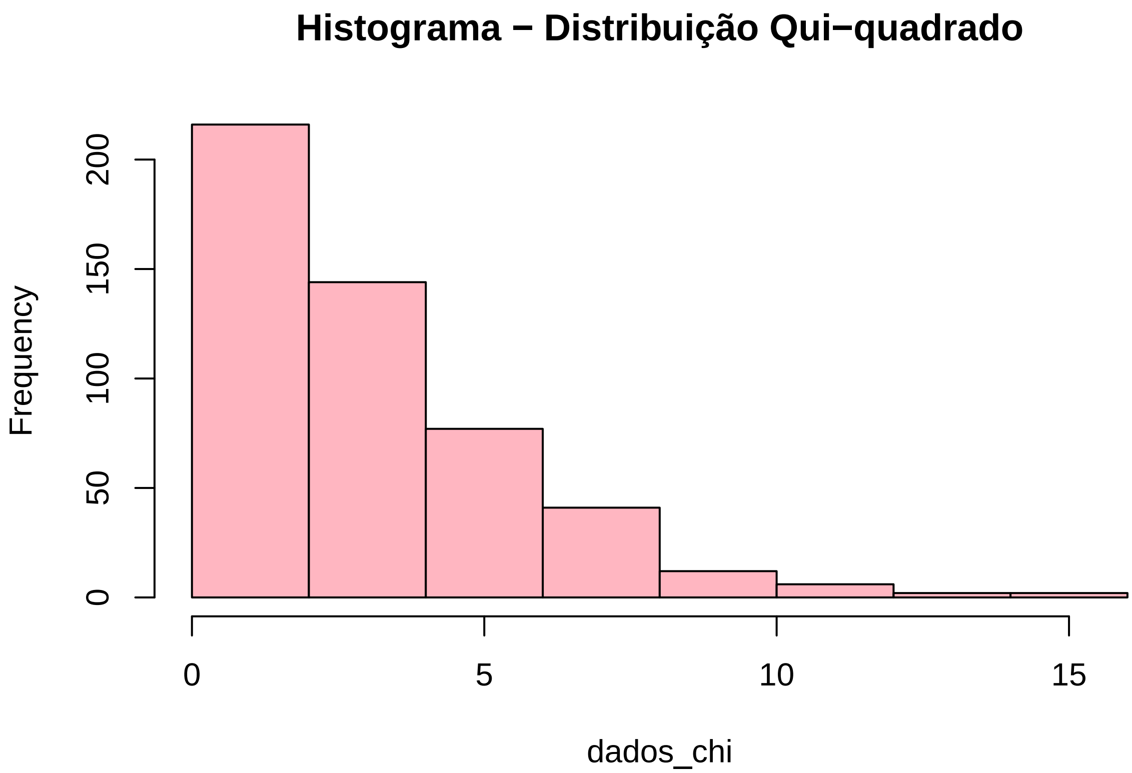



```r
qqnorm(dados_mistos, main = "QQ Plot - Mistura de Normais")
qqline(dados_mistos, col = "orange", lwd = 2)
```

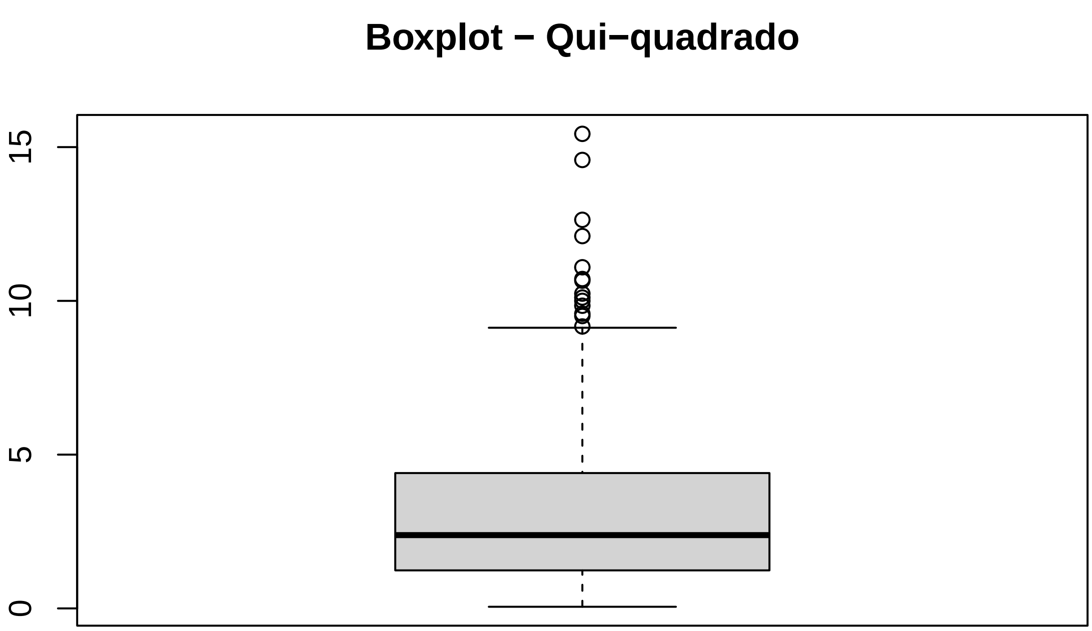



*Observação:* As caudas do gráfico QQ começam a se afastar da linha, o que pode gerar dúvida sobre a normalidade.

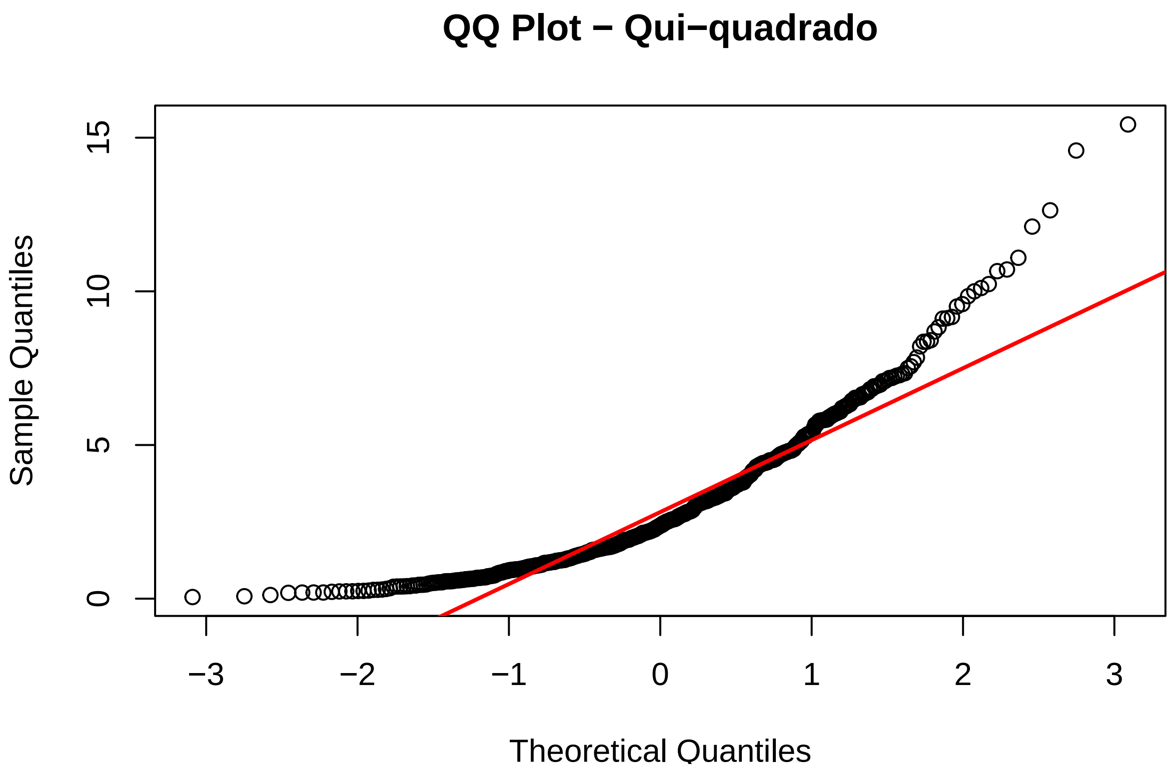
```r
set.seed(30)
dados_chi <- rchisq(500, df = 3)

hist(dados_chi, main = "Histograma - Distribuição Qui-quadrado", col = "lightpink")
```




```r
boxplot(dados_chi, main = "Boxplot - Qui-quadrado")
```

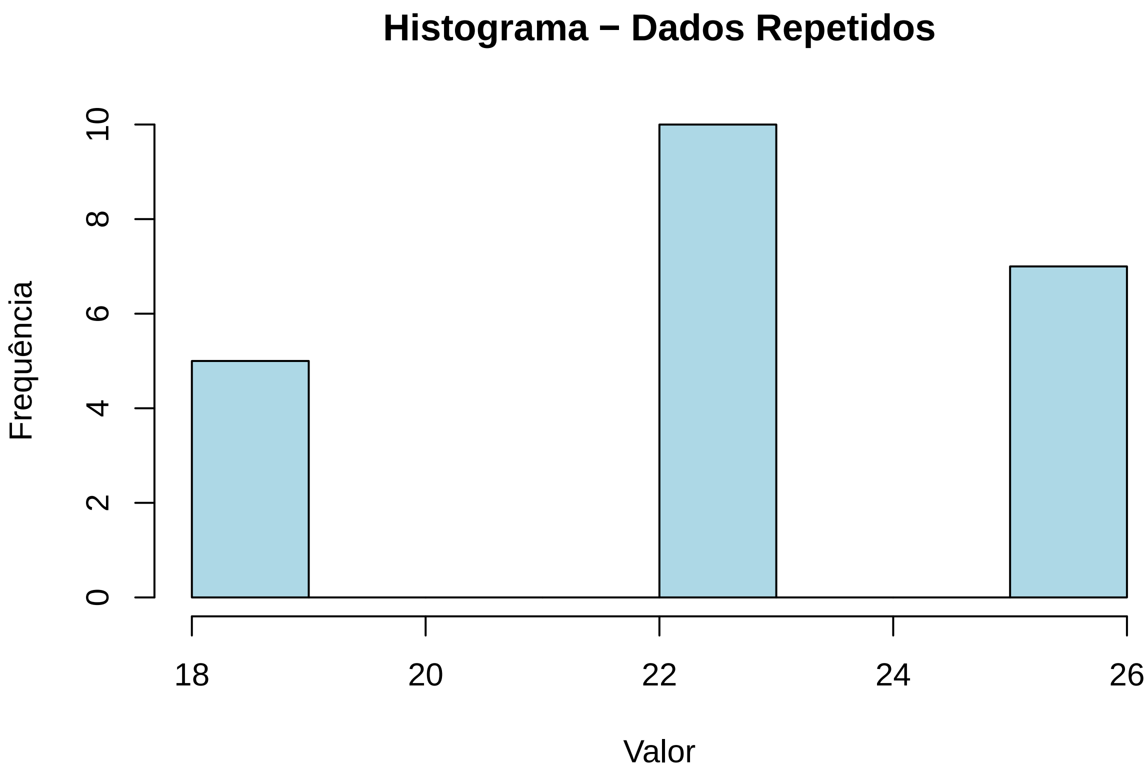



```r
qqnorm(dados_chi, main = "QQ Plot - Qui-quadrado")
qqline(dados_chi, col = "red", lwd = 2)
```

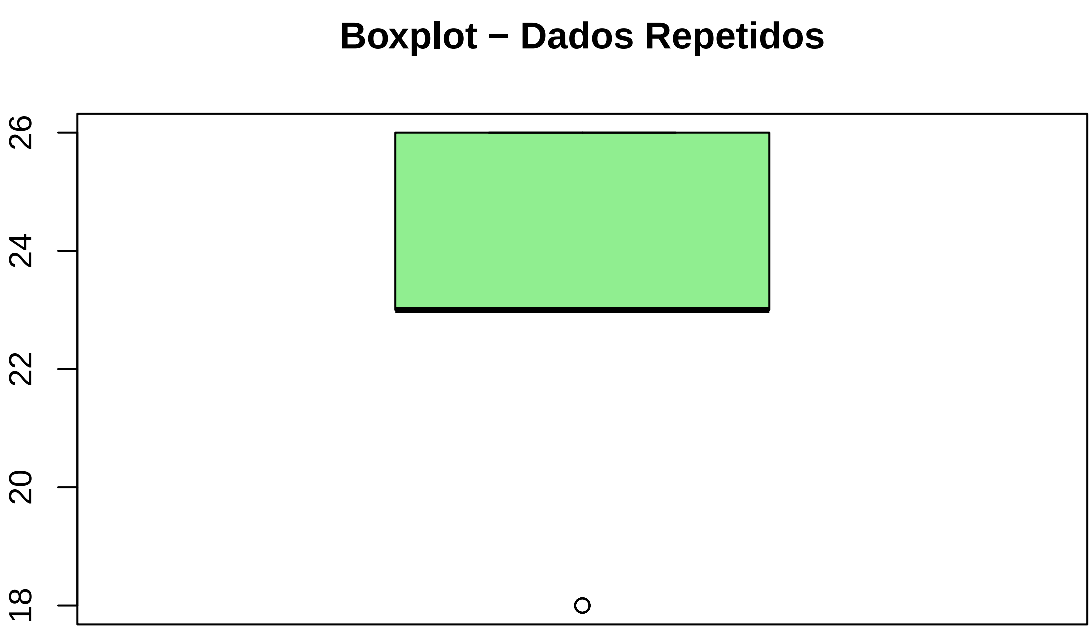



*Observação:* A forte curvatura do gráfico QQ indica clara violação da normalidade (assimetria à direita).

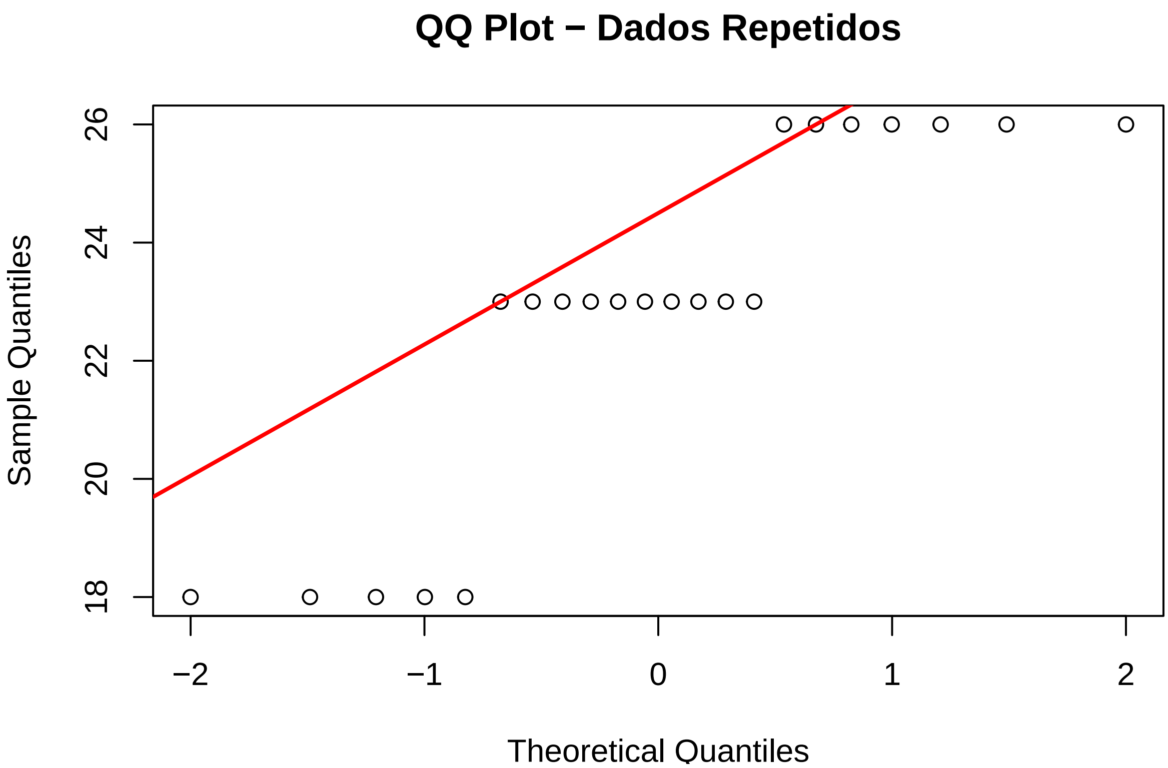
```
set.seed(42)
dados_repetidos <- c(rep(18, 5), rep(23, 10), rep(26,7))

# Histograma
hist(dados_repetidos, main = "Histograma - Dados Repetidos", col = "lightblue", xlab = "Valor", ylab
```




```
# Boxplot
boxplot(dados_repetidos, main = "Boxplot - Dados Repetidos", col = "lightgreen")
```




```
# QQ Plot
qqnorm(dados_repetidos, main = "QQ Plot - Dados Repetidos")
qqline(dados_repetidos, col = "red", lwd = 2)
```




O gráfico QQ é uma ferramenta visual poderosa para avaliar se um conjunto de dados segue uma distribuição normal. Em conjunto com histogramas e boxplots, permite identificar desvios, outliers e assimetrias de maneira clara.

# Chapter 18

# Estatística Inferencial

A **estatística inferencial** é uma área essencial da estatística que permite fazer generalizações sobre uma população com base em dados coletados de uma amostra. Um passo fundamental nesse processo é o **cálculo amostral**, que determina o tamanho ideal da amostra para garantir a validade dos resultados, e esse cálculo depende diretamente do **tipo de teste estatístico** que será utilizado, pois diferentes testes exigem diferentes parâmetros, como variabilidade, efeito esperado e poder do teste.

Ao conduzir uma análise inferencial, formulam-se **hipóteses**: a **hipótese nula (H)**, que representa a ausência de efeito ou diferença, e a **hipótese alternativa (H)**, que sugere a existência de um efeito ou diferença. Define-se também um **nível de significância ()**, geralmente 0,05 (5%), que representa a probabilidade de rejeitar H quando ela é verdadeira (**erro tipo I**). O **valor-p**, calculado a partir dos dados, indica a probabilidade de obter um resultado tão extremo quanto o observado, supondo que H seja verdadeira. Com base na comparação entre o valor-p e o nível de significância, aplica-se o **critério de decisão**: se o valor-p for menor que , rejeita-se H, o que indica **evidências estatísticas suficientes para apoiar a hipótese alternativa**.

## 18.1 Etapas de um Teste Estatístico

Para realizar qualquer **teste estatístico**, seguimos uma sequência básica de passos:

1. **Escrevemos as hipóteses do teste**, começando com a **hipótese nula (H)**, que geralmente afirma que não há efeito ou diferença, e a **hipótese alternativa (H)**, que propõe o contrário.
2. **Definimos o nível de significância ()**, que é a margem de erro aceitável. Esse valor, geralmente 0,05 (ou 5%), já foi considerado no **cálculo do tamanho da amostra**, garantindo que os resultados tenham confiabilidade.
3. **Utilizamos um recurso computacional**, como um software estatístico, para calcular o **valor-p**, que mostra a probabilidade de observarmos os dados coletados caso a hipótese nula seja verdadeira.
4. **Comparamos o valor-p com o nível de significância**. Se o valor-p for menor que , **rejeitamos a hipótese nula**. Caso contrário, não rejeitamos.
5. Por fim, **chegamos à conclusão do teste**, que nos diz se há ou não **evidência estatística** suficiente para apoiar a hipótese alternativa.

## 18.2 Erros Tipo I e Tipo II

Em um **teste estatístico**, tomamos uma decisão com base nos dados coletados de uma amostra: **rejeitar** ou **não rejeitar a hipótese nula (H)**. Por isso, é fundamental realizar um **cálculo amostral adequado**, selecionar cuidadosamente a **técnica de amostragem** e estar atento às possíveis **fontes de vieses**, garantindo que a amostra seja representativa e que os resultados do teste sejam confiáveis.

Entretanto, essa decisão envolve **riscos de erro**, que são inerentes ao processo justamente porque estamos baseando nossas conclusões em uma amostra e não na população inteira. Esses erros se dividem em dois tipos:

- **Erro Tipo I () – Falso Positivo**
  Ocorre quando **rejeitamos H mesmo ela sendo verdadeira**.
  É como afirmar que existe um efeito quando, na verdade, não existe.

