

Estatística + R

Ana Paula Fernandes (DESCO/UFTM)

Atualizado em: 18/04/2025

Sumário

1	Bem-vindos!	5
2	Introdução	7
2.1	Atividade 1	7
3	Definições Iniciais	9
3.1	Estatística Descritiva	9
3.2	Estatística Inferencial	10
3.3	Conexão com a Probabilidade	11
3.4	População x Amostra	11
4	Tamanho da Amostra	13
4.1	Passos básicos para o cálculo do tamanho da amostra	13
4.2	Exemplo de cálculo simples	13
4.3	Atividade 2	14
5	Ambiente computacional	17
5.1	Plano A: Instalação R e RStudio	18
5.2	Plano B: R e RStudio online	18
6	Trabalhando no RStudio	21
7	Primeiros exercícios no R	25
7.1	Exemplo 1	25
7.2	Exemplo 2	25
7.3	Exemplo 3	26
8	Tipos de variáveis	29
8.1	Atividade 3	29
9	Estatística descritiva	31
9.1	Medidas de tendência central (ou posição)	31
9.2	Medidas de dispersão (ou variabilidade)	31
9.3	Medida relativa de variabilidade	31
9.4	Funções do R	32
9.5	Atividade 4	32
10	Importando banco de dados	33
10.1	Importando um banco csv	33
10.2	Importando um banco xls	35
10.3	Exemplo 1	35
10.4	Exemplo 2	36
10.5	Exemplo 3	36

11 Instalando pacotes	37
11.1 Exemplo 1	38
11.2 Exemplo 2	38
11.3 Exemplo 3	38
11.4 Exemplo 4	38
11.5 Atividade 5	39
12 Gráficos	41
12.1 Histograma	42
12.2 Boxplot	44
12.3 Atividade 6	46
13 Distribuição de Probabilidade!	47
13.1 Distribuição Normal	47
13.2 Diagnóstico de Normalidade (QQ)	51
13.3 Atividade 7	54

Capítulo 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.

Capítulo 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 teóricos, lidar com recursos visuais (gráficos) e, de fato, transformar o conteúdo 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 cursos das exatas. Assim, a primeira tarefa é sempre desconstruir essa ideia.

A estatística é **MULTIDISCIPLINAR**, 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 algoritmos das redes sociais? Ou das sugestões de filmes e músicas que aparecem no seu *streaming* favorito? Ou no ranque de busca realizada 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 epidemiológicos, 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 escrever 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 Psicologia

- 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>

Capítulo 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**, 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.

Capítulo 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

Para calcular o tamanho da amostra, precisamos de algumas informações:

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:

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

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%)

4.2 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:

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

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

4.2.1 Ferramentas para cálculo de amostra

Hoje, você não precisa se preocupar em fazer esses cálculos manualmente. Existem várias ferramentas online que facilitam esse processo. Por exemplo:

- 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, com base em diferentes tipos de testes estatísticos. Esse 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
- **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

Além disso, ferramentas de **Inteligência Artificial**, como o **Chat GPT**, podem ser utilizadas para automatizar esses cálculos ou até mesmo explicar os conceitos de forma mais simples. Com a IA, você pode rapidamente encontrar a fórmula correta ou usar calculadoras online sem precisar entender todos os detalhes matemáticos.

4.2.2 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.

4.3 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>

Capítulo 5

Ambiente computacional

Existem vários softwares que são dedicados a análise estatística que vão do maravilhoso SPSS (da IMB) às planilhas eletrônicas (como o Excel). Para citar alguns algumas dessas ferramentas:

Softwares pagos

- SPSS <https://www.ibm.com/br-pt/spss>
- Stata <https://www.stata-brasil.com/software/stata.html>
- SAS https://www.sas.com/pt_br
- JMP <https://www.jmp.com/>
- Prisma <https://software.com.br/p/prism>
- Minitab <https://osbsoftware.com.br/produto/minitab-statistical-software>
- Excel (Microsoft)

Softwares livre

- Jamovi <https://www.jamovi.org>
- OpenStat <https://openstat.info>

Linguagens computacionais

- R <https://www.r-project.org>
- Python <https://www.python.org/>

Concentraremos nossas forças na utilização do R, que é uma linguagem computacional que foi desenvolvida especificamente para análise estatística. Saiba um pouco mais o motivo dessa escolha: <https://blog.cursor.com/posts/2021-07-23-por-que-usar-r/>

Assim, vamos preparar o ambiente computacional para realizarmos nossas análises.

E para ficar claro:

- **R é uma linguagem computacional** (não se preocupe, não vamos programar!)
- **RStudio é um software** onde executaremos códigos R, é o que o pessoal da computação denomina de ambiente de desenvolvimento (IDE).¹

¹Integrated Development Environment - Ambiente de Desenvolvimento Integrado

5.1 Plano A: Instalação R e RStudio

No laboratório de informática da UFTM o R e a IDE RStudio estão instaladas nos computadores que utilizamos nas nossas aulas práticas, no entanto, nem sempre há tempo de desenvolver todas as atividades em sala de aula, assim, fica a sugestão para que o estudante faça a instalação do R e da IDE RStudio em seus computadores.

O RStudio é propriedade da empresa Posit (desde outubro de 2022), em seu site são dadas as instruções:

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

Ou, veja diretamente no site <https://posit.co/download/rstudio-desktop>

É importante baixar e instalar as versões do R e RStudio que sejam compatíveis com seu computador.

Essa etapa de preparação do ambiente computacional é de suma importância para o andamento da disciplina, para que as atividades sejam executadas, mas pode ser que você enfrente algum tipo dificuldade na instalação, então faça o quanto antes!

Se der tudo certo, ao clicar no ícone do RStudio, uma tela parecida como será apresentada:

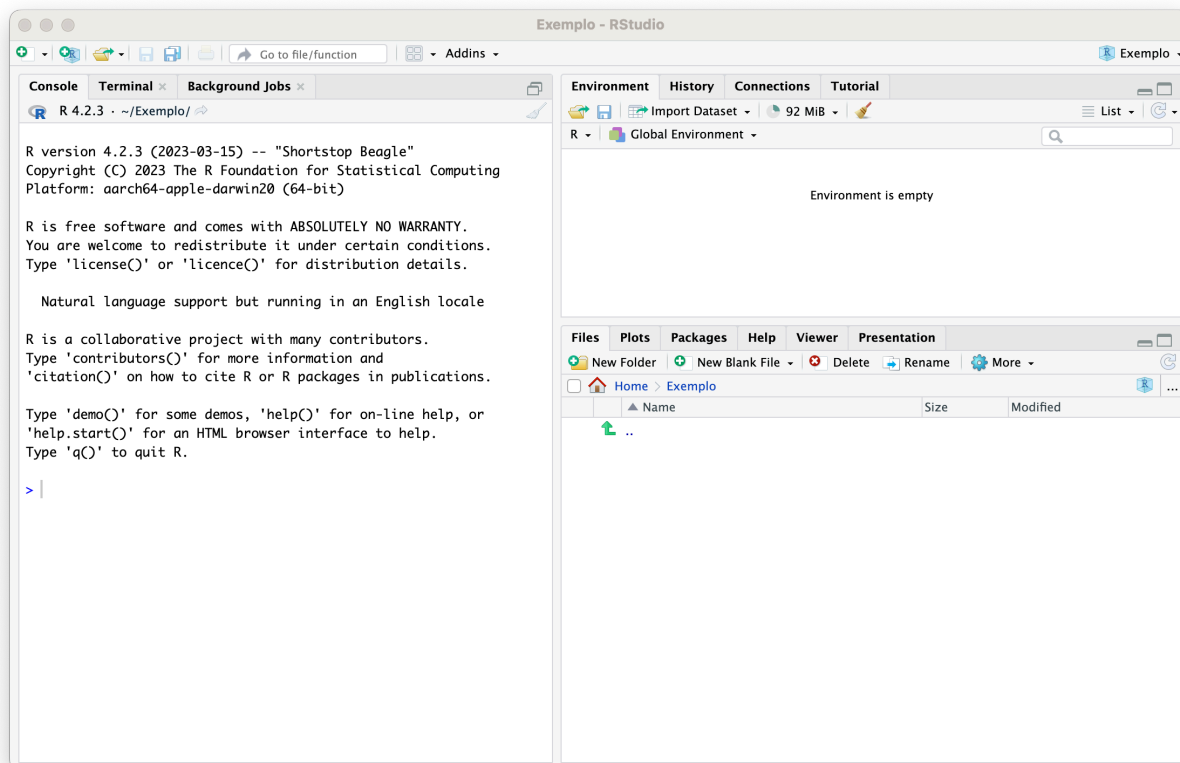


Figura 5.1: Figura: Tela inicial do RStudio

Se nada der certo, temos o plano B.

5.2 Plano B: R e RStudio online

O plano B é tão bom, mas tão bom, que poderia ser considerado plano A, no entanto, é preciso estar conectado à Internet, o RStudio será executado online na nuvem da Posit.

Se você tem uma boa conexão de Internet, fica a sugestão para usar o plano B.