  > Exemplo na medicina: concluir que um novo medicamento reduz a pressão arterial quando ele não tem efeito real, o que pode levar à aprovação de um tratamento ineficaz.

  > Exemplo no esporte: afirmar que um programa de treinamento melhora o desempenho dos atletas quando ele não traz benefício real, gerando investimentos desnecessários.

  > Exemplo na psicologia: dizer que uma terapia cognitivo-comportamental reduz a ansiedade, mesmo sem efeito comprovado, gerando falsas expectativas nos pacientes.

  Essa é a chance de um **falso positivo**, e o **nível de significância** (geralmente 0,05) representa a probabilidade de cometer esse erro.

- **Erro Tipo II () – Falso Negativo**
  Ocorre quando **não rejeitamos H mesmo ela sendo falsa**.
  É como deixar passar um efeito real sem detectá-lo.

  > Exemplo na medicina: não identificar que o medicamento é eficaz, rejeitando seu uso quando ele realmente funciona.

  > Exemplo no esporte: não perceber a melhora no desempenho causada pelo programa de treinamento, deixando de adotá-lo e prejudicando os atletas.

  > Exemplo na psicologia: não detectar que a terapia é eficaz para reduzir a ansiedade, levando à rejeição de um tratamento que poderia beneficiar os pacientes.

  Isso é um **falso negativo**, e é a probabilidade de cometer esse erro.

## 18.3 Poder do Teste Estatístico

- O **poder do teste** é a probabilidade de **detectar um efeito real quando ele realmente existe**, ou seja, **rejeitar H quando H é falsa**.

- O poder é calculado como:
  **Poder = 1 -**

Um teste com **alto poder** (geralmente desejado acima de 80%) tem menos chance de cometer erro tipo II, o que significa maior capacidade de detectar diferenças reais quando elas existem. O poder depende do **tamanho da amostra**, do **nível de significância**, da **variabilidade dos dados** e da **magnitude do efeito** que se deseja identificar.

## 18.4 Tabela: Erros, Decisões e Poder do Teste

| Situação Real | Decisão: Rejeitar H | Decisão: Não Rejeitar H |
|---|---|---|
| H é verdadeira | Erro Tipo I () | Decisão correta |
| H é falsa | Decisão correta (*poder*) | Erro Tipo II () |

- **Poder do teste:** Probabilidade de rejeitar H quando H é falsa (ou seja, evitar o erro tipo II).

## 18.5 Tamanho do Efeito

### 18.5.1 O que é o Tamanho do Efeito?

O **tamanho do efeito** é uma medida que indica **o quanto uma diferença ou relação observada nos dados é relevante na prática**.

Enquanto o **valor-p** nos informa se um resultado é estatisticamente significativo (ou seja, se é improvável que tenha ocorrido por acaso), o tamanho do efeito **complementa essa análise mostrando a magnitude real do efeito observado**.

### 18.5.2 Exemplos para Entender Melhor

Imagine que um novo remédio reduza a pressão arterial em média em **1 mmHg** comparado ao tratamento padrão.
Com uma amostra muito grande, essa pequena diferença pode ser estatisticamente significativa (*valor-p* $<$ 0,05), mas **clinicamente irrelevante**. O tamanho do efeito, neste caso, mostra que, apesar do resultado ser significativo, **o impacto prático é muito pequeno**.

Suponha um estudo que compara dois métodos de treinamento de força e encontra uma diferença média de **0,5 kg** no aumento de carga máxima entre os grupos após 8 semanas.
Com uma amostra grande, essa diferença pode ser estatisticamente significativa, mas **na prática, esse ganho é muito pequeno** para justificar a troca de método. O tamanho do efeito indica que, embora exista diferença detectada, **ela tem pouco impacto real no desempenho atlético**.

Considere uma pesquisa que compara níveis de ansiedade entre terapia online e presencial, com diferença média de **1 ponto** (em escala de 0 a 100).
Mesmo que essa diferença seja estatisticamente significativa, **não representa uma mudança clinicamente relevante** no estado emocional dos pacientes. O tamanho do efeito ajuda a perceber que a diferença entre as modalidades pode ser mínima na prática.

## 18.6 Medidas Comuns de Tamanho do Efeito

A escolha da medida depende do tipo de teste e das variáveis analisadas.

### 18.6.1 d de Cohen (teste t)

Usado para comparar médias entre dois grupos, mede a diferença em unidades de desvio padrão.

| Valor do d | Interpretação |
|---|---|
| 0.2 | Efeito pequeno |
| 0.5 | Efeito médio |
| 0.8 | Efeito grande |

### 18.6.2 C de Cramer, para tabela 2x2 (teste qui-quadrado)

Mede a associação entre variáveis categóricas (varia de 0 a 1).

| Valor de C | Interpretação |
|---|---|
| 0.1 | Associação fraca |
| 0.3 | Associação moderada |
| 0.5 | Associação forte |

### 18.6.3 Coeficiente de Correlação de Pearson (r)

Mede a força da relação linear entre duas variáveis quantitativas.

| Valor de r | Interpretação |
|---|---|
| 0.1 | Correlação fraca |
| 0.3 | Correlação moderada |
| 0.5 | Correlação forte |
| 0 | Nenhuma correlação |
| ś1 | Correlação perfeita |

## 18.7 Por que o Tamanho do Efeito é Importante?

Um resultado estatisticamente significativo **nem sempre indica relevância prática**. O tamanho do efeito responde à pergunta: **"Esse resultado faz diferença no mundo real?"**

## 18.8 Testes

Nas próximas seções, abordaremos os testes de **comparação entre grupos** com base em uma variável quantitativa, os testes de **associação entre variáveis categóricas** nominais, e os testes de **correlação**, que podem envolver variáveis quantitativas ou qualitativas ordinais. Essa abordagem visa facilitar a compreensão e a interpretação dos resultados estatísticos em variados contextos, ressaltando sempre a importância de avaliar a relevância prática dos achados para uma tomada de decisão mais informada.

# Chapter 19

# Testes de Comparação entre Grupos

Os testes de comparação são usados para verificar se existem diferenças significativas entre grupos em relação a uma variável quantitativa. Essas comparações podem variar conforme a relação entre os grupos, o número de grupos e o tipo de dado.

## 19.1 Comparações Pareadas (Dependentes)

Quando os grupos comparados são relacionados ou "ligados" de alguma forma, dizemos que a comparação é **pareada** ou entre grupos **dependentes**. Isso acontece, por exemplo, quando medimos a mesma amostra em dois momentos diferentes — antes e depois de uma intervenção — ou quando comparamos pares de indivíduos relacionados, como gêmeos.

**Exemplos:**

Avaliar a pressão arterial dos pacientes antes e depois de um tratamento.

Medir a força muscular dos atletas antes e após um programa de treinamento.

Aplicar um teste de ansiedade em pacientes antes e depois de uma terapia.

## 19.2 Comparações Não-Pareadas (Independentes)

As comparações são **não-pareadas** ou entre grupos **independentes** quando os grupos não têm relação entre si, ou seja, os participantes de um grupo não pertencem ao outro. Cada observação é independente das outras.

**Exemplos:**

Comparar a pressão arterial entre pacientes que receberam medicamento e pacientes que receberam placebo.

Comparar a velocidade média de corrida entre dois times diferentes.

Comparar níveis de estresse entre grupos que fizeram terapia presencial e terapia online.

## 19.3 Número de Grupos na Comparação

A comparação pode ser feita entre **dois grupos** ou **mais de dois grupos**:

- **Dois grupos:** Por exemplo, comparar o desempenho entre dois métodos de ensino.
- **Mais de dois grupos:** Por exemplo, comparar o efeito de três diferentes dietas no peso corporal.

## 19.4 Testes Paramétricos e Não Paramétricos

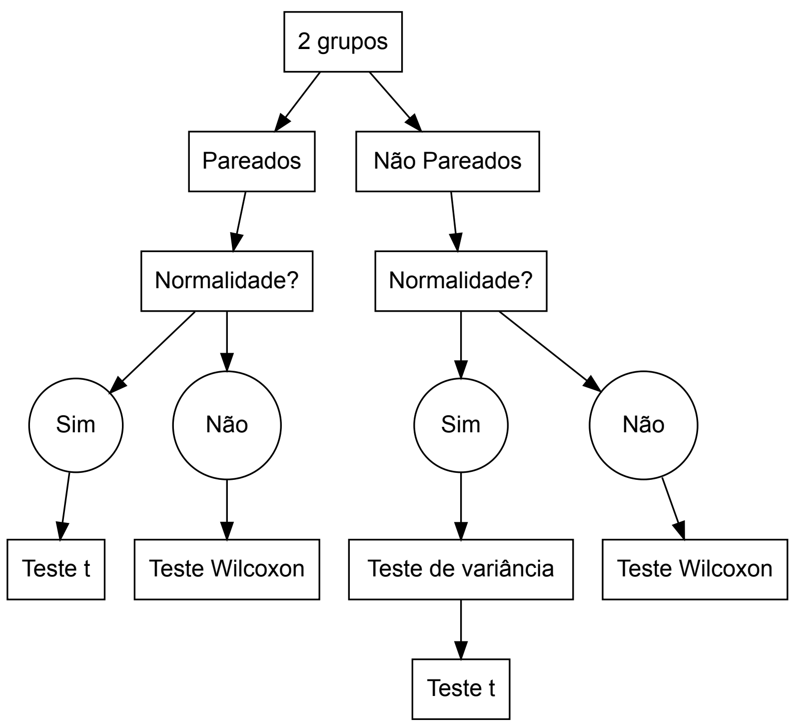
Para escolher o teste estatístico adequado, é importante considerar as características dos dados:

- **Testes Paramétricos:**
  Utilizados quando os dados seguem certas condições, como distribuição normal e variância homogênea, permitindo uma análise mais precisa.
- **Testes Não Paramétricos:**
  Indicados quando os dados não atendem às condições dos testes paramétricos, como em distribuições assimétricas ou dados ordinais, oferecendo uma alternativa robusta para comparação.

# Chapter 20

# Comparação entre dois grupos

O fluxograma abaixo apresenta a sequência de etapas para a escolha do teste estatístico mais adequado na comparação entre dois grupos, considerando o tipo de comparação entre os dois grupos (pareada ou não pareada) e a verificação da normalidade dos dados.




Exemplos didáticos para comparação entre dois grupos são mostrados a seguir.

## 20.1 Teste t pareado

O teste t (pareado ou não pareado) é um teste **paramétrico**, por isso, antes de aplicar o teste, recomenda-se verificar a normalidade dos dados.

O teste t pareado avalia se há diferença entre as médias de dois conjuntos dependentes (ex: antes e depois em um mesmo grupo).

**Hipóteses:**

H: A média antes é igual à média depois (não há efeito).

H: A média antes é diferente da média depois (há efeito).

```r
# Simulando dados: pressão antes e depois
set.seed(123)
antes <- rnorm(20, mean=120, sd=10)
depois <- antes + rnorm(20, mean=-5, sd=8) # espera-se uma redução média de 5

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

```
##
##  Paired t-test
##
## data:  antes and depois
## t = 3.644, df = 19, p-value = 0.001726
## alternative hypothesis: true mean difference is not equal to 0
## 95 percent confidence interval:
##  2.302671 8.517443
## sample estimates:
## mean difference
##        5.410057
```

No R, fazer o teste t pareado é muito fácil. Você só precisa usar uma linha: `t.test(antes, depois, paired = TRUE)`.

Basta colocar os seus dois conjuntos de dados (por exemplo, as medidas antes e depois) e escrever **`paired = TRUE`** para avisar ao R que os dados são do mesmo grupo em dois momentos.

Assim, o R faz todo o cálculo para você!

Ao realizar o teste t pareado no R, a saída apresenta várias informações importantes:

- **t**: Este é o valor do teste t calculado a partir dos dados. Ele indica o quanto a diferença média observada entre os grupos se afasta do que seria esperado se não houvesse diferença real.
- **df**: Significa "degrees of freedom" (graus de liberdade), que neste caso é igual ao número de pares menos 1 (aqui, 19).
- **p-value**: É a probabilidade de observarmos uma diferença igual ou maior que a encontrada, caso a hipótese nula (de que não há diferença) seja verdadeira. Um p-valor pequeno (por exemplo, menor que 0,05) indica que a diferença é estatisticamente significativa.
- **alternative hypothesis**: Mostra qual hipótese alternativa está sendo testada – neste caso, que a diferença média entre "antes" e "depois" é diferente de zero.
- **95 percent confidence interval**: Indica o intervalo de valores no qual a verdadeira diferença média está com 95% de confiança. Aqui, podemos dizer que a diferença média entre os grupos está provavelmente entre 2,30 e 8,52.
- **mean difference**: É a diferença média observada nos dados, neste exemplo, aproximadamente 5,41.

**Resumo:**

O resultado sugere que há uma diferença significativa entre "antes" e "depois" ($p = 0{,}0017$), com a diferença média estimada em 5,41 unidades e um intervalo de confiança que não inclui zero.

> **Observação**: Se p é maior que 0,05, não rejeitamos H, nesse caso a conclusão seria que não há evidência de diferença significativa.

## 20.2 Wilcoxon pareado

O teste de Wilcoxon para amostras pareadas é a alternativa não paramétrica ao teste t pareado, sendo utilizado quando os dados não seguem uma distribuição normal.

**Hipóteses:**

> H: As distribuições dos pares são iguais (não há diferença entre antes e depois).

H: As distribuições dos pares são diferentes (há diferença entre antes e depois).

Formalmente, o teste verifica se a **distribuição das diferenças entre os pares** é simétrica em torno de zero. Quando essa condição de simetria é atendida, podemos interpretar o teste como uma comparação das **medianas** das duas situações (por exemplo, antes e depois).

**Hipóteses:**

H: A mediana das diferenças entre os pares é igual a zero (não há diferença entre antes e depois).

H: A mediana das diferenças entre os pares é diferente de zero (há diferença entre antes e depois).

*Observação: Esta formulação das hipóteses em termos de mediana é válida quando as diferenças apresentam distribuição simétrica.*

```
wilcox.test(antes, depois, paired = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  antes and depois
## V = 179, p-value = 0.004221
## alternative hypothesis: true location shift is not equal to 0
```

Ao fazer o teste de Wilcoxon pareado no R, a resposta traz as seguintes informações:

- **V**: É o valor da estatística do teste de Wilcoxon, calculado a partir das diferenças entre os pares. Não precisamos interpretar esse número diretamente; ele serve para o cálculo do p-valor.
- **p-value**: É a chance de observarmos uma diferença igual ou maior que a encontrada, caso não exista diferença real entre os grupos. Se o p-valor for menor que 0,05, dizemos que há diferença significativa entre "antes" e "depois".
- **alternative hypothesis**: Mostra qual hipótese alternativa foi testada. Nesse caso, que a posição central (mediana) dos dois momentos não é igual.

No R, por padrão, a função `wilcox.test()` não mostra o intervalo de confiança para a mediana das diferenças. No entanto, você pode pedir para calcular o intervalo de confiança usando o argumento **`conf.int = TRUE`**:

```
wilcox.test(antes, depois, paired = TRUE, conf.int = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  antes and depois
## V = 179, p-value = 0.004221
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##  2.291397 8.680443
## sample estimates:
## (pseudo)median
##       5.545154
```

Observações sobre o intervalo de confiança do teste de Wilcoxon

**95 percent confidence interval:**
O intervalo de confiança de 95% vai de 2,29 até 8,68. Isso significa que, com 95% de confiança, a verdadeira mediana da diferença entre "antes" e "depois" está entre esses valores.
Como esse intervalo **não inclui o zero**, temos mais uma indicação de que existe diferença significativa entre os dois momentos.

**(pseudo)median:**
O valor de 5,55 indica a mediana das diferenças observadas nos dados.
É chamado de "pseudo-median" porque, no teste de Wilcoxon, esse valor é uma estimativa robusta do deslocamento central entre os pares.

O termo "pseudo-median" aparece no teste de Wilcoxon porque, nesse teste, a maneira de calcular a "mediana" é um pouco diferente da mediana comum.

**Mediana comum:** É o valor do meio quando colocamos todos os números em ordem.

**Pseudo-median:** No teste de Wilcoxon, em vez de pegar só o valor do meio, o cálculo usa todos os pares possíveis de diferenças entre os grupos. Ele faz uma média especial desses valores do meio. Por isso, chama-se "pseudo" (uma "falsa" ou "quase" mediana).

## 20.3 Teste t não pareado

Se você olhar no fluxograma apresentado no início do capítulo, você verá que antes de executar o teste t para amostras não pareadas, é necessário executar o teste de variância.

### 20.3.1 Teste de variância

**É um pré-requisito para o teste t não pareado** Verifica se as variâncias de dois grupos não pareados podem ser consideradas iguais ou não.

**Hipóteses:**

H: As variâncias dos dois grupos são iguais.

H: As variâncias dos dois grupos são diferentes.

**Interpretação:**

Se $p < 0{,}05$, rejeitamos H e concluímos que as variâncias são diferentes

No teste t acrescentado o argumento `var.equal=FALSE` ou `var.equal=F`.

Se $p > 0{,}05$, consideramos as variâncias iguais.

No teste t acrescentado o argumento `var.equal=TRUE` ou `var.equal=T`

**Importante:**

O argumento **`var.equal`** dentro de **`t.test()`** define qual versão do teste será utilizada: se `var.equal = TRUE`, o teste assume variâncias iguais; se `var.equal = FALSE`, é usada a correção de Welch, que não assume homogeneidade. A escolha incorreta desse parâmetro pode comprometer a validade do p-valor obtido, afetando a conclusão do teste.

**Explicação didática**

Imagine que você quer comparar as médias de duas turmas. A Turma A tem poucos alunos e notas parecidas; a Turma B tem muitos alunos, mas as notas variam bastante.

Se você comparar as médias assumindo que ambas têm a mesma variação, o resultado pode ser injusto.

A correção de Welch resolve isso: ela ajusta o cálculo do teste para considerar que as turmas são diferentes — olhando para a variação e o tamanho de cada grupo separadamente. Assim, o teste fica mais confiável quando os grupos são desiguais.

---

O teste t para amostras não pareadas compara médias de dois grupos independentes.

**Hipóteses:**

H: A média do grupo A é igual à média do grupo B.

H: As médias dos grupos são diferentes.

```r
# Simulando dados
set.seed(456)
grupoA <- rnorm(30, mean=50, sd=7)
grupoB <- rnorm(30, mean=55, sd=7)
```

```r
# Teste de variância (pré-requisito do t independente)
var.test(grupoA, grupoB)
```

```
##
##  F test to compare two variances
##
## data:  grupoA and grupoB
## F = 1.8842, num df = 29, denom df = 29, p-value = 0.09347
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.896797 3.958627
## sample estimates:
## ratio of variances
##           1.884167
```

```r
# Teste t independente
t.test(grupoA, grupoB, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  grupoA and grupoB
## t = -2.4453, df = 58, p-value = 0.01753
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.1306016 -0.8110054
## sample estimates:
## mean of x mean of y
##   51.6222   56.0930
```

**Descrição dos resultados do teste F para comparação de variâncias**

- **F = 1.8842**: Este é o valor da estatística F, que compara as duas variâncias. Ele é calculado dividindo a variância do grupo com maior variância pela variância do outro grupo.
- **num df = 29, denom df = 29**: Esses são os graus de liberdade, relacionados ao tamanho de cada grupo (n - 1), neste caso ambos com 30 participantes.
- **p-value = 0.09347**: Este é o valor de significância. Ele mostra a probabilidade de observarmos uma diferença entre as variâncias tão grande quanto a encontrada, caso as variâncias dos grupos fossem realmente iguais. Como o p-valor é maior que 0,05, não podemos dizer que as variâncias são diferentes de forma significativa.
- **alternative hypothesis: true ratio of variances is not equal to 1**: Isso mostra que o teste está verificando se as variâncias são diferentes (não iguais).
- **95 percent confidence interval: 0.896797 a 3.958627**: Com 95% de confiança, o verdadeiro valor da razão entre as variâncias está entre aproximadamente 0,90 e 3,96. Como esse intervalo inclui o valor 1, não há diferença significativa entre as variâncias.
- **ratio of variances = 1.884167**: Esse é o valor observado da razão entre as variâncias dos dois grupos. O grupo com maior variância tem uma variância cerca de 1,88 vezes maior que o outro grupo.

**Resumo:**
O teste F mostrou que não há diferença estatisticamente significativa entre as variâncias dos grupos A e B, pois o p-valor é maior que 0,05 e o intervalo de confiança inclui o valor 1.

**Descrição dos resultados do teste t para duas amostras não pareadas**

- **t = -2.4453**: Este é o valor da estatística t, que indica o quanto as médias dos grupos A e B diferem em relação à variação dos dados.
- **df = 58**: São os graus de liberdade do teste, relacionados ao tamanho das amostras.
- **p-value = 0.01753**: O p-valor representa a probabilidade de encontrar uma diferença igual ou maior que a observada entre as médias, caso realmente não haja diferença entre os grupos. Como o p-valor é

menor que 0,05, podemos considerar que há diferença estatisticamente significativa entre as médias de grupoA e grupoB.

- **alternative hypothesis: true difference in means is not equal to 0**: O teste verifica se existe diferença entre as médias dos dois grupos (hipótese alternativa de diferença).
- **95 percent confidence interval: -8.13 a -0.81**: Com 95% de confiança, a diferença real entre as médias está entre -8,13 e -0,81. Como esse intervalo não inclui o zero, reforça a indicação de diferença significativa.
- **mean of x (grupoA): 51.62**
  **mean of y (grupoB): 56.09**
  As médias dos grupos mostram que o grupo B teve, em média, um valor maior que o grupo A.

**Resumo:**
O teste t indicou que existe uma diferença estatisticamente significativa entre as médias dos grupos A e B. O grupo B apresentou média maior que o grupo A, e essa diferença dificilmente ocorreu ao acaso.

## 20.4 Wilcoxon não pareado

O teste é também conhecido como Wilcoxon-Mann-Whitney. É a alternativa não paramétrica ao t independente para comparar dois grupos independentes.

**Hipóteses:**

H: As distribuições dos grupos são iguais.

H: As distribuições dos grupos são diferentes.

```
wilcox.test(grupoA, grupoB)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  grupoA and grupoB
## W = 316, p-value = 0.04789
## alternative hypothesis: true location shift is not equal to 0
```

**Descrição dos resultados do teste de Wilcoxon para duas amostras não pareadas**

- **W = 316**: Este é o valor da estatística de Wilcoxon, que representa a soma dos postos atribuídos aos valores dos grupos ao comparar suas distribuições.
- **p-value = 0.04789**: O valor de p indica a probabilidade de observar uma diferença igual ou maior entre os grupos, caso não exista diferença real. Como é menor que 0,05, isso sugere que há diferença estatisticamente significativa entre os grupos A e B.
- **alternative hypothesis: true location shift is not equal to 0**: O teste está avaliando se há diferença (deslocamento) entre as localizações centrais dos dois grupos (medianas diferentes).

**Resumo:**
O teste de Wilcoxon mostrou que há uma diferença estatisticamente significativa nas distribuições dos grupos A e B, indicando que eles provavelmente apresentam medianas diferentes.

## 20.5 Teste bilateral vs. teste unilateral

### 20.5.1 O que são?

- **Teste bilateral (bicaudal ou two-sided):**
  - Verifica se há diferença **em qualquer direção** entre os grupos ou condições.
  - Exemplo: "Será que a média do grupo A é **diferente** da média do grupo B?" (pode ser maior ou menor).
- **Teste unilateral (caudal ou one-sided):**
  - Verifica se há diferença **em uma direção específica**.

- Exemplo: "Será que a média do grupo A é **maior** que a média do grupo B?" ou "Será que é **menor**?".

---

### 20.5.2 Formulação das hipóteses

- **Teste bilateral:**
  - Hipótese nula ($H_0$): não há diferença ($_A = _B$)
  - Hipótese alternativa ($H_1$): há diferença ($_A \; _B$)
- **Teste unilateral (maior):**
  - Hipótese nula ($H_0$): $_A \; _B$
  - Hipótese alternativa ($H_1$): $_A > _B$
- **Teste unilateral (menor):**
  - Hipótese nula ($H_0$): $_A \; _B$
  - Hipótese alternativa ($H_1$): $_A < _B$

---

### 20.5.3 Influência no cálculo do p-valor

- **Teste bilateral:**
  O p-valor representa a probabilidade de encontrar um resultado tão extremo quanto o observado **em ambas as direções** (para mais ou para menos).
  É mais rigoroso, pois considera desvios para cima e para baixo.
- **Teste unilateral:**
  O p-valor considera apenas **uma direção** (maior ou menor).
  Geralmente, o p-valor do teste unilateral é a metade do bilateral, para o mesmo dado e direção, tornando o teste **mais sensível** a detectar diferenças naquela direção, mas exige justificativa teórica para ser usado.

---

### 20.5.4 O que muda nas funções `t.test()` e `wilcox.test()`

Nas duas funções, você controla o tipo de teste pelo argumento `alternative`:

- **Teste bilateral (padrão):**
  - `alternative = "two.sided"`
- **Teste unilateral (maior):**
  - `alternative = "greater"`
- **Teste unilateral (menor):**
  - `alternative = "less"`

**Exemplo:**

```
# Teste t bilateral (padrão)
t.test(x, y, alternative = "two.sided")

# Teste t unilateral (x maior que y)
t.test(x, y, alternative = "greater")

# Teste de Wilcoxon unilateral (x menor que y)
wilcox.test(x, y, alternative = "less")
```

Ou seja: - **O argumento `alternative` define se o teste é bilateral ou unilateral.** - Isso altera a formulação das hipóteses, o cálculo do p-valor e a interpretação do resultado.

## 20.6 Exercício 1

**Comparação da velocidade dos Pokémons verdes e amarelos**

1. **Baixe o banco de dados**
   Use o banco de dados: Pokemon.csv
   Importe o arquivo **Pokemon.csv** para o RStudio.

2. **Pergunta do exercício**
   Teste se existe diferença significativa entre a velocidade (**speed**) dos Pokémons verdes (**green**) e amarelos (**yellow**).

3. **Formulação das hipóteses de acordo com os testes**

   Hipótese nula ($H_0$):

   Hipótese alternativa ($H_1$):

4. **Nível de significância**
   Considere $= 0, 05$.

5. **Passos para executar o teste adequado**

   a) Separe os dados dos Pokémons de cor verde e de cor amarela.

   b) Verifique a normalidade dos dados de velocidade de cada grupo (gráfico QQ).

   c) Escolha o teste mais apropriado:
      - Se as duas amostras seguirem distibuição Normal teste de variância teste t
      - Se uma das amostras não seguir distibuição Normal teste de Wilcoxon

   d) Execute o teste, compare o p-valor com e conclua

6. **Conclusão**

   - Se $p$-valor $<$ 0,05: rejeite a hipótese nula e conclua que há diferença significativa nas velocidades.
   - Se $p$-valor 0,05: não rejeite a hipótese nula e conclua que não há diferença significativa nas velocidades.

**Resposta** Posit.cloud

## 20.7 Exercício 2

Considere a tabela de resultados publicados no artigo Resposta da Pressão Arterial ao Esforço em Adolescentes: Influência do Sobrepeso e Obesidade:

| Grupo | GSO Meninas (n = 24) | GE Meninas (n = 24) | p |
|---|---|---|---|
| **Idade** (anos) | 12,1 ś 1,3 | 12,0 ś 1,5 | 0,86 |
| **Peso** (kg) | 59,3 ś 12,9 | 38,8 ś 9,3** | <0,0001 |
| **Estatura** (m) | 1,53 ś 0,09 | 1,46 ś 0,10* | 0,02 |
| **IMC** (kg/mš) | 25,2 ś 3,8 | 17,9 ś 2,3** | <0,0001 |
| **RT/S** (mm) | 0,85 ś 0,19 | 1,43 ś 0,40** | <0,0001 |
| **PAS repouso** (mmHg) | 114 ś 12 | 108 ś 10 | 0,07 |
| **PAD repouso** (mmHg) | 66 ś 6 | 67 ś 8 | 0,51 |
| **PAM repouso** (mmHg) | 82 ś 7 | 81 ś 8 | 0,67 |
| **FC** (bpm) | 84 ś 10 | 87 ś 9 | 0,34 |

*Teste t-Student para amostras independentes;* p 0,05; **p 0,01.
Diferenças entre médias do grupo de sobrepeso e obesos vs eutróficos.

IMC - índice de massa corporal (peso/estaturaš);

RT/S - relação tríceps/subescapular;

PAS - pressão arterial sistólica;

PAD - pressão arterial diastólica;

PAM - pressão arterial média;

FC - frequência cardíaca média.*

**Contexto:**
A tabela acima apresenta características antropométricas e hemodinâmicas de meninas do grupo de sobrepeso/obesidade (GSO) e do grupo eutrófico (GE). Os autores utilizaram o teste t de Student para amostras independentes.

**Pergunta:**
Escolha uma variável da tabela e formule as hipóteses nula e alternativa para a comparação entre os grupos GSO e GE. Depois, descreva a conclusão dos autores com base no valor de p apresentado na tabela.

---

**Exemplo de resposta:**

**Variável escolhida:** Peso (kg)

**Hipóteses:**

- Hipótese nula (H):
  As médias de peso (kg) dos grupos GSO e GE são iguais.
- Hipótese alternativa (H):
  As médias de peso (kg) dos grupos GSO e GE são diferentes.

**Nível de significância:** $= 0{,}05$

**Valor de p na tabela:** $p < 0{,}0001$

**Conclusão:**
Como o valor de p é menor que 0,05, rejeita-se a hipótese nula. Portanto, os autores concluem que existe diferença estatisticamente significativa entre as médias de peso dos grupos GSO e GE, sendo o peso significativamente maior no grupo GSO.

**Você pensou nisso?** O peso é a variável intrínseca à comparação, pois define os grupos.

---

## 20.8 Exercício 3

O artigo Sintomas de estresse pré-competitivo em atletas adolescentes de handebol utilizou como instrumento a Lista de Sintomas de Estresse Pré-competitivo Infanto-juvenil (LSSPCI), uma escala do tipo Likert (De Rose Jr., 1998). A LSSPCI contém 31 sintomas, para os quais cada atleta responde em uma escala de 1 (nunca) a 5 (sempre) sobre a frequência com que cada situação ocorreu nas 24 horas anteriores à competição. Os escores dos sintomas são somados para cada atleta, resultando em um escore total de estresse. Analise os resultados da Tabela 1 do artigo:

**Tabela 1. Médias, desvios-padrão, valores p da diferença de médias e mediana dos sintomas de estresse medidos pela LSSPCI em atletas adolescentes de handebol, segundo o sexo (n = 97)**

| Sintomas de estresse – LSSPCI | Meninos | Meninas | Valor p |
|---|---|---|---|
| Meu coração bate mais rápido que o normal | 2,2 ś 1,0 (2,0) | 2,5 ś 0,9 (2,5) | 0,1 |
| Suo bastante | 2,5 ś 1,2 (2,0) | 2,4 ś 1,3 (2,0) | 0,4 |
| Fico agitado (a) | 3,0 ś 0,9 (3,0) | 3,2 ś 1,3 (3,0) | 0,5 |

| Sintomas de estresse – LSSPCI | Meninos | Meninas | Valor p |
|---|---|---|---|
| Fico preocupado (a) com críticas das pessoas | 2,9 ś 1,5 (3,0) | 2,6 ś 1,4 (2,0) | 0,2 |
| Sinto muita vontade de fazer xixi | 2,0 ś 1,3 (1,0) | 2,0 ś 1,2 (2,0) | 0,7 |
| Fico preocupado (a) com meus adversários | 2,7 ś 1,2 (3,0) | 2,9 ś 1,4 (3,0) | 0,5 |
| Bebo muita água | 2,5 ś 1,2 (3,0) | 2,9 ś 1,5 (3,0) | 0,3 |
| Roo (como) as unhas | 2,5 ś 1,9 (2,0) | 2,1 ś 1,5 (1,0) | 0,2 |
| Fico empolgado (a) | 3,5 ś 1,3 (4,0) | 3,6 ś 1,5 (4,0) | 0,6 |
| Fico aflito (a) | 2,3 ś 1,7 (2,0) | 2,8 ś 1,3 (3,0) | 0,1 |
| Tenho medo de competir mal | 2,6 ś 1,4 (3,0) | 2,9 ś 1,4 (3,0) | 0,4 |
| Demoro muito para dormir | 2,8 ś 2,4 (3,0) | 2,3 ś 2,0 (2,0) | 0,05 |
| Tenho dúvidas sobre minha capacidade de competir | 2,4 ś 1,8 (2,0) | 2,4 ś 1,2 (2,0) | 0,7 |
| Sonho com a competição | 2,4 ś 1,2 (2,0) | 1,9 ś 1,3 (1,0) | 0,04 |
| Fico nervoso (a) | 3,1 ś 1,6 (3,0) | 3,3 ś 1,2 (3,0) | 0,6 |
| Fico preocupado (a) com o resultado da competição | 3,1 ś 1,4 (3,0) | 3,7 ś 1,4 (4,0) | 0,1 |
| Minha boca fica seca | 2,4 ś 1,3 (2,0) | 2,3 ś 1,4 (2,0) | 0,6 |
| Sinto muito cansaço no fim do treino | 2,6 ś 1,3 (2,0) | 2,7 ś 1,3 (2,0) | 0,9 |
| A presença de meus pais na competição me preocupa | 3,1 ś 1,7 (2,0) | 2,6 ś 1,6 (2,0) | 0,1 |
| Falo muito sobre a competição | 3,0 ś 1,5 (3,0) | 3,2 ś 1,3 (3,0) | 0,6 |
| Tenho medo de perder | 2,9 ś 1,4 (3,0) | 3,1 ś 1,4 (3,0) | 0,4 |
| Fico impaciente | 2,4 ś 1,1 (2,0) | 2,6 ś 1,3 (2,5) | 0,5 |
| Não penso em outra coisa a não ser na competição | 2,2 ś 1,2 (2,0) | 2,7 ś 1,4 (2,0) | 0,2 |
| Não vejo a hora de competir | 3,2 ś 1,4 (3,0) | 3,3 ś 1,4 (3,0) | 0,7 |
| Fico emocionado (a) | 1,8 ś 1,3 (1,0) | 2,3 ś 1,3 (2,0) | 0,1 |
| Fico ansioso (a) | 3,6 ś 1,2 (3,0) | 3,4 ś 1,5 (4,0) | 0,7 |
| No dia da competição acordo mais cedo do que o normal | 3,1 ś 1,7 (3,0) | 3,0 ś 1,6 (3,0) | 0,8 |
| Tenho medo de decepcionar as pessoas | 2,7 ś 1,4 (3,0) | 3,0 ś 1,5 (3,0) | 0,3 |
| Sinto-me mais responsável | 3,1 ś 1,3 (3,0) | 2,7 ś 1,3 (3,0) | 0,2 |
| Sinto que as pessoas exigem muito de mim | 2,5 ś 1,4 (2,0) | 2,4 ś 1,4 (2,0) | 0,8 |
| Tenho medo de cometer erros na competição | 3,4 ś 1,4 (3,0) | 3,6 ś 1,3 (4,0) | 0,4 |

Média ś DP (mediana); Teste de Wilcoxon não pareado; $p < 0,05$.

**Com base na Tabela 1, qual sintoma de estresse apresentou diferença estatisticamente significativa entre meninos e meninas segundo o teste de Wilcoxon? Explique como interpretar esse resultado considerando o contexto da escala LSSPCI.**

## 20.9 Exercício 4

O artigo Qualidade de vida de estudantes de Psicologia avaliou a qualidade de vida de acadêmicos de psicologia e correlacionou-a com fatores sociodemográficos. Participaram 310 alunos de psicologia que responderam a um questionário sociodemográfico e ao The Medical Outcomes Study 36-item Short-Form Health Survey (SF-36) para avaliar a qualidade de vida. Veja os resultados do teste t na Tabela 2:

**Tabela 2: Comparação entre gêneros nas dimensões de qualidade de vida**

| Dimensões | Variável | Média | DP | t | p-valor |
|---|---|---|---|---|---|
| Capacidade funcional | Masculino | 89,74 | 14,49 | 2,44 | 0,119 |
| | Feminino | 86,95 | 11,6 | | |

| Dimensões | Variável | Média | DP | t | p-valor |
|---|---|---|---|---|---|
| Aspectos físicos | Masculino | 79,74 | 25,84 | 0,76 | 0,383 |
| | Feminino | 75,82 | 31,84 | | |
| Dor | Masculino | 74,93 | 22,05 | 4,33 | 0,038 |
| | Feminino | 68,26 | 22,00 | | |
| Estado geral de saúde | Masculino | 75,21 | 20,36 | 1,15 | 0,284 |
| | Feminino | 72,32 | 17,67 | | |
| Vitalidade | Masculino | 61,38 | 22,47 | 1,50 | 0,222 |
| | Feminino | 57,92 | 18,57 | | |
| Aspectos sociais | Masculino | 71,34 | 27,61 | 0,06 | 0,804 |
| | Feminino | 70,43 | 24,31 | | |
| Aspecto emocional | Masculino | 63,16 | 40,67 | 0,38 | 0,536 |
| | Feminino | 59,51 | 39,85 | | |
| Saúde mental | Masculino | 70,55 | 19,97 | 3,05 | 0,082 |
| | Feminino | 65,7 | 18,80 | | |

**Com base na Tabela 2, em qual dimensão da qualidade de vida foi observada diferença estatisticamente significativa entre estudantes do sexo masculino e feminino? O que esse resultado sugere sobre a experiência dos alunos nesses grupos?**

# Chapter 21

# Cálculo amostral no R

O cálculo do tamanho da amostra é fundamental para garantir a validade estatística de um estudo. No R, esse cálculo pode ser realizado de forma prática utilizando diversos pacotes.

## 21.1 Principais Pacotes para Cálculo Amostral

- **pwr**: Utilizado para análise de poder estatístico e cálculo do tamanho da amostra para diferentes testes (t-test, ANOVA, correlação, etc.).
- **epiDisplay**: Oferece funções para cálculos amostrais em estudos epidemiológicos.
- **epiR**: Bastante utilizado em epidemiologia, com funções para diferentes desenhos de estudo.
- **samplesize**: Possui métodos para cálculo de tamanho amostral em diferentes contextos.

## 21.2 Exemplo de cálculo amostral para teste t

```r
# Instale o pacote se necessário
# install.packages("pwr")
library(pwr)

# Cálculo do tamanho da amostra para detectar um efeito de tamanho médio (d = 0.5)
# com poder de 80% e nível de significância de 5%
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.05, type = "two.sample")
```

O uso de funções e pacotes específicos no R torna o cálculo amostral mais acessível, permitindo maior precisão no planejamento de estudos. Sempre leve em conta os parâmetros do seu estudo, como efeito esperado, variância, poder e nível de significância.

No exemplo os seguintes parâmetros foram utilizados:

- **d = 0.5**
  *Tamanho do efeito (Effect size)*: Representa a magnitude da diferença que se espera detectar entre as médias dos grupos, em unidades de desvio padrão. Quanto maior o efeito, mais fácil é detectá-lo com uma amostra menor.
  - O tamanho do efeito pode ser estimado a partir de dados prévios, estudos semelhantes ou com base em uma suposição informada.

  - Categorias comuns: pequeno (~0,2), médio (~0,5) e grande (~0,8).
- **power = 0.8**
  *Poder estatístico (Power)*: É a probabilidade de detectar uma diferença real quando ela de fato existe (ou seja, de rejeitar corretamente a hipótese nula se ela for falsa).
  - Poder = 1 - , onde  é a probabilidade de erro tipo II (falso negativo).

  - O valor padrão mais utilizado é 0,80 (ou 80%).
- **sig.level = 0.05**
  *Nível de significância ()*: É a probabilidade de rejeitar a hipótese nula quando ela é verdadeira (erro tipo I, falso positivo).
  - O valor padrão é 0,05 (ou 5%), indicando que aceitamos até 5% de chance de um falso positivo.
- **type = "two.sample"**
  Indica que estamos comparando dois grupos independentes (teste t para duas amostras).

**Interpretação dos Resultados**

Ao rodar o comando acima, o R retorna o tamanho mínimo de amostra necessário **em cada grupo** para que seja possível detectar uma diferença de tamanho de efeito `d = 0.5` com 80% de poder e 5% de nível de significância, usando um teste t para duas amostras.

Por exemplo, o resultado pode ser algo assim:

```
     Two-sample t test power calculation

              n = 63.76561
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

**Explicação:**

- *n 64*: Você precisaria de pelo menos **64 participantes em cada grupo** para ter 80% de chance de detectar uma diferença de tamanho médio (0,5 desvios padrão) entre os grupos, com risco de 5% de um falso positivo.