1. Acesse <https://posit.cloud>
2. Faça o login (eu, por exemplo, acesso com meu usuário do Google)
3. A seguinte tela será apresentada

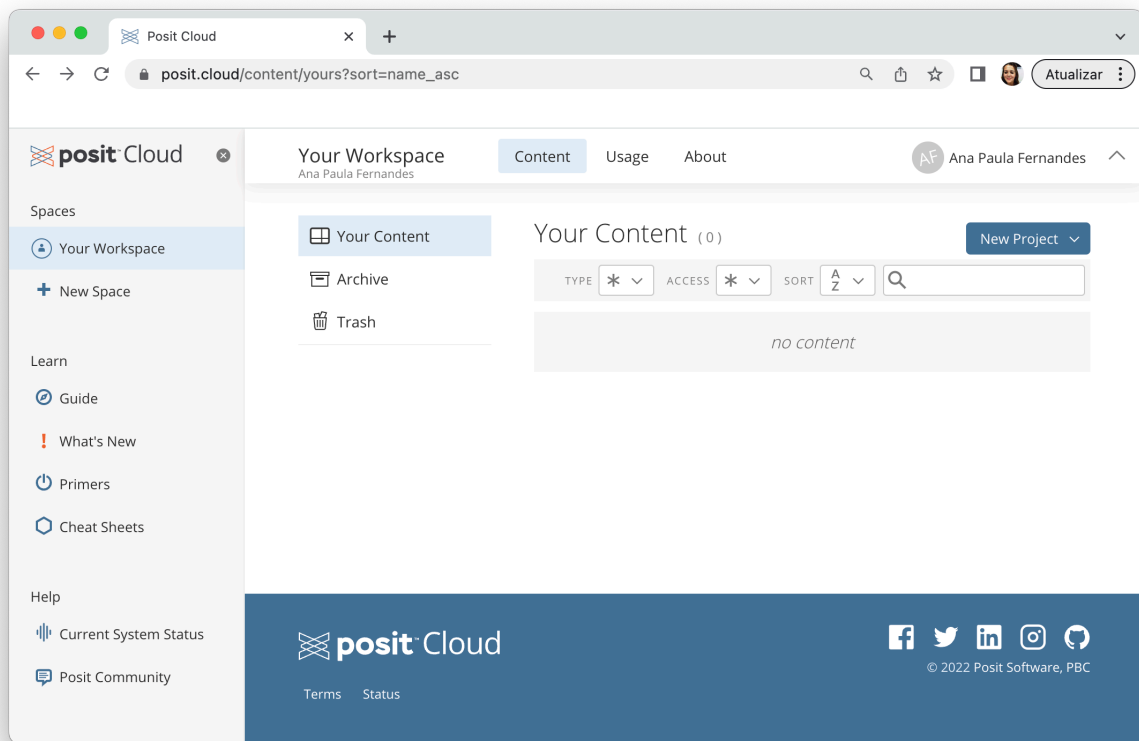


Figura 5.2: Figura: Tela inicial na nuvem da Posit

4. Crie um projeto RStudio, selecionando a opção *New Project*, e em seguida, escolhendo *New RStudio Project*.

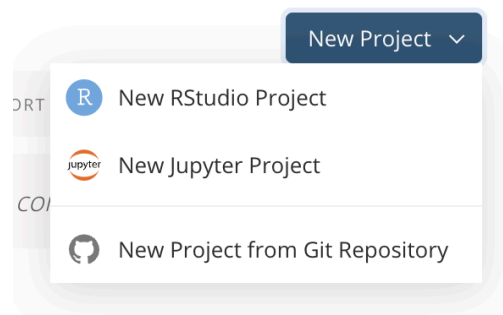


Figura 5.3: Figura: Botão de criação de um novo projeto

Se deu tudo certo, você verá a seguinte tela:

O melhor de usar a nuvem da Posit é que tudo armazenado por lá, isso quer dizer que suas análises ficam gravadas em um lugar seguro.

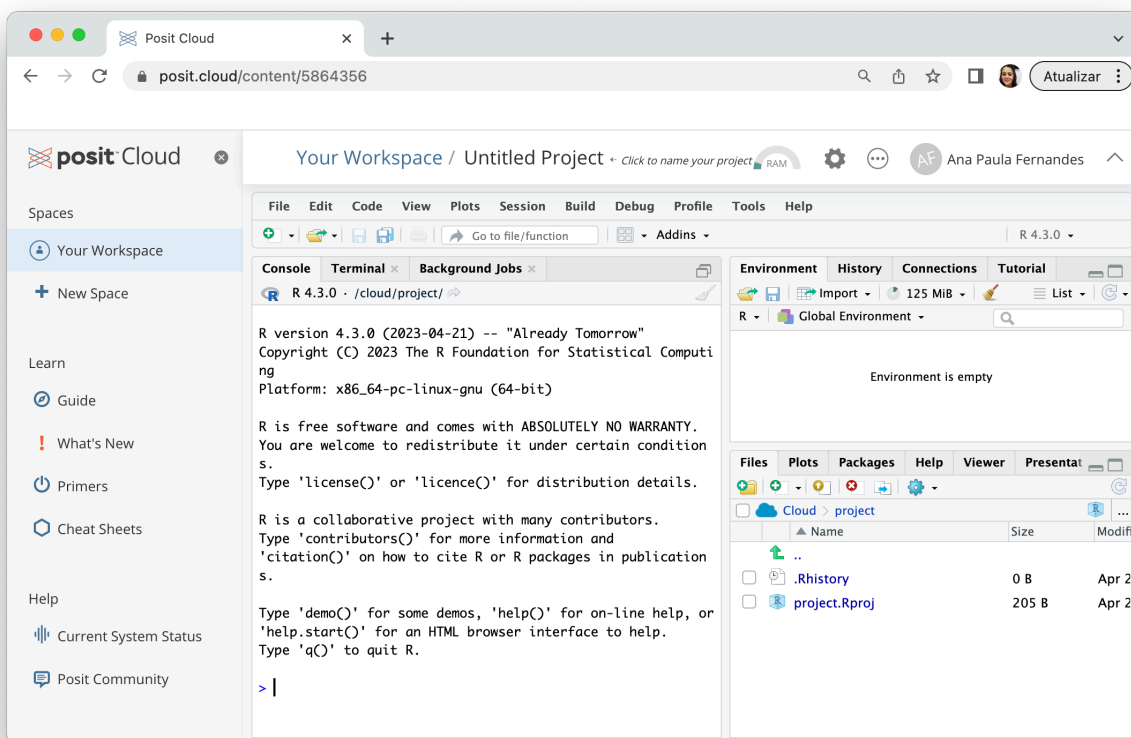


Figura 5.4: Tela inicial do RStudio online na nuvem da Posit

Capítulo 6

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, ajuda, janela de visualização e 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-o 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*.