**Considerações**

- O tamanho da amostra depende diretamente do tamanho do efeito esperado, do poder estatístico desejado e do nível de significância escolhido.
- Quanto maior o efeito esperado, menor precisa ser a amostra.
- Quanto maior o poder ou menor o nível de significância, maior será o tamanho da amostra necessária.
- Use dados prévios, literatura ou cálculos exploratórios para definir o tamanho do efeito.

## 21.3 O que é o d de Cohen?

O d de Cohen é uma **medida padronizada** do tamanho do efeito para comparar a diferença entre duas médias, levando em conta a variabilidade dos dados. Ele é muito utilizado para quantificar o quão grande é a diferença entre dois grupos em termos de desvios padrão.

### 21.3.1 Fórmula Geral

$$d = \frac{X_1 \;\; X_2}{s_p}$$

Onde:

- $X_1$ = média do grupo 1

- $X_2$ = média do grupo 2

- $s_p$ = desvio padrão combinado (pooled) dos dois grupos

**Como calcular o desvio padrão combinado**

$$s_p = \frac{(n_1 \ 1)s_1^2 + (n_2 \ 1)s_2^2}{n_1 + n_2 \ 2}$$

Onde:

- $n_1$, $n_2$ = tamanhos das amostras dos grupos 1 e 2
- $s_1$, $s_2$ = desvios padrão dos grupos 1 e 2

### 21.3.2 Exemplo prático em R

Suponha:

- Grupo 1: média = 10, desvio padrão = 2, n = 30
- Grupo 2: média = 8, desvio padrão = 2.5, n = 35

```r
media1 <- 10
media2 <- 8
sd1 <- 2
sd2 <- 2.5
n1 <- 30
n2 <- 35

# Desvio padrão combinado
sp <- sqrt( ((n1-1)*sd1^2 + (n2-1)*sd2^2) / (n1 + n2 - 2) )

# d de Cohen
d <- (media1 - media2) / sp
d
```

```
## [1] 0.8758557
```

**Valores típicos para d de Cohen:**

- **0.2**: efeito pequeno
- **0.5**: efeito médio
- **0.8**: efeito grande

Assim, quanto maior o valor de d, maior a diferença entre os grupos em relação à variabilidade dos dados.

**Explicação didática:**

Pense no **d de Cohen** como uma régua que compara a diferença entre as médias de dois grupos, levando em conta o quanto os dados variam dentro desses grupos.

- Se d é pequeno, significa que a diferença entre as médias dos grupos é pequena em relação à "bagunça" (variação) dos dados – ou seja, os grupos são parecidos.
- Se d é grande, significa que a diferença entre as médias é grande em comparação com a variação dos dados – ou seja, os grupos são bem diferentes.

**Resumindo:**

Um valor alto de d indica que os grupos são realmente diferentes entre si, enquanto um valor baixo mostra que eles são muito parecidos, considerando o quanto os dados "dispersam" dentro de cada grupo.

## 21.4 Exemplos práticos

A seguir, apresento exemplos práticos para Medicina, Psicologia e Educação Física.

### 21.4.1 Medicina: Comparação de dois tratamentos para dor

**Cenário:**
Um estudo clínico quer comparar dois medicamentos para dor crônica.

- Tamanho do efeito esperado (Cohen's d): 0.2 (pequeno)
- Poder desejado: 0.80
- Nível de significância: 0.05

```r
library(pwr)
# Calculando o tamanho da amostra necessária para cada grupo
pwr.t.test(d = 0.2, power = 0.8, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 393.4057
##               d = 0.2
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Tamanho da amostra: 394 em cada grupo

### 21.4.2 Psicologia: Intervenção para redução de ansiedade

**Cenário:**
Um psicólogo quer testar se uma nova terapia reduz a ansiedade em relação ao tratamento padrão.

- Tamanho do efeito esperado (Cohen's d): 0.6 (médio/grande)
- Poder desejado: 0.80
- Nível de significância: 0.05

```r
library(pwr)
# Calculando o tamanho da amostra necessária para cada grupo
pwr.t.test(d = 0.6, power = 0.8, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 44.58577
##               d = 0.6
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Tamanho da amostra: 45 em cada grupo

### 21.4.3 Educação Física: Efeito de um programa de treinamento na performance de corrida

**Cenário:**
Um preparador físico quer comparar a performance de alunos submetidos a dois programas de treinamento

diferentes.

- Tamanho do efeito esperado (Cohen's d): 0.8 (grande)

- Poder desejado: 0.80
- Nível de significância: 0.05

```r
library(pwr)
# Calculando o tamanho da amostra necessária para cada grupo
pwr.t.test(d = 0.8, power = 0.8, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 25.52458
##               d = 0.8
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

Tamanho da amostra: 26 em cada grupo

- O resultado de cada função indica o **número mínimo de participantes em cada grupo** para atingir o **poder estatístico** desejado.

## 21.5 Como calcular o poder do teste quando você já tem o tamanho da amostra

O pacote `pwr` do R não serve apenas para calcular o tamanho da amostra necessário. Ele também permite **descobrir o poder do teste** caso você já saiba quantos participantes terá em cada grupo.

### 21.5.1 Exemplo

Suponha que você fez um estudo com **26 participantes em cada grupo** e espera encontrar um **tamanho de efeito grande (d = 0.8)**. O nível de significância que você escolheu é o padrão, **0,05**. Você quer saber: *com essa amostra, qual é o poder do meu teste*?

Você pode usar:

```r
pwr.t.test(n = 26, d = 0.8, power = NULL, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 26
##               d = 0.8
##       sig.level = 0.05
##           power = 0.8074866
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

**Explicação dos parâmetros**

- `n = 20`: tamanho da amostra em cada grupo já definido.
- `d = 0.8`: tamanho do efeito (Cohen's d), aqui considerado grande.
- `sig.level = 0.05`: nível de significância (chance máxima de um falso positivo).
- `type = "two.sample"`: teste t para dois grupos independentes.

- `power = NULL`: ao deixar o parâmetro `power` em NULL, você está dizendo ao R para calcular **qual é o poder do teste** com esses valores.

Nesse caso o poder do teste é de 70%.

**O que acontece nesse cálculo?**

O R resolve a equação para o poder estatístico usando os valores que você forneceu. Ou seja, ele vai informar a **probabilidade de detectar uma diferença real (de tamanho d = 0.8) entre os grupos, com 26 participantes em cada grupo**, considerando o nível de significância especificado.

**Por que isso é útil?**

- Se o poder calculado for **menor que 0,80** (80%), pode ser interessante aumentar a amostra para garantir maior chance de detectar diferenças reais.
- Se o poder já for **maior que 0,80**, seu estudo tem boa sensibilidade para o efeito esperado.

**Resumindo**

- Você pode usar `pwr.t.test` para calcular o poder do teste, **bastando informar `n` e deixar `power = NULL`**.
- Isso é útil para analisar experimentos já realizados ou planejar com base em restrições de amostra.

## 21.6 Cálculo do Tamanho da Amostra para Testes t Pareado e Wilcoxon

Além do teste t para amostras independentes, você pode calcular o tamanho da amostra para outras situações, como o teste t pareado e os testes não paramétricos de Wilcoxon (para amostras pareadas ou não pareadas). Para os testes não paramétricos, recomenda-se aumentar o tamanho da amostra em 15% em relação ao cálculo do teste t correspondente.

### 21.6.1 Teste t Pareado

**Cenário:**
Você quer testar se uma intervenção reduz o nível de estresse dos mesmos participantes antes e depois do tratamento.

- Tamanho do efeito esperado (d de Cohen): 0.5

- Poder desejado: 0.80

- Nível de significância: 0.05

```r
library(pwr)
# Tamanho da amostra para teste t pareado
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.05, type = "paired")
```

```
##
##      Paired t test power calculation
##
##               n = 33.36713
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

Observe o argumento **type = "paired"** indicando que o cálculo é para amostras pareadas.

### 21.6.2 Teste de Wilcoxon para Amostras Não Pareadas

**Cenário:**
Você deseja comparar dois grupos de pacientes com dados não normalmente distribuídos.

- Tamanho do efeito esperado (d de Cohen): 0.5
- Poder desejado: 0.80
- Nível de significância: 0.05

```r
library(pwr)
# Tamanho da amostra para teste t de duas amostras
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

```r
# Acrescentar 15% para o teste de Wilcoxon
63.76 * 1.15
```

```
## [1] 73.324
```

Acrescentando 15% ao cálculo amostral realizado com o `pwr.t.test`, são necessárias 74 amostras. Os testes não paramétricos, como o teste de Wilcoxon, geralmente apresentam menor poder estatístico em comparação aos testes paramétricos. Para compensar essa diferença e obter um poder semelhante ao dos testes paramétricos, recomenda-se aumentar o tamanho da amostra em cerca de 15%. Assim, o teste não paramétrico passa a ter uma chance semelhante de detectar um efeito real, caso ele exista.

### 21.6.3 Teste de Wilcoxon Pareado

**Cenário:**
Você avalia o desempenho dos mesmos alunos antes e depois de um programa de treinamento, com dados não normalmente distribuídos.

- Tamanho do efeito esperado (d de Cohen): 0.5
- Poder desejado: 0.80
- Nível de significância: 0.05

```r
library(pwr)
# Tamanho da amostra para teste t pareado
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.05, type = "paired")
```

```
##
##      Paired t test power calculation
##
##               n = 33.36713
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