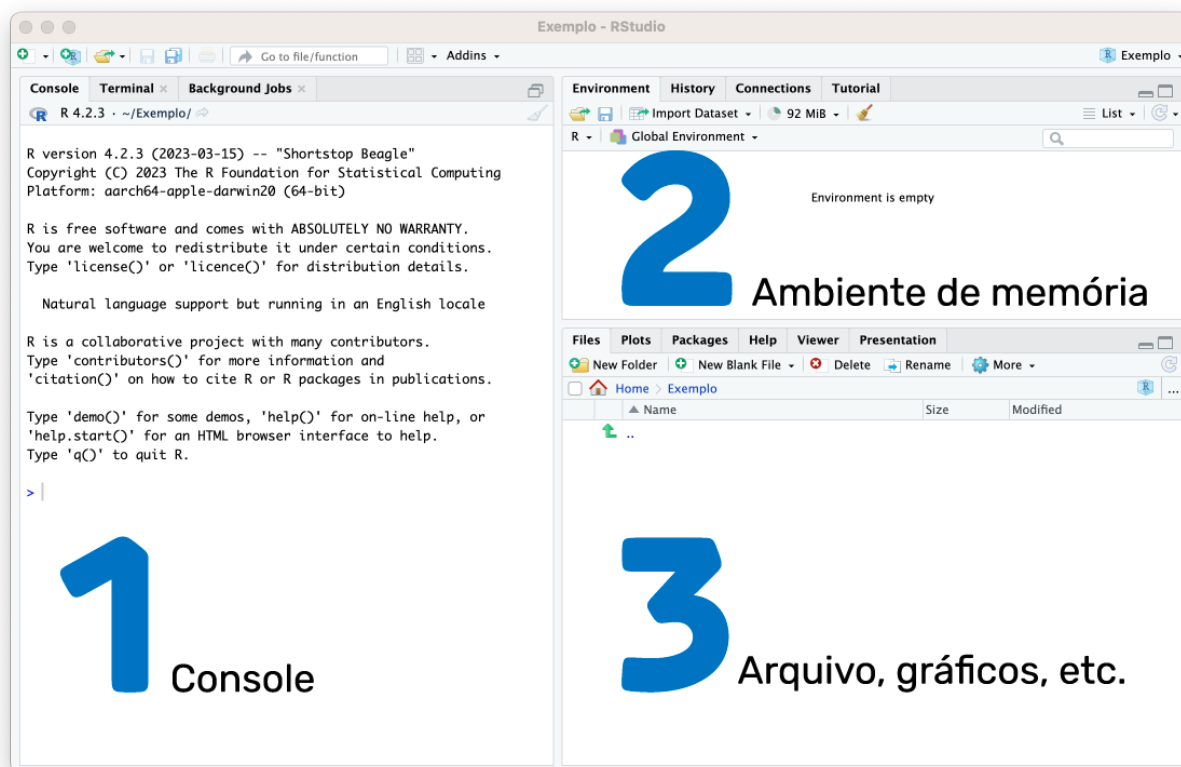


Figura 6.1: Figura: Identificação das áreas do RStudio

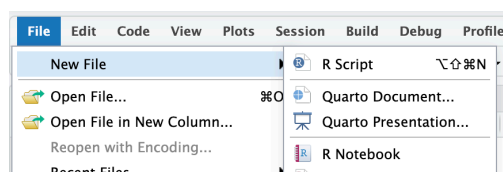


Figura 6.2: Figura: Como abrir um novo arquivo de script R

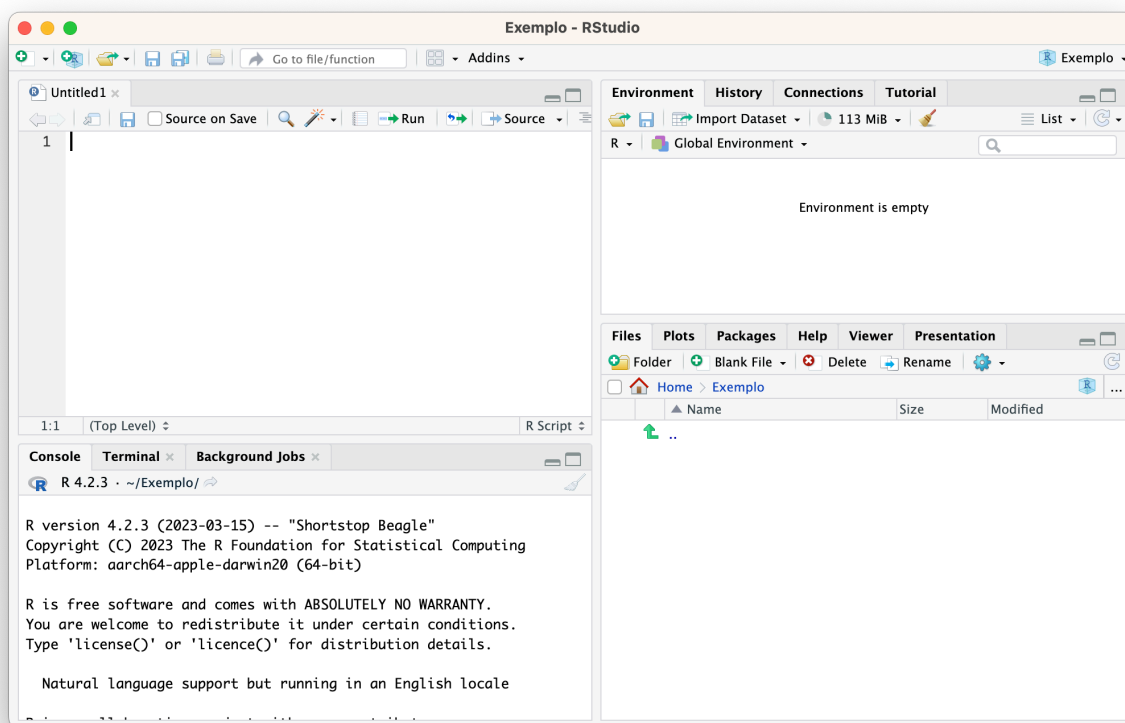


Figura 6.3: Figura: Como abrir um novo arquivo de script R

Capítulo 7

Primeiros exercícios no R

Nos capítulos 5 e 6 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 códigos 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.

7.1 Exemplo 1

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

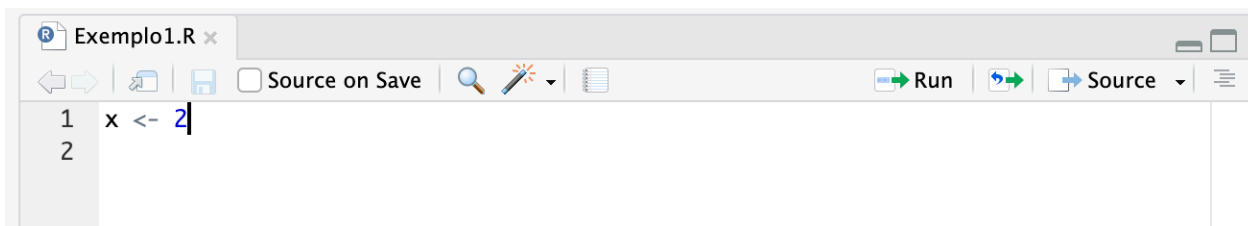


Figura 7.1: Figura: Primeiro exemplo de código R

- 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 atribuímos 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.

7.2 Exemplo 2

Execute o seguinte código 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.

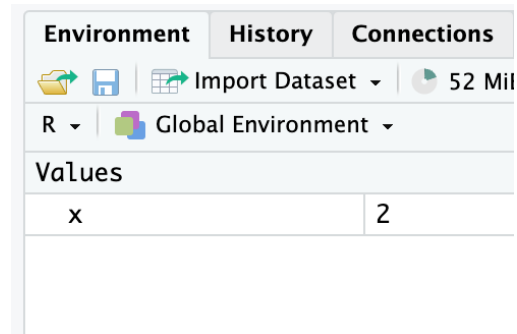


Figura 7.2: Figura: O objeto x é registrado na memória do R, armazenando o valor igual a 2

- 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.

7.3 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 (CTRL+A) e execute todo o código de uma única vez.

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



```
R 4.3.0 · /cloud/project/
> # 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
>
```

Figura 7.3: 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.

Capítulo 8

Tipos de variáveis

A natureza das variáveis (ou dados) são

- **Quantitativa** - expressa quantidade
 - Discreta - assume valores inteiros (contagem)
 - Contínua - assume qualquer valor de um dado intervalo (mensuração)
- **Qualitativa** - expressa qualidade (categorias, rótulos)
 - Nominal - categorias que não podem ser ordenadas
 - Ordinal - categorias que podem ser ordenadas, existe uma graduação entre as categorias.

Veja o vídeo do Prof. Heitor no **Canal Pesquisa**: https://youtu.be/_oc37Ea_tl8!

O procedimento estatístico que iremos realizar depende da natureza da variável que estamos analisado, por exemplo:

- Na estatística descritiva:
 - as variáveis qualitativas são representadas por sua frequência absoluta ou percentual;
 - as variáveis quantitativas são representadas por medidas resumo, como por exemplo, média e desvio padrão.
- Na estatística inferencial:
 - o teste Qui-quadrado é o teste de hipótese que tem como objetivo verificar se existe ou não associação entre as categorias de duas variáveis, como estamos falando de categorias, esse teste é aplicado a variáveis de natureza qualitativa.
 - o teste de correlação de Pearson, mede a força da correlação linear entre duas variáveis, é aplicado à variáveis quantitativas.

8.1 Atividade 3

Veja o artigo *Estado nutricional, tempo de internação e mortalidade em pacientes submetidos à cirurgia cardíaca em um hospital na cidade de Maceió* <https://www.rasbran.com.br/rasbran/article/view/1724/443> publicado na RASBRAN, Revista da Associação Brasileira de Nutrição em 2023 (<https://www.rasbran.com.br/>).

- A Tabela 1 - Características clínicas dos pacientes submetidos à cirurgia cardíaca, é uma tabela descritiva para amostra que foi analisada na pesquisa.
 - Classifique as variáveis (características) em qualitativa e quantitativa e observe atentamente como elas foram resumidas (Em porcentagens? Pela média e o desvio padrão?).

- A Tabela 2 - Associação entre estado nutricional, sexo, idade e tempo de internação hospitalar entre os pacientes submetidos à cirurgia cardíaca e Tabela 3 - Associação entre evolução clínica, sexo, idade, tempo de internação hospitalar e estado nutricional entre os pacientes submetidos à cirurgia cardíaca são tabelas que mostram o resultado de um teste de hipótese (estatística inferencial).
 - Qual teste estatístico foi aplicado? Qual o objetivo deste teste?

Capítulo 9

Estatística descritiva

Discutimos em sala de aula as medidas resumo

9.1 Medidas de tendência central (ou posição)

- Média
- Mediana e quartis
- Moda

Veja o vídeo do Canal Pesquisa <https://youtu.be/ot0aDB-grDY>

9.2 Medidas de dispersão (ou variabilidade)

- Amplitude (maior - menor)
- Variância
- Desvio padrão (DP)
- Distância interquartil (terceiro quartil - primeiro quartil)

Veja o vídeo do Canal Pesquisa <https://youtu.be/sISPcOIcwXs>

IMPORTANTE Para resumir os dados quantitativos devemos usar uma medida de tendência central e uma medida de variabilidade, assim escolhemos a forma mais ADEQUADA entre: média (desvio padrão) ou mediana (primeiro quartil; terceiro quartil)

9.3 Medida relativa de variabilidade

- Coeficiente de variação (CV) - quociente entre o desvio padrão e a média, geralmente expressamos em porcentagem (ou seja, multiplicamos essa divisão por 100%).
 - O CV é um indicador da variabilidade de um conjunto de dados.
 - O CV indica em % o quanto os dados que estamos analisando são homogêneos ou heterogêneos.
 - Um CV é considerado baixo (indicando um conjunto de dados razoavelmente homogêneo) quando for menor ou igual a 25%. Entretanto, esse padrão varia de acordo com a aplicação.
- * Por exemplo, em medidas vitais (batimento cardíaco, temperatura corporal, etc) espera-se um CV muito menor do que 25% para que os dados sejam considerados homogêneos. Fonte: <http://www.leg.ufpr.br/~silvia/CE001/node24.html>

- Pode ser difícil classificar um coeficiente de variação como baixo, médio, alto ou muito alto, no entanto, o CV é útil na comparação de duas variáveis de natureza diferentes.

9.4 Funções do R

Supondo que o objeto `x <- c(valor 1, valor 2, ..., valor n)` está na memória do R.

Medida resumo	Função básica do R
média	<code>mean(x)</code>
mediana	<code>median(x)</code>
primeiro quartil	<code>quantile(x,0.25)</code>
terceiro quartil	<code>quantile(x,0.75)</code>
moda	<code>table(x)</code>
menor valor / mínimo	<code>min(x)</code>
maior valor / máximo	<code>max(x)</code>
resumo das medidas	<code>summary(x)</code>
amplitude	<code>range(x)</code>
variância	<code>var(x)</code>
desvio padrão	<code>sd(x)</code>
amplitude interquartil	<code>IQR(x)</code>
coeficiente de variação	<code>sd(x)/mean(x)</code>

- *use `sort(table(x))`, use a função `sort()` para ordenar as ocorrências da menor para a maior, a maior ocorrência é a moda!
- use a função `summary(x)` para obter, menor valor, média, mediana, primeiro e terceiro quartil e maior valor.

Viu como calcular é fácil? Então, tenha em mente que o mais **importante é interpretar** essas medidas, ou seja, descrever o que essas medidas revelam sobre a amostra em estudo.

9.5 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 variação:
- Escreva sobre o conjunto média e desvio padrão:
- Escreva sobre conjunto mediana e quartis:
- Escreva sobre o coeficiente de variação:
- 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.

Capítulo 10

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 análises estatísticas, aqui estão algumas delas:

- DataSus: <https://datasus.saude.gov.br/transferecia-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>

10.1 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 ambiente de memória, localize **Import Dataset**, ao clicar nessa opção você terá o seguinte:

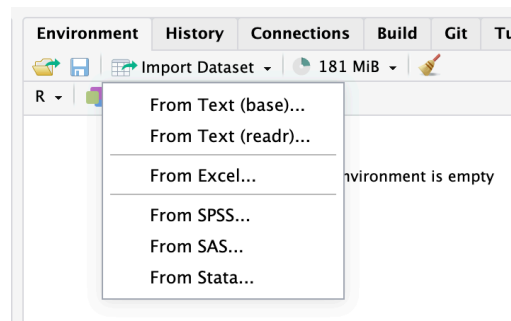


Figura 10.1: Figura: Importando banco de dados

- 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** indique onde (no seu computador) está localizado o arquivo a ser importado. A seguinte tela será apresentada:

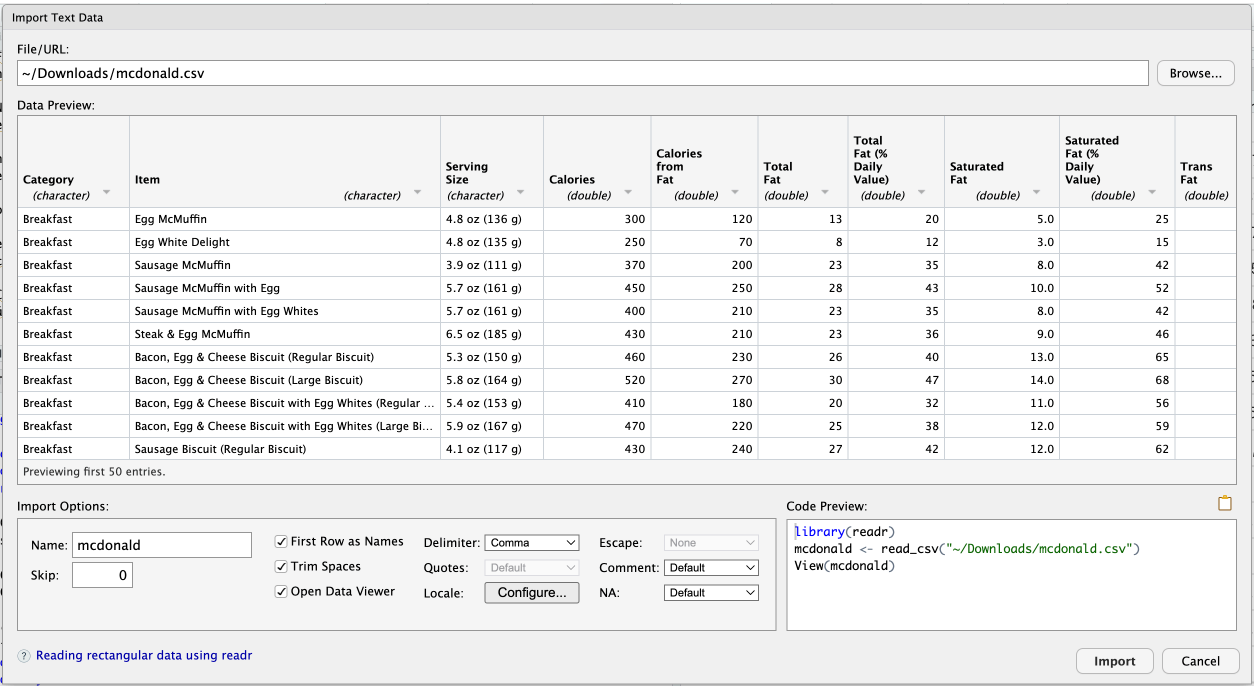


Figura 10.2: Figura: Prévia dos dados

- 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.

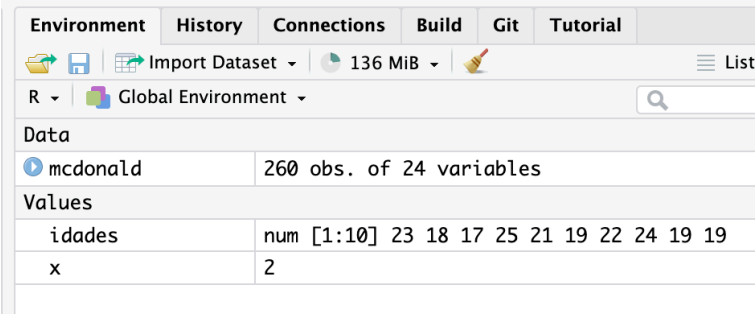


Figura 10.3: Figura: Import dataset

- 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 aos nomes e tipos computacionais das variáveis, e ao clicar sobre o nome do objeto, o banco será aberto!

10.2 Importando um banco xls

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

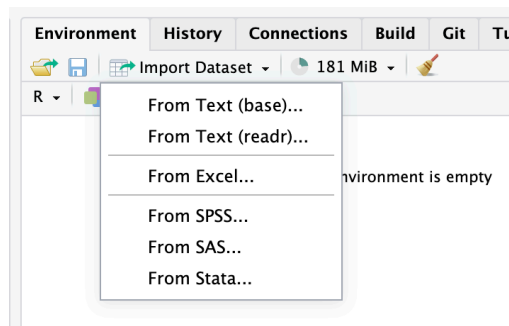


Figura 10.4: Figura: Importando banco de dados

- 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:

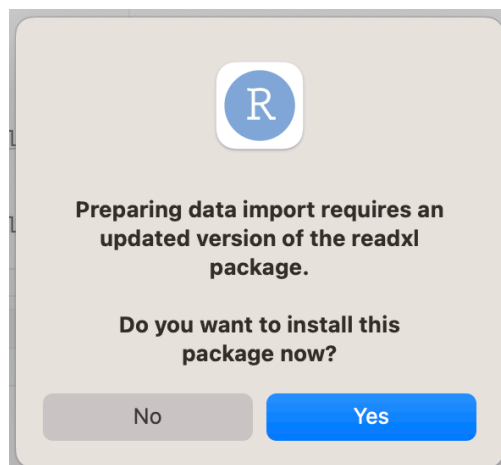


Figura 10.5: Figura: Aviso para instalação de pacote

10.3 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)
```

10.4 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)
```

10.5 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.

Capítulo 11

Instalando pacotes

Quando instalamos nosso ambiente computacional R e RStudio, instalamos uma versão básica, onde apenas os recursos básicos do R estão disponíveis, 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 biblioteca 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.danielsjoberg.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 parênteses (%) é mostrado a porcentagem de cada categoria.
- **variáveis quantitativas:** Median é a mediana e entre parênteses (IQR - de InterQuantile Range) estão o primeiro e terceiro quartil respectivamente.

11.1 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})")
)
```

11.2 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)
```

11.3 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 referenciá-la entre aspas: `Serving Size`
dadosSelecionados2 <- select(mcdonald, -Item, -`Serving Size`)

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

11.4 Exemplo 4

Como selecionar 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 referenciá-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/>

11.5 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!