```
##
## NOTE: n is number of *pairs*
```

```
# Acrescentar 15% para o teste de Wilcoxon
34 * 1.15
```

```
## [1] 39.1
```

Acrescentando 15% ao cálculo amostral realizado com o `pwr.t.test`, são necessárias aproximadamente 40 amostras.

# Chapter 22

# Memes de Estatística: p-valor

O p-valor é um dos conceitos mais populares (e polêmicos) da estatística. A seguir, veja alguns memes famosos sobre p-valor, acompanhados de interpretações e tradução para o português.

Memes ajudam a ilustrar de forma divertida como o p-valor é frequentemente mal interpretado ou supervalorizado. Rejeitar H0 não é tudo: pense criticamente sobre seus resultados!

## 22.1 Quando você encontra um p-valor baixo e sabe que deve rejeitar H0

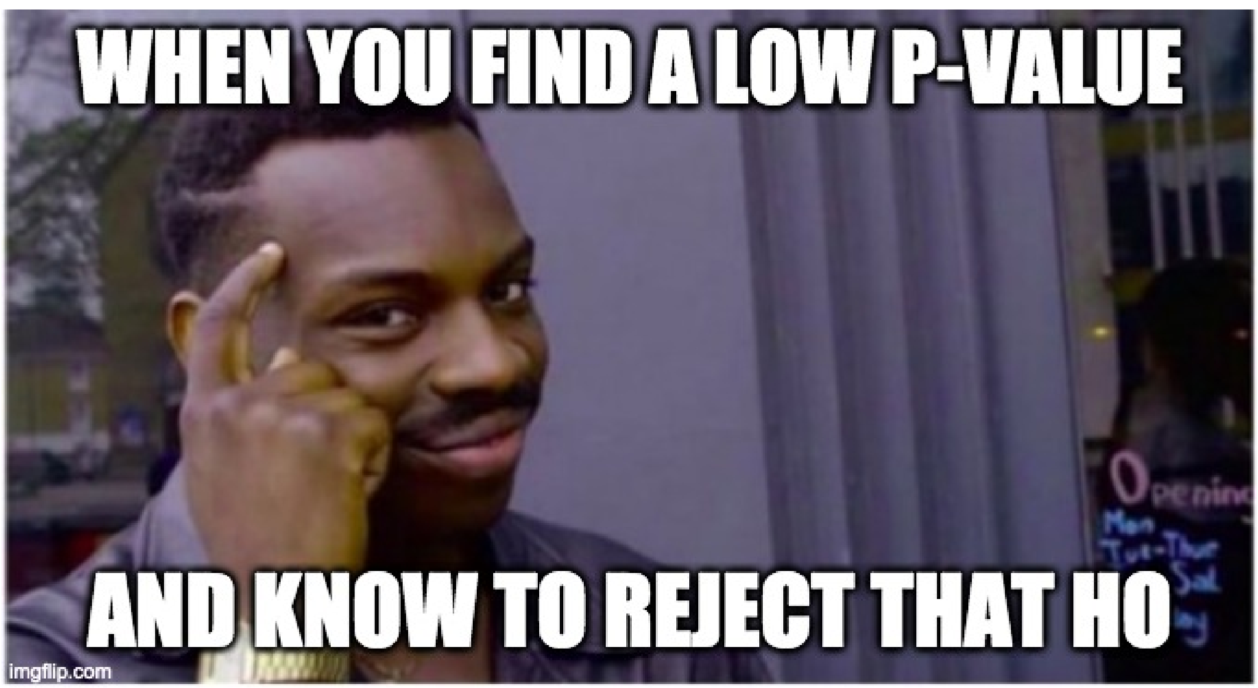



Figure 22.1: Fonte: https://library.fiveable.me/ap-stats/unit-6/concluding-test-for-population-proportion /study-guide/THZeUpkm11DAwnNb6p4g

**Tradução:**
Quando você encontra um p-valor baixo e sabe que deve rejeitar a hipótese nula.

**Interpretação:**
O meme brinca com o entusiasmo de muitos pesquisadores ao encontrar um p-valor baixo (menor que 0,05),

pois sabem que isso permite rejeitar a hipótese nula (H0). No entanto, é importante lembrar que rejeitar H0 não garante que o resultado seja relevante do ponto de vista prático ou científico. O p-valor indica o quanto os resultados observados seriam inesperados caso a hipótese nula fosse verdadeira, mas não informa o tamanho ou a relevância prática dessa diferença encontrada.

---

## 22.2 Homem-Aranha apontando

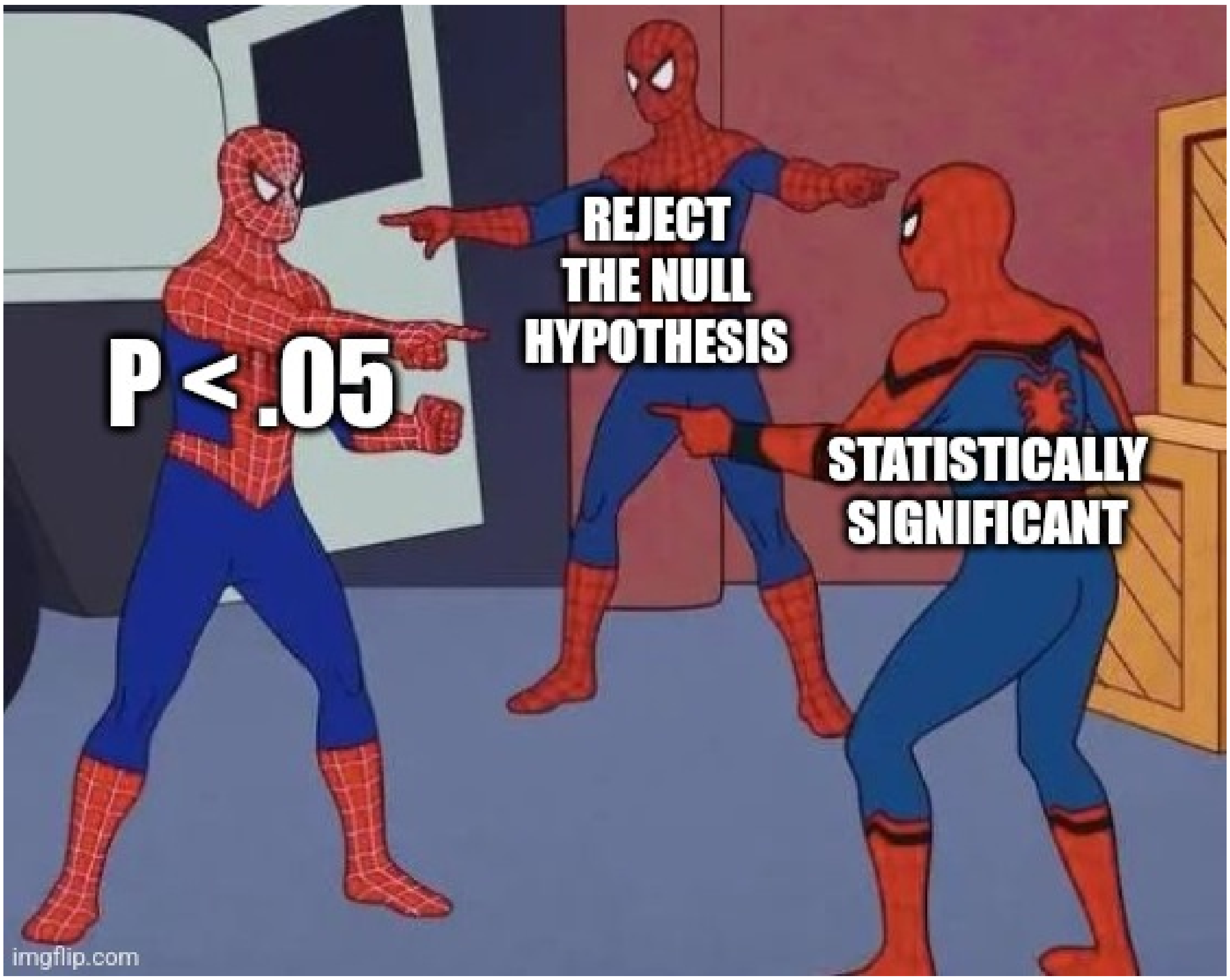



Figure 22.2: Fonte: https://danawanzer.github.io/stats-with-jamovi/alpha-p-values.html

**Tradução:**
P < 0,05
Rejeitar a hipótese nula
Estatisticamente significativo

**Interpretação:**
Esse meme mostra como as pessoas frequentemente confundem ou usam como sinônimos as expressões "p < 0,05", "rejeitar a hipótese nula" e "significância estatística". Na prática, são conceitos relacionados, mas não exatamente iguais: um p-valor menor que o nível de significância leva à rejeição de H0, o que é chamado de resultado estatisticamente significativo. Mas lembre-se: um resultado pode ser estatisticamente significativo e, mesmo assim, não ter importância prática no mundo real.

---

## 22.3 Namorado distraído


Figure 22.3: Fonte: https://imgflip.com/i/2b00xk

**Tradução:**
p-valor 0,083
>
Nível de significância 0,05

**Interpretação:**
O pesquisador está andando tranquilamente com sua fiel companheira, o nível de significância 0,05, mas não resiste e joga aquele olhar para o p-valor 0,083 que acabou de passar. Mesmo sabendo que, tecnicamente, 0,083 é "maior" do que 0,05 e não deveria chamar tanta atenção, ele fica tentado a inventar desculpas para dar uma chance àquele resultado quase-significativo: "ah, mas foi quase!", "quem nunca, né?". No fundo, esse é o típico caso do pesquisador que se empolga com resultados que chegaram perto do valor de corte, tentando transformar um "quase" em uma grande descoberta. Mas é importante lembrar: resultado quase significativo ainda não é significativo!

---

## 22.4 Willy Wonka e o p-valor menor que 0,05

**Tradução:**
Ah, você encontrou um p-valor menor que 0,05?
Por favor, me conte tudo sobre sua grande descoberta.

**Interpretação:**
Willy Wonka ironiza a empolgação de quem encontra um p-valor abaixo de 0,05, sugerindo que nem sempre isso significa uma grande descoberta. O meme alerta para a importância de interpretar o p-valor no contexto


Figure 22.4: Fonte: https://naturallyspeaking.blog/2015/06/03/episode-25-the-problem-with-p-values/

do estudo, considerando tamanho de efeito, relevância prática e outros fatores além da simples significância estatística. O próprio Wonka pediria: "me convença de que seu resultado importa de verdade!"

---

## 22.5 Cantada Nerd

**Tradução**:
Ei, garota, seu p deve ser maior que 0,05, porque eu falho em te rejeitar.

**Interpretação**:
Esse meme faz uma brincadeira romântica usando a linguagem estatística: em vez de rejeitar a hipótese nula, a pessoa está dizendo que não consegue rejeitar a garota, assim como um p-valor acima de 0,05 geralmente significa que não rejeitamos H0. É o clássico "não posso te rejeitar, cientificamente falando"! Uma piada para conquistar corações e estatísticos ao mesmo tempo.

---

## 22.6 P-valor não é tudo

**Tradução:**
Cena de duas pessoas no carro, uma com expressão de frustração ao ouvir "p-values", enquanto a outra está alegre ou satisfeita.

**Interpretação:**
A graça do meme está justamente no contraste entre frustração e alegria: enquanto uma pessoa demonstra cansaço ou desapontamento ao ouvir sobre p-values de novo, a outra parece satisfeita só com isso. O meme brinca com as diferentes expectativas na análise estatística, há quem se contente apenas com a significância estatística ($p < 0,05$), enquanto outros esperam uma análise completa, considerando também o tamanho do efeito, intervalos de confiança, e a real relevância dos achados. É um lembrete de que estatística vai além do p-value!

---

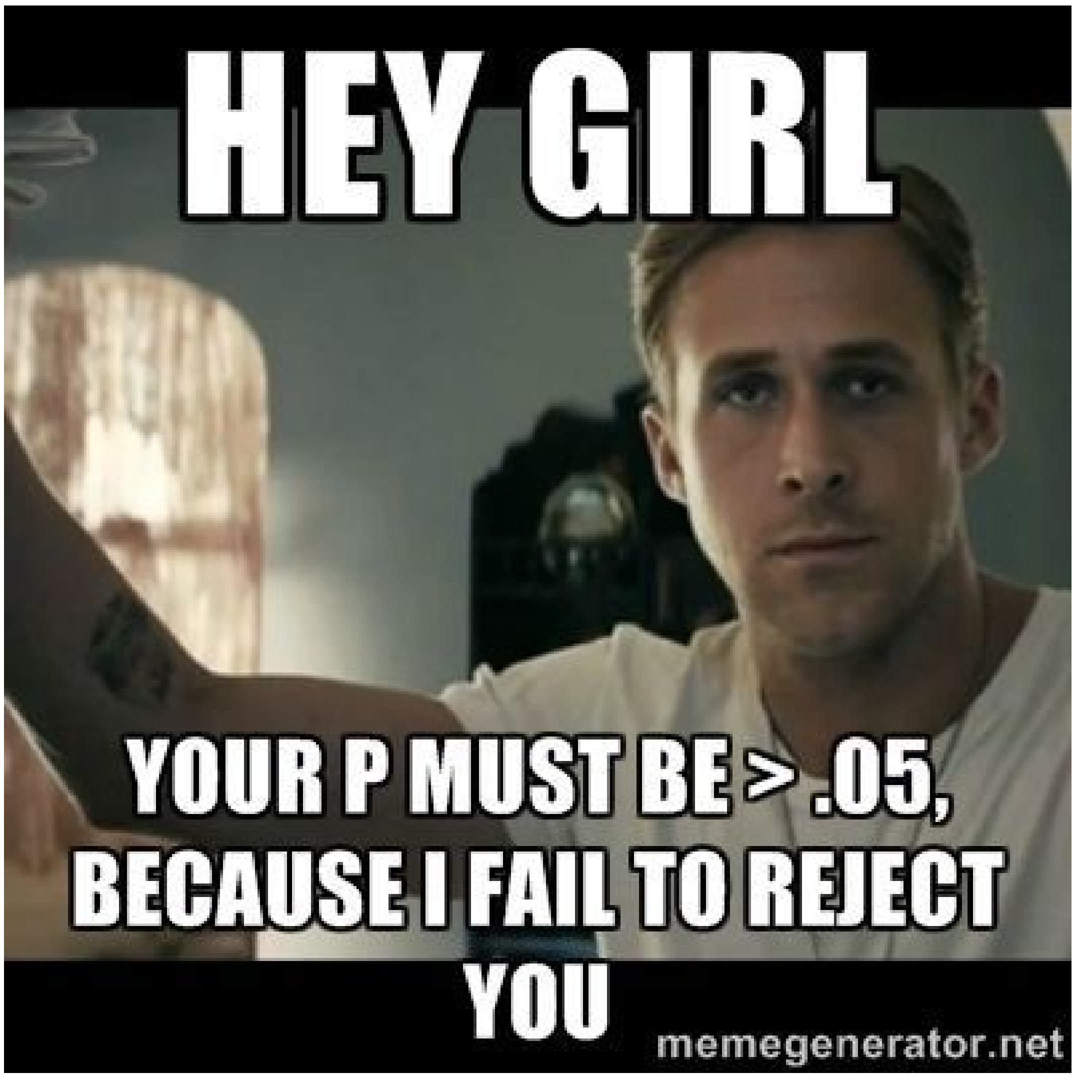



Figure 22.5: Fonte: https://varshitasher.medium.com/memorizing-p-value-interpretation-through-coronavirus-76e36606b5ff

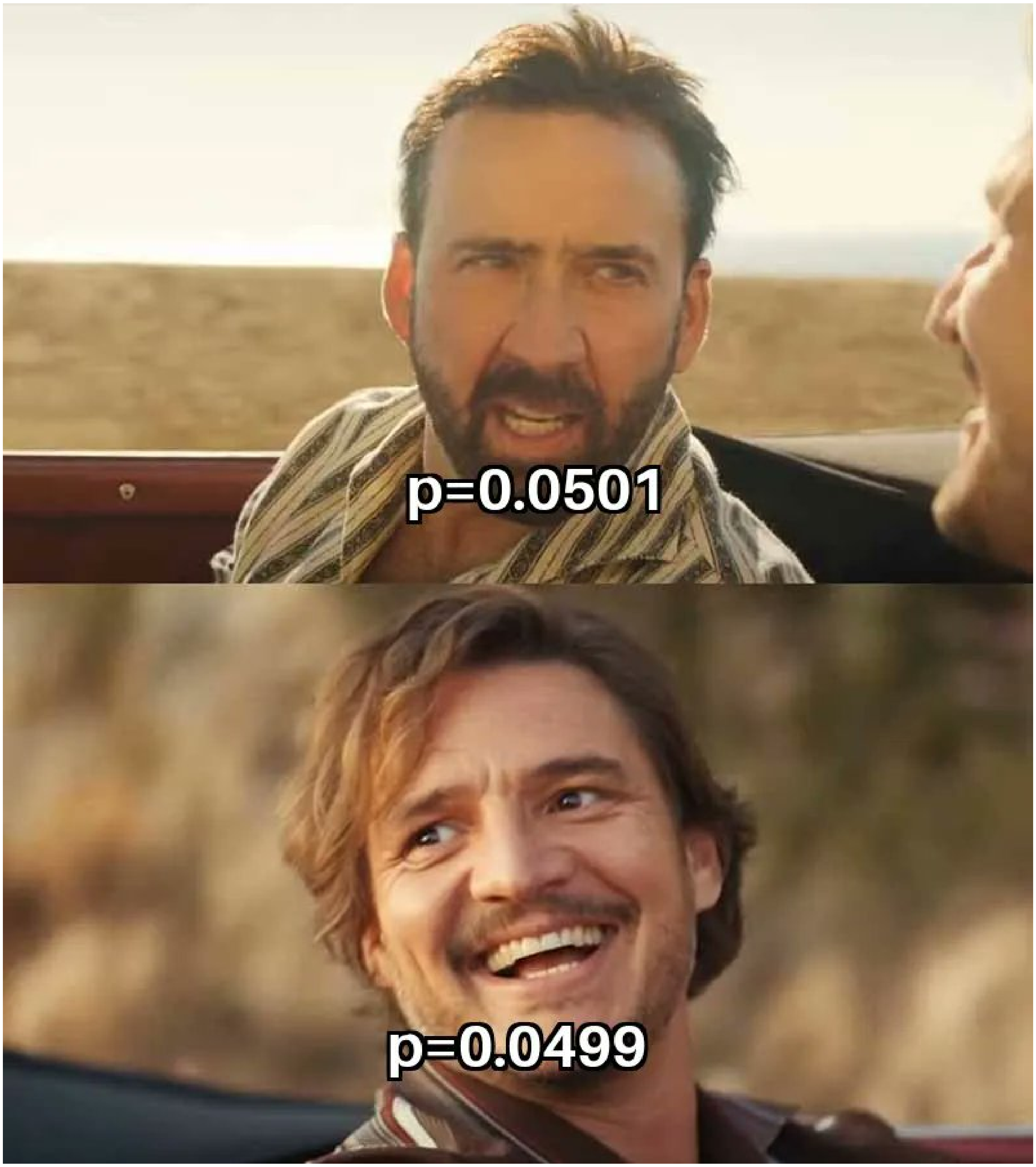
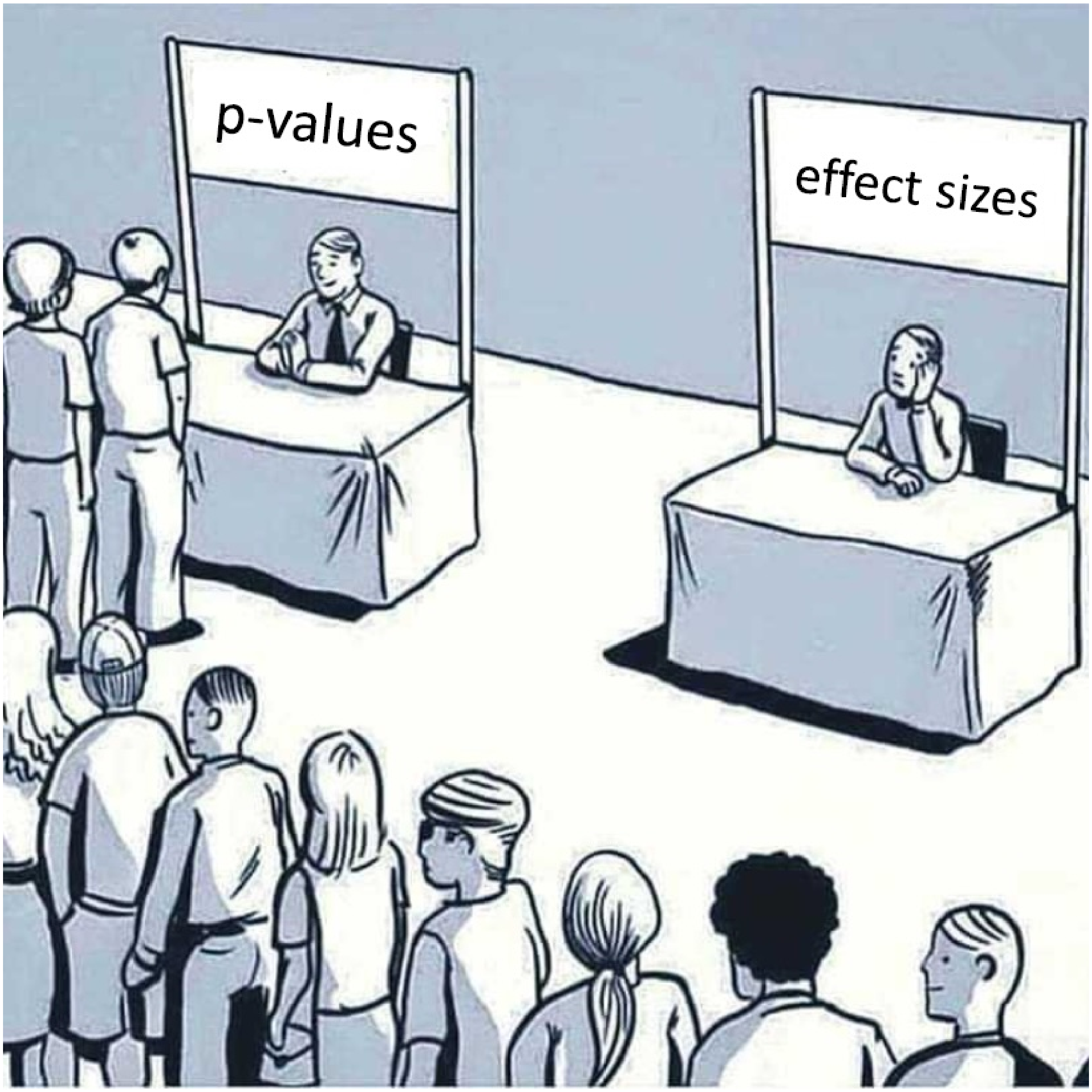



Figure 22.6: Fonte: https://x.com/data_question/status/1675416773665890304

## 22.7 Coitado do tamanho do efeito




**Tradução:**
À esquerda, uma fila enorme se forma na barraca "p-values". À direita, a barraca "effect sizes" está vazia, quase ouvindo grilos.

**Interpretação:**
Este meme é o retrato da triste (e hilária) realidade dos artigos científicos: todo mundo corre para saber se o valor de p é menor que 0,05, como se fosse a última coca-cola do deserto. Já o tamanho do efeito, que realmente diz se o resultado é relevante, é solenemente ignorado. Moral estatística: enquanto os p-values lotam (mesmo que tragam resultados irrelevantes), o effect size fica sozinho no canto, esperando alguém notar sua real importância. Quem entende estatística sabe: tamanho do efeito merece atenção!


Figure 22.7: Fonte: https://lovestats.wordpress.com/dman/survey-research-statistics-meme/

## 22.8 Senhora estatística

**Tradução:**
Em cima: "ONE DOES NOT SIMPLY" ("SIMPLESMENTE NÃO SE...") Embaixo: "REPORT P-VALUES WITHOUT EFFECT SIZES" ("RELATA P-VALORES SEM TAMANHOS DE EFEITO")

**Interpretação:**
Esse meme faz referência ao clássico da cultura pop (Senhor dos Aneis) para lembrar que estatística não é terra sem lei: não basta encontrar um $p < 0{,}05$ e sair comemorando. Relatar apenas o valor de p é igual a contar só metade da fofoca: falta o contexto! O tamanho do efeito é quem diz se a diferença é daquelas que mudam a vida ou se é só uma diferença microscópica e irrelevante. Em outras palavras, estatística de verdade combina significado estatístico (p-value) com significado prático (effect size). Não faça estatística "pela metade"!

---

## 22.9 Pobi do gatinho

**Tradução:**
Primeira cena: Pessoa abraça um cachorro chamado "P-VALUES". Atrás, um gato (effect sizes) observa, excluído.
Demais cenas: Close no gato "effect sizes", visivelmente triste, segurando um brinquedo, ignorado.

**Interpretação:**
O meme mostra o drama dos tamanhos de efeito: sempre deixados de lado, enquanto os p-values recebem todo o carinho e atenção. É como se o cachorro (p-value) ganhasse petiscos só por latir, enquanto o gato (effect size), que realmente entrega o conteúdo, só observa de longe. Estatisticamente, é como celebrar um resultado significativo sem saber se ele realmente importa no mundo real. O gato chora, e o revisor do artigo também!

---

Esses memes ilustram, com bom humor, que o p-valor sozinho não conta toda a história. Um resultado estatisticamente significativo não garante relevância prática. O tamanho do efeito é fundamental para avaliar se uma diferença observada tem impacto real, e não apenas estatístico. Ao analisar dados, valorize tanto o p-valor quanto o tamanho do efeito para conclusões mais robustas e informativas.


Figure 22.8: Fonte: https://x.com/Jente_O/status/1698942647493153251

# Chapter 23

# Pacotes `pwr`, `rstatix` e `effsize`

Nesse capítulo, abordaremos como calcular tamanho de amostra, poder estatístico e tamanho do efeito em diferentes testes estatísticos usando os pacotes `pwr`, `rstatix` e `effsize`. O foco será nos testes t (pareado e não pareado) e Wilcoxon (pareado e não pareado), com exemplos práticos e orientações sobre interpretação dos resultados.

> **Importante:** Antes de começar, certifique-se de instalar o pacote `pwr`, `rstatix` e `effsize`:

```
install.packages(c("pwr", "rstatix", "effsize"))
```

## 23.1 Cálculos de tamanho de amostra, poder e efeito com o pacote `pwr`

O pacote `pwr` permite calcular:

- **Tamanho de amostra** necessário para detectar um determinado efeito;
- **Poder estatístico** de um teste para amostras de tamanho fixo;
- **Menor tamanho de efeito** que pode ser detectado para um dado poder e amostra.

### 23.1.1 Exercício 1: Tamanho da amostra para teste t não pareado

Calcule o tamanho da amostra necessário para detectar um efeito moderado (d = 0.5), com poder de 0.8 e nível de significância de 0.05, em um teste t bilateral para duas amostras independentes.

```
library(pwr)
# Tamanho da amostra para teste t não pareado
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

### 23.1.2 Exercício 2: Poder do teste t não pareado para n fixo

Se você dispõe de 26 observações em cada grupo, qual é o poder do teste para detectar um efeito de 0.5, com sig.level = 0.05?

```
# Poder do teste t não pareado para n = 26 por grupo
pwr.t.test(n = 26, d = 0.5, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 26
##               d = 0.5
##       sig.level = 0.05
##           power = 0.4240344
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

---

### 23.1.3 Exercício 3: Menor efeito detectável no teste t não pareado

Com 26 observações por grupo, poder de 0.8 e sig.level de 0.05, qual o menor tamanho de efeito detectável?

```
# Tamanho do efeito detectável
pwr.t.test(n = 26, power = 0.8, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 26
##               d = 0.7923522
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

---

### 23.1.4 Exercício 4: Repita para outros testes

#### 23.1.4.1 a) Teste t pareado

```
# Tamanho da amostra para teste t pareado
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.05, type = "paired")
```

```
##
##      Paired t test power calculation
##
##               n = 33.36713
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

```
# Poder para n = 26 pares
pwr.t.test(n = 26, d = 0.5, sig.level = 0.05, type = "paired")
```

```
##
##      Paired t test power calculation
##
##               n = 26
##               d = 0.5
##       sig.level = 0.05
##           power = 0.6881801
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

```r
# Tamanho do efeito detectável
pwr.t.test(n = 26, power = 0.8, sig.level = 0.05, type = "paired")
```

```
##
##      Paired t test power calculation
##
##               n = 26
##               d = 0.5717051
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

#### 23.1.4.2 b) Wilcoxon não pareado

```r
# Tamanho da amostra para Wilcoxon não pareado (aproximação pelo teste t)
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 63.76561
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

> **Dica:** O pacote `pwr` não possui funções específicas para testes não-paramétricos como Wilcoxon. Por isso, costuma-se usar o cálculo para teste t como aproximação, aumentando o tamanho da amostra em cerca de 15% (multiplique o valor obtido por 1,15), já que testes não-paramétricos geralmente requerem amostras maiores para o mesmo poder estatístico.

```r
# Recalculando o tamanho da amostra
63.76561*1.15
```

```
## [1] 73.33045
```

> O *tamanho do efeito para o teste de Wilcoxon* pode ser aproximado *por d Œ 0,86*, sendo d o tamanho do efeito de Cohen, desde que as distribuições dos grupos sejam aproximadamente normais e com variâncias semelhantes (Lehmann, 2006; Noether, 1987).

```r
# Poder da amostra para Wilcoxon não pareado (aproximação pelo teste t)
pwr.t.test(n = 26, d = 0.5*0.86, sig.level = 0.05, type = "two.sample")
```

```
##
```

```
##      Two-sample t test power calculation
##
##               n = 26
##               d = 0.43
##       sig.level = 0.05
##           power = 0.3304646
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

#### 23.1.4.3 c) Wilcoxon pareado

```r
# Tamanho da amostra para Wilcoxon pareado (aproximação pelo teste t pareado)
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.05, type = "paired")
```

```
##
##      Paired t test power calculation
##
##               n = 33.36713
##               d = 0.5
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number of *pairs*
```

```r
# Poder da amostra para Wilcoxon não pareado (aproximação pelo teste t)
pwr.t.test(n = 26, d = 0.5*0.86, sig.level = 0.05, type = "two.sample")
```

```
##
##      Two-sample t test power calculation
##
##               n = 26
##               d = 0.43
##       sig.level = 0.05
##           power = 0.3304646
##     alternative = two.sided
##
## NOTE: n is number in *each* group
```

> **Atenção:** O **d de Cohen** foi criado para testes paramétricos, como o teste t. Para testes não paramétricos (Wilcoxon), utilize medidas específicas, pois a interpretação do d não se aplica ao contexto de ranks.

---

### 23.1.5 Argumentos principais da função `pwr.t.test()`

A função `pwr.t.test()` possui os seguintes argumentos principais:

- `n`: tamanho da amostra em cada grupo (ou número de pares, para teste pareado)
- `d`: tamanho do efeito (diferença padronizada entre grupos)
- `power`: poder estatístico desejado
- `sig.level`: nível de significância (geralmente 0.05)
- `type`: tipo de teste t: `"two.sample"`, `"paired"` ou `"one.sample"`

> **Dica:** Deixe como `NULL` o parâmetro que você deseja calcular. - Para calcular o tamanho do efeito: `d = NULL` - Para calcular o poder: `power = NULL` - Para calcular o tamanho da amostra: `n = NULL`

**Exemplo para calcular o tamanho do efeito:**

```
pwr.t.test(n = 26, d = NULL, power = 0.8, sig.level = 0.05, type = "two.sample")
```

Neste exemplo, o argumento `d` é o valor a ser calculado.

---

### 23.1.6 Grupos Desbalanceados: usando `pwr.t2n.test()`

Quando os grupos têm tamanhos diferentes, use `pwr.t2n.test()` para calcular o poder.

```
# Exemplo: grupo 1 com 95, grupo 2 com 30 observações
library(pwr)
pwr.t2n.test(n1 = 95, n2 = 30, d = 0.5, sig.level = 0.05)
```

```
##
##      t test power calculation
##
##              n1 = 95
##              n2 = 30
##               d = 0.5
##       sig.level = 0.05
##           power = 0.6586749
##     alternative = two.sided
```

- **n1**: tamanho do primeiro grupo
- **n2**: tamanho do segundo grupo

**Atenção:** Grupos desbalanceados podem comprometer o poder do teste, aumentar a variância e dificultar a interpretação dos resultados. Sempre que possível, busque amostras equilibradas.

---

## 23.2 Tamanho do efeito em testes de Wilcoxon

O tamanho do efeito complementa a análise estatística, indicando a magnitude da diferença entre grupos. Para testes paramétricos (como o t), usa-se o **d de Cohen**. Para testes não paramétricos (Wilcoxon), utilize medidas apropriadas, como **r de Wilcoxon** e **Delta de Cliff**.

### 23.2.1 Por que NÃO usar d de Cohen no Wilcoxon?

- O d de Cohen pressupõe distribuição normal e comparação direta de médias/desvios.
- O Wilcoxon compara **ranks**, não médias.
- Aplicar o d de Cohen em dados de Wilcoxon pode gerar interpretações erradas.

### 23.2.2 Medidas recomendadas para Wilcoxon:

- **r de Wilcoxon:** Calculado com o pacote `rstatix`.
- **Cliff's Delta:** Disponível no pacote `effsize`.

---

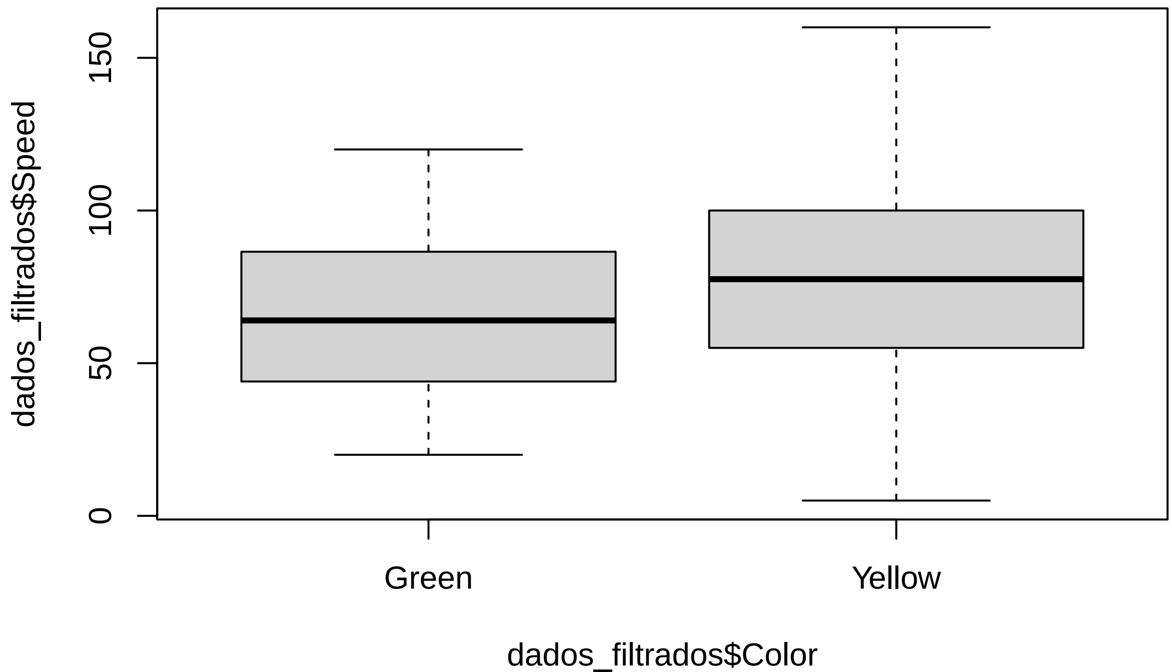
#### 23.2.2.1 r de Wilcoxon (Wilcoxon rank-sum, não pareado)

```
library(rstatix)
library(readr)

# Importe o banco de dados Pokemon
Pokemon <- read_csv("Pokemon.csv")
```

```r
# Subconjunto com Pokémons verdes ou amarelos
dados_filtrados <- subset(Pokemon, Color %in% c("Green", "Yellow"))

# Verifique os dados após o filtro
print(table(dados_filtrados$Color))
```

```
##
##  Green Yellow
##     79     64
```

```r
# Boxplot
boxplot(dados_filtrados$Speed ~ dados_filtrados$Color)
```




```r
# Teste de Wilcoxon não pareado
wilcox.test(dados_filtrados$Speed ~ dados_filtrados$Color)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dados_filtrados$Speed by dados_filtrados$Color
## W = 1920.5, p-value = 0.01363
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Tamanho do efeito r de Wilcoxon para Green vs Yellow
wilcox_effsize(dados_filtrados, Speed ~ Color)
```

```
## # A tibble: 1 x 7
##   .y.   group1 group2 effsize    n1    n2 magnitude
## * <chr> <chr>  <chr>    <dbl> <int> <int> <ord>
## 1 Speed Green  Yellow   0.206    79    64 small
```

```r
levels(dados_filtrados$Color)
```

```
## NULL
```

A tabela apresentada é resultado da função `wilcox_effsize()` do pacote `rstatix` e resume o tamanho do efeito (r de Wilcoxon) para a comparação entre dois grupos ("Green" e "Yellow") em relação à variável

"Speed". Veja como interpretar cada coluna:

- **.y.**: variável de interesse analisada, neste caso, "Speed" (velocidade dos Pokémons).
- **group1**: primeiro grupo comparado ("Green").
- **group2**: segundo grupo comparado ("Yellow").
- **effsize**: valor ABSOLUTO do tamanho de efeito calculado, aqui 0.206.

- O resultado apresentado na coluna `effsize` é sempre positivo, independentemente da ordem dos grupos definidos na variável categórica.
- **n1**: número de observações no grupo 1 (79 Pokémons verdes).
- **n2**: número de observações no grupo 2 (64 Pokémons amarelos).
- **magnitude**: classificação qualitativa do tamanho de efeito ("small", ou seja, efeito pequeno).

**Interpretação:**
A diferença de velocidade entre Pokémons verdes e amarelos é estatisticamente significativa (W = 1920.5, p-value = 0.01363), porém de **pequena magnitude** (r = 0.206). Isso indica que, embora exista uma diferença, ela é discreta do ponto de vista prático.

> No caso de teste pareado, acrescente o argumento `paired = TRUE`:
>
> ```
> wilcox_effsize(dados, antes ~ depois, paired = TRUE)
> ```

---

##### 23.2.2.2 Delta de Cliff

```
library(effsize)

# Calcule o Delta de Cliff
cliff.delta(dados_filtrados$Speed ~ dados_filtrados$Color)
```

```
##
## Cliff's Delta
##
## delta estimate: -0.2403085 (small)
## 95 percent confidence interval:
##       lower       upper
## -0.41405995 -0.04966082
```

O **Delta de Cliff** estimado foi de **-0,24**, classificado como efeito **pequeno** ("small"). O intervalo de confiança de 95% vai de aproximadamente **-0,41** a **-0,05**, indicando que, com alta confiança, o efeito verdadeiro é negativo e pequeno.

- **Sinal negativo:** Indica que o grupo "Yellow" tende a apresentar velocidades maiores do que o grupo "Green". Ou seja, ao comparar aleatoriamente um Pokémon amarelo com um verde, é mais provável que o amarelo tenha uma velocidade superior.
- **Magnitude pequena:** O valor absoluto de -0,24 mostra que a diferença entre os grupos existe, mas é de pouca relevância prática.
- **Intervalo de confiança não inclui zero:** Como o intervalo vai de -0,41 a -0,05, podemos afirmar que existe uma diferença real entre os grupos, embora seja pequena.
- **Interpretação prática:** Apesar de haver uma diferença estatística, ela não é marcante. É importante avaliar se essa diferença tem relevância biológica ou prática no contexto do seu estudo.

O Delta de Cliff indica que Pokémons amarelos tendem a ser um pouco mais rápidos do que os verdes, mas a diferença é pequena do ponto de vista prático.

### 23.2.3 Comparação entre o r de Wilcoxon e o Delta de Cliff

Quando realizamos testes não paramétricos para comparar grupos, como o teste de Wilcoxon (Mann-Whitney ou Wilcoxon pareado), é importante complementar o resultado do teste com uma medida de tamanho de efeito. As duas opções mais comuns são o **r de Wilcoxon** e o **Delta de Cliff**. Veja a seguir uma comparação entre elas:

### 23.2.4 r de Wilcoxon

- **Definição:** Mede a magnitude da diferença entre grupos com base nos postos (ranks) dos dados. O cálculo é semelhante ao coeficiente de correlação de Pearson, mas aplicado a dados não paramétricos.
- **Variação dos valores:** O r de Wilcoxon varia de -1 a 1.
  - Valores próximos de 0 indicam pouca diferença entre grupos.
  - Valores próximos de -1 ou 1 indicam diferenças muito grandes.
  - O sinal indica a direção da diferença.

**Qualificação dos valores**, segundo Cohen (1988):

| Valor absoluto de $r$ | Interpretação |
|---|---|
| aproximadamente 0,10 | Pequeno |
| aproximadamente 0,30 | Moderado |
| maior ou igual 0,50 | Grande |

**Exemplos:**

- **Efeito pequeno:**
  Se o valor absoluto de r estiver próximo de **0,10**, a diferença entre os grupos é pequena.
  *(Exemplo: |r| = 0,10  efeito pequeno)*

- **Efeito moderado:**
  Se o valor absoluto de r estiver próximo de **0,30**, a diferença entre os grupos é moderada.
  *(Exemplo: |r| = 0,30  efeito moderado)*

- **Efeito grande:**
  Se o valor absoluto de r for igual ou superior a **0,50**, a diferença é grande.
  *(Exemplo: |r| = 0,55  efeito grande)*

### 23.2.5 Delta de Cliff ()

- **Definição:** Mede a probabilidade de um valor de um grupo ser maior do que de outro grupo, subtraída da probabilidade contrária. É totalmente não paramétrico e não depende de distribuição, sendo adequado para variáveis ordinais ou dados assimétricos.
- **Variação dos valores:** O Delta de Cliff varia de -1 a 1.
  - Valor 0: não há diferença entre os grupos.
  - Valor positivo: o grupo 1 tende a ter valores maiores que o grupo 2.
  - Valor negativo: o grupo 2 tende a ter valores maiores que o grupo 1.
- **Como interpretar os valores do Delta de Cliff?**
  O **Delta de Cliff** mostra o tamanho da diferença entre dois grupos. O mais comum é considerar o valor absoluto de  (ou seja, ignorar se é positivo ou negativo e olhar só para o tamanho do número).

**Classificação da magnitude do efeito**, segundo Romano et al. (2006)

| Valor absoluto de | Interpretação |
|---|---|
| menor que 0.147 | Desprezível |
| de 0.147 e menor que 0.33 | Pequeno |
| de 0.33 e menor que 0.474 | Médio |
| maior ou igual a 0.474 | Grande |

**Exemplos:**

- **Efeito desprezível:**
  Se o valor absoluto do delta for menor que **0,147**, a diferença entre os grupos é desprezível.
  *(Exemplo: || = 0,10  efeito desprezível)*
- **Efeito pequeno:**
  Se o valor absoluto do delta for igual ou maior que **0,147** e menor que **0,33**, a diferença é pequena.
  *(Exemplo: || = 0,20  efeito pequeno)*
- **Efeito médio:**
  Se o valor absoluto do delta for igual ou maior que **0,33** e menor que **0,474**, a diferença é média.
  *(Exemplo: || = 0,40  efeito médio)*
- **Efeito grande:**
  Se o valor absoluto do delta for igual ou maior que **0,474**, a diferença é grande.
  *(Exemplo: || = 0,50  efeito grande)*

**Resumindo:**
Quanto mais próximo de zero, menor a diferença. Quanto mais próximo de 1 (ou -1), maior a diferença entre os grupos.

> **Dica:** Sempre use o valor absoluto, ou seja, olhe apenas para o tamanho do número, sem se preocupar se ele é positivo ou negativo.

### 23.2.6 Qual é melhor usar?

- **Ambos** são válidos e amplamente aceitos para dados não paramétricos.
- O **r de Wilcoxon** é intuitivo se você já está acostumado com o coeficiente de correlação, e é facilmente interpretável em contextos onde se deseja uma analogia ao r de Pearson.
- O **Delta de Cliff** é mais robusto em situações com muitos empates, diferentes tamanhos de grupo ou dados ordinais, e fornece uma interpretação mais direta da diferença de probabilidades entre grupos.
- **Recomendação prática:**
  - Para estudos com dados ordenados, amostras desbalanceadas ou muitos empates, o Delta de Cliff pode ser preferido.
  - Se você busca uma medida análoga ao r de Pearson para facilitar comparações, use o r de Wilcoxon.
  - Em muitos casos, apresentar ambos os valores enriquece a interpretação dos resultados.

Não existe uma medida “melhor” de forma absoluta; a escolha depende do contexto do estudo e das características dos dados. Ambas são úteis para interpretar a relevância prática das diferenças identificadas em testes não paramétricos.

> **Nota:** Para a comparação entre os grupos *Green* e *Yellow* na variável *Speed*, o tamanho de efeito calculado pelo **r de Wilcoxon** foi **0,206** (efeito pequeno), enquanto o **delta de Cliff** foi **-0,240** (efeito pequeno, IC 95%: -0,414 a -0,050). É esperado que os valores numéricos dessas métricas diferem, pois são baseados em cálculos estatísticos distintos, mas ambos apontam para uma diferença de pequena magnitude entre os grupos.

No entanto, é importante destacar que, neste caso, o valor negativo do delta de Cliff está condizente com o observado no boxplot: o grupo *Yellow* tende a apresentar valores de *Speed* ligeiramente maiores que o grupo *Green*, o que é evidenciado pela direção negativa do delta.

Esse alinhamento entre o resultado numérico do delta de Cliff e a visualização gráfica reforça a importância de sempre analisar os dados de forma complementar, utilizando tanto medidas estatísticas quanto representações visuais (como boxplots). A visualização gráfica pode revelar padrões, tendências e assimetrias que enriquecem a interpretação do tamanho de efeito e facilitam a comunicação dos resultados.

# Chapter 24

# Teste de Normalidade

Antes de aplicar muitos testes estatísticos, é importante verificar se os dados seguem uma distribuição normal (ou "distribuição gaussiana"). Essa etapa é fundamental porque vários testes paramétricos (como o teste t de Student e a ANOVA) assumem a normalidade dos dados. Quando essa condição não é atendida, a escolha do teste estatístico pode mudar para opções não paramétricas.

## 24.1 Hipóteses nos Testes de Normalidade

Todo teste de normalidade avalia as seguintes hipóteses:

- **Hipótese Nula (H):** Os dados seguem uma distribuição normal.
- **Hipótese Alternativa (H):** Os dados **não** seguem uma distribuição normal.
- Se o valor de p for menor que 0,05: **Não rejeitamos** a hipótese nula. Os dados **podem ser tratados como normais**.
- Se o valor de p for maior ou igual 0,05: **Rejeitamos** a hipótese nula. **Os dados não seguem distribuição normal**.

## 24.2 Principais Testes de Normalidade

Abaixo estão os testes de normalidade mais conhecidos, quando são mais recomendados e os pacotes em R:

| Teste | Quando Usar | Pacote no R |
|---|---|---|
| **Shapiro-Wilk** | Amostras pequenas e médias ($n < 50$ até ~ 2000) | `stats` |
| **Kolmogorov-Smirnov (K-S)** | Amostras maiores; pode comparar duas distribuições, mas menos potente para normalidade | `stats` |
| **Lilliefors** | Versão do K-S ajustado para média e desvio amostrais | `nortest` |
| **Anderson-Darling** | Mais sensível nas caudas da distribuição; recomendado para médias e grandes amostras | `nortest` |
| **Jarque-Bera** | Teste baseado em assimetria (skewness) e curtose (kurtosis); comum em econometria | `tseries` |
| **D'Agostino-Pearson** | Avalia assimetria e curtose em conjunto; recomendado para médias e grandes amostras | `moments` |

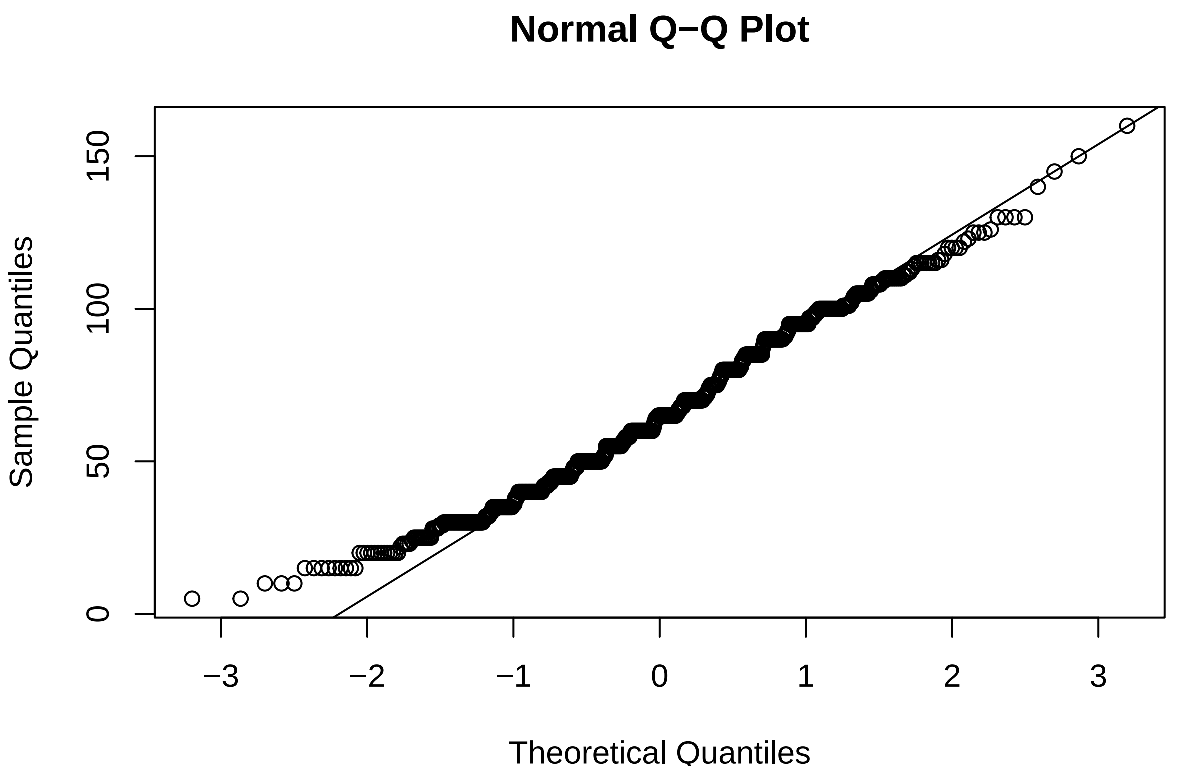
### 24.2.1 Dicas de Uso

- Para **amostras pequenas** ($n < 50$): prefira o **Shapiro-Wilk** (função `shapiro.test()` do pacote base `stats`).

- Para **amostras médias a grandes**: considere **Anderson-Darling** (`ad.test()` do pacote `nortest`) ou **D'Agostino-Pearson** (`agostino.test()` do pacote `moments`).
- O **Kolmogorov-Smirnov** (`ks.test()`) é mais geral, mas menos recomendado quando a média e desvio padrão são estimados dos próprios dados.
- O **Lilliefors** (`lillie.test()` do pacote `nortest`) é uma boa alternativa ao K-S para dados amostrais.
- Para análises em econometria ou séries temporais, o **Jarque-Bera** (`jarque.bera.test()` do pacote `tseries`) é bastante utilizado.

### 24.2.2 Exemplo Prático em R