Capítulo 12

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)

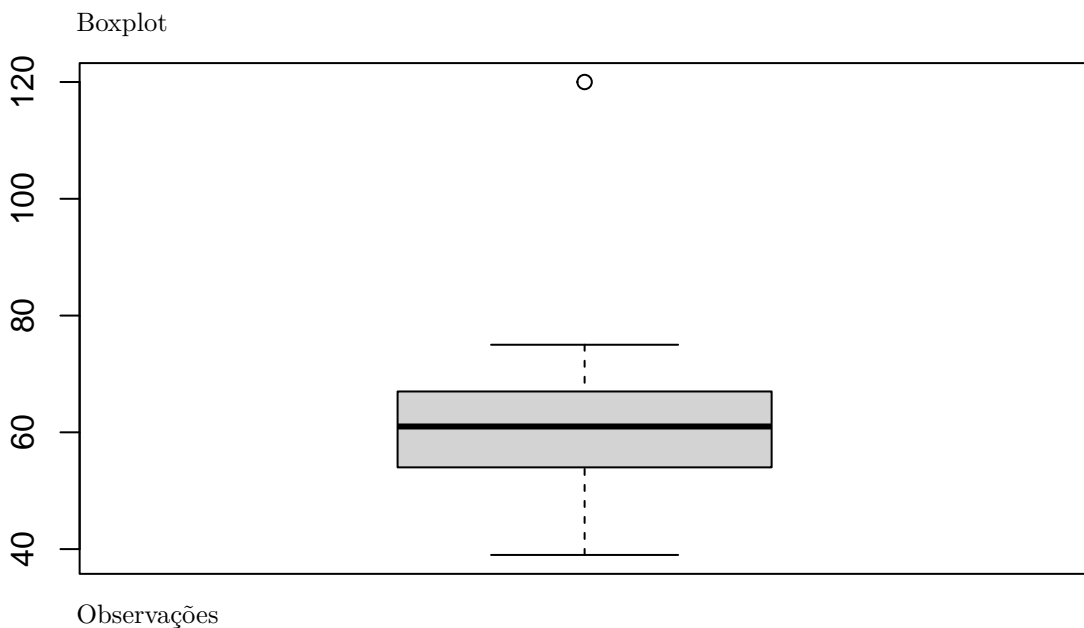
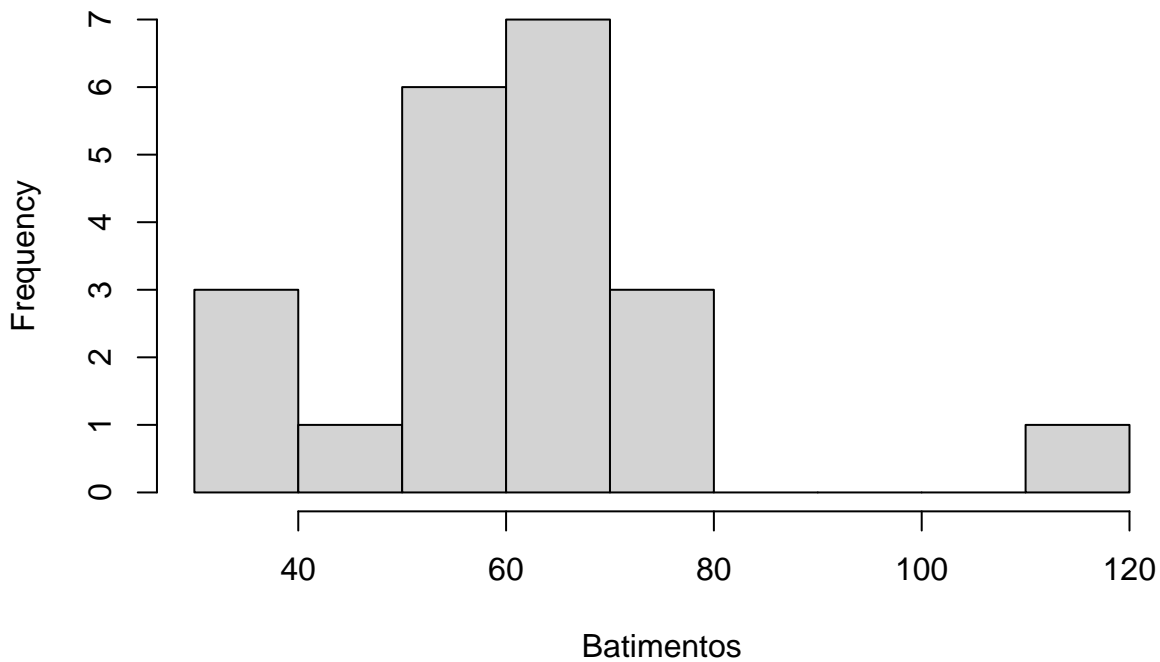
# Para fazer o Histograma de Batimentos
hist(Batimentos)

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

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

Histograma

Histogram of Batimentos



- 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

12.1 Histograma

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

O histograma pode nos dar uma noção do tipo de **distribuição de probabilidade** 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 **amplitude 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 frequências (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
<i>n</i>	21

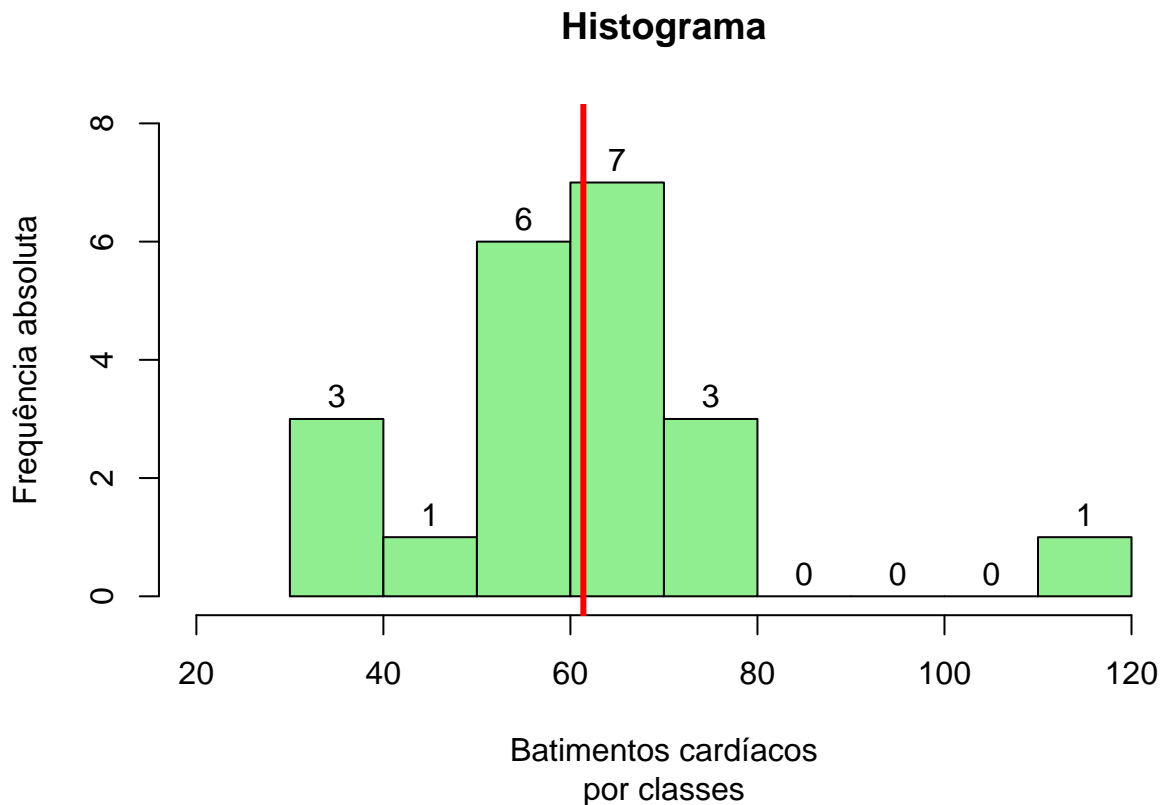
- Por meio do histograma ou da tabela podemos concluir que a classe modal (moda) é a classe de 60 - 70 batimentos;
- A frequência foi apresentada em termos absolutos mais pode ser transformada em frequência percentual.
- Quando estamos aprendendo a fazer um histograma manualmente, primeiro construímos essa tabela de frequência, 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$mids, hisBat$count, labels=hisBat$count, adj = c(0.5,-0.5))

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

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



12.2 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:

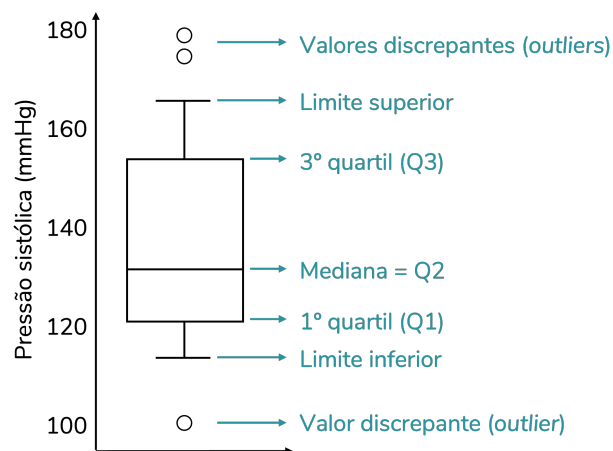


Figura 12.1: Figura: “Anatomia de um boxplot”

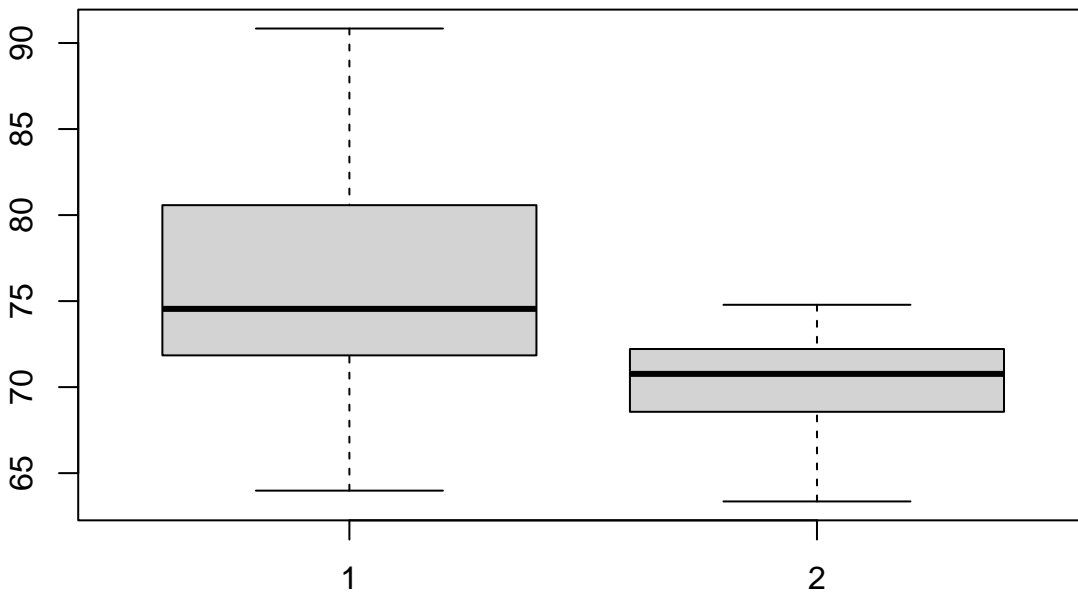
Essa figura foi retirada do site da Prof. Fernanda <https://fernandafperes.com.br/blog/interpretacao-boxplot/> (uma excelente referência para estudar estatística!)

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)
```

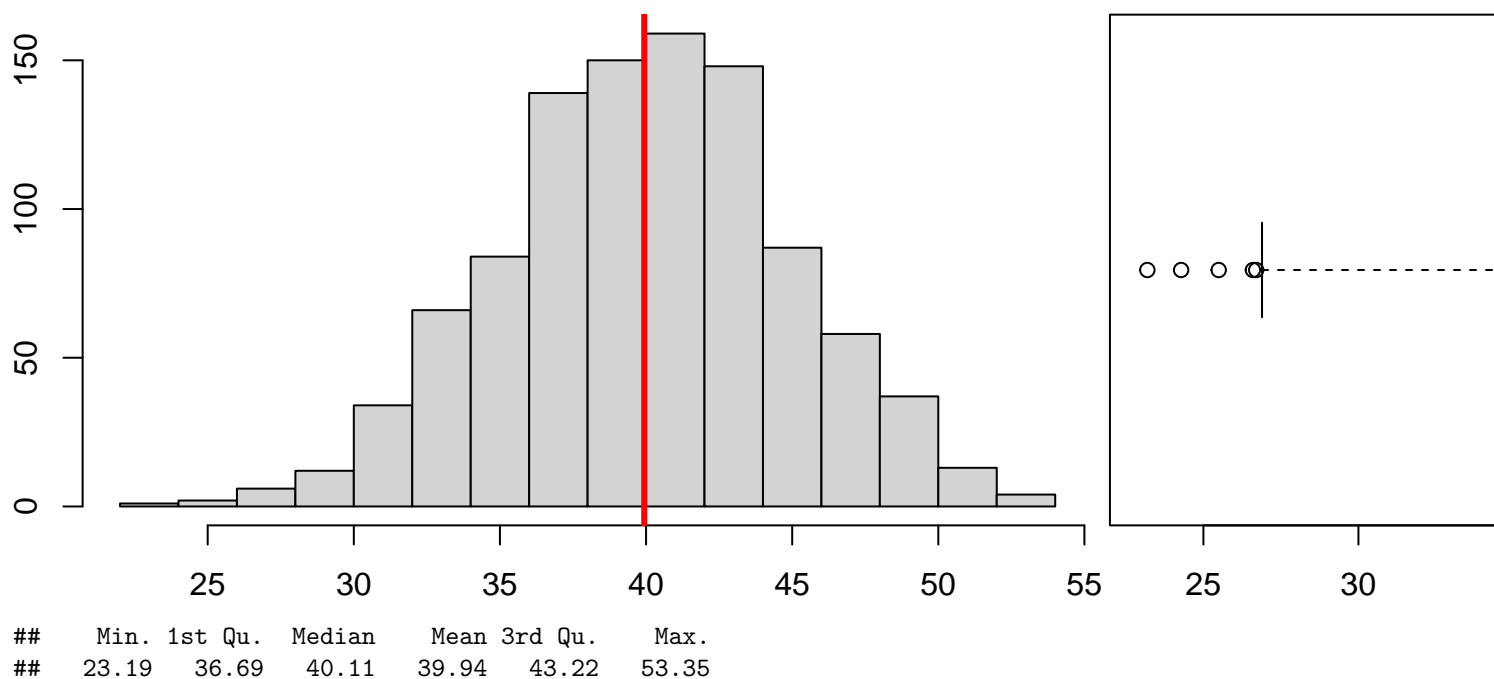
É uma forma de comparar dois grupos em relação a uma medida, por exemplo os batimentos cardíacos de grupo de homens e de mulheres