```r
#install.packages("nortest")
#install.packages("moments")

# Importe o banco de dados Pokemon
library(readr)
Pokemon <- read_csv("Pokemon.csv")

# Veja o QQ
qqnorm(Pokemon$Speed)
qqline(Pokemon$Speed)
```




```r
# Shapiro-Wilk
shapiro.test(Pokemon$Speed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Pokemon$Speed
## W = 0.98574, p-value = 1.757e-06
```

```r
# Anderson-Darling
library(nortest)
ad.test(Pokemon$Speed)
```

```
##
##  Anderson-Darling normality test
##
## data:  Pokemon$Speed
## A = 3.0665, p-value = 1.064e-07
```

```r
# Lilliefors
lillie.test(Pokemon$Speed)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Pokemon$Speed
## D = 0.06446, p-value = 1.662e-07
```

```r
# D'Agostino
library(moments)
agostino.test(Pokemon$Speed)
```

```
##
##  D'Agostino skewness test
##
## data:  Pokemon$Speed
## skew = 0.27799, z = 3.02335, p-value = 0.0025
## alternative hypothesis: data have a skewness
```

**Resumo**

- Sempre comece analisando a normalidade dos seus dados.
- Escolha o teste de acordo com o tamanho da amostra e o objetivo da análise.
- **Use gráficos (histograma, boxplot e principalmente o QQ) junto com testes estatísticos para uma análise mais completa.**

> **Importante:** Testes de normalidade são sensíveis ao tamanho da amostra. Em amostras muito grandes, pequenas desvios podem resultar em p-valores baixos; em amostras pequenas, a potência dos testes é baixa.

## 24.3 Assimetria e Curtose: O que são e por que são importantes?

Quando analisamos dados, é importante entender não só a média e o desvio padrão, mas também o **formato da distribuição** dos valores. Duas medidas ajudam muito nisso: **assimetria** e **curtose**. Vamos entender cada uma de forma simples:

---

### 24.3.1 O que é Assimetria (Skewness)?

A **assimetria** mostra se os dados estão “puxados” para algum lado, ou seja, se a distribuição tem uma cauda maior para a direita ou para a esquerda.

- **Assimetria zero:** Os dados estão distribuídos de forma equilibrada ao redor da média (exemplo: distribuição normal perfeita).
- **Assimetria positiva (à direita):** A cauda é mais longa à direita, ou seja, há mais valores extremos acima da média.
- **Assimetria negativa (à esquerda):** A cauda é mais longa à esquerda, há mais valores extremos abaixo da média.

**Por que isso importa?**
A distribuição Normal é um exemplo de distribuição simétrica.

### 24.3.2 O que é Curtose (Kurtosis)?

A **curtose** indica o quanto a distribuição é "pontuda" ou "achatada" em relação à Distribuição Normal.

- **Curtose alta (leptocúrtica):** Distribuição muito "pontuda", com muitas observações próximas da média, mas também mais valores extremos (outliers).
- **Curtose baixa (platicúrtica):** Distribuição mais achatada, com menos valores extremos.
- **Curtose média (mesocúrtica):** Parecida com a distribuição normal.

**Por que isso importa?**
A presença de muita curtose (valores muito altos ou baixos) pode indicar que há mais outliers do que o esperado, o que também pode afetar os resultados de testes estatísticos.

- **Assimetria:** mostra se a distribuição "pende" para um lado.
- **Curtose:** mostra se a distribuição é "pontuda" e cheia de extremos, ou mais "achatada".

Essas duas medidas ajudam a entender melhor o comportamento dos seus dados e garantem que você escolha os testes estatísticos mais adequados para a sua análise!

# Chapter 25

# Comparação de mais de dois grupos

Assim como na comparação entre dois grupos, ao desejarmos comparar mais de dois grupos em uma análise estatística, é fundamental selecionar o teste mais adequado de acordo com as características dos dados. Os testes podem ser **paramétricos** ou **não paramétricos**, e a escolha correta depende da verificação prévia de alguns pressupostos, como normalidade, homogeneidade de variâncias e independência das observações.

## 25.1 Revisando as hipóteses da comparação entre dois grupos

- **Paramétrico (Teste t)**
  - **Não pareado (t de Student):**
    - H: As médias dos dois grupos são iguais ( = )
    - H: As médias dos dois grupos são diferentes (  )
  - **Pareado (t pareado):**
    - H: A média das diferenças entre os pares é igual a zero (_d = 0)
    - H: A média das diferenças entre os pares é diferente de zero (_d  0)
- **Não paramétrico**
  - **Não pareado (Mann-Whitney/Wilcoxon rank-sum):**
    - H: As distribuições dos dois grupos são iguais
    - H: As distribuições dos dois grupos são diferentes
  - **Pareado (Wilcoxon signed-rank):**
    - H: A distribuição das diferenças entre os pares é simétrica em torno de zero
    - H: A distribuição das diferenças entre os pares não é simétrica em torno de zero

Na comparação entre **dois grupos**, tanto nos testes paramétricos quanto nos não paramétricos, a hipótese nula (H) geralmente afirma que as médias (ou distribuições) dos dois grupos são iguais, enquanto a hipótese alternativa (H) aponta que elas são diferentes. Ou seja, o teste busca identificar uma diferença específica entre os dois grupos analisados.

## 25.2 Hipóteses para a comparação entre mais de dois grupos

- **Paramétrico (ANOVA)**
  - **Não pareado (ANOVA one-way):**
    - H: As médias dos grupos são todas iguais ( = = ... = _k)
    - H: Pelo menos uma média de grupo é diferente das outras
  - **Pareado (ANOVA de medidas repetidas):**
    - H: As médias dos tratamentos (ou condições) são iguais
    - H: Pelo menos uma média de tratamento é diferente das outras
- **Não paramétrico**
  - **Não pareado (Kruskal-Wallis):**
    - H: As distribuições dos grupos são todas iguais
    - H: Pelo menos uma distribuição de grupo é diferente das outras

- **Pareado (Friedman):**
  - H: As distribuições dos tratamentos (ou condições) são todas iguais
  - H: Pelo menos uma distribuição de tratamento é diferente das outras

Na comparação entre **mais de dois grupos** (por exemplo, usando ANOVA ou Kruskal-Wallis), a hipótese nula (H) é que **todas** as médias (ou distribuições) dos grupos são iguais. Por outro lado, a hipótese alternativa (H) não especifica qual grupo é diferente, mas sim que **pelo menos um dos grupos se difere dos demais**. Ou seja, a rejeição da hipótese nula indica que existe pelo menos uma diferença, mas não revela imediatamente entre quais grupos essa diferença ocorre.

> **Atenção:** É importante notar que, no contexto de mais de dois grupos, a hipótese alternativa não identifica quais grupos são diferentes, apenas aponta que existe pelo menos uma diferença.

## 25.3 Testes Paramétricos

- **Não Pareados:** O teste mais utilizado é a **ANOVA de uma via**, que compara as médias de três ou mais grupos independentes.
- **Pareados:** Utiliza-se a **ANOVA para medidas repetidas** quando as medições são feitas nos mesmos indivíduos em diferentes condições ou tempos.

### 25.3.1 Pré-requisitos para ANOVA (dados pareados ou não)

1. **Normalidade dos resíduos:** Os resíduos do modelo devem seguir uma distribuição normal.
2. **Homogeneidade de variâncias:** As variâncias dos grupos devem ser semelhantes.
3. **Esfericidade** (apenas para medidas repetidas): A variância das diferenças entre todas as combinações de pares de grupos deve ser semelhante.
4. **Independência das observações:** Para grupos independentes.

Se algum desses pré-requisitos não for atendido, é recomendado utilizar testes não paramétricos.

## 25.4 Testes Não Paramétricos

- **Não Pareados:** O teste de **Kruskal-Wallis** é utilizado para comparar mais de dois grupos independentes.
- **Pareados:** O teste de **Friedman** é usado para comparar três ou mais grupos pareados.

## 25.5 ANOVA de uma via no R

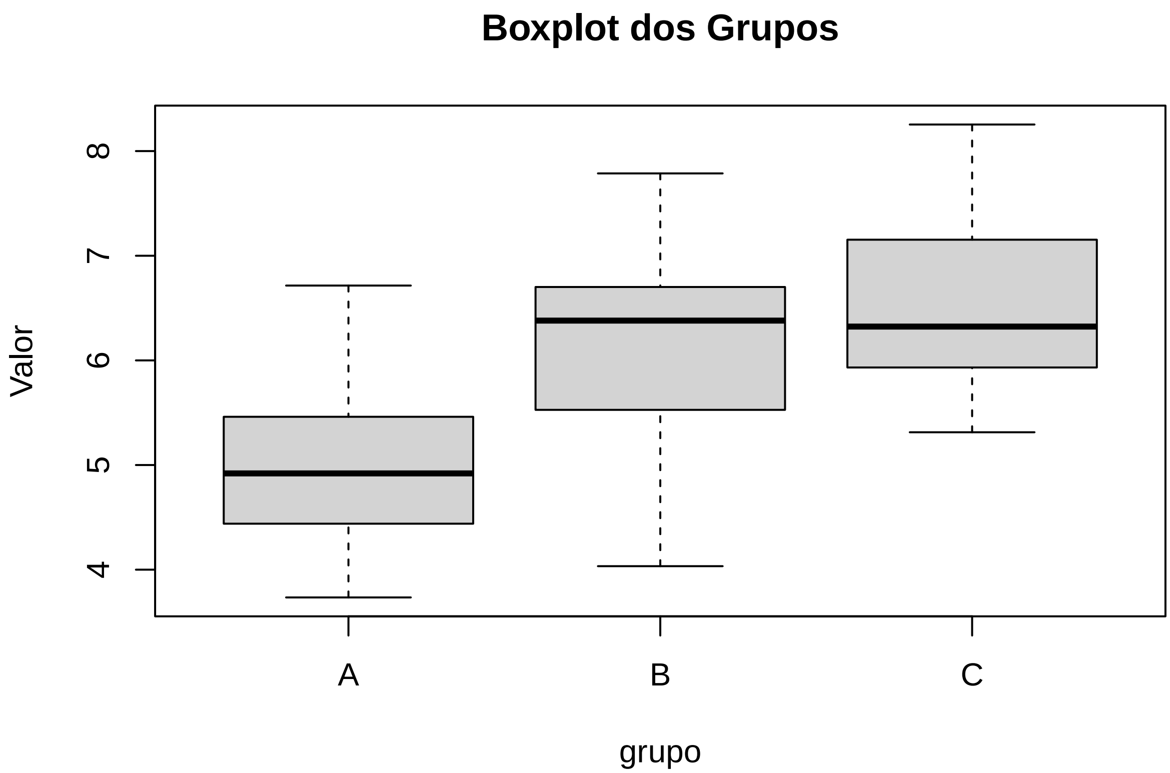
A seguir, apresentamos um exemplo de ANOVA de uma via e do teste de Kruskal-Wallis, incluindo a verificação dos pré-requisitos.

### 25.5.1 Simulação de Dados

```
set.seed(123)
grupo <- factor(rep(c("A", "B", "C"), each = 10))
valor <- c(rnorm(10, mean = 5, sd = 1),
           rnorm(10, mean = 6, sd = 1),
           rnorm(10, mean = 7, sd = 1))
dados <- data.frame(grupo, valor)
```

### 25.5.2 Visualização dos Dados

```
boxplot(valor ~ grupo, data = dados, main = "Boxplot dos Grupos", ylab = "Valor")
```

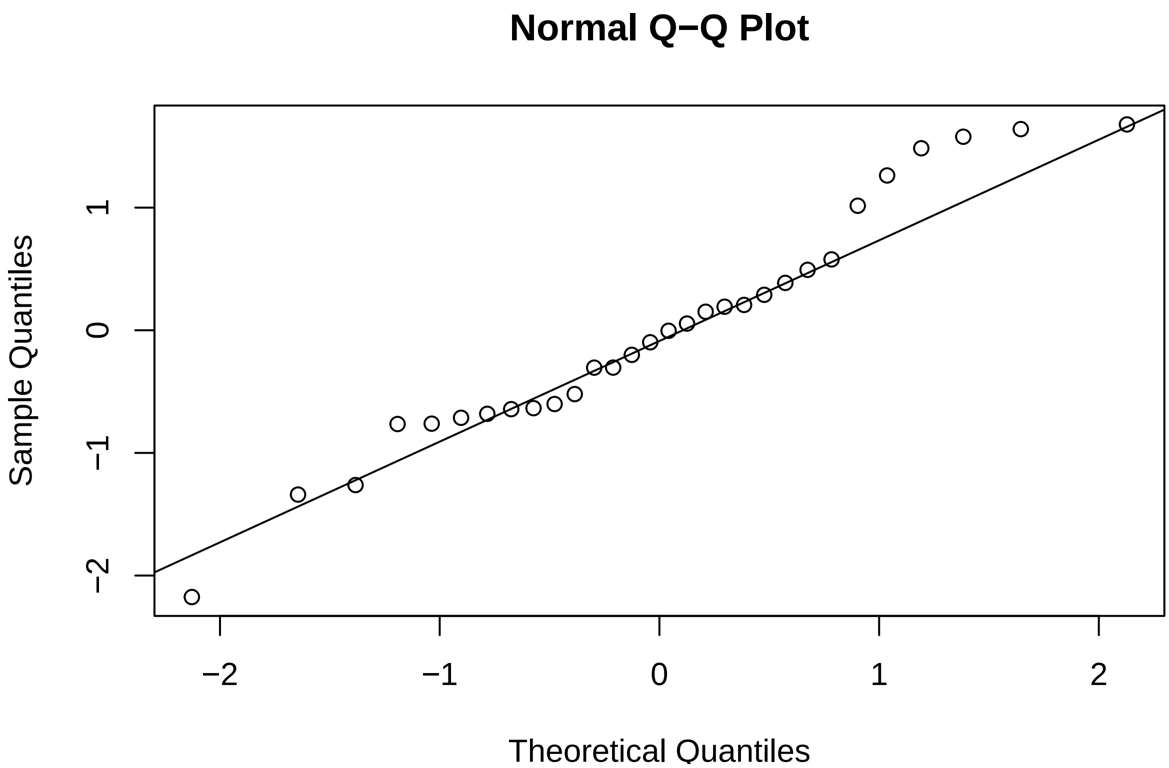



### 25.5.3 Verificação dos Pré-requisitos para ANOVA

```
modelo_aov <- aov(valor ~ grupo, data = dados)
residuos <- residuals(modelo_aov)

shapiro.test(residuos)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuos
## W = 0.96213, p-value = 0.3507
```

```
qqnorm(residuos)
qqline(residuos)
```




- A função `aov` no R é utilizada para realizar **análise de variância (ANOVA)**.
- **modelo_aov <- aov(valor ~ grupo, data = dados)**
  Esta linha ajusta um modelo de ANOVA de uma via, avaliando se a média da variável `valor` difere entre os diferentes níveis do fator `grupo`, usando os dados do data frame `dados`.
- **residuos <- residuals(modelo_aov)**
  Esta linha extrai os resíduos do modelo ajustado, ou seja, as diferenças entre os valores observados e os valores previstos pelo modelo. A análise dos resíduos é fundamental para verificar os pressupostos da ANOVA, como a normalidade.

  Neste exemplo, tanto o teste de normalidade quanto a inspeção visual do gráfico QQ sugerem que os resíduos seguem uma distribuição normal.

#### 25.5.3.2 Homogeneidade de variâncias

```
# Instale 'car' se necessário: install.packages("car")
library(car)
leveneTest(valor ~ grupo, data = dados)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2   5e-04 0.9995
##       27
```

- O *teste de Levene* avalia a homogeneidade das variâncias entre os grupos, que é um dos pré-requisitos para a ANOVA.
- **Valor de p (Pr(>F)) = 0.9995:** O valor de p é muito maior que 0,05, indicando que **não há evidências para rejeitar a hipótese nula de igualdade das variâncias**.

  Nesse exemplo, as variâncias dos grupos podem ser consideradas homogêneas. Portanto, o pré-requisito de homogeneidade de variâncias para a ANOVA foi atendido.

### 25.5.4 ANOVA de Uma Via

```r
summary(modelo_aov)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## grupo        2  12.24   6.122   6.435 0.00518 **
## Residuals   27  25.68   0.951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- **Valor de p (Pr(>F)) = 0.00518:** O valor de p é menor que 0,05, indicando que existem diferenças estatisticamente significativas entre as médias dos grupos analisados.
- Rejeita-se a hipótese nula de igualdade das médias. Isso significa que pelo menos um dos grupos difere significativamente dos demais. Recomenda-se realizar *testes post hoc (por exemplo, Tukey)* para identificar quais grupos apresentam diferenças entre si.

#### 25.5.4.1 Teste Post Hoc (se ANOVA for significativa)

```r
TukeyHSD(modelo_aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = valor ~ grupo, data = dados)
##
## $grupo
##          diff         lwr      upr     p adj
## B-A 1.1339963  0.05254072 2.215452 0.0384525
## C-A 1.5008155  0.41935988 2.582271 0.0052283
## C-B 0.3668192 -0.71463643 1.448275 0.6812485
```

O teste de Tukey compara as médias dos grupos dois a dois, após a ANOVA indicar diferença significativa entre eles. Os resultados mostram as diferenças entre as médias dos grupos (`diff`), os limites inferior (`lwr`) e superior (`upr`) do intervalo de confiança de 95%, e o valor de p ajustado (`p adj`).

**Resultados:**

- **B vs A:**
  Diferença = 1.13; IC 95% = [0.05, 2.22]; p = 0.038
  O grupo B tem média significativamente maior que o grupo A.
- **C vs A:**
  Diferença = 1.50; IC 95% = [0.42, 2.58]; p = 0.005
  O grupo C tem média significativamente maior que o grupo A.
- **C vs B:**
  Diferença = 0.37; IC 95% = [-0.71, 1.45]; p = 0.681
  Não há diferença significativa entre os grupos C e B.

  Os grupos B e C apresentam médias significativamente maiores do que o grupo A. Não foi observada diferença significativa entre os grupos B e C.

## 25.6 Kruskal-Wallis no R

Se os pré-requisitos da ANOVA não forem atendidos, utilize o Kruskal-Wallis:

```r
kruskal.test(valor ~ grupo, data = dados)
```

```
##
##  Kruskal-Wallis rank sum test
```

```
##
## data:  valor by grupo
## Kruskal-Wallis chi-squared = 8.1858, df = 2, p-value = 0.01669
```

- **Valor de p (p-value) = 0.01669:** O valor de p é menor que 0,05, indicando que há diferença estatisticamente significativa entre pelo menos dois dos grupos analisados.

- Rejeita-se a hipótese nula de que as distribuições dos grupos são todas iguais. Ou seja, pelo menos um dos grupos apresenta distribuição diferente dos demais. Para identificar especificamente quais grupos diferem entre si, é recomendada a realização de testes post hoc apropriados (por exemplo, Dunn ou pairwise Wilcoxon com ajuste para múltiplas comparações).

### 25.6.1 Teste Post Hoc para Kruskal-Wallis

```r
# Instale PMCMRplus se necessário: install.packages("PMCMRplus")
library(PMCMRplus)
kwAllPairsDunnTest(valor ~ grupo, data = dados)
```

```
##   A     B
## B 0.058 -
## C 0.021 0.611
```

O teste de Dunn é um teste pós-hoc não-paramétrico, utilizado após o teste de Kruskal-Wallis para identificar quais pares de grupos diferem significativamente entre si.

Os *p-valores* apresentados pelo teste compara todos os pares possíveis entre os grupos (neste exemplo: A, B e C):

- **A vs. B**: p = 0.058
- **A vs. C**: p = 0.021
- **B vs. C**: p = 0.611

  Portanto, os resultados sugerem que apenas os grupos A e C, para a variável analisada, são estatisticamente diferentes entre si.

## 25.7 Por que não é indicado comparar os grupos dois a dois diretamente?

Quando se tem mais de dois grupos, pode parecer tentador realizar múltiplos testes de comparação entre pares de grupos (por exemplo, vários testes t para todos os pares possíveis). No entanto, esse procedimento **não é recomendado**, pois aumenta significativamente o risco de cometer um **erro do tipo I** (falso positivo).