```
# 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)
```

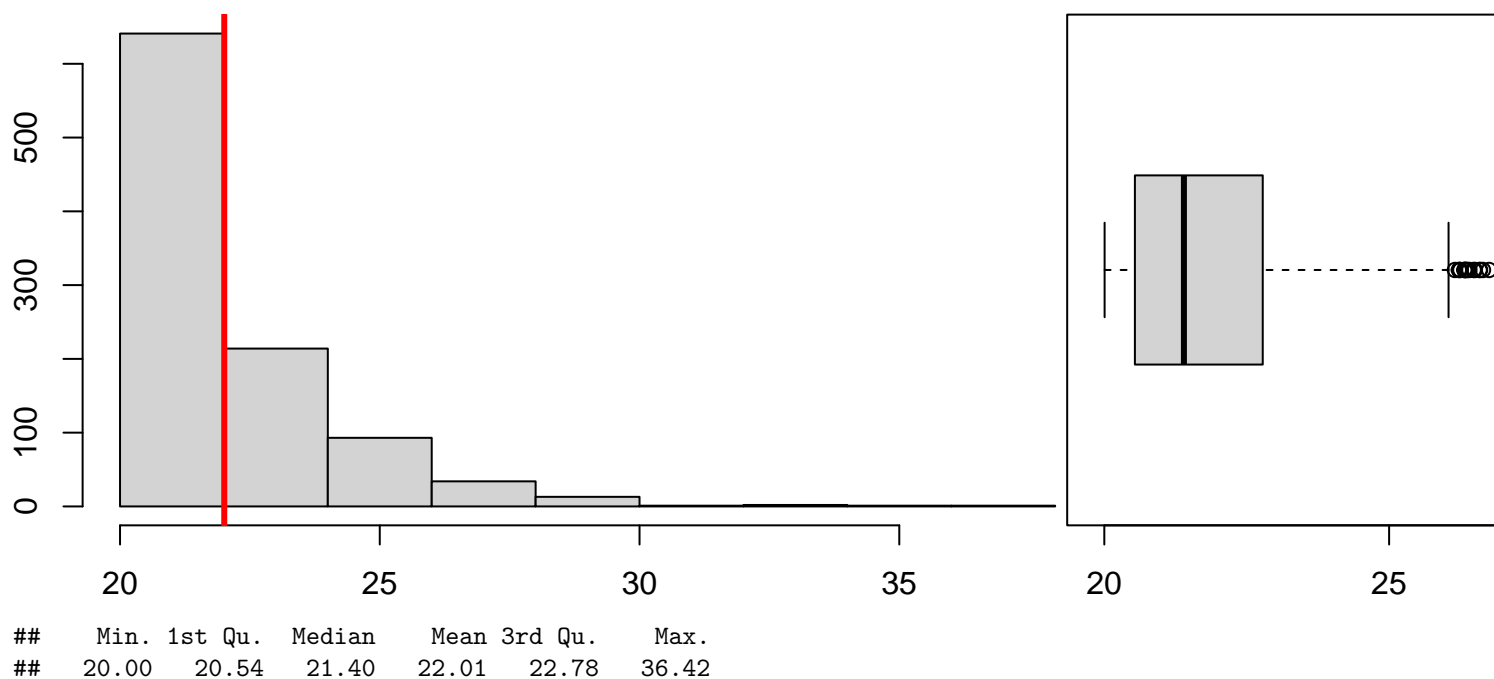


O boxplot também pode nos informar se uma distribuição de probabilidade é simétrica ou não. Analise os gráficos abaixo, veja a conexão entre histograma e boxplot.

Distribuição SIMÉTRICA



Distribuição ASSIMÉTRICA



12.3 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.

Capítulo 13

Distribuição de Probabilidade!

Uma distribuição de probabilidade é um modelo matemático que associa um dado valor de uma variável a sua probabilidade de ocorrer.

A **distribuição Normal** (ou Gaussiana) é uma das distribuições de probabilidade mais utilizadas na estatística.

Outras distribuições de probabilidade: Binomial, Poisson, Exponencial, Uniforme, Qui-quadrado, T-Student, Gama, Weibull, Lognormal. . .

13.1 Distribuição Normal

A distribuição Normal pode ser usada para modelar muitos conjuntos de medidas na natureza, na indústria e nos negócios. Por exemplo, a pressão sanguínea sistólica dos humanos, a vida útil de televisões de plasma e até mesmo custos domésticos (LARSON, 2015).

A distribuição Normal é definida por dois parâmetros: média e desvio padrão.

Características da Distribuição Normal

- Tem forma de um sino
- É simétrica em relação a média
- O valor da média, mediana e moda são iguais

Para toda distribuição de probabilidade

- A área total abaixo da curva é igual a 1
- A área representa a probabilidade

Teoricamente o comportamento da Distribuição Normal é dado por:

Onde:

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

Regra empírica 68% - 95% - 99,7%

Se os dados seguem distribuição Normal podemos afirmar que

- 68% dos dados concentram-se no intervalo: $[\text{média} - 1\text{dp} ; \text{média} + 1\text{dp}]$
- 95% dos dados concentram-se no intervalo: $[\text{média} - 2\text{dp} ; \text{média} + 2\text{dp}]$
- 99,7% dos dados concentram-se no intervalo: $[\text{média} - 3\text{dp} ; \text{média} + 3\text{dp}]$

Isso significa dizer que eventos que estão fora do intervalo $[\text{média} - 3\text{dp} ; \text{média} + 3\text{dp}]$ são eventos raros!

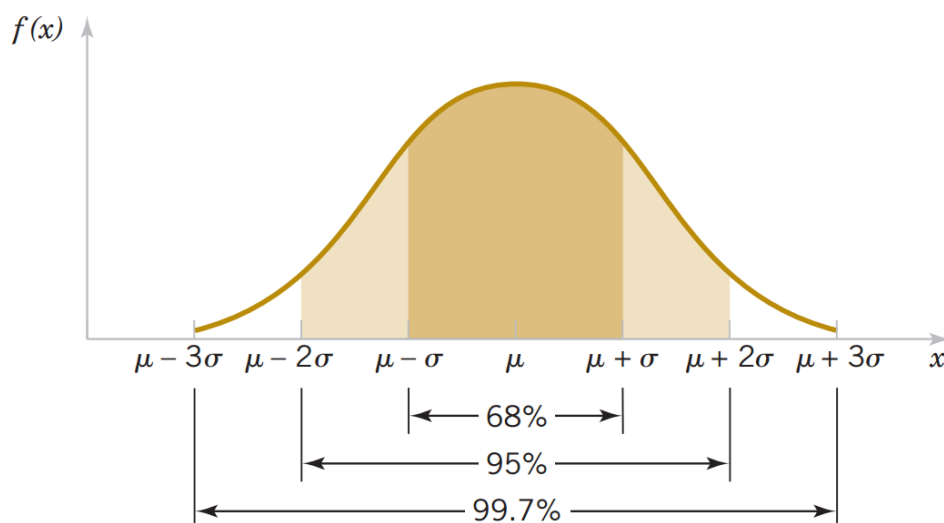


Figura 13.1: Figura: Distribuição Normal teórica (Fonte: <https://www.inf.ufsc.br/~andre.zibetti/probabilidade/figures/normal.PNG>)

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

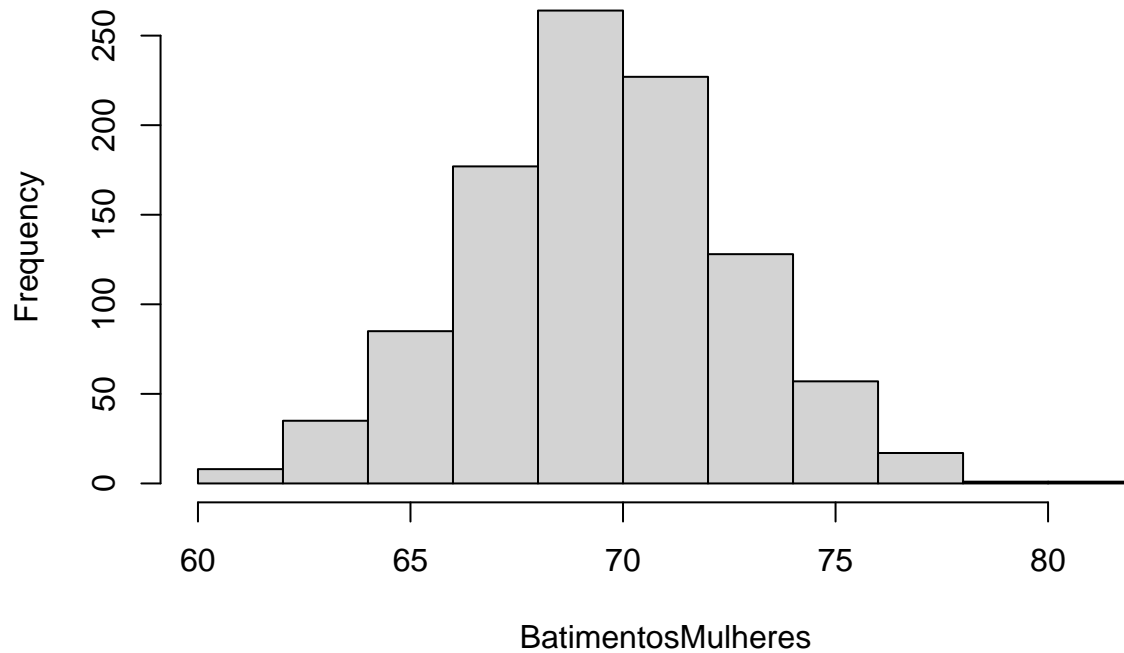
```
# semente de geração de números aleatórios
set.seed(1)

# 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)

# Arredondamento com nenhuma casa depois da vírgula
BatimentosMulheres <- round(BatimentosMulheres, 0)

# histograma
hist(BatimentosMulheres)
```


Histogram of BatimentosMulheres



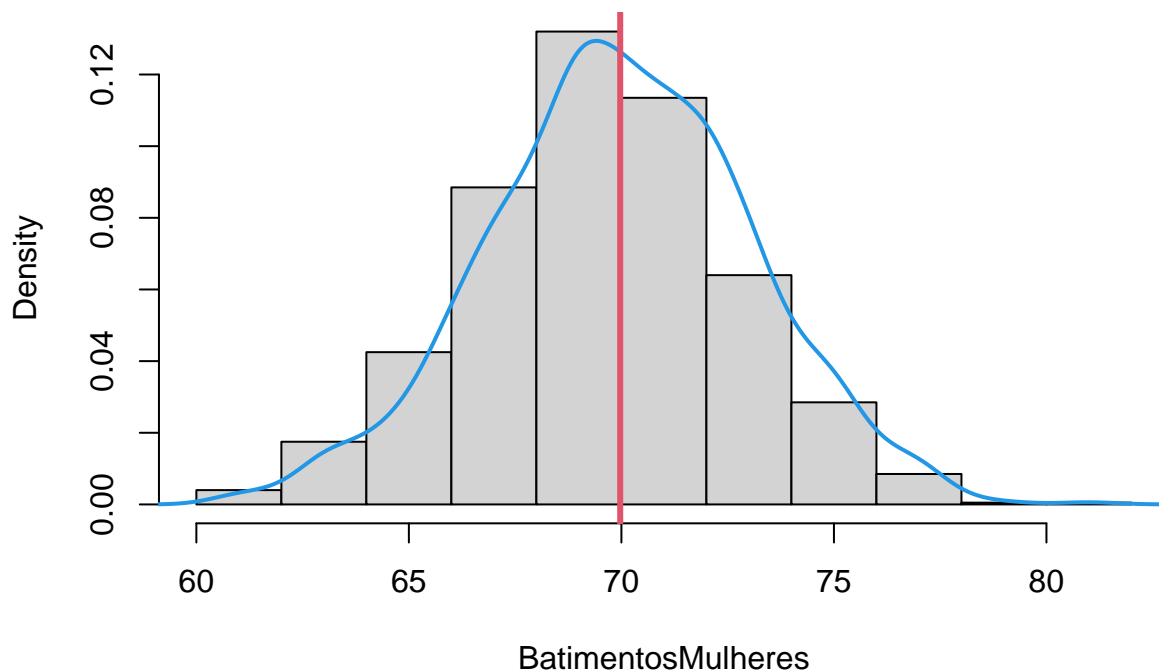
```
# Classes e frequências do histograma
hist(BatimentosMulheres)$breaks
```

```
## [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)

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

Histogram of BatimentosMulheres



```
# 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()** é usada 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 70bpm e desvio padrão igual a 3bpm.

Calcule as probabilidades:

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

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

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

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

```
dnorm(70, 70, 3)
```

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

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

```
# 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 73bpm $P(64 < x < 76)$:

```
# 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 79bpm $P(61 < x < 79)$:

```
# 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 90bpm $P(x > 90)$:

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

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

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

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

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

13.2 Diagnóstico de Normalidade (QQ)

Uma forma visual para **verificarmos a normalidade** dos dados é a através do **gráfico QQ**.

A ideia desse gráfico é comparar a distribuição da nossa amostra com uma **distribuição Normal padrão**.

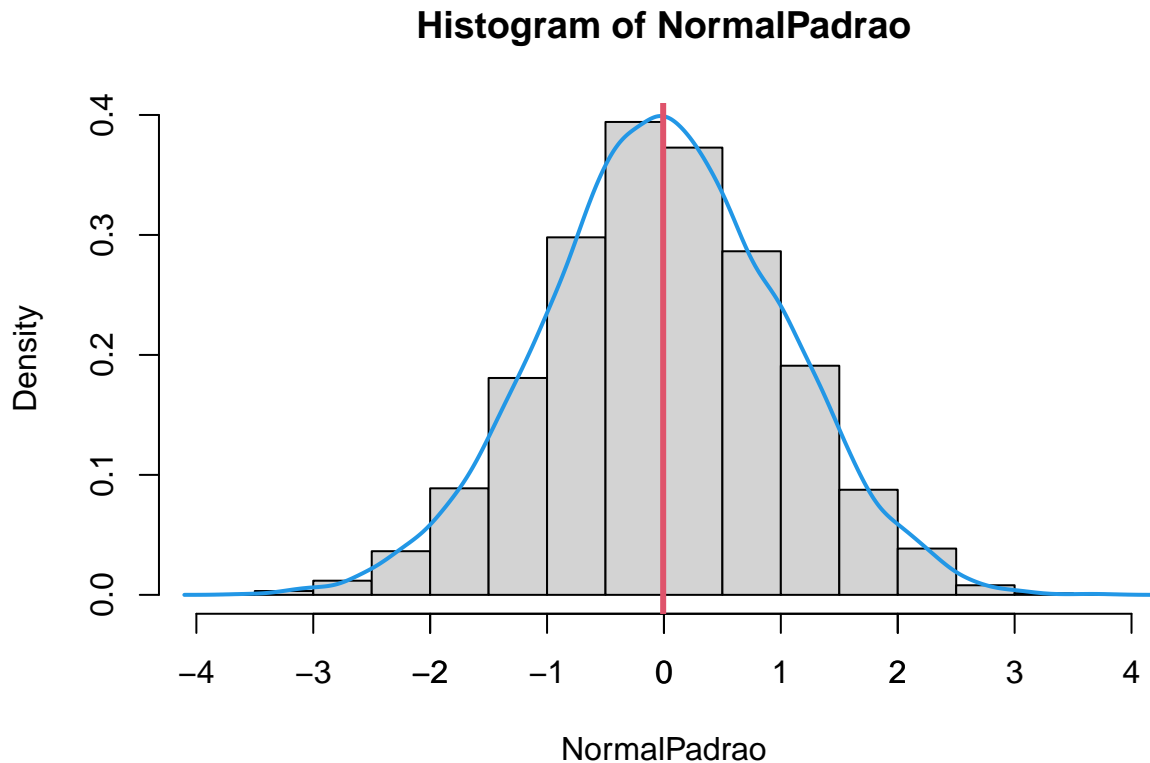
A característica da distribuição Normal Padrão é que ela tem média igual a zero e desvio padrão igual a 1.

- Qualquer distribuição Normal pode ser convertida em uma distribuição Normal Padrão (por meio do cálculo escore z), é mais o menos o que o gráfico QQ faz.
- O z escore é dado por: $z = (\text{valor} - \text{media})/\text{dp}$

Veja um exemplo no R para uma distribuição normal padrão

```
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)
```



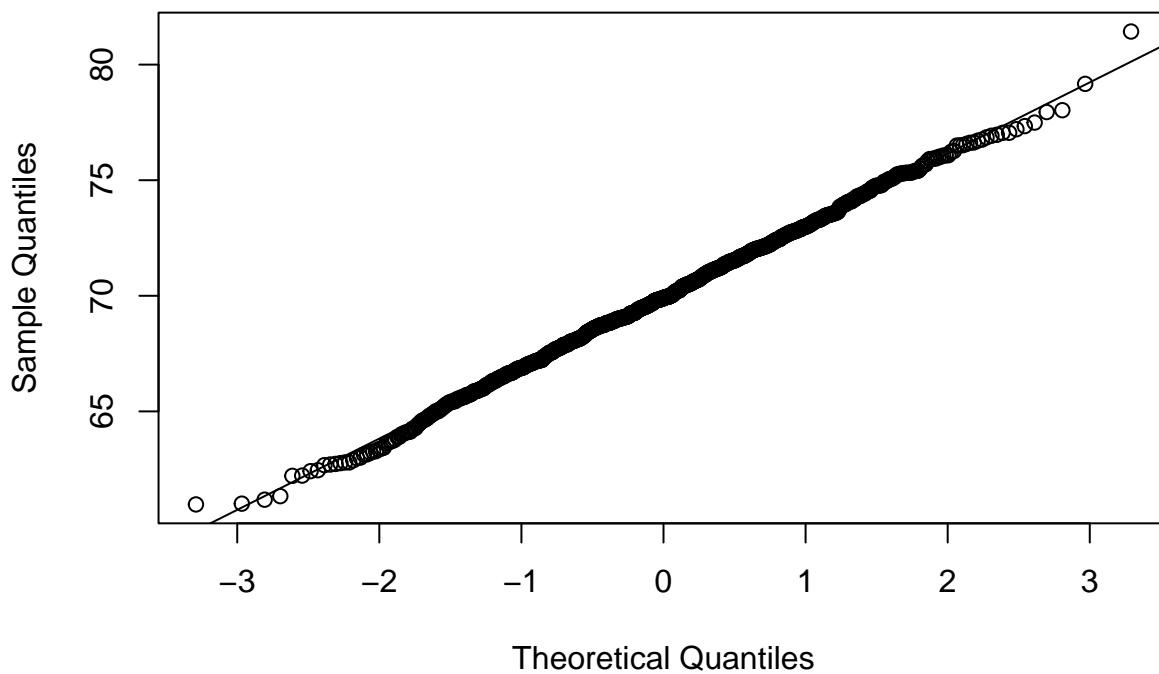
- o intervalo -1 a 1 em torno da média: [média - 1dp ; média + 1dp]
- o intervalo -2 a 2 em torno da média: [média - 2dp ; média + 2dp]
- o intervalo -3 a 3 em torno da média: [média - 3dp ; média + 3dp]

Um gráfico QQ compara os quantis de uma amostra com os quantis de uma distribuição teórica normal. Se os pontos no gráfico seguirem uma linha reta, isso sugere que os dados são normalmente distribuídos.

Distribuição normal é usada para dados contínuos!

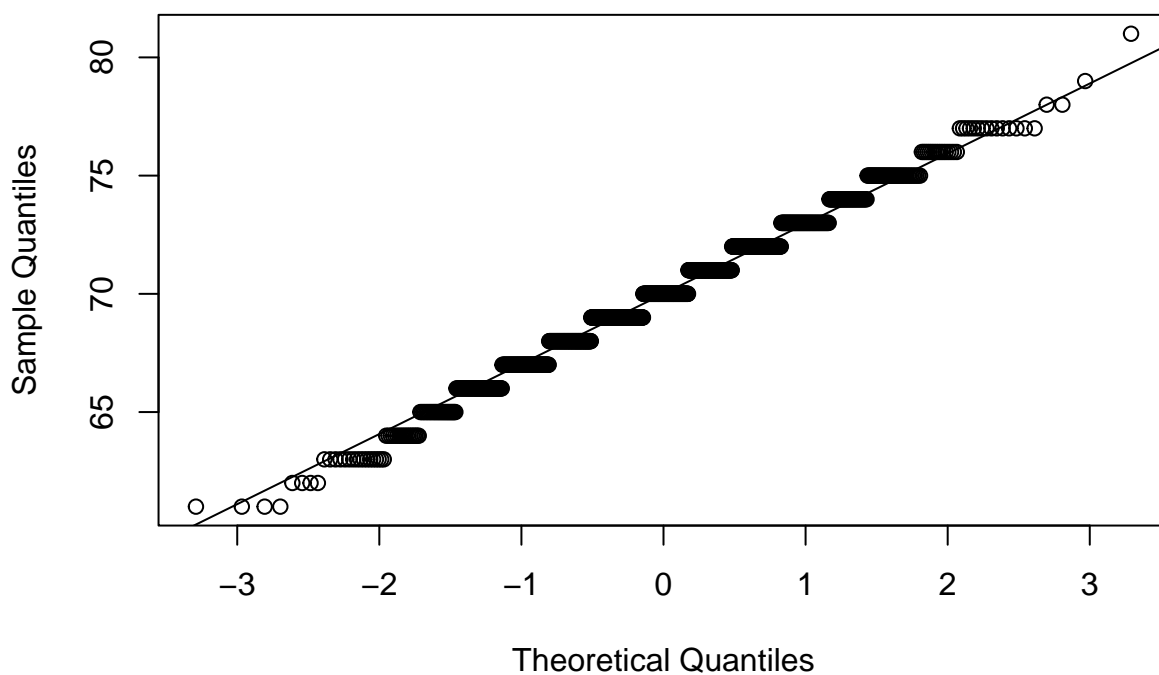
```
# Gráfico QQ
set.seed(1)
BatimentosMulheres <- rnorm(1000, 70, 3)
qqnorm(BatimentosMulheres)
qqline(BatimentosMulheres)
```

Normal Q-Q Plot



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

Normal Q-Q Plot



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

13.3 Atividade 7

Para o banco de dados escolhido para as atividade 5 e 6, faça gráficos QQ para as variáveis quantitativas e verifique visualmente se elas seguem distribuição Normal.