Cada teste realizado tem uma determinada probabilidade de indicar uma diferença por acaso (erro tipo I), geralmente 5% se  = 0,05. Ao fazer muitos testes independentes, a probabilidade de encontrar pelo menos um resultado "significativo" apenas por acaso aumenta. Esse fenômeno é chamado de **inflacionamento da taxa de erro tipo I**.

Por isso, para comparação de mais de dois grupos, utiliza-se primeiro um teste global (como ANOVA ou Kruskal-Wallis). Se o resultado for significativo, aí sim são realizados testes post hoc, que já incluem correções para múltiplas comparações (como o teste de Tukey), controlando o risco de erro tipo I.

## 25.8 ANOVA para Dados Repetidos no R

Vamos analisar o desempenho de 6 alunos em 3 provas diferentes (Prova 1, Prova 2 e Prova 3). Cada aluno fez todas as provas, ou seja, temos medidas repetidas!

### 25.8.1 Gerar dados simulados

```
set.seed(123)
n <- 20
dados <- data.frame(
  aluno = factor(1:n),
  prova1 = round(rnorm(n, mean = 7, sd = 1), 1),
  prova2 = round(rnorm(n, mean = 7.5, sd = 1), 1),
  prova3 = round(rnorm(n, mean = 8, sd = 1), 1)
)
head(dados)
```

```
##   aluno prova1 prova2 prova3
## 1     1    6.4    6.4    7.3
## 2     2    6.8    7.3    7.8
## 3     3    8.6    6.5    6.7
## 4     4    7.1    6.8   10.2
## 5     5    7.1    6.9    9.2
## 6     6    8.7    5.8    6.9
```

### 25.8.2 Converter para formato longo

```
library(tidyr)
dados_long <- pivot_longer(
  dados,
  cols = c("prova1", "prova2", "prova3"),
  names_to = "prova",
  values_to = "nota"
)
head(dados_long)
```

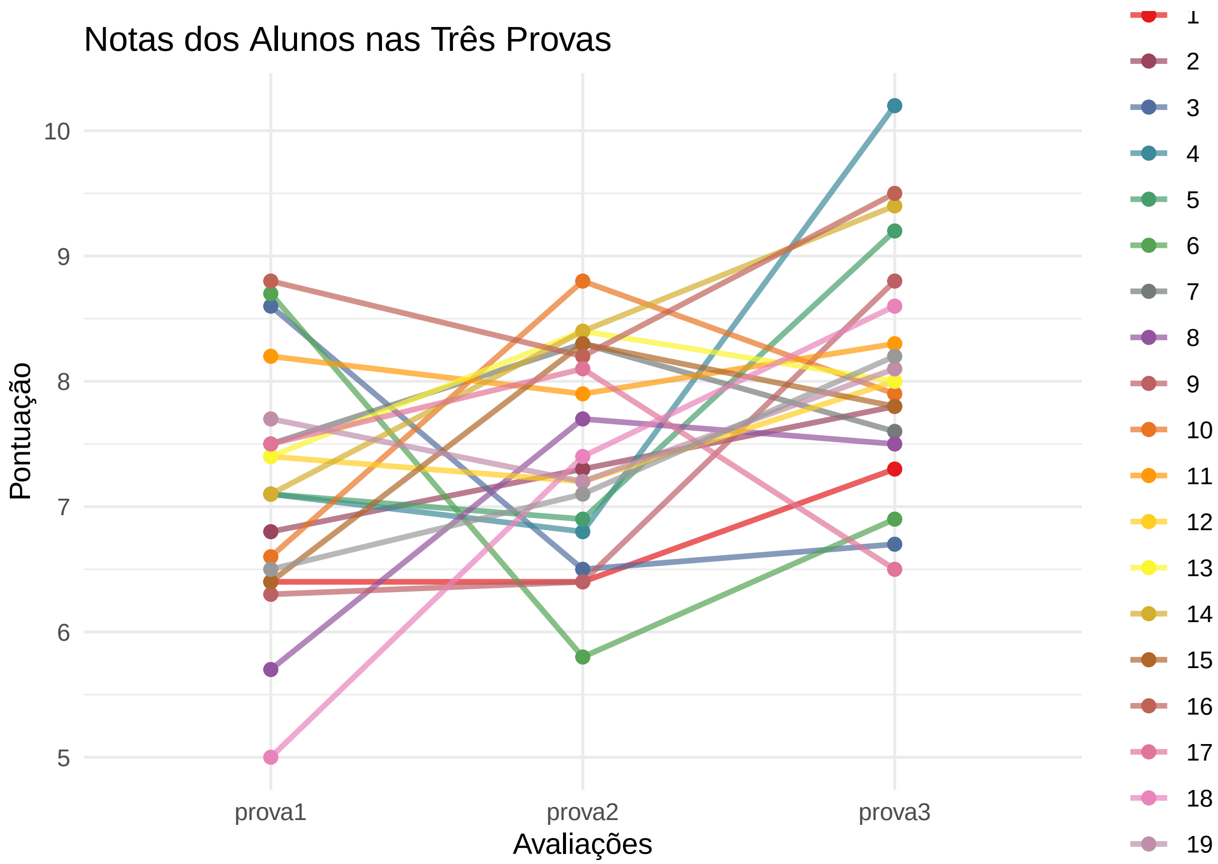
```
## # A tibble: 6 x 3
##   aluno prova   nota
##   <fct> <chr>  <dbl>
## 1 1     prova1   6.4
## 2 1     prova2   6.4
## 3 1     prova3   7.3
## 4 2     prova1   6.8
## 5 2     prova2   7.3
## 6 2     prova3   7.8
```

Transformar os dados para o formato longo é fundamental para análises de medidas repetidas porque permite que o R identifique corretamente as repetições de cada aluno ao longo das diferentes provas. Sem isso, não é possível fazer ANOVA de medidas repetidas de forma adequada!

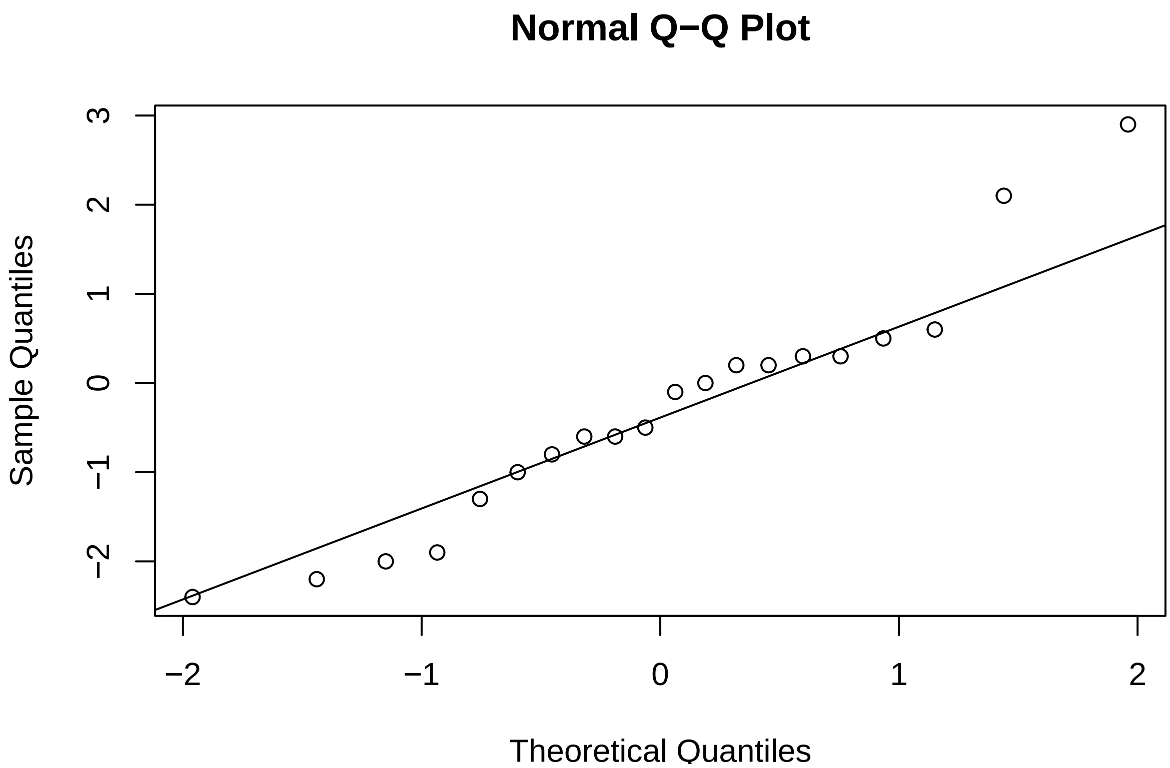
### 25.8.3 Visualização (opcional)

```
library(ggplot2)
library(RColorBrewer)

# Use uma paleta de cores apropriada para até 20 alunos
cores <- colorRampPalette(brewer.pal(9, "Set1"))(20)

ggplot(dados_long, aes(x = prova, y = nota, group = aluno, color = aluno)) +
  geom_line(alpha = 0.7, size = 1) +
  geom_point(size = 2) +
  labs(x = "Avaliações", y = "Pontuação") +
```

```
  scale_color_manual(values = cores) +
  labs(title = "Notas dos Alunos nas Três Provas", color = "Aluno") +
  theme_minimal() +
  theme(legend.position = "right")
```



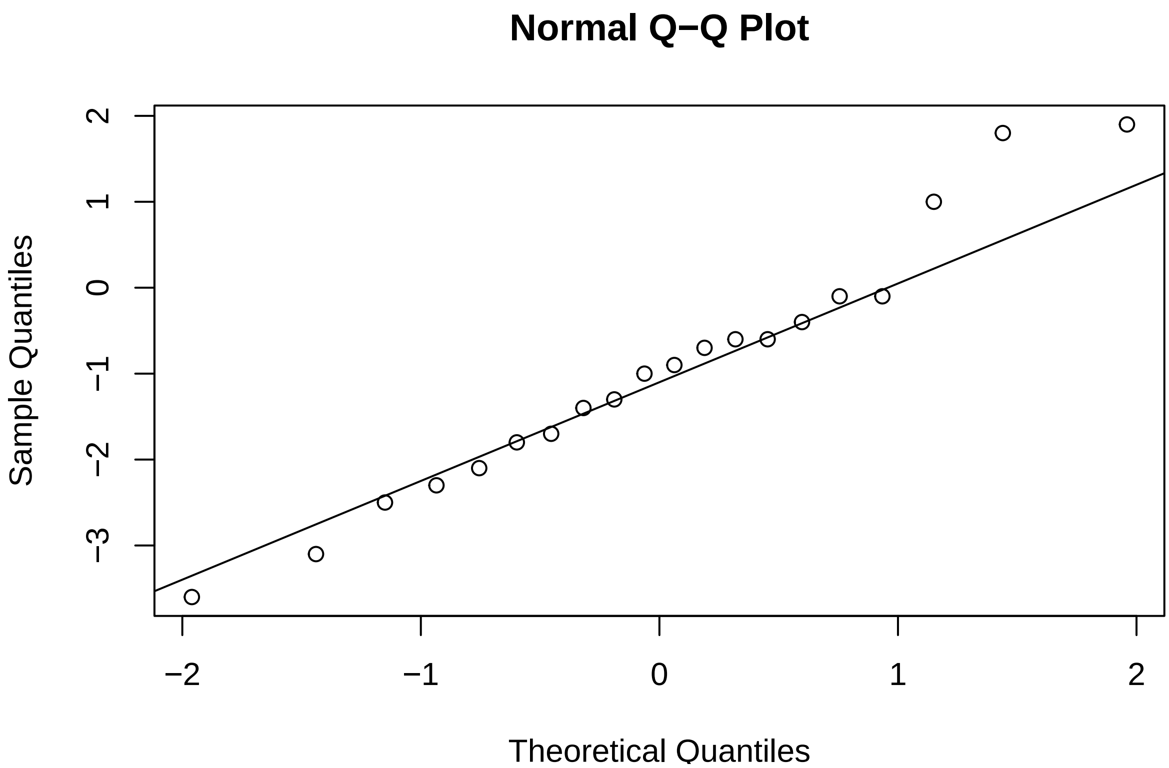

```
dif12 <- dados$prova1 - dados$prova2
dif13 <- dados$prova1 - dados$prova3
dif23 <- dados$prova2 - dados$prova3

qqnorm(dif12)
qqline(dif12)
```




```r
shapiro.test(dif12)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dif12
## W = 0.94596, p-value = 0.3099
```

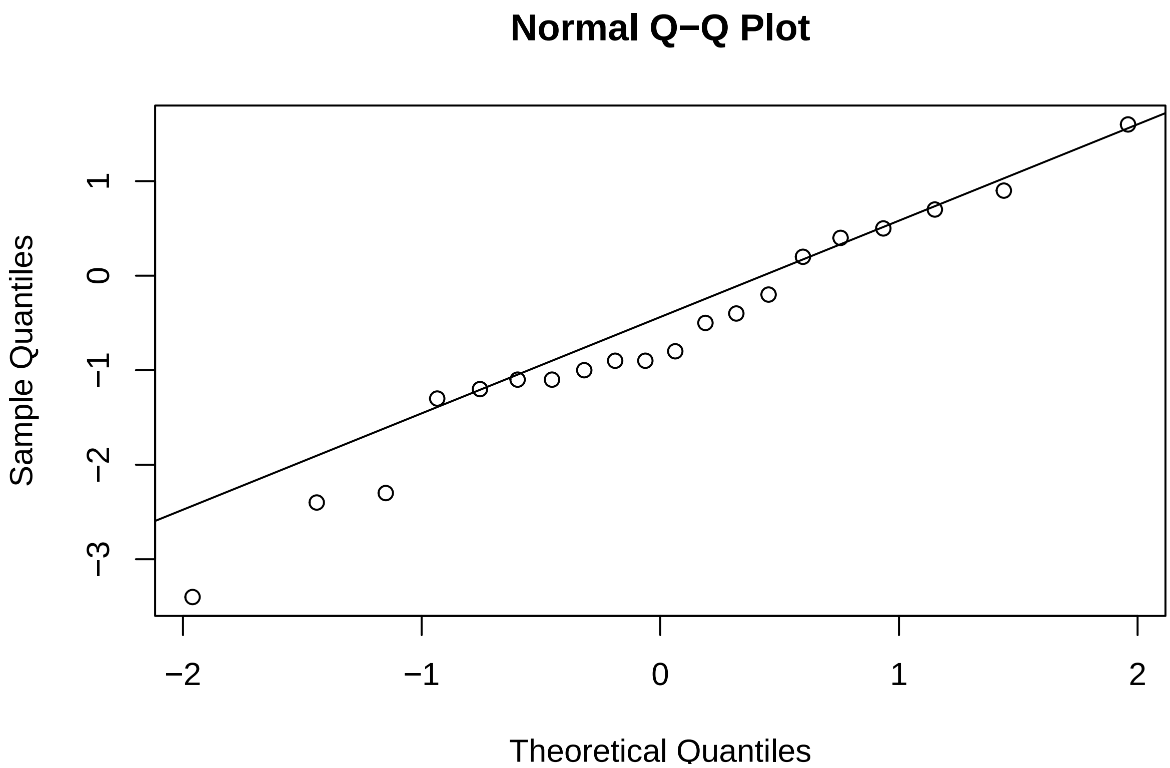
```r
qqnorm(dif13)
qqline(dif13)
```




```r
shapiro.test(dif13)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dif13
## W = 0.96929, p-value = 0.7398
```

```r
qqnorm(dif23)
qqline(dif23)
```




```r
shapiro.test(dif23)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  dif23
## W = 0.97175, p-value = 0.7914
```

As diferenças seguem distribuição normal.

### 25.8.5 ANOVA de medidas repetidas com afex (testa e corrige esfericidade)

A função `aov_ez()`, do pacote `afex` realiza toda a ANOVA de medidas repetidas de forma prática e automática.

```r
# Instale se necessário: install.packages("afex")
library(afex)
modelo_afex <- aov_ez(
  id = "aluno",
  dv = "nota",
  within = "prova",
  data = dados_long
)
modelo_afex
```

```
## Anova Table (Type 3 tests)
##
## Response: nota
##   Effect          df  MSE       F  ges p.value
## 1  prova 1.91, 36.27 0.94 5.53 ** .168    .009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1
##
## Sphericity correction method: GG
```

- **Effect**: O fator analisado (`prova`), ou seja, se as notas variam entre as provas.
- **df**: Graus de liberdade. Note que aparecem valores decimais (1.91, 36.27) porque foi aplicada a correção de Greenhouse-Geisser (GG), devido à violação da esfericidade.
- **MSE**: Erro quadrático médio.
- **F**: Estatística F da ANOVA.
- **ges**: Generalized Eta Squared, uma medida de tamanho de efeito.
- **p.value**: Valor de p associado ao teste F. Neste caso, $p = 0.009$ indica que há diferença significativa entre as médias das provas ($p < 0.05$).
- **Observação sobre esfericidade:** A linha `Sphericity correction method: GG` informa que o pressuposto de esfericidade foi violado (teste de Mauchly $p < 0.05$) e, por isso, os graus de liberdade e o valor de p foram corrigidos automaticamente pelo método de Greenhouse-Geisser.

  A análise de variância (ANOVA) mostrou que o fator "prova" teve um efeito significativo sobre as notas ($F(1.91, 36.27) = 5.53$, $p = 0.009$, ges $= 0.168$), indicando que as médias das notas diferem entre as provas.

### 25.8.6 Pós-hoc: Comparações múltiplas entre as provas

```
# Instale se necessário: install.packages("emmeans")
library(emmeans)
emm <- emmeans(modelo_afex, ~ prova)
pairs(emm, adjust = "bonferroni") # ou "holm", "sidak", etc.
```

```
##  contrast        estimate    SE df t.ratio p.value
##  prova1 - prova2   -0.315 0.300 19  -1.050  0.9212
##  prova1 - prova3   -0.975 0.326 19  -2.993  0.0225
##  prova2 - prova3   -0.660 0.269 19  -2.455  0.0717
##
## P value adjustment: bonferroni method for 3 tests
```

- **prova1 vs prova2:** Não há diferença significativa ($p = 0.9212$).
- **prova1 vs prova3:** Diferença significativa ($p = 0.0225$). As médias dessas provas são estatisticamente diferentes.
- **prova2 vs prova3:** Não há diferença significativa ($p = 0.0717$).

  A análise revelou que existe diferença significativa nas notas entre as provas. Especificamente, a única diferença significativa foi entre as provas 1 e 3, indicando que as médias dessas avaliações diferem estatisticamente. Entre as demais provas, não foram observadas diferenças significativas após o ajuste de Bonferroni.

## 25.9 Teste de Friedman

```
# O teste de Friedman compara as três provas considerando a repetição por aluno
friedman.test(as.matrix(dados[, c("prova1", "prova2", "prova3")]))
```

```
##
##  Friedman rank sum test
##
## data:  as.matrix(dados[, c("prova1", "prova2", "prova3")])
## Friedman chi-squared = 10.101, df = 2, p-value = 0.006405
```

### 25.9.1 Pós-teste: Comparações Múltiplas (Wilcoxon pareado)

```
pairwise.wilcox.test(
  dados_long$nota,
  dados_long$prova,
  paired = TRUE,
```

```r
  p.adjust.method = "bonferroni"
)
```

```
##
##  Pairwise comparisons using Wilcoxon signed rank test with continuity correction
##
## data:  dados_long$nota and dados_long$prova
##
##        prova1 prova2
## prova2 0.729  -
## prova3 0.033  0.087
##
## P value adjustment method: bonferroni
```

Após o teste de Friedman indicar que há diferença nas notas das provas, fazemos um pós-teste para descobrir entre quais provas há diferença.

- Apenas **entre Prova 1 e Prova 3 existe diferença significativa** ($0.033 < 0.05$).
- Entre Prova 1 e 2, e entre Prova 2 e 3, não há diferença significativa (valores maiores que 0.05).


- **Cohen, J. (1988).** Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Lawrence Erlbaum Associates.
- **Field, A. (2013).** Discovering Statistics Using IBM SPSS Statistics (4th ed.). Sage.
- **Romano, J., Kromrey, J. D., Coraggio, J., & Skowronek, J. (2006).** Appropriate statistics for ordinal level data: Should we really be using t-test and Cohen's d for evaluating group differences on the NSSE and other surveys? *Annual Meeting of the Florida Association of Institutional Research*, Jacksonville, FL, USA.
- **Cliff, N. (1993).** Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin, 114*(3), 494-509. https://doi.org/10.1037/0033-2909.114.3.494